Wide Range Depth Estimation from Binocular Light Field Camera

Feng Dai[1]
fdai@ict.ac.cn
Xianyu Chen[1,2]
chenxianyu@ict.ac.cn
Yike Ma[1]
ykma@ict.ac.cn
Guoqing Jin[1]
jinguoqing@ict.ac.cn
Qiang Zhao ✉[1]
zhaoqiang@ict.ac.cn

[1] Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

[2] University of Chinese Academy of Sciences
Beijing, China

Abstract

Light field camera have been developed to capture spatial and angular information of rays. But limited by the structure of micro-lens, it acts as a short-baseline multi-view camera. Foreground objects have accurate depth map in light field while depth information is missing in the background. To avoid such a drawback in the existing light field depth estimation system, we propose a binocular light field camera and introduce long-baseline stereo matching in it. The system can estimate wide range depth of scene by merging complementary depth map of far scene into depth map from light field camera. We firstly estimate an relative depth map from light field and stereo matching respectively, and present calibration methods that normalize both depth maps to the real depth space. Then we model depth fusion problem as Markov random field which can be solved by graph cuts efficiently. Experiments show that our system have a wider depth sensing ability than either single light field camera or traditional binocular camera.

1 Introduction

Depth map estimation from images is a long standing problem in computer vison, which has usage in many applications. By providing angular and spatial information of light rays, light field images offer an easy way for depth map estimation. Currently some algorithms [15] [13] [14] have been proposed to estimate depth map from a single light field image by leveraging the special structure of light field data. However because the effective baselines of light field cameras are much small, these methods can only reconstruct high quality depth map for near scene. If the scene is far from the light field cameras, the recovered depth map has low accuracy.

In this paper, we estimate the depth map from a binocular light field camera system. As depth map estimation from light field image performs well for near scene and stereo

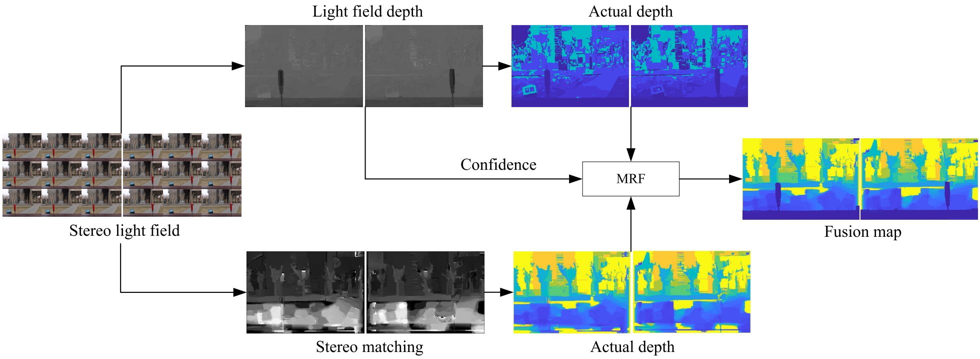

Figure 1: Pipeline of depth map fusion process.

matching performs well for far scene, our method fuses the depth maps from these two methods. Thus our method inherits the advantages of both methods and can improve the accuracy of the depth map. Specifically, our binocular camera system captures two light field images that have large baseline. Then we estimate depth maps using the above two methods. After calibrating the camera system, we transform the estimated depth maps into those with real depth values. Finally we fuse the transformed depth maps using Markov Random Field mehtod, which choses final result from different depth maps based on which value is more confident. Experimental Result shows that the final depth map has high quality for both near scene and far scene. The proposed method is more robust than either stereo matching or depth estimation from single light field image.

2 Related Work

2.1 Light Field Camera Depth

The micro-lens structure of light field camera is first proposed by Ng Ren[10]. It can record both spatial and angular information of light in a single shot. Disparity in light field can be viewed as a special pattern in the epipolar plane images. Methods, like [18][15], calculate depth based on epipolar plane images using line fitting or structure tensor. The advanced light field depth methods will use either corresponding points or focal shifting, the characteristic of light field[13][14]. For example, Wang's algorithm[14] mainly deals with the occlusion problem in light field and models it in angular space. In its implementation, it explicitly detects edge in the center view of angular space and use a color consistency constraint to filter the incorrect depth of occlusion. The algorithm also includes a regularization step, which uses initial depth[13] as data term, correspondence cue and refocus cue as smooth terms. Other works [16, 19] compute the angular super-resolution of light field images, which can improve the quality of the output of light field depth estimation methods.

2.2 Stereo Matching

Various binocular systems have been proposed to for depth map refinement or image enhancement. General stereo matching is a pair of pin-hole cameras. It has a complete and effective taxonomy, including cost computation, cost aggregation, optimization and refinement, proposed by Scharstein and Szeliski[11]. Semi-global matching[5] is one of effective traditional method, which use mutual information to approximate global pixel matching. Ensemble method, like [9], uses absolute differences and census[4] patch feature to measure cost jointly. The method also introduces a cross-based aggregation method[17] and semi-global matching[5] is then used to optimize and propagate cost from high confidence areas to noisy areas. Finally, some refinement tricks are introduced to enhance the quality of disparity.

Currently there are many stereo matching methods using deep learning network to improve the effect[7][12][6]. At first, [7] use convolutional neural network to generate pixel-wise cost for traditional cost function substitution. Akihito and Marc propose a network for predicting penalties in SGM framework to achieve better result[12]. Luo etc. use a siamese architecture to model the stereo problem and merge features from left and right image by inner product[8]. Kendall etc.[6] further develop [8] by concatenate features directly and apply a regularization network at output.

2.3 Fusion Binocular Systems

There are also some special binocular systems consisting of heterogeneous cameras. Zhu[20] and Gandhi[3] built depth fusion systems by combining a time-of-flight range sensor and two pin-hole camera. Cameras in these systems estimate the same range of depth, so it aims at improving quality of depth map. After they capture images and depth information of scene, they merge two depth maps by seed-growing algorithm or graph optimization, respectively. Alam[1] present a hybrid system including a light field camera and a regular camera. The device can capture a light field image and a high-resolution image simultaneously. Then he calculates the warping relation between regular image and all sub-aperture images. The light field is up-sampled using these warping cues, which results in a high-resolution rays set and accurate depth map in post light field process. Although it explores the warping information, it is still a short-baseline camera and the sensing ability of depth range is not increased. In summary, these binocular systems are designed for specific purposes, but none of them consider the wide depth range application.

3 Wide Range Depth Estimation

We propose a binocular light field system for wide range depth estimation. In this section, we introduce how to obtain a depth map by this system. Firstly we calculate a relative depth map from light field and stereo matching respectively. Because the output of these two algorithms may have different scales, we then propose a calibration method for obtaining camera parameters, and map both relative depth maps to real depth based on those parameters. Finally, we fuse the two processed depth map together.

3.1 Light Field Camera Scene Depth

We extract relative depth by light field depth algorithm[14][1]. To achieve an accurate mapping, we develop a procedure to calibrate the light field camera including a curve fitting step, which actually convert the result of depth algorithm to real scene depth.

[1] Please note that although we use this specific light field estimation method, other methods can also be used.

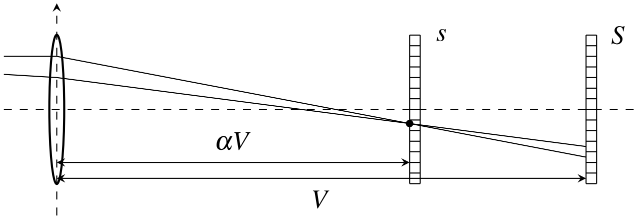

Figure 2: Set a new virtual image plane s in the proper position and render a refocused image with rays from light field. Changing focal equals to change value α, the ratio of virtual image plane s to real image plane S .

3.1.1 Real Depth from Light field Depth

We assume that every single view in a light field image is captured by a pin-hole camera model which simply contains one convex lens. The imaging formula of this optic model is

$$\frac{1}{u}+\frac{1}{v}=\frac{1}{f}, \tag{1}$$

where u is object distance, v is image distance and f is focal of the convex lens.

In Ng's paper[10], he proposed a ray selecting method to achieve digital refocusing. He set a variable α to control the distance of virtual image plane.

$$v=\alpha V. \tag{2}$$

When rays from the same scene point add up together, the virtual image plane will have sharp point rather than a blur blob. So α is actually a value indicating real depth in the scene. From (1) and (2), we can get the formula of depth mapping.

$$u=\frac{\alpha V f}{\alpha V-f}. \tag{3}$$

The result α in above calculation is a relative depth. To avoid the scale problem in fusion, we develop a calibration method to map α to real depth. From the formula (3), the calculation need at least the parameter V and f. However, for most cameras, manufacturer will not make many key parameters of camera known to the public. And some key parameters may change by users and affect the depth mapping. So we calibrate light field cameras by fitting the camera model to ground truth and estimate the parameter in different camera setting.

3.1.2 Calibration and Depth d_{lf} Calculation

To generate the ground truth, we place a chess board in front of a fixed light field camera. Then we move it away from camera gradually and take a picture of chess board for every 10 cm. The distance between camera and the board is also recorded. We use algorithm[14] to calculate depth of the board. Depth value within the chess board area will be cut out for estimation. We then aggregate pixels in these areas and calculate an average as the estimation at that certain distance. We plot these points in Fig.5 in our experiment. It is easy to find that the result is not linear to the real depth. Inspired by the formula (3), we choose a similar form to fitting the depth mapping function

$$d_{lf}(\alpha)=\frac{\alpha}{a\alpha+b}, \tag{4}$$

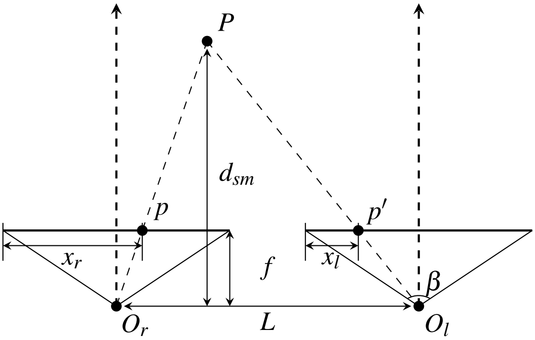

Figure 3: Conversion from disparity to depth.

where d_{lf} is the final real depth, and α is the relative depth we estimated. Since light field depth affects final real depth fusion directly, we use a robust and accurate light field depth algorithm[14] in our method. And it finally outputs a continues value α of light field shearing, see Fig.2. We apply the algorithm to both light field images directly and get a depth observation α.

We use non-linear least square method to estimate the parameter in the mapping function. The range of mapping function is only available within the distance we recorded. Depth beyond the definition domain is incorrect because we have no data to calibrate it. One reason is that the calibrated distant is limited, and the other reason is that the baseline of light field camera is short and scene view far away have no disparity in light field. These two reasons constrain the ability of light field depth estimation.

3.2 Stereo Matching Scene Depth

Depth from stereo matching can estimate scene information far from camera. We calculate depth map from stereo matching. Then we derive a formula to convert relative depth to real depth. We again propose a calibration method to obtain missing parameter of the camera in the formula.

3.2.1 Real Depth from Stereo Matching

Light field camera outputs a set of multi-view images and we use center view in stereo matching. We adopt method in [9] to calculate relative depth. In stereo matching, the result is disparity between views in left camera and right camera. Disparity is not linear to the real depth. So we derive a depth formula for this conversion.

In a binocular system, two cameras are displaced horizontally and set to face the same direction, see Fig.3. Gap between cameras is set to L, which is the key parameter affecting the ability of depth sensing. We assume two cameras have the same horizontal field of view and the resolution in horizontal direction. The disparity of scene point P is the distance between image point p and p' in the same image space, $d = x_r - x_l$. Then we can derive the real depth formula from triangle similarity as

$$\frac{L}{d_{sm}} = \frac{L-d}{d_{sm}-f}, \tag{5}$$

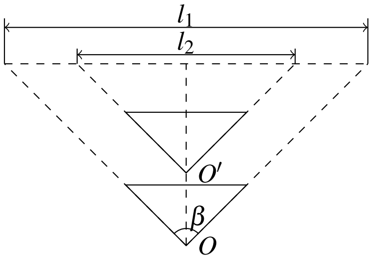

Figure 4: Estimating β to calculate focal.

where d_{sm} is the real depth of the scene point P.

3.2.2 Depth d_{sm} Calculation

Further, the equivalent focal f is sill unknown because of special structure of light field camera. We develop a method to estimate the focal by measuring the horizontal field of view β. We mount the light field camera on a camera slider and fix a ruler in front of the camera. Camera will shot at various positions and the field of view will be recorded by the length l_1 and l_2 of ruler appeared in the images, see Fig.4. We can calculate the horizontal field of view β as

$$\beta = 2\tan^{-1}\left(\frac{l_1 - l_2}{2 l_{OO'}}\right). \tag{6}$$

We calculate horizontal field of view β in such indirect way because we do not know exact position of optic center in this camera. The relative distance help us figure out the horizontal field of view β without knowing optic center.

And next, we can get the focal f in (5) using β,

$$f = \frac{r}{\tan\frac{\beta}{2}}. \tag{7}$$

Based on (5) and (7), we get

$$d_{sm}(d) = \frac{rL}{d\tan\frac{\beta}{2}}, \tag{8}$$

where d is the disparity between left and right images, which is calculated by [9], and r is the resolution in horizontal direction.

3.3 Depth Fusion of Short and Wide Baseline camera

To merge depth from different algorithms smoothly, we need an algorithm that can globally optimize the depth map. Besides, depth should not be modified during the optimization. Otherwise, depth map will no longer guarantee the accuracy and continuity of depth value. So we model the fusion problem as a label choice in Markov random field and solve it by Graph Cut[2].

We have two depth maps after both light field depth conversion and stereo depth conversion. For each pixel in the original scene, depth should be defined by one of these two maps.

The selection of depth map of pixels depends on confidence of foreground and background. As we mentioned in Section 3.1, depth which is out of calibration range or beyond the sense ability of short baseline camera, is not trustworthy. Objects far from cameras will leave no disparity in light field, but distinctive in stereo matching. Therefore, curve of the mapping function (4) of light field depth to real depth slows down as real depth grows, see Fig.5. We take the gradient of the mapping function (4) as observation of confidence of light field depth. Depth gradient $\mathrm{d}(d_{lf}) = b/(a\alpha+b)^2$ is along α in (4). And we define the confidence function in MRF as

$$c = 1 - \tan^{-1}\left(\frac{kb}{(a\alpha+b)^2}\right), \tag{9}$$

where k is a user setting scale factor, which is 0.001 in our experiment.

The data term in MRF model should be defined by confidence of light field depth because of its limited definition domain.

$$E_{unary}(p) = \begin{cases} 1-c & \text{if } p \text{ is defined by } d_{lf} \\ c & \text{if } p \text{ is defined by } d_{sm} \end{cases}. \tag{10}$$

The smooth term can constrain the propagation of depth between dissimilar pixels and suppress depth noise within an object. We simply use color gradient of image in the smooth term

$$E_{smooth}(p,p') = \|I(p) - I(p')\|_2, \tag{11}$$

where I is image and p is label of depth of one pixel.

The final cost function is a standard MRF model. We will minimize the cost function using graph cut [2] to get a final label.

$$E(I) = \sum_{p\in I} E_{unary}(p) + \sum_{p'\in Near\{p\}} E_{smooth}(p,p'). \tag{12}$$

By picking out depth value from corresponding depth map, we will finally get a wide range depth map.

4 Experiments

4.1 Experiment Data Introduction

In our practical installation, we mount a Lytro Illum camera on a slider rail and make sure the direction of camera is perpendicular to slider. We restrict the gap between two captures to 23 centimeters, which is essential for the ability to sense far scene. We use wide angle setting in camera so that two views will have a larger overlap area. We also use the manual mode to reduce influence of color difference.

Our algorithm performs well if scenes have both close and far objects. Considering there is no existing dataset that has binocular light field images, we choose to take images by our camera. Then we preprocess them using Lytro Power Tools to get multi views from camera raw files. Our experiment data contains four pairs of light field image both indoor and outdoor.

Besides, we also take images for light field calibration. It contains 50 images in which distance between camera and chess board varies from 30 to 500 centimeters. The calibration images should be unique to one single light field camera and one set of camera setting.

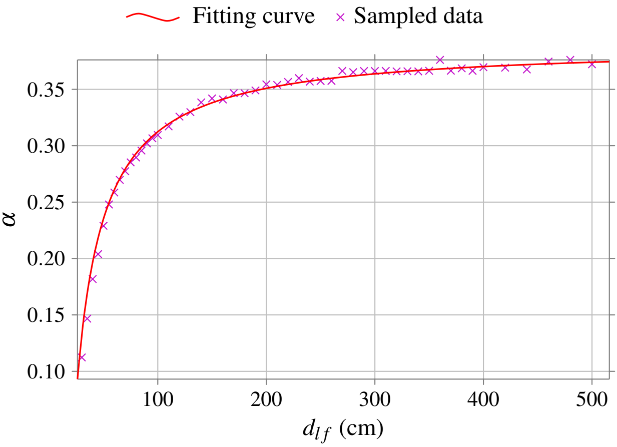

Figure 5: Light field depth calibration and curve fitting for estimate the real depth from α. The generation process is described in section 3.1.

.

4.2 Wide Range Scene Depth Result

We use our calibration images to obtain some key parameters for light field depth normalization. After depth estimation, we fit formula (4) to the sampled calibration data. Both the fitted curve and calibration data are plotted in Fig.5. The parameters in (4), a is -0.03487 and b is 0.01389. It is clear that the curve fits calibration data well. Also, as the chess board moves away from camera, the value α gradually is getting indistinguishable and depth of this area is not trustworthy.

Our lens focal are set to wide angle. In our measurement, horizontal field of view β under our camera experiment setting is 60 degrees and horizontal resolution is 541 pixels. So the available depth estimation range is from 18.5 centimeters to 100.3 meters. As Fig. 6 shows, some foreground objects have wrong depth because of lack of disparity in left and right images. The missing depth values are filled with result from light depth map.

In Fig.6, we display results of each stages. Dark blue color means objects are very close to scene point while light colors, yellow and light blue label far scenes. We can see that the final depth map takes the content from that generated by light field depth estimation method for nearby objects, while takes the content from that generated by stereo matching for far away objects. This validates that the depth maps generated from different methods have complementary depth-sensing ability. Light field depth method can detect occlusion in front of the camera accurately. It will not be affected by the other camera at all. Especially, if some objects move toward our binocular light field cameras, our system can still capture it and calculate its distance.

5 Conclusion and Future Work

In this paper, we roughly explore depth map fusion of binocular light field, and we find that it can really expand the range of depth calculation. A major obstacle of the fusion is the calibration of light field camera and the binocular system. We may not have much knowledge

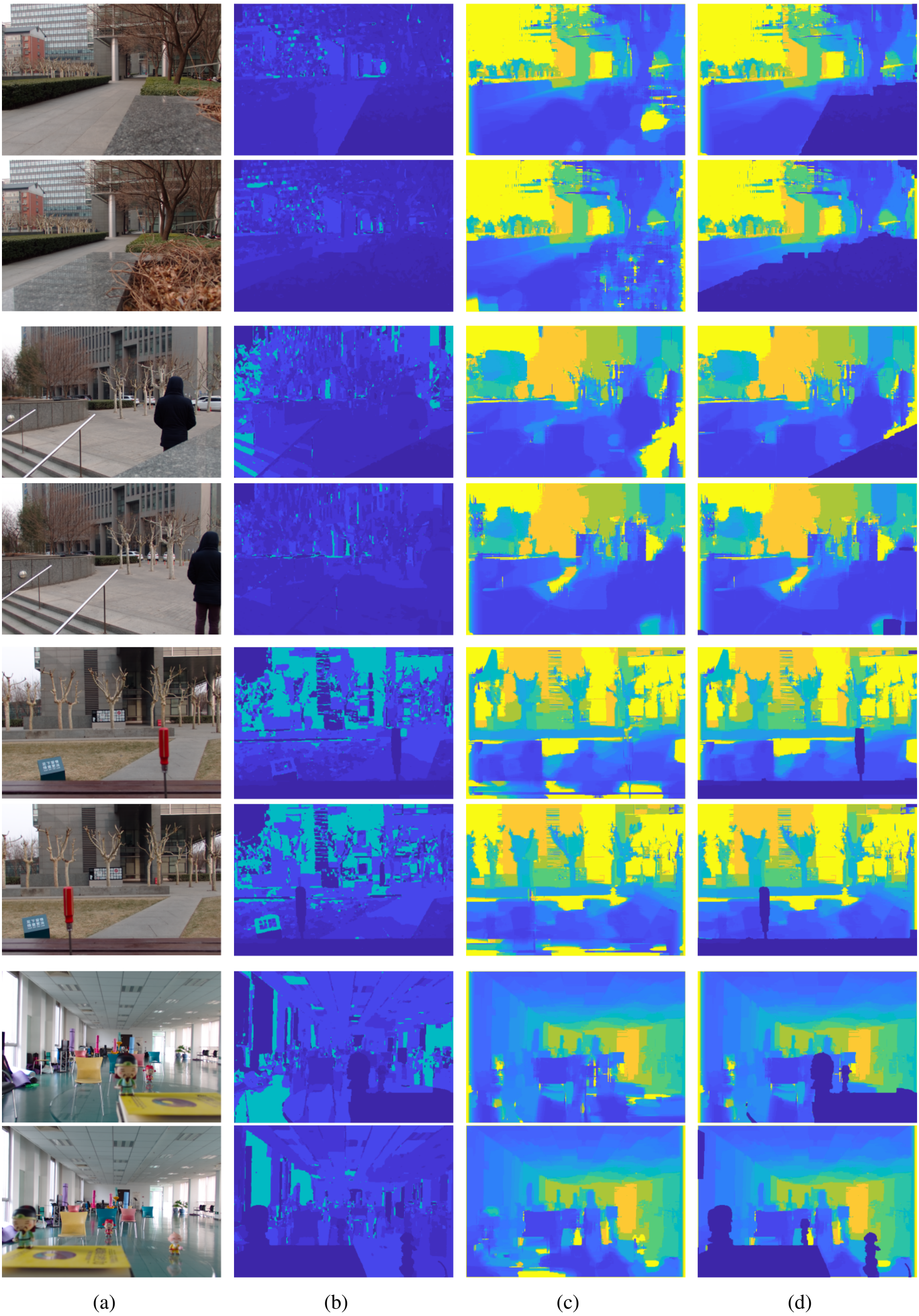

Figure 6: Outputs of different stages. Every two rows are pairs of stereo light field images. Pixels of close scene points are darker than far ones. (a) Stereo images, (b) Light field depth map[14], (c) Stereo matching[9], (d) Fusion map.

of inside structure of the camera. So we use curve fitting to estimate the transform from light field depth to real depth. Besides, we define a depth confidence of light field and model the fusion process in Markov random field. The final map show that depth of close objects is well measured.

In the future, we might study how to integrate fusion procedure into light field depth calculation and develop a more intrinsic method. The quality of depth map can also be improved further.

Acknowledgments

This work is supported by National Key R&D Program of China (2016YFB0801203), National Natural Science Foundation of China (61702479, 61771458, U1703261, 61571424), and the Science and Technology Service Network Initiative of the Chinese Academy of Sciences (KFJ-STS-ZDTP-018).

References

[1] M. Zeshan Alam and Bahadir K. Gunturk. Hybrid stereo imaging including a light field and a regular camera. *Signal Processing and Communication Application Conference*, pages 1293–1296, 2016.

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (11):1222–1239, 2002.

[3] Vineet Gandhi, Jan ÄŇech, and Radu Horaud. High-resolution depth maps based on tof-stereo fusion. *IEEE International Conference on Robotics and Automation*, pages 4742–4749, 2012.

[4] H HirschmÃijller and D Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582, 2009.

[5] Heiko HirschmÃijller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2007.

[6] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. *IEEE International Conference on Computer Vision*, 2017.

[7] Yann Lecun. Stereo matching by training a convolutional neural network to compare image patches. *The Journal of Machine Learning Research*, pages 2287–2318, 2016.

[8] Wenjie Luo, Alexander G. Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

[9] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang. On building an accurate stereo matching system on graphics hardware. *IEEE International Conference on Computer Vision Workshops*, pages 467–474, 2012.

[10] Ren Ng, Marc Levoy, Mathieu Bredif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenopic camera. *Stanford University Computer Science Tech Report*, 2005.

[11] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47 (1-3):7–42, 2002.

[12] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6640–6649, 2017.

[13] Michael Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. *IEEE International Conference on Computer Vision*, pages 673–680, 2013.

[14] Ting Chun Wang, Alexei A. Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. *IEEE International Conference on Computer Vision*, pages 3487–3495, 2016.

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

[16] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu. Light field reconstruction using deep convolutional network on epi. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1638–1646, 2017.

[17] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(7):1073–1079, 2009.

[18] 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, 2017.

[19] M. Zhao, G. Wu, Y. Li, X. Hao, L. Fang, and Y. Liu. Cross-scale reference-based light field super-resolution. *IEEE Transactions on Computational Imaging*, 2018.

[20] Jiejie Zhu, Liang Wang, Ruigang Yang, and James Davis. Fusion of time-of-flight depth and stereo for high accuracy depth maps. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.