ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2025-04-24

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Zubair, M., Nunes, P., Conti, C. & Soares, L. D. (2024). Light field view synthesis using deformable convolutional neural networks. In 2024 Picture Coding Symposium, PCS 2024, Proceedings. (pp. 1-5). Taichung, Taiwan: IEEE.

Further information on publisher's website:
10.1109/PCS60826.2024.10566360

Publisher's copyright statement:
This is the peer reviewed version of the following article: Zubair, M., Nunes, P., Conti, C. & Soares, L. D. (2024). Light field view synthesis using deformable convolutional neural networks. In 2024 Picture Coding Symposium, PCS 2024, Proceedings. (pp. 1-5). Taichung, Taiwan: IEEE., which has been published in final form at https://dx.doi.org/10.1109/PCS60826.2024.10566360. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Light Field View Synthesis using Deformable Convolutional Neural Networks

Muhammad Zubair, Paulo Nunes, Caroline Conti, and Luís Ducla Soares
Instituto Universitário de Lisboa (ISCTE-IUL), Instituto de Telecomunicações, Lisbon, Portugal
Email: {muhammad.zubair, paulo.nunes, caroline.conti, lds}@lx.it.pt

***Abstract*—Light Field (LF) imaging has emerged as a technology that can simultaneously capture both intensity values and directions of light rays from real-world scenes. Densely sampled LFs are drawing increased attention for their wide application in 3D reconstruction, depth estimation, and digital refocusing. In order to synthesize additional views to obtain a LF with higher angular resolution, many learning-based methods have been proposed. This paper follows a similar approach to Liu et al. [1] but using deformable convolutions to improve the view synthesis performance and depth-wise separable convolutions to reduce the amount of model parameters. The proposed framework consists of two main modules: i) a multi-representation view synthesis module to extract features from different LF representations of the sparse LF, and ii) a geometry-aware refinement module to synthesize a dense LF by exploring the structural characteristics of the corresponding sparse LF. Experimental results over various benchmarks demonstrate the superiority of the proposed method when compared to state-of-the-art ones. The code is available at https://github.com/MSP-IUL/deformable_lfvs.**

***Index Terms*—light field view synthesis, deformable convolution, depth-wise separable convolution, geometry-aware network**

I. Introduction

Light Field (LF) imaging is a technology that records both spatial and angular information of light rays. Due to the angular information, LF imaging can be used for 3D reconstruction [2], depth estimation [3], and post-capture image refocusing [4]. Conventional LF acquisition devices include camera arrays [5], computer-controlled gantries [6], and plenoptic cameras [7], [8]. However, LF acquisition from camera arrays and computer-controlled gantries is expensive and time consuming. Plenoptic cameras perform LF sampling by inserting a micro lens array between the main lens and the sensor, thus, due to the limitation of the sensor resolution, there is an inevitable trade-off between spatial and angular resolutions in densely sampled LFs. Thus, a plethora of learning-based methods have shown promising results in LF view synthesis tasks [9], [10], [11], [12], [13]. In this context, Convolutional Neural Network (CNN) LF view synthesis methods can be roughly divided into two categories: i) depth-based, and ii) non-depth-based. While depth-based methods can have depth estimation errors and occlusions, especially in case of wide-baseline LFs, non-depth methods upsample the angular information without considering the geometry of the scene.

This work was supported by the Fundação para a Ciência e a Tecnologia (FCT)/Ministério da Ciência, Tecnologia e Ensino Superior (MCTES) through national funds under Project UIDB/50008/2020 and Project PTDC/EEI-COM/7096/2020.

Depth-based methods warp existing LF views to novel views according to estimated depth/disparity information. Kalantari et al. [9] used two sequential CNNs to estimate disparity and color information simultaneously for high quality LF views synthesis at arbitrary locations. However, the correlation between synthesized views was ignored due their arbitrary locations. Jin et al. [14] used a CNN-based depth estimation module for explicitly modeling the scene geometry and a warping module for novel view synthesis. This method introduces a novel loss function to promote the preservation of the LF parallax structure. A common method for depth-dependent methods is feature extraction of input views by using a CNN. However, for randomly sampled Sub-Aperture Image (SAI) inputs, the angular position always varies, which makes it difficult for the network to provide sampling and target positions. In order to handle this issue, Jin et al. [15] synthesize the LF in a coarse-to-fine manner, where 4D disparity maps are estimated by using the Plane-Sweep Volumes (PSVs). Confidence maps are then utilized to evaluate and merge the warped views. The final densely-sampled LF is refined by a refinement network. Liu et al. [1] proposed a CNN-based LF view synthesis method using Multi-Representation View Reconstruction (MRVR) and Geometry-Assisted Refinement (GAR) modules, where the first module explores various alternative LF representations to synthesize dense LF views, while the second module enhances the view synthesis performance by exploring the structural characteristics of the LF scene.

Non-depth-based methods synthesize LF views by exploring local LF information, such as neighboring views, Epipolar Plane Images (EPIs) or EPI volumes, through traditional signal processing [16], [17], or learning-based methods [18], [19], [20]. Yoon et al. [18] used two cascaded CNN networks to upsample the spatial and angular resolutions simultaneously, where each SAI is first upsampled and then novel views are generated between the upsampled SAIs. Yang et al. [20] proposed an end-to-end LF angular super-resolution network that explicitly uses structure and scene information, which is divided into three parts: i) a feature extraction network, ii) an upsampling network, and iii) a refinement network. A LF Residual Block (LF-ResBlock) is used to extract EPI and scene features. The upsampling network is used to generate the intermediate LF and the refinement network, consisting of LF-ResBlocks that exploit the LF structure and scene information from the intermediate LFs, is used to synthesize the final LF views.

The new LF view synthesis method proposed in this paper is derived from Liu's et al. framework [1] introducing a rich LF representation by using deformable convolutions [21] for the Angular Synthesis Branch (ASB) and depth-wise separable convolutions for the Spatial Synthesis Branch (SSB), while keeping the Pseudo-Temporal Synthesis Branch (PTSB) intact. In [1], the SAI-array-based synthesis pipeline has complex CNN branches with large receptive fields to cover the large spatial context for the disparity estimation network. CNNs are inherently limited to model geometric transformations due to the fixed kernel geometric structure [21]. To alleviate this dilemma of the disparity network with large receptive fields in depth dependent view synthesis tasks, the proposed method uses 2D deformable CNNs [21] that add a 2D offset to each grid sampling point and allow the sampling grid to deform freely. Additionally, the proposed view synthesis method also uses depth-wise separable convolutions for the GAR module to reduce the number of model parameters. Chollet et al. [22] claimed that depth-wise separable convolutions have a smaller number of parameters with same computational efficiency because traditional convolution connects all input channels to all output channels in a single operation, while depth-wise separable convolution first processes each input channel separately and then combines the results through point-wise convolution. Therefore, the main contributions of this paper are as follows:

- We designed 2D deformable convolutional layers to model the disparity estimation network instead of multi-branches or cascaded CNN layers with large receptive fields.
- We designed depth-wise separable convolutional layers for the SSB and GAR module to reduce the number of model parameters.

II. Proposed Method

A LF can be expressed by a 4D function $L(u, v, x, y)$, where (u, v) and (x, y) denote, respectively, the angular and spatial coordinates, that can be represented as a $U \times V$ array of LF views (or SAIs), where each LF view, $L(u, v, :, :)$, has a spatial resolution of $X \times Y$. The goal of the proposed method, presented in the following subsections, is to synthesize a high angular resolution array of $U \times V$ LF views, denoted as $\hat{L}(u, v, x, y)$, from a low angular resolution set of $U' \times V'$ LF views, denoted as the reference LF, $L_{ref}(u', v', x, y)$.

A. Multi-Representation View Synthesis (MRVS)

The MRVS module consists of three view synthesis branches: i) ASB, ii) SSB, and iii) PTSB (see Fig. 1). Each view synthesis branch uses as input the reference LF in a different representation format, respectively: (i) the original low angular resolution set of LF views, L_{ref}, (ii) a low resolution lenslet image, M_{ref}, and (iii) a low temporal resolution Pseudo Video Sequence (PVS), P_{ref}. The M_{ref} is composed of a $X \times Y$ array of $U' \times V'$ Micro-Images (MIs) that are constructed by extracting one pixel from the same relative (x, y) position from L_{ref}. P_{ref} is obtained by rearranging the SAIs from L_{ref} into a video sequence in a zig-zag manner.

The output of these view synthesis branches are then concatenated and fed into a feature fusion network to obtain high angular resolution LF views, $L(u, v, x, y)$. The MRVS can be described by the following equation:

$$\begin{aligned} \widetilde{L}(u, v, x, y) &= f\left(\left(L_{ref}, M_{ref}, P_{ref}\right), \Theta^*\right) \\ \Theta^* &= \underset{\Theta}{\operatorname{argmin}} \left\| L_{GT}(u, v, x, y) - L(u, v, x, y) \right\| \end{aligned} \quad (1)$$

In (1), $L_{GT}(u, v, x, y)$ represents the Ground Truth (GT) LF, $f(\cdot)$ represents the mapping from low angular LF to high angular LF view synthesis, $\| \cdot \|$ is the L_1 loss function and Θ represents the set of network parameters.

1) Angular Synthesis Branch (ASB): In ASB, a disparity-based warping approach is adopted to explore the LF angular geometry correlation between adjacent SAIs. To estimate the disparity, differently from [1], which is more focused on multi-branch convolution networks, we use seven 2D deformable convolution layers to model geometric transformations with more degrees of freedom instead of pure translations. Conventional convolution operations use fixed filters, which limits their ability to capture geometric variations or deformations in the input data. The proposed 2D deformable disparity network has kernel sizes $\{11, 7, 5, 5, 3, 3, 3\}$, with each layer being followed by a ReLU activation function, except the last one. The resulting disparity is then passed to the backward warping module to generate a high angular resolution LF from the set of $U' \times V'$ SAIs in the low resolution array L_{ref} as follows:

$$\widetilde{L}^{(\mathbf{u}')}(\mathbf{x}, \mathbf{u}) = L_{ref}\left(\mathbf{x} + \mathbf{d}(\mathbf{x}, \mathbf{u}) \cdot (\mathbf{u} - \mathbf{u}'), \mathbf{u}'\right) \quad (2)$$

In (2), $\widetilde{L}^{(\mathbf{u}')}(:, \mathbf{u})$ represents a novel SAI at angular position $\mathbf{u} = (u, v)$ that is synthesized from the SAI with angular position $\mathbf{u}' = (u', v')$ in L_{ref}. $L_{ref}(:, \mathbf{u}')$ represents the SAI from L_{ref} that appears at angular position $\mathbf{u}' = (u', v')$; $\mathbf{x} = (x, y)$ corresponds to the spatial coordinates. The output of this branch, $\widetilde{L}_{ASB} \in \mathbb{R}^{U \times V \times X \times Y \times \left(U' \times V'\right)}$, corresponds to rich LF angular information.

2) Spatial Synthesis Branch (SSB): In SSB, a depth-wise separable convolution network is used to explore the local-spatial correlation inside MIs from the low resolution LF lenslet image, M_{ref}. The network consists of six 3D depth-wise separable convolution layers because depth-wise separable convolution splits the computation into two steps: (i) depth-wise convolution that applies a single convolutional filter per each input channel, and (ii) point-wise convolution that is used to create a linear combination of the output of the depth-wise convolution and one pixel shuffle layer. The pixel shuffle layer is used to upsample the number of views. Further, four 2D Residual Blocks (ResBlocks) are used to fuse the learned local-spatial features from MIs and reordered into SAIs. The output of this branch is a high angular resolution LF, $\widetilde{L}_{SSB} \in \mathbb{R}^{U \times V \times X \times Y}$.

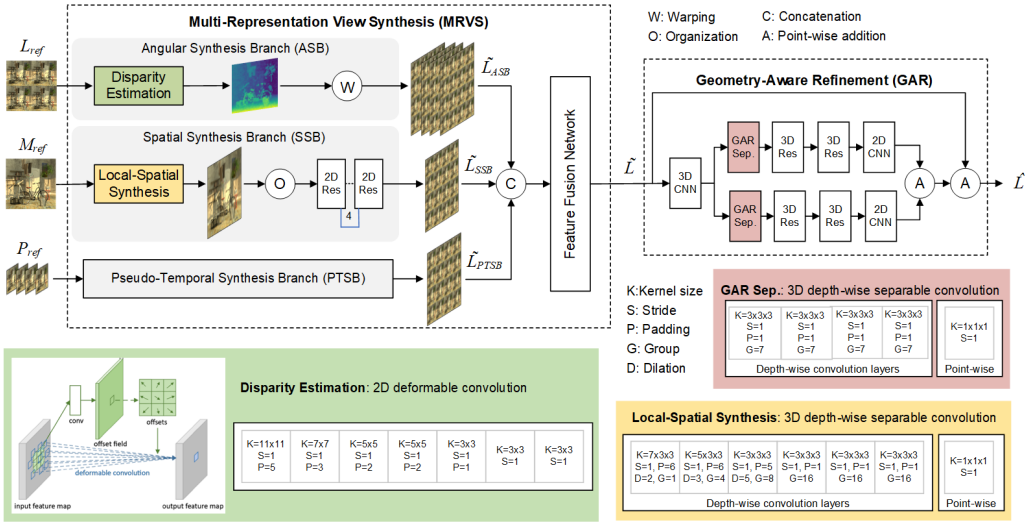


Fig. 1: Architecture of the proposed view synthesis method.

3) Pseudo-Temporal Synthesis Branch (PTSB): The PTSB module is borrowed from [1] and is used to further exploit the angular correlation of LFs as pseudo-temporal information. The low resolution PVS P_{ref} is fed into the PTSB network, which consists of two dilated 3D convolution layers and four 3D convolution layers. Further, four 3D ResBlocks and one 3D convolution network are then followed to restore more texture details and reshaped into SAIs to obtain a high angular resolution LF, $\widetilde{L}_{PTSB} \in \mathbb{R}^{U \times V \times X \times Y}$.

4) Feature Fusion Network: The angular geometry information in $\widetilde{L}_{ASB}$, the spatial information in $\widetilde{L}_{SSB}$, and correlation between adjacent SAIs in $\widetilde{L}_{PTSB}$, are then concatenated and fed to a feature fusion network, as in [1]. This network consists of three 2D convolution layers, to extract further spatial features, and four 3D ResBlock convolutional layers, that are used to fuse the high dimensional features from the three multi-representation branches. The output of this process is the Dense image Stack (DeS), $\widetilde{L}(u, v, x, y) \in \mathbb{R}^{U \times V \times X \times Y}$.

B. Geometry-Aware Refinement (GAR)

This module enhances the LF views extracted from the DeS using a 3D convolution with kernel-size 1 to constitute a bidirectional view stack structure by stacking SAIs from horizontal and vertical directions. Then 3D depth-wise separable convolution network is proposed to enhance the view synthesis from the stacked SAIs. Further, two 3D ResBlocks, followed by a 2D CNN with three layers, are used to improve texture details. At the end, one by one summation strategy is used to produce the high angular resolution LF views, $\hat{L}(u, v, x, y)$.

III. Experimental Results

A. Datasets and Training Details

Two end-to-end network models were trained on the following datasets (see Table I): (i) HCInew (20 scenes) – synthetic LFs [23], and (ii) UCSD Training – real-world LFs captured with a Lytro Illum camera [9]. Each LF was cropped to the central 7×7 views and their 4 corner views were extracted to be used as 2×2 low angular resolution LF input (only the Y channel was considered). Each view was cropped into 96×96 patches to generate the training samples. The batch-size was set to 1 and the networks were trained using the L_1 loss function and optimized using the Adam method [24]. The learning rate was initialized to 10^{-4} and halved every 5×10^3 epochs. For testing, three real-world and two synthetic LF datasets were adopted (Table I); the average Y Peak Signal-to-Noise Ratio (PSNR-Y) and the Structural Similarity Index Measure (SSIM) for the synthesized views were used.

TABLE I: Training and testing LF datasets.

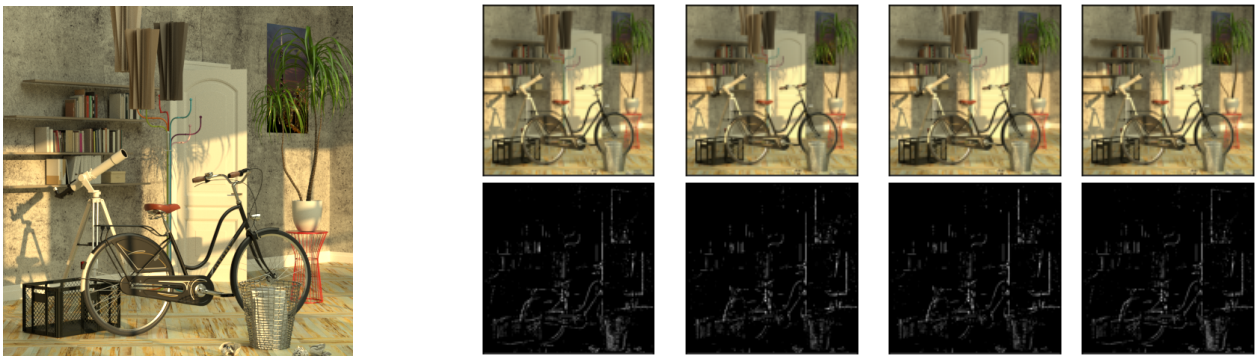
LF dataset	Type	GT depth	Angular resolution	Spatial resolution	Training scenes	Testing scenes
HCInew [23]	Synthetic	✓	9×9	512×512	20	4
HCIold [25]	Synthetic	✓	9×9	768×768, 1024×576	-	5
UCSD Training [9]	Real	-	14×14	376×541	100	-
UCSD Testing [9]	Real	-	14×14	376×541	-	30
INRIA Occlusions [26]	Real	-	14×14	376×541	-	25
INRIA Reflective [26]	Real	-	14×14	376×541	-	15

B. Comparative Analysis

Four state-of-the-art view synthesis methods are compared with proposed method: (i) geometry-assisted multi-representation view reconstruction (GA-MRVR) [1], (ii) learning LF angular super resolution via geometry-aware network (LFASR-Geo) [14], (iii) deep coarse-to-fine dense LF reconstruction with flexible sampling and geometry aware fusion (FS-GAF) [15], and (iv) light field angular super-resolution based on structure and scene information (LFASR-SSI) [20]. Tables II and III show that the proposed method has higher performance in terms of PSNR-Y and SIMM on all benchmark methods, except for FS-GAF on synthetic datasets, which uses PSVs for depth inference. Fig. 2 and Fig. 3 illustrate the qualitative results and Table IV shows the number of model parameters for each method under analysis.

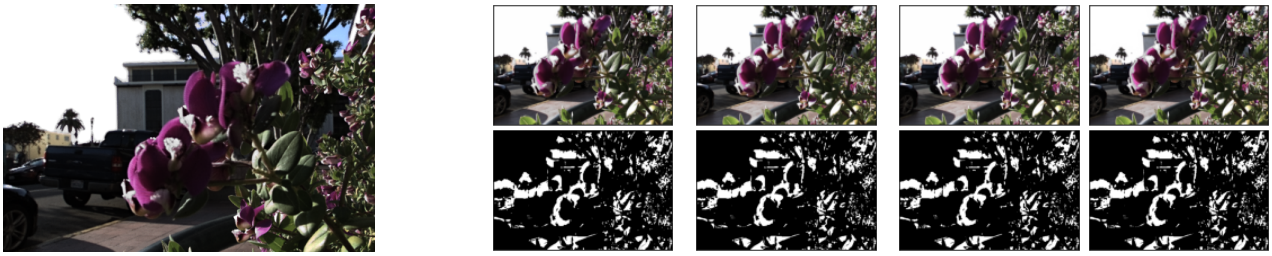
C. Ablation Study

This section presents several experiments to provide deep insight into the proposed method. For this, different convolution neural network variants and LF representations are



(a) (b) (c) (d) (e)

Fig. 2: Qualitative results on a synthetic LF: (a) GT; (b) GA-MRVR [1]; (c) LFASR-Geo [14]; (d) FS-GAF [15]; (e) Proposed.



(a) (b) (c) (d) (e)

Fig. 3: Qualitative results on a real-world LF: (a) GT; (b) GA-MRVR [1]; (c) LFASR-Geo [14]; (d) FS-GAF [15]; (e) Proposed.

TABLE II: Quantitative results on synthetic datasets.

Test sets	GA-MRVR [1]		LFASR-Geo [14]		FS-GAF [15]		LFASR-SSI [20]		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
HCInew	34.31	0.935	33.76	0.932	**37.14**	**0.966**	33.24	0.911	34.50	0.937
HCIold	40.84	0.954	40.84	0.954	**41.80**	**0.974**	40.53	0.957	41.25	0.958
Average	37.58	0.945	37.30	0.943	**39.47**	**0.970**	36.89	0.934	37.88	0.947

TABLE III: Quantitative results on real-world datasets.

Test sets	GA-MRVR [1]		LFASR-Geo [14]		FS-GAF [15]		LFASR-SSI [20]		Proposed	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
30scenes	42.91	**0.987**	42.53	0.985	42.75	0.986	42.52	0.986	**43.19**	**0.987**
Occlusion	39.06	0.981	38.36	0.977	38.51	0.979	38.63	0.980	**39.26**	**0.982**
Reflective	39.04	0.962	38.20	0.955	38.35	0.955	38.41	0.960	**39.20**	**0.963**
Average	40.34	0.977	39.70	0.972	39.87	0.973	39.85	0.975	**40.55**	**0.977**

TABLE IV: Number of model parameters ($\times 10^6$).

GA-MRVR [1]	LFASR-Geo [14]	FS-GAF [15]	LFASR-SSI [20]	Proposed
2.4996	1.0113	1.5353	0.7145	2.3700

used. In Var-I, we use deformable convolution networks for disparity estimation and backward warping. In this depth-dependent method, we use a SAI LF representation. In Var-II, we use the 3D convolution on MIs and upsample the angular resolution by using a pixel shuffle layer. In order to find the LF angular-spatial correlation, we use the PVS representation with 3D convolution in Var-III. We also test a Convolution Block Attention Module (CBAM) in depth-dependent method of SAI representation in Var-IV. During these experiments, we find that a "rich" LF representation plays a vital role in LF view synthesis. In Var-V, we use SAI, MIs and PVS LF representation and and deformable convolution networks. However, these approaches revealed to be too complex in terms of number of model parameters. Therefore, in the proposed solution, we take advantage of a "rich" LF representation and use different deformable convolution neural networks to achieve high angular resolution, with a similar number of model parameters to competing methods. Table V shows the overall results of this ablation study.

TABLE V: Ablation results on HCInew scenes.

Ablation Variants	Method	LF representation	HCInew	
			PSNR	SSIM
Var-I	2D deformable convolution network	SAIs	34.48	0.936
Var-II	3D convolution network	MIs	34.13	0.932
Var-III	3D convolution network	PVS	33.87	0.925
Vari-IV	2D deformable convolution and CBAM	SAIs	34.02	0.933
Vari-V	2D deformable convolution network	SAIs, MIs, PVS	32.83	0.919
Proposed network	2D deformable and depth-wise separable convolution networks	SAIs, MIs, PVS	34.50	0.937

IV. Conclusions

This paper addressed the problem of LF view synthesis by combining deformable convolutions and depth-wise separable convolutions in a multiple-representation view synthesis (MRVS) module and a geometry-aware refinement (GAR) module. The first module extracts structural characteristics of LF by exploring spatial geometry information in SAI images, abundant angular information in lenslet image, and adjacent correlation of SAIs in PVS. The GAR module improves the reconstruction performance by restoring more texture detail and fusing rich LF structural information. The experimental results show the advantages of the proposed approach relatively to benchmark solutions. In the future, we will extend the proposed approach to handle more challenging scenes, notably with wide baselines.

REFERENCES

[1] 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. [Online]. Available: https://doi.org/10.1016/j.knosys.2023.110390

[2] J. Wu, Z. Lu, D. Jiang, Y. Guo, H. Qiao, Y. Zhang, T. Zhu, Y. Cai, X. Zhang, K. Zhanghao, H. Xie, T. Yan, G. Zhang, X. Li, Z. Jiang, X. Lin, L. Fang, B. Zhou, P. Xi, J. Fan, L. Yu, and Q. Dai, "Iterative tomography with digital adaptive optics permits hour-long intravital observation of 3D subcellular dynamics at millisecond scale," *Cell*, vol. 184, no. 12, pp. 3318–3332.e17, 2021.

[3] 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.

[4] 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, 2019.

[5] 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," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, 2005.

[6] V. Vaish and A. Adams, "The (new) stanford light field archive." 2018. [Online]. Available: http://lightfield.stanford.edu/acq.html

[7] "Lytro," 2016. [Online]. Available: https://www.lytro.com/

[8] "Raytrix," 2016. [Online]. Available: https://www.raytrix.de

[9] 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.

[10] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-End View Synthesis for Light Field Imaging with Pseudo 4DCNN," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11206 LNCS, pp. 340–355, 2018.

[11] 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, 2019.

[12] L. Liu, Z. Nian, and C. Jung, "DCM-CNN: Densely connected multiloss convolutional neural networks for light field view synthesis," *IEEE Access*, vol. 8, pp. 78 542–78 552, 2020.

[13] N. Meng, H. K. 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.

[14] 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. 11 141–11 148, 2020.

[15] 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.

[16] 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.

[17] 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.

[18] 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," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015-February, pp. 57–65, 2016.

[19] 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," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11210 LNCS, pp. 138–154, 2018.

[20] J. Yang, L. Wang, L. Ren, Y. Cao, and Y. Cao, "Light field angular super-resolution based on structure and scene information," *Applied Intelligence*, vol. 53, no. 4, pp. 4767–4783, 2023.

[21] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable Convolutional Networks," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 764–773, 2017.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 1800–1807, 2017.

[23] 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.

[24] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.

[25] 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.

[26] (2016) Stanford lytro light field archive. kernel description. [Online]. Available: http://lightfields.stanford.ed