PVBLiF: A Pseudo Video-Based Blind Quality Assessment Metric for Light Field Image

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

PVBLiF: A Pseudo Video-Based Blind Quality Assessment Metric for Light Field Image

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

***Abstract*—Going beyond traditional 2D imaging is not only an emerging trend of imaging technology, but also the key to a more immersive user experience. Light Field Image (LFI) is a typical high-dimensional imaging format, and the quality evaluation of which is very challenging but necessary. In this paper, we propose a novel Pseudo Video-based Blind quality assessment metric for Light Field image (PVBLiF). In contrast to most previous Light Field Image Quality Assessment (LF-IQA) metrics, in which different types of 2D representations derived from LFI are used for quality assessment indirectly, our metric exploits a more intuitive 3D representation, named Pseudo Video Block Sequence (PVBS), to evaluate the perceptual quality of LFI. For this purpose, we first divide the LFI into a massive number of non-overlapping PVBSs, which simultaneously contain spatial and angular information of LFI. Then, we propose a novel network (named PVBSNet) based on Convolutional Neural Networks (CNNs) to extract the spatio-angular features of PVBS and further evaluate the PVBS quality. The proposed PVBSNet consists of four stages: multi-information division, intra-feature extraction, cross-feature fusion, and quality regression. Finally, a Saliency- and Variance-guided Pooling (SVPooling) method is presented to integrate all the PVBS quality into the overall quality of LFI. The proposed PVBLiF metric has been extensively evaluated on three widely-used LFI datasets: Win5-LID, NBU-LF1.0, and SHU. Experimental results demonstrate that our proposed PVBLiF metric outperforms state-of-the-art metrics and is capable of highly approximating the performance of human observers. The source code of PVBLiF is publicly available at https://github.com/ZhengyuZhang96/PVBLiF.**

***Index Terms*—Light field, pseudo video, blind quality assessment, CNN, spatio-angular features.**

I. Introduction

IMMERSIVE imaging technologies, including light field, 360-degree panoramic, and volumetric images/videos, aim to increase the audience presence and improve the immersive visualization, which is difficult to be achieved with traditional 2D imaging. With the recent availability of hand-held light field cameras, Light Field Image (LFI) has received extensive attention from both academia and industry, further offering the possibility for a wide range of applications. Theoretically, LFI records all the information of light rays as they travel in free space, which was first defined as a 7D plenoptic function [1], [2] and further predigested to a 4D model [3] by assuming that light is wavelength- and time-invariant and unobstructed. As a result, LFI is described via a biplane parameterization $L(u, v, h, w)$, where (u, v) denote the angular coordinates and (h, w) denote the spatial coordinates. However, despite a series of simplifications, LFI is still very complicated, with inherently high-dimensional characteristics different from traditional 2D images.

From initial acquisition to end-user visualization, LFI unavoidably suffers from various perceptual quality impairments [4], [5], primarily including compression, reconstruction, and display distortions. A well-performing LFI quality monitoring indicator, known as Light Field Image Quality Assessment (LF-IQA), ensures a pleasing visual experience for human viewers. The quality of LFI can be assessed by two approaches: subjective and objective. Subjective approaches directly collect the perceptual assessment opinions from human viewers about the attributes of LFI. Thus, subjective approaches are widely considered as the most reliable way to assess the visual quality. However, collecting subjective opinions is a cumbersome and costly task, and it is impossible to automate in real-time. To this end, objective approaches, based on computational models, are designed to approximate the subjective opinions in an efficient and inexpensive manner. Many previous studies [6], [7], [8] have shown that traditional 2D/3D/multi-view IQA metrics [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] fail to accurately evaluate the perceptual quality of LFI that contains both spatial and angular information. Thanks to the efforts of researchers on LF-IQA in recent years, several landmark quality assessment metrics for LFI have been proposed. The existing objective LF-IQA metrics generally fall into three categories relying on the amount of reference information used: Full-Reference (FR), Reduced-Reference (RR), and No-Reference/Blind (NR). The FR/RR LF-IQA metrics require the presence of the undistorted LFI or the reduced version thereof. Instead, the NR LF-IQA metrics are more practical in real-world scenarios where access to reference information is expensive or even impossible. In addition, recently Convolutional Neural Networks (CNNs) have been shown to be effective in modeling high-dimensional LFIs due to their powerful discriminative ability. Therefore, we focus on the CNN-based NR LF-IQA metric in this paper.

The mainstream idea of existing NR LF-IQA metrics is to evaluate the LFI quality with its 2D representations, $e.g.$, Sub-Aperture Image (SAI), Refocused Image (RI), Epipolar

This work was supported in part by the National Natural Science Foundation of China under grants 62101344, 62171294, in part by the Natural Science Foundation of Guangdong Province, China under grants 2022A1515010159, 2020A1515010959, in part by the Key Project of DEGP under grants 2018KCXTD027, in part by the Key Project of Shenzhen Science and Technology Plan under Grant 20220810180617001.

Corresponding author: Shishun Tian

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

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

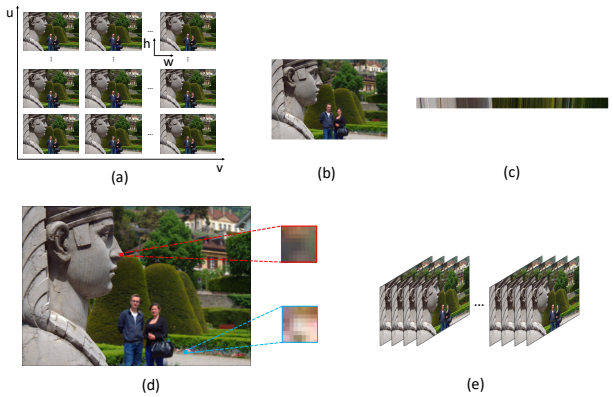


Fig. 1. Different representations of a sample LFI [19]. (a) SAI array; (b) RI; (c) EPI; (d) MLI; (e) PVS.

Plane Image (EPI), and MicroLens Image (MLI), as shown in Fig. 1 (a)-(d). Nonetheless, whether the perceptual quality of 4D LFI can be adequately reflected by these 2D representations is still unclear. Alternatively, assessing the LFI quality based on the raw 4D LFI or its 3D representations is more reliable. However, when combined with CNNs, implementing 4D CNNs for the raw 4D LFI is computationally expensive and hardware-unfriendly. Instead, evaluating the LFI quality with its 3D representations is considered as the best compromise for the CNN-based NR LF-IQA metric.

Pseudo Video Sequence (PVS) is a 3D representation of LFI commonly used in LFI compression algorithms [20], [21], [22], [23], [24]. By arranging each view as a frame and displaying all views in a certain order, the perceptual quality degradation caused by spatio-angular distortions can be more easily perceived in PVS, as shown in Fig. 1 (e). As a result, PVS is widely served as the passive mode of LFI visualization in the subjective LF-IQA experiments [25], [26], [27]. Despite the importance of PVS, most existing objective LF-IQA metrics evaluate the LFI quality without considering this 3D representation. Thus, it is necessary to explore the usage of PVS in the objective LF-IQA metrics.

In order to develop a CNN-based metric based on PVS, two challenges are faced. First, although PVS is a video-like 3D representation with the same structure as conventional videos, the feature extraction networks for video quality assessment ($e.g.$, [28], [29], and [30]) are not suitable for PVS. The main reason is that the temporal dimension of conventional video only contains 1D motion information, while the temporal dimension of PVS consists of the scrambled 2D angular information of LFI. Based on this characteristic, an objective LF-IQA metric needs to extract the spatio-angular features instead of temporal-spatial features in PVS, which has not been studied in the quality assessment domain. Second, due to the limited size of the existing LFI datasets, directly using PVS for training can easily lead to the over-fitting problem. In addition, it is suboptimal to extract spatio-angular features directly from PVS, because the angular variation in PVS is much smaller than the spatial variation in PVS. A feasible solution is to partition PVS into a large number of 3D blocks in the spatial domain, where the generated blocks are named Pseudo Video Block Sequences (PVBSs). Here, such a block-based strategy can highly enlarge the training set and balance the spatial and angular information. In this case, the overall LFI quality is usually obtained by averaging the quality of all blocks [31], [32]. However, we believe that not all regions contain enough information to reflect the quality degradation of the whole LFI. Furthermore, the average pooling method ignores that different regions of the same scene contribute differently to the overall quality perception of human vision. Therefore, a post-processing pooling method based on the mechanism of human visual perception should be developed.

The aforementioned analyses provide some guidance and inspiration for us to design a new NR LF-IQA metric. In summary, the main contributions of this paper are listed as follows.

1) Instead of using 2D representations, we use the PVBS, a more intuitive 3D representation for quality assessment. Based on the PVBS's characteristics, a novel CNN-based PVBSNet is presented, in which four stages are designed to fully exploit the spatio-angular features and evaluate the quality: multi-information division, intra-feature extraction, cross-feature fusion, and quality regression.

2) Considering that different regions of the same scene have different effects on the human perception, a Saliency- and Variance-guided Pooling (SVPooling) method is proposed to simulate human visual attention on LFI and reweight the importance of different PVBSs to obtain the overall LFI quality.

3) A novel Pseudo Video-based Blind quality assessment metric for Light Field image, abbreviated as PVBLiF, is proposed in this paper. Extensive experimental results on three representative LFI datasets show that the proposed PVBLiF metric outperforms the state-of-the-art LF-IQA metrics by a large margin, while having low computational complexity.

The rest of this paper is organized as follows. Section II introduces the related works. Section III describes the proposed metric in detail. Section IV provides experimental results and discussions. Finally, conclusions will be drawn in Section V.

II. Related Works

As mentioned before, the existing objective LF-IQA metrics can be classified into FR, RR, and NR categories. The FR LF-IQA metrics [6], [7], [33], [34], [35], [36], [37], [38], [39], [40] assess the quality of the distorted LFI when the originally pristine information is available. For example, KRIQE [7] exploits the gradient magnitude and phase congruency of the key reference and distorted RIs for LFI quality evaluation. MDFM [34] extracts the multi-order derivative information in reference and distorted SAIs to assess the LFI quality. The RR LF-IQA metrics estimate the distorted LFI quality with reduced information derived from the reference LFI. For instance, Paudyal $et\ al.$ [41] propose a RR LF-IQA metric based on the depth map similarity between the distorted and reference LFIs.

Compared to the FR/RR LF-IQA metrics, the NR LF-IQA metrics evaluate the distorted LFI quality without any

reference information, which are more applicable in most real-world scenarios. In general, existing NR LF-IQA metrics can be further divided into two types: Natural Scene Statistics-based (NSS-based) and CNN-based.

Among NSS-based NR LF-IQA metrics, some metrics [42], [43], [44], [45], [46] extract NSS features from the original 2D representations of LFI, $e.g.$, SAI, RI, EPI, and MLI. Luo $et\ al.$ [42] utilize the information entropy of SAIs and naturalness distribution of MLI to measure the spatial quality and angular consistency of LFI, respectively. Shi $et\ al.$ [43] combine the naturalness distribution features of the light field cyclopean image array and the global and local features of EPIs to measure the LFI quality. Ak $et\ al.$ [44] extract the structural features in EPIs using convolutional sparse coding and histogram of oriented gradients. In VBLFI [45], the NSS and energy features are extracted from the mean difference image and SAIs using curvelet transform. DSA [46] is considered as the updated version of VBLFI, which additionally measures the angular deterioration based on EPIs. However, these 2D representations always ignore some primitive characteristics of LFI. For example, SAI contains only spatial information but lacks angular information; Only scene within the refocused slope can be clearly displayed in RI; EPI considers only 1D information from both spatial and angular domains. Instead of using these original 2D representations, some NSS-based NR LF-IQA metrics [47], [8], [48], [49] utilize Tucker decomposition to decompose LFI into 2D principal components for quality assessment. For example, BELIF [47] decomposes the light field cyclopean image array and extracts naturalness distribution features and statistical structural features for quality assessment. Tensor-NLFQ [8] adopts Tucker decomposition on SAIs for dimensionality reduction, and then measures the LFI quality with naturalness, frequency, and structural similarity properties. TSSV-LFIQA [48] processes the LFI with Tucker decomposition and analyzes the sharpness and distribution information of tensor slice and the percentage of singular value. PVRI [49] evaluates the angular quality from the structure, motion and disparity information of the decomposed PVS, and measures the spatial quality from the depth and semantic information of RIs. The motivation behind these metrics is that the LFI data has considerable redundancy due to its narrow parallax. Although the redundancy in LFI can be greatly reduced after using Tucker decomposition, some angular nuances that lead to quality degradation may be overlooked in the meantime. More recently, compared to the previous works that focus on extracting 2D NSS features, Xiang $et\ al.$ [50] extract naturalness distribution and energy features of distorted LFI in 4D frequency domain.

For CNN-based NR LF-IQA metrics, Guo $et\ al.$ [51] present a deep CNN model based on SAI fusion and global context perception modules. ALAS-DADS [52] designs a depth-wise separable convolution-based CNN model to evaluate the LFI quality. However, these two metrics take the whole LFI as input which may suffer from the problem of over-fitting and high computational complexity. To alleviate these problems, DeLFIQE [31] proposes to evaluate the LFI quality via discriminative EPI patches. Our previous work DeeBLiF [32] also adopts a patch-based method, where the quality of spatio-angular patches is evaluated. Finally, both DeLFIQE and DeeBLiF obtain the overall LFI quality via averaging the quality of all patches. However, these patch-based methods do not take into account that different patches may contribute differently to the overall quality.

In summary, there are two main drawbacks in the existing NR LF-IQA metrics. First, it is still unclear whether the perceptual quality of LFI is well inherited by the widely-used 2D representations. Thus, we use the PVBS, a more intuitive 3D representation for quality evaluation. Based on the characteristics of PVBS, a CNN-based PVBSNet is proposed to fully extract the spatio-angular features and further evaluate the PVBS quality. Second, the average pooling used in the patch/block-based metric is not an optimal solution. Therefore, we propose a SVPooling method to reweight the importance of different PVBSs to obtain the overall LFI quality. The above components constitute our proposed PVBLiF metric, which essentially differentiate our work from previous NR LF-IQA metrics.

III. Proposed Metric

The overall framework of the proposed PVBLiF metric is illustrated in Fig. 2. It mainly consists of three components: PVBS generation, PVBSNet, and SVPooling. Given a SAI array of LFI, we first transform it into PVS form and then generate a large number of PVBSs. After that, the spatio-angular features of each PVBS are exploited and further quantified into a quality score. Finally, the SVPooling method is applied for integrating all the PVBS quality into the overall quality of the input LFI. Note that PVBSNet is trained with PVBSs in an end-to-end manner, while the SVPooling method is used for post-processing in the test phase and no trainable parameters are involved. All components are described in the following subsections.

A. PVBS generation

As introduced before, the 2D representations used in existing LF-IQA metrics more or less ignore the primitive characteristics of LFI. Instead, PVS is a 3D representation of LFI, which highly preserves the spatio-angular information of the original LFI. Derived from the idea of PVS, we exploit a 3D representation, named PVBS, to evaluate the LFI quality. The generation of PVBS and the motivation behind it are detailed below.

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 channel number. First, to reduce the computational complexity, the central $A\times A$ SAIs are extracted and further transformed into PVS form (denoted as $\mathcal{P}_C \in \mathbb{R}^{(A*A)\times H\times W\times C}$), by arranging each SAI as a frame and displaying all SAIs in a certain order. Generally, there are several orders for creating a PVS [53]. In our implementation, we create the PVS using raster order [53], $i.e.$, horizontally scanning the SAI array from left to right and row by row, and starting from the SAI on the left superior corner. This generation order facilitates the subsequent extraction of spatio-angular features, which will be described in

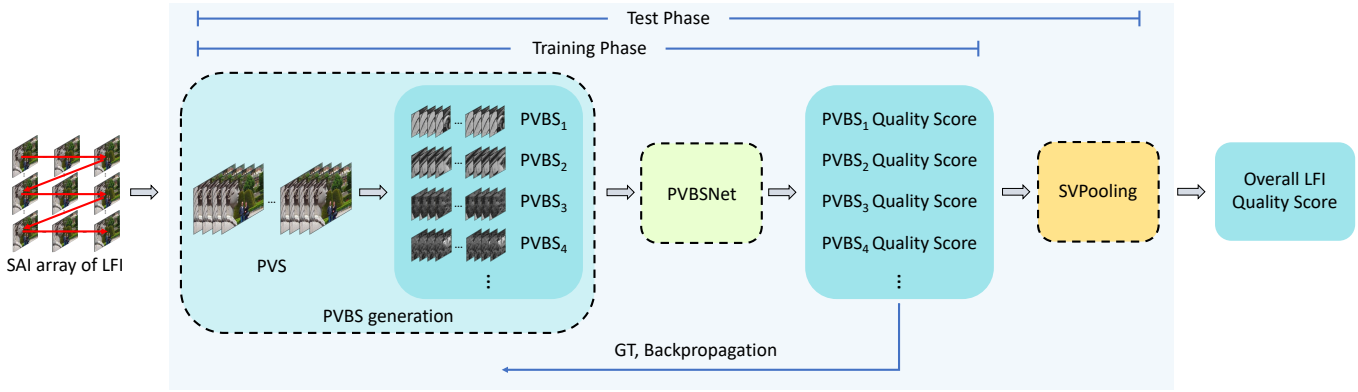


Fig. 2. Overall framework of the proposed PVBLiF metric.

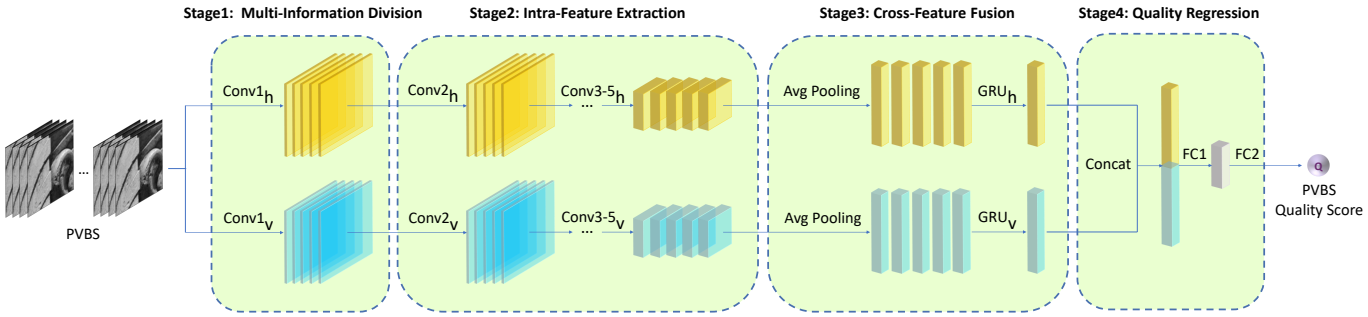


Fig. 3. Framework of the proposed PVBSNet. The top and bottom streams are used for extracting the horizontal and vertical angular features, respectively.

detail in the following subsection. Then, we convert $\mathcal{P}_C$ from RGB color space to YCbCr color space as recommended in many literature ($e.g.$, [54], [55]). After that, only the luminance component Y is used for subsequent feature extraction and quality prediction. Let $\mathcal{P}_{CY} \in \mathbb{R}^{(A*A)\times H\times W}$ denote the Y component of PVS, the above operations are described in Eq. (1)-(2).

$$\mathcal{P}_C = PVS(Central(\mathcal{L})) \tag{1}$$

$$\mathcal{P}_{CY} = \{YCbCr(\mathcal{P}_C)\}_Y \tag{2}$$

where $Central(\cdot)$, $PVS(\cdot)$, and $YCbCr(\cdot)$ denote the selection of central SAIs, the PVS transformation, and the YCbCr color space conversion, respectively.

The generated PVS is a series of SAIs arranged in a specific order, where individual SAIs contain the spatial information and the changes between different SAIs reflect the angular information. However, it is difficult to evaluate the LFI quality directly using PVS, because the spatial variation is much greater than the angular variation, and the slight angular variation will be submerged. In addition, taking each PVS as an individual sample for training may suffer from the over-fitting problem since such a large amount of data requires deeper networks [51]. Therefore, we split PVS into a large number of non-overlapping blocks in the spatial domain. Each generated block is named PVBS and denoted as $\mathcal{PB} \in \mathbb{R}^{(A*A)\times S\times S}$, where $A*A$ and $S\times S$ are the length and the spatial resolution of PVBS, respectively. The generation of PVBS is described in Eq. (3).

$$\mathcal{P}_{CY} = \{\mathcal{PB}_1, \mathcal{PB}_2, ..., \mathcal{PB}_K\} \tag{3}$$

where $\mathcal{PB}_k$ denotes the k-th PVBS and K is the total number of PVBSs.

In this paper, A and S are set to 5 and 32 to achieve the best trade-off between precision and efficiency. The impact of different parameter settings on the performance of the proposed metric will be further discussed in Section IV. Although each PVBS contains only a small amount of information derived from the original LFI, it is still a 3D representation with reduced spatial information of PVS while preserving the structural information of PVS. In other words, this representation still fully reflects the perceptual quality of the original LFI. Further, evaluating the LFI quality with PVBS can be viewed as a 3D extension of previous patch-based methods [31], [32], inheriting the advantage of low computational complexity.

B. PVBSNet

Previous studies on LF-IQA [43], [49] have demonstrated that the potential interaction between spatial and angular information, called spatio-angular information, is more closely related to the LFI quality than the spatial information. Although the 3D structure of PVBS has great potential in simulating how LFI is presented to human eyes, two challenges are encountered when extracting spatio-angular features for quality assessment. First, during the generation of PVBS, the original 2D angular information of LFI is scrambled as it is transformed into the 1D temporal information of PVBS. Second, extracting spatio-angular features from a 3D

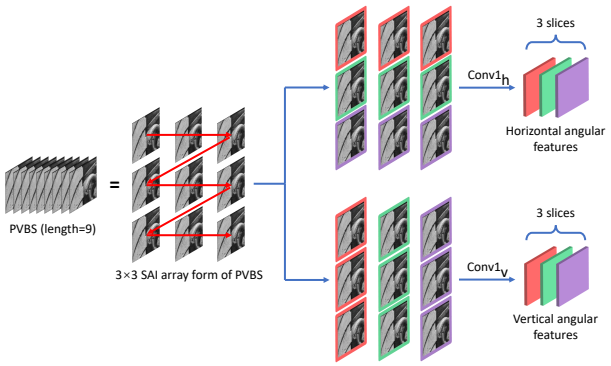


Fig. 4. Stage 1: Multi-information division. For better visualization, A is set to 3 for all illustrations with $A\times A$ angular resolution.

TABLE I
NETWORK CONFIGURATION OF THE PROPOSED PVBSNET.

Stage	Layer Name	Output Size	Kernel Size	Dilation	Stride	Padding
Stage 1	Conv1_h	$64\times A\times S\times S$	$A\times1\times1$	$1\times1\times1$	$A\times1\times1$	$0\times0\times0$
			Batch Normalization, Leaky ReLU			
	Conv1_v	$64\times A\times S\times S$	$A\times1\times1$	$A\times1\times1$	$1\times1\times1$	$0\times0\times0$
			Batch Normalization, Leaky ReLU			
Stage 2	$\text{Conv2}_\text{h/v}$	$64\times A\times S\times S$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv3}_\text{h/v}$	$128\times A\times S\times S$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv4}_\text{h/v}$	$256\times A\times \frac{S}{2}\times \frac{S}{2}$	$1\times3\times3$	$1\times1\times1$	$1\times2\times2$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
	$\text{Conv5}_\text{h/v}$	$256\times A\times \frac{S}{2}\times \frac{S}{2}$	$1\times3\times3$	$1\times1\times1$	$1\times1\times1$	$0\times1\times1$
			Batch Normalization, Leaky ReLU			
Stage 3	Avg Pooling	$256\times A$	$1\times \frac{S}{2}\times \frac{S}{2}$	$1\times1\times1$	$1\times1\times1$	$0\times0\times0$
	$\text{GRU}_\text{h/v}$	256×1	Gated Recurrent Unit (GRU)			
Stage 4	Concat	512×1	Concatenation layer			
	FC1	128×1	Fully Connected (FC) layer			
	FC2	1×1	Fully Connected (FC) layer			

representation to evaluate quality in an end-to-end manner is non-trivial.

To address the above challenges, we propose a CNN-based network, named PVBSNet, to extract the spatio-angular features of PVBS and further evaluate the PVBS quality. Fig. 3 illustrates the framework of the proposed PVBSNet, where the top and bottom streams are used for extracting the horizontal and vertical angular features, respectively. In addition, the network is systematically divided into four stages according to different functions: multi-information division, intra-feature extraction, cross-feature fusion, and quality regression. Note that the network is trained and tested in an end-to-end manner, and its configurations are detailed in TABLE I. The four stages will be introduced in turn below.

Stage 1: Multi-information division. From the original LFI to the PVS in raster order, the relationship between horizontal adjacent views is stored while the relationship between vertical adjacent views is destroyed. For this reason, video quality assessment methods are not suitable for PVS, nor for PVBS. The multi-information division stage is proposed to handle this challenge, which aims to divide the scrambled temporal information of PVBS into horizontal and vertical information. To achieve this goal, we notice that the horizontal angular information is recorded in the A adjacent views of PVBS while the vertical angular information is recorded in every A views of PVBS, where $A\times A$ denotes the angular resolution of the original LFI. Thus, we propose to adopt special 3D CNN kernels to separate the regular temporal information into horizontal and vertical angular information, as shown in Fig. 4. Specifically, based on the raster order used in generating PVBS, a 3D convolutional kernel with $A\times1\times1$ size and $A\times1\times1$ stride is used for extracting horizontal angular features, while a 3D convolutional kernel with $A\times1\times1$ size and $A\times1\times1$ dilation is used for extracting vertical angular features. As a result, horizontal and vertical angular features are extracted from the scrambled temporal information of PVBS. Given a PVBS $\mathcal{PB}$, let $\mathcal{F}^1$ represent the features output from stage 1, the above process is described in Eq. (4).

$$\mathcal{F}^1_m = Conv1_m(\mathcal{PB}) \tag{4}$$

where $m \in \{h, v\}$, h and v denote the horizontal and vertical angular streams, respectively. $Conv1_m(\cdot)$ denotes the 3D convolutional layer used in stage 1.

Stage 2: Intra-feature extraction. After separating horizontal and vertical angular information in stage 1, we obtain horizontal angular features $\mathcal{F}^1_h$ and vertical angular features $\mathcal{F}^1_v$. The resulting horizontal (or vertical) angular features include A slices, where slice is defined as the spatio-horizontal (or spatio-vertical) features of all SAIs from a certain row (or column). As shown in the right side of Fig. 4, the feature maps in different colors are considered to be different slices. Considering that 2D CNNs aim to extract the spatial information and integrate it into the channel information, therefore, in the second stage, we propose to extract the so-called spatio-angular features for each slice with a sequence of repeated 2D CNNs. Specifically, a 4-layer network with $1\times3\times3$ kernel size and $1\times1\times1$ dilation is used. To avoid losing too much spatial information, we only set the stride to $1\times2\times2$ to halve the spatial size in $Conv4_{h/v}$. The operations adopted in stage 2 are described in Eq. (5).

$$\mathcal{F}^N_m = ConvN_m(\mathcal{F}^{N-1}_m) \tag{5}$$

where $\mathcal{F}^N_m$ denotes the spatio-angular features output from the N-th 2D convolutional layer $ConvN_m(\cdot)$, $N \in \{2, 3, 4, 5\}$. Specifically, $\mathcal{F}^5_m$ represents the output of stage 2.

Stage 3: Cross-feature fusion. The third stage aims to establish the relationship between the spatio-angular features of each slice and fuse them into unified features. Due to the inherently narrow baseline of different views in LFI, there are only subtle differences between different slices. In addition, all slices are arranged in order, $i.e.$, from the first row (or column) to the last row (or column). Therefore, inspired by [56], we consider it as a temporal-memory issue and propose to model the long-term dependencies for fusing. Here, we perform the Gated Recurrent Unit (GRU) [57], a recurrent neural network model with gate control, to integrate the features of different slices into unified features. Before adopting the GRU model, we convert each spatio-angular feature map into a feature vector using an average pooling layer. Let $\mathcal{F}^G$ represent the

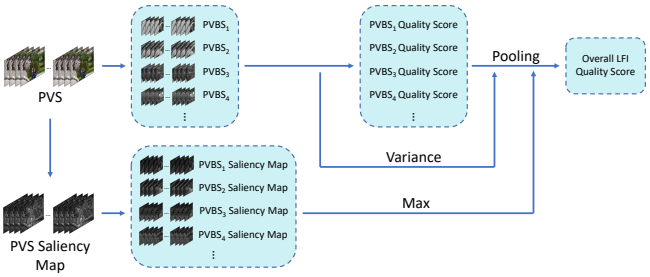


Fig. 5. Pipeline of the proposed SVPooling method.

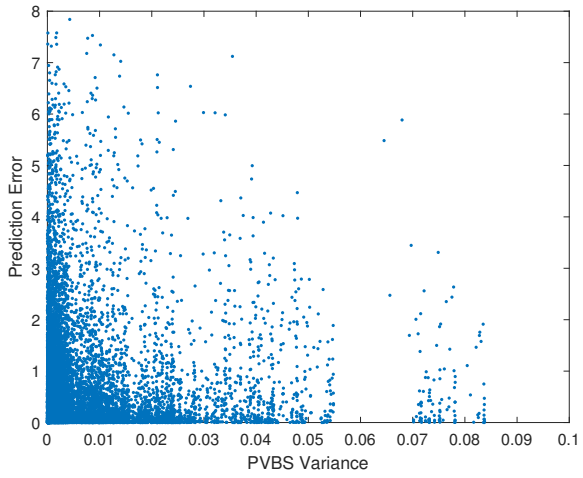


Fig. 6. Scatter plots of all PVBSs in terms of quality prediction error against variance on the Win5-LID dataset [19].

feature vector output from stage 3, the process of stage 3 is described in Eq. (6).

$$\mathcal{F}_m^G = GRU_m(AP(\mathcal{F}_m^5)) \tag{6}$$

where $GRU_m(\cdot)$ denotes the GRU model, $AP(\cdot)$ denotes the average pooling layer.

Stage 4: Quality regression. In the final stage, we quantify the extracted spatio-angular features into quality score following [32]. Specifically, the concatenation operation is applied for $\mathcal{F}_h^G$ and $\mathcal{F}_v^G$, followed by two Fully Connected (FC) layers to predict the quantified quality score. The generation of PVBS quality score $\mathcal{Q}$ is described in Eq. (7).

$$\mathcal{Q} = FC2(FC1(Concat(\mathcal{F}_h^G, \mathcal{F}_v^G))) \tag{7}$$

where $Concat(\cdot)$, $FC1(\cdot)$, and $FC2(\cdot)$ represent the concatenation, first FC, and second FC layers, respectively.

C. SVPooling method

As the result of PVBSNet, the PVBS quality reflects the local quality of LFI since it derives from the partial information of LFI. To obtain the overall LFI quality, a pooling method is required. Previous metrics [31], [32] employ average pooling to convert all the local quality into the overall LFI quality. In other words, all regions of the same scene are considered to contribute equally to the quality perception of human vision. However, as concluded in many previous studies ($e.g.$, [58], [59]), human eyes naturally pay more attention to certain regions when observing a scene. In addition, we consider that some regions have insufficient information to reflect the LFI quality degradation. Therefore, in this subsection, we propose a Saliency- and Variance-guided Pooling (SVPooling) method to obtain the overall LFI quality, as illustrated in Fig. 5.

1) Saliency. First, we reweight the contribution of different PVBSs according to their saliency maps. In this paper, the saliency map of PVBS is generated by applying saliency detection approach to all views of SAI array and converting them into PVBS form. Specifically, we choose SDSP [60] for saliency detection. Further, human eyes pay more attention to a certain area, often because there are some noteworthy points in it. Based on this principle, we consider the maximum value of PVBS saliency map as the PVBS weight. Let $\mathcal{S}$ and $\mathcal{W}$ denote the saliency map and weight of PVBS, respectively, the generation of $\mathcal{W}$ is described in Eq. (8).

$$\mathcal{W} = Max\{\mathcal{S}\} \tag{8}$$

where $Max\{\cdot\}$ represents the max operation.

2) Variance. Then, we delve into the relationship between the prediction accuracy of PVBS quality and the variance of PVBS, by assuming that PVBSs with complex textures may better reflect their quality. To prove our conjecture, we conducted an experiment on the Win5-LID dataset [19] and visualized the experimental results as suggested in [61]. Fig. 6 shows the scatter plots of all PVBSs in terms of quality prediction error against variance. Here, the prediction error is calculated as the square error between the quality prediction and its corresponding Mean Opinion Score (MOS). It can be found that low-variance PVBSs are not reliable for evaluating the quality of PVBSs as their prediction errors are quite scattered. In other words, although some low-variance PVBSs have accurate predictions, it is difficult to discern the prediction quality of low-variance PVBSs. In contrast, the error distribution of high-variance PVBSs is more concentrated in low prediction error, implying that high-variance PVBSs should be biased when pooling the overall LFI quality. For this reason, we propose to discard the low-variance PVBSs directly to avoid introducing inaccurate predictions from low-variance PVBSs. In this way, the predicted overall LFI quality will be dominated by the quality of high-variance PVBSs. In order to distinguish between low-variance and high-variance PVBSs, for simplicity, we divide all PVBSs into two equal parts sorted by their variances. Let Ω denote the assemble of high-variance PVBSs, the generation of Ω is described in Eq. (9)-(10).

$$\mathcal{T} = Median\{V_{\mathcal{PB}_1}, V_{\mathcal{PB}_2}, ..., V_{\mathcal{PB}_K}\} \tag{9}$$

$$\mathcal{PB}_k \in \Omega, \quad if\ V_{\mathcal{PB}_k} > T \tag{10}$$

where $\mathcal{PB}_k$ and $V_{\mathcal{PB}_k}$ denote the k-th PVBS and its variance, respectively, $\mathcal{T}$ denotes the threshold, $Median\{\cdot\}$ represents the median operation.

TABLE II
OVERALL PERFORMANCE COMPARISON ON THE WIN5-LID, NBU-LF1.0, AND SHU DATASETS. **BOLD** AND UNDERLINE REPRESENT THE BEST AND SECOND BEST PERFORMANCE, RESPECTIVELY.

Metric Types	Metrics	Win5-LID			NBU-LF1.0			SHU		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
NR 2D-IQA	PIQE [63]	0.4820	0.3920	0.9220	0.2561	0.1779	1.0000	0.7780	0.7996	0.6836
	OG [64]	0.5311	0.3960	0.8195	0.4279	0.3446	0.8122	0.8924	0.8558	0.4905
	FRIQUEE [65]	0.5803	0.4978	0.7814	0.4843	0.2791	0.7771	0.9107	0.8891	0.4367
	GWH-GLBP [66]	0.4752	0.3391	0.8364	0.4947	0.3550	0.7780	0.6964	0.6102	0.7677
	NIQE [67]	0.6246	0.4482	0.7584	0.4792	0.3701	0.7948	0.9187	0.8920	0.4247
	BRISQUE [68]	0.6263	0.4559	0.7530	0.4969	0.3750	0.7910	0.9012	0.8747	0.4614
	HyperIQA [69]	0.6571	0.5032	0.7254	0.5876	0.4858	0.7286	0.9259	0.8934	0.4062
	GraphIQA [70]	0.6685	0.5567	0.7146	0.6563	0.5624	0.6758	0.9306	0.8099	0.2659
NR multi-view-IQA	MNSS [71]	0.4053	0.2470	0.8901	0.2653	0.1692	0.8714	0.3325	0.1097	1.0293
	NIQSV [72]	0.3311	0.3130	0.9045	0.4063	0.3556	0.8330	0.4990	0.0806	0.9698
	NIQSV+ [73]	0.3030	0.2174	0.9159	0.3512	0.2339	0.8354	0.4362	0.0757	0.9390
NR 3D-IQA	Xu's [15]	0.5704	0.4345	0.7917	0.4088	0.2814	0.8231	0.8477	0.8177	0.5976
	SINQ [16]	0.6051	0.5075	0.7410	0.5276	0.4374	0.7633	0.9189	0.8955	0.4209
	BSVQE [17]	0.6402	0.5770	0.7430	0.5324	0.4574	0.7571	0.7912	0.7165	0.6577
FR LF-IQA	MDFM [34]	0.7303	0.6768	0.6625	0.8444	0.8138	0.4749	0.8275	0.8543	0.6149
	Min's [6]	0.7350	0.6645	0.6794	0.7112	0.6577	0.6476	0.8496	0.8470	0.5745
	Meng's [35]	0.6924	0.6347	0.7001	0.8367	0.7819	0.4944	0.9282	0.9203	0.4037
NR LF-IQA	BELIF [47]	0.5912	0.5119	0.7572	0.7161	0.6892	0.6291	0.8976	0.8671	0.4784
	Tensor-NLFQ [8]	0.7595	0.7345	0.6327	0.7624	0.7261	0.5856	0.8649	0.8630	0.5424
	PVRI [49]	0.7217	0.6827	0.6530	0.8017	0.7603	0.5271	0.9167	0.8771	0.4237
	NR-LFQA [43]	0.6952	0.6275	0.6750	0.8327	0.8036	0.4895	0.9390	0.9347	0.3729
	VBLFI [45]	0.6844	0.6116	0.7041	0.8179	0.7660	0.5027	0.9220	0.8992	0.4100
	DSA [46]	0.7754	0.7328	0.6150	0.8482	0.8093	0.4578	0.9313	0.9195	0.3904
	4D-DCT-LFIQA [50]	0.8267	0.8079	0.5512	0.8381	0.8213	0.4906	0.9400	0.9320	0.3691
	DeeBLiF [32]	0.8427	0.8186	0.5160	0.8583	0.8229	0.4588	0.9548	0.9419	0.3185
	PVBLiF (ours)	**0.8749**	**0.8580**	**0.4660**	**0.9060**	**0.8883**	**0.3746**	**0.9554**	**0.9501**	**0.3160**

3) Pooling. Finally, we obtain the overall LFI quality score using only high-variance PVBSs and their corresponding weights, as described in Eq. (11).

$$\mathbb{Q} = \frac{\sum_{\Omega} \mathcal{Q} * \mathcal{W}}{\sum_{\Omega} \mathcal{W}} \tag{11}$$

where $\mathbb{Q}$ is the overall quality score of the input LFI. Note that no trainable parameters are involved in SVPooling, which is considered as a post-processing method in our PVBLiF metric.

D. Training

In the training phase, we train our PVBSNet on a TITAN Xp GPU for 70 epochs with the mini-batch Stochastic Gradient Descent (SGD) optimizer, where the weight momentum and decay are set to 0.9 and 0.001, respectively. The quality evaluation is performed using the last epoch model. An initial learning rate of 0.001 is adopted and multiplied by 0.1 every 30 epochs. With the batch size of 8, the network is trained from scratch. All the PVBSs from the same LFI use the same MOS as their training Ground Truth (GT). Our metric involves two hyperparameters that determine the length and the spatial resolution of the generated PVBS, A and S, are set to 5 and 32, respectively. The commonly-used Mean Square Error (MSE) loss is employed to measure the difference between the predicted PVBS quality and its corresponding GT. The goal of the training phase is to minimize the loss and update the trainable parameters, as described in Eq. (12)-(13).

$$\mathcal{LOSS} = \frac{1}{B}\sum_{b=1}^{B}(f(\mathcal{PB}_b; w) - G_b)^2 \tag{12}$$

$$w' = \min_{w}(\mathcal{LOSS}) \tag{13}$$

where $f(\mathcal{PB}_b; w)$ represents the quality of the b-th PVBS predicted using network weights w, and G_b is the corresponding GT. B denotes the batch size, w' denotes the updated network weights.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To verify the effectiveness of our proposed PVBLiF metric, we conduct extensive experiments on three publicly available LFI datasets, including Win5-LID [19], NBU-LF1.0 [25], and SHU [26]. For all datasets, all SAIs of real-world and synthetic LFIs are with spatial resolutions 434×625 and 512×512, respectively. Therefore, each real-world LFI can produce 247 PVBSs, whereas each synthetic LFI can generate 256 PVBSs.

TABLE III
MEDIAN SROCC PERFORMANCE OF ALL COMBINATIONS IN LEAVE-TWO-FOLD-OUT CROSS-VALIDATION ON THE WIN5-LID AND NBU-LF1.0 DATASETS. **BOLD** AND UNDERLINE REPRESENT THE BEST AND SECOND BEST PERFORMANCE, RESPECTIVELY.

Metric Types	Metrics	Datasets	
		Win5	NBU
NR 2D-IQA	PIQE [63]	0.4034	0.1660
	OG [64]	0.4032	0.3570
	FRIQUEE [65]	0.5247	0.2297
	GWH-GLBP [66]	0.3577	0.3632
	NIQE [67]	0.4903	0.3805
	BRISQUE [68]	0.4752	0.3691
	HyperIQA [69]	0.5355	0.4905
	GraphIQA [70]	0.5459	0.5976
NR multi-view-IQA	MNSS [71]	0.2410	0.1425
	NIQSV [72]	0.2527	0.3632
	NIQSV+ [73]	0.1908	0.2122
NR 3D-IQA	Xu's [15]	0.4589	0.2662
	SINQ [16]	0.5466	0.4761
	BSVQE [17]	0.6342	0.4698
FR LF-IQA	MDFM [34]	0.6879	0.8121
	Min's [6]	0.6737	0.6703
	Meng's [35]	0.6433	0.7820
NR LF-IQA	BELIF [47]	0.5344	0.7084
	Tensor-NLFQ [8]	0.7576	0.7462
	PVRI [49]	0.7257	0.7676
	NR-LFQA [43]	0.6626	0.8138
	VBLFI [45]	0.6806	0.8053
	DSA [46]	0.7491	0.8311
	4D-DCT-LFIQA [50]	0.8219	<u>0.8316</u>
	DeeBLiF [32]	<u>0.8307</u>	0.8212
	PVBLiF (ours)	**0.8714**	**0.9009**

The Win5-LID dataset consists of 6 real-world reference scenes and 4 synthetic reference scenes. There are 220 distorted LFIs subjected to 6 types of distortions, including HEVC, JPEG2000, Linear interpolation (LN), and Nearest Neighbor interpolation (NN) with 5 distortion levels, and two CNN models with only 1 distortion level. The subjective experiments are conducted using interactive mode and Double-Stimulus Continuous Quality Scale (DSCQS) method. The MOS ranged from 1 (very annoying) to 5 (imperceptible) is provided.

The NBU-LF1.0 dataset provides 14 reference scenes, of which 8 are from the real world and 6 are synthetic. The dataset contains 210 distorted LFIs disturbed with 5 types of distortions, $i.e.$, NN, Bicubic Interpolation (BI), learning based reconstruction (EPICNN), disparity map based reconstruction (Zhang), and spatial super-resolution reconstruction (VDSR). Each distortion type has 3 distortion levels. Based on DSCQS, subjective experiments are carried out in a combination of passive and interactive modes. The MOS on a 5-point discrete scale is provided.

The SHU dataset comprises of 240 distorted LFIs derived from 8 real-world reference scenes, with 5 distortion types and 6 distortion levels. These distortion types include JPEG, JPEG2000, Gaussian blur (GAUSS), white noise, and motion blur. The passive mode and DSCQS method are adopted in the subjective evaluation. Different to the above two datasets, the SHU dataset provides the MOS ranged from 0 (bad) to 5 (excellent).

Most existing LFI datasets for quality assessment contain only hundreds of distorted LFIs generated from a small number of reference LFIs (always less than 20). If we randomly split the whole dataset into training and test sets, these two sets may contain distorted LFIs derived from the same reference LFI. Therefore, in order to ensure that the training set and test set are completely independent, we use leave-two-fold-out cross-validation to conduct the experiments following [50], [32]. Specifically, for each dataset, all the distorted LFIs generated from the same reference scene are considered as one fold. In this way, a LFI dataset containing K reference scenes will be divided into K folds of distorted LFIs, and each fold contains all distorted versions of only one reference scene. Then we randomly select K-2 folds for training and report the performance on the remaining 2 folds. There are a total of $K(K\text{-}1)/2$ possible combinations. Hence, we conduct the experiments based on all possible combinations and take the average result as the reported performance.

Additionally, we adopt three standard criteria to evaluate the performance, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Square Error (RMSE). Among them, PLCC is used to evaluate the linear relationship, SROCC focuses on the monotonicity while RMSE measures the prediction accuracy. Before the calculation of PLCC and RMSE, a five-parameter logistic mapping function [62] is employed, as shown in Eq. (14).

$$f(p) = \beta_1\{\frac{1}{2} - \frac{1}{1+e^{\beta_2(p-\beta_3)}}\} + \beta_4 p + \beta_5 \qquad (14)$$

where $\beta_{1\ldots5}$ are the fitting parameters, p and $f(p)$ denote the objective prediction and its nonlinear mapping result, respectively.

A. Overall performance comparison

This subsection compares the proposed PVBLiF metric with eight NR 2D-IQA metrics (including PIQE [63], OG [64], FRIQUEE [65], GWH-GLBP [66], NIQE [67], BRISQUE [68], HyperIQA [69], and GraphIQA [70]), three NR multi-view-IQA metrics (including MNSS [71], NIQSV [72], and NIQSV+ [73]), three NR 3D-IQA metrics (including Xu's [15], SINQ [16], and BSVQE [17]), three FR LF-IQA metrics (including MDFM [34], Min's [6], and Meng's [35]), and eight NR LF-IQA metrics (including BELIF [47], Tensor-NLFQ [8], PVRI [49], NR-LFQA [43], VBLFI [45], DSA [46], 4D-DCT-LFIQA [50], and DeeBLiF [32]). For 2D-IQA metrics and multi-view-IQA metrics, we employ these IQA metrics on all SAIs of LFI and take the average result as the final performance. For 3D-IQA metrics, we measure the quality of every two horizontal adjacent SAIs of LFI, and report the average performance. Note that the performance of all

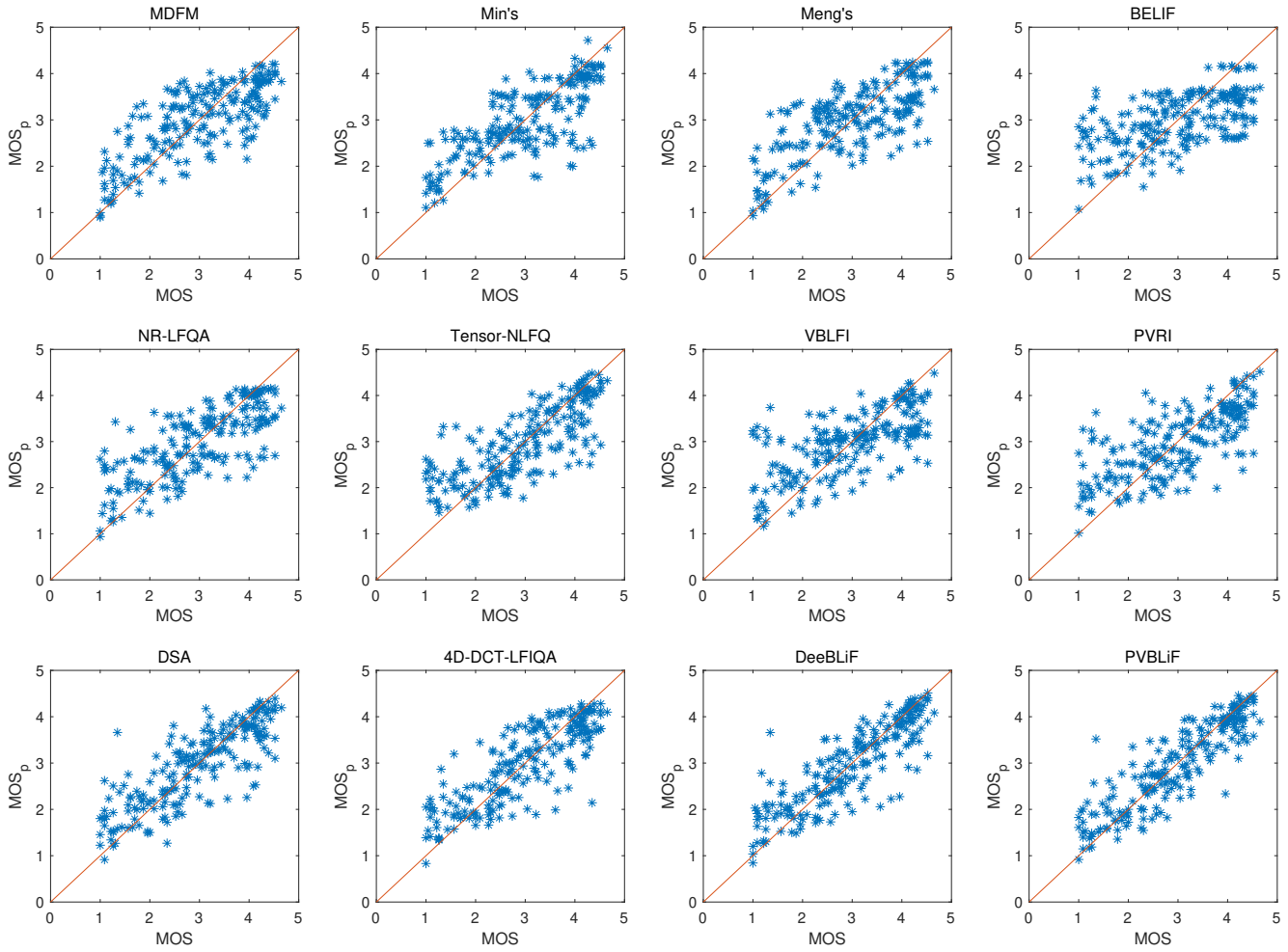


Fig. 7. Scatter plots of the objective prediction versus the subjective MOS on the Win5-LID dataset.

metrics is reported using leave-two-fold-out cross-validation for fair comparison. Specifically, the learning-based metrics are trained with K-2 folds and tested on the remaining 2 folds, while the non-learning-based metrics are directly executed on the same test set. The reported performance of all metrics is reproduced using the released code and default parameter settings from their authors to avoid bias.

TABLE II exhibits the overall performance of the proposed metric compared with state-of-the-art IQA metrics on the Win5-LID, NBU-LF1.0, and SHU datasets. From the table we can find that traditional 2D/3D/multi-view IQA metrics cannot perceive the quality of LFI accurately, because these metrics do not consider the effect of angular consistency degradation on human perception. In contrast, FR/NR LF-IQA metrics can better reflect the LFI quality owing to the additional consideration of angular quality deterioration. Among them, our previous metric DeeBLiF outperforms other existing metrics probably due to the powerful representative ability of the CNN features. This phenomenon also occurs in NR 2D-IQA metrics, where two CNN-based metrics, $i.e.$, HyperIQA and GraphIQA, achieve better performance than other handcrafted feature-based metrics. However, most existing LF-IQA metrics (including DeeBLiF) assess the LFI quality by evaluating the quality of its 2D representations, as a solution to the inherently high-dimensional attributes of LFI. Compared to previous methods, our proposed PVBLiF metric consistently achieves the best performance on three datasets. Moreover, our metric shows significant improvement over the second best metric on the Win5-LID and NBU-LF1.0 datasets. The reason may be that the Win5-LID and NBU-LF1.0 datasets contain more reconstruction distortions in LFI while the SHU dataset mainly includes common distortions in 2D images. These experimental results fully demonstrate the effectiveness of the proposed 3D representation ($i.e.$, PVBS) for LFI quality evaluation.

Additionally, we provide the scatter plots of the objective prediction versus the subjective MOS for more intuitive visualization, as shown in Fig. 7. Due to the space constraint, we only show the scatter plots of the FR/NR LF-IQA metrics on the Win5-LID dataset. MOS_p and MOS denote the prediction and its corresponding MOS, respectively. It can be found that the prediction of our metric is more in line with the subjective MOS, further demonstrating the validity of our metric.

In addition to showing the average results of all combinations in leave-two-fold-out cross-validation, the distribution of all combination results can reflect the metric generalization

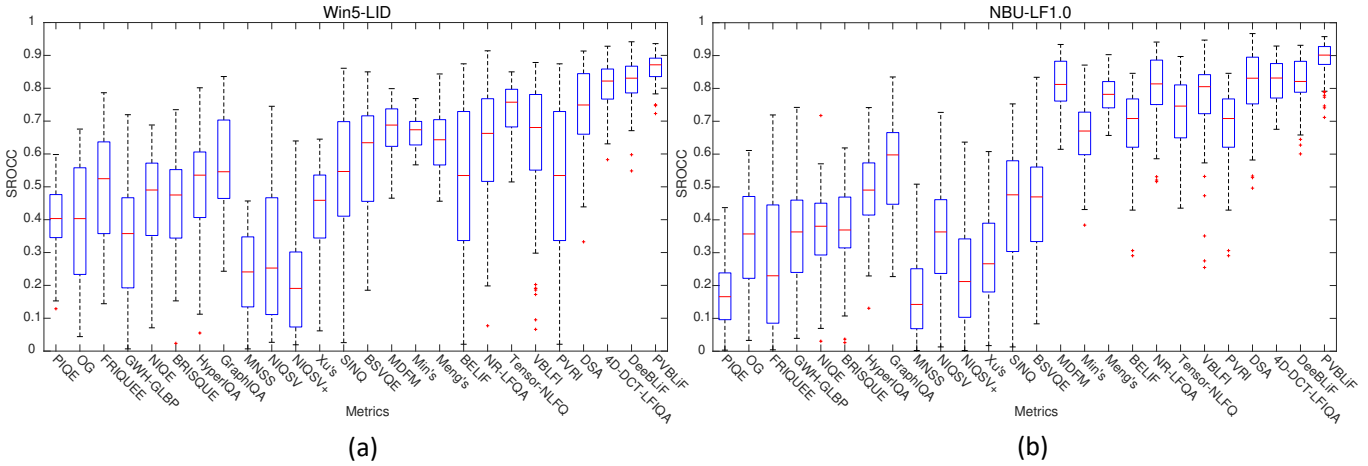


Fig. 8. Box plots of SROCC distribution in leave-two-fold-out cross-validation. (a) Win5-LID dataset; (b) NBU-LF1.0 dataset.

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

Datasets	models	PLCC	SROCC	RMSE
Win5-LID	w/o stage 1	0.8064	0.7772	0.5754
	w/o stage 2	0.8005	0.7821	0.5823
	w/o stage 3	0.8714	0.8529	0.4911
	PVBLiF	**0.8749**	**0.8580**	**0.4660**
NBU-LF1.0	w/o stage 1	0.8019	0.7811	0.4843
	w/o stage 2	0.8345	0.8149	0.4887
	w/o stage 3	0.8997	0.8842	0.3905
	PVBLiF	**0.9060**	**0.8883**	**0.3746**

TABLE V
PERFORMANCE WITH AND WITHOUT THE SVPOOLING METHOD ON THE WIN5-LID AND NBU-LF1.0 DATASETS. **BOLD** REPRESENTS THE BEST PERFORMANCE.

Datasets	Saliency	Variance	PLCC	SROCC	RMSE
Win5-LID			0.8668	0.8521	0.4762
	✓		0.8694	0.8562	0.4739
		✓	0.8706	0.8540	0.4714
	✓	✓	**0.8749**	**0.8580**	**0.4660**
NBU-LF1.0			0.9018	0.8840	0.3823
	✓		0.9045	**0.8883**	0.3775
		✓	0.9040	0.8848	0.3853
	✓	✓	**0.9060**	**0.8883**	**0.3746**

to some extent. Fig. 8 provides the box plots of SROCC distribution on the Win5-LID and NBU-LF1.0 datasets. Boxes with smaller sizes and higher positions indicate better generalization and performance, respectively. The red line in the box denotes the median SROCC result. The quantitative median SROCC performance is shown in TABLE III. We can see that the proposed PVBLiF metric achieves better generalization and performance than other state-of-the-art IQA metrics.

TABLE VI
PERFORMANCE DEPENDENCY OF HYPERPARAMETER A AND S ON THE WIN5-LID DATASET. **BOLD** REPRESENTS THE BEST PERFORMANCE.

A value	S value	PLCC	SROCC	RMSE
3	16	0.8433	0.8243	0.5177
	32	0.8559	0.8411	0.4989
	48	0.8490	0.8303	0.5106
	64	0.8428	0.8232	0.5185
5	16	0.8567	0.8428	0.4960
	32	**0.8749**	0.8580	**0.4660**
	48	0.8733	0.8551	0.4727
	64	0.8604	0.8445	0.4904
7	16	0.8614	0.8435	0.5069
	32	0.8721	0.8564	0.4705
	48	0.8678	0.8503	0.4798
	64	0.8637	0.8460	0.5006
9	16	0.8591	0.8450	0.4911
	32	0.8740	**0.8607**	0.4669
	48	0.8697	0.8536	0.4775
	64	0.8641	0.8478	0.4897

B. Ablation study

The proposed PVBSNet contains four stages with different functions, and the efficacy of each stage deserves further investigation. For this goal, the ablation studies of multi-information division (stage 1), intra-feature extraction (stage 2), and cross-feature fusion (stage 3) are performed on the Win5-LID and NBU-LF1.0 datasets, and the results are reported in TABLE IV. Here, we do not delve into the efficacy of the stage 4, as it is deemed necessary in our metric. From the table we can see that without stage 1, it is difficult for CNNs to learn discriminative spatio-angular features from the scrambled temporal information in PVS, thus the quality assessment performance drops dramatically. Likewise, stage 2 also shows a crucial role in extracting deep feature for quality assessment. The incorporation of stage 3 promotes a marginal improvement in the final performance.

TABLE VII
PERFORMANCE COMPARISON OF DIFFERENT GENERATION ORDERS ON THE WIN5-LID DATASET.

Generation Orders	PLCC	SROCC	RMSE
Serpentine	0.8573	0.8434	0.4928
Spiral	0.8606	0.8446	0.4887
Raster (ours)	0.8749	0.8580	0.4660

TABLE VIII
PERFORMANCE COMPARISON OF DIFFERENT SALIENCY MODELS ON THE WIN5-LID AND NBU-LF1.0 DATASETS.

Datasets	Saliency Models	PLCC	SROCC	RMSE
Win5-LID	w/o	0.8668	0.8521	0.4762
	Wavelet	0.8731	0.8557	0.4677
	Cluster	0.8727	0.8572	0.4693
	SDSP	0.8749	0.8580	0.4660
NBU-LF1.0	w/o	0.9018	0.8840	0.3823
	Wavelet	0.9039	0.8863	0.3850
	Cluster	0.9042	0.8855	0.3787
	SDSP	0.9060	0.8883	0.3746

The proposed PVBLiF assesses the LFI quality by evaluating the quality of each PVBS, thus a post-processing method is required to pool all the PVBS quality into the overall LFI quality. Most previous works, including our previously proposed DeeBLiF, adopt the commonly-used average pooling to obtain the LFI quality. In this subsection, to verify the effectiveness of the proposed SVPooling method and its two components, Saliency and Variance, we conduct the corresponding experiments on the Win5-LID and NBU-LF1.0 datasets. The experimental results are reported in TABLE V. From the table we can see that using Saliency or Variance alone for post-processing can improve performance to a certain extent. The combination of Saliency and Variance, $i.e.$, the proposed SVPooling, yields the best quality assessment performance. These experimental results show that both Saliency and Variance parts help to obtain a more accurate overall quality. This may be because the SVPooling method is able to filter out the low-variance PVBSs with inaccurate quality predictions, and reweight each PVBS quality according to its visual saliency.

C. Hyperparameter dependency

As mentioned before, our metric involves two manual hyperparameters that determine the length and the spatial resolution of the generated PVBS, $i.e.$, A and S. In this subsection, we investigate the performance dependency of different hyperparameter values in our metric. TABLE VI exhibits the experimental results of different hyperparameter values on the Win5-LID dataset. When the hyperparameter S is fixed, it is observed that setting a small value for A leads to relatively poor performance, whereas similar performance can be achieved by setting A to 5 or higher. One possible explanation is that the selection of A directly determines the amount of angular information, and the lack of angular information will directly affect the prediction accuracy. Besides, experimental results also show that our proposed metric can be fully exploited with moderate angular information. When the hyperparameter A is fixed, we can see that adopting a moderate S value achieves better performance than a smaller or larger value. The reason to this phenomenon could be that as the value of S increases, the spatial information of a single PVBS increases while the number of training samples decreases, and vice versa. A more reasonable trade-off can be obtained with a moderate S value, resulting in better performance. Based on the above analyses, we set A and S to 5 and 32, respectively.

In our metric, we use raster order to generate PVBS due to the availability of the regular horizontal and vertical information. To investigate the impact of different generation orders, we conduct comparative experiments on the Win5-LID dataset with two other widely-used generation orders [53], i.e., serpentine and spiral orders, as shown in TABLE VII. It can be seen that using serpentine or spiral order results in a performance degradation compared to raster order. The reason behind this is that with raster order, the subsequent multi-information division can extract effective angular information, which facilitates learning the influence of different distortions on reference LFIs.

In addition, it is necessary to investigate the dependency of different saliency models in the SVPooling method. Here, three saliency models are adopted to validate the effectiveness of SVPooling, including Wavelet [74], Cluster [75], and SDSP [60]. The experiments are conducted on the Win5-LID and NBU-LF1.0 datasets, as shown in TABLE VIII. It can be found that on both datasets, using different saliency models can improve performance to varying degrees, implying that the SVPooling method has a relatively weak dependence on the selected saliency model.

TABLE IX
PERFORMANCE OF TRAINING ON THE NBU-LF1.0 AND SHU DATASETS, AND TESTING ON THE WIN5-LID DATASET.

Training datasets	PLCC	SROCC	RMSE
NBU-LF1.0 (NN)	0.8084	0.7113	0.4238
SHU (JPEG2000)	0.8470	0.7874	0.4570

D. Cross-dataset validation

To investigate the cross-dataset robustness of the proposed PVBLiF, we conduct the cross-dataset experiments in this subsection. Specifically, we train our metric on the NBU-LF1.0 and SHU datasets, and report the performance on the Win5-LID dataset. Note that both Win5-LID and NBU-LF1.0 datasets contain the NN distortion, while both Win5-LID and SHU datasets include the JPEG2000 distortion. As shown in TABLE IX, even when trained on other dataset, our proposed PVBLiF still achieves competitive performance, which proves that our metric has strong cross-dataset robustness.

TABLE X
PLCC PERFORMANCE OF DIFFERENT DISTORTION TYPES ON THE WIN5-LID AND NBU-LF1.0 DATASETS. **BOLD** AND UNDERLINE REPRESENT THE BEST AND SECOND BEST PERFORMANCE, RESPECTIVELY.

Metric Types	Metrics	Win5-LID				NBU-LF1.0				
		HEVC	JPEG2000	LN	NN	NN	BI	EPICNN	Zhang	VDSR
NR 2D-IQA	PIQE [63]	0.6971	0.8257	0.4681	0.3047	0.2176	0.3402	0.3732	0.3615	0.8244
	OG [64]	0.6701	0.6104	0.6003	0.4147	0.4127	0.5743	0.6583	0.4729	0.8514
	FRIQUEE [65]	0.8090	0.7663	0.6810	0.5342	0.5301	0.7105	0.6978	0.3338	0.6564
	GWH-GLBP [66]	0.7104	0.6002	0.6442	0.4606	0.5709	0.5950	0.7109	0.5233	0.6913
	NIQE [67]	0.8361	0.7317	0.6669	0.4331	0.4600	0.5225	0.7018	0.5357	0.9171
	BRISQUE [68]	0.8280	0.7629	0.6731	0.5055	0.4981	0.5634	0.7097	0.6310	0.8969
	HyperIQA [69]	0.8944	0.8333	0.6203	0.4707	0.5564	0.6771	0.8271	0.6442	0.9123
	GraphIQA [70]	0.9530	0.8723	0.6373	0.4326	0.6177	0.7111	0.8219	0.7332	0.9463
NR multi-view-IQA	MNSS [71]	0.5843	0.6507	0.6765	0.5558	0.5551	0.8017	0.5937	0.7666	0.5614
	NIQSV [72]	0.4851	0.3694	0.4391	0.3162	0.2678	0.4105	0.5998	0.3559	0.6348
	NIQSV+ [73]	0.4598	0.4213	0.4536	0.3536	0.2986	0.4169	0.4911	0.4119	0.5226
NR 3D-IQA	Xu's [15]	0.6545	0.7278	0.6702	0.5848	0.4585	0.5123	0.6745	0.4920	0.8817
	SINQ [16]	0.7135	0.7928	0.8133	0.6982	0.6514	0.6252	0.7412	0.6390	0.8586
	BSVQE [17]	0.7558	0.7301	0.7300	0.6990	0.6407	0.5759	0.6484	0.4442	0.9078
NR LF-IQA	BELIF [47]	0.8062	0.7275	0.7172	0.7219	0.9244	0.8732	0.6707	0.6886	0.8749
	Tensor-NLFQ [8]	0.8909	0.8340	0.8543	0.8446	0.8517	0.9199	0.8395	0.7135	0.9223
	PVRI [49]	0.8352	0.8702	0.7677	0.7890	0.8396	0.8967	0.8551	0.8268	0.9149
	NR-LFQA [43]	0.7641	0.8098	0.7731	0.7920	0.9544	0.9519	0.9157	0.7108	0.8850
	VBLFI [45]	0.8037	0.8273	0.8151	0.7261	0.9056	0.9276	0.8729	0.7072	0.9526
	DSA [46]	0.8673	0.8647	0.8338	0.8542	0.9472	0.9243	0.8670	0.8010	0.9342
	4D-DCT-LFIQA [50]	0.9001	0.9365	0.8803	0.8534	0.9386	0.9389	0.8183	**0.9048**	0.9317
	DeeBLiF [32]	0.9389	0.9254	**0.9021**	0.9207	0.9610	0.9499	0.9395	0.6659	0.9487
	PVBLiF (ours)	**0.9768**	**0.9388**	0.8941	**0.9286**	**0.9688**	**0.9636**	**0.9498**	0.9011	**0.9605**

E. Robustness against distortion types

Since LFI may suffer from different types of distortions during the image processing chain, we aim to investigate the robustness against distortion types of our metric in this subsection. TABLE X shows the experimental results for individual distortion type on the Win5-LID and NBU-LF1.0 datasets. Two CNN-based distortions in Win5-LID dataset are not considered here since they have only one distortion level. Limited by the paper space, we only show the PLCC results, while the results of SROCC and RMSE are similar. The table shows that the proposed PVBLiF not only achieves the best or second best performance in all distortion types, but also yields significant improvements in some distortion types, such as HEVC. These experimental results demonstrate that the proposed metric has strong robustness when confronting different distortion types.

F. Statistical significance test

In the subsection, we perform the F-test to compare the statistical significance between any two LF-IQA metrics following [73]. Specifically, we first calculate the residuals between the predictions MOS_p and their corresponding subjective scores MOS for each metric,

$$Res^M = MOS_p^M - MOS^M \tag{15}$$

where Res^M denotes the residuals of metric M.

Since leave-two-fold-out cross-validation is adopted in our experiments, Res^M contains the residual results of all training-test splits for a comprehensive significance test. Then the $vartest2$ function in Matlab is used to calculate the significance relationship of any two metrics, in which the confidence level is set to 95%. The statistical significance results between any two LF-IQA metrics on three datasets are shown in Fig. 9, where "1", "0", and "-1" indicate that the row metric has better, competitive, and worse statistical performance than the column metric, respectively. From the figure, we can see that on the SHU dataset, the proposed PVBLiF metric significantly outperforms most FR/NR LF-IQA metrics, only our previous metric DeeBLiF has competitive statistical performance to our PVBLiF metric. On the Win5-LID and NBU-LF1.0 datasets, our PVBLiF metric has significantly better statistical performance than all state-of-the-arts. These demonstrate the superiority of our metric in LFI quality evaluation.

G. Time complexity

Before employing an IQA metric to practical applications, we need to consider its time complexity. For fair comparison, all the metrics are executed using the same hardware configurations (CPU: Intel (R) i7-10700 2.90GHz; GPU: NVIDIA GeForce RTX 3080 10G; Memory: 64G RAM). The test time represents the runtime for testing a single LFI, excluding

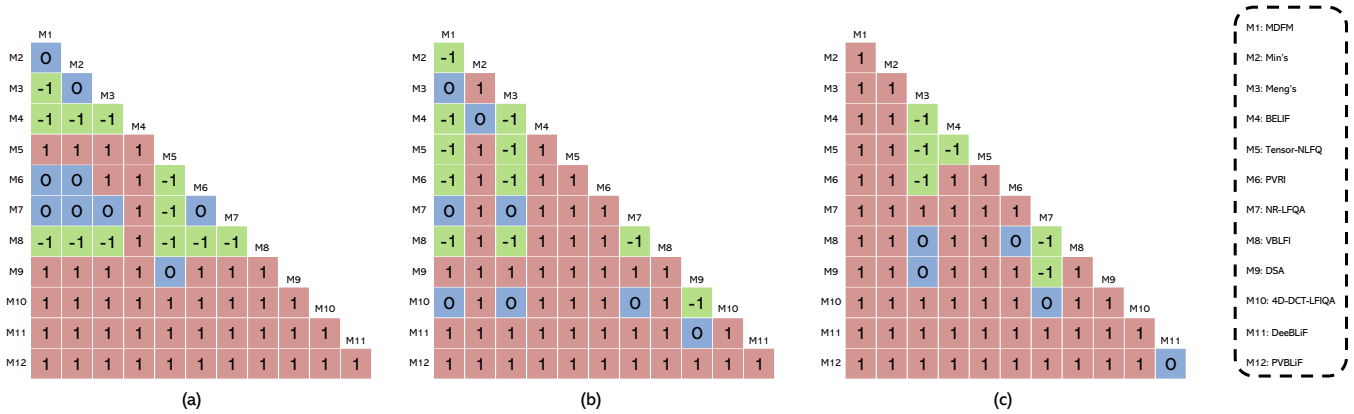


Fig. 9. Statistical significance between any two LF-IQA metrics. (a) Win5-LID dataset; (b) NBU-LF1.0 dataset; (c) SHU dataset. Here, "1", "0", and "-1" indicate that the row metric has better, competitive, and worse statistical performance than the column metric, respectively.

TABLE XI
PERFORMANCE COMPARISON OF THE TEST TIME AGAINST SROCC PERFORMANCE ON THE WIN5-LID DATASET. **BOLD** AND UNDERLINE REPRESENT THE BEST AND SECOND BEST PERFORMANCE, RESPECTIVELY. METRICS WITH AND WITHOUT * DENOTE THE RUNTIME USING GPU AND CPU, RESPECTIVELY.

Metric Types	Metrics	Win5-LID	
		Test time (s/LFI)	SROCC
NR 2D-IQA	PIQE [63]	5.1466	0.3920
	OG [64]	2.9768	0.3960
	FRIQUEE [65]	881.2774	0.4978
	GWH-GLBP [66]	4.8472	0.3391
	NIQE [67]	6.0077	0.4482
	BRISQUE [68]	3.1543	0.4559
	HyperIQA [69]	10.7484	0.6246
	HyperIQA* [69]	5.7384	0.6246
	GraphIQA [70]	5.4899	0.6263
	GraphIQA* [70]	2.8354	0.6263
NR multi-view-IQA	MNSS [71]	8.0746	0.2470
	NIQSV [72]	1.8744	0.3130
	NIQSV+ [73]	2.5747	0.2174
NR 3D-IQA	Xu's [15]	62.4815	0.4345
	SINQ [16]	187.7167	0.5075
	BSVQE [17]	169.4089	0.5770
FR LF-IQA	MDFM [34]	**0.8537**	0.6768
	Min's [6]	3.9845	0.6645
	Meng's [35]	30.4872	0.6347
NR LF-IQA	BELIF [47]	107.8814	0.5119
	Tensor-NLFQ [8]	697.6515	0.7345
	PVRI [49]	71.3578	0.6827
	NR-LFQA [43]	225.2069	0.6275
	VBLFI [45]	65.6667	0.6116
	DSA [46]	198.5443	0.7328
	4D-DCT-LFIQA [50]	169.2623	0.8079
	DeeBLiF [32]	4.8533	0.8186
	DeeBLiF* [32]	2.1794	0.8186
	PVBLiF (ours)	8.3795	**0.8580**
	PVBLiF* (ours)	3.9382	**0.8580**

the time of data loading and model initialization. All the CNN-based metrics are reported using Pytorch, while other metrics are performed using Matlab. For CNN-based metrics, the time complexity of using CPU only and GPU are reported independently. For other metrics, only time complexity of using CPU is reported. TABLE XI shows the test time against the SROCC performance on the Win5-LID dataset, where metrics with and without * denote the runtime using GPU and CPU, respectively. From the table we can find that most NR LF-IQA metrics are time-consuming, because they rely on extracting multiple handcrafted features to ensure the prediction accuracy. In contrast, the proposed PVBLiF achieves the best performance with low time complexity, probably because our metric is a 3D extension of patch-based methods, inheriting the advantage of low computational complexity.

V. CONCLUSION

In this paper, we propose a novel Pseudo Video-based Blind quality assessment metric for Light Field image (PVBLiF). First, considering the limitations of 2D representations in reflecting LFI quality, a more intuitive 3D representation, Pseudo Video Block Sequence (PVBS), is used for the quality assessment of LFI. To achieve this goal, we exploit the potential of CNN structure and propose a novel CNN-based network, named PVBSNet, to extract the spatio-angular features of PVBS and predict the PVBS quality. Further, since different regions of the same scene have different visual effects on the human perception, we present a Saliency- and Variance-guided Pooling (SVPooling) method to integrate all the PVBS quality into the overall LFI quality. Finally, to validate the superiority of the proposed PVBLiF metric, we conduct extensive experiments with 25 existing 2D/3D/multi-view/LF IQA metrics on three publicly available LFI datasets. Experimental results demonstrate that our metric outperforms the state-of-the-art metrics and shows great potential to simulate the human perception. In the future, we aim to investigate the impact of training strategies and pre-trained models on LF-IQA metrics.

REFERENCES

[1] G. Arun, "The light field," *J. Math. Phys.*, vol. 18, pp. 51-151, 1936.

[2] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Proc. Comput. Models Visual Process.*, MIT Press, 1991, pp. 3-20.

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

[4] Y. Liu, L. Fang, D. Gutierrez, Q. Wang, J. Yu, and F. Wu, "Introduction to the issue on light field image processing," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 923-925, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202-211, 2009.

[10] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imag.*, vol. 2016, no. 16, pp. 1-6, 2016.

[11] P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, "No-reference image quality assessment based on statistics of local ternary pattern," in *Proc. IEEE Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2016, pp. 1-6.

[12] P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, "No-reference image quality assessment using orthogonal color planes patterns," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3353-3360, 2018.

[13] T. Chinen, J. Ballé, C. Gu, et al., "Towards a semantic perceptual image metric," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2018, pp. 624-628.

[14] Z. Zhou, J. Li, Y. Quan, and R. Xu, "Image quality assessment using kernel sparse coding," *IEEE Trans. Multimedia*, vol. 23, pp. 1592-1604, 2020.

[15] 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.

[16] 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.

[17] 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, 2018.

[18] S. Ling, J. Li, Z. Che, X. Min, G. Zhai, and P. L. Callet, "Quality assessment of free-viewpoint videos by quantifying the elastic changes of multi-scale motion trajectories," *IEEE Trans. Image Process.*, vol. 30, pp. 517-531, 2021.

[19] 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.

[20] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2015, pp. 4733-4737.

[21] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, "Pseudo-sequence-based light field image compression," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2016, pp. 1-4.

[22] G. Wang, W. Xiang, M. Pickering, and C. W. Chen, "Light field multi-view video coding with two-directional parallel inter-view prediction," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5104-5117, 2016.

[23] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo-sequence-based 2-D hierarchical coding structure for light-field image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1107-1119, 2017.

[24] H. Amirpour, M. Pereira, and A. Pinheiro, "High efficient snake order pseudo-sequence based light field image compression," in *Proc. Data Compress. Conf. (DCC)*, 2018, pp. 397-397.

[25] 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. (ISNCC)*, 2019, pp. 1-5.

[26] 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.

[27] 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.

[28] J. Wu, Y. Liu, W. Dong, G. Shi, and W. Lin, "Quality assessment for video with degradation along salient trajectories," *IEEE Trans. Multimedia,* vol. 21, no. 11, pp. 2738-2749, 2019.

[29] P. Chen, L. Li, L. Ma, J. Wu, and G. Shi, "RirNet: Recurrent-in-recurrent network for video quality assessment," in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp. 834-842.

[30] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3D convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 4447-4451.

[31] 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.

[32] 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.

[33] 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.

[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] 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.

[36] 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, 2020.

[37] 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.

[38] 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.

[39] 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," *Frontiers Comput. Neurosci.*, vol. 15, pp. 768021, 2022.

[40] 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.

[41] 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.

[42] Z. Luo, W. Zhou, L. Shi, and Z. Chen, "No-reference light field image quality assessment based on micro-lens image," in *Proc. Picture Coding Symp. (PCS)*, 2019, pp. 1-5.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

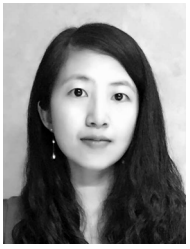
[50] J. Xiang, G. Jiang, M. Yu, Z. Jiang, and Y.-S. Ho, "No-reference light field image quality assessment using four-dimensional sparse transform," *IEEE Trans. Multimedia*, vol. 25, pp. 457-472, 2023.
[51] Z. Guo, W. Gao, H. Wang, J. Wang, and S. Fan, "No-reference deep quality assessment of compressed light field images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2021, pp. 1-6.
[52] 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.
[53] C. Conti, L. D. Soares, and P. Nunes, "Dense light field coding: A survey," *IEEE Access*, vol. 8, pp. 49244-49284, 2020.
[54] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1-14, 2011.
[55] X. Shang, J. Liang, G. Wang, H. Zhao, C. Wu, and C. Lin, "Color-sensitivity-based combined PSNR for objective video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1239-1250, 2019.
[56] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2019, pp. 2351-2359.
[57] K. Cho, B. V. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2014, pp. 1724-1734.
[58] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284-2298, 2007.
[59] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270-4281, 2014.
[60] L. Zhang, Z. Gu, and H. Li, "SDSP: A novel saliency detection method by combining simple priors," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2013, pp. 171-175.
[61] L.-M. Po, M. Liu, W. Y. F. Yuen, Y. Li, X. Xu, C. Zhou, P. H. W. Wong, K. W. Lau, and H.-T. Luk, "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, 2019.
[62] 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.
[63] 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.
[64] 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.
[65] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *J. Vis.*, vol. 17, no. 1, pp. 32-32, 2017.
[66] 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, 2016.
[67] 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, 2012.
[68] 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, 2012.
[69] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 3664-3673.
[70] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Trans. Multimedia*, 2022, doi: 10.1109/TMM.2022.3152942.
[71] 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, 2020.
[72] 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.
[73] 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, 2018.
[74] N. Imamoglu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 96-105, 2013.
[75] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766-3778, 2013.

Zhengyu Zhang received the B.E. degree in electronic and information science and technology from Guangzhou University, Guangzhou, China, and the M.E. degree in electronics and communication engineering from Shenzhen University, Shenzhen, China, in 2018 and 2021, respectively. 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 Telecommunications of Rennes Laboratory. His research interests include image/video quality assessment, 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.

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 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.