Deep Learning Based Depth Estimation and Reconstruction of Light Field Images

Jae-Seong Yun* and Jae-Young Sim*†
* Department of Electrical Engineering, UNIST, Ulsan, Korea
† Graduate School of Artificial Intelligence, UNIST, Ulsan, Korea
E-mails: {jsyun,jysim}@unist.ac.kr

***Abstract*—Light field imaging is one of the most promising methods to capture realistic 3D scenes. In this paper, we propose a deep learning network composed of two sub-networks performing depth estimation and light field image reconstruction, respectively. We simultaneously train the two sub-networks by employing a loss function combining the reconstruction loss of the reconstruction network and the estimation loss of the depth estimation network. Experimental results demonstrate that the proposed method accurately estimates the disparity maps of light field images and also faithfully reconstructs light field images.**

I. Introduction

With increasing demand for augmented and virtual realities, numerous methods for 3D space capturing have been proposed. One of the most promising methods is light field imaging. To capture the light field in 3D space, camera array systems and microlens array camera systems were proposed. The camera array systems spatially distribute multiple cameras in grid structures, and the microlens array camera systems employ microlens array in front of the sensor. The light field images captured by the camera array systems are easily converted to those captured by the microlens array camera systems and vise versa.

One of the applications of light field images is the depth estimation using multiple images from different viewpoints, which uses epipolar plane images (EPIs) [1], [2], [3], angular patches [4], [5], [6], [7] and deep learning [8], [9], [10], [11], [12]. The horizontal EPIs are generated by collecting the horizontal lines from the images by fixing the y image coordinate at a fixed view coordinate of camera. The vertical EPIs are generated in a similar manner. As shown in Fig. 1, each line in EPIs is obtained from the corresponding pixels at multiple viewpoints. Since the slope of line depends on the depth of pixel, the depth map can be estimated from EPIs. Angular patches are generated by collecting all corresponding pixels from images at all viewpoints with a given candidate disparity value. If the candidate disparity value is correct, the intensity of all pixels in an angular patch is identical, and this characteristics of angular patches can be used to estimate the depth map.

In this paper, we propose an U-net[13] based network which estimates the depth map and then reconstructs light field images from the estimated depth map and the center view image. We first generate four stacks of images generated by collecting horizontal, vertical and two diagonal view directions of light field images. Using the generated stacks of images, the network estimates the depth map. Then, the four stacks of images are reconstructed using the center view image and the estimated depth map. The network is trained by minimizing the estimation error of the depth map and the reconstruction error of the image stacks. We evaluate the performance of the proposed method using 4D Light Field Benchmark dataset[14]. It provides 16 training and 12 testing light field images captured by a camera array system with 9×9 grid. The main contribution of this paper is the reconstruction of light field images at different viewpoints from the center view image and the estimated depth map.

Fig. 1: An example of light field image with horizontal and vertical EPIs.

II. Related Work

We present the related work of depth estimation for light field images.

A. Epipolar Plane Image Based Methods

Wanner and Goldluecke[1] estimated the depth map from light field images by measuring the local direction of line in EPIs. Kim *et al.*[2] proposed the confidence of the initially estimated depth value in EPIs and refined the initially estimated depth values to get the final depth values. If a certain pixel has a low edge confidence, the depth value is re-estimated at a

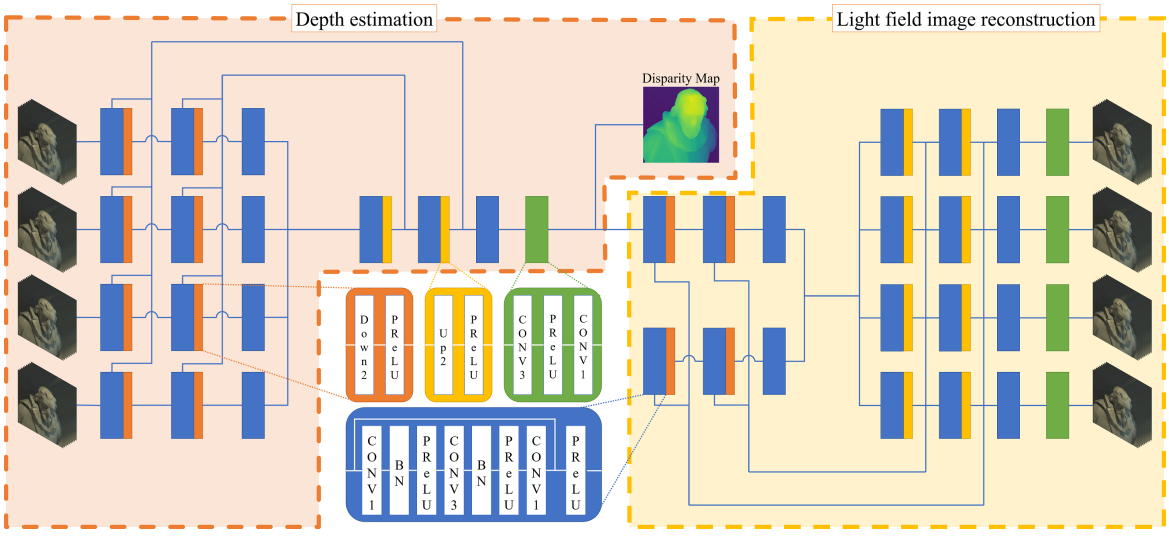

Fig. 2: The structure of the proposed network.

coarser image resolution. Tao *et al.*[3] computed a depth value which has optimal defocus and correspondence response when shearing EPIs with the candidate depth value.

B. Angular Patch Based Methods

Chen *et al.*[4] computed an optimal depth value for each pixel which minimizes the consistency of angular patches between center view image and the others. Wang *et al.*[5] detected the occlusion boundaries using the characteristic that the edges at occluded regions have the same orientation in images and angular patches, then they regularized an energy function estimating the depth map using the occlusion boundaries. Jeon *et al.*[6] proposed the correspondence matching based methods at sub-pixel accuracy by shifting a pixel in the Fourier domain. Williem *et al.*[7] proposed an angular entropy metric which measures the entropy in the angular patch, and estimated the depth map by minimizing the angular entropy metric whether there is occlusion or not.

C. Deep Learning Based Methods

Heber and Pock[8] extracted the patches centered at a query pixel from the horizontal and vertical EPIs, and used the convolutional neural network to estimate the depth value of the query pixel. Heber *et al.*[9] stacked images at different view points along the horizontal and vertical directions to create the EPI volumes, and estimated the depth maps using the convolutional neural network. Shin *et al.*[10] generated the four stacks of images by collecting images along the horizontal, vertical and two diagonal directions of view points, respectively, and estimated the depth map using multi-stream networks which take the four image stacks as inputs. Alperovich *et al.*[11] proposed Encoder-Decoder networks to estimate the depth map and to separate light field images into specular and diffusion layers. Zhou *et al.*[12] proposed a network which uses the stack of 115 differently focused center view images to estimate the depth map.

III. Proposed Methods

Numerous methods based on deep learning trained networks with additional information such as EPIs[8], [9] and focal stack[12] to improve the performance of depth estimation. However, these methods only consider the difference between the estimated depth map and the ground truth depth map, and does not consider the reliability of the estimated depth map. In this paper, we measure not only the difference between the estimated depth map and the ground truth but the reliability of the estimated depth map using the reconstructed light field images which are obtained from the estimated depth map and the center view image.

A. Network Design

The light field image is represented by using a 4D plenoptic function $I(x, y, s, t)$, where (x, y) are the image coordinates and (s, t) are the view coordinates of camera, respectively. Using the 4D plenoptic function, the relationships between the center view image $I(x, y, 0, 0)$ and the images at other viewpoints are given by

$$I(x, y, 0, 0) = I(x + D(x, y) \cdot s, y + D(x, y) \cdot t, s, t), \quad (1)$$

where $D(x, y)$ is the disparity of the pixel at (x, y) in $I(x, y, 0, 0)$. When the disparity $D(x, y)$ and the center view image $I(x, y, 0, 0)$ are given, non-center view images are reconstructed based on the Eq. (1). We reconstruct the light field images with the estimated disparity map and the center view image, and estimates the disparity map of a given light field image more reliably using the reconstructed light field images.

Fig. 2 shows the structure of the proposed network composed of the estimation and reconstruction networks. Similar to [10], we first generate the four image stacks of light field images, S_{0°, S_{45°, S_{90° and S_{135°, by collecting the images along the horizontal, vertical, left and right diagonal directions,

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

Fig. 3: The results of the light field image reconstruction. Top and bottom rows show the ground truth and the reconstructed light field images, respectively, with the horizontal and vertical EPIs. (a) Cotton, (b) Boxes, (c) Sideboard, and (d) Dino.

respectively. Each image stack is represented as 3D tensors where the shape of tensor is (X, Y, S) where (X, Y) is the resolution of image and S is the number of views. The four image stacks are fed to the estimation network which has the U-net[13] structure to preserve the information of high resolution images. The estimation network extracts the features using the residual blocks with bottle neck design. Since the image stacks are represented as 3D tensors, the 3D convolution layers are needed for feature extraction, however we use 2D convolution layers to extract features to reduce high computational complexity and GPU memory requirement. To apply the 2D convolution layers, we change the images into gray scale and consider the number of views in the image stacks as the dimension of features. From the extracted features, we apply a convolution layer with the kernel size 1 and the linear activation function to get the final disparity map. The reconstruction network has a similar structure with the estimation network while it takes the estimated disparity map and the center view image as input to reconstruct the four image stacks. To train both networks simultaneously, we combine the loss function of the two networks as

$$L = l(\hat{D}, D) + \lambda \sum_{d=\{0°,45°,90°,135°\}} l(\hat{S}_d, S_d), \qquad (2)$$

where $\hat{D}$ and D denote the estimated and ground truth disparity maps, respectively, and $\hat{S}_d$ and S_d denote the estimated and ground truth image stacks, respectively. The $l(\cdot, \cdot)$ is the mean absolute error (MAE) and λ is a weighting parameter.

TABLE I: The quantitative performance comparison in terms of MSE (multiplied by 100) and runtime in seconds.

Algorithm	MSE		Runtime	
	Median	Average	Median	Average
[1]	5.723	8.240	8.789	8.406
[5]	2.803	6.690	10614.535	10508.469
[6]	7.963	9.128	994.311	1009.756
[7]	2.667	3.730	822.272	832.081
[10]	1.280	2.521	2.032	2.041
[12]	1.913	5.242	85.045	88.194
Proposed	1.344	2.608	0.344	0.355

B. Details of Learning

To train the proposed network, we make $10,000$ mini-batches for one epoch where the batch size is 16 where each batch consists of the patches with the size of 32×32. The patches are extracted from random positions of randomly selected light field images among all the training samples. We apply the batch normalization[15] to all the convolution layers and use the PReLU[16] as an activation function for the convolution layers. We use the Adam optimizer[17] and set the learning rate to 10^{-4}. The weight parameter λ in Eq. (2) is set to 1. We implemented the proposed network in TensorFlow[18] which was trained for several days on a NVIDIA Titan RTX.

IV. Experimental Results

We evaluated the performance of the proposed method using 4D Light Field Benchmark dataset[14]. The dataset consists of 28 light field images and each light field image has 9×9 viewpoints. We trained the proposed network using 16 light field images. To measure the quantitative performance of the

depth estimation, we measured the MSE (Mean Squared Error) multiplied by 100 for all testing light field images. To show the qualitative results of the light field reconstruction, we used 4 light field images as shown in Fig. 3.

Fig. 3 shows the reconstructed results of the light field images. Note that we put gray-scale images as input to the network, and the reconstructed light field images are also gray-scale. As shown in EPIs, the light field image is not accurately reconstructed when there are occlusions due to the lack of information of the center view image. However, the overall structures of the objects are successfully reconstructed.

We also compared the depth estimation performance of the proposed method compared with six existing methods: [1], [5], [6], [7], [10], [12]. Table I shows the MSE multiplied by 100 and the runtime in seconds. In terms of MSE, the proposed method outperforms the five methods [1], [5], [6], [7], [12], and shows a comparable performance to the state-of-the-art method [10]. Note that the runtime of the algorithms were reported by the authors and measured under different computing environments, but we see that the proposed method outperforms yields the shortest runtime among the compared methods. Fig. 4 visualizes the estimated disparity maps and MSE maps. As shown in the figure, the proposed method successfully estimates the disparity maps, especially on the homogeneous regions, e.g., the walls behind the objects.

We performed the experiments of light field image reconstruction and depth estimation with four image stacks, but the proposed method can be applied with one or two image stacks as well, e.g. horizontal or vertical image stacks. When there are fewer numbers of image stacks, we expect that the performance decreases as [10] reported.

V. Conclusions

In this paper, we proposed a deep learning network which performs both of depth estimation and light field image reconstruction. To estimate the disparity maps more accurately, we combined the loss functions of the estimation network and the reconstruction network which are trained simultaneously. Experimental results showed that the proposed method not only provided comparable performance of depth estimation to the state-of-the-art algorithms but also successfully reconstructed the light field images from the estimated disparity maps and the center view images.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea within the Ministry of Science and ICT (MSIT) under Grant 2020R1A2B5B01002725, and in part by the 2020 Research Fund (1.200033.01) of UNIST (Ulsan National Institute of Science and Technology).

References

[1] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4d light fields," in *Proc. IEEE CVPR*, June 2012, pp. 41–48.

[2] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, "Scene reconstruction from high spatio-angular resolution light fields." *ACM Trans. Graph.*, vol. 32, no. 4, pp. 73–1, 2013.

[3] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *Proc. IEEE ICCV*, Dec 2013, pp. 673–680.

[4] C. Chen, H. Lin, Z. Yu, S. B. Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proc. IEEE CVPR*, June 2014, pp. 1518–1525.

[5] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE ICCV*, 2015, pp. 3487–3495.

[6] 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 *Proc. IEEE CVPR*, June 2015, pp. 1547–1555.

[7] Williem, I. K. Park, and K. M. Lee, "Robust light field depth estimation using occlusion-noise aware data costs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2484–2497, Oct 2018.

[8] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE CVPR*, June 2016.

[9] S. Heber, W. Yu, and T. Pock, "Neural epi-volume networks for shape from light field," in *Proc. IEEE ICCV*, Oct 2017.

[10] C. Shin, H.-G. Jeon, Y. Yoon, I. So Kweon, and S. Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE CVPR*, June 2018.

[11] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldluecke, "Light field intrinsics with a deep encoder-decoder network," in *Proc. IEEE CVPR*, June 2018.

[12] W. Zhou, E. Zhou, Y. Yan, L. Lin, and A. Lumsdaine, "Learning depth cues from focal stack for light field depth estimation," in *Proc. IEEE ICIP*, Sep. 2019, pp. 1074–1078.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Springer MICCAI*, 2015, pp. 234–241.

[14] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Proc. ACCV*, 2016, pp. 19–34.

[15] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, p. 448–456.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE ICCV*, 2015, pp. 1026–1034.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

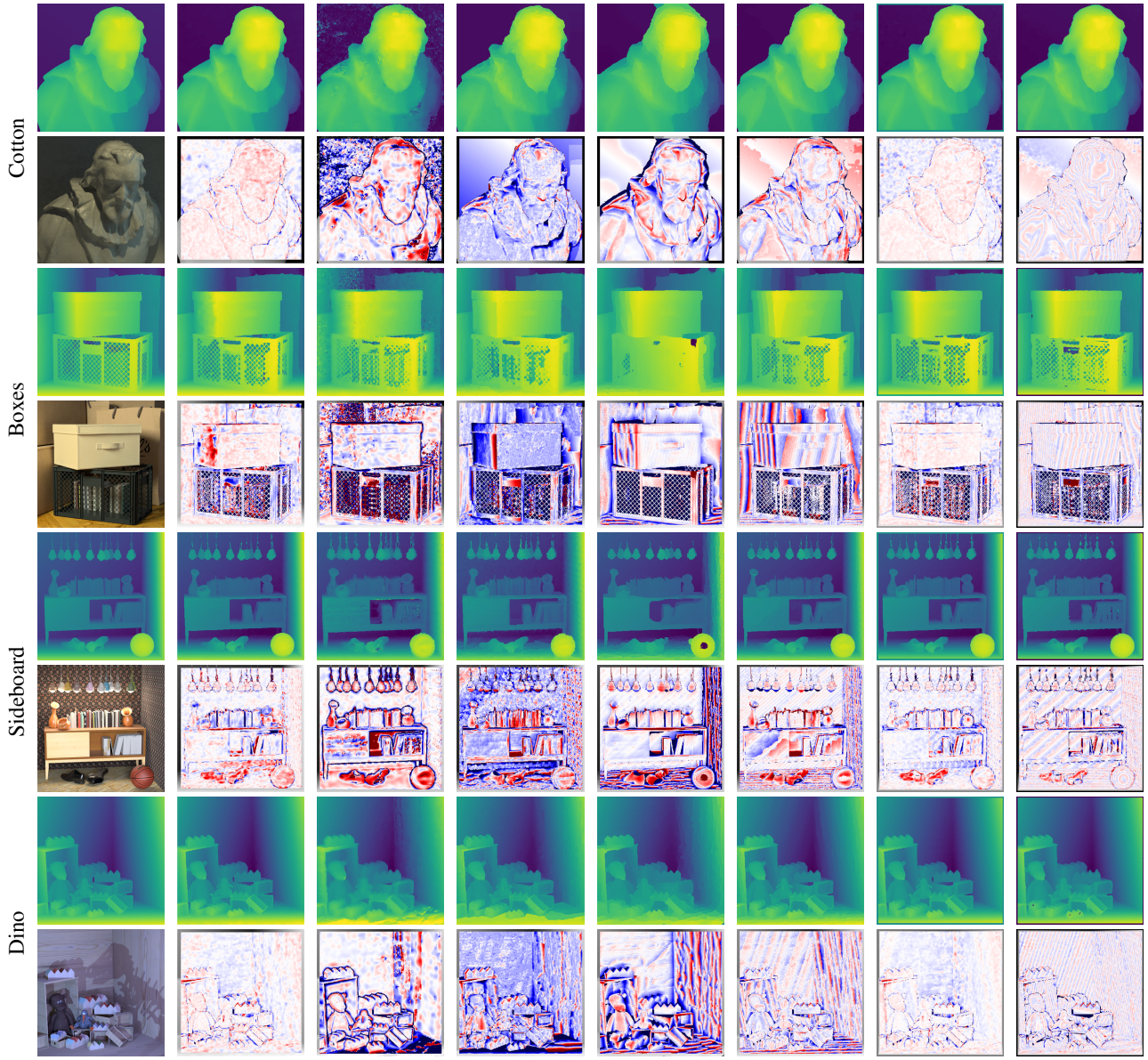

Fig. 4: The comparative results of depth estimation. (a) The ground truth disparity maps (top) and the center view images (bottom). (b-h) The estimated disparity maps (top) obtained by using the compared methods, and the corresponding MSE maps (bottom) where the white, red, and blue represent the estimated disparity values are correct, too far and too close, respectively.