IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 July 2024, accepted 12 August 2024, date of publication 20 August 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3446592

RESEARCH ARTICLE

Light Field Reconstruction With Dual Features Extraction and Macro-Pixel Upsampling

AHMED SALEM[1,2], EBRAHEM ELKADY[3,4], HATEM IBRAHEM[5], JAE-WON SUH[3], AND HYUN-SOO KANG[1], (Member, IEEE)

[1]School of Information and Communication Engineering, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, South Korea
[2]Electrical Engineering Department, Faculty of Engineering, Assiut University, Asyut 71526, Egypt
[3]School of Electronics Engineering, College of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si 28644, South Korea
[4]Information Technology Department, Faculty of Computers and Information, Assiut University, Asyut 71526, Egypt
[5]Department of Computer Science, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada

Corresponding authors: Ahmed Salem (ahmeddiefy@cbnu.ac.kr) and Hyun-Soo Kang (hskang@cbnu.ac.kr)

This work was supported in part by the Chungbuk National University BK21 Program, in 2023; the National Research Foundation of Korea grant funded by Korean Government (Ministry of Science and ICT) (MSIT) under Grant 2022R1A5A8026986; and in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant 2023R1A2C1006944.

ABSTRACT Dense multi-view image reconstruction has been a focal point of research for an extended period, with recent surges in interest. The utilization of multi-view images offers solutions to numerous challenges and amplifies the effectiveness of various applications including 3D reconstruction, de-occlusion, depth sensing, saliency detection, and identifying salient objects. This paper introduces an approach to reconstructing high-density light field (LF) images, addressing the inherent challenge of balancing angular and spatial resolution caused by limited sensor resolution. We introduce an innovative approach to reconstructing LF images through a CNN-based network that combines spatial and epipolar features in both initial and deep feature extraction phases. Our network utilizes angular information during upsampling and employs dual feature extraction to effectively analyze horizontal and vertical epipolar data. Weight sharing within the CNN block between horizontal and vertically transposed stacks enhances quality while preserving model compactness. The outcomes of experiments carried out on real-world and synthetic datasets demonstrate the effectiveness of our method, showcasing its superior performance in both inference speed and reconstruction quality when compared to state-of-the-art (SOTA) techniques.

INDEX TERMS Light field reconstruction, based view synthesis, angular super-resolution, convolutional neural network.

I. INTRODUCTION

Unlike traditional cameras that capture 2D images, light field (LF) cameras capture multi-view images, preserving the directions of incoming light rays and 3D geometry information efficiently [1], [2]. This additional angular information enhances 3D representation quality, benefiting applications such as 3D reconstruction and display [3], object segmentation [4], virtual reality applications [5], image post-refocus [6], [7], depth inference [8], [9], de-occlusion [10], reflectance estimation [11], [12].

Different methodologies exist for acquiring LF images [13]. Early approaches used multi-camera system [14] for single-shot capture and time-sequential systems [15] with a computer-controlled gantry and single camera for multiple shots, but these are complex and expensive. Recently, plenoptic cameras (e.g., RayTrix [5]) have advanced LF imaging by using a micro-lens array between the sensor and primary lens to capture additional angular information. Plenoptic cameras capture densely sampled LF images in one shot, but face a trade-off between angular and spatial resolutions due to limited sensor resolution. Since the product of these resolutions cannot exceed the sensor resolution, achieving high spatial resolution in densely sampled LF

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

images is costly. This trade-off presents challenges for practical applications of LF imaging.

To address this problem, some researchers investigate LF spatial super-resolution [16], aims to generate high-resolution images from low-resolution inputs, while others explore LF angular super-resolution [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], enhancing the angular resolution by reconstructing densely sampled views from sparse input views.

Learning-based methods for densely sampled LF construction are categorized into depth-based [17], [18], [19], [20], [21], [22], [23] and non-depth-based [24], [25], [26], [27], [28], [29], [30], [31], [32]. Depth-based methods estimate disparity maps to synthesize novel views, performing well in regions with large disparities but struggling with small disparities and textureless areas. Non-depth-based methods use local LF information without explicit depth, excelling in small disparity regions but struggling with large disparities. Success relies on effectively exploiting intrinsic LF relationships. Various representations, such as sub-aperture images (SAIs), epipolar-plane images (EPIs), and macro-pixel images (MacPIs), contribute to diverse approaches as illustrated in Fig. 1.

LF sampling's grid-like structure necessitates a deep exploration of angular information across views. Key considerations include distinct epipolar information within the grid, revealing parallax relationships, and complementary angular domain information that enhances sub-pixel details. Therefore, we propose a CNN with dual-feature extraction and macro-pixel upsampling. The initial part of our network delves into spatio-epipolar information while preserving the grid-like structure. The subsequent part focuses on exploring spatio-angular information for angular upsampling. We encapsulate our contributions in the following manner.

1) Unlike previous methods that use the same LF representation for feature extraction and upsampling, we perform feature extraction on SAIs and upsampling on the MacPI representation.

2) Our dual feature extraction mechanism captures both horizontal and vertical epipolar information within a single CNN block. By reorganizing view stacks to emphasize spatial information and sharing weights between horizontally and vertically transposed stacks, we enhance quality and model compactness.

3) Experiments on real-world and synthetic datasets demonstrate that our approach outperforms SOTA methods in reconstruction quality across most datasets.

The subsequent sections are organized as follows: Section II reviews relevant literature, Section III outlines the proposed approach, Section IV details experiments and ablation studies, and Section V concludes the paper.

II. RELATED WORK

Researchers have employed various methods to reconstruct a densely sampled LF from a limited set of input views, categorized into depth-based and non-depth-based methods. Below, we briefly discuss these approaches.

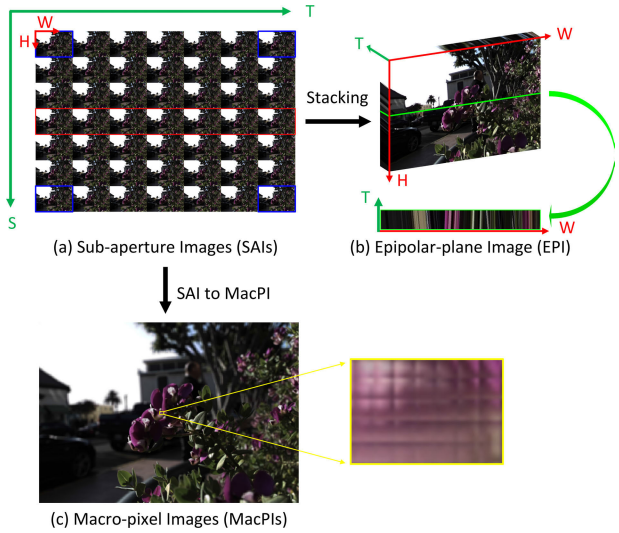


FIGURE 1. Different LF subspaces. (a) Sub-aperture Images (SAIs). (b) Epipolar-plane Image (EPI). (c) Macro-pixel Images (MacPIs).

A. DEPTH-BASED LF RECONSTRUCTION

Wanner and Goldluecke [17] analyzed 4D LF reconstruction using EPI to estimate the disparity map, which is then used to synthesize novel views. Zhang et al. [18] employed a phase-based approach to reconstruct LF from a micro-baseline stereo pair, using disparity-assisted phase-based synthesis. However, this method is susceptible to artifacts in occluded regions. Expanding upon the patch-based synthesis technique, Zhang et al. [19] extended patch-based synthesis by representing LF views as overlapping layers with varying depths.

Recent deep learning approaches use CNNs for LF angular SR. Typically, these systems comprise two sub-networks: one for depth estimation and another for view refinement. The whole process can be described as:

1) Depth estimation:
$$D(x,u) = f_d\left(L(x,u')\right)$$
Here, $D(x,u)$ represents the depth map estimated using the function f_d from the input LF views $L(x,u')$.
2) Warping based on estimated depth:
$$W(x,u,u') = L\left(x + D(x,u)(u-u'),u'\right)$$
Where $W(x,u,u')$ denotes a novel view at angular position u produced by warping an input view at u'.
3) View refinement:
$$\hat{L}(x,u) = W(x,u,u') + f_r(W(x,u,u'))$$
The final view $\hat{L}(x,u)$ is obtained by adding the wrapped view and the refined view using the function f_r.

Following this paradigm, Kalantari et al. [20] proposed a learning-based method that incorporates disparity and color estimation to reconstruct novel views. However, the quality of these reconstructed views degrades, especially in occluded and textureless regions where depth estimation is particularly challenging. Salem et al. [21] introduced a dual-disparity vector technique within a two-stage neural network. This approach enhances the reconstruction quality of Kalantari's

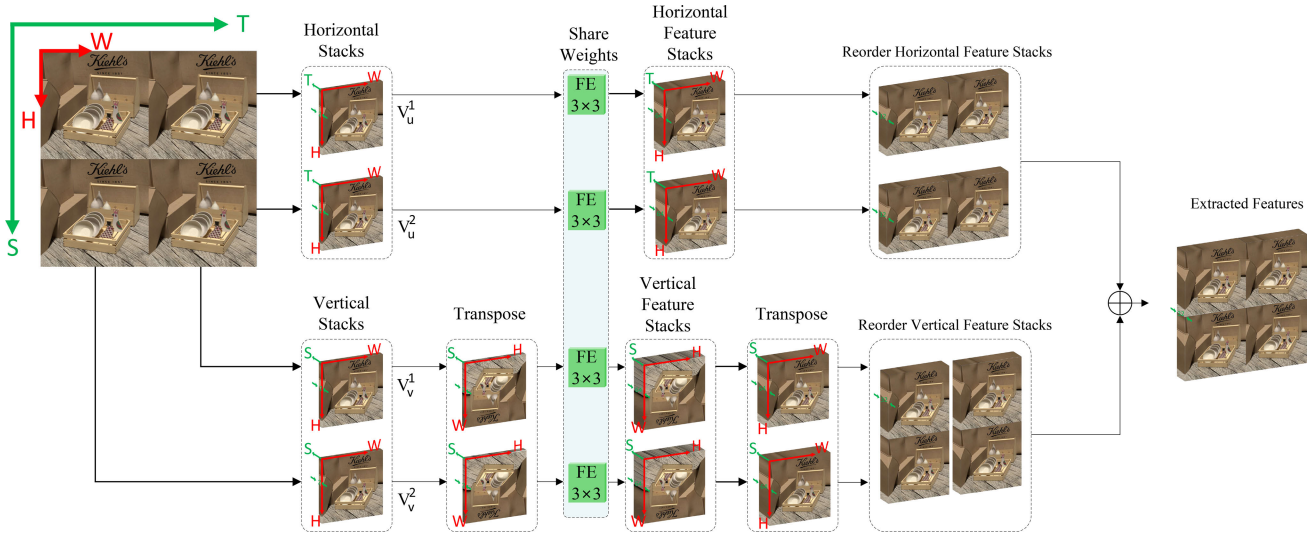


FIGURE 2. Dual features extraction mechanism.

method and speeds up the process using a predefined discrete cosine filter. Jin et al. [22] developed a depth estimator with a large receptive field and a refinement module to blend novel views. They later introduced a versatile reconstruction network that estimates confidence and disparity maps for view synthesis and image-warping [23].

Depth-based methods rely heavily on accurate depth estimation, which is challenging in occluded regions, leading to ghosting artifacts and photo consistency issues between synthesized views.

B. NON-DEPTH-BASED LF RECONSTRUCTION

Non-depth-based methods bypass explicit depth estimation, using traditional and CNN-based approaches. Mitra and Veeraraghavan [24] introduced a patch-based reconstruction technique, representing patches as a Gaussian Mixture Model. Shi et al. [25] exploited LF sparsity in the Fourier domain, using boundary and diagonal viewpoints for LF reconstruction, though this method requires a specific LF capture pattern. Vagharshakyan et al. [26] applied a sparse representation of EPI in the shearlet transform domain, using an iterative regularization algorithm to reconstruct semi-transparent scenes.

CNN-based methods have also been developed for LF reconstruction without depth information.The process can be described as follows:

$$\hat{L}(x,u) = f(L(x,u'),\theta)$$

where f represents the function that reconstructs $\hat{L}(x,u)$ from input LF views $L(x,u')$ and θ represents the network parameters learned during training.

Following this paradigm, Yoon et al. [27] generated new views from two adjacent views using a deep network. Zhu et al. [28] introduced a CNN-LSTM network to enhance LF spatial and angular dimensions. Wu et al. [29] addressed spatial-angular information asymmetry with a blur-restoration-deblur strategy in the EPI domain, though this method is challenged by significant disparities. Later, they improved this approach by merging sheared EPIs [30]. Wang et al. [31] proposed a disentangling mechanism to separate spatial, angular, and epipolar information for LF angular SR. Liu et al. [32] achieved ASR using spatial-angular feature extraction and MacPI upsampling, offering an effective structure for restoring angular resolution. Salem et al. [33] simplified the LF reconstruction problem by converting the 4D LF into a 2D MacPI image. This transformation effectively reduces the complexity from a 4D to a 2D domain.

III. METHODOLOGY

A. PROBLEM FORMULATION

LF is represented as a 4D function, denoted as $L(u,v,x,y) \in \mathbb{R}^{S\times T\times H\times W}$, where $S\times T$ represents the angular resolution, and $H\times W$ represents the spatial resolution, as shown in Fig. 1. Let $I_{LR} \in \mathbb{R}^{S\times T\times H\times W}$ denote a sparsely sampled (i.e., low angular resolution) LF image, comprising $S\times T$ sub-aperture images, each with a spatial resolution of $H\times W$. Given I_{LR}, the objective of this research is to reconstruct a high angular resolution equivalent, $I_{HR} \in \mathbb{R}^{\alpha S\times \alpha T\times H\times W}$, where α is the angular scale factor, with $\alpha > 1$. After the reconstruction process, the number of novel views synthesized to increase the angular resolution is $(\alpha U\times \alpha V - U\times V)$ views. In this work, we convert LF images from RGB to the YCbCr color space and operate solely on the Y channel. The Cb and Cr channels are first upsampled using bicubic interpolation and then integrated with the Y channel to generate the final I_{HR} LF image.

B. DUAL FEATURES EXTRACTION

To build an effective feature extraction stage, it's important to understand the LF information, which includes angular, spatial, and epipolar aspects. Our network uses spatial and

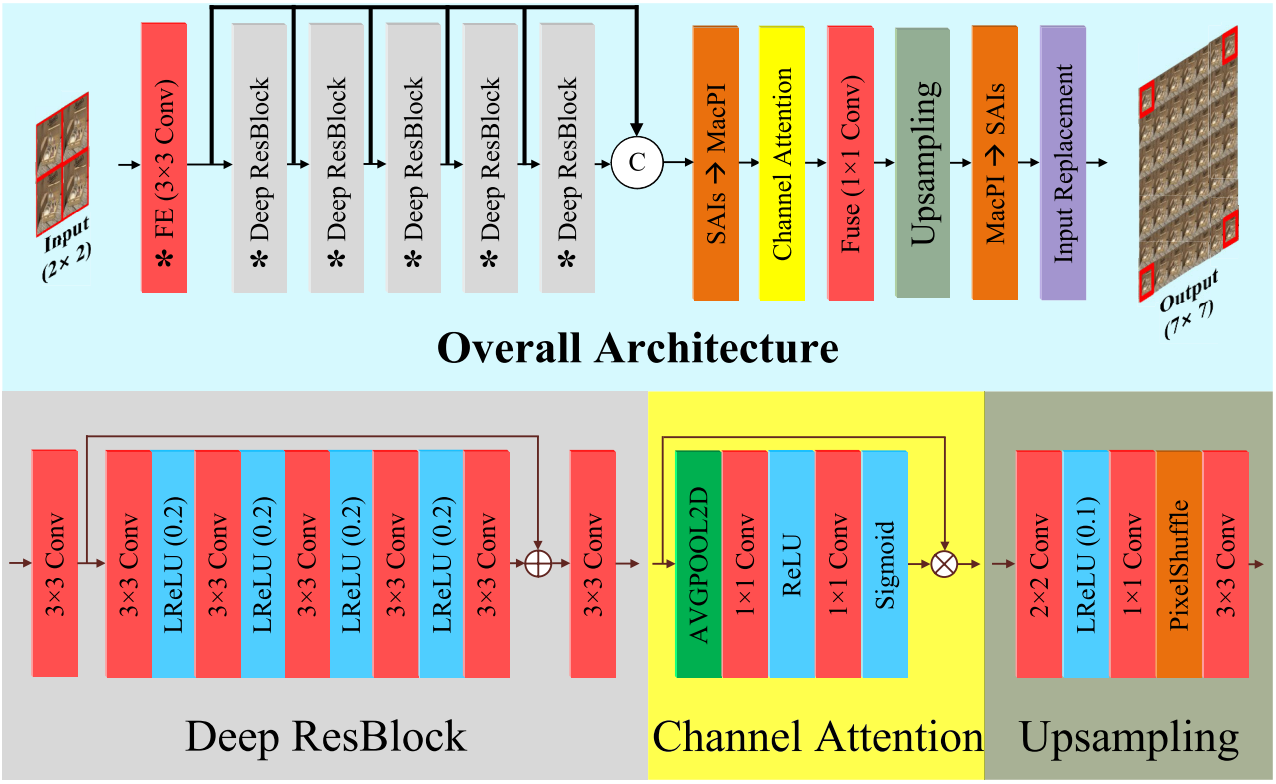


FIGURE 3. An overview of the proposed network framework. The (*) symbol indicates the application of dual feature extraction in the block.

epipolar information during feature extraction and uses angular information during upsampling.

In our framework, instead of separating the LF information, we create view image stacks to utilize the combined epipolar and spatial information. By stacking view images along the horizontal or vertical angular direction, we obtain view image stacks $V_u^i = L(i, :, :, :) \in \mathbb{R}^{V\times H\times W}$ and $V_v^i = L(:, i, :, :) \in \mathbb{R}^{U\times H\times W}$. As illustrated in Fig. 2, slicing the view image V_v^i in the w spatial direction results in 2D slices $I_w^v \in \mathbb{R}^{U\times H}$, which are EPIs. To enhance efficiency, we share parameters during the extraction process by transposing the vertical stacks before extraction, making them behave like horizontal stacks, and then transposing them back after extraction. The feature extraction process can be described as follows:

1) Horizontal feature extraction:
$$I_{\mathrm{HF}} = \mathrm{FE}(I_{\mathrm{LR}}(i, :, :, :)), \quad i \in \{1, \ldots, S\}$$
2) Vertical feature extraction:
$$I_{\mathrm{VF}} = \mathrm{FE}(I_{\mathrm{LR}}(:, j, :, :)), \quad j \in \{1, \ldots, T\}$$
3) Combining features:
$$I_F = I_{\mathrm{HF}} + I_{\mathrm{VF}}, \quad I_F \in \mathbb{R}^{S\times T\times H\times W\times C}$$

Here, combining the features ensures that the network leverages information from both horizontal and vertical directions. This comprehensive approach captures more details and relationships between different views, leading to improved accuracy and balanced representation. FE is the initial feature extraction layer using a 2D convolution with a 3×3 kernel, as shown in Fig. 2.

We apply the same dual feature extraction mechanism to the Deep-ResBlocks. The input to the first Deep-ResBlock, $I_F \in \mathbb{R}^{S\times T\times H\times W\times C}$, is arranged into horizontal and vertical feature stacks, represented as $I_F(i, :, :, :) \in \mathbb{R}^{TC\times H\times W}$ and $I_F(:, j, :, :, :) \in \mathbb{R}^{SC\times H\times W}$, respectively. The process is as follows: DFE_1 represents the first Deep-ResBlock, as shown in Fig. 3:

1) Horizontal deep feature extraction:
$$I_{\mathrm{DHF}} = \mathrm{DFE}_1(I_{\mathrm{F}}(i, :, :, :)), \quad i \in \{1, \ldots, S\}$$
2) Vertical feature extraction:
$$I_{\mathrm{DVF}} = \mathrm{DFE}_1(I_{\mathrm{F}}(:, j, :, :)), \quad j \in \{1, \ldots, T\}$$
3) Combining features:
$$I_{\mathrm{DF}} = I_{\mathrm{DHF}} + I_{\mathrm{DVF}}, \quad I_{\mathrm{DF}} \in \mathbb{R}^{S\times T\times H\times W\times C}$$

In previous work DistgASR [31], the 4D LF was organized as an H × W array of macro-pixels (MacPIs) with three separate feature extractors designed for spatial, angular, and epipolar information. While this approach reduced processing complexity, it ignored the coupling between spatial, angular, and epipolar information. In our approach, we process spatial and epipolar information together to utilize the coupled information more effectively. While in [32], the angular dimensions of the extracted features were stacked and processed by a 3D-UNet, resulting in features with dimensions $(UV) \times C \times H \times W$. This approach not only ignores the epipolar information but also makes it difficult for the network

to decouple spatial and angular information simultaneously. In our approach, we divide the processing into two steps. First, we process horizontal and angular information separately, and then we combine them.

C. NETWORK STRUCTURE

Our network, shown in Fig. 3, is designed to take a sparsely sampled LF, denoted as $I_{LR} \in \mathbb{R}^{S\times T\times H\times W}$ as input and reconstruct a densely sampled LF represented as: $I_{HR} \in \mathbb{R}^{\alpha S\times \alpha T\times H\times W}$. The network's workflow can be divided into several stages: 1) Initial Feature Extraction (FE): The network begins with an Initial Feature Extraction layer that processes the input I_{LR} to produce feature maps $I_{FE} \in \mathbb{R}^{S\times T\times H\times W\times C}$. 2) Deep Feature Extraction (DFE): Following the initial extraction, the network uses several Deep-ResBlocks to further explore and refine the features. These blocks effectively capture spatial and epipolar information from the input. Both the Initial Feature Extraction layer and the Deep-ResBlocks use a dual feature extraction, indicated by the asterisk (*) in Fig. 3. This means they extract features in both horizontal and vertical directions, (as shown in Fig. 2). 3) Combining Features: The features extracted from the FE layer and each DFE layer are concatenated together and reshaped into a MacPIs for further processing. 4) Channel Attention and Fusion: The combined features pass through a Channel Attention Block (CA), which helps the network focus on the most important features. A 1×1 convolution layer then merges this information, preparing it for the next stage. 5) Angular Upsampling: The upsampling block increases the resolution of the combined features, resulting in a higher-resolution image array resulting in $I_{HR} \in \mathbb{R}^{\alpha S\times \alpha T\times H\times W}$, which is subsequently reshaped back into the SAI format $I_{HR} \in \mathbb{R}^{\alpha S\times \alpha T\times H\times W}$, now with more densely sampled data. 6) Final Output: The network performs an input replacement step to finalize the high-resolution output I_{HR}.

D. NETWORK BUILDING BLOCKS

In this subsection, we offer a comprehensive explanation of the key elements that constitute our network.

1) INITIAL FEATURE EXTRACTION (FE)

Our feature extraction process starts with the initial feature extraction phase. It takes $I_{LR} \in \mathbb{R}^{S\times T\times H\times W}$ as input and produces I_{FE} as output:

$$I_{FE} = H_{FE}(I_{LR}), \quad I_{FE} \in \mathbb{R}^{S\times T\times H\times W\times C}$$

Here, H_{FE} represents the 3×3 convolutional layer that extracts the initial features using the dual feature extraction mechanism as outlined earlier. This layer captures essential information from the input data.

2) DEEP FEATURE EXTRACTION (DFE)

In the second phase, the network employs N Deep-ResBlocks to enhance the feature details. These blocks densely connect spatial and epipolar information from the input. The process is as follows:

$$I^n_{DF} = H^n_{DFE}(I^{(n-1)}_{DF}),\ I^n_{DF} \in \mathbb{R}^{S\times T\times H\times W\times C},\ n \in \{1,\ldots,N\}$$

Each Deep-ResBlock H^n_{DFE} refines the features extracted from the previous block. Both the initial feature extraction layer and the Deep-ResBlocks use the dual feature extraction mechanism, capturing information in both horizontal and vertical directions.

The Deep-ResBlock is constructed using 2D convolutions with a kernel size of 3×3 and LeakyReLU activations with a negative slope of 0.1, as illustrated in Fig. 3.

3) FEATURE CONCATENATION AND RESHAPING

To strengthen the network's ability to learn hierarchical representations, we concatenate features from both the initial and deep feature extraction stages:

$$I_{CAT} = H_{CAT}\left(I_{FE}, I^1_{DFE}, \ldots, I^N_{DFE}\right), \quad I_{CAT} \in \mathbb{R}^{S\times T\times H\times W\times (N+1)C}$$

Here, H_{CAT} denotes the concatenation process. These concatenated features are reshaped into MacPI representation, organizing spatial and angular information jointly. Following similar approaches as [31] and [32], we use a PixelShuffle layer to convert the 4D LF features into the MacPIs representation:

$$I_{\mathrm{CAT}} \in \mathbb{R}^{SH\times TW\times (N+1)C}$$

4) FEATURE ATTENTION AND FUSION

Following [37], we incorporate a channel attention mechanism to selectively enhance informative features while suppressing less relevant ones through adaptive weighting.

$$I_{CA} = f(W_U \cdot \delta(W_D \cdot \mathrm{AvgPool}(I_{CAT}))) \times I_{CAT}$$

Here, f and δ are sigmoid and ReLU activation functions, respectively, and W_U and W_D are 1×1 convolution kernels. This mechanism focuses the network on the most relevant features. A 1×1 convolution layer then fuses these features:

$$I_{Fuse} \in \mathbb{R}^{SH\times TW\times C}$$

5) ANGULAR UPSAMPLING

We adopt the downsample-upsample approach to perform angular upsampling [31]. The process involves several steps, detailed below:

1) Downsampling:
   A convolution layer with a kernel size of $S\times T$ is used to downsample the fused features from $\mathbb{R}^{SH\times TW\times C}$ to $\mathbb{R}^{H\times W\times C}$.
2) Channel Expansion:
   A 1×1 convolutional layer is applied to expand the channels to $\mathbb{R}^{H\times W\times \alpha^2 STC}$, where α is the scaling factor.
3) Pixel Shuffling:
   A pixel-shuffle operator is employed to rearrange the expanded features. This step is followed by a 3×3 convolutional layer to generate a high-resolution with dimensions $\mathbb{R}^{SH\times TW\times 1}$.
4) Reshaping and Replacement:
   The resulting MacPI image is reshaped back into the SAI format. The corner images in the output are replaced with the original input images to maintain consistency.

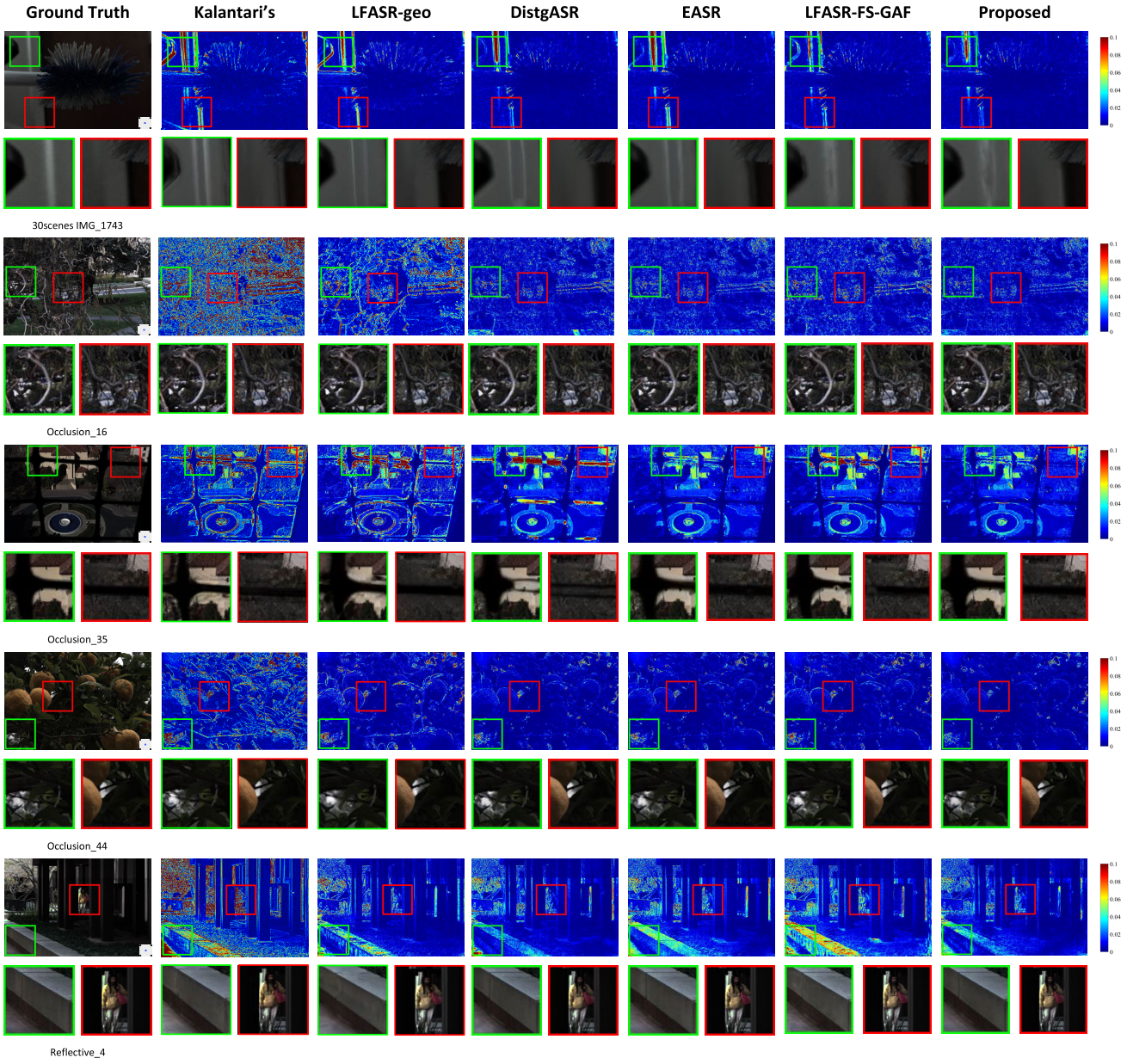


FIGURE 4. Visual comparison between our proposed approach and SOTA methods for 2 × 2 → 7 × 7 ASR on real-world datasets.

IV. EXPERIMENTS

This section starts with an exploration of the datasets used and delves into the implementation details of our model. Following this, we perform a comprehensive comparative analysis between our proposed model and SOTA methods, highlighting key differences and advantages. Next, we present detailed ablation studies to illustrate the impact of the introduced modifications on performance, providing insights into the effectiveness of each component.

A. IMPLEMENTATION DETAILS

Following the approach described in [23], our experiments include both real-world and synthetic LF datasets. Specifically, we use the 30scenes [20] and STFlytro [34] datasets for real-world data, while HCInew [35] and HCIold [36] serve as our synthetic datasets. The division of training and testing data aligns with the method in [23], using 100 real-world scenes and 20 synthetic scenes for training. For testing, we use 30 scenes from the 30scenes dataset, and 40 scenes from the STFlytro dataset, divided into 25 and 15 scenes from the Occlusion and Reflective categories respectively. For the synthetic datasets, we select 4 scenes from HCInew and 5 scenes from HCIold.

These datasets incorporate several critical factors essential for evaluating LF reconstruction methods. Specifically, the synthetic datasets are particularly valuable for assessing the performance of methods on large-baseline LFs, as evidenced by their substantial disparity ranges shown in Table 1. These datasets also feature high-resolution textures, which are crucial

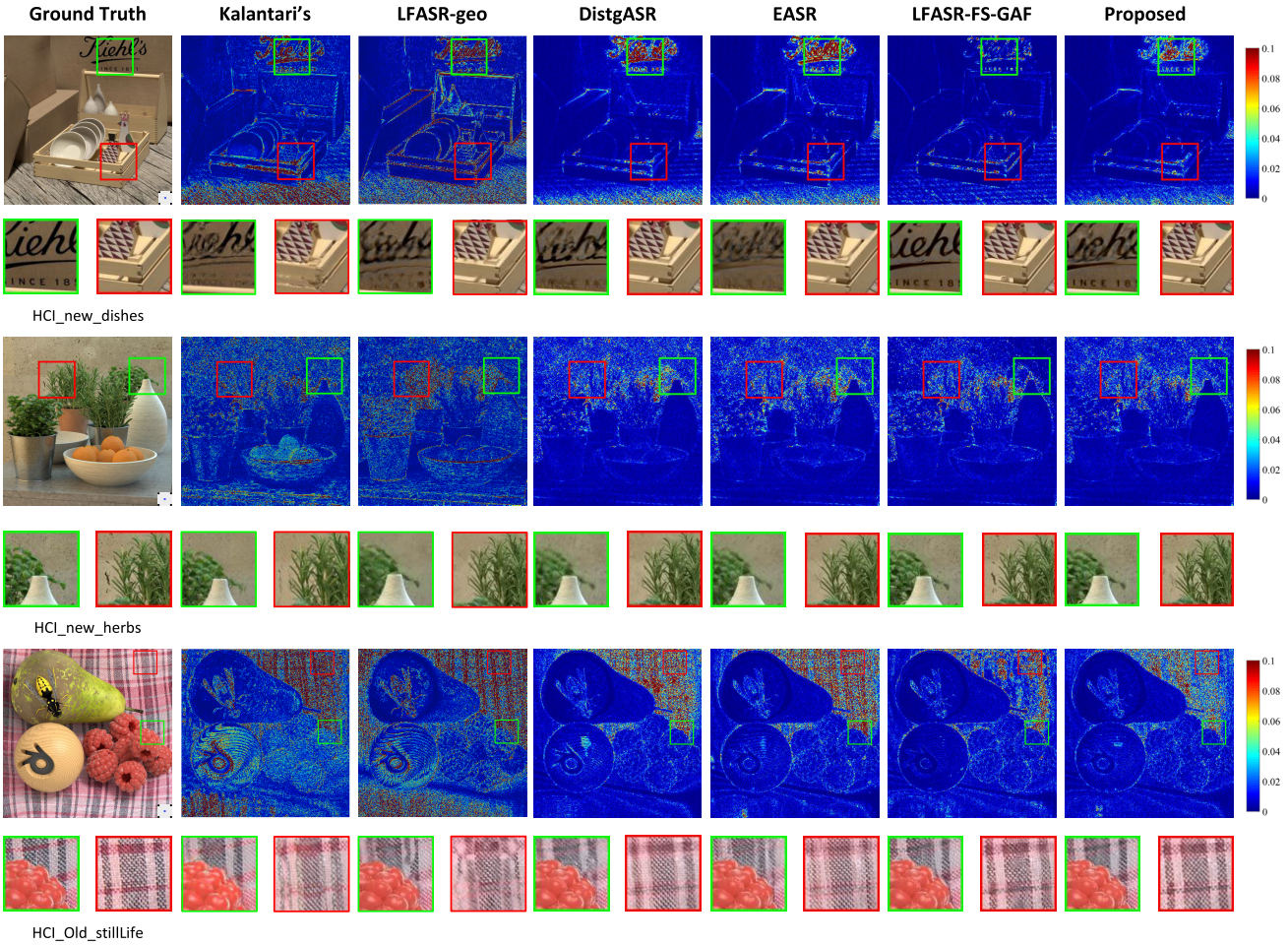


FIGURE 5. Visual comparison between our proposed approach and SOTA methods for 2 × 2 → 7 × 7 ASR on synthetic datasets.

TABLE 1. Details of the used datasets.

	Dataset	Disparity Range	Texture Contrast	Resolution
Synthetic	HCIold	[-3, 3]	4.834 ± 3.847	9×9×768×768
	HCInew	[-4, 4]	4.124 ± 2.782	9×9×512×512
Real-world	30s	[-1, 1]	5.797 ± 3.351	14×14×540×374
	occlusions	[-1, 1]	7.825 ± 5.447	14×14×540×374
	Reflective	[-1, 1]	4.377 ± 2.576	14×14×540×374

for evaluating the methods' ability to preserve high-frequency details. To quantify the complexity of textures within these scenes, we report the texture contrast, calculated on the center view using the Gray-Level Co-occurrence Matrix [39]. A higher texture contrast indicates scenes with more complex textures. Conversely, the real-world datasets provide a means to evaluate performance under natural illumination conditions and practical camera distortions.

These datasets typically exhibit much smaller disparity ranges compared to synthetic datasets. However, the portability of the Lytro camera allows for the capture of diverse outdoor scenes with intricate real-world textures, the texture contrast is very high indicating more complex textures as highlighted in Table 1. By addressing these varied aspects, these datasets facilitate a comprehensive evaluation of LF reconstruction methods, ensuring that they are accurately tested across a wide spectrum of scenarios and challenges.

for simplicity, we employ a $2 \times 2 \rightarrow 7 \times 7$ ASR setting. To create training and test samples, we conduct angular cropping on the central 7×7 SAIs and use the 2×2 corner views to rebuild the remaining views. Throughout the training, each SAI is cropped into 64×64 patches. To enhance model robustness, we use data augmentation techniques such as vertical flipping, 90-degree rotation, and random horizontal flipping.

Our network was trained with N = 5 for the number of the Deep Res-Blocks, a batch size of 4, employing optimization through the Adam method [38] setting $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The L_1 loss is more effective at reconstructing sharp edges than the L_2 loss [32]. Therefore, we use the L1 loss to constrain the constructed novel views and the ground truth. For a training pair $\{I_{GT}, I_{LR}\}$, where I_{GT} is the LF ground truth image, I_{LR}

is the sparsely sampled input, and f represents our proposed network. The loss function is defined as:

$$\text{Loss} = \|f(I_{\text{LR}}) - I_{\text{GT}}\|_1$$

The starting learning rate was configured at 2×10^{-4} and was reduced by half every 25 epochs. The training process concluded after 80 epochs. All experiments were conducted on a PC equipped with Nvidia GeForce RTX 4090 GPU.

For quantitative evaluation, we utilized PSNR and SSIM values calculated on the Y channel images. Initially, PSNR and SSIM values were computed for the reconstructed views (a total of 45 views for $2 \times 2 \rightarrow 7 \times 7$ ASR) and averaged these values to get a score for each scene. The overall score for a dataset was then determined by averaging the scores of all its scenes.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

We compare the proposed model with SOTA LF reconstruction methods, including Kalantari et al. [20], Jin et al. [22], LFASR-FS-GAF [23], DistgASR [31], and EASR [32]. To assess performance quantitatively, we employ PSNR and SSIM metrics. The PSNR and SSIM values for all methods are computed using the Y channel images from the synthesized views. To assess performance visually, we employ error maps between the ground truth Y channel images of different views and the Y channel images of the synthesized views from all the methods. The quantitative results of our reconstruction task ($2 \times 2 \rightarrow 7 \times 7$) are presentedin Table 2. It's evident that our model outperforms others on real-world test datasets. While it ranks second-best behind the depth-based LF reconstruction method LFASR-FS-GAF [23] on the HCInew test dataset, our model achieves superior performance on the HCIold dataset for synthetic datasets. Specifically, compared to the non-depth-based methods DistgASR [31] and EASR [32], our model surpasses them on the real-world test dataset by 0.14 dB and 0.37 dB, respectively, on the 30-scenes dataset. Additionally, it surpasses them by 0.74 dB and 0.40 dB on the occlusion dataset, and by 0.67 dB and 0.43 dB on the reflective dataset. Furthermore, our model outperforms both on average by 0.81 dB and 0.55 dB on the synthetic datasets.

Compared to the depth-based methods Kalantari et al. [20], Jin et al. [22], LFASR-FS-GAF [23], our model surpasses them on the real-world test dataset by 2.41 dB, 1.27 dB, and 1.06 dB, respectively, on the 30-scenes dataset. Additionally, it surpasses them by 2.95 dB, 1.67, and 1.69 dB on the occlusion dataset, and by 1.69 dB, 1.32, and 1.43 dB on the reflective dataset. Furthermore, our model outperforms both Kalantari et al. [20], and Jin et al. [22], on HCInew by 2.78 dB and 1.03 dB respectively. Additionally, it surpasses them by 4.29 dB and 2.03 dB on the HCIold dataset. Compared to LFASR-FS-GAF [23], we surpass them by 1.07 dB on the HCIold dataset. Still, because they formulate plane-sweep volumes (PSVs) for deducing depth, enabling them to perform better on the large-disparity HCInew dataset, However, it faces challenges in effectively utilizing sub-pixel correspondences in scenarios with small disparities as seen in real-world datasets. Additionally, our model achieves the best SSIM among all the methods. Our framework learns occlusion relationships, spatial texture structures, and geometric consistency among neighboring views implicitly. By leveraging more information due to dual feature extraction in LF angular reconstruction, we achieve a higher quality of reconstruction, particularly in occluded regions.

Our proposed method achieves a model size of 2.08M parameters, which is smaller compared to other non-depth-based methods with sizes of 6.44M [32] and 2.74M [31]. Given that our approach and depth-based methods operate in fundamentally different ways, a direct comparison of computational complexity may not fully capture the efficiency of each method. Nevertheless, our model's smaller size contributes to its efficiency in terms of memory usage and processing speed. Additionally, the design of our CNN-based network supports scalable performance with reduced computational overhead in real-world applications.

Figures 4 and 5 depict the visual comparison among various methods. The error maps reveal that our approach produces views that closely resemble the ground truth, preserving intricate structures. However, it's worth noting that for the HCInew dataset, LFASR-FS-GAF [23] outperforms our method. The zoomed-in areas demonstrate that our approach effectively preserves detailed textures during the synthesis of new views, whereas the compared methods exhibit varying degrees of blurring or artifacts.

C. ABLATION INVESTIGATION

In this subsection, we conduct multiple experiments to validate the efficacy of the proposed method. The primary contribution of this method is dual feature extraction; therefore, we conduct four variants to demonstrate its effectiveness and the fifth variant for the Channel Attention layer, as shown in Fig. 6 compared to our proposed method.

1) INTIAL FEATURS FROM EACH SAI IMAGE

Instead of employing dual feature extraction for the initial features, we developed an alternative approach as our first variant where the initial features are individually extracted from each SAI. This is achieved by first applying a shared weight 3×3 convolution to each SAI. Subsequently, we continue to utilize dual feature extraction within the Deep ResBlock for further processing, as shown in Fig. 6 (Model with FE on SAIs). Table 3 demonstrates that the PSNR values of this variant (SAIs_init) experience a decrease of 0.69 dB, 0.47 dB, and 0.60 dB on the 30-scenes, Occlusions, and Reflective datasets, respectively, when compared with our method. This indicates the effectiveness of dual feature extraction with the initial feature extraction.

2) DEEP FEATURS FROM A SINGLE DIRECTION

We developed the second and third variants by selecting a single direction, either horizontal or vertical, for extracting deep features in the Deep ResBlocks instead of employing dual feature extraction. In the second variant, we stacked

TABLE 2. Quantitative comparisons (using PSNR/SSIM) between our proposed approach and SOTA methods for 2 × 2 → 7 × 7 ASR. The best results are bolded while the second-best results are underlined.

Test Sets	Kalantari *et al.* [20]	Jin *et al.* [22]	LFASR-FS-GAF [23]	DistgASR [31]	EASR [32]	Ours
30scenes	41.40 / 0.982	42.54 / 0.986	42.75 / 0.986	43.67 / **0.995**	43.44 / **0.995**	**43.81 / 0.995**
Occlusions	37.25 / 0.972	38.53 / 0.979	38.51 / 0.979	39.46 / 0.991	39.80 / 0.992	**40.20 / 0.992**
Reflective	38.09 / 0.953	38.46 / 0.959	38.35 / 0.957	39.11 / 0.978	39.35 / 0.981	**39.78 / 0.982**
Average	38.91 / 0.969	39.84 / 0.975	39.87 / 0.974	40.75 / 0.988	40.86 / 0.989	**41.26 / 0.990**
HCInew	32.85 / 0.909	34.60 / 0.937	**37.14** / 0.966	34.70 / 0.974	35.86 / 0.975	35.63 / **0.978**
HCIold	38.58 / 0.944	40.84 / 0.960	41.80 /0.974	42.18 / 0.978	41.54 / 0.971	**42.87 / 0.988**
Average	35.72 / 0.927	37.72 / 0.949	**39.47** / 0.970	38.44 / 0.976	38.70 / 0.973	39.25 / **0.983**

TABLE 3. Ablation results on the real-world dataset for (2 × 2 → 7 × 7) task. The best results are in bold.

Variants	#params	30scenes	Occlusion	Reflective
SAIs_init	2.08M	43.12/0.994	39.31/0.991	39.18/0.979
H	2.08M	43.45/0.995	39.66/0.992	39.35/0.981
V	2.08M	43.42/0.995	39.62/0.991	39.34/0.981
w/o WS	3.93M	43.70/0.995	40.05/0.992	39.64/0.982
w/o CA	2.05M	43.47/0.995	39.81/0.992	39.26/0.981
Ours	2.08M	**43.81/0.995**	**40.20/0.992**	**39.78/0.982**

the SAIs in the horizontal direction only, as illustrated in Fig. 6 (Model with Horizontal Deep ResBlocks). For the third variant, we stacked the SAIs in the vertical direction only, as depicted in Fig. 6 (Model with Vertical Deep ResBlocks). Table 3 demonstrates that the PSNR values of the horizontal variant (H) experience a decrease of 0.36 dB, 0.54 dB, and 0.43 dB on the 30-scenes, Occlusions, and Reflective datasets, respectively, when compared with our method. Similarly, for the vertical variant (V), there is also a decrease of 0.39 dB, 0.58 dB, and 0.44 dB. This shows that adding more information in the reconstruction process helps increase the quality of the reconstruction.

3) W/O WEIGHT SHARING BETWEEN HORIZONTAL AND VERTICAL DIRECTIONS

In the dual feature extraction process, we transposed the vertical stack and applied weight sharing between the horizontal and transposed vertical stacks. For the fourth variant, we eliminated weight sharing between the horizontal and vertical directions and processed each stack separately, as shown in Fig. 6 (Model with Horizontal & Vertical Deep ResBlocks). As seen from Table 3, there is a significant disparity in the number of parameters between this variant (i.e., w/o WS) and our method, favoring our approach. Additionally, it experiences a reduction of 0.11 dB, 0.15 dB, and 0.14 dB on the 30-scenes, Occlusions, and Reflective test sets, respectively, when compared with our method. This underscores the efficiency of our proposed approach. It also demonstrates that the weight-sharing strategy can offer additional regularization for our dual features, aiding them in better conforming to the LF parallax structure.

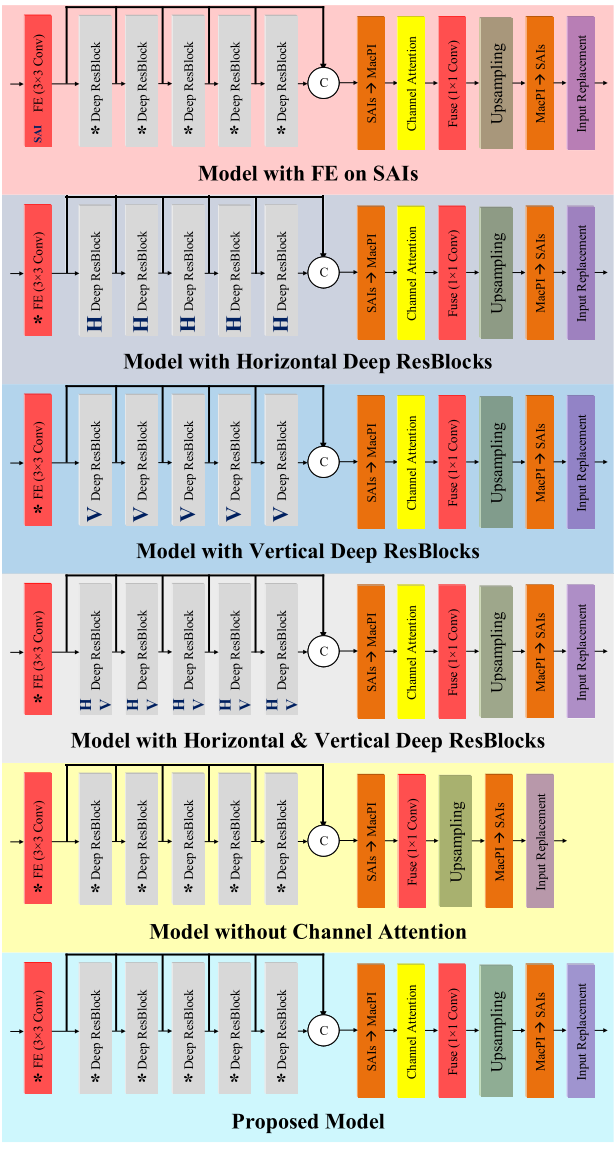


FIGURE 6. Different variants validate the efficacy of the proposed model. The (*) symbol indicates the application of dual feature extraction in the block.

4) CHANNEL ATTENTION LAYER

The final part of our ablation study was conducted to demonstrate the effectiveness of the Attention layer. For this

purpose, we removed the Attention layer and directly fused the features of the Deep ResBlocks, as shown in Fig. 6 (Model without Channel Attention). In this variant (i.e., w/o CA), the PSNR values experienced decreases of 0.34 dB, 0.39 dB, and 0.52 dB on the 30-scenes, Occlusions, and Reflective datasets, respectively, compared to our method, as shown in Table 3. This illustrates the model's capability to selectively enhance informative features while suppressing less relevant ones through adaptive weighting.

In summary, our framework, unlike previous variants, offers several benefits, including the addition of extra information for reconstruction, reduction in size complexity, preservation of the parallax structure, and enhancement of angular reconstruction quality.

V. CONCLUSION

In this paper, we introduce a novel LF reconstruction method by designing a CNN-based network that integrates spatial and epipolar features during both the initial and deep feature extraction stages, while leveraging angular information during the upsampling stage. Our key contributions include a dual feature extraction method that combines epipolar and spatial information and a parameter-sharing mechanism that maintains model efficiency and compactness. Together, these components contribute to our SOTA performance, as validated by extensive experiments demonstrating superior numerical and visual results.

In future work, we aim to advance the network by integrating diagonal processing alongside traditional horizontal and vertical processing. Diagonal processing is intended to capture patterns and relationships that extend beyond the constraints of these axes, thereby enhancing the handling of complex data relationships and contributing to more accurate and robust results. Additionally, we will implement shearing techniques to address challenges related to occlusions, reflections, and reconstructions involving larger disparity ranges. LF shearing entails applying a shear transformation to EPIs to mitigate disparities between views and reduce angular aliasing. This technique has the potential to improve the quality of reconstructed views and diminish ghosting effects, resulting in clearer and more detailed reconstructions. Lastly, we will address the current lack of statistical analysis by incorporating confidence intervals, p-values, and other statistical measures to rigorously validate our results and substantiate the claims of performance improvements.

REFERENCES

[1] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computer Models Visual Processing*. Cambridge, MA, USA: MIT Press, 1991, pp. 3–20.

[2] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, Aug. 1996, pp. 31–42.

[3] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013.

[4] K. Yucer, A. Sorkine-Hornung, O. Wang, and O. Sorkine-Hornung, "Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction," *ACM Trans. Graph.*, vol. 35, no. 3, pp. 1–15, Jun. 2016.

[5] A. Raytrix. (2017). *3D Light Field Camera Technology*. [Online]. Available: http://raytrix.de/products

[6] C. Zhang, G. Hou, Z. Zhang, Z. Sun, and T. Tan, "Efficient auto-refocusing for light field camera," *Pattern Recognit.*, vol. 81, pp. 176–189, Sep. 2018.

[7] 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.

[8] T. Yan, F. Zhang, Y. Mao, H. Yu, X. Qian, and R. W. H. Lau, "Depth estimation from a light field image pair with a generative model," *IEEE Access*, vol. 7, pp. 12768–12778, 2019.

[9] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19809–19818.

[10] Y. Wang, T. Wu, J. Yang, L. Wang, W. An, and Y. Guo, "DeOccNet: Learning to see through foreground occlusions in light fields," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 118–127.

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

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

[13] 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, Oct. 2017.

[14] B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," in *Proc. ACM SIGGRAPH Papers*, Jul. 2005, p. 765.

[15] S. G. Laboratory. *The (New) Stanford Light Field Archive*. Accessed: Nov. 23, 2023. [Online]. Available: http://lightfield.stanford.edu

[16] Y. Wang, L. Wang, Z. Liang, J. Yang, R. Timofte, and Y. Guo, "NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results," 2023, *arXiv:2304.10415*.

[17] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[18] Z. Zhang, Y. Liu, and Q. Dai, "Light field from micro-baseline image pair," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3800–3809.

[19] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S.-M. Hu, "PlenoPatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 5, pp. 1561–1573, May 2017.

[20] 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, Nov. 2016.

[21] A. Salem, H. Ibrahem, and H.-S. Kang, "Dual disparity-based novel view reconstruction for light field images using discrete cosine transform filter," *IEEE Access*, vol. 8, pp. 72287–72297, 2020.

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

[23] 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, doi: 10.1109/TPAMI.2020.3026039.

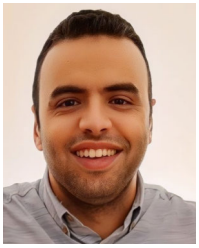
[24] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 22–28.

[25] L. Shi, H. Hassanieh, A. Davis, D. Katabi, and F. Durand, "Light field reconstruction using sparsity in the continuous Fourier domain," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 1–13, Dec. 2014.

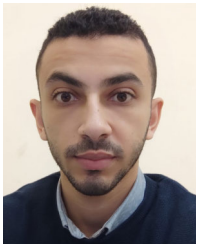
[26] S. Vagharshakyan, R. Bregovic, and A. Gotchev, "Light field reconstruction using shearlet transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 133–147, Jan. 2018.

[27] 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. Workshop (ICCVW)*, Dec. 2015, pp. 57–65.

[28] H. Zhu, M. Guo, H. Li, Q. Wang, and A. Robles-Kelly, "Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 3019–3033, Jun. 2021.
[29] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1681–1694, Jul. 2019.
[30] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared EPI structure for light field reconstruction," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3261–3273, Jul. 2019.
[31] Y. Wang, L. Wang, G. Wu, J. Yang, W. An, J. Yu, and Y. Guo, "Disentangling light fields for super-resolution and disparity estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 425–443, Jan. 18, 2023, doi: 10.1109/TPAMI.2022.3152488.
[32] 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*, early access, Oct. 10, 2022, doi: 10.1109/TMM.2022.3211402.
[33] A. Salem, H. Ibrahem, and H.-S. Kang, "Light field reconstruction using residual networks on raw images," *Sensors*, vol. 22, no. 5, p. 1956, Mar. 2022.
[34] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein. (2016). *Stanford Lytro Light Field Archive*. [Online]. Available: http://lightfields.stanford.edu
[35] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.
[36] S. Wanner, S. Meister, and B. Goldlucke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. 18th Int. Workshop Vis., Modeling Vis.* Lugano, Switzerland: The Eurographics Association, 2013, pp. 225–226.
[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image superresolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[39] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst. Man, Cybern.*, vols. SMC-3, no. 6, pp. 610–621, Nov. 1973.



AHMED SALEM received the B.Eng. degree in electrical engineering (electronics and communication) from Assiut University, Egypt, in 2012, the M.Eng. degree in electronics and communication engineering from the Egypt–Japan University of Science and Technology, Egypt, in 2016, and the Ph.D. degree in information and communication engineering from the School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si, South Korea, in 2022. He is currently a Postdoctoral Fellow with the Department of Information and Communication Engineering, Chungbuk National University. His research interests include multimedia, image processing, machine learning, deep learning, and computer vision.



EBRAHEM ELKADY received the B.S. degree in computers and information (information technology) from Assiut University, Asyut, Egypt, in 2017. He is currently pursuing the combined master's and Ph.D. degree with the School of Electronics Engineering, Chungbuk National University, Cheongju-si, South Korea. His research interests include computer vision, image processing, deep learning, and machine learning.

HATEM IBRAHEM received the B.Eng. degree in electrical engineering (electronics and communication) from Assiut University, Egypt, in 2013, and the Ph.D. degree in information and communication engineering from the School of Electrical and Computer Engineering, Chungbuk National University, Cheongju-si, South Korea, in 2023. He is currently a Postdoctoral Fellow with the Department of Computer Science, Toronto Metropolitan University. His research interests include multimedia, image processing, machine learning, deep learning, and computer vision.

JAE-WON SUH received the B.S. degree in electronic engineering from Chungbuk National University, Cheongju-si, South Korea, in 1995, and the M.S. and Ph.D. degrees in information and communications engineering from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 1997 and 2003, respectively. He joined LG Elite Mobile Multimedia Laboratories, Seoul, South Korea, in 2003. From 2003 to 2004, he was involved in development of the mobile phone. Since 2004, he has been with Chungbuk National University, where he is currently a Professor of electronic engineering with the School of Electrical and Computer Engineering, Chungbuk National University. His research interests include bio-medical signal processing, content-based signal representation, data hiding, digital video and audio coding, error resilience video coding, advanced video coding techniques, scalable video coding, multi-view coding, depth map estimation, and light field image.

HYUN-SOO KANG (Member, IEEE) received the B.S. degree in electronic engineering from Kyungpook National University, Republic of Korea, in 1991, and the M.S. and Ph.D. degrees in electrical and electronics engineering from KAIST, in 1994 and 1999, respectively. From 1999 to 2005, he was with Hynix Semiconductor Company Ltd., the Electronics and Telecommunications Research Institute (ETRI), and Chung-Ang University. In March 2005, he joined the College of Electrical and Computer Engineering, Chungbuk National University, Chungbuk, Republic of Korea. His research interests include image compression and image processing.