A New Benchmark Database and Objective Metric for Light Field Image Quality Evaluation

Zhengyu Zhang, Shishun Tian, Jinjia Zhou, Luce Morin, Lu Zhang

A New Benchmark Database and Objective Metric for Light Field Image Quality Evaluation

Zhengyu Zhang, Shishun Tian, Jinjia Zhou, Luce Morin, and Lu Zhang

***Abstract*—Light Field Image (LFI) records both angular and spatial information and provides immersive experiences for observers by rendering a scene from multiple perspectives. To cope with the resolution limitations of capture hardware, LFI angular reconstruction and spatial super-resolution are two widely-used methods, but they can also induce some special types of distortions, especially when two methods are adopted in combination. To this end, new challenges have been brought in assessing the quality of these distorted LFIs. In this paper, firstly, we conduct subjective experiments to evaluate the distorted LFI quality and present a novel perceptual quality assessment database with the associated subjective quality scores. Specifically, the proposed database focuses on the distortions introduced by deep learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. Besides, in the case of multiple distortions, the adoption order of two distortions is taken into consideration. Further, our database presents three types of LFIs that suffer from distortions: real-world, dense synthesis, and sparse synthesis. As a result, the quality of distorted LFIs was subjectively assessed by 32 valid observers using the Pairwise Comparison (PC) protocol. Secondly, we develop a novel objective No-Reference (NR) metric for LFI quality evaluation, based on the features extracted from spatial gradients, angular-spatial statistics, and binocular disparity. Finally, a benchmark of the proposed metric and numerous state-of-the-art quality assessment metrics on the proposed database is presented. Experimental results demonstrate the superiority of the proposed metric over most existing metrics in various aspects. The proposed database and metric will be publicly available at https://github.com/ZhengyuZhang96/IETR-LFI.**

***Index Terms*—Light field image, quality assessment database, multiple distortions, pairwise comparison, no-reference metric.**

I. Introduction

THE realization of immersive experience has gradually become attainable with the advent of Apple Vision Pro. To achieve this goal, researchers are delving into a variety of studies related to immersive media [1], [2], [3], [4], [5]. Among them, Light Field Image (LFI) achieves immersive experiences by recording not only the spatial content of a scene, but also the angular discrepancy across multiple viewpoints [6]. In order to facilitate actual use, Levoy and Hanrahan [7] first simplified LFI into a four-parameter biplane model $\boldsymbol{L} = L(u, v, h, w)$, in which (u, v) and (h, w) denote angular and spatial planes, respectively. Based on this model assumption, High Density Camera Array (HDCA) and Time-Sequential Capture (TSC) system were the early mainstream methods for capturing LFIs [8]. Unfortunately, obtaining LFIs with high angular resolution by these means is expensive and laborious. Although the emergence of light field cameras [9] has greatly reduced the acquisition cost, the captured LFIs face a trade-off between angular resolution and spatial resolution due to the limited number of imaging sensors [10]. Instead of directly capturing high-resolution LFIs, angular reconstruction and spatial super-resolution are two feasible methods to increase the LFI resolution in the post-processing stage [11]. However, these methods inevitably lead to some unusual distortions in LFIs that differ from those common in 2D images, ultimately affecting the visual experience of human eyes. To this end, Light Field Image Quality Assessment (LFIQA) plays a crucial role in providing dependable feedback to control the effect of distortions in LFIs.

Typically, LFIQA research can be classified into two categories: subjective experiments and objective metrics. Subjective experiments explore the potential quality relationships between different distorted LFIs and establish standard databases for benchmarking objective metrics. At present, significant efforts have been invested in subjective experiments to study the distortion influence caused by angular reconstruction in LFIs [12], [13]. Besides, some subjective experiments investigated the impact of LFI spatial super-resolution on human visual perception [14]. However, existing subjective experiments and their resulting databases have not thoroughly explored the relationship between human visual perception and LFIs distorted by these two types of generative methods [15], [16]. The limitations are summarized as follows:

1) With the widespread adoption of deep learning technologies, many deep learning-based LFI angular reconstruction and spatial super-resolution methods are proposed and achieve automatic feature learning through backpropagation. As a result, the associated distortions are also more intricate and difficult to assess. Nevertheless, existing databases do not adequately cover this aspect.

2) Current researches only investigate the impact of angular reconstruction or spatial super-resolution in LFIs individually. However, in practice, these two methods may be implemented

This work was supported in part by the National Natural Science Foundation of China under grants 62101344, in part by the Natural Science Foundation of Guangdong Province, China under grants 2022A1515010159, and in part by the China Scholarship Council.

Zhengyu Zhang is with the School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, China, and also with the Univ Rennes, INSA Rennes, CNRS, IETR-UMR 6164, F-35000 Rennes, France. (e-mail: zhengyuzhang23@outlook.com)

Shishun Tian is with the Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, China. (e-mail: stian@szu.edu.cn)

Jinjia Zhou is with the Graduate School of Science and Engineering, Hosei University, Koganei 184-8584, Japan. (e-mail: zhou@hosei.ac.jp)

Luce Morin and Lu Zhang are with the Univ Rennes, INSA Rennes, CNRS, IETR-UMR 6164, F-35000 Rennes, France. (e-mail: luce.morin@insa-rennes.fr; lu.ge@insa-rennes.fr)

Corresponding author: Shishun Tian

in combination and induce multiple distortions. Further, the adoption order of angular reconstruction and spatial super-resolution may also affect the degradation degree of LFI quality. These all deserve further investigation.

3) Studying how different types of LFIs are affected by different types of distortions can help to better understand the Human Visual System (HVS). In other words, the limited consideration of the diversity of LFI types restricts the comprehension of HVS to some extent. However, most existing databases only consider a single type of LFIs that suffers from different distortions.

4) The absence of exploration in the above three aspects results in a lack of standard databases and corresponding objective metrics. In addition, there is a pressing need to establish a benchmark of existing objective metrics to facilitate a comprehensive effectiveness analysis for practical applications.

To fill the aforementioned gaps in the research of LFIQA, in this paper, we conduct subjective experiments by employing the Pairwise Comparison (PC) protocol for quality scoring, and present a novel LFIQA database with the associated subjective quality scores. Different from previous databases, our database focuses on LFIs that are only disturbed by deep learning-based LFI angular reconstruction and spatial super-resolution methods. A total of 12 types of distortions are considered, including both individual and multiple distortions. Note that when multiple distortions are present, the adoption order of two distortions is also taken into account. In addition, we select three types of source LFIs, including real-world, dense synthesis, and sparse synthesis, and evaluate their perceptual quality after suffering different distortions. Based on the proposed database, we then design a new objective metric, namely SAB, which is capable of accurately predicting LFI quality without reference information. Considering the timeliness requirement in some practical scenarios, we further develop a lightweight version of the proposed metric (SAB-light), which significantly reduces the time complexity while slightly sacrificing the predictive accuracy. Finally, quantitative experiments on the proposed database are conducted, and a relatively comprehensive benchmark of the proposed metric and other state-of-the-art objective metrics for quality assessment is provided. Experimental results demonstrate that the proposed metric performs better than most state-of-the-art metrics. In summary, the main contributions of this work are listed as follows:

1) We perform subjective experiments to investigate the influence of LFI distortions induced by deep learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. Further, three different types of LFIs are considered to ensure the diversity of image types. As a result, a novel perceptual quality assessment database with subjective scores is established.

2) We propose a novel objective No-Reference (NR) metric to evaluate the quality of distorted LFIs by exploring quality-aware features from Spatial gradient, Angular-spatial statistics, and Binocular disparity, which is abbreviated as SAB. To meet the real-time needs of some practical purposes, a lightweight version of the proposed SAB metric is further presented.

3) We conduct quantitative experiments to build a benchmark of the two proposed metrics and a variety of existing LFIQA metrics on the proposed database. Experimental results demonstrate the superiority of the proposed SAB metric and its lightweight version over other state-of-the-art metrics.

The rest of this paper is structured as follows. Section II introduces the related works. Section III describes the subjective experiments and the established database. Section IV presents the proposed objective metric. A benchmark of objective metrics on our database is provided in Section V. Finally, conclusions are drawn in Section VI.

II. Related works

A. Benchmark Databases for LFIQA

By collecting opinion scores in subjective experiments, several publicly available benchmark databases for LFIQA have been developed, as summarized in TABLE I. In addition, several special databases related to the quality evaluation of LFIs have also been proposed recently, such as KULF-TT53 (display-specific turntable-based LFIQA database) [17] and WLFI (stitched wide field of view LFIQA database) [18]. In the following, we will introduce the databases in TABLE I one by one in chronological order.

Paudyal *et al.* [19] established the first quality assessment database for LFIs, namely the SMART database. The database includes 16 reference LFIs captured by the second-generation light field camera (Lytro Illum) from real-world scenes. All reference LFIs are subject to 4 types of distortions related to image compression: JPEG, JPEG2k, HEVC, and Sparse Set Disparity Coding (SSDC), and each distortion type has 4 distortion levels. Thus, there are 256 distorted LFIs in total. The main disadvantage of the SMART database is that only compression-related distortions are evaluated.

Adhikarla *et al.* [12] proposed a LFIQA database (MPI-LFA database), which consists of 14 pristine LFIs and 336 distorted LFIs spanning 6 distortion types with 6 distortion levels. The 6 distortions of this database involve 3 categories: HEVC for compression, Gaussian blur for acquisition or display, and Quantized Depth maps (DQ), OPTical flow estimation (OPT), LINEAR interpolation (LINEAR), Nearest Neighbor interpolation (NN) for angular reconstruction. However, since the MPI-LFA database captures reference content using the TSC system, all included LFIs are 3D with only one angular dimension instead of 4D with two angular dimensions, which limits a comprehensive study on the human visual perception of LFIs.

Viola and Ebrahimi [20] presented the VALID database to subjectively evaluate the quality of distorted LFIs. The database contains 5 real-world reference LFIs [21], based on which 100 distorted LFIs are generated by 5 distortion types with 4 distortion levels. All distortions are compression-related, including 2 off-the-shelf compression solutions (HEVC and VP9) and 3 relatively state-of-the-art LFI compression algorithms ([22], [23], and [24]). Insufficient content coverage is the main shortcoming of the VALID database.

The Win5-LID database proposed by Shi *et al.* [13] collects subjective quality scores when delivering a windowed 5 degree

TABLE I
SUMMARY OF EXISTING LFIQA DATABASES. DISTORTIONS INDUCED BY DEEP LEARNING-BASED METHODS ARE MARKED WITH *. NOTE THAT A → B REFERS TO THE MULTI DISTORTIONS INDUCED BY ADOPTING THE A METHOD AND THEN THE B METHOD.

Name	Size	Distortion type	Distortion category	Mul. Dis.	No. SRCs	Ang. Res.	Spa. Res.	No. Obs.
SMART	256	JPEG, JPEG2k, HEVC, SSDC	Compression	No	16 (real-world)	15×15	434×625	19
MPI-LFA	336	HEVC	Compression	No	14 (5 real-world and 9 dense synthesis)	1×101	960×720	40
		Gaussian blur	Acquisition or Display					
		DQ, OPT, LN, NN	Angular reconstruction					
VALID	100	HEVC, VP9	Compression	No	5 (real-world)	15×15	434×625	28
		[22], [23], [24]	LFI compression					
Win5-LID	220	HEVC, JPEG2k	Compression	No	10 (6 real-world and 4 dense synthesis)	9×9	434×625 or 512×512	23
		LN, NN	Angular reconstruction					
		EPICNN*, LF-SYN*	LFI angular reconstruction					
NBU-LF1.0	210	NN, BI	Angular reconstruction	No	14 (8 real-world and 6 dense synthesis)	9×9	434×625 or 512×512	22
		VDSR*	Spatial super-resolution					
		EPICNN*, Zhang	LFI angular reconstruction					
SHU	240	JPEG, JPEG2k	Compression	No	8 (real-world)	15×15	434×625	20
		Gaussian blur, White noise, Motion blur	Acquisition or Display					
LFDD	480	JPEG, JPEG2k, BPG, VP9, AV1, AVC, HEVC	Compression	No	8 (dense synthesis)	9×9	512×512	16
		Gaussian blur, Impulse noise, Barrel, Pincushion, Unsharp mask	Acquisition or Display					
Bakir's	160	HEVC, JEM	Compression	No	10 (real-world)	15×15	434×625	18
		LF-SYN*, [22]	LFI compression					
Ours	120	HDDRNet*, LFASR*	LFI angular reconstruction	Yes	10 (4 real-world, 3 dense synthesis and 3 sparse synthesis)	9×9	434×625 or 512×512	32
		LF-IINet*, DistgSSR*	LFI spatial super-resolution					
		HDDRNet* → LF-IINet*, HDDRNet* → DistgSSR*, LFASR* → LF-IINet*, LFASR* → DistgSSR*	LFI angular reconstruction → LFI spatial super-resolution					
		LF-IINet* → HDDRNet*, LF-IINet* → LFASR*, DistgSSR* → HDDRNet*, DistgSSR* → LFASR*	LFI spatial super-resolution → LFI angular reconstruction					

of freedom experience to observers. There are 10 reference LFIs, of which 6 are real-world [21] and the rest are dense synthetic [25]. The Win5-LID database consists of 4 distortion types (HEVC, JPEG2k, LN, and NN) with 5 distortion levels, and 2 deep learning-based distortion types (EPICNN [26] and LF-SYN [27]) with only 1 distortion level. As a result, the database has 220 distorted LFIs. In this database, the distortions caused by super-resolution methods are not taken into account.

Huang *et al.* [14] proposed a reconstruction distortion oriented LFIQA database (NBU-LF1.0 database), in which 5 types of distortions associated with reconstruction are considered: NN and BI for angular reconstruction, VDSR [28] for spatial super-resolution, EPICNN [26] and Zhang [29] for LFI angular reconstruction. The NBU-LF1.0 database has 210 distorted LFIs, which are generated by 8 real-world reference LFIs [21] and 6 dense synthetic reference LFIs [25]. In the NBU-LF1.0 database, the distortions obtained by deep learning-based algorithms are considered somewhat "old-fashioned".

Shan *et al.* [30] proposed the SHU database for the evaluation of LFI quality, consisting of 240 distorted LFIs and 8 pristine real-world LFIs [21]. This database considers 5 widely-used traditional distortion types, of which JPEG and JPEG2k are compression-related, while Gaussian blur, White noise, and Motion blur are related to acquisition or display. Each distortion type has 6 distortion levels. For the SHU database, the insufficient diversity of image types and the lack of distortions caused by more advanced algorithms restrict its broad applicability.

The LFDD database is a dense synthesis LFIQA database presented by Zizien and Fliegel [31], which contains 480 distorted LFIs disturbed from 8 reference LFIs [25]. Similar to the SHU database, the LFDD database focuses on distortion types that are caused by compression techniques (JPEG, JPEG2k, BPG, VP9, AV1, AVC, and HEVC) and acquisition or display processes (Gaussian blur, Impulse noise, Barrel, Pincushion, and Unsharp mask). Each distortion type contains 5 distortion levels. Despite including more distorted LFIs, the LFDD database still faces the same limitations as the SHU database, *i.e.*, a limited range of image types and deep learning-based methods for obtaining distortions are not considered.

Bakir *et al.* [32] proposed a LFIQA database to evaluate the quality of LFIs affected by compression artifacts. The database involves 10 real-world reference LFIs. In addition, 160 distorted LFIs subject to 4 distortion types and 4 distortion levels are considered, including 2 compression solutions (HEVC and JEM) and 2 LFI compression algorithms (LF-SYN [27] and [22]). The extensive utilization of this database is restricted since it only considers compression artifacts.

In addition to the limitations specific to each database

mentioned above, none of the existing databases cover the hypothesis of multiple distortions occurring within a single LFI. In fact, due to the low resolution of source LFIs, a combination of angular reconstruction and spatial super-resolution is a commonly-employed strategy in LFI post-processing. To fill this gap, our proposed database focuses on evaluating the quality of LFIs disturbed by multiple distortions, especially the distorted LFIs produced by the alternating use of state-of-the-art deep learning-based LFI angular reconstruction and spatial super-resolution methods. Moreover, three different types of LFIs are considered in our database to enhance the diversity of test images.

B. Objective Metrics for LFIQA

On the basis of the establishment of benchmark databases, a substantial quantity of objective LFIQA metrics have been designed to approximate the subjective quality scores [33]. Currently, state-of-the-art objective LFIQA metrics can be roughly divided into three categories: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR).

The FR LFIQA metrics evaluate the quality of distorted LFIs when their reference versions are provided. Tian *et al.* assessed the LFI quality by exploring multi-order derivative information in [34] and radial symmetry information in [35], respectively. Fang *et al.* [36] utilized the gradient magnitude information for quality assessment. Min *et al.* [37] estimated the quality of LFIs with multi-view structure matching and near-edge mean square error. Meng *et al.* designed quality-aware features based on Gaussian operator and structural similarity in [38], and gradient magnitude and phase congruency in [39], respectively. Huang *et al.* first extracted the multi-scale information in the contourlet domain to evaluate the LFI quality in [40], and further improved the predictive accuracy by capturing 3D-Gabor information in [41]. Ma *et al.* [42] measured the quality of distorted LFIs using natural scene statistics and structural information. Our previous work EDDMF [43] combined a Convolutional Neural Network-based (CNN-based) model and a local discrepancy extraction module for FR LFIQA.

The RR LFIQA metrics utilize only partial or indirect reference information to evaluate the degradation of LFI quality. Paudyal *et al.* [44] calculated the structural similarity between the depth maps of reference and distorted LFIs as predicted quality. Xiang *et al.* [45] performed quality evaluation by calculating the similarity between the generated pseudo LFIs and distorted LFIs in the wavelet domain.

The NR LFIQA metrics evaluate the LFI quality without accessing any reference information. Shi *et al.* extracted naturalness and structural features in [46] and further measured the angular consistency in [8]. Based on Tucker decomposition, Zhou *et al.* [47] predicted the LFI quality with global naturalness, local frequency, and structural similarity features. Xiang *et al.* also adopted Tucker decomposition to handle the high-dimensional LFIs to facilitate LFIQA in [48]. Further, Xiang *et al.* [49] proposed to combine naturalness, energy, and angular consistency features for LFI quality evaluation. In another work [50], Xiang *et al.* estimated the quality of distorted LFIs in the 4D frequency domain. Pan *et al.* [51] focused on the utilization of sharpness, information distribution, and singular value features in LFIQA. Chai *et al.* [52] considered texture and structure information. Our previous work SATV-BLiF [53] exploited textural variations in LFIs for quality assessment. The aforementioned NR metrics are all based on handcrafted features, while several state-of-the-art works have adopted deep learning-based technologies in NR LFIQA. Qu *et al.* [54] developed a CNN-based model using depth-wise separable convolutions. Alamgeer *et al.* [55] extracted deep features with the input in the frequency domain. Zhao *et al.* [56] proposed to evaluate the LFI quality by evaluating the quality of LFI patches. Our previous work DeeBLiF [57] followed the 2D patch-wise idea and designed CNN-based models for extracting angular and spatial information separately. Further, the 3D Pseudo Video Sequence (PVS) representation of LFIs was exploited in [58], and a combination of CNNs and Transformer was developed in [59].

Fig. 1. Illustration of the central view of SRCs in the proposed database.

Although numerous objective LFIQA metrics have been devised over the years, their effectiveness in assessing multiple distortions remains unclear due to the lack of corresponding standard databases. This fact motivates us to build a benchmark of existing objective metrics on our proposed database. In addition, we notice that most existing LFIQA metrics lack a comprehensive analysis of LFIs to design efficient quality-aware features, thereby suffering from either limited predictive accuracy or high computational complexity. In contrast, the handcrafted feature-based NR metric proposed in this paper extracts features from spatial gradients, angular-spatial statistics, and binocular disparity, providing an overall representation of LFI quality degradation.

III. Proposed Database

A. Source Sequences

The selection of Source Sequences (SRCs) plays a crucial role in establishing a high-quality database. In order to cover a variety of practical application scenarios and enhance the accuracy and generalization of subjective test results, it is

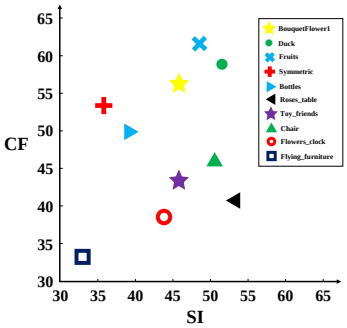


Fig. 2. Distribution of the SI and CF values of the selected SRCs.

necessary to ensure the diversity of SRCs. In this work, we consider the diversity of SRCs from two aspects: type diversity and content diversity. For the type diversity, three types of SRCs are involved, including real-world, dense synthesis, and sparse synthesis. For content diversity, several attributes are considered in the selection of SRCs, such as Spatial Information (SI), Colorfulness (CF), contrast, and brightness.

Specifically, the SRCs of the proposed database consist of 10 reference LFIs, and we provide an illustration of the central view of the selected SRCs in Fig. 1. Among them, $BouquetFlower1$, $Duck$, $Fruits$, and $Symmetric$ are captured in real-world scenarios [60], $Bottles$, $Roses\_table$, and $Toy\_friends$ are dense synthesis [61], while $Chair$, $Flowers\_clock$, and $Flying\_furniture$ are sparse synthesis [61]. The source content has a rich range of coverage, including but not limited to vibrant/dull colors, small/large objects, distant/close-up views, and indoor/outdoor scenes. Further, following [13], [14], we provide the SI and CF values of our selected SRCs to demonstrate the content diversity, as shown in Fig. 2. Note that the SI and CF values are calculated according to the ITU-T Rec. P.910 [62] and Hasler's method [63], respectively. Although each real-world reference LFI in [21] provides 15×15 angular views, we only keep the central 9×9 angular views to avoid introducing edge views with inherent noise. As a result, the resolution of real-world reference LFIs and other synthesis reference LFIs are 9×9×434×625 and 9×9×512×512, respectively.

B. Hypothetical Reference Circuits

To investigate the impact of distortions introduced by deep learning-based LFI angular reconstruction and spatial super-resolution methods individually and multiplely, we define 12 Hypothetical Reference Circuits (HRCs) in the proposed database. Given a source LFI, we introduce distortions by halving its angular/spatial resolution and then restoring its original resolution using angular reconstruction/spatial super-resolution methods. Specifically, two angular reconstruction methods (HDDRNet [64] and LFASR [65]) and two spatial super-resolution methods (LF-IINet [66] and DistgSSR [67]) are used. These methods have different technical focuses based on deep learning, and their descriptions are as follows:

HDDRNet proposed by Meng *et al.* [64] is a two-stage scheme using a deep convolutional network with dense connections for LFI angular reconstruction. By progressively performing spatial-angular restoration and high-frequency texture refinement, the model utilizes the geometry information encoded in angular views to reconstruct LFIs.

LFASR proposed by Jin *et al.* [65] is a totally end-to-end LFI angular reconstruction approach based on CNNs. This approach employs a depth estimation module to capture intrinsic scene geometry information, and utilizes warping and blending modules to produce novel angular views.

LF-IINet proposed by Liu *et al.* [66] develops an intra-inter view interaction network to achieve LFI spatial super-resolution. The two parallel branches of the model extract global inter-view information and exploit correlations among intra-view information, respectively, and further interact with each other for better representation learning.

DistgSSR proposed by Wang *et al.* [67] employs a generic mechanism to disentangle LFIs from different dimensions. It is achieved through the utilization of a set of domain-specific convolutions. Subsequently, these disentangled features are combined and applied to perform spatial super-resolution for LFIs.

In our database, we employ these four methods on source LFIs individually and multiplely, resulting in 12 HRCs that involve 4 individual distortions (HDDRNet, LFASR, LF-IINet, and DistgSSR) and 8 multiple distortions (HDDRNet → LF-IINet, HDDRNet → DistgSSR, LFASR → LF-IINet, LFASR → DistgSSR, LF-IINet → HDDRNet, LF-IINet → LFASR, DistgSSR → HDDRNet, and DistgSSR → LFASR), as summarized in TABLE I. Here, the methods on the left and right of the arrow "→" represent the methods used successively. Further, using the two methods in different orders during the generation of multiple distortions are considered as different distortion types. Unlike most existing LFIQA databases, the proposed database no longer considers the severity levels of each distortion type, but focuses more on the relationship between different distortion types. This is mainly because we have observed that most objective metrics are able to distinguish different severity levels within the same distortion type [8], [39], [48]. As a result, our database consists of 120 distorted LFIs.

C. Methodology of Subjective Experiments

According to the ITU-T Rec. P.910 [62], there are some recommended testing protocols for subjective quality scoring, mainly including Absolute Category Rating (ACR), Degradation Category Rating (DCR), and Pairwise Comparison (PC). In the ACR and DCR protocols, the observers are required to rate the images directly based on the given scoring table. However, as concluded in [12], these protocols are considered to be insensitive to subtle but noticeable degradation of LFI quality, which always results in observer confusion regarding the selection of the desired rating. Instead, the PC protocol simplifies decision-making by comparing stimuli in pairs and giving relative preferences rather than absolute ratings [68]. Therefore, we adopt the PC protocol in our subjective test.

In each trial, a pair of LFIs is presented horizontally side by side on the same screen to the observers, who are then required to make a two-alternative-forced-choice about the one they

perceive as higher quality. In order to obtain a full PC count matrix, all pairs of LFIs should be displayed in both possible orders ($e.g.$, AB, BA) following the PC implementation in [62]. The adopted PC protocol places higher requirements on the display because two LFIs need to be displayed at the same time. Unfortunately, the current coverage of light field displays is very limited, and it is difficult to display all LFI information on a traditional display at once. We thus develop a Graphical User Interface (GUI) that only displays one view of each LFI, while observers can change their viewing perspective by moving the mouse [13], which simulates the process of eye movement when observing a LFI. Such an interactive method provides observers with greater freedom to perceive the angular and spatial differences between two LFIs and thereby make accurate choices. Note that there is no time limit for observer interactions, but once a selection is made it cannot be changed. Before the formal test, we offer 10 pre-training pairs from other scenarios to familiarize observers with the usage of the interactive GUI.

The proposed database contains 120 distorted LFIs derived from 10 reference LFIs, $i.e.$, each LFI has 12 distorted versions. In the PC protocol, the reference version is treated as one stimuli and hidden in all pairs, thus there are 156 testing pairs for each reference LFI. Since the fixed display order may lead to a decrease in the observer's accuracy in judging latter pairs, we shuffle the display order for each observer. However, according to the three different types of LFIs in our database, we divide all testing pairs into three parts and display them separately. This is based on the consideration that repeatedly switching viewing different types of LFIs may affect the observer's perception. To minimize the negative impact of visual fatigue, observers can take breaks at any time during the test. Nonetheless, observers still need to take a mandatory 5-minute break after completing each part. Generally, the total duration required for each observer to complete all tests will not exceed 90 minutes.

The viewing conditions in our experiments are set based on the ITU-T Rec. P.910 [62] to ensure that the results are reliable. In our subjective test, all LFIs are presented at their original spatial resolution without any reshaping or scaling operations. Moreover, we strictly follow the principle in subjective experiments: each pixel on the image is displayed by a single pixel on the monitor. Therefore, the used display device is the LENOVO T2424PA monitor with a resolution of 1920×1080, which ensures that a pair of LFIs can be displayed horizontally side by side. The region outside the testing pairs is filled with a constant gray color.

Finally, a total of 32 valid observers participated in our subjective experiments, including 21 males and 11 females, ranging in age from 21 to 29. None of the observers had any prior knowledge in the field of image quality evaluation. The normal (or corrected-to-normal) visual acuity and normal color vision of all observers were examined using the Snellen and Isihara charts, respectively.

D. Processing and Analysis of Subjective Scores

After completing the PC test, the ranking results of all observers are typically arranged in a count matrix, and the final step is to convert this matrix into a scalar quality score for each distorted LFI. We perform this conversion according to [69]. Specifically, let $\boldsymbol{C}$ denote a D-by-D count matrix, which can be described as:

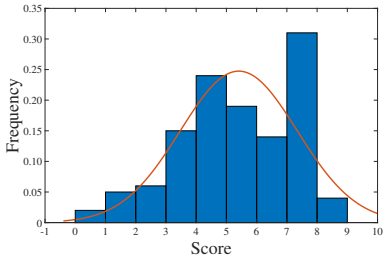


Fig. 3. Distribution of all quality scores in our database.

$$\boldsymbol{C} = \begin{bmatrix} 0 & c_{12} & \dots & c_{1D} \\ c_{21} & 0 & \dots & c_{2D} \\ \dots & \dots & \dots & \dots \\ c_{D1} & c_{D2} & \dots & 0 \end{bmatrix} \tag{1}$$

where D=13 in our database, representing the sum of one reference and all distorted versions in a single scene. The element c_{ij} denotes the count of cases that the quality of distorted LFI i was selected as better than that of distorted LFI j. Then the probability matrix $\boldsymbol{P}$ can be empirically estimated as:

$$\boldsymbol{P} = \begin{bmatrix} 0 & \hat{p}_{12} & \dots & \hat{p}_{1D} \\ \hat{p}_{21} & 0 & \dots & \hat{p}_{2D} \\ \dots & \dots & \dots & \dots \\ \hat{p}_{D1} & \hat{p}_{D2} & \dots & 0 \end{bmatrix} \tag{2}$$

where the element $\hat{p}_{ij}$ denotes the probability that the quality of distorted LFI i was selected as better than that of distorted LFI j, which can be described as:

$$\hat{p}_{ij} = \frac{c_{ij}}{c_{ij} + c_{ji}}, \quad i \neq j \tag{3}$$

Then, we adopt the Thurstone Case V model [70] as the observer model, in which the perceptual quality of distorted LFI i is assumed as a normal distribution g_i:

$$g_i \sim N(q_i, \sigma_i) \tag{4}$$

where the true quality score q_i is assumed to be the distribution mean, and the standard deviation σ_i denotes the combination of inter-observer and intra-observer variances. Thus, the distance between two distorted versions i and j is also considered as a normal distribution g_{ij}:

$$g_{ij} \sim N(q_{ij}, \sigma_{ij}) \tag{5}$$

where $g_{ij} = g_i - g_j$, $q_{ij} = q_i - q_j$, and $\sigma_{ij} = \sigma_i^2 + \sigma_j^2$.

In addition, the probability of choosing the quality of distorted LFI i over that of distorted LFI j can be calculated using the cumulative normal distribution based on g_{ij}:

$$\hat{p}_{ij} \approx \boldsymbol{P}(g_i > g_j) = \boldsymbol{P}(g_{ij} > 0) = \Phi(q_{ij}, \sigma_{ij}) \tag{6}$$

where $\hat{p}_{ij}$ is approximated as $\boldsymbol{P}(g_i > g_j)$. As shown in Eq. (6), given a $\hat{p}_{ij}$ and a σ_{ij}, we can obtain q_{ij}, the distance of true quality scores between distorted LFIs i and j, using the inverse of the cumulative normal distribution function $\Phi(\cdot)$.

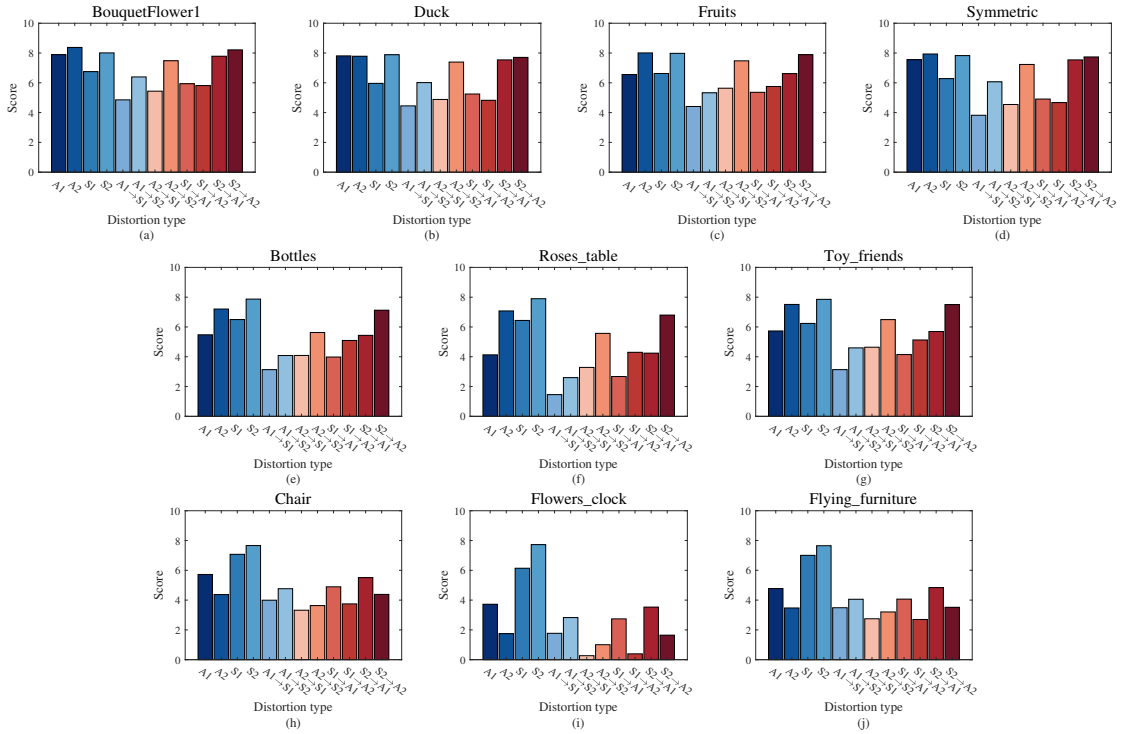


Fig. 4. Quality score summary of 12 involved distortions for different scenes. (a) BouquetFlower1; (b) Duck; (c) Fruits; (d) Symmetric; (e) Bottles; (f) Roses_table; (g) Toy_friends; (h) Chair; (i) Flowers_clock; (j) Flying_furniture.

According to the assumption of the Thurstone Case V model [70], σ_{ij} is a constant, whose selection generally satisfies the case that a probability of 0.75 corresponds to a distance score of 1 Just-Objectionable-Difference (JOD) unit. As a result, σ_{ij} is set to 1.4826 [69].

Based on the above assumptions, the probability that the quality of distorted LFI i was selected over that of distorted LFI j in c_{ij} trials of the total trial number of $n_{ij} = c_{ij} + c_{ji}$ can be described by the binomial distribution:

$$\boldsymbol{P}(\hat{q}_{ij}|c_{ij},n_{ij}) = \binom{n_{ij}}{c_{ij}} \Phi(\hat{q}_{ij},\sigma_{ij})^{c_{ij}} (1-\Phi(\hat{q}_{ij},\sigma_{ij}))^{n_{ij}-c_{ij}} \tag{7}$$

and the estimated quality score $\hat{q}_2$,...,$\hat{q}_n$ can be calculated using the maximum likelihood estimation:

$$\arg\max_{\hat{q}_2,\ldots,\hat{q}_n} \prod_{i,j} \ln(\boldsymbol{P}(\hat{q}_{ij}|c_{ij},n_{ij})) \tag{8}$$

Here, due to the use of the natural logarithm function, most of the obtained quality scores are negative, ranging in [-7.7319, 0.3751]. A LFI with a negative score indicates that it has worse visual quality than its reference, while a LFI with a positive score means that its quality is enhanced. To avoid any ambiguity and confusion caused by negative scores, we directly add an offset value of 8 to all scores to ensure they are positive. Thus, the final quality scores of our proposed database range in [0.2681, 8.3751].

Then, we conduct statistical analyses on the obtained subjective quality scores. To ensure the clarity of expression and simplicity of diagrams, in the following, we will refer to the two involved LFI angular reconstruction methods, HDDRNet and LFASR, as A1 and A2, respectively. Similarly, we will refer to the two involved LFI spatial super-resolution methods, LF-IINet and DistgSSR, as S1 and S2, respectively.

1) Distribution of all quality scores. Fig. 3 presents the quality score distribution of distorted LFIs in our database. It can be observed that the quality scores of distorted LFIs have a wide and reasonable distribution, further demonstrating the rationality of the distortion type selection and subjective experiment design in our database.

2) Quality score summary of different scenes. Fig. 4 provides the quality score summary of 12 involved distortion types for different scenes, where (a)-(d) are real-world scenes, (e)-(g) and (h)-(j) are dense and sparse synthetic scenes, respectively. The figure shows that the same distortion has different effects on LFIs in different scenarios, resulting in varying degrees of visual quality degradation. These results are consistent with the conclusion obtained by Adhikarla *et al.* in [12], that is, the LFI quality is scene-dependent. Nonetheless, we can find that when suffering from distortions, real-world scenes tend to have better visual perception than computer-synthesized scenes. A possible explanation is that real-world scenes contain more complex textures and content, which presents favorable conditions for generative tasks. Further, we can also see that the visual effect of sparse synthetic scenes is generally poorer than that of dense synthetic scenes. This phenomenon may be attributed to the fact that existing LFI angular reconstruction methods (including HDDRNet and LFASR used in our database) mainly focus on LFIs with narrow disparities, and have limited reconstruction performance for LFIs with large disparities.

3) Relationship between the impact of individual and multiple distortions. In Fig. 5, we summarize the quality score

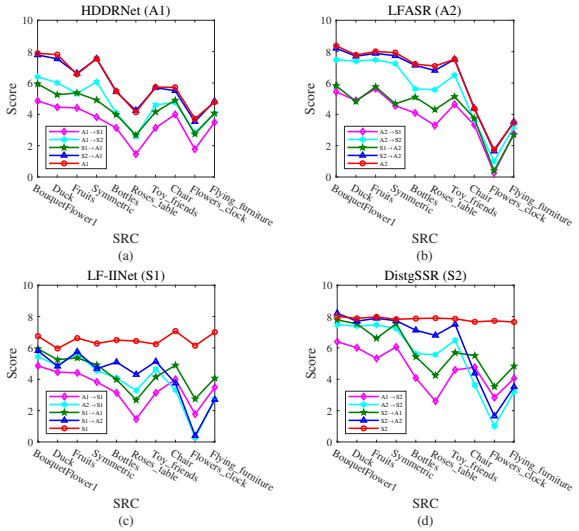


Fig. 5. Quality score distribution with respect to different generative methods used in our database. (a) HDDRNet; (b) LFASR; (c) LF-IINet; (d) DistgSSR.

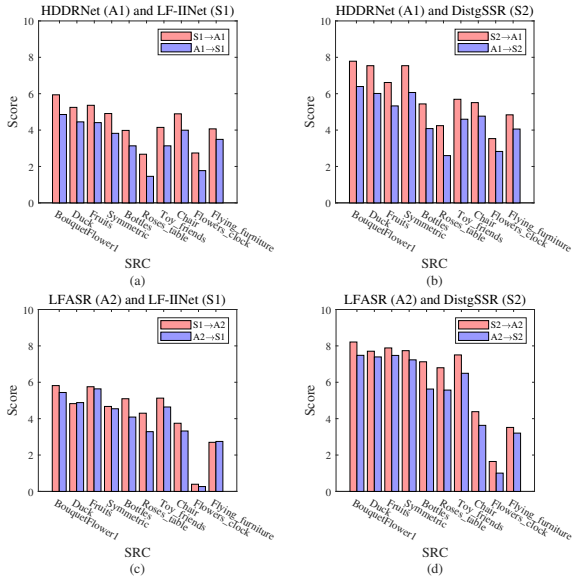


Fig. 6. Quality score distribution with respect to multiple distortions induced by two different methods. (a) HDDRNet and LF-IINet; (b) HDDRNet and DistgSSR; (c) LFASR and LF-IINet; (d) LFASR and DistgSSR.

distribution with respect to different generative methods used in our database, where (a)-(d) are related to HDDRNet (A1), LFASR (A2), LF-IINet (S1), and DistgSSR (S2) methods, respectively. For the distortions induced by the HDDRNet method (Fig. 5(a)), the individual distortion A1 has similar visual quality scores as the multiple distortion S2→A1, while causing significantly less damage to visual perception compared to other three multiple distortions (A1→S1, A1→S2, and S1→A1). Similar conclusions can be drawn in the case of LFASR (Fig. 5(b)). However, when the two LFI spatial super-resolution methods LF-IINet (Fig. 5(c)) and DistgSSR (Fig. 5(d)) are used, the individual distortion consistently exhibits higher visual quality scores than all multiple distortions. Overall, these score distributions are consistent with our expectation, $i.e.$, whenever LFIs suffer an additional distortion, its visual quality should decrease accordingly.

4) Impact of different adoption orders of two distortions in multiple distortions. The combined use of LFI angular reconstruction and spatial super-resolution methods introduces multiple distortions in LFIs, in which the impact of different adoption orders of two distortions is investigated in this work. As shown in Fig. 6, (a)-(d) present the quality score distribution with respect to multiple distortions induced by two different methods respectively: (a) HDDRNet and LF-IINet; (b) HDDRNet and DistgSSR; (c) LFASR and LF-IINet; (d) LFASR and DistgSSR. We can clearly see that adopting a LFI spatial super-resolution method followed by a LFI angular reconstruction method generally obtains a better visual effect than adopting the reverse order. We believe it is mainly because LFIs contain richer spatial information than angular information (or have higher spatial resolution than angular resolution). This inherent attribute enables the LFI spatial super-resolution task to utilize more spatial information to expand spatial resolution and achieve better visual perception, which benefits the subsequent LFI angular reconstruction task. On the contrary, if the LFI angular reconstruction method is adopted first, it would lead to limited reconstruction performance and poor visual effects due to inadequate information. The above findings provide certain guidance for practical applications involving multiple distortions.

IV. Proposed Metric

The high-dimensional characteristics of LFIs lead to intricate visual perception when LFIs are subject to distortions. Therefore, we need to comprehensively evaluate the visual effect of distorted LFIs from multiple aspects. In this work, we present a novel objective NR quality assessment metric for LFIs. As shown in Fig. 7, the proposed metric exploits the quality-aware features of distorted LFIs from three aspects: spatial gradient, angular-spatial statistics, and binocular disparity. Finally, all extracted features are combined and converted into quantified quality scores using the Support Vector Regression (SVR).

A. Spatial Gradient Feature Extraction

The spatial information in LFIs has similar characteristics to that in traditional 2D images, and it tends to suffer from blur-like distortions caused by spatial-related tasks, such as spatial super-resolution. These kinds of distortions often modify the local anisotropy and smooth the edge sharpness of images. Considering the well-documented effectiveness of image gradients in assessing blur-like distortions [71], [72], we utilize the Relative Gradient Orientation (RGO) and Relative Gradient Magnitude (RGM) to measure the local anisotropy and edge sharpness, respectively, in order to evaluate the LFI quality.

Specifically, for a given LFI $\boldsymbol{L} = L(u, v, h, w)$, we first convert it into a stack of grayscale Sub-Aperture Images (SAIs) $\boldsymbol{S} = \{S_{u,v}(h, w)\}$ focusing on spatial information. For simplicity, we denote each single SAI as $S(h, w)$ without u

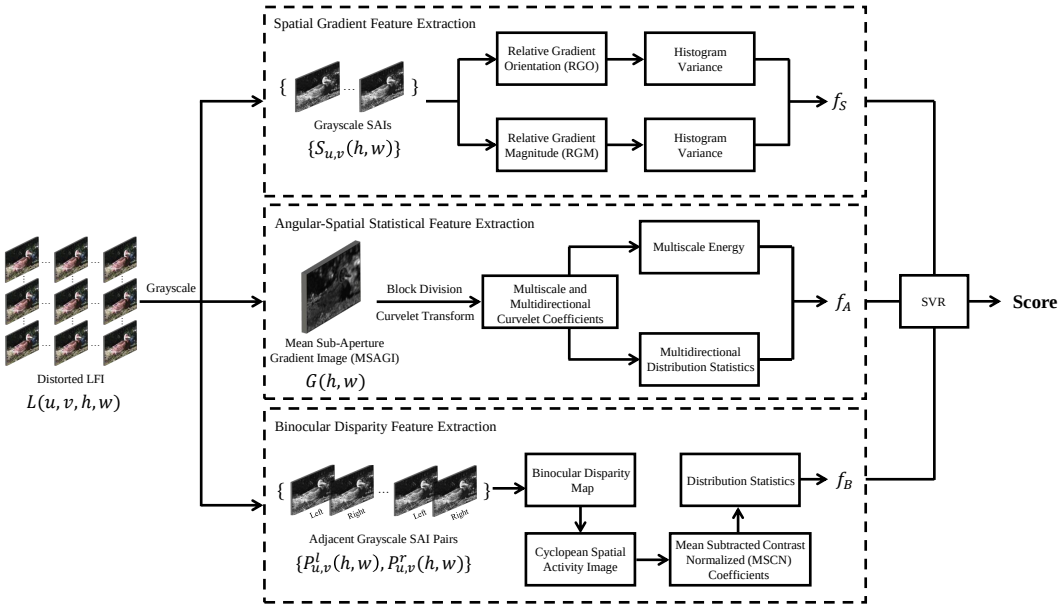


Fig. 7. Block diagram of the proposed metric.

and v in this subsection. Firstly, the Gradient Orientation (GO) of $S(h,w)$ is calculated as:

$$\angle\nabla S(h,w) = \arctan(\frac{S_y(h,w)}{S_x(h,w)}) \tag{9}$$

where $S_x(h,w)$ and $S_y(h,w)$ represent the approximate values of the directional derivatives in the horizontal and vertical directions, respectively. Similarly, the local average GO of $S(h,w)$ is calculated as:

$$\angle\nabla \bar{S}(h,w) = \arctan(\frac{\bar{S}_y(h,w)}{\bar{S}_x(h,w)}) \tag{10}$$

where the horizontal and vertical average directional derivatives, $\bar{S}_x(h,w)$ and $\bar{S}_y(h,w)$, are calculated on the 3×3 neighboring pixels (*i.e.*, M=N=3), with a set of relative coordinate offsets Ω={(-1,-1), (-1,0), (-1,1), (0,-1), (0,0), (0,1), (1,-1), (1,0), (1,1)}. Accordingly, $\bar{S}_x(h,w)$ and $\bar{S}_y(h,w)$ are described as:

$$\bar{S}_x(h,w) = \frac{1}{M \times N} \sum_{(m,n)\in\Omega} S_x(h-m, w-n) \tag{11}$$

$$\bar{S}_y(h,w) = \frac{1}{M \times N} \sum_{(m,n)\in\Omega} S_y(h-m, w-n) \tag{12}$$

The RGO and RGM of $S(h,w)$ are then calculated as:

$$\angle\nabla S(h,w)_{RGO} = \angle\nabla S(h,w) - \angle\nabla \bar{S}(h,w) \tag{13}$$

$$|\nabla S(h,w)|_{RGM} = \sqrt{\begin{array}{r}(S_x(h,w) - \bar{S}_x(h,w))^2 \\ +(S_y(h,w) - \bar{S}_y(h,w))^2\end{array}} \tag{14}$$

Subsequently, inspired by Ruderman's work [73], we utilize the histogram variance of RGO and RGM as the quality-aware features to measure the spatial degradation of $S(h,w)$:

$$V_{RGO} = var(hist(\angle\nabla S(h,w)_{RGO})) \tag{15}$$

$$V_{RGM} = var(hist(|\nabla S(h,w)|_{RGM})) \tag{16}$$

In addition to obtain the V_{RGO} and V_{RGM} of $S(h,w)$ with the original size, we also consider these two variances from the down-sampled version of $S(h,w)$, denoted as V'_{RGO} and V'_{RGM}, respectively. Finally, we combine all these variances to obtain the spatial gradient features f_S:

$$f_S = [V_{RGO}, V_{RGM}, V'_{RGO}, V'_{RGM}] \tag{17}$$

Note that the spatial gradient features of the whole LFI are obtained by averaging the spatial gradient features of all SAIs.

B. Angular-Spatial Statistical Feature Extraction

The angular distortions induced by LFI angular reconstruction methods affect visual perception in an implicit manner. Generally, angular distortions are perceived in conjunction with spatial information, as they are revealed only when refocusing or changing viewpoints [59], [74]. Therefore, we advocate prioritizing angular-spatial distortions rather than angular distortions. Following [50], given a grayscale LFI, we first calculate the horizontal and vertical angular differences $D^{hor}_{u,v}(h,w)$ and $D^{ver}_{u,v}(h,w)$ to mitigate the impact of the extensive information redundancy among different SAIs:

$$D^{hor}_{u,v}(h,w) = S_{u+1,v}(h,w) - S_{u,v}(h,w) \tag{18}$$

$$D^{ver}_{u,v}(h,w) = S_{u,v+1}(h,w) - S_{u,v}(h,w) \tag{19}$$

The Mean Sub-Aperture Gradient Image (MSAGI), denoted as $G(h,w)$, is then calculated as:

$$G_{u,v}(h,w) = \sqrt{D^{hor}_{u,v}(h,w)^2 + D^{ver}_{u,v}(h,w)^2} \tag{20}$$

$$G(h,w) = \underset{u,v}{avg}(G_{u,v}(h,w)) \tag{21}$$

where $\underset{u,v}{avg}(\cdot)$ is the average operation in the angular domain.

Curvelet transform offers multiscale and multidirectional representations with its efficient sparse expression capabilities,

making it a powerful analysis tool in the field of image processing. Also, Curvelet transform is a commonly-used decomposition method in quality assessment research [42], [49]. Thus, we convert the MSAGI into multiscale and multi-directional Curvelet coefficients $C_{s,d}$ based on a set of image blocks with size $B \times B$:

$$C_{s,d} = \sum_{0 \le h,w < B} G(h,w)\overline{\Phi_{s,d}(h,w)} \tag{22}$$

where s and d are the parameters related to scale and direction. $\overline{\Phi_{s,d}(h,w)}$ denotes the conjugate function of the basis Curvelet transform. $C_{s,d}$ contains 5 scales and each scale has different directional information. $C_{1,d}$ and $C_{5,d}$ record low-frequency information and high-frequency information, respectively, with only one direction. $C_{4,d}$ contains the richest detailed information with 64 directions. In our implementation, B is set to 256.

Then, we extract statistical features based on $C_{s,d}$ for quality assessment. First, considering that distortions typically disrupt the energy distribution at different scales, we calculate the multiscale energy as quality-aware features f^s_{eng}:

$$f^s_{eng} = \log_{10}|C_{s,d}| \tag{23}$$

where $s \in \{1,2,3,4,5\}$. Thus, the length of f^s_{eng} is 5.

Besides, the distribution changes of multidirectional information also reflect the LFI quality to a certain extent. To this end, we compute the multidirectional distribution statistics on $C_{4,d}$, as it consists of the richest 64 directional information. Specifically, we divide the 64 directions into four equal parts in sequential order, and calculate the kurtosis and skewness of the first three discriminative parts, resulting in f_{kur} and f_{ske}:

$$f_{kur} = [kur(C_{4,1-16}), kur(C_{4,17-32}), kur(C_{4,33-48})] \tag{24}$$

$$f_{ske} = [ske(C_{4,1-16}), ske(C_{4,17-32}), ske(C_{4,33-48})] \tag{25}$$

Consequently, the angular-spatial statistical features f_A are generated by combining f^s_{eng}, f_{kur}, and f_{ske}:

$$f_A = [f^s_{eng}, f_{kur}, f_{ske}] \tag{26}$$

Note that, f_A is an 11-dimensional feature vector.

C. Binocular Disparity Feature Extraction

As concluded in Section III-D, LFIs with different disparities will produce different visual effects when distortion is introduced. Therefore, we propose to consider the binocular visual effect and extract the binocular disparity features, which provide a complementary description for the degradation of angular-spatial information in LFIs. With a given grayscale LFI, we first combine each pair of horizontally adjacent SAIs and generate a large number of pseudo stereoscopic image pairs, denoted as $\{P^l_{u,v}(t), P^r_{u,v}(t)\}$, where t=(h,w) are spatial coordinates. For each stereopair, we generate the binocular disparity map $d(t)$ by maximizing the structural similarity score [75]. Then, following [76], we define $O^l(t)$ as a $A \times A$ (A is set to 17 as default [76]) neighborhood window of the left image $P^l(t)$, and $\varepsilon(O^l(t))$ is the spatial activity of $O^l(t)$:

$$\varepsilon(O^l(t)) = \log_2(\sigma^2(O^l(t)) + 1) + \delta \tag{27}$$

where δ is set to 0.01 to ensure the stability of the subsequent calculations. $O^r(t)$ and $\varepsilon(O^r(t))$ of the right image $P^r(t)$ can be defined in a similar manner. After that, we calculate the cyclopean spatial activity image $I(t)$ as:

$$I(t) = \frac{\varepsilon(O^l(t)) \cdot P^l(t) + \varepsilon(O^r(t+d(t))) \cdot P^r(t+d(t))}{\varepsilon(O^l(t)) + \varepsilon(O^r(t+d(t)))} \tag{28}$$

The generated cyclopean spatial activity image is capable of capturing and preserving the differences between left and right views, especially the visual discontinuity regions between the two views. Then we calculate the saliency-weighted Mean Subtracted Contrast Normalized (MSCN) coefficients $\hat{I}(t)$ of the cyclopean spatial activity image $I(t)$ as:

$$\hat{I}(t) = \frac{1}{1+|\nabla d(t)|} \cdot \frac{I(t) - \mu(I(t))}{\sigma(I(t)) + 1} \tag{29}$$

where $|\nabla d(t)|$ denotes the gradient magnitude of $d(t)$. $\mu(\cdot)$ and $\sigma(\cdot)$ represent the calculations of local mean and standard deviation, respectively. Also, based on the consideration that the changes of the statistical distributions of MSCN coefficients are associated with the quality deterioration [77], we compute the kurtosis and skewness of $\hat{I}(t)$ as the binocular disparity features f_B:

$$f_B = [kur(\hat{I}(t)), ske(\hat{I}(t))] \tag{30}$$

The average binocular disparity features of all stereopairs are calculated as the binocular disparity features of the whole LFI.

D. Quality Score Prediction

After the aforementioned feature extractions, we obtain the overall quality-aware features $f_{SAB} = [f_S, f_A, f_B]$ for the proposed SAB metric, and $f_{SAB-light} = [f_S, f_A]$ for its lightweight version. Finally, the quality score Q_p is predicted by the SVR with a radial basis function kernel, *i.e.*, $Q_p = SVR(f_{SAB})$, where $SVR(\cdot)$ denotes the trained SVR model.

V. Experiment and Benchmark

A. Experimental Settings

In our experiments, we adopt leave-two-fold-out cross-validation as the train-test split strategy to avoid scene overlap between the training and test sets. Specifically, we first partition all distorted LFIs into K folds based on their reference scenes, where K denotes the number of SRCs and is set to 10 in our database. Subsequently, in each split, K-2 folds are used for training the model and the remaining 2 folds are used for testing the performance. The average result of $K(K$-1)/2 splits is reported as the final performance. To measure the relationship between predicted and subjective scores, Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC), and Root Mean Square Error (RMSE) are adopted, focusing on linear relationship, monotonicity, and predictive accuracy, respectively. Before calculating PLCC and RMSE, we employ a five-parameter nonlinear function for score-mapping as suggested in [78].

TABLE II
COMPARISON OF PERFORMANCE AND RUNNING TIME ON THE PROPOSED DATABASE. THE DEEP-LEANING-BASED AND HANDCRAFTED FEATURE-BASED METRICS ARE MARKED WITH AND WITHOUT *. THE BEST PERFORMANCE OF RR/FR METRICS AND NR METRICS ARE MARKED IN **BOLD**, RESPECTIVELY.

Metric Types	Metrics	PLCC	SRCC	RMSE	Time (s)
RR LFIQA	LF-IQM [44]	0.4506	0.3409	1.4498	589.7851
	RRLFIQA-4DDWT [45]	0.8500	0.8251	0.8223	128.1106
FR LFIQA	EDDMF* [43]	**0.9469**	**0.9234**	**0.5171**	1.8775
	MDFM [34]	0.6367	0.4864	1.1822	**0.8537**
	Fang's [36]	0.8935	0.8712	0.7100	1.1574
	Min's [37]	0.8987	0.8617	0.7204	3.9845
	Meng's [38]	0.7684	0.6677	1.0075	30.4872
	KRIQE [39]	0.8815	0.8557	0.7523	115.7841
NR LFIQA	DNNF-LFIQA* [55]	0.8807	0.8213	0.7835	3.7410
	DeLFIQE* [56]	0.7349	0.6795	1.0807	15.4193
	DeeBLiF* [57]	0.8813	0.8434	0.7653	4.8533
	PVBLiF* [58]	0.8948	0.8743	0.7091	8.3795
	ASEM-BLiF* [59]	0.9023	0.8490	0.7094	5.2377
	BELIF [46]	0.7915	0.7335	0.9887	107.8814
	NR-LFQA [8]	0.8043	0.7624	0.9065	225.2069
	Tensor-NLFQ [47]	0.7705	0.6787	1.0375	697.6515
	PVRI [48]	0.8520	0.8148	0.8366	71.3578
	DSA [49]	0.8739	0.8386	0.7992	198.5443
	4D-DCT-LFIQA [50]	0.8080	0.7583	0.9644	169.2623
	TSSV-LFIQA [51]	0.8054	0.7494	0.9661	44.6696
	NR-LF-QAE [52]	0.7876	0.7427	0.9533	254.1088
	SATV-BLiF [53]	0.8494	0.8064	0.8552	3.8812
	SAB-light (ours)	0.8838	0.8661	0.7485	**3.6741**
	SAB (ours)	**0.9144**	**0.8934**	**0.6693**	133.3048

B. Benchmark Establishment and Performance Comparison

In this subsection, we establish a relatively comprehensive benchmark on the proposed database. The objective metrics involved in this benchmark include the two proposed metrics and a total of 22 existing objective LFIQA metrics covering multiple types: two RR LFIQA metrics (LF-IQM [44] and RRLFIQA-4DDWT [45]), six FR LFIQA metrics (EDDMF [43], MDFM [34], Fang's [36], Min's [37], Meng's [38], and KRIQE [39]), and fourteen NR LFIQA metrics (DNNF-LFIQA [55], DeLFIQE [56], DeeBLiF [57], PVBLiF [58] ASEM-BLiF [59], BELIF [46], NR-LFQA [8], Tensor-NLFQ [47], PVRI [48], DSA [49], 4D-DCT-LFIQA [50], TSSV-LFIQA [51], NR-LF-QAE [52], and SATV-BLiF [53]). Among them, EDDMF, DNNF-LFIQA, DeLFIQE, DeeBLiF, PVBLiF, and ASEM-BLiF are deep learning-based, while other metrics are based on handcrafted features. For all metrics, we reproduce their performance on the same hardware configurations using the code from their authors.

TABLE IV exhibits the benchmark of objective metrics in terms of quality assessment performance and running time on the proposed database. Although most LFIQA metrics share a common emphasis on the degradation of angular-spatial information, they show significant differences in quality assessment capabilities. These findings indicate that different quality assessment metrics exhibit different sensitivities to multiple distortions, and also highlight the guiding significance

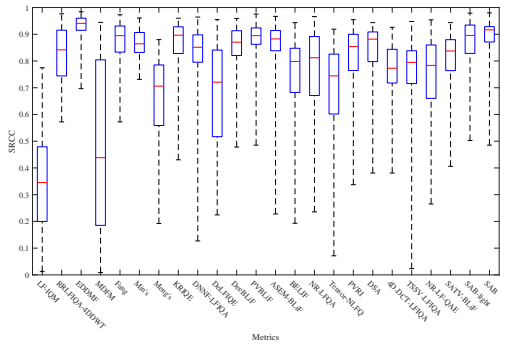


Fig. 8. Box plots of SRCC distributions on the proposed database.

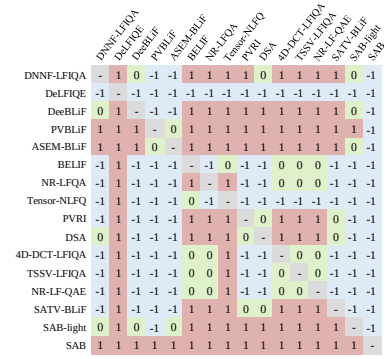


Fig. 9. Statistical significance between any two NR LFIQA metrics.

of our database in the selection of objective LFIQA metrics. Typically, due to the availability of reference information, RR/FR metrics show superior quality evaluation performance than NR metrics. In addition, the NR metrics based on deep learning generally outperform those based on handcrafted features. However, despite the use of handcrafted features, the proposed SAB metric still performs better than most FR/RR metrics or deep learning-based metrics, without incorporating any reference information. This may be because our metric considers the potential impact of LFI distortions from multiple aspects. Further, both our SAB metric and its lightweight version achieve superior performance compared to the metrics that also utilize handcrafted features. In addition to the quality evaluation performance, the running time of each metric is also provided in TABLE IV. We can see that the handcrafted feature-based NR LFIQA metrics tend to be more time-consuming compared to the deep learning-based ones, but this comparison is somewhat unfair because deep learning-based methods require significant extra time and more powerful parallel computing platforms to train a model. Among the NR metrics based on handcrafted features, our SAB metric has no clear advantage in running time although it has the best quality evaluation performance. However, the SAB-light metric achieves an optimal trade-off by effectively reducing running time without significantly compromising predictive accuracy.

Then, we provide the box plots of SRCC distributions for all metrics on the proposed database in Fig. 8. For a specific box, shorter length and higher vertical location represent better stability and predictive accuracy, respectively. Accordingly, we

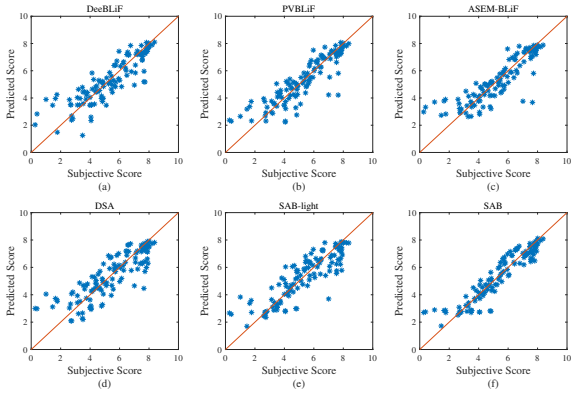


Fig. 10. Scatter plots of predicted scores versus subjective scores for the six best-performing NR LFIQA metrics on the proposed database. (a) DeeBLiF; (b) PVBLiF; (c) ASEM-BLiF; (d) DSA; (e) SAB-light; (f) SAB.

can see that the proposed SAB metric has remarkable stability and quality evaluation ability, while its lightweight version also achieves competitive results.

Additionally, a F-test is performed to study the statistical significance relationship between any two NR LFIQA metrics. Following the implementation in [58], statistical significance results are provided in Fig. 9, where "-1", "0", and "1" indicate that the row metric is statistically worse, competitive, and better than the column metric, respectively. From the figure, we can see that our SAB metric statistically outperforms all existing NR LFIQA metrics. Besides, the SAB-light metric performs significantly better than all handcrafted feature-based NR LFIQA metrics, while obtaining competitive results compared to most deep learning-based NR LFIQA metrics.

To further demonstrate the superiority of the proposed metric in quality evaluation, Fig. 10 presents the scatter plots of predicted scores versus subjective scores for the six best-performing NR metrics, including DeeBLiF, PVBLiF, ASEM-BLiF, DSA, SAB-light, and SAB. We can see that the scatter plots of our SAB metric show closer proximity to the red line ($i.e.$, perfect predictions) with fewer outliers compared to that of other metrics. All these results provide strong evidence of the effectiveness of the proposed metric.

C. Ablation Studies

In order to investigate the effectiveness of different features in our metric, ablation studies are conducted on the proposed database, as shown in TABLE III. We can see that both f_S and f_A show significant efficacy in quality assessment. The combination of these two features constitutes the SAB-light metric and achieves remarkable quality evaluation performance. Further, it could also be observed that relying solely on f_B for quality assessment does not yield satisfactory results, as f_B only includes 2 quality-aware values. Intuitively, too few features may fail to provide a comprehensive description of LFI distortions. However, the incorporation of f_B with any other features consistently leads to noticeable performance improvements, indicating its capability to serve as discriminative supplementary information. Finally, by combining f_S, f_A, and f_B, the SAB metric is formed, resulting in the best quality evaluation performance.

TABLE III
ABLATION STUDIES OF DIFFERENT FEATURE COMBINATIONS ON THE PROPOSED DATABASE. THE BEST PERFORMANCE IS IN **BOLD**.

f_S	f_A	f_B	PLCC	SRCC	RMSE
✓			0.8114	0.7661	1.0026
	✓		0.7961	0.7445	0.9531
		✓	0.4882	0.3797	1.4500
✓	✓		0.8838	0.8661	0.7485
✓		✓	0.8509	0.7987	0.8798
	✓	✓	0.8328	0.7947	0.8681
✓	✓	✓	**0.9144**	**0.8934**	**0.6693**

TABLE IV
COMPARISON OF SRCC PERFORMANCE OF DIFFERENT DISTORTION TYPES ON THE PROPOSED DATABASE. THE DEEP-LEANING-BASED AND HANDCRAFTED FEATURE-BASED METRICS ARE MARKED WITH AND WITHOUT *. THE BEST PERFORMANCE OF RR/FR METRICS AND NR METRICS ARE MARKED IN **BOLD**, RESPECTIVELY.

Metric Types	Metrics	S1	S2	A1	A2
RR LFIQA	LF-IQM [44]	0.4445	0.3962	0.4699	0.4571
	RRLFIQA-4DDWT [45]	0.8589	0.8885	0.7543	0.7624
FR LFIQA	EDDMF* [43]	**0.8906**	**0.9152**	0.8430	**0.9389**
	MDFM [34]	0.3460	0.4448	0.4788	0.4995
	Fang's [36]	0.8255	0.8793	0.7697	0.8257
	Min's [37]	0.8034	0.8747	**0.8634**	0.8807
	Meng's [38]	0.5297	0.6372	0.6404	0.6954
	KRIQE [39]	0.7729	0.8316	0.8055	0.9025
NR LFIQA	DNNF-LFIQA* [55]	0.8053	0.7837	0.8238	0.8772
	DeLFIQE* [56]	0.5817	0.6822	0.7841	0.7905
	DeeBLiF* [57]	0.7546	0.7991	0.7673	0.9095
	PVBLiF* [58]	0.8004	0.8346	0.7953	**0.9197**
	ASEM-BLiF* [59]	0.7826	0.8211	0.7468	0.9087
	BELIF [46]	0.6638	0.6582	0.7727	0.8521
	NR-LFQA [8]	0.7298	0.7600	0.7851	0.7667
	Tensor-NLFQ [47]	0.5997	0.6628	0.7465	0.8026
	PVRI [48]	0.7791	0.7716	0.8220	0.6800
	DSA [49]	0.8400	0.8066	0.8211	0.9001
	4D-DCT-LFIQA [50]	0.7253	0.7007	0.7207	0.8446
	TSSV-LFIQA [51]	0.8036	0.7457	0.7947	0.8896
	NR-LF-QAE [52]	0.7298	0.8036	0.8182	0.8300
	SATV-BLiF [53]	0.7684	0.7015	0.7907	0.8117
	SAB-light (ours)	0.8527	0.8378	**0.9081**	0.8863
	SAB (ours)	**0.8545**	**0.8389**	0.8987	0.8508

D. Comparison of Different Distortion Types

As the proposed database consists of 12 different distortion types, it would be interesting to investigate the robustness against different distortion types of the two proposed metrics as well as other state-of-the-art metrics. However, there are only 10 distorted LFIs for each distortion type in our database. If leave-two-fold-out cross-validation is adopted, only 2 distorted LFIs will be allocated in the test set, which will not be able to verify the quality assessment performance of objective metrics. Therefore, given a single distortion type, we define all distortion types derived from it as the same distortion type. For example, the "A1", "A1S1", "A1S2", "S1A1", and "S2A1"

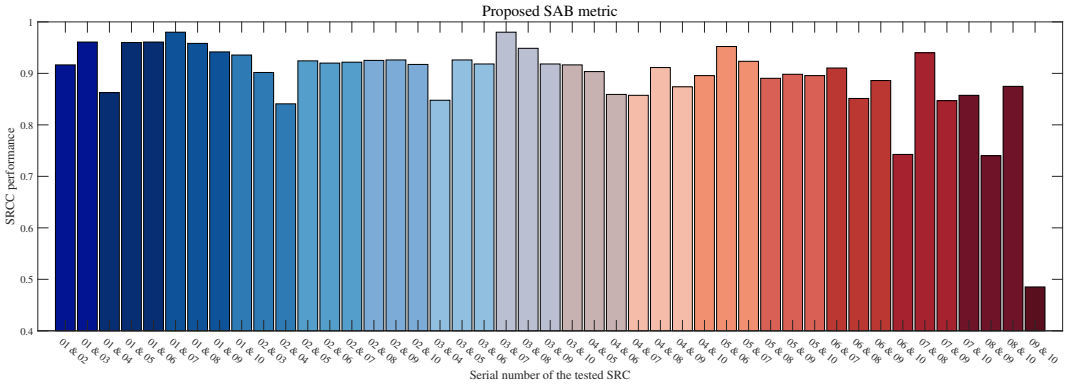
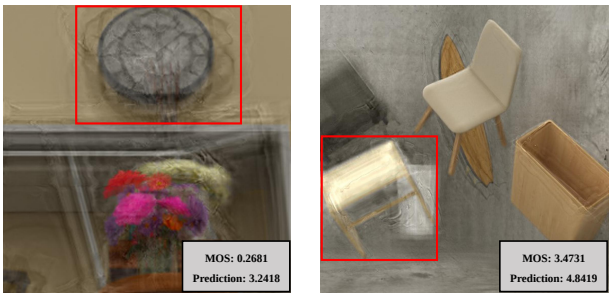


Fig. 11. Summary of the SRCC performance of different train-test splits in leave-two-fold-out cross-validation on the proposed database.

TABLE V
PERFORMANCE OF TRAINING ON THREE DIFFERENT DATABASES, AND TESTING ON THE PROPOSED DATABASE.

Metrics	Training Databases	PLCC	SRCC	RMSE
SAB-light	SHU	0.6438	0.6668	1.4739
	SMART	0.7001	0.6933	1.3754
	VALID	0.6885	0.7241	1.3968
SAB	SHU	0.6671	0.6633	1.4349
	SMART	0.6941	0.6823	1.3865
	VALID	0.6888	0.7273	1.3963



R9A2S1 R10A2

Fig. 12. Illustrative examples of inaccurate predictions. Severe artifacts can be observed in red boxes.

distortion types all belong to the "A1" distortion type. Based on this classification principle, the proposed database has 4 single distortion types: "S1", "S2", "A1", and "A2".

TABLE IV shows the experimental result comparison of SRCC performance of different distortion types on the proposed database. Due to the space constraint, we only provide the SRCC performance for brevity, while the PLCC and RMSE performance show similar results. We can see that our proposed SAB metric and its lightweight version perform better than other NR LFIQA metrics in most cases, and even achieve superior performance over FR/RR LFIQA metrics on certain distortion types ($e.g.$, A1), which fully demonstrates the robustness and effectiveness of the proposed metrics. Nonetheless, it can be found that there is a performance gap between the SAB metric and the SAB-light metric in terms of A2 distortion. One possible explanation is that due to the underexplored information between long-distance viewpoints, binocular disparity features exploited from adjacent horizontal viewpoints may have a counterproductive effect on the evaluation of angular distortions, which deserves further investigation and optimization.

E. Cross-Database Evaluation

In this subsection, we perform the cross-database experiments by training the proposed metrics on three different databases (including SHU [30], SMART [19], and VALID [20]) and testing their performance on the proposed database. We choose these three training databases for diversity considerations: SHU has the richest distortion levels; SMART only involves LFIs that suffer from compression-related distortions; VALID is a small-scale database. The cross-database experimental results are shown in TABLE V. It can be found that two proposed metrics achieve competitive quality evaluation performance even when trained on other three databases with different characteristics, which strongly verifies the effectiveness and robustness of our designed features. Moreover, we can also see that the SAB metric and its lightweight version have similar performance, indicating that the cross-database evaluation ability mainly benefits from spatial gradient features f_S and angular-spatial statistical features f_A.

F. Limitation Analysis

Although the effectiveness of the proposed SAB metric has been fully demonstrated, it does have certain limitations, which offer valuable insights and inspiration for future endeavors.

First, we summarize the SRCC performance of different train-test splits in leave-two-fold-out cross-validation on the proposed database, as shown in Fig. 11. We can easily find that the proposed SAB metric consistently performs well on most train-test splits, except for the test set consisting of distorted LFIs from SRC09 and SRC10 scenarios. A possible reason is that the sparse angular disparity in these two scenarios leads to severe artifacts when introducing LFI angular reconstruction, but our SAB metric does not cope well in this case. In addition, the lack of sufficient sparse angular disparity LFI data (only the SRC08 scenario is included in the training set) further limits the effectiveness of our metric. Illustrative examples

of inaccurate predictions are provided in Fig. 12, where only the central viewpoint of each LFI is displayed due to space limitations. From the figure, we can observe severe artifacts at the edges of several objects, in which case the performance of our metric is limited. The practical value of our metric can be further promoted by addressing this shortcoming.

Second, the proposed metric analyzes the degradation of LFI quality from three relatively independent perspectives. Considering that the high-dimensional LFI information is presented to the human eye in a holistic manner, additional consideration of the holistic human visual perception may help to further improve the quality evaluation performance of our metric.

The final limitation is related to the time complexity. Although the proposed SAB metric achieves superior quality evaluation performance compared to the lightweight version, its time complexity also increases exponentially. The main time consumption lies in extracting binocular disparity features from all pseudo stereoscopic image pairs. Therefore, it would be interesting to define and explore some representative pseudo stereoscopic image pairs for quality assessment, which would help in developing more efficient and effective quality evaluation metrics.

VI. Conclusion

In this paper, we fill the gap in subjective experimental studies of multiple distortions in LFIs, and therefore propose a new perceptual quality assessment database for LFIs. Specifically, the proposed database focuses on the distortions induced by deep learning-based LFI angular reconstruction and spatial super-resolution methods, individually and multiplely. The impact of different adoption orders of the two distortions in multiple distortions is also studied in our database. Further, our database collects three different types of LFIs to present a comprehensive study on the human visual perception of LFIs. As a result, 10 reference scenes and 120 distorted LFIs subject to various distortion types are included in our database. Through the participation of 32 observers in subjective tests, we obtain subjective quality scores and conduct a thorough statistical analysis on these scores. In addition, by exploring quality-aware features from spatial gradient, angular-spatial statistics, and binocular disparity, we propose a novel NR LFIQA metric in this paper, while developing its lightweight version with the consideration of efficiency. Finally, a benchmark of the two proposed metrics and numerous state-of-the-art LFIQA metrics on the proposed database is established. Experimental results demonstrate the superiority of our metrics over other existing NR metrics. In the future, we will focus on developing LFIQA metrics on unlabeled large-scale LFI databases ($e.g.$, NTIRE 2023 [79]) in an unsupervised manner.

References

[1] R. Mekuria, K. Blom, and P. Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 828-842, 2017.

[2] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1778-1791, 2022.

[3] Y. Fang, X. Sui, J. Wang, J. Yan, J. Lei, and P. Le Callet, "Perceptual quality assessment for asymmetrically distorted stereoscopic video by temporal binocular rivalry," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 8, pp. 3010-3024, 2021.

[4] B. Hu, S. Wang, L. Li, J. Leng, Y. Yang, and X. Gao, "Hierarchical discrepancy learning for image restoration quality assessment," *Signal Process.*, vol. 198, pp. 108595, 2022.

[5] Y. Liu, Z. Ni, S. Wang, H. Wang, and S. Kwong, "High dynamic range image quality assessment based on frequency disparity," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4435-4440, 2023.

[6] W. Wen, K. Wei, Y. Fang, and Y. Zhang, "Visual quality assessment for perceptually encrypted light field images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2522-2534, 2021.

[7] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31-42.

[8] L. Shi, W. Zhou, Z. Chen, and J. Zhang, "No-reference light field image quality assessment based on spatial-angular measurement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4114-4128, 2019.

[9] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1-11, 2005.

[10] C. Conti, L. D. Soares, and P. Nunes, "Dense light field coding: A survey," *IEEE Access*, vol. 8, pp. 49244-49284, 2020.

[11] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926-954, 2017.

[12] V. K. Adhikarla, M. Vinkler, D. Sumin, R. K. Mantiuk, K. Myszkowski, H.-P. Seidel, and P. Didyk, "Towards a quality metric for dense light fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3720-3729.

[13] L. Shi, S. Zhao, W. Zhou, and Z. Chen, "Perceptual evaluation of light field image," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 41-45.

[14] Z. Huang, M. Yu, G. Jiang, K. Chen, Z. Peng, and F. Chen, "Reconstruction distortion oriented light field image dataset for visual communication," in *Proc. Int. Symp. Net. Comp. Commun.*, 2019, pp. 1-5.

[15] E. Shafiee and M. G. Martini, "Datasets for the quality assessment of light field imaging: Comparison and future directions," *IEEE Access*, vol. 11, pp. 15014-15029, 2023.

[16] W. Ellahi, T. Vigier, and P. L. Callet, "Analysis of public light field datasets for visual quality assessment and new challenges," in *Proc. Eur. Light Field Image Workshop*, 2019, pp. 2-6.

[17] K. Javidi, M. G. Martini, and P. A. Kara, "KULF-TT53: A display-specific turntable-based light field dataset for subjective quality assessment," *Electronics*, vol. 12, no. 23, pp. 4868, 2023.

[18] Y. Cui, G. Jiang, M. Yu, Y. Chen, and Y.-S. Ho, "Stitched wide field of view light field image quality assessment: Benchmark database and objective metric," *IEEE Trans. Multimedia*, vol. 26, pp. 5092-5107, 2023.

[19] P. Paudyal, F. Battisti, M. Sjöström, R. Olsson, and M. Carli, "Toward the perceptual quality evaluation of compressed light field images," *IEEE Trans. Broadcast.*, vol. 63, no. 3, pp. 507-522, 2017.

[20] I. Viola and T. Ebrahimi, "VALID: Visual quality assessment for light field images dataset," in *Proc. Int. Conf. Quality Multimedia Exper.*, 2018, pp. 1-3.

[21] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. Int. Conf. Quality Multimedia Exper.*, 2016, pp. 1-2.

[22] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4562-4566.

[23] W. Ahmad, R. Olsson, and M. Sjostrom, "Interpreting plenoptic images as multi-view sequences for improved compression," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4557-4561.

[24] I. Tabus, P. Helin, and P. Astola, "Lossy compression of lenslet images from plenoptic cameras combining sparse predictive coding and JPEG 2000," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 4567-4571.

[25] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19-34.

[26] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6319-6327.

[27] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1-10, 2016.

[28] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646-1654.

[29] S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, "Micro-lens-based matching for scene recovery in lenslet cameras," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1060-1075, 2018.

[30] L. Shan, P. An, C. Meng, X. Huang, C. Yang, and L. Shen, "A no-reference image quality assessment metric by multiple characteristics of light field images," *IEEE Access*, vol. 7, pp. 127217-127229, 2019.

[31] A. Zizien and K. Fliegel, "LFDD: Light field image dataset for performance evaluation of objective quality metrics," in *Proc. Appl. Digit. Image Process. XLII*, vol. 11510, 2020, Art. no. 115102U.

[32] N. Bakir, S. A. Fezza, W. Hamidouche, K. Samrouth, and O. Déforges, "Subjective evaluation of light field image compression methods based on view synthesis," in *Proc. Eur. Signal Process. Conf.*, 2019, pp. 1-5.

[33] IEEE recommended practice for the quality assessment of light field imaging, document IEEE 3333.1.4-2022, IEEE Standards Association, 2022.

[34] Y. Tian, H. Zeng, L. Xing, J. Chen, J. Zhu, and K.-K. Ma, "A multi-order derivative feature-based quality assessment model for light field image," *J. Vis. Commun. Image Represent.*, vol. 57, pp. 212-217, 2018.

[35] Y. Tian, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, "A light field image quality assessment model based on symmetry and depth features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 2046-2050, 2020.

[36] Y. Fang, K. Wei, J. Hou, W. Wen, and N. Imamoglu, "Light filed image quality assessment by local and global features of epipolar plane image," in *Proc. IEEE Int. Conf. Multimedia Big Data*, 2018, pp. 1-6.

[37] X. Min, J. Zhou, G. Zhai, P. L. Callet, X. Yang, and X. Guan, "A metric for light field reconstruction, compression, and display quality evaluation," *IEEE Trans. Image Process.*, vol. 29, pp. 3790-3804, 2020.

[38] C. Meng, P. An, X. Huang, C. Yang, and D. Liu, "Full reference light field image quality evaluation based on angular-spatial characteristic," *IEEE Signal Process. Lett.*, vol. 27, pp. 525-529, 2020.

[39] C. Meng, P. An, X. Huang, C. Yang, L. Shen, and B. Wang, "Objective quality assessment of lenslet light field image based on focus stack," *IEEE Trans. Multimedia*, vol. 24, pp. 3193-3207, 2021.

[40] H. Huang, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "Light field image quality assessment using contourlet transform," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, 2021, pp. 1-5.

[41] H. Huang, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, "A spatial and geometry feature-based quality assessment model for the light field images," *IEEE Trans. Image Process.*, vol. 31, pp. 3765-3779, 2022.

[42] J. Ma, X. Zhang, C. Jin, P. An, and G. Xu, "Light field image quality assessment using natural scene statistics and texture degradation," *IEEE Trans. Circuits Syst. Video Technol.*, 2023, doi: 10.1109/TCSVT.2023.3297016.

[43] Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, "EDDMF: An efficient deep discrepancy measuring framework for full-reference light field image quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 6426-6440, 2023.

[44] P. Paudyal, F. Battisti, and M. Carli, "Reduced reference quality assessment of light field images," *IEEE Trans. Broadcast.*, vol. 65, no. 1, pp. 152-165, 2019.

[45] J. Xiang, P. Chen, Y. Dang, R. Liang, and G. Jiang, "Pseudo light field image and 4D Wavelet-transform-based reduced-reference light field image quality assessment," *IEEE Trans. Multimedia*, 2023, doi: 10.1109/TMM.2023.3273855.

[46] L. Shi, S. Zhao, and Z. Chen, "BELIF: Blind quality evaluator of light field image with tensor structure variation index," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 3781-3785.

[47] W. Zhou, L. Shi, Z. Chen, and J. Zhang, "Tensor oriented no-reference light field image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4070-4084, 2020.

[48] J. Xiang, M. Yu, G. Jiang, H. Xu, Y. Song, and Y.-S. Ho, "Pseudo video and refocused images-based blind light field image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2575-2590, 2021.

[49] J. Xiang, G. Jiang, M. Yu, Y. Bai, and Z. Zhu, "No-reference light field image quality assessment based on depth, structural and angular information," *Signal Process.*, vol. 184, pp. 108063, 2021.

[50] J. Xiang, G. Jiang, M. Yu, Z. Jiang, and Y.-S. Ho, "No-reference light field image quality assessment using four-dimensional sparse transform," *IEEE Trans. Multimedia*, vol. 25, pp. 457-472, 2023.

[51] Z. Pan, M. Yu, G. Jiang, H. Xu, and Y.-S. Ho, "Combining tensor slice and singular value for blind light field image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 672-687, 2021.

[52] X. Chai, F. Shao, Q. Jiang, X. Wang, L. Xu, and Y.-S. Ho, "Blind quality evaluator of light field images by group-based representations and multiple plane-oriented perceptual characteristics," *IEEE Trans. Multimedia*, 2023, doi: 10.1109/TMM.2023.3268370.

[53] Z. Zhang, S. Tian, W. Zou, Y. Zhang, L. Morin, and L. Zhang, "Blind quality assessment of light field image based on spatio-angular textural variation," in *Proc. IEEE Int. Conf. Image Process.*, 2023, pp. 2385-2389.

[54] Q. Qu, X. Chen, V. Chung, and Z. Chen, "Light field image quality assessment with auxiliary learning based on depthwise and anglewise separable convolutions," *IEEE Trans. Broadcast.*, vol. 67, no. 4, pp. 837-850, 2021.

[55] S. Alamgeer and M. C. Q. Farias, "Deep learning-based light field image quality assessment using frequency domain inputs," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper.*, 2022, pp. 1-6.

[56] P. Zhao, X. Chen, V. Chung, and H. Li, "DeLFIQE—A low-complexity deep learning-based light field image quality evaluator," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-11, 2021.

[57] Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, "DeeBLiF: Deep blind light field image quality assessment by extracting angular and spatial information," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 2266-2270.

[58] Z. Zhang, S. Tian, W. Zou, L. Morin, and L. Zhang, "PVBLiF: A pseudo video-based blind quality assessment metric for light field image," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 6, pp. 1193-1207, 2023.

[59] Z. Zhang, S. Tian, Y. Zhang, W. Zou, L. Morin, and L. Zhang, "Blind perceptual quality assessment of LFI based on angular-spatial effect modeling," *IEEE Trans. Broadcast.*, 2023, doi: 10.1109/TBC.2023.3308329.

[60] "INRIA Lytro image dataset." [Online]. Available: https://www.irisa.fr/temics/demos/lightField/LowRank2/datasets/datasets.html

[61] J. Shi, X. Jiang, and C. Guillemot, "A framework for learning depth from a flexible subset of dense and sparse light field views," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5867-5880, 2019.

[62] Subjective Video Quality Assessment Methods for Multimedia Applications, document ITU-T Rec. P.910, International Telecommunication Union, 2023.

[63] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," *Proc. SPIE*, vol. 5007, pp. 87-96, 2003.

[64] N. Meng, H. K.-H. So, X. Sun, and E. Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873-886, 2021.

[65] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11141-11148.

[66] G. Liu, H. Yue, J.Wu, and J. Yang, "Intra-inter view interaction network for light field image super-resolution," *IEEE Trans. Multimedia*, vol. 25, pp. 256-266, 2021.

[67] Y. Wang, L. Wang, G. Wu, J. Yang, W. An, J. Yu, and Y. Guo, "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425-443, 2023.

[68] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5959-5974, 2022.

[69] M. Pérez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, "From pairwise comparisons and rating to a unified quality scale," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 1139-1151, 2019.

[70] L. L. Thurstone, "A law of comparative judgement," *Psychol. Rev.*, vol. 34, no. 4, pp. 273-286, 1927.

[71] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Process., Image Commun.*, vol. 40, pp. 1-15, 2016.

[72] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-reference quality assessment of screen content images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 1–14, 2018.

[73] D. L. Ruderman, "The statistics of natural images," *Netw.: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517-548, 1994.

[74] P. A. Kara, A. Cserkaszky, A. Barsi, T. Papp, M. G. Martini, and L. Bokor, "The interdependence of spatial and angular resolution in the quality of experience of light field visualization," in *Proc. Int. Conf. 3D Immersion (IC3D)*, 2017, pp. 1-8.

[75] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Process., Image Commun.*, vol. 28, no. 9, pp. 1143-1155, 2013.

[76] L. Liu, B. Liu, C.-C. Su, H. Huang, and A. C. Bovik, "Binocular spatial activity and reverse saliency driven no-reference stereopair quality assessment," *Signal Process., Image Commun.*, vol. 58, pp. 287-299, 2017.

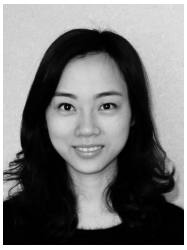
[77] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao,"Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929-943, 2020.

[78] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2015. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/vqeg-home.

[79] Y. Wang, L. Wang, Z. Liang, et al. "NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1320-1335.

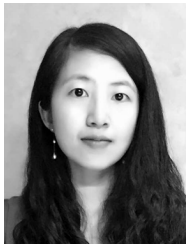
Zhengyu Zhang received the B.E. degree from Guangzhou University, Guangzhou, China, the M.E. degree from Shenzhen University, Shenzhen, China, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2018, 2021, and 2024, respectively. He is currently a lecturer with the School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, China. His research interests include image/video quality assessment, light field imaging, visual perception, and deep learning.

Shishun Tian received the B.Sc. degree from Sichuan University, Chengdu, China, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2012, 2015, and 2019, respectively. He is currently an Assistant Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include image quality assessment, visual perception, and machine learning.



Jinjia Zhou received B.E. degree from Shanghai Jiao Tong University, China, in 2007. She received M.E. and Ph.D. degrees from Waseda University, Japan, in 2010 and 2013, respectively. From 2013 to 2016, she worked as a junior researcher at Waseda University, Fukuoka, Japan. Following that, she served as an Associate Professor and co-director of the English-based graduate program at Hosei University from 2016 to 2020. Additionally, from 2017 to 2021, she held the position of senior visiting scholar at the State Key Laboratory of ASIC & System, Fudan University, China. From 2020 to 2021, she was also appointed as a specially appointed Associate Professor at Osaka University. Currently, she is an Associate Professor at Hosei University. Her research interests focus on algorithms and VLSI architectures for multimedia signal processing.

Luce Morin is currently a Full-Professor with National Institute of Applied Science (INSA Rennes), University of Rennes, France, and a Researcher with the Institut d'Electronique et Technologies du numéRique (IETR), within the VAADER research team. She has authored or coauthored over 90 scientific papers in international journals and conferences. Her research activities deal with computer vision, 3D reconstruction, image and video compression, and representations for 3D videos and multiview videos.



Lu Zhang is an associate professor at National Institute of Applied Sciences (INSA) of Rennes in France. She received the B.S degree from Southeast University and the M.S. degree from Shanghai Jiaotong University in China in 2004 and 2007, respectively. From October 2009 to November 2012, she was a PhD student of the LISA and CNRS IRCCyN labs in France, working on the model observers for the medical image quality assessment. She received the Excellent Doctoral Dissertation of France awarded by IEEE France Section, SFGBM, AGBM and GdR CNRS-Inserm Stic-Santé in 2013. Then she worked on the quality of experience (QoE) in telemedicine before she joined INSA in September 2013, as a member of the VAADER research group of the IETR lab. She is a board member of the international VQEG (Video Quality Experts Group). She is elected as a Multimedia Signal Processing Technical Committee (MMSP TC) Member and EURASIP TAC (Technical Area Committees) VIP (Visual Information Processing) Member for the period of 2022-2024. She works on human perception understanding, image quality assessment, saliency prediction, image analysis and coding.