COMPUTATIONAL NEUROSCIENCE FOR PERCEPTUAL QUALITY ASSESSMENT

EDITED BY: Guangtao Zhai, Vinit Jakhetiya, Ke Gu, Lu Zhang and Xiongkuo Min

PUBLISHED IN: Frontiers in Neuroscience, Frontiers in Psychology and Frontiers in Computational Neuroscience

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-950-8
DOI 10.3389/978-2-88974-950-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL NEUROSCIENCE FOR PERCEPTUAL QUALITY ASSESSMENT

Topic Editors:
Guangtao Zhai, Shanghai Jiao Tong University, China
Vinit Jakhetiya, Indian Institute of Technology Jammu, India
Ke Gu, Beijing University of Technology, China
Lu Zhang, Institut national des sciences appliquées de Rennes, France
Xiongkuo Min, Shanghai Jiao Tong University, China

Citation: Zhai, G., Jakhetiya, V., Gu, K., Zhang, L., Min, X., eds. (2022). Computational Neuroscience for Perceptual Quality Assessment. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-950-8

Table of Contents

EDITORIAL
published: 28 March 2022
doi: 10.3389/fnins.2022.876969

Editorial: Computational Neuroscience for Perceptual Quality Assessment

Xiongkuo Min[1]*, *Ke Gu*[2], *Lu Zhang*[3], *Vinit Jakhetiya*[4] *and Guangtao Zhai*[1]

[1] *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China,* [2] *Faculty of Information Technology, Beijing University of Technology, Beijing, China,* [3] *Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France,* [4] *Department of Computer Science Engineering, Indian Institute of Technology, Jammu, India*

Keywords: computational neuroscience, perceptual quality assessment, perception, multimedia, image processing

OPEN ACCESS

Edited and reviewed by:
Rufin VanRullen,
Centre National de la Recherche Scientifique (CNRS), France

****Correspondence:***
Xiongkuo Min
minxiongkuo@sjtu.edu.cn

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *16 February 2022*
Accepted: *01 March 2022*
Published: *28 March 2022*

Citation:
Min X, Gu K, Zhang L, Jakhetiya V and Zhai G (2022) Editorial: Computational Neuroscience for Perceptual Quality Assessment. Front. Neurosci. 16:876969. doi: 10.3389/fnins.2022.876969

Editorial on the Research Topic

Computational Neuroscience for Perceptual Quality Assessment

Quality assessment aims to measure the degree of delight or annoyance of the users when experiencing an application or service. With the quick improvement of content acquisition, processing, transmission, and display techniques, the end-users are expecting and demanding continuously improved quality of experience (QoE) from the service providers. To guarantee a good QoE to end-users, perceptual quality assessment is introduced and widely studied in recent years (Brunnström et al., 2013; Zhai and Min, 2020; Min et al., 2022). Since the ultimate receiver of the processed signal is usually human, it is reasonable and beneficial to take human perception properties into consideration. Though we still have limited knowledge of the intrinsic neuroscience working mechanism of human perception, it is worthwhile to study and take inspiration from neuroscience and utilize these properties for computational modeling of perceptual quality.

Many of the current quality assessment models have already attempted to include human perception properties at some level, however, the majority of these models only take simplified concepts of human perception, and use "black box" machine learning techniques to model the QoE. The rapid development of neuroscience and computer science have provided opportunities for deeper explorations of the intrinsic neuroscience working mechanism of quality perception, and to utilize computational neuroscience theories and models for more efficient and explainable quality assessment. Specifically, on one hand the underlying biological bases of human perception especially those related to quality perception can be further explored on the basis of the recent advancement of neurobiology. While on the other hand, it is worthwhile to seek better ways to apply the relevant neuroscience working mechanisms for quality assessment and to build more accurate brain-inspired computational quality assessment models.

This Research Topic is a collection of articles concerning computational neuroscience studies for perceptual quality assessment and the potential applications in artificial systems. The final list of accepted articles can be categorized into four groups: 1. Neuroscience studies of human perception, especially those related to quality perception; 2. Neuroscience inspired perceptual quality modeling; 3. Perceptual quality assessment for emerging and advanced multimedia technologies; 4. Applications of perceptual quality modeling. The below is an overview and discussion of the accepted articles.

NEUROSCIENCE STUDIES OF HUMAN PERCEPTION, ESPECIALLY THOSE RELATED TO QUALITY PERCEPTION

In recent years, a large amount of perceptual quality assessment studies has taken human perception properties into consideration, since human is usually the ultimate judger of signal quality. To study the intrinsic neuroscience working mechanism of human perception, subjective neuroscience and perceptual studies are usually necessary.

The influence of audio on perceptual QoE has been studied and verified by some previous studies (You et al., 2010; Akhtar and Falk, 2017; Min et al., 2017, 2020a,b). In this Research Topic, Sun and Hines give an overview for the audiology and cognitive science researches which study how cognitive processes influence the quality of listening experience. Moreover, they also propose to introduce these mechanisms from audiology and cognitive science into the current QoE framework, through which we can better incorporate cognitive load in speech listening. Pieper et al. use electroencephalogram and some other questionnaire-based subjective measures to study if noise-canceling technologies can reduce the influence of external distractions and free up mental resources. Results partially verify that an assumed lower mental load is observed in no noise and noise-canceling environment compared to that of in the noise environment. Han et al. study the influence of the refresh rate of a display on the motion perception response. Moreover, they introduce an objective visual electrophysiological assessment model to better select the display parameters.

NEUROSCIENCE INSPIRED PERCEPTUAL QUALITY MODELING

Full understanding of the intrinsic neuroscience working mechanism of human perception is difficult in the current stage, however it is worthwhile to study and take inspiration from neuroscience and utilize these properties for computational modeling of perceptual quality.

Over the last two decades, many perceptual quality assessment models have been proposed (Wang et al., 2004; Brunnström et al., 2013; Min et al., 2018a,b, 2022; Zhai and Min, 2020), and many of them have taken inspirations from neuroscience. Song et al. introduce a blind quality assessment model for authentically distorted images by considering both distortion degree and intelligibility. Specifically, they analyze the relation between intelligibility and image quality, and then incorporate such intelligibility into a highly generalizable image quality prediction model. Feng et al. introduce an end-to-end cross-domain feature similarity guided deep neural network for perceptual quality assessment. This model is built based on the observation that features for the object recognition task and features for the quality prediction task are highly correlated in terms of characteristics of the human visual system. Experimental results have verified the effectiveness of the proposed model.

PERCEPTUAL QUALITY ASSESSMENT FOR EMERGING AND ADVANCED MULTIMEDIA TECHNOLOGIES

Recently, a growing number of emerging and advanced multimedia technologies or systems have invaded into our daily lives, for example light field, virtual reality, etc. Such emerging multimedia applications also call for new quality perception models, since traditional quality perception models are not good at such contents.

In this Research Topic, Meng et al. propose a light field image quality assessment model by predicting the global angular-spatial distortion of macro-pixels as well as the local angular-spatial quality of the focus stack. Wang et al. present a quality metric for depth-image-based rendering images by jointly measuring the synthesized image's colorfulness, texture structure, and depth structure. Hu et al. first introduce a method to simulate the wrap-around artifact on the artifact-free MRI image to increase the quantity of MRI data, and then propose an image restoration method to reduce the wrap-around artifact.

APPLICATIONS OF PERCEPTUAL QUALITY MODELING

The research of perceptual quality modeling applications has also aroused increasing attention in recent years, since perceptual quality modeling can play an important role in the quality control and optimization of multimedia communication systems. In this Research Topic, Lei et al. first introduce a new quality assessment database for swimming pool images, and then propose an objective swimming pool image quality measure by detecting the main target and integrating multiple quality-aware features. Yu et al. first construct a new image database by collecting 1,000 pictures from the official social network accounts of nine well-known universities, as well as the corresponding number of page views.

We hope that readers find this Research Topic useful, timely and informative, in addressing the important topics in Computational Neuroscience for Perceptual Quality Assessment.

AUTHOR CONTRIBUTIONS

All authors wrote, and equally contributed to the article, and approved the submitted version.

REFERENCES

Akhtar, Z., and Falk, T. H. (2017). Audio-visual multimedia quality assessment: a comprehensive survey. *IEEE Access* 5, 21090–21117. doi: 10.1109/ACCESS.2017.2750918

Brunnström, K., Beker, S. A., De Moor, K., Dooms, A., Egger, S., Garcia, M. N., et al. (2013). *Qualinet White Paper on Definitions of Quality of Experience*.

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., and Chen, C. W. (2018a). Blind quality assessment based on pseudo-reference image. *IEEE Trans. Multimedia* 20, 2049–2062. doi: 10.1109/TMM.2017.2788206

Min, X., Gu, K., Zhai, G., Yang, X., Zhang, W., Le Callet, P., et al. (2022). Screen content quality assessment: overview, benchmark, and beyond. *ACM Comput. Surv*. 54, 1–36. doi: 10.1145/3470970

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018b). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast*. 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhai, G., Gu, K., and Yang, X. (2017). Fixation prediction through multimodal analysis. *ACM Trans. Multimedia Comput. Commun. Appl*. 13, 6:1–6:23. doi: 10.1145/2996463

Min, X., Zhai, G., Zhou, J., Farias, M. C., and Bovik, A. C. (2020a). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Proces*. 29, 6054–6068. doi: 10.1109/TIP.2020.2988148

Min, X., Zhai, G., Zhou, J., Zhang, X. P., Yang, X., and Guan, X. (2020b). A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. Image Proces*. 29, 3805–3819. doi: 10.1109/TIP.2020.2966082

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proces*. 13, 600–612. doi: 10.1109/TIP.2003.819861

You, J., Reiter, U., Hannuksela, M. M., Gabbouj, M., and Perkis, A. (2010). Perceptual-based quality assessment for audio-visual services: a survey. *Signal Proces. Image Commun*. 25, 482–501. doi: 10.1016/j.image.2010.02.002

Zhai, G., and Min, X. (2020). Perceptual image quality assessment: a survey. *Sci. China Inform. Sci*. 63, 211301. doi: 10.1007/s11432-019-2757-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 21 October 2021
doi: 10.3389/fncom.2021.746549

Simulation and Mitigation of the Wrap-Around Artifact in the MRI Image

Runze Hu [1], Rui Yang [1], Yutao Liu [2] and Xiu Li [1]*

[1] Department of Information Science and Technology, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, [2] School of Computer Science and Technology, Ocean University of China, Qingdao, China

Magnetic resonance imaging (MRI) is an essential clinical imaging modality for diagnosis and medical research, while various artifacts occur during the acquisition of MRI image, resulting in severe degradation of the perceptual quality and diagnostic efficacy. To tackle such challenges, this study deals with one of the most frequent artifact sources, namely the wrap-around artifact. In particular, given that the MRI data are limited and difficult to access, we first propose a method to simulate the wrap-around artifact on the artifact-free MRI image to increase the quantity of MRI data. Then, an image restoration technique, based on the deep neural networks, is proposed for wrap-around artifact reduction and overall perceptual quality improvement. This study presents a comprehensive analysis regarding both the occurrence of and reduction in the wrap-around artifact, with the aim of facilitating the detection and mitigation of MRI artifacts in clinical situations.

Keywords: magnetic resonance imaging (MRI), wrap-around artifact, deep learning, image quality (IQ), image restoration

OPEN ACCESS

Edited by:
Xiongkuo Min,
University of Texas at Austin,
United States

Reviewed by:
Yucheng Zhu,
Shanghai Jiao Tong University, China
Wei Sun,
Shanghai Jiao Tong University, China

****Correspondence:***
Yutao Liu
liuyutao2008@gmail.com

Received: *24 July 2021*
Accepted: *15 September 2021*
Published: *21 October 2021*

Citation:
Hu R, Yang R, Liu Y and Li X (2021) Simulation and Mitigation of the Wrap-Around Artifact in the MRI Image.
Front. Comput. Neurosci. 15:746549.
doi: 10.3389/fncom.2021.746549

1. INTRODUCTION

Magnetic resonance imaging (MRI) has become one of the most essential means in disease diagnostics and management. It deepens our understanding of the pathology involved in the development and progression of the disease. An MRI image is generally constructed using the Fourier transform (FT) method. The MRI signal is obtained by the interaction between the hydrogen atoms and the external electromagnetic fields. This signal is then encoded into the phase information and frequency information that are subsequently utilized to construct the spatial frequency map, also known as the K-space. The inverse Fourier transform (iFT) can be used to reconstruct the K-space data into the human-interpretable image (Gallagher et al., 2008). Although the MRI technique possesses numerous merits in clinical trials, such as radiation-free and high-contrast imaging, artifacts occur throughout the entire image acquisition process, from the MRI signal generation to the image display, which can significantly deteriorate the perceptual quality of the MRI image and subsequently affect the reliability of diagnosis (Bellon et al., 1986; Liu et al., 2018, 2019, 2020b; Zhai et al., 2020). Thus, it is crucial to effectively detect and eliminate artifacts of MRI image. This study hereby deals with one of the most common artifacts of MRI, namely the wrap-around artifact (also known as the aliasing artifact). We propose a novel artifact reduction framework to reduce the wrap-around artifact of the MRI image while improving the image perceptual quality.

Wrap-around artifact occurs when the scanned area of the human body exceeds the predefined field of view (FOV). These areas outside the FOV cannot be properly encoded relative to their actual position and are wrapped back into the opposite side of the image, resulting in the wrapped information reappearing on the other side of the image and subsequently cannot be distinguished from the objects inside the FOV. The wrap-around artifact can be further classified into frequency-related and phase-related. During the imaging process, there are a number of classical ways to mitigate the wrap-around artifact (Chen et al., 2013). The frequency-related artifact can be mitigated by the oversampling scheme that increases the density of the K-space frequency data and thus increases the FOV. As for the wrap-around in the phase-encode direction, we can swap phase and frequency directions such that the phase direction is oriented in the smallest direction. This method is straightforward while maintains the same spatial resolution. However, it may induce other artifacts to the MRI image, i.e., chemical shift artifact. Another method for reducing the phase-related artifact is to double the FOV in the phase direction, yet it may lower the spatial resolution. These remedies are only operational during the process of MR imaging. However, radiologists generally face post-operated (reconstructed) MRI image without knowing the occurrence of the artifact in the imaging processing. Eliminating the wrap-around artifact from the post-operated MRI image has remained a major deterrent to clinical adoption.

Numerous efforts for MRI artifacts reduction have been made in the last decades. Yang et al. (2001) proposed a maximum likelihood-based method to remove the ringing artifact, in which the prior knowledge of MRI, i.e., the sampled low-frequency data points, was adopted to deduce the high-frequency data in the K-space. This method aims to increase the high-frequency information and thus alleviate the artifact. Lee (1998) designed a Bayesian framework with the regularization scheme to reduce the MRI artifacts. This framework deduces the posterior probability of the output image by the likelihood of sampled spatial information and the local spatial structure of the input image. Yatchenko et al. (2013) mitigated the ringing artifact by computing the average edge-normal and edge tangential derivatives in the edge area of the image. In Guo and Huang (2009), a k-means-based method was proposed to remove the MRI artifact. The maximum likelihood method was at first employed to detect the artifact of the image. Then, the detected structures were fitted to a k-means model to map the neighboring pixel values and the estimation region. Sebastiani and Barone (1995) proposed to use the Markov random field to model the errors arisen in the truncation and characteristics of the Fourier series. The modeled errors can be utilized to implement artifact removal.

In addition to these model-based approaches, recent years have seen the prosperity of the deep learning-based techniques for the MRI artifact reduction. Lee et al. (2017) proposed a multi-scale deep neural network to remove the wrap-around artifact. This neural network estimates the area of a wrap-around artifact based on the distorted magnitude and the phase information of the input image. The removal of wrap-around artifact can be achieved by subtracting the estimated artifact area from the input image. Yang et al. (2017) proposed a de-aliasing strategy based on the conditional generative adversarial networks. The adversarial loss of this model incorporates three typical losses, i.e., the pixel-wise loss, frequency information loss, and perceptual loss, in order to better learn the texture and edge information, thereby improving the quality of the MRI image reconstructed from undersampled k-space data. The work in Hyun et al. (2018) presented a deep learning-based sample strategy to reconstruct the MR image from the undersampled k-space data while enhancing the image quality. This strategy adopts the uniform sampling method to obtain phase information of the image so that the details of the corrupted area of the image are preserved after Fourier transform. Consequently, the deep learning model can effectively learn the features of the wrap-around artifact.

Although the abovementioned model and learning-based methods have shown great potential in reducing the MRI artifacts, their capability for clinical practice is restricted. The reason is fourfold. First, many methods, i.e., Yang et al. (2001) and Guo and Huang (2009), directly manipulate the k-space data, which could inadvertently remove the non-artifact information, such as the anatomical or pathological details. Second, some deep learning methods, i.e., Yang et al. (2017), are based on the generative adversarial network (GAN), where the MRI image is synthesized from the given samples. This strategy is not very reliable since the synthesized MRI data may contain fake information, which can complicate the pathologic diagnosis. Third, in the context of Bayesian framework, such as Lee (1998) and Sebastiani and Barone (1995), reconstructing the MRI image from the undersampled k-space data is practically an ill-posed problem, and the rate of convergence of these methods remains questionable. Finally, one of the main limitations of the learning-based method is the scarcity of MRI data. Nevertheless, given the sensitivity and confidentiality of clinical data, it is rather difficult to obtain adequate MRI data, which severely restricts the development of learning-based methods.

We herein propose a novel wrap-around artifact reduction framework to address the aforementioned issues. The proposed framework comprises two stages, namely, artifact simulation and image enhancement. For the artifact simulation, we design an artifact occurrence mechanism to simulate the characteristics of the wrap-around artifact. Two parameters are designed to describe the characteristics of the wrap-around artifact. The first parameter determines the size of the wrapped area indicating how much area of the MRI image is corrupted by the artifact. The second parameter describes the intensity of the wrapped area, which is closely related to the distortion level of the MRI image. A large intensity may completely contaminate the wrapped area, of the MRI image, resulting in the difficulty of artifact removal. These two parameters work jointly to simulate the wrap-around artifact.

For the image enhancement, we propose a deep neural network (DNN) to remove the wrap-around artifact while improving the overall perceptual quality of the MRI image (Min et al., 2020a,b). The proposed DNN is based on the U-net network owing to its powerful performance in medical image processing. The DNN composes of two phases, i.e., artifact estimation and deep elimination. In the artifact estimation, a U-net-based network is trained by pairing artifacted MRI images with the corresponding artifact patterns. This enables the network to accurately estimate the wrapped area of the artifacted MRI image, which can be subsequently utilized to assist the network training at the second phase of deep elimination. As for the deep elimination, an end-to-end U-net based network is built, in which the inputs are the artifacted MRI images and the outputs are the artifact-free MRI images. In this phase, the loss function is dedicatedly designed based on the binary cross entropy (BCE) loss and the mean squared error (MSE) loss in order to maximize the performance of artifact elimination. These two phases work cooperatively to remove the wrap-around artifact while improving the image quality. Experiments, in terms of quantitative metrics and qualitative visualizations, demonstrate the high potential of the proposed method in the reduction of the wrap-around artifact.

The rest of this study is organized as follows. Section 2 details the proposed framework. Section 3 presents the experiments and detailed analysis regarding the wrap-around artifact removal. Finally, we conclude the work of this study in section 4.

2. METHODOLOGY

In section 2, we first propose a technique to simulate the wrap-around artifact on the MRI image. Then, a dataset is formed by pairing the artifact-free MRI image with the artifacted MRI image obtained from the proposed simulation technique. At last, the dataset is employed to train a deep learning network to implement the removal of the wrap-around artifact.

2.1. Artifact Simulation

For the MRI image with the wrap-around artifact, two factors affect the perceptual quality of the image, including the size and intensity of the wrapped area. Therefore, we generate the wrap-around artifact based on these two factors. Given an artifact-free MRI image $\boldsymbol{I} \in \mathbb{R}^{M\times N}$, we produce the artifact layer $\hat{\boldsymbol{I}}$ by horizontally shifting the pixels in $\boldsymbol{I}$ as

$$\hat{\boldsymbol{I}}(\hat{m},\hat{n}) = \begin{cases} 0, & d+\hat{n} \leq N \\ \boldsymbol{I}(\hat{m}, d+\hat{n}-N)\cdot r, & \text{otherwise} \end{cases} \tag{1}$$

where $\hat{\boldsymbol{I}}(\hat{m},\hat{n})$ indicates the pixel of $\hat{\boldsymbol{I}}$ located at $(\hat{m},\hat{n})$, $d \in [1, N]$ is the shift distance of $\boldsymbol{I}$ meaning that $\boldsymbol{I}$ is shifted horizontally by d columns, and $r > 0$ determines the intensity of the artifact layer. This study only considers the horizontal shift, implying that the wrap-around artifact only appears on either the right side or the left side of the image. However, it is straightforward to apply the proposed method to the situation of vertical shift.

After obtaining the artifact layer, the wrap-around artifact can be produced by directly adding the image and the artifact layer together. However, doing so will change the image contrast as the pixel value of the wrapped area is increased after the summation. Such a change will increase the difference between the light and dark areas of the image leading to that light areas become lighter and dark areas become darker. Consequently, the simulated artifact is inconsistent with the clinical practice. Herein, we propose a technique to circumvent these problems. Let $\boldsymbol{F}$ and $\hat{\boldsymbol{F}}$ be the binary patterns of $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$, respectively, satisfying

$$\boldsymbol{F}(m,n) = \begin{cases} 1, & \boldsymbol{I}(m,n) > 0 \\ 0, & \text{otherwise} \end{cases};\ \hat{\boldsymbol{F}}(m,n) = \begin{cases} 1, & \hat{\boldsymbol{I}}(m,n) > 0 \\ 0, & \text{otherwise} \end{cases}.$$

We first overlay the image $\boldsymbol{I}$ with the artifact layer $\hat{\boldsymbol{I}}$ as

$$\boldsymbol{I}_r = (\boldsymbol{I} + \hat{\boldsymbol{I}}) \odot \boldsymbol{F}, \tag{2}$$

where $\odot$ indicates the element-wise multiplication. The summation of $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$ will lead to the artifact appearing on the blank area of the image. This kind of artifact information does not corrupt the image, and can be easily eliminated by applying the binary pattern of $\boldsymbol{I}$. Therefore, this study is only interested in the artifacts that contaminate the image information of $\boldsymbol{I}$. We apply $\boldsymbol{F}$ in Equation (2) to remove the artifact on the blank area of the image.

Then, the wrapped area of the image can be calculated by $\boldsymbol{V} = (\boldsymbol{F} + \hat{\boldsymbol{F}}) \odot \boldsymbol{F}$. The elements in $\boldsymbol{V}$ involves three different values, such as 0, 1, and 2. The pixels in the wrapped area are marked as 2. Therefore, we can obtain the location and size of the wrapped area by counting the number of elements of 2 in $\boldsymbol{V}$. Following this, we calculate the brightness ratio between the non-wrapped area and the wrapped area in the original artifact-free image. When artifacts are generated, we maintain the same ratio to avoid the problem of uneven brightness. Let I_1 and I_2 be the summation of the brightness in the unwrapped-area and wrapped-area of the image, written by

$$\begin{aligned} I_1 &= \sum \boldsymbol{I}(m,n),\ \text{for}\ \boldsymbol{V}(m,n) = 1 \\ I_2 &= \sum \boldsymbol{I}(m,n),\ \text{for}\ \boldsymbol{V}(m,n) = 2. \end{aligned} \tag{3}$$

The final wrapped-around artifact is generated as

$$\boldsymbol{I}_s(m,n) = \begin{cases} \boldsymbol{I}_r(m,n)\cdot \frac{I_2}{I_1}, & \boldsymbol{V}(m,n) = 2 \\ \boldsymbol{I}_r(m,n), & \text{otherwise.} \end{cases} \tag{4}$$

The complete process of the wrap-around artifact simulation is presented in **Figure 1A**. The proposed simulation technique allows us to overcome the problem of data shortage. Hence, we can produce adequate artifact resources to facilitate the development of the artifact reduction technique. Toward this end, a deep learning-based method for the reduction of the wrap-around artifact is proposed.

2.2. Problem Formulation

In general, the observed image $\boldsymbol{Y}$ can be represented using a discrete linear model, written by

$$\boldsymbol{W}\boldsymbol{I}_s + \epsilon = \boldsymbol{Y} \approx \boldsymbol{I}, \tag{5}$$

where $\boldsymbol{I}_s$ and $\boldsymbol{I}$ are the artifacted and artifact-free images, respectively. $\boldsymbol{W}$ is a linear operator representing various operations against the image quality, i.e., the convolution operation in the K-space for image deblurring or the non-local

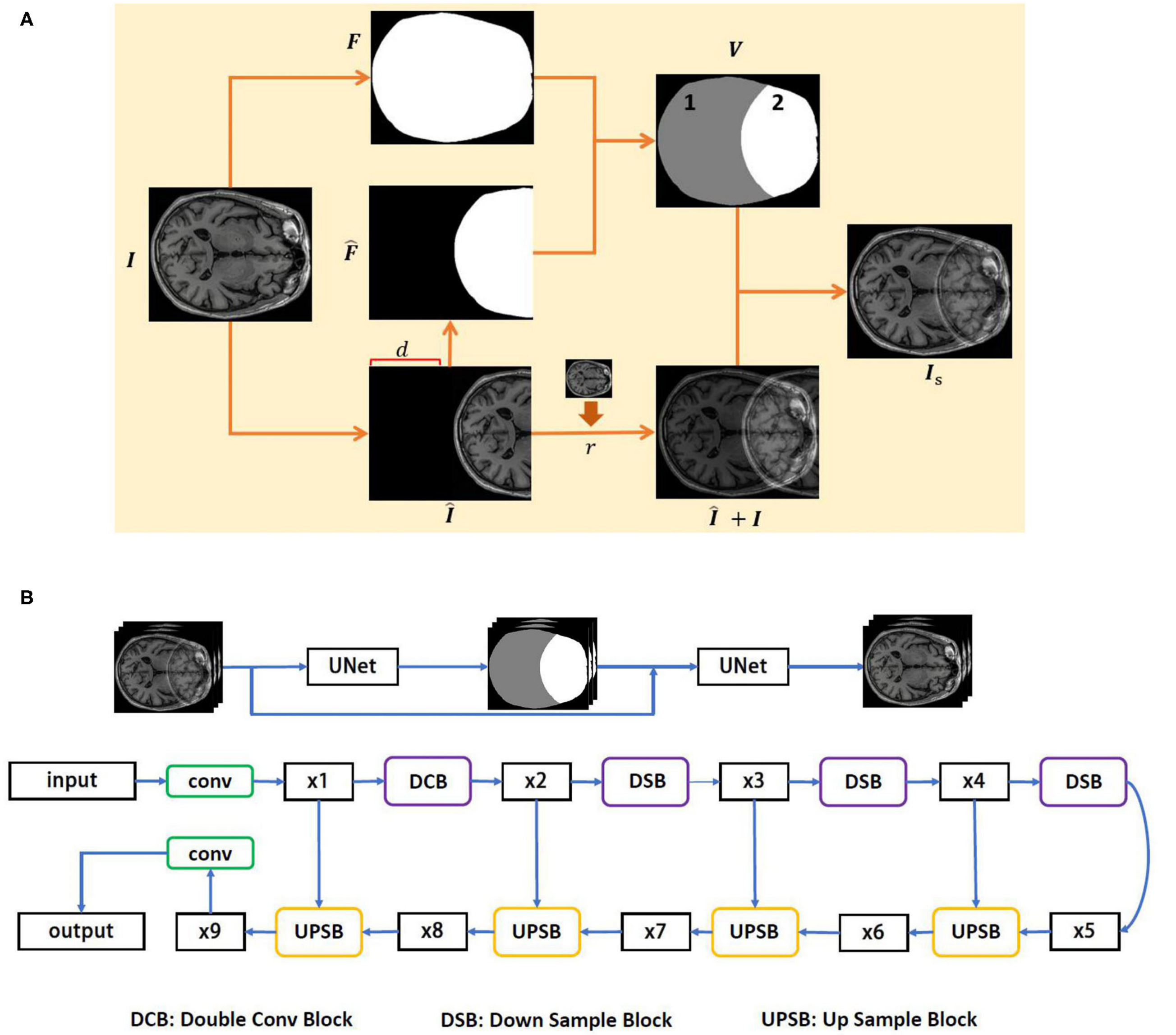

FIGURE 1 | The proposed wrap-around artifact reduction framework. **(A)** The process of the wrap-around artifact simulation. **(B)** The architecture of the DUARN method.

means filtering for image denoising. ϵ is a bias term. Our purpose is to solve $\boldsymbol{W}$ in Equation (2), which is an ill-posed inverse problem that the solution of $\boldsymbol{W}$ is generally underdetermined. The priori knowledge of $\boldsymbol{I}_s$, is therefore, required in order to constrain the solution space of $\boldsymbol{W}$. In other words, we hope to find $\boldsymbol{W}$ such that

$$L = \frac{1}{2}||\boldsymbol{I} - \boldsymbol{Y}||_2^2 + \lambda R(\boldsymbol{I}_s) \qquad (6)$$

reaches minimum, where $\frac{1}{2}||\boldsymbol{I} - \boldsymbol{Y}||_2^2$ is known as data term. The regularization term $\lambda R(\boldsymbol{I}_s)$ with the regularization parameter λ is utilized to alleviate the problem of ill-posedness and $R(\boldsymbol{I}_s)$ generally involves l_q-norms. Equation (6) can be solved by learning-based method, such as the gradient descent method, to iteratively minimize the difference between $\boldsymbol{I}$ and $\boldsymbol{Y}$ to the local minimum.

2.3. Learning-Based Artifact Reduction

A deep learning-based method is herein proposed to solve Equation (6). The proposed method is constructed by two U-net networks; thus, we name it as Dual U-net Artifact Reduction Network (DUARN). The two U-nets correspond to two phases of DUARN, namely, artifact estimation and deep elimination. In the first phase, we train a U-net by pairing the artifacted MRI images with the binary artifact pattern aiming at accurately predicting the artifact area from the input artifacted image. The BCE loss is adopted in the network training owing to its powerful performance of binary image prediction. In the second phase, we rewrite Equation (6) as

$$\boldsymbol{L} = a\boldsymbol{L}_{MSE} \odot \boldsymbol{P}_1 + b\boldsymbol{L}_{BCE} \tag{7}$$

where $\boldsymbol{P}_1$ is a binary pattern of the artifacted image obtained from the first phase of DUARN. When the pixel of $\boldsymbol{P}_1$ is in the area of artifact, it is equal to 1 otherwise equal to 0. We use $\boldsymbol{P}_1$ to maximize the learning efficiency of the second deep neural network. Since this study only deals with the wrap-around artifact, we assume that there is no other artifact existing outside the wrapped area of the image. Hence, when the network learns from the loss function, we set a relatively small weight for those losses outside the wrap-around artifact area, thereby improving the efficiency and accuracy of the network. $\boldsymbol{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{I}_i - \boldsymbol{Y}_i)^2$ indicates the pixel-wise image domain MSE loss, where $\boldsymbol{I}_i$ and $\boldsymbol{Y}$ are the $i-$th pixel value of $\boldsymbol{I}$ and $\boldsymbol{Y}$. $\boldsymbol{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i \log(p(\boldsymbol{Y}_i)) + (1 - \boldsymbol{Y}_i) * \log(1 - p(\boldsymbol{Y}_i))$ refers to the BCE loss that minimizes the average probability error between the target and predicted images for each pixel. Herein, we adopt the BCE loss to penalize the misalignment of boundaries. a and b are small positive real numbers, satisfying $a + b = 1$. Empirically, we set $a = 0.75$ and $b = 0.25$.

The architecture of DUARN is illustrated in **Figure 1B**. The DUARN contains two U-nets, each of which involves four scales, such as 64, 128, 256, and 512. In the input layer, 64 filters with kernel size of 3×3 and ReLU as an activation function are applied. Following the input layers, there are four convolution layers (encoder) and four transposed convolution layers (decoder) with each followed by batch normalization and ReLU layers. The skip connection between the 2×2 strided convolution (downscaling) and 2×2 transposed convolution (upscaling) are employed in order to supplement the reconstruction details with different level of features. Finally, a 1×1 convolution layer is used to predict a single channel image as the output of the network.

3. EXPERIMENTS AND ANALYSIS

In this section, we evaluate the proposed DUARN method with respect to its quantitative and qualitative performance. First, we enlist the help of radiologist to select 140 artifact-free and high perceptual quality MRI images (T1-weighted). The invited radiologist who has over 5 years of clinical experience in the brain radiology. Following the MRI data acquisition, we simulate the wrap-around artifact on the 140 artifact-free MRI images using the method proposed in section 2.1. We generate five different degrees of the wrap-around artifacts corresponding to five distortion levels of the image, in which the distortion level of 1, 2, 3, 4, and 5 indicate minor artifact, mild artifact, moderate artifact, severe artifact, and non-diagnostic as suggested by Liu et al. (2017, 2020a) and Liu and Li (2020). The simulation process of the artifact is carried out under the guidance of the radiologist who visually assesses the quality of each simulated image and recommends the parameter values of d and r in Equation (1) to ensure the generated image matches the desired distortion level. Examples of the simulated MRI images are presented in **Figure 2** with the distortion level ranging from 1 to 5. The ground truth image is also provided on the left of **Figure 2**. It is observed that the wrap-around artifact on the minor artifacted MRI image is insignificant in terms of the area size and the intensity of the artifact. Images of such a quality may still be useful if the diagnostic area of interest is outside the artifact. Correspondingly, the MRI images with severe and non-diagnostic artifact can hardly be useful under any clinical situations.

We then produce the dataset of artifacted MRI images for training the proposed DUARN method. Since we have 140 artifact-free images, the produced dataset yields to a total number of 700 artifacted images. The dataset is split into two non-overlapped parts, i.e., training data and testing data with the standard ratio of 80/20%. We train the DUARN on the training data and test it on the testing data. We at first train the first U-net of DUARN, where the Adam optimizer is adopted with the initial learning rate of 0.0001, batch size of 2, and momentum of 0.8. When the training process is complete, we train the second U-net using the artifacted MRI image and the output of the first U-net as its inputs. The Adam optimizer is also applied to the second U-net with the initial learning rate of 0.00001, batch size of 1, and momentum of 0.9. The early stopping scheme is employed in the training process of both U-nets for the prevention of overfitting. In addition, the conventional data augmentation techniques, such as image flipping, rotating, and brightness adjustment, are adopted to boost the network performance.

In order to vividly demonstrate the performance of the proposed method, we compare it quantitatively and qualitatively with the state-of-the-art artifact reduction method in Tamada et al. (2020). Tamada et al. (2020) proposed an artifact reduction method, namely motion artifact reduction based on convolutional neural network (MARC) method, to remove the motion ghost from the MRI images. In the MARC method, a convolutional neural network (CNN)-based network was trained to extract the artifact components from the artifacted images. The artifacts can be, therefore, removed by subtracting the extracted artifact component from the input image. The targeted artifact in Tamada et al. (2020) is similar to the wrap-around artifact since both of them belong to the aliasing of the image. We adapt the MARC method to implement the wrap-around artifact reduction and present the results of the MARC method and DUARN method in **Figure 3**. As can be observed, both methods are capable of eliminating the wrap-around artifact to a certain extent while the qualitative performance of the DUARN method is notably better than the MARC method, especially for those high distortion level image, i.e., 2nd and 5th images. In addition, we noticed that although the MARC method can alleviate the wrap-around artifacts, noise may be introduced into the images, resulting in further degradation of image quality. This is inconsistent with our purpose of obtaining high-quality artifact-free MRI image. On the contrary, the DUARN method can maintain the high perceptual quality after the artifact removal.

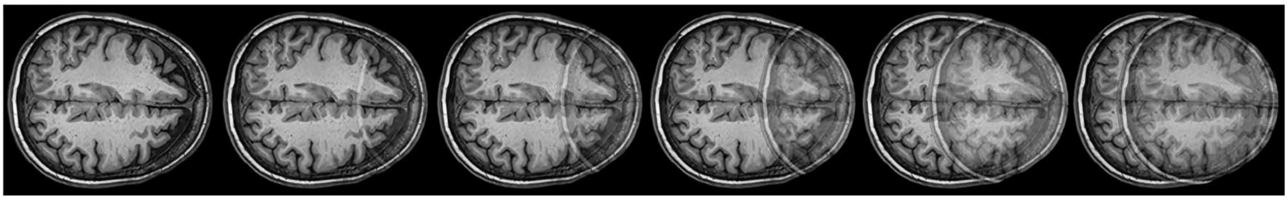

FIGURE 2 | MRI images with different degrees of wrap-around artifact. From the left to right is the ground truth image, minor artifact, mild artifact, moderate artifact, severe artifact, and non-diagnostic.

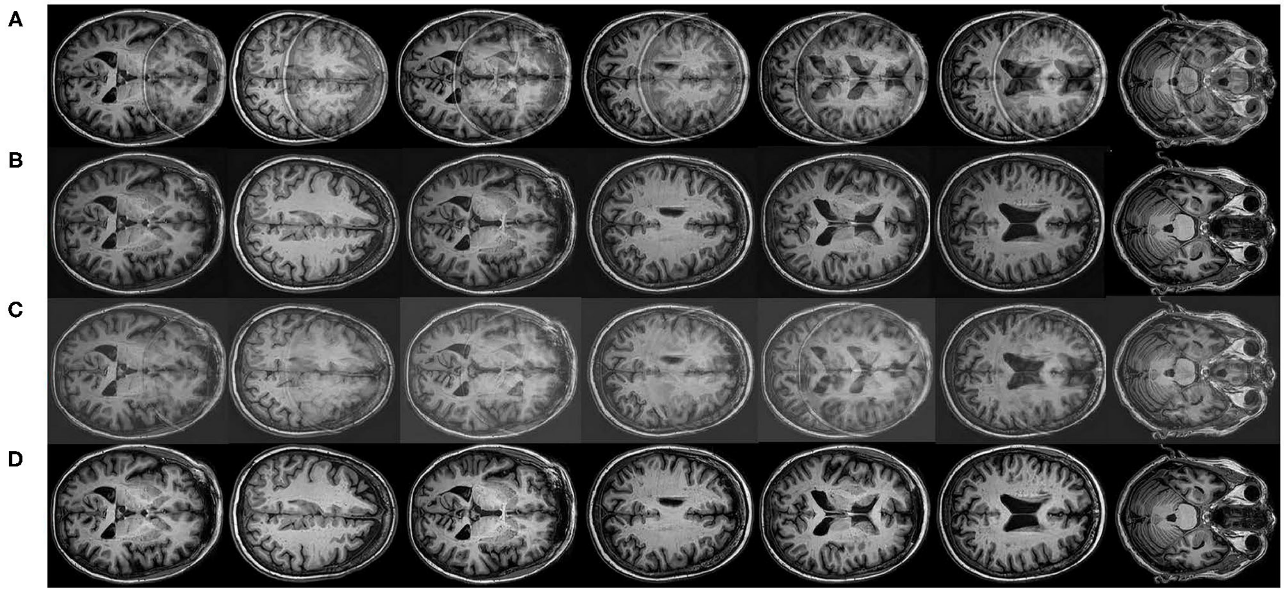

FIGURE 3 | Qualitative visualizations. **(A)** Artifacted MRI images involving mild, moderate, severe, and non-diagnostic wrap-around artifacts. **(B)** Reconstructed MRI images from the DUARN method. **(C)** Reconstructed MRI images from the MARC method. **(D)** Ground truth images.

TABLE 1 | The SSIM and PSNR from DUARN and MARC methods.

	Minor artifact	Mild artifact	Moderate artifact	Severe artifact	Non-diagnostic	Overall
SSIM						
MARC	0.9033	0.8949	0.9006	0.8766	0.8739	0.8899
DUARN1	0.9221	0.9292	0.9339	0.9384	0.9401	0.9327
DUARN2	0.8941	0.8901	0.8867	0.9012	0.8993	0.8943
DUARN	0.9536	0.95577	0.9594	0.9654	0.9684	**0.9605**
PSNR						
MARC	22.2786	20.7544	21.0742	17.6049	15.9149	19.5254
DUARN1	23.4502	23.6031	23.0125	24.5327	26.0182	24.1233
DUARN2	20.1963	19.7167	19.5369	22.3562	22.6943	20.9001
DUARN	24.3683	24.1942	24.0889	25.7176	27.6271	**25.1992**

DUARN1 and DUARN2 indicate the DUARN method with the MSE loss and the BCE loss, respectively. Overall indicates the average values of SSIM and PSNR for all the 140 testing images. The highest performance values on each evaluation index are highlighted with boldface.

Finally, we evaluate the quantitative performance of the DUARN method by quantifying the quality of the reconstructed MRI image. Numerous image quality metrics have been proposed in the last decades each with their respective merits (Zhang et al., 2011; Mittal et al., 2012; Min et al., 2017, 2019, 2020c). In this study, we adopt two widely used metrics to measure the MRI image quality, including the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM) (Wang et al., 2004). **Table 1** tabulates the PSNR and SSIM from the DUARN and MRAC methods. The testing data contain 140 images with 28 images for each artifact type. We calculate the average PSNR and SSIM for each artifact type, and the overall in **Table 1** refers to the average PSNR and SSIM for all the 140 testing images. As can be observed, the DUARN model achieves superior performance in the evaluation of all types of artifacts. More importantly, when the degree of image distortion increases, the performance of MARC method shows a clear downward trend. Comparatively, the DUARN method can still maintain a robust performance and even has a slight upward trend. This implies that the capability of the DUARN method will not be affected by the distortion level of the image. Such a feature is essential because clinical trials often face artifacted MRI images with various distortion levels, which may exceed the scope of the test samples. An artifact removal technique with stable performance can exert promising application value in practice. The DUARN method can be also combined with other image enhancement techniques, such as contrast stretching and histogram equalization, to further improve the perceptual quality of the reconstructed MRI image. This can be considered in the future work. In addition, since the DUARN method combines two losses of the BCE loss and the MSE loss in the network training, we are interested in the individual contribution of each loss in the performance of the proposed method. Toward this end, we introduce each loss to the network training of the DUARN method and quantify the performance of each loss by the PSNR and SSIM. The experimental results are presented in **Table 1**, where DUARN1 and DUARN2 indicate the DUARN method with the MSE loss and the BCE loss, respectively. As observed, the MSE loss brings more contributions in the DUARN method, and the combination of these two losses earns the best performance, which evidences that the BCE loss and the MSE loss play complementary roles in the DUARN method.

4. CONCLUSION

This study deals with the wrap-around artifact of the MRI image, wherein two contributions are made. We first propose a simulation technique to generate the wrap-around artifact on the MRI image. The design of the proposed method is based on the image quality assessment scheme and with the assistance of an experienced radiologist, which allows the simulated artifact resources to match clinical situations. Then, we propose a novel artifact reduction technique, based on the deep neural network, to implement the elimination of the wrap-around artifact. This technique composes two U-net networks corresponding to two phases, such as artifact estimation and deep elimination. Dedicated losses are designed in order to maximize the effectiveness of artifact removal while improving the perceptual quality of the reconstructed MRI image. Extensive experiments are carried out to evaluate the quantitative and qualitative performance of the proposed method, with the results demonstrating the superiority of the proposed method against the state-of-the-art method.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

RH, YL, and RY developed the theoretical formalism, performed the analytic calculations, and carried out the experiments. XL supervised the project. All authors contributed to the final version of the manuscript.

FUNDING

This work was supported in part by the China Postdoctoral Science Foundation under Grant 2019M650686.

REFERENCES

Bellon, E. M., Haacke, E. M., Coleman, P. E., Sacco, D. C., Steiger, D. A., and Gangarosa, R. E. (1986). MR artifacts: a review. *Am. J. Roentgenol*. 147, 1271–1281. doi: 10.2214/ajr.147.6.1271

Chen, L., Bao, L., Li, J., Cai, S., Cai, C., and Chen, Z. (2013). An aliasing artifacts reducing approach with random undersampling for spatiotemporally encoded single-shot MRI. J. *Magn. Reson*. 237, 115–124. doi: 10.1016/j.jmr.2013.10.005

Gallagher, T. A., Nemeth, A. J., and Hacein-Bey, L. (2008). An introduction to the fourier transform: relationship to MRI. *Am. J. Roentgenol*. 190, 1396–1405. doi: 10.2214/AJR.07.2874

Guo, W., and Huang, F. (2009). "Adaptive total variation based filtering for MRI images with spatially inhomogeneous noise and artifacts," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro (IEEE)* (Boston, MA), 101–104. doi: 10.1109/ISBI.2009.5192993

Hyun, C. M., Kim, H. P., Lee, S. M., Lee, S., and Seo, J. K. (2018). Deep learning for undersampled MRI reconstruction. *Phys. Med. Biol*. 63:135007. doi: 10.1088/1361-6560/aac71a

Lee, D., Yoo, J., and Ye, J. C. (2017). "Deep residual learning for compressed sensing MRI," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017) (IEEE)* (Melbourne, VIC), 15–18. doi: 10.1109/ISBI.2017.7950457

Lee, S.-J. (1998). "Improved method for reduction of truncation artifact in magnetic resonance imaging," in *Applications of Digital Image Processing XXI (International Society for Optics and Photonics)*, Vol. 3460, 587–598. doi: 10.1117/12.323214

Liu, Y., Gu, K., Li, X., and Zhang, Y. (2020a). Blind image quality assessment by natural scene statistics and perceptual characteristics. ACM Trans. Multimedia Comput. *Commun. Appl*. 16, 1–91. doi: 10.1145/3414837

Liu, Y., Gu, K., Wang, S., Zhao, D., and Gao, W. (2019). Blind quality assessment of camera images based on low-level and high-level statistical features. *IEEE Trans*. Multimedia 21, 135–146. doi: 10.1109/TMM.2018.2849602

Liu, Y., Gu, K., Zhai, G., Liu, X., Zhao, D., and Gao, W. (2017). Quality assessment for real out-of-focus blurred images. *J. Vis. Commun. Image Represent*. 46, 70–80. doi: 10.1016/j.jvcir.2017.03.007

Liu, Y., Gu, K., Zhang, Y., Li, X., Zhai, G., Zhao, D., et al. (2020b). Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. IEEE Trans. *Circuits Syst*. Video Technol. 30, 929–943. doi: 10.1109/TCSVT.2019.2900472

Liu, Y., and Li, X. (2020). No-reference quality assessment for contrast-distorted images. *IEEE Access* 8, 84105–84115. doi: 10.1109/ACCESS.2020.2991842

Liu, Y., Zhai, G., Gu, K., Liu, X., Zhao, D., and Gao, W. (2018). Reduced-reference image quality assessment in free-energy principle and sparse representation. *IEEE Trans*. Multimedia 20, 379–391. doi: 10.1109/TMM.2017.2729020

Min, X., Ma, K., Gu, K., Zhai, G., Wang, Z., and Lin, W. (2017). Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Trans. Image Process*. 26, 5462–5474. doi: 10.1109/TIP.2017.2735192

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Trans. Multim*. 21, 2319–2333. doi: 10.1109/TMM.2019.2902097

Min, X., Zhai, G., Zhou, J., Farias, M. C., and Bovik, A. C. (2020a). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Process*. 29, 6054–6068. doi: 10.1109/TIP.2020.2988148

Min, X., Zhai, G., Zhou, J., Zhang, X.-P., Yang, X., and Guan, X. (2020b). A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. Image Process*. 29, 3805–3819. doi: 10.1109/TIP.2020.2966082

Min, X., Zhou, J., Zhai, G., Le Callet, P., Yang, X., and Guan, X. (2020c). A metric for light field reconstruction, compression, and display quality evaluation. *IEEE Trans. Image Process*. 29, 3790–3804. doi: 10.1109/TIP.2020.2966081

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process*. 21, 4695–4708. doi: 10.1109/TIP.2012.2214050

Sebastiani, G., and Barone, P. (1995). Truncation artifact reduction in magnetic resonance imaging by Markov random field methods. *IEEE Trans. Med. Imaging* 14, 434–441. doi: 10.1109/42.414607

Tamada, D., Kromrey, M.-L., Ichikawa, S., Onishi, H., and Motosugi, U. (2020). Motion artifact reduction using a convolutional neural network for dynamic contrast enhanced MR imaging of the liver. *Magn. Reson. Med. Sci*. 19:64. doi: 10.2463/mrms.mp.2018–0156

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., et al. (2017). Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE Trans. Med. Imaging* 37, 1310–1321. doi: 10.1109/TMI.2017.2785879

Yang, S., Hu, Y.-H., Nguyen, T. Q., and Tull, D. L. (2001). Maximum-likelihood parameter estimation for image ringing-artifact removal. IEEE *Trans. Circuits Syst. Video Technol*. 11, 963–973. doi: 10.1109/76.937440

Yatchenko, A. M., Krylov, A. S., and Nasonov, A. V. (2013). Deringing of MRI medical Images. Pattern Recogn. image Anal. 23, 541–546. doi: 10.1134/S1054661813040184

Zhai, G., Zhu, Y., and Min, X. (2020). Comparative perceptual assessment of visual signals using free energy features. *IEEE Trans. Multimedia*. doi: 10.1109/TMM.2020.3029891

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process*. 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 21 October 2021
doi: 10.3389/fnins.2021.739138

IE-IQA: Intelligibility Enriched Generalizable No-Reference Image Quality Assessment

Tianshu Song[1], Leida Li[2,3], Hancheng Zhu[4] and Jiansheng Qian[1]**

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [2] School of Artificial Intelligence, Xidian University, Xi'an, China, [3] Pazhou Lab, Guangzhou, China, [4] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China

Image quality assessment (IQA) for authentic distortions in the wild is challenging. Though current IQA metrics have achieved decent performance for synthetic distortions, they still cannot be satisfactorily applied to realistic distortions because of the generalization problem. Improving generalization ability is an urgent task to make IQA algorithms serviceable in real-world applications, while relevant research is still rare. Fundamentally, image quality is determined by both distortion degree and intelligibility. However, current IQA metrics mostly focus on the distortion aspect and do not fully investigate the intelligibility, which is crucial for achieving robust quality estimation. Motivated by this, this paper presents a new framework for building highly generalizable image quality model by integrating the intelligibility. We first analyze the relation between intelligibility and image quality. Then we propose a bilateral network to integrate the above two aspects of image quality. During the fusion process, feature selection strategy is further devised to avoid negative transfer. The framework not only catches the conventional distortion features but also integrates intelligibility features properly, based on which a highly generalizable no-reference image quality model is achieved. Extensive experiments are conducted based on five intelligibility tasks, and the results demonstrate that the proposed approach outperforms the state-of-the-art metrics, and the intelligibility task consistently improves metric performance and generalization ability.

Keywords: image quality assessment, NR-IQA, intelligibility, distortion, generalization, semantic

OPEN ACCESS

Edited by:
Guangtao Zhai,
Shanghai Jiao Tong University, China

Reviewed by:
Yucheng Zhu,
Shanghai Jiao Tong University, China
Weisheng Li,
Chongqing University of Posts and Telecommunications, China
Wei Sun,
Shanghai Jiao Tong University, China

***Correspondence:**
Leida Li
ldli@xidian.edu.cn
Jiansheng Qian
qianjsh@cumt.edu.cn

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *10 June 2021*
Accepted: *08 September 2021*
Published: *21 October 2021*

Citation:
Song T, Li L, Zhu H and Qian J (2021) IE-IQA: Intelligibility Enriched Generalizable No-Reference Image Quality Assessment. Front. Neurosci. 15:739138. doi: 10.3389/fnins.2021.739138

1. INTRODUCTION

Image quality assessment (IQA) plays a vital role in image acquisition, compression, enhancement, retrieval, etc. The existing IQA metrics are mainly designed for synthetic distortions and cannot be applied to wild images satisfactorily due to the limited generalization ability. Fundamentally, image quality embodies two aspects: distortion and intelligibility (Abdou and Dusaussoy, 1986). Most IQA algorithms only focus on the distortion measurement and the intelligibility aspect is rarely investigated. In this paper, we mainly investigate the role of intelligibility in building a highly generalizable IQA model.

Intelligibility refers to the ability of an image to provide information to a person or a machine (Abdou and Dusaussoy, 1986), that is, the degree to which the image could be understood. Distortions affect image intelligibility, and accordingly, intelligibility is indicative of image quality when humans make judgments. Traditional handcrafted feature-based IQA metrics mainly focus on distortions and cannot commendably describe image intelligibility. Deep learning-based methods learn the IQA task in a data-driven manner, and consequently do not directly pay attention to image intelligibility, either.

Since the most essential function of image is to convey information, when distortions seriously undermine the expression of information, the intelligibility will also become low, which in turn indicates poor image quality. Real-world images are typically contaminated by complicated distortions, which lead to different degrees of intelligibility. **Figure 1** explains how intelligibility indicates image quality. **Figures 1A,B** both suffer from severe motion blur, and both contain human as the main content. The human face in **Figure 1A** is too blurred to be recognized, whereas a woman's face in **Figure 1B** can still be easily identified. Thus, **Figure 1B** has higher intelligibility and accordingly higher quality score. The distortion in **Figure 1C** is not heavier than **Figure 1D**, but **Figure 1D** is easier to be recognized; hence, **Figure 1D** has higher intelligibility and accordingly higher quality score. Finally, **Figures 1E,F** was mainly underexposed with locally overexposed. The main content in **Figure 1E** is illegible, whereas **Figure 1F** can still be distinguished as a singing stage with performers. Therefore, the quality of **Figure 1F** is better than that of **Figure 1E**. It can be concluded from **Figure 1** that images with similar distortions may have significantly different quality due to different degrees of intelligibility. Therefore, a robust quality assessment metric should also take intelligibility into account, especially for severe distortions.

Motivated by the above facts, this paper presents a new framework to achieve highly generalizable image quality assessment by integrating intelligibility and distortion measure. The intelligibility of an image can be represented from different perspectives, such as "whether the content of the image is recognizable," "which category does the main object in the image belong to," and "what scene does the image show." The results of these questions are all important information conveyed by the image, and through the mining of these questions, we can obtain descriptions of image intelligibility. These questions can be described by popular computer vision tasks, such as image classification, scene recognition, object detection, and instance segmentation. Therefore, we calculate intelligibility features based on these semantic tasks. Then, we propose a bilateral network to combine the distortion features and intelligibility features. Further, we design different feature selection strategies for different semantic understanding tasks. This produces highly generalizable intelligibility features. The distortion network is applied to extract distortion features that are complementary to those intelligibility features. With the bilateral network, highly generalizable intelligibility features with rich semantic information can be fused with distortion features, producing the final IQA model.

The contributions of this work are summarized as follows:

- We propose a new framework for designing highly generalizable image quality models by integrating intelligibility and distortion, two fundamental aspects of image quality. In the proposed framework, intelligibility features can be extracted based on popular semantic tasks, such as image recognition, scene classification, and object detection.
- We propose a bilateral network with an intelligibility enhanced module to fuse intelligibility features with distortion features for building a robust IQA model. A feature selection strategy is proposed to extract intelligibility features instead of doing direct training. This strategy can avoid the risk of damaging generalizable features.

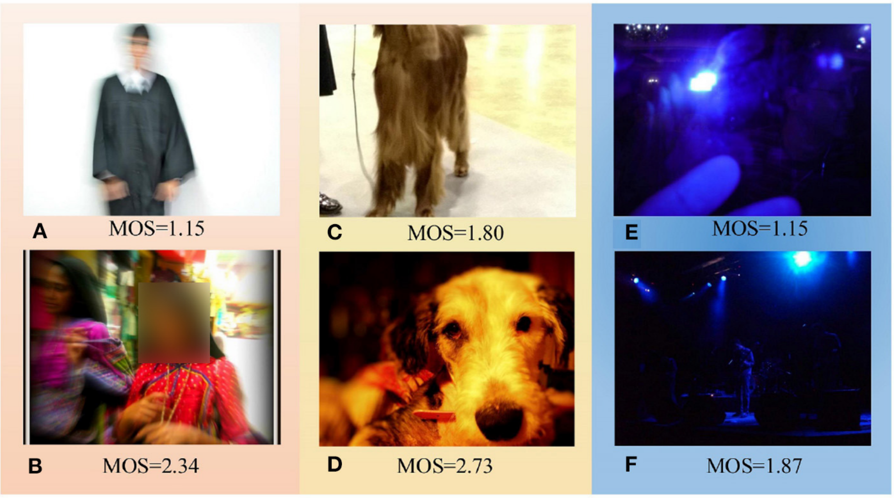

FIGURE 1 | Relation between intelligibility and image quality. **(A–F)** Compared to images in the first row, images in the second row have higher intelligibility and accordingly higher mean opinion score (MOS). Images are from the KonIQ-10k (Hosu et al., 2020) dataset. The range of MOS is [1, 5], and higher MOS means better quality.

- We have verified the effectiveness of the proposed method through extensive experiments and compared with the state-of-the-arts. The experimental results demonstrate that the proposed model can achieve significantly better generalization performance.

2. RELATED WORK

Early no-reference IQA (NR-IQA) metrics typically train a regressor to obtain quality scores based on handcrafted features. For example, BLIINDS-II (Saad et al., 2012), BRISQUE (Mittal et al., 2012), and BIQI (Moorthy and Bovik, 2010) designed features meticulously through natural scene statistics (NSS). NFERM (Gu et al., 2014) incorporated features inspired by the free energy theory, human visual system, and NSS. CORNIA (Ye et al., 2012) and HOSA (Xu et al., 2016) trained large-scale visual codebooks from natural image to make predictions. The above handcrafted feature-based IQA models are usually limited in handling the diversified scenes and distortion types in real-world images.

With the boom of deep learning, convolutional neural networks have been widely applied in IQA. Early attempts utilized relatively shallow networks (Kang et al., 2014; Kang et al., 2015; Kottayil et al., 2016) to extract features for assessing synthetic distortions. Then, deeper networks were utilized to handle more complex distortions (Bosse et al., 2017; Kim and Lee, 2017; Ma et al., 2017; Yan et al., 2019; Zhai et al., 2020; Zhang J. et al., 2020). It is widely acknowledged that large datasets are needed for training deep neural networks. However, so far the largest IQA dataset only has 11,125 images, which are still limited. Thus, recent deep IQA metrics (Bianco et al., 2018; Varga et al., 2018; Zhang W. et al., 2020) utilize networks pre-trained on large-scale computer vision tasks and then fine-tune on them. For example, Bianco et al. (2018) made fine-tuning on the model pre-trained on subset of ImageNet (Imagenet large scale visual recognition challenge, 1.3M images) (Russakovsky et al., 2015) and Places-205 (Wang et al., 2015) (2.5M images). Varga et al. (2018) made fine-tuning on deep pre-trained network (ResNet101 He et al., 2016) to learn the distribution of mean opinion score (MOS). Zhang W. et al. (2020) utilized two different networks to evaluate synthetic and authentic distortions, respectively, and the authentic network was fine-tuned on pre-trained network (VGG16, Simonyan and Zisserman, 2015). Make fine-tuning on pre-trained model of recognition task is a suboptimal method because IQA task is different from recognition tasks. Recognition tasks should be robust to distortions while IQA should distinguish distortions. Though fine-tuning with IQA images can improve IQA performance, the generalizable features trained with large-scale dataset were damaged during further training. And due to the small sample property of IQA, generalization ability of new features is still unsatisfying and cannot be adopted to real-world applications.

Until recently, the generalization problem of IQA models began to receive attention. Zhu et al. (2020) adopted meta-learning to learn the prior knowledge of distortions in synthetic distortions and then fine-tune on authentic distortions to achieve better generalization ability. Hosu et al. (2020) built a large dataset (KonIQ-10k: 10,073 images) for model training and obtained better generalization performance. Su et al. (2020) incorporated semantic features and multi-scale content features to handle challenges of distortion diversity and content variation. The above methods have achieved better generalization performance than earlier metrics, but their generalization ability is still far from ideal and further explorations are needed. In this paper, we work toward this direction by proposing a new framework to address the generalization problem, where the intelligibility property of images is investigated.

3. PROPOSED METHOD

3.1. Relation Between Intelligibility and Quality

As aforementioned, image intelligibility can be described by semantic understanding tasks. The most popular one is the classification task on Imagenet Large Scale Visual Recognition Challenge, which has 1.3 M images belonging to 1,000 classes (Russakovsky et al., 2015). Therefore, we take the deep convolutional neural network (DCNN) trained on this task as an example. The output of the classification network is a probability distribution $o_i, i = 1, 2, ..., 1,000$, and $1,000$ is the total number of classes. The prediction confidence c can be obtained by

$$c = max(o_i), i = 1, 2, ..., 1000. \quad (1)$$

The confidence c in Equation (1) also represents the top1-probability. If the intelligibility of an image is high, the model may easily recognize the category and the top1-probability may be notably high. When the intelligibility is low, the model will be unconfident of its predictions and the top1-probability also tends to be low.

To have an intuitive understanding of the above characteristic, we compare the average classification confidence score obtained from images of different quality. First, we divide images from an IQA dataset into several groups according to their MOS values in ascending order. (Specifically, MOS are divided into 6 equal intervals of $[m_i, m_{i+1}]$ where $i = 1 - 6$, $m_1 = min(MOS)$, $m_7 = max(MOS)$.) Then, we utilize an image classification model trained on ImageNet to obtain the confidence score of images in each group. Finally, we calculate the average confidence score of each quality group, and illustrate them in **Figure 2A**. We can observe that images with poor quality tend to have lower prediction confidence than those with high quality. That is, image quality does have a significant impact on intelligibility.

In this paper, we are more interested in how intelligibility indicates image quality. Therefore, we do another experiment by dividing images according to the prediction confidence and compare the average MOS value of different confidence intervals. The results are presented in **Figure 2B**. Furthermore, to show the relation more intuitively, we also show sample images in **Figures 2C–H** that corresponds to the six ascending bins of **Figure 2B**. We can observe from **Figures 2B–H** that intelligibility

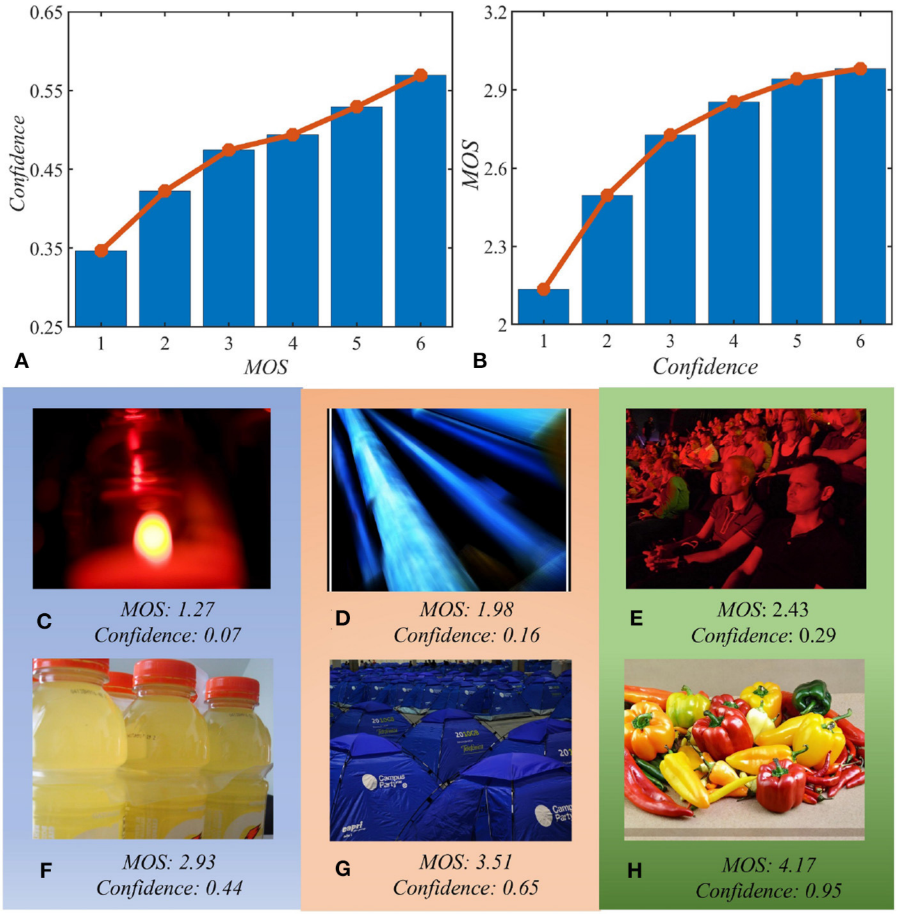

FIGURE 2 | Relation between recognition confidence and image quality. **(A)** Image quality affects recognition confidence; **(B)** recognition confidence indicates image quality; **(C–H)** representative images with different prediction confidence and mean opinion score (MOS). Panels **(C–H)** correspond to six ascending bins of B whose MOS increases with confidence. All results are obtained from the KonIQ-10k dataset based on EfficientNet-B0 network (Tan and Le, 2019).

described by image recognition task can distinctly indicate image quality.

3.2. Our Framework

In this paper, we propose an intelligibility enriched IQA (IE-IQA) framework, as illustrated in **Figure 3**. In our framework, we propose a bilateral network to integrate intelligibility features and conventional distortion features. Since intelligibility can be represented using different image understanding tasks, it is reasonable to utilize features from these tasks as intelligibility features. However, IQA is different from image understanding tasks, and directly utilizing features of image understanding tasks may lead to negative transfer, which has been proved by many transfer learning researches (Pan and Yang, 2010; Cao et al., 2018; Zhang J. et al., 2018). Since intelligibility is vital to our framework, utilizing features that are most relevant to intelligibility is a better way. Thus, we first propose a feature selection module to pick out more relevant features, and then fuse them with distortion features through an intelligibility enhanced module.

The distortion backbone with parameter θ^* in **Figure 3** is denoted as G_{θ^*}, which is adopted for extracting distortion features g_j from image I. The intelligibility backbone F_θ with parameter θ is adopted for extracting intelligibility features f_j. We select the most important features f_j' from f_j and then fuse them with distortion features g_j (denoted as $f_j' \Leftrightarrow g_j$) to obtain quality score q through a regressor $R_{\theta'}$. The whole process is explained as follows:

$$\begin{cases} f_j = F_\theta(I), \\ f_j^{'} \leftarrow f_j, \\ g_j = G_{\theta^*}(I), \\ q = R_{\theta'}(f_j' \Leftrightarrow g_j). \end{cases} \tag{2}$$

In this paper, four extensively studied semantic understanding tasks are utilized to obtain intelligibility features, including image recognition on subset of ImageNet (Russakovsky et al., 2015), scene classification on Places-365 (Zhou et al., 2017), object detection and instance segmentation on MS-COCO (Lin et al., 2014). In addition, we also utilize a relevant unrecognizability prediction task, which predicts the unrecognizable degree of an image. This task is trained on the VizWiz-QualityIssues dataset (Chiu et al., 2020), containing images with labels of the unrecognizable degree. Even if intelligibility features of heavily

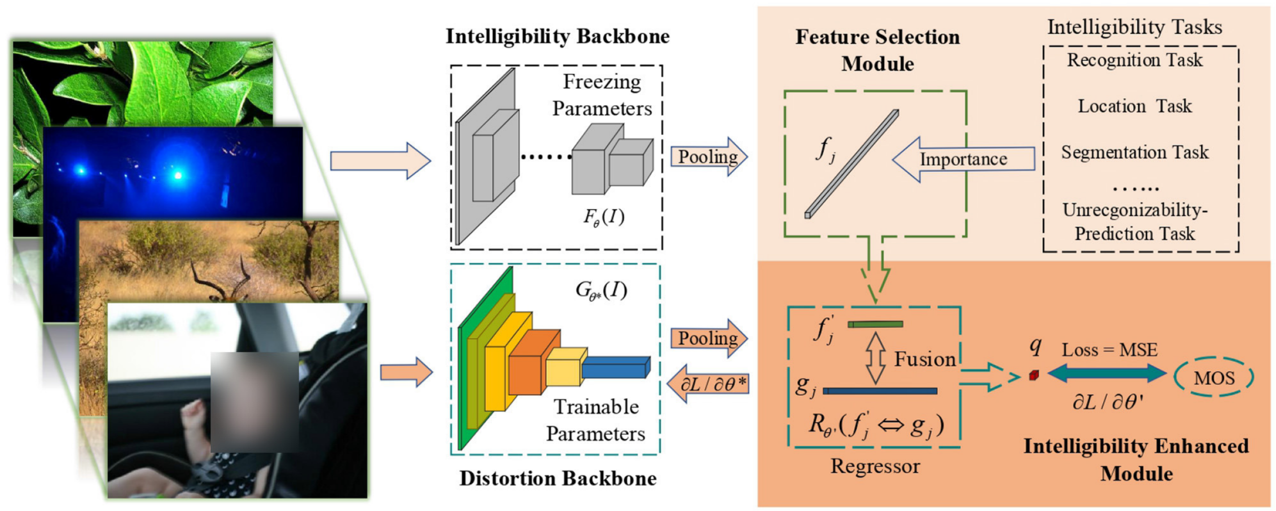

FIGURE 3 | Proposed framework of IE-IQA. Our framework contains intelligibility and distortion backbone, and colorful blocks in the distortion backbone are trainable while gray blocks in the intelligibility backbone are not trainable. An intelligibility enhanced module is adopted to fuse distortion features with intelligibility features obtained from the proposed feature selection module.

distorted images cannot obtain desired results in original tasks, they can still be distinguished from features of high-quality images, which is beneficial to the IQA task.

In our framework, the distortion backbone works in a data-driven manner to search for the best distortion features, and the intelligibility backbone is guaranteed to obtain features with high generalization ability and rich semantic information. To achieve these goals, we propose to freeze parameters θ of the intelligibility backbone during the training process while keeping parameters $\theta*$ in the distortion backbone trainable. On the one hand, the distortion network loads the pre-trained model trained on ImageNet. Though the pre-trained model has decent generalization ability, we still need to train the feature extractor with image quality data so that the network can adapt to IQA task and obtain better performance. Therefore, we make parameters of the distortion backbone trainable. On the other hand, training the intelligibility backbone may be problematic. High level features of image understanding tasks are rich in semantic information which is generalizable. If we train the intelligibility network using the IQA data, the generalization ability of intelligibility features (which are already generalizable) may be destroyed. Therefore, we freeze the intelligibility backbone to handle this problem.

In the proposed intelligibility enhanced module, we tried several feature fusion strategies: (1) utilize one/two/three fully connected (FC) layers to regress the quality score and fuse intelligibility features to different FC layer with add/multiply/concatenate operation; (2) utilize other layers to align intelligibility features with distortion features and then use other FC layers to regress the quality score; (3) utilize auxiliary layers and loss fuction to train intelligibility features along with strategy-(1) or strategy-(2); (4) replace low-dimensional features with sparse selected features (features that are not selected are set to zero) and then utilize strategy-(1) or strategy-(2). In implementation, we have found that these strategies achieve similar results. Due to the feature selection module, it is easy to combine lower dimensional intelligibility features and simple strategy can obtain satisfying results. The loss function we utilized is the mean square error (MSE).

3.3. Feature Selection

During the feature fusion process, we propose strategies to select intelligibility features. For a specific semantic understanding task, only a part of neural units and corresponding features in a DCNN are significantly activated during the inference process, while others are not vital to the final prediction and intelligibility (Hu et al., 2016; Zhang Q. et al., 2018; Zhou et al., 2019). Since introducing too many features are not conducive (even harmful in many transfer learning experiments) to IQA performance and generalization, we design feature selection strategies for different tasks based on contribution and sensitivity. Contribution-based strategy chooses features with greater contributions to predictions while sensitivity-based strategy chooses features that predictions are more sensitive to.

3.3.1. Contribution-Based Strategy

We propose to select features that have prominent contributions to final predictions. Theoretically, this strategy is not limited to any specified network as long as the network can be separated into a backbone and one FC layer. In fact, this kind of network architecture is very common in the image classification and scene recognition. Specifically, the output of backbone can be denoted by $f_j, j = 1, 2..., N_d$, where N_d is the dimension of features and the output-logits of the FC layer can be described by

$$z_i = \sum_{j=1}^{N_d} w_{ij} \times f_j + b_i, \tag{3}$$

where w_{ij}, b_i, z_i are weights, bias and logits of the FC layer, $i = 1, 2, ..., C$, and C is the number of total classes. The feature selection strategy is shown in **Algorithm 1**. In **Algorithm 1**, we locate the top1-probability first. Then, we calculate the contribution of each dimension of feature f_j by

$$contrb_j = abs(w_{i_{max},j} \times f_j). \tag{4}$$

Algorithm 1 | Feature selection strategy based on contributions.

```
Inputs: Output features of the backbone $f_j, j = 1, 2, ...., N_d$;
weights of the FC layer $w_{ij}$;
the number of total classes $C$; selected percentage $k$.
Output: The selected features $f_j^{'}$.
    // Obtain top1-probability index $i_{max}$:
        $i_{max} = argmax(z_i)$;
    // Calculate contributions of different features $contrb_j$:
        $contrb_j = abs(w_{i_{max},j} \times f_j)$;
    // Calculate the number of selected features $N_s$:
        $N_s = int(N_d \times k\%)$;
    // Sort $contrb_j$ in descending order and obtain index $ind_j$:
        $ind_j = argsort(contrb_j)$;
    // Select features of top-$N_s$ contributions:
        $f_j^{'} = f_j[sort(ind_j[1 : N_s])]$;
Return: $f_j^{'}$.
```

In Equation (4), the contributions of features are determined by both weights and activation values. Finally, features that contribute significantly to the top1-probability are selected.

3.3.2. Sensitivity-Based Strategy

Some networks have several non-linear FC layers and it is not easy to measure their contributions. Consider the unrecognizability prediction task for example. First, we train a model with the backbone of EfficientNet-B0 (Tan and Le, 2019) and three FC layers with RELU to regress the unrecognizability score. Then, we adopt a sensitivity-based method to select features and the sensitivity can be obtained by gradients. Specifically, the input feature is $f_j, j = 1, 2, ...N_d$ and the FC layers with active function are represented by function F. The unrecognizability score s can be obtained by

$$s = F(f_j). \tag{5}$$

The sensitivity of features can be described by

$$grad_j = abs(\partial s / \partial f_j). \tag{6}$$

Equation (6) represents the importance of features through partial derivatives, which is widely used in sensitivity analysis and model interpreting (Garson, 1991; Dimopoulos et al., 1995). After obtaining the importance of features, the selected number is calculated. Then, the index of sorted features can be obtained through

$$ind_j = argsort(grad_j), \tag{7}$$

where "*argsort*" means that sort the sequence and return corresponding index (it is the same with "*argsort*" in **Algorithm 1**). Finally, a selection operation is executed.

In contrast to directly merging all intelligibility features with distortion features, fusing features with lower dimension after feature selection exhibits better performance and generalization ability during the test process. Different from attention mechanism, the proposed feature selection strategy can reduce the dimension of the intelligibility feature and does not need any additional module or further training.

4. EXPERIMENTS

4.1. Datasets

In our experiments, five image quality datasets with authentic distortions are adopted, including KonIQ-10k (Hosu et al., 2020), Smartphone Photography Attribute and Quality (SPAQ) (Fang et al., 2020), LIVE in the Wild Image Quality Challenge (LIVEW) (Ghadiyaram and Bovik, 2016), CID2013 (Virtanen et al., 2015), and BID (Ciancio et al., 2011). Specifically, the KonIQ-10k dataset has 10,073 labeled images selected from a massive public database YFCC100M (Thomee et al., 2016), and the labels are obtained from 1.2 million ratings. The SPAQ dataset contains 11,125 labeled images obtained from 66 smartphones with exchangeable image file format data tags and rich opinion annotations. The annotations include MOS, attribute scores (such as brightness, noisiness, and sharpness) as well as scene category labels. LIVEW contains 1,162 labeled images and CID2013 contains 480 images from eight scenes. Different from the other four datasets, the BID dataset focuses on blur images and contains 586 images.

4.2. Implementation and Evaluation Protocol

In our experiments, the distortion network adopts the backbone of EfficientNet-B0 and the intelligibility network for the image recognition task is EfficientNet-B0 as well. EfficientNet-B0 consists of one convolutional layer followed by seven mobile inverted bottleneck modules, and then another convolutional layer followed by global average pooling. EfficientNet-B0 has an input size of 224 × 224 and 5.3 M parameters, and the dimension of its output feature is 1280. Network for scene classification task is ResNet-18 (He et al., 2016), and object detection is Faster-RCNN (Ren et al., 2017) with ResNet50-FPN (Lin et al., 2017) backbone. The instance segmentation task is DeeplabV3+ (Chen et al., 2018) with the backbone of ResNet101. During the training process, SGD optimizer with initial learning-rate 0.03 is utilized (we train FC layers first and then utilize

TABLE 1 | Pearson's linear correlation coefficient (PLCC)/Spearman's rank order correlation coefficient (SRCC) results of cross-dataset test.

PLCC/SRCC	KonIQ-10k	SPAQ	LIVEW	CID	BID
BIQI Moorthy and Bovik, 2010	0.637/0.595	0.622/0.661	0.492/0.471	0.612/0.599	0.478/0.493
NFERM Gu et al., 2014	0.725/0.689	0.697/0.711	0.551/0.540	0.708/0.680	0.529/0.530
BRISQUE Mittal et al., 2012	0.689/0.647	0.660/0.682	0.576/0.554	0.553/0.533	0.589/0.597
BLINDSII Saad et al., 2012	0.440/0.447	0.466/0.460	0.331/0.319	0.278/0.301	0.393/0.401
GWH-GLBP Li et al., 2016	0.549/0.514	0.614/0.628	0.464/0.435	0.071/0.002	0.477/0.483
FISBLIM Gu et al., 2013	0.375/0.347	0.566/0.569	0.376/0.289	-0.219/-0.234	0.392/0.344
CORNIA Ye et al., 2012	0.773/0.738	0.727/0.766	0.672/0.639	0.599/0.538	0.692/0.688
HOSA Xu et al., 2016	0.791/0.761	0.743/0.771	0.677/0.652	0.684/0.664	0.694/0.679
NSSADNN Yan et al., 2019	/	/	0.813/0.745*	0.825/0.748*	/
MEON Ma et al., 2017	/	/	0.693/0.688*	0.703/0.701*	/
BIECON Kim and Lee, 2017	/	/	0.613/0.595*	0.620/0.606*	/
DeepRN (ResNet101) Varga et al., 2018	0.880/0.867	/	0.750/0.726	/	/
DeepBIQ (InceptionV2) Bianco et al., 2018	0.911/0.907	/	0.821/0.804	/	/
HyperNet Su et al., 2020	0.917/**0.906**	0.843/0.846$^+$	NA/0.785	0.808/0.782$^+$	NA/**0.819**
MetaIQA Zhu et al., 2020	0.876/0.846	0.804/0.822	0.748/0.716	0.726/0.682	0.740/0.738
WaDIQaM-NR Bosse et al., 2017	0.657/0.631	0.675/0.702	0.521/0.523	0.584/0.495	0.499/0.538
DBCNN Zhang W. et al., 2020	0.892/0.868	0.827/0.836	0.802/0.775	0.788/0.758	0.769/0.769
		Our Results			
IE-IQA (*w/ recognition task*)	**0.921**/0.900	**0.863/0.859**	**0.839/0.829**	0.815/0.788	**0.822**/0.817
IE-IQA (*w/ classification task*)	0.920/0.900	0.862/0.858	0.835/0.828	0.818/0.795	0.819/0.813
IE-IQA (*w/ detection task*)	**0.921**/0.901	0.862/0.857	0.835/0.826	0.819/0.800	0.816/0.810
IE-IQA (*w/ segmentation task*)	0.917/0.900	0.862/0.857	0.825/0.826	**0.827/0.801**	0.812/0.809
IE-IQA (*w/ unrecognization task*)	0.920/ 0.902	**0.863**/0.858	0.835/**0.829**	0.819/0.794	0.816/0.813

The model is trained on 80% images of KonIQ-10k and directly tested on rest 20% KonIQ-10k images and other datasets. Results with "" are obtained after fine-tuning on the dataset and reported in original papers. Results with NA of HyperNet (only three datasets) are reported in original papers (Su et al., 2020) and results with "+" are obtained from the released model. Best results are in bold.*

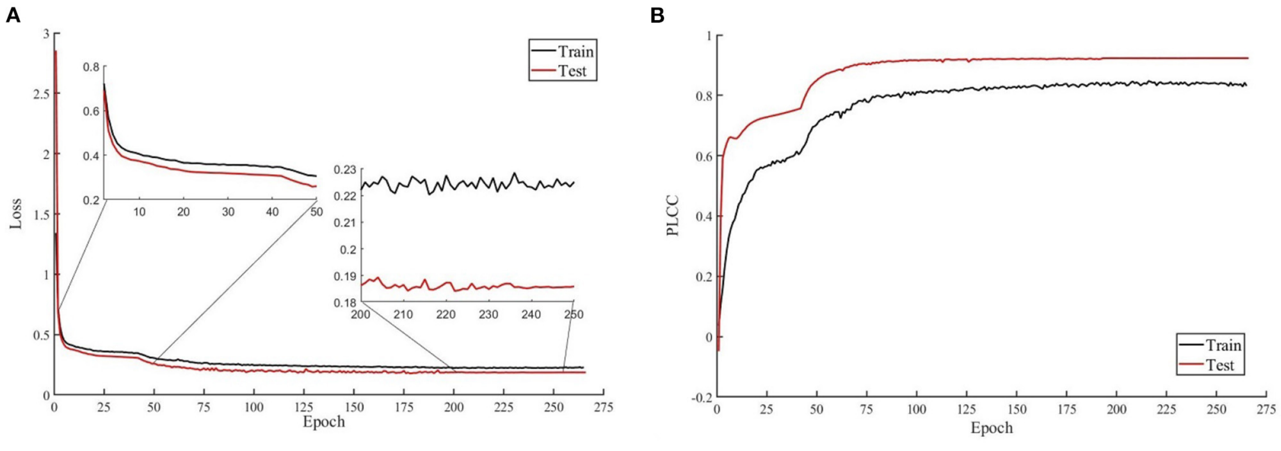

FIGURE 4 | Loss and Pearson's linear correlation coefficient (PLCC) during training and test. **(A)** Loss of training and test. Two enlarged subfigures shows results of epochs 2–50 and epochs 200–250. **(B)** PLCC of training and test. The model is trained with recognition task on KonIQ-10k.

warm-up strategy when training the distortion backbone). For all of our experiments, we first resize images into 244 × 244, then we randomly crop them to 224 × 224 with a randomly horizontal flip to augment training images. During the test process, we directly resize test images into 224 × 224 and then predict once, which is more efficient in real applications. We tried different selection ratios of 1, 5, 20, and 50%. The final selection ratio of the recognition task, class task, detection task,

TABLE 2 | Pearson's linear correlation coefficient (PLCC)/Spearman's rank order correlation coefficient (SRCC) results of cross-dataset test.

PLCC/SRCC	SPAQ	KonIQ-10k	LIVEW	CID	BID
NFERM Gu et al., 2014	0.832/0.823	0.455/0.447	0.591/0.542	0.437/0.342	0.578/0.570
BRSIQUE Mittal et al., 2012	0.833/0.822	0.446/0.433	0.593/0.553	0.499/0.504	0.589/0.578
CORNIA Ye et al., 2012	0.867/0.859	0.532/0.516	0.663/0.621	0.552/0.465	0.676/0.673
HOSA Xu et al., 2016	0.873/0.866	0.559/0.534	0.682/0.650	0.593/0.536	0.681/0.670
Baseline Fang et al., 2020	0.909/0.908	0.532/0.523^{+}	0.564/0.517^{+}	0.518/0.569^{+}	0.574/0.566^{+}
MT-S Fang et al., 2020	**0.921/0.917**	0.486/0.485^{+}	0.539/0.493^{+}	0.342/0.389^{+}	0.530/0.529^{+}
HyperNet Su et al., 2020	0.917/0.915	0.679/0.645	0.695/0.680	0.624/0.585	0.648/0.647
MetalQA Zhu et al., 2020	0.871/0.870	0.722/0.686	0.765/0.731	0.737/0.695	0.743/0.735
		Our Results			
IE-IQA (*w/ recognition task***)**	0.918/0.913	0.768/0.710	0.779/0.764	0.743/0.713	0.744/0.742
IE-IQA (*w/ classification task***)**	0.917/0.915	0.761/0.720	0.764/0.758	0.737/0.702	0.737/0.737
IE-IQA (*w/ detection task***)**	0.920/0.916	**0.777/0.728**	**0.782/0.772**	0.742/0.702	**0.748/0.749**
IE-IQA (*w/ segmentation task***)**	0.918/0.914	0.775/0.724	0.781/0.768	**0.752/0.737**	0.744/0.746
IE-IQA (*w/ unrecognization task***)**	0.920/0.916	0.770/0.721	0.774/0.764	**0.752**/0.725	0.747/0.746

The model is trained on 80% images of Smartphone Photography Attribute and Quality (SPAQ) and directly tested on rest 20% SPAQ images and other datasets. Results with "+" are obtained from the released model. HyperNet are retrained with image size of 244 × 244. Best results are in bold.

TABLE 3 | Pearson's linear correlation coefficient (PLCC)/Spearman's rank order correlation coefficient (SRCC) results on intra-dataset tests.

Dataset	KonIQ-10k	SPAQ	LIVEW	CID2013	RBID
NFERM Gu et al., 2014	0.725/0.689	0.832/0.823	0.562 /0.517	0.825/0.823	0.585/0.559
BRISQUE Mittal et al., 2012	0.689/0.647	0.833/0.822	0.574/0.557	0.810/0.814	0.617/0.594
CORNIA Ye et al., 2012	0.773/0.738	0.867/0.859	0.692/0.655	0.822/0.803	0.712/0.695
HOSA Xu et al., 2016	0.791/0.761	0.873/0.866	0.703/0.667	0.835/0.833	0.716/0.684
NSSADNN Yan et al., 2019	/	/	0.813*/0.745*	0.825*/0.748*	/
MEON Ma et al., 2017	/	/	0.693*/0.688*	0.703* / 0.701*	/
BIECON Kim and Lee, 2017	/	/	0.613*/0.595*	0.620*/0.606*	/
Baseline Fang et al., 2020	0.908/0.889	0.909*/0.908*	0.825/0.794	0.876/0.881	0.802/0.794
WaDIQaM-NR Bosse et al., 2017	0.805*/0.797*	/	0.680*/0.671*	0.729*/0.708*	0.742*/0.725*
HyperNet Su et al., 2020	0.917*/**0.906***	0.914/0.909	**0.882*/0.859***	/	**0.878*/0.869***
DBCNN Zhang W. et al., 2020	0.892/0.868	0.915*/0.911*	0.869*/0.851*	/	0.859*/0.845*
MetalQA Zhu et al., 2020	0.887*/0.850*	0.871/0.870	0.835*/0.802*	0.784*/0.766*	0.777/0.746
IE-IQA (*w/ recognition task***)**	**0.921**/0.900	**0.918/0.913**	0.868/0.838	**0.934/0.934**	0.838/0.837

Results with ∗ are obtained from published papers. Other results are obtained from retrained model. Best results are marked in bold.

segmentation task, and unrecognization task are 5, 5, 20, 50, and 50%, respectively.

Our evaluation criteria are two widely used correlation coefficients: Pearson's linear correlation coefficient (PLCC) and Spearman's rank order correlation coefficient (SRCC).

4.3. Performance Comparison

This paper aims to propose a highly generalizable NR-IQA model, thus we train our model in one dataset and then test on other datasets directly without doing any fine-tuning. For comparison, we also re-train some popular handcrafted feature-based methods, such as BRISQUE, CORNIA, HOSA, and deep learning-based methods, including DBCNN (Zhang W. et al., 2020), MetaIQA (Zhu et al., 2020), and WaDIQaM-NR (Bosse et al., 2017) (codes are publically available) with the same setting. All results trained on KonIQ-10k are shown in **Table 1**. The middle group in **Table 1** shows deep learning-based methods, and the results of methods without public codes are obtained from the original papers. The bottom group shows our results.

From **Table 1**, we can observe that our framework with five intelligibility tasks can consistently achieve the best cross-dataset performance for most cases. It should be emphasized that our models are only trained with KonIQ-10k (80% images) and directly tested on other datasets without any fine-tuning. Though NSSADNN, MEON, and BIECON made fine-tuning on the target dataset, our generalization performance can still maintain a significant advantage.

Efficient-B0 has 5.3M parameters, which is less than ResNet18 (11.7 M parameters, the backbone of MetaIQA), ResNet50 (26

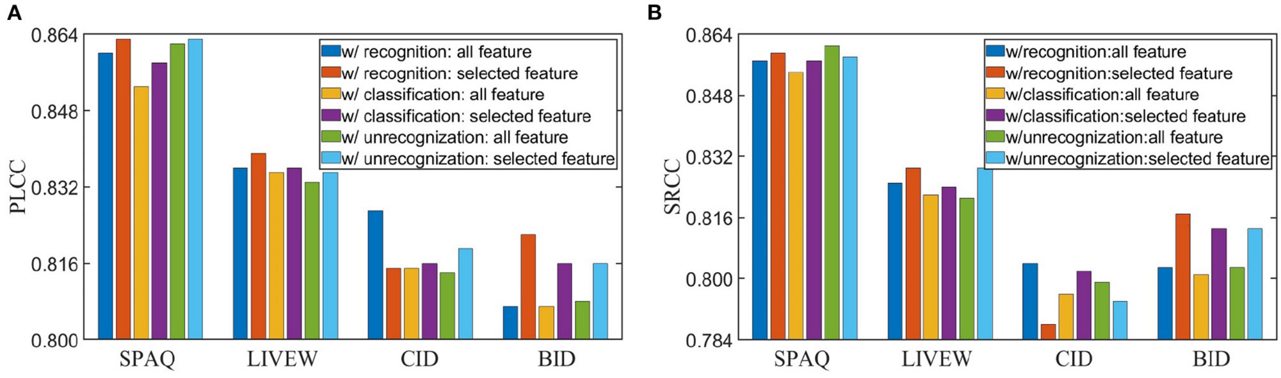

FIGURE 5 | Performance comparison for different tasks with/without feature selection strategy on KonIQ-10k. **(A)** Pearson's linear correlation coefficient (PLCC) results. **(B)** Spearman's rank order correlation coefficient (SRCC) results. The recognition and classification task utilize contribution-based strategy, and the unrecognizability task utilizes gradient-based strategy.

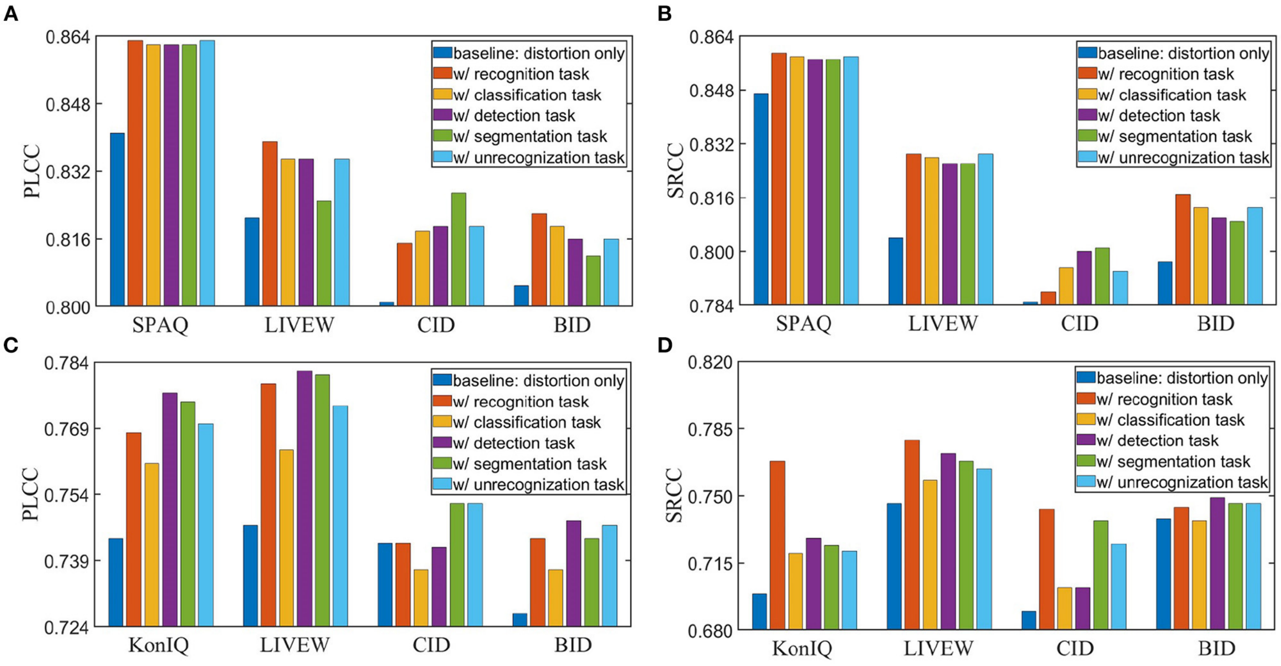

FIGURE 6 | Ablation study of intelligibility enriched IQA (IE-IQA). **(A)** Pearson's linear correlation coefficient (PLCC) results of models trained with 80% KonIQ-10k. **(B)** Spearman's rank order correlation coefficient (SRCC) results of models trained with 80% KonIQ-10k. **(C)** PLCC results of models trained with 80% Smartphone Photography Attribute and Quality (SPAQ). **(D)** SRCC results of models trained with 80% SPAQ.

M parameters, the backbone of HyperNet), and ResNet101 (44.5 M parameters, the backbone of DeepRN). Efficient-B0 is easy to converge, and we show the loss and PLCC results during training and test in **Figure 4**. We can observe from **Figure 4** that the test loss decreases with the training loss and the test performance increases with training performance. This means that the network is trained without overfitting.

To make a further comparison, we also train our methods on SPAQ and perform cross-dataset tests on the other four datasets. The results are shown in **Table 2**.

The model "Baseline" in Fang et al. (2020) means the baseline model (ResNet50) and "MT-S" means the model jointly trained with MOS and scene labels (The SPAQ dataset has scene category labels). We can observe that compared to MT-S, our method can achieve similar performance on the training dataset. However, by combining intelligibility features, the generalization performance of the proposed method is apparently much better.

Comparing **Table 2** with **Table 1**, we can observe that models trained on KonIQ-10k have better cross-dataset performance. One possible reason is the source of images. The SPAQ dataset

TABLE 4 | Results of training the distortion network from scratch on 80% KonIQ-10k.

PLCC	KonIQ-10k(20%)	SPAQ	LIVEW	CID	BID
Only distortion	0.784	0.756	0.638	0.676	0.645
W/recognition	0.814	**0.812**	**0.689**	**0.714**	**0.706**
W/classification	0.814	0.763	0.653	0.695	0.659
W/detection	0.812	0.740	0.644	0.682	0.640
W/segmentation	**0.826**	0.758	0.676	0.688	0.684
W/unrecognization	0.811	0.749	0.651	0.691	0.666

Best results are in bold.

TABLE 5 | Results of training the distortion network from scratch on 80% Smartphone Photography Attribute and Quality (SPAQ).

PLCC	SPAQ(20%)	KonIQ-10k	LIVEW	CID	BID
only distortion	0.878	0.568	0.605	0.665	0.598
w/recognition	0.883	0.591	**0.628**	**0.702**	0.628
w/classification	**0.884**	0.585	0.625	0.695	**0.631**
w/detection	0.882	**0.598**	0.626	0.698	0.623
w/segmentation	0.883	0.592	0.627	0.697	0.629
w/unrecognization	0.881	0.585	0.621	0.686	0.615

Best results are in bold.

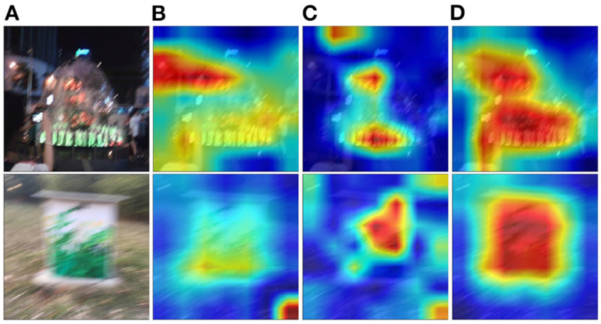

FIGURE 7 | Visualization results of Grad-CAM. **(A)** Original images; **(B)** heat-maps of the baseline model; **(C)** heat-maps of the intelligibility network; **(D)** heat-maps of proposed model with image recognition task.

is obtained from smartphones only, while the image sources of KonIQ-10k are more diversified. Another possible reason is that the image size of the SPAQ dataset is very large (4000 × 3000 is common) and our model has an input size of 224 × 224. Small size input may lose much information and the interpolation algorithm may bring new distortions.

Another phenomenon observed from **Tables 1**, **2** is that the proposed method achieves slightly worse generalization performance on the BID/CID databases than the other three datasets. The BID dataset focuses on blur images and the CID dataset consists of limited scenes of images (eight scenes). This may lead to a more pronounced distribution discrepancy between CID/BID and the training datasets.

Though our metric aims to achieve high generalization ability, we still make further experiments on intra-dataset tests. The results are listed in **Table 3**. We can summarize from **Table 3** that our metric can achieve state-of-the-arts intra-dataset performance. Though HyperNet achieves better performance for some cases, it needs to evaluate crop 25 patches during evaluating, costing much more time than the proposed metric. For example, when evaluating 1,000 images with the resolution of 1024 × 768 (batchsize = 1, using one TITANXp GPU and Intel Xeon E5-2630V4 CPU), HyperNet costs 2,040 s, while the proposed metric only costs 84 s.

To explore how feature selection strategy affects prediction results, we make a comparison of the results with/without feature selection strategy, and show them in **Figure 5**. The results show that removing noisy features and utilizing features having significant influence on final predictions tend to achieve higher performance and better generalization ability with only one exception (recognition task) on the CID dataset. One possible reason is that the CID dataset has only eight specific scenes and many images in CID contain the same objects. In this situation, selected features may not provide rich distinguished information for evaluating quality of images with similar contents.

To demonstrate the effectiveness of intelligibility features, we make ablation studies and show the results in **Figure 6**. The baseline means the model with distortion backbone alone. From **Figure 6**, we can observe that intelligibility features do improve both performance and generalization ability. Therefore, it is necessary to combine both intelligibility aspect and distortion aspect in IQA metrics.

During the training process, the distortion network loads the pre-trained model, and some semantic information and intelligibility features may have already existed in the pre-trained model. To further investigate the effects of original intelligibility features on the distortion network, we train the distortion network from scratch. Then we fuse the intelligibility network with the distortion network. The results are shown in **Tables 4**, **5**. From the tables, we can observe that the introduced intelligibility network still benefits the performance of the whole framework even the distortion network is not pre-trained.

To explore how intelligibility affects quality assessment results intuitively, we utilize the method of Grad-CAM (Selvaraju et al., 2017) to investigate which area of an image affects the prediction most. Examples are shown in **Figure 7**, where red areas have more conspicuous influence to the prediction than blue areas. As shown in **Figure 7**, the intelligibility features do play an important role in the quality assessment. The baseline model with distortion network only (**Figure 6B**) cannot effectively locate salient objects which people may pay attention to. The intelligibility features (**Figure 6C**) alone mainly focus on relatively local regions and cannot well utilize global information of images. In contrast, the proposed model (**Figure 6D**) not only meticulously locate salient objects (important for intelligibility), but also pay more attention to wider areas, which catches global information. It is widely acknowledged that both global and local information are vital to IQA metrics (Fang et al., 2018); hence, from this point of view, it is not hard to understand that by combining the intelligibility features, our model can achieve better performance.

5. CONCLUSIONS

In this paper, we first analyzed the relation between intelligibility and image quality. The results reveal that intelligibility is indicative of image quality. Therefore, we proposed a new framework, i.e., Intelligibility-Enriched-IQA, to combine intelligibility with conventional distortion measure. Feature selection strategy was proposed to select the most important intelligibility features, which alleviates negative transfer and avoids damaging highly generalizable features. Extensive experimental results show the effectiveness of proposed method, and our model achieves state-of-the-art performance in terms of the generalization ability. These results demonstrate that introducing intelligibility is a promising way in building highly generalizable IQA metrics.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

TS and LL contributed to conception and design of the study. TS performed the experiment and wrote the first draft of the manuscript. LL, HZ, and JQ wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grants 62171340, 61771473, 61991451, and 61379143, the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024, the Fundamental Research Funds for the Central Universities under Grant JBF211902, the Science and Technology Plan of Xi’an under Grant 20191122015KYPT011JC013, the Natural Science Foundation of Jiangsu Province under Grants BK20181354 and BK20200649, and the Six Talent Peaks High-level Talents in Jiangsu Province under Grant XYDXX-063.

REFERENCES

Abdou, I. E., and Dusaussoy, N. J. (1986). “Survey of image quality measurements,” in *Proceedings of 1986 ACM Fall Joint Computer Conference, ACM ’86* (Washington, DC: IEEE Computer Society Press), 71–78.

Bianco, S., Celona, L., Napoletano, P., and Schettini, R. (2018). On the use of deep learning for blind image quality assessment. *Signal Image Video Process*. 12, 355–362. doi: 10.1007/s11760-017-1166-8

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2017). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process*. 27, 206–219. doi: 10.1109/TIP.2017.2760518

Cao, Z., Long, M., Wang, J., and Jordan, M. I. (2018). “Partial transfer learning with selective adversarial networks,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT), 2724–2732. doi: 10.1109/CVPR.2018.00288

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *2018 Proceedings of European Conference on Computer Vision (ECCV)* (Munich), 833–851. doi: 10.1007/978-3-030-01234-2_49

Chiu, T., Zhao, Y., and Gurari, D. (2020). “Assessing image quality issues for real-world problems,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3643–3653. doi: 10.1109/CVPR42600.2020.00370

Ciancio, A., Targino da Costa, A. L. N. T., da Silva, E. A. B., Said, A., Samadani, R., and Obrador, P. (2011). No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process*. 20, 64–75. doi: 10.1109/TIP.2010.2053549

Dimopoulos, Y., Bourret, P., and Lek, S. (1995). Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Process. Lett*. 2, 1–4. doi: 10.1007/BF02309007

Fang, Y., Yan, J., Li, L., Wu, J., and Lin, W. (2018). No reference quality assessment for screen content images with both local and global feature representation. *IEEE Trans. Image Process*. 27, 1600–1610. doi: 10.1109/TIP.2017.2781307

Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. (2020). “Perceptual quality assessment of smartphone photography,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3674–3683. doi: 10.1109/CVPR42600.2020.00373

Garson, G. D. (1991). Interpreting neural-network connection weights. *AI Expert* 6, 46–51.

Ghadiyaram, D., and Bovik, A. C. (2016). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process*. 25, 372–387. doi: 10.1109/TIP.2015.2500021

Gu, K., Zhai, G., Liu, M., Yang, X., Zhang, W., Sun, X., et al. (2013). “FISBLIM: a five-step blind metric for quality assessment of multiply distorted images,” in *SiPS 2013 Proceedings* (Taipei), 241–246. doi: 10.1109/SiPS.2013.6674512

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2014). Using free energy principle for blind image quality assessment. *IEEE Trans. Multimedia* 17, 50–63. doi: 10.1109/TMM.2014.2373812

He, K., Zhang, X., Ren, S., and Sun, J. (2016). “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. (2020). KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process*. 29, 4041–4056. doi: 10.1109/TIP.2020.2967829

Hu, H., Peng, R., Tai, Y.-W., and Tang, C.-K. (2016). Network trimming: a data-driven neuron pruning approach towards efficient deep architectures. *arXiv [Preprint]* arXiv: 1607.03250.

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). “Convolutional neural networks for no-reference image quality assessment,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Columbus, OH), 1733–1740. doi: 10.1109/CVPR.2014.224

Kang, L., Ye, P., Li, Y., and Doermann, D. (2015). “Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks,” in *2015 IEEE International Conference on Image Processing (ICIP)* (Quebec City, QC), 2791–2795. doi: 10.1109/ICIP.2015.7351311

Kim, J., and Lee, S. (2017). Fully deep blind image quality predictor. *IEEE J. Select. Top. Signal Process*. 11, 206–220. doi: 10.1109/JSTSP.2016.2639328

Kottayil, N. K., Cheng, I., Dufaux, F., and Basu, A. (2016). A color intensity invariant low-level feature optimization framework for image quality assessment. *Signal Image Video Process*. 10, 1169–1176. doi: 10.1007/s11760-016-0873-x

Li, Q., Lin, W., and Fang, Y. (2016). No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Process. Lett*. 23, 541–545. doi: 10.1109/LSP.2016.2537321

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). “Feature pyramid networks for object detection,” in *2017 IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI), 936–944. doi: 10.1109/CVPR.2017.106

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft COCO: common objects in context," in *2014 Proceedings of European Conference on Computer Vision (ECCV)* (Zurich), 740–755. doi: 10.1007/978-3-319-10602-1_48

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. (2017). End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Process*. 27, 1202–1213. doi: 10.1109/TIP.2017.2774045

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process*. 21, 4695–4708. doi: 10.1109/TIP.2012.2214050

Moorthy, A. K., and Bovik, A. C. (2010). A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett*. 17, 513–516. doi: 10.1109/LSP.2010.2043888

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowledge Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell*. 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Saad, M. A., Bovik, A. C., and Charrier, C. (2012). Blind image quality assessment: a natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process*. 21, 3339–3352. doi: 10.1109/TIP.2012.2191563

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice), 618–626. doi: 10.1109/ICCV.2017.74

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *2015 International Conference on Learning Representations (ICLR)* (San Diego, CA).

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., et al. (2020). "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3664–3673. doi: 10.1109/CVPR42600.2020.00372

Tan, M., and Le, Q. V. (2019). EfficientNet: rethinking model scaling for convolutional neural networks. *arXiv [preprint]* arXiv: 1905.11946.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. (2016). YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 64–73. doi: 10.1145/2812802

Varga, D., Saupe, D., and Sziranyi, T. (2018). "DeepRN: a content preserving deep architecture for blind image quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)* (San Diego, CA), 1–6. doi: 10.1109/ICME.2018.8486528

Virtanen, T., Nuutinen, M., Vaahteranoksa, M., Oittinen, P., and Hakkinen, J. (2015). CID2013: a database for evaluating no-reference image quality assessment algorithms. *IEEE Trans. Image Process*. 24, 390–402. doi: 10.1109/TIP.2014.2378061

Wang, L., Guo, S., Huang, W., and Qiao, Y. (2015). Places205-vggnet models for scene recognition. *arXiv [Preprint]* arXiv: 1508.01667.

Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., and Doermann, D. (2016). Blind image quality assessment based on high order statistics aggregation. *IEEE Trans. Image Process*. 25, 4444–4457. doi: 10.1109/TIP.2016.2585880

Yan, B., Bare, B., and Tan, W. (2019). Naturalness-aware deep no-reference image quality assessment. *IEEE Trans. Multimedia* 21, 2603–2615. doi: 10.1109/TMM.2019.2904879

Ye, P., Kumar, J., Kang, L., and Doermann, D. (2012). "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Providence, RI), 1098–1105.

Zhai, G., Zhu, Y., and Min, X. (2020). Comparative perceptual assessment of visual signals using free energy features. *IEEE Trans. Multimedia*. doi: 10.1109/TMM.2020.3029891. [Epub ahead of print].

Zhang, J., Ding, Z., Li, W., and Ogunbona, P. (2018). "Importance weighted adversarial nets for partial domain adaptation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT), 8156–8164. doi: 10.1109/CVPR.2018.00851

Zhang, J., Min, X., Zhu, Y., Zhai, G., Zhou, J., Yang, X., et al. (2020). Hazdesnet: an end-to-end network for haze density prediction. *IEEE Trans. Intell. Transport. Syst*. doi: 10.1109/TITS.2020.3030673. [Epub ahead of print].

Zhang, Q., Wu, Y. N., and Zhu, S.-C. (2018). "Interpretable convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8827–8836. doi: 10.1109/CVPR.2018.00920

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. (2020). Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circ. Syst. Video Technol*. 30, 36–47. doi: 10.1109/TCSVT.2018.2886771

Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2019). Interpreting deep visual representations via network dissection. *IEEE Trans. Pattern Anal. Mach. Intell*. 41, 2131–2145. doi: 10.1109/TPAMI.2018.2858759

Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., and Oliva, A. (2017). Places: an image database for deep scene understanding. *J. Vis*. 17:296. doi: 10.1167/17.10.296

Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. (2020). "MetaIQA: deep meta-learning for no-reference image quality assessment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14131–14140. doi: 10.1109/CVPR42600.2020.01415

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 28 October 2021
doi: 10.3389/fnins.2021.766396

Learning to Predict Page View on College Official Accounts With Quality-Aware Features

Yibing Yu[1,2]*, Shuang Shi[3], Yifei Wang[3], Xinkang Lian[3], Jing Liu[3] and Fei Lei[3]

[1] The Communist Youth League Committee, Beijing University of Technology, Beijing, China, [2] School of Economics and Management, Beijing University of Technology, Beijing, China, [3] Faculty of Information Technology, Beijing University of Technology, Beijing, China

At present, most of departments in colleges have their own official accounts, which have become the primary channel for announcements and news. In the official accounts, the popularity of articles is influenced by many different factors, such as the content of articles, the aesthetics of the layout, and so on. This paper mainly studies how to learn a computational model for predicting page view on college official accounts with quality-aware features extracted from pictures. First, we built a new picture database by collecting 1,000 pictures from the official accounts of nine well-known universities in the city of Beijing. Then, we proposed a new model for predicting page view by using a selective ensemble technology to fuse three sets of quality-aware features that could represent how a picture looks. Experimental results show that the proposed model has achieved competitive performance against state-of-the-art relevant models on the task for inferring page view from pictures on college official accounts.

Keywords: page view, quality-aware features, selective ensemble, human visual system, college official accounts

OPEN ACCESS

Edited by:
Xiongkuo Min,
University of Texas at Austin,
United States

Reviewed by:
Yutao Liu,
Tsinghua University, China
Weiling Chen,
Fuzhou University, China
Qiuping Jiang,
Ningbo University, China

****Correspondence:***
Yibing Yu
yuyibing@bjut.edu.cn

Specialty section:
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

Received: *29 August 2021*
Accepted: *27 September 2021*
Published: *28 October 2021*

Citation:
Yu Y, Shi S, Wang Y, Lian X, Liu J and Lei F (2021) Learning to Predict Page View on College Official Accounts With Quality-Aware Features.
Front. Neurosci. 15:766396.
doi: 10.3389/fnins.2021.766396

1. INTRODUCTION

With the popularization and development of the Internet, the official accounts have attracted extensive attention. The majority of college departments now own accounts because it has become the main channel for publishing notices and posting news. Page view is a very significant indicator for college official accounts, capable of visually showing the popularity of an article. If we can predict the page views, it is of great help to improve the attention of audience for articles. The number of views on articles is influenced by the content of pictures. To this end, we explore the quality-aware features of pictures and attempt to predict page views in the official accounts based on image processing technology in this paper.

In recent years, with the development of image processing technology, there are many great contributions in multimedia telecommunication domain (Geng et al., 2011; Kang et al., 2019; Moroz et al., 2019; Su et al., 2019; Wu et al., 2019; Yildirim, 2019), education and teaching (Richard, 1991; Greenberg et al., 1994; Rajashekar et al., 2002; Yaman and Karakose, 2016), and environmental perception and protection, such as air pollution detection (Gu et al., 2020a,c, 2021b; Liu et al., 2021), $PM_{2.5}$ monitoring (Gu et al., 2019, 2021a), air quality forecast (Gu et al., 2018, 2020b), and distance education (Zheng et al., 2009). Among them, picture quality assessment (PQA) has been receiving a lot of attention as an important part of image processing technology. With a variety of PQA models available from Wang et al. (2004), how to achieve evaluation results that are consistent with the subjective PQA of human beings is crucial. Usually, subjective experiments are performed by human observers who score the pictures, and the final reliable results obtained are taken as the ground truth (Gu et al., 2014, 2015a). However, the method mentioned above is time consuming and complicated, so the focus of relevant scientific research has shifted to the design of objective PQA algorithms implemented by computers. The objective PQA algorithm has the characteristics of convenience, high-speed, repeatable, batch processing, and real-time, which make up for the deficiency of the subjective PQA method.

The objective PQA approach establishes a mathematical model that is combined with the subjective human visual system (HVS) to realize the evaluation of picture quality. According to the amount of information provided by the reference picture, the existing objective PQA methods can be divided into: full reference (FR) PQA method, reduced reference (RR) PQA method, and no reference (NR) PQA method. Among them, the FR PQA method is the most reliable and technically mature evaluation method. It has a complete original picture and allows a one-to-one correspondence comparison of the distorted picture with the pixels of the original picture. Instead, RR PQA method requires only partial original picture information, researchers like Liang and Weller (2016) and Wu et al. (2013) put forward a series of novel RR PQA algorithms. The FR PQA algorithm and the RR PQA algorithm combine the visual features of the picture to quantify the difference between the original picture and the distorted picture, so as to get the quality of pictures.

In official accounts, the original picture information is not available, so it is particularly important to propose PQA algorithm. Most of the current NR PQA methods were proposed based on two steps, which are feature extraction proposed by Gu et al. (2017b) and the support vector machine (SVM) proposed by Smola and Schölkopf (2004) that can find out the underlying relationship between the selected features and human subjective evaluations. No reference method is a situation where none of the information contained in any reference picture or video is used to draw quality conclusions. Since the picture is not available in most cases, more and more metrics were proposed for NR PQA method. Nowadays, the advanced method (e.g., BRISQUE) is a universal blind PQA model based on Natural Scene Statistics (NSS) proposed by Mittal et al. (2012). Natural scene pictures belong to a small domain of Internet picture signals that follow predictable statistical laws. Specifically, the natural scene pictures captured by high-quality devices obey the Gaussian-like distribution, while the pictures with distortion (such as blur, noise, watermarks, color transformation, etc.) do not follow the Bell curve law. Based on this theory, the features of NSS can be used as an effective and robust natural PQA tool. In recent years, a large number of studies based on NSS have been carried out, such as the MSDDs presented by Jiang et al. (2018), Bliinds-II constructed by Saad et al. (2012), BLIQUE-TMI created by Jiang et al. (2019b), GMLF designed by Xue et al. (2014), and DIIVINE presented by Moorthy and Bovik (2011), which is capable of assessing the quality of distorted pictures across multiple distortion categories, etc. In addition, Ruderman (1994) investigated the data rules of natural pictures, which provides a basis for evaluating the perceptual quality of pictures. The local features of pictures can perfectly reflect the perceptual quality of pictures.

Due to the fact that most of the audiences get the information from official accounts from vision, we also introduce into the approach based on the HVS. Advances in brain science and neuroscience studied by Friston et al. (2006) have encouraged scholars to explore new fields of machine vision. Eye movement research is also of significance to the visual perception of brain science. Jiang et al. (2019a), Kim et al. (2019), Lin et al. (2019), Tang et al. (2020), Zhang et al. (2020), Jiang et al. (2021), Wang et al. (2021) had carried out a lot of research work. Brain science research have shown that the brain produces an intrinsic model to explain the process of perception and understanding, and that the free energy generated during this cognitive process can reflect the difference between picture signals and internal descriptions. By modeling important physiological and psychological visual features, Xu et al. (2016) discussed the mechanism related to free energy in the human brain and proposed an efficient PQA method by using JPEG and JPEG2000 compression, Jiang et al. (2020) presented a new FR-SIQM method by measuring and fusing the degradations on hierarchical features. Besides, Gu et al. (2015b) designed the NFSDM in an alternative way of extracting features. On the basis of the NFSDM approach, the NFERM is combined with HVS to reduce the number of extracted by half, further improving the accuracy of the evaluation.

Based on image processing technology, this paper investigates a large collection of quality-aware features of pictures to predict the page view that reflects the popularity of articles. To accomplish this goal, the authors do a lot of work to collect the pictures published by the WeChat official accounts of nine universities in Beijing in recent months, and establish a new picture database consisting of 1,000 pictures. In addition, we collect three groups of features from the Official Accounts Picture Quality Database (OAPQD) and use the selective ensemble technique proposed for NSS, HVS, and histogram feature analysis to fuse these features, allowing them to fit the correlation between page view and the quality of pictures. The results of experiments show that these features are able to predict the page view of articles, and that the method of using the three groups of features can more accurately fit the correlation.

The structure of this paper is as follows. In section 2, we describe the construction of the OAPQD dataset. In section 3, the three features and the selective ensemble method that can fuse them are presented separately. We conduct the comparison experiment on the OAPQD to analyze the magnitude of the seven features on fitting the page view in section 4. Section 5 gives the concluding remarks.

2. THE DATASET

With the development of information and network technology, traditional media were gradually replaced by digital new media, such as WeChat official account, which has been widely used by all walks of life. Currently, most universities use official accounts as the platform for campus culture construction. In order to better explore the reasons why articles are popular on public accounts, we focus mainly on the page view of articles. To this end, we first subscribed to the WeChat official accounts of nine well-known universities in Beijing, then selected the pictures inserted in the articles that were published by the accounts in the past months, based on which a new database is created. To be specific, the most researched and representative pictures are extracted from the selected article. Simultaneously, the number of page views corresponding to the selected article is recorded, with a maximum of 100,000 and a minimum of 253. We selected a picture from a large number of articles published by official accounts of schools every day, and we have collected 1,276 pictures altogether. However, not each of the above pictures has research value. In these pictures, the selection criteria are first based on the picture content and type, and then exclude extreme special cases, such as the case where the picture quality is very poor but the number of clicks is very high. Finally, 1,000 most representative pictures were selected to form the picture data set. **Figure 1** shows the subset of OAPQD.

By observing the data set we constructed, we find that there is a positive correlation between picture quality and page view. As shown in **Figure 2**, there are three pictures from left to right. The picture on the left is the most colorful and clear among the three pictures, giving a better visual experience with 41,000 hits. The intermediate picture is of poor quality, with only 7,466 clicks. The picture on the far right is the least visually appealing and thus logically the least clicked picture with only 1,052.

3. METHODOLOGY

The specific features can well reflect the page view of pictures, but the fitting accuracy of using a certain characteristic feature alone is relatively low. In this section, we will introduce the three groups of complementary features extracted based on natural scene analysis, histogram, and free energy theory, and further describe a selective ensemble approach capable of fusing the 99 features.

3.1. NSS-Based Feature Extraction

The first group is composed of 36 features (f_{01}-f_{36}), which were proposed on the basis of NSS theory. Bovik (2010) suggested that natural pictures have regular statistical characteristics, therefore, the statistical features of natural scenes can be considered as an effective and powerful tool for PQA. In general, complex image textures affect the perceptual level of distortion, and the local brightness normalization can greatly reduce the correlation between adjacent pixels of the original picture and the distorted picture. Thus, the classic spatial NSS model is first used to preprocess the picture to remove the local mean value, and then the picture is segmented and normalized to extract the mean subtracted contrast normalized coefficient of natural scene pictures. The Mean Subtracted Contrast Normalized (MSCN) coefficients vary in different ways due to distinct distortions. On the basis of this variation, the type of picture distortion and the perceived quality of pictures can be predicted. The pixel intensity of natural pictures follows a Gaussian distribution, which can be represented by a Bell curve. In order to clearly observe the differences in data distribution between different distortion types and natural pictures, we use the generalized Gaussian distribution (GGD) to fit the distribution of MSCN. The sign of the transformed picture coefficients are regular, but Mittal et al. suggested that the existence of distortion affects this above correlation structure. In order to research the correlation information between connected pixels, the zero-mode AGGD is used to model the inner product of MSCN adjacent coefficient. The moment matching-based approach proposed by Lasmar et al. (2009) can estimate the parameters of the AGGD. Then we calculate the adjacent pairs of coefficients from the horizontal, vertical, and diagonal directions to obtain the 16 parameters, respectively. Low-resolution pictures are obtained from each picture through low-pass filtering and downsampling with a factor of 2. We measure the MSCN parameters fitted by GGD and the 16 parameters generated by AGGD according to the above two scales. Once all the work mentioned above is done, the first feature set consisting of 36 features is obtained.

3.2. Histogram-Based Feature Extraction

The second group consists of 40 features (f_{37}-f_{76}), illustrating the main features of the HVS introduced from biology in image processing. Since the visual information in picture is often redundant, the understanding of the HVS is mainly related to its basic features, such as contour, zero cross, and so on. Gradient magnitude (GM) feature can reflect the intensity of local luminance variations. The local maximum GM pixels can reflect small details and textural change of pictures, which is the main element of contour. GM has been widely used for PQA methods, such as FSIM proposed by Zhang et al. (2011), GMSD constructed by Xue et al. (2013), PSIM designed by Gu et al. (2017a), and ADD-GSIM established by Gu et al. (2016), where picture quality is evaluated only by the similarity of gradient magnitude. Besides, on the basis of GM method, Min et al. (2019b) first proposed a picture dehazing algorithm, then a novel objective index named DHQI was presented by Min et al. (2019a) can be utilized to evaluate DHAs or optimize practical dehazing systems. Finally, a blind PQA method was introduced by Min et al. (2018) has a superior performance. Generally, GM is calculated using linear filter convolution, where the typical filters are mainly Sobel, Prewitt, Roberts, etc. Unlike the GM operator, isotropic measurements on the second spatial derivative of pictures show the strongest brightness variation. The Laplacian of Gaussian (LOG) operator reflects the intensity contrast of a small spatial neighborhood, and Marr and Hildreth (1980) proposed that it can model the receptive fields of retinal ganglion cells. The LOG operator and the GM operator adopt the anisotropic calculation method without angular preference to obtain the local picture structure from different angles. They can represent the structural information of pictures, especially the local contrast features, and therefore can be used to form the semantic information of pictures. Finally, the picture local quality prediction is achieved by using these two operators mentioned above.

3.3. Free Energy-Based Feature Extraction

The 23 features (f_{77}-f_{99}) extracted in the third group are inspired by the free energy principle and the structural degradation

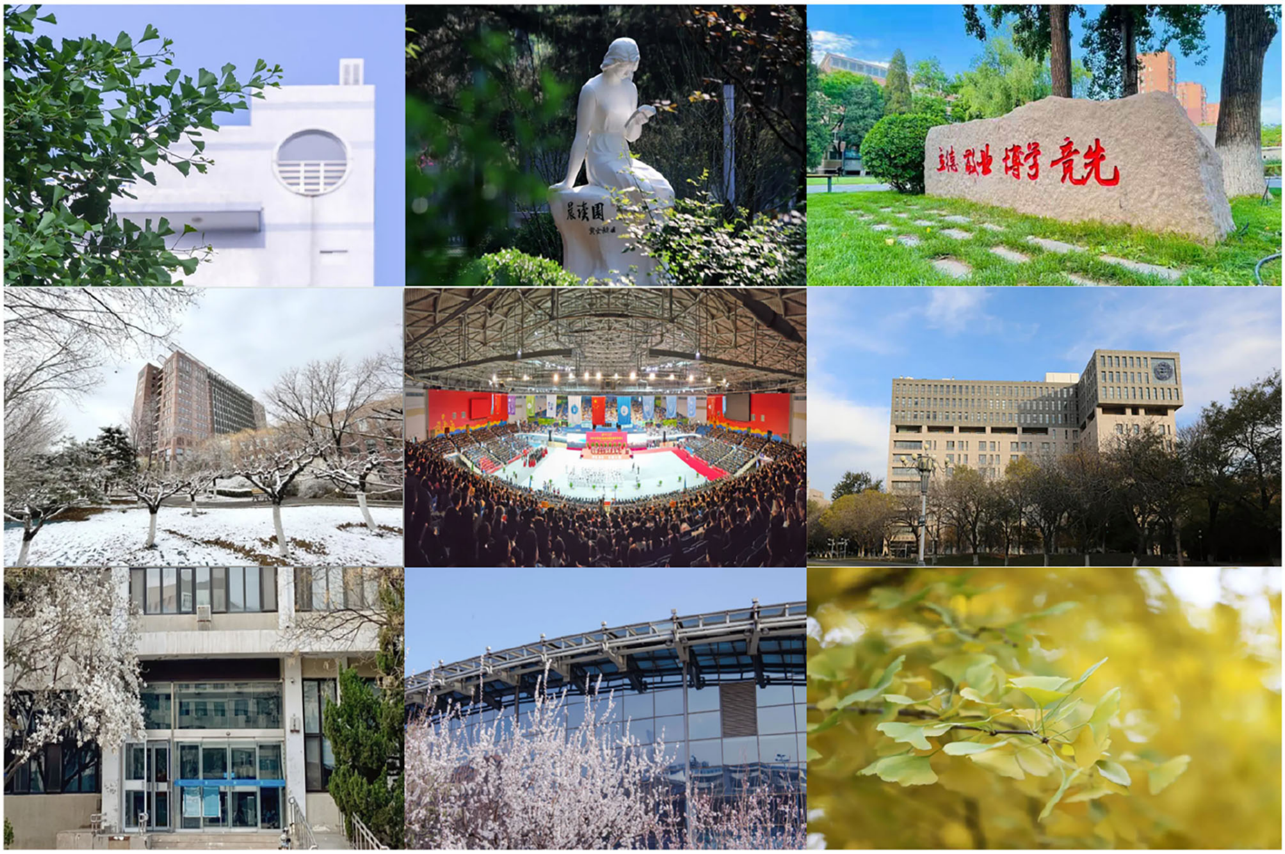

FIGURE 1 | Representative nine pictures from the OAPQD data set, the content of above mainly includes architecture, landscape, people, text content, meeting scene, etc.

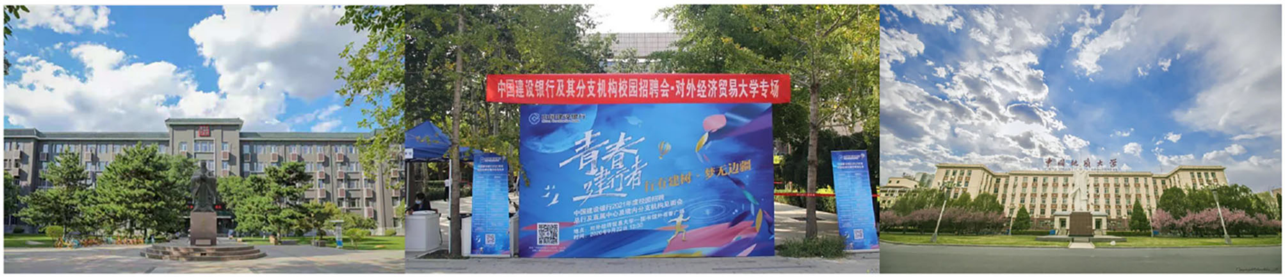

FIGURE 2 | The quality of the three pictures in the OAPQD decreases gradually from left to right.

model (SDM). A basic premise of the free energy theory is that an internal generative model can be used to estimate the gap between the viewing scene and the corresponding brain prediction. It measures the difference between the probability distribution of environmental quantities acting on the system and an arbitrary distribution encoded by its configuration. Since this process is very closely related to the quality of human visual perception, it can be used for the PQA method. The free energy of pictures can be approximated by the AR model as the total description length of pictures data. In an effective RR SDM proposed by Gu et al. (2015b), we observe the structural degradation after low-pass filtering of the picture. The spatial frequency of input picture I has different degrees of decrease. We first define the local mean and variance of I with a two-dimensional circularly symmetric Gaussian weighting function. The linear dependence between the free energy and the structural degradation information provides an opportunity to characterize distorted pictures in the absence of the information of original picture. Furthermore, the NFEQM is added to the third group as feature f_{99} due to its excellent performance in noisy and blurred pictures.

3.4. Selective Ensemble-Based Page View Inference

A single picture feature does not represent the picture quality well, which will lead to its poor fitting of the relationship between features and page view. To solve this question, we consider an ensemble learning approach which can produce strong generalization to improve the fitting accuracy. This content has become a hot research topic in the international machine learning community, so there are more and more methods presents by scholars, such as the geometric structural ensemble (GSE) learning framework approach presented by Zhu et al. (2018). Zhou et al. (2002) suggested that the presence of high-dimensional selective ensemble methods based on direct merging is prone to overfitting or some of these features may be overlooked in the fitting process. On the basis of this theory, we adopt the method of selective ensemble to further enhance the performance of our presented approach in this paper.

It is natural to combine features to derive a more effective preprocessing method, so as to better remove random details caused by the varying viewing method and picture resolution in different but supplementary domains. We combine the three features two by two and last fuse the three by using a selective ensemble technique proposed by Gu et al. (2020b) and Chen et al. (2021), so as to make an experimental comparison with the accuracy of the fit using single features. The following seven categories can be generated based on the number of features: (1) BRISQUE; (2) GMLF; (3) NFERM; (4) BRISQUE+GMLF; (5) BRISQUE+NFERM; (6) GMLF+NFERM; (7) BRISQUE+GMLF+NFERM. The experimental results show that the number of fused features affects the linearity of the results, where the method that fuses three features has the best performance and the single feature has the worst accuracy in fitting the correlation between picture quality and page view.

4. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we carry out the comparison experiment on the OAPQD, so as to understand the degree of seven features on fitting the page view of articles. In this process, we select the two classical metrics to evaluate the performance of experiments.

In order to further predict page view with quality-aware features, experiments are conducted on the OAPQD dataset consisting of 1,000 pictures selected from the WeChat official accounts of nine universities in Beijing. The pictures from the dataset used for testing are rich in content and variety, and can improve well the hit of pictures published in the college official accounts. This provides a certain foundation for our proposed method. In the experimental analysis section, we use two commonly statistical indicators as the metrics to assess the performance, which are the Pearson linear correlation coefficient (PLCC) and the Spearman rank order correlation coefficient (SRCC). The PLCC is a linear correlation coefficient with scale invariance, which indicates the degree of similarity between picture features and page view. The PLCC is defined as

$$PLCC = \frac{\sum_i (q_i - \bar{q}) \cdot (o_i - \bar{o})}{\sqrt{\sum_i (q_i - \bar{q})^2 \cdot \sum_i (o_i - \bar{o})^2}} \tag{1}$$

where o_i and $\bar{o}$ represent the features of the ith picture and its overall mean value, and q_i and $\bar{q}$ are the page view of ith picture and its mean value. Before using the PLCC metric for evaluation, we employ the nonlinear regression equation proposed by Sheikh et al. (2006), which is given by

$$p(x) = \alpha_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\alpha_2 (x - \alpha_3)}} \right] + \alpha_4 x + \alpha_5 \tag{2}$$

where $p(x)$ represents the predicted score, α_i $(i = 1,2,3,4,5)$ is the parameter of the generation fitting, and x is the original prediction score. While the SRCC represents the strength of the monotonic relationship predicted by the algorithm, it can be calculated by

$$SRCC = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^{N} d_i^2 \tag{3}$$

where N is the number of pictures in the dataset, and d_i is the difference between the ranking of *ith* picture in features and page view. The value range of PLCC and SRCC is $[-1, 1]$. The closer the absolute value of these two indicators is to 1, the stronger the correlation between picture features and page view, where >0 means a positive correlation and <0 means a negative correlation. In the regression problem, the closer the value is to 1, the higher the accuracy of the algorithm.

We extract three types of features, where the first set of feature coefficients has the characteristic statistical property of varying due to the distortion. Quantifying these variations allows obtaining the type of picture distortion while enabling the prediction of page view. The second group of features is composed of 40 local contrast features, GM and LOG, which can detect changes in the semantic structure of the picture due to variations of luminance for the purpose of predicting the page view of the article. The third set of features consists of 23 features based on free energy and structure degradation information. In addition, they are inspired by the HVS and the free energy theory, which fill the gap in the NR PQA method due to the lack of prior knowledge.

The features mentioned above can reflect the page view well, and based on this, we use selective ensemble technology to fuse features in different ways for comparison experiments. The results of comparison experiments show that the method that fuses all the three features together obtain the largest data value and the highest accuracy of the results, followed by the method of fusing two features. The experimental data is placed inside **Table 1**, where the values obtained by the best-performing

TABLE 1 | The Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SRCC) values of seven feature fusion methods on the dataset.

Algorithm	PLCC	SRCC
BRISQUE (direct use)	0.0156	0.0347
GMLF (direct use)	0.0034	0.0343
NFERM (direct use)	0.0683	0.0146
BRISQUE (re-train)	0.3925	0.2707
GMLF (re-train)	0.3911	0.3340
NFERM (re-train)	0.3983	0.2782
BRISQUE+GMLF	0.4545	0.3577
BRISQUE+NFERM	0.4454	0.3054
GMLF+NFERM	0.4387	0.3655
BRISQUE+GMLF+ NFERM	**0.4764**	**0.3863**

The top data values are given in bold.

method are given in bold. In **Table 1**, it can be seen that the values of PLCC and SRCC are very approximate when using a single algorithm. It is not difficult to find that GMLF has gained the best results (on average) of SRCC, which is sensitive to pictures with gradient features. However, **Table 1** reports the low correlation performance on SRCC when combined with the features from BRISQUE and NFERM. It also can be seen that the more the number of fused picture features, the better the fit to the relation between features and page view. Meanwhile, it shows a certain degree of similarity between the features and click-through rate. This method proposed in this paper can provide guidance for the management of college official accounts. For example, the insertion of high-definition and high-quality pictures into published articles can increase the visibility of the articles.

5. CONCLUSION

In this paper, we have studied the connection between picture features and the popularity of articles published in college official accounts. We elaborately select 1,000 pictures from the official accounts of nine universities, construct a picture database named OAPQD, and record the clicks of corresponding articles. Three groups of features extracted from different angles can reflect the features, and the stacked selective ensemble technology is used to fuse them for comparison experiments. The experimental results show that the method integrating three groups of 99 features at the same time has the highest accuracy in fitting the page view. Therefore, in future publicity work, the selection of pictures is very meaningful for the popularity of official account articles. For the publicity department of the college, they can import our method to predict the page views of their articles and use these data parameters to adjust picture quality or change diffusion strategy. All of these measures can improve the visibility of official accounts to some extent.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

YY conceived the framework of the paper and completed the main content of the paper. SS collected a large number of references to provide a strong background basis for the paper. YW was mainly responsible for the revision of the thesis. XL participated in the revision and content supplement of the article. JL revised the layout of the article and checked for grammatical errors. FL checked the final version of the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Social Science Foundation of China (15BJY048).

REFERENCES

Bovik, A. C. (2010). *Handbook of Image and Video Processing*. Academic Press.

Chen, W., Gu, K., Zhao, T., Jiang, G., and Callet, P. L. (2021). Semi-reference sonar image quality assessment based on task and visual perception. *IEEE Trans. Multimedia* 23, 1008–1020. doi: 10.1109/TMM.2020.2991546

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *J. Physiol.* 100, 70–87. doi: 10.1016/j.jphysparis.2006.10.001

Geng, B., Yang, L., Xu, C., Hua, X.-S., and Li, S. (2011). "The role of attractiveness in web image search," in *Proceedings of the 19th ACM International Conference on Multimedia* (Scottsdale, AZ), 63–72.

Greenberg, R., Magisos, M., Kolvoord, R., and Strom, R. (1994). "Image processing for teaching: a national dissemination program," in *Proceedings of 1st International Conference on Image Processing, Vol. 1* (Austin, TX), 511–514.

Gu, K., Li, L., Lu, H., Min, X., and Lin, W. (2017a). A fast reliable image quality predictor by fusing micro-and macro-structures. *IEEE Trans. Indus. Electron.* 64, 3903–3912. doi: 10.1109/TIE.2017.2652339

Gu, K., Liu, H., Xia, z., Qiao, J., Lin, W., and Thalmann, D. (2021a). Pm $_{2.5}$ monitoring: use information abundance measurement and wide and deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* 2, 4278–4290. doi: 10.1109/TNNLS.2021.3105394

Gu, K., Liu, M., Zhai, G., Yang, X., and Zhang, W. (2015a). Quality assessment considering viewing distance and image resolution. *IEEE Trans. Broadcast.* 61, 520–531. doi: 10.1109/TBC.2015.2459851

Gu, K., Qiao, J., and Li, X. (2019). Highly efficient picture-based prediction of PM2. 5 concentration. *IEEE Trans. Indus. Electron.* 66, 3176–3184. doi: 10.1109/TIE.2018.2840515

Gu, K., Qiao, J., and Lin, W. (2018). Recurrent air quality predictor based on meteorology- and pollution-related factors. *IEEE Trans. Indus. Informatics* 14, 3946–3955. doi: 10.1109/TII.2018.2793950

Gu, K., Tao, D., Qiao, J.-F., and Lin, W. (2017b). Learning a no-reference quality assessment model of enhanced images with big data. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 1301–1313. doi: 10.1109/TNNLS.2017.2649101

Gu, K., Wang, S., Zhai, G., Lin, W., Yang, X., and Zhang, W. (2016). Analysis of distortion distribution for pooling in image quality prediction. *IEEE Trans. Broadcast.* 62, 446–456. doi: 10.1109/TBC.2015.2511624

Gu, K., Xia, Z., and Qiao, J. (2020a). Deep dual-channel neural network for image-based smoke detection. *IEEE Trans. Multimedia* 22, 311–323. doi: 10.1109/TMM.2019.2929009

Gu, K., Xia, Z., and Qiao, J. (2020b). Stacked selective ensemble for PM2.5 forecast. *IEEE Trans. Instrum. Meas.* 69, 660–671. doi: 10.1109/TIM.2019.2905904

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2014). Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Trans. Broadcast.* 60, 555–567. doi: 10.1109/TBC.2014.2344471

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2015b). Using free energy principle for blind image quality assessment. *IEEE Trans. Multimedia* 17, 50–63. doi: 10.1109/TMM.2014.2373812

Gu, K., Zhang, Y., and Qiao, J. (2020c). Vision-based monitoring of flare soot. *IEEE Trans. Instrum. Meas.* 69, 7136–7145. doi: 10.1109/TIM.2020.2978921

Gu, K., Zhang, Y., and Qiao, J. (2021b). Ensemble meta-learning for few-shot soot density recognition. *IEEE Trans. Indus. Informat.* 17, 2261–2270. doi: 10.1109/TII.2020.2991208

Jiang, Q., Peng, Z., Yue, G., Li, H., and Shao, F. (2021). No-reference image contrast evaluation by generating bidirectional pseudoreferences. *IEEE Trans. Indus. Informat.* 17, 6062–6072. doi: 10.1109/TII.2020.3035448

Jiang, Q., Shao, F., Gao, W., Chen, Z., Jiang, G., and Ho, Y.-S. (2019a). Unified no-reference quality assessment of singly and multiply distorted stereoscopic images. *IEEE Trans. Image Process.* 28, 1866–1881. doi: 10.1109/TIP.2018.2881828

Jiang, Q., Shao, F., Lin, W., Gu, K., Jiang, G., and Sun, H. (2018). Optimizing multistage discriminative dictionaries for blind image quality assessment. *IEEE Trans. Multimedia* 20, 2035–2048. doi: 10.1109/TMM.2017.2763321

Jiang, Q., Shao, F., Lin, W., and Jiang, G. (2019b). Blique-TMI: blind quality evaluator for tone-mapped images based on local and global feature analyses. *IEEE Trans. Circuits Syst. Video Technol.* 29, 323–335. doi: 10.1109/TCSVT.2017.2783938

Jiang, Q., Zhou, W., Chai, X., Yue, G., Shao, F., and Chen, Z. (2020). A full-reference stereoscopic image quality measurement via hierarchical deep feature degradation fusion. *IEEE Trans. Instrum. Meas.* 69, 9784–9796. doi: 10.1109/TIM.2020.3005111

Kang, H., Ko, J., Park, H., and Hong, H. (2019). Effect of outside view on attentiveness in using see-through type augmented reality device. *Displays* 57, 1–6. doi: 10.1016/j.displa.2019.02.001

Kim, H., Yi, S., and Yoon, S.-Y. (2019). Exploring touch feedback display of virtual keyboards for reduced eye movements. *Displays* 56, 38–48. doi: 10.1016/j.displa.2018.11.004

Lasmar, N.-E., Stitou, Y., and Berthoumieu, Y. (2009). "Multiscale skewed heavy tailed model for texture analysis," in *2009 16th IEEE International Conference on Image Processing (ICIP)* (Cairo), 2281–2284.

Liang, H., and Weller, D. S. (2016). Comparison-based image quality assessment for selecting image restoration parameters. *IEEE Trans. Image Process.* 25, 5118–5130. doi: 10.1109/TIP.2016.2601783

Lin, C. J., Prasetyo, Y. T., and Widyaningrum, R. (2019). Eye movement measures for predicting eye gaze accuracy and symptoms in 2d and 3d displays. *Displays* 60, 1–8. doi: 10.1016/j.displa.2019.08.002

Liu, H., Lei, F., Tong, C., Cui, C., and Wu, L. (2021). Visual smoke detection based on ensemble deep cnns. *Displays* 69:102020. doi: 10.1016/j.displa.2021.102020

Marr, D., and Hildreth, E. (1980). Theory of edge detection. *Proc. R. Soc. Lond. Ser. B. Biol. Sci.* 207, 187–217. doi: 10.1098/rspb.1980.0020

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast.* 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhai, G., Gu, K., Yang, X., and Guan, X. (2019a). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transport. Syst.* 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019b). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Trans. Multim.* 21, 2319–2333. doi: 10.1109/TMM.2019.2902097

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 4695–4708. doi: 10.1109/TIP.2012.2214050

Moorthy, A. K., and Bovik, A. C. (2011). Blind image quality assessment: from natural scene statistics to perceptual quality. *IEEE Trans. Image Process.* 20, 3350–3364. doi: 10.1109/TIP.2011.2147325

Moroz, M., Garzorz, I., Folmer, E., and MacNeilage, P. (2019). Sensitivity to visual speed modulation in head-mounted displays depends on fixation. *Displays* 58, 12–19. doi: 10.1016/j.displa.2018.09.001

Rajashekar, U., Panayi, G., Baumgartner, F., and Bovik, A. (2002). The siva demonstration gallery for signal, image, and video processing education. *IEEE Trans. Educ.* 45, 323–335. doi: 10.1109/TE.2002.804392

Richard, W. (1991). An educational image processing/machine vision system. *IEEE Trans. Educ.* 34, 129–132. doi: 10.1109/13.79893

Ruderman, D. L. (1994). The statistics of natural images. *Network* 5:517. doi: 10.1088/0954-898X_5_4_006

Saad, M. A., Bovik, A. C., and Charrier, C. (2012). Blind image quality assessment: a natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.* 21, 3339–3352. doi: 10.1109/TIP.2012.2191563

Sheikh, H. R., Sabir, M. F., and Bovik, A. C. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* 15, 3440–3451. doi: 10.1109/TIP.2006.881959

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Su, H., Jung, C., Wang, L., Wang, S., and Du, Y. (2019). Adaptive tone mapping for display enhancement under ambient light using constrained optimization. *Displays* 56, 11–22. doi: 10.1016/j.displa.2018.10.005

Tang, X.-T., Yao, J., and Hu, H.-F. (2020). Visual search experiment on text characteristics of vital signs monitor interface. *Displays* 62:101944. doi: 10.1016/j.displa.2020.101944

Wang, X., Jiang, Q., Shao, F., Gu, K., Zhai, G., and Yang, X. (2021). Exploiting local degradation characteristics and global statistical properties for blind quality assessment of tone-mapped HDR images. *IEEE Trans. Multim.* 23, 692–705. doi: 10.1109/TMM.2020.2986583

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wu, H.-N., Wang, X.-M., Yu, L.-K., Yuan, T., and Kuai, S.-G. (2019). Rendering a virtual light source to seem like a realistic light source in an electronic display: a critical band of luminance gradients for the perception of self-luminosity. *Displays* 59, 44–52. doi: 10.1016/j.displa.2019.07.001

Wu, J., Lin, W., Shi, G., and Liu, A. (2013). Reduced-reference image quality assessment with visual information fidelity. *IEEE Trans. Multimedia* 15, 1700–1705. doi: 10.1109/TMM.2013.2266093

Xu, L., Lin, W., Ma, L., Zhang, Y., Fang, Y., Ngan, K. N., et al. (2016). Free-energy principle inspired video quality metric and its use in video coding. *IEEE Trans. Multimedia* 18, 590–602. doi: 10.1109/TMM.2016.2525004

Xue, W., Mou, X., Zhang, L., Bovik, A. C., and Feng, X. (2014). Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Trans. Image Process.* 23, 4850–4862. doi: 10.1109/TIP.2014.2355716

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2013). Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Process.* 23, 684–695. doi: 10.1109/TIP.2013.2293423

Yaman, O., and Karakose, M. (2016). "Development of image processing based methods using augmented reality in higher education," in *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)* (Istanbul), 1–5.

Yildirim, C. (2019). Cybersickness during vr gaming undermines game enjoyment: a mediation model. *Displays* 59, 35–43. doi: 10.1016/j.displa.2019.07.002

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhang, Y., Tu, Y., and Wang, L. (2020). Effects of display area and corneal illuminance on oculomotor system based on eye-tracking data. *Displays* 63:101952. doi: 10.1016/j.displa.2020.101952

Zheng, T., Pan, W., and Jia, B. (2009). "Study on the application of pattern recognition technology in distance education system," in *2009 Second International Conference on Computer and Electrical Engineering, Vol. 1* (Dubai), 474–477.

Zhou, Z.-H., Wu, J., and Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artif. Intell.* 137, 239–263. doi: 10.1016/S0004-3702(02)00190-X

Zhu, Z., Wang, Z., Li, D., Zhu, Y., and Du, W. (2018). Geometric structural ensemble learning for imbalanced problems. *IEEE Trans. Cybern.* 50, 1617–1629. doi: 10.1109/TCYB.2018.2877663

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 01 November 2021
doi: 10.3389/fnins.2021.771533

Working With Environmental Noise and Noise-Cancelation: A Workload Assessment With EEG and Subjective Measures

Kerstin Pieper[1]*, Robert P. Spang[1], Pablo Prietz[1], Sebastian Möller[1,2], Erkki Paajanen[3], Markus Vaalgamaa[3] and Jan-Niklas Voigt-Antons[1,2]

[1] Quality and Usability Lab, Institute of Software Engineering and Theoretical Computer Science, Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany, [2] German Research Center for Artificial Intelligence, Berlin, Germany, [3] Tampere Wireless Headset Audio Lab, Finland Research Center, Huawei Technologies Oy (Finland) Co., Ltd., Tampere, Finland

OPEN ACCESS

Edited by:
Xiongkuo Min,
University of Texas at Austin,
United States

Reviewed by:
Yucheng Zhu,
Shanghai Jiao Tong University, China
Huiyu Duan,
Shanghai Jiao Tong University, China

***Correspondence:**
Kerstin Pieper
kerstin.pieper@tu-berlin.de

Specialty section:
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

Received: 06 September 2021
Accepted: 23 September 2021
Published: 01 November 2021

Citation:
Pieper K, Spang RP, Prietz P, Möller S, Paajanen E, Vaalgamaa M and Voigt-Antons J-N (2021) Working With Environmental Noise and Noise-Cancelation: A Workload Assessment With EEG and Subjective Measures.
Front. Neurosci. 15:771533.
doi: 10.3389/fnins.2021.771533

As working and learning environments become open and flexible, people are also potentially surrounded by ambient noise, which causes an increase in mental workload. The present study uses electroencephalogram (EEG) and subjective measures to investigate if noise-canceling technologies can fade out external distractions and free up mental resources. Therefore, participants had to solve spoken arithmetic tasks that were read out via headphones in three sound environments: a quiet environment (*no noise*), a noisy environment (*noise*), and a noisy environment but with active noise-canceling headphones (*noise-canceling*). Our results of brain activity partially confirm an assumed lower mental load in *no noise* and *noise-canceling* compared to *noise* test condition. The mean P300 activation at Cz resulted in a significant differentiation between the *no noise* and the other two test conditions. Subjective data indicate an improved situation for the participants when using the noise-canceling technology compared to "normal" headphones but shows no significant discrimination. The present results provide a foundation for further investigations into the relationship between noise-canceling technology and mental workload. Additionally, we give recommendations for an adaptation of the test design for future studies.

Keywords: mental workload, ambient noise, noise-canceling, event-related potentials, EEG frequency, subjective measures

1. INTRODUCTION

In flexible working surroundings like landscape offices, business trips, or even the home office, people have to deal with noisy environments. It is hardly avoidable to be distracted by, e.g., other conversations, traffic noise, or screaming kids while focusing on the actual task. The combination of stressful influences and task difficulty increases the workload for the person. The interaction of task characteristics and the person's capacity influences the amount of mental load a person is able to allocate in a task (Choi et al., 2014). Additionally, environmental stressors decrease task performance and lead to motivational deficits (Evans and Stecker, 2004). In task solving, which requires cognitive resources, keeping the demand on an appropriate level is important.

Especially while working, a balance is necessary between the work and any parallel (and potentially distracting) tasks to stay focused over longer time (Teigen, 1994). There is evidence for a relationship between the development of mental disorders and continuous high levels of workload, as well as for decreased satisfaction and well-being (van Daalen et al., 2009).

Mobile solutions which help to stay focused are frequently used to improve the situation for the working person. One option to directly reduce environmental auditory noise without changing the working environment is headphones with active noise-canceling. These technologies use a basic principle of wave optics called destructive interference. A signal superimposes the incoming noise signal, which has the same amplitude but the opposite phase (Kuo et al., 2006). Thereby, noise-canceling headphones offer an individual and a mobile solution for noise suppression. In the present study, we examine to what extent noise affects the workload level in task solving and whether noise-canceling technologies can reduce the workload compared to the use of headphones in normal mode in otherwise identical circumstances.

In the present experiment, the participants performed the same cognitive task in three different noise environments, which serve as test conditions. The within-subject test design should deliver insights about differences in the workload level between conditions and changes over time for each condition separately. In the *no noise* condition, the quiet environment should allow the participants to focus on the task. We suggest the mental load to be on a mid-level in this test condition. The ambient noise presentation was assumed to increase mental load due to a higher need for resources to stay focused. This effect should become sharper in the *noise* condition as the persons were directly exposed to the ambient noise. With the activation of the noise-canceling feature in the *noise-canceling* condition, the workload level was supposed to be lower compared to the *noise* condition and slightly increased compared to *no noise* condition.

For measuring workload, we employed subjective and EEG measurement of brain. Subjective measures primarily assess the participants' reactions to experimental manipulation and thereby give valuable insights about the person's state at the moment of the measurement. EEG and, in general, physiological measures offer the advantage of a recording over an experiment's whole duration. The resulting continuous signal enables the detection of stimulus-related reactions and also the observation of changes over time. Therefore, the combination of measures is assumed to give more complex insights as one measure alone. We suggested the delivered findings from subjective measures to give a fundamental differentiation between conditions into the person's mental and affective state. EEG should deliver information about the brain's underlying processes, which cause differentiation in the level of mental workload for the three test conditions.

Research about workload and its underlying processes was extensively studied for decades, but it is still an elusive concept. There are different considerations about how to define workload and how it interacts with other mental processes. In an early concept, given by Kahnemann, "mental effort" is described as a capacity that is invested in task processing or demanded by a task. The extend of effort invested in task solving is less influenced by the task solver's intention, but rather it is regularized by the task demand (Kahneman, 1973). Later on, Wickens describes "Workload" as the interrelation between the task demand and the humans' limited mental resources needed for solving it. Depending on the complexity of one or more tasks, multiple resources are required. These resources are multidimensional and can be differentiated in several "stages" and "modalities" (Wickens, 1979, 2008), whereas the resulting load is a global (mental-) "load" on the human (Rasmussen, 1979; Wickens, 2008). Especially subjective measures have a high operator acceptance because of paying attention to the opinion of the participant (Hill et al., 1992). Since it appears that emotions are related to the perceived workload of a task and the other way around (Jeon et al., 2011; Chaouachi and Frasson, 2012) we wanted to investigate aspects of the emotional state of the participants in the current test design. With higher ratings for negatively related emotional items, we suggest a higher perceived workload.

In several studies, it was shown that both subjective measurements and EEG measurements show sensitivity for workload (Parasuraman, 1990; Hankins and Wilson, 1998; Borghini et al., 2014). On the one hand, it is of interest to confirm the results of one measurement with the other measurement results. However, it is also suggested that subjective meaning conscious ratings deliver deviating observations as unconscious activation in the brain. We expect additional and possibly more detailed observations from the study of brain activity.

EEG data can be investigated regarding mental processing and workload, considering event-related potentials (ERPs) and power spectral densities of frequency bands. In the frequency domain, we investigate the spectral power of frequency bands. The EEG frequency bands of interest are delta, theta, alpha, beta, and gamma. We define the frequency range 0.1–4 Hz corresponding to delta (delta is categorized differently but often in the range between 0.3 and 4.5 Hz; see, e.g., Feinberg et al., 1987; Anderson and Horne, 2003; Knyazev, 2012), 4–8 Hz corresponding to theta, 8–12 Hz corresponding to alpha, 12–30 Hz corresponding to beta, and 30–40 Hz corresponding to gamma (gamma frequency range is referred to as $< 30Hz$; see, e.g., Knyazev, 2012). An increase in delta activation was observed in pilots during flying operations with rising cognitive demand (Harmony et al., 1996; Wilson, 2002). The theta band is suggested to be associated with memory processes and mental workload (Klimesch, 1999). Reduced alpha in combination with higher theta power is suggested to occur when workload increases (Brouwer et al., 2012). With increasing task difficulty and thereby with increasing cognitive load, the frontal-midline theta responds with a maximum at frontal central electrode positions (Ishihara and Yoshii, 1972; Gevins et al., 1998). The alpha band is the dominant frequency in the human scalp EEG (Klimesch, 1999). Alpha band power response to workload showed a varying behavior. In a visual spatial task, alpha at parietal-temporal-occipital region decreased with task difficulty (Gevins et al., 1998). In a following experiment in which a memory component extended the task, it was shown that alpha total power increased with task difficulty (Murata, 2005). In an experiment by (Yu et al., 2009) with a mental arithmetic task, an alpha decrease and a beta increase at parietal and occipital sites was shown. Additionally, beta power activation seems to be related with cognitive processing (Ray and Cole, 1985). Studies investigating the gamma band suggest an increased activation with raised task difficulty (Gevins et al., 1998; Knoll et al., 2011). Other findings suggest gamma (40 Hz) activity reported an activation in a selective attention task of auditory stimuli at the auditory cortex (Tiitinen et al., 1993). It is also known to be more generally associated with sensory processing and cognitive processes with a wide distribution on the scalp (Başar-Eroglu et al., 1996).

Based on these insights, we suggested the highest delta, theta, beta, and gamma power spectral density in the *noise* condition and in the *no noise* condition the lowest. In the *noise-canceling* condition, it was suggested to be on a mid-level. For alpha power spectral density, it was suggested to be in lower in the *no noise*, on a mid-level in the *noise-canceling*, and smallest in the *noise* condition.

In the time domain, the stimulus-locked ERPs were investigated. Based on a body of literature, we suggested differences in the P300 that is a positive component of the ERPs, which peaks 300 ms after a stimulus onset (Duncan et al., 2009). It shows sensitivity to workload, a maximum characteristic over midline scalp sites, and has a centro-parietal distribution. The P300 component is often divided in two parts: the P3a and the P3b. Whereby, the P3a appears as a response to novelty of a stimulus and as an orienting response, and the P3b shows a sensitivity for task-relevant processing and decision-making processes (Friedman et al., 2001). The amplitude of P300 has been reported to be an indicator of different levels of difficulty (Wickens et al., 1977; Kramer et al., 1987) and thereby workload (Ullsperger et al., 2001). With increasing workload, the amplitude of P300 is suggested to remain smaller, whereby the latency of the component remains higher (Duncan et al., 2009). Studies of the P300 were conducted primarily in conjunction with a classical Oddball paradigm. With our task design, we deviate from the classic oddball like it was done in studies about the discrimination of different workload levels (Ullsperger et al., 2001; Allison and Polich, 2008). The main task, solving mental arithmetics, is complicated with different levels of noise intensities. This scenario is comparable to attending an online meeting in a noisy environment while recording thoughts into a protocol. The mental demand caused by the recall of numbers and the calculation of the arithmetics is suggested to elicit a P300. The spoken arithmetic equations and the environmental noise address the same sensory modality, which means a higher workload in this channel. So we suggest the P300 amplitude to be smallest in the *noise* and most extensive in the *no noise* condition. As the *noise-canceling* technology suppresses the ambient noise, making it easier to focus as in the *noise* condition, we suggested the amplitude to be on a mid-level (between the other two conditions).

In the following, we explain the task and the whole test setup. The following part reports the results and delivers the base for the subsequent discussion, including limitations and suggestions for future work. In the final section, we provide a conclusion.

2. MATERIALS AND METHODS

2.1. Task

The fundamental task of each trial was to solve an arithmetic equation. These consisted of two numbers and the four basic operators: addition, subtraction, multiplication, and division. The two numbers and the result were in the range of 1–200, and they were all integer numbers. These tasks were presented auditory via headphones to the participant.

We decided against providing the participants with a fixed time for answering because the difficulty among the tasks varied heavily. In pre-tests, we observed participants to develop answering strategies to cope better with the demanding situation. For example, the task $3 + 2$ was more intuitive to solve in a short time for most participants, whereas many people took a long time to solve $23 * 7$. While testing a constant time given for all sorts of tasks, we noticed that participants sometimes typed in the answer of an easy task but waited until the time was almost over to provide themselves with a short break. If we are now interested in, e.g., the total amount of tasks solved correctly throughout a condition, this avoiding behavior will bias our findings. Furthermore, the short, unplanned breaks might impact the perceived workload as well. Hence, we decided to create a machine learning-based algorithm to predict the ideal time needed to solve the task for each participant. Since it is not the focus of this paper to describe the algorithms in-depth, we present here only the fundamental idea of the model: Using general features from the arithmetic task at hand, e.g., the operator and the digit-span, as well as the previous performance of the participant, we predicted the time per task individually. This allowed us to address the individual abilities of each subject but also to not allow for any headroom in the time given.

2.2. Subjective Measures

We chose the NASA Task Load Index (NASA-TLX) questionnaire to assess the participants' perceived workload. It is a sensitive indicator of workload because participants describe their personal impressions from their individual viewpoint (Hart and Staveland, 1988). It is a widely acknowledged multi-dimensional rating scale, which was adapted in several studies (Hart, 2006) to obtain workload estimates. In the present study, ratings from six dimensions (mental demand, physical demand, temporal demand, frustration, effort, and performance) were averaged without individual weights. We decided on the unweighted version as it is easier to apply, and the sensitivity seems to be similar as with adding the weighting process (Hart, 2006). Since our approach was to get information about the participants' affective states, we used the Self-Assessment Manikin (SAM). The pictorial assessment is easy to explain and covers essential aspects of a person's affective reaction to a stimulus (Bradley and Lang, 1994). The participant could rate with three items: pleasure (from 1 = satisfied to 9 = unsatisfied), arousal (from 1 = excited to 9 = unexcited), and dominance (from 1 = controlled to 9 = controlling). Additionally, a scale for assessing subjectively experienced effort was deployed. We referred to this scale as "subjective rating scale (SRS)." The scale is an adaption from the SEA scale (Eilers et al., 1986) and measures the subjectively experienced effort for performing the task. We transferred the original scale into a numeric rating scale with equal intervals starting from 1 ("little effort") to 7 ("extreme effort").

2.3. EEG

EEG data were continuously recorded from 14 standard scalp locations according to the 10–20 system (Oz, O1, O2, P3, P4, Pz, Cz, C3, C4, Fz, F3, F4, T3, T4). Since high-density EEG measurements are often time consuming and unpleasant for the test person due to the high number of electrodes, we aimed for a reduced test setup that still delivers informative value. Kumar and Kumar (2016) measured cognitive load by using EEG and found reliable results with 14 channels (similar done by Anderson et al., 2011). An even reduced number of channels was used in studies by Brouwer et al. (2012) and Hogervorst et al. (2014), in which they investigated workload with not more than seven channels successfully. Given that the expected effects get visible at different regions throughout the whole scalp characteristics and considering potential noisy channels, we decided against a minimum but a reduced setup of 14 channels.

2.4. Test Setup

The hearing ability of each participant was tested to ensure a comparable experience of the auditory stimuli for every participant. This was done using an audiometry tool (model MA 33; MAICO Diagnostics GmbH, Berlin, Germany). Baseline instructions and both tasks were deployed in PsychoPy (Peirce et al., 2019) running on a ThinkPad X1 Carbon Ultrabook (Lenovo Ltd., Hongkong). All visual stimuli were presented on a Fujitsu (model: DY24W-7) monitor. The acoustic representation of the mathematical equations was generated by the Win TTS API (German language) and provided to the participants via Sony WHX-1000X M3 headphones in 70 dB SPL. In two out of the three condition blocks, the noise was presented to the participants via four loudspeakers (model PM 0.4) from Fostex (Foster Electric Co., Ltd., Tokyo, Japan) with 76 dB SPL. They were mounted on stands at a height of 1.0 m, placed at a 1.5 m distance to the participants and at a 90° angle to each other. The audio file of the background noise was controlled from a notebook (model Vaio VPCF13C5E; Sony Corporation, Tokyo, Japan) with the expansion card (model HDSP I/O ExpressCard; RME Intelligent Audio Solutions, Audio AG, Haimhausen, Germany). Noise consisted of a combination of very frequent numbers (using the same TTS voice as for the tasks; partially overlapping from different directions), environmental noise (recordings from cars, public streets, cafe chatter; from every direction, not overlapping), and speech snippets (excerpts of German podcasts and news broadcasts; partially overlapping from different directions).

EEG data, stimulus marker (e.g., keypresses of the participant), and stimulus data from PsychoPy was time-synchronized and recorded as one combined data stream via Labstreaminglayer Framework in Lab Recorder running on a ThinkPad X1 Carbon Ultrabook (Lenovo Ltd., Hongkong, China). To access EEG data in Labstreaminglayer, we used g.USBamp App[1] (Pre-release 30.04.2019). For streaming stimulus marker from PsychoPy, we used pylsl[2] (version 1.13.1).

EEG was assessed via wet Ag/AgCl electrodes placed in a head cap, a driver box for 16 channels and the g.USBamp amplifier by g.tec (g.tec medical engineering GmbH, Schiedlberg, Austria).

EEG data and stimulus marker data (e.g., keypresses of the participant) from PsychoPy were time-synchronized and recorded as one combined data stream via Labstreaminglayer Framework in Lab Recorder running on a notebook ThinkPad X1 Carbon Ultrabook (Lenovo Ltd., Hongkong). **Figure 1A** shows the apparatus with all measures, the stimulus presentation, and arrangement of loudspeakers. Also the environmental noise in the setup is illustrated in **Figure 1B**. For analyzing the EEG data, we used the open-source Python package MNE (Gramfort et al., 2013) (version 0.19.1). Statistical analysis was computed with the open-source package Pingouin (Vallat, 2018) (version 0.3.7).

2.5. Procedure

The experiment was conducted in a quiet standardized test room adhering to ITU-T Rec. P.910[3] and P.911[4]. The participants were seated in a chair with a comfortable and upright seating position for the whole duration of the test. In preparation for the EEG measurement, the experimenter placed a flexible cap with plugged-in electrodes on the participant's head and inserted a water-based conductive gel in every electrode. The preparation was completed by equipping the participants with headphones. **Figure 2** illustrates the participants' seating position and the arrangement of applied electrodes on the scalp. To compare individual responses to different sound environments, the participants had to perform the task in three different conditions: *no noise* (quiet environment), *noise* (noise environment), and *noise-canceling* (noise environment with the noise-canceling function of headphones). The order of the three condition blocks was randomized. Additionally, the order of the condition was counterbalanced [6 possible combinations; count of every combination ($M = 4.67$, $SD: 1.03$)]. Each block followed the same procedure: First, the participant had to perform the three subjective measure ratings: NASA-TLX, SAM, and the subjective rating scale. Each questionnaire was presented in a separate view on display in front of them. The ratings were submitted by moving a slider for each item. After the subjective measures, the main task block started and had a duration of 30 min. After the main task, the participants were again asked to rate their state with the subjective measure questionnaires. Between the condition blocks, the participants should rest for 5 min.

2.6. Participants

In total, 29 persons aged 21–64 years ($M = 34.62$, $SD = 12.62$) participated. The gender distribution was nearly balanced (male: 15, female: 14). All participants stated that they were employed, studying at university, or both at that moment. Participants were recruited via a university participant database. All of them confirmed that they had a normal or corrected to normal vision

[1] https://github.com/labstreaminglayer/App-g.Tec/releases/tag/gusbamp

[2] https://github.com/chkothe/pylsl

[3] https://www.itu.int/rec/T-REC-P.910-200804-I/en

[4] https://www.itu.int/rec/T-REC-P.911-199909-I!Cor1/en

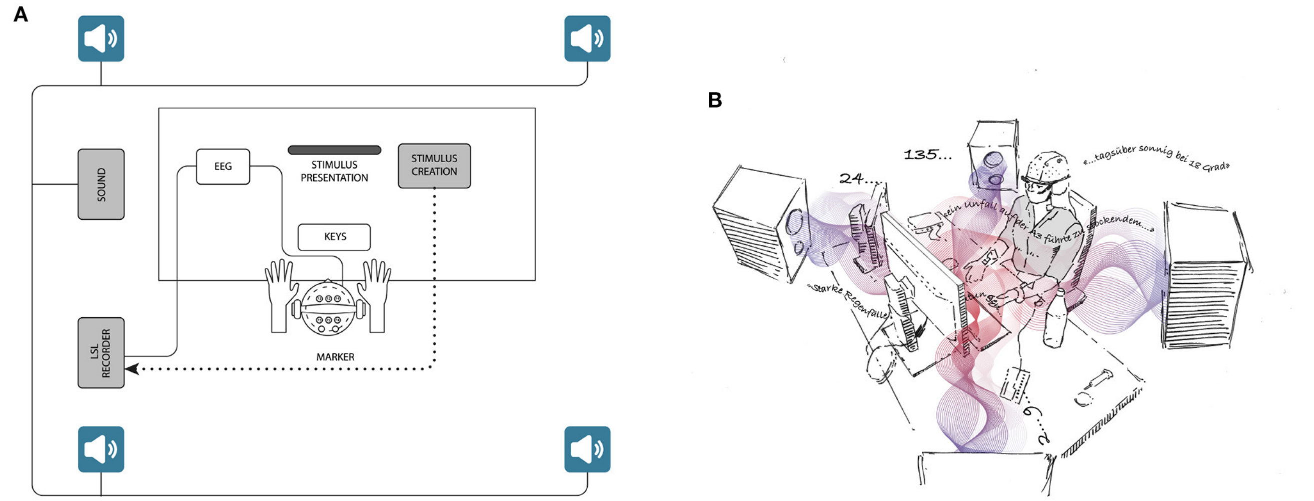

FIGURE 1 | (A) Schematic presentation of the whole test setup, including measurements, stimulus presentation, and the acoustical setup. Four loudspeakers were mounted on stands at the height of 1.0 m, placed at a 1.5 m distance to the participant and in at a 90° angle to each other. **(B)** Illustration of the environmental noise situation in the *noise* and *noise-canceling* condition. It consisted of a combination of frequent numbers, environmental noise, and speech snippets.

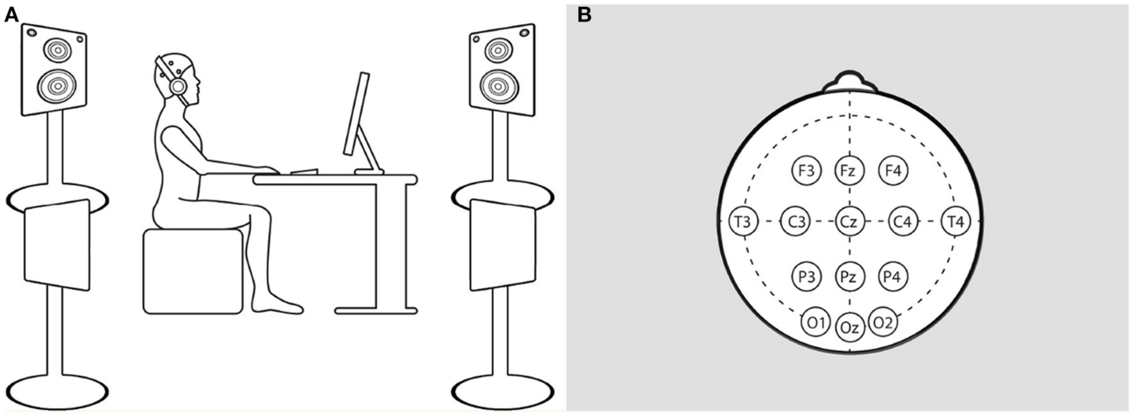

FIGURE 2 | Schematic presentation of the participants' seating position **(A)** and the electrode setup of EEG cap **(B)**. It should be noted that on the left illustration **(A)**, the height of the loudspeakers was adjusted for display purposes but, as mentioned, was actually 1.0 m from the floor. **(B)** The distribution of applied electrodes on the scalp.

and average hearing ability. None of the participants showed a hearing impairment in the performed audiogram (see section 2.5). For participation, the persons got monetary compensation; 15 euro per hour and a bonus depending on their performance in solving the arithmetic tasks (beginning with 50% performance score: 5 euro to max. 10 euro in case of about 100%). This was done to try to get the participants more motivated to gain correct results. The participants were informed about the experiment beforehand, and they agreed to it by signing the informed consent sheet. The study abides by the standards specified in the Declaration of Helsinki. The Ethics Committee of the Faculty IV of Technical University Berlin evaluated the procedure retrospectively and declared that all ethical aspects of the study design follow the Guideline of the German Research Foundation (date of assessment: 17.02.2021; fast track code: FR_2021_01retro).

2.7. Data Analysis

2.7.1. Subjective Measures

The ratings for each item of the subjective measures were collected before and after the task in each test condition block. Values from the beginning of the test condition block were subtracted from ratings after to gain baseline corrected values. The differences between conditions were of interest. By normalizing the values, we measured only changes in ratings induced by the task and the noise situation. All "before" ratings were performed in an equal quiet surrounding. Although the test condition block order was randomized for every participant to avoid sequence effects, we could not exclude that specific block sequences might affect the dependent variable. Therefore, we considered the block order as within factor. Consequential, a two-way repeated measurements ANOVA with the within factors “condition” and “block order” was computed for each questionnaire item (Subjective rating scale: 1 item, SAM: 3 items, and NASA-TLX: 6 items). Significant main effects are reported with a Greenhouse-Geisser corrected *p*-value. The *post-hoc* comparisons were corrected with a Bonferroni–Holm adjustment.

2.7.2. EEG Acquisition and Processing

EEG was continuously recorded from 16 channels (for details, see section 2.5) and sampled with 265*Hz*. The impedance between EEG electrodes and the scalp was kept under 5 kω. One data set had to be excluded from analysis due to an incorrect time synchronization of the stimulus marker stream coming from Psychopy and the measured EEG data stream. In total, data from 28 subjects were included in the EEG analysis. The raw data were filtered with a fir (Filter design: Firwin) band-pass filter from 0.1 to 45*Hz* with a Hamming window. The filter length was 8, 449 samples (33.004*s*). Additionally, a fir (Filter design: Firwin) band stop filter with a Hamming window with 50*Hz* was applied. The aim was to exclude high-frequency line-noise coming from electrical equipment and to remove slow drifts. Channels with extreme noise were detected by visual inspection and removed from the data set and interpolate afterward. Interpolation of bad channels in MNE is done with the spherical spline method (Perrin et al., 1989) that computes the missing signal based on the location and the data of the remaining channels. Afterward, the filtered data were re-referenced to an average reference.

2.7.3. Artifact Rejection With SSP

For removing noise coming from eye movements (EOG) and heart activity (ECG), we chose Signal-Space Projection (SSP) (Uusitalo and Ilmoniemi, 1997). Therefore, we defined one channel that showed the corresponding artifacts’ most characteristic behavior. The computation of the SSP projectors was done on the filtered and re-referenced continuous data of one participant and per condition separately. For the calculation of the EOG projector, the data were band-pass filtered from 1 to 10 Hz (filter design: Firwin; Window: Hann window) to remove DC offset and distinguish blinks from saccades. On the basis of blink detection creating events, the SSP projectors were computed. After that, the data were filtered in one contiguous segment from 1 to 35 Hz [filter design: Firwin; Window: Hamming window; Filter length: 2, 560 samples (10.000*s*)]. For creating the ECG projector, data were band-pass filtered from 5 to 35 Hz [filter design: Firwin; Window: Hann window; Filter length: 2,560 samples (10.000*s*)]. After the computation of the ECG projector, the data were filtered with a band-pass filter [filter design: Firwin; Window: Hann window; Filter length: 2, 560 samples (10.000*s*)] in one contiguous segment from 1 to 35 Hz. The computed EOG and ECG SSP projectors were saved and applied in further processing steps.

2.7.4. Time Frequency Analysis

For analysis in the frequency domain, we segmented the continuous data into epochs. The beginning audio output of the math equation served as stimulus onset. Depending on the number of solved tasks a participant reached per block, the number of trials varied. All these events were used to create epochs. Every trial epoch started 200*ms* before the event and ended 3800*ms* after the stimulus onset. The 3800*ms* after stimulus onset corresponds to the maximum audio output duration of the longest equation. On the epoched data, we applied the formerly calculated SSP projectors to remove ECG and EOG artifacts. We then calculated the periodograms from 0.1 to 45*Hz* using Welch’s method (Welch, 1967) with a sliding hamming window and a window size of 1.0*s* (256 samples), which were then averaged for each channel and epoch. The calculation was done for every participant and each condition block separately. The resulting power spectral densities (PSDs) were normalized by dividing each power value by the total power (per condition block). The correction was done for each participant individually to consider inter-individual variations. Afterward, we aggregated the PSDs of the corresponding EEG frequency bands: delta: 0.1–4 Hz, theta: 4–8 Hz, alpha: 8–12 Hz, beta: 12–30 Hz, and gamma: 30–45 Hz. To compare for differences between conditions, we calculated the mean activity at channels in the region of interest for the corresponding frequency bands: Delta: frontal, central, temporal; Theta: frontal, central; Alpha: parietal, temporal, occipital; Beta: parietal, occipital.

2.7.5. Event-Related Potentials

For investigation of ERPs, epochs of 200 ms before and 800*ms* after the stimulus onset were created. We chose a more prolonged epoch duration. We decided to do so as the stimulus is continuous with multiple information to be processed. Therefore, the characteristic ERP component of interest could occur delayed. We aggregated data from all epochs to the averaged evoked response for each condition block and channel. We calculated the average activation of a specific time of interest for every condition to investigate differences in particular components. As the P300 is known to be prominent at midline electrodes, we included the channels Fz, Cz, and Pz separately in our analysis. The respective time interval considered for P300 is between 250 and 400 ms. Due to the formerly mentioned suspected occurrence delay, we also considered the time interval between 400 and 800 ms in our analysis.

2.7.6. Statistical Analysis of EEG Data

Of the corresponding data, we investigated differences between the three tests conditions. We computed a Mauchly’s to test if the data met the assumption of sphericity. We tested for normal distribution of the data with the Shapiro–Wilk test. If data were not normally distributed, we conducted a Friedman’s test to investigate differences between conditions. If the data met the assumption of the normal distribution, we conducted a repeated-measures ANOVA with the main effect test condition. *Post-hoc* comparisons were calculated with a Wilcoxon signed-rank test.

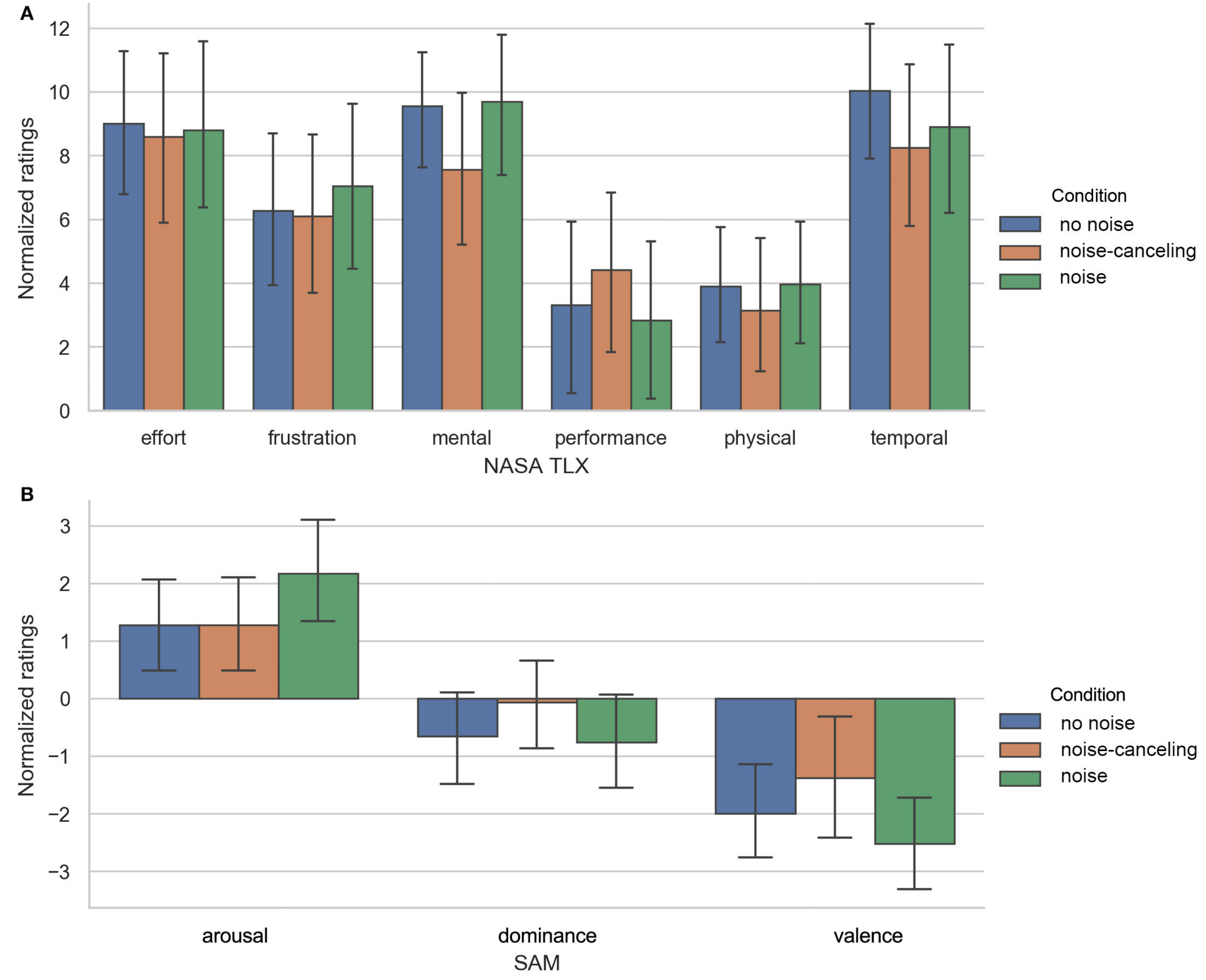

FIGURE 3 | (A) Normalized values for NASA TLX. Ratings from *before* were subtracted from ratings *after*. Error bars show 95% confidence interval. **(B)** Normalized values for SAM. Error bars show 95% confidence interval.

3. RESULTS

3.1. Subjective Data

The differences of ratings between before and after one main task block for NASA TLX and SAM are shown in **Figure 3**. Results of the repeated measures ANOVA of normalized values showed no significant main effect condition but a significant main effect "block order" for SAM item "valence" [$F_{(2,\,56)} = 4.413$; $p = 0.018$; $\eta_p^2 = 0.136$]. The *post-hoc* comparison showed significantly ($p = 0.022$) lower ratings in the second compared to the first block and lower values in the last ($p = 0.032$) compared to the first block.

Although we found no other significant comparisons, on a descriptive level, the normalized values of NASA-TLX, **Figure 3A** show lower absolute values for negatively associated items, e.g., "effort" and "frustration" in the *noise-canceling* condition. This goes along with higher ratings in the positive-related item "performance" (NASA TLX) and lower decreases in "dominance," "valence" (SAM) (see **Figure 3B**) and the rating of the subjective rating scale.

Table 1 shows M and SD of ratings for every item before and after the main task. The values suggest that participants felt more mentally loaded or rather more uncomfortable after the task than before. Out of that, we can assume that the main task seems to be demanding in every condition. This observation is supported by the normalized ratings of the items "mental," "effort," and "temporal" (NASA TLX), which reached the highest scores of the NASA TLX. As the comparisons between conditions are not significant, it seems as if the participants invested similar effort in all test conditions. Interestingly, the participants reported the highest temporal pressure in the *no noise* condition.

3.2. EEG Data

Due to the individual response times for each equation and the therefore varying overall number of quotations per condition, the

TABLE 1 | Mean and standard deviation of subjective measure ratings before and after the performed task.

		Condition											
		no noise				**noise-canceling**				***noise***			
		Before		**After**		**Before**		**After**		**Before**		**After**	
Questionnaire	**Item**	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
NASA TLX	Mental	6.14	5.36	15.69	4.09	9.07	5.92	16.62	3.65	6.97	5.23	16.66	3.94
	Physical	2.66	2.91	6.55	5.68	3.93	4.63	7.07	6.37	2.79	3.03	6.76	5.65
	Temporal	6.86	6.40	16.90	2.62	9.31	6.47	17.55	2.26	7.66	6.37	16.55	3.51
	Performance	8.62	6.29	11.93	5.27	8.24	5.57	12.66	4.70	9.14	5.26	11.97	5.68
	Effort	7.31	6.46	16.31	3.49	8.14	6.36	16.72	2.45	8.41	6.62	17.21	2.97
	Frustration	8.17	6.38	14.45	4.76	8.52	5.84	14.62	3.86	8.03	5.98	15.07	3.99
SAM	Valence	5.86	1.77	3.86	2.15	5.79	1.97	4.41	1.88	6.24	1.62	3.72	1.98
	Arousal	4.24	1.92	5.52	2.03	4.48	1.90	5.76	2.05	3.90	1.76	6.07	1.81
	Dominance	4.72	1.67	4.07	1.85	4.76	1.62	4.69	1.77	4.76	1.46	4.00	1.58
SRS		2.79	1.80	5.55	1.27	3.48	1.94	5.79	1.18	3.21	1.88	5.90	1.21

count of trials per condition considered for analysis varied: *no noise* condition ($M = 253.97$, $SD = 78.17$, $min = 129$, $max = 406$); *noise-canceling* condition ($M = 267.21$, $SD = 82.82$, $min = 111$, $max = 430$), and *noise* condition ($M = 259.62$, $SD = 75.89$, $min = 121$, $max = 410$).

3.2.1. Frequency

From the repeated measures ANOVA results for the frequency band delta, we cannot reject the null hypothesis in favor of the alternate hypothesis [$F_{(2,\ 54)} = 1.488$, $p = 0.235$, $\eta_{\mathrm{p}}^2 = 0.052$]. There is no significant differences between the average values of the frequency band conditions. Similar results were obtained for theta [$F_{(2,\ 54)} = 0.426$, $p = 0.655$, $\eta_{\mathrm{p}}^2 = 0.016$], Beta [$F_{(2,\ 54)} = 0.708$, $p = 0.497$, $\eta_{\mathrm{p}}^2 = 0.026$] and Gamma [$F_{(2,\ 54)} = 1.933$, $p = 0.155$, $\eta_{\mathrm{p}}^2 = 0.067$].

For alpha, the p values of Shapiro–Wilk tests were partly significant ($p < 0.001$, $p < 0.079$, $p = 0.12$) for one of the three levels of the condition-factor. As the assumption normal distribution is violated we computed a Friedman test, which revealed no significant difference between the conditions $\chi^2_{(2)} = 3.5$, $p = 0.174$. **Figure 4** shows the topography plots for every condition and every frequency band, respectively. On a descriptive level, the intensities show the hypothesized behavior of higher power in Theta and Delta, frequency band for *noise* compared to the other two conditions. Also Alpha power spectral density seems to be smallest in the *noise* condition. But these observations aren't supported by statistical significant results. The spectral power densities in Beta and Gamma frequency band show only very minimal changes in intensity between the conditions.

3.2.2. Event-Related Potentials

3.2.2.1. Time Interval 250–400 ms After Stimulus Onset

From the repeated measures ANOVA results for Fz, we cannot reject the null hypothesis in favor of the alternate hypothesis [$F_{(2,54)} = 1.601$, $p = 0.211$, $\eta_{\mathrm{p}}^2 = 0.056$]. We conclude that the mean P300 activation at Fz does not significantly differ between the conditions. The measure of effect size (partial eta squared; $\eta_{\mathrm{p}}^2 = 0.056$) suggests that there is a negligible effect of the conditions on the P300 activation. Mauchly's test of sphericity for Cz revealed a significant p-value ($p = 0.006$). Hence the data did not meet the assumption of sphericity. As the assumptions of the repeated measures ANOVA were violated, we ran a Friedman's test to investigate a main effect of the condition. The Friedman's Test showed a significant difference between the three conditions [$\chi^2_{(2)} = 12.214$, $p = 0.002$]. A pair-wise comparison using Wilcoxon signed-rank tests between the conditions revealed significant differences between the *noise* and the *no noise* condition ($W = 76$, $p = 0.004$), with a higher amplitude for *no noise* ($Mdn = 0.04$) compared to *noise* ($Mdn = -0.10$). This difference is remarkable visible in **Figure 5A**. Also the difference between the *noise-canceling* ($Mdn = -0.58$) and the *no noise* ($Mdn = 0.04$) condition reached statistical significance ($W = 82$, $p = 0.006$). The difference between *noise* and *noise-canceling* is not statistically significant ($W = 198$, $p = 0.909$). We conclude that the data at Pz is non-normally distributed as the p-values of Shapiro–Wilk tests are partly significant ($p < 0.001$, $p < 0.001$, $p = 0.15$) for two of the three levels of the condition factor. As the assumptions of the repeated measures ANOVA were violated, we ran a Friedman's test to investigate a main effect of the condition. The Friedman's test showed no significant difference between the three conditions, $\chi^2_{(2)} = 2$, $p = 0.368$.

3.2.2.2. Time Interval 400–800 ms After Stimulus Onset

From the repeated measures ANOVA results for Fz, we cannot reject the null hypothesis in favor of the alternate hypothesis [$F_{(2,\ 54)} = 1.194$, $p = 0.311$, $\eta_{\mathrm{p}}^2 = 0.042$]. We conclude that the mean activation at Fz does not significantly differ between the conditions.

As the assumptions of the repeated measures ANOVA is violated (evidence for a violation of the assumption of sphericity

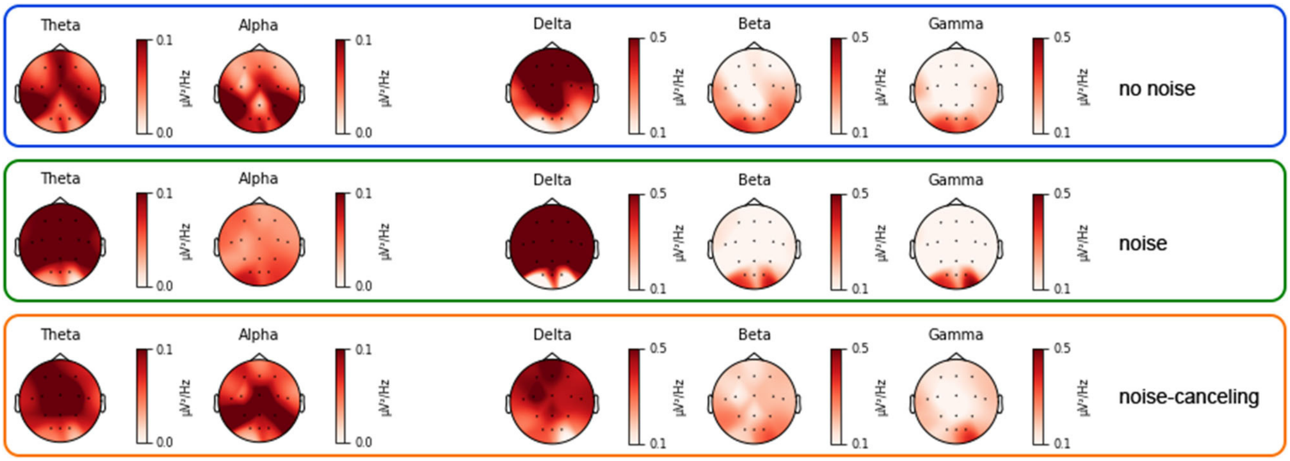

FIGURE 4 | Topographies of the different frequency bands for *noise*, *no noise*, and *noise-canceling* condition. The red colored intensity shows the normalized power spectral density values. The scaling of the intensities was adjusted according to the frequency bands.

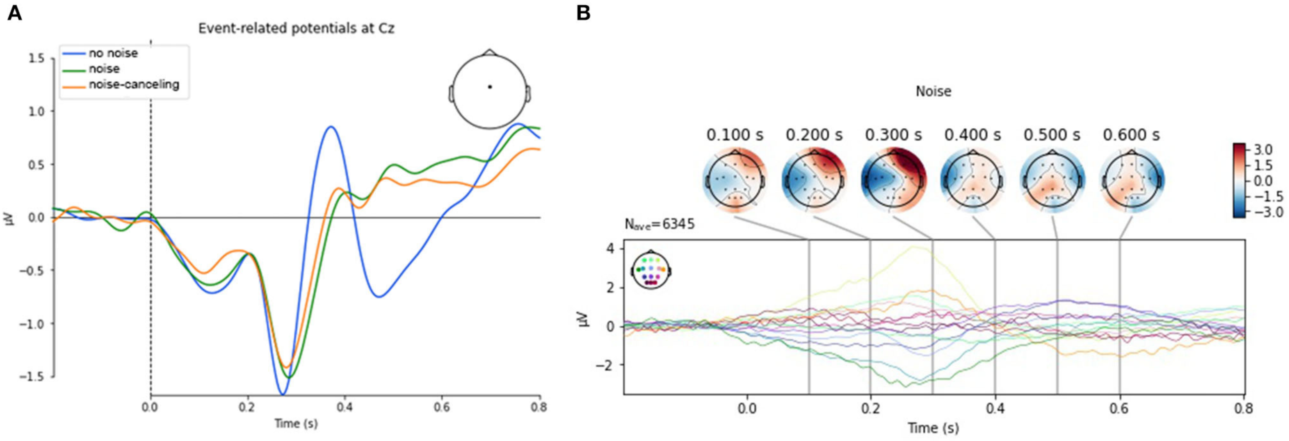

FIGURE 5 | (A) Event-related potentials at Cz. The characteristic signal course of component of P300 the amplitude is higher in *no noise* compared to the other two conditions. **(B)** Topography plot of event-related potentials aggregated from all sensors exemplary for the *noise* condition. Sensor positions are color-coded. The activity range is given from −3 to 3μV. After 400*ms* after stimulus, the activation is increasingly prominent at the centro-parietal area of the scalp that indicates a P300 response.

through Mauchly's test at $p < 0.001$), we ran a Friedman's test to investigate the main effect of the condition at Cz. The Friedman's test shows a significant difference between the three conditions [$\chi^2_{(2)} = 7.358$, $p = 0.025$]. *Post-hoc* tests using a Wilkoxon signed-rank test shows that the activation in the *noise* condition ($Mdn = 0.28$) is higher than in the *no noise* condition ($Mdn = -0.08$). This differences is statistically significant ($Z = 107$, $p = 0.029$). However, the differences between *noise* and *noise-canceling* ($Mdn = 0.17$) are statistically non-significant ($Z = 138$, $p = 0.139$). Likewise, the difference in the *no noise* and the *noise-canceling* is not significant either ($Z = 151$, $p = 0.236$).

As the assumptions of the repeated measures ANOVA for Pz are violated [evidence for a violation of the assumption of sphericity through Mauchly's test ($p < 0.001$)], we ran a Friedman's test to investigate a main effect of the condition. The Friedman's test shows a non-significant difference between the three conditions [$\chi^2_{(2)} = 1.786$, $p = 0.409$]. All results can also be found in **Table 2** for a better overview. **Figure 5B** shows the topographies of ERPs from all sensors exemplary for the *noise* condition. Additional plots of event-related potentials at midline electrodes can be found in **Figures S1–S3**. Topographies of the evoked responses of all electrodes can be found in **Figure S4**.

4. DISCUSSION

4.1. Subjective Assessment

The significant main effect of block order for the item "valence" (SAM) indicates that the participants felt less positive and

TABLE 2 | Statistical results of event-related potentials (ERPs) with different periods and the three midline sensors.

Time	Sensor	Main effect	*Post-hoc*			Median		
			n - nn	***n - nc***	***nn - nc***	**nn**	**n**	**nc**
250–400 ms	Fz	$F_{(2, 54)} = 1.601, p = 0.211, \eta_p^2 = 0.056$						
	Cz	$\boldsymbol{\chi^2_{(2)} = 12.214, p = 0.002}$	$\boldsymbol{W = 76, p = 0.004}$	$W = 198, p = 0.909$	$\boldsymbol{W = 82, p = 0.006}$	0.04	−0.10	−0.58
	Pz	$\chi^2_{(2)} = 2, p = 0.368$						
400–800 ms	Fz	$F_{(2,54)} = 1.194, p = 0.311, \eta_p^2 = 0.042$						
	Cz	$\boldsymbol{\chi^2_{(2)} = 7.358, p = 0.025}$	$\boldsymbol{Z = 107, p = 0.029}$	$Z = 138, p = 0.139$	$Z = 151, p = 0.236$	−0.08	0.28	0.17
	Pz	$\chi^2_{(2)} = 1.786, p = 0.409$						

In case of significant results, values are marked in bold. For significant post-hoc comparisons, median values of every condition are given.

therefore were potentially more uncomfortable in the later stages of the experiment. According to descriptive values of items "temporal" and "mental," the participants seemed to perceive high mental demand and time pressure. Based on descriptive values (higher ratings in positive items and lower ratings in negative items during *noise-canceling*, see section 3.1), we conclude that the subjectively experienced increase of, e.g., mental demand or stress is lower in the *noise-canceling* condition. However, this assumption has to be further substantiated in future studies. Surprisingly, we could not find a similar observation in the *no noise* condition. A possible explanation could be an unexpected psychological influence reported verbally by few participants: in a quiet environment, they felt more uncomfortable than in a noisy environment. Potentially, the feeling of being observed increased the pressure to perform well due to no excuse for making mistakes. The presence of an experimenter is necessary to guide and maintain the experiment procedure especially when working with physiological measures. More importantly, we ensure to minimize the feeling of being observed by placing the experimenters to not look at the participants' screen and by emphasizing that the experimenter did not observed them directly throughout the experiment. This should mimic a general office situation with colleagues nearby but without direct monitoring.

The fact that the differences between the silent (*no noise*) and the noisy (*noise* and *noise-canceling*) conditions indicated in the data did not reach significance could be justified in the relatively short test block duration. A prolongation of each test condition duration could increase the already visible (but not significant) differences between the conditions. Whereby, this adjustment could also cause other influence factors like fatigue, which are hard to control.

4.2. Brain Activity

The analysis of the spectral power of frequency bands delivered no statistical evidence for differences between conditions. The differences in activity between the conditions that are indicated in **Figure 4** reach no significant value.

The analysis of ERPs, however, revealed interesting results. We suggested a decrease in amplitude of the P300 peak amplitude with increasing noise levels. Our results support the assumption that the P300 amplitude is highest in the *no noise* condition at electrode Cz compared to the other two. The difference between the *no noise* and *noise* condition, as well as between *no noise* and *noise-canceling* was found to be statistically significant in the time 250– 400 ms after stimulus onset. After the peak in amplitude in the *no noise* condition, the signal drops rapidly with a negative peak around 500*ms* after stimulus onset. In the other noisy conditions, the activation stays positive and even slightly increases. The difference reaches significance between the *no noise* and the *noise* condition. The formerly mentioned observation that the P300's latency remains larger with a higher mental workload could explain this behavior, which would support our hypothesis of higher mental demand in the *noise* condition. The topography of the ERP, as shown in **Figure 5B**, supports this assumption.

A significant discrimination between *noise-canceling* and *noise* could not be found in the ERPs. This lack of differentiation could have several reasons. The present experiment presented the target stimuli (arithmetic equations) directly on both ears via headphones. The distracting stimuli was present as ambient sound but also detectable for both ears similar. Additionally, the target and distraction stimuli were complex as they consisted of speech and environmental noise, which varies in frequency and inter-stimulus intervals. This frequency and time-varying presentation of stimuli could affect the brain activity-related components in several ways. For example, a jittering in stimulus presentation is known to reduce the peak amplitude of ERPs. Furthermore, the resulting timing effect of ERPs can shift and be later compared to non-jittered stimulus presentations due to jittering in timing. Due to the aforementioned multiple ways how the ERP components can be influenced, an interpretation of a substantial influence is difficult.

Further investigation should focus on more significant discrimination between the two noisy test situations. Our approach resulted in overall high demand (according to ratings in NASA item "mental") in all conditions, making it hard to discriminate between the different experimental manipulations. One reasonable modification would be choosing a visual first task, for example, reading, and adding a second task like detecting specific auditory events ("auditory oddball"; for more details, see Duncan et al., 2009). This would address two modalities and reduce the demand in one channel while keeping the overall demand high. The second task would demand additional attentional resources. This testing paradigm would still be comparable to a real-life working situation (e.g., mobile working on a business trip) in which a person is focused on the work but must not miss important announcements. Additionally, this approach has the advantage of clear differentiation in brain characteristics, becoming more straightforward than the present task.

4.3. Limitations

The study has some already mentioned limitations regarding the setup and the stimuli, which should be addressed in further studies. Regarding the processing of the EEG data, the main focus was on the sensor-based analytic. It would be possible also to consider doing the performed data analysis on source signals obtained by a source reconstruction. Of course, that approach is limited due to the number of used electrodes in the current setup. It would be advisable to increase the number of electrodes in total or focus on specific cortical areas known to show the observed effects.

4.4. Conclusion

The current study aimed to investigate differences in the mental load of participants in varying environmental noise situations. Moreover, it was of interest if noise-canceling headphones help to reduce mental load while focusing on a task. We suggested finding indications that the *noise* condition results in a higher mental load than the other two conditions. The noise-canceling technology was suggested to improve the user's situation in terms of mental load and stress. Additionally, we assumed that in the *no noise* condition, the participants felt less loaded as in the *noise-canceling* condition. We found evidence in subjective data that valence decreases from the beginning to the end of the experiment.

The ERPs of electrical brain activity resulted in significant differentiation between the *no noise* and the other two test conditions. The mentioned adjustment of the setup and the analysis could lead to a stronger delimitation of the two noisy situations. The findings of the current work provide a foundation for the investigation of noise-cancelation and its potential improvement of the working situation in noisy surroundings.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the raw data access is restricted by a disclosure statement with the funder. Requests to access the datasets should be directed to kerstin.pieper@tu-berlin.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Faculty IV—Electrical Engineering and Computer Science at the Technical University of Berlin. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KP, SM, EP, MV, and J-NV-A contributed to the design of the study. KP, RS, and PP implemented the stimuli, carried out data collection, and conducted data analyses. KP wrote the main manuscript text. All authors reviewed and contributed to the final manuscript.

FUNDING

This research was funded by Tampere Wireless Headset Audio Lab, Finland Research Center, Huawei Technologies Oy (Finland) Co., Ltd., Tampere, Finland.

ACKNOWLEDGMENTS

We acknowledge support by the German Research Foundation and the Open Access Publication Fund of TU Berlin.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.771533/full#supplementary-material

REFERENCES

Allison, B. Z., and Polich, J. (2008). Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. *Biol. Psychol.* 77, 277–283. doi: 10.1016/j.biopsycho.2007.10.014

Anderson, C., and Horne, J. A. (2003). Prefrontal cortex: links between low frequency delta EEG in sleep and neuropsychological performance in healthy, older people. *Psychophysiology* 40, 349–357. doi: 10.1111/1469-8986.00038

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., and Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Comput. Graph. Forum* 30, 791–800. doi: 10.1111/j.1467-8659.2011.01928.x

Başar-Eroglu, C., Strüber, D., Schürmann, M., Stadler, M., and Başar, E. (1996). Gamma-band responses in the brain: A short review of psychophysiological correlates and functional significance. *Int. J. Psychophysiol.* 24, 101–112. doi: 10.1016/S0167-8760(96)00051-7

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci. Biobehav. Rev.* 44, 58–75. doi: 10.1016/j.neubiorev.2012.10.003

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Therapy Exp. Psychiatry* 25, 49–59. doi: 10.1016/0005-7916(94)90063-9

Brouwer, A.-M., Hogervorst, M. A., van Erp, J. B. F., Heffelaar, T., Zimmerman, P. H., and Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. *J. Neural Eng.* 9:045008. doi: 10.1088/1741-2560/9/4/045008

Chaouachi, M., and Frasson, C. (2012). "Mental workload, engagement and emotions: an exploratory study for intelligent tutoring systems," in *Intelligent Tutoring Systems*, eds S. A. Cerri, W. J. Clancey, G. Papadourakis, and K. Panourgia (Berlin; Heidelberg: Springer), 65–71. doi: 10.1007/978-3-642-30950-2_9

Choi, H.-H., van Merriënboer, J. J. G., and Paas, F. (2014). Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educ. Psychol. Rev*. 26, 225–244. doi: 10.1007/s10648-014-9262-6

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., et al. (2009). Event-related potentials in clinical research: guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clin. Neurophysiol*. 120, 1883–1908. doi: 10.1016/j.clinph.2009.07.045

Eilers, K., Nachreiner, F., and Hänecke, K. (1986). Entwicklung und überprüfung einer skala zur erfassung subjektiv erlebter anstrengung. *Z. Arbeitswissenschaft* 4, 214–224.

Evans, G. W., and Stecker, R. (2004). Motivational consequences of environmental stress. *J. Environ. Psychol*. 24, 143–165. doi: 10.1016/S0272-4944(03)00076-8

Feinberg, I., Floyd, T. C., and March, J. D. (1987). Effects of sleep loss on delta (0.3–3 Hz) EEG and eye movement density: new observations and hypotheses. *Electroencephalogr. Clin. Neurophysiol*. 67, 217–221. doi: 10.1016/0013-4694(87)90019-8

Friedman, D., Cycowicz, Y. M., and Gaeta, H. (2001). The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. *Neurosci. Biobehav. Rev*. 25, 355–373. doi: 10.1016/S0149-7634(01)00019-7

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Hum. Factors* 40, 79–91. doi: 10.1518/001872098779480578

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci*. 7:267. doi: 10.3389/fnins.2013.00267

Hankins, T. C., and Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat. Space Environ. Med*. 69, 360–367.

Harmony, T., Fernández, T., Silva, J., Bernal, J., Díaz-Comas, L., Reyes, A., et al. (1996). EEG delta activity: an indicator of attention to internal processing during performance of mental tasks. *Int. J. Psychophysiol*. 24, 161–171. doi: 10.1016/S0167-8760(96)00053-0

Hart, S. G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proc. Hum. Fact. Ergon. Soc. Annu. Meet*. 50, 904–908. doi: 10.1177/154193120605000909

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (task load index): results of empirical and theoretical research," in *Advances in Psychology*, Vol. 52 (North-Holland: Elsevier), 139–183. doi: 10.1016/S0166-4115(08)62386-9

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklade, A. L., and Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Hum. Factors* 34, 429–439. doi: 10.1177/001872089203400405

Hogervorst, M. A., Brouwer, A.-M., and van Erp, J. B. F. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Front. Neurosci*. 8:322. doi: 10.3389/fnins.2014.00322

Ishihara, T., and Yoshii, N. (1972). Multivariate analytic study of EEG and mental activity in Juvenile delinquents. *Electroencephalogr. Clin. Neurophysiol*. 33, 71–80. doi: 10.1016/0013-4694(72)90026-0

Jeon, M., Yim, J.-B., and Walker, B. N. (2011). "An angry driver is not the same as a fearful driver: effects of specific negative emotions on risk perception, driving performance, and workload," in *Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI '11* (Salzburg: Association for Computing Machinery), 137–142. doi: 10.1145/2381416.2381438

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev*. 29, 169–195. doi: 10.1016/S0165-0173(98)00056-3

Knoll, A., Wang, Y., Chen, F., Xu, J., Ruiz, N., Epps, J., et al. (2011). "Measuring cognitive workload with low-cost electroencephalograph." in *Human-Computer Interaction –INTERACT 2011*, eds P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler (Berlin; Heidelberg: Springer), 568–571. doi: 10.1007/978-3-642-23768-3_84

Knyazev, G. G. (2012). EEG delta oscillations as a correlate of basic homeostatic and motivational processes. *Neurosci. Biobehav. Rev*. 36, 677–695. doi: 10.1016/j.neubiorev.2011.10.002

Kramer, A. F., Sirevaag, E. J., and Braune, R. (1987). A psychophysiological assessment of operator workload during simulated flight missions. *Hum. Factors* 29, 145–160. doi: 10.1177/001872088702900203

Kumar, N., and Kumar, J. (2016). Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Proc. Comput. Sci*. 84, 70–78. doi: 10.1016/j.procs.2016.04.068

Kuo, S., Mitra, S., and Gan, W. S. (2006). Active noise control system for headphone applications. *IEEE Trans. Control Syst. Technol*. 14, 331–335. doi: 10.1109/TCST.2005.863667

Murata, A. (2005). An attempt To evaluate mental workload using wavelet transform of EEG. *Hum. Factors* 47, 498–508. doi: 10.1518/001872005774860096

Parasuraman, R. (1990). "Event-related brain potentials and human factors research," in *Event-Related Brain Potentials: Basic Issues and Applications* (New York, NY: Oxford University Press), 279–300.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y

Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalogr. Clin. Neurophysiol*. 72, 184–187. doi: 10.1016/0013-4694(89)90180-6

Rasmussen, J. (1979). "Reflections on the concept of operator workload," in *Mental Workload: Its Theory and Measurement*, ed N. Moray (Boston, MA: Springer), 29–40. doi: 10.1007/978-1-4757-0884-4_4

Ray, W. J., and Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science* 228, 750–752. doi: 10.1126/science.3992243

Teigen, K. H. (1994). Yerkes-Dodson: a law for all seasons. *Theory Psychol*. 4, 525–547. doi: 10.1177/0959354394044004

Tiitinen, H. T., Sinkkonen, J., Reinikainen, K., Alho, K., Lavikainen, J., and Näätänen, R. (1993). Selective attention enhances the auditory 40-Hz transient response in humans. *Nature* 364, 59–60. doi: 10.1038/364059a0

Ullsperger, P., Freude, G., and Erdmann, U. (2001). Auditory probe sensitivity to mental workload changes –an event-related potential study. *Int. J. Psychophysiol*. 40, 201–209. doi: 10.1016/S0167-8760(00)00188-4

Uusitalo, M. A., and Ilmoniemi, R. J. (1997). Signal-space projection method for separating MEG or EEG into components. *Med. Biol. Eng. Comput*. 35, 135–140. doi: 10.1007/BF02534144

Vallat, R. (2018). Pingouin: statistics in Python. *J. Open Source Softw*. 3:1026. doi: 10.21105/joss.01026

van Daalen, G., Willemsen, T. M., Sanders, K., and van Veldhoven, M. J. P. M. (2009). Emotional exhaustion and mental health problems among employees doing "people work": the impact of job demands, job resources and family-to-work conflict. *Int. Arch. Occupat. Environ. Health* 82, 291–303. doi: 10.1007/s00420-008-0334-0

Welch, P. (1967). The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust*. 15, 70–73. doi: 10.1109/TAU.1967.1161901

Wickens, C. D. (1979). "Measures of workload, stress and secondary tasks," in *Mental Workload: Its Theory and Measurement*, ed N. Moray (Boston, MA: Springer), 79–99. doi: 10.1007/978-1-4757-0884-4_6

Wickens, C. D. (2008). Multiple resources and mental workload. *Hum. Factors* 50, 449–455. doi: 10.1518/001872008X288394

Wickens, C. D., Isreal, J., and Donchin, E. (1977). "The event related cortical potential as an index of task workload," in *Proceedings of the Human Factors Society Annual Meeting* (Sage, CA; Los Angeles, CA: SAGE Publications), 282–286. doi: 10.1177/107118137702100404

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *Int. J. Aviat. Psychol*. 12, 3–18. doi: 10.1207/S15327108IJAP1201_2

Yu, X., Zhang, J., Xie, D., Wang, J., and Zhang, C. (2009). Relationship between scalp potential and autonomic nervous activity during a mental arithmetic task. *Auton. Neurosci*. 146, 81–86. doi: 10.1016/j.autneu.2008.12.005

Conflict of Interest: EP and MV were employed by the company Tampere Wireless Headset Audio Lab, Finland Research Center, Huawei Technologies Oy (Finland) Co., Ltd., Tampere, Finland. PP was employed by Pupil Labs GmbH, Berlin, Germany.

The remaining authors declare that this study received funding from Tampere Wireless Headset Audio Lab, Finland Research Center, Huawei Technologies Oy (Finland) Co., Ltd., Tampere, Finland. The funder was not involved in collection, analysis, and interpretation of data.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 03 November 2021
doi: 10.3389/fnins.2021.761610

DIBR-Synthesized Image Quality Assessment With Texture and Depth Information

Guangcheng Wang[1], *Quan Shi*[1]*, *Yeqin Shao*[1] *and Lijuan Tang*[2]

[1] *School of Transportation and Civil Engineering, Nantong University, Nantong, China,* [2] *School of Electronics and Information, Jiangsu Vocational College of Business, Nantong, China*

Accurately predicting the quality of depth-image-based-rendering (DIBR) synthesized images is of great significance in promoting DIBR techniques. Recently, many DIBR-synthesized image quality assessment (IQA) algorithms have been proposed to quantify the distortion that existed in texture images. However, these methods ignore the damage of DIBR algorithms on the depth structure of DIBR-synthesized images and thus fail to accurately evaluate the visual quality of DIBR-synthesized images. To this end, this paper presents a DIBR-synthesized image quality assessment metric with Texture and Depth Information, dubbed as TDI. TDI predicts the quality of DIBR-synthesized images by jointly measuring the synthesized image's colorfulness, texture structure, and depth structure. The design principle of our TDI includes two points: (1) DIBR technologies bring color deviation to DIBR-synthesized images, and so measuring colorfulness can effectively predict the quality of DIBR-synthesized images. (2) In the hole-filling process, DIBR technologies introduce the local geometric distortion, which destroys the texture structure of DIBR-synthesized images and affects the relationship between the foreground and background of DIBR-synthesized images. Thus, we can accurately evaluate DIBR-synthesized image quality through a joint representation of texture and depth structures. Experiments show that our TDI outperforms the competing state-of-the-art algorithms in predicting the visual quality of DIBR-synthesized images.

Keywords: depth-image-based-rendering, image quality assessment, colorfulness, texture structure, depth structure

OPEN ACCESS

Edited by:
Ke Gu,
Beijing University of Technology, China

Reviewed by:
Guanghui Yue,
Shenzhen University, China
Suiyi Ling,
Xlaomi Inc., France
Wei Sun,
Shanghai Jiao Tong University, China

****Correspondence:***
Quan Shi
sq@ntu.edu.cn

Specialty section:
This article was submitted to
Perception SciencePerception
Science,
a section of the journal
Frontiers in Neuroscience

Received: *20 August 2021*
Accepted: *11 October 2021*
Published: *03 November 2021*

Citation:
Wang G, Shi Q, Shao Y and Tang L
(2021) DIBR-Synthesized Image
Quality Assessment With Texture and
Depth Information.
Front. Neurosci. 15:761610.
doi: 10.3389/fnins.2021.761610

1. INTRODUCTION

With the advent of the 5G era and the advancement of 3-dimensional display technology, video technology moves from "seeing clearly" to the ultra-high definition and immersive virtual reality era of "seeing the reality." Free-viewpoint videos (FVVs) have broad applications in entertainment, education, medical treatment, military applications for its ability to provide users with visual information of integrity, immersion, and interactivity (Selzer et al., 2019; Yildirim, 2019). Thus, FVV is also regarded as the vital research direction of next-generation video technologies (Tanimoto et al., 2011). Due to hardware conditions, cost, and bandwidth constraints, it is feasible to collect a certain number of viewpoint images in realistic environments. Still, it is often impractical to collect a full range of 360-degree viewpoint images. Therefore, it is necessary to synthesize virtual viewpoint images from existing reference viewpoint images by relying on virtual viewpoint synthesis techniques (Wang et al., 2020, 2021; Li et al., 2021a; Ling et al., 2021).

Because depth-image-based-rendering (DIBR) technologies only require a texture image and its corresponding depth map to generate the image at any viewpoint, it becomes the most popular virtual viewpoint synthesis technique (Luo et al., 2020). Unfortunately, because the performance of existing DIBR algorithms is not perfect, some distortions are often introduced during the warping and rendering processes, as shown in **Figure 1**. The quality of DIBR-synthesized images directly influences the visual experience in FVV-related applications, determining whether these applications can be successfully put into use. Hence, studying the quality evaluation methods for virtual viewpoint synthesis has important practical significance.

Image quality assessment (IQA) has been a crucial frontier research direction in image processing in recent decades. Massive IQA algorithms for natural images have been proposed, divided into full-reference, reduced-reference, and no-reference according to the required full, partial, and no information of the reference image. For instance, Wang et al. (2004) proposed a full-reference IQA metric based on comparing the structural information between the reference and distorted images, namely Structural SIMilarity (SSIM). Zhai et al. (2012) quantify psychovisual quality of images based on free-energy interpretation of cognition in brain theory. Min et al. (2018) proposed a pseudo-reference image (PRI) based IQA framework, which is different from the traditional full-reference IQA framework. The standard full-reference IQA framework assumes that the reference image is a high visual quality image. In contrast, the framework proposed by Min et al. assumes that the reference image suffers the most severe distortion in related applications. Based on the PRI-based IQA framework, Min et al. measures the similarity between the distorted image's and the PRI's structures to estimate blockiness, sharpness, and noisiness.

In recent years, researchers have realized that IQA algorithms for natural images have difficulty in estimating the geometric distortion prevalent in DIBR-synthesized images. For this problem, Bosc et al. (2011) calculated the difference map between the synthesized image and the reference image based on SSIM and adopted a threshold strategy to detect the disoccluded area in the synthesized image. Then, the quality score of a synthesized image is obtained by measuring the average structural similarity of the disoccluded region. Conze et al. (2012) used SSIM to generate a similarity map between the reference image and the synthesized image and further extracted the texture, gradient direction, and image contrast weighting maps based on the obtained similarity map to predict the synthesized image quality score. Stankovic et al. designed the Morphological Wavelet Peak signal-to-noise ratio (MW-PSNR) for assessing the synthesized image quality (Dragana et al., 2015b). Meanwhile, the authors proposed a simplified version of MW-PSNR called MW-PSNR-reduce (Dragana et al., 2015b), which only uses the PSNR value of the higher-level scale image to predict the synthesized image quality. For better performance, Stankovic et al. adopted morphological pyramid decomposition to replace the morphological wavelet decomposition in the above-mentioned MW-PSNR (Dragana et al., 2015b) and MW-PSNR-reduce (Dragana et al., 2015b), which successively produce MP-PSNR (Dragana et al., 2015a) and MP-PSNR-reduce (Dragana et al., 2016). Although these methods for the synthesized images have better performance than the IQA algorithms devised for natural images, their performance still misses the actual requirements.

Over the past few years, researchers have been aware of a close relationship between quantifying the local geometric distortion and the quality assessment of DIBR-synthesized images and the screen content images (Gu et al., 2017b). Gu et al. (2018a), Li et al. (2018b), Jakhetiya et al. (2019), and Yue et al. (2019) have arranged the idea in the design of DIBR-synthesized IQA methods, respectively. In literature (Gu et al., 2018a), Gu et al. adopted an autoregression (AR)-based local description operator to estimate the local geometric distortion. Specifically, the authors measure the local geometric distortion by calculating the reconstruction error between the synthesized image and its AR-based prediction. In literature (Jakhetiya et al., 2019), assumed that the geometric distortion behavior is similar to the outliers and further proved this hypothesis using ROR statistics based on the three-Sigma rule. Based on this view, the authors highlight the local geometric distortion through a median filter and further fuse these prominent distortions to assess the synthesized image quality.

Moreover, based on the local geometric distortion measurement, Yue et al. (2019)'s and Li et al. (2018b)'s methods introduce global sharpness estimation to predict the synthesized image quality. Yue et al. (2019) considered three major DIBR-related distortions, including the disoccluded

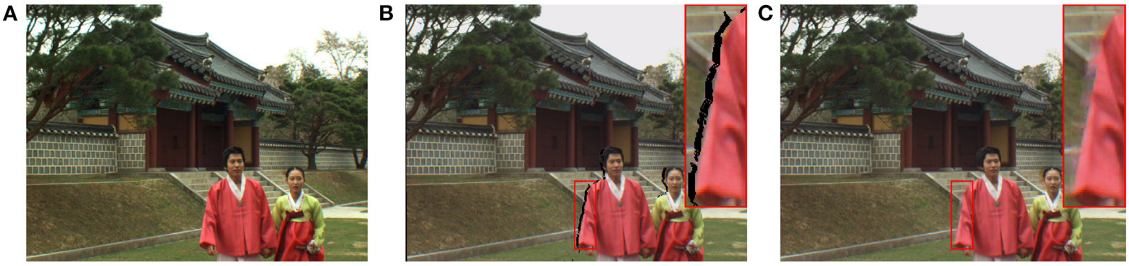

FIGURE 1 | Examples of the local geometric distortion and the color deviation distortion in the synthesized images. **(A)** is the ground-truth image. **(B,C)** are the synthesized images, which includes the local geometric distortion and the color deviation distortion compared to the ground-truth image.

region, the stretching region, and global sharpness. The authors first detect disoccluded regions by analyzing the local similarity. Then, the stretching regions are determined by combining the local similarity analysis and a threshold solution. Finally, the authors measure inter-scale self-similarity to estimate global sharpness. Li et al. (2018b) designed a SIFT-flow warping based disoccluded region detection algorithm. Then, the geometric distortion is measured by combining with the size and distortion intensity of local disoccluded areas. Moreover, a reblurring-based solution is developed to capture blur distortion. We find two critical problems from the above-mentioned DIBR-synthesized IQA methods. First, these methods ignore the influence of color deviation distortion on the visual quality of DIBR-synthesized images. Second, These methods only focus on estimating the geometric distortion and blur distortion from textured images without considering the local geometric distortion's adverse effects on the synthesized image's depth structure.

Inspired by these findings, we present a newly synthesized image quality assessment metric that combines Texture and Depth Information, namely TDI. Specifically, we adopt the colorfulness module proposed by Hasler and Suesstrunk (2003) to extract the color features of a synthesized image and its reference image (i.e., the ground-truth image) and then calculate the feature error to estimate the color deviation distortion. We perform discrete wavelet transform on the texture information of the synthesized image and its reference image and further calculate the similarity of the high-frequency subbands of a pair of synthesized and reference images. The similarity result is used to estimate the local geometric distortion and global sharpness. Meanwhile, we use SSIM to compute the structural similarity between the depth maps of a pair of synthesized and reference images to represent the effects of the local geometric distortion and blur distortion on the depth of field of the synthesized image. In addition, TDI develops a linear weighting scheme to fuse the obtained features. We verify the performance of our TDI metric on the public IRCCyN/IVC DIBR-synthesized image database Bosc et al. (2011), and the experimental results prove that our TDI metric performs better than the competing state-of-the-art (SOTA) IQA algorithms. Compared with the existing works, the highlights of the proposed algorithm mainly include two aspects: (1) we integrate the color deviation distortion caused by DIBR algorithms into the development of DIBR-synthesized view quality perception model; (2) This paper estimates the quality degradation brought by the local geometric distortion and blur distortion from the texture and depth information of the synthesized view.

The remaining chapters of this paper are organized as follows. Section 2 introduces the proposed TDI in detail. Section 3 compares our TDI with SOTA IQA metrics for natural and DIBR-synthesized images. Section IV summarizes the whole research.

2. PROPOSED METHOD

The design philosophy of our TDI is based on quantifying the local geometric distortion, global sharpness, and color deviation distortion. After extracting the corresponding features, a linear weighting strategy fuses the above features to infer the final quality score. **Figure 2** shows the framework of the proposed TDI.

2.1. Color Deviation Distortion Estimation

The human visual system (HVS) is susceptible to color, so the measurement of color deviation distortion has a direct impact on the visual experience (Gu et al., 2017a; Liao et al., 2019). As shown in **Figure 1**, compared to the high-quality reference image, the synthesized image has the color deviation distortion. However, since it is not the main distortion in the synthesized image, most existing DIBR-synthesized IQA algorithms ignore the impact of the color deviation distortion on the visual experience. To more accurately evaluate the synthesized image quality, this paper takes the measurement of color deviation distortion into account in the proposed TDI metric. In the literature (Hasler and Suesstrunk, 2003), Hasler and Suesstrunk devised a highly HVS-related image colorfulness estimation based on psychophysical category scale experiments. The image colorfulness estimation model is specifically defined as follows:

$$C = (\sigma_{rg}^2 + \sigma_{yb}^2)^{\frac{1}{2}} + 0.3 \cdot (\mu_{rg}^2 + \mu_{yb}^2)^{\frac{1}{2}}, \tag{1}$$

where σ_{rg}, σ_{yb}, μ_{rg} and μ_{yb} are the variance and mean of the *rg* and *yb* channels, respectively. The calculation method of *rg* and *yb* channels is shown in formula 2.

$$rg = R - G, yb = \frac{1}{2}(R + G) - B \tag{2}$$

Then, we calculate the absolute value of the colorfulness difference between a synthesized image and its associated reference image (i.e., formula 5) as the quantized result of the color deviation distortion that existed in the synthesized image.

$$Q_1 = |C_{syn} - C_{ref}|, \tag{3}$$

where C_{syn} and C_{ref} represent the colorfulness of the synthesized image and its reference image, respectively.

2.2. Local Geometric Distortion and Global Sharpness Measurement

The proposed TDI extracts structural features from the texture image and its corresponding depth image and designs a linear pooling strategy for information fusion to achieve a more accurate measurement of the local geometric distortion and global sharpness. This part explains in detail how TDI extracts structure features from texture and depth images.

2.2.1. Structure Feature Extracting From Texture Domain

We first use the Cohen-Daubechies-Fauraue 9/7 filter (Cohen et al., 1992) to perform discrete wavelet transform on the synthesized and reference images. **Figure 3** shows some examples of high-frequency wavelet subbands (i.e., HL, LH, and HH subbands) of two synthesized images and their reference image. From **Figure 3**, we observe that the geometric distortion regions (such as the red box area) of the synthesized and reference

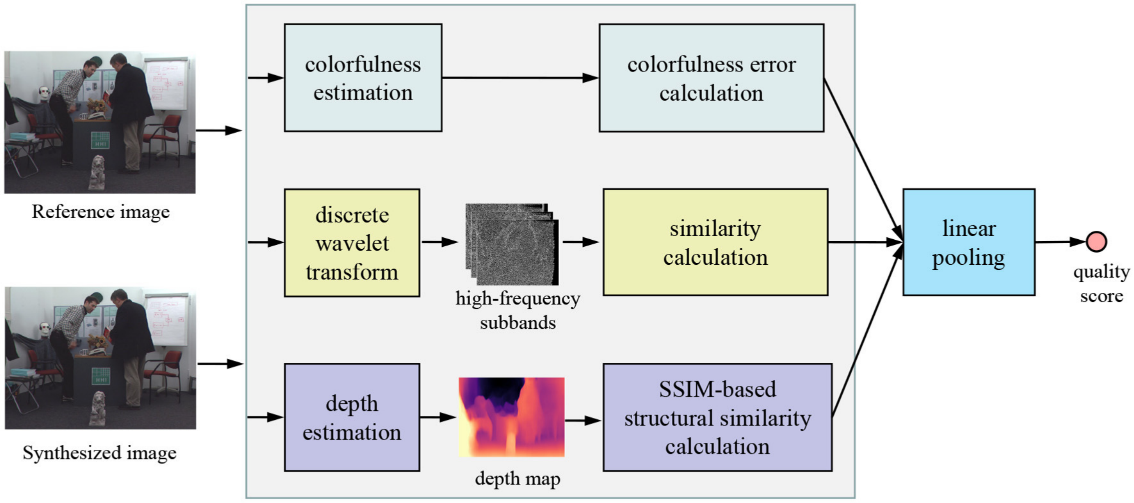

FIGURE 2 | Framework of the proposed TDI metric for predicting the quality of DIBR-synthesized images.

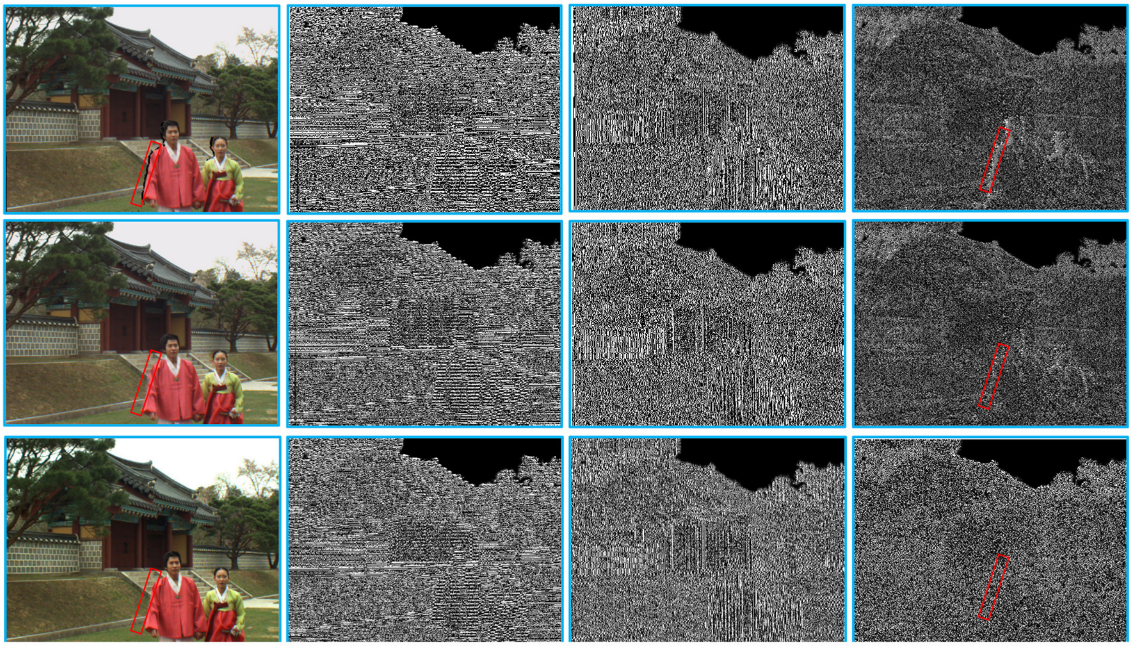

FIGURE 3 | Examples of the high-frequency wavelet subbands (i.e., HL, LH, and HH subbands) of two synthesized images and their reference image. From left to right, the images in each row are a synthesized/reference image and its corresponding HL, LH, and HH wavelet subbands. Note that the synthesized image of the first row has only the warping process.

images in the HH subbands differ significantly. Motivated by this, we measure the local geometric distortion by computing the similarity between the HH subbands of a pair of synthesized and reference images, which is defined as follows:

$$Q_2 = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{2 \cdot HH_{syn}(i) \cdot HH_{ref}(i) + \epsilon}{HH_{syn}(i) + HH_{ref}(i) + \epsilon}\right], \tag{4}$$

where HH_{syn} and HH_{ref} represent the HH subbands of a synthesized image and its corresponding reference image. i and N are the pixel index and the number of pixels of a given image, respectively. A small constant ϵ avoids the risk of zero denominator. Moreover, since blur distortion usually causes loss of high-frequency information in images, the energy of high-frequency wavelet subbands has been widely used for no-reference image sharpness estimation (Vu and Chandler, 2012; Wang et al., 2020). Therefore, the developed similarity between

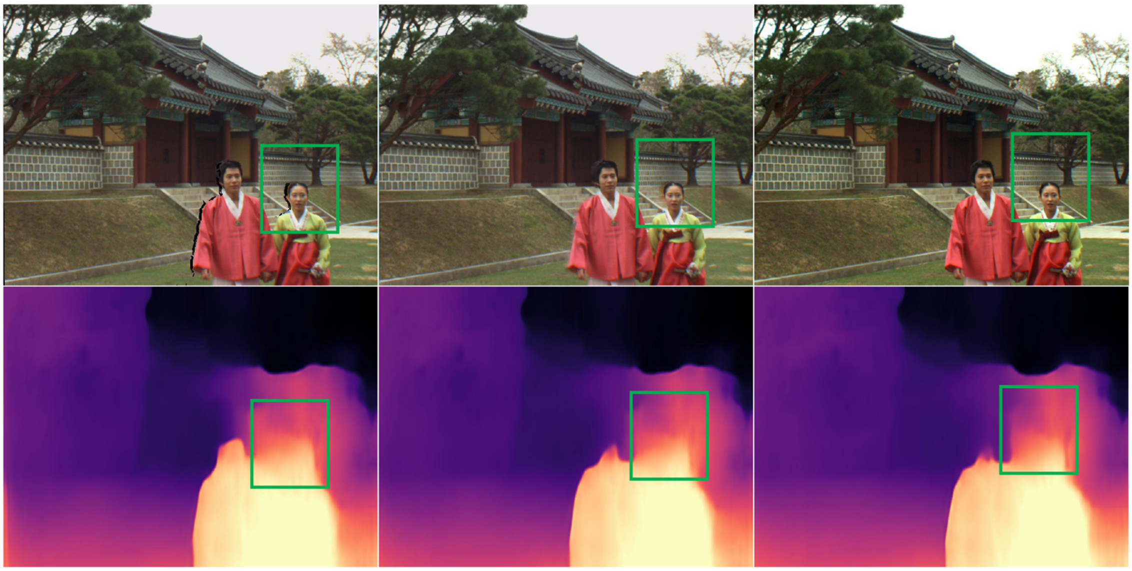

FIGURE 4 | Examples of the depth maps of two synthesized images and their reference image. From top to bottom, the images in each column are a synthesized/reference image and its corresponding depth map. Note that the synthesized image of the first column has only the warping process.

the HH subbands of the synthesized image and its reference image can also effectively estimate the global sharpness of the DIBR-synthesized image.

2.2.2. Structure Feature Extracting From Depth Domian

Considering that local geometric distortion and global sharpness damage the structural information of the synthesized view in the texture domain and affect the depth structure of the synthesized view. Thus, we measure the structural similarity between the depth maps of a pair of synthesized and reference views in the depth domain to estimate the depth degradation introduced by the local geometric distortion and blur distortion. The depth map prediction algorithm computes the depth map at the virtual viewpoint. At present, massive deep learning-based depth image estimation algorithms have been proposed (Atapour-Abarghouei and Breckon, 2018; Li et al., 2018a; Zhang et al., 2018; Godard et al., 2019). In our TDI, we employ Clément Godard's depth prediction network for estimating the depth maps of the DIBR-synthesized image and its reference image. **Figure 4** shows some examples of the depth maps of two synthesized images and their ground-truth image estimated by Clément Godard's method. From the green box area in **Figure 4**, it can be easily observed that the local geometric distortion is very destructive to the depth structure of the synthesized image. So the geometric distortion contained in a synthesized image can be effectively estimated by measuring the structural similarity between the depth maps of a pair of synthesized and reference images. In particular, the structural similarity between the depth maps of a synthesized image and its reference image is computed as follows:

$$Q_3 = \frac{1}{N}\sum_{i=1}^{N}(SSIM(D_{syn}(i), D_{ref}(i))), \tag{5}$$

where D_{syn} and D_{ref} represent the depth maps of a synthesized image and its reference image predicted by Clément Godard's algorithm. SSIM is an image quality evaluation index based on the structural similarity between the reference and distorted images (Wang et al., 2004; Jang et al., 2019).

2.3. Linear Pooling Scheme

To evaluate the visual quality of DIBR-synthesized views more efficiently, this paper extracts three features from the texture and depth domains to estimate the color deviation distortion, the local geometric distortion, and global sharpness. Since the features Q_1, Q_2, and Q_3 are complementary, we propose a novel linear pooling scheme to fuse the texture and depth information to form the final TDI model. A smaller Q_1 value shows the difference between the colorfulness of the synthesized image and its reference image is smaller. That is, the quality of the synthesized image is higher. The Q_2 and Q_3 are the texture and depth structure similarity between a pair of synthesized and reference images, respectively. The values of Q_2 and Q_3 are higher, indicating that the quality of a pair of synthesized and reference views is more similar. That is, the quality of the synthesized image is better. With this fact, a linear pooling scheme is developed to fuse the obtained features, which is defined as follows:

$$S = -\frac{\alpha}{1+\alpha+\beta}\cdot Q_1 + \frac{1}{1+\alpha+\beta}\cdot Q_2 + \frac{\beta}{1+\alpha+\beta}\cdot Q_3, \tag{6}$$

where the parameters α and β are used to adjust the contribution of Q_1, Q_2, and Q_3. In section 3, we detail the selection of parameters α and β.

TABLE 1 | Performance comparison of 21 SOTA IQA measures on the IRCCyN/IVC database (Bosc et al., 2011).

Metric	Type	SRCC	PLCC	RMSE
PSNR	Natural Images	0.3095	0.3976	0.6109
SSIM (Wang et al., 2004)	Natural Images	0.4368	0.4850	0.5823
IW-SSIM (Wang and Li, 2011)	Natural Images	0.4053	0.5831	0.5409
ADD-SSIM (Gu et al., 2016)	Natural Images	0.4672	0.5512	0.5556
PSIM (Gu et al., 2017a)	Natural Images	0.4576	0.5315	0.5640
NIQE (Mittal et al., 2013)	Natural Images	0.3739	0.4374	0.5987
IL-NIQE (Zhang et al., 2015)	Natural Images	0.5348	0.4998	0.5767
ARISM (Gu et al., 2015)	Natural Images	0.3728	0.3994	0.6104
BIQME (Gu et al., 2018b)	Natural Images	**0.6770**	**0.7271**	**0.4571**
MW-PSNR (Dragana et al., 2015b)	DIBR-synthesized Images	0.5757	0.5622	0.5506
MP-PSNR (Dragana et al., 2015a)	DIBR-synthesized Images	0.6227	0.6174	0.5238
MP-PSNR-reduce (Dragana et al., 2016)	DIBR-synthesized Images	0.6634	0.6772	0.4899
NIQSV+ (Tian et al., 2018)	DIBR-synthesized Images	0.6668	0.7114	0.4679
APT (Gu et al., 2018a)	DIBR-synthesized Images	0.7157	0.7307	0.4546
CLGM (Yue et al., 2019)	DIBR-synthesized Images	0.6528	0.6750	0.4620
STD (Wang et al., 2021)	DIBR-synthesized Images	0.7729	0.7901	0.4082
LMS (Zhou et al., 2019)	DIBR-synthesized Images	0.8050	0.7690	0.3940
IDEA (Li et al., 2021b)	DIBR-synthesized Images	—	0.7796	—
GANs-NRM (Ling et al., 2020)	DIBR-synthesized Images	**0.8070**	**0.8260**	**0.3860**
OUT (Jakhetiya et al., 2019)	DIBR-synthesized Images	0.7036	0.7678	0.4266
TDI (Pro.)	DIBR-synthesized Images	**0.7905**	**0.7992**	**0.4002**

The best performance in each type is highlighted in bold.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this part, we construct experiments on the IRCCyN/IVC database to test the performance of the proposed TDI method and other SOTA IQA algorithms.

3.1. Experimental Setup

3.1.1. Competing IQA Metrics

In this paper, we collect twenty SOTA IQA algorithms for natural images and DIBR-synthesized images as competing algorithms. The competing IQA metrics designed for natural images include PSNR, SSIM (Wang et al., 2004), IW-SSIM (Wang and Li, 2011), ADD-SSIM (Gu et al., 2016), PSIM (Gu et al., 2017a), NIQE (Mittal et al., 2013), ILNIQE (Zhang et al., 2015), ARISM (Gu et al., 2015), and BIQME (Gu et al., 2018b). The competing IQA methods devised for DIBR-synthesized images consist of MW-PSNR (Dragana et al., 2015b), MP-PSNR (Dragana et al., 2015a), MP-PSNR-reduce (Dragana et al., 2016), NIQSV+ (Tian et al., 2018), APT (Gu et al., 2018a), CLGM (Yue et al., 2019), STD (Wang et al., 2021), LMS (Zhou et al., 2019), IDEA (Li et al., 2021b), GANs-NRM (Ling et al., 2020), and OUT (Jakhetiya et al., 2019).

3.1.2. Testing Dataset

In this paper, we test the performance of the proposed TDI metric and twenty SOTA IQA algorithms on the public IRCCyN/IVC database (Bosc et al., 2011). The IRCCyN/IVC DIBR-synthesized image database contains 12 reference images and its corresponding 84 synthesized images generated via seven DIBR algorithms. In the subjective experiment, the authors adopt the absolute category rating-hidden reference method to mark DIBR-synthesized images. The images in the IRCCyN/IVC image dataset are from three free-view sequences (i.e., “Book Arrival,” “Lovebird,” and “Newspaper”) with a resolution of 1,024 × 768.

TABLE 2 | Ablation experiments about the proposed components.

Metric	SRCC	PLCC	RMSE
Q1	0.4412	0.4971	0.5777
Q2	0.6126	0.6133	0.5259
Q3	0.4470	0.5346	0.5627
TDI (overall model)	0.7905	0.7992	0.4002

3.1.3. Performance Benchmarking

In this paper, three commonly used indicators, including Spearman Rank-order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE), are used to evaluate the performance of the proposed TDI metric and other competing IQA algorithms devised for natural images and DIBR-synthesized images. The SRCC index evaluates the monotonic consistency between subjective scores and objective scores predicted by IQA metrics. The PLCC and RMSE indicators evaluate the accuracy of the scores predicted by IQA algorithms. The larger values of SRCC and PLCC, and the smaller value of RMSE, indicate the

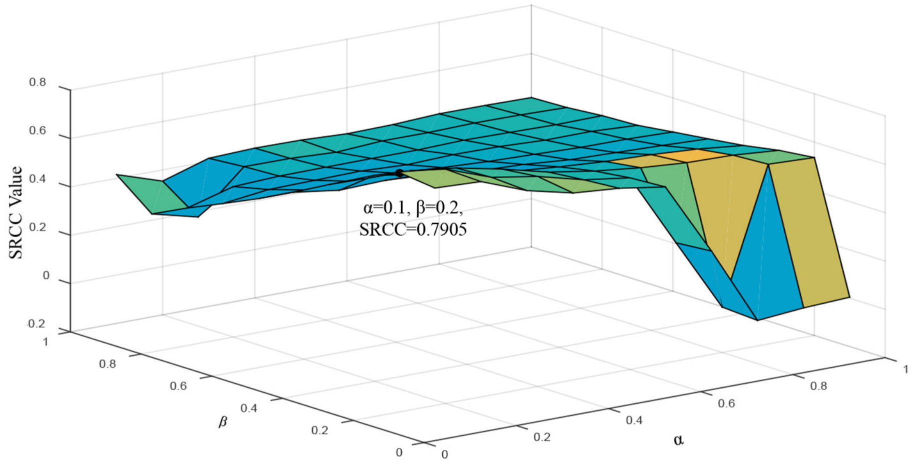

FIGURE 5 | The impact of the parameters α and β on the robustness of the proposed TDI metric.

performance of the corresponding IQA metric is better. The PLCC is defined as follows:

$$PLCC = \frac{\sum_i (a_i - \bar{a})(l_i - \bar{l})}{\sqrt{\sum_i (a_i - \bar{a})^2 \sum_i (l_i - \bar{l})^2}}, \tag{7}$$

where a_i and $\bar{a}$ are the estimated quality score of the i-th synthesized image and the average value of all a_i, respectively. l_i and $\bar{l}$ are the subjective quality label of the i-th synthesized image and the average value of all l_i, respectively. The SRCC is computed as follows:

$$SRCC = 1 - \frac{6\sum_{q=1}^{Q} d_q^2}{Q(Q^2-1)}, \tag{8}$$

where Q is the number of pairs of predicted quality scores and subjective quality labels. d_q represents the ranking difference between the predicted quality scores and the subjective quality labels in each group. Before calculating the above indicators, we need to map the quality scores of all IQA methods to the same range through a non-linear logistic function (Min et al., 2020a,b), which is defined as follows:

$$f(x) = \tau_1\left(\frac{1}{2} - \frac{1}{1+e^{\tau_2(x-\tau_3)}}\right) + \tau_4 x + \tau_5, \tag{9}$$

where τ_1, τ_2, τ_3, τ_4, and τ_5 are the fitting parameters. x and $f(x)$ are the quality scores predicted by IQA algorithms and their corresponding non-linear mapping results, respectively.

3.2. Performance Comparisons With SOTA IQA Metrics

As shown in **Table 1**, our TDI metric achieves SRCC value of 0.7905, PLCC value of 0.7992, and RMSE value of 0.4002 on the IRCCyN/IVC dataset, which outperforms most competing IQA metrics designed for natural images and DIBR-synthesized images. In terms of SRCC, the performance of our proposed method is very close to that of the best-performing GANs-NRM. From **Table 1**, we observe two important conclusions:

1. The performance of the IQA algorithms for natural images on IRCCyN/IVC is far inferior to the IQA methods designed for DIBR-synthesized images. The SRCC, PLCC, and RMSE values of the best BIQME (Gu et al., 2018b) on the IRCCyN/IVC dataset (Bosc et al., 2011) are 0.6770, 0.7271, and 0.4571, respectively, and its SRCC value still does not reach 0.7. Regarding SRCC, PLCC and RMSE, the proposed TDI metrics are 16.77, 9.92, and 12.45% higher than the top BIQME methods, respectively.
2. The APT (Gu et al., 2018a) and OUT (Jakhetiya et al., 2019) metrics, existing best performing IQA algorithms on the IRCCyN/IVC (Bosc et al., 2011) database based on geometric distortion quantization, achieve SRCC value of 0.7157, PLCC value of 0.7678, and RMSE value of 0.4266, respectively. Our proposed TDI metric increases the values of SRCC, PLCC, and RMSE by 10.45, 4.09, and 6.19% on this result. Experiments show that the proposed TDI metric, combining colorfulness, texture structure, and depth structure, can efficiently predict DIBR-synthesized image quality.

3.3. Ablation Study

In this part, we conduct some ablation experiments to verify the contributions of the proposed key components (i.e., Q_1, Q_2, and Q_3). **Table 2** shows the test results of the components Q_1, Q_2, Q_3, and the overall module on the public IRCCyN/IVC data set. From the results, we observe the performance of the overall TDI model is far superior to each component, which shows that the proposed sub-modules can complementally evaluate the quality of the synthesized view. That is, the fusion of texture and depth information is of great significance to the view synthesis quality perception. Moreover, we further analyze the influence of the parameters α and β in equation (6) on the robustness of the proposed TDI metric, and the experimental results are shown in **Figure 5**. Obviously, when the parameters α and β are smaller, the performance of the proposed TDI metric is better, that is, compared to the components Q_1 and Q_3, the component Q_2 is more important, which is also in line with the test results in **Table 2**. According to the robustness analysis, the parameters α and β are set to 0.1 and 0.2, respectively, to optimize the proposed TDI module.

3.4. Applications in Other Fields

With the rapid development of computer vision, the three-dimensional-related technologies can be implemented in numerous practical applications. The first application is abnormality detection in industry, especially the smoke detection in industrial scenarios which has received an amount of attention from researchers in recent years (Gu et al., 2020b, 2021b; Liu et al., 2021). The process of abnormality detection relies on images, therefore combining three-dimensional technology with this can make the image acquisition equipment obtain a more accurate, intuitive and realistic image information, so as to enable the staff to monitor the abnormal situation in time and then avoid bad things from happening. The second application is atmospheric pollution monitoring and early warning (Gu et al., 2020a, 2021a; Sun et al., 2021). The three-dimensional visualized images contain more detailed information, thus enabling efficient and accurate air pollution monitoring. The third application field is three-dimensional vision and display technologies (Gao et al., 2020; Ye et al., 2020). Compared with the ordinary two-dimensional screen display, three-dimensional technology can make the image is no longer confined to the plane of the screen (Sugita et al., 2019), as if it can come out of the screen, so that the audience has a feeling of immersion. The fourth application is road traffic monitoring (Ke et al., 2019). Three-dimensional technology can monitor the traffic flow information of major intersections in an all-round and intuitive way. All in all, there are several advantages of DIBR technology, so it is necessary to extend this technology to different fields.

4. CONCLUSION

This paper presents a novel DIBR-synthesized image quality assessment algorithm based on texture and depth information fusion, dubbed as TDI. First, in the texture domain, we evaluate the visual quality of the synthesized images by extracting the differences in colorfulness and HH wavelet subband between the synthesized image and its reference image. Then, in the depth domain, we estimate the impact of the local geometric distortion on the quality of the synthesized views by calculating the structural similarity between the depth maps of a pair of synthesized and reference views. Finally, a linear pooling model is developed to fuse the above features to predict DIBR-synthesized image quality. Experiments on the IRCCyN/IVC database show that the proposed TDI algorithm outperforms each sub-module and most competing SOTA image quality assessment methods designed for natural and DIBR-synthesized images.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

QS and YS designed and instruct the experiments. GW wrote the code for the experiments. GW, QS, and LT carried out the experiments and wrote the manuscript. YS and LT collected and analyzed the experiment data. All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This research was funded by the National Natural Science Foundation of China, grant no. 61771265.

REFERENCES

Atapour-Abarghouei, A., and Breckon, T. P. (2018). "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 2800–2810.

Bosc, E., Pepion, R., Le Callet, P., Koppel, M., Ndjiki-Nya, P., Pressigout, M., et al. (2011). Towards a new quality metric for 3-d synthesized view assessment. *IEEE J. Sel. Top. Signal. Process*. 5, 1332–1343. doi: 10.1109/JSTSP.2011.2166245

Cohen, A., Daubechies, I., and Feauveau, J.-C. (1992). Biorthogonal bases of compactly supported wavelets. *Commun. Pure Appl. Math*. 45, 485–560. doi: 10.1002/cpa.3160450502

Conze, P.-H., Robert, P., and Morin, L. (2012). Objective view synthesis quality assessment. *Int. Soc. Opt. Eng*. 8288:8256–8288. doi: 10.1117/12.908762

Dragana, S.-S., Dragan, K., and Patrick, L. C. (2015a). "Dibr synthesized image quality assessment based on morphological pyramids," in *2015 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)* (Lisbon), 1–4.

Dragana, S.-S., Dragan, K., and Patrick, L. C. (2015b). "Dibr synthesized image quality assessment based on morphological wavelets," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)* (Costa Navarino), 1–6.

Dragana, S.-S., Dragan, K., and Patrick, L. C. (2016). Multi-scale synthesized view assessment based on morphological pyramids. *J. Electr. Eng*. 67, 3–11. doi: 10.1515/jee-2016-0001

Gao, Z., Zhai, G., Deng, H., and Yang, X. (2020). Extended geometric models for stereoscopic 3d with vertical screen disparity. *Displays* 65:101972. doi: 10.1016/j.displa.2020.101972

Godard, C., Aodha, O. M., Firman, M., and Brostow, G. (2019). "Digging into self-supervised monocular depth estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 3827–3837.

Gu, K., Jakhetiya, V., Qiao, J.-F., Li, X., Lin, W., and Thalmann, D. (2018a). Model-based referenceless quality metric of 3d synthesized images using local image description. *IEEE Trans. Image Process*. 27, 394–405. doi: 10.1109/TIP.2017.2733164

Gu, K., Li, L., Lu, H., Min, X., and Lin, W. (2017a). A fast reliable image quality predictor by fusing micro- and macro-structures. *IEEE Trans. Ind. Electr*. 64, 3903–3912. doi: 10.1109/TIE.2017.2652339

Gu, K., Liu, H., Xia, Z., Qiao, J., Lin, W., and Daniel, T. (2021a). Pm2.5 monitoring: use information abundance measurement and wide and deep learning. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 4278–4290. doi: 10.1109/TNNLS.2021.3105394

Gu, K., Tao, D., Qiao, J.-F., and Lin, W. (2018b). Learning a no-reference quality assessment model of enhanced images with big data. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 1301–1313. doi: 10.1109/TNNLS.2017.2649101

Gu, K., Wang, S., Zhai, G., Lin, W., Yang, X., and Zhang, W. (2016). Analysis of distortion distribution for pooling in image quality prediction. *IEEE Trans. Broadcast*. 62, 446–456. doi: 10.1109/TBC.2015.2511624

Gu, K., Xia, Z., and Junfei, Q. (2020a). Stacked selective ensemble for pm2.5 forecast. *IEEE Trans. Instrum. Meas*. 69, 660–671. doi: 10.1109/TIM.2019.2905904

Gu, K., Xia, Z., Qiao, J., and Lin, W. (2020b). Deep dual-channel neural network for image-based smoke detection. *IEEE Trans. Multimedia* 22, 311–323. doi: 10.1109/TMM.2019.2929009

Gu, K., Zhai, G., Lin, W., Yang, X., and Zhang, W. (2015). No-reference image sharpness assessment in autoregressive parameter space. *IEEE Trans. Image Process*. 24, 3218–3231. doi: 10.1109/TIP.2015.2439035

Gu, K., Zhang, Y., and Qiao, J. (2021b). Ensemble meta-learning for few-shot soot density recognition. *IEEE Trans. Ind. Inform*. 17, 2261–2270. doi: 10.1109/TII.2020.2991208

Gu, K., Zhou, J., Qiao, J.-F., Zhai, G., Lin, W., and Bovik, A. C. (2017b). No-reference quality assessment of screen content pictures. *IEEE Trans. Image Process*. 26, 4005–4018. doi: 10.1109/TIP.2017.2711279

Hasler, D., and Suesstrunk, S. E. (2003). "Measuring colorfulness in natural images," in *Human Vision and Electronic Imaging VIII, Vol. 5007*, eds B. E. Rogowitz and T. N. Pappas (Santa Clara, CA: SPIE), 87–95.

Jakhetiya, V., Gu, K., Singhal, T., Guntuku, S. C., Xia, Z., and Lin, W. (2019). A highly efficient blind image quality assessment metric of 3-d synthesized images using outlier detection. *IEEE Trans. Ind. Inform*. 15, 4120–4128. doi: 10.1109/TII.2018.2888861

Jang, C. Y., Kim, S., Cho, K.-R., and Kim, Y. H. (2019). Performance analysis of structural similarity-based backlight dimming algorithm modulated by controlling allowable local distortion of output image. *Displays* 59, 1–8. doi: 10.1016/j.displa.2019.05.001

Ke, R., Li, Z., Tang, J., Pan, Z., and Wang, Y. (2019). Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow. *IEEE Trans. Intel. Trans. Syst*. 20, 54–64. doi: 10.1109/TITS.2018.2797697

Li, B., Dai, Y., and He, M. (2018a). Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. *Pattern Recognit*. 83, 328–339. doi: 10.1016/j.patcog.2018.05.029

Li, L., Huang, Y., Wu, J., Gu, K., and Fang, Y. (2021a). Predicting the quality of view synthesis with color-depth image fusion. *IEEE Trans. Circ. Syst. Video Technol*. 31, 2509–2521. doi: 10.1109/TCSVT.2020.3024882

Li, L., Zhou, Y., Gu, K., Lin, W., and Wang, S. (2018b). Quality assessment of dibr-synthesized images by measuring local geometric distortions and global sharpness. *IEEE Trans. Multimedia* 20, 914–926. doi: 10.1109/TMM.2017.2760062

Li, L., Zhou, Y., Wu, J., Li, F., and Shi, G. (2021b). Quality index for view synthesis by measuring instance degradation and global appearance. *IEEE Trans. Multimedia* 23:320–332. doi: 10.1109/TMM.2020.2980185

Liao, C.-C., Su, C.-W., and Chen, M.-Y. (2019). Mitigation of image blurring for performance enhancement in transparent displays based on polymer-dispersed liquid crystal. *Displays* 56, 30–37. doi: 10.1016/j.displa.2018.11.001

Ling, S., Li, J., Che, Z., Min, X., Zhai, G., and Le Callet, P. (2021). Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories. *IEEE Trans. Image Process*. 30, 517–531. doi: 10.1109/TIP.2020.3037504

Ling, S., Li, J., Che, Z., Wang, J., Zhou, W., and Le Callet, P. (2020). Re-visiting discriminator for blind free-viewpoint image quality assessment. *IEEE Trans. Multimedia*. doi: 10.1109/TMM.2020.3038305. [Epub ahead of print].

Liu, H., Lei, F., Tong, C., Cui, C., and Wu, L. (2021). Visual smoke detection based on ensemble deep cnns. *Displays* 69:102020. doi: 10.1016/j.displa.2021.102020

Luo, G., Zhu, Y., Weng, Z., and Li, Z. (2020). A disocclusion inpainting framework for depth-based view synthesis. *IEEE Trans. Pattern Anal. Mach. Intell*. 42, 1289–1302. doi: 10.1109/TPAMI.2019.2899837

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., and Chen, C. W. (2018). Blind quality assessment based on pseudo-reference image. *IEEE Trans. Multimedia* 20, 2049–2062. doi: 10.1109/TMM.2017.2788206

Min, X., Zhai, G., Zhou, J., Farias, M. C. Q., and Bovik, A. C. (2020a). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Process*. 29, 6054–6068. doi: 10.1109/TIP.2020.2988148

Min, X., Zhou, J., Zhai, G., Le Callet, P., Yang, X., and Guan, X. (2020b). A metric for light field reconstruction, compression, and display quality evaluation. *IEEE Trans. Image Proc*. 29, 3790–3804. doi: 10.1109/TIP.2020.2966081

Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal. Process. Lett*. 20, 209–212. doi: 10.1109/LSP.2012.2227726

Selzer, M. N., Gazcon, N. F., and Larrea, M. L. (2019). Effects of virtual presence and learning outcome using low-end virtual reality systems. *Displays* 59, 9–15. doi: 10.1016/j.displa.2019.04.002

Sugita, N., Sasaki, K., Yoshizawa, M., Ichiji, K., Abe, M., Homma, N., et al. (2019). Effect of viewing a three-dimensional movie with vertical parallax. *Displays* 58, 20–26. doi: 10.1016/j.displa.2018.10.007

Sun, K., Tang, L., Qian, J., Wang, G., and Lou, C. (2021). A deep learning-based pm2.5 concentration estimator. *Displays* 69:102072. doi: 10.1016/j.displa.2021.102072

Tanimoto, M., Tehrani, M. P., Fujii, T., and Yendo, T. (2011). Free-viewpoint tv. *IEEE Signal. Process. Mag*. 28, 67–76. doi: 10.1109/MSP.2010.939077

Tian, S., Zhang, L., Morin, L., and Déforges, O. (2018). Niqsv+: a no-reference synthesized view quality assessment metric. *IEEE Trans. Image Process*. 27, 1652–1664. doi: 10.1109/TIP.2017.2781420

Vu, P. V., and Chandler, D. M. (2012). A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal. Process. Lett*. 19, 423–426. doi: 10.1109/LSP.2012.2199980

Wang, G., Wang, Z., Gu, K., Jiang, K., and He, Z. (2021). Reference-free dibr-synthesized video quality metric in spatial and temporal domains. *IEEE Trans. Circ. Syst. Video Technol*. doi: 10.1109/TCSVT.2021.3074181

Wang, G., Wang, Z., Gu, K., Li, L., Xia, Z., and Wu, L. (2020). Blind quality metric of dibr-synthesized images in the discrete wavelet transform domain. *IEEE Trans. Image Process*. 29, 1802–1814. doi: 10.1109/TIP.2019.2945675. [Epub ahead of print].

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., and Li, Q. (2011). Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Process*. 20, 1185–1198. doi: 10.1109/TIP.2010.2092435

Ye, P., Wu, X., Gao, D., Deng, S., Xu, N., and Chen, J. (2020). Dp3 signal as a neuro-indictor for attentional processing of stereoscopic contents in varied depths within the 'comfort zone'. *Displays* 63:101953. doi: 10.1016/j.displa.2020.101953

Yildirim, C. (2019). Cybersickness during vr gaming undermines game enjoyment: a mediation model. *Displays* 59, 35–43. doi: 10.1016/j.displa.2019.07.002

Yue, G., Hou, C., Gu, K., Zhou, T., and Zhai, G. (2019). Combining local and global measures for dibr-synthesized image quality evaluation. *IEEE Trans. Image Process*. 28, 2075–2088. doi: 10.1109/TIP.2018.2875913

Zhai, G., Wu, X., Yang, X., Lin, W., and Zhang, W. (2012). A psychovisual quality metric in free-energy principle. *IEEE Trans. Image Process*. 21, 41–52. doi: 10.1109/TIP.2011.2161092

Zhang, L., Zhang, L., and Bovik, A. C. (2015). A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Process*. 24, 2579–2591. doi: 10.1109/TIP.2015.2426416

Zhang, Z., Xu, C., Yang, J., Tai, Y., and Chen, L. (2018). Deep hierarchical guidance and regularization learning for end-to-end depth estimation. *Pattern Recognit*. 83:430–442. doi: 10.1016/j.patcog.2018.05.016

Zhou, Y., Li, L., Ling, S., and Le Callet, P. (2019). Quality assessment for view synthesis using low-level and mid-level structural representation. *Signal. Process. Image Commun*. 74, 309–321. doi: 10.1016/j.image.2019.03.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

MINI REVIEW
published: 05 January 2022
doi: 10.3389/fpsyg.2021.767840

Listening Effort Informed Quality of Experience Evaluation

Pheobe Wenyi Sun and Andrew Hines*

QxLab, School of Computer Science, University College Dublin, Dublin, Ireland

Perceived quality of experience for speech listening is influenced by cognitive processing and can affect a listener's comprehension, engagement and responsiveness. Quality of Experience (QoE) is a paradigm used within the media technology community to assess media quality by linking quantifiable media parameters to perceived quality. The established QoE framework provides a general definition of QoE, categories of possible quality influencing factors, and an identified QoE formation pathway. These assist researchers to implement experiments and to evaluate perceived quality for any applications. The QoE formation pathways in the current framework do not attempt to capture cognitive effort effects and the standard experimental assessments of QoE minimize the influence from cognitive processes. The impact of cognitive processes and how they can be captured within the QoE framework have not been systematically studied by the QoE research community. This article reviews research from the fields of audiology and cognitive science regarding how cognitive processes influence the quality of listening experience. The cognitive listening mechanism theories are compared with the QoE formation mechanism in terms of the quality contributing factors, experience formation pathways, and measures for experience. The review prompts a proposal to integrate mechanisms from audiology and cognitive science into the existing QoE framework in order to properly account for cognitive load in speech listening. The article concludes with a discussion regarding how an extended framework could facilitate measurement of QoE in broader and more realistic application scenarios where cognitive effort is a material consideration.

Keywords: Quality of Experience (QoE), cognitive load, listening effort, subjective test, QoE framework

OPEN ACCESS

Edited by:
Guangtao Zhai,
Shanghai Jiao Tong University, China

Reviewed by:
Xiongkuo Min,
University of Texas at Austin,
United States
Yucheng Zhu,
Shanghai Jiao Tong University, China

****Correspondence:***
Pheobe Wenyi Sun
wenyi.sun@ucdconnect.ie

Specialty section:
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Psychology

Received: *31 August 2021*
Accepted: *31 October 2021*
Published: *05 January 2022*

Citation:
Sun PW and Hines A (2022) Listening
Effort Informed Quality of Experience
Evaluation.
Front. Psychol. 12:767840.
doi: 10.3389/fpsyg.2021.767840

1. INTRODUCTION

Quality of experience (QoE) is a paradigm that assesses media quality by mimicking human judgement. The goal is to understand and quantify how consumers perceive media quality. Instead of using the measurable signal parameters, QoE researchers evaluate the quality of a multimedia event based on reported quality ratings from participants in subjective experimental studies. To void the biases from the interpersonal differences, a mean opinion score (MOS) is used to represent an averaged perceived quality. The subjective ratings from experiments are also used to develop signal-based QoE prediction models (also called objective models). Such models are expected to predict quality judgements for multimedia application. Thus, the QoE evaluation approach has been widely adopted to rapidly test the perceptual effect of new products and services.

Despite the wide applicability of QoE evaluation methods, current QoE evaluations for naturalistic multimedia consumption scenarios, when a person is listening to podcasts while driving for example, are limited. They lack the consideration of a person's comprehension, engagement, effort, and other mental status. The current QoE framework, a conceptual model that characterizes how QoE forms, adopts a simple filtering structure that collapse all the interactions of different influencing factors to a single outcome—people's internal comparison between their expectation of the signal properties and what they actually perceive—which can be observed from the subjective quality judgement. Such framework has been widely adopted and works well for many scenarios. For instance, the telecommunication industry uses it to analyse the quality impact of a change in network capacity or system parameters. However, how the cognitive processes affect the multimedia QoE are not addressed by the framework nor by the evaluation methods.

As the multimedia consumption scenarios become more complex, the cognitive aspects of the experience need to be taken into account. QoE evaluation methods applicable to more natural scenarios are important to understand the impact of potential technological changes. Although cognitive aspects are highly personal and are hard to be modeled, the theories and the empirical studies in cognitive science can provide us with practical tools to systematically evaluate the impacts of the cognitive processes. This paper reviews the existing QoE framework as well as the cognitive listening methods and models from the audiology and cognitive psychology domains. The paper then discusses the potential ways to integrate cognitive effort into the existing QoE framework. While this paper uses listening effort as a focus, this review prompts consideration of broader and more realistic QoE framework for application scenarios where cognitive effort is a factor.

2. THE EXISTING QOE FRAMEWORK AND ITS LIMITS

2.1. The QoE Framework

The QoE framework is a conceptual model that describes a QoE formation mechanism for any multimedia consumption scenario. It can be applied as a template to characterize a quality judgement formation for an experience. The QoE framework identifies the QoE formation pathways, the QoE observables, and the QoE influencing factors (see **Figure 1**). Quality of Experience (QoE) describes a person's satisfactory level of a perceptual event (Brunnström et al., 2013). It results from the fulfillment of expectations. The satisfactory level of a perceptual experience can

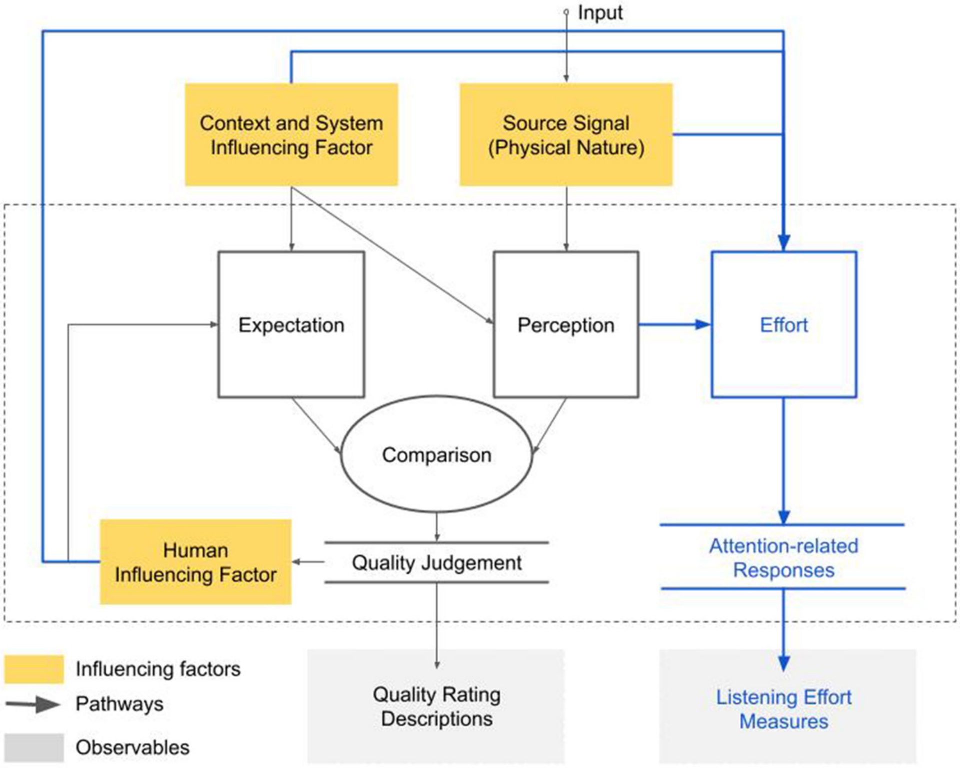
Input
Context and System Influencing Factor
Source Signal (Physical Nature)
Expectation
Perception
Effort
Comparison
Human Influencing Factor
Quality Judgement
Attention-related Responses
Influencing factors
Pathways
Observables
Quality Rating Descriptions
Listening Effort Measures

FIGURE 1 | The QoE framework adapted from the QoE whitepaper (Brunnström et al., 2013) where the QoE formation pathways (lines with arrows), the QoE observables (gray boxes), and the QoE influencing factors (orange boxes) are identified. The elements in the existing framework are denoted in black and the expanded parts are in blue. The existing model assumes that the QoE is the outcome of comparing the expected event and the perceived event (see the mechanistic diagrams in black). Both expectation and perception are influenced by different influencing factors. The influencing factors are grouped to four categories (orange boxes). The perceived quality is observed by the subjective rating and/or description of an event (gray box at the bottom).

be reflected by people's quality judgement. Therefore descriptions and ratings are used as the observables to indicate the latent state of interest—the perceived QoE.

Building on the QoE formation mechanism, *influencing factors* are classified that contribute to either the formation of one's expectation or the perceived event via *formation pathways* (the black lines with arrows in **Figure 1**). For example, the context of media consumption can influence one's expectation (Sackl et al., 2017), e.g., for a free vs. paid telephone call, or listening-only radio vs. conversational telephone call (Moller et al., 2011). Other factors such as noise and network conditions also affect the perceived event. All the possible QoE influencing factors are grouped to four categories in the QoE framework: signal, context, system, and human factors (Brunnström et al., 2013), each has its own pathway that ultimately contributes to the formation of QoE (see the orange boxes in **Figure 1**). The identified categories of the QoE influencing factors provide a structural guideline for researchers to analyse the quality impact of any factors of interest in a variety of scenarios. Together with the QoE formation pathways and the observables, researchers can design subjective experimental procedures that yield quantitative QoE measures.

2.2. QoE Evaluation in Practice

The two commonly used QoE evaluation approaches, the "descriptive" and the "integrated" (Katz and Nicol, 2019) approaches, conform well with the observables in the QoE framework. The *descriptive* (or *performance*) approach uses the verbal descriptions as QoE evaluation. The focus of the experiential aspects will shift across different application scenarios using this approach. For example, descriptions of the noise and intelligibility levels are useful to evaluate the QoE of a voice call; comments regarding the perceived origin of a sound or how it blends with the rest of the environment are useful in a spatial sound scenario. The *integrated* approach, to the contrary, uses a single numerical value to represent the impression of an overall QoE. For instance, the basic audio quality (BAQ) test (ITU-R, 2015a,b; Schöffler, 2017) uses the mean opinion scores (MOS) for QoE. Using a uni-dimensional representation for QoE makes the comparison of different experiences easier, and hence, making it an efficient solution for rapid evaluations in industry. While acknowledging that experience is a high dimensional concept, the QoE framework provides guidelines to evaluate QoE that is repeatable experimentally and useful for media technology development and evaluation.

2.3. The Overlooked Impact of Cognitive Processes

The cognitive processes are modeled in the QoE framework through the pathways connecting the human influencing factors (orange box in bottom left of **Figure 1**). The human influencing factors comprise factors such as mood, motivation, language, or prior experience (Brunnström et al., 2013). The human influencing factors only contribute to expectation formation, not the downstream QoE formation as human influencing factors are considered to be either temporarily volatile (such as mood and motivation) or personal (such as language proficiency or prior experience). In order to model a QoE evaluation that is representative and relevant for a large population, the effect of the transient factors needs to be dampened in the model. To realize this, QoE evaluation protocols (ITU-T, 1996) recommend implementing a variety of mechanisms to minimize the effect of the human influencing factors such as accent familiarity, voice preferences, fatigue, or boredom. Studies in both audiology and cognitive neuroscience (Pichora-Fuller et al., 2016; Peelle, 2018; Herrmann and Johnsrude, 2020) show that the effort expended on our cognitive process has a substantial impact on perceived experience. Increased listening effort is found to reduce the ability to memorize (Murphy et al., 2000; Rabbitt, 2007; Heinrich et al., 2008; Heinrich and Schneider, 2011), and thereafter comprehension can be adversely affected (Piquado et al., 2012; Ward et al., 2016) due to less context information available from the memory to help decode the current information. A sustained high listening effort is found to lead to lower arousal levels (Aston-Jones and Cohen, 2005) and reduced affective responses (Francis and Love, 2020) such as fatigue (Hockey, 2011) and boredom (Elpidorou, 2018). The strenuous cognitive process is also found to have negative impact on behaviors such as slower response time (Phillips, 2016), inferior task performance (Wingfield et al., 2006; Hornsby, 2013; Lemke and Besser, 2016; Phillips, 2016), or withdrawal from listening task (Lemke and Besser, 2016; Herrmann and Johnsrude, 2020) and social interactions (Mick et al., 2014; Shukla et al., 2020). Several neurological evidences [such as EEG (Hunter and Pisoni, 2018), fMRI (Kuchinsky et al., 2013), and pupil dilation (Aston-Jones and Cohen, 2005; Adank, 2012)] have showed distinct patterns when listeners are exposed to challenging auditory material, indicating the recruitment of different cognitive resources in astute listening scenarios. These findings indicate that the adverse effect of heavy auditory cognition is not only relevant to the population who are diagnosed with hearing impairment, but also relevant to anyone who needs to engage with listening in their day-to-day activities as the recruitment of other cognitive resources can directly affect the allocation of attention and therefore the task performance.

From a multimodal perspective, the existing pathways in the QoE framework are not exhaustive in modeling the effect of different source signals. The combined effect of audio and visual input signals have been shown to produce shifts in attention in various studies (Talsma et al., 2006; Rapela et al., 2012; Chao et al., 2020). Although the multimodal integration is still an active area of study in neuroscience (Koelewijn et al., 2010; Fu et al., 2020), the consideration of audio-visual interaction is shown to be useful for attention and saliency modeling to improve existing QoE prediction (Min et al., 2015, 2020; Zhu et al., 2020).

Attentional saliency, comprehension, fatigue level, task performance, and emotional status are important building blocks for understanding QoE in realistic listening scenarios, and these aspects cannot be captured and fully understood by the quality judgement alone via the standard QoE observable adopted by the community. The existing QoE framework lacks an explicit systematic model to guide effective studies exploring the impact of the cognitive processes on QoE. The attentional control can be influenced by the source signals (e.g., multimodal interaction) as well as by the human influencing factor (e.g., mental capacity). This study will focus on the latter and use the uni-modal input signal as an example to show how studies from cognitive hearing and perception theory could provide complementary learning to supplement the existing QoE framework.

3. INTEGRATING LISTENING EFFORT INTO EXISTING QOE FRAMEWORK

To integrate listening effort into the QoE framework model, we consider three questions: (i) what contributes to the increase in the cognitive effort; (ii) how increased effort affects QoE; (iii) how to quantify the effect of effort on QoE. These questions correspond to the three core component in the QoE framework: influencing factors, QoE pathways, and the observables.

This section addresses each question and discuss how each component in the existing QoE framework can be adapted with reference to two cognitive hearing models: the Framework for understanding Effortful Listening (FUEL) (Pichora-Fuller et al., 2016) and the Model of Listening Engagement (MoLE) (Herrmann and Johnsrude, 2020). They also draw on the more general cognitive load models (the load theory Murphy et al., 2016 and the mental capacity model Kahneman, 1973).

3.1. Influencing Factors

Listening effort increases along with the listening demand (McGarrigle et al., 2014) as more attentional resources need to be allocated to meet the demand. The FUEL (Pichora-Fuller et al., 2016) model categorizes the sources of listening effort as source, transmission, listener, message, and context factors. These categories all have their counterparts in the QoE framework. **Table 1** illustrates how different sources of listening effort can be mapped to different influencing factor categories in the FUEL and the QoE framework. The middle column highlights that all four QoE influencing factor categories contribute to the effort formation. The overlapping factors of concern in both frameworks indicate that the existing QoE framework has already incorporated the main factors that lead to listening effort. The next step is to analyse whether the cognitive effect of these influencing factors can be modeled by the QoE formation pathways.

TABLE 1 | Sources of listening effort and their corresponding influencing factor categories in the QoE framework and the FUEL.

Factors	QoE	FUEL
Voice degradation	System	Transmission
Bandwidth limit	System	Transmission
Noise	System	Transmission
Reverberation	System	Transmission
Multi-talker	Signal	Source & context
Spatial separation	Signal	Source & context
Synthesized voice	Signal	Source
Sustained speech	Context	Source
Voice similarity	Signal	Source
Foreign language	Signal & context	Message & context
Reward	Human	Motivation
Hearing loss	Human	Listener

3.2. Pathways

The formation pathways in a model identify the possible mechanisms through which the influencing factors can follow to impact an outcome. Although the formation pathways are not concrete, they are depicted in the models to guide research protocol designs wishing to evaluate the effect of factors of interest. The implications of increased listening effort are the result of complex combinations of interactions. The existing QoE formation pathways collapse the contributions of influencing factors to an internal comparison, which limits the capacity to capture the wider cognitive effects that make up our listening experience. Cognitive hearing studies (McGarrigle et al., 2014; Pichora-Fuller et al., 2016; Herrmann and Johnsrude, 2020) indicate that multiple effort formation pathways exist during speech listening. When a speech signal is being processed at an early stage, with presence of noise for instance, effort arises when listeners inhibit the irrelevant signals and keep attentive to the target signals. However, sometimes a higher load level helps people to concentrate (Mick et al., 2014; Murphy et al., 2016; Herrmann and Johnsrude, 2020). At a later stage when the speech signal is being processed semantically, effort increases when the content topic is obscure and more context information needs to be recalled from memory to aid comprehension. Effort is also be influenced by the demands of concurrent tasks (Skowronek and Raake, 2014) as attention needs to be constantly reallocated depending on the dynamics of a subtask. This pathway is particularly relevant to the design of technology and multimedia applications where people increasingly consume multimedia while multi-tasking in day-to-day scenarios.

It has yet to be shown whether the effect of multiple effort formation pathways can be simplified to a single pathway. Therefore, we show multiple potential effort formation pathways so that systematic investigations into the cognitive impact can be designed. Multiple pathways might result in different experiential implications in addition to the quality judgement, thus additional measurements that capture different aspects of an experience need to be recorded to compare the differences in the perceptual experiences.

3.3. Observables

The observables are used by researchers to infer the impact of influencing factors. The choice of the observables depends on the outcome of interest and the corresponding formation pathways. For instance, the corresponding observables for the percept (Johnsrude and Rodd, 2016), cognitive activity, and the mental capacity as a result of listening effort can be the self-reported responses, neuroimaging, and concurrent task performance. As multiple listening effort formation pathways might exist, a single observable (i.e., a quality judgement) may not be sufficient to capture the QoE. Initiatives in the QoE domain (Engelke et al., 2017) already attempt to use other observables to give a broader definition of QoE. We will next summarize the various listening effort observables in use and discuss how different types of observable account for different aspects of an experience.

The most direct observables for listening effort are the self-reported ratings or descriptions. Ratings are more commonly adopted as they are both scalable and easier to process. The NASA-TLX mental effort scale (Hart and Staveland, 1988), for example, is a mature instrument that asks subjects to rate on different relevant aspects such as fatigue, stress, and task difficulty to gauge one's overall cognitive load (Rubio et al., 2004). Another example of a self-reported measure asks subjects to estimate the duration they can sustain a task to gauge the cognitive load while listening (Pichora-Fuller et al., 2016). However, due to the retrospective nature of these self-reported measures, such measures are susceptible to memory and descriptive biases.

Behavioral responses are also used to indicate effort. These include the memory recall, speech comprehension (observed after the task), or attention-related task performance (observed during the task). The Span Test (Conway et al., 2005) is a well established working memory test where participants are asked to read a series of sentences and to recall the last word from each sentence. It is used to indirectly evaluate listening effort based on the assumption of working memory capacity (Baddeley, 2000). In a demanding listening scenario, an increase in the allocated cognitive resources to comprehend the signal will adversely impact information recall capacity. Another popular experimental paradigm is the dual-task method where participants conduct a parallel task simultaneously to force the division of attention. In this case, an increase in the listening effort is indicated by a performance reduction in the concurrent task (Hunter, 2020). The dual-task paradigm is based on the assumption that attention allocated to one task will leave less spare cognitive capacity to process another task (Kahneman, 1973; Beatty, 1977; Sweller, 1994; Schnotz and Kürschner, 2007) leading to an observable reduced performances in the less attended task.

Psychophysiological changes are also used to indicate the effort involved in a listening task. Some physiological observables (e.g., pupil dilation, cardiac responses, skin conductance, and hormonal changes) are the result of sympathetic or parasympathetic responses to stress or effort (de Waard, 1996; Peelle, 2018). Thus, they are regarded as indirect measures for listening effort. Observables captured around the brain area (such as the activity intensity and the differences in the activated brain regions) are also used as indicators of listening effort. For example, an increase in the alpha band power in the electroencephalography signal can be observed when there is signal degradation or an increased demand for information storage (Piquado et al., 2012; Pichora-Fuller et al., 2016; Hunter, 2020). An increase in activity is found in the cingulo-opercular network from the functional magnetic resonance imaging when listeners are exposed to less intelligible signals (Wild et al., 2012; Erb et al., 2013; Vaden et al., 2013; Eckert et al., 2016). The psychophysiological observables are highly susceptible to many other internal and external factors such as environment temperature and mental status. Yet the high resolution in time makes them the preferred instruments for event-related analysis.

Identifying the potential and appropriate observables is critical in order to select the methods that will capture how effort affects different aspects of our experience. Using multiple observables is also recommended to reduce the structural interference in data analysis (Kahneman, 1973; Pichora-Fuller et al., 2016). The theoretical and empirical cognitive psychology literature provides a broad selection of observables to complement the commonly-used self-reported measures in the QoE community. It also prompts looking beyond the existing QoE framework to consider pathways to better capture different impacts of listening effort in naturalistic scenarios.

4. CONCLUSION AND FUTURE DIRECTION

This review introduced the QoE framework model used by the media technology community to assign in designing and selecting the appropriate methods to empirically evaluate quality of experience. We introduced the rationale behind the framework and explained the structural influencing factors, pathways and observables. The limited capability within the framework to capture and quantify how effort interacts with QoE was highlighted. With a focus on listening effort, this paper reviewed multiple listening effort formation pathways from the cognitive science domain to complement the existing QoE formation pathway. A review of literature and methods drawn from the audiology and cognitive science domains, illustrated how the QoE framework could be expanded and QoE experimental methods could be applied to naturalistic listening scenarios where the cognitive process plays a significant part in QoE formation. Pathways and observables beyond self-reported quality ratings were reviewed. We believe the review warrants adding a cognitive dimension to QoE framework. It would allow for more direct comparisons of different subjective experiments. It would encourage the community to design subjective experiments that consider the impact of less explored cognitive processes. Furthermore, subjective experiments guided by such framework should provide new insights into the more nuanced experiential aspects of our multimedia consumption experience.

More generally, the review highlights the flexibility within the framework for extension and the potential to capture a better understanding of audio influence within wider QoE studies, e.g., listening effort impacting video or immersive QoE. This review also presents an opportunity to apply a similar approach beyond listening, identifying new pathways and observables within the QoE framework, for visual, haptic or multimodal interactions.

AUTHOR CONTRIBUTIONS

PS and AH both contributed to writing, development, and editing. Both authors contributed to the article and approved the submitted version.

FUNDING

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 12/RC/2289_P2.

REFERENCES

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: two activation likelihood estimation (ALE) meta-analyses. *Brain Lang*. 122, 42–54. doi: 10.1016/j.bandl.2012.04.014

Aston-Jones, G., and Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci*. 28, 403–450. doi: 10.1146/annurev.neuro.28.061604.135709

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends Cogn. Sci*. 4, 417–423. doi: 10.1016/S1364-6613(00)01538-2

Beatty, J. (1977). *Activation and Attention*. Los Angeles, CA: California Univ Los Angeles Dept of Psychology.

Brunnström, K., Beker, S. A., de Moor, K., Dooms, A., Egger, S., Garcia, M.-N., et al. (2013). *Qualinet White Paper on Definitions of Quality of Experience*. Technical report, Novi Sad.

Chao, F. Y., Ozcinar, C., Wang, C., Zerman, E., Zhang, L., Hamidouche, W., et al. (2020). "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2020* (London: IEEE).

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., and Engle, R. W. (2005). Working memory span tasks: a methodological review and user's guide. *Psychonomic Bull. Rev*. 12, 769–786. doi: 10.3758/BF03196772

de Waard, D. (1996). *The Measurement of Drivers'Mental Workload* (Ph.D. thesis). University of Groningen.

Eckert, M. A., Teubner-Rhodes, S., Vaden, K. I., and Jr. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear Hear*. 37(Suppl 1):101S. doi: 10.1097/AUD.0000000000000300

Elpidorou, A. (2018). The bored mind is a guiding mind: toward a regulatory theory of boredom. *Phenomenol. Cogn. Sci* 17, 455–484. doi: 10.1007/s11097-017-9515-1

Engelke, U., Darcy, D. P., Mulliken, G. H., Bosse, S., Martini, M. G., Arndt, S., et al. (2017). Psychophysiology-Based QoE assessment: a survey. *IEEE J. Select. Top. Signal Proc*. 11, 6–21. doi: 10.1109/JSTSP.2016.2609843

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2013). The brain dynamics of rapid perceptual adaptation to adverse listening conditions. *J. Neurosci*. 33, 10688–10697. doi: 10.1523/JNEUROSCI.4596-12.2013

Francis, A. L., and Love, J. (2020). Listening effort: are we measuring cognition or affect, or both? *Wiley Interdiscip. Rev. Cogn. Sci*. 11, e1514. doi: 10.1002/wcs.1514

Fu, D., Weber, C., Yang, G., Kerzel, M., Nan, W., Barros, P., et al. (2020). What can computational models learn from human selective attention? a review from an audiovisual unimodal and crossmodal perspective. *Front. Integr. Neurosci*. 14:10. doi: 10.3389/fnint.2020.00010

Hart, S. G., and Staveland, L. E. (1988). Development of nasa-tlx (task load index): results of empirical and theoretical research. *Adv. Psychol*. 52, 139–183. doi: 10.1016/S0166-4115(08)62386-9

Heinrich, A., and Schneider, B. A. (2011). Elucidating the effects of ageing on remembering perceptually distorted word pairs. *Q. J. Exp. Psychol*. 64, 186–205. doi: 10.1080/17470218.2010.492621

Heinrich, A., Schneider, B. A., and Craik, F. I. (2008). Investigating the influence of continuous babble on auditory short-term memory performance. *Q. J.Exp. Psychol*. 61, 735–751. doi: 10.1080/17470210701402372

Herrmann, B., and Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hear Res*. 397:108016. doi: 10.1016/j.heares.2020.108016

Hockey, R. (2011). *The Psychology of Fatigue: Work, Effort and Control*. New York, NY: Cambridge University Press. 1–272.

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear Hear*. 34, 523–534. doi: 10.1097/AUD.0b013e31828003d8

Hunter, C. R. (2020). Tracking cognitive spare capacity during speech perception with EEG/ERP: effects of cognitive load and sentence predictability. *Ear Hear*. 41, 1144–1157. doi: 10.1097/AUD.0000000000000856

Hunter, C. R., and Pisoni, D. B. (2018). Extrinsic cognitive load impairs spoken word recognition in high-and low-predictability sentences. *Ear Hear*. 39, 378–389. doi: 10.1097/AUD.0000000000000493

ITU- (2015b). *BS.1534 Method for the subjective assessment of intermediate quality level of audio systems*. Technical report.

ITU-R. (2015a). *BS.1116-3 Methods for the subjective assessment of small impairments in audio systems*. Technical report, ITU.

ITU-T. (1996). *P.800: Methods for subjective determination of transmission quality*. Technical report, Int. Telecomm. Union.

Johnsrude, I. S., and Rodd, J. M. (2016). "Chapter 40. Factors that increase processing demands when listening to speech," in *Neurobiology of Language*, eds G. Hickok, S. L. Small (Academic Press), 491–502. doi: 10.1016/B978-0-12-407794-2.00040-7

Kahneman, D. (1973). *Attention and Effort, Vol. 1063*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Katz, B. F., and Nicol, R. (2019). "Binaural spatial reproduction," in *Sensory Evaluation of Sound, Chapter 11*, ed N. Zacharov (Boca Raton, FL: CRC Press Taylor & Francis Group), 349–388.

Koelewijn, T., Bronkhorst, A., and Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: a review of audiovisual studies. *Acta Psychol*. 134, 372–384. doi: 10.1016/j.actpsy.2010.03.010

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., et al. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology* 50, 23–34. doi: 10.1111/j.1469-8986.2012.01477.x

Lemke, U., and Besser, J. (2016). Cognitive load and listening effort: concepts and age-related considerations. *Ear Hear*. 37, 77S–84S. doi: 10.1097/AUD.0000000000000304

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., et al. (2014). Listening effort and fatigue: What exactly are we measuring? A british society of audiology cognition in hearing special interest group 'white paper'. *Int. J. Audiol*. 53, 433–445. doi: 10.3109/14992027.2014.890296

Mick, P., Kawachi, I., and Lin, F. R. (2014). The association between hearing loss and social isolation in older adults. *Otolaryngol. Head Neck Surg*. 150, 378–384. doi: 10.1177/0194599813518021

Min, X., Zhai, G., Hu, C., and Gu, K. (2015). "Fixation prediction through multimodal analysis," in *2015 Visual Communications and Image Processing (VCIP)* (Singapore: IEEE), 1–4.

Min, X., Zhai, G., Member, S., Zhou, J., Zhang, X.-P., Yang, X., et al. (2020). A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. Image Proc*. 29:2020. doi: 10.1109/TIP.2020.2966082

Moller, S., Chan, W.-Y., Cote, N., Falk, T., Raake, A., and Waltermann, M. (2011). Speech quality estimation: models and trends. *IEEE Signal Proc. Mag*. 28, 18–28. doi: 10.1109/MSP.2011.942469

Murphy, D. R., Craik, F. I. M., Li, K. Z. H., and Schneider, B. A. (2000). Comparing the effects of aging and background noise on short-term memory performance. *Psychol. Aging* 15, 323–334. doi: 10.1037/0882-7974.15.2.323

Murphy, G., Groeger, J. A., and Greene, C. M. (2016). Twenty years of load theory–Where are we now, and where should we go next? *Psychonomic Bull. Rev*. 23, 1316–1340. doi: 10.3758/s13423-015-0982-5

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear Hear*. 39, 204–214. doi: 10.1097/AUD.0000000000000494

Phillips, N. A. (2016). The implications of cognitive aging for listening and the framework for understanding effortful listening (FUEL). *Ear Hear*. 37, 44S–51S. doi: 10.1097/AUD.0000000000000309

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., et al. (2016). Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear*. 37, 5S–27S. doi: 10.1097/AUD.0000000000000312

Piquado, T., Benichov, J. I., Brownell, H., and Wingfield, A. (2012). The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *Int. J. Audiol*. 51, 576–583. doi: 10.3109/14992027.2012.684403

Rabbitt, P. M. A. (2007). Channel-capacity, intelligibility and immediate memory. *Q. J. Exp. Psychol*. 20, 241–248. doi: 10.1080/14640746808400158

Rapela, J., Gramann, K., Westerfield, M., Townsend, J., and Makeig, S. (2012). Brain oscillations in switching vs. focusing audio-visual attention. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*. 2012, 352–355. doi: 10.1109/EMBC.2012.6345941

Rubio, S., Díaz, E., Martín, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl. Psychol*. 53, 61–86. doi: 10.1111/j.1464-0597.2004.00161.x

Sackl, A., Schatz, R., and Raake, A. (2017). More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services. *Quality User Exp*. 2, 1–27. doi: 10.1007/s41233-016-0004-z

Schnotz, W., and Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educ. Psychol. Rev*. 19, 469–508. doi: 10.1007/s10648-007-9053-4

Schöffler, M. (2017). *Overall Listening Experience - a new Approach to Subjective Evaluation of Audio* (Ph.D. thesis).

Shukla, A., Harper, M., Pedersen, E., Goman, A., Suen, J. J., Price, C., et al. (2020). Hearing loss, loneliness, and social isolation: a systematic review. *Otolaryngol. Head Neck Surg*. 162, 622–633. doi: 10.1177/0194599820910377

Skowronek, J., and Raake, A. (2014). Assessment of cognitive load, speech communication quality and quality of experience for spatial and non-spatial audio conferencing calls. *Speech Commun*. 66, 154–175. doi: 10.1016/j.specom.2014.10.003

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instruct*. 4, 295–312. doi: 10.1016/0959-4752(94)90003-5

Talsma, D., Doty, T. J., and Woldorff, M. G. (2006). Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17, 679–690. doi: 10.1093/cercor/bhk016

Vaden, K. I. Jr, Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., and Eckert, M. A. (2013). The cingulo-opercular network provides word-recognition benefit. *J. Neurosci*. 33, 18979. doi: 10.1523/JNEUROSCI.1417-13.2013

Ward, C. M., Rogers, C. S., Van Engen, K. J., and Peelle, J. E. (2016). Effects of age, acoustic challenge, and verbal working memory on recall of narrative speech. *Exp. Aging Res*. 42, 97–111. doi: 10.1080/0361073X.2016.1108785

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *J. Neurosci*. 32, 14010–14021. doi: 10.1523/JNEUROSCI.1528-12.2012

Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., and Cox, C. L. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *J. Am. Acad. Audiol*. 17, 487–497. doi: 10.3766/jaaa.17.7.4

Zhu, Y., Zhai, G., Min, X., and Zhou, J. (2020). The prediction of saliency map for head and eye movements in 360 degree images. *IEEE Trans. Multimedia* 22, 2331–2344. doi: 10.1109/TMM.2019.2957986

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Neuroscience

ORIGINAL RESEARCH
published: 07 January 2022
doi: 10.3389/fnins.2021.757679

Assessing the Effect of the Refresh Rate of a Device on Various Motion Stimulation Frequencies Based on Steady-State Motion Visual Evoked Potentials

Chengcheng Han[1], Guanghua Xu[1,2], Xiaowei Zheng[1], Peiyuan Tian[1], Kai Zhang[1], Wenqiang Yan[1,2], Yaguang Jia[1] and Xiaobi Chen[1]*

[1] School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China, [2] State Key Laboratory for Manufacturing System Engineering, Xi'an Jiaotong University, Xi'an, China

OPEN ACCESS

Edited by:
Lu Zhang,
Institut National des Sciences Appliquées de Rennes, France

Reviewed by:
Yasue Mitsukura,
Keio University, Japan
Yunyong Punsawad,
Walailak University, Thailand

****Correspondence:***
Guanghua Xu
ghxu@xjtu.edu.cn

Specialty section:
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

Received: *12 August 2021*
Accepted: *15 December 2021*
Published: *07 January 2022*

Citation:
Han C, Xu G, Zheng X, Tian P, Zhang K, Yan W, Jia Y and Chen X (2022) Assessing the Effect of the Refresh Rate of a Device on Various Motion Stimulation Frequencies Based on Steady-State Motion Visual Evoked Potentials. Front. Neurosci. 15:757679. doi: 10.3389/fnins.2021.757679

The refresh rate is one of the important parameters of visual presentation devices, and assessing the effect of the refresh rate of a device on motion perception has always been an important direction in the field of visual research. This study examined the effect of the refresh rate of a device on the motion perception response at different stimulation frequencies and provided an objective visual electrophysiological assessment method for the correct selection of display parameters in a visual perception experiment. In this study, a flicker-free steady-state motion visual stimulation with continuous scanning frequency and different forms (sinusoidal or triangular) was presented on a low-latency LCD monitor at different refresh rates. Seventeen participants were asked to observe the visual stimulation without head movement or eye movement, and the effect of the refresh rate was assessed by analyzing the changes in the intensity of their visual evoked potentials. The results demonstrated that an increased refresh rate significantly improved the intensity of motion visual evoked potentials at stimulation frequency ranges of 7–28 Hz, and there was a significant interaction between the refresh rate and motion frequency. Furthermore, the increased refresh rate also had the potential to enhance the ability to perceive similar motion. Therefore, we recommended using a refresh rate of at least 120 Hz in motion visual perception experiments to ensure a better stimulation effect. If the motion frequency or velocity is high, a refresh rate of≥240 Hz is also recommended.

Keywords: motion perception, refresh rate, brain computer interface (BCI), steady-state motion visual evoked potential (SSMVEP), electroencephalogram (EEG)

INTRODUCTION

An accurate presentation of stimuli is a prerequisite for accurate results of visual perception experiments. However, most modern monitors present motion objects at discrete locations, showing an approximate continuous motion process (Chapiro et al., 2019). Motion blur will occur when the motion speed or motion frequency is too high; it violates the assumption of smooth motion and causes stimulus distortion and false experimental results (Tourancheau et al., 2009; Watson and Ahumada, 2011). Since motion blur is positively related to speed and inversely related to the refresh rate of a device, increasing the refresh rate of the monitor used in experiments has become a mainstream choice in visual motion perception research. However, hardware cost and software compatibility limit the maximum refresh rate that can be employed. Therefore, it is an important part of visual motion perception research to analyze the effect of the monitor refresh rate and select the monitor with an appropriate refresh rate according to the stimulus.

Many studies have been conducted on the effect of the refresh rate of a device on visual motion perception with psychological methods. For example, Mulholland et al. (2015) used the contrast thresholds method to evaluate the effect of the refresh rate on temporal summation, and a CRT (cathode-tube-ray) monitor with refresh rates of 60 and 160 Hz was used as the visual stimulation device. Kime et al. (2016) evaluated perceptual performance with a digital micromirror device (DMD) with a high refresh rate of 1,000 Hz and a normal refresh rate of 60 Hz. Denes et al. (2020) asked participants to observe motion stimuli at different speeds (15, 30, and 45 deg/s) on liquid crystal displays (LCDs) with different refresh rates (50–165 Hz) and evaluated the display quality with participative just-noticeable differences (JND) indicators. However, due to differences in motion stimulus parameters or evaluation indicators, different studies have reached different conclusions on the effect of the refresh rate of a device on the improvement of visual motion perception. In particular, most reports use participative psychological methods or sampling theory to study the effect of the refresh rate of a device on visual motion perception (Kuroki et al., 2006, 2007; Noland, 2014); hence, it is time-consuming to conduct a continuous quantitative analysis of the interaction between the refresh rate of a device and motion frequency (or motion velocity) due the vast number of values possible for many parameters (Emoto et al., 2014). These problems limit refresh rate research in visual motion perception, and it is difficult to provide suitable suggestions for the selection of a refresh rate in visual experiments.

Brain computer interface (BCI) is a technology that directly converts brain activity into external instructions (Wolpaw et al., 2002; Nicolasalonso and Gomezgil, 2012), and many researchers have noticed the potential of BCI in visual perception and medical diagnosis (Norcia et al., 2015; Nakanishi et al., 2017; Overbeek et al., 2018; Guger et al., 2021). Among them, the steady-state visual evoked potential (SSVEP) (Middendorf et al., 2000) method uses visual stimuli of a specific frequency to induce steady-state potentials. The SSVEP method is a relatively mature electroencephalogram (EEG)-BCI technology (Guger et al., 2001; Bashashati et al., 2007), it has the characteristics of a high SNR (signal-noise ratio). Compared with the broadband distribution of noise signals, the SSVEP response has a narrow-band distribution. By defining a specific frequency, researchers can record the subtle differences of different visual stimuli responses in a short period of time. Moreover, researchers can measure the SSVEP response without noting obvious behavior, control the influence of decision criteria after the sensory or perceptual coding stage (de Lissa et al., 2020), and provide a quantitative method for visual perception research on refresh rates (Gembler et al., 2017; Nagel et al., 2018; Başaklar et al., 2019).

In addition to the commonly used flicker or pattern-reversal stimulation methods, motion stimulation can also elicit a steady-state response, which can be called steady-state motion visual evoked potentials (SSMVEPs) (Xie et al., 2011); In recent years, some researchers have tried to use the SSMVEP method to analyze the effect of the refresh rate of a device on visual motion perception. For example, Khoei et al. (2018) found that coherent trajectory SSMVEP stimuli (3 Hz) induced stronger responses at high refresh rates, and they suggested that a display with higher refresh rates (≥240 Hz) should be used to induce visual perception cortical responses. Chai et al. (2020) used the SSMVEP paradigm (8–15 Hz) to induce visual cortical responses with monitors that had refresh rates of 60 and 144 Hz. However, they reported that the refresh rate of the monitor had no significant effect on the improvement of the evoked response. These studies use flicker or size scaling as the stimulus targets, resulting in changes in brightness that interfere with the ability to achieve an accurate motion perception response. Moreover, the frequency range of the above SSMVEP experiment was limited, and the effect of the refresh rate of the monitor on high-frequency or high-speed motion was not analyzed. The design flaws of pattern and frequency in the SSMVEP paradigm made the research results incomprehensive.

The goal of this study is to offer an analysis method of the effect of the refresh rate of a device on visual motion perception using broadband flicker-free SSMVEP (Han et al., 2018). The flicker-free SSMVEP paradigm utilizes the contraction and expansion of the checkerboard texture, which has the characteristics of low flicker and concentrated spectral peaks; also, it is convenient for the analysis of response changes under different conditions. In this study, the frequency of the stimulus is set to 7–28 Hz, the motion form of the paradigm is modulated by sine waves and triangle waves, and the monitor refresh rates are 60, 120, and 240 Hz. By analyzing the difference in induced response intensity under different refresh rates, stimulation frequencies and motion forms, we comprehensively evaluate the effect of the refresh rate. Considering that the multiparameter experiment is time-consuming and easily induces visual fatigue, this study uses the sweep method to linearly modulate the stimulus frequency, quickly induce a continuous broadband visual response, and avoid interference from the evoked potential response.

MATERIALS AND METHODS

Participants

Seventeen healthy participants (with normal or corrected-to-normal vision) participated in the experiment in this study (including 7 women; age 20–25 years, average age 22 years). Before the test, all experiment participants received training to familiarize themselves with the experimental process. All participants were asked to sign informed written consent following a protocol approved by the institutional review board of Xi'an Jiaotong University and that conformed to the Declaration of Helsinki.

Environment and Data Acquisition

The visual stimulator was an ASUS PG258Q 24.5-inch LCD monitor (1,920 × 1,080 pixels, 543.7 × 302.6 mm, the actual width of each pixel was approximately 0.28 mm, and the maximum supportable refresh rate was 240 Hz). The experiment was carried out in a quiet room with general lighting. All participants were asked to sit in comfortable armchairs 65 cm in front of the LCD monitor.

The EEG signals were recorded with ZhenTec NT1 (ZhenTec Intelligence Ltd., China). The electrodes were arranged according to the international 10–20 electrode system. A total of 6 electrodes were arranged. These electrodes were placed in the occipital region (POz, PO3, PO4, Oz, O1, and O2), the reference channel was set in the unilateral earlobe (A1), and the ground channel was set in the middle of the forehead (Fpz). The acquisition device sampled EEG signals at a frequency of 1,200 Hz, the bandpass filter was set at 2–100 Hz, and the notch filter was set at 48–52 Hz. The impedance of all electrodes was kept below 5 kOhms.

Paradigm Design and Experiment Process

Our paradigm design utilized motion checkerboard patterns to construct flicker-free SSMVEP visual stimuli paradigm (Han et al., 2018), motion checkerboard paradigm have the characteristics of low contrast and low visual fatigue (Xie et al., 2016; Zheng et al., 2019, 2020b), it can avoid the effects of fatigue on the response of evoked potential. The motion checkerboard pattern consisted of eight concentric rings, and each ring was divided into 24 equal sectors of black and white squares. In the experiment, participants were asked to gaze at motion stimuli without head movement or eye movement. In order to avoid interference from surrounding stimuli, the single-target stimulation paradigm was used. Since the evoked visual potential is most affected by the parameters in the visual field center stimulus, the experiment results of the paradigm can ensure the accuracy of the analysis conclusions.

The motion displacement curve of the stimulus was modulated by a sinusoidal sweep signal (chirp) or triangular sweep signal, and the frequency increased linearly to induce a continuous wide-band steady-state visual potential. Taking a sinusoidal motion stimulus as an example, the expression of the displacement curve was constructed as

$$\mathbf{y(t)} = \mathbf{A}\,\cos(2\pi(\frac{\mathbf{a}}{\mathbf{2}}\mathbf{t}+\frac{\mathbf{f_0}}{\mathbf{2}})\mathbf{t}+\varphi_0) \tag{1}$$

where A is the motion amplitude, a is the frequency change rate, φ_0 is the initial phase, and f_0 is the start motion reversal frequency. The motion reversal frequency, which indicates the frequency of motion direction conversion, is twice the frequency of a whole period of motion. The stimulation parameter setting of the SSMVEP paradigm is shown in **Figure 1**, the viewing angle of the motion stimulus was set at 5°, the motion amplitude was set at 0.6°, the initial phase was set at 0°, the duration of the stimulation trial was set at 8.5 s, the frequency change rate was set at about 2.47 Hz/s, the start motion reversal frequency was set at 7 Hz and the end frequency was set at 28 Hz corresponding to an average motion velocity of 8.4 deg/s (2*0.6°*7) to 33.6 deg/s (2*0.6°*28).

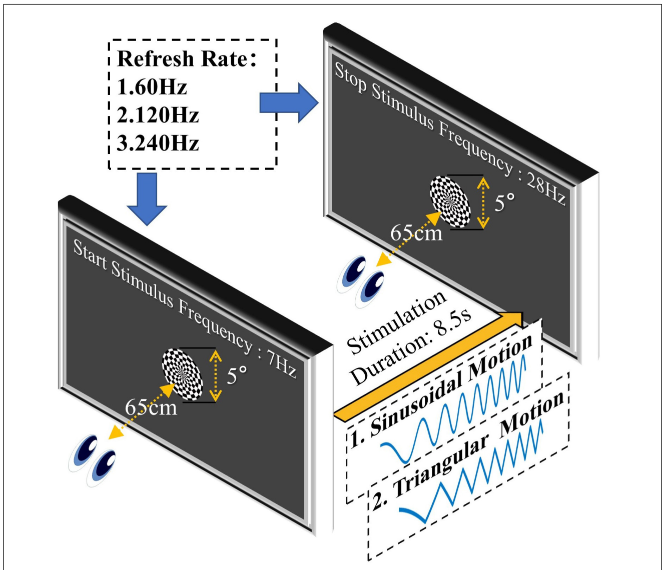
Refresh Rate:
1.60Hz
2.120Hz
3.240Hz
Stop Stimulus Frequency : 28Hz
5°
65cm
Start Stimulus Frequency : 7Hz
5°
65cm
Stimulation Duration: 8.5s
1. Sinusoidal Motion
2. Triangular Motion

FIGURE 1 | Stimulation parameter settings of the paradigm. The motion stimulus was modulated by sinusoidal or triangular sweep signals. The duration of stimulation was set at 8.5 s, the start frequency was set at 7 Hz, the end frequency was set at 28 Hz, the viewing angle of the stimulus was set at 5°, and the motion amplitude was set at 0.6°.

The motion reversal process is an important inducing factor for SSMVEPs, and the frequency of the SSMVEP generally takes the motion reversal frequency as the fundamental frequency. Therefore, the stimulus frequency mentioned in this study is equal to the motion reversal frequency.

To stabilize the visual evoked potential in advance, all participants watched the motion stimulus with the start frequency for 1 s before the formal experiment began. The experiment process is shown in **Figure 2**. In order to ensure the stability of the stimulation frequency, a photoelectric trigger device was used to test the visual paradigm before the formal experiment. The test results showed that only a few display frames have time deviations, and the error does not exceed 10-ms. When the formal experiment began, the stimulation frequency began to change. The duration of stimulation was 8.5 s, and the rest interval was 5 s. The experiment block with the same parameters was repeated 5 times. The motion paradigm was developed using MATLAB (MathWorks, Natick, United States) and Psychophysics Toolbox Version 3.

Signal Analysis

Preprocessing of Electroencephalogram Data

A bandpass filter of 2–100 Hz and a 48–52 Hz notch filter were utilized to eliminate high-frequency interference, low-frequency drifts and power frequency interference of EEG signals. The five blocks were averaged to an 8.5-s data epoch for the next step in signal processing.

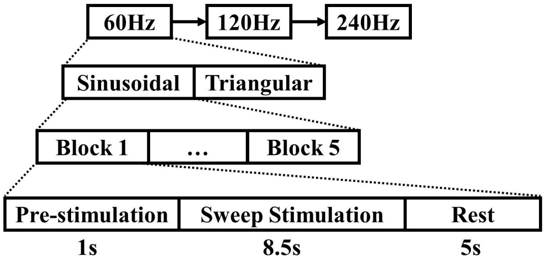

FIGURE 2 | Experiment process. The duration of stimulation was 8.5 s, the rest interval was 5 s, and the experiment block with the same parameters was repeated 5 times.

Canonical Correlation Analysis

Although Fourier transform is widely used for frequency detection with single-channel EEGs, it might still be sensitive to noise if the signal to be analyzed is from a single channel. Canonical correlation analysis (CCA) is an array signal processing method that can be used to calculate the underlying correlation between two sets of variables, it finds a pair of linear transforms for two sets and then maximizes their correlation.

CCA has been widely applied for frequency detection in multichannel visual-based BCIs (Lin et al., 2007; Zhang et al., 2020) due to its high efficiency, high robustness, high signal-to-noise ratio, and simple implementation (Bin et al., 2009; Kalunga et al., 2013; Nakanishi et al., 2015). Therefore, CCA was implemented to detect frequency components in our research.

Suppose that there are N frequencies f_1,f_2,. . .,f_N that we need to analyze. To detect the stimulation frequency, two sets of signals are introduced into CCA. One set comprises the EEG signals X from several different recording channels. The other set comprises frequency signals Y_i (i = 1, . . ., N), denotes the reference signal and is constructed as

$$\mathrm{Y_i} = \begin{pmatrix} \sin(2\pi \mathrm{f_i} \mathrm{n}) \\ \cos(2\pi \mathrm{f_i} \mathrm{n}) \end{pmatrix}, \mathrm{t} = \frac{1}{\mathrm{Fs}}, \frac{2}{\mathrm{Fs}}, \cdots, \frac{\mathrm{K}}{\mathrm{Fs}} \tag{2}$$

where F_s is the sampling rate and K is the number of sampling points. In this study, only the corresponding responses under different visual stimulations needed to be analyzed; therefore, the reference signals Y_i were only composed of sinusoid and cosinusoid pairs at the same frequency of the stimulus.

CCA can be used to find a pair of weight vectors W_x and W_{yi} to maximize the canonical correlation coefficient between linear transformations $X = X^T W_x$ and $Y_i = Y_i{}^T W_{yi}$ by the following optimization problem:

$$\underset{\mathbf{w_x}, \mathbf{w}_{\mathbf{y}i}}{\mathrm{M}ax}\ \rho(\mathbf{x}, \mathbf{y_i}) = \frac{\mathbf{E}[\mathbf{w_x^T} \mathbf{X} \mathbf{Y_i^T} \mathbf{w}_{\mathbf{y}i}]}{\sqrt{\mathbf{E}[\mathbf{w_x^T} \mathbf{X} \mathbf{X^T} \mathbf{w_x}] \mathbf{E}[\mathbf{w}_{\mathbf{y}i}^\mathbf{T} \mathbf{Y_i} \mathbf{Y_i^T} \mathbf{w}_{\mathbf{y}i}]}} \tag{3}$$

where E represents the calculation of the expected value, ρ is the canonical correlation coefficient, and x and y_i are the first pair of canonical variables. $\rho(x,y_i)$ corresponds to the maximum canonical correlation coefficient between x and y_i. When each canonical correlation coefficient of fi (i = 1, . . ., N) is calculated separately, the CCA response coefficient spectrum can be drawn by the maximum ρ of N canonical correlations.

This study used sliding window CCA spectrum analysis for time-frequency analysis. First, the 8.5-s EEG data in each block were superimposed in the time domain. Then, the EEG data were segmented according to a 0.75-s time window and a 0.25-s overlap length. A total of 32 segments were generated in this case, the frequency change range of each segment is about 0.66 Hz. Finally, CCA calculation was performed on the segmented data to obtain the correlation coefficient value. The response frequency corresponding to each segmented data was the average scanning stimulation frequency of the time window. The frequency range of the CCA coefficient spectrum analysis was set from 5 to 40 Hz, and the frequency interval was 0.2.

Statistical Analysis

Two-way repeated measures analysis of variance (ANOVA) and one-way repeated measures ANOVA were used in this study to analyze the difference and agreement between different refresh rates and stimulation frequencies. *Post hoc* comparisons with the Bonferroni correction method for multiple comparisons were also used when necessary.

Before two-way or one-way repeated measures ANOVA was performed, outliers were removed by the studentized residual analysis, and the Shapiro-Wilk test was used to test whether each group of data obeyed a normal distribution. Mauchly's test of sphericity was performed before repeated measures ANOVA was conducted. If Mauchly's test of sphericity was violated, the data were corrected by the Greenhouse-Geisser estimates of sphericity. Two-way and one-way repeated measures ANOVA were carried out by SPSS (Version 22.0 IBM, Armonk, United States).

RESULTS

Visual Evoked Potential Average Response Analysis

This subsection qualitatively analyzed the effect of refresh rate on the intensity of evoked response. First, the CCA coefficient spectrum analysis was preformed, which could present the response distribution of each subject under different stimulus conditions. Then the appropriate response frequency was selected to perform frequency response analysis, and the average evoked response intensity trend of all subjects was obtained. Finally, by dividing common EEG rhythms, the effect of refresh rate on the evoked response intensity under different frequency stimuli was presented.

Canonical Correlation Analysis Coefficient Spectrum Analysis of the Average Stimulus Response

The CCA coefficient spectrum analysis of the average stimulus response of all participants is presented in **Figure 3**. **Figures 3A,B** shows the sinusoidal motion and the triangular motion stimulation response, respectively.

The results of spectrum analysis demonstrate that the Sweep-SSMVEP paradigm evoked "single fundamental peak" responses.

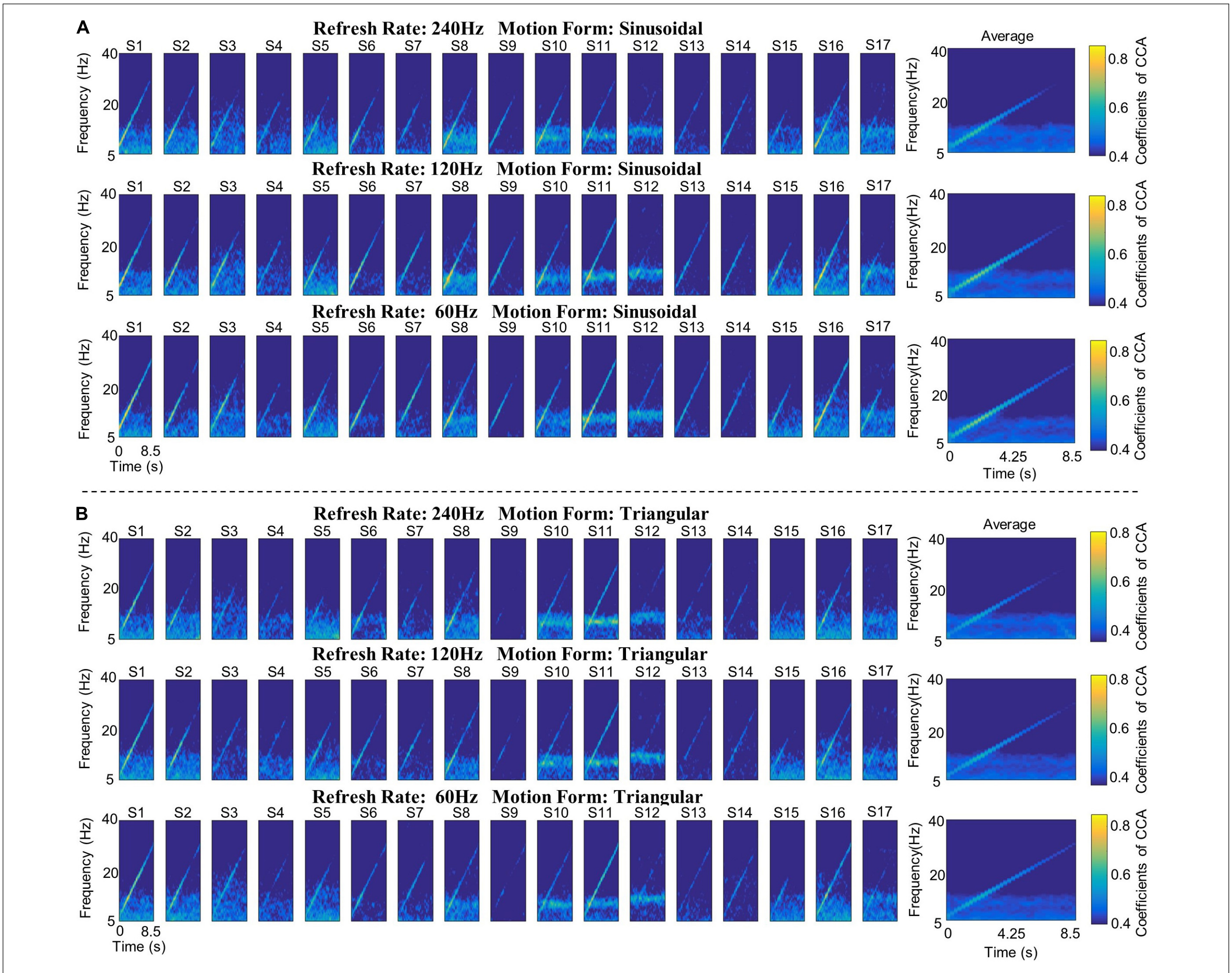

FIGURE 3 | The CCA coefficient spectrum of the Sweep-SSMVEP paradigm. **(A)** The sinusoidal motion stimulation response. **(B)** The triangular motion stimulation response. Each row represents the stimulus response at the same refresh rate, each column (column 1–17) represents the stimulus response of the same participant, and the last column represents the average response of all participants under different refresh rates. In the CCA coefficient spectrum, the vertical axis indicates the response frequency, the horizontal axis indicates the stimulation duration, and the color indicates the value of the CCA coefficients.

In other words, the fundamental frequency components of the Sweep-SSMVEP response (7–28 Hz) were prominent, whereas the higher-order harmonics were barely invisible. In addition, different motion forms and different refresh rates had little effect on high-order harmonic harmonics; therefore, in the subsequent analysis, the fundamental frequency response components were mainly considered evaluation indices.

Frequency Response Analysis of Fundamental Frequency

The average fundamental frequency responses of all participants are presented in **Figure 4**. **Figures 4A,B** show the frequency response of sinusoidal motion stimulation and triangular motion stimulation, respectively. To compare the effect of the refresh rate of the monitor on the evoked potential response under different frequencies, the stimulation frequencies were divided into three ranges according to the EEG rhythm: alpha wave (7–14 Hz), low beta wave (14–21 Hz) and middle beta wave (21–28 Hz).

The changing trend of the fundamental frequency response of the Sweep-SSMVEP paradigm was similar to that of the flicker SSVEP paradigm. The response amplitude reached the highest value when the stimulation frequency was approximately 10 Hz and then dropped as the stimulation frequency increased. Furthermore, the results of frequency responses demonstrated that refresh rates of visual motion stimulation significantly influence the intensity of the evoked potential and that the law of effect is also related to the frequency or form of stimulation. The results show that the sinusoidal motion stimulation response intensities under refresh rates of 120 Hz (Average CCA: 0. 0.4623) and 240 Hz (Average CCA: 0. 0.4771) were both higher than that under a refresh rate of 60 Hz (Average CCA: 0.4226) with an

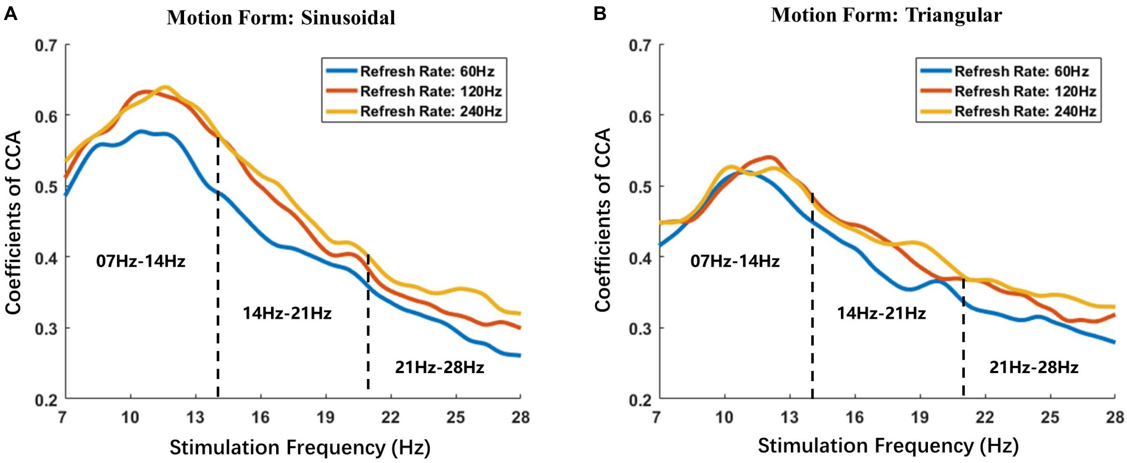

FIGURE 4 | Fundamental average frequency responses of all participants. **(A)** The frequency response of sinusoidal motion stimulation. **(B)** The frequency response of triangular motion stimulation. The horizontal axis indicates the stimulation frequency at the corresponding time, and the vertical axis indicates the CCA coefficient under the corresponding stimulation frequency. The blue solid lines depict the average visual response of all participants to stimulation with a refresh rate of 60 Hz, the red solid line depicts the average visual response to stimulation with a refresh rate of 120 Hz and the yellow solid line depicts the average visual response to stimulation with a refresh rate of 240 Hz. The black dotted line depicts the frequency divisions.

average increase of 8.8 and 12.4%, respectively. Moreover, the triangular motion stimulation response intensity under refresh rates of 120 Hz (Average CCA: 0.4162) and 240 Hz (Average CCA: 0.4236) were also both higher than that under a refresh rate of 60 Hz (Average CCA: 0.3915), with an average increase of 6.5 and 8.8%, respectively.

Average Response Intensity in Different Stimulation Frequency Bands

As shown in the average stimulation response boxplot (**Figure 5**), the alpha wave (7–14 Hz), low beta wave (14–21 Hz), and middle beta wave (21–28 Hz) responses of all participants were averaged. To distinguish parameters, a different color was used to indicate different refresh rates. The box plot results suggest that, in general, a high refresh rate can induce a higher visual potential response than a low refresh rate, and sinusoidal motion stimulation can induce a higher visual potential response than triangular motion stimulation. These data were used in subsequent statistical analyses.

The Effect of the Refresh Rate of the Monitor on the Sinusoidal Visual Motion Stimulation Response

The two-way repeated measures ANOVA was applied in this subsection to analyze the effect of the refresh rate on the sinusoidal visual motion stimulation response. First, it was necessary to determine the interaction effect of the refresh rate and stimulation frequency, that is, to find out whether the refresh rate have a differentiated effect under different frequency stimulations. When the interaction effect between refresh rate and stimulation frequency was determined, one-way repeated measures ANOVA was used to perform simple effect analysis in each frequency band, respectively, which could determine the response intensity significant difference under different refresh rates. If the one-way repeated measures ANOVA show that the refresh rate will have significant different effects on the evoked response in a certain stimulation frequency range, then the *post hoc* comparisons analysis could be further carried out to determine the refresh rate response intensity difference between each other, finally obtained specific statistical analysis results.

Analysis of the Interaction Effect of the Refresh Rate and Stimulation Frequency

The CCA coefficient data of the sinusoidal stimulation response satisfied the conditions of two-way repeated measures ANOVA, and the distribution of response data obeyed a normal distribution and satisfied the sphericity property [Mauchly's test of sphericity, $\chi^2(9) = 4.17$, $P = 0.043 > 0.05$].

The outcomes of the analysis suggest that the interaction effect of the refresh rate and stimulation frequency had a statistically significant effect on the evoked response to sinusoidal motion stimulation [$F(4, 56) = 3.30$, $P = 0.017 < 0.05$, $\eta_p{}^2= 0.19$]. Therefore, it was possible to analyze evoked response changes with different refresh rates under three frequency band sinusoidal motion stimulations separately.

The Simple Effect Analysis of Refresh Rate in Each Frequency Band

One-way repeated measures ANOVA was used to analyze the simple effect of refresh rate in each frequency band. Mauchly's test of sphericity was also used to evaluate whether the sphericity assumption was violated. The results showed that the CCA coefficient data of 7–14 Hz sinusoidal stimulation responses [$\chi^2(2) = 6.83$, $P = 0.033 < 0.05$] and 21–28 Hz stimulation responses [$\chi^2(2) = 6.16$, $P = 0.046 < 0.05$] violated Mauchly's test of sphericity, and the CCA coefficient data of 14–21 Hz stimulation responses [$\chi^2(2) = 2.41$, $P = 0.300 > 0.05$] were not violated. Then, the Greenhouse-Geisser estimates of sphericity

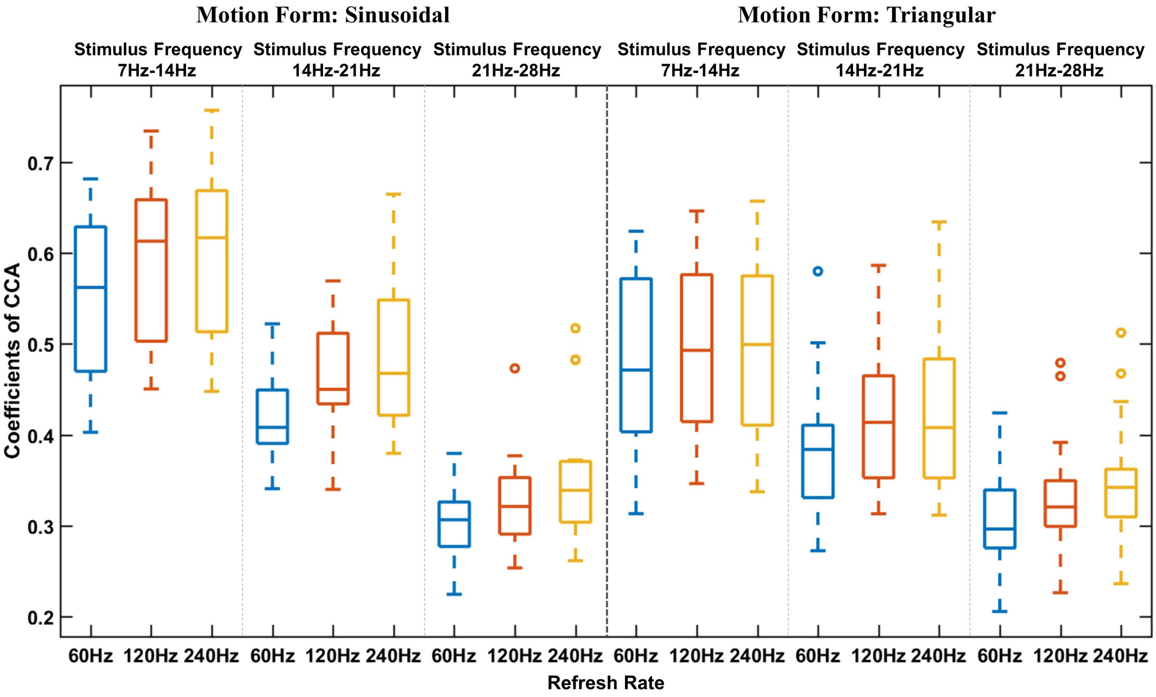

FIGURE 5 | Average stimulation response boxplot of different frequencies. The horizontal axis indicates the refresh rate, and the vertical axis indicates the average CCA coefficient. The black dotted line is used to separate different motion forms, and the blue dotted line is used to separate different stimulation frequencies.

were used to correct the CCA coefficient data ($\varepsilon_{7-14\ Hz} = 0.87$, $\varepsilon_{21-28\ Hz} = 0.73$).

The outcomes of one-way repeated measures ANOVA suggested that the refresh rate had a statistically significant simple effect on the evoked response under sinusoidal stimulation in the 7–14 Hz frequency band [$F(1.46, 23.43) = 17.24$, $P < 0.001$, $\eta_p{}^2= 0.52$], 14–21 Hz frequency band [$F(2, 32) = 15.16$, $P < 0.001$, $\eta_p{}^2= 0.49$] and 21–28 Hz frequency band [$F(1.452, 20.33) = 12.188$, $P = 0.01$, $\eta_p{}^2 = 0.47$].

Post hoc Comparisons of the Refresh Rate in Each Frequency Band

The differences in evoked responses under different refresh rates in each stimulation frequency band were compared by *post hoc* comparisons. The results are shown in **Table 1**. The column of differences of evoked responses indicates the difference between different refresh rates. The asterisk in the column of significance indicates that the difference is statistically significant at the level of $\alpha = 0.05$.

Figure 6 is a graphical display of the data in **Table 1**. **Figure 6A** shows the histogram of sinusoidal motion stimulation-evoked responses, including the mean and standard deviation. Different colors are used to indicate different refresh rates. **Figure 6B** shows the relative proportion of evoked responses under different refresh rates in each stimulation frequency. The relative average CCA coefficient of evoked responses under a refresh rate of 60 Hz was set at 1, which allowed the calculation and application of the relative average CCA coefficient under 120 and 240 Hz.

The results of *post hoc* comparisons demonstrate that the sinusoidal motion stimulation response intensities under refresh rates of 120 and 240 Hz were both higher than that under a refresh rate of 60 Hz with an average increase of 8.8 and 12.4%, respectively, and the differences were statistically significant. The response intensity was also higher under a refresh rate of 240 Hz than that under 120 Hz refresh rate, with an average increase of 3.3%, but the difference was not statistically significant.

Furthermore, the stimulation frequency and refresh rate had a significant interactive effect on the visual evoked potential response. As shown in **Figure 6B**, the response intensity differences between the 60 Hz refresh rate and 120 Hz refresh rate under all frequencies of stimulation were remarkable. However, the response intensity difference between the 120 Hz refresh rate and the 240 Hz refresh rate varied drastically with stimulation frequency. That is, the response intensity difference was minor at lower frequency stimulation; as stimulation frequency increased, the difference became larger, and the overall effect trend of the refresh rate on response intensity became linear.

The Effect of the Refresh Rate on the Triangular Visual Motion Stimulation Response

The analysis of the effect of the refresh rate on the triangular visual motion stimulation response was similar to the analysis process of sinusoidal motion stimulation, the methods are as follows: interaction effect analysis, simple effect analysis and *post hoc* comparisons.

TABLE 1 | The *post hoc* comparison results of sinusoidal motion stimulation with different refresh rates.

Stimulation frequency	Refresh rate	Average CCA coefficient mean (SD)	Differences of evoked responses		Significance
			Mean (S.E.)	95% CI	
7–14 Hz	60 Hz	0.546 (0.092)	120–60 Hz: 0.047 (0.009)	[0.023 0.072]	*P < 0.001**
	120 Hz	0.594 (0.094)	240–60 Hz: 0.052 (0.012)	[0.019 0.085]	*P = 0.002**
	240 Hz	0.599 (0.100)	240–120 Hz: 0.005 (0.007)	[−0.014 0.024]	*P = 1*
14–21 Hz	60 Hz	0.418 (0.053)	120–60 Hz: 0.047 (0.009)	[0.022 0.072]	*P < 0.001**
	120 Hz	0.465 (0.059)	240–60 Hz: 0.062 (0.013)	[0.026 0.097]	*P = 0.001**
	240 Hz	0.480 (0.079)	240–120 Hz: 0.015 (0.012)	[−0.018 0.047]	*P = 0.73*
21–28 Hz	60 Hz	0.295 (0.032)	120–60 Hz: 0.019 (0.005)	[0.005 0.033]	*P = 0.006**
	120 Hz	0.314 (0.035)	240–60 Hz: 0.038 (0.009)	[0.012 0.063]	*P = 0.004**
	240 Hz	0.332 (0.054)	240–120 Hz: 0.018 (0.008)	[−0.03 0.039]	*P = 0.10*

**p < 0.05.*

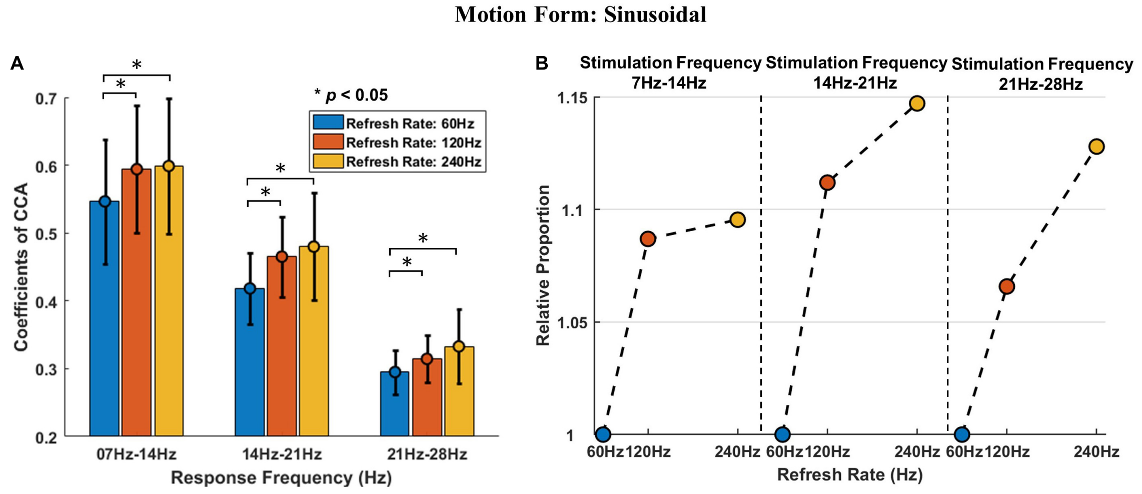

FIGURE 6 | The *post hoc* comparison results and relative proportion comparison of sinusoidal motion stimulation. **(A)** The histogram of sinusoidal motion stimulation-evoked responses, including the mean and standard deviation. The asterisks indicate that the difference is statistically significant at the level of α = 0.05. **(B)** The relative proportion of sinusoidal stimulation-evoked responses under different refresh rates in each stimulation frequency range. The vertical axis indicates the relative proportion, the horizontal axis indicates the refresh rate, and the black dotted thin line is used to separate different stimulation frequencies.

Analysis of the Interaction Effect of the Refresh Rate and Stimulation Frequency

For triangular motion stimulation, the data satisfied the sphericity property [Mauchly's test of sphericity, $\chi^2(9) = 8.21$, $P = 0.52 > 0.05$], which mean the data of triangular wave stimulation responses met the conditions of two-way repeated measurement ANOVA.

The outcomes of two-way repeated measures ANOVA also demonstrated that the interaction effect of the refresh rate and stimulation frequency was statistically significant [$F(4, 56) = 2.532$, $P = 0.05$, $\eta_p{}^2 = 0.15$].

Simple Effect Analysis of the Refresh Rate in Each Frequency Band

The results showed that the CCA coefficient of triangular stimulation response data at frequencies of 7–14 Hz [$\chi^2(2) = 0.55$, $P = 0.76 > 0.05$], 14–21 Hz [$\chi^2(2) = 0.19$, $P = 0.91 > 0.05$] and 14–21 Hz [$\chi^2(2) = 4.88$, $P = 0.087 > 0.05$] did not violate Mauchly's test of sphericity.

The outcomes of one-way repeated measures ANOVA demonstrated that the refresh rate had a statistically significant simple effect on the evoked response at the 14–21 Hz frequency band [$F(2, 30) = 10.63$,$P < 0.001$, $\eta_p{}^2$= 0.415] and 21–28 Hz frequency band [$F(2, 28) = 13.07$, $P < 0.001$, $\eta_p{}^2$= 0.483], and the simple effect of the refresh rate on the 7–14 Hz evoked response was not statistically significant [$F(2, 32) = 2.456$, $P = 0.1 > 0.05$, $\eta_p{}^2 = 0.13$].

Post hoc Comparisons of the Refresh Rate in Each Frequency Band

Consequently, only *post hoc* comparisons of responses to 14–21 Hz and 21–28 Hz stimulations were performed. The analysis results of triangular motion stimulation with different refresh rates are shown in **Table 2**. The asterisk in the column of significance indicates that the difference is statistically significant at the level of α = 0.05.

Figure 7 is a graphical display of the data in **Table 2**. **Figure 7A** shows the histogram of triangular motion stimulation-evoked

TABLE 2 | The *post hoc* comparison results of sinusoidal motion stimulation with different refresh rates.

Stimulation frequency	Refresh rate	Average CCA coefficient mean (SD)	Differences of evoked responses		Significance
			Mean (S.E.)	95% CI	
7–14 Hz	60 Hz	0.374 (0.060)	120–60 Hz: 0.034	–	–
	120 Hz	0.407 (0.065)	240–60 Hz: 0.038	–	–
	240 Hz	0.412 (0.071)	240–120 Hz: 0.004	–	–
14–21 Hz	60 Hz	0.374 (0.060)	120–60 Hz: 0.034	[0.009 0.058]	*P = 0.006**
	120 Hz	0.407 (0.065)	240–60 Hz: 0.038	[0.015 0.064]	*P = 0.001**
	240 Hz	0.412 (0.071)	240–120 Hz: 0.004	[−0.021 0.03]	*P = 1*
21–28 Hz	60 Hz	0.295 (0.038)	120–60 Hz: 0.022	[0.008 0.035]	*P = 0.002**
	120 Hz	0.316 (0.038)	240–60 Hz: 0.037	[0.014 0.061]	*P = 0.002**
	240 Hz	0.332 (0.050)	240–120 Hz: 0.016	[−0.005 0.037]	*P = 0.18*

*$p < 0.05$.

responses, including the mean and standard deviation. **Figure 7B** shows the relative proportion of evoked responses under different refresh rates in each stimulation frequency range.

Similar to the effect of refresh rate under sinusoidal motion stimulation, the results demonstrate that the triangular motion stimulation response intensity under refresh rates of 120 and 240 Hz were also both higher than that under a refresh rate of 60 Hz, with an average increase of 6.5 and 8.8%, respectively. However, this statistical significance only occurred in the triangular motion stimulation with middle (14–21 Hz) and high (21–28 Hz) frequencies. The response intensity under the 240 Hz refresh rate response intensity was also higher than that under the refresh rate of 120 Hz, with an average increase of 2.1%, and the difference was not statistically significant.

The interactive effect of stimulation frequency and refresh rate under triangular motion stimulation was also similar to the interactive effect under sinusoidal motion stimulation. The response intensity differences between the 60 Hz refresh rate and the 120 Hz refresh rate were notable, whereas the response intensity difference between the 120 Hz refresh rate and the 240 Hz refresh rate was minor under lower frequency stimulation. As the stimulation frequency increased, the difference became larger. This conclusion means that the increase in refresh rate can improve the response intensity of motion stimulation and enhance the perceptual ability of visual motion. This conclusion further confirms that the refresh rate enhances motion perception.

These conclusions from the analysis of sinusoidal and triangular motion stimulation responses mean that an increased refresh rate can improve the response intensity of motion stimulation and enhance the perceptual ability of visual motion.

The Effect of the Motion Form in Each Refresh Rate Group

The analysis of the effect of the motion form was similar to the analysis process of the effect of the refresh rate, the methods are as follows: interaction effect analysis and simple effect analysis. Besides, it is worth pointing out that the *post hoc* comparison was not applicable, because there were only two motion parameters in simple effect analysis.

Analysis of the Interaction Effect of the Motion Form and Stimulation Frequency

Similar to the above analysis, the data or corrected data of different motion forms in each refresh rate group were tested to meet the conditions of two-way and one-way repeated measurement ANOVA.

The result of Mauchly's test of sphericity showed that the CCA coefficient data of the refresh rate of 60 Hz [$\chi^2(2) = 7.84$, $P = 0.02 < 0.05$] did not violate Mauchly's test of sphericity, and the data of the refresh rates of 14–21 Hz [$\chi^2(2) = 2.59$, $P = 0.27 > 0.05$] and 14–21 Hz [$\chi^2(2) = 2.02$, $P = 0.364 > 0.05$] violated Mauchly's test of sphericity. After data correction ($\varepsilon_{60Hz} = 0.87$), two-way repeated measures ANOVA was performed on the data.

The outcomes of two-way repeated measures ANOVA suggest that the interaction effect of motion form and stimulation frequency under each refresh rate was statistically significant [$F_{60\ Hz}$ (1.42, 17.11) = 3.994, $P_{60\ Hz} = 0.048 < 0.05$, $\eta_p{}^2 = 0.25$; $F_{120\ Hz}$ (2, 24) = 6.69, $P_{120\ Hz} = 0.005 < 0.05$, $\eta_p{}^2 = 0.358$; $F_{240\ Hz}$ (2, 24) = 5.78, $P_{240\ Hz} = 0.009 < 0.05$, $\eta_p{}^2 = 0.325$].

The outcomes of the simple effect analysis of motion forms in each stimulation frequency band are shown in **Table 3**. The column of differences of evoked responses indicates the difference between sinusoidal stimulation and triangular stimulation. The asterisk in the column of significance in **Table 3** indicates that the difference is statistically significant at the level of $\alpha = 0.05$.

Figure 8 is a graphical display of the data in **Table 3**. **Figure 8** shows the mean and standard deviation of stimulation-evoked responses in each refresh rate group.

The outcomes of simple effect analysis of motion forms suggest that the simple effects of motion forms at 14–21 Hz were statistically significant under a 60 Hz refresh rate, the simple effects of motion forma at 7–4 Hz and 14–21 Hz were statistically significant under a 120 Hz refresh rate, the simple

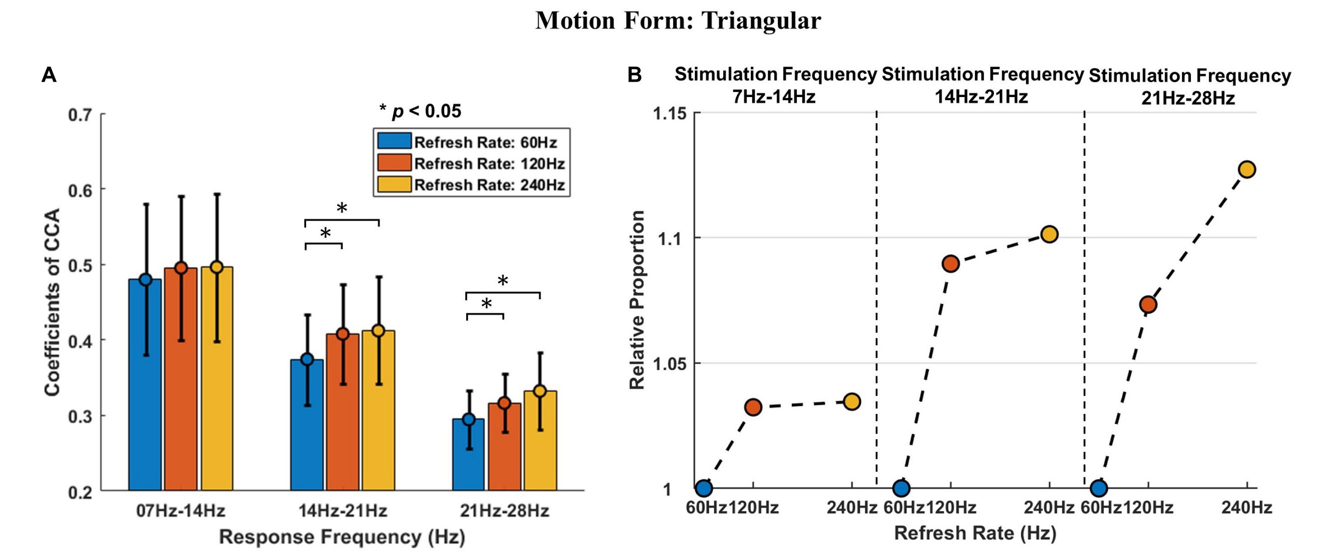

FIGURE 7 | The *post hoc* comparison results and relative proportion comparison of triangular motion stimulation. **(A)** The histogram of triangular motion stimulation-evoked responses, including the mean and standard deviation. The asterisks indicate that the difference is statistically significant at the level of α = 0.05. **(B)** The relative proportion of triangular stimulation-evoked responses under different refresh rates in each stimulation frequency range. The vertical axis indicates the relative proportion, the horizontal axis indicates the refresh rate, and the black dotted thin line is used to separate different stimulation frequencies.

TABLE 3 | The simple effect analysis results of motion forms.

Refresh rate	Stimulation frequency	Motion form	Average CCA coefficient mean (SD)	Differences of evoked responses (sinusoidal-triangular)		Significance
				Mean	95% CI	
60 Hz	7–14 Hz	Sinusoidal	0.54 (0.091)	0.067	[−0.009 0.14]	*P = 0.079*
		Triangular	0.46 (0.1)			
	14–21 Hz	Sinusoidal	0.42 (0.05)	0.049	[0.008 0.089]	*P = 0.022**
		Triangular	0.36 (0.06)			
	21–28 Hz	Sinusoidal	0.3 (0.028)	0.007	[−0.027 0.041]	*P = 0.65*
		Triangular	0.29 (0.039)			
120 Hz	7–14 Hz	Sinusoidal	0.58 (0.094)	0.099	[0.03 0.17]	*P = 0.008**
		Triangular	0.47 (0.095)			
	14–21 Hz	Sinusoidal	0.46 (0.061)	0.059	[0.008 0.11]	*P = 0.026**
		Triangular	0.4 (0.065)			
	21–28 Hz	Sinusoidal	0.32 (0.031)	0.003	[−0.032 0.037]	*P = 0.87*
		Triangular	0.31 (0.037)			
240 Hz	7–14 Hz	Sinusoidal	0.58 (0.099)	0.102	[0.027 0.18]	*P = 0.01**
		Triangular	0.47 (0.098)			
	14–21 Hz	Sinusoidal	0.47 (0.08)	0.072	[0.008 0.14]	*P = 0.029**
		Triangular	0.4 (0.071)			
	21–28 Hz	Sinusoidal	0.34 (0.055)	0.013	[−0.037 0.063]	*P = 0.58*
		Triangular	0.32 (0.036)			

**p < 0.05.*

effects of motion forms at 7–14 Hz and 14–21 Hz were statistically significant in the 240 Hz refresh rate group.

However, it is worth pointing out that although the simple effect of motion forms at the frequency of 7–14 Hz and the refresh rate of 60 Hz was not statistically significant ($P = 0.079 > 0.05$), the difference between the sinusoidal stimulation response and triangular stimulation response was noteworthy. The reason for this outcome may be the volatility of variance due to the sample size.

These results demonstrate that the visual evoked response intensity of sinusoidal motion stimulation is significantly different from that of triangular motion stimulation. However,

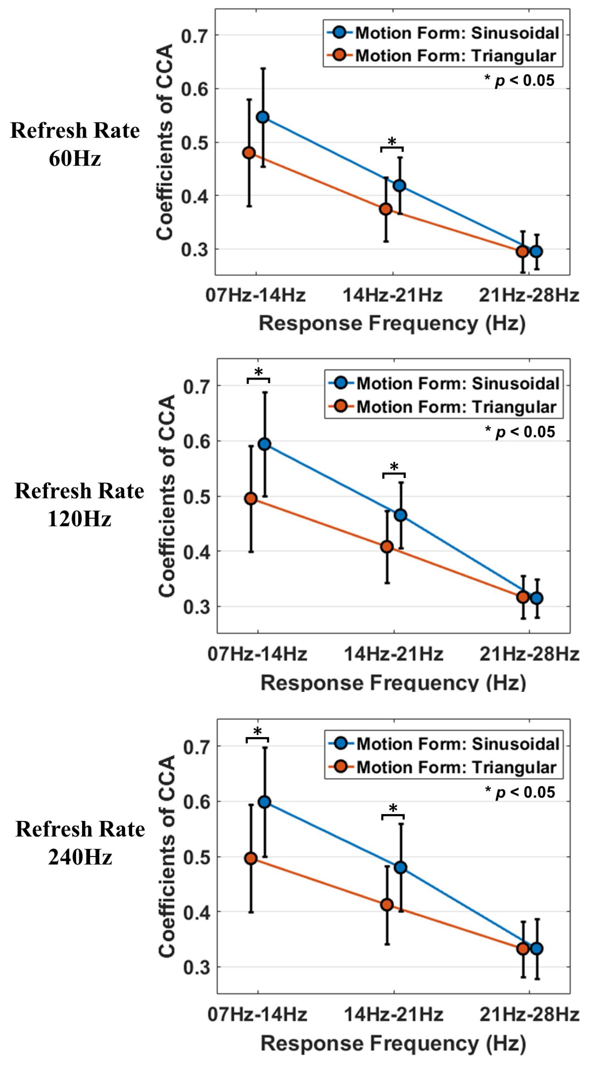

FIGURE 8 | The simple effect analysis results of motion forms in each refresh rate group. Different colors are used to indicate different motion forms, and the asterisks indicate that the difference is statistically significant at the level of α = 0.05. The vertical axis indicates the coefficients of CCA, and the horizontal axis indicates the stimulation frequency.

the difference is also related to the stimulation frequency; that is, the response difference is significant at a frequency in low and middle ranges (7–14 Hz and 14–21 Hz) whereas it is not significant at a high-range frequency (21–28 Hz).

Difference of Various Motion Forms

The relative proportion of the response intensity difference between different motion forms is shown in **Figure 9**. The response intensity difference was calculated by subtracting the average response of the triangular stimulation from the average response of sinusoidal motion stimulation, then setting the relative difference of the 60 Hz rate refresh to 1, and calculated and achieved relative proportion of other relative difference under 120 and 240 Hz.

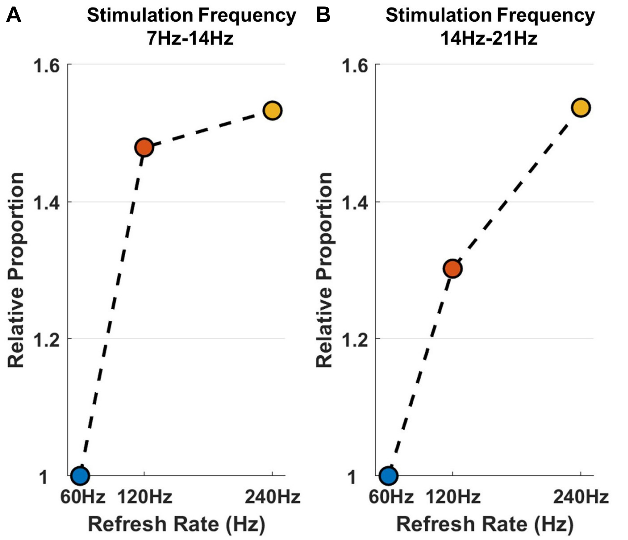

FIGURE 9 | The relative proportion of the response intensity difference between different motion forms. **(A)** The difference at the frequency of 7–14 Hz. **(B)** The difference at the frequency of 14–21 Hz. The horizontal axis and different colors indicate stimulation frequencies, and the vertical axis indicates the relative proportion of the response intensity difference between different motion forms.

Figure 9A shows the difference at frequencies of 7–14 Hz, and **Figure 9B** shows the difference at frequencies of 14–21 Hz. Since the difference in the response intensity of different motion form stimulations at a high frequency was not significant, the relative difference at frequencies of 21–28 Hz is not presented in **Figure 9**.

At frequencies of 7–14 Hz, the response difference under refresh rates of 240 and 120 Hz between sinusoidal motion and triangular motion rises by 48 and 53% on average, respectively, compared with that under a refresh rate of 60 Hz. At frequencies of 14–21 Hz, the response difference rises by 30 and 54% on average, respectively. The results demonstrate that increasing the refresh rate can increase the difference in motion visual evoked potential between sinusoidal stimulation and triangular stimulation; in particular, this effect is prominent when the motion stimulation frequency is not high. This conclusion further confirms that the refresh rate enhances motion perception. However, since data were limited by the maximum refresh rate in this study, the difference in high-frequency stimulation between different motion forms was insignificant, and it cannot be indicated that increasing the refresh rate can effectively improve the perception and ability to distinguish of high-frequency motion.

DISCUSSION

Selection of Monitor Parameters

CRT monitors have the characteristics of low latency and high stability, and they have long been the standard equipment used in visual perception research (Wiens and Öhman, 2007).

However, LCD monitors have gradually become mainstream equipment with the improvement of production technology; they are more energy-efficient and compact and show little or no visual flicker. Many studies have proven that the performance of LCD monitors is also close to that of CRT displays (Kihara et al., 2010; Lagroix et al., 2012; Bognár et al., 2016; Zhang et al., 2018; Rohr and Wagner, 2020). Therefore, the LCD monitor was chosen as the experimental equipment in this study.

A low delay response time and high refresh rate are both effective measures to improve the performance of motion stimulation (Claypool and Claypool, 2007, 2009; Spjut et al., 2019; Denes et al., 2020). The effect of the refresh rate on visual perception was a major concern in this study. To avoid interference from latency factors, a low-latency LCD monitor (ASUS ROG PG258Q) with multiple optional refresh rates was chosen as the experimental equipment. When the overdrive setting parameter of the monitor was set to "normal," the average gray-to-gray (GtG) delay response time of this monitor was approximately 4.9, 3.3, and 2.9 ms when the refresh rate reached 60, 120, and 240 Hz, respectively.[1] The GtG delay response times were all less than the refresh time, and the ghosting artifacts caused by the latency of response time were slight, so the negative impact of the delay response time was not considered.

The ultralow motion blur (ULMB) function of the monitor was not enabled in the experiment. Although this function helps reduce motion blur to a degree (Zhang et al., 2018), it is a technology only available in high-end monitoring and causes flicker sensation and visual fatigue. In the experimental SSMVEP paradigm, the stimulus was internal texture motion in a circle with fixed size and position, the participants were required to gaze at the target stimulus without eye movement, and the positive impact of the ULMB technique was further limited.

The Effect of the Refresh Rate on the Intensity of Steady-State Response, Which Can Be Called Steady-State Motion Visual Evoked Potentials

SSMVEPs are induced by the perception of stable frequency visual motion stimulation. In previous studies, it was found that the SSMVEP has the characteristic of a single peak, the evoked response energy is concentrated (Han et al., 2018). This characteristic makes the steady-state motion paradigm very suitable as a "probe" for non-invasive visual perception research. At present, a vision assessment method using steady-state motion visual evoked potential has been proposed (Zheng et al., 2019) to achieve an objective and quantitative assessment of visual acuity. The experimental results show that the correlation and agreement between objective SSMVEP and subjective FrACT (Freiburg Visual Acuity and Contrast Test) acuity were all good (Zheng et al., 2020a), demonstrating good performance in visual perception detection for the motion visual stimulation paradigm.

Due to the refresh interval between display frames, the lower the rendering refresh rate is, the greater the possibility of causing motion blur and dispersion, which will have a negative impact on the elicitation of visual potentials. Therefore, the change in the response intensity of visual evoked potentials can be used to measure whether the display system can correctly present the motion stimulus. However, it is important to note that the amplitude of EEG-based SSVEPs or SSMVEP is very unreliable at very high frequencies (>40 Hz) or very low frequencies (<2 Hz), and the assessment method in this study cannot be used in this case.

The results of the study demonstrate that the refresh rate has a significant positive effect on the perception response of visual motion stimulation at different frequencies. The intensity of the visual evoked potential under high refresh rate stimulation is always higher than the intensity of that under low refresh rate stimulation. Similar to the results of previous research literature (DoVale, 2017), there is a range in which a plateau of slow growth is observed, the effect of refresh rate has obvious diminishing returns. The positive effect of refresh rate is most significant when the refresh rate is increased from 60 to 120 Hz, and then the positive effect gradually gets into the realms of diminishing returns as the refresh rate range continues to increase above 120 Hz. This trend has no concern with the form of motion stimulation.

In addition, the diminishing return is also related to the stimulation frequency, and the attenuation effect of improving the evoked response at low-frequency stimulation is more obvious than that at high-frequency stimulation. In other words, under high frequency (21–28 Hz) stimulation, the increment of response between 120 and 60 Hz refresh rates is similar to that between 240 and 120 Hz refresh rates. However, the increment of the response between the 120 and 60 Hz refresh rates was much larger than that between the 240 and 120 Hz refresh rates under low-frequency (7–14 Hz) stimulation.

The reason for this phenomenon may be related to the adequacy of the spatiotemporal sampling of the stimulus motion. Adelson and Bergen (1985) developed a model of motion detection in which spatiotemporal filtering is used to detect motion energy of luminance-defined motion. Fujii et al. (2018) reported smoother motion in high frame rate content should activate the central nerve of vision more effectively because it produces more motion energy than low frame rate stimulus based on these models. This mechanism explains why there is a significant interaction between the refresh rate and the stimulation frequency in our experiment. In other words, high-frequency motion under low refresh rate has poor smoothness, increasing the refresh rate in this case can obviously improve the motion energy of visual stimulation, but increasing the refresh rate is of little significance for smoother motion.

It has been long known that the mammalian visual system is highly sensitive to motion, even when presented briefly. The

[1] https://www.tftcentral.co.uk/reviews/asus_rog_swift_pg258q.htm

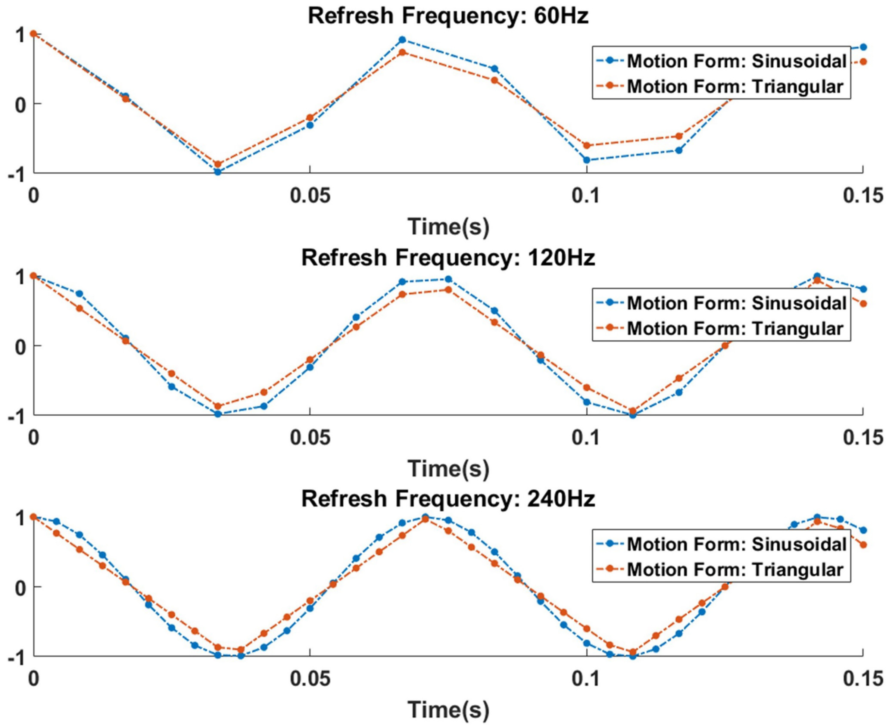

FIGURE 10 | The similarity of sinusoidal and triangular motion at different refresh rates. The vertical axis indicates displacement of stimulation, and the horizontal axis indicates stimulation time.

middle temporal visual area is a region of the extrastriate visual cortex in primates that has been demonstrated to be critical for motion vision. Area MT has among the shortest response latencies in the extrastriate cortex (Schmolesky et al., 1998; Born and Bradley, 2005). In 2011, researchers used three refresh rates to investigate how changes in the CRT (cathode ray tube) temporal stimulus affect cortical responses in tree shrew V1 (the primary visual cortex), they find that refresh rate had a large impact on firing rate and the amplitude of LFP (120 Hz > 90 Hz > 60 Hz). Since mean firing rate is positively correlated with refresh rate, V1 acts like a high-pass filter for sparse noise stimuli as a function of refresh rate (Potter et al., 2014). Furthermore, researchers found the minimum timescale for motion encoding by ganglion cells of cat retinal was 4.6 ms and depended non-linearly on temporal frequency in 2011 (Borghuis et al., 2019). These anatomical evidences from retinal nerves to higher visual cortex nerves demonstrated that the perception frequency of human vision for continuous motion may be much higher than previously speculated. Therefore, consider of the display hardware burden, we choose 120 Hz refresh rate as a conservative estimate of the optimal motion presentation parameters.

The Effect of the Refresh Rate on the Ability to Distinguish Motion Forms

The visual evoked responses caused by various motion forms are different (Teng et al., 2011; Grgič et al., 2016; Labecki et al., 2016). In this study, the experiments verified that there were significant differences in the intensity of visual evoked potentials between sinusoidal and triangular motion stimulation, and the response of sinusoidal motion stimulation was higher than that of triangular motion stimulation in general. The reason may be the difference in continuity in the motion reversal process. The motion reversal process is an important way to induce SSMVEPs, and a continuous and clear motion reversal process can improve the evoked response. In the triangular motion stimulation, the absolute value of speed always remains constant, and the rendering points are evenly distributed in the motion trajectory. In the sinusoidal motion stimulation, the rendering points are more concentrated around the reversal position, and the motion reversal process is more continuous, so the inducing effect of sine motion stimulation is superior.

The results of this analysis show that the difference in evoked potential response intensity between different motion forms increases with the refresh rate. In other words, the increase in refresh rate can improve the ability to distinguish between similar visual motions. This conclusion further demonstrates the positive effect of the rate refresh on the perception response to visual motion stimulation. Moreover, the difference in response intensity between different motion stimulations is also affected by the stimulation frequency. The difference is significant under low-frequency stimulation, but as the frequency increases, the difference decreases until it is not significant. Therefore, the changes in response difference under medium- and low-frequency stimuli were mainly analyzed to evaluate the effect of the refresh rate.

The reason for this phenomenon is also obvious, as shown in **Figure 10**. In this figure, the results of a 14 Hz (motion reversal frequency) motion stimulation process of sinusoidal and triangular structures at different refresh rates are depicted. It is difficult to distinguish the displacement details of different forms under a low refresh rate. With the increase in the refresh rate, the details of displacement are gradually improved, and different motion forms can be distinguished.

Refresh Rate Selection With Different Stimulation Frequencies

The above analysis determined the effect of the refresh rate on the response intensity of evoked potentials with different stimulation frequency bands. Within the frequency range (7–28 Hz) set by the experiment, the response intensity and motion discrimination ability at a refresh rate of 60 Hz are significantly lower than those at refresh rates of 120 and 240 Hz. This result suggests that a monitor with a refresh rate of 60 Hz has a limited ability to present fast motion stimulation. Therefore, unless special displays such as VR devices must be used or the frequency of motion stimulation is very low, the findings study signify that a monitor with a high refresh rate (120 Hz or above) should be chosen to ensure accurate motion presentation in visual perception or BCI experiments.

Furthermore, although the positive effect gradually enters the realm of diminishing returns as the refresh rate range continues to increase above 120 Hz, the decay trend is not significant when the stimulation frequency is high, and choosing a monitor with an ultrahigh refresh rate (240 Hz or above) is also of considerable significance. It is recommended, if conditions permit, to choose a monitor with an ultrahigh refresh rate according to the motion frequency or speed.

This study proposes a visual motion perception assessment method based on visual electrophysiological signals. A flicker-free Sweep-SSMVEP paradigm was designed and utilized to assess the effect of the refresh rate on motion stimulation of different frequencies. The results demonstrate that the refresh rate had a positive effect and improved visual motion perception, and the refresh rate also had a significant interaction between the refresh rate and stimulation frequency. In future studies, we will examine the impact of motion perception with eye movement on visual evoked potentials and improve the assessment method to make it suitable for visual motion perception at extreme frequencies or extreme velocities.

Visual Fatigue and Limitation

In the research of visual perception, long-term viewing of strong stimuli may cause adaptation effect (Heinrich and Bach, 2001; Priebe et al., 2002) and visual fatigue (Cao et al., 2014), resulting in changes of the evoked potential amplitude that interfere with the ability to achieve an accurate motion perception response.

Therefore, we improved the stimulation pattern to minimize the limitation of visual fatigue. The steady-state motion reversal stimulation was used as stimulation pattern in the visual motion perception experiment, the steady-state motion reversal stimulation can overcome the high susceptibility to adaptation (Heinrich and Bach, 2003) and also has a good long-term fatigue resistance (Xie et al., 2016; Zheng et al., 2020b). Furthermore, the sweep signal was used to modulated the motion stimulation, which greatly reduced the experiment time. Therefore, the total visual stimulation time of each subject is less than 5 min, there would be no obvious fatigue problems during this period of time.

However, visual fatigue still limits the development of this research. The experiment only uses the common refresh rate of 60, 120, and 240 Hz. Although the results show that the refresh rate of 60–120 Hz has a significant effect on the motion visual response, in order to control the duration of experiment, accurate segmentation of the refresh rate is not performed which leads to an inability to determine the influence trend detail of refresh rate.

CONCLUSION

The implications of this study are that it proposes an objective, reliable, visual electrophysiological method and assesses the effect of the refresh rate on motion stimulation at different frequencies with the method. The results demonstrated that an increase in the refresh rate significantly improved the intensity of sinusoidal motion visual evoked potentials at the three stimulation frequency ranges of 7–14 Hz [$F(1.46, 23.43) = 17.24$, $P < 0.001$, $\eta^2 = 0.52$], 14–21 Hz [$F(2, 32) = 15.16$, $P < 0.001$, $\eta^2 = 0.49$], and 21–28 Hz [$F(1.452, 20.33) = 12.188$, $P = 0.01$, $\eta^2 = 0.47$]. The intensity of the response at refresh rates of 240 and 120 Hz increased by 8.8 and 12.4% on average, respectively, compared with that at a refresh rate of 60 Hz. There was a significant interaction between the refresh rate and sinusoidal motion frequency [$F(4, 56) = 3.30$, $P = 0.017 < 0.05$, $\eta^2 = 0.19$], and the effect of the refresh rate more easily reached diminishing returns at lower frequencies. Furthermore, the increased refresh rate also had the potential to enhance the ability to perceive similar motion. Therefore, a refresh rate of at least 120 Hz is recommended for motion visual perception experiments to ensure a better stimulation effect, if the motion frequency or velocity is high, a refresh rate of 240 Hz or higher is also recommended.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of Xi'an Jiaotong University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CH and GX contributed to conception and design of the study. CH wrote the first draft of the manuscript and performed the statistical analysis. PT and XZ wrote sections of the manuscript. KZ, WY, YJ, and XC organized the database. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was supported by the Key Research and Development Program of Shaanxi Province of China (2021GXLH-Z-008), the Science and Technology Plan Project of Xi'an (20KYPT0001-10), the Science and Technology Plan Project of Xi'an (21RGSF0018), and the Xi'an Key Laboratory of Brain Computer Interaction and Neuroregulation Digital Medicine.

REFERENCES

Adelson, E. H., and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* 2, 284–299. doi: 10.1364/josaa.2.000284

Başaklar, T., Tuncel, Y., and Ider, Y. Z. (2019). Effects of high stimulus presentation rate on EEG template characteristics and performance of c-VEP based BCIs. *Biomed. Phys. Eng. Express* 5:035023. doi: 10.1088/2057-1976/ab0cee

Bashashati, A., Fatourechi, M., Ward, R. K., and Birch, G. E. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *J. Neural Eng.* 4, R32–R57. doi: 10.1088/1741-2560/4/2/R03

Bin, G., Gao, X., Yan, Z., Hong, B., and Gao, S. (2009). An online multi-channel SSVEP-based brain–computer interface using a canonical correlation analysis method. *J. Neural Eng.* 6:046002. doi: 10.1088/1741-2560/6/4/046002

Bognár, A., Csibri, P., András, C. M., and Sáry, G. (2016). LCD monitors as an alternative for precision demanding visual psychophysical experiments. *Perception* 45, 1070–1083. doi: 10.1177/0301006616651954

Borghuis, B. G., Tadin, D., Lankheet, M. J., Lappin, J. S., and van de Grind, W. A. (2019). Temporal limits of visual motion processing: psychophysics and neurophysiology. *Vision* 3:5. doi: 10.3390/vision3010005

Born, R. T., and Bradley, D. C. (2005). Structure and function of visual area MT. *Annu. Rev. Neurosci.* 28, 157–189. doi: 10.1146/annurev.neuro.26.041002.131052

Cao, T., Wan, F., Wong, C. M., da Cruz, J. N., and Hu, Y. (2014). Objective evaluation of fatigue by EEG spectral analysis in steady-state visual evoked potential-based brain-computer interfaces. *Biomed. Eng. Online* 13:28. doi: 10.1186/1475-925X-13-28

Chai, X., Zhang, Z., Guan, K., Zhang, T., Xu, J., and Niu, H. (2020). Effects of fatigue on steady state motion visual evoked potentials: optimised stimulus parameters for a zoom motion-based brain-computer interface. *Comput. Methods Programs Biomed.* 196:105650. doi: 10.1016/j.cmpb.2020.105650

Chapiro, A., Atkins, R., and Daly, S. A. (2019). Luminance-aware model of judder perception. *ACM Trans. Graph.* 38, 1–10. doi: 10.1145/3338696

Claypool, K. T., and Claypool, M. (2007). On frame rate and player performance in first person shooter games. *Multimed. Syst.* 13, 3–17.

Claypool, M., and Claypool, K. (2009). "Perspectives, frame rates and resolutions: it's all in the game," in *Proceedings of the 4th International Conference on Foundations of Digital Games*, Orlando, FL.

de Lissa, P., Caldara, R., Nicholls, V., and Miellet, S. (2020). In pursuit of visual attention: SSVEP frequency-tagging moving targets. *PLoS One* 15:e0236967. doi: 10.1371/journal.pone.0236967

Denes, G., Jindal, A., Mikhailiuk, A., and Mantiuk, R. K. (2020). A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution. *ACM Trans. Graph.* 39, 133:1–133:17.

DoVale, E. (2017). High frame rate psychophysics: experimentation to determine a JND for frame rate. *SMPTE Mot. Imaging J.* 126, 41–47. doi: 10.5594/jmi.2017.2749919

Emoto, M., Kusakabe, Y., and Sugawara, M. (2014). High-frame-rate motion picture quality and its independence of viewing distance. *J. Disp. Technol.* 10, 635–641.

Fujii, Y., Seno, T., and Allison, R. S. (2018). Smoothness of stimulus motion can affect vection strength. *Exp. Brain Res.* 236, 243–252. doi: 10.1007/s00221-017-5122-1

Gembler, F., Stawicki, P., Rezeika, A., Saboor, A., Benda, M., and Volosyak, I. (2017). "Effects of monitor refresh rates on c-VEP BCIs," in *International Workshop on Symbiotic Interaction*, eds J. Ham, A. Spagnolli, B. Blankertz, L. Gamberini, and G. Jacucci (Cham: Springer).

Grgič, R. G., Calore, E., and de'Sperati, C. (2016). Covert enaction at work: recording the continuous movements of visuospatial attention to visible or imagined targets by means of Steady-State Visual Evoked Potentials (SSVEPs). *Cortex* 74, 31–52. doi: 10.1016/j.cortex.2015.10.008

Guger, C., Schlogl, A., Neuper, C., Walterspacher, D., Strein, T., and Pfurtscheller, G. (2001). Rapid prototyping of an EEG-based brain-computer interface (BCI). *IEEE Trans. Neural Syst. Rehabil. Eng.* 9, 49–58. doi: 10.1109/7333.918276

Guger, C., Spataro, R., Hebb, A. O., Krusienski, D., and Prabhakaran, V. (2021). *Breakthrough BCI Applications in Medicine*. Lausanne: Frontiers Media SA.

Han, C., Xu, G., Xie, J., Chen, C., and Zhang, S. (2018). Highly interactive brain–computer interface based on flicker-free steady-state motion visual evoked potential. *Sci. Rep.* 8:5835. doi: 10.1038/s41598-018-24008-8

Heinrich, S. P., and Bach, M. (2001). Adaptation dynamics in pattern-reversal visual evoked potentials. *Doc. Ophthalmol.* 102, 141–156.

Heinrich, S. P., and Bach, M. (2003). Adaptation characteristics of steady-state motion visual evoked potentials. *Clin. Neurophysiol.* 114, 1359–1366. doi: 10.1016/s1388-2457(03)00088-9

Kalunga, E., Djouani, K., Hamam, Y., Chevallier, S., and Monacelli, E. (2013). "SSVEP enhancement based on canonical correlation analysis to improve BCI performances," in *Proceedings of the IEEE AFRICON Conference, 2013* (Piscataway, NJ: IEEE).

Khoei, M. A., Galluppi, F., Sabatier, Q., Pouget, P., Cottereau, B. R., and Benosman, R. (2018). Faster is better: visual responses to motion are stronger for higher refresh rates. *bioRxiv* [Preprint]. doi: 10.1101/505354

Kihara, K., Kawahara, J.-I., and Takeda, Y. (2010). Usability of liquid crystal displays for research in the temporal characteristics of perception and attention. *Behav. Res. Methods* 42, 1105–1113. doi: 10.3758/brm.42.4.1105

Kime, S., Galluppi, F., Lagorce, X., Benosman, R. B., and Lorenceau, J. (2016). Psychophysical assessment of perceptual performance with varying display frame rates. *J. Disp. Technol.* 12, 1372–1382. doi: 10.1109/jdt.2016.2603222

Kuroki, Y., Nishi, T., Kobayashi, S., Oyaizu, H., and Yoshimura, S. (2006). "3.4: Improvement of motion image quality by high frame rate," in *Proceedings of the SID Symposium Digest of Technical Papers* (Hoboken, NJ: Wiley Online Library).

Kuroki, Y., Nishi, T., Kobayashi, S., Oyaizu, H., and Yoshimura, S. (2007). A psychophysical study of improvements in motion—image quality by using high frame rates. *J. Soc. Inf. Disp.* 15, 61–68.

Labecki, M., Kus, R., Brzozowska, A., Stacewicz, T., Bhattacharya, B. S., and Suffczynski, P. (2016). Nonlinear origin of SSVEP spectra—a combined experimental and modeling study. *Front. Comput. Neurosci.* 10:129. doi: 10.3389/fncom.2016.00129

Lagroix, H. E., Yanko, M. R., and Spalek, T. M. (2012). LCDs are better: psychophysical and photometric estimates of the temporal characteristics of CRT and LCD monitors. *Atten. Percept. Psychophys.* 74, 1033–1041. doi: 10.3758/s13414-012-0281-4

Lin, Z., Zhang, C., Wu, W., and Gao, X. (2007). Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs. *IEEE Trans. Biomed. Eng.* 54(6 Pt 2), 1172–1176.

Middendorf, M., Mcmillan, G., Calhoun, G., and Jones, K. S. (2000). Brain-computer interfaces based on the steady-state visual-evoked response. *IEEE Trans. Rehabil. Eng.* 8, 211–214. doi: 10.1109/86.847819

Mulholland, P. J., Zlatkova, M. B., Redmond, T., Garway-Heath, D. F., and Anderson, R. S. (2015). Effect of varying CRT refresh rate on the measurement of temporal summation. *Ophthalmic Physiol. Opt.* 35, 582–590. doi: 10.1111/opo.12227

Nagel, S., Dreher, W., Rosenstiel, W., and Spüler, M. (2018). The effect of monitor raster latency on VEPs, ERPs and brain–computer interface performance. *J. Neurosci. Methods* 295, 45–50. doi: 10.1016/j.jneumeth.2017.11.018

Nakanishi, M., Wang, Y., Wang, Y.-T., and Jung, T.-P. (2015). A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. *PLoS One* 10:e0140703. doi: 10.1371/journal.pone.0140703

Nakanishi, M., Wang, Y.-T., Jung, T.-P., Zao, J. K., Chien, Y.-Y., Diniz-Filho, A., et al. (2017). Detecting glaucoma with a portable brain-computer interface for objective assessment of visual function loss. *JAMA Ophthalmol.* 135, 550–557. doi: 10.1001/jamaophthalmol.2017.0738

Nicolasalonso, L. F., and Gomezgil, J. (2012). Brain computer interfaces, a review. *Sensors* 12, 1211–1279.

Noland, K. (2014). *The Application of Sampling Theory to Television Frame Rate Requirements*. *BBC Research & Development White Paper*, Vol. 282. London: BBC.

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., and Rossion, B. (2015). The steady-state visual evoked potential in vision research: a review. *J. Vis.* 15:4. doi: 10.1167/15.6.4

Overbeek, B. U., Eilander, H. J., Lavrijsen, J. C., and Koopmans, R. T. (2018). Are visual functions diagnostic signs of the minimally conscious state? An integrative review. *J. Neurol.* 265, 1957–1975. doi: 10.1007/s00415-018-8788-9

Potter, M. C., Wyble, B., Hagmann, C. E., and McCourt, E. S. (2014). Detecting meaning in RSVP at 13 ms per picture. *Atten. Percept. Psychophys.* 76, 270–279. doi: 10.3758/s13414-013-0605-z

Priebe, N. J., Churchland, M. M., and Lisberger, S. G. (2002). Constraints on the source of short-term motion adaptation in macaque area MT. I. The role of input and intrinsic mechanisms. *J. Neurophysiol.* 88, 354–369. doi: 10.1152/jn.00852.2001

Rohr, M., and Wagner, A. (2020). How monitor characteristics affect human perception in visual computer experiments: CRT vs. LCD monitors in millisecond precise timing research. *Sci. Rep.* 10:6962. doi: 10.1038/s41598-020-63853-4

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., et al. (1998). Signal timing across the macaque visual system. *J. Neurophysiol.* 79, 3272–3278. doi: 10.1152/jn.1998.79.6.3272

Spjut, J., Boudaoud, B., Binaee, K., Kim, J., Majercik, A., McGuire, M., et al. (2019). "Latency of 30 ms benefits first person targeting tasks more than refresh rate above 60 Hz," in *Proceedings of the SIGGRAPH Asia 2019 Technical Briefs*, Brisbane, QLD, 110–113.

Teng, F., Chen, Y., Choong, A. M., Gustafson, S., Reichley, C., Lawhead, P., et al. (2011). Square or sine: finding a waveform with high success rate of eliciting SSVEP. *Comput. Intell. Neurosci.* 2011:364385. doi: 10.1155/2011/364385

Tourancheau, S., Le Callet, P., Brunnström, K., and Andrén, B. (2009). "Psychophysical study of LCD motion-blur perception," in *Proceedings of the Human Vision and Electronic Imaging XIV*, eds S. Tourancheau, P. Le Callet, K. Brunnström, and B. Andrén (Bellingham, WA: International Society for Optics and Photonics).

Watson, A. B., and Ahumada, A. J. (2011). Blur clarified: a review and synthesis of blur discrimination. *J. Vis.* 11:10. doi: 10.1167/11.5.10

Wiens, S., and Öhman, A. (2007). "Probing unconscious emotional processes: on becoming a successful masketeer," in *Handbook of Emotion Elicitation and Assessment*, eds J. A. Coan and J. J. B. Allen (New York, NY: Oxford University Press), 65–90.

Wolpaw, J. R., Birbaumer, N., Mcfarland, D. J., Pfurtscheller, G., and Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clin. Neurophysiol.* 113, 767–791.

Xie, J., Xu, G., Jing, W., Feng, Z., and Zhang, Y. (2011). Steady-state motion visual evoked potentials produced by oscillating Newton's rings: implications for brain-computer interfaces. *PLoS One* 7:e39707. doi: 10.1371/journal.pone.0039707

Xie, J., Xu, G., Wang, J., Li, M., Han, C., and Jia, Y. (2016). Effects of mental load and fatigue on steady-state evoked potential based brain computer interface tasks: a comparison of periodic flickering and motion-reversal based visual attention. *PLoS One* 11:e0163426. doi: 10.1371/journal.pone.0163426

Zhang, G.-L., Li, A.-S., Miao, C.-G., He, X., Zhang, M., and Zhang, Y. (2018). A consumer-grade LCD monitor for precise visual stimulation. *Behav. Res. Methods* 50, 1496–1502. doi: 10.3758/s13428-018-1018-7

Zhang, Y., Xie, S. Q., Wang, H., and Zhang, Z. (2020). Data analytics in steady-state visual evoked potential-based brain–computer interface: a review. *IEEE Sens. J.* 21, 1124–1138.

Zheng, X., Xu, G., Wang, Y., Han, C., Du, C., Yan, W., et al. (2019). Objective and quantitative assessment of visual acuity and contrast sensitivity based on steady-state motion visual evoked potentials using concentric-ring paradigm. *Doc. Ophthalmol.* 139, 123–136.

Zheng, X., Xu, G., Zhang, Y., Liang, R., Zhang, K., Du, Y., et al. (2020b). Anti-fatigue performance in SSVEP-based visual acuity assessment: a comparison of six stimulus paradigms. *Front. Hum. Neurosci.* 14:301. doi: 10.3389/fnhum.2020.00301

Zheng, X., Xu, G., Wu, Y., Wang, Y., Du, C., Wu, Y., et al. (2020a). Comparison of the performance of six stimulus paradigms in visual acuity assessment based on steady-state visual evoked potentials. *Doc. Ophthalmol.* 141, 237–251.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Subjective and Objective Quality Assessment of Swimming Pool Images

Fei Lei, Shuhan Li, Shuangyi Xie and Jing Liu*

Faculty of Information Technology, Beijing University of Technology, Beijing, China

As the research basis of image processing and computer vision research, image quality evaluation (IQA) has been widely used in different visual task fields. As far as we know, limited efforts have been made to date to gather swimming pool image databases and benchmark reliable objective quality models, so far. To filled this gap, in this paper we reported a new database of underwater swimming pool images for the first time, which is composed of 1500 images and associated subjective ratings recorded by 16 inexperienced observers. In addition, we proposed a main target area extraction and multi-feature fusion image quality assessment (MM-IQA) for a swimming pool environment, which performs pixel-level fusion for multiple features of the image on the premise of highlighting important detection objects. Meanwhile, a variety of well-established full-reference (FR) quality evaluation methods and partial no-reference (NR) quality evaluation algorithms are selected to verify the database we created. Extensive experimental results show that the proposed algorithm is superior to the most advanced image quality models in performance evaluation and the outcomes of subjective and objective quality assessment of most methods involved in the comparison have good correlation and consistency, which further indicating indicates that the establishment of a large-scale pool image quality assessment database is of wide applicability and importance.

Keywords: image quality assessment, subjective/objective quality assessment, swimming pool image database, main target extraction, multi-feature fusion

OPEN ACCESS

Edited by:
Xiongkuo Min,
University of Texas at Austin,
United States

Reviewed by:
Yutao Liu,
Tsinghua University, China
Wei Sun,
Shanghai Jiao Tong University, China
Weiling Chen,
Fuzhou University, China

****Correspondence:***
Fei Lei
leifei@bjut.edu.cn

Specialty section:
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

Received: *30 August 2021*
Accepted: *08 November 2021*
Published: *11 January 2022*

Citation:
Lei F, Li S, Xie S and Liu J (2022)
Subjective and Objective Quality
Assessment of Swimming Pool
Images. Front. Neurosci. 15:766762.
doi: 10.3389/fnins.2021.766762

1. INTRODUCTION

The acquisition of underwater images plays a significant role in the research of underwater rescue and biometric tracking at swimming pools in Fei et al. (2012), Alshbatat et al. (2020), and Pleština et al. (2020). However, since the underwater environment is always complicated and variable, this would lead to can result in inaccurate judgments if the unprocessed images extracted from the swimming pool are analyzed directly. Image quality assessment (IQA) has contributed significantly to the study of plentiful many visual signal applications (Wang, 2011), including image transmission, enhancement, and restoration, so the underwater image quality evaluation of swimming pools will open up the possibility for future visual research tasks. Nevertheless, to the best of our knowledge, limited efforts have been made so far to gather a database of swimming pool images and to identify a reliable benchmark for objective quality models.

In recent years, a large number of IQA approaches have been proposed, which mainly contain subjective and objective evaluation methods. Human beings, as the ultimate recipients of visual signals, have the highest voice in judging best ability to judge the quality of images. But subjective assessment methods involving humans are somewhat expensive, time-consuming, and not very useful for practical applications. Therefore, it is urgent necessary to design an objective evaluation method that can simulate the human visual system (HVS) to automatically measure the image quality. So far, these objective IQA approaches can be classified into the following three categories based on the degree of reference to the original image information: full reference (FR) method, reduce reference (RR) method, and no reference (NR) method. In the methods proposed by Gu et al. (2017a), FR IQA method, requires all the information of the original image. After decades of development, it has formed a relatively complete theoretical system and a mature evaluation framework. As the opposite of Unlike the FR method, NR IQA does not require any information of on the original image. Since it is not easy to obtain the original image in some cases, this method has attracted the attention of scholars in recent years (Gu et al., 2015b; Min et al., 2018), and RR method, which is involved in Chen et al. (2021), can obtain some information of the image. This method evaluates the image quality by comparing the difference between the extracted reference image and the partial information of the distorted image.

The most reliable FR IQA methods in the early days are have traditionally been mean square error (MSE) and peak signal-to-noise ratio (PSNR), which are statistical measurements based on image pixels. Although these methods are simple and easy to understand, the results obtained from their evaluation are very different from vary based on the subjective perceived quality of the images. Since then, there are a large number of researchers. There has been significant work carried out toward working on quality assessment models that simulate the human visual system, such as Chandler and Hemami (2007), and so on. One of the most popular algorithms based on HVS is structural similarity (SSIM) presented by Wang et al. (2004), which focuses on extracting the information of brightness, contrast, and structure from reference images. Afterwards, many extensions of the SSIM have been put forward successively. Inspired by the natural scene statistics (NNS) pointed out by Simoncelli and Olshausen (2001), Sheikh et al. resolved the IQA question from the viewpoint of information theory, and they put forward the information fidelity criterion (IFC) mentioned in Sheikh et al. (2005) and its extension version, which is called as the visual information fidelity (VIF) index in Sheikh and Bovik (2006). Zhang et al. (2011) proposed another impactive evaluation algorithm named the feature similarity (FSIM), which selects phase consistency information and gradient information as its two features. Blind parameter algorithm solves the important problem that the original image cannot be obtained. The traditional FR IQA algorithm proposes many gradient evaluation functions from the perspective of image sharpness, such as Brenner gradient function, Tenengrad gradient function, and Laplacian gradient function, etc. Through methods mentioned above can judge the level of image sharpness to a certain extent, there may be major errors for different types of images or scenes. After that, image quality assessment methods based on Natural scene statistics (NSS) emerged. The most typical model is dubbed blind/referenceless image spatial quality evaluator (BRISQUE), an RR IQA method in the spatial domain, which was proposed in Mittal et al. (2012). Other experts and scholars have also made great contributions to this kind of very practical algorithm. Gu et al. (2018) and Gu et al. (2014) have provided corresponding solutions to problems such as huge data and diverse distortion based on the RR IQA model. With the advent of the era of big data, a series of deep learning network structures have shown great advantages in the application of image processing, such as environmental protection (Gu et al., 2020a, 2021b; Liu et al., 2021), $PM_{2.5}$ forecast (Gu et al., 2019, 2021a), and air quality prediction (Gu et al., 2020b). Extensive Considerable attention from researchers has been given to evaluating image quality with deep learning (Hou et al., 2015; Liu et al., 2019) in the past few years. There is no need to define image features as, it relies on a unique deep structure to learn important features of the distorted image so as to predict the image quality score. In recent years, many scholars have improved the IQA methods mentioned above, so there are a large number of IQA methods with high accuracy and stability.

Despite the success of plentiful many IQA methods, there is still a long way to go when it comes to studying a new complex pool environment. To this end, in this paper, we created a large pool database in the first step, and then we proposed the MM-IQA model for the pool environment to objectively evaluate the quality of the database. Finally, we conducted the comparison experiments among available FR IQA and NR IQA methods on the swimming pool image database and, analyzed the advantages and disadvantages of different algorithms;, and the results show that the database is effective and valuable, which and can be used for the future visual research of the pool environment.

The rest of this paper is organized as follows. Section 2 first introduces the swimming pool underwater image dataset. In section 3, we propose an image quality evaluation method based on main object extraction and multi-feature fusion and introduce the quality evaluation method for comparison in experimental parts. Experiments and analysis conducted on our proposed database are reported in section 4. Finally, we conclude our paper in section 5.

2. SWIMMING POOL IMAGE DATASET

Although IQA has made great progress in many areas involving underwater images, very little research has been done in the last decades specifically for the particular scene of swimming pools. In order to make the underwater images of swimming pools more objective to restore the real scene and better reflect the underwater information, so as to meet the actual research needs, we construct a novel and appropriative database of swimming pool images in this paper, which are taken at different shooting angles, locations, and different brightnesses.

The process of creating the database will be described at length in the following sections.

2.1. Original Image Creation and Filtering

We selected two natatoria for data collection on the spot, one of which is the swimming pool of Ordos Stadium in Inner Mongolia, and the other is the swimming pool of North China University of Science and Technology. We used the same equipment to collect pool images and choose cameras with different angles to acquire images in order to construct a more effective dataset. As the acquisition process is continuous, the similarity of these collected photos is high. Therefore, we filter the 3,000 frames collected when selecting the reference images to obtain more images with different features. In addition, our dataset includes images of simulated drowning pools and pools without people. It is worth noting that, to further ensure the standardization of the IQA database, all reference images are selected according to the uniform size of the original image. To sum up, our presented pool underwater database includes 150 raw images with a resolution of 1920 × 1080, as shown in **Figure 1**.

2.2. Distortion Type and Distortion Level

Digital images often differ from the real environment;, for a particular scene, the distortion type should be judged first. After determining the distortion type, the performance of quality evaluation in subsequent research can be improved (Min et al., 2019a,b). There are many types of image distortion, including blurring, JPEG compression, noise injection, etc. Actually, the damaged image is complex, mainly reflected in many types of distortion, distortion degree, and so on, which requires us to fully consider all possible situations. The, integrated learning method has been proposed accordingly (Gu et al., 2017a). Considering that we are still in the early stages of this research area, we chose only one distortion type to process the database. The type of distortion chosen here is JPEG compression, which is a common lossy compression format for images. The compression process can be divided into five steps: image segmentation, color space transformation, discrete cosine transformation, data quantization, and coding.

We use the inwriter command in Matlab to generate JPEG compressed images, by setting the parameter Q, we can get images with compression levels of 10, 20, 30, 40, and 50 (distortion), as shown in **Figure 2**. In this way, we have a quality evaluations database in swimming pools.

2.3. Subjective Evaluation Process

In fact, when people evaluate the quality of an image, many factors should be taken into consideration, including not only the factors of the image itself, but also the psychological factors of subjects and the external environment. The distance between the observer and the image is studied in Gu et al. (2015a). According to ITU-R BT.500-1 in Union (2002), our subjective viewing test experiment is conducted with a single-stimulus method. In this process, we select 16 inexperienced subjects, most of whom are college students from various professional fields. The an interactive system is designed by using MATLAB, so as to automatically display the images and collect the original subjective scores, which are represented by x'_{ab}. To reduce the influence of memory on opinion scores, the presentation order is provided randomly to the observers who are asked to give their overall sensation of quality on a continuous quality scale of 1 to 5. **Table 1** summarizes many critical parameters of the subjective testing environment.

We calculate all of the gathered differential mean opinion scores (DMOS) after the viewing test experiment. Here, we denote the subjective assessment score on the distorted image I_b as a and the number of distorted images as b, where $a = \{1, ..., 16\}$ and $b = \{1, ..., 1500\}$. In addition, we set x_{ab} to indicate the score of the primitive images. Then, the following steps are shown below:

- Outliers screening: Due to the large number of test pictures, it is impossible for subjects to maintain a high level of attention at all times, which can lead to outliers. To solve this issue, we adopt the method proposed by Ponomarenko et al. (2009) to screen the outliers of the scores. Specifically, we treat this value with caution when the original DOMS value of an image is outsides the standard deviation of the mean score of this image.
- Differential scores: Subtracting the score of original images from its reference image, which can be expressed as $D_{ab} = x_{ab} - x'_{ab}$.
- Average score: The DMOS value for the image is defined as $\frac{1}{N_A}\sum_a D_{ab}$, where N_A is the number of subjects.

3. METHODOLOGY

The objective evaluation method of image quality, which realizes the accurate and automatic perception of image quality through specific formulas, replaces the subjective visual system of human eyes. In the past decades, a large number of evaluation criterions have been put forward to assess the quality of images. In this section, we will start with a detailed introduction of the MM-IQA algorithm, followed by an overview of some classic quality evaluation algorithms involved in comparison.

Higher recognition speed is desirable for underwater visual research in swimming pools, especially when it involves underwater tasks such as target recognition and tracking and rescue assistance, which often requires high speeds of recognition. Therefore, we put forward an image quality evaluation method based on main target area extraction and multi-feature fusion for swimming pool images. To begin with, because the sensitivity of vision to distortion varies in different areas, the main target area is separated from the large- scale reference image and distortion image of the swimming pool image. Then, the brightness, contrast, and gradient information extracted from the small-scale image are fused into local structure information. Finally, we obtained the image quality evaluation results by structural fusion of the two scales.

3.1. Main Target Extraction

It is known to all accepted that the information of the outside world is huge while the processing capacity of the human sensory nervous system is limited. Human visual processing can be naturally divided into two stages: the self-processing process

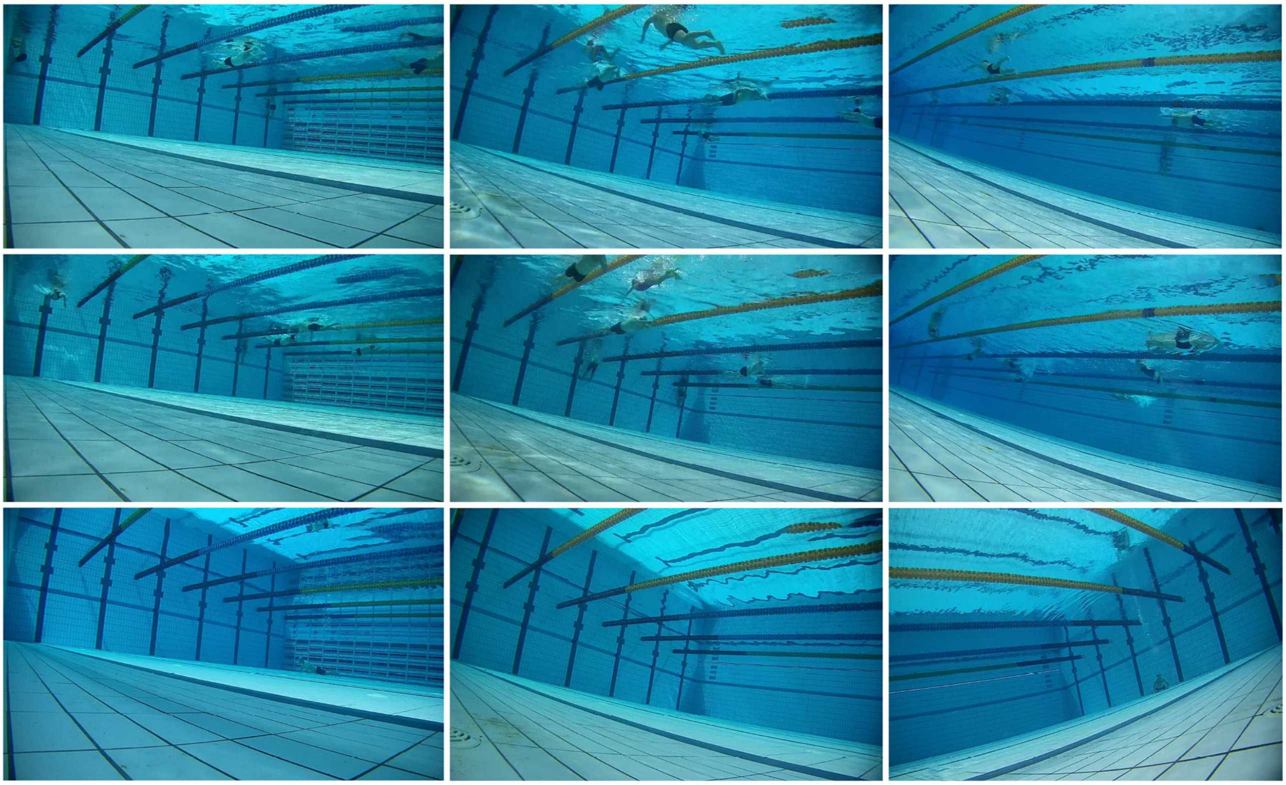

FIGURE 1 | Nine lossless color images in the swimming pool database.

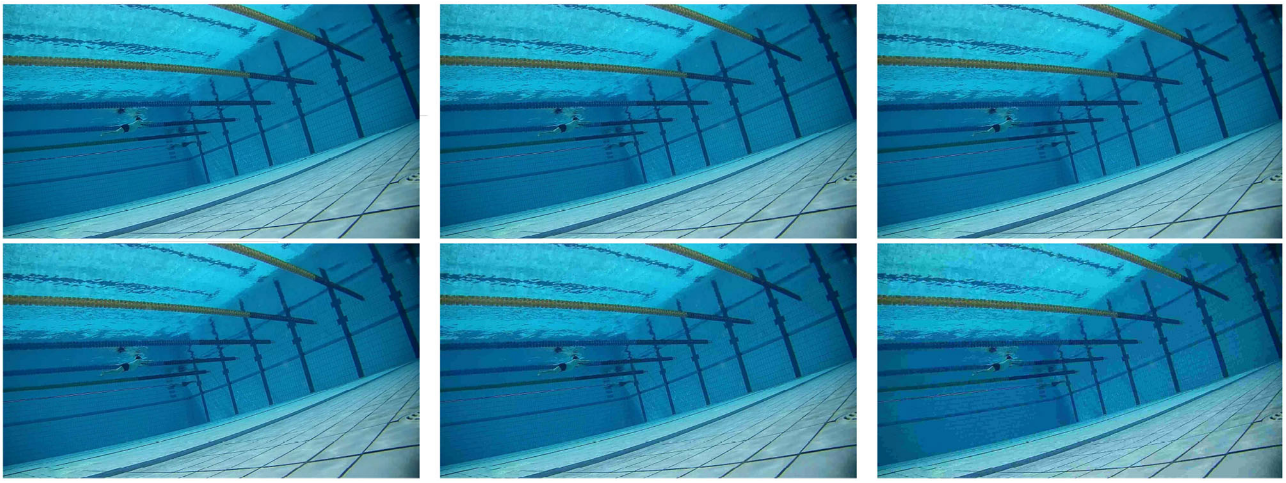

FIGURE 2 | One original image and its five distorted images vary from 10 to 50.

of distributed attention, parallel processing, and automatic feature registration;, and then, the controlled processing process of attention concentration and feature integration. Zhang et al. (2020), Tang et al. (2020), and Emoto (2019) noted, based on their observations, that the HVS tends to focus on interesting areas of the images when viewing and judging the quality of each distorted image. Furthermore, numerous studies have shown that in computer vision tasks, the method of dividing the target region into the main region first and then studying the main region can greatly accelerate the detection speed. In the paper, different degrees of distortion do not affect the location of prominent targets (pool wall, swimmer,

TABLE 1 | Subjective experimental conditions and parameters.

Method	Single-stimulus(ss)
Evaluation scales	Continuous quality scale from 1 to 5
Color depth	24
Image coder	Joint Picture Group(JPG)
Subject	Sixteen inexperienced subjects
Image resolution	1,920 × 1,080
Viewing distance	Four times the image height
Room illuminance	Dark

drowning person) in the pool. Therefore, we believe that the main target area extraction can be used as a contributor to improve the performance of the pool environmental quality assessment algorithm.

The last decade has witnessed the development and expansion of the extraction of the main target region, which has been applied in various researches studies, e.g., Image quality evaluation (Gu et al., 2016a), target tracking (Gongguo et al., 2020), and target recognition (Gu et al., 2021b). In 2007, Hou and Zhang (2007) proposed a significance detection method based on spectral residuals. After a series of operations including number spectrum analysis, spectral residuals extraction, and spatial domain mapping of the input image, the region where the main target is located was finally obtained. Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT) are known to us for the characteristics of fast detection speed and high frequency information accessibility. And the improved versions of this method these methods are used in our model for extracting the main target contour of the image. Before processing the image in frequency domain, we transformed the pixel coordinates of two-dimensional images of the spatial domain into the spectral coordinates of the frequency domain by using Fourier transform. Hence, the FFT of image $f(a, b)$ can be defined as:

$$F(\mu,\theta) = \frac{1}{PQ}\sum_{a=0}^{P-1}\sum_{b=0}^{Q-1} f(a,b)e^{-j2\pi(\frac{\mu a}{P}+\frac{\mu b}{Q})} \tag{1}$$

where P, Q represent the size information of the image, a and b are the spatial variables of the image, and μ and θ are the frequency variables of the image.

The spectrum of the image $h(x)$ is divided into amplitude spectrum $\mathcal{A}(f)$ and phase spectrum $\mathcal{P}(f)$. In order to suppress the influence of noise in the process of image acquisition, we stretched the amplitude spectrum to get keep the energy of different pixel values in a small gap interval. Then, We normalized the stretched $\mathcal{A}'(f)$ to get $\bar{\mathcal{A}}(f) = \frac{\sum \mathcal{A}(f)}{\sum \mathcal{A}'(f)}\mathcal{A}'(f)$, the spectral residual $\mathcal{R}(f)$ can be computed by subtracting the product of $\bar{\mathcal{A}}(f)$ and δ from $\bar{\mathcal{A}}(f)$. By using IFFT, the main target region map is constructed in the spatial domain. The values of each pixel in the primary target area are then squared to indicate the estimation error. Finally, smooth the saliency map was smoothed with a Gaussian filter $g(x)$ to achieve a better visual effect. The whole process is as follows:

$$\begin{aligned}
\mathcal{A}(f) &= \log(|\mathcal{F}[h(x)]|),\\
\mathcal{P}(f) &= \varphi(\mathcal{F}[h(x)]),\\
\mathcal{A}^{'}(f) &= \mathcal{A}^{\gamma}(f),\\
\mathcal{R}(f) &= \bar{\mathcal{A}}(f) - \delta\bar{\mathcal{A}}(f),\\
P_{mt} &= g(x)\cdot\mathcal{F}^{-1}[e^{\mathcal{R}(f)+\mathcal{P}(f)}]^2
\end{aligned} \tag{2}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the FFT operator and the IFFT operator, δ is the 7 × 7 identity matrix for mean filtering.

Considering that the difference of the main target area is mainly reflected in the target contour, we further extract the contour information. We select the similarity between the reference image and the distorted image as the contour information, which is a simple and effective method.

$$Con(x,y) = \frac{2P_{Mt_x}\cdot P_{Mt_y}+C1}{P_{Mt_x}^2+P_{Mt_x}^2+C1} \tag{3}$$

where the constants C1 is set to increase the stability when the denominator is close to zero.

In addition, we found that different areas of the pool contributed differently to the quality of the human perceived image. For example, it is easier to draw conclusions by observing the tiles on the pool walls and the swimmers when the distortion is low. Therefore, location information is also essential for similarity evaluation. We use $P_{Mt_w} = (Mt_x \cdot g(x)) \cup (Mt_y \cdot g(x))$ to weight the global similarity;, $g(x)$ is a gaussian matrix whose function is to eliminate noise. After adding location information, we can get the final global structure G_s :

$$G_s = \frac{\sum_{\Omega} Con(x,y)^{\psi}\cdot P_{Mt_w}(x,y)}{\sum_{\Omega} P_{Mt_w}(x,y)} \tag{4}$$

where Ω are the whole spatial domain, and parameter ψ is used to adjust the relative importance of global structure.

3.2. Multi-Feature Fusion

The pool environment is complex and easily affected by the external environment. Generally speaking, the fusion of a variety of information can make up for the deficiency, which will make the experimental results more complete and convincing (Gu et al., 2020b, 2021a). So, in order to better describe the distortion degree of the pool image, we compare the reference image with the distorted image from local brightness, local contrast, and local clarity. The characteristic of vision is non-linear, it being too bright or too dark will cause varying degrees of damage to the quality of the image. As the bottom feature of image, brightness feature will directly affect the result of image quality evaluation (Mantel et al., 2016). The basic information of the image or pixel can be obtained from the brightness characteristics. When the brightness value is lower than a certain value, the details of an image will become difficult to observe, and the image quality will also deteriorate if the image is overexposed. The average intensities of reference image x and distorted image y are calculated, respectively:

$$\mu_x = \frac{1}{N}\sum_{i=0}^{N} x_i, \mu_y = \frac{1}{N}\sum_{i=0}^{N} y_i \tag{5}$$

where μ_x and μ_y represent the local brightness of reference and distorted pool images, respectively. And then, for luminance comparison, the similarity measurement method has been used between μ_x and μ_y:

$$P_{l(x,y)} = \frac{2\mu_x \cdot \mu_y + C2}{\mu_x^2 + \mu_y^2 + C2} \tag{6}$$

where the constants C2 has the same function as C1.

As the key to the visual effect, contrast reflects the sharpness of the image and the depth of the grooves in the texture. Generally speaking, high contrast is of great help to image clarity, detail performance, and gray level performance. On the contrary, a low image contrast usually causes the whole image to be blurred. Signal contrast is mainly obtained by estimating the standard deviation (square root of variance) of the image, and the standard deviation of discrete signal is calculated as:

$$\begin{aligned} \sigma_x &= [\tfrac{1}{N-1} \textstyle\sum_{i=0}^{N} (x_i - \mu_x)]^{\frac{1}{2}}, \\ \sigma_y &= [\tfrac{1}{N-1} \textstyle\sum_{i=0}^{N} (x_i - \mu_y)]^{\frac{1}{2}} \end{aligned} \tag{7}$$

where σ_x and σ_y represent the local brightness of reference and distorted pool images, respectively. Similarly, for contrast comparison, the similarity measurement method has also been used between σ_x and σ_y:

$$P_{c(x,y)} = \frac{2\sigma_x \cdot \sigma_y + C3}{\sigma_x^2 + \sigma_y^2 + C3} \tag{8}$$

where the constant is C3 has the same function as C1 and C2.

Besides contrast and brightness, sharpness feature is another important image feature, which includes sharpness of image plane and sharpness of image edge. More attention has been paid to the edge of the image when it comes to sharpness feature (Tao et al., 2014; Sheng et al., 2015), which also makes up for the lack of contrast sensitivity in this aspect of contrast. Image edge is a set of pixels connected by the boundary between two regions of an image. We can use gradient feature to fully describe the information of image edge structure and contrast change. Commonly used operators for calculating gradients include the Sobel operator, the Prewitt operator, and the Scharr operator. Here, we used the Scharr gradient operator to extract gradient information of reference image x and distorted image y, respectively:

$$S_h = \begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix} \times \frac{1}{16}, S_v = \begin{bmatrix} 3 & -10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} \times \frac{1}{16} \tag{9}$$

where S_h and S_v are separately represent the Scharr convolution masks along the horizontal and vertical directions, which are used for gradient extraction of the image. We can obtained the gradient magnitudes of x and y, denoted as s_x and s_y, which are given by:

$$s_x = \sqrt{(S_h * x)^2 + (S_v * x)^2}, s_y = \sqrt{(S_h * y)^2 + (S_v * y)^2} \tag{10}$$

where symbol "$*$" indicates the convolution operation. Then the difference between s_x and s_y can be written as:

$$P_{s(x,y)} = \frac{2s_x \cdot s_y + C4}{s_x^2 + s_y^2 + C4} \tag{11}$$

where the constant is C4 has the same function as C1, C2, and C3.

By structure-fusion of the three local features of brightness, contrast, and sharpness in the small-scale range with the main target region extraction in the large-scale range, we obtained the final MM-IQA metric:

$$MM - IQA = \frac{\sum_{\Omega} [P_s + w_1 P_l \cdot P_c]^{\theta} Con(x,y)^{\psi} \cdot P_{Mt_w}}{\sum_{\Omega} P_{Mt_w}} \tag{12}$$

where $P_s + w_1 P_l \cdot P_c$ presents a fusion of three local features, w_1 is a weight parameter, and θ has the same function as ψ.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Performance Measures

This section will conduct a wide range of experiments on our constructed database to assess the accuracy of these methods mentioned above. The swimming pool image database is a large-scale IQA database with 1500 images generated from 150 pristine images, having 5 five distortion levels and 1 one distortion type, therefore it is chosen as the testing bed. As per the suggestion given by Corriveau (2017), we first map the prediction outputs of each IQA metrics to subjective scores using non-linear regression with the five-parameter logistic function, which is regarded as:

$$S(q) = \tau_1 \left\{ \frac{1}{2} - \frac{1}{1 + e^{(q - \tau_3)\tau_2}} \right\} + q\tau_4 + \tau_5 \tag{13}$$

where q and $S(q)$ are the input and mapped scores, and the regression model parameters τ_1 to τ_5 are to be determined during the curve fitting process.

Then, we evaluate the IQA index using five commonly used performance indicators, where the Spearman rank order correlation coefficient (SROCC) and the Kendall rank order correlation coefficient (KROCC) are applied for evaluating to evaluate the monotonicity of prediction. The third index is Pearson linear correlation coefficient (PLCC), which estimates the prediction accuracy by measuring the correlation between the MOS and objective fractions after non-linear regression. Finally, in order to evaluate the prediction consistency, we also use the Root mean square error (RMSE) and the Mean absolute error (MSE) between S(q) and q.

4.2. Methods for Comparison

In this paper, we used the classical and the latest FR IQA method and part of NR IQA method to conduct a comparative experiment with MM-IQA in the underwater database of swimming pools. The methods involved in the experiment are shown below:

- The MSE, PSNR, and SSIM proposed by Wang et al. (2004), are the benchmark IQA methods that are widely used in image processing researches.

- NQM in Damera-Venkata et al. (2000), quantifies the effects of linear frequency distortion and noise injection on HVS.
- FSIM and FSIMc from Zhang et al. (2011), apply phase congruency and gradient magnitude to represent the local quality of the image based on the fact that the HVS understands images mainly from the low-level features of the images.
- IGM in Wu et al. (2013), who decomposes the reference image into a predicted part and a disordered part according to the Bayesian prediction model. In addition, the PSNR and SSIM values are used to measure the noise energy of these two parts, respectively. Finally, we combine the two results to obtain the overall mass score.
- MS-SSIM pointed out by Wang et al. (2003), performs the SSIM in different scales and integrates their outputs with psychophysical weights.
- VIF and VIFP, quantify the Shannon information shared between the reference and distorted images in Sheikh and Bovik (2006) by using a unified information fidelity criterion based on NSS, distortion, and HVS modeling.
- MAD presented by Chandler (2010), combines two different strategies based on detection and appearance. When the quality of the image is high, local brightness and contrast masking can be used to estimate the perceptual distortion based on detection, while variations in local statistics of spatial frequency components are used to estimate appearance-based perception distortion in low-quality images.
- GSI developed by Liu et al. (2012), emphasizes on the similarity of gradient sizes plays which play an important role in scene understanding.
- GMSD is designed by Xue et al. (2014), and predicts visual quality score by using the standard deviation of the similarity graph of the gradient amplitude between the reference image and the distorted image, which meets both the time and efficiency requirements.
- VSI presented by Zhang et al. (2014), which would integrate visual saliency into IQA metrics.
- ADD1 and ADD2 in Gu et al. (2016b), new aggregation models in IQA, which proposed via analyzing the distortion distribution of image content and distortion effects.
- PSIM from Gu et al. (2017a), combines two scales of the GM similarities, both of which are color information similarity, and a reliable perceptual-based pooling, respectively.
- BRISQUE in Mittal et al. (2012), an NR IQA method based on natural scene statistics who that uses scene statistics of local normalized luminance coefficient to quantify distortion.
- NIQE pointed out by Mittal et al. (2013), is proved to be a simple and efficient quality assessment algorithm who that calculates the deviation only and only relies on the statistical rules in natural images without training the artificially assessed distorted images.
- SISBLIM proposed by Gu et al. (2014), takes the multi-distortion image problem as the research object and evaluates image quality from six parts: noise estimation, image deionizing, blur measure, JPEG-quality evaluator, joint effects' prediction, and HVS-based fusion.
- NIQMC from Gu et al. (2017b), an NR IQA based on the concept of information maximization who that considers both local and global information to generate the quality fraction of the contrast distortion image.
- ASIQE presented in Gu et al. (2017c), which quantifies the effects of image complexity, screen content statistics, overall brightness quality and detail sharpness on HVS, is commonly used to evaluate the quality of screen content images.

TABLE 2 | Performance comparison of FR-IQA metrics on the pool image database.

Metrics	SROCC	KROOC	PLCC	MSE	RMSE
MSE	0.8659	0.6591	0.4662	0.3616	0.4773
PSNR	0.8695	0.6591	0.4662	0.3537	0.4714
SSIM	0.8779	0.6940	0.5064	0.3416	0.4570
NQM	0.8546	0.6379	0.4527	0.3748	0.4955
VIF	0.8817	0.6931	0.5054	0.3380	0.4502
IGM	0.8842	0.6888	0.5014	0.3337	0.4457
FSIM	0.8835	0.6918	0.5037	0.3376	0.4469
FSIMc	0.8834	0.6885	0.5004	0.3376	0.4472
MS-SSIM	**0.8859**	**0.6976**	**0.5097**	**0.3321**	**0.4426**
MAD	0.8740	0.6734	0.4842	0.3489	0.4638
GSI	0.8820	0.6830	0.4942	0.3389	0.4497
GMSM	0.8840	0.6819	0.4948	0.3348	0.4461
GMSD	0.8833	0.6749	0.4859	0.3359	0.4474
PAMSE	0.8802	0.6740	0.4879	0.3394	0.4529
VSI	0.8799	0.6954	0.5107	0.3400	0.4534
SWGSSIM	0.8829	0.6769	0.4869	0.3375	0.4481
ADD1	0.8859	0.6944	0.5077	0.3327	0.4427
ADD2	0.8838	0.6753	0.4876	0.3358	0.4465
PSIM	0.8838	0.7197	0.5314	0.3348	0.4465
MM-IQA	**0.8934**	**0.7508**	**0.5675**	**0.3246**	**0.4287**

TABLE 3 | Performance comparison of RR-IQA metrics on the pool image database.

Metrics	SROCC	KROOC	PLCC	MSE	RMSE
BRISQUE	0.6540	0.5392	0.3783	0.5529	0.7219
SISBLIM-SM	**0.8901**	**0.7535**	**0.5734**	**0.3343**	**0.4348**
SISBLIM-WM	0.8861	0.7384	0.5560	0.3341	0.4432
NIQE	0.8787	0.7549	0.5702	0.3559	0.4555
ASIQE	0.8630	0.6851	0.5013	0.3612	0.4821
MM-IQA	**0.8934**	**0.7508**	**0.5675**	**0.3246**	**0.4287**

4.3. Overall Performance Evaluation

In order to better verify the effect of objective IQA method and subjective consistency, we test and calculate the objective IQA algorithm on a subjective IQA database. **Tables 2**, **3** illustrate the performance results of PLCC, SROCC, KROCC, RMSE, MSE of FR IQA, and NR IQA on the new pool database, respectively. At the bottom of this these two tables is the performance of MM-IQA method shown in bold, and the best models for both FR IQA and NR IQA algorithms used for comparison are also shown in bold.

The performance of the same quality evaluation algorithm varies from different databases. For the FSIM algorithm, the result of SROCC in the swimming pool image database is 0.8835, while the SROCC result of the same algorithm in the LIVE database is 0.9634, which is pointed out by Sheikh (2003). In addition, due to the good correlation between subjective score and objective evaluation results, our proposed database can also be used to compare the performance of some IQA algorithms, e.g., the extended algorithms MSSSIM obtains better performance than SSIM. We can transform the pool images into grayscale for further study in that pool images are always single singular in color. In this regard, we can conclude from the results that FSIM using gray scale images achieves better results than FSIMc. Surprisingly, the non-parametric algorithms also perform the task of visual evaluation better on the pool database, and even some of the non-parametric algorithms perform better than the mature parametric algorithms. In terms of the overall experimental results, the large-scale IQA database created in this paper shows good consistency in testing different IQA algorithms, which also proves the effectiveness of the database.

5. DISCUSSION AND CONCLUSION

As an interactive form of information, images are playing an increasingly important role in the field of multimedia. Yet the amount or importance of the information conveyed by images is not only related to the content and the format of images, but also to the image quality. In general, the higher the quality of the image, the more information people can receive and perceive by looking at the image. In At present, IQA method is becoming more and more important in the field of image processing and computer vision, and is widely used in different practical scenarios.

As a new research field, the swimming pool image research has also been more and more people's attention been gathering increasing attention in recent years, at present there are a lot of swimming pool water to carry on many areas in which to ask research questions, such as swimming pool environment anomaly detection, swimming pool body posture recognition, swimming pool, target tracking, etc., and the image quality is the basis of all vision problems, so the establishment of the swimming pool image database is very necessary. After establishing the database, we evaluated the subjective and objective image quality, respectively, then used three correlation indices, SROCC, KROCC, and PLCC, to describe the consistency between the subjective IQA approach and the objective IQA method, and finally measured the error of the objective image quality score with MOS by using MSE and RMSE. The results of the experiment show that the subjective and objective evaluation can match well, but as the swimming pool environment is easily disturbed by the external environment (such as light, shade, and water ripples). In the future, we will select more distortion types to process the images in our database and further consider the characteristics of the swimming pool environment, so as to seek a more appropriate IQA model and make contributions to the practical research.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

FL conceived the framework of the paper and implementation and wrote the manuscript. SL assisted in algorithm conception and interpretation of the results. SX participated in the revision and content supplement of the article. JL revised the layout of the article and checked for grammatical errors.

REFERENCES

Alshbatat, A. I. N., Alhameli, S., Almazrouei, S., Alhameli, S., and Almarar, W. (2020). "Automated vision-based surveillance system to detect drowning incidents in swimming pools," in *2020 Advances in Science and Engineering Technology International Conferences (ASET)* (Dubai), 1–5. doi: 10.1109/ASET48392.2020.9118248

Chandler, D. M., and Hemami, S. S. (2007). VSNR: a wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. Image Process*. 16, 2284–2298. doi: 10.1109/TIP.2007.901820

Chandler, L. D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19:011006. doi: 10.1117/1.3267105

Chen, W., Gu, K., Zhao, T., Jiang, G., and Callet, P. L. (2021). Semi-reference sonar image quality assessment based on task and visual perception. *IEEE Trans. Multimedia* 23, 1008–1020. doi: 10.1109/TMM.2020.2991546

Corriveau, P. (2017). *Video Quality Experts Group*. Boca Raton, FL: CRC Press. doi: 10.1201/9781420027822-11

Damera-Venkata, N., Kite, T. D., Geisler, W. S., Evans, B. L., and Bovik, A. C. (2000). Image quality assessment based on a degradation model. *IEEE Trans. Image Process*. 9, 636–650. doi: 10.1109/83.841940

Emoto, M. (2019). Depth perception and induced accommodation responses while watching high spatial resolution two-dimensional tv images. *Displays* 60, 24–29. doi: 10.1016/j.displa.2019.08.005

Fei, L., Xinying, Z., and Yi, W. (2012). "Real-time tracking of underwater moving target," in *Proceedings of the 31st Chinese Control Conference* (Hefei: IEEE), 3984–3988.

Gongguo, X., Ganlin, S., and Xiusheng, D. (2020). Sensor scheduling for ground maneuvering target tracking in presence of detection blind zone. *J. Syst. Eng. Electron*. 31, 692–702.

Gu, K., Li, L., Lu, H., Min, X., and Lin, W. (2017a). A fast reliable image quality predictor by fusing micro-and macro-structures. *IEEE Trans. Indus. Electron*. 64, 3903–3912. doi: 10.1109/TIE.2017.2652339

Gu, K., Lin, W., Zhai, G., Yang, X., Zhang, W., and Chen, C. W. (2017b). No-reference quality metric of contrast-distorted images based on information maximization. *IEEE Trans. Cybernet*. 47, 4559–4565. doi: 10.1109/TCYB.2016.2575544

Gu, K., Liu, H., Xia, Z., Qiao, J., Lin, W., and Thalmann, D. (2021a). $\mathrm{Pm}_2.5$ monitoring: use information abundance measurement and wide and deep learning. *IEEE Trans. Neural Netw. Learn. Syst*. 32, 4278–4290. doi: 10.1109/TNNLS.2021.3105394

Gu, K., Liu, M., Zhai, G., Yang, X., and Zhang, W. (2015a). Quality assessment considering viewing distance and image resolution. *IEEE Trans. Broadcast*. 61, 520–531. doi: 10.1109/TBC.2015.2459851

Gu, K., Qiao, J., and Li, X. (2019). Highly efficient picture-based prediction of PM2.5 concentration. *IEEE Trans. Indus. Electron*. 66, 3176–3184. doi: 10.1109/TIE.2018.2840515

Gu, K., Tao, D., Qiao, J.-F., and Lin, W. (2018). Learning a no-reference quality assessment model of enhanced images with big data. *IEEE Trans. Neural Netw. Learn. Syst*. 29, 1301–1313. doi: 10.1109/TNNLS.2017.2649101

Gu, K., Wang, S., Yang, H., Lin, W., Zhai, G., Yang, X., et al. (2016a). Saliency-guided quality assessment of screen content images. *IEEE Trans. Multimedia* 18, 1098–1110. doi: 10.1109/TMM.2016.2547343

Gu, K., Wang, S., Zhai, G., Lin, W., Yang, X., and Zhang, W. (2016b). Analysis of distortion distribution for pooling in image quality prediction. *IEEE Trans. Broadcast*. 62, 446–456. doi: 10.1109/TBC.2015.2511624

Gu, K., Xia, Z., and Qiao, J. (2020a). Deep dual-channel neural network for image-based smoke detection. *IEEE Trans. Multimedia* 22, 311–323. doi: 10.1109/TMM.2019.2929009

Gu, K., Xia, Z., and Qiao, J. (2020b). Stacked selective ensemble for PM2.5 forecast. *IEEE Trans. Instrument. Measure*. 69, 660–671. doi: 10.1109/TIM.2019.2905904

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2014). Hybrid no-reference quality metric for singly and multiply distorted images. *IEEE Trans. Broadcast*. 60, 555–567. doi: 10.1109/TBC.2014.2344471

Gu, K., Zhai, G., Yang, X., and Zhang, W. (2015b). Using free energy principle for blind image quality assessment. *IEEE Trans. Multimedia* 17, 50–63. doi: 10.1109/TMM.2014.2373812

Gu, K., Zhang, Y., and Qiao, J. (2021b). Ensemble meta-learning for few-shot soot density recognition. *IEEE Trans. Indus. Inform*. 17, 2261–2270. doi: 10.1109/TII.2020.2991208

Gu, K., Zhou, J., Qiao, J.-F., Zhai, G., Lin, W., and Bovik, A. C. (2017c). No-reference quality assessment of screen content pictures. *IEEE Trans. Image Process*. 26, 4005–4018. doi: 10.1109/TIP.2017.2711279

Hou, W., Gao, X., Tao, D., and Li, X. (2015). Blind image quality assessment via deep learning. *IEEE Trans. Neural Netw. Learn. Syst*. 26, 1275–1286. doi: 10.1109/TNNLS.2014.2336852

Hou, X., and Zhang, L. (2007). "Saliency detection: a spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (Minneapolis), 1–8. doi: 10.1109/CVPR.2007.383267

Liu, A., Lin, W., and Narwaria, M. (2012). Image quality assessment based on gradient similarity. *IEEE Trans. Image Process*. 21, 1500–1512. doi: 10.1109/TIP.2011.2175935

Liu, D., Cheng, B., Wang, Z., Zhang, H., and Huang, T. S. (2019). Enhance visual recognition under adverse conditions via deep networks. *IEEE Trans. Image Process*. 28, 4401–4412. doi: 10.1109/TIP.2019.2908802

Liu, H., Lei, F., Tong, C., Cui, C., and Wu, L. (2021). Visual smoke detection based on ensemble deep cnns. *Displays* 69:102020. doi: 10.1016/j.displa.2021.102020

Mantel, C., Søgaard, J., Bech, S., Korhonen, J., Pedersen, J. M., and Forchhammer, S. (2016). Modeling the quality of videos displayed with local dimming backlight at different peak white and ambient light levels. *IEEE Trans. Image Process*. 25, 3751–3761. doi: 10.1109/TIP.2016.2576399

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast*. 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhai, G., Gu, K., Yang, X., and Guan, X. (2019a). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transport. Syst*. 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019b). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Trans. Multimedia* 21, 2319–2333. doi: 10.1109/TMM.2019.2902097

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process*. 21:4695. doi: 10.1109/TIP.2012.2214050

Mittal, A., Soundararajan, R., and Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett*. 20, 209–212. doi: 10.1109/LSP.2012.2227726

Pleština, V., Papić, V., and Turić, H. (2020). "Swimming pool segmentation in pre-processing for tracking water polo players," in *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (Istanbul), 1–4. doi: 10.1109/ICECCE49384.2020.9179299

Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F. (2009). Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Adv. Modern Radioelectron*. 10, 30–45.

Sheikh, H., Bovik, A., and de Veciana, G. (2005). An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process*. 14, 2117–2128. doi: 10.1109/TIP.2005.859389

Sheikh, H. R. (2003). *Live Image Quality Assessment Database*. Available online at: http://live.ece.utexas.edu/research/quality

Sheikh, H. R., and Bovik, A. C. (2006). Image information and visual quality. *IEEE Trans. Image Process*. 15, 430–444. doi: 10.1109/TIP.2005.859378

Sheng, J., Xing, M., Zhang, L., Mehmood, M. Q., and Yang, L. (2015). Isar cross-range scaling by using sharpness maximization. *IEEE Geosci. Remote Sens. Lett*. 12, 165–169. doi: 10.1109/LGRS.2014.2330625

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci*. 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

Tang, X. T., Yao, J., and Hu, H. F. (2020). Visual search experiment on text characteristics of vital signs monitor interface. *Displays* 62:101944. doi: 10.1016/j.displa.2020.101944

Tao, Y., Zheng, X., Xuan, H., Wei, Z., Wang, W., and Pelli, D. G. (2014). A method for the evaluation of image quality according to the recognition effectiveness of objects in the optical remote sensing image using machine learning algorithm. *PLoS ONE* 9:e86528. doi: 10.1371/journal.pone.0086528

Union, I. T. (2002). *Methodology for the Subjective Assessment of the Quality of Television Pictures*. ITU-R Recommendation BT.

Wang, Z. (2011). Applications of objective image quality assessment methods [applications corner]. *IEEE Signal Process. Mag*. 28, 137–142. doi: 10.1109/MSP.2011.942295

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process*. 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). "Multiscale structural similarity for image quality assessment," in *Thrity-Seventh Asilomar Conference on Signals, Systems & Computers* (Pacific Grove, CA), 1398–1402. doi: 10.1109/ACSSC.2003.1292216

Wu, J., Lin, W., Shi, G., and Liu, A. (2013). Perceptual quality metric with internal generative mechanism. *IEEE Trans. Image Process*. 22, 43–54. doi: 10.1109/TIP.2012.2214048

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2014). Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Process*. 23, 684–695. doi: 10.1109/TIP.2013.2293423

Zhang, L., Shen, Y., and Li, H. (2014). VSI: a visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Process*. 23, 4270–4281. doi: 10.1109/TIP.2014.2346028

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: a feature similarity index for image quality assessment. *IEEE Trans. Image Process*. 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhang, Y., Tu, Y., and Wang, L. (2020). Effects of display area and corneal illuminance on oculomotor system based on eye-tracking data. *Displays* 63:101952. doi: 10.1016/j.displa.2020.101952

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 14 January 2022
doi: 10.3389/fnins.2021.767977

Cross-Domain Feature Similarity Guided Blind Image Quality Assessment

Chenxi Feng[1], *Long Ye*[2]* *and Qin Zhang*[2]

[1] Key Laboratory of Media Audio and Video, Ministry of Education, Communication University of China, Beijing, China, [2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

This work proposes an end-to-end cross-domain feature similarity guided deep neural network for perceptual quality assessment. Our proposed blind image quality assessment approach is based on the observation that features similarity across different domains (e.g., Semantic Recognition and Quality Prediction) is well correlated with the subjective quality annotations. Such phenomenon is validated by thoroughly analyze the intrinsic interaction between an object recognition task and a quality prediction task in terms of characteristics of the human visual system. Based on the observation, we designed an explicable and self-contained cross-domain feature similarity guided BIQA framework. Experimental results on both authentical and synthetic image quality databases demonstrate the superiority of our approach, as compared to the state-of-the-art models.

Keywords: cross-domain feature similarity, image quality assessment, deep learning, transfer learning, human visual system

OPEN ACCESS

Edited by:
Guangtao Zhai,
Shanghai Jiao Tong University, China

Reviewed by:
Xiongkuo Min,
University of Texas at Austin,
United States
Huiyu Duan,
Shanghai Jiao Tong University, China
Wei Sun,
Shanghai Jiao Tong University, China

****Correspondence:***
Long Ye
yelong@cuc.edu.cn

Specialty section:
This article was submitted to
Perception Science,
a section of the journal
Frontiers in Neuroscience

Received: *31 August 2021*
Accepted: *08 November 2021*
Published: *14 January 2022*

Citation:
Feng C, Ye L and Zhang Q (2022)
Cross-Domain Feature Similarity
Guided Blind Image Quality
Assessment.
Front. Neurosci. 15:767977.
doi: 10.3389/fnins.2021.767977

1. INTRODUCTION

Objective image quality assessment (IQA) aims to enable computer programs to predict the perceptual quality of images in a manner that is consistent with human observers, which has become a fundamental aspect of modern multimedia systems (Zhai and Min, 2020). Based on how much information the computer program could access from the pristine (or reference) image, objective IQA could be categorized into full-reference IQA (FR-IQA) (Wang et al., 2003, 2004; Sheikh and Bovik, 2006; Larson and Chandler, 2010a; Li et al., 2011; Zhang et al., 2011, 2014; Liu et al., 2012; Chang et al., 2013; Xue et al., 2013), reduced-reference IQA (RR-IQA) (Wang and Simoncelli, 2005; Wang and Bovik, 2011; Rehman and Wang, 2012), and no-reference (or blind) IQA (NR-IQA/BIQA) (Kim and Lee, 2016; Liu et al., 2017; Ma et al., 2017a; Lin and Wang, 2018; Pan et al., 2018; Talebi and Milanfar, 2018; Sun et al., 2021). The absence of reference information in most real-world multimedia systems calls for BIQA methods, which are more applicable but also more difficult.

Deep neural network (DNN) has significantly facilitated various image processing tasks (Fang et al., 2017; Park et al., 2018; Casser et al., 2019; Ghosal et al., 2019) in recent years due to its powerful capacity in feature abstraction and representation. It is also worth noting that the success of deep-learning techniques is derived from large amounts of training data, which is often leveraged to adjust the parameters in the DNN architecture to guarantee that both the accuracy and generalization ability are satisfying. Unfortunately, image quality assessment is typically a small-sample problem since the annotation of the ground-truth quality labels calls for time-consuming subjective image quality experiments (Zhang et al., 2018a). Inadequate quality annotations severely restrict the performance of DNN-based BIQA models in terms of both accuracy and generalization ability.

In order to address the problem caused by limited subjective labels, data augmentation is firstly employed to increase the training labels (e.g., Kang et al. (2014)) proposed to split the image with quality labels into multiple patches. and each of the patches is assigned with a quality score which is the same with the whole image. However, some distortion types are inhomogeneous, i.e., the perceptual quality of local patches might differ from the overall quality of the whole image. Therefore, transfer learning has gained more attention to relieve the small-sample problem (Li et al., 2016). Specifically, the BIQA framework is comprised of two stages: which are pre-training and fine-tuning. In the pre-training stage, the parameters in the DNN architecture are trained by other image processing tasks such as object recognition, whilst in the fine-tuning stage, images with subjective labels are employed as training samples. Such a transfer-learning scheme is feasible since the low-level feature extraction procedure across different image processing tasks are shared (Tan et al., 2018).

More recently, various sources of external knowledge are incorporated to learn a better feature representation for the BIQA issue. For example, hallucinated reference (Lin and Wang, 2018) is generated via a generative network and employed to guide the quality-aware feature extraction. The distortion identification is incorporated as the auxiliary sub-task in MEON model (Ma et al., 2017b), by which the distortion type information is transparent to the primary quality prediction task for better quality prediction. Visual saliency is employed in Yang et al. (2019) to weight the quality-aware features more reasonably. Semantic information is also employed for better understanding of the intrinsic mechanism of quality prediction, e.g., multi-layer semantic features are extracted and aggregated through several statistical structures in Casser et al. (2019). An effective hyper network is employed in Su et al. (2020) to generate customized weights from the semantic feature for quality prediction, i.e., the quality perception rule differs as the image content changes.

Unlike other studies, this paper employs the cross-domain feature similarity as an extra restraint for better quality-aware feature representation. Specifically, the transfer-learning based BIQA approach is pre-trained in one domain (say, object recognition in the semantic domain) and is fine-tuned in the perceptual quality domain with similar DNN architectures, we have observed that the cross-domain (Semantic vs. Quality) feature similarity would, in turn, contribute to the quality prediction task (as shown in **Figure 1**).

By thoroughly analyzing the intrinsic interaction between object recognition task and quality prediction task, we think the phenomenon represented in **Figure 1** is sensible. As shown in **Figure 2**, previous works (Larson and Chandler, 2010b) have revealed that human observers would take different strategies to assess the perceptual quality when viewing images with different amounts of degradation: when judging the quality of a distorted image containing near-threshold distortions, one tends to rely primarily on visual detection of any visible local differences, in such a scenario, semantic information is instructive for quality perception since distortion in the semantic-sensitive area would contribute more in the quality decision and vice versa. On the other hand, when judging the quality of a distorted image with clearly visible distortions, one would rely much less on visual detection and much more on the overall image appearance, in such a scenario, the quality decision procedure is much more independent with semantic information.

Considering the effectiveness of cross-domain feature similarity (CDFS), this work leverages CDFS as an extra restraint to improve the prediction accuracy of BIQA models. As shown in **Figure 3**, the parameters in our CDFSNet are updated according to both the basic loss and the extra loss, which would restrain the network yielding quality predictions as similar as the ground-truth label whilst maintaining that the CDFS also correlates well with the perceptual quality, in such a manner that, the accuracy of the DNN architecture would get improved according to the experimental results presented in section 3.

Compared to the aforementioned works, the superiority of the cross-domain feature similarity guided BIQA framework is embodied in the following aspects:

(1) The proposed cross-domain feature similarity is self-contained for transfer-learning based BIQA models since the transfer-learning procedure itself is comprised of the training in two different domains (i.e., object recognition and quality prediction). Therefore, no extra annotation procedure (such as distortion identification in Ma et al., 2017b and visual saliency in Yang et al., 2019) is needed.

(2) The proposed cross-domain feature similarity is more explicable since it is derived from the intrinsic characteristic of interactions between semantic recognition and quality perception.

(3) In addition to general-purpose IQA, the performance of our proposed CDFS guided BIQA framework is also evaluated on other specific scenarios such as screen content (Xiongkuo et al., 2021) and dehazing oriented (Min et al., 2018b, 2019) IQA. The experimental results indicate that CDFS guided BIQA has significant potential toward diverse types of BIQA tasks (Min et al., 2020a,b).

The rest part of the paper is organized as follows: Section 2 illustrates the details of our CDFS-based BIQA framework and section 3 shows the experimental results; Section 4 is the conclusion.

2. MATERIALS AND METHODS

2.1. Problem Formulation

Let x denote the input image, conventional DNN based BIQA works usually leverage an pre-trained DNN architecture $f(\cdot;\theta)$ (with learnable parameters θ) to predict the perceptual quality of x via $\hat{q} = f(x;\theta)$, where $\hat{q}$ denotes the prediction of perceptual quality q.

Our work advocates employing the cross-domain feature similarity to supervise the update of parameters in a quality prediction network. Specifically, let $f(\cdot;\theta_{Smtc})$ denotes the DNN with fixed and pre-trained parameters oriented toward semantic recognition, and $f(\cdot;\theta_{Qlty})$ denotes the DNN with learnable

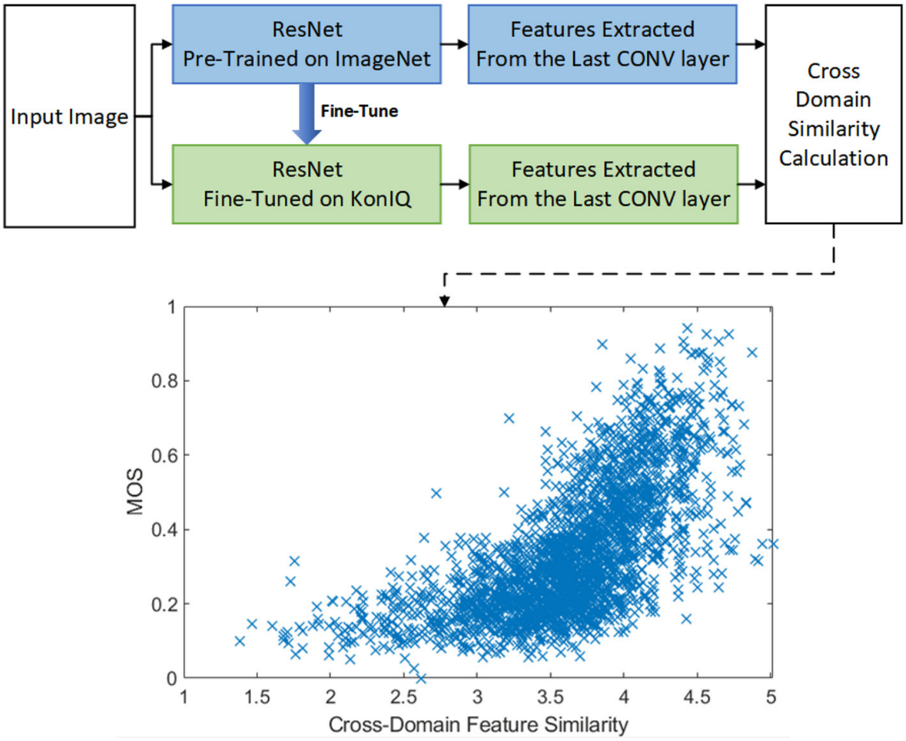

FIGURE 1 | The overall framework of our proposed CDFS guided BIQA approach. As shown in the lower part, the cross-domain feature similarity is highly correlated with the perceptual quality. The 'cross-domain similarity calculation' is obtained by: (1) Extractinged the features from the last convolutional layer of pre-trained ResNet (denoted as R_s) and fine-tuned ResNet (denoted as R_q); (2) Calculatinge the similarity matrix W according to Equation 1; (3) Obtaining the eigen values of W by $\vec{v} = eig(W)$; (4) The similarity Sim is calculated by $Sim = \frac{1}{std(\vec{v})}$, in which $std(\cdot)$ denotes the standard deviation operator.

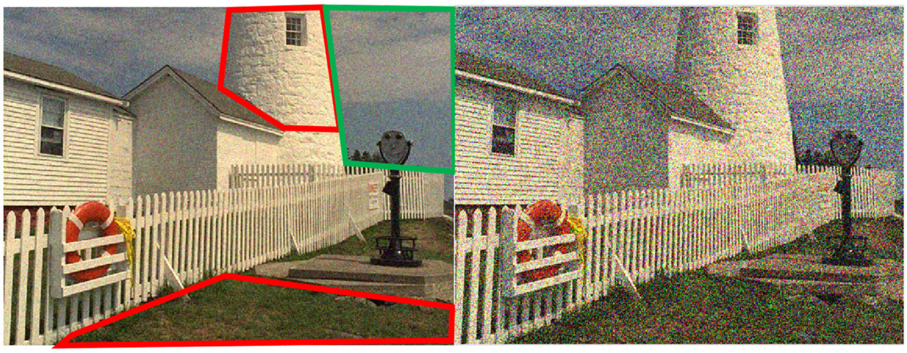

FIGURE 2 | Illustration of different strategies that the human visual system would take to assess the perceptual quality when viewing images with different amounts of degradation. Specifically, when judging the quality of a distorted image containing near-threshold distortions (Left), one tends to rely primarily on visual detection of any visible local differences, e.g., the distortions in red boxed are slighter than that in the green box even though the noise intensity is the same. On the other hand, when judging the quality of a distorted image with clearly visible distortions (Right), one would rely much less on visual detection and much more on overall image appearance e.g., the distortions in each image area are roughly the same.

parameters oriented toward quality prediction. It should be noticed that $f(\cdot; \theta_{Smtc})$ and $f(\cdot; \theta_{Qlty})$ share the same architectures whilst having own different parameters. This work attempts to further improve the quality prediction accuracy by analyzing the similarity between the features extracted for different tasks, i.e., features extracted for semantic recognition $ft_s = f(x; \theta_{Smtc})$, and features extracted for quality regression $ft_q = f(x; \theta_{Qlty})$.

Given three-dimensional features ft_s and ft_q with size $[C, H, W]$, where C, H, W denotes the channel size, height, and width of the features, respectively, ft_s and ft_q are firstly reshaped

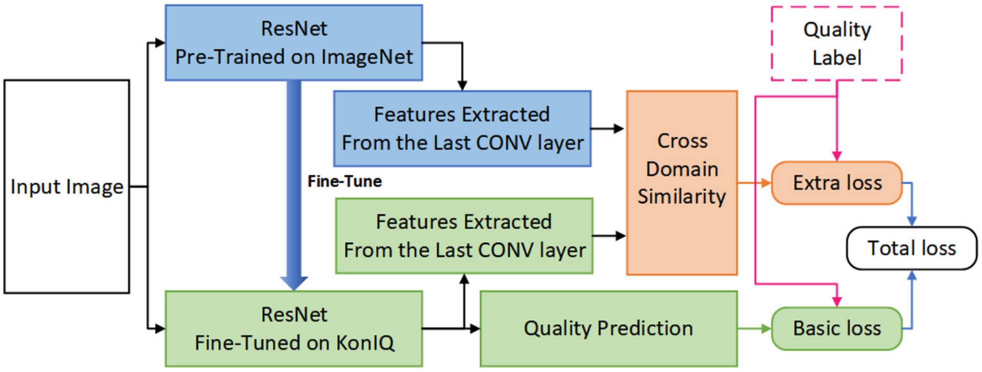

FIGURE 3 | The overall pipeline of our proposed CDFS-based IQA approach.

into R_q and R_s with size $[C, H \times W]$. The similarity Sim between R_q and R_s is obtained via the following steps.

Step 1, employ linear regression to express R_q via R_s, i.e., $R_q = W \times R_s + e$, where W denotes the weighting matrix and e denotes the prediction error of linear regression. Therefore, W could be obtained by

$$W = (R_s^T \times R_s)^{-1} \times R_s^T \times R_q \tag{1}$$

Step 2, a learnable DNN architecture $g(\cdot;\gamma)$ is employed to yield the similarity between ft_s and ft_q given W, i.e., $Sim = g(W;\gamma)$

2.2. Network Design

The architecture of our proposed network is shown in **Figure 4**, which mainly consists of a semantically oriented feature extractor, perceptual-quality oriented feature extractor, and cross-domain feature similarity predictor. More details are described as follows.

2.2.1. Semantic Oriented Feature Extractor

The DNN pre-trained in large-scale object recognition datasets (e.g., ImageNet Deng et al., 2009) are leveraged as the semantic oriented feature extractor.

Specifically, this work employs the activations of the last convolutional layers in ResNet50 to represent the semantic-aware features ft_s of an specific image, i.e., $ft_s = f(x;\theta_{Smtc})$.

It is worth noting that θ_{Smtc} is fixed during the training stage since the proposed DNN framework will be fine-tuned in IQA datasets in which the semantic label is unavailable.

2.2.2. Perceptual-Quality Oriented Feature Extractor

The architecture of perceptual-quality oriented feature extractor $f(\cdot;\theta_{Qlty})$ is quite similar with semantic oriented feature extractor. However, the parameters θ_{Qlty} in $f(\cdot;\theta_{Qlty})$ are learnable and independent with θ_{Smtc}.

The quality-aware features $ft_q = f(x;\theta_{Qlty})$ are further leveraged to aggregate the prediction of subjective quality score, i.e., $\hat{q} = h(ft_q;\delta)$, in which q denotes the subjective quality score (MOS), $\hat{q}$ is the prediction of q, and $h(\cdot;\delta)$ stands for the MOS prediction network given quality-aware features with learnable parameters δ.

2.2.3. Cross-Domain Feature Similarity Predictor

As illustrated in section 1, the cross-domain feature similarity would contribute to the prediction of perceptual quality. However, directly evaluating the similarity between ft_s and ft_q via Minkowski-Distance or Wang-Bovik metric (Wang et al., 2004) is not as efficient, as shown in **Figure 6**. We think the invalidation of the Wang-Bovik metric is mainly attributed to its pixel-wise sensitivity, i.e., any turbulence during the parameter initializing and updating of the DNN framework would result in a significant difference between ft_s and ft_q.

To this end, this work proposes to depict the cross-domain feature similarity through a global perspective. Specifically, the similarity is derived from the weighting matrix W which is employed to reconstruct ft_q given ft_s via linear regression. Since the W is derived from the features amongst all channels, it is less likely to suffer from the instability of the DNN during initializing and updating. The experiments reported in section 3.3 also demonstrate the superiority of our proposed similarity measurement for cross-domain features. In our CDFS-guided BIQA framework, the CDFS is incorporated as follows:

Linear regression is employed for the reconstruction and the weighting matrix W could be obtained according to equation 1 and ***Step 1*** in section 2.1

A stack of convolutional layers (denoted as g($\cdot$;γ)) is followed to learn the cross-domain feature similarity given W.

During the training stage, the cross-domain similarity is employed as a regularization item to supervise the quality prediction network.

2.2.4. Loss Function

The loss function L of our proposed network is designed as

$$L_1 = argmin_{[\theta_{Qlty},\delta]} \left\| q - h(f(x;\theta_{Qlty});\delta) \right\| \tag{2}$$

$$L_2 = argmin_{[\theta_{Qlty},\gamma]} \left\| q - g(W;\gamma) \right\| \tag{3}$$

and

$$L = L_1 + \lambda L_2 \tag{4}$$

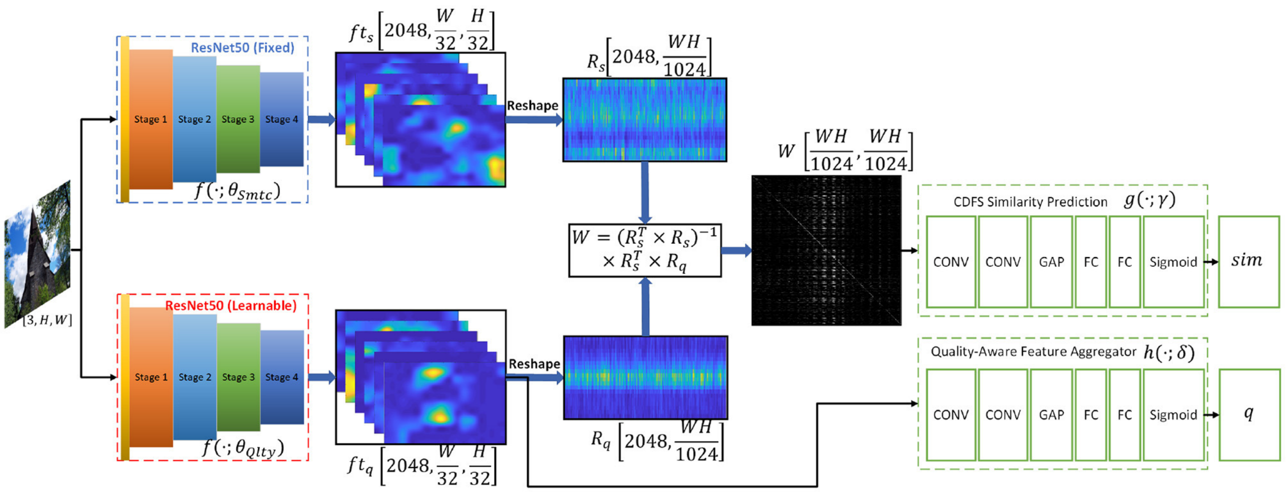
$ft_s\left[2048, \frac{W}{32}, \frac{H}{32}\right]$
ResNet50 (Fixed)
Stage 1
Stage 2
Stage 3
Stage 4
$f(\cdot;\theta_{Smtc})$
Reshape
$R_s\left[2048, \frac{WH}{1024}\right]$
$W\left[\frac{WH}{1024}, \frac{WH}{1024}\right]$
$W = (R_s^T \times R_s)^{-1} \times R_s^T \times R_q$
CDFS Similarity Prediction $g(\cdot;\gamma)$
CONV
CONV
GAP
FC
FC
Sigmoid
sim
$[3,H,W]$
ResNet50 (Learnable)
Stage 1
Stage 2
Stage 3
Stage 4
$f(\cdot;\theta_{Qlty})$
Reshape
$R_q\left[2048, \frac{WH}{1024}\right]$
$ft_q\left[2048, \frac{W}{32}, \frac{H}{32}\right]$
Quality-Aware Feature Aggregator $h(\cdot;\delta)$
CONV
CONV
GAP
FC
FC
Sigmoid
q

FIGURE 4 | The detailed architecture of our proposed Cross-Domain Feature Similarity Guided Network. The "CONV" denotes convolutional layers followed by batch normalization and ReLU layer, the "FC" denotes the fully-connected layer, and the 'GAP' denotes the global average pooling layer.

where $\|\cdot\|$ denotes the $L1$ norm operator, W is calculated according to equation 1, and λ is a hyper parameter controlling the weights of L_1 and L_2.

2.3. Implementation Details

We use ResNet50 (He et al., 2016) as the backbone model for both the semantically oriented feature extractor and the perceptual-quality oriented feature extractor. As aforementioned, the pre-trained model on ImageNet (Deng et al., 2009) is used for network initialization. During the training stage, the θ_{Smtc} is fixed whilst θ_{Qlty} is learnable. In our network, the last two layers of the origin ResNet50, i.e., an average pooling layer and a fully connected layer, are removed to output features ft_s and ft_q.

For quality regression, a global average pooling (GAP) layer is used to pool the features ft_q into one-dimensional vectors, then three fully -connected (FC) layers are followed with size 2048-1024-512-1 and activated by ReLu, except for the last layer (activated by sigmoid).

The $g(\cdot;\gamma)$ in cross-domain feature similarity predictor is implemented by 3 three stacked convolutional layers, a GAP layer, and three FC layers. The architectures of convolutional layers are $in(1) - out(32) - k(1) - p(0)$, $in(32) - out(64) - k(3) - p(1)$, and $in(64) - out(128) - k(3) - p(1)$, respectively, where $in(\alpha) - out(\beta) - k(x) - p(y)$ denotes the input channel size and output channel size is α and β, the kernel size is x, and the padding size is y. Each of the convolutional layers is followed by a batch normalization layer and a ReLu layer. The GAP layer and the FC layers are the same with quality regression except that the size of FC layers is 128-512-512-1.

The experiment is conducted on Tesla V100P GPUs, while the DNN modules are implemented by Pytorch. The size of minibatch is 24. Adam (Kingma and Ba, 2014) is adopted to optimize the loss function with weight decay 5×10^{-4} and learning rate 1×10^{-5} for parameters in baseline (ResNet) and 1×10^{-4} for other learnable parameters. As mentioned, the parameters in semantic oriented feature extractor is are fixed, i.e., the learning rate is 0 for θ_{Smtc}.

3. EXPERIMENTAL RESULTS

3.1. Datasets and Evaluation Metrics

Three image databases including KonIQ-10k (Hosu et al., 2020), LIVE Challenges (LIVEC) (Ghadiyaram and Bovik, 2015), and TID2013 (Ponomarenko et al., 2015) are employed to validate the performance of our proposed network. The KonIQ-10k and LIVEC are authentically distorted image databases containing 10,073 and 1,162 distorted images, respectively, and the TID2013 is a synthetic image database containing 3,000 distorted images.

Two commonly used criteria, Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC), are adopted to measure the prediction monotonicity and the prediction accuracy. For each database, 80% images are used for training, and the others are used for testing. The synthetic image database is split according to reference images. All the experiments are under five times random train-test splitting operation, and the median SRCC and PLCC values are reported as final statistics.

3.2. Comparison With the State-of-the-Art Methods

Ten BIQA methods are selected for performance comparison, including five hand-crafted based (BRISQUE Mittal et al., 2012, ILNIQE Xu et al., 2016, HOSA Zhang et al., 2015, BPRI Min et al., 2017a, BMPRI Min et al., 2018a) and five DNN-based approaches (SFA Li et al., 2018, DBCNN Zhang et al., 2018b, HyperIQA Su et al., 2020, SDGNet Yang et al., 2019). The experimental results are shown as in **Table 1**.

As shown in **Table 1**, our method outperforms all the SOTA methods on the two authentic image databases in terms of SRCC. As for PLCC measurement, our method achieves

TABLE 1 | Performance comparison in terms of PCLL and SRCC on KonIQ, LIVEC, and TID2013, respectively.

SRCC	**KonIQ**	**LIVEC**	**TID2013**
BRISQUE	0.665	0.608	0.572
ILNIQE	0.507	0.432	0.521
HOSA	0.671	0.640	0.688
BPRI	–	–	0.899
BMPRI	–	–	**0.929**
SFA	0.856	0.812	–
DBCNN	0.875	0.851	–
HyperIQA	0.906	0.859	–
SGDNet	0.903	0.851	0.843
DeepFL	0.877	0.734	0.858
ours	**0.918**	**0.865**	0.899
PLCC	**KonIQ**	**LIVEC**	**TID2013**
BRISQUE	0.681	0.645	0.651
ILNIQE	0.523	0.508	0.648
HOSA	0.694	0.678	0.764
BPRI	–	–	0.892
BMPRI	–	–	**0.947**
SFA	0.872	0.833	–
DBCNN	0.884	0.869	–
HyperIQA	0.917	**0.882**	–
SGDNet	0.920	0.872	0.861
DeepFL	0.887	0.769	0.876
ours	**0.928**	0.875	0.880

Values in bold represents the highest value.

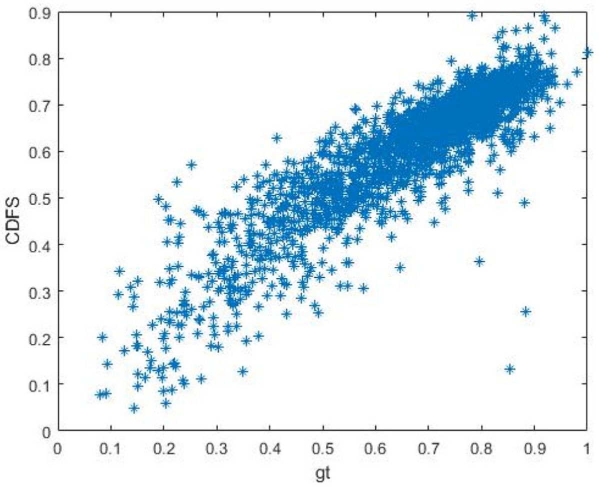

FIGURE 5 | The scatter plot of CDFS vs. MOS on KonIQ.

TABLE 2 | Ablation results in terms of SRCC and PLCC on KonIQ.

Modules	**BaseLine**	**+SP_wang**	**+SP_W**
SRCC	0.842	0.895	0.918
Gain(%)	–	6.3	9.0
PLCC	0.849	0.913	0.928
Gain(%)	–	7.5	9.3

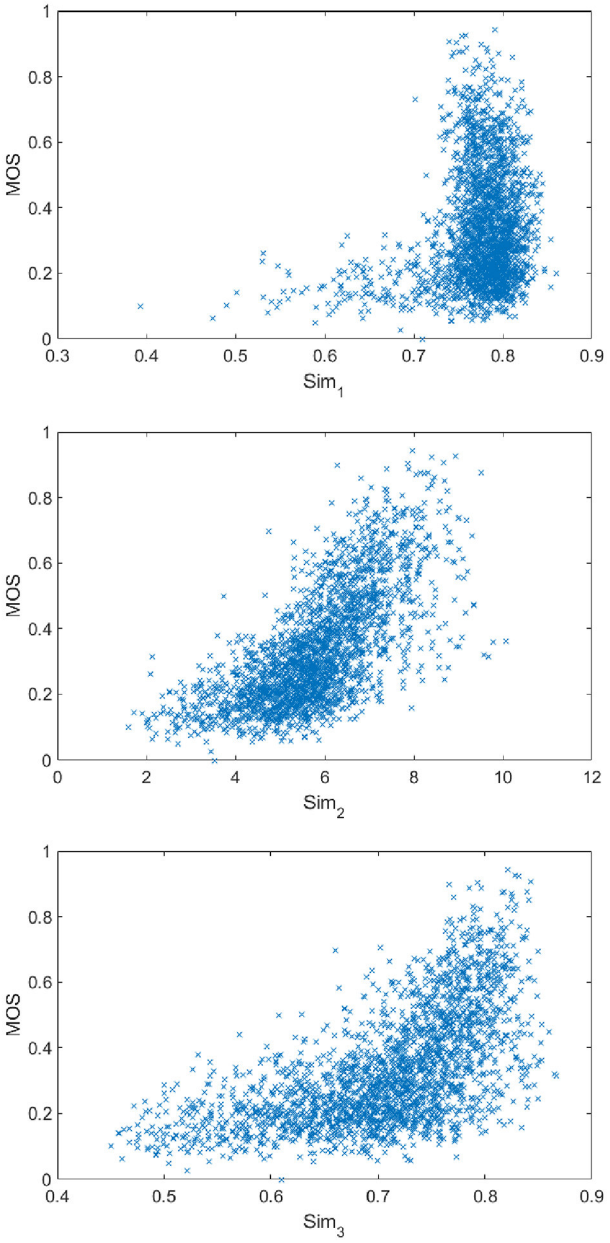

FIGURE 6 | The scatter plot of Sim_1, Sim_2, and Sim_3 vs. MOS on KonIQ.

the best performance on KonIQ and competing (the second) performance on LIVEC. This suggests that calculating cross-domain feature similarity for quality prediction refinement is effective. Though we do not especially modify the networks for synthetic image feature extraction, the proposed network has achieved competing performance in TID2013. Specifically, the proposed approach achieves the second-highest performance in terms of SRCC and the third-highest performance in terms of PLCC on TID2013.

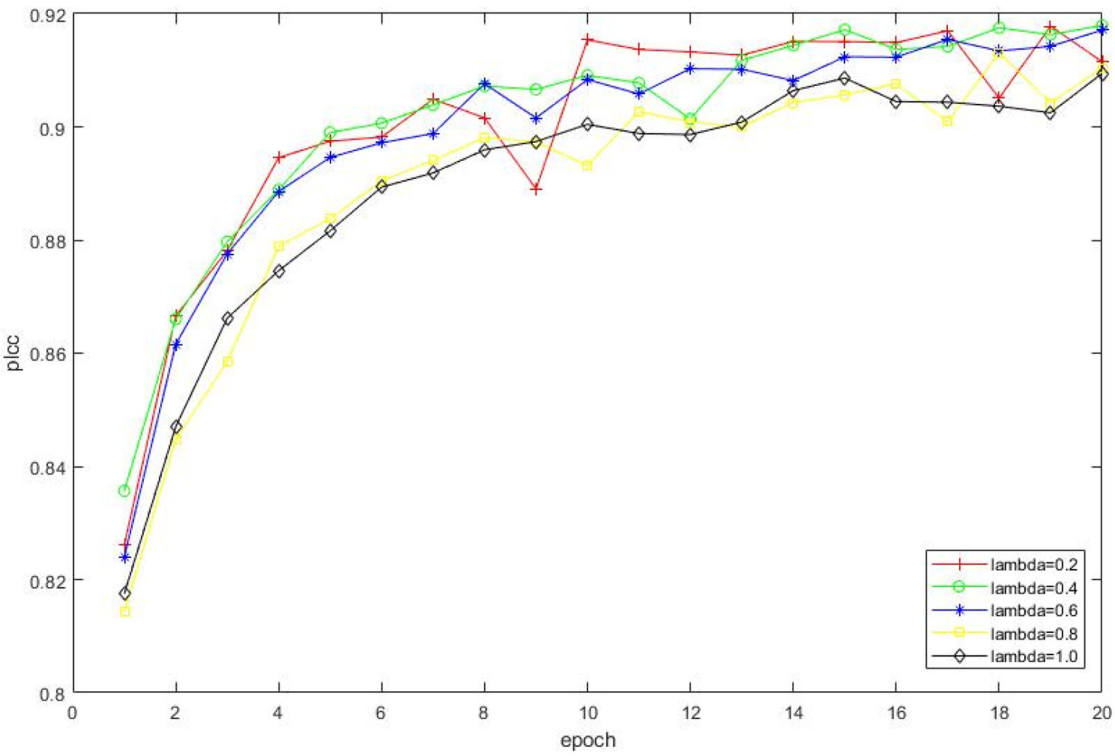

FIGURE 7 | Impact on selections of different λ. The experimental result is conducted on KonIQ, and a total of 20 epochs are involved.

3.3. Cross-Domain Feature Similarity Visualization

In order to further illustrate the superiority of our proposed CDFS, we firstly present the scatter plot of CDFS vs. MOS on KonIQ in **Figure 5**, indicating the CDFS is well correlated with perceptual quality.

In addition, we also investigate several non-learnable approaches for calculating CDFS: (1) $Sim_1 = mean(\frac{2 \times ft_s \times ft_q + C}{ft_s^2 + ft_q^2 + C})$, where C denotes the constant to avoid numerical singularity; and (2) $Sim_2 = std(eig(W))$; (3) $Sim_3 = mean(\frac{2 \times \vec{v} \times \vec{1} + C}{\vec{v}^2 + \vec{1}^2 + C})$, and $\vec{v} = eig(W)$, in which $\vec{1}$ denotes the vectors with the same size as $\vec{v}$ whilst whose elements are all 1.

Therefore, the calculation of Sim_1 is directly comparing the difference between ft_s and ft_q, and the calculation of Sim_2 and Sim_3 is based on the W derived according to equation 1. As shown in **Figure 6**, Sim_2 and Sim_3 is more correlated with the subjective score, demonstrating that measuring the cross-domain feature similarity based on W is more effective.

3.4. Ablation Study

Ablation study is conducted on KonIQ-10k to validate the efficiency of our proposed components, including the ResNet50 backbone (BaseLine), the similarity predictor (SP) obtained by Wang-Bovik metric (SP_wang, similar as Sim_1 in section 3.3), and the similarity predictor derived from the weighting metric W (SP_W). The results are shown in **Table 2**, indicating that incorporating a cross-domain similarity predictor could significantly improve the accuracy of quality prediction. Our proposed similarity measurement has achieved a great PLCC improvement (1.8%) compared to SP_wang and a more significant SRCC improvement (2.7%).

The impact of λ in equation 4 is also investigated, i.e., we set $\lambda = [0.2, 0.4, 0.6, 0.8, 1.0]$, respectively and observe the corresponding performance as shown in **Figure 7**. Therefore, we select $\lambda = 0.4$ for performance comparison and the following experiments.

3.5. Cross-Database Validation

In order to test the generalization ability of our network, we train the model on the entire KonIQ-10k and test on the entire LIVEC. The four most competing IQA models in terms of generalization ability are involved in the comparison, which are PQR (Zeng et al., 2017), DBCNN, HyperIQA, and DeepFL. The validation results are shown in **Table 3**, indicating the generalization ability of our approach is higher than existing SOTA methods for assessing authentically distorted images.

However, if the network is trained on KonIQ-10k and directly applied for a synthetic image database, its generalization ability is not satisfactory, and the SRCC on TID2013 is only 0.577. That is mainly because the distortion mechanisms between synthetic and authentically distorted image databases are widely different. Training the network solely on authentically

TABLE 3 | Cross data base validation (Trained on KonIQ-10k and Tested on LIVEC).

Modules	DeepFL	DBCNN	HyperIQA	PQR	Ours
SRCC	0.704	0.755	0.770	0.785	0.817
Gain(%)	–	7.2	9.4	11.5	16.1

TABLE 4 | SRCC and PLCC performance on CCT, DHQ, and SHRQ.

		SRCC	PLCC
CCT	20-%Test	0.9655	0.9672
	100-%Test	0.5758	0.6193
DHQ	20-%Test	0.9533	0.9223
	100-%Test	0.6819	0.6678
SHRQ	20-%Test	0.8875	0.9082
	100-%Test	0.4233	0.4761

-distorted image databases could not learn the specific synthetic distortion patterns such as JPEG compression, transmission errors, or degradation caused by denoising, *etc.*

3.6. Further Validation on Other Specific IQA Tasks

In order to further validate the robustness of our BIQA framework toward other specific IQA tasks, the performance of CDFS guided BIQA network is evaluated on CCT (Min et al., 2017b), DHQ (Min et al., 2018b), and SHRQ (Min et al., 2019). The CCT contains 1,320 distorted images with various types of images including natural scene images (NSI), computer graphic images (CGI), and screen content images (SCI); The DHQ contains 1,750 dehazed images generated from 250 real hazy images.; The SHRQ database consists of two subsets, namely: regular and aerial image subsets, which include 360 and 240 dehazed images created from 45 and 30 synthetic hazy images using 8 eight image dehazing algorithms, respectively.

The training pipeline is similar with section 3.1, i.e., 80% of the CCT, DHQ, or SHRQ are involved as the training set and the other 20% is the testing set. Considering that the scale of the subset is not adequate for the training of DNN, we merge the subsets in each datasets. For example, the NSI, CGI, and SCI are merged as the training set of CCT.

As shown in **Table 4**, the predictions of our CDFS guided BIQA framework shows significant consistency with subjective scores, indicating that our proposed BIQA approach is feasible to be generalized into other types of IQA tasks.

Furthermore, if the network is trained on KonIQ-10k and directly applied on CCT, DHQ, and SHRQ, the accuracy is not satisfactory, as shown in **Table 4**. Such phenomenon is similar to the cross-database validation results discussed in section 3.5, indicating that training the network solely on authentically-distorted natural image databases could not sufficiently learn the quality-aware features for CGI, SCI, etc.

4. CONCLUSION

This work aims to evaluate the perceptual quality based on cross-domain feature similarity. The experimental results on KonIQ, LIVEC, and TID2013 demonstrate the superiority of our proposed methods.

We would further investigate such CDFS-incorporated BIQA framework in the following aspects: (1) investigating more efficient approaches of CDFS measurement; (2) investigating more types of DNN baselines in addition to ResNet.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

CF established the BIQA framework and adjusted the architecture for better performance. LY and CF conducted the experiments and wrote the manuscripts. QZ designed the original method, and provided resource support (e.g., GPUs) for this manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work is supported by the National Key R&D Program of China under Grant No. 2021YFF0900503, the National Natural Science Foundation of China under Grant Nos. 61971383 and 61631016, and the Fundamental Research Funds for the Central Universities.

ACKNOWLEDGMENTS

We would like to thank Li Fang, Wei Zhong, and Fei Hu for some swell ideas.

REFERENCES

Casser, V., Pirk, S., Mahjourian, R., and Angelova, A. (2019). Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos. *Proc. AAAI Conf. Artif. Intell.* 33, 8001–8008. doi: 10.1609/aaai.v33i01.33018001

Chang, H.-W., Yang, H., Gan, Y., and Wang, M.-H. (2013). Sparse feature fidelity for perceptual image quality assessment. *IEEE Trans. Image Proc.* 22, 4007–4018. doi: 10.1109/TIP.2013.2266579

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255.

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). "RMPE: regional multi-person pose estimation," IN *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2334–2343.

Ghadiyaram, D., and Bovik, A. C. (2015). Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Proc.* 25, 372–387. doi: 10.1109/TIP.2015.2500021

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). Dialoguegcn: a graph convolutional neural network for emotion recognition in conversation. *arXiv preprint* arXiv:1908.11540. doi: 10.18653/v1/D19-1015

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

Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. (2020). Koniq-10k: an ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Proc.* 29, 4041–4056. doi: 10.1109/TIP.2020.2967829

Kang, L., Ye, P., Li, Y., and Doermann, D. (2014). "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 1733–1740.

Kim, J., and Lee, S. (2016). Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.* 11, 206–220. doi: 10.1109/JSTSP.2016.2639328

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint* arXiv:1412.6980.

Larson, E. C., and Chandler, D. M. (2010a). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19, 011006.

Larson, E. C., and Chandler, D. M. (2010b). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* 19, 011006. doi: 10.1117/1.3267105

Li, D., Jiang, T., Lin, W., and Jiang, M. (2018). Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Trans. Multimedia* 21, 1221–1234. doi: 10.1109/TMM.2018.2875354

Li, S., Zhang, F., Ma, L., and Ngan, K. N. (2011). Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Trans. Multimedia* 13, 935–949. doi: 10.1109/TMM.2011.2152382

Li, Y., Po, L.-M., Feng, L., and Yuan, F. (2016). "No-reference image quality assessment with deep convolutional neural networks," in *2016 IEEE International Conference on Digital Signal Processing (DSP)* (Beijing: IEEE), 685–689.

Lin, K.-Y., and Wang, G. (2018). "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City: IEEE), 732–741.

Liu, T.-J., Lin, W., and Kuo, C.-C. J. (2012). Image quality assessment using multi-method fusion. *IEEE Trans. Image Proc.* 22, 1793–1807. doi: 10.1109/TIP.2012.2236343

Liu, X., Van De Weijer, J., and Bagdanov, A. D. (2017). "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 1040–1049.

Ma, K., Liu, W., Liu, T., Wang, Z., and Tao, D. (2017a). dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans. Image Proc.* 26, 3951–3964. doi: 10.1109/TIP.2017.2708503

Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. (2017b). End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Proc.* 27, 1202–1213. doi: 10.1109/TIP.2017.2774045

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., and Chen, C. W. (2017a). Blind quality assessment based on pseudo-reference image. *IEEE Trans. Multimedia* 20, 2049–2062. doi: 10.1109/TMM.2017.2788206

Min, X., Ma, K., Gu, K., Zhai, G., Wang, Z., and Lin, W. (2017b). Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Trans. Image Proc.* 26, 5462–5474. doi: 10.1109/TIP.2017.2735192

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018a). Blind image quality estimation via distortion aggravation. *IEEE Trans. Broadcast.* 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhai, G., Gu, K., Yang, X., and Guan, X. (2018b). Objective quality evaluation of dehazed images. *IEEE Trans. Intell. Transport. Syst.* 20, 2879–2892. doi: 10.1109/TITS.2018.2868771

Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., et al. (2019). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Trans. Multimedia* 21, 2319–2333. doi: 10.1109/TMM.2019.2902097

Min, X., Zhai, G., Zhou, J., Farias, M. C., and Bovik, A. C. (2020a). Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Proc.* 29, 6054–6068. doi: 10.1109/TIP.2020.2988148

Min, X., Zhou, J., Zhai, G., Le Callet, P., Yang, X., and Guan, X. (2020b). A metric for light field reconstruction, compression, and display quality evaluation. *IEEE Trans. Image Proc.* 29:3790–3804. doi: 10.1109/TIP.2020.2966081

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Proc.* 21, 4695–4708. doi: 10.1109/TIP.2012.2214050

Pan, D., Shi, P., Hou, M., Ying, Z., Fu, S., and Zhang, Y. (2018). "Blind predicting similar quality map for image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6373–6382.

Park, S.-J., Son, H., Cho, S., Hong, K.-S., and Lee, S. (2018). "Srfeat: single image super-resolution with feature discrimination," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 439–455.

Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal Proc. Image Commun.* 30, 57–77. doi: 10.1016/j.image.2014.10.009

Rehman, A., and Wang, Z. (2012). Reduced-reference image quality assessment by structural similarity estimation. *IEEE Trans. Image Proc.* 21, 3378–3389. doi: 10.1109/TIP.2012.2197011

Sheikh, H. R., and Bovik, A. C. (2006). Image information and visual quality. *IEEE Trans. Image Proc.* 15, 430–444. doi: 10.1109/TIP.2005.859378

Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., et al. (2020). "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 3667–3676.

Sun, W., Min, X., Zhai, G., and Ma, S. (2021). Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *arXiv preprint* arXiv:2105.14550.

Talebi, H., and Milanfar, P. (2018). Nima: Neural image assessment. *IEEE Trans. Image Proc.* 27, 3998–4011. doi: 10.1109/TIP.2018.2831899

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks* (Rhodes: Springer), 270–279.

Wang, Z., and Bovik, A. C. (2011). Reduced-and no-reference image quality assessment. *IEEE Signal Process Mag.* 28, 29–40. doi: 10.1109/MSP.2011.942471

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., and Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. *Hum. Vision Electron. Imaging* 5666, 149–159. doi: 10.1117/12.597306

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2* (Pacific Grove, CA: IEEE), 1398–1402.

Xiongkuo, M., Ke, G., Guangtao, Z., Xiaokang, Y., Wenjun, Z., Callet, P. L., et al. (2021). *Screen Content Quality Assessment: Overview, Benchmark, and Beyond*. ACM Computing Surveys.

Xu, J., Ye, P., Li, Q., Du, H., Liu, Y., and Doermann, D. (2016). Blind image quality assessment based on high order statistics aggregation. *IEEE Trans. Image Proc.* 25, 4444–4457. doi: 10.1109/TIP.2016.2585880

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2013). Gradient magnitude similarity deviation: a highly efficient perceptual image quality index. *IEEE Trans. Image Proc.* 23, 684–695. doi: 10.1109/TIP.2013.2293423

Yang, S., Jiang, Q., Lin, W., and Wang, Y. (2019). "Sgdnet: an end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proceedings of the 27th ACM International Conference on Multimedia* (Nice), 1383–1391.

Zeng, H., Zhang, L., and Bovik, A. C. (2017). A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint* arXiv:1708.08190.

Zhai, G., and Min, X. (2020). Perceptual image quality assessment: a survey. *Sci. China Inf. Sci*. 63, 211301. doi: 10.1007/s11432-019-2757-1

Zhang, L., Shen, Y., and Li, H. (2014). Vsi: a visual saliency-induced index for perceptual image quality assessment. *IEEE Trans. Image Proc*. 23, 4270–4281. doi: 10.1109/TIP.2014.2346028

Zhang, L., Zhang, L., and Bovik, A. C. (2015). A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Proc*. 24, 2579–2591. doi: 10.1109/TIP.2015.2426416

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). Fsim: a feature similarity index for image quality assessment. *IEEE Trans. Image Proc*. 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

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

Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. (2018b). Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circ. Syst. Video Technol*. 30, 36–47. doi: 10.1109/TCSVT.2018.2886771

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Computational Neuroscience

ORIGINAL RESEARCH
published: 20 January 2022
doi: 10.3389/fncom.2021.768021

Image Quality Evaluation of Light Field Image Based on Macro-Pixels and Focus Stack

Chunli Meng, Ping An, Xinpeng Huang, Chao Yang and Yilei Chen*

Key Laboratory for Advanced Display and System Application, Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai, China

Due to the complex angular-spatial structure, light field (LF) image processing faces more opportunities and challenges than ordinary image processing. The angular-spatial structure loss of LF images can be reflected from their various representations. The angular and spatial information penetrate each other, so it is necessary to extract appropriate features to analyze the angular-spatial structure loss of distorted LF images. In this paper, a LF image quality evaluation model, namely MPFS, is proposed based on the prediction of global angular-spatial distortion of macro-pixels and the evaluation of local angular-spatial quality of the focus stack. Specifically, the angular distortion of the LF image is first evaluated through the luminance and chrominance of macro-pixels. Then, we use the saliency of spatial texture structure to pool an array of predicted values of angular distortion to obtain the predicted value of global distortion. Secondly, the local angular-spatial quality of the LF image is analyzed through the principal components of the focus stack. The focalizing structure damage caused by the angular-spatial distortion is calculated using the features of corner and texture structures. Finally, the global and local angular-spatial quality evaluation models are combined to realize the evaluation of the overall quality of the LF image. Extensive comparative experiments show that the proposed method has high efficiency and precision.

Keywords: light field, objective image quality assessment, focus stack, macro-pixels, corner

Edited by:
Guangtao Zhai,
Shanghai Jiao Tong University, China

Reviewed by:
Yucheng Zhu,
Shanghai Jiao Tong University, China
Huiyu Duan,
Shanghai Jiao Tong University, China
Xiongkuo Min,
University of Texas at Austin,
United States

****Correspondence:***
Ping An
anping@shu.edu.cn

Received: *31 August 2021*
Accepted: *15 December 2021*
Published: *20 January 2022*

Citation:
Meng C, An P, Huang X, Yang C and Chen Y (2022) Image Quality Evaluation of Light Field Image Based on Macro-Pixels and Focus Stack. Front. Comput. Neurosci. 15:768021. doi: 10.3389/fncom.2021.768021

INTRODUCTION

Light field (LF) imaging technology is designed to record rich scenario information. Compared with ordinary two-dimensional (2D) images and binocular stereoscopic images, LF images are favored in researches like immersive stereoscopic display and object recognition because of their particular characteristics of dense view and post-focusing (Huang et al., 2016; Ren et al., 2017a). For these applications, image quality degradation will directly affect the perception of the immersive experience and the accuracy of object recognition. However, the quality assessment of LF images is different from that of ordinary image types. It involves analyzing the complex imaging structure relationships among dense multi-view LF images. Therefore, it is beneficial to consider the characteristics of LF images, such as the relationship between dense viewpoints, perception of human eyes to the structure of multi-view images, to accurately evaluate the quality. Traditional image quality evaluation models are not suitable for LF because they do not consider the special characteristics of LF images. It is of great significance for the development of LF to build an objective quality evaluation model that effectively utilizes the characteristics of LF images.

The characteristics of LF images are reflected in its various expressions. The dense viewpoints of an LF image, hereinafter referred to as subaperture images (SAIs), represent spatial information of the captured scenes from different visual angles. Adjacent SAIs have strong texture similarity, which enables the compression operation to be better realized. Compression algorithms of LF images can alleviate the problem of inconvenience in transmission caused by a large amount of data of LF images. Furthermore, the reconstruction algorithms play an excellent role in recovering the loss of spatial resolution or angular resolution in the LF image processing. The compression and reconstruction algorithms are mainly based on the multiple representations of LF images: hexagonal lenslet image, rectangular decoded image, SAIs, focus stack, and epipolar plane images (EPIs) (Huang et al., 2019a; Wu et al., 2019). All of the above representations can reflect the angular and spatial characteristics of LF images. Although both compression and reconstruction operations promote the practical application of LF images, they inevitably bring the problem of quality degradation. Moreover, the performance of these algorithms varies a lot, so the criteria to check out the optimal one are necessary.

For situations where SAIs are used to evaluate the quality of LF images, Tian et al. (2018) presented a multi-order derivative feature-based model using the multi-order derivative features extracted on the SAIs of LF images. However, their analysis remains in the texture aspect of spatial information, lacking the analysis of the connection between the angular and spatial information. As an LF image can be regarded as a low-rank 4D tensor, Shi et al. (2019) adopted the tensor structure of the cyclopean image array from the LF to explore the angular-spatial characteristic. Zhou et al. (2020) used tensor decomposition of view stack in four directions to extract the spatial-angular features. To explore the angular-spatial characteristics of LF images, Min et al. (2020) averaged the structural matching degree of all viewpoints to compute the spatial quality and analyzed the amplitude spectrum of near-edge mean square error along viewpoints to express the angular quality. Xiang et al. (2020) computed the mean difference image from SAIs to describe the depth and structural information of LF images, and it used a curvelet transform to reflect the multi-channel characteristics of the human visual system.

The focus stack is constructed by stacking the refocused images from the perspective of depth, which reflects both the texture and depth information of LF images. Meng et al. (2019) compared different objective metrics under SAIs and the focus stack, which verified the superiority of the refocus characteristic of LF images. Meng et al. (2019) utilized the LF angular-spatial and human visual characteristics and verified the effectiveness of the assumed optimal parallax range. Meng et al. (2021) built a key refocused image extraction framework based on the maximal spatial information contrast and the minimal angular information variation to reduce the redundancy of quality evaluation in the focus stack.

The depth feature makes the LF more popular in object detection, three-dimensional reconstruction, and other applications. Paudyal et al. (2019) compared different depth extraction strategies and assessed the quality of LF through the structural similarity of the depth map. It is proven that the depth information is effective in reflecting the distortion degree of LF images, but Paudyal et al. (2019) ignored the texture structure information of LF images. Therefore, some studies have attempted to combine depth features with the features from SAIs to achieve better prediction results. Shan et al. (2019) combined the ordinary 2D features of SAIs and sparse gradient dictionary of LF depth map. Tian et al. (2020) performed radial symmetric transformation on the luminance components of all dense viewpoints to extract symmetric features and used depth maps to measure the structural consistency between viewpoints, which explored the way humans perceive structures and geometries.

To preferably explore the angular-spatial characteristics of LF, many pieces of research are devoted to take advantage of various LF expressions. For the form of uniting multiple representations, Luo et al. (2019) used the global entropy and uniform local binary pattern features of a lenslet image to evaluate the angular consistency, and adopted the information entropy of SAIs to measure spatial quality. Fang et al. (2018) calculated the change in visual quality by combining the gradient amplitude of SAIs and EPIs.

In addition to traditional methods, as deep learning exhibits excellent performance in other aspects of image processing, some teams have worked to fill the research gap of deep learning in the quality evaluation of LF images. Zhao et al. (2021) proposed an LF-IQA method based on the multi-task convolutional neural network (CNN), in which the EPI patches were taken as the input of the CNN model and the model followed ResNet in the convolution layer. Lamichhane et al. (2021) proposed an LF-IQA metric based on a CNN that measures the distortion of the saliency map. Lamichhane et al. (2021) confirmed that there is a strong correlation between the distortion levels of normalized images and the corresponding saliency maps. Guo et al. (2021) proposed a deep neural network-based approach, in which the relationship among SAIs was obtained by SAI fusion and global context perception models. To solve the problem of insufficient databases, they proposed a ranking-based method to generate pseudo-labels to pre-train the quality assessment network, and then fine-tuned the model at small-scale data sets with real labels.

This paper attempts to build a quality evaluation index that comprehensively considers the angular-spatial characteristics of LF images and human vision characteristics. The angular information of LF is directly expressed in the form of macro-pixel, which has been widely used in LF compression (Schiopu and Munteanu, 2018). Macro-pixels can be simply used to compare changes in angular information and do not involve a complex analysis of texture. For lenslet images, the array of pixels beneath each microlens is named as a macro-pixel. As shown in **Figure 1**, the second line is the enlarged local macro-pixels of the referenced lenslet image and the corresponding distorted macro-pixels. The enlarged part of the lenslet image contains 7 × 7 macro-pixels, and each macro-pixel contains 9 × 9 pixels. It can be seen from **Figure 1C** that luminance and chrominance have changed in the distorted macro-pixels. Hence,

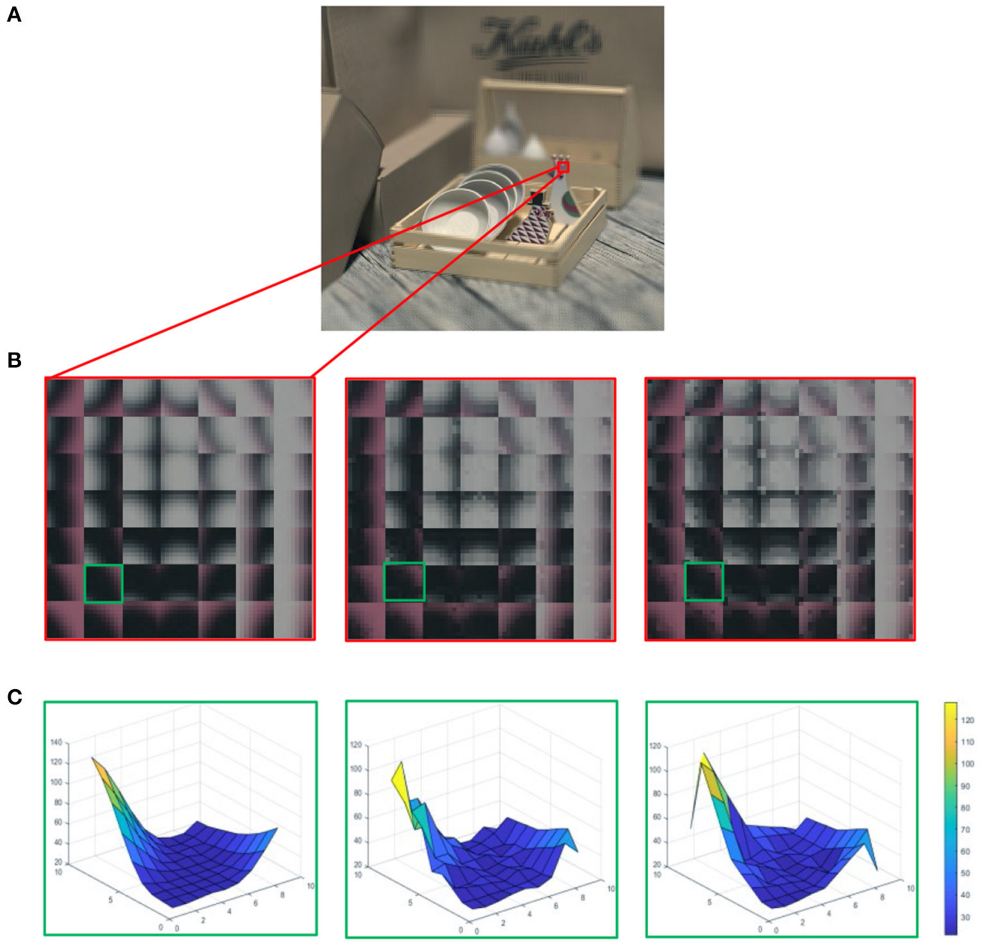

FIGURE 1 | **(A)** The referenced light field (LF) image in the form of decoded lenslet. **(B)** The first column is the enlarged local macro-pixels from **(A)**, and the other two columns correspond to macro-pixels with different degrees of distortion, which increased from left to right. **(C)** Each column corresponds to the grid distribution of gray values of a single macro-pixel in the green block in **(B)**.

we first utilize the angular information of all spatial positions to globally analyze the angular-spatial quality of LF images. As for spatial information, texture structure is an important and a direct means for human eyes to perceive image quality. Ingeniously, the focus stack not only reflects the texture structure information but also partly maps the angular information. Min et al. (2018) mentioned that quality degradations can cause local image structure changes, and Min et al. (2017a,b) mentioned that corners and edges are presumably the most important image features that are sensitive to various image distortions. Therefore, we construct a local LF angular-spatial quality evaluation model based on the focus stack through the measurement of corner and texture structures. Finally, the abovementioned two clues are combined to represent the overall quality of LF images. The contributions of this paper mainly include the following three points.

- A prediction framework of global angular-spatial distortion of LF images is established on the lenslet images. First, the distortion of angular information is calculated by averaging the changes in luminance and chrominance of each macro-pixel. All the evaluated values are arranged according to the corresponding spatial coordinates, forming an array of predicted values of angular distortion. Then, the visual saliency of the central SAI, which reflects the spatial information distribution with human visual characteristics, is introduced to pool an array of predicted values of angular distortion to obtain the predicted value of global distortion.
- An evaluation framework of local angular-spatial distortion of LF images is built on the principal components of the focus stack. The loss of the focalizing structure and the distortion of spatial texture structure are analyzed on the principal components through the corner similarity and texture similarity, respectively. The final local distortion is evaluated by fusing the predicted values of the focalizing structure and texture structure.
- The proposed method is compared with multiple objective metrics in the stitched multi-view image framework, and their results are analyzed with three subjective LF-IQA databases to verify their effectiveness and robustness.

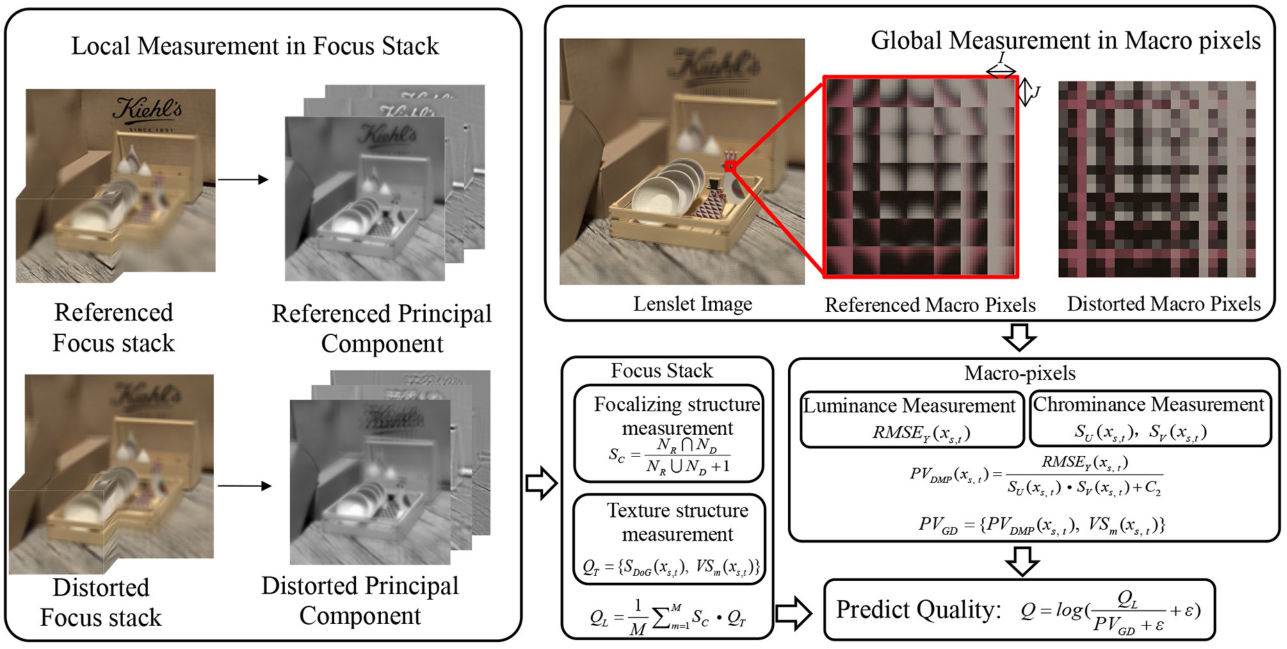
Local Measurement in Focus Stack
Referenced Focus stack
Referenced Principal Component
Distorted Focus stack
Distorted Principal Component
Global Measurement in Macro pixels
I
J
Lenslet Image
Referenced Macro Pixels
Distorted Macro Pixels
Focus Stack
Focalizing structure measurement
$S_C = \frac{N_R \cap N_D}{N_R \cup N_D + 1}$
Texture structure measurement
$Q_T = \{S_{DoG}(x_{s,t}),\ VS_m(x_{s,t})\}$
$Q_L = \frac{1}{M}\sum_{m=1}^{M} S_C \bullet Q_T$
Macro-pixels
Luminance Measurement
$RMSE_Y(x_{s,t})$
Chrominance Measurement
$S_U(x_{s,t}),\ S_V(x_{s,t})$
$PV_{DMP}(x_{s,t}) = \frac{RMSE_Y(x_{s,t})}{S_U(x_{s,t}) \bullet S_V(x_{s,t}) + C_2}$
$PV_{GD} = \{PV_{DMP}(x_{s,t}),\ VS_m(x_{s,t})\}$
Predict Quality: $Q = log(\frac{Q_L}{PV_{GD} + \varepsilon} + \varepsilon)$

FIGURE 2 | The proposed LF-IQA framework based on the angular-spatial feature information.

MATERIALS AND METHODS

Although the angular-spatial characteristics of LF are reflected in various expressions of LF, it is still a great challenge to extract and calculate the angular-spatial characteristics of LF. The lenslet images not only macroscopically reflect the global angular-spatial information of the LF images, but also microscopically reflect the angular information distribution. Inspired by this, we intend to start from the macro-pixels of the lenslet images to evaluate the angular distortion at the micro level, and then use the feature of spatial information to pool the predicted values of angular distortion. In consideration of the lack of analysis of useful texture and edge structure in the scene, which has a great influence on the quality perception, in the calculation of global distortion of LF images, the study in this paper will combine with other LF representations to supplement its deficiency. As each refocused image in the focus stack contains both angular-spatial information and texture structure, this paper chooses to analyze the texture and edge structure of the LF images with the focus stack.

According to the abovementioned analysis, we propose an evaluation method to comprehensively predict the distortion of LF images from both global and local aspects. The distribution of global and local distortion is analyzed from the lenslet images and focus stack, respectively. As illustrated in **Figure 2**, the global distortion in lenslet images is analyzed at each macro-pixel through the luminance and chroma channels. After then, we utilize the visual salient feature of spatial information to assign different weights to the measured values of each distorted macro-pixel, so as to realize the fusion of spatial information and angular information. Moreover, human visual characteristic has been taken into account in the calculation of visual saliency. As the single macro-pixel of a lenslet image lacks the texture and edge information of the objects in the scene, we complement the global distortion measurement by analyzing the principal components in the focus stack. The prediction processes of global and local distortion are described in sections The Prediction of Global Angular-Spatial Distortion and The Evaluation of Local Angular-Spatial Quality, respectively, and the two complementary prediction frameworks are fused in section The Evaluation of Union Angular-Spatial Quality.

The Prediction of Global Angular-Spatial Distortion

A lenslet image is composed of an array of macro-pixels embedded with angular information. The array of macro-pixels reflects the distribution of angular-spatial distortion macroscopically, while a single macro-pixel reflects the distribution of angular distortion microscopically. The size of a lenslet image is $S \times T$ units of macro-pixels, and the size of a macro-pixel is $I \times J$, where $S \times T$ is the spatial resolution of LF images, and $I \times J$ is the angular resolution of LF images.

As it can be seen from **Figures 1A,B**, the distortion of macro-pixels is manifested as the changes in luminance and chrominance. **Figure 1C** describes the grid distribution of referenced and different distorted macro-pixels, which reflects the influence of distortion on the angular information. Considering that a single macro-pixel involves all the angular information of the corresponding spatial position, we first compute the angular distortion within each macro-pixel.

As a single macro-pixel does not involve the complex texture and edge structure of the objects in the scene, we decided to study the variation of luminance information and chroma information in each macro-pixel. Without considering the image texture structure information, the root mean squared error (*RMSE*) method can simply and accurately calculate the error between referenced and distorted macro-pixels. As people are more sensitive to the change of luminance than that of chrominance (Su, 2013), we mainly measure the distortion of each macro-pixel on the luminance channel. Specifically, Equation (1) expresses the *RMSE* of luminance ($RMSE_Y$) of the referenced macro-pixel (Y_R) and the distorted macro-pixel (Y_D):

$$RMSE_Y(x_{s,t}) = \sqrt{\frac{1}{I \cdot J}\left(\sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J}\left(Y_R(x_{i,j}) - Y_D(x_{i,j})\right)^2\right)} \quad (1)$$

where $x_{s,t}$ is the pixel value on the spatial coordinate (s, t). $x_{i,j}$ is the pixel value on the angular coordinate (i, j). I and J are the angular resolutions, in this paper, $I = 9$, $J = 9$.

In addition to the variation of luminance information in the macro-pixel array, the distortion of chroma information will also affect the perception of the overall quality of images. As macro-pixels have no texture and edge structure of objects in the scene, the measurement of chroma distortion of macro-pixels can be simpler and more direct. Considering that the chrominance information has a much smaller impact on the overall quality than the luminance, we adopt the similarity measurement method that is widely used in objective assessment methods, as given in Equations (2) and (3). The chrominance information is analyzed in the YUV color space. The similarity map of each macro-pixel is averaged to calculate the quality value of the corresponding spatial position (s, t).

$$S_U(x_{s,t}) = \frac{1}{I \cdot J}\left(\sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J}\frac{2U_R(x_{i,j}) \cdot U_D(x_{i,j}) + C_1}{{U_R}^2(x_{i,j}) + {U_D}^2(x_{i,j}) + C_1}\right) \quad (2)$$

$$S_V(x_{s,t}) = \frac{1}{I \cdot J}\left(\sum\nolimits_{i=1}^{I}\sum\nolimits_{j=1}^{J}\frac{2V_R(x_{i,j}) \cdot V_D(x_{i,j}) + C_1}{{V_R}^2(x_{i,j}) + {V_D}^2(x_{i,j}) + C_1}\right) \quad (3)$$

where S_U and S_V are the color similarity of U and V channels. U_R and V_R are referenced macro-pixels of U and V channels, and U_D and V_D are distorted macro-pixels of U and V channels. The constant C_1 is used to maintain the stability of the similarity measurement function (Zhang et al., 2011), we fixed $C_1 = 1$ through the experiments.

The smaller $RMSE_Y$ between the referenced and distorted macro-pixel signifies the smaller error of the luminance components between them, while the greater chrominance similarity represents the smaller chroma error. For each macro-pixel, we use Equation (4) to fuse the predicted values of luminance and chrominance components. The values of $RMSE_Y$ are in the range of 0–255, to make the contribution of chroma less to the overall distortion prediction than the luminance, we set C_2 to 0.01, so that the range of chroma error is 0.99–100.

$$PV_{DMP}(x_{s,\,t}) = \frac{RMSE_Y(x_{s,\,t})}{S_U(x_{s,\,t}) \cdot S_V(x_{s,\,t}) + C_2} \quad (4)$$

where $PV_{DMP}(x_{s,t})$ is the fused prediction value of the distorted macro-pixel in the spatial coordinate (s, t), $s \epsilon [1, S]$, $t \epsilon [1, T]$. S and T are the spatial resolution, in this paper, $S = 434$, $J = 625$. The PV_{DMP} values arranged in spatial coordinates form an array of predicted values of angular distortion.

To integrate the angular information and spatial information of LF images in the process of image quality assessment, we intend to pool the predicted values of angular distortion using the spatial information. The exciting thing is that the corresponding spatial coordinates of macro-pixels reflect the significance of the texture and contour of the LF images. As the central SAI is the main perspective from which humans observe the scenes, we choose to use the features of the central SAI to pool an array of predicted values of angular distortion. The visual saliency map of the central SAI, which reflects the spatial information distribution with human visual characteristics, is introduced to pool the predicted values of all distorted macro-pixels, as given in Equation (5):

$$PV_{GD} = \frac{\sum_{s=1}^{S}\sum_{t=1}^{T} PV_{DMP}\left(x_{s,\,t}\right) \cdot VS_m\left(x_{s,\,t}\right)}{\sum_{s=1}^{S}\sum_{t=1}^{T} VS_m\left(x_{s,\,t}\right)} \quad (5)$$

where PV_{GD} is the predicted value of global angular-spatial distortion of LF images. $\mathrm{VS}_m\ (x_{s,t}) = \max\ [VS_r(x_{s,t}), VS_d(x_{s,t})]$, $VS_r(x_{s,t})$, and $VS_d(x_{s,t})$ are visual saliency maps of the central SAIs of referenced and distorted LF images, respectively. In this paper, we use the simple saliency model in Zhang et al. (2013), which integrates the frequency prior, color prior, and location prior and has been proven to be a simple and an effective visual saliency model that simulates the perceptual characteristics of human eyes to the images (Zhang et al., 2014).

The Evaluation of Local Angular-Spatial Quality

As mentioned earlier, the prediction of global angular-spatial distortion lacks direct measurements of the texture and edge structure of the objects in the scenes. This section aims to complement the global distortion measurement by analyzing the principal components in the focus stack. The focus stack consists of a series of refocused images arranged in the direction of depth. A refocused image is obtained by shifting and summing the SAIs at a given slope. Therefore, the refocused images only contain the local angular-spatial information of LF images. Specifically, the distortion of the angular information is directly manifested as the loss of the focalizing structure in the focus stack, while the distortion of the spatial information is manifested as various forms of destruction of the texture and edge structure in the scenes.

The loss of the focalizing structure is reflected as the disorder of the focus state. As shown in **Figure 3A**, the red and green boxes correspond to the cross and vertical sections of the focus stack. The sections of the referenced focus stack show that the focalizing structure is orderly, while the focusing state of the distorted focus stack is chaotic. Specifically, the foremost focusing position of the referenced focus stack is located on the wood plate, while the forefront refocused slice of the distorted focus stack is not in the focus state. Moreover, **Figure 3B** shows that the backmost refocus

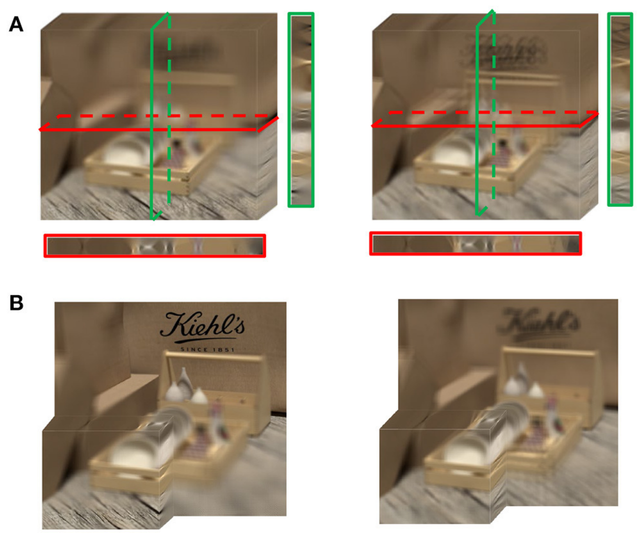

FIGURE 3 | Focus stack. **(A)** The focus stack of reference and distortion from left to right. The red and green boxes are the cross and vertical sections of the focus stack, respectively. **(B)** The partial focus stack of reference and distortion from left to right.

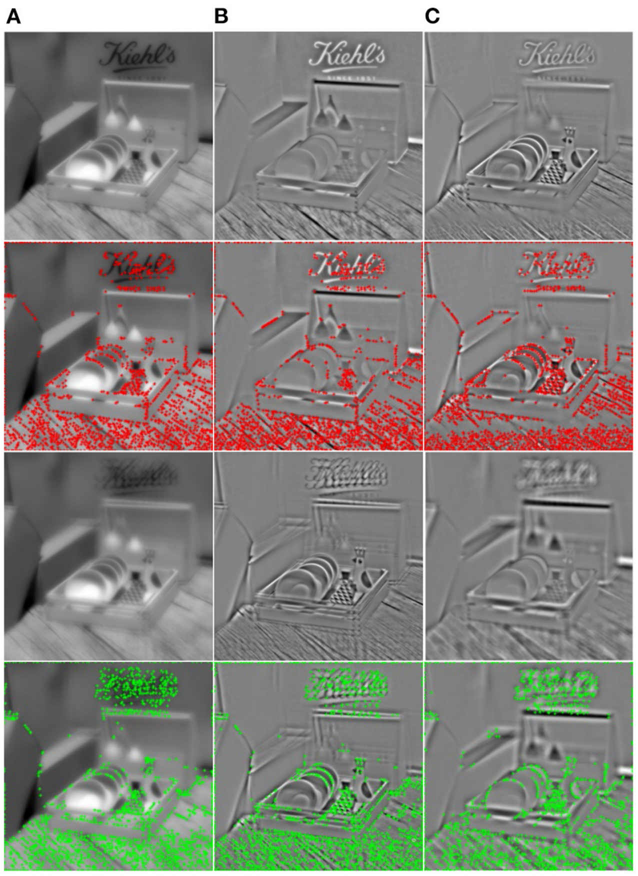

FIGURE 4 | The first and third rows are the principal components of the referenced and distorted focus stack, respectively. The second and fourth rows are the corners of referenced and distorted principal components, respectively. **(A)** The first principal component; **(B)** the second principal component; and **(C)** the third principal component.

slice of the referenced focus stack focused on the text, while the corresponding distorted refocus slice was not focused on the text that should be focused due to the angular-spatial distortion. In a word, the energy distribution of the distorted focus stack is scattered throughout the whole depth range, and the original focalizing structure is destroyed.

We also noticed from **Figure 3** that there is a defocused blur in the unfocused parts of the focus stack. When human eyes focus on a point of the scene, the object points at other depths of the field become blurred. The focus stack simulates the human eyes' habit of viewing a scene, so a defocused blur is inevitably introduced. To alleviate the effect of a defocused blur, we attempt to use principal component analysis (PCA) to extract the main components from the focus stack, as shown in the first and third rows of **Figure 4**. As we have analyzed the effect of chrominance on the prediction of global distortion (section Databases for Validation), the principal components are extracted only in the grayscale of the focus stack (Ren et al., 2017b).

Principal component analysis is a means of dimension reduction. The advantage is that PCA not only reduce the calculation amount for the focus stack but also alleviate the influence of a defocused blur in the analysis of the focalizing structure. By sorting the eigenvalues and corresponding eigenvectors of the covariance matrix of gray refocused slices in the focus stack, the focus stack can be rearranged according to the proportion of information content. As for the number of selected principal components, the experimental comparison and analysis are conducted (section 4.6). In this paper, the first three principal components are selected to predict the local angular-spatial quality for accuracy and simplicity.

For the principal components of the focus stack, we analyze the loss of focalizing structure and texture damage caused by the angular-spatial distortion. Firstly, the corner structure based on phase congruency (PC-corner) is used to evaluate the focalizing structure loss. As shown in the second and fourth rows of **Figure 4**, the PC-corner operator detects the features as points in an image with a high-phase component order in the Fourier domain, and it is not affected by luminance, contrast, and scale. The PC-corner feature operator can detect a wide range of features, such as angle, line, and texture information of images.

The corner response function is developed based on the covariance matrix of PC (Kovesi, 2003), as given in Equation (6):

$$CM = \begin{bmatrix} {PC_x}^2 & PC_x \cdot PC_y \\ PC_x \cdot PC_y & {PC_y}^2 \end{bmatrix} \tag{6}$$

where PC_x and PC_y are PC-corner at horizontal and vertical directions. The phase consistency utilizes the log-Gabor filter of multi-scale and multi-direction. The final covariance matrix is normalized with the orientations used in the log-Gabor filter. In this paper, we use three scales ($n = 1, 2, 3$) and six orientations ($\theta = 0, \pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6$).

Being different from the structural loss of ordinary image, the structural loss of the focus stack includes the reduction and increment of structure due to the angular-spatial distortion. Therefore, we use the form of Equation (7) to calculate the corner similarity S_C between referenced and distorted principal components.

$$S_C = \frac{N_R \bigcap N_D}{N_R \bigcup N_D + 1} \tag{7}$$

where N_R and N_D are the number of corners in referenced and distorted principal components, respectively. ∩ is the intersection of N_R and N_D, and ∪ is the union of N_R and N_D. The constant 1 is added to avoid the denominators being 0.

Secondly, in addition to assessing the loss of the focalizing structure, the angular-spatial distortion can also lead to an obvious texture damage of the focus stack. Similar to the evaluation of focalizing structure, the prediction of texture distortion is conducted on the principal components of the focus stack. The vertebrate retina can be mathematically represented by the Laplacian of Gaussian, which is an effective method of texture calculation reflecting the characteristics of human vision. Considering that the waveform distribution of DoG algorithm is similar to that of Laplacian of Gaussian, and the complexity of DoG is much smaller, we choose DoG to calculate the texture feature.

The DoG is the difference of the image signal $I(x_{s,t})$ convolved with the two different Gaussian scales $\sigma 1, \sigma 2$:

$$L(x_{s,t}, \sigma_1) = G(x_{s,t}, \sigma_1) * I(x_{s,t}) \tag{8}$$

$$L(x_{s,t}, \sigma_2) = G(x_{s,t}, \sigma_2) * I(x_{s,t}) \tag{9}$$

$$DoG(x_{s,t}) = L(x_{s,t}, \sigma_1) - L\left(x_{s,t}, \sigma_2\right) \tag{10}$$

where L $(x_{s,t}, \sigma_1)$ and L $(x_{s,t}, \sigma_2)$ are convolutions of the image signal I $(x_{s,t})$ with Gaussian functions at the two different Gaussian scales $(\sigma 1, \sigma 2)$.

Equation (11) was initially used in the calculation of structure similarity (SSIM) (Wang et al., 2004), and then widely used for the distance calculation of feature similarity (FSIM) in objective assessment methods. Hence, the texture similarity of referenced and distorted principal components is calculated by Equation (11).

$$S_{DoG}(x_{s,t}) = \frac{2DoG_R(x_{s,t}) \cdot DoG_D(x_{s,t}) + C_3}{DoG_R{}^2(x_{s,t}) + DoG_D{}^2(x_{s,t}) + C_3} \tag{11}$$

where DoG_R and DoG_D are differences of Gaussian feature of referenced and distorted principal components, respectively. The constant C_3 is used to maintain the stability of the similarity measurement function, we fixed $C_3 = 0.1$ through the experiments.

Concretely, the similarity map of DoG is pooled through the feature of visual saliency to obtain the quality of texture Q_T, as given in Equation (12). The calculation method of visual saliency is the same (as mentioned in section Databases for Validation):

$$Q_T = \frac{\sum_{s=1}^{S} \sum_{t=1}^{T} S_{DoG}\left(x_{s,t}\right) \mathbf{.}\, VS_m\left(x_{s,t}\right)}{\sum_{s=1}^{S} \sum_{t=1}^{T} VS_m\left(x_{s,t}\right)} \tag{12}$$

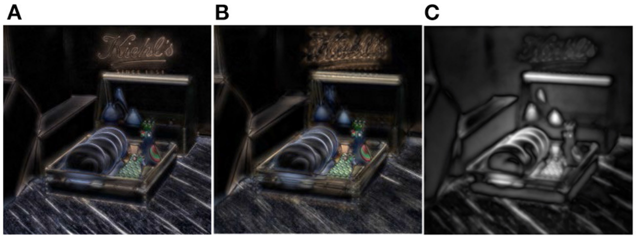

FIGURE 5 | The light flow in the focus stack and the visual saliency map. **(A)** The light flow of referenced focus stack. **(B)** The light flow of distorted focus stack. **(C)** The visual saliency map based on the sum of light flow of focus stack.

We define the light flow in the focus stack as the sum of the differences between adjacent refocus slices. The feature of visual saliency VS_m is computed with the light flow of the focus stack, as shown in **Figure 5** and Equation (13).

$$VS_m = \max(VS_{Lif\text{-}R}, VS_{Lif\text{-}D}) \tag{13}$$

where VS_{Lif-R} and VS_{Lif-D} are visual saliency maps of the light flow of referenced and distorted focus stack, respectively.

Finally, the local angular-spatial quality Q_L is obtained by averaging the fused quality of the focalizing structure and texture. M in Equation (14) is the number of principal components, which is analyzed in section 4.6 at different M values.

$$Q_L = \frac{1}{M} \sum\nolimits_{m=1}^{M} S_C \,\mathbf{.}\, Q_T \tag{14}$$

The Evaluation of Union Angular-Spatial Quality

According to sections Databases for Validation and Performance Analysis of Image Quality Metrics, a smaller PV_{GD} value indicates the smaller global distortion, which corresponds to the higher global quality, while a smaller Q_L value indicates the smaller local quality. The overall quality of LF images is calculated by fusing the predicted value of global angular-spatial distortion $PV_{GD}$and local angular-spatial quality Q_L. Considering that PV_{GD} and Q_L are inversely and directly proportional to the overall quality, respectively, we use Equation (15) to calculate the overall quality of the LF images.

$$Q = log(\frac{Q_L}{PV_{GD} + \varepsilon} + \varepsilon) \tag{15}$$

where log operation is added to increase the linearity of the results, which conforms to the human eyes' ability to recognize the light intensity (Min et al., 2020). PV_{GD} is given by Equation (5), and Q_L is given by Equation (14). ε is a constant for equation stability, which is set as 0.0001.

TABLE 1 | The detailed information of LF-IQA databases used in the experiment.

Database	Distortion types		Distortion levels
SHU	Traditional Distortion	JPEG	QLs: 1, 10, 15, 20, 50, 90
		JPEG2000:	CRs: 10, 70, 150, 200, 250, 400, 600
		Motion Blur:	MLs: 10, 20, 60, 100, 150, 200
		Gaussian Blur:	SDs: 0.5, 2, 4, 5, 10, 20
		White Noise:	SDs: 0.05, 0.1, 0.3, 0.5, 1, 2
VALID-10bit	Video & LF Compression	HEVC, VP9, Ahmad et al., 2017; Tabus et al., 2017; Zhao and Chen, 2017	Bpp: 0.005, 0.02, 0.1, 0.75.
NBU-LF1.0	LF Reconstruction	NN, BI, EPICNN	RFs: 5, 3, 2
		DR	RFs: 7, 5, 3
		VDSR	RFs: 2, 3, 4

RESULTS

Databases for Validation

Resource identification initiative. To verify the performance of the proposed method, experiments were conducted on three subjective quality assessment databases of LF images, including the database of traditional distortion types: SHU (Shan et al., 2019), video compression, and LF compression types: VALID-10bit (Viola and Ebrahimi, 2018), and LF reconstruction types: NBU-LF1.0 (Huang et al., 2019b). The detailed information of these databases is listed in **Table 1**.

1) *SHU database: traditional distortion types*. The SHU database is composed of 8 referenced LF images and 240 distorted LF images. There are five distortion types, including the classical compression artifacts (JPEG and JPEG2000) and other distortions (motion blur, Gaussian blur, and white noise). Each type of distortion has six distortion levels. The database is visualized by pseudo-sequence video of SAIs to the subjects.
2) *VALID-10bit database: video compression and LF compression distortion types*. There are two general compression schemes (HEVC and VP9) and three compression schemes specifically designed for LF (Ahmad et al., 2017; Tabus et al., 2017; Zhao and Chen, 2017). For each compression type, 4 levels of compression are introduced, and a total of 100 compressed LFs are included in this data set. It has five referenced LF contents and is evaluated in the passive methodology. For the passive evaluation, the perspective views were shown as animation and followed by the refocused views (Viola et al., 2017).
3) *NBU-LF1.0 data set: reconstruction distortion types*. It includes five LF reconstruction schemes: neighbor interpolation (NN), bicubic interpolation (BI), learning-based reconstruction (EPICNN), disparity-map-based reconstruction (DR), and spatial super-resolution reconstruction (SSRR). It has 14 referenced LF contents and 210 distorted LF images. Each reconstruction type has three levels of reconstruction.

To reduce the complexity, the number of multiple views selected from the databases in **Table 1** is 9 × 9, and the image resolution is 434 × 625.

Performance Analysis of Image Quality Metrics

There are three main representations of LF with whole global information: EPIs, lenslet images, and SAIs. First of all, the oblique texture structure in EPIs is not similar to the texture structure of objects in ordinary images, which is not conducive to the realization of traditional image quality evaluation methods. Except for the statistical IQA method at pixel-level, such as peak signal-to-noise ratio (PSNR), most traditional image quality evaluation methods cannot take the advantage of their simulation in image structure and human visual characteristics. Secondly, lenslet images have discontinuities of scene texture due to the angular information, which is not conducive to the application of algorithms based on human visual characteristics. Thirdly, SAIs can be regarded as a matrix of 2D images distributed in different angular directions. The superiority of traditional algorithms can be developed in the stitched SAIs, which is due to the fact that the stitched SAIs can be seen as a large 2D image with texture redundancy. Hence, we decide to apply the traditional algorithms to the stitched SAIs to carry out the following comparison experiments.

In general, the objective evaluation includes three categories according to their dependence on the reference image: full reference (FR), reduced reference (RR), and no reference (NR) (Wang and Bovik, 2006). In **Table 2**, the performance of the proposed MPFS is broadly compared with the classical FR, RR, and NR metrics over three subjective LF-IQA databases. The metrics mainly include classical traditional IQA metrics and the state-of-the-art LF-IQA metrics. 2D FR IQA metrics include PSNR, SSIM (Wang et al., 2004), multi-scale SSIM (MS-SSIM) (Wang et al., 2003), information weighting SSIM (IW-SSIM) (Wang and Li, 2010), FSIM (Zhang et al., 2011), FSIM based on Riesz transforms (RFSIM) (Zhang et al., 2010), noise quality measure (NQM) (Damera-Venkata et al., 2000), gradient similarity (GSM) (Liu et al., 2011), visual signal noise ratio (VSNR) (Chandler and Hemami, 2007), most apparent distortion (MAD) (Larson and Chandler, 2010), gradient magnitude similarity deviation (GMSD) (Xue et al., 2013), and HDRVDP (Mantiuk et al., 2011). Sparse feature fidelity (SFF) (Chang et al., 2013), universal image quality index (UQI) (Wang and Bovik, 2002), visual saliency-induced index (VSI) (Zhang et al., 2014), 2D RR IQA metrics include wavelet-domain natural image statistic model (WNISM) (Wang and Simoncelli, 2005), wavelet-based contourlet transform (WBCT) (Gao et al., 2008), and contourlet (Tao et al., 2009). Multi-view FR IQA metrics include morphological pyramids PSNR (MP-PSNR) (Sandić-Stanković et al., 2015), morphological wavelets PSNR (MW-PSNR) (Sandić-Stanković et al., 2015), MW-PSNRreduc (Sandić-Stanković et al., 2015), and 3DSwIM (Battisti et al., 2015). LFI FR IQA metrics include the algorithms in Min et al. (2020) and Meng et al. (2020). LFI NR IQA metrics include BELIF (Shi et al., 2019),

TABLE 2 | The performance comparison of classical IQA indexes on three benchmark databases.

	Database	SHU				VALID-10bit				NBU-LF1.0				Overall	
	Metric	**RMSE**	**PLCC**	**SROCC**	**KROCC**	**RMSE**	**PLCC**	**SROCC**	**KROCC**	**RMSE**	**PLCC**	**SROCC**	**KROCC**	**WSROCC**	**MSROCC**
2D-FR	PSNR	0.6316	0.8190	0.8859	0.7315	0.4122	0.9036	0.8868	0.7158	0.5982	0.7627	0.7609	0.5640	0.8383	0.8295
	SSIM	0.6422	0.8121	0.8262	0.6567	0.3481	0.9323	0.9273	0.7614	0.6197	0.7424	0.7223	0.5218	0.8049	0.8253
	MS-SSIM	0.5192	0.8817	0.8909	0.7150	0.3155	0.9447	0.9348	0.7793	0.5447	0.8083	0.8125	0.6078	0.8689	0.8794
	IW-SSIM	0.5129	0.8848	0.8892	0.7181	0.2781	0.9573	0.9441	0.7957	0.5461	0.8071	0.8045	0.6032	0.8668	0.8793
	FSIMc	0.5362	0.8733	0.8928	0.7168	0.2907	0.9533	0.9477	0.8006	0.5351	0.8157	0.8106	0.6055	0.8714	0.8837
	RFSIM	0.5977	0.8397	0.8473	0.6686	0.5738	0.8028	0.7915	0.6006	0.7877	0.5242	0.5352	0.3857	0.7180	0.7247
	NQM	0.6507	0.8065	0.8129	0.6330	0.7043	0.6815	0.6675	0.4867	0.7369	0.6044	0.5938	0.4264	0.7028	0.6914
	GSM	0.6381	0.8148	0.8209	0.6410	0.4159	0.9018	0.8686	0.7139	0.6890	0.6671	0.6583	0.4914	0.7675	0.7826
	VSNR	0.6255	0.8228	0.8408	0.6547	0.5425	0.8260	0.8049	0.6234	0.6199	0.7422	0.7497	0.5497	0.7995	0.7985
	MAD	0.5311	0.8759	0.8652	0.6869	0.2744	0.9585	0.9327	0.7776	0.4798	0.8549	0.8583	0.6614	0.8748	0.8854
	GMSD	0.5353	0.8737	0.8782	0.7003	0.2604	0.9627	0.9465	0.8037	0.5669	0.7902	0.7900	0.5916	0.8569	0.8716
	HDRVDP	0.6668	0.7955	0.7754	0.5935	0.4254	0.8970	0.8799	0.6963	0.7358	0.6059	0.5247	0.3744	0.6987	0.7267
	SFF	0.4594	0.9087	0.9196	0.7597	0.3299	0.9394	0.9245	0.7662	0.5554	0.7997	0.8009	0.6050	0.8752	0.8817
	UQI	0.8322	0.6544	0.6004	0.4424	0.4148	0.9024	0.8578	0.7049	0.7729	0.5493	0.5630	0.4066	0.6329	0.6737
	VSI	0.5755	0.8524	0.8556	0.6819	0.5122	0.8466	0.8191	0.6438	0.7044	0.6481	0.6399	0.4774	0.7666	0.7715
2D-RR	WNISM	0.7477	0.7338	0.7250	0.5578	0.3341	0.9378	0.9394	0.7846	0.8057	0.4911	0.4710	0.3229	0.6670	0.7118
	WBCT	0.7582	0.7248	0.7617	0.5861	0.5122	0.8466	0.8191	0.6438	0.6869	0.6697	0.6393	0.4636	0.7254	0.7400
	Contourlet	0.6985	0.7728	0.7498	0.5812	0.4473	0.8854	0.8704	0.6919	0.6595	0.7012	0.6605	0.4786	0.7376	0.7602
Multi- view FR	MP-PSNR	0.5983	0.8393	0.8599	0.6694	0.3633	0.9260	0.9239	0.7614	0.6885	0.6678	0.6611	0.4799	0.7956	0.8150
	MW-PSNR	0.5970	0.8401	0.8548	0.6658	0.3597	0.9275	0.9219	0.7561	0.6600	0.7007	0.6934	0.5019	0.8054	0.8234
	MW-PSNRreduc	0.6452	0.8101	0.8337	0.6433	0.3833	0.9172	0.9100	0.7369	0.7034	0.6494	0.6492	0.4653	0.7771	0.7976
	3DSwIM	0.5958	0.8408	0.8849	0.7135	0.2762	0.9579	0.9513	0.8185	0.7594	0.5709	0.5506	0.3890	0.7693	0.7956
LFI NR	BELIF	0.4847	0.8985	0.8697	0.6953	0.2431	0.9643	0.9454	0.8211	0.7072	0.6489	0.5983	0.4304	0.7798	0.8045
	Tensor-NLFQ	0.3494	0.9469	0.9392	0.8020	0.3163	0.9476	0.9074	0.7586	0.6603	0.6988	0.6064	0.4318	0.8063	0.8177
	VBLFI	0.4025	0.9354	0.9135	0.7613	0.2268	0.9705	0.9414	0.8042	0.5568	0.7934	0.7439	0.5549	0.8538	0.8663
LFI FR	Min et al., 2020	0.5951	0.8412	0.8460	0.6745	0.3335	0.9380	0.8524	0.7052	0.6843	0.6728	0.6659	0.4773	0.7784	0.7881
	Meng et al., 2020	0.4291	0.9208	0.9067	0.7427	0.2692	0.9601	0.9484	0.8043	0.5823	0.7770	0.7040	0.5133	0.8369	0.8530
	MPFS	0.3436	0.9500	0.9534	0.8183	0.2207	0.9734	0.9599	0.8305	0.4336	0.8833	0.8754	0.6908	0.9248	0.9296

Tensor-NLFQ (Zhou et al., 2020), and VBLIF (Xiang et al., 2020).

This paper used four IQA indexes to measure the fitting of the degree of objective scores and subjective scores. The Pearson linear correlation coefficient (PLCC) and the RMSE denote the accuracy of correlation between mean opinion scores (MOS) and predict scores. The Spearman rank order correlation coefficient (SROCC) and the Kendall rank order correlation coefficient (KROCC) can measure the prediction monotonicity of IQA metrics.

Table 2 presents the performance of classical objective metrics on SHU, VALID-10bit, and NBU-LF1.0 databases, where the values in bold indicate the best performance. The results show that the proposed MPFS method consistently fits well with MOS in both accuracy and monotonicity over the databases of traditional distortion, compressed distortion, and reconstructed distortion.

It can be seen from **Table 2** that the performance of traditional algorithms varies in different databases. Although these three databases contain different distortion types, their effects on angular and spatial information are reciprocal. First of all, some traditional algorithms perform well in the VALID-10bit database. This may be due to the fact that angular and spatial distortions in the VALID-10bit database are evenly distributed. Secondly, although the distortion of the SHU database is not derived from LF processing, it is still difficult to estimate the effects of these distortions on LF contents. For example, traditional algorithms do not take advantages they should have for traditional types of distortion. This is due to the fact that traditional algorithms fail to consider the relationship between the angular and spatial quality. In addition, most objective metrics cannot achieve good results in the NBU-LF1.0 database. This may be due to the complex distribution of angular-spatial distortion, for example, the cross effects of angular-spatial distortion vary greatly in different perspectives.

The performance of the multi-view algorithms is similar to that of the traditional 2D algorithms. They perform well when the distribution of the angular-spatial distortion is not complex, but worse for the NBU-LF1.0 database containing the distortion of reconstructed types. It somewhat indicates that the angular-spatial distortion caused by reconstruction algorithms is more complex.

The NR LF-IQA models were trained with 80% contents from each data set used in this paper, and 20% of contents were used

TABLE 3 | PLCC performance of different distortion types on VALID-10bit, SHU and NBU-LF1.0 databases.

	Database	VALID-10bit					SHU					NBU-LF1.0				
Type	**Metric**	**HEVC**	**VP9**	**P1**	**P2**	**P3**	**GB**	**JPEG2k**	**JPEG**	**MB**	**WN**	**NN**	**BI**	**EPICNN**	**DR**	**VDSR**
2D-FR	PSNR	0.9522	0.9392	0.9282	0.9361	0.8569	0.9198	**0.9502**	0.9752	0.8674	0.9570	0.7740	0.9345	0.8794	0.7030	0.7176
	SSIM	0.9493	0.9407	0.9531	0.9453	0.9289	0.9133	0.8697	0.9724	0.8446	0.9420	0.7951	0.8654	0.8502	0.4157	0.8415
	MS-SSIM	0.9625	0.9522	0.9444	0.9464	0.9360	0.9070	0.9321	0.9725	0.8983	0.9548	0.7695	0.9083	0.9294	0.6854	0.9056
	IW-SSIM	0.9727	0.9674	0.9567	0.9561	**0.9599**	0.9366	0.9375	0.9688	0.9430	0.9549	0.7409	0.9108	0.9360	0.7219	0.6393
	FSIMc	0.9667	0.9651	0.9569	0.9619	0.9409	0.9394	0.9389	**0.9797**	0.9134	0.9157	0.7810	0.9201	0.9213	0.6561	0.8912
	RFSIM	0.9368	0.9219	0.9220	0.7790	0.8378	0.8057	0.8593	0.9162	0.6631	0.9439	0.9189	0.8742	0.2057	0.8104	0.6917
	NQM	0.7686	0.6794	0.7272	0.6573	0.6725	0.7450	0.8479	0.8890	0.5832	0.9322	0.7128	0.8002	0.6220	0.7248	0.5475
	GSM	0.9761	0.9555	0.9677	0.9367	0.8530	0.8351	0.8257	0.9377	0.5277	0.9316	**0.9507**	0.8943	0.7124	**0.8552**	0.6360
	VSNR	0.8820	0.8273	0.8644	0.8747	0.8119	0.8363	0.6665	0.8758	0.6889	0.8569	0.8079	0.8498	0.8144	0.6629	0.7619
	MAD	0.9793	0.9674	**0.9774**	0.9504	0.9366	0.8769	0.9174	0.9186	0.8498	0.9551	0.9095	**0.9501**	**0.9429**	0.8496	0.8973
	GMSD	0.9782	0.9701	0.9731	**0.9738**	0.9520	0.9210	**0.9637**	0.9716	0.9260	0.9009	0.7216	0.9170	0.9265	0.7432	**0.9314**
	HDRVDP	0.9530	0.8827	0.9135	0.9016	0.8796	0.7197	0.8695	0.9523	0.5510	0.9600	0.8910	0.9418	**0.9396**	**0.8500**	0.7857
	SFF	0.9646	0.9528	0.9646	0.9678	0.8787	0.8799	0.9408	0.9734	0.8470	0.9308	0.7845	**0.9462**	0.9271	0.7273	**0.9466**
	UQI	0.9699	0.9680	**0.9749**	0.8785	0.9207	0.6885	0.4614	0.2193	0.5023	0.8736	0.7082	0.8691	0.1932	0.7449	0.0975
	VSI	0.9669	0.9503	0.9668	0.7954	0.8796	0.8413	0.8489	0.9525	0.5385	0.9378	**0.9355**	0.8994	0.7243	**0.8505**	0.6240
2D-RR	WNISM	0.9651	0.9537	0.9522	0.9282	0.9038	0.8924	0.6937	0.8170	0.8839	0.8508	0.7289	0.6830	0.7778	0.4444	0.8648
	WBCT	0.9128	0.8492	0.9105	0.9079	0.8648	0.8075	0.7910	0.7716	0.7744	0.9101	0.5781	0.8303	0.9144	0.6609	0.8089
	Contourlet	0.9288	0.9007	0.9373	0.9231	0.8498	0.8579	0.8528	0.7922	0.7789	0.9471	0.7039	0.8098	0.9218	0.6773	0.8650
Multi-view FR	MP-PSNR	**0.9818**	0.9766	0.9725	0.9701	0.9508	0.8475	0.8758	0.9391	0.7919	0.8190	0.8414	0.8441	0.6679	0.7210	0.7039
	MW-PSNR	0.9709	0.9619	0.9641	0.9610	0.9435	0.8221	0.8760	0.9622	0.6799	0.9074	0.8137	0.8917	0.6805	0.7601	0.6554
	MW-PSNRreduc	0.9784	0.9760	**0.9749**	0.9626	0.9539	0.7508	0.8706	0.9512	0.6076	0.8345	0.8159	0.8427	0.6096	0.7392	0.6788
	3DSwIM	0.9801	**0.9780**	0.9728	0.9640	0.9459	**0.9522**	0.9458	0.8893	0.9344	0.9139	0.8997	0.8746	0.8392	0.8333	0.8720
LFI NR	BELIF	–	–	–	–	–	0.9045	0.8308	0.9585	0.9388	**0.9665**	0.9026	0.9100	0.7182	0.7520	0.9134
	Tensor-NLFQ	–	–	–	–	–	0.9399	0.9284	**0.9849**	0.9411	**0.9749**	**0.9243**	0.8819	0.8430	0.8096	0.7926
	VBLIF	–	–	–	–	–	**0.9578**	0.7452	0.9694	**0.9632**	**0.9854**	0.8820	0.8905	0.8421	0.7051	0.8885
LFI FR	Min et al., 2020	0.9338	0.9667	0.9540	**0.9801**	**0.9616**	0.9288	**0.9643**	0.9397	0.9534	0.9581	0.7851	0.8303	0.7428	0.7492	0.9219
	Meng et al., 2020	**0.9825**	**0.9739**	0.9665	0.9486	0.9473	**0.9647**	0.8398	0.9772	**0.9815**	0.9586	0.8258	0.8812	0.8758	0.1380	0.9234
	MPFS	**0.9828**	**0.9789**	**0.9765**	**0.9702**	**0.9722**	0.9480	0.9456	**0.9774**	**0.9682**	0.9505	0.8766	**0.9677**	**0.9435**	0.7765	**0.9389**

for prediction. The optimal training parameters were obtained by multiple adjustments, and the result of each adjustment was the median value of 1,000 experiments. It can be seen from **Table 2** that they achieved preferable results at the first two databases, but perform worse for the reconstruction distortions with complex angular-spatial artifacts.

For the FR LF-IQA, the concept of optimal parallax range of human eyes is introduced into the focus stack to calculate the quality of LF images. Meng et al. (2019) used some camera parameters provided by the EPFL database (Honauer et al., 2016) when calculating the optimal parallax range, while some databases do not have these parameters. Therefore, in combination with the experiments of refocusing factors in section 4.7, we set the focusing range of Meng et al. (2020) as [−3, 3] over all databases for the sake of fairness. Min et al. (2020) computed the quality of LF images through the global–local spatial quality and the angular consistency measurement. It is necessary to note that the angular resolution of all databases is set as 9 × 9 in the comparison experiment for fairness. Therefore, the performance of both Meng et al. (2020) and Min et al. (2020) presented in **Tables 2**, **3** is not optimal.

It should be known that the performance of the same objective algorithm is slightly different in different databases. As suggested in Wang and Li (2010) and Zhang et al. (2014) we analyze the objective IQA metrics with the weighted average results across all databases for the overall performance. The weighted average $\overline{\rho}$ is computed as follows:

$$\overline{\rho}=\frac{\sum_i \rho_i \cdot \omega_i}{\sum_i \omega_i} \tag{16}$$

where ρ_i $(i = 1, 2, 3, 4)$ is the fitting performance for each database. The weight coefficient of each database depends on the number of distorted images in the respective database. **Table 2** presents the overall performance and the ranking of weighted-average SROCC of LF-IQA metrics over all databases.

The last two columns in **Table 2** are the weight-average SROCC (WSROCC) and the mean SROCC (MSROCC) for each objective metric over all databases, respectively. It can be seen that MPFS performs much better than the other metrics on the WSROCC and the MSROCC.

Robustness Against Distortion Types

The robustness of the proposed objective IQA model against various distortion types is verified. **Table 3** presents the performance comparison of classical objective models on the abovementioned three databases, covering various distortion types. Specifically, the VALID-10bit database contains two classical video compression schemes and three compression schemes specialized for LF images. The SHU database contains classical compression distortion and display distortion, and the NBU-LF1.0 database contains a variety of reconstructed distortion types specialized for LF images.

In **Table 3**, the values in bold indicate the first three best PLCC values for each distortion type. The performance of different objective algorithms for different distortion types is analyzed through PLCC, which can reflect the fitting accuracy of two sets of data. The results show that many algorithms have the optimal scope of application, and can only be sensitive to some specific distortion types. For example, most algorithms have a good predicted effect on the compressed distortion types in the VALID-10bit database, but are not effective for the reconstructed distortion types in the NBU-LF1.0 database or the traditional distortion types in the SHU database. The reason may be that the angular and spatial distortion in the VALID-10bit database is evenly distributed, while the cross effects of angular and spatial distortion of the other two databases vary greatly in different perspectives. The proposed method cannot achieve the best prediction for each distortion, but it performs relatively stable for all distortion types. The robustness of MPFS is superior to other metrics.

The Validity of the MPFS Model

The proposed MPFS method has two applications: the prediction of global angular-spatial distortion and the evaluation of local angular-spatial quality. The prediction framework of global angular-spatial distortion is established on the lenslet images. The angular distortion is first predicted at each macro-pixel. Then, the visual saliency of the central SAI is introduced to combine the angular and spatial information. The evaluation framework of local angular-spatial quality utilized the PC-corner and DoG algorisms to evaluate the loss of the focalizing structure and texture structure on the principal components of the focusing stack, respectively.

Table 4 compares the performance of the proposed MPFS in three cases: only the prediction framework of global angular-spatial distortion, only the local angular-spatial quality framework, and the combination of global and local frameworks. It can be seen that both local and global frameworks are effective in the VALID-10bit database, and they have reverse effects on the other two databases. The local angular-spatial quality evaluation framework based on the focus stack is more effective for both the spatial texture distortion and the focalizing structure loss caused by the angular distortion. Because the global framework is mainly based on the prediction of angular distortion, it will be mediocre when the distribution of the angular-spatial distortion is more complex. But the combination of the two frameworks works well, benefiting from their complementarity. Besides, **Table 5** lists the time complexity of the proposed MPFS method. The listed time under each data set is calculated by averaging the run time of all LF images. Although the size of some LF images in the NBU-LF1.0 database is slightly different from those in the other two databases, the running time is similar.

TABLE 4 | Performance of individual case on SHU, VALID-10bit, and NBU-LF1.0 databases.

Database	VALID-10bit		SHU		NBU-LF1.0	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
Local	0.9662	0.9551	0.8477	0.8286	0.8462	0.8356
Global	0.9493	0.9412	0.8610	0.8633	0.7854	0.7780
Local_Global	0.9734	0.9599	0.9500	0.9534	0.8833	0.8754

TABLE 5 | Time complexity on SHU, VALID-10bit, and NBU-LF1.0 databases.

Database	VALID-10bit	SHU	NBU-LF1.0
Time (second)	76.4793	74.0516	74.0665

TABLE 6 | Performance of individual features on SHU, VALID-10bit, and NBU-LF1.0 databases.

Database	VALID-10bit		SHU		NBU-LF1.0	
Features	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PC-corner	0.9643	0.9528	0.8339	0.8203	0.8336	0.8301
PC-corner-DoG	0.9662	0.9551	0.8477	0.8286	0.8462	0.8356
PC-corner-DoG-Y	0.9664	0.9534	0.8896	0.8785	0.8757	0.8684
PC-corner-DoG-YUV	0.9734	0.9599	0.9500	0.9534	0.8833	0.8754

The Validity of Individual Quality Component

After analyzing the contributions of the local/global angular-spatial quality framework, **Table 6** presents various features used in the proposed MPFS algorithm, the first two features measure the loss of focalizing structure and texture structure in the local angular-spatial quality framework. It can be seen that the combination of PC-corner and DoG features can better evaluate the angular-spatial distortion of the focus stack. However, due to the complex distribution of angular-spatial distortion, it does not work well in the SHU database.

In addition to the PC-corner and DoG features, **Table 6** also presents the performance after adding the luminance and chrominance features. These two features improve the accuracy of the evaluation algorithm. It can be seen that the chroma information contributes greatly to improve the performance of the proposed method in the SHU database because of the high chromaticity distortion of JPEG.

The Impact of Principal Components on the MPFS Model

The order of the principal components of the focus stack is obtained by sorting the eigenvalues and the corresponding eigenvectors of its covariance. The eigenvectors with larger eigenvalues reflect a larger amount of information. As can be seen from **Figure 4**, the first-order principal component reflects most of the low-frequency information in the focus stack, in which the

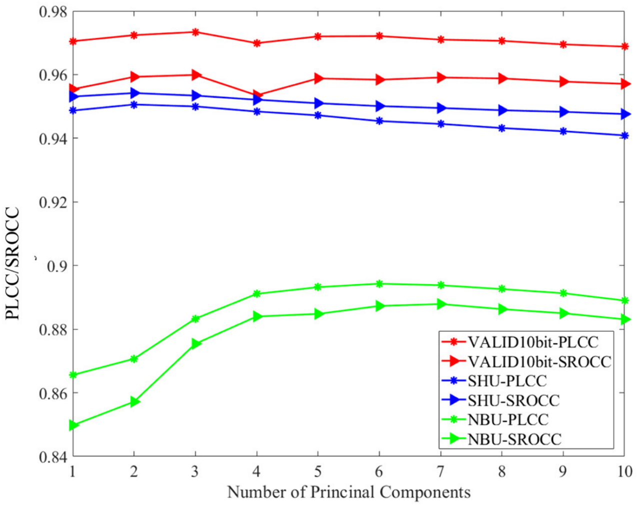

FIGURE 6 | The distribution of Pearson linear correlation coefficient (PLCC)/Spearman rank order correlation coefficient (SROCC) of the MPFS method at different numbers of the principal components in the focus stack over the three databases.

TABLE 7 | PLCC and SROCC of different refocus scopes on VALID-10bit, SHU, and NBU-LF1.0 databases.

Database	VALID-10bit		SHU		NBU-LF1.0	
Scope	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
[-1, 1]	0.9694	0.9544	0.9475	0.9503	0.8802	0.8727
[-2, 2]	0.9691	0.9560	0.9475	0.9499	0.8802	0.8721
[-3, 3]	0.9734	0.9599	0.9500	0.9534	0.8833	0.8754
[-4, 4]	0.9723	0.9583	0.9473	0.9514	0.8735	0.8582

TABLE 8 | PLCC and SROCC of different refocus intervals on VALID-10bit, SHU, and NBU-LF1.0 databases.

Database	VALID-10bit		SHU		NBU-LF1.0	
Scope	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
[-3, 3, 10]	0.9655	0.9488	0.9433	0.9475	0.8743	0.8642
[-3, 3, 15]	0.9734	0.9599	0.9500	0.9534	0.8833	0.8754
[-3, 3, 20]	0.9718	0.9583	0.9495	0.9525	0.8765	0.8616

defocused blur of the focus stack is mainly distributed in the first-order principal component. The other principal components mainly reflect the high-frequency information of the focus stack, and the distortion of focalizing structure is obvious in the higher-order principal components.

Although the PCA is carried out in the local angular-spatial quality evaluation framework, we analyze the impact of different numbers of principal components on the overall algorithm due to the complementarity of the two frameworks. **Figure 6** describes the distribution of PLCC/SROCC of the proposed MPFS method at different numbers of the principal components in the focus stack over the three databases. It can be seen that the variation trend of the final evaluation results over the three databases is inconsistent with an increase of the number of principal components, which is related to the completely different distortion types of the three databases. We finally choose the first three principal components to calculate the local angular-spatial quality for accuracy and simplicity.

The Impact of Refocusing Factors on the MPFS Model

The evaluation framework of local angular-spatial quality is based on the focus stack, while the refocusing factors will affect the evaluated final results. Specifically, the refocusing factors contain the refocus scope and refocus step. This paper conducts the refocus operation in the spatial domain. The refocused images are obtained by the LFFiltShiftSum function in LFToolbox0.4, which acts on shifting and summing the SAIs within a given slope scope to obtain the focus stack. Different slopes correspond to different depth planes. A step between the two slopes determines the number of refocused images within the given refocus scope.

Table 7 lists the PLCC and SROCC in multiple refocus scopes over the three databases. We set 15 intervals for all to refocus scopes in **Table 7**, that is, 16 refocus images are obtained. **Table 8** illustrates the effect of different intervals on the local angular-spatial quality under the optimal refocused scope in **Table 7**.

The results show that the optimal refocus scope of the focus stack is [−3, 3] in the local angular-spatial quality evaluation framework, and the optimal number of refocusing intervals is 15. However, the change of the refocus scope and step cannot cause a great influence, which indicates that the local angular-spatial quality framework based on the focus stack is relatively stable.

DISCUSSION

The quality evaluation for LF images is a new challenge due to the abundant scene information and the complex imaging structure. The existing objective methods are mainly carried out on the classical representations of LF images, especially SAIs, focus stack, and EPIs. It should be noted that different LF representations usually place different emphasis on the distribution of angular and spatial information. Comparatively speaking, the lenslet image and EPIs directly reflect the distortion of angular information, while the focus stack and SAIs directly reflect the distortion of spatial information. The advantages of angular-spatial information distribution in each representation can be better utilized by combining these LF representations, but the disadvantage is increased computational complexity.

The key to quality evaluation of LF images lies on how to combine the human visual perception and the LF angular-spatial characteristics. In this paper, we propose a new LF quality evaluation method through the global angular-spatial quality framework based on macro-pixels and the local angular-spatial quality framework based on the focus stack. The global angular-spatial quality framework evaluates the distortion of luminance and chrominance at each macro-pixel, primarily representing the angular distortion. Then, the visual saliency of human eyes to spatial texture structure is introduced to pool an array of predicted values of angular distortion. However, although the macro-pixel array reflects the global information of LF images, the single macro-pixel lacks the texture information of objects in the scene. Fortunately, the focus stack can help to measure the damage of spatial texture structure and the loss of the focalizing structure caused by the angular distortion. Therefore, a local angular-spatial quality framework based on the principal component of the focus stack is adopted to complement the global framework. The losses of the focalizing structure and texture structure are analyzed through the PC-corner similarity and DoG texture similarity, respectively. Extensive experimental results show that better performance can be obtained by combining the complementary local/global angular-spatial quality evaluation framework.

In the future work, we decide to explore ways to reduce the computational complexity of evaluating global angular-spatial distortion distribution, such as introducing the random sampling mechanism into the distortion prediction of macro-pixels. Moreover, how to achieve better integration of LF angular-spatial characteristics and human visual characteristics under the condition of low computational complexity is still a challenge for the quality evaluation of LF images. The application of human visual characteristics in this paper is divided into two types. First, the global framework uses the saliency distribution of spatial information as the weight to realize the integration of the distribution of angular distortion and spatial structure. Second, feature extraction operators of PC-corner and DoG, which simulate human visual characteristics, are, respectively, applied to the calculation of focalizing structure and texture structure. In general, the application of human visual characteristics in the quality evaluation of LF images mainly lies on the fusion of angular and spatial distortion prediction, or the feature extraction in the prediction of angular distortion and spatial distortion. It is difficult to achieve the perfect fusion of LF angular-spatial characteristics and human visual characteristics in the traditional algorithms, while the deep learning methods have strong ability to learn the relationship between the angular information and spatial information, as well as the relationship between the human visual characteristics and LF angular-spatial characteristics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Visual quality Assessment for Light field Images Dataset (VALID), https://mmspg.epfl.ch/VALID.

AUTHOR CONTRIBUTIONS

CM performed the experiments and wrote the first draft of the manuscript. PA provided mentorship into all aspects of the research. PA and XH modified the content of the manuscript. All authors contributed ideas to the design and implementation of the proposal, read, and approved the final version of the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China, under (Grant Nos. 62020106011, 62001279, 62071287, and 61901252), Science and Technology Commission of Shanghai Municipality under (Grant 20DZ2290100).

REFERENCES

Ahmad, W., Olsson, R., and Sjöström, M. (2017). "Interpreting plenoptic images as multi-view sequences for improved compression," in *2017 IEEE International Conference on Image Processing* (ICIP). doi: 10.1109/ICIP.2017.8297145

Battisti, F., Bosc, E., Carli, M., Le Callet, P., and Perugia, S. (2015). Objective image quality assessment of 3D synthesized views. *Signal Process.* 30, 78–88. doi: 10.1016/j.image.2014.10.005

Chandler, D. M., and Hemami, S. S. (2007). VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transact. Image Process.* 16, 2284–2298. doi: 10.1109/TIP.2007.901820

Chang, H. W., Yang, H., Gan, Y., and Wang, M. H. (2013). Sparse feature fidelity for perceptual image quality assessment. *IEEE Transact. Image Process.* 22, 4007–4018. doi: 10.1109/TIP.2013.2266579

Damera-Venkata, N., Kite, T. D., Geisler, W. S., Evans, B. L., and Bovik, A. C. (2000). Image quality assessment based on a degradation model. *IEEE Transact. Image Process.* 9, 636–650. doi: 10.1109/83.841940

Fang, Y., Wei, K., Hou, J., Wen, W., and Imamoglu, N. (2018). "Light filed image quality assessment by local and global features of epipolar plane image," in *2018 IEEE Fourth International Conference on Multimedia Big Data* (BigMM). doi: 10.1109/BigMM.2018.8499086

Gao, X., Lu, W., Li, X., and Tao, D. (2008). Wavelet-based contourlet in quality evaluation of digital images. *Neurocomputing* 72, 378–385. doi: 10.1016/j.neucom.2007.12.031

Guo, Z., Gao, W., Wang, H., Wang, J., and Fan, S. (2021). "No-reference deep quality assessment of compressed light field images," in *2021 IEEE International Conference on Multimedia and Expo* (ICME). doi: 10.1109/ICME51207.2021.9428383

Honauer, K., Johannsen, O., Kondermann, D., and Goldluecke, B. (2016). "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Asian Conference on Computer Vision* (Cham: Springer). doi: 10.1007/978-3-319-54187-7_2

Huang, F. C., Chen, K., and Wetzstein, G. (2016). The light field stereoscope immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Transact. Graphics* 35, 60.1-60.12. doi: 10.1145/2766922

Huang, X., An, P., Cao, F., Liu, D., and Wu, Q. (2019a). Light-field compression using a pair of steps and depth estimation. *Optics Express* 27, 3557–3573. doi: 10.1364/OE.27.003557

Huang, Z., Yu, M., Jiang, G., Chen, K., Peng, Z., and Chen, F. (2019b). "Reconstruction distortion oriented light field image dataset for visual communication," In *2019 International Symposium on Networks, Computers and Communications* (ISNCC). doi: 10.1109/ISNCC.2019.8909170

Kovesi, P. (2003). "Phase congruency detects corners and edges," in *The Australian Pattern Recognition Society Conference: DICTA* (Sydney).

Lamichhane, K., Battisti, F., Paudyal, P., and Carli, M. (2021). "Exploiting saliency in quality assessment for light field images," in *2021 Picture Coding Symposium* (PCS) (pp. 1-5). doi: 10.1109/PCS50896.2021.9477451

Larson, E. C., and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electr. Imaging* 19:011006. doi: 10.1117/1.3267105

Liu, A., Lin, W., and Narwaria, M. (2011). Image quality assessment based on gradient similarity. *IEEE Transact. Image Process.* 21, 1500–1512. doi: 10.1109/TIP.2011.2175935

Luo, Z., Zhou, W., Shi, L., and Chen, Z. (2019). "No-reference light field Image quality assessment based on micro-lens image," in *2019 Picture Coding Symposium* (PCS). doi: 10.1109/PCS48520.2019.8954551

Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transact. Graphics* 30, 1–14. doi: 10.1145/2010324.1964935

Meng, C., An, P., Huang, X., and Yang, C. (2019). "Objective quality assessment for light field based on refocus characteristic," in *International Conference on Image and Graphics* (Cham: Springer). doi: 10.1007/978-3-030-34113-8_17

Meng, C., An, P., Huang, X., Yang, C., and Liu, D. (2020). Full reference light field image quality evaluation based on angular-spatial characteristic. *IEEE Signal Process. Lett.* 27, 525–529. doi: 10.1109/LSP.2020.2982060

Meng, C., An, P., Huang, X., Yang, C., Shen, L., and Wang, B. (2021). Objective quality assessment of lenslet light field image based on focus stack. *IEEE Transact Multimedia*. doi: 10.1109/TMM.2021.3096071

Min, X., Gu, K., Zhai, G., Liu, J., Yang, X., and Chen, C. W. (2017a). Blind quality assessment based on pseudo-reference image. *IEEE Transact. Multimedia* 20, 2049–2062. doi: 10.1109/TMM.2017.2788206

Min, X., Ma, K., Gu, K., Zhai, G., Wang, Z., and Lin, W. (2017b). Unified blind quality assessment of compressed natural, graphic, and screen content images. *IEEE Transact Image Process.* 26, 5462–5474. doi: 10.1109/TIP.2017.2735192

Min, X., Zhai, G., Gu, K., Liu, Y., and Yang, X. (2018). Blind image quality estimation via distortion aggravation. *IEEE Transact. Broadcasting*. 64, 508–517. doi: 10.1109/TBC.2018.2816783

Min, X., Zhou, J., Zhai, G., Le Callet, P., Yang, X., and Guan, X. (2020). A metric for light field reconstruction, compression, and display quality evaluation. *IEEE Transact. Image Process.* 29, 3790–3804. doi: 10.1109/TIP.2020.2966081

Paudyal, P., Battisti, F., and Carli, M. (2019). Reduced reference quality assessment of light field images. *IEEE Transact. Broadcasting* 65, 152–165. doi: 10.1109/TBC.2019.2892092

Ren, M., Liu, R., Hong, H., Ren, J., and Xiao, G. (2017a). Fast object detection in light field imaging by integrating deep learning with defocusing. *Appl. Sci.* 7:1309. doi: 10.3390/app7121309

Ren, W., Wu, D., Jiang, J., Yang, G., and Zhang, C. (2017b). "Principle component analysis based hyperspectral image fusion in imaging spectropolarimeter," in Second *International Conference on Photonics and Optical Engineering* (International Society for Optics and Photonics). doi: 10.1117/12.2256805

Sandić-Stanković, D., Kukolj, D., and Le Callet, P. (2015). "DIBR synthesized image quality assessment based on morphological wavelets," in *2015 Seventh International Workshop on Quality of Multimedia Experience* (QoMEX). doi: 10.1109/QoMEX.2015.7148143

Schiopu, I., and Munteanu, A. (2018). "Macro-pixel prediction based on convolutional neural networks for lossless compression of light field images," in *2018 25th IEEE International Conference on Image Processing* (ICIP). doi: 10.1109/ICIP.2018.8451731

Shan, L., An, P., Meng, C., Huang, X., Yang, C., and Shen, L. (2019). A no-reference image quality assessment metric by multiple characteristics of light field images. *IEEE Access* 7, 127217–127229. doi: 10.1109/ACCESS.2019.2940093

Shi, L., Zhao, S., and Chen, Z. (2019). "BELIF: Blind quality evaluator of light field image with tensor structure variation index," in *2019 IEEE International Conference on Image Processing* (ICIP). doi: 10.1109/ICIP.2019.8803559

Su, Q. (2013). *Method for Image Visual Effect Improvement of Video Encoding and Decoding. U.S. Patent No. 8,457,196*. Washington, DC: U.S. Patent and Trademark Office.

Tabus, I., Helin, P., and Astola, P. (2017). "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000," in *2017 IEEE International Conference on Image Processing* (ICIP). doi: 10.1109/ICIP.2017.8297147

Tao, D., Li, X., Lu, W., and Gao, X. (2009). Reduced-reference IQA in contourlet domain. *IEEE Transact. Syst.* 39, 1623–1627. doi: 10.1109/TSMCB.2009.2021951

Tian, Y., Zeng, H., Hou, J., Chen, J., Zhu, J., and Ma, K. K. (2020). A light field image quality assessment model based on symmetry and depth features. *IEEE Transact. Circuits Syst. Video Technol.* 31, 2046–2050. doi: 10.1109/TCSVT.2020.2971256

Tian, Y., Zeng, H., Xing, L., Chen, J., Zhu, J., and Ma, K. K. (2018). A multi-order derivative feature-based quality assessment model for light field image. *J. Visual Commun. Image Represent.* 57, 212–217. doi: 10.1016/j.jvcir.2018.11.005

Viola, I., and Ebrahimi, T. (2018). "VALID: Visual quality assessment for light field images dataset," in *2018 Tenth International Conference on Quality of Multimedia Experience* (QoMEX). doi: 10.1109/QoMEX.2018.8463388

Viola, I., Rerábek, M., and Ebrahimi, T. (2017). "Impact of interactivity on the assessment of quality of experience for light field content," in *2017 Ninth International Conference on Quality of Multimedia Experience* (QoMEX). doi: 10.1109/QoMEX.2017.7965636

Wang, Z., and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Process. Lett.* 9, 81–84. doi: 10.1109/97.995823

Wang, Z., and Bovik, A. C. (2006). Modern image quality assessment. *Synthesis Lectures Image Video Multimedia Process.* 2, 1–156. doi: 10.2200/S00010ED1V01Y200508IVM003

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transact. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wang, Z., and Li, Q. (2010). Information content weighting for perceptual image quality assessment. *IEEE Transact. Image Process.* 20, 1185–1198. doi: 10.1109/TIP.2010.2092435

Wang, Z., and Simoncelli, E. P. (2005). Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In Human vision and electronic imaging X. *Int. Soc. Optics Photonics*. 5666, 149–159. doi: 10.1117/12.597306

Wang, Z., Simoncelli, E. P., and Bovik, A. C. (2003). "Multiscale structural similarity for image quality assessment. *Computers* 2, 1398–1402. doi: 10.1109/ACSSC.2003.1292216

Wu, G., Liu, Y., Fang, L., and Chai, T. (2019). Lapepi-net: A Laplacian pyramid EPI structure for learning-based dense light field reconstruction. *arXiv preprint arXiv:*1902.06221.

Xiang, J., Yu, M., Chen, H., Xu, H., Song, Y., and Jiang, G. (2020). "Vblfi: Visualization-based blind light field image quality assessment," in *2020 IEEE International Conference on Multimedia and Expo* (ICME). doi: 10.1109/ICME46284.2020.9102963

Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transact. Image Process.* 23, 684–695. doi: 10.1109/TIP.2013.2293423

Zhang, L., Gu, Z., and Li, H. (2013). "SDSP: A novel saliency detection method by combining simple priors," in *2013 IEEE International Conference on Image Processing*. doi: 10.1109/ICIP.2013.6738036

Zhang, L., Shen, Y., and Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transact. Image Process.* 23, 4270–4281. doi: 10.1109/TIP.2014.2346028

Zhang, L., Zhang, L., and Mou, X. (2010). "RFSIM: A feature based image quality assessment metric using Riesz transforms," in 2*010 IEEE International Conference on Image Processing*. doi: 10.1109/ICIP.2010.5649275

Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transact. Image Process.* 20, 2378–2386. doi: 10.1109/TIP.2011.2109730

Zhao, P., Chen, X., Chung, V., and Li, H. (2021). "Low-complexity deep no-reference light field image quality assessment with discriminative EPI patches focused," in *2021 IEEE International Conference on Consumer Electronics* (ICCE). doi: 10.1109/ICCE50685.2021.9427654

Zhao, S., and Chen, Z. (2017). "Light field image coding via linear approximation prior," in *2017 IEEE International Conference on Image Processing* (ICIP). doi: 10.1109/ICIP.2017.8297146

Zhou, W., Shi, L., Chen, Z., and Zhang, J. (2020). Tensor oriented no-reference light field image quality assessment. *IEEE*

Transact. Image Process. 29, 4070–4084. doi: 10.1109/TIP.2020.2969777

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership