EPI-Guided Cost Construction Network for Light Field Disparity Estimation

Tun Wang[1], Rongshan Chen[1,2], Ruixuan Cong[1,2], Da Yang[1,2], Zhenglong Cui[1,2],
Fangping Li[1], Hao Sheng[1,2,3*]
[1] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China
[2] Beihang Hangzhou Innovation Institute Yuhang, Xixi Octagon City,
Yuhang District, Hangzhou 310023, P.R.China
[3] Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, P.R.China
{wangtun,rongshan,congrx,da.yang,zhenglong.cui,shenghao}@buaa.edu.cn, dclifp@126.com

Abstract

Recent learning-based light field (LF) disparity estimation methods construct cost volume by sequentially shifting each sub-aperture image (SAI) with a series of predefined offsets. They only use the visual information of SAIs and lose the geometry of LF. In this paper, we design a simple network that can cleverly integrate EPI features with cost volume to estimate the disparity. Firstly, we propose an efficient EPI extraction module to use abundant line characteristics. Secondly, we offer an EPI-Cost volume construction module that can create volume guided by the EPI line and the color consistency of images. Finally, after completing it, we adopt an intervolume fusion module to considerably correlate the validity of EPI lines in both directions. Experimental results show the proposed method achieves state-of-the-art performance in the quantitative and qualitative evaluation of the UrbanLF-Syn dataset.

1. Introduction

Light field (LF) can record light in different directions and describe scenes with richer specialties than traditional images [32]. With so much information, LF images can capture the concave or convex details and the reflection of the surface characteristics. It helps achieve more accurate results in disparity estimation. The development of LF disparity estimation can also improve the performance of computer vision systems [16], the accuracy of object detection, recognition, classification, and etc.

LF images contain the scene multi-view characteristics, making disparity estimation one of the most important research and application directions [10]. As Fig. 1 shows, LF is usually represented by two planes that have four parametric dimensions: it carries two-dimensional angular resolution (u, v) and two-dimensional spatial resolution (x, y) [12] [29]. EPI can be obtained by simultaneously fixing x and u dimensions or y and v dimensions. Further, we can represent disparity θ by the slope in the EPI. Recent methods based on cost volume construction only use color consistency and can not combine EPI structure with it. In addi-

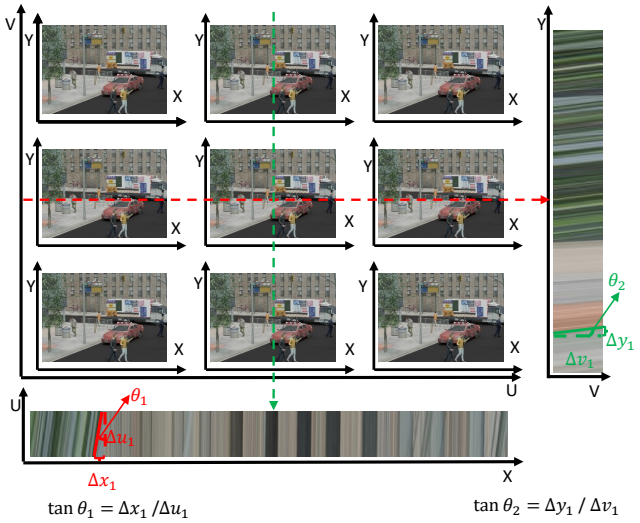

Figure 1. Horizontal and vertical EPI schematic representation and an illustration of the relationship between disparity and EPI. On the left is the 4D LF structure diagram, where (x, y) represents the pixel coordinate, and (u, v) represents the perspective coordinate. As shown at the bottom of the figure, when the value of v and y is fixed, and the EPI structure of the horizontal image is obtained. In a similar way, when the value of u and x is fixed, and the EPI structure of the vertical image is obtained. Disparity information θ can be represented by the slope of EPI: $\tan\theta = \Delta x(y)/\Delta u(v)$.

*Corresponding author.

tion, the scene contains various image noises and highly reflective target objects. The quality of matching will be disturbed by this noise and occlusion. However, EPI has a line structure that utilizes the relationship between the views, and it is currently a valid way to deal with noise [9] [20]. Its slope also corresponds to disparity, which is quite a reasonable way to estimate disparity [2]. Besides, compared with matching pixel points, EPI has the advantage of line structure, and the matching process is simple and accurate.

In this paper, we use EPI lines to guide the cost volume construction, which not only takes advantage of the relationship between EPI but also overcomes the problem of noise and occlusion during matching to a certain extent. Firstly, we design an efficient two-branch EPI extraction module to extract horizontal and vertical line features. Secondly, guided by these characteristics, we construct two special cost volumes in a specific order and match them. Thirdly, the horizontal and vertical volumes are merged using the intervolume attention mechanism. Finally, we use the 3D CNN to aggregate and regress it to get the final disparity map.

The contributions of this paper are as follows:

- We combine the EPI line features and the color consistency to construct a cost volume called EPI-Cost, and it can improve matching quality by using the structure of simple lines.
- We propose a new basic network architecture based on the cost volume construction method, it consists of four parts: EPI extraction module, EPI-Cost construction module, intervolume fusion module, cost aggregation and regression module.
- The experimental results prove that our approach achieves state-of-the-art performance in the quantitative and qualitative evaluation of UrbanLF-Syn dataset.

2. Related Work

In this section, we review the major works in LF disparity estimation. And most existing methods are classified into EPI-based and cost volume-based.

2.1. EPI-Based Methods

The EPI method estimates disparity by analyzing LF data structure [23]. By analyzing the structural characteristics of LF, the four-dimensional is projected onto the two-dimensional EPI image, and disparity estimation is transformed into the problem of finding the line slope by using the proportional relationship between the gradient and disparity [34] [11].

Bolles et al. [3] first propose the concept and acquisition method of EPI. In EPI, linear fitting detects edges, crests, and troughs to reconstruct three-dimensional structures. Wanner et al. [28] use structure tensor to extract the direction of lines in EPI images, propose a global consistency labeling algorithm, and obtain disparity through global optimization. Johannsen et al. [17] present a technique that uses EPI patches to compose a dictionary with a corresponding known disparity. Wanner et al. [29] use a structured tensor to compute the slope of each line in vertical and horizontal EPIs. Zhang et al. [34]locate the optimal slope of each line segmentation on EPIs using the locally linear embedding. Shin et al. [21] propose the EPINet, the first end-to-end network using CNN to extract EPI geometry disparity from LF images. Li et al. [13] extract oriented relation features between the center pixel and its neighborhood from EPI patches. Hassan et al. [7] improve EPINet and design a lighter and faster network architecture that can quickly calculate disparity. Li et al. [14] combine EPI feature extraction with the transformer and propose a valid EPI extraction network.

2.2. Cost Volume-Based Methods

According to the imaging principle of the LF camera, SAI can be regarded as the set of scene images taken by the virtual camera array from different angles [30] [18]. Therefore, the multi-view stereo matching method can be applied to LF disparity estimation. The matching cost is to construct the loss amount on the basis of the difference between the central and remaining perspectives [6] [33].

Yu et al. [31] explore the geometry of 3D lines in ray space to improve the triangulation of LF and stereo matching. Since this method uses stereo matching to estimate the disparity per pixel, it could be more effective on data sets with a small range of disparities. Liu et al. [15] propose a method to acquire LF video and image disparity maps based on stereo matching. Firstly, the LF data is rendered and enhanced. Secondly, the local disparity map is obtained based on the fourier phase shift theory. Finally, the accurate disparity map is obtained based on the multi-label optimization algorithm of graph cutting. Tsai et al. [24] first apply cost volume to deep learning networks and design the primary four-step backbone network, which can estimate disparity in LF. Chen et al. [5] develop a new way with four branches and use the attention mechanism to establish the relationship among cost volumes. Wang et al. [26] [27] solve the occlusion problem in light field images by masking and combining the occlusion module with the cost volume. Meanwhile, it is constructed by means of encoding and decoding, which significantly reduces time consumption. Chen et al. introduce a disparity before identifying pixels and propose a pixel-based matching cost function (PMCF). Chao et al. [4] use a larger memory space to construct a more refined cost volume and propose an elaborate loss function. Wang et al. [25] divide the light field into

four regions to build a four-branch cost volume to reduce computational consumption and get a more accurate result.

The EPI-based approaches take an amount of computation to acquire disparity information, and their real-time performance is low. The cost volume-based methods match pixels, but they only use the feature of color consistency of LF images. However, our approach uses color consistency information and employs the EPI line characteristics.

3. Methodology

In this section, the network structure is described in detail. As Fig. 3 shows, it represents the overall network architecture. Our proposed approach is divided into four parts: Firstly, the efficient EPI extraction module is designed to generate EPI line features. Secondly, we use these characteristics to guide the construction of cost volume, which is matched by simple lines. Thirdly, the intervolume fusion module associates branches of EPI-Cost volume and learns from each other. Finally, the aggregation and regression module is used to obtain the final disparity map.

3.1. EPI Extraction Module

For disparity estimation, current methods use the SPP module to extract features from different scales or SAIs and provide the hierarchical context information about the region [8] [35]. But it only depends on the context information, and having useful features in large textureless or specular areas is challenging. Unlike present feature extraction methods, we utilize the characteristics between perspectives to overcome these errors. The EPI-based method can mitigate the misestimation caused by factors such as illumination and specular reflection to some extent. So we propose a module combining visual image information with view messages to extract line characteristics.

The input is horizontal and vertical center images of the SAI, the size of which is $B \times H \times W \times C \times U(V)$ after data preprocessing. Firstly, to take advantage of the relationship between perspective and structure, we convert input images I_{input} to I_h and I_v, which are shaped as $(B \times H) \times C \times W \times U$ and $(B \times W) \times C \times H \times V$. Secondly, we design a two-branch feature extraction structure, and the horizontal and vertical images are convolved separately. This module can effectively use the correlation between perspectives and color consistency to learn the EPI characteristics. Finally, the residual training method is used to thoroughly combine the initial, intermediate, and final features, which steadily improves the effect of the model by deepening the number of layers. Each network layer is composed of $Conv2d-BatchNorm2d-Softplus-Tanh-Conv2d-BatchNorm2d-Softplus-Tanh$ labeled as C_f^2, and the number of channels is respectively 16, 32, and 64 [25]. Because data is zero-centered and a zigzag path is not easy to occur when updating parameters, $Tanh$ can help to speed up training [22] and $Softplus$ can also solve the dead ReLU problem [1]. After the above steps, we can get accurate EPI line characteristics $F_{h(v)}$. The specific formula is as follows:

$$F_h(F_v) = C_f^2(I_h(I_v)) + C_f^2(C_f^2(I_h(I_v))) \tag{1}$$

where, I_h and I_v are horizontal and vertical images, C_f^2 is network layer.

3.2. EPI-Cost Volume Construction Module

Traditional methods construct matching cost using the warp-and-concat approach. Given a disparity range, they warp features according to their view coordinate and concatenate all warped features to generate cost volumes [26]. However, current methods only use the information of pixels for matching without using the structure of lines. Because the line structures are simple and efficient, the volumes easily match. In addition, with the relationship between EPI slope and disparity, the results are accurate and reasonable. As Fig. 3 shows, compared with the previous methods, our approach combines the EPI line and color consistency to guide the surrounding views to warp to the center and the cost volume is easy to match under the guide of EPI line.

Specific steps are as follows: Firstly, because all disparity values of adjacent views in the UrbanLF-Syn dataset are $[-0.47, 1.55]$, we select nine equally spaced disparity values from the range. Values respectively are $[-0.47, -0.22, 0.03, 0.28, 0.54, 0.79, 1.04, 1.29, 1.55]$ labeled as d. Secondly, we manually shift the horizontal EPI lines along the v direction with different disparity levels to get the F_{EPI}^h. Similarly, vertical EPI lines are carried out to get the F_{EPI}^v. After this operation, the later part of the network can directly study EPI line information at different spatial positions by using a relatively small receptive field. The specific formula is as follows:

$$F_{EPI}^h = con_{d=d_{min}}^{d_{max}} F_h(0, (v_c – v) \times d) \tag{2}$$

$$F_{EPI}^v = con_{d=d_{min}}^{d_{max}} F_v((u_c – u) \times d, 0) \tag{3}$$

where con is concatenating operation, $d_{min} = -0.47$ is the smallest disparity value and $d_{max} = 1.55$ is the biggest. We shift horizontal and vertical EPI lines according to their offset $v_c – v(u_c – u)$ and disparity ranges d. Secondly, we concatenate the shifted EPI lines in the feature dimension into 5D cost volume, which includes color consistency, EPI line characteristics, and disparity information. Finally, we get the cost volume $Cost_{EPI}^h$ and $Cost_{EPI}^v$ are both sized $B \times (N \times C) \times H \times W \times 9$. The specific formula is as follows:

$$Cost_{EPI}^{h(v)} = con_0^N(F_{EPI}^{h(v)}) \tag{4}$$

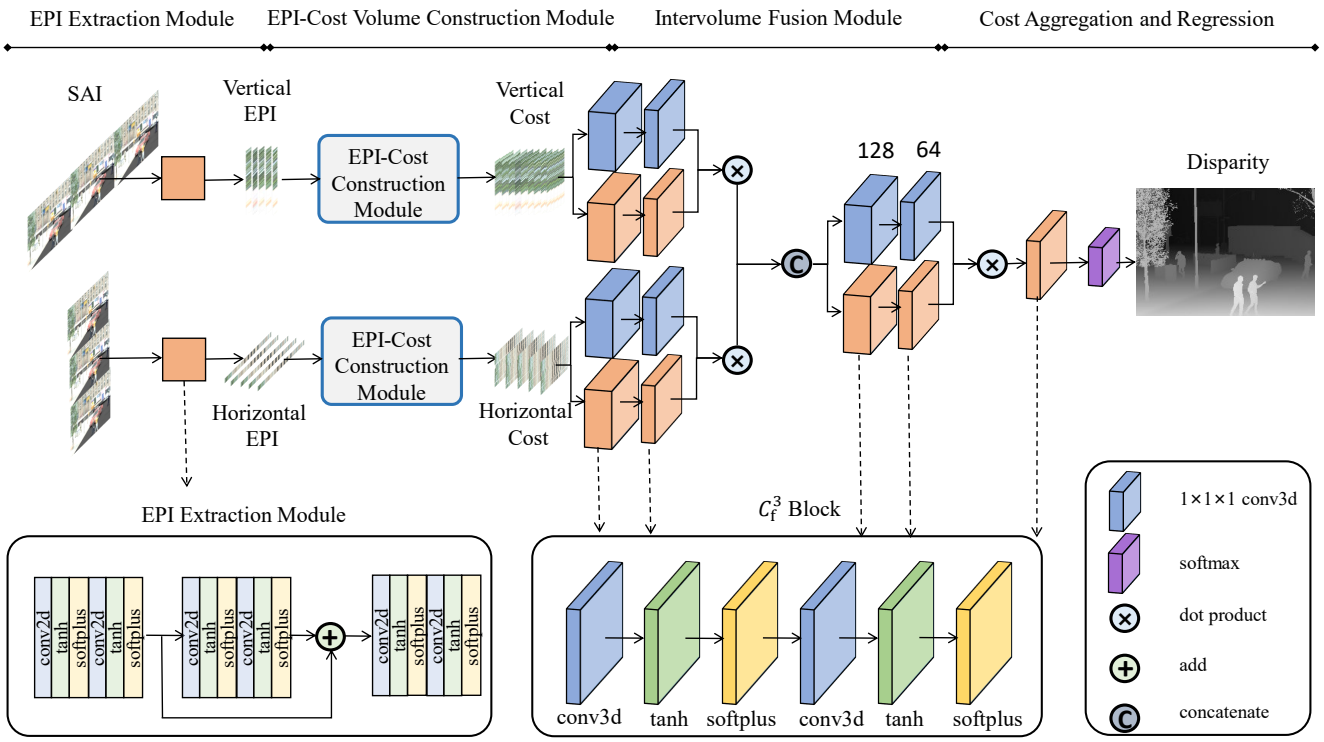

Figure 2. The EPI-Guided cost construction network. There are four steps in our network, including EPI extraction module, EPI-Cost volume construction module, the intervolume fusion module and the aggregation and regression module.

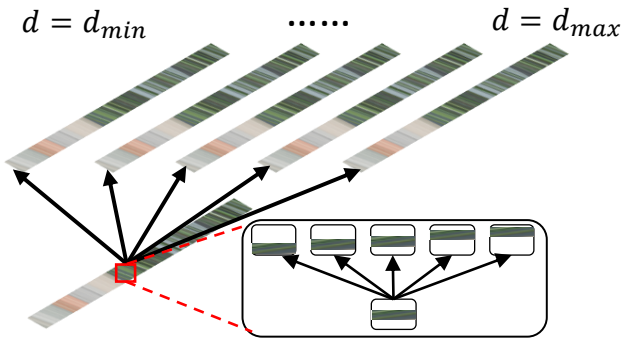

Figure 3. An illustration of how to construct EPI-Cost volume. We shift the horizontal EPI features with different disparity levels and concatenate them in order of disparity from smallest to largest. The same goes for vertical EPI.

3.3. Intervolume Fusion Module

After the EPI-Cost volume construction module, $Cost_{EPI}^{h(v)}$ from horizontal and vertical are obtained. Both horizontal and vertical cost volumes contain essential disparity information, and they need to complement each other. In this module, we further fuse them to integrate information effectively. Especially if the slope feature of EPI line in the horizontal direction is not apparent but is evident in the vertical, it needs to be supplemented by the other direction information. Previous method [5] proposes an inter-branch fusion module that can fuse the features from different branches. Inspired by this, we design an intervolume fusion module that facilitates their interactions.

The specific steps are as follows: Firstly, we extract local and global attention weights for two cost volumes. We use 3D $conv + bn$ with the kernel size is $1 \times 1 \times 1$ to extract local attention and use the adaptive average pooling layer to get global attention. The cost map labeled as $Cost_{map}^{h(v)}$ can be calculated by using 3D convolution written as C_{att}^3 with the kernel size is $3 \times 3 \times 3$. The formula is as follows:

$$Cost_{map}^{h(v)} = C_{att}^3(H_{att}(Cost_{EPI}^{h(v)})) \tag{5}$$

where, H_{att} represents local and global attention. Secondly, we concatenate the cost map of two branches and use a 2D convolution written as C_{att}^2 with the kernel size is 3×3 to fuse information. The formula is as follows:

$$Att = C_{att}^2(con(Cost_{map}^h, Cost_{map}^v))) \tag{6}$$

where Att represents the horizontal and vertical attention weights, and the channel is 2. Finally, the two EPI-Cost

volumes are multiplied by attention maps and concatenated. The specific is as follows:

$$Cost_{att} = con(Att^h \odot Cost^h_{EPI}, Att^v \odot Cost^v_{EPI}) \quad (7)$$

where $Cost_{att}$ is the final EPI-Cost volume, Att^h is the first dimension of Att and Att^v is the second dimension.

3.4. Cost Aggregation and Regression

Obtained the EPI-Cost volume $Cost_{att}$, spatial attention [5] is adopted to enhance the features of the cost volume further. Firstly, we use one $1 \times 1 \times 1$ convolution to extract the attention map and reduce the channel numbers from $views \times C$ to $4 \times C$. Secondly, we cascade three convolution modules with a kernel size of $3 \times 3 \times 3$ for cost aggregation and obtain the normalized probability of each disparity value. Each convolution module comprises $Conv3d{-}BatchNorm3d{-}Softplus{-}Tanh{-}Conv3d{-}BatchNorm3d - Softplus - Tanh$ named C^3_f, and the number of channels is respectively 128, 128, and 1. Finally, we calculate the final predicted disparity by the sum of each disparity with its normalized probability and regressing. The specific calculation formula is as follows:

$$D_{final} = \sum_{d=d_{min}}^{d_{max}} d \times (softmax(C^3_f(Cost_{att}))) \quad (8)$$

where D_{final} reprents final disparity map, d is disparity range from the smallest disparity value $d_{min} = -0.47$ to the biggest $d_{max} = 1.55$.

4. Experiments

In this section, firstly, we introduce the UrbanLF-Syn dataset and describe the implementation details. Secondly, we compare our EPI-Cost to several state-of-the-art methods in Sec.4.3. Finally, more experiments are conducted to testify to the effectiveness of our model. Including 1)ablation experiments are performed in Sec.4.4 to verify the validity of each module used. 2)We do a cross-data set test in Sec.4.5 to demonstrate the universality and robustness of our model.

4.1. Datasets

To validate the effectiveness of our method, we conduct experiments on a synthetic dataset: the UrbanLF-Syn dataset [19].

UrbanLF-Syn dataset is a publicly available synthetic LF dataset, partitioned into three sub-sets: 'Train,' 'Val,' and 'Test.' It contains 230 scenes with LF resolution $480{\times}640{\times}9{\times}9$, and the ground truth is provided. $480{\times}640$ is the spatial resolution, and $9{\times}9$ represents the angular resolution. Its ground truth disparity range from -0.47 to 1.55 pixels between adjacent views. We use 170 training images for training the network and choose the 30 validation images for validating.

4.2. Implementation Details

We exploit the patch-wise training by randomly sampling patches of size 32×32 from the 170 training images. In addition to this operation, we also take other data enhancement operations, such as converting the image to grayscale, scale enhancement, transposing, rotating, color transformation, brightness and contrast adjustment, etc.

During the training phase, we select the batch size of 16. So after we process the initial images, the shape of each input is $16 \times 32 \times 32 \times 9 \times 2$, and the last dimension value represents horizontal and vertical. Our network is trained in a supervised manner with an L1 loss and is optimized using the Adam method with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is set to 1×10^{-3}. We save a model every 2000 epochs. And finally, the training is stopped after 1×10^5 iterations and takes about 14 days. Our model is implemented in PyTorch and trained on Nvidia RTX 1080Ti GPU.

For performance evaluation, we use the standard evaluation metrics in LF disparity estimation, including the mean square error (MSE$\times$100) and bad pixel ratio (BadPix(α)). BadPix(α) measures the percentage of incorrectly estimated pixels whose absolute errors exceed a predefined threshold (e.g.,$\alpha = 0.07, 0.03, 0.01$). The metrics (MSE, BP) are calculated only on the central view image with a cropping of 15 pixels at each border. The specific calculation formula is as follows:

$$MSE \times 100 = 100 \times \frac{1}{h \times w} \sum_{p=1}^{h \times w} (D_p - D_p^{gt})^2 \quad (9)$$

$$BadPix(\alpha) = 100 \times \frac{1}{h \times w} \sum_{p=1}^{h \times w} (|D_p - D_p^{gt}| > \alpha) \quad (10)$$

where, h and w respectively represent the length and width of the image, p represents pixel coordinates, D_p is predicted pixel p disparity value and D_p^{gt} is ground truth.

4.3. Comparison with State-of-the-Arts

To prove the advancement of our method for LF disparity estimation, we compare EPI-Cost with other state-of-the-art methods, including EPINet [21], LFattNet [5], AttMLFNet [5], OACC-Net [26], SubFocal [4]. To ensure the experiment's fairness and validity, we train on RTX 1080Ti GPU for two weeks and select the best MSE model during the training phases.

Quantitative Results. Tab. 1 shows our proposed method achieves state-of-the-art performance in the quantitative and qualitative evaluation of the UrbanLF-Syn dataset. We pick the first eight images of the verification

Method	Image11				Image27				Image34			
	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1
EPINet [21]	0.94	26.88	37.87	71.31	1.10	19.18	43.46	79.55	2.01	28.38	43.04	69.89
LFattNet [5]	0.93	12.63	23.96	59.78	0.46	12.83	26.47	69.43	1.76	19.37	33.36	67.84
AttMLFNet [5]	0.80	14.33	35.99	77.06	0.43	14.20	28.85	66.42	1.32	18.43	32.73	66.93
OACC-Net [26]	0.84	13.07	24.34	58.37	0.83	14.87	29.54	74.42	1.87	18.30	32.62	67.16
SubFocal [4]	0.91	13.37	27.08	68.84	1.65	19.12	31.37	68.99	1.91	30.21	40.90	64.71
Ours	**0.74**	**11.04**	**22.04**	**59.77**	**0.30**	**12.57**	**21.56**	**50.47**	**1.21**	**13.37**	**28.81**	**56.35**
	Image50				Image54				Image68			
	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1
EPINet	2.43	19.68	31.78	65.99	1.75	16.82	27.98	64.37	1.43	14.89	39.62	76.91
LFattNet	1.47	**14.06**	30.98	65.64	0.97	20.05	36.61	71.43	0.73	14.08	**21.24**	**53.90**
AttMLFNet	0.99	16.73	29.12	66.51	0.89	17.15	27.98	64.49	1.09	14.16	39.34	73.46
OACC-Net	1.19	16.93	27.73	65.86	0.94	16.76	29.35	65.91	1.15	13.65	37.78	74.52
SubFocal	0.85	17.45	**25.34**	70.15	0.88	16.11	29.64	**53.38**	0.64	13.41	37.52	73.17
Ours	**0.83**	16.64	26.62	**56.35**	**0.60**	**15.61**	**26.59**	54.49	**0.48**	**11.63**	24.16	59.91
	Image69				Image70				Average			
	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1	MSE	BP7	BP3	BP1
EPINet	0.65	17.75	26.87	65.53	1.38	22.10	37.74	61.66	1.77	25.83	36.41	64.40
LFattNet	1.04	13.69	24.02	63.36	1.09	17.72	34.60	66.89	1.33	19.28	30.00	59.27
AttMLFNet	1.45	18.65	26.88	68.39	1.56	22.23	36.81	60.87	1.79	14.61	36.83	63.71
OACC-Net	0.63	17.81	26.95	65.78	0.90	14.94	27.12	62.63	0.90	14.68	39.26	64.75
SubFocal	0.61	12.20	27.54	65.92	0.91	13.41	27.29	65.64	0.88	13.27	27.92	58.82
Ours	**0.60**	**11.15**	**24.01**	**55.50**	**0.83**	**13.18**	**24.39**	**58.67**	**0.81**	**12.51**	**23.65**	**55.39**

Table 1. Quantitative results on UrbanLF-Syn dataset. We calculate mean square errors ($MSE \times 100$) and BadPix(0.07, 0.03, 0.01) by different methods on UrbanLF-Syn. The best results are shown in bold.

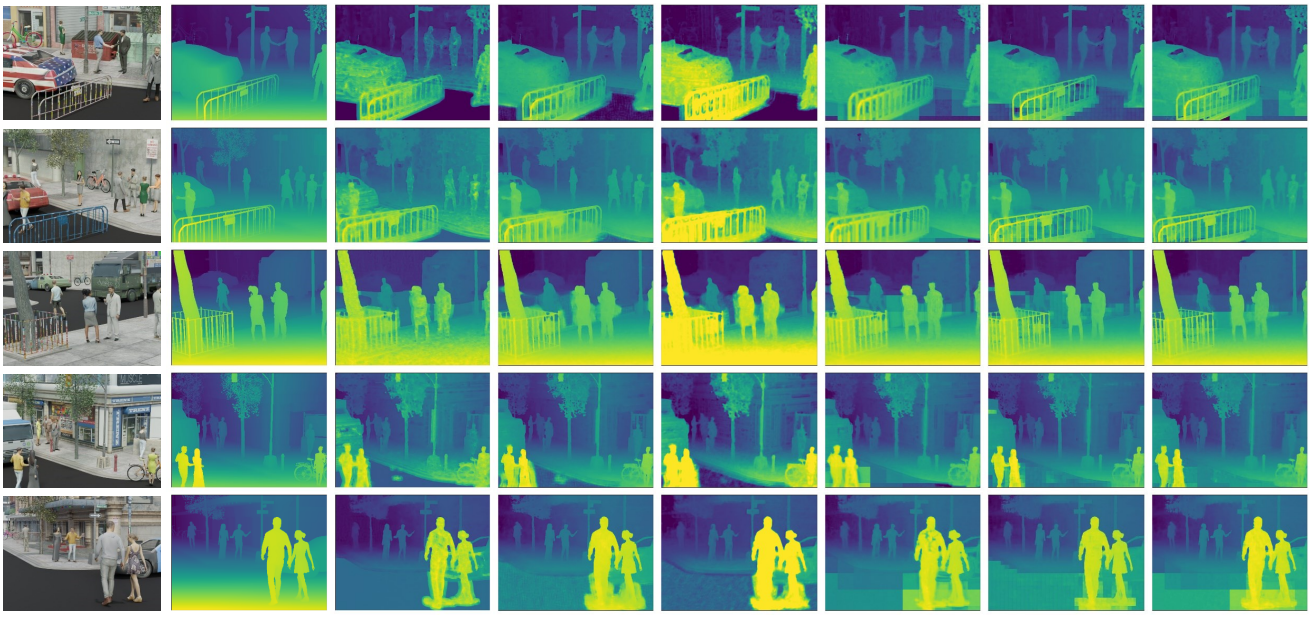

Figure 4. Qualitative results on UrbanLF-Syn dataset.

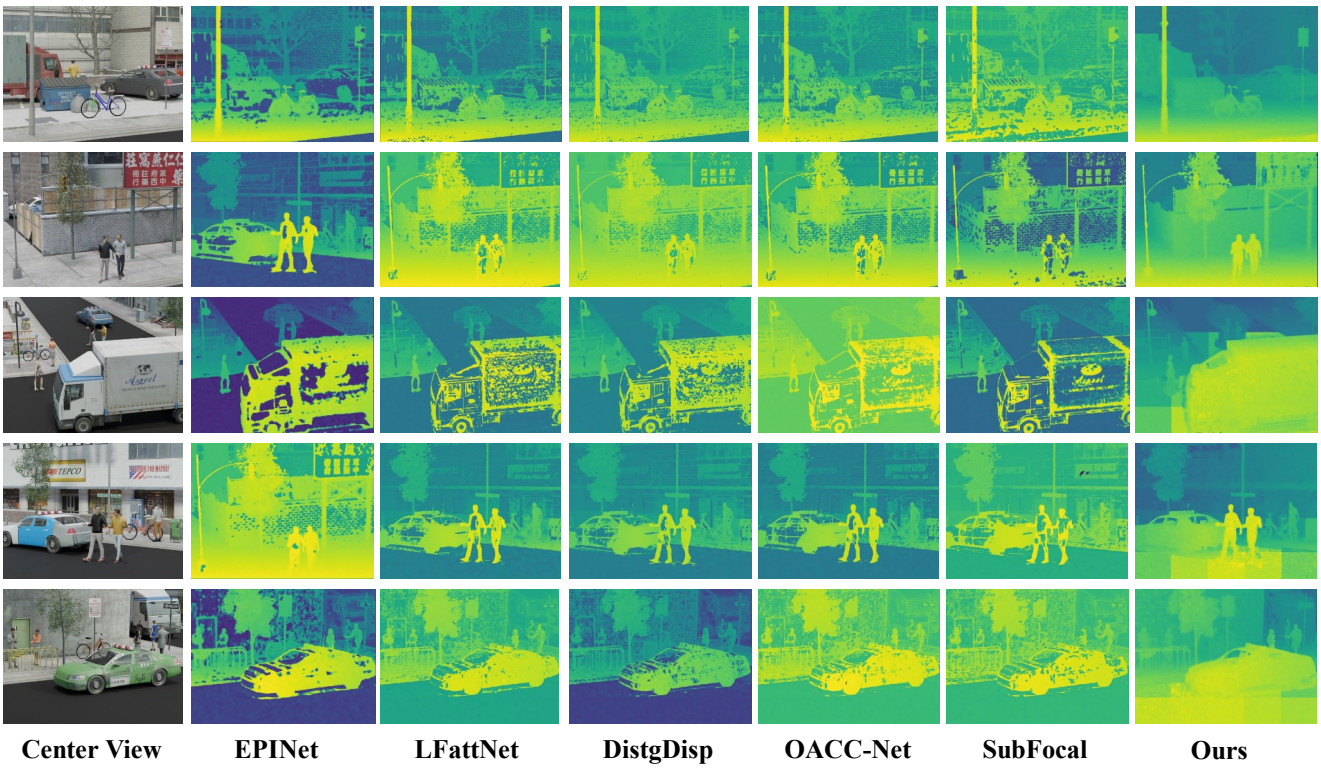

Figure 5. Cross dataset test results on UrbanLF-Syn dataset. The models are trained on HCI 4D Light Field Dataset.

Method	Average			
	MSE	BP7	BP3	BP1
HRDE	**0.388**	6.717	13.970	30.492
CBPP	0.394	**5.907**	**12.628**	**27.385**
MS3D	0.559	7.917	14.664	31.066
SF-Net	0.712	8.311	14.595	28.007
EPI-Cost	0.738	14.327	27.041	57.946
UOAC	0.915	17.294	31.194	58.458
MTLF	1.156	13.518	22.852	46.933
MultiBranch	2.776	43.402	64.915	86.350

Table 2. The benchmark in the average comparison of the whole validation and testing images in the 3rd workshop on the light field for computer vision LFNAT. The best results are shown in bold.

Method	Average			
	MSE	BP7	BP3	BP1
Use ReLU	0.87	12.98	24.45	55.67
Previous Cost	1.33	19.28	30.00	59.27
No Attention	1.54	15.32	24.75	63.40
No Horizontal	1.05	14.05	23.93	64.43
No Vertical	1.16	14.91	24.33	63.85
Ours	**0.81**	**12.51**	**23.65**	**55.39**

Table 3. Ablation results on UrbanLF-Syn dataset. The first is our feature extraction network architecture is replaced with the previous SPP module, which is composed of Conv2d-BatchNorm2d-ReLU-Conv2d-BatchNorm2d-ReLU. The second is we use the previous method of cost volume construction. The third is we removed the interaction module. The fourth and fifth are removing a branch. The best results are shown in bold.

set to display and calculate the average of 30 images. Our approach achieves state-of-the-art performance in terms of MSE$\times$100 and a relatively high ranking in terms of BadPix(0.07, 0.03, 0.01). Tab. 2 shows the average comparison of the whole testing images in the 3rd workshop on the light field for computer vision LFNAT.

Visual Comparison. Fig. 4 shows some visual results of predicted disparity. Compared with other methods, it can be observed that the disparities estimated by our approach are more precise to textured regions, and our model can preserve more details. For example, the fence and the complex background can be estimated more accurately. In addition, the prediction results for the untextured region are also close to the true value.

Method	Average			
	MSE	BP7	BP3	BP1
EPINet [21]	12.826	41.268	49.164	**66.439**
LFattNet [5]	15.679	51.449	60.947	75.333
DistgDisp [27]	15.185	46.569	59.266	78.954
OACC-Net [26]	16.215	52.395	62.930	79.566
SubFocal [4]	17.326	54.127	62.904	76.288
Ours	**12.753**	**39.732**	**47.286**	70.149

Table 4. Cross dataset test results on UrbanLF-Syn dataset. We compare in UrbanLF-Syn about the training in HCInew. The best results are shown in bold.

4.4. Ablation Study

To verify the validity of our experiment, we design several comparative experiments to prove it. We replace each module to demonstrate the significance of every part of our approach. Firstly, the EPI extraction module is replaced with the previous SPP module, which is composed of $Conv2d-BatchNorm2d-ReLU-Conv2d-BatchNorm2d-ReLU$. Secondly, we use the previous method of cost volume construction instead of EPI-Cost. Finally, to prove that horizontal and vertical EPI branches can interact, we also use a single-branch EPI feature extraction method or remove the interaction module.

As Tab. 3 shows, we calculate the average 30 validation set. Each module in our model is very effective. If we replace any module, the results of both MSE and BadPix(0,07, 0.03, 0.01) are higher than our approach.

4.5. Cross Dataset Test Results

To prove that EPI-Cost also works well in big disparity images, we train our model on the HCI 4D Light Field Dataset and test it on the UrbanLF-Syn dataset. We compare our model with other state-of-the-art methods, including EPINet [21], LFattNet [5], DistgDisp [27], OACC-Net [26], SubFocal [4]. All of these open-source models are completely trained on HCI 4D Light Field Dataset. Fig. 5 shows our model preserves more details and works well in big disparity images. Tab. 4 shows, our approach achieves 1st place in terms of MSE$\times$100 and BadPix(0.03, 0.01) and 2nd place in terms of BadPix(0.07).

5. Conclusion

In this paper, we propose a method called EPI-Cost which can let EPI line information guide to construct cost volumes and improve matching quality. It overcomes the shortcoming that cost volume only uses the color consistency of the image. Experimental results show the proposed method achieves state-of-the-art performance in the quantitative and qualitative evaluation on UrbanLF-Syn dataset.

Acknowledgement

This study is partially supported by the National Key R&D Program of China (No.2022YFC3803600), the National Natural Science Foundation of China (No.61872025), and the Open Fund of the State Key Laboratory of Software Development Environment (No.SKLSDE-2021ZX-03). Thank you for the support from HAWKEYE Group.

References

[1] Koushik Biswas, Sandeep Kumar, Shilpak Banerjee, and Ashish Kumar Pandey. Tanhsoft–a family of activation functions combining tanh and softplus. *arXiv preprint arXiv:2009.03863*, 2020. 3

[2] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):287–300, 2016. 2

[3] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision*, 1(1):7–55, 1987. 2

[4] Wentao Chao, Xuechun Wang, Yingqian Wang, Liang Chang, and Fuqing Duan. Learning sub-pixel disparity distribution for light field depth estimation. *arXiv preprint arXiv:2208.09688*, 2022. 2, 5, 6, 8

[5] Jiaxin Chen, Shuo Zhang, and Youfang Lin. Attention-based multi-level fusion network for light field depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1009–1017, 2021. 2, 4, 5, 6, 8

[6] Kang Han, Wei Xiang, Eric Wang, and Tao Huang. A novel occlusion-aware vote cost for light field depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8022–8035, 2021. 2

[7] Ali Hassan, Mårten Sjöström, Tingting Zhang, and Karen Egiazarian. Light-weight epinet architecture for fast light field disparity estimation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2022. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 3

[9] Stefan Heber, Wei Yu, and Thomas Pock. Neural epi-volume networks for shape from light field. In *Proceedings of the IEEE international conference on computer vision*, pages 2252–2260, 2017. 2

[10] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 19–34. Springer, 2017. 1

[11] Zhicong Huang, Xuemei Hu, Zhou Xue, Weizhu Xu, and Tao Yue. Fast light-field disparity estimation with multi-disparity-scale cost aggregation. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 6320–6329, 2021. 2

[12] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 1

[13] Kunyuan Li, Jun Zhang, Rui Sun, Xudong Zhang, and Jun Gao. Epi-based oriented relation networks for light field depth estimation. *arXiv preprint arXiv:2007.04538*, 2020. 2

[14] Peng Li, Jiayin Zhao, Jingyao Wu, Chao Deng, Haoqian Wang, and Tao Yu. Opal: Occlusion pattern aware loss for unsupervised light field disparity estimation. *arXiv preprint arXiv:2203.02231*, 2022. 2

[15] Fei Liu, Guangqi Hou, Zhenan Sun, and Tieniu Tan. High quality depth map estimation of object surface from light-field images. *Neurocomputing*, 252:3–16, 2017. 2

[16] Huimin Lu, Yujie Li, Tomoki Uemura, Hyoungseop Kim, and Seiichi Serikawa. Low illumination underwater light field images reconstruction using deep convolutional neural networks. *Future Generation Computer Systems*, 82:142–148, 2018. 1

[17] Yaoxiang Luo, Wenhui Zhou, Junpeng Fang, Linkai Liang, Hua Zhang, and Guojun Dai. Epi-patch based convolutional neural network for depth estimation on 4d light field. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part III 24*, pages 642–652. Springer, 2017. 2

[18] In Kyu Park, Kyoung Mu Lee, et al. Robust light field depth estimation using occlusion-noise aware data costs. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2484–2497, 2017. 2

[19] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. Urbanlf: a comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7880–7893, 2022. 5

[20] Hao Sheng, Shuo Zhang, Xiaochun Cao, Yajun Fang, and Zhang Xiong. Geometric occlusion analysis in depth estimation using integral guided filter for light-field image. *IEEE Transactions on Image Processing*, 26(12):5758–5771, 2017. 2

[21] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4748–4757, 2018. 2, 5, 6, 8

[22] Yi-Yuan Tang, Yinghua Ma, Junhong Wang, Yaxin Fan, Shigang Feng, Qilin Lu, Qingbao Yu, Danni Sui, Mary K Rothbart, Ming Fan, et al. Short-term meditation training improves attention and self-regulation. *Proceedings of the national Academy of Sciences*, 104(43):17152–17156, 2007. 3

[23] Ivana Tosic and Kathrin Berkner. Light field scale-depth space transform for dense depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 435–442, 2014. 2

[24] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12095–12103, 2020. 2

[25] Xucheng Wang, Chenning Tao, and Zhenrong Zheng. Occlusion-aware light field depth estimation with view attention. *Optics and Lasers in Engineering*, 160:107299, 2023. 2, 3

[26] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation (supplemental material). 2, 3, 5, 6, 8

[27] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 2, 8

[28] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2012. 2

[29] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013. 1, 2

[30] W Williem and In Kyu Park. Robust light field depth estimation for noisy scene with occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4404, 2016. 2

[31] Zhan Yu, Xinqing Guo, Haibing Lin, Andrew Lumsdaine, and Jingyi Yu. Line assisted light field triangulation and stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2792–2799, 2013. 2

[32] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 1

[33] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12926–12934, 2020. 2

[34] Yongbing Zhang, Huijin Lv, Yebin Liu, Haoqian Wang, Xingzheng Wang, Qian Huang, Xinguang Xiang, and Qionghai Dai. Light-field depth estimation via epipolar plane image analysis and locally linear embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):739–747, 2016. 2

[35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 3