Depth Estimation from Light Field Images via Convolutional Residual Network

Ji-Hun Mun, Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST),
123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, Republic of Korea
E-mail: {jhm, hoyo}@gist.ac.kr

***Abstract*—Estimating the depth map from multi-view images or light field images is an essential component in 3D geometry analysis. Conventionally, the depth map is estimated from integrated local and global information of stereoscopic images. However, the estimated depth map is not accurate due to the depth discontinuity and homogeneity. To solve this problem, we propose a light field depth estimation method based on the convolutional residual network. The discontinuity problem in the depth map is handled by computing depth cost maps in the residual network. In addition, we consider a phase shifted light field image in the loss function to acquire a robust depth map in the homogeneous region of the light field images. Experimental results demonstrate that our network outperforms other neural network architectures in terms of the depth map accuracy.**

Keywords—Depth estimation, Residual network, Light field

I. INTRODUCTION

Accurate depth estimation is an important step to understand the 3D geometric relationship in many applications. Especially, depth estimation from the light field images is a critical issue in computer vision. Finding corresponding points between neighboring images is an essential operation in depth estimation. Since light field images consist of a set of horizontal or vertical sub-aperture images, we can estimate the depth map from light field images in several different ways.

In computer vision applications, the accurate depth map is used for scene recognition [1], 3D modeling [2], and robotics [3]. There are various depth estimation algorithms for stereo images. In order to estimate an precise depth map from the stereo images, we should consider geometric and color consistency conditions. For a proper stereo matching operation, we need image rectification and color correction.

In order to extract the image characteristics, we can use geometrical image analysis for precise depth map estimation [4]. Currently, depth estimation methods based on deep learning show good performance in image classification [5], image denoising, visual simultaneous localization and mapping (SLAM), and depth estimation [6].

The main goal of this paper is to generate the depth map from light field images using a neural network. Estimating the depth map from dense light field images is a highly ill-posed problem. To avoid this problem, we need preprocessing operations, such as labeling, view point shifting, and scaling, which are used in the network training step. In order to minimize the number of training parameters, we adopt the architecture of the residual bottleneck network, as shown in Fig. 1.

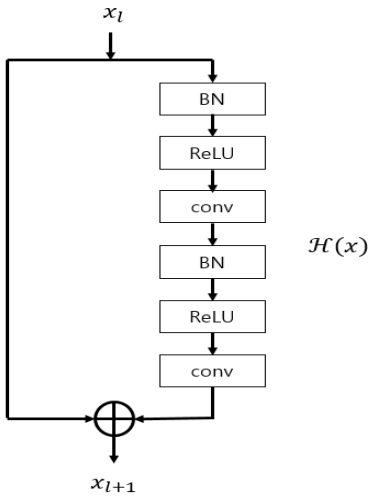

Fig. 1. Architecture of residual bottleneck network

The residual network updates the parameters not only for stacked layers $\mathcal{H}(x)$ but also for directly connected input data x_l. Among many different types of residual networks, we adopt the full pre-activated residual network for depth estimation. In the original residual network, the activation function is located in both inside layers and an additional part. However, the original architecture shows worse results in terms of the error rate than the full pre-activated architecture [7].

Most depth estimation methods, including deep learning-based methods, suffer from inaccuracy in depth discontinuity regions. Training parameters in the learning process are heavily affected by the occlusion of objects. In order to handle this problem, we can use several sub-aperture images. Since the microlens array provides horizontal sub-aperture images, our depth estimation network takes advantage of them during network training.

In this paper, we propose a new depth estimation method from light field images. The depth discontinuity problem is handled by selecting an optimal depth cost value from a set of sub-aperture depth maps. We can also obtain viewpoint shifted images by applying the phase shift operation. In order to estimate the depth map in homogeneous regions, we can use the phase shifted images.

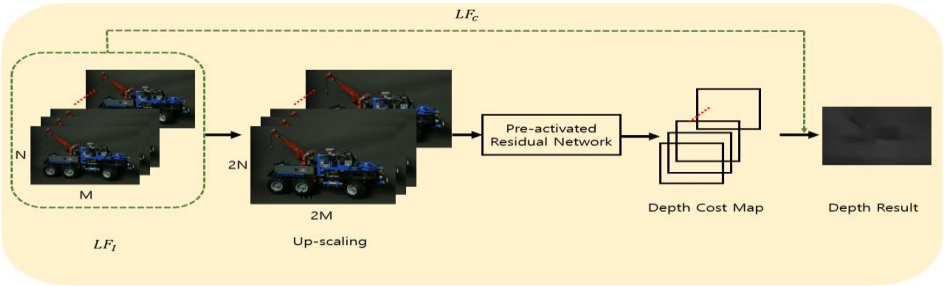

Fig. 2. The pipeline of our depth estimation model

II. Light Field Depth Estimation Network

In this section, we explain details of our depth estimation method. Fig. 2 represents the overall architecture of our method. In order to train our model, we use the light field datasets [8]. Since the datasets consist of color and depth images, we can train our network using both of them. The discontinuous regions in the estimated depth map are handled via light field sub-aperture images. In order to obtain accurate depth values near the discontinuous regions, we find an optimal cost value among depth cost maps.

A. Network structure

The architecture of our convolutional residual network is shown in Fig. 3. Our network is based on the ResNet 34-layer architecture [9]. Since a deeper ResNet layer, such as 100-layer, does not show any critical difference, compared to a shallow layer in terms of the error rate, we built our architecture on the 34-layer. In order to prevent the small resolution problem in the neural network, we first up-scaled the input images. Throughout large scaled input images, our network can find more meaningful characteristics than for the case of small scale images.

We partially added a convolution layer at the end of the each block $B_i (i \in \{1, 2, 3, 4\})$ to generate a high dimensionality of the depth map. The original ResNet 34-layer considers at most 512 convolution layer. However, from B_1 to B_4, we newly added a convolution block at the end of the original layer structure. That blocks have the same depth level with the initial layer of continuous convolution block. The increased depth level preserves the spatial information in the image and it help generate an precise depth cost maps as indicated in Fig. 2.

The kernel size that is used in the original convolution layer is 3 × 3 and the size of the last layer kernel is 1 × 1. ResNet 18-layer and 34-layer only use 3 × 3 kernel size, but the last part of each block in our network kernel size is 1 × 1 for every convolution layer. Our model considers a bottleneck network as represented in Fig. 1. The output of the basic model can be defined as:

$$x_{l+1} = \mathcal{H}\big(x_l \cdot W(m)\big) + x_l \tag{1}$$

where $W(m)$ indicates the weight of model which obtained by training, x_l is input data and $\mathcal{H}$ is the network of stacked layer. Among the light field sub-aperture images, the center sub-aperture image LF_c is used to create a final depth map in our network.

B. Loss function

While training the network with many datasets, estimated depth values are compared with ground truth depth values. For the accurate depth value selection, network based depth estimation method considers the cost value. In order to get the final depth map, our model effort to minimize the color dissimilarity between neighbor light field sub-aperture image LF_i and center sub-aperture image LF_c as defined in (2). Where the k indicates the range of the neighbor sub-aperture images.

$$C_{photometric} = \sum_{k=1}^{N} \|LF_c - LF_k\|^2 \tag{2}$$

During the photometric consistency is computed between the center sub-aperture image and neighbor sub-aperture images, we consider phase shifted center sub-aperture image. As the light field image usually has very narrow baseline distance, we can shift the pixel coordinate at the frequency domain instead of the spatial domain as indicated in (3).

$$\mathcal{F}(LF(x, y + \Delta)) = \mathcal{F}\{LF(x, y)\} e^{2\pi k \Delta} \tag{3}$$

where $\mathcal{F}$ indicates the 2D Fourier transform function, and Δ represents the shifting ratio of the center sub-aperture image in the frequency domain. Shifted sub-aperture images can be recovered in the spatial domain via the inverse Fourier transform.

The recovered image has different 2D coordinate value compare to the original center sub-aperture image. In order to compute the $L2$ regularization value, we use the shifted image which obtained by the shifting ratio Δ as defined in (4). This term help minimize the noise artifact in the homogeneous regions. In (4), LF_Δ represents the phase shifted center sub-aperture image with a ratio of Δ. Due to the shifted regions have different coordinate values with the center sub-aperture image, those are very helpful for exact cost value computation.

$$C_{smooth} = \|LF_c - LF_\Delta\|^2 \tag{4}$$

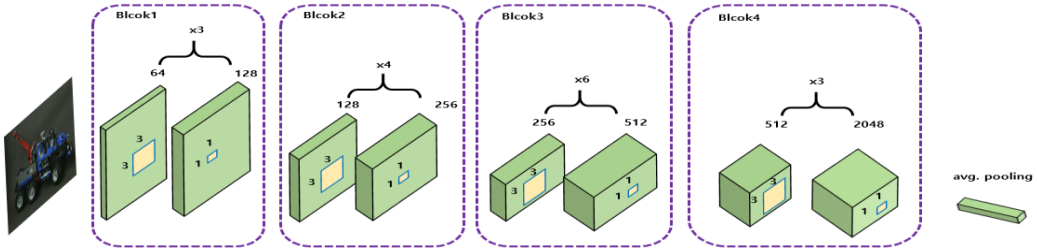

Fig. 3. The Pipeline of our depth estimation model

In order to handle the discontinuous and homogeneous regions efficiently, we consider photometric and smooth terms simultaneously. Thus, we made a network to minimize the final loss function which is defined in (5). The weighting factor α controls the equilibrium of two terms.

$$C = \sum_{k=1}^{N} [\alpha \cdot C_{photometric}^{k} + (1-\alpha) \cdot C_{smooth}] \qquad (5)$$

III. Experiment Results

We train our model through the light field datasets those are captured by Lytro Illum B01. The training datasets are composed of texture and depth map. After training the our model using light field datasets, we validate our network via light field and indoor test datasets.

In order to improve the performance of the training result, we apply online data augmentation with rotation of 30 degrees, horizontal and vertical flip, and random scaling. In addition, we use the fixed earning rate of 0.05 for every convolution layer.

We compare the generated depth map throughout the our model with widely used neural network models to show the superiority of our result. In order to evaluate the generated depth map quality, we use the following evaluation metrics:

- Average relative error (rel): $\frac{1}{n}\sum_i \frac{|\bar{x}_i - x_i|}{x_i}$
- root mean squared error (rms): $\sqrt{\frac{1}{n}\sum_i (\tilde{x}_i - x_i)^2}$
- root mean squared error in log space (rms(log)): $\sqrt{\frac{1}{n}\sum_i (log\tilde{x}_i - logx_i)^2}$
- average log_{10} error (log10): $\frac{1}{n}\sum_i |\log_{10} y_i - \log_{10} y_i^*|$
- accuracy with threshold (δ): % of $\tilde{x}_i$ s.t. $\max(\frac{\tilde{x}_i}{x_i}, \frac{x_i}{\tilde{x}_i}) = \delta$

$$\delta_{1,2,3} = 1.25, 1.25^2, 1.25^3$$

Since our model is proposed based on the ResNet 34-layer, we compare the performance of our model with different layer structure such as 18-layer and 50-layer. In addition, we compare the performance of proposed network with state-of-the-art depth estimation method. Due to our network is designed for light field depth estimation with the neighbor sub-aperture images, we verify the performance of our network using light field datasets and NYUD v2 datasets.

A. *Light field dataset*

For the fair comparison of the proposed method in the light field datasets, we only compare the our model with different ResNet layer models. Since the used light field datasets are captured in indoor condition, training weight values are appropriate to estimate a depth map of indoor scene and the results are represented in Fig. 4.

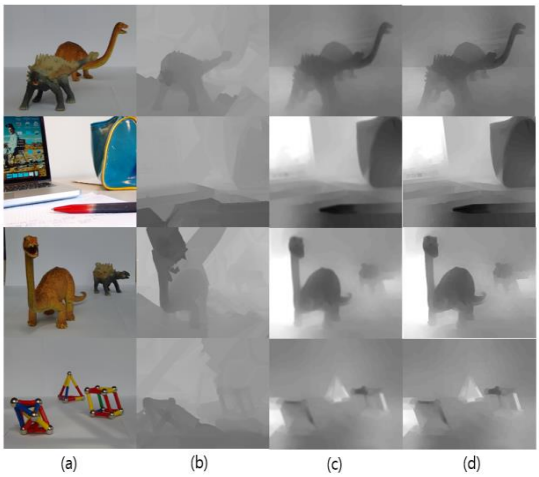

Fig. 4. Depth estimation results of light field scene, (a) input data, (b) ground truth, (c) ResNet 18-layer, (d) proposed network

Depending on the number of light field neighbor sub-aperture images *N* and phase shifting ratio Δ, the artifacts of the discontinuous region and the homogeneous region are changed. Even though we use many sub-aperture images for depth estimation, the depth map quality is not always guaranteed. The distance between center sub-aperture image and neighbor sub-aperture affect to the depth map quality.

In order to analyze the artifact of a number of sub-aperture in depth map quality, we evaluate the result of the depth map in term of bad pixel ratio (BPR) in Table 1. In the same manner, the effect of shifting ratio also considered. Thus, we analyze the relationship between phase shift ratio and number of sub-aperture images simultaneously.

From the BPR result of light field scenes, the proposed network shows better performance than the original ResNet 18-layer model. Most of the experiment results show accurate depth values when we apply the *N*=3 and Δ= 0.5 parameters.

TABLE I
LIGHT FIELD SCENES BPR COMPARISON RESULTS WITH FIXED PARAMETERS (N AND Δ)

Sequences	ResNet 18-Layers			Proposed		
	N=2 Δ=0.3	N=3 Δ=0.5	N=4 Δ=0.7	N=2 Δ=0.3	N=3 Δ=0.5	N=4 Δ=0.7
Diplodocus1	39.57%	**37.07%**	39.76%	32.12%	**31.41%**	33.23%
Desktop	28.18%	**25.17%**	29.12%	20.88%	**18.82%**	24.09%
Diplodocus2	42.88%	41.29%	**40.35%**	36.29%	**34.14%**	37.18%
Magnets1	38.01%	**33.22%**	39.27%	31.18%	**28.04%**	35.31%

B. NYUD v2 dataset

We have also compared the performance of the proposed model with the original shallow and deep ResNet network. The estimated depth maps are indicated in Fig. 5.

The ResNet 18-layer and 50-layer based depth map have a blurring issue in result depth map when it compared with the proposed network depth map result. As already proven that in [7], deep network layers shows better performance than shallow network layers. As explained, 18-layer result of Fig. 5 (c) shows more blurring artifact than 50-layer result of Fig. 5(d).

Since our model is designed on the ResNet 34-layer, it shows competitive performance than original ResNet 34-layer. However, the estimated depth still has unsharpened depth regions near the object boundary, especially in the detailed regions. To numerically analyze the depth map quality, we compute quality evaluation metrics in Table 2.

ResNet 18 and ResNet 50 indicate original ResNet architecture based depth quality evaluation results. As represented in Fig. 5 (c) and Fig. 5 (d) results, the original ResNet results have smudged edge depth values compare to the proposed results. For that reason, proposed depth map quality always outperform the other network-based depth map results.

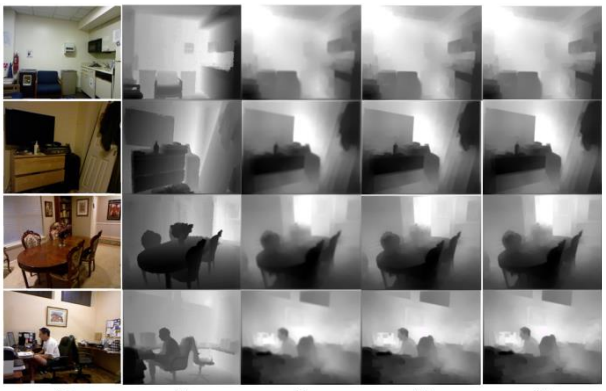

Fig. 5. Depth estimation from light field scene, (a) input data, (b) ground truth, (c) ResNet 18-layer, (d) ResNet 50-layer, (e) Proposed network

TABLE II
COMPARISON RESULTS ON NYUD V2 DATASET

Method	rel	rms	log10	δ_1	δ_2	δ_3
ResNet 18	0.251	0.750	0.313	0.688	0.802	0.836
ResNet 50	0.231	0.709	0.288	0.753	0.833	0.873
Eigen[9]	0.177	0.665	**0.242**	0.792	0.977	**1.070**
Proposed	**0.156**	**0.571**	0.255	**0.828**	**0.981**	1.013

However, our network is slightly weaker in *log10* and δ_3 than [10] results. Even though deeper ResNet is much stronger than the shallow network, our architecture has a better result than ResNet 50.

IV. CONCLUSION

In this paper, we propose the ResNet-based light field depth estimation network. In order to handle depth discontinuous regions, we use the residual network with the neighbor sub-aperture images. In addition, smooth term alleviate the homogeneous region problem in depth map by using phase shifted sub-aperture images. In the future work, we will improve the depth map quality via unsupervised network. That network will bring out more robust depth estimation results without considering the type of training datasets.

ACKNOWLEDGMENT

This work was supported by the 'Civil-Military Technology Cooperation Program' grant funded by the Korea government.

REFERENCES

[1] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Proc. Eur. Conf. Comput. Vis.*, pp. 746-760, Sept. 2012.

[2] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic Photo Pop-up," in *ACM SIGGRAPH*, vol. 24, no. 3, pp. 577-584, Aug. 2005.

[3] R. Hadsell, P. Sermanet, J. Ben, A. Erkan, M. Scoffier, K. Kavukcuglu, U. Muller, and Y. LeCun, "Learning Long-range Vision for Autonomous Off-road Driving, *Journal of Field Robotics*, vol. 26, no.2 pp.120-144, Feb. 2009.

[4] A. Gupta, A. A. Efros, and M. Hebert, "Blocks World Revisited: Image Understanding using Qualitative Geometry and Mechanics," *in Proc. Eur. Conf. Comput. Vis*., pp. 482-496, 2010.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Image Net Classification with Deep Convolutional Neural Networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, no. 1, pp. 1097-1105, Dec. 2012.

[6] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified Depth and Semantic Prediction from a Single Image," in *Proc. IEEE Conf. Comp. Vis. Patt. Recog.*, June 2015.

[7] K. He, X. Zhang, S. Ren, and Jian Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conf. Compt. Vis. Pattet. Recog.*, pp.770-778, Dec. 2016.

[8] M. Rerabek and T. Ebrahimi, "New Light Field Image Dataset," 8th International Workshop on Quality of Multimedia Experience, 2016.

[9] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2650-2658, Dec. 2015.