Blind Perceptual Quality Assessment of LFI Based on Angular-Spatial Effect Modeling

Zhengyu Zhang, Shishun Tian, Yuhang Zhang, Wenbin Zou, Luce Morin, Lu Zhang

Blind Perceptual Quality Assessment of LFI Based on Angular-Spatial Effect Modeling

Zhengyu Zhang, Shishun Tian, *Member, IEEE*, Yuhang Zhang, *Graduate Student Member, IEEE*, Wenbin Zou, Luce Morin, *Member, IEEE*, and Lu Zhang

***Abstract*—By recording scenes from multiple viewpoints, Light Field Image (LFI) encompasses both angular and spatial information, thereby offering users a more immersive experience. Since LFIs may be distorted at various stages from acquisition to visualization, Light Field Image Quality Assessment (LFIQA) is of vitally important to monitor the potential impairments of LFI quality. However, existing objective LFIQA metrics fail to establish a reasonable correlation between spatial and angular information in LFIs, especially ignoring the imbalance problem of large spatial variations and subtle angular variations, which results in unsatisfactory quality evaluation performance. To alleviate this imbalance, in this paper, we propose a novel Blind LFIQA metric based on Angular-Spatial Effect Modeling, abbreviated as ASEM-BLiF. Specifically, the proposed metric consists of two branches. In the principal branch, we first present an Angular Effect Modeling (AEM) module to capture the angular information independently of spatial information. Based on AEM, we further design an Angular-Spatial Quality Learning (ASQL) module to model the local angular-spatial effect and establish the global relationship between different local regions for quality assessment via Transformer. In the auxiliary branch, a Discriminative Region Selection (DRS) module is proposed for auxiliary learning to improve the learning efficiency and prediction accuracy from a local perspective. Moreover, we present a Dynamic Weighting Loss (DWLoss) to achieve an optimal balance between principal and auxiliary learning throughout training. To demonstrate the effectiveness of the proposed metric, extensive experiments are conducted on five publicly available LFIQA databases with a variety of metrics. The experimental results show that compared to our previous work DeeBLiF, the current state-of-the-art LFIQA metric, our proposed ASEM-BLiF metric achieves 5.67%, 7.75%, 5.96%, 4.44%, and 0.33% SROCC performance improvements in quality assessment on the Win5-LID, NBU-LF1.0, LFDD, VALID10bit, and SHU databases, respectively. The code will be publicly available.**

Manuscript received 15 May 2023; revised 9 August 2023; accepted 10 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62101344 and Grant 62171294; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2022A1515010159 and Grant 2020A1515010959; in part by the Key Project of DEGP under Grant 2018KCXTD027; in part by the Key Project of Shenzhen Science and Technology Plan under Grant 20220810180617001; and in part by the China Scholarship Council. *(Corresponding author: Shishun Tian.)*

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

Shishun Tian, Yuhang Zhang, and Wenbin Zou are with the Guangdong Key Laboratory of Intelligent Information Processing, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518000, China (e-mail: stian@szu.edu.cn; zhangyuhang2019@email.szu.edu.cn; wzou@szu.edu.cn).

Digital Object Identifier 10.1109/TBC.2023.3308329

***Index Terms*—Light field image, quality assessment, blind, angular-spatial effect, auxiliary learning.**

I. Introduction

THE EMERGENCE of Light Field Image (LFI) enables a wide range of immersive broadcasting scenes [1], [2], from which many attrIEEEdraft,active applications are emerged, such as post-capture image editing [3], [4], de-occlusion [5], [6], and reflectance estimation [7], [8]. By encoding intensity and directions of light rays into a 4D representation [9], the LFI can be described by a biplane model $L(u, v, h, w)$. In this model, (u, v) denote different angular viewpoints to record the same scene, while (h, w) record the spatial information of each viewpoint, *i.e.*, Sub-Aperture Image (SAI). In the process of compression [10], reconstruction [11], and display [12], LFIs containing extra angular information inevitably suffer from various distortions, which are quite different from those in other image types [13], [14]. This further affects the quality of the user's visual experience [15]. Light Field Image Quality assessment (LFIQA) has thus become an imperative in monitoring the visual quality of LFIs.

At present, quality assessment can be classified into two categories: subjective and objective. Subjective methods directly collect the human ratings for each viewed image, and are thus treated as the most reliable way to obtain the quantitative perceptual quality [16]. Nevertheless, subjective methods are extremely time-consuming and labor-intensive, and difficult to be employed in real-time systems. Instead, objective metrics design computational models to evaluate the perceptual quality automatically, with the aim of being effective substitutes for subjective methods. In recent years, a wealth of objective LFIQA metrics have been proposed [17]. However, despite the remarkable achievements, the prediction accuracy of objective LFIQA metrics is still far from ideal due to various factors. First, the characteristics of LFIs vary slightly depending on the capture hardware (*e.g.*, LFIs captured by the multi-camera array [18] have larger angular disparity than that captured by the light field camera [19]), which increases the difficulty of designing a general LFIQA metric. Further, LFIs can be visualized in different representations due to its high-dimensional nature, which also brings more challenges to LFIQA. More importantly, the inherently narrow parallax of LFIs results in subtle differences between adjacent angular views. This characteristic causes complex and idiosyncratic visual effect, and

thus sets LFIQA metrics apart from traditional 2DIQA metrics [20], [21], [22], [23], [24], [25], 3DIQA metrics [26], [27], [28], [29], and multi-view IQA metrics [30], [31], [32].

Existing objective LFIQA metrics can be categorized into Full-Reference (FR), Reduced-Reference (RR), and No-Reference/blind (NR). Among them, the FR/RR LFIQA metrics are fully/partially based on the undisturbed information derived from reference LFIs. In contrast, the blind LFIQA metrics without using reference information are more feasible for most real-world scenarios. In the domain of blind LFIQA, many previous studies [33], [34], [35] have shown that the blind LFIQA metrics need to consider the additional effect of angular discontinuity on spatial quality. The hand-crafted feature-based blind LFIQA metrics typically extract Natural Scene Statistic (NSS) features from angular and spatial aspects, respectively, and then combine these features to assess the perceptual quality of LFIs. Obviously, these metrics fail to establish a deep relationship between angular and spatial information for quality evaluation. More recently, some researchers (*e.g.*, [36], [37]) attempted to extract angular and spatial features in a unified deep learning-based framework, aiming to establish a deeper relationship between these two features. For this goal, current deep blind LFIQA metrics generally adopt two pipelines. One pipeline is to first extract spatial features from each viewpoint image (*i.e.*, SAI), and then fuse all spatial features based on angular characteristics. This pipeline has two disadvantages. First, the computational complexity is very high since spatial feature extraction is employed for all SAIs. Second, original angular information will be lost after spatial feature extraction. The other pipeline is to simultaneously extract angular and spatial features from the low-dimensional representations of LFIs, such as Epipolar Plane Image (EPI). The weakness of this pipeline also lies in two aspects. First, whether the quality of LFIs can be adequately reflected by its low-dimensional representations is still unclear. Second, the resulting features are dominated by spatial information since spatial variations are much larger than angular variations. Although angular information is important, existing metrics based on the above two pipelines more or less ignore the angular effect on spatial quality, resulting in unsatisfactory quality evaluation performance. Therefore, how to effectively and elegantly model the angular-spatial effect in the design of deep blind LFIQA metrics still deserves further investigation.

Enlightened by the above analyses, in this paper, we develop a new deep blind LFIQA metric by effectively modeling the angular-spatial effect of LFIs. Different from previous works, the underlying design principle of the proposed metric is to minimize the imbalance of large spatial variations and subtle angular variations, and preserve the angular effect in the original information for subsequent angular-spatial effect modeling. Specifically, we first model the angular effect of different viewpoints without introducing spatial information, to prevent the loss of angular information. After that, we model the local angular-spatial effect and establish the global relationship between different local regions to predict the global quality. In addition, angular-spatial effect is more pronounced in some local regions [38], [39], and the quality of these regions is more consistent with the global quality, which facilitates the quality-aware feature learning. To this end, we propose to utilize the discriminative local regions for auxiliary learning to improve the learning efficiency and prediction accuracy. Further, considering that the relationship between principal and auxiliary learning directly affects the training outcome, a reasonable weighting scheme throughout training is a prerequisite for achieving better result. The contributions of this paper are summarized as follows.

- We propose a novel blind LFIQA metric named ASEM-BLiF, which effectively models the angular-spatial effect by addressing the imbalance problem between angular and spatial information. Specifically, we first present an Angular Effect Modeling (AEM) module to capture the angular information independently of spatial information, and then model the angular-spatial effect for quality assessment via Angular-Spatial Quality Learning (ASQL).
- We design an auxiliary learning branch based on Discriminative Region Selection (DRS) to improve the learning efficiency and prediction accuracy from a local perspective. Further, a Dynamic Weighting Loss (DWLoss) is presented to balance the relationship between principal and auxiliary learning throughout the training process.
- We conduct extensive experiments on five representative LFIQA databases with the state-of-the-arts. The experimental results demonstrate that the proposed metric performs better than the existing LFIQA metrics by a significant margin, while having a faster running time than most existing blind LFIQA metrics.

The reminder of this paper is organized as follows. Section II describes the related works. Section III introduces the proposed metric in detail. Section IV exhibits the experimental results. In Section V, conclusions will be drawn.

II. Related Works

A. Representations of LFIs

As aforementioned, the LFI can be described by a biplane model $L(u, v, h, w)$, but it is still difficult to imagine this 4D format. A solution to this challenge is to observe the underlying data along with a subset of dimensions. To this end, an LFI can be visualized in several low-dimensional representations [40], [41], and they are summarized as follows.

- Sub-Aperture Image (SAI). Taking the uv plane as a set of camera views and the hw plane as their focal plane, an LFI can be represented as a 2D array of pinhole views, and each view is a 2D image called SAI.
- MicroLens Image (MLI). By collecting all rays from different viewpoints of the uv plane approaching to the hw plane, an LFI can be visualized as a 2D image with high spatial resolution.
- Epipolar Plane Image (EPI). By vertically stacking the rows h from SAIs in a fixed angular row u, one can obtain a 2D horizontal EPI. A 2D vertical EPI can be obtained in a similar manner.

- Pseudo Video Sequence (PVS). PVS is a 3D representation of the LFI, created by arranging each SAI as a frame and displaying all SAIs in a certain order.
- Refocused Image (RI). A 2D refocused image containing focus and defocus regions is generated by superimposing multi-views with a specific slope related to the disparity.

B. Quality Assessment of LFIs

The FR/RR LFIQA metrics [38], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53] evaluate the perceptual quality of distorted LFIs by using the full/partial reference information. For example, Fang et al. [43] propose a FR LFIQA metric by calculating the similarity between the gradient magnitudes of reference and distorted SAIs and EPIs. Paudyal et al. [52] present a RR metric for LFIQA, in which the depth map similarity between reference and distorted LFIs is exploited as the predicted quality score. However, the practical application of FR/RR LFIQA metrics is very limited because reference LFIs are often not available, the research of blind LFIQA metrics has thus received more attention.

The traditional handcrafted feature-based blind LFIQA metrics [33], [34], [35], [39], [54], [55], [56], [57], [58], [59], [60], [61] generally extract angular and spatial NSS features, and then utilize non-linear regression models [62] to produce the quality score. For example, Shi et al. [54] design a Blind quality Evaluator of Light Field image (BELIF), in which the principal component of cyclopean image array is firstly generated, then the naturalness and structural similarity index are extracted to assess the spatial and angular quality degradation, respectively. Shi et al. [55] further propose a blind LFIQA metric named NR-LFQA, in which the spatial and angular quality are measured based on the naturalness distribution features of the cyclopean image array and the global and local features of EPIs, respectively. Zhou et al. [33] present a Tensor oriented-based blind LFIQA metric called Tensor-NLFQ, which adopts Tucker decomposition to obtain the principal components of four oriented SAI stacks, the global naturalness and local frequency features are extracted to evaluate the spatial quality, and the structural similarity distributions are used to measure the angular consistency. Xiang et al. [57] propose a Visualization-based Blind quality assessment metric for LFIs (VBLFI), which calculates the mean difference image of LFIs and evaluates the perceptual quality using Curvelet transform. Based on VBLFI, Xiang et al. [58] additionally measure the angular quality deterioration on EPIs. Further, Xiang et al. [39] propose a blind LFIQA metric (PVRI) based on PVS and RIs, in which the angular quality is measured from the structure, motion and disparity information of PVS, and the spatial quality is evaluated from the depth and semantic information of RIs. Pan et al. [34] also employ Tucker decomposition on LFIs, and then utilize the sharpness and distribution information of tensor slice and the percentage of singular value to measure the quality deterioration in spatial and angular domains, respectively. Chai et al. [61] perform quality assessment by measuring angular consistency and spatial-angular features with texture and structure descriptors. Although the handcrafted feature-based blind LFIQA metrics take into account the angular inconsistency, they have limitation in establishing a deep connection between angular and spatial information for measuring quality degradation.

With the explosive development of deep learning, some deep blind LFIQA metrics [36], [37], [63], [64], [65] have been designed to extract angular and spatial features in a unified framework. As mentioned before, existing deep blind LFIQA metrics often follow two pipelines. Guo's metric [63] is the representative work of the first pipeline, in which spatial feature extraction is first employed on each SAI, and then angular information is exploited to fuse spatial features of different SAIs. However, the first pipeline is not only computationally inefficient, but also suffers from the loss of angular information. On the contrary, several existing deep blind LFIQA metrics adopt the second pipeline, which simultaneously extracts angular and spatial features from the low-dimensional representations of LFIs. For example, Zhao et al. [36] propose a Deep Light Field Image Quality Evaluator (DeLFIQE), in which 2D discriminative EPI patches containing both angular and spatial information are first generated, and then a Convolutional Neural Network-based (CNN-based) model is designed for feature extraction and quality assessment. Similarly, Alamgeer and Farias [37] propose a deep blind LFIQA metric named DNNF-LFIQA, which extracts CNN features from 2D horizontal and vertical EPIs in the frequency domain. Our previous work DeeBLiF [64] first generates 2D spatio-angular patches from LFIs, and then extracts angular and spatial features via a two-stream CNN model for quality evaluation. Fu et al. [65] develop a Stereo vision-based LFIQA metric called SvLFIQA. This metric explores the local quality and local significance of light field cyclopean image patches to predict the LFI quality. However, these metrics employing the second pipeline still have two disadvantages: First, the relationship between the generated low-dimensional representations and the LFI quality is unclear. Second, spatial information is dominant in feature extraction and quality degradation learning because spatial variations are much larger than angular variations, which leads to insufficient measurement of angular distortions.

To sum up, it can be found that all the existing deep blind LFIQA metrics take into account angular and spatial information. However, compared to previous blind LFIQA metrics (including our previous work DeeBLiF), several innovations are incorporated into the proposed ASEM-BLiF metric: First and foremost, existing metrics suffer from the imbalance problem between angular and spatial information. To alleviate this imbalance, ASEM-BLiF presents a more efficient manner for angular-spatial effect modeling, in which angular effect is modeled without introducing spatial information, and then the relationship between angular and spatial information is subsequently established for LFI quality evaluation. Besides, ASEM-BLiF utilizes local information for auxiliary learning to improve the learning efficiency and prediction accuracy, which is neglected in existing metrics. Moreover, a fixed learning function is widely adopted in state-of-the-art metrics, but ASEM-BLiF exploits a dynamic learning scheme to achieve better training outcome. Finally, comprehensive experiments

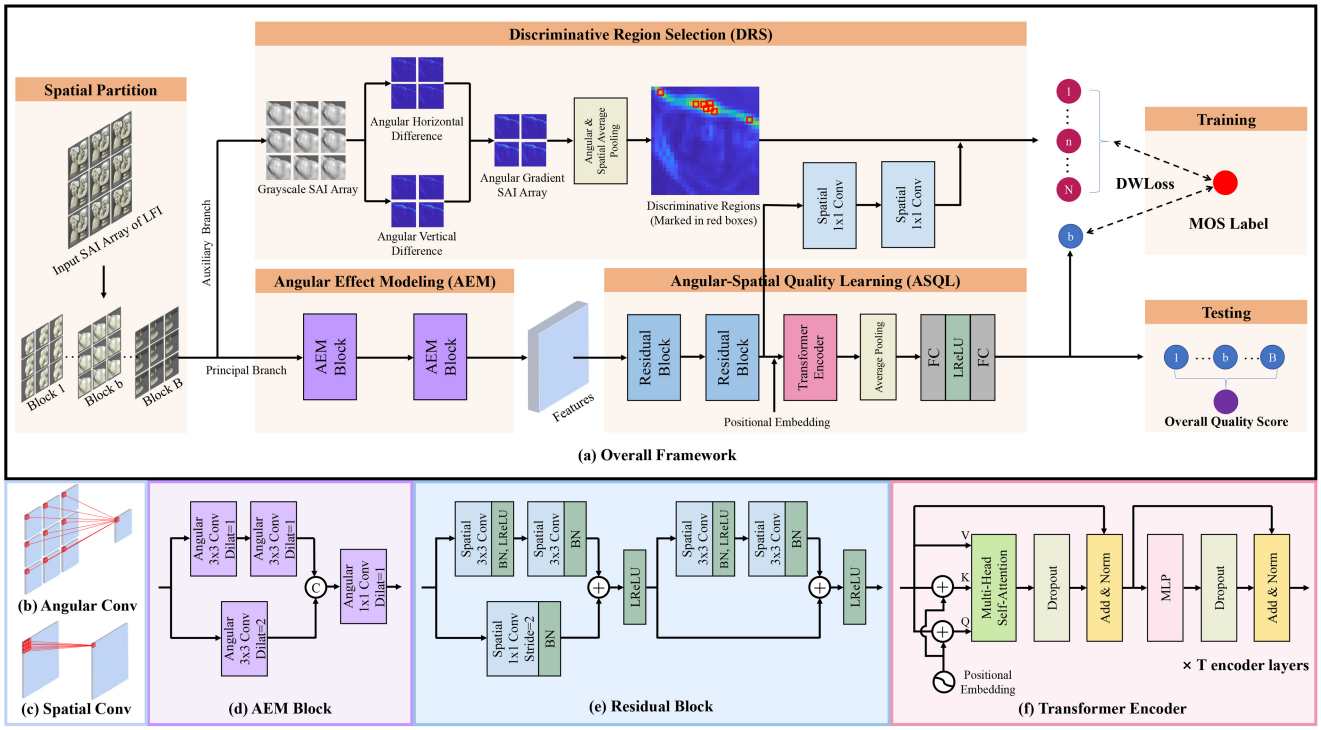


Fig. 1. Overview of the proposed ASEM-BLiF metric. For better visualization, the angular resolution of the input SAI array of LFI is set to 3×3.

are conducted on five LFIQA databases, which demonstrate the superiority of ASEM-BLiF in various aspects.

III. Proposed Metric

Fig. 1 presents the overview of the proposed ASEM-BLiF metric, which mainly consists of four components: AEM, ASQL, DRS, and DWLoss. Among them, the AEM module is used to capture the angular information independently of spatial information, while the ASQL module is designed for modeling angular-spatial effect and performing quality prediction. The design of the AEM and ASQL modules is motivated by the fact that subtle LFI angular variations are easily affected by large LFI spatial variations, which results in the loss of angular information and the reduction of the discrimination ability against angular distortions. Further, due to the lack of the consideration of local information in the quality learning process, the DRS module is explored to construct the auxiliary learning branch, aiming to improve the learning efficiency and prediction accuracy from a local perspective. Finally, considering that a reasonable weighting scheme promotes better training results, DWLoss is developed to balance the relationship between principal and auxiliary learning throughout the training process.

Let $\mathcal{L} \in \mathbb{R}^{U\times V\times H\times W\times C}$ denote the input SAI array of LFI, where $U\times V$ and $H\times W$ are the angular and spatial resolutions, respectively, and C denotes the RGB color channels. Inspired by [64], we first spatially partition $\mathcal{L}$ into B overlapping blocks to ensure a sufficient training set, denoted as $\mathcal{LB}_b \in \mathbb{R}^{U\times V\times S\times S\times C}$, $b = 1, 2, \ldots, B$, where S is the spatial size of the block. Then two branches are designed, the principal branch participates in both training and test stages, while the auxiliary branch is used only in the training stage. In the principal branch, the angular information of the input block is first captured by the AEM module, followed by the ASQL module for local angular-spatial effect modeling and quality prediction from a global perspective. In the auxiliary branch, the DRS module is utilized to selectively obtain discriminative regions with large angular sparsity, whose quality is subsequently predicted and exploited for auxiliary learning. During training, a DWLoss is further employed to achieve an optimal balance between the learning of principal and auxiliary branches. In the test stage, the overall LFI quality score is obtained by averaging the predicted quality score of all blocks. All components are described in the following subsections.

A. Angular Effect Modeling (AEM)

Subtle angular variation is one of the most distinctive characteristics of LFIs [66]. With the final goal of modeling angular-spatial effect for quality assessment, the design of AEM module is inspired by the following two points to handle the angular information: First, to minimize the imbalance of large spatial variations and subtle angular variations, the raw angular information should be captured individually rather than captured simultaneously with spatial information. Second, angular hierarchical features extracted from different angular disparities are beneficial to encode the angular information with subtle variations.

As shown in Fig. 1(a), the proposed AEM module consists of two AEM blocks (Fig. 1(d)). In each AEM block, two subbranches with angular convolutions (Fig. 1(b)) are designed in parallel to extract angular hierarchical features without considering any spatial information. Note that, due to the low

angular resolution (typically 9×9) of LFIs, the padding of all angular convolutions is set to zero to avoid introducing irrelevant information. Specifically, as shown in Fig. 1(b), the top sub-branch encodes the small angular disparity by utilizing two 3×3 angular convolutions with a dilation rate of 1, while the bottom sub-branch uses a 3×3 angular convolution with a dilation rate of 2 to encode the large angular disparity. The design motivation is to ensure that both sub-branches have the same output size, which facilitates the subsequent combination. Finally, features of these two sub-branches are concatenated and fused by a 1×1 angular convolution. Given an input block $\mathcal{LB} \in \mathbb{R}^{U\times V\times S\times S\times C}$, the above process is described as,

$$\boldsymbol{f}_A = \boldsymbol{\phi_A}(\mathcal{LB}) \tag{1}$$

where $\boldsymbol{f}_A \in \mathbb{R}^{S\times S\times C}$ denotes the output features of the AEM module $\boldsymbol{\phi_A}(\cdot)$.

B. Angular-Spatial Quality Learning (ASQL)

The main objective of ASQL module is to model the angular-spatial effect and further predict the global quality. Firstly, two residual blocks (Fig. 1(e)) based on the widely-used ResNet [67] are performed to model the local angular-spatial effect. The motivation is that the angular information from multiple viewpoints is captured in the channel dimension of $\boldsymbol{f}_A$, so the angular-spatial effect can be modeled by conventional spatial convolutions (Fig. 1(c)) which simultaneously extract spatial and channel features. Note that such a feature extraction part can be easily extended by using other mainstream backbones. Here, we use only a small number of spatial convolutions for feature extraction to prevent spatial information from dominating in the angular-spatial effect. Let the above feature extraction denote as $\boldsymbol{\phi_L}(\cdot)$,

$$\boldsymbol{f}_L = \boldsymbol{\phi_L}(\boldsymbol{f}_A) \tag{2}$$

where $\boldsymbol{f}_L \in \mathbb{R}^{m\times m\times d}$ denotes the generated local angular-spatial features. Here, m is equal to $S/4$ since we adopt two residual blocks and each block halves the spatial resolution. d denotes the depth of feature maps.

However, using a small number of spatial convolutions with small receptive field inevitably lacks the consideration of global content information [68]. Different local regions may have different angular-spatial effect and local quality [38], thus it is necessary to establish the relationship between the local regions and the global quality. To this end, we subsequently adopt the Transformer derived from DETR [69] to achieve this goal. Specifically, we convert the generated local angular-spatial features $\boldsymbol{f}_L$ into a sequence of features with size $m^2\times d$, where m^2 and d represent the sequence length and feature depth, respectively.

$$\boldsymbol{f}_L = [\boldsymbol{f}_1; \boldsymbol{f}_2, \ldots, ; \boldsymbol{f}_{m^2}], \quad \boldsymbol{f}_L \in \mathbb{R}^{m^2\times d} \tag{3}$$

To mine the self-attention relationship between the local regions, Multi-head Self-Attention (MSA) [69] extended from the standard Query-Key-Value-Self-Attention (QKV-SA) architecture is adopted. First, the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ matrices are separately obtained from the input sequence $\boldsymbol{f}_L$. Note that, since the position information of local regions is lost after the operation in Eq. (3), an extra learnable positional embedding $\boldsymbol{p}_e \in \mathbb{R}^{m^2\times d}$ is added to the sequence before generating the $\boldsymbol{Q}$ and $\boldsymbol{K}$ matrices [69],

$$\boldsymbol{Q} = (\boldsymbol{f}_L + \boldsymbol{p}_e)\boldsymbol{W}_q, \quad \boldsymbol{W}_q \in \mathbb{R}^{d\times d'} \tag{4}$$

$$\boldsymbol{K} = (\boldsymbol{f}_L + \boldsymbol{p}_e)\boldsymbol{W}_k, \quad \boldsymbol{W}_k \in \mathbb{R}^{d\times d'} \tag{5}$$

$$\boldsymbol{V} = \boldsymbol{f}_L\boldsymbol{W}_v, \quad \boldsymbol{W}_v \in \mathbb{R}^{d\times d'} \tag{6}$$

where d' is set to d/Y, Y is the head number of MSA, $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ are the linear projection matrices of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. Then, let $\boldsymbol{x}_0$ denote the combination of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, we obtain $MSA(\boldsymbol{x}_0)$ as,

$$SA_y(\boldsymbol{x}_0) = Softmax\left(\boldsymbol{Q}\boldsymbol{K}^{\mathsf{T}}/\sqrt{d'}\right)\boldsymbol{V} \tag{7}$$

$$MSA(\boldsymbol{x}_0) = Concat(SA_1(\boldsymbol{x}_0); \cdots ; SA_Y(\boldsymbol{x}_0))\boldsymbol{W}_{ln} \tag{8}$$

where $\boldsymbol{W}_{ln} \in \mathbb{R}^{d\times d}$ is the linear projection matrix used in the final step before obtaining the MSA output.

Based on MSA, the standard Transformer encoder consisting of T encoder layers is designed linearly. As shown in Fig. 1(f), each encoder layer contains MSA, Multi-Layer Perception (MLP), Dropout (Drop), Addition, and Normalization (Norm) [70] operations. As a result, we generate the self-attention features $\boldsymbol{x}_T$, and further obtain the global features $\boldsymbol{f}_G$ by Average (Avg) pooling,

$$\boldsymbol{x}'_t = Norm(Drop(MSA(\boldsymbol{x}_{t-1})) + \boldsymbol{x}_{t-1}), \ t = 1, \ldots, T \tag{9}$$

$$\boldsymbol{x}_t = Norm\big(Drop\big(MLP\big(\boldsymbol{x}'_{t-1}\big)\big) + \boldsymbol{x}'_{t-1}\big), \ t = 1, \ldots, T \tag{10}$$

$$\boldsymbol{f}_G = Avg(\boldsymbol{x}_T), \quad \boldsymbol{f}_G \in \mathbb{R}^{1\times d}, \ \boldsymbol{x}_T \in \mathbb{R}^{m^2\times d} \tag{11}$$

Finally, two Fully Connected (FC) layers and one LeakyReLU (LReLU) activation layer are adopted for producing global quality score $\boldsymbol{Q}_G$ of the input block,

$$\boldsymbol{Q}_G = FC\big(LReLU\big(FC\big(\boldsymbol{f}_G\big)\big)\big) \tag{12}$$

The above processes model the local angular-spatial effect and establish the global relationship between different local regions to predict the global quality, which are viewed as the principal branch of the proposed metric.

C. Discriminative Regions Selection (DRS)-Based Auxiliary Learning

In addition to the learning of the global quality prediction in the principal branch, we further design an auxiliary branch to exploit the local angular-spatial effect $\boldsymbol{f}_L$ for learning from a local perspective. As shown in Fig. 1(a), we first predict the local quality score for each spatial position of $\boldsymbol{f}_L$ with two 1×1 convolutions,

$$\boldsymbol{Q}_L = Conv_{1\times 1}\big(\boldsymbol{f}_L\big) \tag{13}$$

where $\boldsymbol{Q}_L \in \mathbb{R}^{m\times m}$ denotes the predicted local quality scores of different spatial positions.

As widely discussed in many studies [71], [72], [73], [74], when an image suffers from distortions, the quality of different regions in the image will be affected to varying degrees. In addition, since human eyes tend to focus on the anomalies, errors, and incoherent regions of an image, these low-quality

regions are more likely to negatively impact the global quality. Along this vein, we argue that different regions in an LFI contain different angular-spatial effect and corresponding local quality, and only the quality of certain regions has a close correlation with the global LFI quality. Therefore, adopting all region quality as supplementary information for learning will be detrimental to the training process (see Section IV-F). As a result, we propose a DRS module and utilize discriminative regions for auxiliary learning to improve the learning efficiency and prediction accuracy. Since the angular-spatial effect denotes the effect of angular inconsistent on spatial quality, regions with large angular sparsity are regarded as discriminative regions in our metric, which contain strong angular-spatial effect [38], [39]. As shown in Fig. 1(a), for an input block $\mathcal{LB}$, the grayscale SAI array, denoted as $\mathcal{G} \in \mathbb{R}^{U\times V\times S\times S}$, is first generated. Then the angular horizontal difference $\mathcal{D}^{hor}_{u,v}$ and angular vertical difference $\mathcal{D}^{ver}_{u,v}$ between adjacent SAIs are calculated as,

$$\mathcal{D}^{hor}_{u,v} = \mathcal{G}_{u+1,v} - \mathcal{G}_{u,v} \tag{14}$$

$$\mathcal{D}^{ver}_{u,v} = \mathcal{G}_{u,v+1} - \mathcal{G}_{u,v} \tag{15}$$

where $u = 1,\ldots, U\text{-}1$, $v = 1,\ldots, V\text{-}1$, and $\mathcal{G}_{u,v}$ denotes the SAI of (u, v) angular viewpoint in $\mathcal{G}$.

Then the angular gradient SAI array $\boldsymbol{G} \in \mathbb{R}^{(U-1)\times(V-1)\times S\times S}$ is calculated as,

$$\boldsymbol{G}_{u,v} = \sqrt{\mathcal{D}^{hor2}_{u,v} + \mathcal{D}^{ver2}_{u,v}} \tag{16}$$

where $\boldsymbol{G}_{u,v}$ denotes the SAI of (u, v) angular viewpoint in $\boldsymbol{G}$. Then we combine different angular viewpoints in $\boldsymbol{G}$ to $\boldsymbol{G}_a \in \mathbb{R}^{S\times S}$ using angular average pooling $\boldsymbol{\psi}_A(\cdot)$,

$$\boldsymbol{G}_a = \boldsymbol{\psi}_A(\boldsymbol{G}) = \frac{1}{(U-1)(V-1)} \sum_{u=1}^{U-1}\sum_{v=1}^{V-1} \boldsymbol{G}_{u,v} \tag{17}$$

The generated $\boldsymbol{G}_a$ has the same spatial resolution as the input $\mathcal{LB}$ and stores the angular sparsity by recording the gradient intensity, in which regions with larger gradient intensity are more pronounced to have stronger angular-spatial effect. In order to select discriminative regions corresponding to $\boldsymbol{Q}_L$ with spatial resolution $m\times m$, we further reduce the spatial resolution of $\boldsymbol{G}_a$ by applying spatial average pooling $\boldsymbol{\psi}_S(\cdot)$ for non-overlapping $(S/m)\times(S/m)$ regions,

$$\boldsymbol{G}_s = \boldsymbol{\psi}_S(\boldsymbol{G}_a) \tag{18}$$

where $\boldsymbol{G}_s \in \mathbb{R}^{m\times m}$ is the gradient intensity corresponding to $\boldsymbol{Q}_L$.

Let $\boldsymbol{G}_s^{\Omega_N}$ denote the assembly of N ($0< N < m^2$) discriminative regions with the largest gradient intensity in $\boldsymbol{G}_s$, the angular-spatial effect of these regions can be exploited to improve the quality-aware feature learning. Here, N is a hyperparameter and its effect on the performance of the proposed metric will be further discussed in Section IV-F). Correspondingly, the predicted local quality scores of these discriminative regions are denoted as $\boldsymbol{Q}_L^{\Omega_N}$.

D. Dynamic Weighting Loss (DWLoss)

The training of the auxiliary branch is driven by minimizing the differences between the local quality scores $\boldsymbol{Q}_L$ of discriminative regions Ω_N and the Mean Opinion Score (MOS) label $\boldsymbol{Q}_{MOS}$, based on the Mean Square Error (MSE) function,

$$L_a = \frac{1}{N}\frac{1}{B_s}\sum_{n=1}^{N}\sum_{i=1}^{B_s}\left(\boldsymbol{Q}_L^{\Omega_n,i} - \boldsymbol{Q}_{MOS}^{i}\right)^2 \tag{19}$$

where L_a denotes the loss function of the auxiliary branch, B_s represents the batch size.

Similarly, the loss function of the principal branch is defined based on the MSE between global quality scores $\boldsymbol{Q}_G$ and the MOS label $\boldsymbol{Q}_{MOS}$,

$$L_p = \frac{1}{B_s}\sum_{i=1}^{B_s}\left(\boldsymbol{Q}_G^{i} - \boldsymbol{Q}_{MOS}^{i}\right)^2 \tag{20}$$

where L_p denotes the loss function of the principal branch.

Intuitively, the easiest way to construct the final loss function is to add L_p and L_a according to fixed weights [36], [63]. However, this naive way is suboptimal in our proposed metric. This is mainly attributed to the fact that the auxiliary branch is only involved during training, while the final training objective of our metric is the principal branch. In addition, the learning of local angular-spatial effect should serve as the basis for the learning of global quality prediction. Based on such a hypothesis, the auxiliary branch should be biased at the early training stage, while the principal branch should be prominent at the late training stage. As a result, we design a DWLoss to assign dynamic learning weights for principal and auxiliary branches,

$$DWLoss = \frac{Ep_c}{Ep_t}L_p + \left(1 - \frac{Ep_c}{Ep_t}\right)L_a \tag{21}$$

where Ep_c and Ep_t denote the number of current and total epochs, respectively.

At the beginning of the training process, the proposed DWLoss assigns the largest weight to the auxiliary branch, aiming to learn the angular-spatial effect from a local perspective. Subsequently, the weight of the auxiliary branch gradually decreases, while the weight of the principal branch gradually increases. Finally, the global quality prediction is dominant at the end of the training, leading to better training result than the naive way.

E. Implementation Details

We implement the proposed metric using Pytorch library, with the hardware configurations of Intel i7-10700 CPU, NVIDIA GeForce RTX 3080 GPU, and 64G RAM Memory. Our metric is trained for 50 epochs, and the model of the last epoch is used to report the performance. The model parameters are updated using a mini-batch Stochastic Gradient Descent (SGD) optimizer with a weight momentum of 0.9 and a decay of 0.0001 [64]. The model is trained from scratch with the batch size of 8 and the learning rate of 0.001. In order to ensure a large enough training set and avoid excessive overlap between different blocks, we spatially partition each

LFI into 25 blocks (*i.e.*, B=25) with spatial size 112×112 (*i.e.*, S=112). No other data augmentation methods are used except horizontal flipping. The number of the selected discriminative regions is set to 20. The head number of MSA is set to 8, while the encoder layer number T is set to 4. All input blocks cropped from the same LFI use the MOS label of their source LFI as training targets [36].

IV. Experiments

A. Databases

In our experiments, five benchmark LFIQA databases are used, including Win5-LID [75], NBU-LF1.0 [76], LFDD [77], VALID10bit [78], and SHU [79].

- Win5-LID database contains 220 distorted LFIs generated from 6 real-world and 4 synthetic reference scenes. There are 4 distortion types with 5 distortion levels, including HEVC, JPEG2k, Linear interpolation (LN), and Nearest Neighbor interpolation (NN), and 2 CNN-based distortion types with only 1 distortion level. The MOS label ranged from 1 to 5 is provided.
- NBU-LF1.0 database consists of 8 real-world and 6 synthetic reference scenes, based on which 210 distorted LFIs are obtained with 5 distortion types and 3 distortion levels. These distortion types include NN, Bicubic Interpolation (BI), learning based reconstruction (EPICNN), disparity map based reconstruction (Zhang), and spatial super-resolution reconstruction (VDSR). The MOS label on a 5-point discrete scale is presented.
- LFDD database has 8 synthetic reference scenes and 480 distorted LFIs. A total of 12 common distortion types are included, each of which contains 5 distortion levels. The database provides the MOS label ranged from 1 to 5.
- VALID10bit database has 5 reference scenes and 100 distorted LFIs under 5 compression distortions, including HEVC, VP9, and 3 LFI compression methods. Each distortion type has 4 distortion levels. The database contains the MOS label ranged from 1 to 5.
- SHU database contains 8 real-world reference scenes. Based on which, 240 distorted LFIs are generated with 5 distortion types and 6 distortion levels, including JPEG, JPEG2k, Gaussian blur, white noise, and motion blur. The MOS label ranged from 0 to 5 is presented.

For all databases, the distorted LFIs generated from real-world and synthetic reference scenes are of spatial resolutions 434×625 and 512×512, respectively. The distorted LFIs from different databases have different original angular resolutions (ranging from 9×9 to 15×15). However, due to the hardware limitations of light field cameras, the quality of the edge views of LFIs is impacted to some extent. Therefore, we only use the central 9×9 angular views of LFIs for all databases in our metric.

B. Experimental Settings and Evaluation Criteria

K-fold cross-validation is adopted as the train-test split strategy to conduct our experiments. For each database, all distorted LFIs are divided into K folds according to their reference scenes, which guarantees that the scenes in the training and test sets are completely independent. Here, K is set to half of the reference scene number in each database. For example, Win5-LID database has 10 reference scenes, so K is set to 5. As a result, each fold contains the distorted LFIs from two reference scenes. In the experiments, we train the model on K-1 randomly-selected folds and test the performance on the remaining fold. Consequently, we loop through all train-test splits and take the average result as the reported performance.

To evaluate the performance, four standard criteria are adopted, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), and Root Mean Square Error (RMSE). Among them, PLCC evaluates the linear relationship, SROCC and KROCC measure the monotonicity, and RMSE focuses on the prediction accuracy. As suggested in [80], the predicted quality score is passed through a five-parameter non-linear logistic mapping function before calculating PLCC and RMSE,

$$\widetilde{\boldsymbol{Q}} = \beta_1 \left\{ \frac{1}{2} - \frac{1}{1 + e^{\beta_2 (\boldsymbol{Q} - \beta_3)}} \right\} + \beta_4 \boldsymbol{Q} + \beta_5 \tag{22}$$

where β_{1-5} are the fitting parameters, $\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Q}}$ represent the quality prediction and its non-linear mapping result, respectively.

C. Overall Performance Comparison

In this subsection, our proposed ASEM-BLiF metric is compared with plenty of state-of-the-art IQA metrics, including five blind 2DIQA metrics (PIQE [20], NIQE [21], GWH-GLBP [22], BRISQUE [23], and BMPRI [25]), three blind 3DIQA metrics (Xu's [26], SINQ [27], and BSVQE [28]), four FR LFIQA metrics (MDFM [42], Fang's [43], Min's [44], and Meng's [45]), and eleven blind LFIQA metrics (BELIF [54], NR-LFQA [55], Tensor-NLFQ [33], VBLFI [57], DSA [58], PVRI [39], TSSV-LFIQA [34], 4D-DCT-LFIQA [35], DeLFIQE [36], DNNF-LFIQA [37], and DeeBLiF [64]). Among them, DeLFIQE [36], DNNF-LFIQA [37], DeeBLiF [64] and our proposed ASEM-BLiF are deep learning-based metrics, while other metrics are based on handcrafted features. For fair comparison, the learning-based metrics are executed based on K-fold train-test splits, while the non-learning-based metrics are directly performed on the same test sets. For blind 2DIQA metrics, we evaluate the quality of individual SAIs and take the average as the overall LFI quality. For blind 3DIQA metrics, we average the quality of every two horizontal adjacent SAIs as the overall LFI quality. To avoid bias, we reproduce the performance of all metrics on the same hardware configurations (Section III-E), using the released codes/features and default parameter settings from the corresponding papers.

The experimental results are shown in TABLE I. It can be seen that due to the limited consideration of angular inconsistency, traditional 2D/3D IQA metrics perform much worse than LFIQA metrics in quality evaluation. In addition, due to the diversity of reference scenes, distortion types and levels of the adopted LFIQA databases, existing FR/blind LFIQA metrics generally struggle to perform well on all

TABLE I
OVERALL PERFORMANCE COMPARISON ON FIVE LFIQA DATABASES. EXCLUDING OUR PROPOSED METRIC, THE BEST AND SECOND-BEST PERFORMANCE ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY, WHILE THE PERFORMANCE OF OUR PROPOSED METRIC IS SHOWN IN **BOLD**. THE DEEP LEANING-BASED AND HANDCRAFTED FEATURE-BASED METRICS ARE MARKED WITH AND WITHOUT *, RESPECTIVELY

Metric Types	Metrics	Win5-LID			NBU-LF1.0			LFDD			VALID10bit			SHU			Overall		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
blind 2DIQA	PIQE [20]	0.5264	0.4138	0.8281	0.2520	0.1425	0.8616	0.3938	0.3511	1.0177	0.8511	0.8627	0.9684	0.7751	0.7914	0.6920	0.5031	0.4526	0.8916
	NIQE [21]	0.6372	0.4462	0.7486	0.4472	0.3384	0.7987	0.5379	0.4219	0.9415	0.8044	0.7053	0.5333	0.9291	0.9106	0.4064	0.6366	0.5287	0.7482
	GWH-GLBP [22]	0.4717	0.2393	0.8527	0.5436	0.3600	0.7458	0.5101	0.3008	0.9473	0.7709	0.7248	0.5622	0.6227	0.5602	0.8395	0.5515	0.3836	0.8453
	BRISQUE [23]	0.5998	0.4636	0.7680	0.5291	0.3883	0.7621	0.4993	0.3914	0.9828	0.7701	0.6262	0.5674	0.9138	0.8891	0.4399	0.6232	0.5179	0.7704
	BMPRI [25]	0.5114	0.3938	0.8517	0.3312	0.1968	0.8445	0.6017	0.5347	0.8942	0.8675	0.8406	0.4579	0.9081	0.8850	0.4583	0.6205	0.5449	0.7598
blind 3DIQA	Xu's [26]	0.5868	0.4350	0.7884	0.4073	0.3135	0.8163	0.5542	0.5036	0.9309	0.7236	0.6787	0.6464	0.8369	0.8080	0.5961	0.6031	0.5320	0.7995
	SINQ [27]	0.5736	0.4467	0.7708	0.5413	0.4450	0.7502	0.5508	0.4882	0.9249	0.7189	0.5510	0.6013	0.9183	0.9029	0.4305	0.6372	0.5583	0.7476
	BSVQE [28]	0.5854	0.5266	0.7697	0.6830	0.6126	0.6572	0.6724	0.5665	0.8182	0.7327	0.5970	0.6085	0.7550	0.6792	0.7133	0.6796	0.5913	0.7457
FR LFIQA	MDFM [42]	0.7191	0.6519	0.6649	0.8627	0.8329	0.4489	0.5725	0.5311	0.9303	0.9164	0.9054	0.3709	0.8268	0.8537	0.6178	0.7234	0.6949	0.6980
	Fang's [43]	0.8198	0.7969	0.5519	0.8876	0.8466	0.4100	0.5832	0.4943	0.9191	0.9730	0.9578	0.2135	0.8846	0.8852	0.5114	0.7650	0.7189	0.6342
	Min's [44]	0.7290	0.6584	0.6660	0.7109	0.6550	0.6233	0.5550	0.4112	0.9383	0.9757	0.9325	0.2051	0.8541	0.8473	0.5690	0.7029	0.6211	0.7079
	Meng's [45]	0.6842	0.6344	0.7078	0.8518	0.8023	0.4664	0.3304	0.3483	1.0620	0.9579	0.9383	0.2629	0.9282	0.9187	0.4080	0.6452	0.6316	0.7101
blind LFIQA	BELIF [54]	0.5911	0.5402	0.8168	0.7402	0.7236	0.6059	0.7747	0.7217	0.7129	0.8099	0.7852	0.5437	0.8776	0.8513	0.5172	0.7592	0.7200	0.6621
	NR-LFQA [55]	0.7087	0.6341	0.6595	0.8499	0.8169	0.4645	0.6627	0.5919	0.8449	0.7687	0.7339	0.5686	0.9303	0.9300	0.3999	0.7621	0.7134	0.6408
	Tensor-NLFQ [33]	0.7441	0.7259	0.6443	0.7674	0.7270	0.5584	0.8363	0.7933	0.6136	0.8778	0.8553	0.4477	0.8878	0.8877	0.5011	0.8217	0.7934	0.5749
	VBLFI [57]	0.6828	0.6085	0.6999	0.8271	0.7781	0.4924	0.6994	0.6293	0.7970	0.8728	0.8398	0.4879	0.8903	0.8699	0.4725	0.7685	0.7137	0.6417
	DSA [58]	0.8364	0.7999	0.5308	0.8584	0.8131	0.4483	0.7225	0.6827	0.7302	0.8449	0.8202	0.4882	0.9325	0.9098	0.3879	0.8155	0.7798	0.5627
	PVRI [39]	0.7126	0.6713	0.6732	0.8162	0.7686	0.5121	0.7900	0.7314	0.6749	0.8338	0.7778	0.4831	0.8922	0.8351	0.4739	0.8039	0.7507	0.5933
	TSSV-LFIQA [34]	0.7336	0.6759	0.6611	0.8351	0.8033	0.4879	0.6927	0.6438	0.8021	0.8616	0.8187	0.4655	0.9077	0.8931	0.4611	0.7786	0.7381	0.6321
	4D-DCT-LFIQA [35]	0.8390	0.8189	0.5246	0.8331	0.8181	0.4941	0.8326	0.7842	0.6244	0.8535	0.8308	0.4666	0.9363	0.9244	0.3779	0.8554	0.8266	0.5250
	DeLFIQE* [36]	0.5581	0.4261	0.8067	0.8142	0.7655	0.5054	0.6470	0.5809	0.8592	0.4907	0.4254	0.8049	0.6619	0.5453	0.8143	0.6498	0.5654	0.7776
	DNNF-LFIQA* [37]	0.7164	0.6498	0.6826	0.7809	0.7531	0.5532	0.7577	0.7119	0.7363	0.8439	0.7825	0.4902	0.8428	0.7911	0.5676	0.7776	0.7287	0.6440
	DeeBLiF* [64]	0.8586	0.8382	0.5032	0.8574	0.8177	0.4597	0.8830	0.8111	0.5299	0.9283	0.8783	0.3443	0.9538	0.9430	0.3236	0.8916	0.8477	0.4589
	ASEM-BLiF* (Ours)	**0.9072**	**0.8949**	**0.4090**	**0.9087**	**0.8952**	**0.3613**	**0.9190**	**0.8707**	**0.4415**	**0.9399**	**0.9227**	**0.3069**	**0.9492**	**0.9463**	**0.3418**	**0.9227**	**0.8977**	**0.3924**

databases. However, the proposed ASEM-BLiF metric consistently achieves SOTA results on most databases. Especially on the two most complex databases Win5-LID and NBU-LF1.0, which contain both real-world and synthetic LFIs, our metric outperforms all IQA metrics by a significant margin. In most cases, the learning-based blind LFIQA metrics can achieve competitive or even superior performance than the FR LFIQA metrics. However, on the VALID10bit database, the FR LFIQA metrics generally obtain better results than the blind LFIQA metrics. One possible reason is that the VALID10bit database contains only 100 distorted LFIs, which may be insufficient to train a well-performing blind LFIQA model. Nevertheless, our metric still yields competitive performance on the VALID10bit database compared with the FR LFIQA metrics, while outperforming other blind LFIQA metrics significantly. The overall performance further demonstrates the superiority of our metric over other state-of-the-arts.

Moreover, TABLE I shows that compared with our proposed ASEM-BLiF metric, the performance of three existing deep learning-based metrics (*i.e.*, DeLFIQE [36], DNNF-LFIQA [37], and DeeBLiF [64]) is quite limited, and the possible reasons behind this deserve further investigation. First, DeLFIQE executes quality evaluation using a small number of EPI patches, thus its effectiveness is significantly constrained when dealing with LFIs of low angular resolution. In DNNF-LFIQA, each LFI is treated as an individual sample, and the limited size of the LFIQA database inevitably leads to overfitting during the training process. In contrast, DeeBLiF obtains better results by addressing the aforementioned two issues. However, DeeBLiF uses symmetric branches for angular and spatial information respectively, which leads to insufficient measurement of angular distortions. The proposed ASEM-BLiF metric strives to address the imbalance problem between angular and spatial information, thereby achieving superior quality evaluation performance.

Since K-fold cross-validation is adopted to conduct the experiments, each split has different training and test sets, resulting in different results. The average result reported in TABLE I is sensitive to outliers as it is a parametric measure of central tendency. In contrast, the median is a non-parametric measure that is robust to outliers. Therefore, we present the median PLCC/SROCC performance of blind LFIQA metrics for a more comprehensive performance comparison, as shown in Fig. 2. Here, due to the poor performance of DeLFIQE [36], we have removed it for better visualization. It can be seen that our metric still achieves outstanding performance on all databases compared to other metrics when considering the median result.

To further demonstrate the effectiveness of our metric in quality evaluation, we provide the scatter plots of the predicted quality score versus the MOS label for the best-performing handcrafted feature-based metric (4D-DCT-LFIQA [35]) and four deep learning-based metrics (DeLFIQE [36], DNNF-LFIQA [37], DeeBLiF [64], and our proposed ASEM-BLiF). As shown in Fig. 3, three databases with different numbers of predictions are used. The scatter plots of the Win5-LID, NBU-LF1.0, and LFDD databases are exhibited in the top, middle, and bottom rows, with 44, 30, and 120

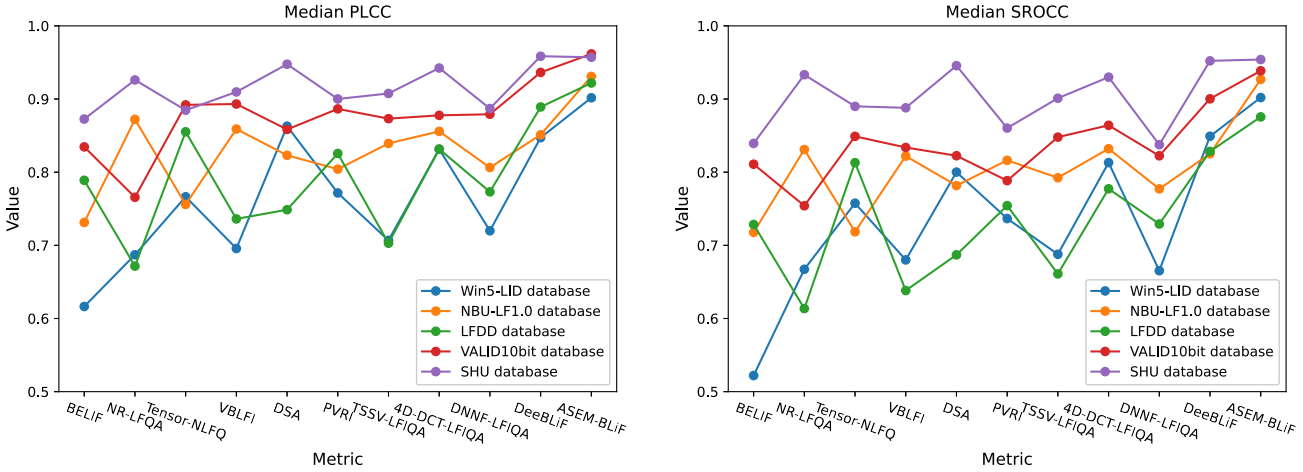


Fig. 2. Median PLCC (left) and SROCC (right) performance comparison of blind LFIQA metrics on five LFIQA databases.

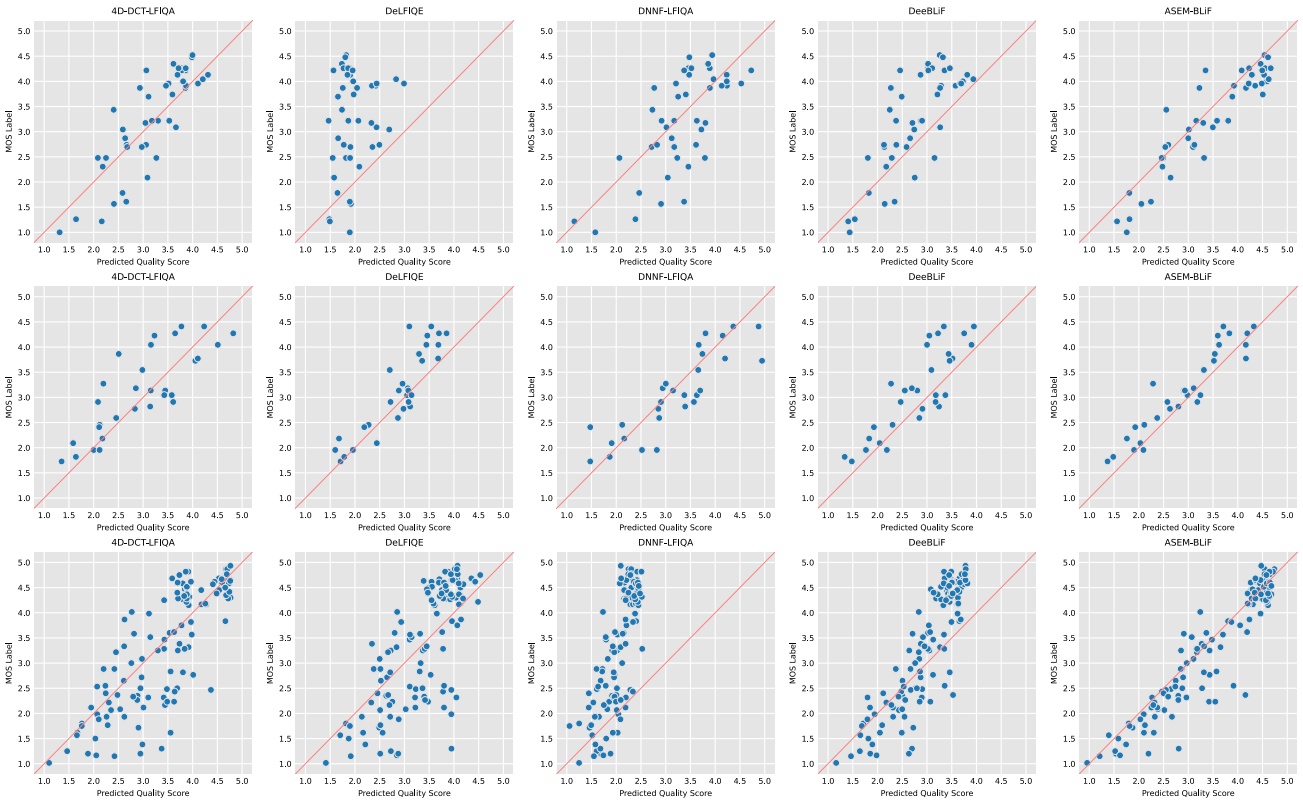


Fig. 3. Scatter plots of the predicted quality score versus the MOS label on the Win5-LID (top row), NBU-LF1.0 (middle row), and LFDD (bottom row) databases, where the red line represents the perfect prediction.

predictions each, respectively. The red line represents the perfect prediction. From the figure we can see that the predictions of our metric are more consistent with the subjective scores than other state-of-the-art metrics, demonstrating its superior ability to simulate the human visual perception.

D. Performance Comparison of Individual Distortion Types

An excellent IQA metric should show efficacy in various distortion types. Therefore, the performance of different IQA metrics for different distortion types is investigated and compared in this subsection. The experiments are performed on the Win5-LID and NBU-LF1.0 databases based on K-fold train-test strategy. Here we do not consider two CNN-based distortions in the Win5-LID database because both of them have only one distortion level. Due to the space constrains, we only provide the SROCC performance in TABLE II, as the PLCC/KROCC/RMSE performance show similar results. It can be observed that the FR LFIQA metrics typically perform better than the blind LFIQA metrics for individual

TABLE II
SROCC PERFORMANCE OF DIFFERENT DISTORTION TYPES ON THE WIN5-LID AND NBU-LF1.0 DATABASES. THE BEST TWO RESULTS ARE MARKED IN **BOLD**. HIT-COUNT TALLIES THE NUMBER OF TIMES EACH IQA METRIC OBTAINING A TOP-TWO RESULT. THE DEEP LEANING-BASED AND HANDCRAFTED FEATURE-BASED METRICS ARE MARKED WITH AND WITHOUT *, RESPECTIVELY.

Metric Types	Metrics	Win5-LID				NBU-LF1.0					Hit-count
		HEVC	JPEG2k	LN	NN	NN	BI	EPICNN	Zhang	VDSR	
blind 2DIQA	PIQE [20]	0.6894	0.8234	0.5661	0.3456	0.4612	0.5185	0.3305	0.3397	0.8694	0
	NIQE [21]	0.7695	0.7059	0.5029	0.3262	0.3469	0.4084	0.6019	0.5203	0.9020	0
	GWH-GLBP [22]	0.6420	0.5453	0.3950	0.4001	0.4694	0.4467	0.5987	0.5153	0.5755	0
	BRISQUE [23]	0.5452	0.5729	0.3572	0.4110	0.3878	0.4786	0.5248	0.5203	0.8612	0
	BMPRI [25]	0.3944	0.7057	0.5309	0.3188	0.3469	0.4466	0.4005	0.3029	0.3878	0
blind 3DIQA	Xu's [26]	0.6238	0.6587	0.5383	0.5178	0.4204	0.4459	0.4970	0.2910	0.8041	0
	SINQ [27]	0.4881	0.5826	0.6548	0.6284	0.5510	0.6546	0.4555	0.5127	0.8531	0
	BSVQE [28]	0.5730	0.5883	0.5649	0.5655	0.7959	0.7774	0.7138	0.3764	0.8939	0
FR LFIQA	MDFM [42]	0.9285	0.8496	0.7393	0.7598	0.9102	0.9324	0.6682	**0.8893**	0.8449	1
	Fang's [43]	0.9696	**0.9144**	**0.9046**	**0.8797**	0.8939	**0.9569**	0.7803	0.7988	0.9347	**4**
	Min's [44]	**0.9818**	**0.9387**	0.7261	0.4311	0.6816	0.7811	0.6459	0.7586	0.8857	2
	Meng's [45]	0.8919	0.7265	0.8294	0.7865	0.8041	**0.9326**	0.7460	0.6355	**0.9755**	2
blind LFIQA	BELIF [54]	0.7547	0.4547	0.6293	0.5715	**0.9265**	0.8467	0.4817	0.6768	0.8204	1
	NR-LFQA [55]	0.7728	0.6783	0.6280	0.5708	**0.9347**	0.8917	0.7095	0.6388	0.8531	1
	Tensor-NLFQ [33]	0.8434	0.8172	0.7931	0.8213	0.7551	0.8755	0.6508	0.6232	0.9020	0
	VBLFI [57]	0.7062	0.7617	0.6598	0.7183	0.9102	0.8999	0.7248	0.4793	0.9184	0
	DSA [58]	0.9016	0.9059	0.8161	0.7820	0.9020	0.8793	0.7552	0.5488	0.9020	0
	PVRI [39]	0.8591	0.7956	0.6512	0.7216	0.7878	0.8344	0.7326	0.8646	0.9102	0
	TSSV-LFIQA [34]	0.8179	0.6732	0.7176	0.7484	0.8694	0.9000	0.7137	0.6766	0.8041	0
	4D-DCT-LFIQA[35]	0.8896	0.8752	0.8004	0.8445	0.8694	0.9323	**0.7980**	**0.8652**	0.9265	2
	DeLFIQE* [36]	0.4949	0.4057	0.3402	0.5458	0.8776	0.9244	0.6808	0.6646	0.7224	0
	DNNF-LFIQA* [37]	0.7538	0.5866	0.7712	0.7270	0.9020	0.8918	0.6702	0.5618	0.7469	0
	DeeBLiF* [64]	0.9648	0.8195	0.7928	0.8306	0.9184	0.8876	0.7248	0.6961	0.8857	0
	ASEM-BLiF* (Ours)	**0.9697**	0.8617	**0.9059**	**0.9027**	0.9184	**0.9326**	**0.8133**	0.7996	**0.9429**	**6**

distortions. A possible explanation is that the availability of reference information can greatly help to distinguish the same distortion type but different distortion levels. In addition, we can also observe that the performance of most blind LFIQA metrics exhibits significant variations as the type of distortion changes. However, our metric robustly achieves competitive performance in most distortion types. Further, in certain types of distortions, our metric significantly outperforms state-of-the-art blind LFIQA metrics, and achieves competitive performance with existing FR metrics, *e.g.*, LN distortion in Win5-LID database and VDSR distortion in NBU-LF1.0 database. The above analyses fully demonstrate the robustness of our metric.

However, the performance of JPEG2k distortion in Win5-LID database and Zhang's distortion in NBU-LF1.0 database is somewhat unsatisfactory. The reasons behind deserve deeper investigation. For JPEG2k distortion in Win5-LID database, JPEG2k distorted LFIs are generated by adopting JPEG2k on each SAI, in which angular information does not directly participate in the generation process. In this case, spatial distortion is much more severe compared to angular distortion. However, our metric focuses on modeling the angular-spatial effect dominated by angular information. This may be the reason for the suboptimal performance of JPEG2k distortion. For Zhang's distortion in NBU-LF1.0 database, Zhang's distortion is a depth estimation-based angular distortion, in which different regions in a scene have different angular reconstruction quality according to the estimated depth. However, our metric is a block-based metric and assumes that all regions are of the same quality during training, which is not conducive to evaluate Zhang's distortion. Our previous work DeeBLiF also performs unsatisfactorily on Zhang's distortion due to the same reason.

E. Performance of Cross-Database Validation

Since the generalization ability is a crucial factor for designing an effective IQA metric, we investigate the cross-database performance of our proposed metric in this subsection. We conduct the cross-database experiments following most previous works [33], [35], [39]. Specifically, we first train our metric on the whole Win5-LID database due to its diversity and complexity in terms of image categories and distortion types. Then, we evaluate the performance on each distortion type that is shared between the Win5-LID database and other databases, including NBU-LF1.0 (NN), LFDD (JPEG2k), VALID10bit (HEVC), and SHU (JPEG2k). The results are shown in TABLE III. We can find that our metric still performs well on HEVC and JPEG2k even when trained on another database, implying a relatively good

TABLE III
PERFORMANCE OF TRAINING ON THE WIN5-LID DATABASE, AND TESTING ON THE NBU-LF1.0, LFDD, VALID10BIT, AND SHU DATABASES

Test databases	PLCC	SROCC	KROCC	RMSE
NBU-LF1.0 (NN)	0.6945	0.5547	0.3942	0.7513
LFDD (JPEG2k)	0.7355	0.7057	0.5151	0.6034
VALID10bit (HEVC)	0.7318	0.8585	0.6773	0.7288
SHU (JPEG2k)	0.8835	0.9061	0.7335	0.2120

TABLE IV
ABLATION STUDIES OF DIFFERENT MODULES ON THE WIN5-LID AND NBU-LF1.0 DATABASES. **BOLD** REPRESENTS THE BEST PERFORMANCE

Databases	Modules	PLCC	SROCC	KROCC	RMSE
Win5-LID	w/o AEM	0.5465	0.5538	0.3968	0.8085
	w/o DRS	0.8849	0.8633	0.6927	0.4552
	w/o Aux. Learn.	0.8816	0.8643	0.7018	0.4498
	w/o DWLoss	0.9003	0.8854	0.7226	0.4248
	ASEM-BLiF (Ours)	**0.9072**	**0.8949**	**0.7311**	**0.4090**
NBU-LF1.0	w/o AEM	0.6645	0.6254	0.4628	0.6531
	w/o DRS	0.9024	0.8712	0.7136	0.3747
	w/o Aux. Learn.	0.9066	0.8773	0.7195	0.3692
	w/o DWLoss	0.9057	0.8878	0.7327	0.3681
	ASEM-BLiF (Ours)	**0.9087**	**0.8952**	**0.7440**	**0.3613**

generalization ability on compression distortions. However, our metric shows relatively poor performance on the NN distortion. The reason for this result has been discussed in [39], that is, the NN distortion is implemented differently in the Win5-LID and NBU-LF1.0 databases. For Win5-LID database, the LFI containing 9×9 SAIs is reconstructed in five different distortion levels, each of which involves 50, 40, 30, 20, and 10 randomly selected SAIs to reconstruct the LFI, respectively. For NBU-LF1.0 database, the 9×9 SAIs are first down-sampled to 5×5, 3×3, and 2×2 fixed SAIs, respectively, and then the NN interpolation is used for reconstruction. We can easily find that the distortion levels between these two databases do not overlap, which leads to relatively poor performance on the NN distortion.

F. Ablation Studies

To further explore the effectiveness of each module of our metric, we conduct ablation experiments in this subsection. TABLE IV reports the experimental results on the Win5-LID and NBU-LF1.0 databases, where w/o is the abbreviation of *without*, and Aux. Learn. denotes the auxiliary learning branch. The table demonstrates a significant decrease in the performance of our metric on both databases when the AEM module is excluded, indicating the crucial role of capturing angular information for LFIQA. In addition, we can also find that using auxiliary learning without DRS module (*i.e.*, w/o DRS) performs even worse than not using auxiliary learning (*i.e.*, w/o Aux. Learn.). A possible explanation is that if all regions are introduced for auxiliary learning, the regions with insufficient angular-spatial information will lead to decreased performance. This demonstrates that the selected discriminative regions are necessary for auxiliary learning. Moreover, incorporating DWLoss into the training process can further promote learning towards better performance. Finally, the combination of all proposed modules culminates in enhanced effectiveness.

TABLE V
PERFORMANCE OF DIFFERENT NUMBER OF N ON THE WIN5-LID AND NBU-LF1.0 DATABASES. **BOLD** REPRESENTS THE BEST PERFORMANCE

Databases	N	PLCC	SROCC	KROCC	RMSE
Win5-LID	1	0.8731	0.8594	0.6916	0.4741
	5	0.8863	0.8683	0.6992	0.4477
	10	0.9015	0.8723	0.7034	0.4232
	20	0.9072	**0.8949**	**0.7311**	0.4090
	50	**0.9085**	0.8889	0.7276	**0.4083**
	100	0.8967	0.8746	0.7051	0.4309
	200	0.8869	0.8650	0.6940	0.4512
	w/o	0.8849	0.8633	0.6927	0.4552
NBU-LF1.0	1	0.8911	0.8787	0.7228	0.3976
	5	0.9059	0.8850	0.7314	0.3614
	10	0.9026	0.8876	0.7321	0.3771
	20	**0.9087**	**0.8952**	**0.7440**	**0.3613**
	50	0.9042	0.8857	0.7268	0.3719
	100	0.9045	0.8796	0.7228	0.3683
	200	0.9005	0.8795	0.7235	0.3807
	w/o	0.9024	0.8712	0.7136	0.3747

G. Hyperparameter Analyses

Since the performance of our metric is affected by the number of the selected discriminative regions, *i.e.*, N, we perform hyperparameter experiments on the Win5-LID and NBU-LF1.0 databases to investigate the impact of its value. As shown in TABLE V, the performance is compared when N is set to 1, 5, 10, 20, 50, 100, 200, and w/o. Here, w/o represents that all regions are incorporated into the learning process. From the table we can see that the performance is relatively poor if the value of N is too large or too small. The main reason could be that insufficient discriminative regions lead to inadequate information for auxiliary training, while excessive discriminative regions introduce too many low-quality predictions, ultimately impeding the learning process. Therefore, in our implementation, we set N to a moderate value of 20 which yields a satisfactory performance on both Win5-LID and NBU-LF1.0 databases.

H. Time Complexity Analyses

Time complexity is an important factor for an IQA metric as it affects the efficiency and practicality of the metric in real-world applications. Therefore, we conduct an analysis of the time complexity of our metric in comparison to other state-of-the-art metrics. All metrics are executed using the same hardware configurations as mentioned in Section III-E. Following [33], we measure the time complexity of each metric by testing a single LFI from the Win5-LID database,

TABLE VI
COMPARISON OF THE RUNTIME VERSUS THE OVERALL SROCC PERFORMANCE. **BOLD** REPRESENTS THE BEST PERFORMANCE. THE DEEP LEANING-BASED AND HANDCRAFTED FEATURE-BASED METRICS ARE MARKED WITH AND WITHOUT *, RESPECTIVELY

Metric Types	Metrics	Runtime (s/LFI)	SROCC
blind 2DIQA	PIQE [20]	4.2890	0.4526
	NIQE [21]	5.7384	0.5287
	GWH-GLBP [22]	3.4198	0.3836
	BRISQUE [23]	2.8354	0.5179
	BMPRI [25]	31.7935	0.5449
blind 3DIQA	Xu's [26]	57.2318	0.5320
	SINQ [27]	172.4818	0.5583
	BSVQE [28]	161.6847	0.5913
FR LFIQA	MDFM [42]	**0.8537**	0.6949
	Fang's [43]	1.1574	0.7189
	Min's [44]	3.9845	0.6211
	Meng's [45]	30.4872	0.6316
blind LFIQA	BELIF [54]	107.8814	0.7200
	NR-LFQA [55]	225.2069	0.7134
	Tensor-NLFQ [33]	697.6515	0.7934
	VBLFI [57]	65.6667	0.7137
	DSA [58]	198.5443	0.7798
	PVRI [39]	71.3578	0.7507
	TSSV-LFIQA [34]	44.6696	0.7381
	4D-DCT-LFIQA [35]	169.2623	0.8266
	DeLFIQE* [36]	15.4193	0.5654
	DNNF-LFIQA* [37]	3.7410	0.7287
	DeeBLiF* [64]	4.8533	0.8477
	ASEM-BLiF* (Ours)	5.2377	**0.8977**

TABLE VII
COMPARISON OF THE TIME CONSUMPTION VERSUS THE SROCC PERFORMANCE OF DEEP BLIND LFIQA METRICS ON THE WIN5-LID DATABASE. **BOLD** REPRESENTS THE BEST PERFORMANCE

Metrics	Pre-Processing & Training time (h)	Runtime (s/LFI)	SROCC
DeLFIQE [36]	1.1518	15.4193	0.4261
DNNF-LFIQA [37]	10.0733	**3.7410**	0.6498
DeeBLiF [64]	**0.9439**	4.8533	0.8382
ASEM-BLiF (Ours)	1.4872	5.2377	**0.8949**

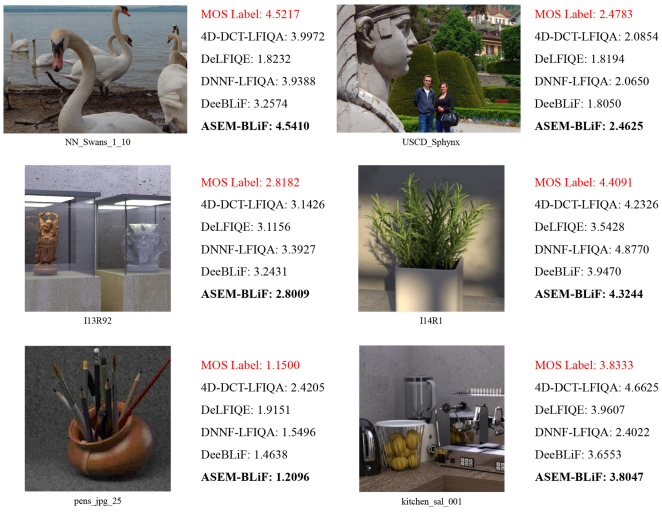


Fig. 4. Illustrative quality predictions of the proposed ASEM-BLiF metric and four state-of-the-art metrics on the Win5-LID (top row), NBU-LF1.0 (middle row), and LFDD (bottom row) databases. **Bold** represents the best prediction.

denoted as the runtime in our experiments. Although the deep learning-based metrics can be accelerated using GPU, we report the runtime of all metrics using CPU only for fair comparison. The handcrafted feature-based metrics are implemented by MATLAB, while the deep learning-based metrics are implemented by Python. TABLE VI shows the runtime versus the overall SROCC performance. We can see that most handcrafted feature-based blind LFIQA metrics are time-consuming and achieve unsatisfactory performance. As comparison, previous deep blind LFIQA metrics have faster running times, but still struggle to perform well in quality evaluation task. However, our proposed metric outperforms other state-of-the-art metrics with a significant margin and a relatively low time complexity, which further demonstrates the effectiveness and efficiency of our metric.

Compared to the handcrafted feature-based metrics, the deep learning-based metrics consume more time on pre-processing and model training, which are also crucial factors in computational complexity that need to be investigated. Therefore, we summarize the time consumption of deep blind LFIQA metrics versus the SROCC performance on the Win5-LID database, as shown in TABLE VII. It can be found that although the training time consumption of our metric is slightly higher than that of DeLFIQE and DeeBLiF, it is still within an acceptable range. Considering the outstanding performance, our proposed metric achieves a better trade-off between computational efficiency and prediction accuracy.

I. Illustrative Examples

To provide a more intuitive example of quality evaluation, we present some illustrative predictions of our proposed metric, along with four state-of-the-art metrics, on the Win5-LID, NBU-LF1.0, and LFDD databases. Due to space limitations, we only display the central viewpoint of each LFI. As shown in Fig. 4, we present a series of LFIs with diverse MOS labels, distortion types and reference scenes, along with the corresponding predictions from five metrics. It can be found that despite the diverse characteristics of LFIs, our proposed metric exhibits superior performance in accurately predicting the LFI quality compared to other metrics.

J. Discussion

The above experimental results have fully demonstrated the superiority of the proposed ASEM-BLiF metric in terms of prediction accuracy, time complexity, and robustness. To delve further in TABLE IV, it can be found that the incorporation of the AEM module contributes most to the final result, while each of the other modules (*e.g.*, DRS and DWLoss)

motivates a better training outcome and slightly improves the final performance. In other words, the performance achieved by ASEM-BLiF mainly attributed to the minimization of the imbalance caused by large spatial variations and subtle angular variations, indicating the importance of angular-spatial effect modeling in the LFIQA task. However, despite the remarkable performance of the proposed metric, two limitations can still be observed. First, the quality of all LFI blocks is assumed to be equally important during training and testing, which may not align with the principles of human visual perception, thus limiting the performance of quality evaluation. Second, the proposed metric is only applicable to 4D LFIs with two angular dimensions due to the use of angular convolutions, but not to 3D LFIs with one angular dimension, such as LFIs in [66].

In addition to the static LFIs, some researchers additionally capture the temporal information of light fields and generate light field videos [81]. In real-world scenarios, visual signals are rarely presented without any auxiliary information. They are often presented alongside other information like audio [82], [83], [84] and text [85], which collectively shape the user-perceived quality of experience. Therefore, in the long run, we argue that the subjective and objective quality evaluation of light fields can take these factors into consideration, to gain a comprehensive understanding of human visual perception.

V. Conclusion

In this paper, we propose a novel blind LFIQA metric by effectively modeling the angular-spatial effect, which is abbreviated as ASEM-BLiF. In comparison to previous works, our metric handles the angular and spatial information in a significantly distinct manner. Specifically, we first present an Angular Effect Modeling (AEM) module to capture the angular information independently of spatial information. Then, we propose an Angular-Spatial Quality Learning (ASQL) module to model the local angular-spatial effect and establish the global relationship between different local regions for quality evaluation. Considering the potential utilization of the local angular-spatial effect for learning, we further design a Discriminative Region Selection (DRS)-based auxiliary learning branch, which serves to enhance both learning efficiency and prediction accuracy. Finally, a Dynamic Weighting Loss (DWLoss) is presented to balance the relationship between principal and auxiliary learning throughout the training process. Experimental results on five widely-used LFIQA databases demonstrate that our metric outperforms state-of-the-art LFIQA metrics by a large margin in quality evaluation, while having higher computational efficiency than most blind LFIQA metrics.

References

[1] P. A. Kara, A. Cserkaszky, M. G. Martini, A. Barsi, L. Bokor, and T. Balogh, "Evaluation of the concept of dynamic adaptive streaming of light field video," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 407–421, Jun. 2018.

[2] Y. Sawahata, Y. Miyashita, and K. Komine, "Estimating angular resolutions required in light-field broadcasting," *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 473–490, Jun. 2021.

[3] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using Bokeh rendering and superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[4] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[5] Z. Pei, X. Chen, and Y.-H. Yang, "All-in-focus synthetic aperture imaging using image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 2, pp. 288–301, Feb. 2018.

[6] X. Wang, J. Liu, S. Chen, and G. Wei, "Effective light field deocclusion network based on Swin transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2590–2599, Jun. 2023, doi: 10.1109/TCSVT.2022.3226227.

[7] T.-C. Wang, M. Chandraker, A. Efros, and R. Ramamoorthi, "SVBRDF-invariant shape and reflectance estimation from light-field cameras," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5451–5459.

[8] M. Zhou, Y. Ding, Y. Ji, S. S. Young, J. Yu, and J. Ye, "Shape and reflectance reconstruction using concentric multi-spectral light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1594–1605, Jul. 2020.

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

[10] X. Huang, Y. Chen, P. An, and L. Shen, "Prediction-oriented disparity rectification model for geometry-based light field compression," *IEEE Trans. Broadcast.*, vol. 69, no. 1, pp. 62–74, Mar. 2023.

[11] G. Wu, Y. Wang, Y. Liu, L. Fang, and T. Chai, "Spatial-angular attention network for light field reconstruction," *IEEE Trans. Image Process.*, vol. 30, pp. 8999–9013, 2021.

[12] P. Paudyal, F. Battisti, P. Le Callet, J. Gutiérrez, and M. Carli, "Perceptual quality of light field images and impact of visualization techniques," *IEEE Trans. Broadcast.*, vol. 67, no. 2, pp. 395–408, Jun. 2021.

[13] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, 2020, Art. no. 211301.

[14] X. Min et al., "Screen content quality assessment: Overview, benchmark, and beyond," *ACM Comput. Surveys*, vol. 54, no. 9, pp. 1–36, 2021.

[15] 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, Sep. 2017.

[16] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T P.910, Int. Telecommun. Union, Geneva, Switzerland, 2022.

[17] S. Alamgeer and M. C. Q. Farias, "A survey on visual quality assessment methods for light fields," *Signal Process. Image Commun.*, vol. 110, Jan. 2023, Art. no. 116873.

[18] B. Wilburn et al., "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.

[19] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Comput. Sci., Stanford Univ., Stanford, CA, USA, Rep. CTSR 2005-02, 2005.

[20] N. Venkatanath, D. Praneeth, B. M. Chandrasekhar, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. Nat. Conf. Commun. (NCC)*, 2015, pp. 1–6.

[21] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[22] Q. Li, W. Lin, and Y. Fang, "No-reference quality assessment for multiply-distorted images in gradient domain," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.

[23] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

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

[25] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[26] X. Xu, Y. Zhao, and Y. Dong, "No-reference stereoscopic image quality assessment based on saliency-guided binocular feature consolidation," *Electron. Lett.*, vol. 53, no. 22, pp. 1468–1470, 2017.

[27] 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, Oct. 2017.

[28] Z. Chen, W. Zhou, and W. Li, "Blind stereoscopic video quality assessment: From depth perception to overall experience," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 721–734, Feb. 2018.

[29] F. Shao, W. Lin, S. Wang, G. Jiang, and M. Yu, "Blind image quality assessment for stereoscopic images using binocular guided quality lookup and visual codebook," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 154–165, Jun. 2015.

[30] S. Tian, L. Zhang, L. Morin, and O. Déforges, "NIQSV: A no reference image quality assessment metric for 3D synthesized views," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2017, pp. 1248–1252.

[31] S. Tian, L. Zhang, L. Morin, and O. Déforges, "NIQSV+: A no reference synthesized view quality assessment metric," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1652–1664, Apr. 2018.

[32] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. L. Callet, "Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views," *IEEE Trans. Broadcast.*, vol. 66, no. 1, pp. 127–139, Mar. 2020.

[33] 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.

[34] 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, Apr. 2021.

[35] 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.

[36] 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.

[37] 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 Exp. (QoMEX)*, 2022, pp. 1–6.

[38] 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.

[39] 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, Jul. 2021.

[40] G. Wu et al., "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Oct. 2017.

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

[42] 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, Nov. 2018.

[43] 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 (BigMM)*, 2018, pp. 1–6.

[44] 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.

[45] 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.

[46] 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, May 2021.

[47] Y. Tian, H. Zeng, J. Hou, J. Chen, and K.-K. Ma, "Light field image quality assessment via the light field coherence," *IEEE Trans. Image Process.*, vol. 29, pp. 7945–7956, 2021.

[48] 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. (BMSB)*, 2021, pp. 1–5.

[49] 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.

[50] C. Meng, P. An, X. Huang, C. Yang, and Y. Chen, "Image quality evaluation of light field image based on macro-pixels and focus stack," *Front. Comput. Neurosci.*, vol. 15, Jan. 2022, Art. no. 768021.

[51] 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.*, early access, Jul. 19, 2023, doi: 10.1109/TCSVT.2023.3297016.

[52] 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, Mar. 2019.

[53] 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*, early access, May 8, 2023, doi: 10.1109/TMM.2023.3273855.

[54] 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. (ICIP)*, 2019, pp. 3781–3785.

[55] 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, Nov. 2020.

[56] A. Ak, S. Ling, and P. L. Callet, "No-reference quality evaluation of light field content based on structural representation of the epipolar plane image," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2020, pp. 1–6.

[57] J. Xiang, M. Yu, H. Chen, H. Xu, Y. Song, and G. Jiang, "VBLFI: Visualization-based blind light field image quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2020, pp. 1–6.

[58] 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, Jul. 2021, Art. no. 108063.

[59] Y. Liu, G. Jiang, Z. Jiang, Z. Pan, M. Yu, and Y.-S. Ho, "Pseudoreference subaperture images and microlens image-based blind light field image quality measurement," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021.

[60] K. Lamichhane, M. Neri, F. Battisti, P. Paudyal, and M. Carli, "No-reference light field image quality assessment exploiting saliency," *IEEE Trans. Broadcast.*, early access, Apr. 3, 2023, doi: 10.1109/TBC.2023.3242150.

[61] 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*, early access, Apr. 19, 2023, doi: 10.1109/TMM.2023.3268370.

[62] C.-C. Chang and C. Lin, "LibSVM: A library for support vector machines," *ACM Trans. Intell. Symp. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[63] Z. Guo, W. Gao, H. Wang, J. Wang, and S. Fan, "No-reference deep quality assessment of compressed light field images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1–6.

[64] 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. (ICIP)*, 2022, pp. 2266–2270.

[65] W. Fu, X. Shen, W. Zhou, A. D. Zhdanov, and C. Geng, "A light field image quality assessment method based on stereo vision," in *Proc. IEEE Int. Conf. Unmanned Syst. (ICUS)*, 2022, pp. 1–8, doi: 10.1109/ICUS55513.2022.9986647.

[66] V. K. Adhikarla et al., "Towards a quality metric for dense light fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3720–3729.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[68] A. Vaswani et al., "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6000–6010.

[69] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[70] J. L. Ba, J. R. Kiros, and G. E. Hinton. "Layer normalization." 2016, *arXiv:1607.06450*.

[71] L.-M. Po et al., "A novel patch variance biased convolutional neural network for no-reference image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1223–1229, Apr. 2019.

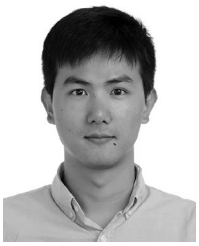
[72] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5462–5474, Nov. 2017.

[73] X. Min et al., "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2319–2333, Sep. 2019.

[74] X. Min, G. Zhai, K. Gu, X. Yang, and X. Guan, "Objective quality evaluation of dehazed images," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 8, pp. 2879–2892, Aug. 2019.

[75] L. Shi, S. Zhao, W. Zhou, and Z. Chen, "Perceptual evaluation of light field image," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 41–45.
[76] 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. Netw. Comput. Commun. (ISNCC)*, 2019, pp. 1–5.
[77] 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, pp. 671–683.
[78] I. Viola and T. Ebrahimi, "VALID: Visual quality assessment for light field images dataset," in *Proc. Int. Conf. Qual. Multimedia Exp. (QoMEX)*, 2018, pp. 1–3.
[79] 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.
[80] 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
[81] M. Broxton et al., "Immersive light field video with a layered mesh representation," *ACM Trans. Graph.*, vol. 39, no. 4, p. 86, 2020.
[82] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, 2020.
[83] X. Min, G. Zhai, J. Zhou, X. Zhang, X. Yang, and X. Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, 2020.
[84] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, pp. 1–23, 2017.
[85] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14071–14081.

Zhengyu Zhang received the B.E. degree in electronic and information science and technology from Guangzhou University, Guangzhou, China, in 2018, and the M.E. degree in electronics and communication engineering from Shenzhen University, Shenzhen, China, in 2021. He is currently pursuing the Ph.D. degree with the National Institute of Applied Sciences, Rennes, France, and also with the Institute of Electronics and Digital Technologies Laboratory. His research interests include image/video quality assessment, visual perception, and deep learning.

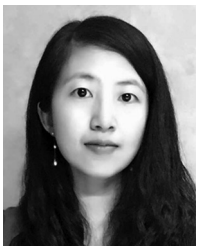


Shishun Tian (Member, IEEE) received the B.Sc. degree from Sichuan University, Chengdu, China, in 2012, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2015, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2019. 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.

Yuhang Zhang (Graduate Student Member, IEEE) received the B.Sc. degree in electronic information engineering from Guangdong Ocean University, Guangdong, China, in 2019. He is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests focus on computer vision, domain adaptation, semantic segmentation, and deep learning.

Wenbin Zou received the M.E. degree in software engineering with a specialization in multimedia technology from Peking University, China, in 2010, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2014. From 2014 to 2015, he was a Researcher with the UMR Laboratoire d'informatique Gaspard-Monge, CNRS, and the Ecole des Ponts ParisTech, France. He is currently an Associate Professor with the College of Electronics and Information Engineering, Shenzhen University, China. His current research interests include saliency detection, object segmentation, and semantic segmentation.

Luce Morin (Member, IEEE) is currently a Full Professor with the National Institute of Applied Science (INSA Rennes), University of Rennes, France, and a Researcher with the Institut d'Electronique et Technologies du numéRique, 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, 3-D reconstruction, image and video compression, and representations for 3-D videos and multiview videos.



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