Geometry-aware light field angular super-resolution using multiple representations

Tariq Saeed Mian [a,1,*]

[a] Department of IS, College of Computer Science & Engineering, Taibah University, Madinah Almunwarah, Saudi Arabia
[1] tmian@taibahu.edu.sa
* corresponding author

ARTICLE INFO

Article history
Received February 21, 2025
Revised June 12, 2025
Accepted June 27, 2025
Available online July 25. 2025

Keywords
Geometry-aware learning
Angular super-resolution
Depth estimation
Convolutional neural networks
Light field representations

ABSTRACT

Light Field Angular Super-Resolution (LFASR) is a critical task that enables applications such as depth estimation, refocusing, and 3D scene reconstruction. Acquiring LFASR from Plenoptic cameras has an inherent trade-off between the angular and spatial resolution due to sensor limitations. To address this challenge, many learning-based LFASR methods have been proposed; however, the reconstruction problem of LF with a wide baseline remains a significant challenge. In this study, we proposed an end-to-end learning-based geometry-aware network using multiple representations. A multi-scale residual network with varying receptive fields is employed to effectively extract spatial and angular features, enabling angular resolution enhancement without compromising spatial fidelity. Extensive experiments demonstrate that the proposed method effectively recovers fine details with high angular resolution while preserving the intricate parallax structure of the light field. Quantitative and qualitative evaluations on both synthetic and real-world datasets further confirm that the proposed approach outperforms existing state-of-the-art methods. This research improves the angular resolution of the light field without reducing spatial sharpness, supporting applications such as depth estimation and 3D reconstruction. The method is able to preserve parallax details and structure with better results than current methods.

1. Introduction

Light Field (LF) imaging is a technology that records both the spatial and angular information of light rays. With the development of commercial LF cameras, LF imaging has attracted increasing attention in industry and academia. LF imaging [1] has emerged as a technology that allows for capturing richer information about a scene. Unlike conventional photography, LF cameras can collect not only the accumulated intensity at each pixel but also light rays from different directions. Due to this angular information, LF photography can achieve effects that are difficult to obtain in conventional photography, such as 3D scene reconstruction [2], [3], depth estimation [4], object view synthesis [5], and digital refocusing [6].

Acquiring LF is a challenging task because plenoptic cameras [7], [8] have a limited number of sensors that are a trade-off between angular and spatial resolution. Earlier LF imaging technology uses a camera array [9] to capture LF images in a single shot or may employ a single camera with a computer-controlled gantry [10] to capture multiple shots in a time-sequential manner. The former method can capture high-angular-resolution LF images by assembling a large number of cameras, whereas the latter

acquisition method is only suitable for static scenes. These acquisition methods are expensive and bulky, making them unsuitable for practical or commercial purposes. To mitigate this problem, researchers have explored learning-based methods, often referred to as novel view synthesis or angular super-resolution, which aim to reconstruct dense angular views from a sparse set of high-spatial-resolution inputs [11], [12]. Learning- based LFASR methods are divided into categories (i) depth-dependent and (ii) non-depth-dependent methods. Depth-dependent methods usually stand on the two-stage framework work, i.e., they first estimate disparity or depth information to synthesize novel views, and then refine and blend novel views through different strategies [13], [14]. Although depth-dependent methods have achieved promising performance, they are still struggling to handle the repeat patterns. texture less regions, and non-Lambertian surfaces, where the scene depth cannot be predicted accurately. Non-depth-based methods are more attractive for LF images that do not adhere to the Lambertian assumption. These methods directly establish the mapping from sparsely sampled LFs to densely sampled ones by circumventing disparity information.

In this study, our primary objective is to reconstruct light fields (LFs) in the angular domain under wide baseline conditions using two main modules: (i) a multi-representation LF reconstruction module and (ii) a geometry-aware refinement module. The reconstruction module comprises two parallel reconstruction pipelines. The first reconstruction pipeline employs a multi-scale residual network with varying receptive fields as a depth estimation network to extract dense features from Sub-Aperture Images (SAIs). The estimated depth maps are passed to a warping module to synthesize initial intermediate views. The second reconstruction pipeline utilizes a 3D U-Net architecture to extract angular features from Micro-Images (MIs). The outputs of both streams are concatenated and passed through a convolutional layer to generate high-angular-resolution LF views. Notably, the depth estimation module is embedded in the physical warping process, as ground-truth disparity maps are unavailable. The key contributions of our proposed approach are as :

- to exploit the structural characteristics of LFs by leveraging multiple LF representations.
- to design a depth estimation network using a multi-scale residual structure to extract a wide range of spatial features from SAIs.
- to develop a 3D U-Net architecture to extract angular features from MIs, capturing spatial-angular correlations to reconstruct high angular resolution LFs.
- to extensive experiments on both synthetic and real-world datasets demonstrate that the proposed method outperforms state-of-the-art techniques in both qualitative and quantitative evaluations.

The remainder of this paper is organized as follows. Section II briefly reviews the related work, and Section III, provides the main components of the proposed method. In Section IV, we present extensive experiments, results, and ablation studies to demonstrate the effectiveness of our approach. Finally, we conclude our paper in Section

The goal of an LFASR (also known as LF view synthesis or reconstruction) is to reconstruct a dense sampled LF from a sparse set of input views. Recent ASR methods are broadly classified into two categories: non-depth-based methods and depth-based methods. A brief review is presented in this section.

In non-depth-based methods, priors are used to reconstruct dense sample LF by using signal processing techniques. Shi *et al.* [15] used the continuous Fourier spectrum for LF reconstruction based on sparsity. In this technique, boundary and diagonal views were used to synthesize the novel views. Vagharshakyan *et al.*[16] designed a novel technique based on the concept of LF sparsification. The shearlet transform was used as the sparse transform, and a restoration technique was proposed for LF reconstruction. These approaches always need a large number of input images with a specific sampling pattern. The drastic success of deep learning techniques in LF image processing also brought a revolution in LF angular SR. Yeung *et al.* [17] presented an alternating convolution of a spatial angular network to reconstruct the densely sampled LF. Wu *et al.* [9] presented a 'blur-restoration-deblur'

framework for angular SR. Meng *et al.* [18] presented a deep high-dimensional dense residual network with 4D convolution for LF reconstruction. Wang *et al.* [19] proposed a general disentangling mechanism and developed a DistgASR network for LF angular SR using four groups of disentangling blocks. Each block separately processes spatial, angular, and EPI information, enabling effective synthesis of dense LF views. Liu *et al.* [20] propose an efficient network that learns LF feature representation from SAIs and upsamples MIs to synthesize the dense LF views. A U-Net architecture is used to exploit spatial–angular correlations (SAC), and a pixel shuffle operator rearranges the expanded features for angular upsampling. However, it is difficult for non-depth-based methods to incorporate the complementary information without alignment among different views, resulting in limited performance. Saleem *et al.* [21] presented a Residual Channel Attention LF (RCA-LF) network for view synthesis, utilizing residual channel attention blocks to enhance feature extraction and restore textures. The architecture employs stacked 2D convolutions with channel attention mechanisms to efficiently learn inter-view relationships. However, it lacks explicit depth information utilization, limiting its performance in challenging scenes. Saleem *et al.* [11] present an LF view synthesis method leveraging deep residual feature extraction and channel attention mechanisms. This framework integrates dual-feature extraction with MIs up-sampling to improve spatial-angular detail and maintain parallax consistency. While achieving state-of-the-art performance, it faces challenges in computational efficiency and effectively handling large disparities. Wang *et al.* [22] uses a residual channel attention mechanism to enhance feature extraction and a classification up-sampling module to improve reconstruction precision. The architecture combines residual attention groups with a classification-guided up-sampling strategy to adaptively refine angular information and address occlusion challenges. However, the method may face limitations in handling extremely sparse angular inputs or generalizing across diverse LF datasets. Wang *et al.* [23] introduce ViewFormer, a Transformer-based framework for LFASR that incorporates view-specific queries to encode both content and spatial coordinates of dense target views. By leveraging a Transformer encoder-decoder architecture and view interpolation, it enables dynamic feature enhancement guided by angular positions. The method achieves state-of-the-art performance on both synthetic and real-world LF datasets, outperforming prior CNN-based approaches. To address the challenge of incomplete Spatial-Angular Correlation (SAC) feature extraction in LFASR, this paper [24] introduces a Deformable Convolutional Network (DCN) that adaptively samples distant correlated pixels. A Multi-Maximum-Offsets Fusion (MMOF) strategy is proposed to further enhance offset accuracy, enabling more precise SAC extraction. The approach significantly improves LF reconstruction quality over existing CNN- and attention-based methods with limited receptive fields. Liu *et al.* [25] introduce an efficient progressive disentangling block (PDistgB) that selectively disentangles LF features in a domain-specific manner through channel-wise splitting. Additionally, angular-domain Transformers are employed to exploit global angular correlations. The proposed method achieves state-of-the-art performance while significantly reducing inference cost compared to conventional disentangling strategies. In this study, Liu *et al.* [10] propose a convolutional Transformer-based framework comprising GLCTNet for global-local feature extraction, DDNet for deep deblurring, and TFNet for texture-aware fusion. Experimental results confirm that the proposed method effectively enhances LF reconstruction quality, particularly in preserving structure and suppressing artifacts, with demonstrated benefits for downstream tasks like depth estimation. Daichuan *et al.*[26] introduce edge-aware LFASR that enhances edge-feature extraction and utilization by combining Sobel-based edge-magnitude maps with a neural network for complete edge representation. A multi-level attention mechanism fuses edge, spatial, and angular features to preserve global structure while refining local edges. Additionally, an edge-aware loss function guides the reconstruction process. Experimental results demonstrate improved PSNR and reduced edge distortion on both real and synthetic datasets, validating the method's effectiveness.

In this method, the depth of the scene is first estimated, and the input images are warped to novel views based on the depth map. The warped images are then blended in different ways to achieve the final novel views. Kalantari *et al.*[27] employs two sequential CNNs to estimate disparity and color values. Their proposed method ignores the correlations between the warped views, leading to limited reconstruction quality and inability to fully exploit the angular information inherent in LF data. Wanner

et al. [28] used EPIs for disparity estimation and then generated novel views of the scene in a variational framework. Shi *et al.* [29] employed pixel-based reconstruction and feature-based modules to construct densely sampled views using the estimated depth maps. Meng *et al.* [30] used two DenseNets to compute the scene depth, a warping confidence map, and a refinement network to synthesize the target views. Wu *et al.* [31] used a CNN-based network to evaluate sheared EPIs, where the sheared value is correlated with the depth. Jin *et al.* [32] employed a CNN-based depth estimation module and blending to reconstruct a high-angular-resolution LF. Zhou *et al.* [33] proposed an encoder-decoder network to estimate the disparity for synthesizing novel views through warping. They utilized a modified ResNet50 to extract expressive representations and employed three subnetworks for disparity estimation, noise filtering, and view rendering. Lui *et al.* [34] propose a method comprising two modules: (i) Multi-Representation View Reconstruction (MRVR), and (ii) Geometry-Assisted Refinement (GAR). The MRVR module extracts dense features from SAIs, MIs, and Pseudo Video Sequences (PVS) through distinct pipelines built on conventional convolutional networks. These pipelines construct a Dense LF Image (DLFI) stack that encapsulates comprehensive spatial–angular and geometric cues. The GAR module further refines this stack via a geometry-aware network operating on a bidirectional view structure, effectively reinforcing angular consistency and significantly elevating the fidelity of LFVS. Zubair *et al.* [35]extended Liu's framework by replacing the traditional CNN in the SAI branch with deformable convolutions for disparity estimation. The use of flexible offsets in deformable convolutions enables the precise modeling of depth variations across views, significantly improving the quality of synthesized views. They use depth-wise separable convolutions for efficient feature extraction from MIs and lightweight refinement in the GAR module.

Chen *et al.* [36] introduce a Cross-Shaped Transformer Network (CSTNet) architecture with a Multiplane-based Cross-view Interaction Mechanism (MCIM) for LFASR. By leveraging Multiplane Feature Fusion (MPFF) and a plane selection strategy inspired by MPI transparency, it enables efficient and geometry-aware view synthesis. Experimental results confirm that CSTNet outperforms existing methods across both real and synthetic LF benchmarks.

Propose a method comprising two modules: (i) Multi-Representation View Reconstruction (MRVR), and (ii) Geometry-Assisted Refinement (GAR). The MRVR module extracts dense features from SAIs, MIs, and Pseudo Video Sequences (PVS) through distinct pipelines built on conventional convolutional networks. These pipelines construct a Dense LF Image (DLFI) stack that encapsulates comprehensive spatial–angular and geometric cues. The GAR module further refines this stack via a geometry-aware network operating on a bidirectional view structure, effectively reinforcing angular consistency and significantly elevating the fidelity of LFVS. Zubair *et al.* [35]

2. Method

LF is a parameterization of two plane parameters, angular $(U; V)$ and spatial $(X; Y)$. U and V represent the angular dimension, while X and Y represent the spatial dimension of LF. For high angular resolution LF, we represent the LF as $L\,(V; U; X; Y)$, where X and Y denote the spatial resolution of the SAIs, and U and V represent the angular resolution. We can construct the angular SR LF from the spatial resolution of X and Y:

$$L^{har}\,(U; V; H; W) = f(L(U; V; H; W) \tag{1}$$

In equation (1),f denotes the network parameter, while $L\,(U; V; H; W)$ represents the sparse set of input views to reconstruct the high angular resolution LFs $L^{har}\,(U; V; H; W)$. Following prior reconstruction methods [37], [38], we consider only the Y channel information in the YCbCr color space as input. This choice avoids the increased model complexity and training difficulty associated with using full RGB channels. The proposed method consists of two modules: (i) multiple-representation-based LF reconstruction and (ii) geometry-aware refinement network. The multiple-representation LF reconstruction module further consists of two reconstruction pipelines: (i) Sub-aperture-based LF reconstruction (SAIRP) and (ii) Micro-images-based LF reconstruction. We employed a multi-scale

residual network with varying receptive fields in the sub-aperture-based reconstruction pipeline, whereas a 3D UNet model was utilized in the MI-based reconstruction. The proposed method is shown in Fig. 1.

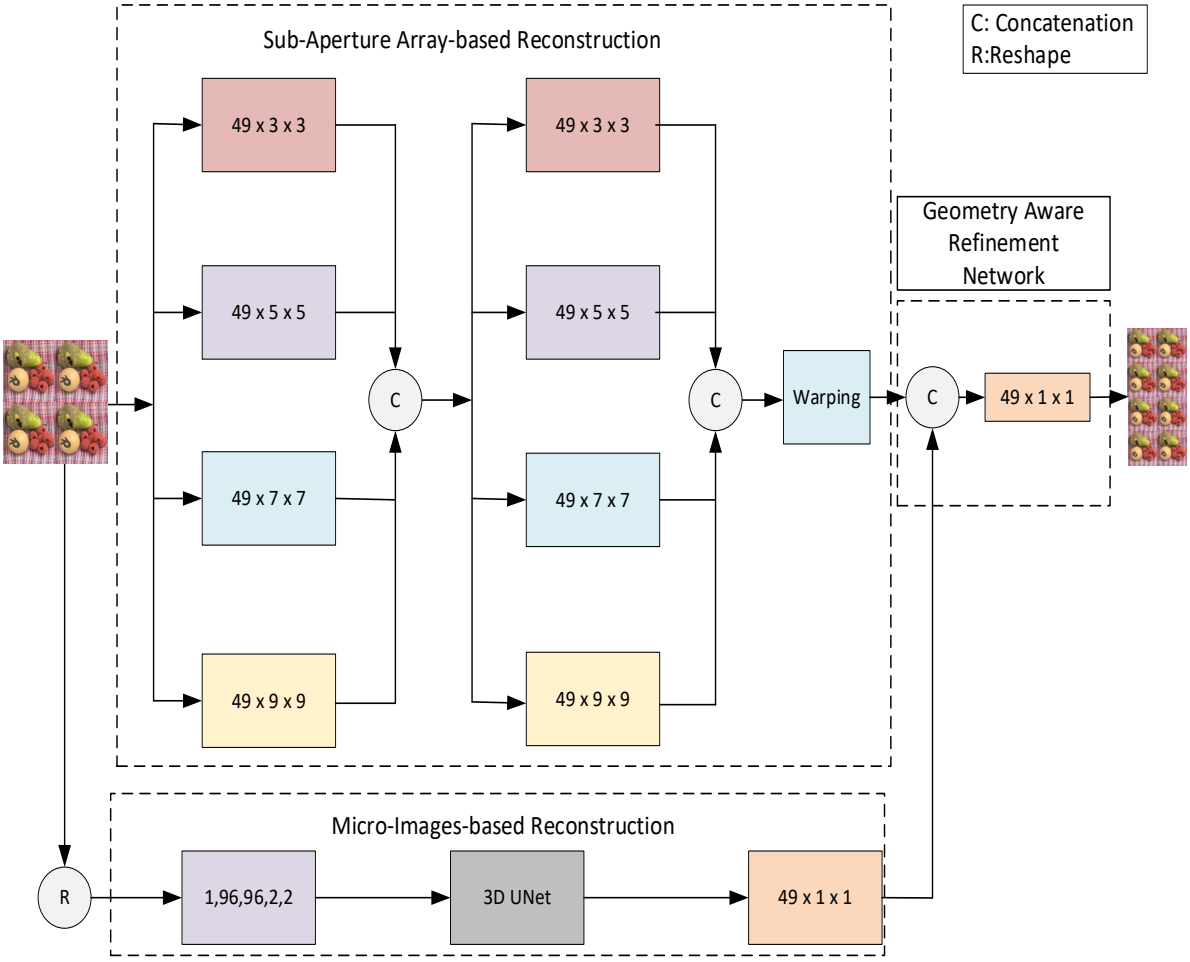

Fig. 1. Proposed method

2.1. Multiple representation-based LF reconstruction

LF representations have rich characteristics for angular reconstruction. An angular reconstruction-based method requires dense data to synthesize wide-baseline views. In this study, we explore two types of LF representations: SAIs and MIs.

2.1.1. Sub-Aperture-based LF Reconstruction (SAIRP)

SAIRP is designed to handle the array of SAIs for the spatial feature extraction of LF images. These features are extracted through a multi-scale residual network. The multi-scale residual network consists of two Multi-Scale Residual Blocks (MSRBs). MSRBs are a type of neural network architecture, designed to improve feature extraction and representation learning by capturing information at multiple scales. These blocks incorporate multiple convolutional filters of different sizes within the same block, allowing the network to capture features at different scales simultaneously. This makes the network more effective at recognizing patterns that vary in size. In our approach, we used kernel sizes of 3, 5, 7, and 9 to extract dense features and then passed them to the warping module. The 4D ray depth maps are represented as D(x, u) and are estimated from the input views:

$$D(x,u) = fd(L^{lar}(x',u')) \quad (2)$$

In the above equations, D represent Disparity estimation network, x for spatial resolution and u for angular resolution, f is a network parameter, d is estimated depth from low angular resolution views

L^{lar} with spatial resolution x' and angular resolution u'.Disparity maps D_t are passed to the warping module to synthesize the dense views. The warping synthesizes the high angular resolution LF views at the target position t, using low angular resolution views ($U \times V$) and disparity maps. The warping process can be mathematically expressed as.

$$L_{HR}^{(t)} = Warp(L_{ref}, D_t), t \in [1 \ldots, U' \times V'] \quad (3)$$

Where Warping(·) represent the back warping operation, $L_{HR}^{(t)}$ represents the synthesized high-angular-resolution LF at the target position t.

2.1.2. Micro-Images-based LF Reconstruction (MIRP)

First, we reshaped the SAIs into MI features and then employed 3D UNet to extract angular features from the MI format. This proposed module explores the spatial-angular correlation. The UNet architecture consists of two convolutional layers and two transposed convolutional layers. We use a 2D convolutional layer with a 1x1 kernel size to reshape the extracted features and then combine them with the output of the SAIRP module. The feature map in our U-Net-based model features 64, 128, and 192 spatial and channel dimensions with upsampling and downsampling operations.

2.2. Geometry-aware refinement network

While SAIRP and MIRP independently extract rich LF features, their outputs are concatenated and processed via a convolution layer to enforce geometric consistency and fuse the multi-dimensional features for high-angular-resolution LF reconstruction.

3. Results and Discussion

This section presents sufficient experimental results to validate the proposed method's performance. Firstly, the implementation details of the proposed method are briefly described. Then, the proposed method is comprehensively evaluated by comparing it with state-of-the-art methods both quantitatively and qualitatively. Subsequently, detailed ablation study experiments are conducted to validate the core modules of our method. Finally, we discuss the limitations of our proposed method.

3.1. Implementation details

The proposed method is implemented using the PyTorch framework. The experimental environment is configured with an Nvidia GTX4090Ti GPU and 128 GB of RAM. The proposed network is trained using an L1 loss and optimized with the Adam optimizer [39], with a batch size of 4. The initial learning rate is set to 2 × 10-4 and decreased by 0.5 every 15 epochs. This paper focuses on reconstructing densely sampled 7×7 LF data from sparsely sampled 2×2 LF data. Therefore, during training data preparation, ground truth (GT) samples are obtained by angularly cropping the central 7 × 7 SAIs of each LF. The input samples are generated using the 2×2 corner SAIs of the GT samples. To save GPU memory, each SAI is cropped into patches with 128×128 pixels during the training process. Several data augmentation strategies are employed to enhance robustness, including horizontal flipping, vertical flipping, and 90-degree rotation. The proposed model has a channel size (C) of 64 and is trained for 80 epochs on both synthetic and real-world datasets.

3.2. Dataset description

We trained two networks: one on synthetic datasets and another on real-world datasets. The network trained on synthetic datasets uses 20 scenes from the HCI new dataset [40]. For evaluation on synthetic scenes, we utilize 4 scenes from the HCI new [40] and 5 scenes from the HCI old dataset [41]. Regarding the real-world training datasets, we employ one hundred LF images provided by Kalantari *et al.*[27] and the Stanford Lytro Archive. These real-world training scenes are captured using the Lytro Illum camera. The performance of the proposed method is evaluated on three real-world datasets: 30 LF scenes from the 30Scenes dataset [27], 25 scenes from the occlusion dataset [42], and 15 scenes from the reflective dataset[42]. Table 1 presents a detailed description of the datasets.

Table 1. Dataset Description

LF datasets	Type	Disparity range	Angular resolution	Spatial resolution	Training scenes	Test scenes
HCI new [28]	Synthetic	[4, 4]	9 × 9	512 × 512	20	4
HCI old [29]	Synthetic	[4, 4]	9 × 9	768 × 768	-	5
Kalantari *et al.* [13]	Real	[1, 1]	14 × 14	376 × 541	100	-
30 Scenes [30]	Real	[1, 1]	14 × 14	376 × 541	-	30
Occlusions [30]	Real	[1, 1]	14 × 14	376 × 541	-	25
Reflective [30]	Real	[1, 1]	14 × 14	376 × 541	_	15

3.3. Comparison with state-of-the-art methods

To prove the efficiency and performance of our proposed methods, we compared it with four state-of-the-art methods [27], [32], [20], [34], and [35]. For a fair comparison, we have retrained the methods on the same training datasets as our method. We use peak-to-signal-noise-ratio (PSNR) and structure-similarity-index-measure (SSIM) metrics for performance evaluation. We calculated the PSNR and SSIM values on the Y channel images for all methods.

Table 2 presents the performance of the proposed method in comparison to state-of-the-art methods on synthetic datasets. The proposed method achieved an average increase in PSNR of 1.564 dB and an average increase in SSIM of 0.0046, demonstrating superior reconstruction quality across diverse scenes.

Table 2. Quantitative results on the synthetic dataset for the Task 2×2 to 7×7

Dataset	Kalantari *et al.* [27]	Jin *et al.* [32]	LF-EASR [20]	Liu *et al.* [34]	Zubair *et al.* [35]	Ours
	PSNR/SSIM	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*
HCI New	32.01/0.928	33.60/0.934	34.08/0.931	34.24/0.935	34.49/0.935	34.51/0.935
HCI old	38.58/0.944	39.90/0.954	40.42/0.966	40.64/0.954	41.22/0.958	41.66/0.961
Average	35.29/0.936	36.75/0.944	37.25/ 0.948	37.44/0.944	37.85/0.945	38.08/0.948

Table 3 shows the performance of the proposed method on real-world datasets. Compared to existing state-of-the-art techniques, the proposed method achieves an average PSNR improvement of 1.95 dB and an average SSIM improvement of 0.009.

Table 3. Quantitative results on a real dataset for the Task 2×2 to 7×7

Dataset	Kalantari *et al.* [27]	Jin *et al.* [32]	LF-EASR [20]	Liu *et al.* [34]	Zubair *et al.* [35]	Ours
	PSNR/SSIM	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*	*PSNR/SSIM*
30scenes	38.90/0.960	39.70/0.986	40.90/0.957	42.91/0.987	43.19/0.987	43.14/0.988
Occlusions	33.89/0.958	34.60/0.964	34.98/0.976	39.06/0.981	39.26/0.982	39.28/0.982
Reflective	36.95/0.925	37.81/0.972	38.56/0.975	39.04/0.962	39.20/0.963	39.24/0.965
Average	36.58/0.947	37.37/0.974	38.15/0.969	40.34/0.977	40.55/0.977	40.55/0.978

We used the error map to find the difference between the synthesized central view and the ground Truth (GT). Fig. 2 shows a visual comparison of the synthesized central view and corresponding error maps for one representative scene. Compared to Jin *et al.* [32], LF-EASR [20], Liu *et al.* [34], and Zubair *et al.* [35], the proposed method produces a more accurate reconstruction with visibly sharper texture details and reduced angular inconsistencies. The error maps further highlight that competing

methods exhibit higher residual errors, particularly along object boundaries and high-parallax regions, as indicated by the red boxes. In contrast, the proposed method yields lower error concentrations and better preserves structural fidelity in challenging areas.

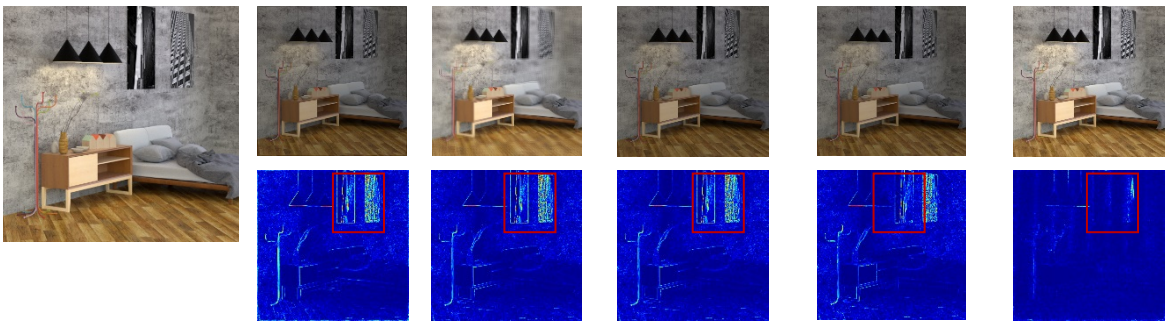

Fig. 2. Visual comparison of the proposed method with state-of-the-art methods on the *Bedroom* scene from the HCI_new dataset

Fig. 3 and Fig. 4 illustrate visual comparisons between the proposed method and existing state-of-the-art approaches on representative scenes from the 30Scenes and Reflective datasets, respectively.

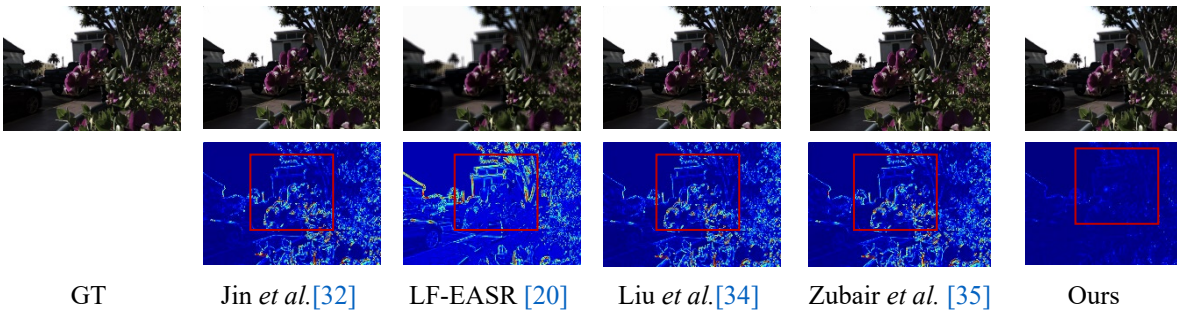

Fig. 3. Visual comparison of the proposed method with state-of-the-art methods on the *IMG-1555_eslf* scene from the 30scenes dataset

In both cases, the error maps clearly show that the proposed method yields fewer reconstruction errors, particularly in regions with high-frequency textures, complex lighting, and reflective surfaces. Specifically, in Fig. 3, competing methods exhibit significant residual errors around object edges and detailed regions (e.g., foliage), as highlighted in the red boxes. In contrast, the proposed method produces a cleaner reconstruction with minimal error concentrations. Similarly, in Fig. 4, the proposed method demonstrates strong robustness against reflective surfaces, maintaining accurate geometry and structure. While other methods suffer from high-intensity errors near shiny surfaces and background objects, the proposed method successfully preserves fine details and suppresses distortion.

Fig. 4. Visual comparison of the proposed method with state-of-the-art methods on the *Reflective-12_eslf* scene from the Reflective dataset

3.4. Angular Consistency

To further validate the effectiveness of the proposed method, we evaluate its ability to preserve angular consistency, which is critical for accurately modeling parallax structure in synthesized LFs. As illustrated in Fig. 5, we extract Epipolar Plane Images (EPIs) from both the synthesized views and their corresponding GT to assess angular coherence. The proposed method produces smoother and more continuous EPI lines, particularly along slanted and high-parallax regions, indicating stronger spatial–angular correlation. In contrast, existing methods, such as those by Liu *et al.* [34] and Zubair *et al.* [35], exhibit broken or distorted EPI structures, particularly near occlusion boundaries and fine geometric edges. These results confirm that our method better captures the geometric continuity and preserves angular consistency in challenging scenes.

3.5. Ablation Study

To assess the significance of each component in the proposed method, we conducted three controlled ablation experiments by removing each module in turn. In Variant I (Var-I), the SAIRP module was excluded while retaining the MIRP and the Geometry-Aware Refinement Network. This configuration resulted in a noticeable decline in reconstruction quality, underscoring the importance of spatial-angular interaction modeling. In Variant II (Var-II), the MIRP was removed while keeping SAIRP and the refinement network. The results indicate a degradation in angular detail synthesis, confirming the relevance of micro-image-based angular feature extraction. In Variant III (Var-III), the Geometry-Aware Refinement Network was excluded, revealing its contribution to enhancing structural consistency and depth-aware reconstruction. These ablation results demonstrate that each module plays a critical role, and their integration yields the best overall performance, as shown in Table 4.

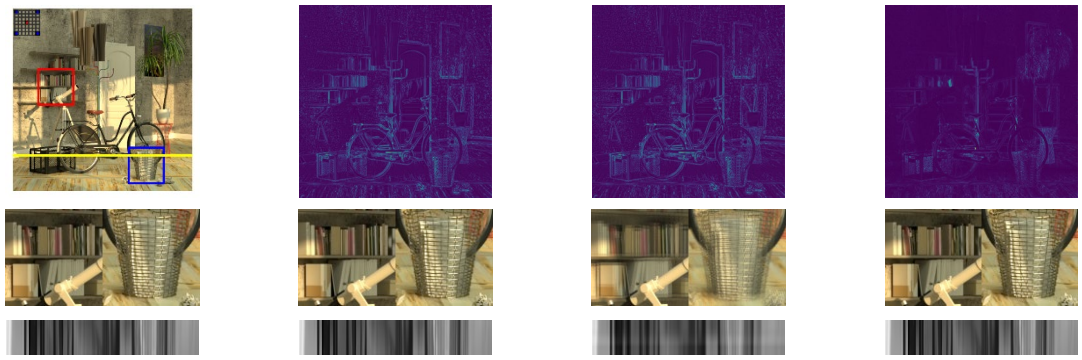

Fig. 5. Angular consistency comparison of the proposed method with the state-of-the-art methods

Table 4. These ablation results

Variants	SAIRP	MIRP	Geometry-aware Refinement Network	Synthetic Datasets
Var-I	×	✓	✓	36.53/0.888
Var-II	✓	×	✓	37.43/0.911
Var-II	✓	✓	×	37.91/0.922

4. Conclusion

In this work, we presented a learning-based end-to-end network for LFASR that extracts features from multiple representations of SAIs and MIs. The proposed method employs a multi-scale residual network with varying receptive fields to extract dense spatial features. At the same time, a 3D U-Net is utilized to capture angular features from the MI representation. This geometry-aware framework effectively synthesizes high-quality LFs and consistently outperforms state-of-the-art methods across multiple benchmark datasets, and is a potential application for rendering. The proposed method still struggles to synthesize high-angular-resolution views of challenging scenes. Future work will focus on non-depth-based methods that utilize advanced learning techniques to synthesize challenging scenes

across diverse datasets. These improvements are expected to benefit applications in virtual reality (VR), augmented reality (AR), and computational photography. The proposed method enables potential application of rendering.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors has received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Tech. SIGGRAPH 1996*, pp. 31–42, 1996, doi: 10.1145/237170.237199.

[2] J. Peng, Z. Xiong, Y. Zhang, D. Liu, and F. Wu, "LF-fusion: Dense and accurate 3D reconstruction from light field images," in *2017 IEEE Visual Communications and Image Processing, VCIP 2017*, Feb. 2018, vol. 2018-Janua, pp. 1–4, doi: 10.1109/VCIP.2017.8305046.

[3] 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, doi: 10.1007/s11263-021-01516-1.

[4] Y. Zhang, W. Dai, M. Xu, J. Zou, X. Zhang, and H. Xiong, "Depth Estimation from Light Field Using Graph-Based Structure-Aware Analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4269–4283, 2020, doi: 10.1109/TCSVT.2019.2954948.

[5] S. Yun, J. Jang, and J. Paik, "Learning-based Light Field View Synthesis Using Multiplane Images," *2023 Int. Conf. Electron. Information, Commun. ICEIC 2023*, pp. 1–3, 2023, doi: 10.1109/ICEIC57457.2023.10049922.

[6] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," *2014 IEEE Int. Conf. Comput. Photogr. ICCP 2014*, pp. 1–9, 2014, doi: 10.1109/ICCPHOT.2014.6831809.

[7] Raytrix, "Raytrix," 2024, [Online]. Available at: https://raytrix.de/raytrix-vision-2-3/.

[8] "'Lytro illum.'"[Online]. Available at: http://clim.inria.fr/IllumDatasetLF/index.html.

[9] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005, doi: 10.1145/1073204.1073259.

[10] D. Liu, Y. Mao, Y. Zuo, P. An, and Y. Fang, "Light Field Angular Super-Resolution Network Based on Convolutional Transformer and Deep Deblurring," *IEEE Trans. Comput. Imaging*, vol. 10, pp. 1736–1748, 2024, doi: 10.1109/TCI.2024.3507634.

[11] A. Salem, E. Elkady, H. Ibrahem, J. W. Suh, and H. S. Kang, "Light Field Reconstruction with Dual Features Extraction and Macro-Pixel Upsampling," *IEEE Access*, vol. PP, p. 1, 2024, doi: 10.1109/ACCESS.2024.3446592.

[12] S. Wang, H. Sheng, D. Yang, Z. Cui, R. Cong, and W. Ke, "MFSRNet: spatial-angular correlation retaining for light field super-resolution," *Appl. Intell.*, vol. 53, no. 17, pp. 20327–20345, 2023, doi: 10.1007/s10489-023-04558-9.

[13] D. Liu, Y. Huang, Y. Fang, Y. Zuo, and P. An, "Multi-Stream Dense View Reconstruction Network for Light Field Image Compression," *IEEE Trans. Multimed.*, vol. 25, pp. 4400–4414, 2023, doi: 10.1109/TMM.2022.3175023.

[14] D. Cai, Y. Chen, X. Huang, and P. An, "Disparity Enhancement-based Light Field Angular," *IEEE Signal Process. Lett.*, vol. PP, no. 8, pp. 1–5, 2024, doi: 10.1109/LSP.2024.3496582.

[15] 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, 2014, doi: 10.1145/2682631.

[16] 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, 2018, doi: 10.1109/TPAMI.2017.2653101.

[17] 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 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11210 LNCS, Springer International Publishing, 2018, pp. 138–154, doi: 10.1007/978-3-030-01231-1_9

[18] N. Meng, H. K. H. So, X. Sun, and E. Y. Lam, "High-Dimensional Dense Residual Convolutional Neural Network for Light Field Reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 873–886, 2021, doi: 10.1109/TPAMI.2019.2945027.

[19] 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, 2022, doi: 10.1109/TPAMI.2022.3152488.

[20] 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. Multimed.*, pp. 1–13, 2022, doi: 10.1109/TMM.2022.3211402.

[21] A. Salem, H. Ibrahem, and H.-S. Kang, "RCA-LF: Dense Light Field Reconstruction Using Residual Channel Attention Networks," *Sensors*, vol. 22, no. 14, p. 5254, Jul. 2022, doi: 10.3390/s22145254.

[22] X. Wang, Z. Wang, and S. You, "Light field angular super resolution based on residual channel attention and classification up-sampling," *Multimed. Tools Appl.*, vol. 84, no. 12, pp. 10945–10967, May 2024, doi: 10.1007/s11042-024-19359-6.

[23] S. Wang, Y. Lu, W. Xia, P. Xia, Z. Wang, and W. Gao, "Light field angular super-resolution by view-specific queries," *Vis. Comput.*, vol. 41, no. 5, pp. 3565–3580, Mar. 2025, doi: 10.1007/s00371-024-03620-y.

[24] D. Li, R. Zhong, and Y. Yang, "Light Field Angular Super-Resolution via Spatial-Angular Correlation Extracted by Deformable Convolutional Network," *Sensors*, vol. 25, no. 4, p. 991, Feb. 2025, doi: 10.3390/s25040991.

[25] G. Liu, H. Yue, and J. Yang, "Efficient Light Field Image Super-Resolution via Progressive Disentangling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern RecognitionProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2024, pp. 6277–6286, doi: 10.1109/CVPRW63382.2024.00631

[26] D. Li and R. Zhong, "Light Field Image Angular Super-Resolution Using Edge Features," *2024 6th Int. Conf. Robot. Comput. Vision, ICRCV 2024*, pp. 132–138, 2024, doi: 10.1109/ICRCV62709.2024.10758555.

[27] N. K. Kalantari, T. C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, doi: 10.1145/2980179.2980251.

[28] 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, 2014, doi: 10.1109/TPAMI.2013.147.

[29] J. Shi, X. Jiang, and C. Guillemot, "Learning fused pixel and feature-based view reconstructions for light fields," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2552–2561, 2020, doi: 10.1109/CVPR42600.2020.00263.

[30] N. Meng, K. Li, J. Liu, and E. Y. Lam, "Light Field View Synthesis via Aperture Disparity and Warping Confidence Map," *IEEE Trans. Image Process.*, vol. 30, no. April, pp. 3908–3921, 2021, doi: 10.1109/TIP.2021.3066293.

[31] 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, 2019, doi: 10.1109/TIP.2019.2895463.

[32] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," *AAAI 2020 - 34th AAAI Conf. Artif. Intell.*, pp. 11141–11148, 2020, doi: 10.1609/aaai.v34i07.6771.

[33] W. Zhou, J. Shi, Y. Hong, L. Lin, and E. Engin Kuruoglu, "Robust dense light field reconstruction from sparse noisy sampling," *Signal Processing*, vol. 186, p. 108121, 2021, doi: 10.1016/j.sigpro.2021.108121.

[34] 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," *Knowledge-Based Syst.*, vol. 267, p. 110390, 2023, doi: 10.1016/j.knosys.2023.110390.

[35] M. Zubair, P. Nunes, C. Conti, and L. D. Soares, "Light Field View Synthesis Using Deformable Convolutional Neural Networks," *2024 Pict. Coding Symp. PCS 2024 - Proc.*, pp. 11–15, 2024, doi: 10.1109/PCS60826.2024.10566360.

[36] R. Chen *et al.*, "Multiplane-based Cross-view Interaction Mechanism for Robust Light Field Angular Super-Resolution," *IEEE Trans. Vis. Comput. Graph.*, vol. 14, no. 8, 2025, doi: 10.1109/TVCG.2025.3564643.

[37] 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, 2022, doi: 10.1109/TPAMI.2020.3026039.

[38] S. Yun, J. Jang, and J. Paik, "Geometry-Aware Light Field Angular Super Resolution Using Multiple Receptive Field Network," *2022 Int. Conf. Electron. Information, Commun. ICEIC 2022*, pp. 2–4, 2022, doi: 10.1109/ICEIC54506.2022.9748458.

[39] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, pp. 1–15, Dec. 22, 2014. [Onlinr]. Available at: https://arxiv.org/abs/1412.6980v9.

[40] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10113 LNCS, no. 3, pp. 19–34, 2017, doi: 10.1007/978-3-319-54187-7_2.

[41] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," *18th Int. Work. Vision, Model. Vis. VMV 2013*, pp. 225–226, 2013, [Online]. Available at: https://diglib.eg.org/items/4e39b2ff-5177-483f-b611-dfc43222f22c.

[42] A. S. Raj, M. Lowney, R. Shah, and G. Wetzstein, "Stanford Lytro Light Field Archive," 2016.[Online]. Available at: https://lightfields.stanford.edu/LF2016.html.