IEEE Open Journal of Signal Processing

Received 26 February 2026; revised 14 July 2026; accepted 11 August 2026. Date of publication 1 September 2026; date of current version 15 September 2026. The review of this article was arranged by Associate Editor Bruno Zatt.

Digital Object Identifier 10.1109/OJSP.2026.3729157

Dual Representation-Based Light Field View Synthesis using Deformable Convolutional and Deep Residual Channel Attention Networks

MUHAMMAD ZUBAIR (Member, IEEE), PAULO NUNES (Member, IEEE), CAROLINE CONTI (Member, IEEE), AND LUÍS DUCLA SOARES (Senior Member, IEEE)

Instituto Universitário de Lisboa (ISCTE-IUL), Instituto de Telecomunicações, 1049-001 Lisbon, Portugal

CORRESPONDING AUTHOR: MUHAMMAD ZUBAIR (e-mail: muhammad.zubair@lx.it.pt).

This work was supported in part by the National Funds Through FCT – Fundação Para a Ciência e a Tecnologia, I.P. and in part by EU Funds through Project/support UID/50008/2025 – Instituto de Telecomunicações.

The code is available on our GitHub page: https://github.com/MSP-IUL/DeformLF-DRCA.

ABSTRACT Light Field (LF) cameras simultaneously capture both intensity values and directional information of light rays in a single exposure, providing a unique perspective for computational photography and 3D geometry perception. However, existing LF cameras are constrained by sensor resolution, limiting their ability to capture high spatial and angular resolutions simultaneously. To mitigate these issues, various learning-based methods have been proposed to increase the angular resolution of captured LFs, known as LF View Synthesis (LFVS). Many of these methods either neglect essential geometric cues or rely on neural networks with large receptive fields, which restrict their ability to accurately exploit LF structural characteristics. In response to these challenges, this paper introduces a dual representation-based LFVS method that employs deformable convolutional and Deep Residual Channel Attention (DRCA) networks. The proposed method includes two main modules: (i) Coarse Light Field View Synthesis (CLFVS) for initial LFVS, and (ii) Coarse-To-Fine Refinement (CFR) for final quality enhancement. The CLFVS module relies on deformable convolutions to adaptively extract LF features using two parallel networks: (i) Spatial Feature Extraction (SPFE) network using a depth-dependent LFVS approach, and (ii) Angular Feature Extraction (AFE) network using a non-depth-dependent LFVS approach. The CFR module refines the CLFVS output using a DRCA network, which employs dense residual connections between residual groups instead of conventional convolutional layers. The DRCA network employs Residual Channel Attention Blocks (RCABs) to model inter-channel dependencies, selectively enhancing meaningful features while suppressing irrelevant ones. The proposed method achieves state-of-the-art performance on synthetic and real-world LF benchmarks.

INDEX TERMS Light field, view synthesis, residual channel attention, spatial-angular correlation, deformable convolution.

I. INTRODUCTION

Light Field (LF) cameras record both intensity values and directional information of light rays from real-world scenes, thus enabling the simultaneous acquisition of spatial and angular data. This comprehensive representation makes LF cameras highly valuable for a wide range of practical applications, such as 3D scene reconstruction [1], [2], [3], depth estimation [4], [5], post-capture refocusing [6], saliency detection [7], and over-segmentation [8]. However, despite their potential, acquiring high quality LF images remains a challenge, which limits their broader practical adoption.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Early LF acquisition methods use either camera arrays [9] to capture LFs in a single shot or a single camera mounted on a computer-controlled gantry [10] to capture views sequentially. Both setups can acquire high-angular-resolution LFs by aggregating data from densely sampled viewpoints, but are costly and bulky, limiting their adoption in commercial systems. Furthermore, gantry-based systems are limited to static scenes due to their sequential capture process, making them unsuitable for dynamic environments.

To overcome these limitations and enable compact, single-shot LF capture, plenoptic cameras (e.g., Lytro [11], Raytrix [12]) were introduced. These cameras use a micro-lens array between the main lens and the image sensor to simultaneously capture spatial and angular information. In such systems, angular resolution depends on the number of sensor pixels under each micro-lens. However, sensor multiplexing introduces a fundamental trade-off between angular and spatial resolution, often resulting in insufficient detail in both dimensions.

Instead of relying on physical LF devices, many learning-based methods [3], [13], [14], [15], [16], [17], [18] have been developed to synthesize dense LFs from a sparse set of input views. These LF View Synthesis (LFVS) methods use different kinds of LF representations. The choice of the LF representation(s) plays a crucial role in LFVS methods, as each representation emphasizes different structural characteristics. In the proposed method, Sub-Aperture Image (SAI) and Micro-Image (MI) representations are selected due to their complementary properties:

- **SAI representation** – A 4D LF captured with a camera array from different viewpoints is naturally represented as a 2D array of 2D (LF) views. When the LF is captured using a plenoptic camera, these 2D views are more commonly referred to as SAIs. Due to their similarity to conventional 2D images, SAIs are typically associated with spatial feature extraction [19], [20].
- **MI representation** – A 4D LF captured with a plenoptic camera is naturally organized as a 2D array of 2D MIs, where each MI captures light rays from the same spatial position but different angular directions. MIs effectively encode angular variations in the scene, facilitating angular feature extraction for view synthesis [1], [21] and depth estimation [2], [22].

To address the LFVS challenges, we previously introduced a LFVS method [15] that processes diverse LF representations through three dedicated branches: (i) Angular Synthesis Branch (ASB), (ii) Spatial Synthesis Branch (SSB), and (iii) Pseudo-Temporal Synthesis Branch (PTSB), which are followed by a Feature Fusion Network (FF-Net) to fuse multi-representation information. In the ASB, a 2D Deformable Convolutional Neural Network (DCNN) extracts features from SAIs. In the SSB and PTSB, features are extracted from MIs and a Pseudo Video Sequence (PVS) representation using a 3D depth-wise separable convolutional and a simple 3D convolutional network, respectively.

Despite its effectiveness, our previous method [15] only partially exploited the Spatial-Angular Correlation (SAC) and did not fully model the LF geometry, especially cross-view dependencies. This limitation underscores the need for improved method with better geometric consistency for LFVS. In this paper, we propose a novel LFVS method based on a dual representation framework and enhanced feature modeling techniques using a hybrid approach that combines two complementary LFVS strategies: (i) a depth-dependent strategy based on the SAI representation, which captures the geometric structure of the scene; and (ii) a non-depth-dependent strategy based on the MI representation, which preserves SAC without relying on explicit geometry.

This hybrid approach processes the LF representations in parallel within a Coarse Light Field View Synthesis (CLFVS) module, which employs 2D DCNNs [23] for adaptive feature extraction. Compared to our earlier work [15], which combined a 2D DCNN with a rigid 3D depth-wise separable convolutional network in the ASB and SSB, the proposed dual-representation design offers greater flexibility in extracting spatial-angular features, resulting in a more robust initial LFVS.

To further improve the synthesized views, a Coarse-To-Fine Refinement (CFR) module is then applied. This module uses a Deep Residual Channel Attention (DRCA) network [24]. Unlike conventional convolutional refinement [14], [20], [15], the proposed DRCA network uses dense residual groups and Residual Channel Attention Blocks (RCABs) to model inter-channel dependencies. This enables the network to amplify relevant features while suppressing irrelevant ones [25], [26], resulting in high fidelity view synthesis. The main contributions of the proposed method can be summarized as follows:

- We design a progressive expansion–contraction strategy to extract dense multi-scale features from the SAI representation using a 2D DCNN in three parallel streams of varying kernel sizes, which effectively capture both local and global geometric structures for depth estimation.
- We design a hierarchical features extraction network using a 2D DCNN with progressively increasing kernel sizes to model SAC from the MI representation. This multi-level receptive field approach ensures the preservation of both structural integrity and angular consistency.
- We propose a DRCA network to select meaningful features more effectively and suppress irrelevant ones to refine the initial coarse synthesized views.

The remainder of the paper is organized as follows: Section II reviews the related work; Section III describes the proposed method; Section IV presents the experimental results and discussion; and, finally, Section V concludes the paper.

II. RELATED WORK

Available LFVS methods can be categorized as either being depth-dependent or non-depth-dependent.

A. DEPTH-DEPENDENT METHODS

Depth-dependent methods synthesize dense LFs by warping the sparse set of input views to target angular positions using estimated depth maps. Kalantari et al. [3] use two sequential CNNs to estimate depth maps and color values. However, their

method does not fully exploit the angular information inherent in LF data. Jin et al. [13] propose two modules for LFVS: (i) a depth-based warping module that explicitly models scene geometry using convolutional layers with large receptive fields and warps the source views to target views based on the estimated depth maps, and (ii) a refinement blending module that exploits SAC among the warped views to obtain the dense LFs. In a subsequent study, Jin et al. [20] address the depth-related challenges identified in earlier work [13] by introducing a coarse-to-fine reconstruction framework. In this method, they construct Plane Sweep Volumes (PSVs) with a predefined disparity range to perform robust disparity estimation, effectively evaluating multiple depth hypotheses by warping the source views across candidate disparity planes. This geometry-aware strategy yields an initial set of coarse novel views, which are then refined by a subsequent network to recover fine details and correct parallax structure. Yun et al. [27] propose a depth estimation module based on hierarchical multiple receptive fields and a blending module for refining the warped views. Shi et al. [28] jointly exploit pixel and feature domains for LFVS. In the pixel domain, occlusions are explicitly addressed using a disparity-dependent interpolation filter, and target views are reconstructed from warped views via convolutional layers. In the feature domain, multi-scale features are warped to the target viewpoint, and fine-scale features are employed to infer the synthesized view. A soft mask is then learned to adaptively merge the outputs from both domains. Gul et al. [29] proposed method is divided into three stages: (i) stereo-feature extraction, to capture spatial correlations, (ii) depth estimation, for geometric alignment, and (iii) adaptive image refinement, to enhance view quality. They use a Convolutional Block Attention Module (CBAM) [30] to emphasize critical features and improve stereo alignment. Liu et al. [31] propose a depth estimation network using convolutional layers having large receptive fields to predict depth maps and warp these depth maps to target positions. For further refinement, an edge-aware inpainting network equipped with an edge pyramid structure is introduced to address artifacts and inaccuracies near edges. Ko et al. [32] extract depth-compensated multi-view features using a trainable depth estimator and then concatenate these multi-view features. The Adaptive Feature Remixing (AFR) is used to perform channel-wise pooling and remix the concatenated features according to the angular coordinates of the input views. These remixed features are adopted to produce blending filters for LFVS. Liu et al. [14] propose two modules: (i) Multi-Representation View Reconstruction (MRVR), and (ii) Geometry-Assisted Refinement (GAR). The MRVR module extracts features from multiple LF representations using three dedicated pipelines of convolutional networks for Intermediate LF View Synthesis (ILFVS). The GAR module refines the ILFVS using a bidirectional horizontal and vertical view stack structure. Chen et al. [33] propose a LFVS framework that combines an explicit-depth-based pipeline and an implicit-depth-based pipeline in a two-branch structure to handle both small and large-disparity regions. They construct PSVs from the sparse input views and apply 3D convolutions to perform feature decoupling, using a single feature extraction process to separate disparity and texture information into different dimensions. The disparity feature drives the explicit pipeline, while the texture feature drives the implicit pipeline. Finally, a disparity-guided combination generates learnable disparity-guided masks that fuse the two LF types, filtering aliasing in the implicit-depth LF while retaining anti-aliasing ability in large-disparity regions. Cai et al. [34] propose a method that targets replication errors caused by occlusion during disparity-based warping. After estimating disparity maps from PSVs, a disparity enhancement network based on adaptive morphological filtering adaptively corrects disparity values in object edge regions to obtain proper sampling positions. A refinement network consisting of Spatial-Angular Information Fusion (SAIF) blocks with pseudo-4D Channel-Attention Residual Blocks (CARBs) restores photo-consistency and fine details, achieving strong results on large-disparity LFs. In summary, existing depth-dependent methods employ large receptive fields [13], [27] or multi-branched convolutional networks [14] for depth estimation. While these methods effectively capture global context, they often struggle to represent fine-grained structural variations and complex depth transitions, particularly within occluded or highly textured regions. To address these limitations, our proposed method uses RCABs for adaptively rescaling features. These blocks allow the network to focus on spatial details and high-frequency structures, leading to superior results in challenging scenes.

B. NON-DEPTH-DEPENDENT METHODS

Non-depth-dependent methods synthesize dense LFs by directly exploiting angular coherence and spatial redundancy within the LF structure. Yoon et al. [35] propose a deep CNN method to perform spatial and angular Super-Resolution (SR) simultaneously. Their method uses an initial CNN to upsample the SAIs for ILFVS and then uses another CNN to refine the ILFVS. Yeung et al. [36] first perform ILFVS by exploring the coarse characteristics through the use of spatial-angular alternating convolutions. Then, the ILFVS is refined by further recovering the fine relations from SAIs via guided residual learning. Liu et al. [37] propose a Densely Connected Multi-loss CNN (DCM-CNN) framework by combining reconstruction loss, angular consistency loss, and perceptual loss for LFVS. Wu et al. [38] use an attention mechanism within a multi-scale reconstruction framework to capture non-local correlations between views. Hu et al. [39] propose a spatial-angular dense feature extraction network using densely connected CNNs to enhance feature reuse and maintain parallax consistency. Wang et al. [21] propose a disentangling mechanism by using domain-specific convolutions in Disentangling Groups (Distg-Groups) and Disentangling Blocks (Distg-Blocks) to extract spatial, angular, and Epipolar Plane Image (EPI) features. The network converts the input SAIs into MIs and EPIs for feature extraction and performs multi-stage information fusion using a 1×1 convolution. Liu et al. [19] propose an Efficient

LF Angular Super-Resolution (LF-EASR) 3D UNet to extract multi-scale spatial-angular features and then perform angular upsampling on MIs to synthesize dense LFs. Mao et al. [40] propose a domain-specialized network for LF spatial-angular SR that decouples the task into stage-tailored modules. In the angular SR stage, a Cross-View Consistency Modulation Module (CVCMM) enhances inter-view coherence using cascaded wavelet-based frequency decomposition over EPI and MI representations to model long-range angular dependencies. In the spatial SR stage, a Detail-Aware State Space Model (DASSM) built on Mamba-style state-space blocks reconstructs fine-grained texture. Finally, a Cross-Domain Integration Module (CDIM) employs 3D convolutions over four LF representation stacks (MI, SAI, and horizontal/vertical EPIs) to fuse multi-representational features and exploit SAC. Liu et al. [41] propose the Implicit and Detail-Enhanced Network (IDNet) for LF spatial-angular SR. It first uses 3D convolutions to jointly extract spatial and angular features. Then, an Implicit Detail Restoration (IDR) module enhances high-frequency details using an Implicit Neural Representation (INR) and a Detail-Enhanced Convolution Block (DECB) that learns gradient priors in multiple directions. Finally, a Multi-Representation Enhancement Block (MREB) processes SAI, MI, and EPI representations to exploit SAC and improve angular consistency. Liang et al. [42] propose a generic Epipolar Transformer mechanism that captures non-local SAC along orthogonal epipolar lines. It stacks Orthogonal Epipolar Modeling Blocks (OEMB), each containing two shared-parameter Epipolar Transformer units for the horizontal and vertical directions. These units use self-attention to model long-range correlations, while a LocalConv layer enhances local spatial features. In summary, the existing non-depth-dependent methods [19] and [21] do not consider geometric cues and solely synthesize dense LFs without considering the underlying scene geometry, degrading spatial-angular fidelity.

III. PROPOSED METHOD

An LF is a 4D function $L(u,\ v,\ x,\ y)$, where $(u,\ v)$ and $(x,\ y)$ denote the angular and spatial coordinates, respectively. The objective of the proposed method is to synthesize a dense LF, $L^{har} \in \mathbb{R}^{U\times V\times X\times Y}$, with angular resolution of $U \times V$ and spatial resolution of $X \times Y$, from a sparse set of input views, $L^{lar} \in \mathbb{R}^{U'\times V'\times X\times Y}$, with angular resolution of $U' \times V'$ (where $U' < U$ and $V' < V$). The process can be formulated as:

$$L^{har}(u,\ v,\ x,\ y) = f\left(L^{lar}\left(u',\ v',\ x, y\right),\ \theta\right) \qquad (1)$$

where $f(\cdot)$ denotes the mapping function and θ represents the learnable network parameters. Following [13], [14], [15], [19], [21], [42], the LF images are converted from RGB to the YCbCr 4:4:4 color space (no chroma subsampling) using the ITU-R BT.601 conversion matrix. The proposed method operates exclusively on the luminance channel (Y_{luma}) of each input view in L^{lar}, since the chrominance channels contain predominantly low-frequency content. For the qualitative RGB figures, each synthesized luminance view is recombined with its Ground-Truth (GT) chrominance (Cb, Cr) channels of the input LF to form a color image. This color reconstruction is used solely for visualization; all reported Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) values are computed on the synthesized Y_{luma} channel only. For the proposed dual LF representation method, we decompose the L^{lar} into a set of MIs, L_m. The modules of the proposed method are detailed below, and its architecture is illustrated in Fig. 1.

A. COARSE LIGHT FIELD VIEW SYNTHESIS (CLFVS)

The CLFVS module uses DCNNs to extract multi-scale and hierarchical features from two different (dual) representations: (i) SAIs, and (ii) MIs. Unlike standard convolutions with fixed sampling grids, the deformable convolutions introduce learnable offsets, allowing the convolutional kernel to dynamically adapt to geometric variations in the input data [23]. This adaptation is particularly important in LFVS, because the same scene point appears at disparity-shifted locations across angular views, according to the scene depth. By predicting offsets at each location, the deformable convolution lets the sampling grid follow these depth-dependent shifts and adapt around occlusion boundaries, thereby aligning the receptive fields with the underlying LF geometry. This way, the proposed method captures SAC more effectively than fixed-grid convolutions, enabling robust and detailed feature extraction. The two CLFVS parallel networks are described below.

- **Spatial Feature Extraction (SPFE)** – The SPFE is designed to employ a disparity-based warping paradigm to model the angular geometry correlation between adjacent SAIs. Since disparity is inversely proportional to depth, with larger disparities corresponding to closer objects and smaller disparities to farther ones, this relationship enables effective depth-dependent modeling within the SAI representation.
  For this, we design a Disparity Estimation Network (DEN) using three parallel streams of 2D deformable convolutional layers to extract multi-scale features from L^{lar}. The first, F_1, and second, F_2, streams extract shallow multi-scale features using 3×3 and 5×5 convolutional kernels, respectively. To extract dense features across multiple receptive fields, the third stream, F_3, progressively expands the kernel sizes from 3×3 to 7×7, increasing the receptive field, thus enhancing the network's ability to model long-range dependencies. Subsequently, the progressive contraction of the kernel size back to 3×3 refines the extracted features by reintroducing fine structural details while mitigating excessive smoothing.
  This expansion-contraction strategy balances local and global features, improving multi-scale feature extraction. The extracted multi-scale features from all three streams are concatenated (C) to produce F_d and then processed through a 1×1 convolutional layer to generate unified disparity maps, F_{dm}, which are subsequently passed to

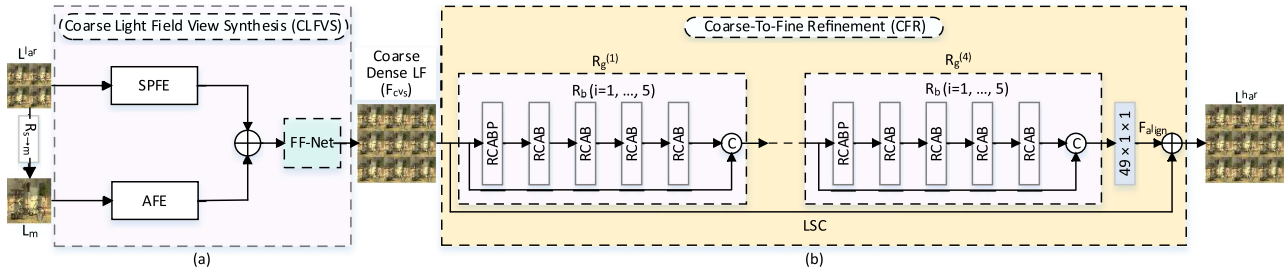


FIGURE 1. Overview of the proposed method consisting of two modules: (a) CLFVS and (b) CFR. CLFVS has two parallel networks: SPFE and AFE, followed by FF-Net. The CFR module uses a DRCA network, which consists of cascaded dense residual blocks within residual groups.

the warping block along with L^{lar}, for synthesizing the initial coarse LF.

The feature extraction process in three streams of DEN can be expressed as:

$$F_1 = \emptyset\left(W_{3\times3} \circledast_{\Delta_{3\times3}} L^{lar} + b_{3\times3}\right) \tag{2}$$

$$F_2 = \emptyset\left(W_{5\times5} \circledast_{\Delta_{5\times5}} L^{lar} + b_{5\times5}\right) \tag{3}$$

$$F_s^{(i)} = \emptyset\left(W_{n\times n}^{(i)} \circledast_{\Delta_{n\times n}} F_s^{(i-1)} + b_{n\times n}^{(i)}\right) \tag{4}$$

$$F_d = C\left(F_1, F_2, F_3\right) \tag{5}$$

where F_d represents the disparity feature maps, i represents the index of the convolutional layer, $n \times n$ represents the kernel sizes of the i^{th} layer, $W_{n\times n}^{(i)}$ denotes the learnable weights of the convolutional kernel, $b_{n\times n}^{(i)}$ represents the bias term, $F_s^{(i)}$ represents the feature maps in each layer in the third stream, where $F_s^{(0)} = L^{lar}$ and $F_3 = F_s^{(5)}$, $\circledast$ denotes the deformable convolution operator with $\Delta_{n\times n}$ learnable offsets, and $\emptyset(\cdot)$ denotes the Leaky ReLU activation function.

The warping block synthesizes the initial coarse LF by shifting pixels from L^{lar} using the predicted disparity maps F_{dm} at the target angular position. For simplicity, angular and spatial positions are denoted as two-dimensional coordinate pairs in bold, where $(\boldsymbol{u}', \boldsymbol{x})$ is a pixel at angular position $\boldsymbol{u'} = (u', v')$ and spatial position $\boldsymbol{x} = (x, y)$. For a given target angular position $\boldsymbol{u} = (u, v)$ and a reference input view at angular position $\boldsymbol{u'} = (u', v')$. The warping operation can be defined as:

$$F_{(\boldsymbol{u},\,\boldsymbol{x})} = L^{lar}\left(\boldsymbol{x} + F_{dm}\left(\boldsymbol{x}\right)\,.\,\left(\boldsymbol{u} - \boldsymbol{u'}\right),\,\boldsymbol{u'}\right) \tag{6}$$

Where $F_{(\boldsymbol{u},\,\boldsymbol{x})}$ represents warped features at angular position $\boldsymbol{u}$ and spatial location $\boldsymbol{x}$. $F_{dm}(\boldsymbol{x})$ is the predicted disparity at pixel $\boldsymbol{x}$, $(\boldsymbol{u} - \boldsymbol{u'})$ is the angular offset between the target and reference views.

The warped features $F_{(\boldsymbol{u},\,\boldsymbol{x})}$ are processed using a 2D deformable convolution with a 3×3 kernel to capture SAC. These features are then reshaped to align angular and spatial coordinates for 3D convolutional processing, denoted as F_{wr}. Three successive 3D convolutional layers are used to extract disparity-aware features, with Leaky ReLU applied after each of the first two layers to introduce non-linearity. The process can be defined as:

$$F_{wr}^{(i)} = \emptyset\left(W_{n\times n\times n}^{(i)} * F_{wr}^{(i-1)} + b_{n\times n\times n}^{(i)}\right) \tag{7}$$

where $F_{wr}^{(0)} = F_{wr}$, $F_{ms} = F_{wr}^{(3)}$ represents the disparity-aware features, '$*$' represents the convolution operation, $n \times n \times n$ represents the kernel sizes of $5 \times 3 \times 3$, $4 \times 3 \times 3$, and $3 \times 3 \times 3$ at the i^{th} layer, $F_{wr}^{(i-1)}$ represents the features from the previous layer, and, finally, $W_{n\times n\times n}^{(i)}$ and $b_{n\times n\times n}^{(i)}$ represent the weight and bias terms at the i^{th} layer ($i = 1,\ 2,\ 3$).

- **Angular Feature Extraction (AFE)** – The AFE is designed to extract hierarchical SAC features using 2D DCNN. As shown in Fig. 2(b), the initial convolutional layer of AFE extracts local features from L_m, using a 3×3 kernel. Subsequently, we expand the receptive fields to 5×5 and 7×7 to capture wider SAC feature representations. Another 3×3 deformable convolutional layer is used to enforce angular consistency. The hierarchical feature extraction mechanism can be described as follows:

$$F_m^{(i)} = \emptyset\left(W_{n\times n}^{(i)} \circledast_{\Delta_{n\times n}} F_m^{(i-1)} + b_{n\times n}^{(i)}\right) \tag{8}$$

where $F_m^{(i)}$ represents the features, extracted by the i^{th} convolutional layer and $F_m^{(i-1)}$ represents the features from the previous layer (with $F_m^{(0)} = L_m$).

We use a PixelShuffle (PS) operation to rearrange the sub-pixel information into a dense grid of MI features and then reshape it into the SAI feature representation, F_{msr}. To preserve dense SAC feature representation, another 3×3 deformable convolution is used at end of the AFE (see Fig. 2(b)).

$$F_h = \emptyset\left(W_{3\times3} \circledast_{\Delta_{n\times n}} F_{msr} + b_{3\times3}\right) \tag{9}$$

where F_h represents the hierarchical angular features.

- **Feature Fusion Network (FF-Net)** – The FF-Net combines the multi-scale spatial features, F_{ms}, and the hierarchical angular features, F_h, to produce fused features, $F_a = F_{ms} + F_h$. These fused features are then processed by two 2D convolutional residual blocks for CLFVS (see Fig. 2(c)). Each residual block in FF-Net can be

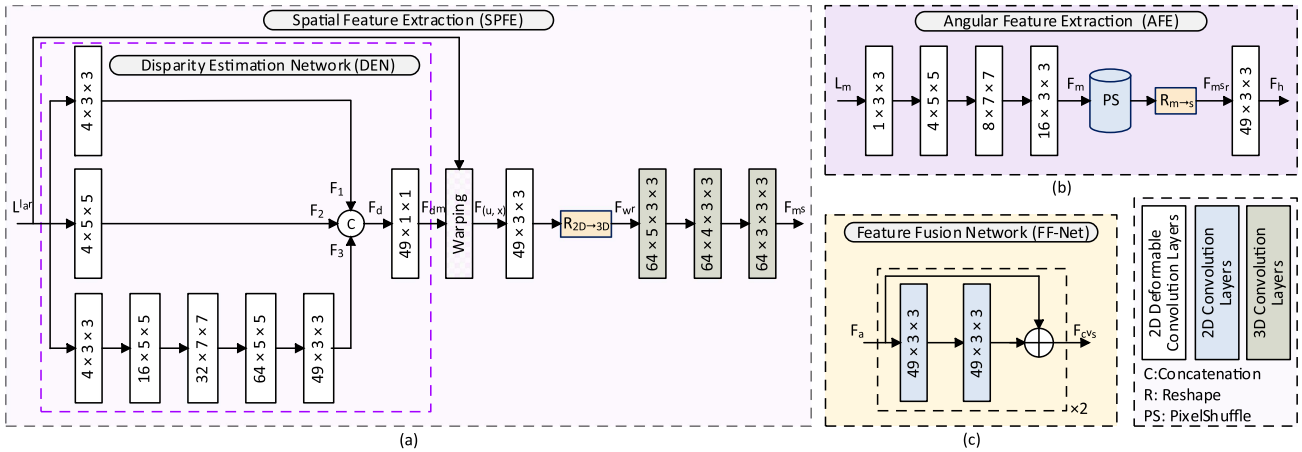


FIGURE 2. Detailed architecture of CLFVS module: (a) SPFE, (b) AFE, and (c) FF-Net.

expressed as:

$$F_{cvs} = \emptyset\left(W^{(2)}_{n\times n}\left(\emptyset\left(W^{(1)}_{n\times n} * F_a + b^{(1)}_{n\times n}\right)\right) + b^{(2)}_{n\times n}\right) + F_a \tag{10}$$

where F_{cvs} represents the coarse dense LF.

B. COARSE-TO-FINE REFINEMENT (CFR)

The CFR module utilizes a DRCA network to refine the coarse dense LF, F_{cvs}. The DRCA network is composed of four residual groups, R_g, where each residual group contains five residual blocks, R_b. The first block, $R^{(1)}_b$, in each group is implemented as a Residual Channel Attention Block with Projection (RCABP) to perform feature alignment. The remaining four blocks, $R^{(2)}_b$ to $R^{(5)}_b$, employ standard RCABs to progressively refine the feature representations within the residual groups (see Fig. 1(b)). The architecture of each RCABP consists of two sequential 3×3 convolutional layers, followed by a Channel Attention (CA) network. The first convolutional layer captures local spatial features, and a Leaky ReLU activation introduces non-linearity. The second convolutional layer refines the extracted features, and the CA network enhances feature representation by first compressing spatial information into a single value per channel through Global Average Pooling (GAP). It then applies two 1×1 convolutions with a non-linearity in between these layers to generate channel-wise attention weights. A sigmoid activation (S) function maps these weights to a bounded range between 0 and 1, enabling the network to assign soft attention scores to each channel. The network multiplies learned attention weights with the original features to emphasize the most informative channels and suppress the irrelevant ones. Feature alignment is then performed via a 1×1 convolutional projection to refine the channel dimensions (see Fig. 3(a)).

The architecture of each standard RCAB, used in the remaining residual blocks, follows a similar architecture but without the convolutional projection (see Fig. 3(b))

All residual groups, $R^{(k)}_g$ ($k = 1, \ldots, 4$), possess the same architecture. The first residual group, $R^{(1)}_g$, receives F_{cvs} as

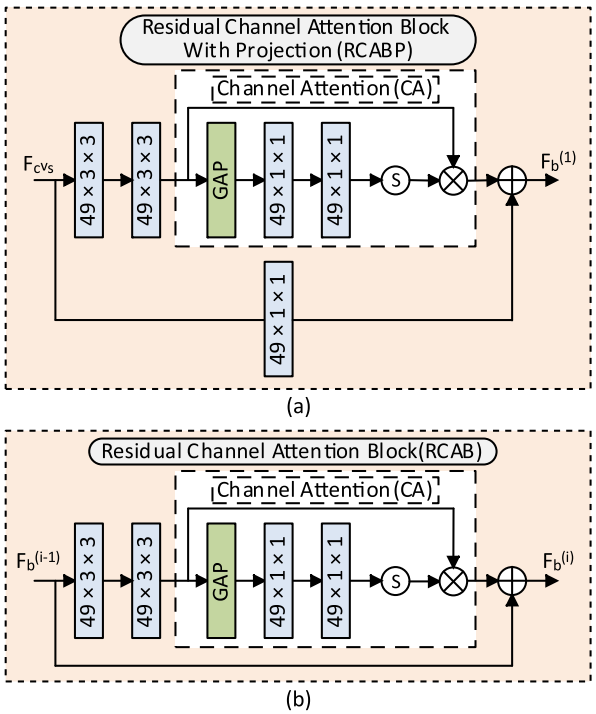


FIGURE 3. The architecture of (a) RCABP to align the features and (b) the standard RCAB to progressively refine the feature representations.

input and passes it to its first residual block, $R^{(1)}_b$, which employs a RCABP, and the output of this block is passed to the following RCAB successively. This process can be expressed as:

$$F^{(i)}_b = R^{(i)}_b\left(F^{(i-1)}_b\right), \; i = 1, \ldots, 5 \tag{11}$$

where $F^{(i)}_b$ denotes the output features of the i^{th} RCAB and $F^{(0)}_b = F_{cvs}$. The output of the last residual block in each residual group $R^{(k)}_b$ is concatenated (C) with the group's initial input features, yielding enriched features. This process can be expressed as:

$$F^{(k)}_g = C\left(F^{(5)}_b, F^{(k-1)}_g\right), \; k = 1, \ldots, 4 \tag{12}$$

where the output of the residual group k, $F_g^{(k)}$, will serve as the input to the next residual group, facilitating a progressive refinement process.

In our proposed DRCA network, this design enables multi-level feature aggregation and efficient feature reuse across residual groups, allowing it to capture both local and global context with detailed LF structure. The output of the final residual group, $R_g^{(4)}$, is processed through a 1×1 convolutional layer followed by a Leaky ReLU activation function for feature alignment, as defined in (13).

$$F_{align} = \emptyset\left(W_{1\times1} * F_g^{(4)} + b_{1\times1}\right) + F_{cvs} \tag{13}$$

The aligned features F_{align} are then fused with the Long-Range Skip Connection (LSC) to preserve global structural consistency and synthesize the dense LF, $L^{har} \in \mathbb{R}^{U\times V\times X\times Y}$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we present the experimental setup, describe the datasets used, and compare the proposed method with state-of-the-art methods to evaluate its effectiveness.

A. EXPERIMENTAL SETUP

The proposed method is implemented in the PyTorch framework. The synthetic and real-world networks are individually trained for 3000 epochs with a batch size of 1, using L1 Charbonnier loss [43] and the Adam optimizer [44]. The initial learning rate is set to 1×10^{-4} and subsequently decreased by a factor of 2 for every 500 epochs. Each training sample consists of angularly aligned patches: a 96×96 window cropped at the same spatial position from all 7×7 SAIs of one LF, with the 2×2 corner views as input and all 49 views as GT. In each epoch, multiple randomly positioned samples are drawn per training LF, with random flips and 90° rotations applied jointly to the spatial and angular dimensions to preserve the LF geometry.

B. DATASET DESCRIPTION

The synthetic network is trained on 20 scenes from the HCI-new dataset [45] and evaluated on 4 scenes from the HCI-new [45] and 5 scenes from the HCI-old [46] datasets. The real-world network is trained on 100 LF scenes from Kalantari et al. [3] and the Stanford Lytro Archive [47], captured using a Lytro Illum camera and evaluated on three real-world datasets: 30 LF scenes from the UCSD dataset [3], 25 scenes from the occlusion dataset [47], and 15 scenes from the reflective dataset [47].

C. RESULTS COMPARISON WITH STATE-OF-THE-ART METHODS

To demonstrate the effectiveness of the proposed method, we compare it with three depth-dependent [13], [14], [15] and three non-depth-dependent [19], [21], [42] state-of-the-art methods. For a fair comparison, we retrain all competing methods from scratch with randomly initialized parameters, using their publicly available code, following the training configuration of their original papers (3000 epochs for the depth-dependent methods and 70 epochs for the non-depth-dependent methods). Performance evaluation is conducted on the Y_{luma} channel images using PSNR and SSIM metrics.

Tables 1 and 2 summarize the quantitative results of the proposed method in comparison with state-of-the-art methods on synthetic and real-world datasets. On synthetic datasets, the proposed method achieves average PSNR gains of 4.44 dB, 0.90 dB, 0.56 dB, 0.58 dB, 0.31 dB, and 0.25 dB compared to LFASR-GEO [13], GA-MRVR [14], Deformable-LFVS [15], LF-EASR [19], Distg-ASR [21], and BasicLFSR-plus [42], respectively. Similarly, on real-world datasets, the proposed method achieves average PSNR gains of 3.12 dB, 0.61 dB, 0.33 dB, 0.53 dB, 0.27 dB, and 0.35 dB over the same methods. While the proposed method achieves the highest average PSNR across the synthetic and real-world datasets, its average SSIM shows slight variations compared to LF-EASR [19] and BasicLFSR-plus [42], reflecting minor differences in structural similarity despite overall gains in fidelity. The highest values are highlighted in bold, and the second-highest values are underlined.

To analyze the visual results, from Figs. 4 to 8, error maps are used to highlight regions of deviation between the predicted central view and their corresponding GT, while PSNR and SSIM scores are also reported to quantitatively assess reconstruction fidelity.

The error maps in Figs. 4 and 5 illustrate that LFASR-GEO [13], GA-MRVR [14], and Deformable-LFVS [15] exhibit localized misaligned structures in the Bedroom scene [45] and fail to synthesize smooth texture details in the StillLife scene [46]. LF-EASR [19], as a non-depth-dependent method, shows noticeable ghosting and texture loss in both the Bedroom and StillLife scenes. Distg-ASR [21] provides a sharp reconstruction of the StillLife scene, but it struggles with the large disparity variations of the Bedroom scene. BasicLFSR-plus [42] synthesizes clean views in both scenes, with low residuals in the high-disparity regions of the Bedroom scene and the fine-texture regions of the StillLife scene. The proposed method maintains a higher degree of structural fidelity across both scenes.

Fig. 6 error maps highlight reconstruction errors in a real-world Car scene [3], focusing on the high-frequency texture details of the car bonnet. The first four methods [13], [14], [15], [19] struggle to synthesize clear texture details, resulting in noticeable over-smoothing artifacts. BasicLFSR-plus [42] also struggles to recover the fine structural details of the car bonnet, where residual errors remain visible in the error map and subtle blurring of the bonnet contours can be observed in the synthesized view. In contrast, Distg-ASR [21] and the proposed method both preserve surface consistency more effectively. While Distg-ASR [21] achieves high fidelity in fine-grained texture reconstruction, our method yields comparable visual results, significantly reducing the error compared to previous depth-dependent methods [13], [14], [15].

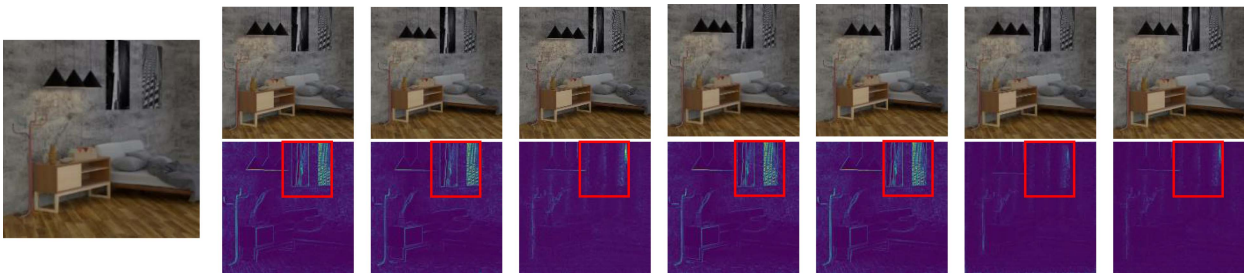
Fig. 7 error maps show the visual comparison on the non-Lambertian scene Reflective_12_eslf [47] from a real-world

TABLE 1. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Synthetic Datasets for Task 2×2 → 7×7

Dataset	LFASR-GEO [13]	GA-MRVR [14]	Deformable-LFVS [15]	LF-EASR [19]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
HCI-new	32.29/0.911	34.28/0.936	34.52/0.936	34.27/0.949	34.16/**0.969**	34.68/0.948	**35.42**/0.950
HCI-old	35.73/0.937	40.82/0.953	41.25/0.958	41.47/0.970	**42.13**/**0.974**	41.72/0.968	41.47/0.960
Average	34.01/0.924	37.55/0.945	37.89/0.955	37.87/0.960	38.14/**0.972**	38.20/0.958	**38.45**/0.955

TABLE 2. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Real-World Datasets for Task 2×2 → 7×7

Dataset	LFASR-GEO [13]	GA-MRVR [14]	Deformable-LFVS [15]	LF-EASR [19]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
UCSD	40.78/0.982	42.91/0.987	43.19/0.987	43.15/0.980	**43.60**/0.986	43.22/0.986	43.43/**0.988**
Reflective	36.25/0.945	39.04/0.962	39.19/**0.963**	38.91/0.958	39.02/0.960	39.18/0.960	**39.58**/**0.963**
Occlusions	36.43/0.971	39.04/0.981	39.45/0.984	39.18/0.979	39.40/0.982	39.36/0.984	**39.82**/**0.985**
Average	37.82/0.966	40.33/0.976	40.61/0.978	40.41/0.972	40.67/0.976	40.59/0.977	**40.94**/**0.979**



(a) (b) 38.76/0.948 (c) 40.18/0.965 (d) 41.37/0.968 (e) 40.32/0.966 (f) 39.98/0.964 (g) 41.54/0.970 (h) 41.58/0.970

FIGURE 4. Qualitative results on the synthetic scene Bedroom (HCI-new [45]): (a) GT; (b) LFASR-GEO [13]; (c) GA-MRVR [14] ; (d) Deformable-LFVS [15]; (e) LF-EASR [19]; (f) Distg-ASR [21]; (g) BasicLFSR-plus [42]; (h) Proposed Method.



(a) (b) 24.19/0.824 (c) 24.80/0.843 (d) 25.64/0.863 (e) 25.72/0.882 (f) 26.12/0.901 (g) 25.99/0.897 (h) 26.02/0.898

FIGURE 5. Qualitative results the synthetic scene StillLife (HCI-old [46]): (a) GT; (b) LFASR-GEO [13]; (c) GA-MRVR [14]; (d) Deformable-LFVS [15]; (e) LF-EASR [19]; (f) Distg-ASR [21]; (g) BasicLFSR-plus [42]; (h) Proposed Method.

dataset. Depth-dependent methods [13], [14], [15] struggle with view-dependent reflections, leading to detailed texture distortions. Non-depth-dependent methods [19], [21] mitigate these effects to some extent, but they still exhibit subtle structural misalignments in reflective regions. BasicLFSR-plus [42] similarly reduces the reflection artifacts but retains residual errors in the reflective region of the error map. The proposed method achieves competitive performance, providing refined texture reconstruction and enhanced consistency on reflective surfaces.

Fig. 8 error maps illustrate a visual comparison on the Occlusions_22_eslf scene [47]. Depth-dependent methods [13], [14], [15] show visible misalignment artifacts at the boundaries of occluded objects. Non-depth-dependent methods [19], [21], [42] also fail to preserve occlusion boundaries, causing blurring and texture inconsistencies in regions with complex depth variations. The proposed method demonstrates a more robust handling of these boundaries, preserving finer texture details and reducing the intensity of the error map in regions of high occlusion.

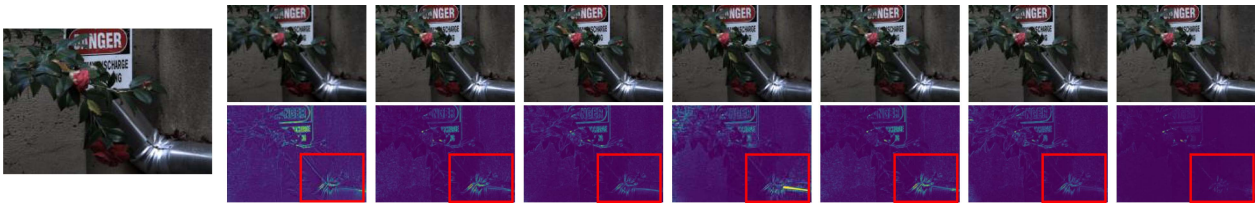
To demonstrate the generalizability of the proposed method across different angular configurations, we present results for the $3\times3\rightarrow7\times7$ and $2\times2\rightarrow8\times8$ settings. We restrict our comparison to the three methods that achieve the strongest overall performance, as this allows the visual differences between the results to be more clearly observed rather than obscured by highlighted error maps.

Tables 3 and 4 present a quantitative comparison of the proposed method against these state-of-the-art methods on synthetic and real-world datasets under the $2\times2\rightarrow8\times8$



(a) (b) 33.75/0.976 (c) 35.83/0.982 (d) 35.95/0.983 (e) 35.88/0.982 (f) 37.18/0.985 (g) 36.48/0.982 (h) 37.04/0.985

FIGURE 6. Qualitative results on the real-world scene Car (UCSD [3]): (a) GT; (b) LFASR-GEO [13]; (c) GA-MRVR [14]; (d) Deformable-LFVS [15]; (e) LF-EASR [5]; (f) Distg-ASR [21]; (g) BasicLFSR-plus [42]; (h) Proposed Method.



(a) (b) 37.04/0.970 (c) 39.55/0.981 (d) 39.62/0.982 (e) 39.30/0.977 (f) 39.42/0.979 (g) 39.48/0.980 (h) 39.70/0.983

FIGURE 7. Qualitative results on the real-world scene Reflective_12_eslf [47]: (a) GT; (b) LFASR-GEO [13]; (c) GA-MRVR [14]; (d) Deformable-LFVS [15]; (e) LF-EASR [19]; (f) Distg-ASR [21]; (g) BasicLFSR-plus [42]; (h) Proposed Method.



(a) (b) 39.38/0.970 (c) 41.53/0.972 (d) 41.62/0.972 (e) 41.55/0.971 (f) 41.61/0.972 (g) 41.60/0.972 (h) 41.72/0.973

FIGURE 8. Qualitative results on the real-world scene Occlusions_22_eslf [47]: (a) GT; (b) LFASR-GEO [13]; (c) GA-MRVR [14]; (d) Deformable-LFVS [15]; (e) LF-EASR [19]; (f) Distg-ASR [21]; (g) BasicLFSR-plus [42]; (h) Proposed Method.

TABLE 3. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Synthetic Datasets for Task 2×2 → 8×8

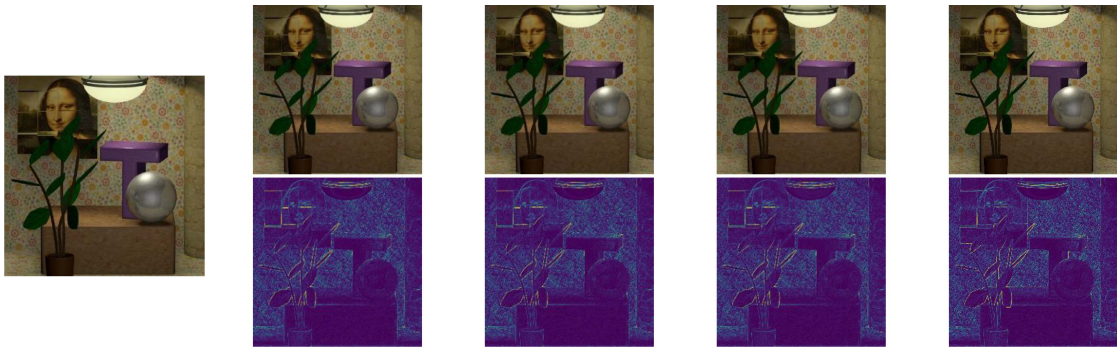
Dataset	Deformable-LFVS [15]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
HCI-new	33.93/0.928	33.72/0.948	34.08/0.944	**34. 72/0.949**
HCI-old	40.95/0.952	**41.65/0.964**	41.28/0.932	41.32/0.960
Average	37.44/0.940	37.69/**0.956**	37.68/0.938	**38.02**/0.955

TABLE 4. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Real-World Datasets for Task 2×2 → 8×8

Dataset	Deformable-LFVS [15]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
UCSD	42.72/0.983	**43.28/0.984**	42.94/0.982	43.22/**0.984**
Reflective	38.87/0.960	38.94/0.959	38.92/0.959	**39.22/0.961**
Occlusions	39.21/0.983	39.08/0.982	39.12/0.982	**39.44/0.984**
Average	40.27/0.975	40.43/0.975	40.33/0.974	**40.63/0.976**

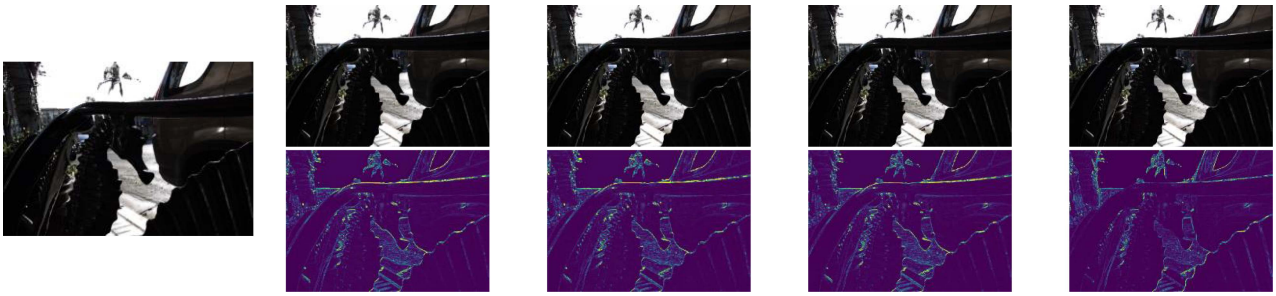
configuration. On synthetic datasets, the proposed method yields average PSNR gains of 0.58 dB, 0.33 dB, and 0.34 dB over Deformable-LFVS [15], Distg-ASR [21], and BasicLFSR-plus [42], respectively. On real-world datasets, it achieves corresponding gains of 0.36 dB, 0.20 dB, and 0.30 dB over the same baselines. In terms of structural similarity, the proposed method attains the highest average SSIM on real-world datasets, while remaining highly competitive on synthetic datasets, underscoring its consistent reconstruction quality across varying angular resolutions and reconstruction configurations.

To further assess the perceptual quality of the synthesized central view under the 2×2→8×8 configuration, we present a qualitative comparison on the challenging scene MonasRoom



(a) (b) 34.86/0.935 (c) 35.01/0.937 (d) 34.96/0.937 (e) 34.95/0.936

FIGURE 9. Qualitative results on the synthetic scene MonasRoom (HCI-old [46]) for task 2×2 → 8×8: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) BasicLFSR-plus [42]; (e) Proposed Method.



(a) (b) 25.52/0.902 (c) 25.45/0.901 (d) 25.51/0.902 (e) 27.61/0.932

FIGURE 10. Qualitative results on the real-world scene Seahorse (UCSD [3]) for task 2 × 2 → 8 × 8: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) BasicLFSR-plus [42]; (e) Proposed Method.

TABLE 5. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Synthetic Datasets for Task 3×3 → 7×7

Dataset	Deformable-LFVS [15]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
HCI-new	38.74/0.984	38.89/0.987	38.58/0.982	**39.52**/**0.988**
HCI-old	42.64/**0.986**	**43.25**/**0.986**	42.58/0.985	42.86/0.979
Average	40.69/0.985	41.07/**0.987**	40.58/0.984	**41.19**/0.984

TABLE 6. Quantitative Comparison of the Proposed Method With State-of-the-Art Methods on Real-World Datasets for Task 3×3 → 7×7

Dataset	Deformable-LFVS [15]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
UCSD	43.98/**0.992**	44.86/0.990	44.28/0.988	**45.38**/0.991
Reflective	41.96/0.968	42.74/0.969	42.86/0.970	**43.42**/**0.975**
Occlusions	42.68/0.988	42.75/**0.990**	42.52/**0.990**	**43.46**/**0.990**
Average	42.87/0.983	43.45/0.983	43.22/0.983	**44.09**/**0.985**

(HCI-old [46]) as shown in Fig. 9. The proposed method achieves comparable performance to the competing methods, with only marginal differences in PSNR and SSIM across all methods, demonstrating that it maintains competitive reconstruction quality even in the presence of high-frequency detail and non-Lambertian surfaces. In contrast, as shown in Fig. 10, the proposed method effectively recovers fine textures and intricate details on Seahorse scene as compared to the other methods.

Tables 5 and 6 present a quantitative comparison of the proposed method against state-of-the-art approaches on synthetic and real-world datasets under the 3×3→7×7 configuration. On synthetic datasets, the proposed method achieves average PSNR gains of 0.50 dB and 0.61 dB over Deformable-LFVS [15] and BasicLFSR-plus [42], respectively, while remaining closely competitive with Distg-ASR [21], trailing by only 0.12 dB on average. On real-world datasets, the proposed method achieves substantially larger gains of 1.22 dB, 0.64 dB, and 0.87 dB over Deformable-LFVS [15], Distg-ASR [21], and BasicLFSR-plus [42], respectively, highlighting its strong generalization to real-world capture conditions. In terms of structural similarity, the proposed method attains the highest average SSIM on real-world datasets, while remaining highly competitive on synthetic datasets. To further evaluate the robustness of the proposed method under a narrower

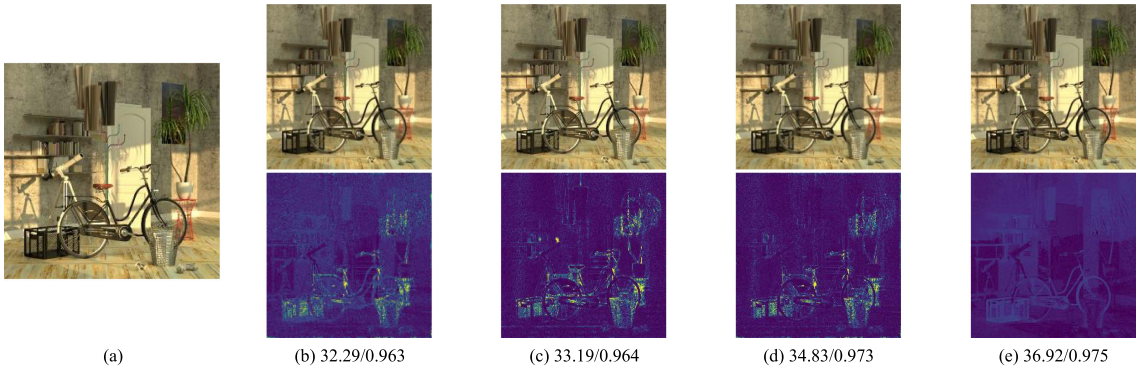


FIGURE 11. Qualitative results on the synthetic scene Bi-cycle (HCI-new [45]) for task 3×3 → 7×7: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) BasicLFSR-plus [42]; (e) Proposed Method.

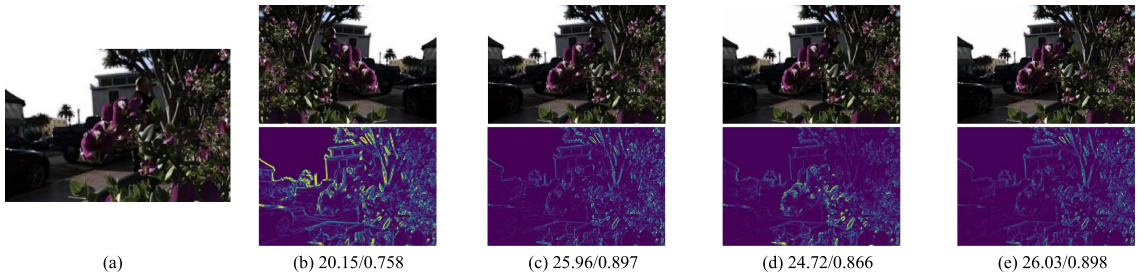


FIGURE 12. Qualitative results on the real-world scene Flower1 (UCSD [3]) for task 3×3 → 7×7: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) BasicLFSR-plus [42]; (e) Proposed Method.

baseline configuration, Figs. 11 and 12 present a qualitative comparison. On the synthetic scene Bi-cycle (HCI-new [45]), the proposed method achieves a substantial improvement over all competing approaches, with visibly reduced error concentrated around fine structural edges such as the bicycle frame and shelving. On the real-world scene Flower1 (UCSD [3]), the proposed method again achieves the best overall performance, clearly outperforming Deformable-LFVS [15], while remaining closely matched with Distg-ASR [21]. These results demonstrate that the proposed method consistently synthesizes high-fidelity views with well-preserved structural and textural detail across both synthetic and real-world scenes.

D. MODEL PARAMETERS, FLOPS AND INFERENCE TIME

We evaluate the efficiency of various methods based on the number of model parameters, Floating-Point Operations (FLOPs), reported as the number of Multiply–Accumulate operations (MACs) following the convention adopted by the compared methods [13], [21], and the inference time measured by including all module invocations during inference. All evaluations are performed on the same desktop PC, featuring a 13th Gen Intel Core^{TM} i9-13900K processor (24 cores, 32 threads, base frequency up to 3.0 GHz, turbo boost up to 5.8 GHz), 64 GB RAM, and an NVIDIA RTX 4090 GPU. Only GPU acceleration is utilized, and inference time is averaged over 10 forward passes with explicit GPU synchronization, using a batch size of 1. The proposed method achieves the lowest FLOPs among all compared methods, while using fewer model parameters than GA-MRVR [14], Distg-ASR [21], and BasicLFSR-plus [42] and maintaining superior reconstruction quality. Although LFASR-GEO [13] has the lowest parameter count, it requires the highest FLOPs and lacks the representational capacity to synthesize high-fidelity dense LFs.

Despite achieving the lowest FLOPs, the proposed method exhibits the highest inference time among the compared methods. This is mainly attributed to the use of deformable convolutions: their adaptive sampling pattern is memory-bound and achieves lower GPU utilization than standard convolutions, resulting, however, in increased latency despite the substantial reduction in the number of operations. A detailed comparison of model parameters, FLOPs, and inference time is provided in Table 7.

To further illustrate the trade-off between model efficiency and reconstruction quality, Fig. 13 visualizes the relationship between average PSNR on synthetic scenes, model parameters, and inference time across all methods.

E. ANGULAR CONSISTENCY

To assess angular consistency, we extract horizontal and vertical EPIs from the synthesized dense LFs and their corresponding GTs. We select two specific methods, Deformable-LFVS [15] and Distg-ASR [21], for comparison due to their high quantitative performance. Fig. 14 shows that both competing methods exhibit noticeable artifacts and parallax distortions in the EPI error maps and zoomed-in regions. In contrast, the proposed method reconstructs EPIs with smooth and continuous structures that are more consistent with the

TABLE 7. Model Parameters (in Millions, M), Flops (in GiGA, G), and Inference Time (in Milliseconds, ms)

	LFASR-GEO [13]	GA-MRVR [14]	Deformable-LFVS [15]	LF-EASR [19]	Distg-ASR [21]	BasicLFSR-plus [42]	Proposed Method
#Params (M)	1.01	2.50	2.32	6.63	2.68	4.12	2.30
FLOPs (G)	72.09	44.48	41.36	36.20	34.84	34.58	19.60
Inference Time (ms)	27.06	27.50	26.60	6.40	15.70	10.60	37.30

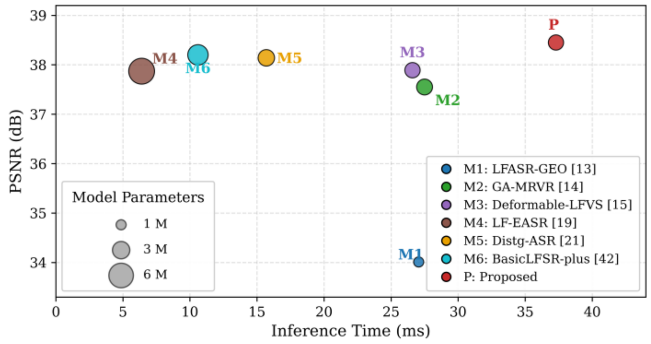


FIGURE 13. Comparison of PSNR and inference time across various methods: circle size indicates the number of model parameters; M1–M5 and P represent method labels.

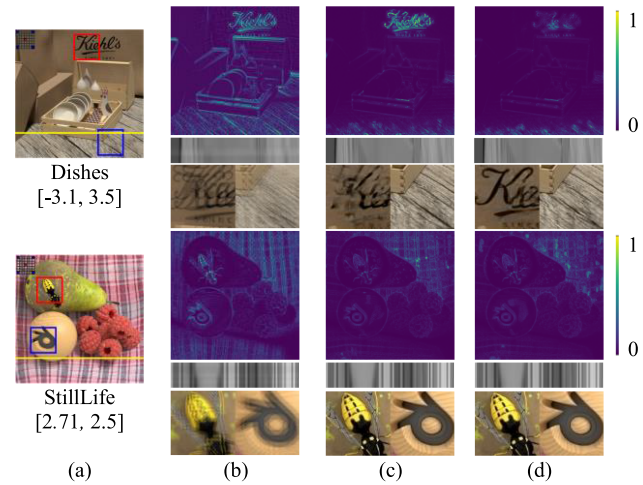


FIGURE 14. Angular consistency analysis on synthetic scenes Dishes (HCI-new [45]) – Top, and StillLife (HCI-old [46]) – Bottom: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) Proposed Method.

GT, demonstrating improved angular coherence and parallax preservation.

F. DEPTH ESTIMATION

Depth estimation serves as a critical benchmark for evaluating the geometric fidelity of synthesized LFs. To assess the effectiveness of the proposed method from a geometric perspective, disparity maps are estimated from the synthesized dense LFs using OACC-Net [48], and are quantitatively evaluated across all synthetic scenes in the HCI benchmark datasets [45], [46]. Two widely adopted metrics are used for accuracy assessment: Root Mean Square Error (RMSE), which quantifies pixel-wise disparity fidelity, and Bad Pixel Rate (BPR) with a disparity error threshold of 0.07 [45], which reflects structural accuracy

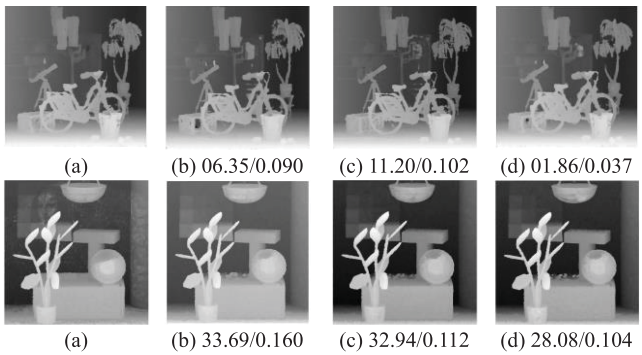


FIGURE 15. Depth estimation results (RMSE/BPR) on synthetic scenes Bicycle (HCI-new [45]) – Top, and MonasRoom (HCI-old [46]) – Bottom: (a) GT; (b) Deformable-LFVS [15]; (c) Distg-ASR [21]; (d) Proposed Method.

by measuring the proportion of pixels exceeding the threshold. As shown in Fig. 15, qualitative and quantitative results are reported on two representative scenes: Bicycle (HCI-new, [45]) and MonasRoom (HCI-old, [46]). The proposed method yields disparity maps with better spatial fidelity, characterized by sharper object boundaries and smoother transitions in homogeneous regions. These observations are further supported by lower RMSE and BPR scores, demonstrating improved geometric accuracy.

These results validate the effectiveness of the proposed method in maintaining angular consistency and leveraging redundant SAC factors, which are essential for synthesizing dense LFs and enabling reliable depth estimation.

G. ABLATION STUDY

We conducted an ablation study to assess the contribution of each component in the proposed method and to validate our design choices by comparing against alternative convolution types, spatial and angular channel attention mechanisms. Table 8 highlights the performance of each variant on synthetic and real-world datasets.

- **Variant I** – This variant employs a 2D DCNN to extract multi-scale features from the SAI representation in the CLFVS module, providing geometric cues to guide view warping at correct angular positions. These features are further refined by the CFR module using a DRCA network, effectively capturing both local and global context for synthesizing dense LFs. However, this variant relies solely on the SAI representation, limiting its ability to incorporate complementary spatial–angular cues available in the MI representation.

TABLE 8. Ablation Results on Synthetic and Real-World Datasets

Variant	Residual groups	Residual blocks	LF representations	FF-Net	CFR module	Model parameters	Synthetic datasets	Real-world datasets
							PSNR/SSIM	PSNR/SSIM
I	4	5	SAI	✗	✓	2.04 M	37.98/0.948	40.70/0.976
II	4	5	MI	✗	✓	1.33 M	36.83/0.932	40.12/0.960
III	4	5	SAI + MI	✓	✗	1.40 M	37.44/0.944	40.56/0.962
IV	4	4	SAI + MI	✓	✓	2.12 M	38.19/0.951	40.82/0.977
V	3	5	SAI + MI	✓	✓	2.00 M	38.04/0.948	40.62/0.976
VI	4	5	SAI	✗	✓	1.87 M	37.56/0.945	40.58/0.973
VII	4	5	MI	✗	✓	1.80 M	36.68/0.932	40.10/0.960
VIII	-	-	SAI + MI	✓	✓	1.62 M	37.58/0.951	40.58/0.965

- **Variant II** – In this variant, 2D DCNN is used to extract hierarchical features from the MI representation, followed by refinement using a DRCA network in CFR module. However, the absence of a multi-scale expansion–contraction strategy limited the network's ability to capture structural cues in the SAI representation, resulting in comparatively reduced performance.
- **Variant III** – This variant uses dual LF representations (i.e., SAI and MI) in the CLFVS module. DCNNs are employed to extract features from each representation, which are then fused using a FF-Net to synthesize dense LFs. However, the exclusion of the CFR module limits the model's ability to enhance angular consistency and fine structural details, despite the benefits of using complementary representations.
- **Variants IV and V** – These variants validate the impact of adjusting the number of residual blocks and groups for feature aggregation and refinement. Variant IV reduces the residual blocks from 5 to 4, while Variant V reduces the residual groups from 4 to 3, allowing us to assess the trade-off between network depth and refinement capacity.
- **Variant VI** – To evaluate the role of deformable convolution in the SPFE network, in Variant VI we replace its 2D deformable convolutions with dilated standard 2D convolutions (i.e., 7 multi-scale convolution layers with a dilation rate of 2) for feature extraction from the SAI representation. Despite the enlarged receptive field provided by dilation, this variant attains lower reconstruction quality because the fixed sampling grid cannot adapt to the disparity-induced geometric shifts across angular views, which the learnable offsets of deformable convolution handle effectively.
- **Variant VII** – To evaluate the role of deformable convolution in the AFE network, in Variant VII we replace the 2D deformable convolutions with standard 3D convolutions for feature extraction from the MI representation. This variant attains slightly lower reconstruction quality despite using more parameters, because the fixed sampling grid of 3D convolution cannot adapt to the disparity-induced geometric shifts across angular views, whereas the learnable offsets of deformable convolution more effectively capture SAC.
- **Variant VIII** – To verify the effectiveness of the proposed DRCA design, in Variant VIII we replace the DRCA with a Spatial-Angular Channel Attention (SACA) network in the CFR module. SACA applies channel attention separately along the spatial and angular dimensions and fuses the two branches, without the residual-group structure. This variant attains lower reconstruction quality because its shallow single-stage design cannot recover fine details in a single attention pass and lacks dense inter-group connections for multi-level feature reuse.

The ablation results confirm that each variant offers a unique contribution. Notably, the integration of dual LF representations with DCNNs for feature extraction and DRCA refinement is essential for achieving high-fidelity LFVS performance.

V. CONCLUSION

In this paper, we proposed a dual representation-based hybrid LFVS method, which integrates deformable convolutional and DRCA networks for synthesizing dense LFs with high fidelity, using two modules: CLFVS and CFR. The CLFVS module explores the LF structural characteristics to synthesize coarse dense LFs, while the CFR module refines them by capturing local and global dependencies. Although the proposed method achieves state-of-the-art results by using dual LF representations, the interplay between these domains necessitates a sophisticated alignment (reshaping) process. Future work will explore the transition toward a unified LF representation to further streamline the synthesis pipeline and potentially yield even higher levels of angular consistency and structural fidelity.

REFERENCES

[1] Q. Zhang, H. Li, X. Wang, and Q. Wang, "3D scene reconstruction with an un-calibrated light field camera," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3006–3026, 2021.
[2] Y. Ding, Z. Chen, Y. Ji, J. Yu, and J. Ye, "Light field-based underwater 3D reconstruction via angular re-sampling," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 881–893, 2023.
[3] 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.
[4] J. Peng, Z. Xiong, Y. Wang, Y. Zhang, and D. Liu, "Zero-shot depth estimation from light field using a convolutional neural network," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 682–696, 2020.
[5] W. Yan, X. Zhang, H. Chen, C. Ling, and D. Wang, "Light field depth estimation based on channel attention and edge guidance," in *Proc. Chin. Autom. Congr.*, 2022, pp. 2595–2600.
[6] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, "Selective light field refocusing for camera arrays using bokeh rendering and superresolution," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 204–208, Jan. 2019.

[7] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light field saliency detection with dual local graph learning and reciprocative guidance," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 4692–4701.
[8] M. F. Hamad, C. Conti, P. Nunes, and L. D. Soares, "Hyperpixels: Flexible 4D over-segmentation for dense and sparse light fields," *IEEE Trans. Image Process.*, vol. 32, pp. 3790–3805, 2023.
[9] B. Wilburn et al., "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.
[10] V. Vaish and A. Adams, "The (NEW) stanford light field archive," 2008. [Online]. Available: http://lightfield.stanford.edu/
[11] Lytro, "LightField forum," 2024. [Online]. Available: http://lightfield-forum.com/lytro/
[12] Raytrix, "Raytrix - 3D light field camera technology," 2024. [Online]. Available: https://raytrix.de/
[13] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 11141–11148.
[14] D. Liu, Z. Tong, Y. Huang, Y. Chen, Y. Zuo, and Y. Fang, "Geometry-assisted multi-representation view reconstruction network for light field image angular super-resolution," *Knowl.-Based Syst.*, vol. 267, 2023, Art. no. 110390.
[15] M. Zubair, P. Nunes, C. Conti, and L. D. Soares, "Light field view synthesis using deformable convolutional neural networks," in *Proc. Picture Coding Symp.*, 2024, pp. 11–15.
[16] G. Liu, H. Yue, and J. Yang, "Efficient light field image super-resolution via progressive disentangling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6277–6286.
[17] Z. Cheng, Y. Liu, and Z. Xiong, "Spatial-angular versatile convolution for light field reconstruction," *IEEE Trans. Comput. Imag.*, vol. 8, pp. 1131–1144, 2022.
[18] M. Zubair, P. Nunes, C. Conti, and L. D. Soares, "Swinscale-LFVS: Parallel feature integration for light field view synthesis," in *Proc. IEEE Int. Conf. Image Process.*, 2025, pp. 1942–1947.
[19] G. Liu, H. Yue, J. Wu, and J. Yang, "Efficient light field angular super-resolution with sub-aperture feature learning and macro-pixel upsampling," *IEEE Trans. Multimedia*, vol. 25, pp. 6588–6600, 2023.
[20] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, "Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1819–1836, Apr. 2022.
[21] Y. Wang et al., "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 2023.
[22] W. Yan, X. Zhang, and H. Chen, "Occlusion-aware unsupervised light field depth estimation based on multi-scale GANs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6318–6333, Jul. 2024.
[23] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
[24] D. W. Jang and R. H. Park, "Densenet with deep residual channel-attention blocks for single image super resolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1795–1803.
[25] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
[26] J. Hu, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[27] S. Yun, J. Jang, and J. Paik, "Geometry-aware light field angular super resolution using multiple receptive field network," in *Proc. Int. Conf. Electron. Inf., Commun.*, 2022, pp. 1–3.
[28] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2552–2561.
[29] M. S. K. Gul, M. U. Mukati, M. Batz, S. Forchhammer, and J. Keinert, "Light-field view synthesis using a convolutional block attention module," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 3398–3402.
[30] S. Woo, J. Park, J. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[31] X. Liu, M. Wang, A. Wang, X. Hua, and S. Liu, "Depth-guided learning light field angular super-resolution with edge-aware inpainting," *Vis. Comput.*, vol. 38, no. 8, pp. 2839–2851, 2022.
[32] K. Ko, Y. J. Koh, S. Chang, and C. S. Kim, "Light field super-resolution via adaptive feature remixing," *IEEE Trans. Image Process.*, vol. 30, pp. 4114–4128, 2021.
[33] Y. Chen, X. Huang, P. An, and Q. Wu, "Enhanced light field reconstruction by combining disparity and texture information in PSVs via disparity-guided fusion," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 665–677, 2023.
[34] D. Cai, Y. Chen, X. Huang, and P. An, "Disparity enhancement-based light field angular super-resolution," *IEEE Signal Process. Lett.*, vol. 32, pp. 81–85, 2025.
[35] Y. Yoon, H. G. Jeon, D. Yoo, J. Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 57–65.
[36] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 138–154.
[37] L. Liu, Z. Nian, and C. Jung, "DCM-CNN: Densely connected multiloss convolutional neural networks for light field view synthesis," *IEEE Access*, vol. 8, pp. 78542–78552, 2020.
[38] 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.
[39] 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, 2021, Art. no. 5014314.
[40] Y. Mao, X. Huang, Y. Chen, D. Liu, P. An, and S. Lee, "Learning a domain-specialized network for light field spatial-angular super-resolution," *IEEE Trans. Vis. Comput. Graph.*, vol. 32, no. 2, pp. 2127–2140, Feb. 2026.
[41] D. Liu, S. Li, Y. Mao, X. Zhou, Z. Xiao, and C. Shan, "Learning implicit and detail-enhanced network for light field image spatial-angular super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 36, no. 2, pp. 1544–1557, Feb. 2026.
[42] Z. Liang et al., "Diving into epipolar transformers for light field super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 48, no. 7, pp. 8726–8743, Jul. 2026.
[43] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[44] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[45] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2017, pp. 19–34.
[46] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. 18th Int. Work. Vis., Model. Vis.*, 2013, pp. 225–226.
[47] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein, "Stanford lytro light field archive," 2016. [Online]. Available: http://lightfields.stanford.edu/LF2016.html
[48] Y. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, "Occlusion-aware cost constructor for light field depth estimation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19777–19786.



MUHAMMAD ZUBAIR (Member, IEEE) received the M.S. degree in computer science from COMSATS University, Islamabad, Pakistan, in 2020. He is currently working toward the fully-funded Ph.D. degree in information science and technology with Instituto Universitário de Lisboa, Lisbon, Portugal. His research interests include light field view synthesis, neural rendering, multi-view image generation, and depth estimation.



PAULO NUNES (Member, IEEE) received the degree, M.Sc., and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico (IST), Universidade de Lisboa, Lisbon, Portugal, in 1992, 1996, and 2007, respectively. He is currently a Senior Researcher with Multimedia Signal Processing Group, Instituto de Telecomunicações, Lisbon. He is also an Associate Professor with the Department of Information Science and Technology, Instituto Universitário de Lisboa, Lisbon. He has coordinated and participated in various national and international (EU) funded projects and acted as a Project Evaluator for the European Commission. He has contributed more than 80 papers to international journals and conferences in his research interests which include 2D/3D image and video processing and coding, namely light field image and video processing and coding. He is a member of the EURASIP Technical Area Committee for Visual Information Processing, acts often as a Reviewer of various ACM, EURASIP/Elsevier, IEEE, IET, SPIE, and Springer conferences and journals, and member of the program and organizing committees of various international conferences.



CAROLINE CONTI (Member, IEEE) received the B.Sc. degree in electrical engineering from Universidade de São Paulo (USP), São Paulo, Brazil, in 2010, and the Ph.D. degree in information science and technology from Instituto Universitário de Lisboa, Portugal, in 2017. She is currently a Researcher with Multimedia Signal Processing Group, Instituto de Telecomunicações and Associate Professor with the Department of Information Science and Technology, Instituto Universitário de Lisboa, Lisbon, Portugal. She has contributed more than 30 papers to international journals and conferences in her research interests which include processing and coding of visual media, including immersive visual technologies, such as light fields, and other emerging sensing modalities, such as event cameras. She has served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and Guest Editor of *Signal Processing: Image Communication* (Elsevier). She actively participates as a Reviewer of various IEEE and EURASIP journals and conferences.



LUÍS DUCLA SOARES (Senior Member, IEEE) received the Licenciatura and Ph.D. degrees in electrical and computer engineering from Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, in 1996 and 2004, respectively. He is currently a Senior Researcher with Multimedia Signal Processing Group, Instituto de Telecomunicações, Lisbon, Portugal. He is also a Full Professor with the Department of Information Science and Technology, Instituto Universitário de Lisboa, Lisbon. He has contributed more than 90 papers to international journals and conferences his research interests include the coding and processing of visual information modalities, including light fields. He has participated in the development of the MPEG-4 Visual standard and in several national and international projects. He is a member of the editorial board of *EURASIP Signal Processing* (Elsevier) journal. He acts as a Reviewer of several IEEE, IET, and EURASIP journals and conferences.