IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 January 2024, accepted 16 April 2024, date of publication 25 April 2024, date of current version 14 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3393576

RESEARCH ARTICLE

LFDA: A Framework for Light Field Depth Estimation With Depth Attention

HYEONGSIK KIM, SEUNGJIN HAN, AND YOUNGSEOP KIM
Department of Electronics and Electrical Engineering, Dankook University, Yongin, Gyeonggi-do 16890, South Korea

Corresponding author: Youngseop Kim (wangcho@dankook.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by Korean Government the Ministry of Science, ICT, and Future Planning (MSIP) under Grant NRF-2020R1A2C2009717.

ABSTRACT Depth estimation in light field imaging is integral for the accurate rendering of 3D scenes and a crucial task in light field applications. However, the development of a model that simultaneously achieves high accuracy and speed in light field depth estimation remains a significant challenge. Existing networks utilizing dilated convolution achieve state-of-the-art speeds, but they often encounter accuracy limitations, particularly in fine-grained details. In this paper, we introduce a fast and accurate method based on depth-wise cross attention. By integrating cross-attention with existing networks, our approach effectively emphasizes local features, thereby overcoming the accuracy limitations commonly encountered. Our method adopts depth attention to compare the center and side views along the epipolar line. As a result of depth attention, the cost volume was aggregated by similarity information that was based on the attention score. This technique not only maintains computational efficiency but also significantly enhances the performance in fine-grained regions by emphasizing the importance of local feature analysis. We validated the efficacy of depth attention in emphasizing local features. Our experiments were conducted using the 4D HCI Benchmark, employing evaluation metrics such as *BadPixel* and MSE. The results demonstrate remarkable performance in estimating fine depth changes, primarily due to the focus on local features, thereby offering a balanced solution in terms of both speed and accuracy. The code is available: https://github.com/syt06007/LFDA.

INDEX TERMS Light field, depth estimation, attention, deep learning.

I. INTRODUCTION

Light field (LF) cameras, unlike conventional cameras, are capable of capturing the real world by encoding it in both spatial and angular resolutions [1]. This unique ability enables the acquisition of sub aperture images (SAIs), which support a variety of applications, including refocusing, 3D reconstruction and also LF display [2], [3], [4]. Since depth estimation is a mid-level process of LF imaging, which is the basis of various LF algorithms, fast and accurate depth information estimation is important [5].

With advances in deep learning and computer vision, a variety of depth estimation methodologies are being explored in the field of LF. Among these, epipolar plane image (EPI) [6], [7], [8], [9], [10] and multi view stereo (MVS) [11], [12], [13], [14] methodologies have been prominent. Traditionally, the EPI approach utilized line orientations in EPIs to estimate depth, leveraging the intuitive information from LF SAIs. However, this method faces challenges in real-world application due to the low spatial resolution of LF and its vulnerability to occlusion, often resulting in lower accuracy [15]. On the other hand, LF depth estimation methodologies have increasingly adopted the MVS approach [16], commonly used in multi-image depth estimation. MVS methodologies involve feature extraction, cost volume construction, cost aggregation and depth regression. These methodologies generally exhibit higher accuracy compared to the EPI method. In LF imaging, the requirement for a large number of SAIs poses a significant computational speed challenge [14]. Dilated convolution [17], as utilized in cost volume construction [14], substantially reduces computational time by addressing this challenge. However, while dilated convolution reduces computational time, it faces limitations associated with the integer dilation rate [18]. This

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

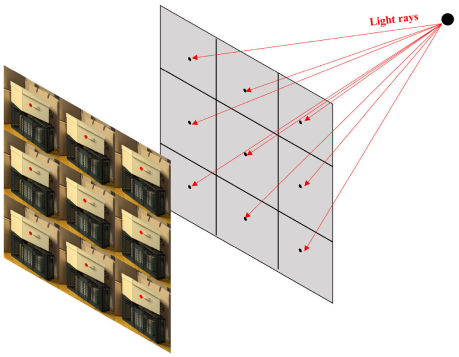

FIGURE 1. Example of SAI array generated light rays on the 3 × 3 micro lens array.

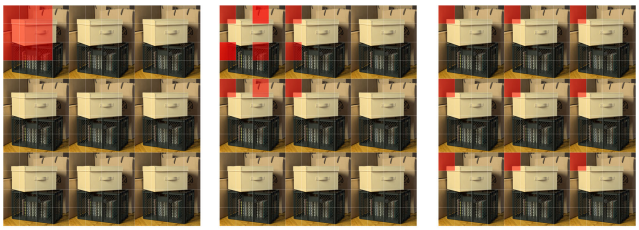

FIGURE 2. Visualization of the dilated convolution filter according to the dilation rate. This figure illustrates a 3 × 3 dilated convolution applied to a 3 × 3 SAI array to create an angular patch. Parts (a) and (b) show aliasing due to an insufficiently small filter gap.

limitation affects the accuracy at the sub-pixel level, making it less precise than other models.

Recently, attention mechanisms [19], initially developed for natural language processing, have been applied in various vision domains [20], [21], [22], [23], [24], including depth estimation. These mechanisms function by calculating the similarity of input vectors through the dot product, similar to how optimal pixel matching is essential for depth estimation. Cross attention enhances depth accuracy by establishing a 3D association between the center and side views along the epipolar line [23]. Despite its potential, the widespread adoption of transformer-based attention in LF depth estimation is limited due to the large number of SAIs involved.

To handle two aforementioned challenges, we introduce the light field depth attention (LFDA), a novel architecture that integrates depth attention into existing networks. Depth attention is conducted with cost volume constructed by dilated convolution. Fig. 1 and Fig. 2 illustrate how angular patches in the SAI are acquired using dilated convolution. When constructing a cost volume, dilated convolution generates angular patches by collecting pixels from the SAIs. By varying the dilation rate across depth candidates within a predefined depth range, the generated angular patches are stacked to construct a cost volume. After this, similarity comparison is performed between center view pixel of SAI array and cost volume. A key advantage of transformer attention is that it leverages cross-attention to build correlations among the SAIs. This method emphasizes local features between images, enhancing both accuracy and computational efficiency in depth estimation [33]. Our paper presents the development of a depth attention layer capable of cost volume matching, transitioning from traditional patch-based pixel matching to an attention layer. Additionally, we aim to overcome existing challenges in LF depth estimation by introducing a novel approach that combines the benefits of attention layers with advanced cost volume construction. The LFDA network enhances the accuracy and speed of traditional methods through the application of modern deep learning approach.

The contributions of LFDA are as follows:

- **Enhancing light field depth estimation through precise pixel matching**: This approach highlights the importance of precise pixel matching, where the depth attention mechanism plays a key role in identifying and emphasizing the most relevant features for accurate depth estimation. As a result, LFDA effectively enhances the accuracy of depth estimation in areas based on local features while maintaining computational efficiency.
- **Utilizing cross attention structure**: LFDA leverages cross attention via a cost volume constructed by dilated convolution, enabling a focused comparison of pixel similarities across all depth candidates. This approach not only allows for more precise pixel matching but also overcomes the challenge of applying attention in the presence of a large number of light field SAIs.

These contributions highlight significant advancements of our model in LF depth estimation, offering improved accuracy while maintaining efficiency.

II. RELATED WORK

In this section, we review the previous approach of LF depth estimation, ranging from traditional methods to deep learning-based methods involving EPI and MVS.

Conventionally, LF depth estimation has been approached through various methodologies. Wanner and Goldluecke [25] proposed a algorithm utilizing 2D structure tensor to estimate the slop of line within EPIs. Tao et al. [26] proposed algorithm that combines defocus and correspondence depth cues. Zhang et al. [27] proposed a spinning parallelogram operator (SPO) that utilize the regions divided in EPI to estimate line orientation. Jeon et al. [16] adopted a multi view stereo method using phase shift on sub-pixel level. Williem and Park [34] proposed a novel method that robust occlusion and noisy scene using angular entropy metric and adaptive refocus response.

Deep learning-based approaches have been introduced in recent years. In particular, learning-based EPI method has been widely used since the beginning of deep learning adoption because it could be used by replacing the existing

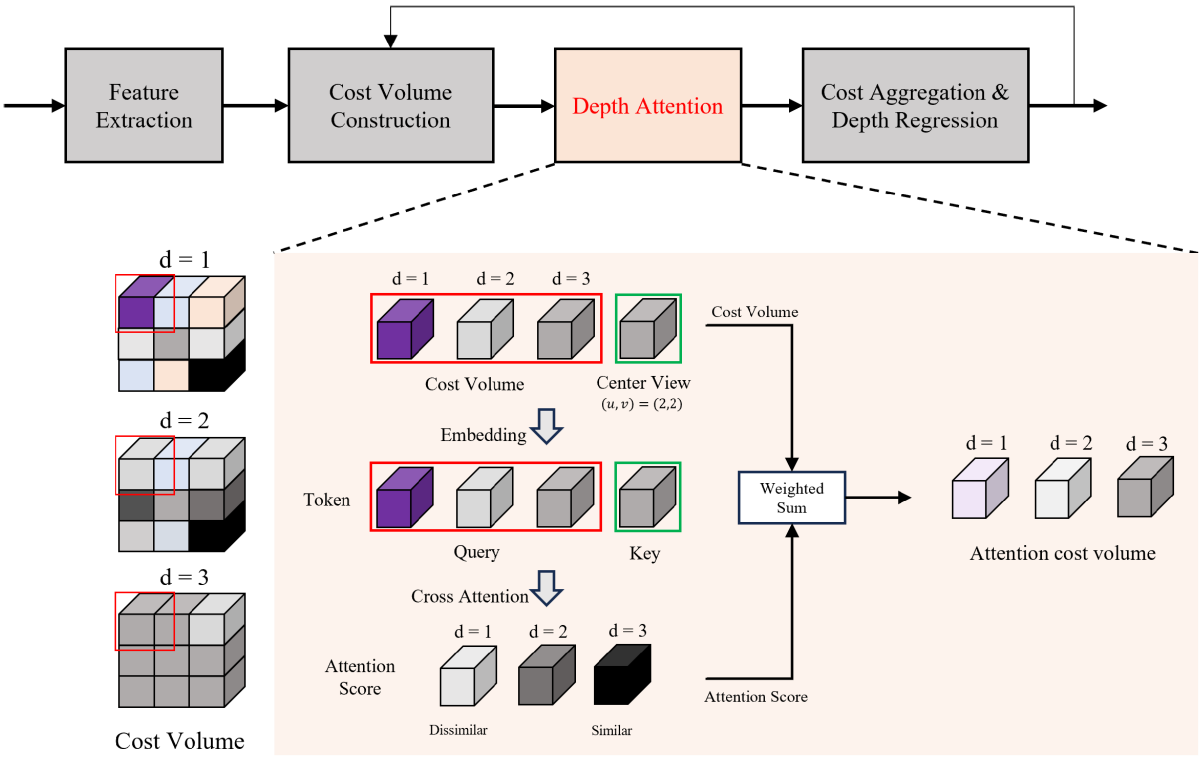

FIGURE 3. A toy example of depth attention using 3 × 3 SAIs. This figure illustrates the depth attention process. It starts with generating the cost volume through depth candidates. Next, the cost volume and the center view are tokenized, followed by conducting a vector dot product. Finally, the obtained attention score is applied as a weighted sum to the cost volume.

methodology with the CNN network. Heber and Pock [8] used end-to-end network to learn each EPI representation seperatly. Also Heber et al. [9] analyze entire EPI using U-shape 3D CNN Network to extract EPI orientation of SAIs. Feng et al. [28] proposed a two-stream network specifically to learn the association between neighboring pixels in EPIs. Anna Alperovich at al. [7] use the autoencoder design for light field encodes horizontal and vertical EPI stacks simultaneously using six stages of residual blocks. Shin et al. [10] proposed multi-stream architecture to analyze each streams of EPI and also they proposed a novel data augmentation for network training. Leistner et al. [29] proposed EPI-shift strategy to retain a small receptive field in wide-baseline EPIs using U-Net [30] architecture. Li et al. [6] proposed oriented relation module to estimate the depth of intersection point on horizontal and vertical EPIs.

Based on deep learning, multi view stereo approach in LF is introduced by Tsai et al. [11]. This method proposed an channel attention based view selection network to utilize all views more effectively and efficiently. Liu at al. [35] proposed novel feature extraction based on dilated-convolution and channel attention for disparity regression. Chen at al. [12] proposed attention-based multi-level fusion network that designed intra and inter branches hierarchically to select views with less occlusions and richer textures. Huang et al. [13] proposed a fast and lightweight disparity estimation model with multi-disparity-scale using sub-network of edge guidance to achieve fast LF depth estimation. Wang et al. [14] proposed occlusion-aware cost constructor with dilated convolution that can achieve fast computational time and high accuracy by occlusion handling via pixel modulation.

Deep learning-based MVS methodologies have achieved high accuracy in various scenarios by learning detailed depth information between each SAIs. Despite this high accuracy, there is still space for improvement in terms of running time, this indicates that although the MVS approach is effective for depth estimation, there is still room for improvement in computational efficiency.

III. METHOD

A. DEPTH ATTENTION

Our method integrates the traditional patch matching depth estimation approach with the existing model, utilizing deep learning techniques. In this section, we first describe the construction of the cost volume, followed by an explanation of the depth-wise cross-attention that emphasizes local features through similarity information on the cost volume. The process of using dilated convolution in the cost volume construction process is mostly similar to the structure of OACC-Net [14]. The overall architecture is illustrated in Fig. 3.

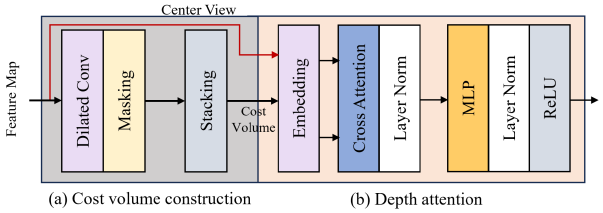

FIGURE 4. Architecture of LFDA mechanism. In this figure depicts detailed structure of (a) cost construction and (b) depth attention.

1) COST VOLUME CONSTRUCTION BY DILATED CONVOLUTION

To estimate depth in LF, we use the feature map $\mathcal{F}_{u_c,v_c}$ from the center view (reference view) as a basis for establishing depth candidates along the epipolar line of the side view (source view), thus constructing the cost volume C. Traditional methods in deep learning-based cost volume construction involve setting minimum and maximum disparities as hyperparameters, employing a shift-and-concat [11] method to warp the side view image across the disparity range.

However, this warping process becomes computationally expensive with a large number of SAIs. To address this challenge, our LF depth estimation approach utilizes a dilated convolution technique tailored to the angular resolution of the SAIs. Dilated convolution, similar to basic convolution but featuring adjustable filter spacing, enables efficient pixel matching across the arranged images $\mathcal{F}_{u,v}$. $\mathcal{F}_{u,v}$ derived from feature extraction is aligned according to angular resolution. Subsequently, patches corresponding to the filter positions are obtained by applying dilation to the CNN filter, sized $U \times V$, equivalent to the angular resolution. The pixels resulting from this process form an array $\mathcal{A}_{h,w,d} \in \mathbb{R}^{UV \times C}$, which are referred to as angular patches.

Then, we construct the volume $C_d \in \mathbb{R}^{HW \times UV \times C}$ by aggregating angular patches for every pixel. Depth candidates $d \in \{d_{min}, \ldots, d_{max}\}$ within the specified range are considered, and the final 4-dimensional cost volume C is constructed through iterative application of CNN for each depth candidate d. The dilation rate, defined as per equation (1), adjusts the distance between filters, effectively replacing image warping. This approach provides a more efficient means of constructing the cost volume compared to traditional image warping techniques.

$$\text{Dilation\_rate}(d) = [H - d, W - d] \tag{1}$$

2) DEPTH-WISE CROSS-ATTENTION

In LF depth estimation, the limitation of dilation rates being integer values means that dilated convolution cannot achieve sub-pixel level shifting [18]. This limitation presents a significant challenge in achieving high accuracy in depth mapping.

The attention module in our approach plays a crucial role in overcoming this limitation. The module calculates vector similarity between the feature map $\mathcal{F}_{u_c,v_c}$ from the center view and the cost volume C. Performing an attention operation on a per-pixel basis, the module enhances depth estimation accuracy. In a manner similar to traditional matching methods, $\mathcal{F}_{u_c,v_c}$ is set as the Key, and the corresponding $C_{d_{min}\cdots d_{max}}$ as the Query. For each $\mathcal{F}_{u_c,v_c}(h, w)$, there are N_d angular patches, each with the same angular resolution (u, v). where N_d is the number of depth candidates. These are then input into the attention layer to obtain an attention score based on vector similarity. This score is multiplied by the candidate pixel to produce a weighted sum. The resulting cost volume utilizes an attention score, assigning higher weights to depths that exhibit better photometric consistency. The attention score $\alpha \in \mathbb{R}^{HW \times UV \times N_d}$ is expressed in the following equation:

$$\alpha = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{2}$$

$$C_{attended} = Concat(\alpha_1 * C_{d_{min}}, \cdots, \alpha_{N_d} * C_{d_{max}}) \tag{3}$$

where "$*$" operator denotes element-wise multiplication, a key aspect of our attention mechanism. This equation illustrates the core mechanism of our attention-based approach.

Fig. 3 illustrates a toy example of the depth attention process using a 3×3 arrayed SAIs, $\mathcal{F}_{u_c,v_c}(h, w)$. A dilated convolution with a UV-sized kernel construct the cost volume. This process generates respective angular patches $\mathcal{A}$ for each depth candidate d, forming the set of angular patches that constitute the cost volume. We set the Query as $\mathcal{A}_{h,w}(u, v, d)$ and the Key as $\mathcal{F}_{u_c,v_c}(h, w)$ for pixels with corresponding (u, v) angular resolution. Following the cross-attention process, the model computes attention scores. These are used for element-wise multiplication with the existing cost volume to create an attention cost volume containing similarity information.

B. LFDA NETWORK DESIGN

In this section, we introduce the overall architecture of our network, which is based on the depth attention concept discussed in the previous section. Our network is comprised of four key stages: feature extraction, cost volume construction, depth attention, and cost aggregation & depth regression. In the feature extraction stage, a residual network module [31] extracts features focusing on depth cues, while the cost volume is constructed via dilated convolution in the cost volume stage. The depth attention stage emphasizes the local feature of the cost volume through similarity comparison. Finally, the cost aggregation & depth regression stage refines the cost volume and generates the depth map. We adopt methodologies presented in previous studies [11], [14] for feature extraction, cost aggregation & depth regression, and, to some extent, for cost volume construction.

More details are as follows:

TABLE 1. Metric scores of models. The best score is expressed in bold.

Model	*BadPix*(0.01)												
	backgammon	dots	pyramids	stripes	boxes	cotton	dino	sideboard	bedroom	bicycle	herbs	origami	Avg.
CAE	17.32	83.70	27.54	39.95	72.69	59.22	61.06	56.92	68.59	59.64	59.24	64.16	55.84
SPO	49.94	58.07	79.20	21.87	73.23	69.05	69.87	73.36	72.37	71.13	86.62	75.58	66.70
Epinet-fcn-m	19.43	35.61	11.42	**11.77**	46.09	25.72	19.39	36.49	31.82	42.83	59.93	42.21	31.90
FastLFnet	39.84	68.15	22.19	63.04	71.82	49.34	56.24	61.96	52.88	59.24	59.98	72.36	56.45
OACC-Net	21.61	**21.02**	3.852	15.24	43.48	10.45	22.11	28.64	21.97	32.74	86.41	32.25	28.32
LFDA(Ours)	**17.21**	23.29	**2.60**	18.47	**38.22**	**4.72**	**14.82**	**23.31**	**15.71**	34.37	**33.09**	**30.60**	**21.37**

Model	MSE×100												
	backgammon	dots	pyramids	stripes	boxes	cotton	dino	sideboard	bedroom	bicycle	herbs	origami	Avg.
CAE	6.074	5.082	0.048	3.556	8.424	1.506	0.382	0.876	0.234	5.135	11.67	1.778	3.730
SPO	4.587	5.238	0.043	6.955	9.107	1.313	0.310	1.024	0.209	5.570	11.23	2.032	3.968
Epinet-fcn-m	**3.705**	1.475	0.007	0.932	5.968	0.197	0.157	0.798	0.204	4.603	9.491	1.478	2.418
FastLFnet	3.986	3.407	0.018	0.984	4.395	0.322	0.189	0.747	0.202	4.715	8.285	2.228	2.456
OACC-Net	3.938	**1.418**	**0.004**	**0.845**	**2.892**	**0.162**	**0.083**	0.542	**0.148**	**2.907**	**6.561**	**0.878**	**1.698**
LFDA(Ours)	4.907	1.638	**0.004**	1.061	3.955	0.548	0.169	**0.499**	0.308	3.447	9.472	1.975	2.332

TABLE 2. Comparison of model size and speed. FLOPs were calculated based on the 32 × 32 image patch.

	Parameters	FLOPs
LFDA	5.00 M	95.06 G
OACC-Net	5.02 M	95.14 G

1) FEATURE EXTRACTION

The feature extraction phase employs CNN layers to process SAIs, focusing on essential visual information for depth estimation, such as texture, color, and shapes. This phase employs residual blocks [31] with a structure comprising Convolution (Conv), Batch Normalization (BN), Leaky ReLU, followed by another Conv and BN sequence. Finally, we can get feature map $\mathcal{F}_{u,v} \in \mathbb{R}^{H\times W\times C}$ where H, W, C denote the height, width, and number of channels, respectively, while u and v represent the angular resolution of the SAIs.

2) COST VOLUME CONSTRUCTION

In our network, dilated convolution plays a key role in efficiently processing SAIs for the cost volume construction. This approach allows handling a range of depth candidates, creating detailed cost volume essential for accurate depth estimation.

To mitigate accuracy reduction due to occlusion, we implement a coarse-to-fine manner. An initial depth map, which is an output of the model, is utilized to identify occluded regions, followed by masking these areas in the side views during cost volume construction. This selective masking, illustrated in Fig. 4(a), refines the cost volume by focusing on relevant areas, enhancing depth estimation accuracy.

This integration of dilated convolution with the occlusion handling strategy demonstrates our ability to manage complex LF data efficiently, ensuring both speed and accuracy in depth information processing.

3) DEPTH ATTENTION

The depth attention stage is a crucial phase in overcoming accuracy issues that may arise from dilated convolution. This layer refines the cost volume attentively by focusing on local features through a depth-wise cross attention mechanism.

Initially, the cost volume and the feature map of the center view are fed into an embedding layer, which tokenizes these two inputs. Subsequently, queries and keys are fed into the cross-attention mechanism, generating and normalizing similarity scores. This process concentrates on local features, a significant advantage of traditional local matching methods, by comparing the similarity between angular patches of the cost volume and pixels of the center view. Subsequently, the inputs are processed through a feed-forward network consisting of a multi-layer perceptron, layer normalization, and ReLU activation. Through the depth attention process, C becomes C_{attended}.

4) COST AGGREGATION AND DEPTH REGRESSION

Depth is estimated through the cost volume C containing similarity information obtained by attention. Aggregation is necessary for depth candidates $C_{d_{min}} \cdots C_{d_{max}}$, hence a 3D convolution with a $3 \times 3 \times 3$ filter size is used. Initially, the number of channels is reduced using a 1×1 CNN layer, followed by the use of 3D CNN residual blocks for feature extraction, gathering depth information to create a 3D cost volume $C \in \mathbb{R}^{D\times H\times W}$. The process culminates in a regression step to generate the final depth map. The regression formula is provided as follows:

$$D_{\text{map}} = \sum_{d=d_{\text{min}}}^{d_{\text{max}}} d \times softmax(C_{\text{attended}}) \tag{4}$$

IV. EXPERIMENTS

In this section, we introduce detailed implementations of network and experiments settings, then we demonstrate aforementioned performance through comparison with other state-of-the-art models and experimental results.

A. MODEL ANALYSES

1) PERFORMANCE EVALUATION

For evaluating the performance of the model, we use metrics including $BadPix(\epsilon)$ with $\epsilon = 0.01$, as defined in [32], which

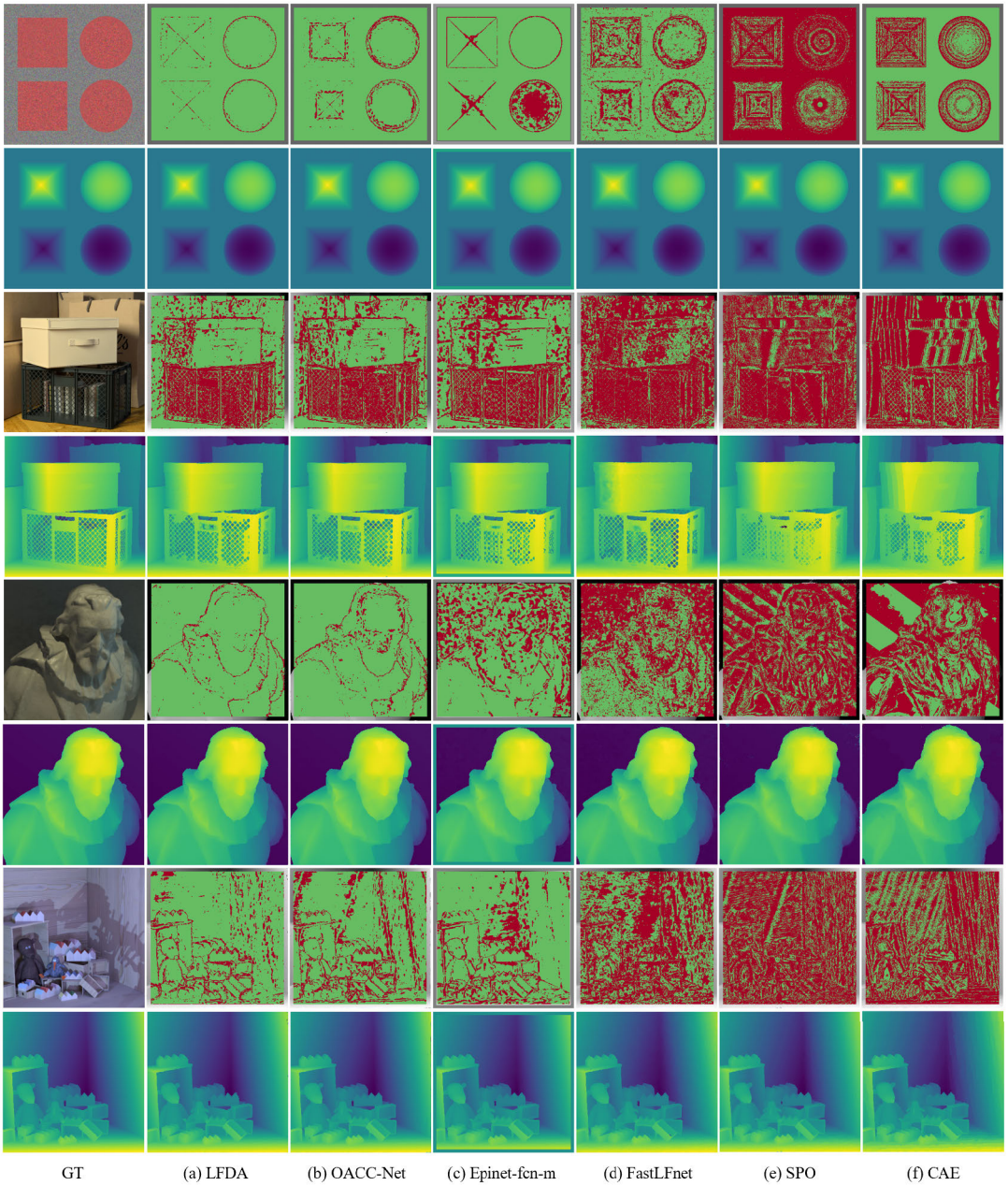

FIGURE 5. Visual comparison of validation scenes "Pyramids", "Boxes", "Cotton", "Dino" with other methods (a) LFDA, (b) OACC-Net, (c) Epinet-fcn-m, (d) FastLFnet, (e) SPO, (f) CAE. The first row of each scene represents the *BadPix*(0.01) error image, and the second row represents the disparity map. In the error image, green areas represent the correct pixel and red areas represent error pixel.

represents the percentage of pixels whose error rate exceeds a certain threshold, and the Mean Square Error (MSE) for a numerical assessment of error across all pixels.

We compared our proposed model with five different state-of-the-art methods, including OACC-Net [14], Epinet-fcn-m [10], FastLFNet [13], SPO [27], and CAE [34], using the metrics *BadPix*(0.01) and MSE. In Table 1, our method shows the best performance in 6 out of 8 scenes when evaluated using the *BadPix*0.01 metric. As illustrated in Fig. 5, the proposed method demonstrates remarkable performance in most scenes, particularly where pronounced textures and intricate depth changes are prevalent. This notable effectiveness, especially in regions with fine-depth change, stems from implementing depth-wise cross-attention in cost volume construction. Comparative experiments reveal the ability to outperform others, including OACC-Net, in texturally detailed and subtly varying depth areas.

Our model excels in the *BadPix*(0.01) metric but underperforms in the MSE metric, particularly in scenes such as 'dots' and 'stripes', which are designed to test noise and occlusion

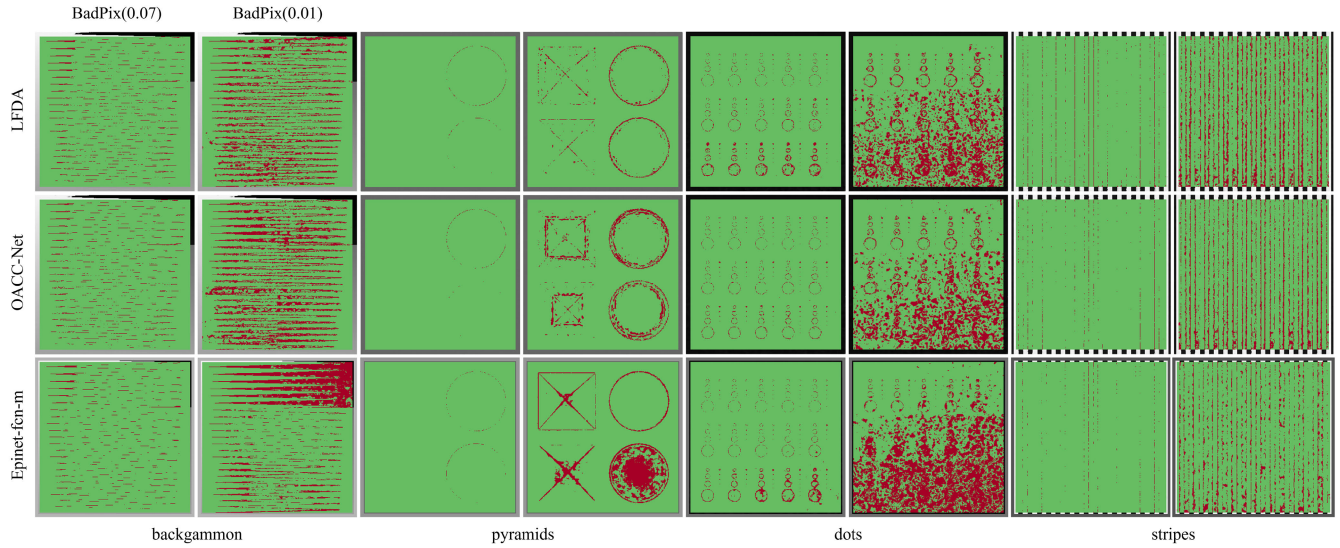

FIGURE 6. Visual comparison of *BadPixel*(0.01) and *BadPixel*(0.07) error maps for four stratified scene. For each scene, the first column is *BadPixel*(0.07) map and the second column is *BadPixel*(0.01) map.

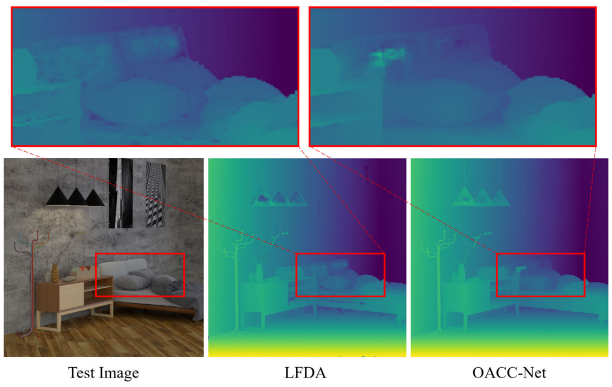

FIGURE 7. The results of depth estimation with 4D HCI test scene "bedroom". Our depth map is better than specific area remarked red box.

handling. This reflects its proficiency in capturing texture but struggles with occlusion and noise. A more detailed analysis of these limitations is presented in the 'Failure Case Analysis' section, emphasizing the balance between local feature detection and global context understanding.

Fig. 7 presents a comparison of the depth maps produced by LFDA and OACC-Net for the test scene 'bedroom'. In this figure, the area highlighted with a red box demonstrates the superior performance of our model in handling fine depth changes. This specific area in the image effectively showcases the strengths of the our approach in handling fine depth change area.

2) PARAMETERS AND FLOPS

Table 2 compares the FLOPs and Parameters of the algorithms. The FLOPs are measured for a 32 by 32 image patch, and both Parameters and FLOPs demonstrate lower memory consumption and computational efficiency compared to existing networks.

TABLE 3. This table presents the average metric values for three models across eight scenes on the 4D HCI benchmark: backgammon, dots, stripes, pyramids, boxes, cotton, dino, sideboard, bedroom, bicycle, herbs and origami.

Model	Metrics Average		
	BadPix(0.01)	BadPix(0.07)	MSE
LFDA	**21.367**	4.279	2.332
OACC-Net	28.315	**3.734**	**1.698**
Epinet-fcn-m	31.898	4.646	2.418

3) FAILURE CASE ANALYSIS

Our proposed model has demonstrated exceptional performance in estimating fine depth changes, particularly in areas rich in local features. However, it has shown less satisfactory results in regions lacking texture or with high noise levels, and near occlusion boundaries. This is evident in the *BadPix* map in Fig. 6 and the lower average scores for *BadPix*(0.07) and MSE in Table 3.

The *BadPix*(0.01) metric is sensitive to errors not only exceeding 7 percent but also to smaller discrepancies above 1 percent. Hence, while the LFDA generally performs well by accurately predicting fine depth changes, its performance under *BadPix*(0.01) is reduced in scenes like "dots" and "stripes", where noise and contrast disrupt local features throughout the image. Furthermore, since *BadPix*(0.07) focuses on errors larger than 7 percent, it tends to highlight areas with more significant error margins over those with less pronounced fine depth changes. This results in LFDA showing suboptimal performance on the *BadPix*(0.07) metric, particularly in regions where global context is crucial.

Additionally, due to the nature of the MSE metric which imposes greater penalties for larger errors, performance of LFDA appears lower in these areas.

This suggests that the model, through its depth attention process, tends to enhance local information, potentially at the expense of global context, indicating a trade-off relationship.

B. DETAILS OF TRAINING

1) DATASET AND EVALUATION METRIC

The LFDA network was trained and validated using the 4D HCI benchmark [32], which comprises entirely synthetic scenes created with 3D graphic tools. Each data consists of SAIs with an angular resolution of 9×9 and a spatial resolution of 512×512. The dataset comprises a total of 16 training scenes, 8 validation scenes, and 4 test scenes (each named additional, test and training).

2) TRAINING PROCESS

The input images are converted to gray scale and cropped 32×32 patches randomly acquired for training. Due to the limited training dataset of only 16 SAIs, we employed data augmentation techniques such as random flipping, rotation, and adjustments in brightness, contrast, and refocusing were employed to ensure sufficient training of the network. In our supervised learning approach, we utilized the Mean Absolute Error (MAE) as the loss function and the Adam optimizer for network optimization. The training was initially planned for 5000 epochs with a batch size of 32, and the learning rate started at 1e-3, reducing by half every 1000 epochs. We effectively use early stopper finally concluding the training at 4830 epochs. The model is trained on an NVIDIA RTX 6000 GPU and Pytorch framework, and takes about two weeks for training.

In the training process, our network design effectively applies a coarse-to-fine manner. Using initial depth maps from an untrained network for occlusion masking can hinder stable convergence of the loss function. To address this, we initially train the network with ground truth (GT) depth maps to block occlusion information. Contrary to expectations, training with GT masks led to early divergence of loss, not stable convergence. Based on this insight, we initially pre-trained the LFDA network for 2100 epochs using conventional depth estimation methods, without considering occlusion. Post this phase, we integrate coarse-to-fine manner training with GT masks.

V. CONCLUSION

In this paper, we proposed a depth attention layer as a solution to balance the trade-off between accuracy and computational speed. The LFDA model applies patch matching via similarity comparisons in attention mechanisms, emphasizing local features within existing networks. The experimental results demonstrated remarkable outcomes in the *BadPix*(0.01) metric, underscoring the effectiveness of our approach in balancing accuracy and speed in depth estimation.

Future research will be directed towards overcoming the challenges associated with the lack of global context, which have contributed to lower MSE scores. The primary goal will be to enhance the model's ability to accurately predict depth in occluded and textureless areas, thereby refining its overall performance.

REFERENCES

[1] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Comput. Models Vis. Process.*, vol. 1, no. 2, pp. 3–20, 1991.

[2] 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.

[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] 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.

[5] 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.

[6] K. Li, J. Zhang, R. Sun, X. Zhang, and J. Gao, "EPI-based oriented relation networks for light field depth estimation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020.

[7] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder–decoder network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9145–9154.

[8] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3746–3754.

[9] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2271–2279.

[10] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4748–4757.

[11] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12095–12103.

[12] J. Chen, S. Zhang, and Y. Lin, "Attention-based multi-level fusion network for light field depth estimation," in *Proc. AAAI Conf. on Artificial Intelligence*, 2021, vol. 35, no. 2.

[13] Z. Huang, X. Hu, Z. Xue, W. Xu, and T. Yue, "Fast light-field disparity estimation with multi-disparity-scale cost aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6300–6309.

[14] 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.

[15] P. Zhou, L. Shi, X. Liu, J. Jin, Y. Zhang, and J. Hou, "Light field depth estimation via stitched epipolar plane images," *IEEE Trans. Vis. Comput. Graphics*, pp. 1–16, 2024.

[16] H.-G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y.-W. Tai, and I. S. Kweon, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1547–1555.

[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Lear. Repres. (ICLR)*, San Juan, Puerto Rico, 2015, pp. 2–4.

[18] W. Chao, X. Wang, Y. Wang, G. Wang, and F. Duan, "Learning sub-pixel disparity distribution for light field depth estimation," *IEEE Trans. Comput. Imag.*, vol. 9, pp. 1126–1138, 2023.

[19] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[20] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Lear. Represent. (ICLR)*, 2020.

[21] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6177–6186.
[22] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, "TransMVSNet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8575–8584.
[23] X. Wang et al., "MVSTER: Epipolar transformer for efficient multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022.
[24] V. Guizilini, R. Ambrus, D. Chen, S. Zakharov, and A. Gaidon, "Multi-frame self-supervised depth with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 160–170.
[25] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 41–48.
[26] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Mali, Dec. 2013, pp. 673–680.
[27] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.
[28] M. Feng, Y. Wang, J. Liu, L. Zhang, H. F. M. Zaki, and A. Mian, "Benchmark data set and method for depth estimation from light field images," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3586–3598, Jul. 2018.
[29] T. Leistner, H. Schilling, R. Mackowiak, S. Gumhold, and C. Rother, "Learning to think outside the box: Wide-baseline light field depth estimation with EPI-shift," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 249–257.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[32] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. 13th Asian Conf. Comput. Vis.*, Taipei, Taiwan. Springer, Nov. 2016, pp. 19–34.
[33] J. Y. Lee, J. DeGol, C. Zou, and D. Hoiem, "PatchMatch-RL: Deep MVS with pixelwise depth, normal, and visibility," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6138–6147.
[34] W. Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4396–4404.
[35] Y. Liu, Y. Pan, K. Luo, Y. Liu, and L. Zhang, "FEAMNet: Light field depth estimation network based on feature extraction and attention mechanism," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2023.

HYEONGSIK KIM is currently pursuing the bachelor's degree in electronics and electrical engineering with Dankook University, Yongin, South Korea. His research interests include computer vision, light field, and multi-view stereo.

SEUNGJIN HAN is currently pursuing the bachelor's degree in electronics and electrical engineering with Dankook University, Yongin, South Korea. His research interests include deep learning, computer vision, image compression, and light field coding.

YOUNGSEOP KIM received the M.S. degree in computer engineering from the University of Southern California, in 1991, and the Ph.D. degree in engineering science from Rensselaer Polytechnic Institute, in 2001. He was a Manager with Samsung SDI, until 2003. He developed the image-processing algorithm for PDP TV while with Samsung. He is currently a Professor with Dankook University, South Korea. His research interests include image/video compression, pattern recognition, light field image processing, stereoscopic codecs, and augmented reality. They include topics such as object-oriented methods for image/video coding, joint source-channel coding for robust video transmission, rate control, video transmission over packet wired or wireless networks, pattern recognition, and image processing. He was a resolution member and an Editor of JPSearch part 2 in JPEG, the Co-Chair of JPXML in JPEG, and the Head of Director (HOD) of Korea. He is also the Editor-in-Chief of Korea Semiconductor and Technology Society.