OPTICS CONTINUUM

Focal stack camera: depth estimation performance comparison and design exploration

ZHENGYU HUANG,* JEFFREY A. FESSLER, AND THEODORE B. NORRIS

[1] *Department of Electrical Engineering and Computer Science, University of Michigan-Ann Arbor, MI 48109, USA*
zyhuang@umich.edu

Abstract: The concept of focal stack cameras based on transparent sensors has been recently demonstrated, enabling depth sensing in a single exposure from a focal stack. This study investigates how the depth estimation accuracy is affected by numerous camera parameters, including aperture size, sensor resolution and number of sensor planes. We investigate the dependence of these parameters on the deep learning based depth estimation performance and make a comparison between the focal stack camera and the light field camera, thus providing guidelines for future focal stack camera design.

1. Introduction

Focal stack photography captures a set of images (a focal stack) of a scene with varying focus positions. These images contain depth-dependent defocus blur and encode 3D information about the scene. Combined with suitable algorithms, the focal stack can be used to track 3D objects [1], estimate depth maps [2–5] and reconstruct light fields [6–8].

In conventional focal stack photography, a focal stack is typically collected by multiple exposures with changing focus position. This approach, however, is only applicable to static scenes with no moving objects. Motion artifacts can be minimized by rapid sequential acquisition of the focal stack images; one method is to use liquid lens [9], which can change focal length quickly, though such lenses suffer from aberrations and hence degraded image quality.

Light field cameras, such as Raytrix and Lytro, provide an alternative way to obtain a focal stack: the camera captures a 4D light field by using a micro-lens array in front of the camera sensor. From the collected light field data, the focal stack can then be computed by digital refocusing using the add-shift algorithm [10]. Light field cameras use the micro-lens array inside the camera body to sort the incident light rays by their propagation directions and maps them to spatially separated pixels, thereby providing angular resolution. As a result, there is an inherent spatial-angular resolution trade-off in the captured light field. Denoting a 4D light field as $L(x, y, u, v)$, where x, y are the spatial coordinates and u, v are the angular coordinates, one sees that for a fixed resolution image sensor, a higher image spatial resolution then forces a lower angular resolution and vice versa. For example, the light field by Lytro Illum camera has a resolution of $376 \times 541 \times 14 \times 14$ and the low spatial resolution is due to the above spatial-angular resolution trade-off.

Recently, a novel focal stack camera based on highly transparent graphene photodetectors was introduced [1,7]; this optical system can capture a focal stack in a single exposure. Such a camera does not require objects to be static as in a conventional focal stack camera, and has no spatial-angular resolution trade-off for light fields reconstructed from focal stacks [7], as in the light field camera. This enables 3D sensing applications in *real time*. Despite several successful demonstrations of focal stack camera applications, the dependence of the focal stack

#472819
Journal © 2022

https://doi.org/10.1364/OPTCON.472819
Received 9 Aug 2022; accepted 12 Aug 2022; published 8 Sep 2022

camera design on its 3D sensing performance has not yet been explored. It is also unknown what the performance trade-offs might be when comparing the focal stack and light field camera approaches.

This paper addresses these questions via a set of numerical experiments. Specifically, we focus on depth estimation performance evaluation, using deep learning based methods. Using focal stacks that are either computed from publicly available light field datasets [5,11,12] or captured experimentally, we train convolutional neural network (CNN) models to estimate depth maps from the input focal stack and study the dependence of the camera parameters, including number of sensor planes, aperture size and sensor resolution, on the depth estimation accuracy. We further compare the system performance with the light field camera and show that focal stacks achieve comparable performance.

The paper is organized as follows: Section 2.1 and 2.2 describe the background and methods for focal stack and light field depth imaging. Section 2.3 describes the network structure we used for estimating the depth from the focal stack and from the light field. Section 2.4 describes the datasets we used for performance evaluation. Section 3. contains the experiment results and analysis.

2. Methods

2.1. Focal stack depth imaging

Figure 1(a) shows a focal stack camera imaging system [1,7,13]. A graphene-transistor based image sensor makes the sensors semitransparent (90%-95% transmission), while still maintaining high responsitivity. According to the thin lens equation:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}, \tag{1}$$

where d_o is the object distance, d_i is the image distance and f is the focal length of the lens, sensors having different distances d_i will thus focus at different depths. By stacking such transparent sensors along the optical axis, a focal stack can be captured with a single exposure. Each image in the focal stack contains depth-dependent defocus blur as illustrated in Fig. 1(b). Specifically, for a camera with an aperture of diameter A, the diameter of the circle of confusion c is given by:

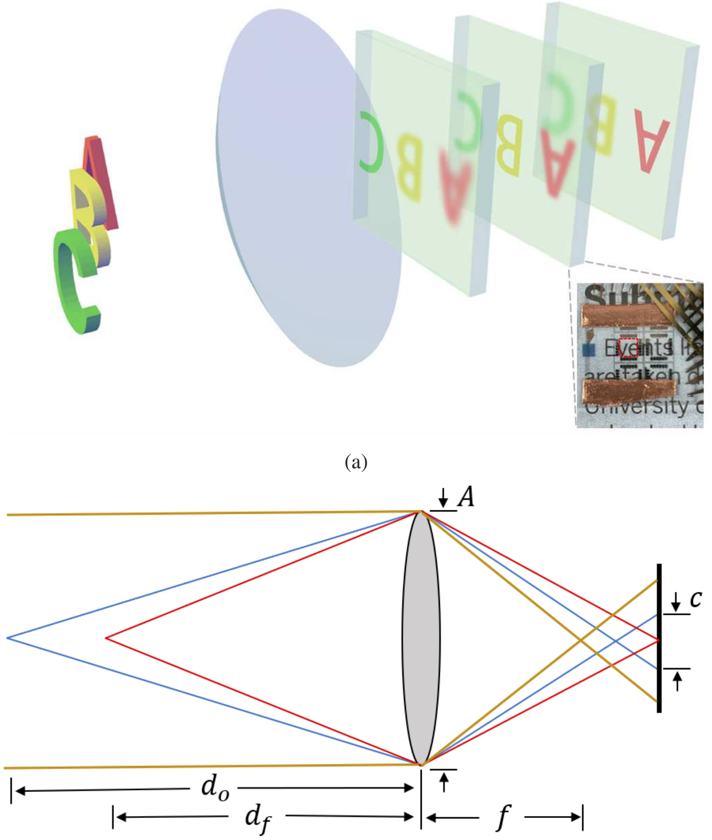
$$c = A\frac{\left|d_o - d_f\right|}{d_o}\frac{f}{d_f - f}, \tag{2}$$

where d_f is the distance from an in-focus object point to the lens (camera focusing distance) and d_o is the distance from an out-of-focal-plane object to the lens.

Several approaches have been developed to estimate depth maps from a focal stack. Nayar et al. [2] used a sum-modified-Laplacian to measure the focus sharpness and fit the focus sharpness by a gaussian distribution to obtain accurate depth. Moeller et al. [3] cast the depth estimation as a nonconvex optimization problem that includes a data fidelity term and a regularization term, which is solved by linearized alternating directions method of multipliers (ADMM) [14]. Sakurikar et al. [4] used a composite focus measure that is a weighted combination of standard focus measures to measure the focus sharpness and showed that it achieves better performance than those using a single individual focus measure. Hazirbas et al. [5] trained a deep neural network for depth estimation from focal stack.

2.2. Light field depth imaging

As one of our goals in this paper is to compare the depth estimation performance of the focal stack camera with the light field camera, this section provides some background on the light field and light field based depth sensing. The radiance of a light ray, at point (x, y, z) in space,



(a)

(b)

Fig. 1. Focal stack imaging system. (a) System schematic. The captured focal stack shows changing defocus blur that depends on the object distance. Inset: a transparent sensor array (active area highlighted in red) overlaid on top of text. (b) Illustration of circle of confusion.

propagating along direction with polar angle θ and azimuthal angle ϕ can be described by the 5D plenoptic function $P(x, y, z, \theta, \phi)$ [15]. In the case of free space propagation, since the radiance of a light ray is constant along its direction of propagation, a 4D light field L is sufficient to fully characterize the light distribution. A 4D light field can be parameterized using two-plane parameterization: the radiance of a light ray intercepting the first parameterization plane at (u, v) and at the second parallel parameterization plane at (x, y) is given by $L(x, y, u, v)$. Such a 4D light field can be thought of as a collection of 2D conventional images $I^{u_0,v_0}(x, y)$, called sub-aperture images, each with a different view point (u_0, v_0). Due to parallax, a point in 3D space maps to different spatial locations in different sub-aperture images. For example, a pixel at (x, y) in view (u, v), if unoccluded, corresponds to the pixel at $(x - D, y)$ in view $(u + 1, v)$, where D is the disparity of the 3D point. The disparity D of the point is directly related to its depth d_o as:

$$D = b \cdot f \cdot \left(\frac{1}{d_o} - \frac{1}{d_f} \right), \tag{3}$$

where b is the separation between sub-aperture images (baseline), f is the focal length, d_f is the camera focusing depth, d_o is the depth of the 3D object.

Identifying such pixel correspondence between views in the light field allows one to estimate the depth. Chen et al. [16] proposed a bilateral consistency metric to evaluate the surface camera light field and then apply a stereo matching algorithm to estimate the depth. Zhang et al. [17] designed a spinning parallelogram to estimate the slope of lines (directly related to disparity) in the epipolar plane image of the light field. Shin et al. [18] trained a neural network, EPINet, to process sub-aperture views along horizontal, vertical, and diagonal directions to regress a depth map. Tsai et al. [19] computed a 4D disparity cost volume and employed an attention mechanism to scale a feature map from each sub-aperture view by its importance and then estimate the depth.

2.3. Network structure

This section describes the neural networks used for estimating the depth from a focal stack and from light field images. Figure 2 shows the neural network structure we used for estimating depth from a focal stack. The input RGB focal stack contains n_F images that are concatenated along the color dimension for a total of $3n_F$ input channels. The network consists of 10 convolution layers with no spatial pooling or up-sampling operations, to preserve fine-details in the final output. Dilated convolutions [20] are used to ensure a large receptive field without significant computation cost. The output from the last convolution layer (after tanh nonlinearity) is further scaled and offset by dataset-dependent constant α and β, respectively to constrain the output to a plausible range.

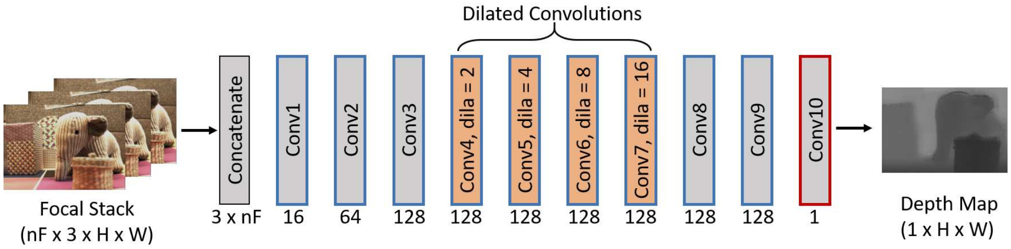


Fig. 2. Network structure for depth estimation from focal stack. All convolutions have filter size of 3 × 3, stride 1, and the output channel number for each layer is indicated beneath. Blue border around a layer indicates that Batch Normalization and leaky ReLU are applied to the output. Red border indicates tanh non-linearity is applied to the output. n_F is the number of images in the focal stack.

We use the EPI-Net [18] for estimating the depth from the light field image. The network has a four-branch structure, where each branch takes in sub-aperture images of the light field along a particular direction (horizontal, vertical, left-diagonal or right-diagonal). Features are extracted from each branch independently using 2D convolutions and then concatenated along the color dimension. Then additional convolutions are used to process the concatenated feature map to predict the final depth map. More details about the network structure can be found in [18].

2.4. Focal stack dataset

We generated focal stack data from three publicly available light field datasets: the HCI light field dataset [11], the DDFF dataset [21] and the CVIA dataset [12]. The HCI light field dataset contains 28 synthetic light fields of resolution $9 \times 9 \times 512 \times 512$, of which 16 light fields in the category ‘additional’ are used as the training data and the remaining 8 light fields are used as the testing data. We synthesized focal stacks using the add-shift algorithm [10], with images focusing at disparity planes evenly distributed in [-3,3]. The DDFF dataset contains 600 training and 120 testing realistic light fields of size $9 \times 9 \times 383 \times 552$ captured by a Lytro light field camera. 480 light fields from the original training data are used in our experiments for training, with the remaining 120 light fields in the original training data used for testing. We synthesized focal stacks, each containing n_F images focusing at disparity planes evenly distributed in [0.020, 0.282]. The CVIA dataset contains 40 light fields of resolution 15 × 15 × 434 × 625 in a distance range of 0.2 to 1.6m using a Lytro camera, of which 32 are used for training, and 8 for testing. We synthesized focal stack with images focusing at disparity planes evenly distributed in [-0.44, 0.17]. All datasets above contain ground truth depth maps for evaluation, either from its synthetic 3D models (HCI synthetic light field), or from depth sensors (DDFF dataset and CVIA dataset).

In addition to using focal stacks generated from the existing light field datasets, we also collected an additional focal stack dataset, which we named as DLSR dataset. Unlike all the above datasets, where the focal stacks are synthesized from the light fields, it consists of focal stacks captured directly using a DLSR camera (Nikon D7200, 35 mm lens) with focal stacking function provided in camera control software 'controlMyNikon'. As such, the focal stacks in the DLSR dataset resemble most closely the focal stack one would capture using the focal stack imaging system shown in Fig. 1(a). We set the step size of the focal stacking in the software to 200, and size of the focal stack n_F to 7, which covers a depth range of approximately 0.4 m to 1.3 m. We repeat the focal stack collection process for 4 aperture size settings (f/3.2, f/5, f/10, f/22). Each setting contains 40 focal stacks of resolution 854 × 1280 after resizing, of which 32 are used as the training data and the remaining 8 are used as the test data. We additionally form DLSR datasets with $n_F = 2$, from the $n_F = 7$ datasets, by using only the 2nd and 6th focus position images, which are used to study the dependence on number of sensor planes. Since the raw captured focal stack exhibits a focus breathing effect due to the change of magnification

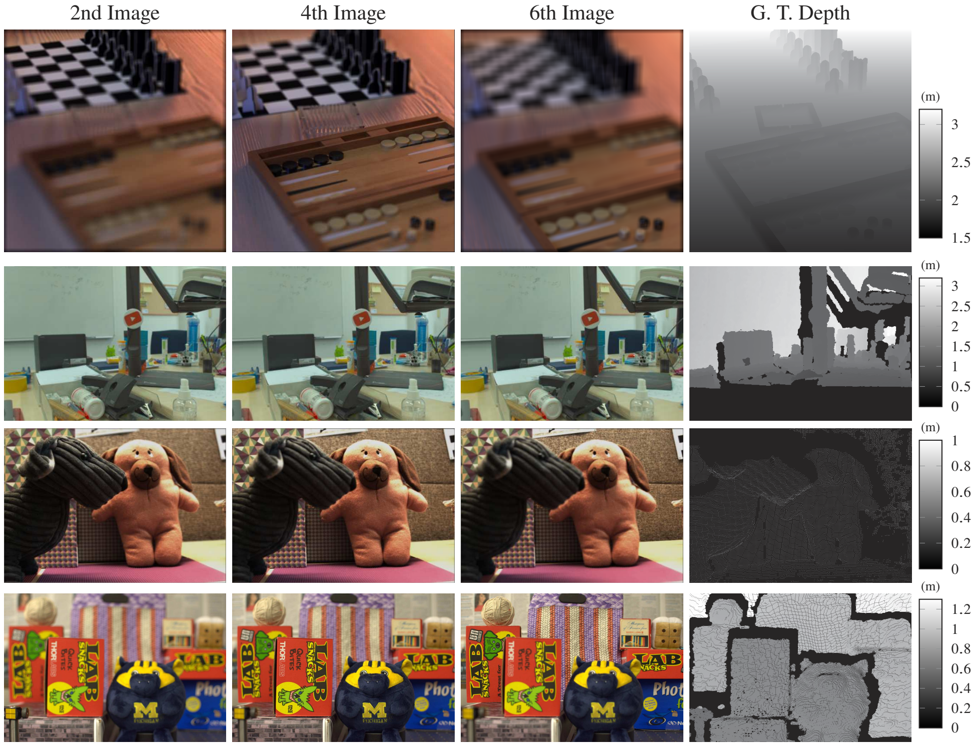


Fig. 3. Example focal stacks showing the 2nd, 4th and 6th images in the stack sequence. Last column shows the ground truth depth maps. Rows correspond to HCI dataset, DDFF dataset, CVIA dataset and DLSR dataset, respectively.

when the 35-mm lens focus is changed, we additionally perform a focal stack alignment process to compensate the magnification change and align the images in the focal stack. We also capture ground truth depth maps for each focal stack, using an Intel RealSense D415 Depth Camera, and register the depth onto the RGB images. More details on the focal stack collection, focal stack alignment and depth registration can be found in the supplementary material. Figure 3 shows example focal stack images of the datasets we use.

3. Experiments

We trained separate networks (Fig. 2) to estimate depth, using focal stack datasets with varying camera parameters (number of sensor planes in focal stack, sensor resolution, aperture size), to study their dependence on the depth estimation accuracy. Finally, we compared the depth performance from the focal stack and the light field. The details of the training setup and experiments are described next.

3.1. Training setup

All networks were trained in Pytorch with L1 loss using Adam optimizer [22] with learning rate 10^{-4}, batch size 4. The input focal stacks/light fields were randomly cropped in the spatial dimension to 125×125. Models were trained till convergence (80k epochs for HCI dataset, 5k epochs for DDFF dataset, 15k epochs for CVIA and DLSR datasets). Given the limited amount of data available for training neural networks, models could be overfitted. This is acceptable since the goal of this paper is to compare the model performances between different settings, and the comparison conclusion should not be sensitive to the degree of overfitting.

We used a Nvidia GTX 1080Ti for model training. Training one epoch of the EPINet and the focal stack based network both take 2s on HCI dataset, one minute on DDFF dataset, 4s on CVIA and DLSR dataset. The entire training takes 2 days on HCI dataset, 3 days on DDFF dataset, 1 day on CVIA and DLSR dataset.

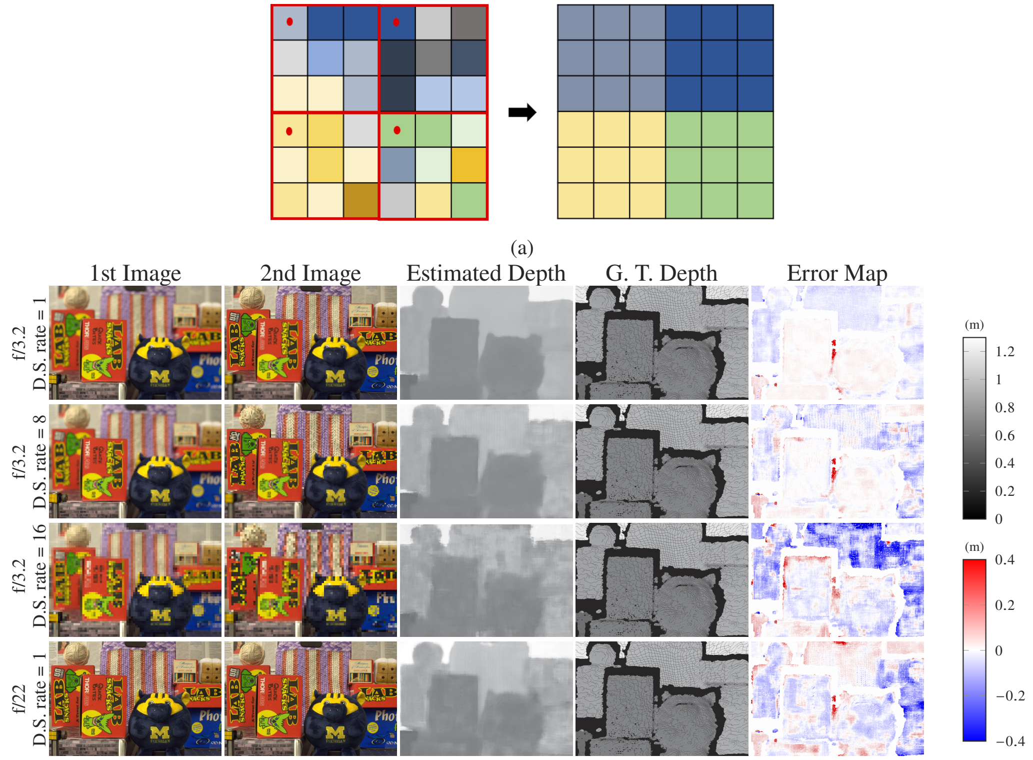
3.2. Sensor resolution dependence

Here we study how the sensor pixel resolution affects the depth estimation performance. Specifically, a down-sample rate of N means reducing the effective resolution of the images in the focal stack by setting the pixel values in every $N \times N$ block to the value of the top left pixel. Figure 4(a) illustrates the downsampling process that mimics the fact that the active sensing areas (individual pixels) of a low resolution sensor are not densely packed in the 2D plane. The first 3 rows of Fig. 4(b) show example focal stacks ($\mathrm{n_F} = 2$) with varying down-sample rates collected with f/3.2 aperture setting, along with the estimated depth maps, which indicate that higher resolution sensors lead to higher quality depth maps as one may intuitively expect. The left column of Fig. 5 shows how the RMSE of the depth estimates depend on the focal stack image resolution. Better resolution images lead to better performance on DDFF, CVIA and DLSR datasets. This trend can be understood as follows: degrading the resolution causes some defocus blur information to be lost (at the extreme of very low resolution, objects at all depths will be equally blurred). In addition, the $\mathrm{n_F} = 7$ result has lower RMSE than that from $\mathrm{n_F} = 2$, especially for a large down-sample rate, indicating that having more focal planes in the focal stack camera is helpful, as expected.

3.3. Aperture size dependence

We next study how the aperture size affects the depth estimation performance. According to Eq. (2), a larger aperture leads to a larger defocus blur, which could potentially affect the depth estimation performance. For focal stacks that are synthesized from a light field (HCI dataset, DDFF dataset, CVIA dataset), changing the aperture size can be realized by refocusing using

OPTICS CONTINUUM



(a)

(b)

Fig. 4. Example focal stacks with different camera parameters in DLSR dataset. (a) Schematic illustrating focal stack generation with down-sample rate = 3. (b) Focal stack examples (n_F = 2) captured with different down-sample rate and aperture setting. The depth estimated from the focal stack, the ground truth depth and the error map are also shown.

only the sub-aperture images that are within the desired aperture window from the light field. For our DLSR datasets, we acquired separate focal stacks with different aperture sizes for each scene. Comparing the 1st and 4th row of Fig. 4(b) shows the effect of reducing the aperture size. The images in the focal stack become sharper as the aperture is reduced and the estimated depth becomes noisier. The right column of Fig. 5 shows quantitatively that decreasing the aperture size increases the RMSE error. This trend can be understood because in the limit of very small aperture size, all images in the focal stack would be the same image with every depth in focus. Comparing the results of n_F = 2 and n_F = 7 with changing aperture size, having more focal planes slightly improves the accuracy in this case.

3.4. Focal stack and light field camera comparison

Here we compare the performance between depth from light field and depth from focal stack on the HCI, DDFF and CVIA datasets. EPINet [18] is used to estimate the depth from light fields. Light fields and focal stacks with the largest possible aperture sizs are used for each dataset. We used n_F = 7 for the focal stack data and used no down-sampling of the focal stack/light field images. Table 1 shows that the depth estimation from focal stack has a disparity RMSE error of 0.018 pixel on the DDFF dataset, which is 33% lower compared to that from the light field. On the CVIA dataset, the focal stack based method also performs better than the light field based method, with 17% lower RMSE. However, the light field based depth estimation performs better

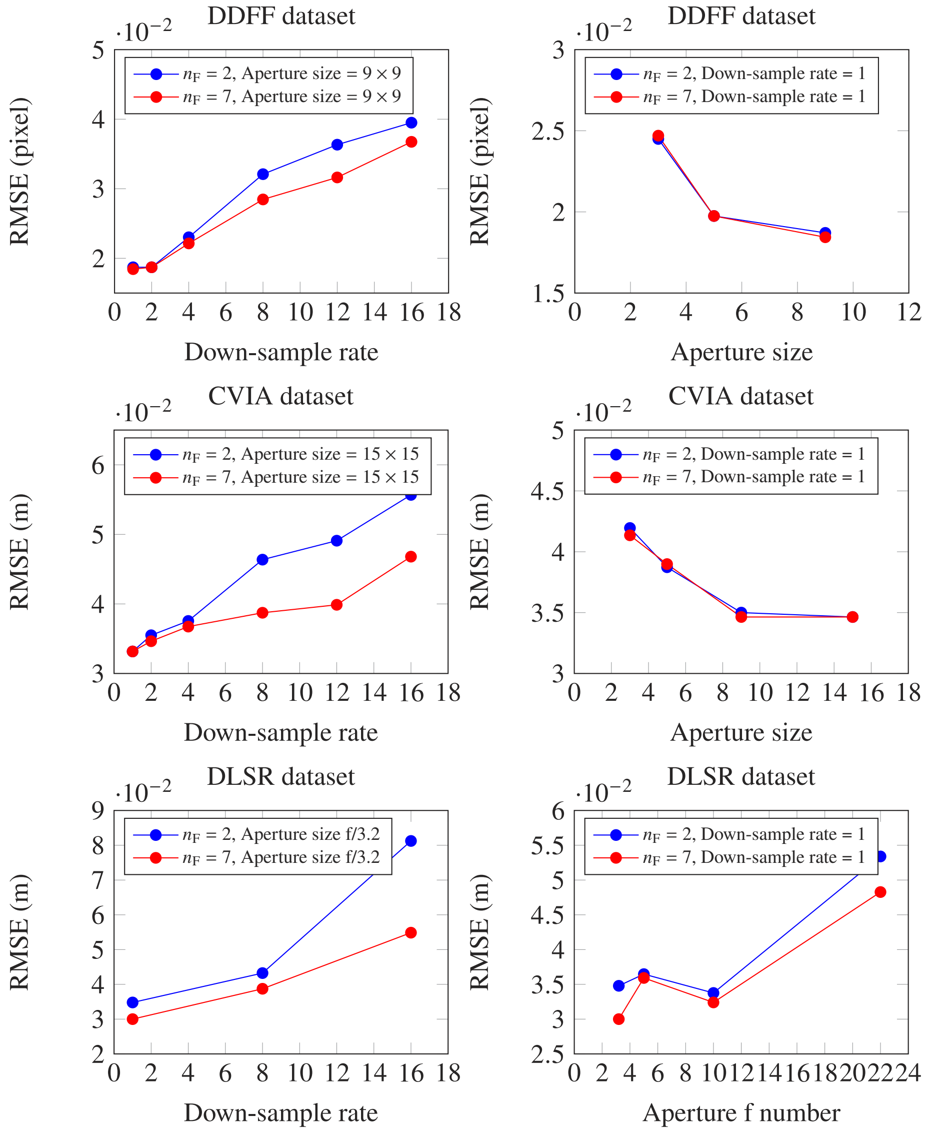


Fig. 5. RMSE of the depth estimated from focal stack images on DDFF dataset, CVIA dataset and DLSR dataset as a function of resolution down-sample rate (left column), aperture size (right column) and number of sensor planes n_F.

OPTICS CONTINUUM

on the HCI dataset, with a disparity RMSE of 0.17 pixel, as opposed to 0.36 pixel for focal stack based method.

Table 1. RMSE of depth map estimated from focal stack and light field. Focal stack of $n_F = 7$ is used. For DDFF and HCI, the RMSE is calculated on the disparity map with unit of pixel. For CVIA, the RMSE is calculated on the depth map with unit of meter. Largest possible aperture is used in all experiments.

	DDFF	CVIA	HCI
Focal Stack	0.018	0.035	0.36
Light Field	0.027	0.042	0.17

To better understand when the focal stack would perform better than a light field camera for depth estimation, Fig. 6 and Fig. 7 show qualitative depth estimation results on the HCI dataset and DDFF datasets. On the HCI dataset (Fig. 6), depth from light field can better resolve the fine structures compared to focal stack method, as can be seen, for example, by comparing the estimated depth maps of sample 1. This is likely because HCI dataset has a large disparity and hence the amount of defocus blur on the out-of-focus object is significant. Unless the object happens to be in focus on one of the image plane, it would be hard to precisely localize the object boundary using the focal stack. On the DDFF dataset, light field based method performs poorly and shows poor estimates on texture-less regions. This is because the maximum disparity in the DDFF dataset is small, and as a result the sub-aperture images in the light field become very similar. This makes it hard to estimate the depth from the light field. On the other hand, the focal stack based method is still able to produce smooth and good depth estimates in this case, by analyzing the small change in the focus sharpness, which is what a CNN excels at. This also suggests that more information is not always better, and the way the information is presented is also important: the light field, which has a larger data size and more information, may not perform better than focal stack on depth estimation, in the cases where the maximum disparity of the scene is small, e.g., small aperture camera, or far away objects. In such cases, it turns out that the more compact representation of the scene in the form of a focal stack is better suited for a neural network to estimate the depth.

Finally, Fig. 8 shows the depth-dependent error of the focal stack based method and the light field based method, obtained by dividing pixels into bins according to their ground truth depth and calculating the RMSE separately for each depth bin. The result shows that both methods exhibit an approximately quadratic error dependence on the object distance. This quadratic trend is reminiscent of the trend for stereo-based disparity estimation, which is given by:$\epsilon_z = \frac{z^2}{\cdot}\epsilon_d$, where ϵ_z is the depth error, z is the depth, b is the baseline, f is the focal length in unit of pixels, ϵ_d is the disparity matching error [23]. Given the similarity between the stereo-based method and focal stack/light field based methods, it is not surprising to see such a similar trend in Fig. 8. Depth-dependent error plots for the CVIA and DLSR datasets are presented in the supplementary material.

OPTICS CONTINUUM

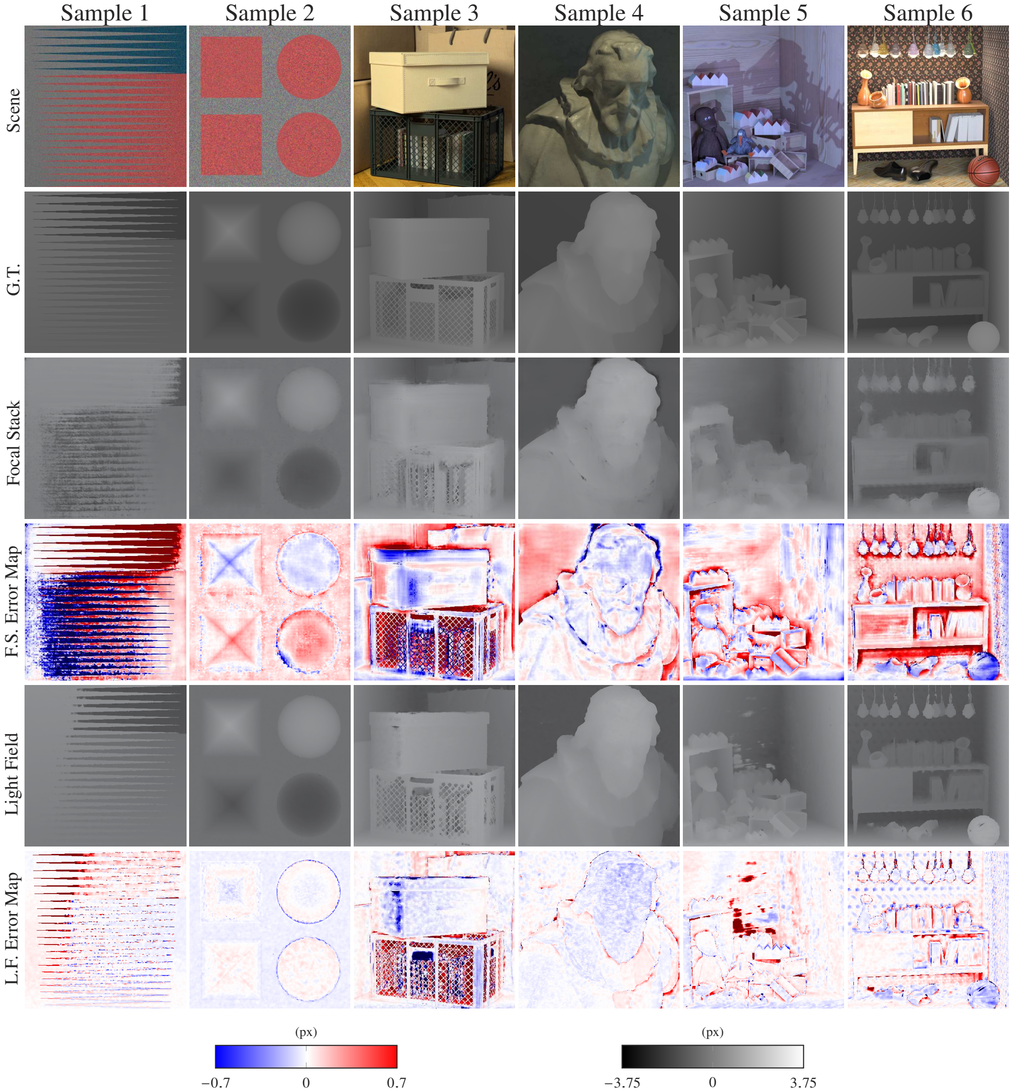


Fig. 6. Qualitative disparity estimation results from light field data and focal stack data on HCI dataset.

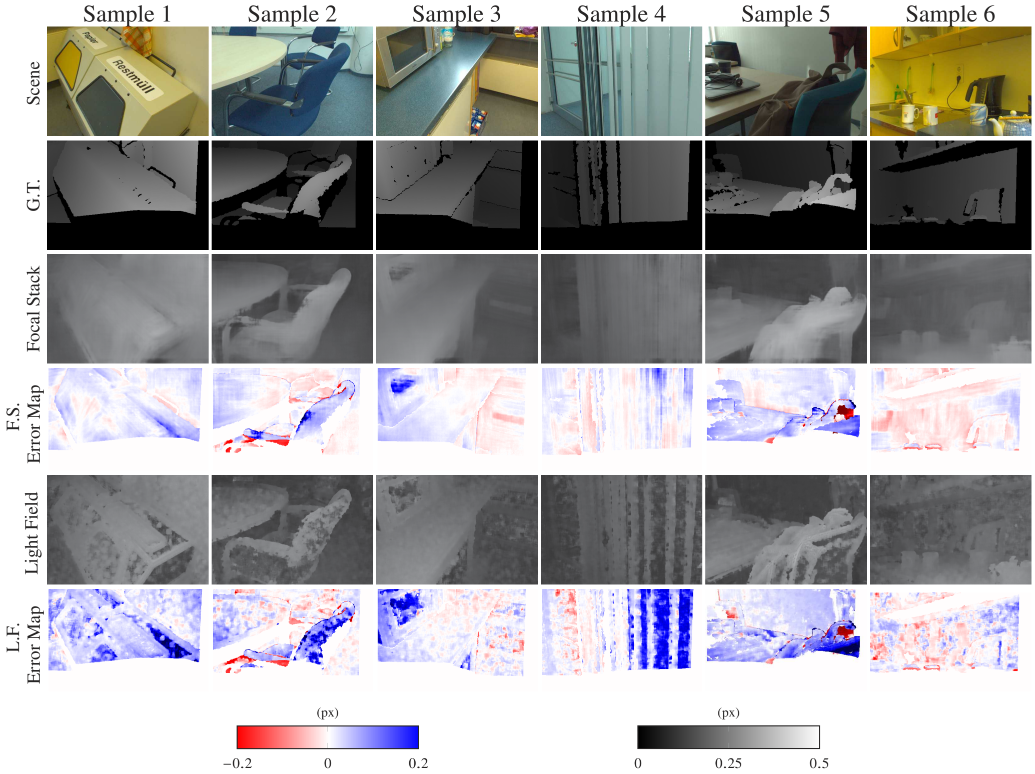


Fig. 7. Qualitative disparity estimation results from light field data and focal stack data on DDFF dataset.

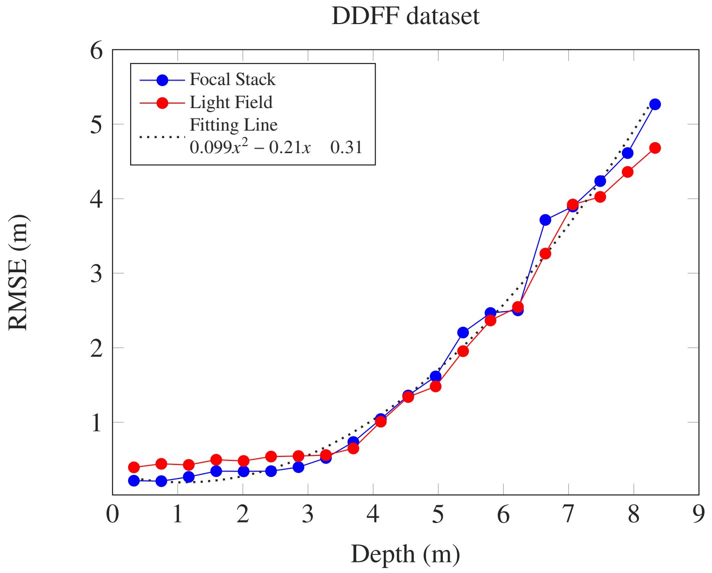


Fig. 8. Depth-dependent error of the estimation using focal stack and light field data. Dotted line: quadratic fitting for the focal stack data.

4. Conclusion

This paper has explored the focal stack camera design parameter space, including the number of focal planes, size of the aperture and sensor resolution, and studied their effects on the depth estimation performance, using three public light field datasets and an experimentally acquired DLSR focal stack dataset. We further compared the focal stack camera performance with the light field camera and showed that which one is better for depth estimation depends on the maximum disparity of the scene. These findings can be helpful for future designs of focal stack cameras.

Funding. W. M. Keck Foundation.

Acknowledgments. The authors gratefully acknowledge financial support from the W. M. Keck Foundation.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. D. Zhang, Z. Xu, Z. Huang, A. R. Gutierrez, C. J. Blocker, C.-H. Liu, M.-B. Lien, G. Cheng, Z. Liu, I. Y. Chun, J. A. Fessler, Z. Zhong, and T. B. Norris, "Neural network based 3D tracking with a graphene transparent focal stack imaging system," Nat. Commun. **12**, 1–7 (2021).
2. S. K. Nayar and Y. Nakagawa, "Shape from focus," IEEE Trans. Pattern Anal. Machine Intell. **16**(8), 824–831 (1994).
3. M. Moeller, M. Benning, C. Schönlieb, and D. Cremers, "Variational depth from focus reconstruction," IEEE Trans. on Image Process. **24**(12), 5369–5378 (2015).
4. P. Sakurikar and P. J. Narayanan, "Composite focus measure for high quality depth maps," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017).
5. C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Asian Conference on Computer Vision*, (Springer, 2018), pp. 525–541.
6. I. Y. Chun, Z. Huang, H. Lim, and J. A. Fessler, "Momentum-Net: Fast and convergent iterative neural network for inverse problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
7. M.-B. Lien, C.-H. Liu, I. Y. Chun, S. Ravishankar, H. Nien, M. Zhou, J. A. Fessler, Z. Zhong, and T. B. Norris, "Ranging and light field imaging with transparent photodetectors," Nat. Photonics **14**(3), 143–148 (2020).
8. Z. Huang, J. A. Fessler, T. B. Norris, and I. Y. Chun, "Light-field reconstruction and depth estimation from focal stack images using convolutional neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (IEEE, 2020), pp. 8648–8652.
9. S. Kuiper and B. Hendriks, "Variable-focus liquid lens for miniature cameras," Appl. Phys. Lett. **85**(7), 1128–1130 (2004).
10. R. Ng, *Digital light field photography*, vol. 7 (Stanford University Stanford, 2006).
11. K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision*, (Springer, 2016), pp. 19–34.
12. S. Pertuz, E. Pulido-Herrera, and J.-K. Kamarainen, "Focus model for metric depth estimation in standard plenoptic cameras," ISPRS J. Photogramm. Remote. Sens. **144**, 38–47 (2018).
13. D. Zhang, Z. Xu, Z. Huang, A. R. Gutierrez, I. Y. Chun, C. J. Blocker, G. Cheng, Z. Liu, J. A. Fessler, Z. Zhong, and T. B. Norris, "Graphene-based transparent photodetector array for multiplane imaging," in *CLEO: Science and Innovations*, (Optical Society of America, 2019), pp. SM4J–2.
14. S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers* (Now Publishers Inc, 2011).
15. E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*, (MIT Press, 1991), pp. 3–20.
16. C. Chen, H. Lin, Z. Yu, S. Bing Kang, and J. Yu, "Light field stereo matching using bilateral statistics of surface cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2014), pp. 1518–1525.
17. S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," Comput. Vis. Image Underst. **145**, 148–159 (2016).
18. C. Shin, H.-G. 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 the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), pp. 4748–4757.
19. Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34 (2020), pp. 12095–12103.
20. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, 2016.

21. C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Asian Conference on Computer Vision*, (Springer, 2018), pp. 525–541.
22. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, (2015), pp. 1–15.
23. D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys, "Variable baseline/resolution stereo," in *2008 IEEE conference on computer vision and pattern recognition*, (IEEE, 2008), pp. 1–8.