EDDMF: An Efficient Deep Discrepancy Measuring Framework For Full-Reference Light Field Image Quality Assessment

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

EDDMF: An Efficient Deep Discrepancy Measuring Framework For Full-Reference Light Field Image Quality Assessment

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

***Abstract*—The increasing demand for immersive experience has greatly promoted the quality assessment research of Light Field Image (LFI). In this paper, we propose an efficient deep discrepancy measuring framework for full-reference light field image quality assessment. The main idea of the proposed framework is to efficiently evaluate the quality degradation of distorted LFIs by measuring the discrepancy between reference and distorted LFI patches. Firstly, a patch generation module is proposed to extract spatio-angular patches and sub-aperture patches from LFIs, which greatly reduces the computational cost. Then, we design a hierarchical discrepancy network based on convolutional neural networks to extract the hierarchical discrepancy features between reference and distorted spatio-angular patches. Besides, the local discrepancy features between reference and distorted sub-aperture patches are extracted as complementary features. After that, the angular-dominant hierarchical discrepancy features and the spatial-dominant local discrepancy features are combined to evaluate the patch quality. Finally, the quality of all patches is pooled to obtain the overall quality of distorted LFIs. To the best of our knowledge, the proposed framework is the first patch-based full-reference light field image quality assessment metric based on deep-learning technology. Experimental results on four representative LFI datasets show that our proposed framework achieves superior performance as well as lower computational complexity compared to other state-of-the-art metrics.**

***Index Terms*—Light field, image quality assessment, full-reference, patch, deep-learning.**

I. Introduction

IN the era of flourishing visual communication, traditional 2D images can no longer satisfy people's yearning for immersive experience. As a novel imaging technology, Light Field Image (LFI) has received a broad spotlight for its potential to provide more immersive experience, such as Virtual Reality (VR) [1] and Augmented Reality (AR) [2]. Different from 2D images, LFI is a high-dimensional imaging format for recording the distribution of light rays, which was originally represented as a 7D function [3], [4]. To facilitate practical applications, the 7D LFI is reduced to 4D representation [5] and described as a biplane model $L(u, v, h, w)$, where (u, v) are the coordinates of angular plane ($i.e.$, camera plane) and (h, w) are the coordinates of spatial plane ($i.e.$, scene plane), as shown in Fig. 1 (a). By fixing each group of angular coordinates (u, v), a 4D LFI can be represented as a Sub-Aperture Image (SAI) array. The SAI array of LFI is a 2D image array as shown in Fig. 1 (b), and it can be regarded as the data obtained by photographing the same scene from an array of viewpoints with narrow parallax. In addition to the SAI array, LFI can be visualized in several other representations, $e.g.$, Refocused Image (RI), Epipolar Plane Image (EPI), Pseudo Video Sequence (PVS), and MicroLens Image (MLI).

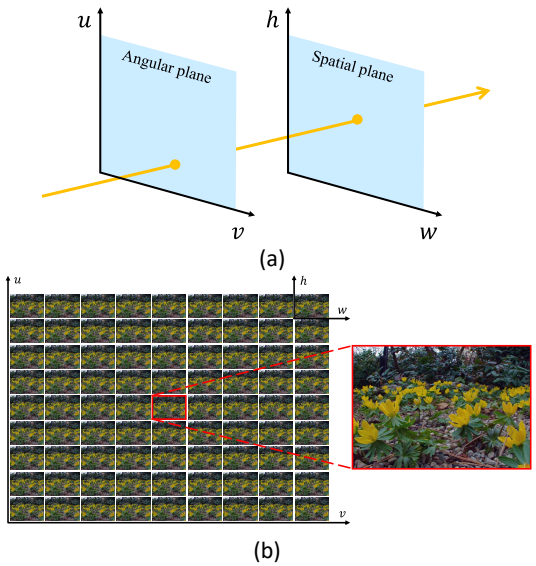


Fig. 1. (a) Biplane model for describing the 4D LFI. (b) Example of SAI array and its enlarged central view [6].

In the image processing chain [7], [8], the quality degradation of LFIs will inevitably be introduced by compression [9] [10], reconstruction [11], [12] and rendering [13] methods, thus it is necessary to evaluate the distortion influence on the LFI quality. Since human eyes are the ultimate recipient of visual information, subjective experiments based on Human Visual System (HVS) are the most reliable way to evaluate

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, and in part by the China Scholarship Council.

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)

Corresponding author: Shishun Tian

the impact of image quality degradation. However, they are extremely time-consuming and resource-intensive. Further, traditional objective Image Quality Assessment (IQA) metrics cannot well deal with the special distortions in LFIs. For example, 2DIQA metrics [14], [15], [16], [17], [18], [19], [20] mainly focus on spatial distortions in 2D images. 3DIQA metrics [21], [22], [23] are designed for image pairs with wide parallax, not for LFIs with narrow parallax. Multiview-IQA metrics [24], [25], [26], [27], [28] tend to measure the view-synthesis distortions such as stretching and black holes. Consequently, an objective LFIQA metric that can accurately evaluate the LFI quality is in great demand. In recent years, researchers have made a lot of efforts on LFIQA and proposed several landmark LFIQA metrics [29]. Existing LFIQA metrics can be generally grouped into three categories based on the availability of reference information: No-Reference (NR), no reference information required; Reduced-Reference (RR), partial reference information required; Full-Reference (FR), full reference information required. Among them, the FR LFIQA metric can obtain the most stable results due to its access to full reference information, which is the focus of this paper.

The incorporation of multiple SAIs from different perspectives in LFIs inevitably results in an enormous amount of data. Currently, most existing LFIQA metrics ($e.g.$, [30], [31], [32]) suffer from high computational complexity due to the use of full LFI data. Many previous quality assessment metrics on 2D images [33], [34], [35] have demonstrated the effectiveness of using patches to evaluate the overall quality. Analogously, two previous works on LFIQA [36], [37] also show that evaluating the LFI quality with patches is a promising approach to reduce the computational complexity. However, these patch-based LFIQA metrics are all NR metrics. Compared to the NR metrics, the FR metrics are undoubtedly more reliable with access to reference information, which motivates us to develop an efficient patch-based FR LFIQA metric in this paper. First, since LFI patches need to reflect the characteristics of high-dimensional LFIs, we exploit two kinds of LFI patches, spatio-angular patches and sub-aperture patches, both of which contain angular and spatial quality degradation information of LFIs. In addition, although spatial and angular information are two different characteristics of LFIs, existing LFIQA metrics generally adopt homogeneous features for both information, $i.e.$, Convolutional Neural Network (CNN) features only or handcrafted features only. However, we argue that they should be treated differently to maximize the efficiency and effectiveness. For angular information, due to the non-intuitive nature of angular information deterioration, it is often perceived in combination with spatial information. Therefore, CNN features with strong discriminative ability are more suitable for angular information. For spatial information, existing handcrafted feature-based FR metrics can handle most spatial distortions well when reference information is available. As a result, a combination of angular-dominant CNN features and spatial-dominant handcrafted features is presented as a more comprehensive description for LFI patches.

Motivated by the above facts, in this paper, we propose the first patch-based FR LFIQA framework adopting CNNs, named Efficient Deep Discrepancy Measuring Framework (EDDMF). The main idea of the proposed EDDMF is to efficiently evaluate the quality degradation of distorted LFIs by measuring the discrepancy between reference and distorted LFI patches. The main contributions of this paper are summarized as follows.

1) To address the efficiency problem when dealing with high-dimensional LFIs, a patch generation module is proposed for preprocessing. Specifically, spatio-angular patches and sub-aperture patches are generated for quality assessment, both of which contain angular and spatial quality degradation information of LFIs. As a result, the computational complexity of the overall framework is significantly reduced.

2) To extract sufficient information from LFI patches to estimate the LFI quality, a hierarchical discrepancy network based on CNNs is designed to extract the angular-dominant hierarchical discrepancy features between reference and distorted spatio-angular patches. In addition, a local discrepancy extraction module is presented to generate the spatial-dominant local discrepancy features from reference and distorted sub-aperture patches, which are exploited as complementary features to evaluate the quality degradation in LFI patches.

3) To fully demonstrate the effectiveness of the proposed EDDMF, extensive experiments are conducted on four representative LFI datasets. Experimental results show that compared with state-of-the-art metrics, the proposed EDDMF achieves superior quality evaluation performance with a relatively low computational complexity.

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

II. Related Works

A. Light Field Processing

Light field image/video enables a wide range of attractive applications because of its abundant information. However, such a large amount of information also brings challenges in various aspects such as light field storage, transmission and display [8]. Therefore, light field compression techniques [38], [39], [40], [41], [42], [43], [44] are of great significance to improve the coding efficiency. Among which, LFI compression methods can be further divided into two categories: lossy [38], [39], [40] and lossless [41], [42]. Currently, lossy compression methods are receiving more attention as they can achieve higher bitrate reduction than lossless compression methods. Unfortunately, both encoding and decoding processes in lossy compression methods inevitably distort the LFI structure to a certain extent, thereby affecting the Quality of Experience (QoE) of end-users. In addition, restricted by the resolution trade-off of the LFI acquisition hardware [8], spatial super-resolution [45], [46] and angular reconstruction [47], [48] are two important means to expand the resolution of LFIs. However, due to various factors such as object occlusion and Lambertian reflectance, these methods may produce some visually discontinuous regions, which also degrade the LFI quality perceived by human eyes. In addition to the aforementioned

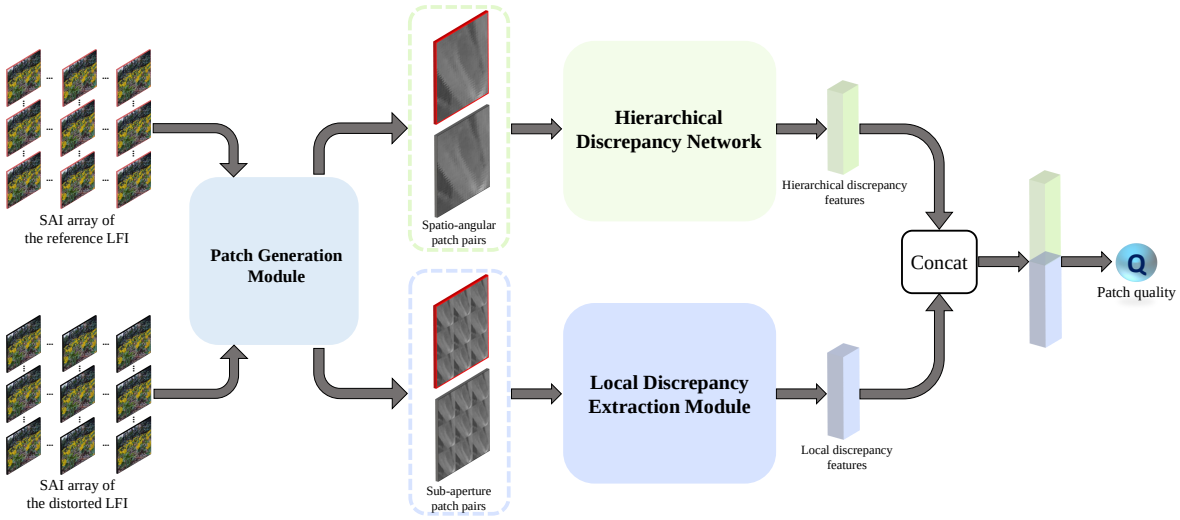


Fig. 2. Overview of the proposed EDDMF. Note that patches with and without red borders denote reference patches and distorted patches, respectively.

quality-impairing processing related to LFI's characteristics, some common image/video distortions may also appear in LFIs and lead to quality deterioration.

B. Light Field Image Quality Assessment

As mentioned before, existing LFIQA metrics can be generally grouped into NR, RR, and FR categories, depending on the amount of reference information involved.

The NR LFIQA metrics evaluate the quality of distorted LFIs without reference information. Most NR LFIQA metrics [49], [50], [30], [51], [52], [53], [54], [55] focus on extracting Natural Scene Statistics (NSS) features to describe the whole LFI and further estimate its quality. BELIF [49] first generates cyclopean images tensor from SAIs, and then explores tensor spatial characteristic features and tensor structure variation index to measure the spatial and angular quality, respectively. NR-LFQA [50] combines the naturalness distribution of cyclopean image array and gradient histogram of EPIs to evaluate the LFI quality. Tensor-NLFQ [30] adopts tensor decomposition on SAI stacks in four directions, and further extracts naturalness statistics features and structural similarity distribution for LFI quality evaluation. VBLFI [51] adopts curvelet transform on the mean difference image and SAIs, and then extracts statistical and energy features to measure the LFI quality. PVRI [52] measures the angular quality from the structure, motion and disparity information of the decomposed PVS, and evaluates the spatial quality from the depth and semantic information of RIs. TSSV-LFIQA [53] assesses the quality deterioration of LFIs using the sharpness and distribution information of tensor slice and the percentage of singular value. PM-BLFIQM [54] measures the LFI quality based on the disparity information of local light fields and the angular consistency of MLI. 4D-DCT-LFIQA [55] extracts naturalness distribution and energy features in 4D frequency domain for LFI quality evaluation.

The RR LFIQA metrics use partial information from reference LFIs to measure the distorted LFI quality. For instance, Paudyal *et al.* [56] proposed a RR LFIQA metric, which computes the depth map similarity between the distorted and reference LFIs as the predicted quality.

The FR LFIQA metrics utilize the differences between reference and distorted LFIs for quality assessment. Existing FR LFIQA metrics [57], [58], [59], [31], [60], [61], [32], [62], [63] generally capture these differences using handcrafted features, and obtain the quality of distorted LFIs via pooling methods. Fang *et al.* [57] calculated the LFI quality by combining the gradient magnitude of SAIs and EPIs. MDFM [58] extracts the multi-order derivative information on SAIs to assess the LFI quality. Min *et al.* [59] computed the global and local quality on SAIs using the view structure matching and the near-edge mean square error, respectively. Meng *et al.* [31] adopted Gaussian operator on the central SAI to measure the spatial quality, and computed the structural similarity between distorted and reference RIs to evaluate the angular quality. In LGF-LFC [60], single-scale and multi-scale log-Gabor filters are applied on SAIs and EPIs to evaluate the local and global quality, respectively. SDFM [61] captures the spatial quality based on the symmetry information of SAIs, and measures the angular quality from the geometry information of EPIs. KRIQE [32] exploits the gradient magnitude and phase congruency of the key RIs for LFI quality evaluation. CTM [62] estimates the LFI quality by applying contourlet transform on SAIs to measure the multi-scale information. As the updated version of CTM, SGFM [63] was recently proposed to further explore the geometry quality of each multiview sequence by applying 3D-Gabor filter.

The naturally massive LFI data not only leads to complicated human visual perception, but also brings great challenges to the efficiency of quality assessment algorithms. Most of the aforementioned works suffer from high computational complexity due to the use of full LFI data. To deal with this issue, Zhao *et al.* [36] pioneeringly proposed a NR metric for LFIQA, in which only a small number of patches are extracted to measure the LFI quality, thus greatly reducing the computational complexity. However, the effectiveness of this metric is significantly constrained when dealing with

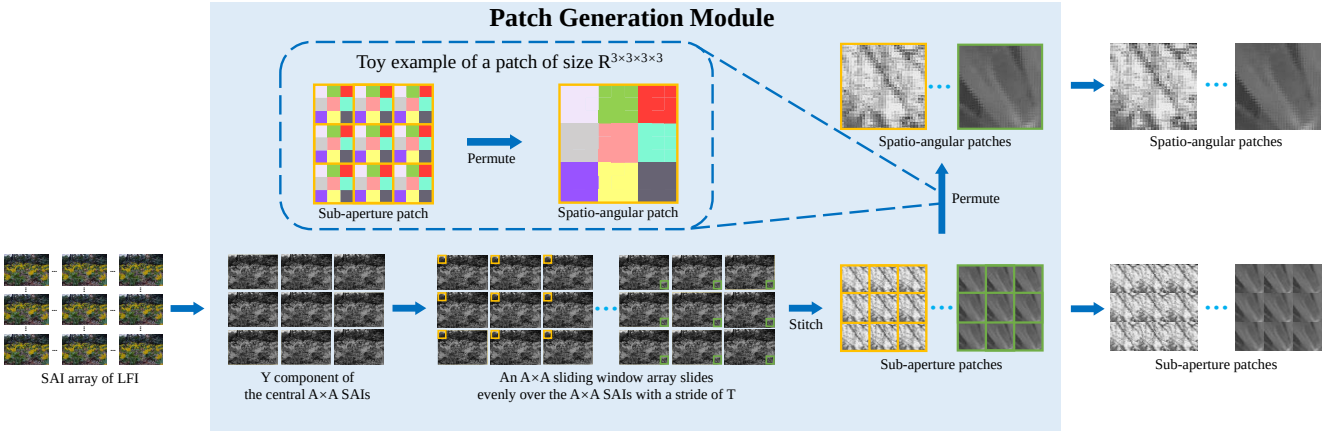


Fig. 3. Pipeline of the proposed patch generation module. For better visualization, A is set to 3 for all illustrations with $A \times A$ angular resolution.

LFIs of low angular resolution. Further, our previous work [37] also presented a patch-based NR LFIQA metric, named DeeBLiF, which achieves competitive performance in most circumstances while maintaining relatively low computational costs. Considering the underutilized potential of patch-based methods in the FR LFIQA research, we develop an efficient patch-based FR LFIQA framework (EDDMF) in this paper, which is extended from the DeeBLiF metric, and the extensions are summarized as follows.

First, DeeBLiF extracts angular and spatial features separately from spatio-angular patches, and obtains spatio-angular features through a naive concatenation operation. The resulting spatio-angular features are not discriminative enough for both angular and spatial quality degradation. On the contrary, the proposed EDDMF evaluates the LFI quality based on two kinds of LFI patches, spatio-angular patches and sub-aperture patches, which focus on measuring the deterioration of angular and spatial quality, respectively. Thus, the proposed EDDMF can learn more discriminative features than DeeBLiF.

Second, DeeBLiF only extracts angular features at a single encoding level, which cannot handle the ever-changing angular effect on spatial quality in LFIs. To this end, EDDMF presents a hierarchical discrepancy network consisting of multiple branches for spatio-angular patches. The generated hierarchical discrepancy features with diverse angular encoding levels are highly discriminative on angular quality degradation.

Third, the implicit spatial information in spatio-angular patches restricts the potential improvement of utilizing spatial information in DeeBLiF. In EDDMF, we further propose a local discrepancy extraction module to extract the local discrepancy features from sub-aperture patches, which are more spatially discriminative and serve as complementary features, thereby providing a more comprehensive description for quality assessment.

Finally, a more in-depth analysis of the proposed framework on four benchmark LFI datasets is provided in this paper, including comprehensive analysis of individual distortion types, time complexity, hyperparameter dependency, cross-dataset and cross-type validation. These were not presented in the original paper of DeeBLiF.

III. The Proposed Framework

The overview of the proposed EDDMF is illustrated in Fig. 2, which contains three main components: patch generation module, hierarchical discrepancy network, and local discrepancy extraction module. First, given a distorted LFI and its corresponding reference LFI, both spatio-angular patches and sub-aperture patches of the reference and distorted LFIs are generated by the patch generation module. Then, each spatio-angular patch pair is composed by a spatio-angular patch and its corresponding reference version, while each sub-aperture patch pair consists of a sub-aperture patch and its corresponding reference version. Note that patches with and without red borders in Fig. 2 denote reference patches and distorted patches, respectively. Second, hierarchical discrepancy features are obtained by the hierarchical discrepancy network using spatio-angular patch pairs. Third, sub-aperture patch pairs are exploited to extract local discrepancy features through the local discrepancy extraction module. Finally, hierarchical discrepancy features and local discrepancy features are combined to predict the patch quality, and then the quality of all patches is pooled into an overall quality of the distorted LFI. The main components are detailed as follows.

A. Patch Generation Module

The feasibility of using patches for 2DIQA has been thoroughly demonstrated in many previous studies [33], [34], [35]. The motivation behind is that most distortions for 2D images are homogeneous distortions, $i.e.$, all patches from the same image include similar information variation caused by the same distortion type and level [64]. Motivated by this fact, we observe that most distortions for LFIs are also homogeneous distortions. Besides, the main difference between 2D images and LFIs is the extra angular information. Therefore, we believe that LFI patches are able to reflect the LFI quality as long as they contain enough angular information. To extract patches that are capable of reflecting the LFI quality, a patch generation module is presented, as shown in Fig. 3.

The input SAI array of LFI is denoted as $L \in \mathbb{R}^{U \times V \times H \times W \times C}$, where $U \times V$ is the angular resolution of the SAI array, $H \times W$ is the spatial resolution of each SAI, C is

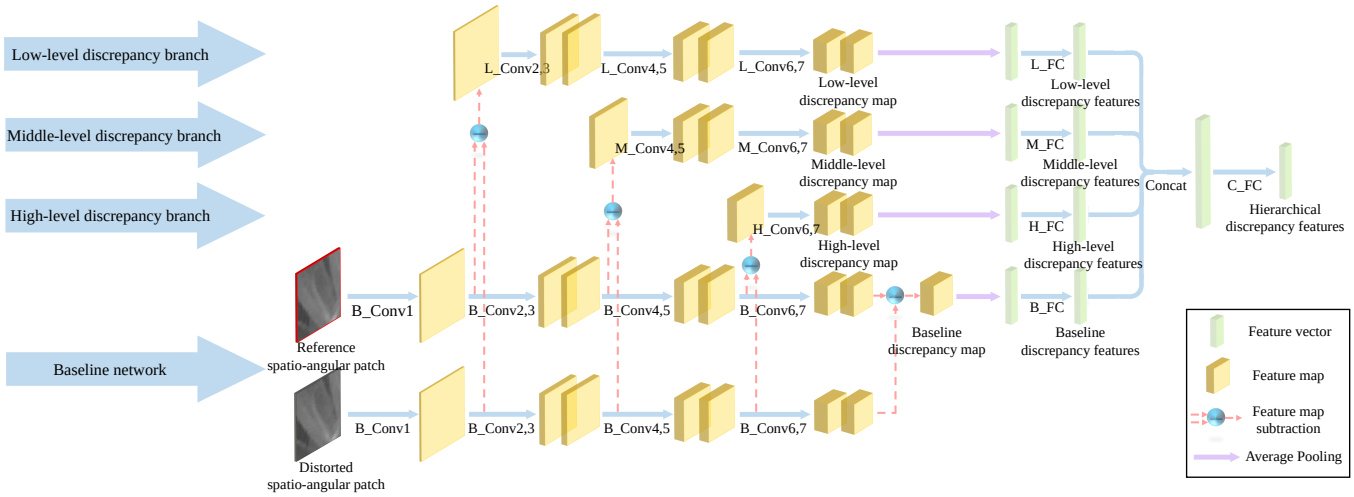


Fig. 4. Architecture of our proposed hierarchical discrepancy network, which consists of a baseline network and three discrepancy branches of different levels.

the channel number. Here, C equals 3 since the original LFI is in the RGB color format. Several previous works [19], [30] have demonstrated that the human eye is more sensitive to the luminance component of an image. Therefore, we convert the input LFI into YCbCr color format and use only the luminance component Y of the central $A\times A$ SAIs for subsequent patch generation, as described in Eq. (1)-(2).

$$L_Y = \{rgb2ycbcr(L)\}_Y \tag{1}$$

$$L_{YC} = Central(L_Y) \tag{2}$$

where $L_Y \in \mathbb{R}^{U\times V\times H\times W}$ and $L_{YC} \in \mathbb{R}^{A\times A\times H\times W}$ denote the luminance component Y of the $U\times V$ SAIs and the central $A\times A$ SAIs of LFI, respectively.

In order to extract both angular and spatial information in LFI, an $A\times A$ sliding window array is defined. Each sliding window is defined as a square of size $S\times S$ to obtain the same amount of spatial information from the vertical and horizontal directions. Then the sliding window array is placed on the $A\times A$ SAIs (*i.e.*, L_{YC}) for evenly cropping and sliding, with a stride of T. For each sliding position, $A\times A$ blocks of size $S\times S$ are obtained and a sub-aperture patch is generated by stitching all blocks. After sliding over L_{YC}, a set of sub-aperture patches $\mathbb{P}_s$ is generated, as described in Eq. (3)-(5).

$$I = \lceil \frac{H-S}{T} \rceil,\ J = \lceil \frac{W-S}{T} \rceil \tag{3}$$

$$P_s^{i,j} = Stitch\{B_{1,1}, ..., B_{Ax,Ay}, ..., B_{A,A}\}^{i,j} \tag{4}$$

$$\mathbb{P}_s = \{P_s^{1,1}, ..., P_s^{i,j}, ..., P_s^{I,J}\} \tag{5}$$

where I and J represent the number of patches generated in the vertical and horizontal directions, respectively. $B_{Ax,Ay} \in \mathbb{R}^{S\times S}$ denotes the Ax-th row and the Ay-th column block cropped from L_{YC}. $P_s^{i,j} \in \mathbb{R}^{A\times A\times S\times S}$ denotes the sub-aperture patch with the i-th vertical index and the j-th horizontal index.

Further, for each sub-aperture patch $P_s^{i,j}$, a permutation operation is applied to exchange its angular coordinates and spatial coordinates to obtain a spatio-angular patch. A toy example of a patch of size $3\times3\times3\times3$ is illustrated in the dashed box in Fig. 3. In this case, the permutation operation can be visualized as grouping all pixels of the same color. After applying the permutation operation to all $P_s^{i,j}$ in $\mathbb{P}_s$, a set of spatio-angular patches $\mathbb{P}_a$ is generated, as described in Eq. (6)-(7).

$$P_a^{i,j} = Permute(P_s^{i,j}) \tag{6}$$

$$\mathbb{P}_a = \{P_a^{1,1}, ..., P_a^{i,j}, ..., P_a^{I,J}\} \tag{7}$$

where $P_a^{i,j} \in \mathbb{R}^{S\times S\times A\times A}$ denotes the spatio-angular patch with the i-th vertical index and the j-th horizontal index.

Finally, $\mathbb{P}_a$ and $\mathbb{P}_s$ are the outputs of the patch generation module. Specifically, $\mathbb{P}_a$ contains $I\times J$ spatio-angular patches while $\mathbb{P}_s$ includes $I\times J$ sub-aperture patches. The data sizes of the original LFI L, the generated $\mathbb{P}_a$ and $\mathbb{P}_s$, and their ratio are calculated in Eq. (8)-(10).

$$D_L = U\times V\times H\times W\times C \tag{8}$$

$$D_{\mathbb{P}_a} = D_{\mathbb{P}_s} = A\times A\times S\times S\times I\times J \tag{9}$$

$$R = \frac{D_{\mathbb{P}_a}}{D_L} = \frac{D_{\mathbb{P}_s}}{D_L} \tag{10}$$

where D_L, $D_{\mathbb{P}_a}$, and $D_{\mathbb{P}_s}$ represent the data sizes of L, $\mathbb{P}_a$, and $\mathbb{P}_s$, respectively. R is the ratio between $D_{\mathbb{P}_a}$ (or $D_{\mathbb{P}_s}$) and D_L. Taking the LFI in Fig. 1 (b) for instance, its angular resolution $U\times V$, spatial resolution $H\times W$, and channel number C are 9×9, 434×625, and 3, respectively. In our implementation, three hyperparameters related to the generation of patches, A, S, and T are set to 5, 32, and 64, respectively. In this case, $R \approx 3\%$, that is, we only use 3% of the original LFI data for quality assessment, which greatly reduces the computational complexity of the whole framework. In addition, since the generated patches consist of angular and spatial information from different locations of the original LFI, they are able to reflect the overall LFI quality.

TABLE I
NETWORK CONFIGURATION OF THE PROPOSED HIERARCHICAL DISCREPANCY NETWORK.

Layer Name	Kernel Size	Channel	Stride	Padding
B_Conv1	3×3	[1, 32]	2	1
	BN, ReLu			
B/L_Conv2	3×3	[32, 64]	2	1
	BN, ReLu			
B/L_Conv3	3×3	[64, 64]	1	1
	BN, ReLu			
$B/L/M\_Conv4$	3×3	[64, 128]	2	1
	BN, ReLu			
$B/L/M\_Conv5$	3×3	[128, 128]	1	1
	BN, ReLu			
$B/L/M/H\_Conv6$	3×3	[128, 256]	2	1
	BN, ReLu			
$B/L/M/H\_Conv7$	3×3	[256, 256]	1	1
	BN, ReLu			
Average Pooling				
$B/L/M/H\_FC$	-	[256, 256]	-	-
$Concat$	-	[4×256, 256]	-	-
C_FC	-	[1024, 128]	-	-

B. Hierarchical Discrepancy Network

The angular information in LFIs essentially distinguishes LFIs from traditional 2D images. Therefore, the key to designing a well-performing LFIQA metric lies in effectively capturing the angular effect on spatial information. Generally, LFI's angular information will be affected to varying degrees when different types and levels of distortion are introduced. For example, JPEG distortion focuses on distorting spatial information, while its effect on angular information is very limited. Conversely, the opposite is true for the interpolation distortion. Thus, we argue that the solidified angular feature extraction fails to accommodate the ever-changing distortion types and levels, thereby resulting in suboptimal performance in quality evaluation.

Given a spatio-angular patch, we notice that shallow convolutional layers capture information between nearest neighbor pixels, $i.e.$, angular features. As the network deepens and the feature map size decreases, information between distant pixels is captured accordingly, $i.e.$, spatial features, which can also be known as spatio-angular features since they are extracted based on the shallow angular features. Inspired by the above observations, we design a hierarchical discrepancy network to extract the hierarchical discrepancy features between reference and distorted spatio-angular patches, in which different branches extract spatio-angular discrepancy features with different angular encoding levels.

Fig. 4 illustrates the architecture of our proposed hierarchical discrepancy network. The proposed network consists of a baseline network and three discrepancy branches, which generate the baseline, low-level, middle-level, and high-level discrepancy maps, respectively. Then each discrepancy map is converted into a feature vector using an average pooling layer, followed by a Fully Connected (FC) layer (denoted as $B/L/M/H\_FC$, respectively). Finally, a FC layer, denoted as C_FC, is applied after concatenating the four feature vectors to generate the final hierarchical discrepancy features. The baseline network contains two sub-networks for extracting features from reference and distorted spatio-angular patches, respectively. Each sub-network includes seven convolutional layers, $i.e.$, $B\_Conv1 - B\_Conv7$. Note that the two sub-networks share the same weights in all convolutional layers. At the end of the baseline network, the feature maps extracted by the two sub-networks are subtracted, $i.e.$, the baseline discrepancy map between reference and distorted spatio-angular patches are obtained. Additionally, three discrepancy branches are exploited to extract the low-level, middle-level, and high-level discrepancy maps, respectively. The three branches start from the subtraction of reference and distorted feature maps generated by B_Conv1, B_Conv3, and B_Conv5, respectively, and have the same structure as the rest of the baseline network.

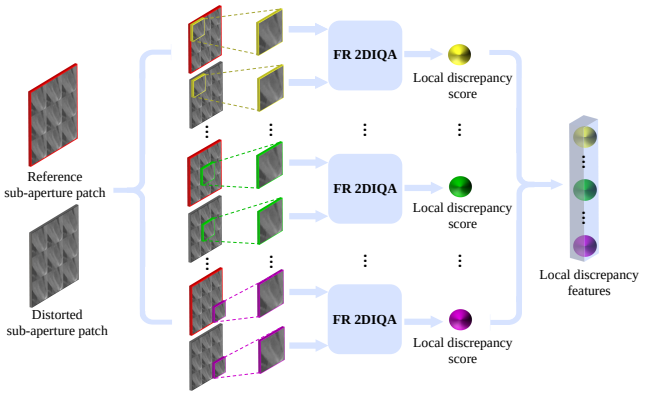


Fig. 5. Pipeline of the proposed local discrepancy extraction module. For better visualization, A is set to 3 for all illustrations with $A\times A$ angular resolution.

As shown in TABLE I, all convolution layers use 3×3 kernel size with 1 padding and each convolutional layer is followed by a Batch Normalization (BN) layer and a ReLu activation function. As the number of convolutional layers increases, the output channel increases while the output feature map size decreases. Specifically, we halve the feature map size by setting the stride to 2 in B_Conv1, B/L_Conv2, $B/L/M\_Conv4$, and $B/L/M/H\_Conv6$, respectively. Therefore, the final feature map size is $1/16$ of the original input patch. Taking a sub-network in the baseline network for example, with an input patch of size $160\times160\times1$, the size and channel of the feature map are changed as follows: $160\times160\times1 - 80\times80\times32 - 40\times40\times64 - 40\times40\times64 - 20\times20\times128 - 20\times20\times128 - 10\times10\times256 - 10\times10\times256$. After applying the average pooling layer, the FC structure $4\times256 - 4\times256 - 1024 - 128$ is adopted to generate hierarchical discrepancy features. Finally, the final output of our proposed hierarchical discrepancy network is 128 dimensions, consisting of spatio-angular discrepancy features with different angular encoding levels.

TABLE II
SUMMARY OF FOUR LFI DATASETS USED IN OUR EXPERIMENTS.

Datasets	Reference LFIs	Distortion types	Distortion levels	Distorted LFIs	Angular resolution	Spatial resolution	Mode	Protocol	MOS	Year
Win5-LID [6]	6 (real-world), 4 (synthesis)	HEVC, JPEG2000, LN, NN, [69], [70]	5 or 1	220	9×9	434×625 (real-world), 512×512 (synthesis)	Interactive	DSCQS	[1,5]	2018
NBU-LF1.0 [66]	8 (real-world), 6 (synthesis)	NN, BI, EPICNN [69], Zhang [71], VDSR [72]	3	210	9×9	434×625 (real-world), 512×512 (synthesis)	Interactive and passive	DSCQS	[1,5]	2019
LFDD [67]	8 (synthesis)	JPEG, JPEG2000, BPG, VP9, AV1, AVC, HEVC, Gaussian noise, Impulse noise, Barrel, Pincushion, Unsharp mask	5	480	9×9	512×512 (synthesis)	Passive	DSIS	[1,5]	2020
SHU [68]	8 (real-world)	JPEG, JPEG2000, Gaussian noise, Motion blur, White noise	6	240	15×15	434×625 (real-world)	Passive	DSCQS	[0,5]	2019

C. Local Discrepancy Extraction Module

As mentioned before, hierarchical discrepancy features are more angularly discriminative due to the hierarchical design based on spatio-angular patches. As a result, the spatial features used for quality assessment are insufficient to some extent. Therefore, we aim to extract local discrepancy features, which focus more on spatial quality degradation and serve as complementary features to hierarchical discrepancy features. Specifically, we propose a local discrepancy extraction module to generate the local discrepancy features between reference and distorted sub-aperture patches, which provide a more comprehensive description of the patch quality degradation.

The pipeline of the proposed local discrepancy extraction module is illustrated in Fig. 5. Given a distorted sub-aperture patch $P_s \in \mathbb{R}^{A\times A\times S\times S}$ and its reference version $\hat{P}_s \in \mathbb{R}^{A\times A\times S\times S}$, they are both composed by $A\times A$ blocks of size $S\times S$, as described in Eq. (4). Since each block contains local information derived from a single SAI, we measure the local discrepancy between reference and distorted blocks to evaluate the impact of spatial quality degradation for each SAI. Specifically, for each distorted block and its reference version, we measure their local discrepancy score by adopting a handcrafted feature-based FR 2DIQA metric instead of the CNN model. The reason behind is that existing handcrafted feature-based FR 2DIQA metrics can deal well with the degradation of spatial information in 2D images. Finally, a total of $A\times A$ local discrepancy scores are generated and further concatenated into a feature vector, named local discrepancy features. In our implementation, we use the representative Structural Similarity Index Metric (SSIM) [65] as the FR 2DIQA metric due to its low computational complexity and excellent performance. The above process can be described in Eq. (11)-(12):

$$S_i = SSIM(\hat{B}_i, B_i) \tag{11}$$

$$F_{local} = Concat(S_1, S_2, ..., S_{A\times A}) \tag{12}$$

where $\hat{B}_i$ and B_i denote the i-th reference and distorted blocks of the sub-aperture patch P_s, respectively, S_i is the local discrepancy score between $\hat{B}_i$ and B_i, F_{local} is the local discrepancy features with $A\times A$ dimensions.

Since the same distortion may have different impacts on different SAIs and each local discrepancy score can only measure the spatial degradation of a single SAI, the angular degradation is neglected. To this end, we combine all the local discrepancy scores to estimate the angular degradation of sub-aperture patches. Although angular information is considered, local discrepancy features are more spatially discriminative and are thus used as complementary features to the angular-dominant hierarchical discrepancy features.

D. Training

In the training stage, we employ mini-batch Stochastic Gradient Descent (SGD) as the optimizer, where the weight momentum and weight decay are set to 0.9 and 0.0001, respectively. Due to the adoption of FC layers in the proposed framework, the size of input patches needs to be fixed. Specifically, we set the central angular resolution of SAIs, the size of each sliding window, and the stride of the sliding window array to 5×5, 32×32, and 64, respectively. Thus, we train our network with a large number of patches on a TITAN Xp GPU. All the patches extracted from the same distorted LFI use the Mean Opinion Score (MOS) of the whole distorted LFI as their training Ground Truth (GT). Although setting a smaller batch size can slightly improve the final performance possibly due to more frequent parameter updates, it also leads to a significant increase in training time. Therefore, the batch size is empirically fixed to 32 for a reasonable trade-off. Following [37], the training set data is trained for 70 epochs with an initial learning rate of 0.001, and the learning rate is divided by 10 at epoch 30 and 60. The proposed framework is trained from scratch with Xavier normal distribution initialization. Note that no data augmentation operation is used in the data preparation stage. Suppose that $(\hat{P}_a^b, P_a^b)$ and $(\hat{P}_s^b, P_s^b)$ denote the b-th spatio-angular patch pair and sub-aperture patch pair in a batch, respectively, and G^b is the corresponding GT. We employ the widely-used L2 Loss to measure the distance between the predicted quality and GT. The learning objective of our proposed framework is to minimize the loss through backpropagation, as described in Eq. (13)-(14).

$$L = \frac{1}{B}\sum_{b=1}^{B}(f(\hat{P}_a^b, P_a^b, \hat{P}_s^b, P_s^b; w) - G^b)^2 \tag{13}$$

$$w' = \min_{w}(L) \tag{14}$$

where B denotes the batch size, $f(\hat{P}_a^b, P_a^b, \hat{P}_s^b, P_s^b; w)$ is the predicted patch quality using network weights w with the input of $(\hat{P}_a^b, P_a^b)$ and $(\hat{P}_s^b, P_s^b)$, w' denotes the updated network weights.

TABLE III
OVERALL PERFORMANCE COMPARISON ON THE WIN5-LID, NBU-LF1.0, LFDD, AND SHU DATASETS. "OVERALL" DENOTES THE WEIGHTED-AVERAGE RESULTS OVER ALL DATASETS, WHERE WEIGHTS ARE PROPORTIONAL TO THE SIZE OF EACH DATASET. THE BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Metric Types	Metrics	Win5-LID			NBU-LF1.0			LFDD			SHU			Overall		
		PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
NR 2DIQA	PIQE [14]	0.4820	0.3920	0.9220	0.2561	0.1779	1.0000	0.4201	0.3746	1.0103	0.7780	0.7996	0.6836	0.4767	0.4307	0.9233
	NIQE [15]	0.6246	0.4482	0.7584	0.4792	0.3701	0.7948	0.5642	0.4449	0.9405	0.9187	0.8920	0.4247	0.6342	0.5252	0.7714
	BRISQUE [16]	0.6263	0.4559	0.7530	0.4969	0.3750	0.7910	0.5103	0.4179	0.9597	0.9012	0.8747	0.4614	0.6116	0.5127	0.7854
FR 2DIQA	NQM [17]	0.6509	0.5908	0.9082	0.6479	0.6614	0.6995	0.4461	0.3810	0.9989	0.8581	0.8618	0.5619	0.6081	0.5727	0.8357
	IFC [18]	0.4945	0.5429	1.0628	0.7440	0.6843	0.6049	0.3683	0.4226	1.0198	0.8853	0.9068	0.4969	0.5689	0.5945	0.8431
	HDR-VDR-2 [19]	0.7315	0.7169	0.6643	0.7827	0.7924	0.6212	0.3986	0.5306	1.0368	0.8444	0.8736	0.5861	0.6255	0.6856	0.7956
NR 3DIQA	SINQ [21]	0.6051	0.5075	0.7410	0.5276	0.4374	0.7633	0.5892	0.4952	0.8959	0.9189	0.8955	0.4209	0.6498	0.5705	0.7429
FR 3DIQA	Chen's [22]	0.8527	0.8278	0.5083	0.8357	0.8048	0.4902	0.4507	0.5051	0.9580	0.8916	0.8960	0.4947	0.6899	0.7031	0.6899
NR multiview-IQA	NIQSV+ [24]	0.3030	0.2174	0.9159	0.3512	0.2339	0.8354	0.3264	0.1944	1.0479	0.4362	0.0757	0.9390	0.3494	0.1812	0.9611
FR multiview-IQA	MW-PSNR [25]	0.3647	0.5436	0.9094	0.3371	0.4730	0.8569	0.6189	0.6095	1.0872	0.8514	0.8763	0.6355	0.5673	0.6276	0.9169
	MW-PSNRr [26]	0.3649	0.5573	0.9093	0.3351	0.4452	0.8576	0.4047	0.6472	1.0237	0.8400	0.8735	0.5885	0.4752	0.6403	0.8807
	MP-PSNR [27]	0.3645	0.5606	0.9094	0.3333	0.4216	0.8581	0.2344	0.6529	1.0540	0.8364	0.8844	0.5986	0.4030	0.6413	0.8955
	MP-PSNRr [26]	0.3648	0.5602	0.9093	0.3334	0.4426	0.8581	0.4321	0.6451	1.0320	0.8139	0.8458	0.7858	0.4809	0.6338	0.9254
NR LFIQA	BELIF [49]	0.5912	0.5119	0.7572	0.7161	0.6892	0.6291	0.7747	0.7072	0.7103	0.8976	0.8671	0.4784	0.7545	0.6999	0.6560
	VBLFI [51]	0.6844	0.6116	0.7041	0.8179	0.7660	0.5027	0.6928	0.6245	0.8238	0.9220	0.8992	0.4100	0.7619	0.7052	0.6559
	NR-LFQA [50]	0.6952	0.6275	0.6750	0.8327	0.8036	0.4895	0.6647	0.5864	0.8374	0.9390	0.9347	0.3729	0.7585	0.7066	0.6459
	Tensor-NLFQ [30]	0.7595	0.7345	0.6327	0.7624	0.7261	0.5856	0.8446	0.7887	0.6025	0.8649	0.8630	0.5424	0.8175	0.7824	0.5926
	4D-DCT-LFIQA [55]	0.8267	0.8079	0.5512	0.8381	0.8213	0.4906	0.8206	0.7699	0.6411	0.9400	0.9320	0.3691	0.8499	0.8204	0.5397
	DeeBLiF [37]	0.8427	0.8186	0.5160	0.8583	0.8229	0.4588	0.8827	0.8086	0.5267	**0.9548**	**0.9419**	**0.3185**	0.8856	0.8409	0.4688
RR LFIQA	LF-IQM [56]	0.3620	0.3438	0.8930	0.4260	0.2700	0.7966	0.4457	0.4287	0.9950	0.3601	0.2960	1.0046	0.4082	0.3558	0.9413
FR LFIQA	MDFM [58]	0.7303	0.6768	0.6625	0.8444	0.8138	0.4749	0.5725	0.5282	0.9276	0.8275	0.8543	0.6149	0.7056	0.6768	0.7290
	Min's [59]	0.7350	0.6645	0.6794	0.7112	0.6577	0.6476	0.5596	0.4094	0.9366	0.8496	0.8470	0.5745	0.6814	0.5949	0.7591
	Meng's [31]	0.6924	0.6347	0.7001	0.8367	0.7819	0.4944	0.3043	0.3493	1.0593	0.9282	0.9203	0.4037	0.6060	0.6021	0.7506
	EDDMF (ours)	**0.8654**	**0.8354**	**0.4839**	**0.9024**	**0.8743**	**0.3866**	**0.9406**	**0.9077**	**0.3794**	0.9443	0.9206	0.3575	**0.9200**	**0.8905**	**0.3961**

E. Patch Quality Pooling

Based on the assumption that the quality of the whole LFI can be reflected in patches, we indirectly evaluate the quality degradation of the distorted LFI by measuring the quality degradation in each patch. Therefore, in the test stage, the output of our proposed EDDMF is the patch quality and a pooling method is required to convert the quality of all patches into an overall quality of the distorted LFI. Since patches from the same LFI are trained with the same GT, we consider all patches in the same LFI to be equally important and use the average pooling to generate an overall quality of the distorted LFI, as shown in Eq. (15).

$$Q = \frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} f(\hat{P}_a^{i,j}, P_a^{i,j}, \hat{P}_s^{i,j}, P_s^{i,j}; w) \qquad (15)$$

where Q denotes the overall quality of the distorted LFI, I and J are the number of patches generated in the vertical and horizontal and directions, respectively, as calculated in Eq. (3). $f(\hat{P}_a^{i,j}, P_a^{i,j}, \hat{P}_s^{i,j}, P_s^{i,j}; w)$ represents the predicted patch quality with the input of a spatio-angular patch pair $(\hat{P}_a^{i,j}, P_a^{i,j})$ and a sub-aperture patch pair $(\hat{P}_s^{i,j}, P_s^{i,j})$.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Datasets, Experimental Settings, and Evaluation Criteria

To validate our proposed EDDMF, we conduct extensive experiments on four publicly available LFI datasets: Win5-LID [6], NBU-LF1.0 [66], LFDD [67], and SHU [68]. TABLE II provides a detailed summary of these four LFI datasets used in our experiments.

The Win5-LID dataset includes 220 distorted LFIs derived from 6 real-world reference LFIs and 4 synthetic reference LFIs, which are subject to 6 types of distortions with different distortion levels. Specifically, these distortion types comprise HEVC, JPEG2000, Linear interpolation (LN), Nearest Neighbor interpolation (NN), and two CNN distortions [69], [70]. Except the two CNN distortions, each of the other four distortions contains 5 distortion levels. The subjective experiment of Win5-LID dataset adopts Double-Stimulus Continuous Quality Scale (DSCQS) protocol and interactive mode, which provides the MOS from 1 (very annoying) to 5 (imperceptible).

The NBU-LF1.0 dataset consists of 8 real-world reference LFIs and 6 synthetic reference LFIs. Each reference LFI is processed by 5 reconstruction distortions: NN, Bicubic Interpolation (BI), learning based reconstruction (EPICNN) [69], disparity map based reconstruction (Zhang) [71], and spatial super-resolution reconstruction (VDSR) [72]. Each distortion has 3 levels. Therefore, the NBU-LF1.0 dataset contains 210 distorted LFIs. The dataset adopts passive and interactive mode and DSCQS protocol to conduct the subjective experiment. The MOS on a 5-point discrete scale is provided.

The LFDD dataset has 8 synthetic reference LFIs and 480 distorted LFIs. A total of 12 common distortion types are involved: JPEG, JPEG2000, BPG, VP9, AV1, AVC, HEVC,

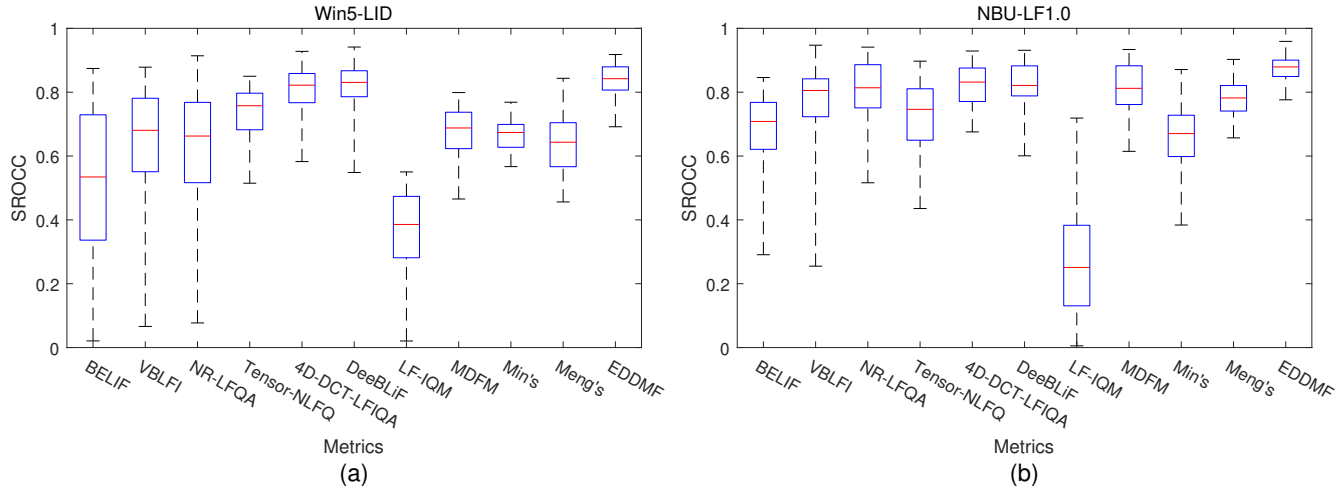


Fig. 6. Box-plots of SROCC distribution of LFIQA metrics. (a) Win5-LID dataset; (b) NBU-LF1.0 dataset.

Gaussian noise, Impulse noise, Barrel, Pincushion, and Unsharp mask. Each distortion type has 5 distortion levels. The passive mode and Double-Stimulus Impairment Scale (DSIS) protocol are adopted for the subjective experiment. The dataset provides the MOS ranged from 1 to 5.

The SHU dataset includes 240 distorted LFIs based on 8 real-world reference LFIs. The dataset includes 5 distortion types: JPEG, JPEG2000, Gaussian blur, Motion blur, and White noise. Each type of distortion has 6 distortion levels. The passive experimental mode and DSCQS protocol are employed for the subjective experiment. The dataset provides the MOS ranged from 0 (bad) to 5 (excellent).

In our experiments, we adopt leave-two-fold-out cross-validation as the train-test split strategy to report the performance. Specifically, for each dataset, we first divide all distorted LFIs into K folds according to their corresponding reference LFIs. Each fold contains all distorted versions of the same reference LFI. Then we use K-2 folds for training and the remaining 2 folds for testing, which ensures that the training and test sets are independent of each other. After going through all train-test partitions, there are $K(K\text{-}1)/2$ combinations in total. Therefore, we conduct experiments based on all combinations and report the average result as the final performance.

In addition, three standard criteria are employed to evaluate the performance of all metrics, including Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Square Error (RMSE). Here, PLCC measures the linear relationship, SROCC focuses on the monotonicity, and RMSE evaluates the predictive accuracy. Higher PLCC and SROCC values represent better performance, while it is opposite for RMSE. As recommended in [73], a five-parameter nonlinear function is adopted for score-mapping process before computing PLCC and RMSE, as shown in Eq. (16).

$$f(p) = \beta_1\left(\frac{1}{2} - \frac{1}{1+e^{\beta_2(p-\beta_3)}}\right) + \beta_4 p + \beta_5 \qquad (16)$$

where p and $f(p)$ denote the prediction and its nonlinear mapping result, respectively, and $\beta_{1\ldots5}$ are optimized to minimize the error between p and its corresponding MOS.

B. Overall Performance Comparison

In this subsection, we conduct comparative experiments to demonstrate the effectiveness of our proposed framework. The proposed framework is compared with twenty-three state-of-the-art IQA metrics, including nine types: three NR 2DIQA metrics (PIQE [14], NIQE [15], BRISQUE [16]), three FR 2DIQA metrics (NQM [17], IFC [18], HDR-VDR-2 [19]), one NR 3DIQA metric (SINQ [21]), one FR 3DIQA metric (Chen's [22]), one NR multiview-IQA metric (NIQSV+ [24]), four FR multiview-IQA metrics (MW-PSNR [25], MW-PSNRr [26], MP-PSNR [27], MP-PSNRr [26]), six NR LFIQA metrics (BELIF [49], VBLFI [51], NR-LFQA [50], Tensor-NLFQ [30], 4D-DCT-LFIQA [55], DeeBLiF [37]), one RR LFIQA metric (LF-IQM [56]) and three FR LFIQA metrics (MDFM [58], Min's [59], Meng's [31]). In our experiments, leave-two-fold-out cross-validation is employed for all metrics to report their performance. For those metrics that can directly predict scores, we report their performance on the same test set as the learning-based metrics for fair comparison. In addition, since 2DIQA metrics and multiview-IQA metrics are designed for 2D images and multiview images, respectively, while 3DIQA metrics are designed for a pair of 2D images, we provide different experimental methods for different types of metrics. For 2DIQA metrics and multiview-IQA metrics, we perform them on each SAI and take the average results of all SAIs as the final performance. For 3DIQA metrics, each adjacent two SAIs in the horizontal direction are regarded as left and right views of the input, and the average performance is reported. The performance of all metrics is reproduced using the features or released codes from their authors.

TABLE III shows the overall experimental results on the Win5-LID, NBU-LF1.0, LFDD, and SHU datasets. We can find that compared to other IQA metrics, our proposed framework achieves superior performance on the Win5-LID, NBU-LF1.0, and LFDD datasets, and yields competitive performance on the SHU dataset. This may be because 2DIQA metrics mainly focus on the quality degradation on 2D images,

TABLE IV
PLCC PERFORMANCE OF DIFFERENT DISTORTION TYPES ON THE WIN5-LID AND NBU-LF1.0 DATASETS. IN EACH CASE, THE BEST TWO RESULTS ARE MARKED IN **BOLD**, AND HIT-COUNT TALLIES THE NUMBER OF TIMES EACH IQA METRIC OBTAINING A TOP-TWO RESULT.

Cases	Metric Types	Metrics	Win5-LID				NBU-LF1.0					Hit-count
			HEVC	JPEG2000	LN	NN	NN	BI	EPICNN	Zhang	VDSR	
Case 1	NR LFIQA	BELIF [49]	0.8062	0.7275	0.7172	0.7219	0.9244	0.8732	0.6707	0.6886	0.8749	0
		VBLFI [51]	0.8037	0.8273	0.8151	0.7261	0.9056	0.9276	0.8729	0.7072	0.9526	0
		NR-LFQA [50]	0.7641	0.8098	0.7731	0.7920	0.9544	0.9519	0.9157	0.7108	0.8850	0
		Tensor-NLFQ [30]	0.8909	0.8340	0.8543	0.8446	0.8517	0.9199	0.8395	0.7135	0.9223	0
		4D-DCT-LFIQA [55]	0.9001	**0.9365**	0.8803	0.8534	0.9386	0.9389	0.8183	0.9048	0.9317	1
		DeeBLiF [37]	0.9389	0.9254	**0.9021**	**0.9207**	**0.9610**	0.9499	0.9395	0.6659	0.9487	3
	RR LFIQA	LF-IQM [56]	0.6172	0.7263	0.8284	0.7158	0.5026	0.6764	0.5976	0.5296	0.5517	0
	FR LFIQA	MDFM [58]	0.9446	0.8878	0.8819	0.9004	0.9327	0.9463	0.9483	**0.9436**	0.9226	1
		Min's [59]	**0.9777**	**0.9465**	0.8798	0.7775	0.8033	0.8796	0.7890	0.8466	**0.9664**	3
		Meng's [31]	0.9129	0.7442	**0.9486**	0.8915	0.8989	**0.9550**	**0.9734**	0.7308	**0.9652**	**4**
		EDDMF (ours)	**0.9662**	0.9259	0.8747	**0.9073**	**0.9699**	**0.9660**	**0.9544**	**0.9204**	0.9534	**6**
Case 2	NR LFIQA	BELIF [49]	0.9669	0.9428	**0.9647**	0.9242	0.8554	0.7225	0.5406	0.6737	0.8566	1
		VBLFI [51]	0.9762	0.9754	0.8676	0.9315	0.9366	**0.9389**	0.8938	0.6749	0.9172	1
		NR-LFQA [50]	**0.9791**	**0.9757**	0.9520	**0.9577**	0.9414	0.9324	**0.9224**	0.7718	0.8773	**4**
		Tensor-NLFQ [30]	0.9409	0.9424	0.8910	0.8481	0.9149	0.8449	0.8064	0.8007	0.9143	0
		4D-DCT-LFIQA [55]	0.9261	0.9603	0.9547	0.9094	0.9405	0.8708	0.7972	**0.8264**	0.8849	1
		DeeBLiF [37]	0.9474	0.9179	0.9320	**0.9616**	**0.9591**	0.9175	0.8952	0.6104	**0.9353**	3
	RR LFIQA	LF-IQM [56]	0.5898	0.7151	0.8332	0.4082	0.3198	0.4834	0.4977	0.4783	0.3826	0
	FR LFIQA	MDFM [58]	0.9367	0.8350	0.8843	0.8198	0.8804	0.8683	0.8874	**0.8244**	0.8627	1
		Min's [59]	0.9652	0.9379	0.8944	0.7965	0.7636	0.8005	0.7308	0.7303	0.9051	0
		Meng's [31]	0.8945	0.6740	**0.9643**	0.8897	0.8347	0.8778	**0.9449**	0.4086	0.9068	2
		EDDMF (ours)	**0.9805**	**0.9795**	0.9078	0.9283	**0.9477**	**0.9557**	0.9136	0.7872	**0.9698**	**5**

but they fail to evaluate the angular consistency between different SAIs in LFIs. Besides, since multiview-IQA metrics are designed to measure view-synthesis distortions such as flickering, stretching, and black holes, they are insensitive to the quality degradation caused by reconstruction and compression methods in LFIs. Further, 3DIQA metrics are proposed to measure the quality of stereoscopic images with wide parallax, they do not perform well on LFIs with narrow parallax. Among the above three types of IQA metrics, it is reasonable that the FR metrics have better performance than the NR metrics due to the available reference information. LF-IQM metric [56] is a RR LFIQA metric whose performance relies heavily on the accuracy of depth map estimation. Although some existing NR or FR LFIQA metrics achieve competitive performance on a certain criterion or a certain dataset, they fail to perform consistently well on all datasets. However, our proposed EDDMF is not only designed based on the characteristics of LFIs, but also evaluates the LFI quality by jointly using hierarchical discrepancy features and local discrepancy features, which obtains more consistent results with HVS on most LFI datasets.

In addition, we provide the box-plots of SROCC distribution of LFIQA metrics on the Win5-LID and NBU-LF1.0 datasets in Fig. 6. Each box represents the SROCC distribution of one metric, and the red line denotes the median performance over all train-test splits. The top and bottom of the blue box denote the lower and upper quartiles, respectively. The top and bottom of the dashed line are the maximum and minimum, respectively. Generally, the blue box with higher red line and position indicates better performance, while the blue box with smaller height denotes better stability. As shown in Fig. 6, we can see that our proposed framework achieves the best performance with strong stability compared to the state-of-the-arts.

C. Robustness Against Different Distortion Types

An excellent IQA metric should show strong robustness with respect to different distortion types. In order to investigate the effectiveness of the proposed method under different experimental conditions, we conduct the experiments on individual distortion types in two different cases:

Case 1: Leave-two-fold-out cross-validation. We use the train-test split strategy that reports overall performance, but each test fold includes only one distortion type. Then the average result of $K(K\text{-}1)/2$ iterations is reported as the performance for each distortion type. In this case, each test set contains all distortion levels of one distortion type for two reference LFIs, which mainly investigates the ability for discriminating different levels of the same LFI.

Case 2: Randomly selection. For each distortion type, we randomly select a distorted LFI from each reference LFI to construct the test set. Then we repeat this 100 times and report the average result as the performance for each distortion type. In this case, each test set includes one random selected

TABLE V
RESULTS OF THE TEST TIME AGAINST OVERALL PLCC PERFORMANCE. THE BEST AND SECOND-BEST RESULTS ARE MARKED IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Metric Types	Metrics	Test Time (s)	PLCC
NR 2DIQA	PIQE [14]	5.1466	0.4767
	NIQE [15]	6.0077	0.6342
	BRISQUE [16]	3.1543	0.6116
FR 2DIQA	NQM [17]	10.1941	0.6081
	IFC [18]	36.3739	0.5689
	HDR-VDR-2 [19]	89.2621	0.6255
NR 3DIQA	SINQ [21]	187.7167	0.6498
FR 3DIQA	Chen's [22]	472.7339	0.6899
NR multiview-IQA	NIQSV+ [24]	2.5747	0.3494
FR multiview-IQA	MW-PSNR [25]	1.0127	0.5673
	MW-PSNRr [26]	1.0039	0.4752
	MP-PSNR [27]	18.4434	0.4030
	MP-PSNRr [26]	10.1670	0.4809
NR LFIQA	BELIF [49]	107.8814	0.7545
	VBLFI [51]	65.6667	0.7619
	NR-LFQA [50]	225.2069	0.7585
	Tensor-NLFQ [30]	697.6515	0.8175
	4D-DCT-LFIQA [55]	169.2623	0.8499
	DeeBLiF [37]	4.8533	0.8856
RR LFIQA	LF-IQM [56]	589.7851	0.4082
FR LFIQA	MDFM [58]	**0.8537**	0.7056
	Min's [59]	3.9845	0.6814
	Meng's [31]	30.4872	0.6060
	EDDMF (ours)	1.8775	**0.9200**

TABLE VI
PERFORMANCE DEPENDENCY OF DIFFERENT HYPERPARAMETERS IN PATCH GENERATION MODULE ON THE WIN5-LID DATASET. HERE, $A\times A$ DENOTES THE CENTRAL ANGULAR RESOLUTION, $S\times S$ DENOTES THE SIZE OF EACH SLIDING WINDOW, T DENOTES THE STRIDE OF THE SLIDING WINDOW ARRAY.

Hyperparameters	Values	PLCC	SROCC	RMSE
$A\times A$	3×3	0.8428	0.8155	0.5178
	5×5	**0.8654**	**0.8354**	**0.4839**
	7×7	0.8635	0.8351	0.4860
	9×9	0.8629	0.8349	0.4891
$S\times S$	8×8	0.8506	0.8287	0.5102
	16×16	0.8589	0.8305	0.4946
	32×32	**0.8654**	**0.8354**	**0.4839**
	64×64	0.8630	0.8232	0.5022
T	32	**0.8668**	**0.8401**	**0.4819**
	48	0.8625	0.8365	0.4883
	64	0.8654	0.8354	0.4839
	80	0.8517	0.8225	0.5046
	96	0.8576	0.8298	0.4942

TABLE VII
PERFORMANCE OF CROSS-DATASET VALIDATION. THE MODEL IS TRAINED AND TESTED USING THE OVERLAPPING DISTORTION TYPES OF TWO LFI DATASETS.

Distortion Types	Training Datasets	Test Datasets	PLCC	SROCC	RMSE
NN	Win5-LID	NBU-LF1.0	0.9724	0.9626	0.2613
	NBU-LF1.0	Win5-LID	0.8230	0.8051	0.5709
JPEG2000	Win5-LID	SHU	0.8375	0.8334	0.4426
	SHU	Win5-LID	0.8624	0.7504	0.5044

distortion level of one distortion type for all reference LFIs, which focuses on testing the robustness to different scenarios.

TABLE IV shows the experimental results against distortion types in two cases on the Win5-LID and NBU-LF1.0 datasets. Due to the space constraint, we only show the PLCC performance of LFIQA metrics. In the table, hit-count tallies the number of times each IQA metric obtaining a top-two result. It can be found that the FR metrics perform better than the NR metrics in $Case$ 1, while it is the opposite in $Case$ 2. One possible reason is that different quality degradation levels of the same LFI are easier to be distinguished when the reference LFI is available, while the NR metrics are more suitable for discriminating the quality of different scenes. However, the proposed EDDMF outperforms state-of-the-art metrics for most distortion types in both cases, which fully demonstrates its robustness and effectiveness.

D. Time Complexity

Time complexity is one of the most important factors in practical applications. TABLE V exhibits the test time against overall PLCC performance among all metrics. Here, test time denotes the runtime for testing a single distorted LFI. For fair comparison, all the metrics are tested with the same CPU-only hardware configuration. We only report the runtime of each metric, excluding the time required for data loading and model initialization. As shown in TABLE V, it is not surprising that our proposed framework has much lower time complexity compared to most existing IQA metrics, because our framework is executed based on the extracted LFI patches which greatly reduce the computational complexity. Further, with such a low computational cost, our proposed framework still achieves the best performance among all state-of-the-art metrics.

E. Hyperparameter Dependency

Since there are several manual setting hyperparameters for generating spatio-angular patches and sub-aperture patches in the patch generation module, it is necessary to study the performance dependency of the proposed framework with respect to different hyperparameters. TABLE VI exhibits the performance dependency of three hyperparameters on the Win5-LID dataset, including $A\times A$, the central angular resolution; $S\times S$, the size of each sliding window; and T, the stride of the sliding window array. Note that the experimental results of each hyperparameter are reported when other hyperparameters are fixed. For hyperparameter A, similar performance can be

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

Test Distortion Types	Training Datasets	PLCC	SROCC	RMSE
NN	Win5-LID	0.8816	0.8147	0.3169
	NBU-LF1.0	0.9699	0.9144	0.1765
BI	Win5-LID	0.9522	0.8986	0.2597
	NBU-LF1.0	0.9660	0.9351	0.2416
EPICNN	Win5-LID	0.9581	0.7903	0.1899
	NBU-LF1.0	0.9544	0.8061	0.2133
Zhang	Win5-LID	0.7802	0.6246	0.3119
	NBU-LF1.0	0.9204	0.8188	0.1952
VDSR	Win5-LID	0.9316	0.8854	0.3005
	NBU-LF1.0	0.9534	0.9168	0.2677
All	Win5-LID	0.7782	0.7467	0.5627
	NBU-LF1.0	0.9024	0.8743	0.3866

TABLE IX
PERFORMANCE OF CROSS-TYPE VALIDATION.

Datasets	Training Types	Test Types	PLCC	SROCC	RMSE
Win5-LID	Real-world	Synthetic	0.7605	0.7436	0.6495
	Synthetic	Real-world	0.7986	0.7131	0.5875
NBU-LF1.0	Real-world	Synthetic	0.8485	0.8343	0.4799
	Synthetic	Real-world	0.8594	0.8459	0.4792

obtained when $A\times A$ is set to 5×5 or larger, while a smaller value results in suboptimal performance, it is possibly because there is not enough angular information to measure the LFI quality. For hyperparameter S, adopting a moderate value, $i.e.$, $S\times S$ is set to 32×32, achieves the best performance. One possible explanation is that as the S value increases, a single patch contains more spatial information while the total number of generated patches decreases. Therefore, adopting a moderate value of S maintains a reasonable trade-off and achieves the best results. For hyperparameter T, the smaller the value set, the more LFI patches can be obtained. Thus, it can be found that a smaller value achieves better performance since more samples are used for training. However, in order to keep a better trade-off between computational complexity and quality evaluation performance, we adopt a moderate value of 64 in our framework.

F. Cross-dataset and Cross-type Validation

In this subsection, we investigate the generalization performance of the proposed framework from two aspects: cross-dataset validation and cross-type validation.

From TABLE II we can see that both Win5-LID and NBU-LF1.0 datasets include the NN distortion, and both Win5-LID and SHU datasets include the JPEG2000 distortion. Thus, we train the framework with one distortion type on one dataset and test it with the same distortion type on another dataset for cross-dataset validation. As reported in TABLE VII, it can be found that our framework still achieves good performance even when trained on other datasets. In addition, since the SHU dataset contains only real-world LFIs while the NBU-LF1.0 dataset contains only 3 levels for each distortion type, the Win5-LID dataset is much more complex than the other two datasets. Thus, it is reasonable to find that training on the Win5-LID dataset achieves better performance than testing on the Win5-LID dataset for both NN and JPEG2000 distortions, which further demonstrates the excellent cross-dataset robustness of our framework.

Additionally, we perform the cross-dataset experiment by training on the whole Win5-LID dataset and testing on the NBU-LF1.0 dataset. To better investigate the cross-dataset performance, we compare the performance of training on the Win5-LID and NBU-LF1.0 datasets respectively, and testing on the NBU-LF1.0 dataset, as shown in TABLE VIII. It can be observed that for most individual distortion types, the model training on the Win5-LID dataset still achieves competitive performance compared to that training on the NBU-LF1.0 dataset. Besides, even when testing on the whole NBU-LF1.0 dataset, the proposed EDDMF still achieves competitive results, which fully demonstrates the robustness and effectiveness in terms of quality evaluation.

For cross-type validation, since both Win5-LID and NBU-LF1.0 datasets contain two types of LFIs, $i.e.$, real-world LFIs and synthetic LFIs, we train the framework with one type and test it with another type. TABLE IX presents the cross-type validation performance on the Win5-LID and NBU-LF1.0 datasets. It can be found that the type of LFIs used for training has a small impact on the performance for both datasets, which verifies the robustness of our proposed framework.

TABLE X
PERFORMANCE OF DIFFERENT FEATURE COMBINATIONS ON THE WIN5-LID, NBU-LF1.0, AND LFDD DATASETS. HERE, $\mathcal{B}$ DENOTES BASELINE DISCREPANCY FEATURES, $\mathcal{H}$ DENOTES HIERARCHICAL DISCREPANCY FEATURES, AND $\mathcal{L}$ DENOTES LOCAL DISCREPANCY FEATURES.

Datasets	Features	PLCC	SROCC	RMSE
Win5-LID	$\mathcal{B}$	0.8036	0.7693	0.5618
	$\mathcal{B}$ + $\mathcal{H}$	0.8605	0.8288	0.4906
	$\mathcal{B}$ + $\mathcal{L}$	0.8480	0.8230	0.5068
	$\mathcal{B}$ + $\mathcal{H}$ + $\mathcal{L}$ (ours)	**0.8654**	**0.8354**	**0.4839**
NBU-LF1.0	$\mathcal{B}$	0.8514	0.7996	0.4706
	$\mathcal{B}$ + $\mathcal{H}$	0.8972	0.8621	0.3934
	$\mathcal{B}$ + $\mathcal{L}$	0.8651	0.8315	0.4574
	$\mathcal{B}$ + $\mathcal{H}$ + $\mathcal{L}$ (ours)	**0.9024**	**0.8743**	**0.3866**
LFDD	$\mathcal{B}$	0.8798	0.7948	0.5311
	$\mathcal{B}$ + $\mathcal{H}$	0.9253	0.8899	0.4244
	$\mathcal{B}$ + $\mathcal{L}$	0.8912	0.8366	0.5089
	$\mathcal{B}$ + $\mathcal{H}$ + $\mathcal{L}$ (ours)	**0.9406**	**0.9077**	**0.3794**

TABLE XI
PERFORMANCE OF DIFFERENT POOLING METHODS ON THE WIN5-LID DATASET.

Pooling Methods	PLCC	SROCC	RMSE
Median	0.8569	0.8251	0.4969
Min	0.8390	0.8190	0.5232
Max	0.7901	0.7434	0.6016
Average	**0.8654**	**0.8354**	**0.4839**

TABLE XII
PERFORMANCE OF DIFFERENT FR 2DIQA METRICS ON THE WIN5-LID DATASET.

Metrics	PLCC	SROCC	RMSE
PNSR	0.8605	0.8308	0.4911
SSIM [65]	**0.8654**	**0.8354**	**0.4839**

G. Ablation Studies

In this subsection, we conduct experiments to investigate the efficacy of different feature combinations in our proposed framework. TABLE X exhibits the experimental results on the Win5-LID, NBU-LF1.0, and LFDD datasets, in which $\mathcal{B}$ denotes baseline discrepancy features, $\mathcal{H}$ denotes hierarchical discrepancy features, and $\mathcal{L}$ denotes local discrepancy features. From the table, we can observe a significant improvement in performance when incorporating $\mathcal{H}$ into $\mathcal{B}$ or $\mathcal{B}$+$\mathcal{L}$, indicating that the angular-dominant hierarchical discrepancy features are quite discriminative for quality assessment. In addition, we can see that incorporating $\mathcal{L}$ into $\mathcal{B}$, *i.e.*, the combination of baseline CNN features and handcrafted features, can also significantly improve the performance. Besides, even combining $\mathcal{L}$ with $\mathcal{B}$+$\mathcal{H}$ can yield slight improvement, which demonstrates the effectiveness of the spatial-dominant local discrepancy features. Moreover, it can be found that the boost with $\mathcal{H}$ is more significant than that with $\mathcal{L}$, indicating that the angular-dominant hierarchical discrepancy features contribute more to quality evaluation than the spatial-dominant local discrepancy features. Finally, the combination of $\mathcal{B}$, $\mathcal{H}$, and $\mathcal{L}$ constitutes our proposed framework and consistently achieves the best performance on the three datasets.

Since our proposed framework aims to evaluate the overall quality of the whole distorted LFI, a patch quality pooling method is required. We explore the effectiveness of four pooling methods in our proposed framework on the Win5-LID dataset, including median, min, max, and average pooling methods. The performance is shown in TABLE XI. It can be found that using average pooling achieves the best performance among the four methods, probably because we adopt the same training target for all patches from the same LFI, which also verifies that our framework can be successfully trained with LFI patches.

In the local discrepancy extraction module, a FR 2DIQA metric is adopted to measure the local discrepancy scores between reference and distorted sub-aperture patches. We conduct experiments on the Win5-LID dataset to investigate the effectiveness of using different FR 2DIQA metrics for generating local discrepancy scores. To maintain the efficiency of the whole framework, we compare two representative lightweight metrics: PSNR and SSIM [65], and report the experimental results in TABLE XII. It can be found that using SSIM to measure local discrepancy scores achieves slightly better performance than using PSNR. The main reason may be that SSIM is able to measure the subtle structural differences between different local regions in LFIs.

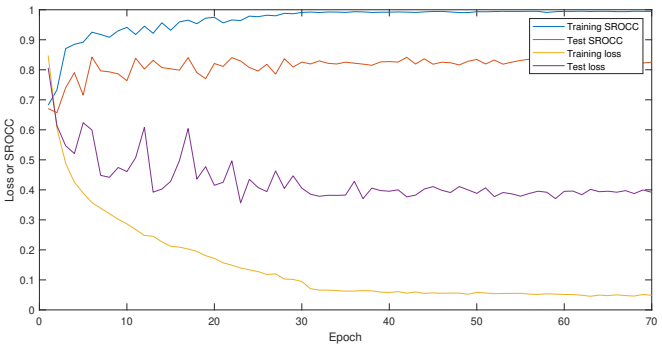


Fig. 7. Sample curves of loss and SROCC performance during training.

H. Training Process

In order to comprehensively show the training process of the proposed framework, we select the test set containing all distorted versions of *Bikes* and *dishes* reference LFIs from the Win5-LID dataset as an example, and visualize its loss and SROCC performance curves during training, as shown in Fig. 7. It can be found that the proposed framework can be trained stably and almost converges after training for 30 epochs, which demonstrates the feasibility of training the framework with LFI patches.

I. Limitation Analysis

Although the proposed EDDMF demonstrates its high efficiency and remarkable performance in terms of quality evaluation in various scenarios, it still has some limitations, which potentially provide valuable insights and inspiration for the follow-up works. First and foremost, we can observe from TABLE III that EDDMF performs significantly better than DeeBLiF on the Win5-LID, NBU-LF1.0, and LFDD datasets, but slightly worse than DeeBLiF on the SHU dataset. A possible explanation is that the SHU dataset involves only spatial distortions, while other three datasets have both angular and spatial distortions. In the case of the SHU dataset, the angular-dominant hierarchical discrepancy features may have a counterproductive effect on the evaluation of LFI quality. Second, the proposed framework is only applicable to 4D LFIs, *i.e.*, LFIs with two angular dimensions, but not to 3D LFIs with one angular dimension, such as LFIs in [74]. Finally, most current LFIs have relatively low angular resolution (typically 9×9) due to the resolution trade-off of the LFI acquisition hardware. The proposed framework demonstrates high efficiency in handling such low angular resolution LFIs. However, as the angular resolution increases, more SAIs need to be introduced to ensure accurate quality evaluation. As a result, the computational complexity of our framework will increase accordingly.

V. CONCLUSION

In this paper, we present an Efficient Deep Discrepancy Measuring Framework (EDDMF) for full-reference light field image quality assessment. Considering the high-dimensional characteristic of LFIs, we propose to evaluate the LFI quality by measuring the quality of LFI patches, which greatly

reduces the computational complexity. Thus, we present a patch generation module to generate spatio-angular patches and sub-aperture patches for quality assessment. Besides, considering that spatial and angular information are two different characteristics of LFIs, we argue that they should be treated differently to maximize the efficiency and effectiveness. To this end, we propose a CNN-based hierarchical discrepancy network to extract the angular-dominant hierarchical discrepancy features between reference and distorted spatio-angular patches. Further, we propose a local discrepancy extraction module to extract the spatial-dominant local discrepancy features between reference and distorted sub-aperture patches, which are regarded as complementary features to provide a more comprehensive description of the quality degradation in patches. Experimental results on four publicly available LFI datasets show that our proposed EDDMF achieves superior performance and lower computational complexity compared with several types of state-of-the-art IQA metrics. In the future, we will consider to delve into LFI's characteristics and explore the potential of CNN structures in LFIQA metrics.

REFERENCES

[1] E. Upenik, I. Viola, and T. Ebrahimi, "A rendering solution to display light field in virtual reality," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2018, pp. 246-250.

[2] I. Schillebeeckx and R. Pless, "Using Chromo-coded light fields for augmented reality," in *Proc. IEEE Virtual Reality (VR)*, 2016, pp. 281-282.

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

[4] 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.

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

[6] 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.

[7] H. Huang, H. Zeng, Y. Tian, J. Chen, J. Zhu, and K. K. Ma, "Light field image quality assessment: An overview," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, 2020, pp. 348-353.

[8] 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.

[9] 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.

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

[11] H. Lv, K. Gu, Y. Zhang, and Q. Dai, "Light field depth estimation exploiting linear structure in EPI," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2015, pp. 1-6.

[12] Z. Hu, H. W. F. Yeung, X. Chen, Y. Y. Chung, and H. Li, "Efficient light field reconstruction via spatio-angular dense network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-14, 2021.

[13] G. Wu, Y. Liu, L. Fang, and T. Chai, "Revisiting light field rendering with deep anti-aliasing neural network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5430-5444, 2021.

[14] N. Venkatanath, D. Praneeth, Bh. 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.

[15] 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.

[16] 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.

[17] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636-650, 2000.

[18] H. R. Sheikh, A. C. Bovik, and G. D. Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117-2128, 2005.

[19] 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.

[20] C. Yang, X. Zhang, P. An, L. Shen, and C.-C.-J. Kuo, "Blind image quality assessment based on multi-scale KLT," *IEEE Trans. Multimedia*, vol. 23, pp. 1557-1566, 2021.

[21] 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.

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

[23] 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.

[24] 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.

[25] D. Sandić-Stanković, D. Kukolj, and P. L. Callet, "DIBR-synthesized image quality assessment based on morphological wavelets," in *Proc. IEEE Int. Conf. Qual. Multimedia Exper. (QoMEX)*, 2015, pp. 1-6.

[26] D. Sandić-Stanković, D. Kukolj, and P. L. Callet, "DIBR-synthesized image quality assessment based on morphological multi-scale approach," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, pp. 1-23, 2016.

[27] D. Sandić-Stanković, D. Kukolj, and P. L. Callet, "Multi-scale synthesized view assessment based on morphological pyramids," *J. Elect. Eng.*, vol. 67, no. 1, pp. 1-9, 2016.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

[33] W. Heng and T. Jiang, "From image quality to patch quality: an image patch model for no-reference image quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 1238-1242.

[34] 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.

[35] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206-220, 2017.

[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] 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.

[38] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, "Content-based light field image compression method with Gaussian process regression," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 846-859, 2020.

[39] Y. Chen, P. An, X. Huang, C. Yang, D. Liu, and Q. Wu, "Light field compression using global multiplane representation and two-step prediction," *IEEE Signal Process. Lett.*, vol. 27, pp. 1135-1139, 2020.

[40] D. Liu, Y. Huang, Y. Fang, Y. Zuo, and P. An, "Multi-stream dense view reconstruction network for light field image compression," *IEEE Trans. Multimedia*, DOI:10.1109/TMM.2022.3175023.

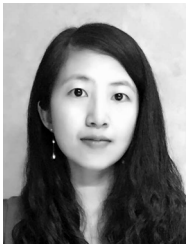
[41] P. Helin, P. Astola, B. Rao, and I. Tabus, "Minimum description length sparse modeling and region merging for lossless plenoptic image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1146-1161, 2017.
[42] J. M. Santos, L. A. Thomaz, P. A. Assunção, L. A. Cruz, L. Távora, and S. M. Faria, "Lossless coding of light fields based on 4D minimum rate predictors," *IEEE Trans. Image Process.*, vol. 31, pp. 1708-1722, 2022.
[43] X. Hu, Y. Pan, Y. Wang, L. Zhang, and S. Shirmohammadi, "Multiple description coding for best-effort delivery of light field video using GNN-based compression," *IEEE Trans. Multimedia*, vol. 35, pp. 690-705, 2021.
[44] N. Mehajabin, M. T. Pourazad, and P. Nasiopoulos, "An efficient pseudo-sequence-based light field video coding utilizing view similarities for prediction structure," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2356-2370, 2021.
[45] M. Rossi and P. Frossard, "Geometry-consistent light field superresolution via graph-based regularization," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4207-4218, 2018.
[46] Z. Cheng, Z. Xiong, and D. Liu, "Light field super-resolution by jointly exploiting internal and external similarities," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2604-2616, 2020.
[47] 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.
[48] K. Han and W. Xiang, "Inference-reconstruction variational autoencoder for light field image reconstruction," *IEEE Trans. Image Process.*, vol. 31, pp. 5629-5644, 2022.
[49] 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.
[50] 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.
[51] 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.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] 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.
[57] 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.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] 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.
[63] 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.
[64] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2200-2211, 2019.
[65] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600-612, 2004.
[66] 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.
[67] 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.
[68] 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.
[69] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 6319-6327.
[70] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1-10, 2016.
[71] S. Zhang, H. Sheng, D. Yang, J. Zhang, and Z. Xiong, "Micro-lens-based matching for scene recovery in lenslet cameras," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1060-1075, 2018.
[72] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1646-1654.
[73] 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.aspx.
[74] 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.

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 Digital Technologies 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 member of the Institut d'Electronique et Technologies du numéRique (IETR) laboratory, within the VAADER research team. She received the M.S. degree from ENSPS Strasbourg in 1989 and the PhD degree from INPG Grenoble in 1993. She prepared her PhD within the LIFIA laboratory on the topic of projective invariants applied to computer vision. From 1993 to 2008, she was an associate professor at University of Rennes and a member of the Temics team in the IRISA laboratory. Since 2008, she has joined INSA Rennes and IETR laboratory. 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 multi-view 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.