EPI-based Oriented Relation Networks for Light Field Depth Estimation

Kunyuan Li
lkyhfut@gmail.com
Jun Zhang*
zhangjun1126@gmail.com
Rui Sun
sunrui@hfut.edu.cn
Xudong Zhang
xudong@hfut.edu.cn
Jun Gao
gaojun@hfut.edu.cn

School of Computer Science and Information Engineering
Hefei University of Technology
Hefei, Anhui, China

Abstract

Light field cameras record not only the spatial information of observed scenes but also the directions of all incoming light rays. The spatial and angular information implicitly contain geometrical characteristics such as multi-view or epipolar geometry, which can be exploited to improve the performance of depth estimation. An Epipolar Plane Image (EPI), the unique 2D spatial-angular slice of the light field, contains patterns of oriented lines. The slope of these lines is associated with the disparity. Benefiting from this property of EPIs, some representative methods estimate depth maps by analyzing the disparity of each line in EPIs. However, these methods often extract the optimal slope of the lines from EPIs while ignoring the relationship between neighboring pixels, which leads to inaccurate depth map predictions. Based on the observation that an oriented line and its neighboring pixels in an EPI share a similar linear structure, we propose an end-to-end fully convolutional network (FCN) to estimate the depth value of the intersection point on the horizontal and vertical EPIs. Specifically, we present a new feature-extraction module, called **Oriented Relation Module (ORM)**, that constructs the relationship between the line orientations. To facilitate training, we also propose a refocusing-based data augmentation method to obtain different slopes from EPIs of the same scene point. Extensive experiments verify the efficacy of learning relations and show that our approach is competitive to other state-of-the-art methods. The code and the trained models are available at https://github.com/lkyahpu/EPI_ORM.git.

1 Introduction

Light field cameras record both 2D spatial and 2D angular information of the observed scene [13]. The lenslet-based light field camera [18], a compact and hand-held light field camera, is able to achieve the dense sampling of the viewpoints by utilizing a micro-lens

* indicates the corresponding author.

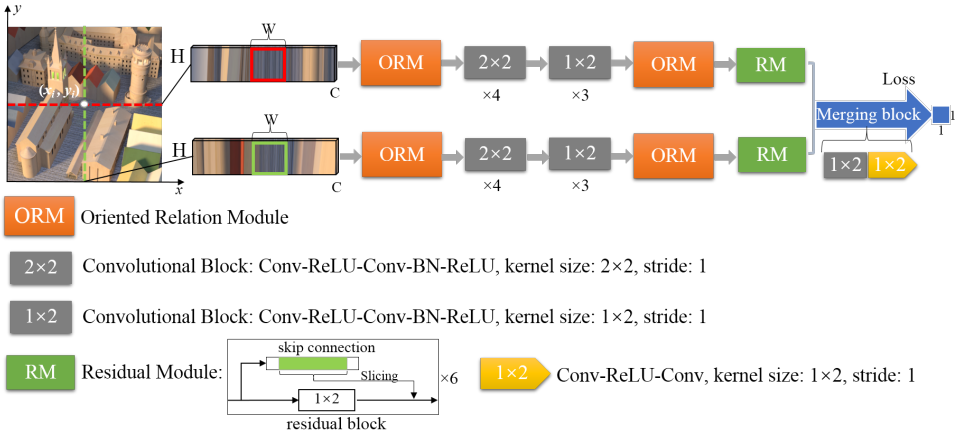

Figure 1: An overview of the proposed network architecture. The input is a pair of EPI patches obtained from the horizontal and vertical EPIs. Each branch consists of two oriented relation modules (ORMs), seven convolutional blocks, and a residual module (RM). The output of the two branches is integrated by a merging block to estimate the depth value of each pixel.

array inserted between the main lens and the photo sensor. The captured 4D light field data implicitly contains geometrical characteristics such as multi-view geometry or epipolar geometry, which has attracted much attention in recent years to improve the performance of depth estimation from light fields.

To visualize light fields and extract light field features, the 4D light field data is often converted into various 2D images such as multi-view sub-aperture images [17], Epipolar Plane Images (EPIs) [13], and focal stacks [14]. Some representative methods [20, 21, 24] exploit different depth cues from sub-aperture images and focal stacks for depth estimation. However, it is difficult to acquire dense and accurate depth maps from the lenslet-based cameras owing to the optical distortions [11] and the narrow baseline [18] between sub-aperture images. Besides, these methods are usually accompanied by heavy computational burdens and carefully-designed optimization measures. To avoid these issues, some methods [4, 15, 23, 27] exploit EPIs that exhibit patterns of oriented lines with constant colors to visualize light fields. Each of these lines corresponds to the projection of a single 3D scene point, and its slope is called disparity [22]. Therefore, one can infer the depth of the corresponding scene point by analyzing the disparity of the oriented line in the EPI. Moreover, the oriented line and its neighboring pixels share the similar linear structure, which is beneficial to estimate the slope of the EPI by constructing the relationship between the center region in the EPI and its neighborhood. Nonetheless, current methods predict depth maps by extracting the optimal slope of EPIs while ignoring the relationship between neighboring pixels in EPIs, which makes the results inaccurate. It has been well recognized that the relation information is capable of offering important visual cues for computer vision tasks, such as spatial and channel relations in semantic segmentation [16] and object detection [8], and temporal relations in activity recognition [29].

In this paper, we propose an end-to-end fully convolutional network to estimate the depth value of the intersection point on the horizontal and vertical EPIs, as shown in Figure 1. We design a Siamese network without sharing weights (i. e. pseudo-Siamese [26]) so

that the convolution weights of the horizontal and vertical EPIs can be learned separately. Specifically, we propose a new feature extraction module, called **Oriented Relation Module (ORM)**, to learn and reason about the relationship between oriented lines in EPIs by extracting oriented relation features between the center pixel and its neighborhood from EPI patches. The proposed method can be considered as the first work on modeling relation features in EPIs, which is novel and different from existing relation models in two aspects: First, existing works [16, 29] focus on modeling temporal relation between frames and spatial relations between pixels. In contrast, our method proposes the geometric relation between line orientations in EPI patches, which is beneficial to extract the accurate slopes of EPIs for light field depth estimation. Second, the proposed method models dependencies between oriented lines, without making any assumptions on their feature distributions and locations. Our network is trained using the 4D light field benchmark dataset [7], where the ground truth disparities are available. However, we find that it is hard to train such a deep network with insufficient data. To mitigate this issue, we propose a data augmentation method by refocusing EPIs so that EPIs with different slopes as well as the corresponding ground truth disparities can be obtained at the same scene point. We show that the newly proposed ORM and EPI-based data augmentation can bring performance boost for light field depth estimation.

2 Related Work

Conventional depth estimation from light fields mainly relies on different assumptions [14, 21] and handcrafted depth features [20, 24] based on sub-aperture images and focal stacks. In this section, we restrict ourselves to methods that exploit EPIs, and review some representative works with relation reasoning.

Light field depth estimation based on EPIs. There exist a few methods that exploit the EPI for light field depth estimation due to its linear structure associated with depth [22]. For example, Wanner *et al*. [23] used a structured tensor to compute the slope of each line in vertical and horizontal EPIs. Zhang *et al*. [27] introduced the Spinning Parallelogram Operator (SPO) to find matching lines in EPIs. The lines with different slopes are located by maximizing the distribution distances of the regions. Zhang *et al*. [28] located the optimal slope of each line segmentation on EPIs by using the locally linear embedding. Differing from these methods, some methods applied CNNs to extract light field features from EPIs. Sun *et al*. [25] presented a data-driven approach to estimate the object depths from an enhanced EPI feature using CNN. Heber and Pock [4] used CNNs for predicting 2D per-pixel hyperplane slope orientations in EPIs. Based on this work, Heber *et al*. [5, 6] improved their work by utilizing an U-shaped network and EPI volumes to predict the depth map. Luo *et al*. [15] designed an EPI-patch based CNN architecture to estimate the depth of each pixel. Feng *et al*. [2] proposed a two-stream network that learns to estimate the depth values of multiple correlated neighborhood pixels from EPI patches. Shin *et al*. [19] introduced a multi-stream network to extract features for epipolar property of four viewpoints with horizontal, vertical and both diagonal directions. This method reaches state-of-the-art results on the 4D light field benchmark [7]. One of the most recent works by Leistner *et al*. [12] shift the light field stack to retain a small receptive field, which improves the performance of depth estimation for large-disparity light fields. Some of these previous methods [2, 6, 12, 15] require data pre-processing and subsequent optimization. In contrast, we present an end-to-end fully convolutional network architecture to predict the depth values of center pixels from the cor-

responding horizontal and vertical EPIs. We explore the similar linear structure information in EPIs and model the relationship between the oriented lines and their neighboring pixels, which help to estimate the slope of the oriented line.

Relation modeling. A few recent papers [8, 16, 29] have shown that relations have been exploited to improve the performance of computer vision tasks. Zhou *et al*. [29] proposed a temporal relation network to learn and reason about temporal dependencies between video frames at multiple time scales. Hu *et al*. [8] proposed an object relation module to model relationships between sets of objects for object detection. Mou *et al*. [16] proposed the spatial and channel relation modules to learn and reason about global relationships between any two spatial positions or feature maps, and then produced relation-augmented feature representations for semantic segmentation. Motivated by these works, we propose a oriented relation module to construct the relationship between the center pixel and its neighborhood in the EPI, which allows the network to explicitly learn the relationship between the line orientations and improve the performance of depth estimation.

3 Proposed Method

In this paper, we present an end-to-end fully convolutional network to predict the depth values of center pixels in EPIs of light fields. Two branches are designed to process the horizontal and vertical EPIs separately. The newly proposed oriented relation module is capable of modeling the relationships between the neighboring pixels in EPIs. A refocusing-based EPI augmentation method is also proposed to facilitate training and improve the performance of depth estimation. An overview of the network architecture is shown in Figure 1.

3.1 EPI Patches for Learning

The light field, indicated as $L(u,v,x,y)$, is generally represented by the two-plane parameterization [13]. Here, (x,y) and (u,v) are spatial and angular coordinates, respectively. The central sub-aperture (center view) image is formed by the rays passing through the optical center of the camera main lens ($u = u_0, v = v_0$). As shown in Figure 2, given a pixel $P(x_i,y_i)$ in the center view image, the horizontal EPI of the row view v_0 can be formulated as $L(u,v_0,x,y_i)$, which is centered at (u_0,x_i). Similarly, the vertical EPI of the column view u_0, with the center at (v_0,y_i), is written as $L(u_0,v,x_i,y)$.

Z can be obtained by analyzing the slope $\frac{\Delta x}{\Delta u}$ of the line [23],

$$\Delta x = -\frac{\Delta u}{Z} f \tag{1}$$

where f is the focal distance and Z is the depth value of the point P. The slope of the oriented line is shown in the EPI patch b of Figure 2.

To learn the slope of the oriented line of $P(x_i,y_i)$, we extract patches of size $H \times W \times C$ from $L(u,v_0,x,y_i)$ and $L(u_0,v,x_i,y)$ as inputs. Here, H and W indicate height and width of the patch, respectively, and C is the channel dimension. The size of the patch is determined by the range of disparities. The proposed network predicts the depth of the center pixel from the pair of EPI patches.

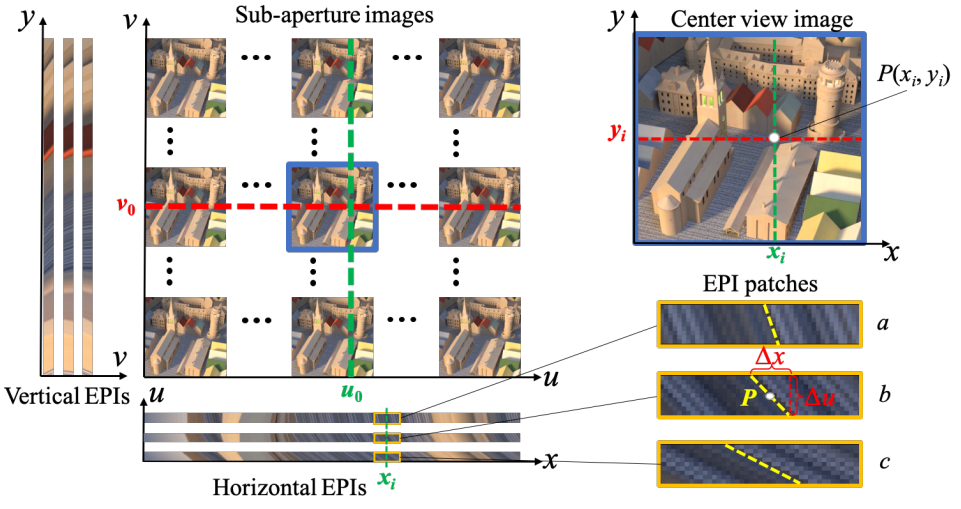

Figure 2: EPIs from the light field: each image in angular coordinates (u,v) yields a sub-aperture view of the scene. Given a pixel $P(x_i,y_i)$ in the spatial coordinate, its horizontal or vertical EPIs are obtained by fixing the view v_0 or u_0, respectively. Three pairs of horizontal and vertical EPIs at different refocused depths are shown. The similar linear structure information between the oriented line marked by the yellow and its neighboring pixels is shown among three EPI patches. The disparity $\frac{\Delta x}{\Delta u}$ of the EPI patch b describes the pixel shift of the scene point P when moving between the views.

3.2 Network Architecture

As shown in Figure 1, the proposed network shares the similar structure with the pseudo-Siamese network proposed in [26], where two branches are designed to learn the weights for the horizontal and vertical EPI patches, respectively. Each branch contains two oriented relation modules (ORMs), a set of seven convolutional blocks, a residual module (RM), and a merging block. The ORM will be discussed in Sect. 3.3. The convolutional block is composed of 'Conv-ReLU-Conv-BN-ReLU'. To handle the small EPI slope, we apply the convolutional filters with size of 2×2 or 1×2 and stride 1 to measure a small depth value. However, detailed information of the EPI slope is lost as the network goes deeper. Inspired by the residual learning [3] that can introduce detailed information of the shallower layer into the deeper layer and effectively improve the network performance, we design a residual module for each branch. The residual module consists of six residual blocks, each of which consists of one convolutional block and one skip connection. We take a slicing operation to implement the skip connection by extracting the center region of the input feature. The final merging block, containing two different convolutional blocks ('Conv-ReLU-Conv-BN-ReLU' and 'Conv-ReLU-Conv'), is used for fusing the horizontal and vertical EPI features to predict the depth value of each pixel.

3.3 Oriented Relation Module

We propose a new Oriented Relation Module (ORM) to reason about the relationship between the center pixel and its neighborhood in each EPI patch. As shown in Figure 3, given

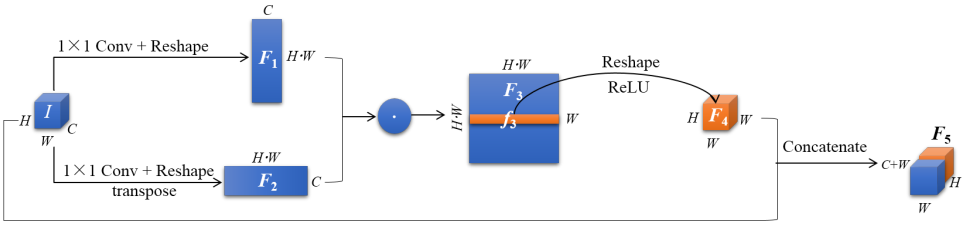

Figure 3: The proposed oriented relation module.

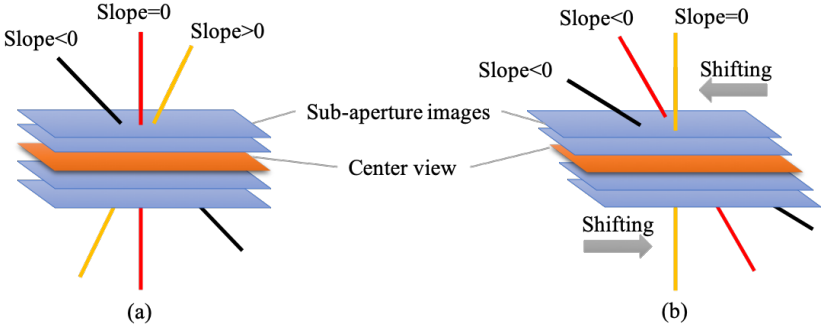

Figure 4: The light field refocusing. (a) before refocusing. (b) after refocusing.

an EPI patch $\boldsymbol{I}$ of size $H \times W \times C$, we apply two single-layer convolutions of 1×1 kernel size to model a compact relationship in the EPI patch. The output features are converted into $\boldsymbol{F_1}$ and $\boldsymbol{F_2}$, respectively, which are followed by a dot product to construct the oriented relation feature $\boldsymbol{F_3}$ of size $(H \cdot W) \times (H \cdot W)$. Furthermore, to obtain the relationship between the center pixel and its neighborhood in $\boldsymbol{F_3}$, we extract the feature $\boldsymbol{f_3}$ of size $W \times (H \cdot W)$ from the relational feature $\boldsymbol{F_3}$. Then, we apply the reshaping and ReLU activation on the feature $\boldsymbol{f_3}$ to obtain a new feature $\boldsymbol{F_4}$ of size $W \times H \times W$. Finally, we concatenate the original EPI patch $\boldsymbol{I}$ with the feature $\boldsymbol{F_4}$ to obtain the output feature $\boldsymbol{F_5}$ of size $W \times H \times (W + C)$.

3.4 EPI Refocusing-based Data Augmentation

To alleviate the problems of insufficient data and overfitting, we propose a new data augmentation method by refocusing EPIs. Differing from general augmentation techniques such as rotation, scaling and flipping [19], we refocus EPIs to generate multiple EPIs focused at different depth levels. The light field refocusing shifts the sub-aperture images to obtain images focused at different depth planes [1]. Figure 4 shows sub-aperture images at the same horizontal or vertical views that are stacked together. Lines with different slopes (i.e. the lines in EPIs) are inserted into the scene points of different depth planes on the sub-aperture images. The line at the focal depth should be vertical (slope = 0), while the other lines are inclined (slope > 0 or slope < 0). Taking the center view as the reference, the disparity shift between sub-aperture images changes the slope of the line. Thus, refocusing at a different depth plane changes the orientation of the structure in the EPI.

We convert the depth information into a disparity shift in every single EPI according to

Eq. 1. The resulting disparity shift $\Delta x(u)$ related to the depth Z is defined following [1],

$$\Delta x(u) = (u_0 - u)\frac{\Delta u}{Z} f \tag{2}$$

Here, we assume that the center view (u_0, v_0) is the reference view. For the sake of simplicity, we also assume that lenslet-based cameras have the same focal length f and the same baseline Δu for the neighboring views. Similarly, we can obtain the disparity shift $\Delta y(v)$. Then we refocus the EPI based on the refocusing principle [18],

$$L(u,v,x,y) = L(u,v,x+\Delta x(u), y+\Delta y(v)) \tag{3}$$

The EPI patches (a,b,c) in Figure 2 show three horizontal EPIs at different refocused depths. Our strategy not only changes the slope of the orientation line but also changes the corresponding ground truth (Eq. 2).

4 Experiments

4.1 Implementation Details

Following previous works [2, 15], we use the 4D light field benchmark [7] as our experimental dataset, which provides highly accurate disparity ground truth and performance evaluation metrics. The dataset includes 24 carefully designed scenes with ground-truth disparity maps. Each scene has 9×9 angular resolution and 512×512 spatial resolution. 16 scenes are used for training and the remaining 8 scenes for testing. We randomly sample the horizontal and vertical EPI patch pairs of size $9 \times 29 \times 3$ from each scene as inputs. To avoid overfitting, we increase the training data to 8 times the original data by the proposed EPI refocusing-based data augmentation.

The bad pixel ratio (BadPix) [7], which denotes the percentage of pixels whose disparity error is larger than 0.07 pixels, as well as the Mean Square Errors (MSE) are computed for evaluation metrics. Given an estimated disparity map d, the ground truth disparity map gt and evaluation region M, BadPix is defined as,

$$\text{BadPix} = \frac{|\{x \in M : |d(x) - gt(x)| > 0.07\}|}{|M|}, \tag{4}$$

and MSE is defined as,

$$\text{MSE} = \frac{\sum_{x \in M} (d(x) - gt(x))^2}{|M|} \times 100. \tag{5}$$

Lower scores are better for both metrics.

We use the Keras library [10] with the mean absolute error (MAE) loss to train the proposed network from scratch. We formularize the depth estimation as a multi-label regression problem to estimate the depth value of a single pixel. Note that the network is trained end-to-end and does not make use of pre- and post-processing complications. We utilize the RMSprop optimizer [30] and set the weight decay rate to $1e-5$ and batch size to 128. Our network training takes one day for $750k$ iterations on an NVIDIA GTX 1080 Ti 11GB GPU. The memory footprint is about 65%.

Metric	Baseline	w/ ORM	w/ EPIR	Full model
BadPix	9.45	7.98	7.41	**5.66**
MSE	2.058	1.621	1.475	**1.393**

Table 1: Effects of ORM and EPIR. Bold: the best.

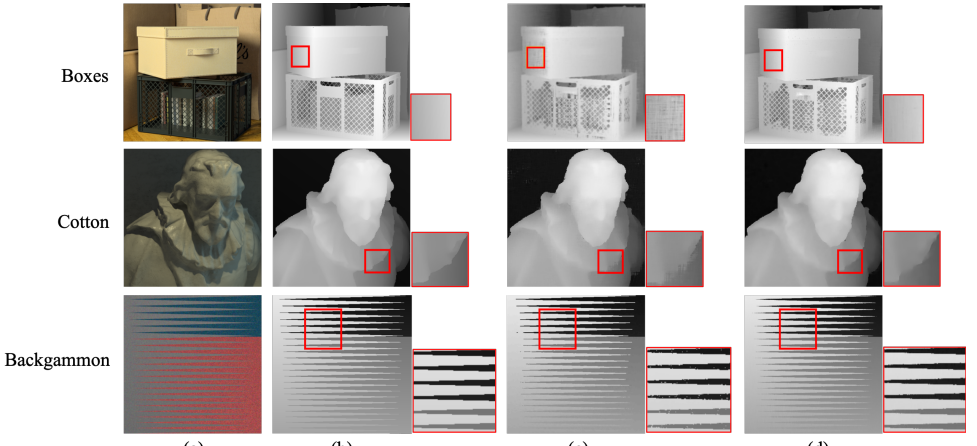

Figure 5: Qualitative comparison of the baseline and the network with the ORM. (a) Original scenes. (b) Ground truth maps. (c) Baseline. (d) Our network with the ORM.

Metric	Baseline	EPIR×2	EPIR×4	EPIR×6	EPIR×8	EPIR×10
BadPix	9.45	8.03	7.78	7.50	**7.41**	7.45
MSE	2.058	1.781	1.511	1.482	**1.475**	1.480

Table 2: Performance in terms of the number of EPI refocusing. Bold: the best.

4.2 Ablation Study

We use the proposed network without the oriented relation module (ORM) and data augmentation based on EPI refocusing (EPIR) as the **Baseline**.

Effect of the oriented relation module. Table 1 shows that the network using the ORM brings a significant improvement over the baseline, which can reduce the BadPix by around 1.5. Figure 5 shows qualitative results for comparison. *Boxes* and *Cotton* show that the ORM can reduce the streaking artifacts and improve the accuracy in weakly textured areas. The occlusion boundaries in *Backgammon* with multiple occlusions can also be better restored through the ORM. In addition, our network with ORM generates smooth depth maps while preserving discontinuity between different objects, yielding the increased MSE by about 21% compared to the baseline.

Effect of EPI refocusing-based data augmentation. From Table 1, we can see that the network using the EPIR is better than the baseline. Moreover, by using both the ORM and the EPIR, the performance is further boosted. To further show the effect of EPI refocusing in the network, we compare the performance by varying the number of refocusing in Table 2. We refocus the training data to the foreground and the background of the original depth plane. From the table, we observe that there are performance gains when increasing the number of refocusing. However, there is no gain from EPIR×8 to EPIR×10 when comparing with Table 2.

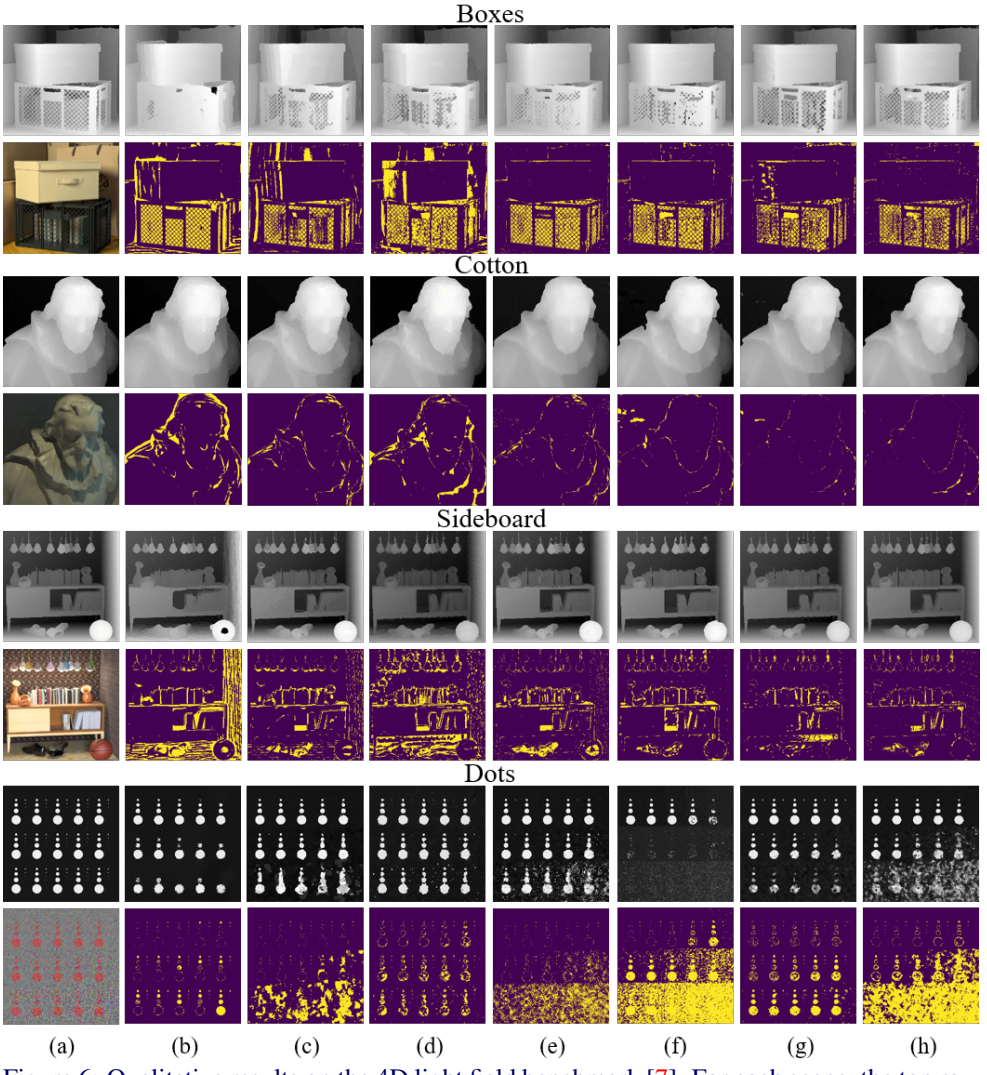

Figure 6: Qualitative results on the 4D light field benchmark [7]. For each scene, the top row shows the estimated disparity maps and the bottom row shows the error maps for BadPix. (a) Ground truth. (b) LF [9]. (c) CAE [24]. (d) LF_OCC [21]. (e) SPO [27]. (f) EPN [15]. (g) EPINET [19]. (h) Ours.

4.3 Comparison with State-of-the-Arts

We compare our approach with other state-of-the-art methods: LF [9], CAE [24], LF_OCC [21], SPO [27], EPN [15], and EPINET [19]. The qualitative comparison is shown in Figure 6. The *Cotton* scene contains smooth surfaces and textureless regions, and the *Boxes* scene consists of occlusions with depth discontinuity. As can be seen from the figure, our approach can reconstruct the smooth surface and the region with sharp depth discontinuity compared to other methods. For the *Sideboard* scene with the complex shape and texture, our approach preserves more details and sharper boundaries by distinguishing the subtle difference of EPI slopes. In addition, our approach obtains better disparity maps in the *Boxes* and *Sideboard*

Scenes	LF [9]	CAE [24]	LF_OCC [21]	SPO [27]	EPN [15]	EPINET [19]	Ours
boxes	24.572	17.885	24.526	15.889	15.304	14.190	13.373
cotton	8.794	3.369	6.548	2.594	2.060	0.810	0.869
dino	21.478	4.968	15.466	2.184	2.877	2.970	2.814
sideboard	23.906	9.845	17.923	9.297	7.997	6.260	5.580
backgammon	4.810	3.924	18.061	3.781	3.328	4.130	2.511
dots	2.441	12.401	5.109	16.274	39.248	9.370	25.930
pyramids	10.949	1.681	2.830	0.356	0.242	0.540	0.240
stripes	35.394	7.872	17.558	14.987	18.545	5.310	5.893

Table 3: Quantitative comparison of different methods using the BadPix metric. The best three results are shown in red, blue, and green, respectively (Best viewed in color).

Scenes	LF [9]	CAE [24]	LF_OCC [21]	SPO [27]	EPN [15]	EPINET [19]	Ours
boxes	16.705	8.424	9.095	9.107	9.314	6.440	4.189
cotton	11.773	1.506	1.103	1.313	1.406	0.270	0.313
dino	1.558	0.382	1.077	0.310	0.565	0.940	0.336
sideboard	4.735	0.876	2.158	1.024	1.744	0.770	0.733
backgammon	15.109	6.074	20.962	4.587	3.699	4.700	1.403
dots	4.803	5.082	2.731	5.238	22.369	3.320	6.754
pyramids	0.243	0.048	0.098	0.043	0.018	0.020	0.016
stripes	17.380	3.556	7.646	6.955	8.731	1.160	1.263

Table 4: Quantitative comparison of different methods using the MSE metric. The best three results are shown in red, blue, and green, respectively (Best viewed in color).

scenes than the recent state-of-the-art method [19], which uses the vertical, the horizontal, the left diagonal and the right diagonal viewpoints as inputs. The number of the viewpoints is almost double that of our approach. Compared with the 28-layer network of 4 branches in [19], our network consists of 30 layers with 2 branches, which makes our trainable parameters be about half of those in [19]. However, our network cannot produce good depth predictions for the *Dots* scene that contains a lot of noise, which is also the common downside of applying EPIs to the CNN-based method (*e.g.* EPN [15]). The reason is that noises may lead to the false straight line estimation of EPI patches. Therefore, one of the future works could introduce global constraints of oriented lines into our model.

Quantitative results are shown in Tables 3 and 4, which show that the proposed approach performs the best in 4 out of 8 scenes. In particular, the proposed approach predicts more accurate disparity values on the *Boxes* and *Backgammon* scenes under multi-occlusions. Note that we do not apply any post-processing for depth optimization while most other methods [9, 15, 21, 24, 27] are accompanied by post optimization.

5 Conclusion

In this paper, we propose an end-to-end fully convolutional network for depth estimation from light fields by exploiting horizontal and vertical EPIs. We introduce a new relational reasoning module to construct the relationship between oriented lines in EPIs. In addition, we propose a new data augmentation method by refocusing the EPIs. We demonstrate the effectiveness of our approach on the 4D light field benchmark [7]. Our approach is competitive with the state-of-the-art methods, and is able to predict more accurate disparity map in some challenging scenes such as *Boxes* and *Sideboard* without any post-processing.

Acknowledgments. This work was supported by the National Natural Science Foundation of China, No. 61876057.

References

[1] Maximilian Diebold and Bastian Goldluecke. Epipolar plane image refocusing for improved depth estimation and occlusion handling. In *Proceedings of Vision, Modelling & Visualization (VMV)*, 2013.

[2] 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 Transactions on Image Processing*, 27(7):3586–3598, 2018.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] S. Heber and T. Pock. Convolutional networks for shape from light field. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3746–3754, 2016.

[5] S. Heber, W. Yu, and T. Pock. Neural EPI-volume networks for shape from light field. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 2271–2279, 2017.

[6] Stefan Heber, Yu Wei, and Thomas Pock. U-shaped networks for shape from light field. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 37.1–37.12, 2016.

[7] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, pages 19–34, 2016.

[8] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei. Relation networks for object detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018.

[9] H. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. Tai, and I. S. Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1547–1555, 2015.

[10] Z. Jiang and G. Shen. Prediction of house price based on the back propagation neural network in the keras deep learning framework. In *Proceedings of International Conference on Systems and Informatics (ICSAI)*, pages 1408–1412, 2019.

[11] Ole Johannsen, Christian Heinze, Bastian Goldluecke, and Christian Perwa. *On the Calibration of Focused Plenoptic Cameras*. Springer Berlin Heidelberg, 2013.

[12] 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 *Proceedings of 2019 International Conference on 3D Vision (3DV)*, pages 249–257, 2019.

[13] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (ACM SIGGRAPH)*, page 3142, 1996.

[14] H. Lin, C. Chen, S. B. Kang, and J. Yu. Depth recovery from light field using focal stack symmetry. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3451–3459, 2015.

[15] Y. Luo, W. Zhou, J. Fang, L. Liang, H. Zhang, and G. Dai. EPI-patch based convolutional neural network for depth estimation on 4D light field. In *Proceedings of International Conference on Neural Information Processing (ICONIP)*, pages 642–652, 2017.

[16] L. Mou, Y. Hua, and X. X. Zhu. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12408–12417, 2019.

[17] Ren Ng, Marc Levoy, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. *Computer Science Technical Report*, 2005.

[18] Ng Ren. *Digital light field photography*. PhD thesis, Stanford University, 2006.

[19] C. Shin, H. 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 *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4748–4757, 2018.

[20] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 673–680, 2013.

[21] T.-C. Wang, A. Efros, and R. Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 3487–3495, 2015.

[22] S. Wanner and B. Goldluecke. Globally consistent depth labeling of 4d light fields. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–48, 2012.

[23] S. Wanner and B. Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):606–619, 2014.

[24] W. Williem and I. K. Park. Robust light field depth estimation for noisy scene with occlusion. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4404, 2016.

[25] Xing Sun, Z. Xu, Nan Meng, E. Y. Lam, and H. K. H. So. Data-driven light field depth estimation using deep convolutional neural networks. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 367–374, 2016.

[26] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4353–4361, 2015.

[27] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016.

[28] Y. Zhang, H. Lv, Y. Liu, H. Wang, X. Wang, Q. Huang, X. Xiang, and Q. 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, 2017.

[29] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 831–846, 2018.

[30] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11119–11127, 2019.