Light-Field Image Super-Resolution Using Convolutional Neural Network

Youngjin Yoon, *Student Member, IEEE*, Hae-Gon Jeon, *Student Member, IEEE*, Donggeun Yoo, *Student Member, IEEE*, Joon-Young Lee, *Member, IEEE*, and In So Kweon, *Member, IEEE*

***Abstract*—Commercial light field cameras provide spatial and angular information, but their limited resolution becomes an important problem in practical use. In this letter, we present a novel method for light field image super-resolution (SR) to simultaneously up-sample both the spatial and angular resolutions of a light field image via a deep convolutional neural network. We first augment the spatial resolution of each subaperture image by a spatial SR network, then novel views between super-resolved subaperture images are generated by three different angular SR networks according to the novel view locations. We improve both the efficiency of training and the quality of angular SR results by using *weight sharing*. In addition, we provide a new light field image dataset for training and validating the network. We train our whole network end-to-end, and show state-of-the-art performances on quantitative and qualitative evaluations.**

***Index Terms*—Convolutional neural network, light-field (LF) image, super-resolution (SR).**

I. Introduction

LIGHT-FIELD (LF) imaging [2] has recently come into the spotlight as the next generation imaging system. LF images contain spatial and angular information of the light ray distribution in space. Thus it can capture a multiview scene in a single photographic exposure. Many studies have shown the LF system's potential in improving the performance of many applications, such as alpha matting [3], saliency detection [4], and single LF image depth estimation [5]–[8].

In order to capture LF images using hand-held devices, a microlens array is placed in front of a camera sensor [9], [10]. The microlens array encodes angular information of the light rays, but results in a trade-off between spatial and angular resolutions in a restricted sensor resolution. This limitation makes it difficult to exploit the advantages of the LF cameras. Therefore, enhancing LF image resolutions is crucial to take full advantage of LF imaging.

Super-resolution (SR) aims at recovering a high-resolution image from a given low-resolution image. Recent SR approaches are mostly based on convolutional neural network [11], [12]. One major benefit is their generalization ability given sufficient training data to fit a model and to cover a wide range of distributions of expected test images. But, these single image SR algorithms cannot be directly applied to a LF image SR problem because the target of LF SR includes the number of subaperture images as well as the number of spatial pixels.

To simultaneously achieve spatial and angular SR, there are some previous studies [13]–[15] using Epipolar plane images (EPI) which are two-dimensional (2-D) slices of constant angular and spatial directions. As the EPI consists only of lines with various slopes, the intrinsic dimension is much lower than its ambient dimension, making image processing and optimization tractable. However, the low quality LF images captured by commercial LF cameras degrade the performance of these approaches. As already discussed in [8], LF images from commercial LF cameras suffer from lens aberration, microlens distortion, and vignetting, having negative impact on EPIs.

To overcome this issue on LF image SR, we propose a data-driven method using supervised learning. In this letter, we introduce a cascade convolution neural network (CNN) framework consisting of a spatial SR network and an angular SR network. In addition, as the existing LF image datasets are too small to train a CNN, we build a new LF image database with a variety of scenes, materials and textures. We train our network end-to-end using our database and demonstrate the state-of-the-art performances through both quantitative and qualitative evaluations.

II. LF Image SR

A. Overview

We propose a new SR method for LF images, which jointly increases the resolution in both the spatial and angular domains. Fig. 1 illustrates the architecture of the proposed model composed of a spatial SR network and an angular SR network, named LF convolution neural network (LFCNN). LFCNN first performs a spatial SR that increases the spatial resolution of the subaperture image, and then performs an angular SR that creates a new view between the subaperture images.

Manuscript received November 23, 2016; revised January 23, 2017 and February 9, 2017; accepted February 10, 2017. Date of publication February 15, 2017; date of current version May 1, 2017. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) under Grant 2010-0028680. The work of H.-G. Jeon was supported in part by the Global Ph.D. Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015H1A2A1034617). This letter was presented in part at the Proceedings of International Conference on Computer Vision Workshop, Santiago, Chile, Dec. 11–18, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Matteo Naccari. *(Corresponding author: In So Kweon.)*

Y. Yoon, H.-G. Jeon, and D. Yoo are with the Robotics and Computer Vision Laboratory, KAIST, Daejeon 34141, South Korea (e-mail: yjyoon@rcv.kaist.ac.kr; hgjeon@rcv.kaist.ac.kr; dgyoo@rcv.kaist.ac.kr).

J.-Y. Lee is with the Robotics and Computer Vision Laboratory, KAIST, Daejeon 34141, South Korea, and also with the Adobe Research, San Jose, CA 95110-2704 USA (e-mail: jolee@adobe.com.

I. S. Kweon is with the Robotics and Computer Vision Laboratory, KAIST, Daejeon 34141, South Korea, and also with the School of Electrical Engineering, KAIST, Daejeon 34141, South Korea (e-mail: iskweon77@kaist.ac.kr)).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2017.2669333

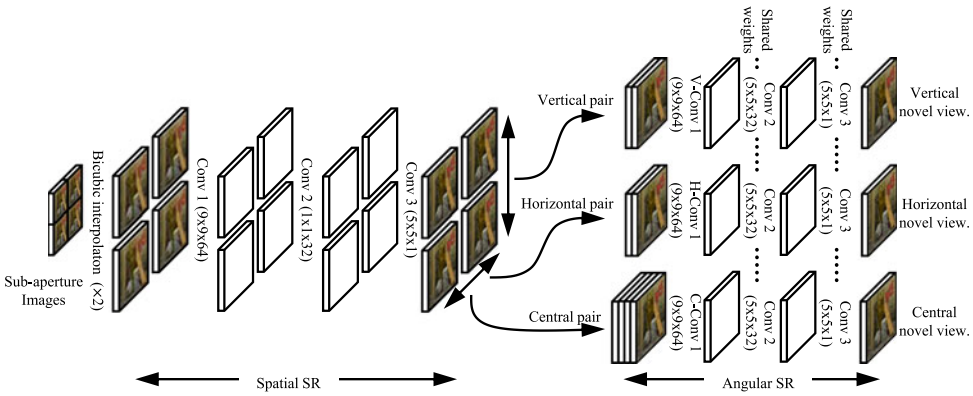

Fig. 1. Architecture of our light field convolution neural network (LFCNN). This is composed of a spatial SR network and the angular SR network. The spatial SR network performs spatial SR for each sub-aperter image. The angular SR network then takes three types of image pair and creates novel views for each type. We apply zero padding to make the same size of input and output images. The convolution filter is described as f × f × c, where f is filter size and c is the number of filters.

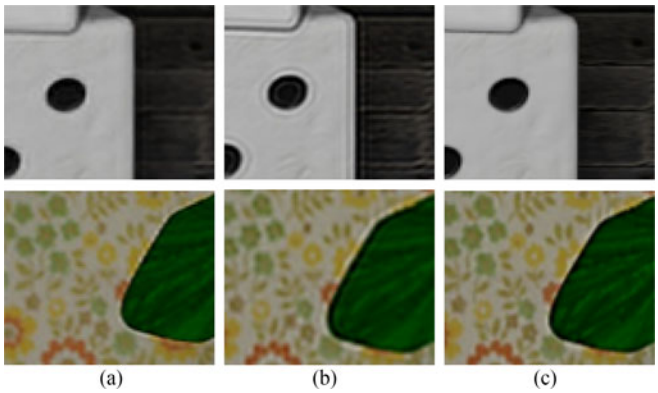

Fig. 2. Qualitative comparison of generated novel views. Compared with our previous model [1] showing ringing artifacts in high-frequency regions, the result of the proposed method has much less artifacts. (a) Proposed. (b) Yoon *et al.* [1]. (c) Ground truth.

Let us suppose, we have four subaperture images to take a look at this process. First, we increase the spatial resolution of these subaperture images by the bicubic interpolation with a desired up-scaling factor. Then, these subaperture images are put into a spatial SR network to enhance the high-frequency components. This spatial SR for the multiple subaperture images is performed independently of each other. The spatial SR network is based on the SR network proposed by Dong *et al.* [11]. Through these processes, we obtain four super-resolved subaperture images in the spatial domain.

Now, we perform the angular SR with the four subaperture images. These four views are inputs to the angular SR network, and the network creates novel views between them. The novel views can be created by three types of input combinations. For example, we can create a novel view between the two views located horizontally, and another novel view between the two views located vertically. Also, we can create a novel central view from all the four views. To this end, we design an angular SR network which takes these three types of input as shown in Fig. 1. Given the four subaperture images, the network creates five novel views including two novel views from the two vertical input pairs, two other novel views from the two horizontal input pairs and a central view from the four inputs. In other words, the network increases the angular resolution from 2 × 2 to 3 × 3 in this example.

B. Spatial SR network

The spatial SR network consists of three convolution layers as illustrated in Fig. 1. The first convolution layer is composed of 64 filters of 9 × 9 size, and the second one has 32 filters of 1 × 1 size. The last convolution layer is composed of a single filter of size 5 × 5. We do not use any pooling layer but apply zero padding in each to preserve the size of intermediate feature maps while minimizing the loss of information. All convolution layers are followed by ReLU [16] except the last convolution layer. To train this network, given N training images, we minimize the mean square loss between the estimation and the ground truth defined as

$$L_{\text{spatial}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\hat{X}^{i}_{\text{spatial}} - X^{i}_{\text{spatial}}\right\|_2^2 \tag{1}$$

where $\hat{X}$ is an estimation while X is the ground truth.

In our previous work [1], we had three separated spatial SR networks which correspond to each of the three types of inputs for the angular SR network. For example, a spatial SR network only takes a vertical input pair, while another spatial SR network only takes a horizontal input pair. In this letter, however, we define a single spatial SR network, which takes each subaperture image independently. Since a single network is learned with all subaperture images, it is more efficient than the previous network in terms of representational capability as well as computation.

C. Angular SR Network

The angular SR aims at augmenting the angular resolution from $S \times S$ views to $(2S-1) \times (2S-1)$ views. The most simplest way to do so is to use the typical interpolation methods such as bilinear or bicubic. However, estimating novel views with these methods do not take disparities into consideration. In contrast, the goal of our angular SR network is to estimate novel views by learning the disparity between the images of an input. Even if we do not supervise the network to directly produce disparities, supervising the network to generate a middle view

TABLE I
QUANTITATIVE EVALUATION ON THE SYNTHETIC HCI DATASET

Methods	PSNR(dB)						SSIM					
	Buddha			Mona			Buddha			Mona		
	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
Bilinear	33.57	33.66	33.78	34.14	34.25	34.32	0.9036	0.9151	0.9242	0.9242	0.9291	0.9320
Bicubic	34.22	34.63	35.14	34.10	34.20	34.25	0.9251	0.9334	0.9466	0.9484	0.9496	0.9512
Mitra and Veeraraghavan [14]	29.29	30.26	31.33	29.59	30.28	30.94	0.7795	0.7994	0.8190	0.7993	0.8171	0.8354
Wanner and Goldluecke [13]	24.43	29.69	36.97	25.40	30.76	37.60	0.7662	0.8691	0.9670	0.8542	0.9324	0.9862
Yoon *et al.* [1] (AngularSR+SpatialSR+FT)	**36.78**	36.86	36.94	37.31	37.40	37.48	0.9571	0.9580	0.9589	0.9667	0.9669	0.9671
Yoon *et al.* [1] (SpatialSR+AngularSR+FT)	36.71	36.84	36.92	**37.46**	37.56	37.64	0.9549	0.9558	0.9565	0.9637	0.9640	0.9644
Proposed (SpatialSR+AngularSR+FT)	36.25	**36.95**	**37.35**	37.03	**37.99**	**38.53**	**0.9579**	**0.9623**	**0.9657**	**0.9833**	**0.9863**	**0.9878**

Our approach significantly outperforms the state-of-the-art methods. Wanner and Goldluecke [13] and Mitra and Veeraraghavan [14] results are obtained by the source code from the authors.

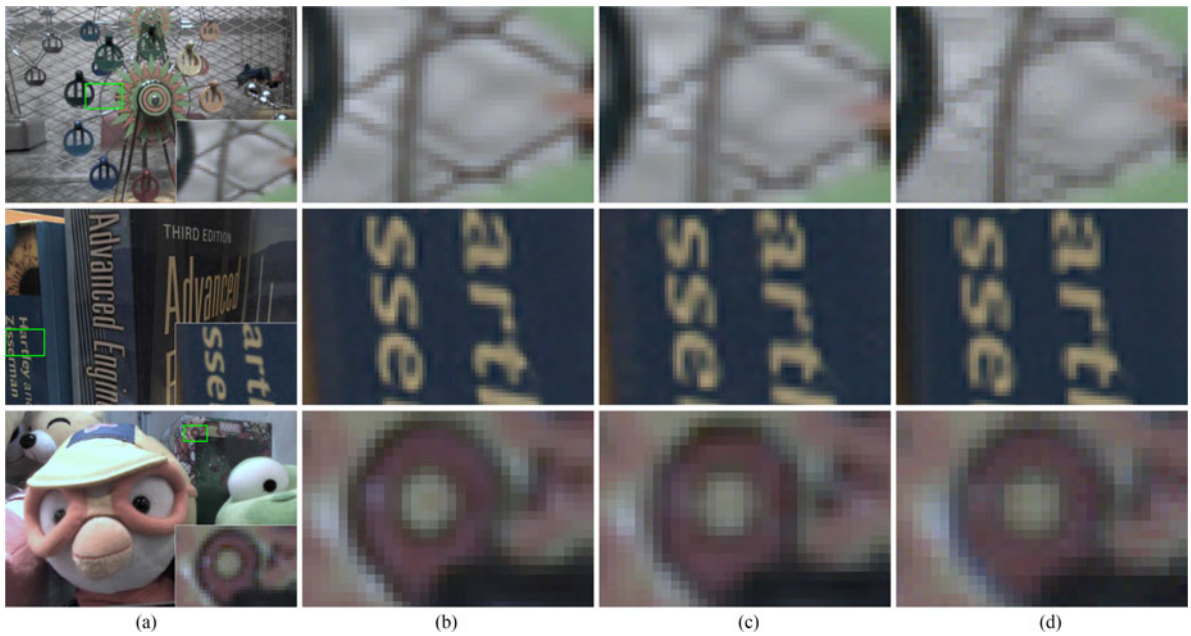

Fig. 3. Qualitative results on real-world images. We compare the proposed approach with Bicubic interpolation and [14]. Each image is a novel view. (a) Ground truth. (b) Proposed. (c) Bicubic. (d) Mitra and Veeraraghavan.

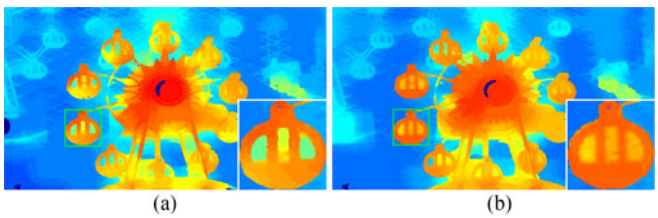

Fig. 4. Comparison of depth estimation. (a) Estimated depth from result images of the proposed approach. (b) Estimated depth from LF original images.

from the two input views makes the network implicitly estimate such information itself.

The angular SR network architecture is illustrated in Fig. 1. Since this network should take three types of input, the architecture begins with three parallel convolution layers corresponding to each input type, V-Conv1 for vertical image pair, H-Conv1 for horizontal image pair, and C-Conv for the four images. Each of these layers is composed of 64 filters of 9×9 size followed by ReLU. A feature map from one of the three parallel layers is then given to the second convolution layer composed of 32 filters of size 5×5, which is followed by the final convolution layer containing a single filter of size 5×5. Note the second and third convolution layers, marked Conv2 and Conv3 in Fig. 1, are *shared* for the three types of inputs. Similar to the spatial SR network, each convolution layer has ReLU except for the last convolution layer. We do not use any pooling layer but apply zero padding in each input of the convolution layers. To train this network, similar to the spatial SR network, we minimize the mean square loss between the estimation and the ground truth defined as

$$L_{\text{angular}} = \frac{1}{M} \sum_{i=1}^{M} \left\| \hat{X}^i_{\text{angular}} - X^i_{\text{angular}} \right\|_2^2 \qquad (2)$$

where M is the number of training input pairs in which vertical, horizontal and central pairs are evenly included.

In our previous work [1], we had three separated angular SR networks which correspond to each of the three types of

input. However, in this letter, we only parallelize the first-level convolution layers for the three types of input and make them share the rest of the convolutions. This architecture enforces the first layer to extract invariant representations from each of the different inputs while the rest of the layers performs the angular SR. Thus, it regularizes the angular SR network to generate consistent SR results from three different inputs. The proposed network has half the number of parameters compared to our previous model [1]. Given a limited training set, less number of parameters make our training procedure tractable. As shown in Fig. 2, while results from [1] show ringing artifacts at edges, our network infers accurate super-resolved images.

D. Training

For the spatial SR network, we synthetically generate blurry images by downsampling and upsampling original images via bicubic interpolation, so the original images are regarded as ground truths. For the angular SR network, we randomly choose odd or even numbered subaperture image pairs as inputs, then a view between a pair is regarded as the ground truth. In this way, we compose a large input pair set in which the three types of input are evenly included. Due to the limited general public utilities memory, we randomly crop 32×32 patches and make minibatches of size 16. Following [11], we transform the color space from red, green, blue (RGB) to YCbCr, and use only the luminance channel for training. In the inference stage, we apply our method to each of YCbCr channels and convert them to the RGB space again.

We independently train each network by minimizing the mean square loss defined as (1) and (2). The filters of the two networks are initialized by a Gaussian distribution with zero mean and standard deviation of 10^{-3}. We employ the Adam optimizer [17] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 10^{-4} which is decreased to 10^{-5} during fine tuning.

III. Experimental Results

In this section, we performed quantitative and qualitative evaluations to demonstrate the validity of the proposed method. We implement the proposed network using Tensorflow [18]. To generate a novel color view, our method takes around average of 0.07 s for 383×552 images taken from a Lytro illum camera on an Intel Core i7 3.6 GHz with GTX Titan X. We trained the network until convergence, which took about 5 h. Source code and dataset are available at https://youngjinyoon.github.io/

A. Quantitative Evaluation

For the quantitative evaluation, we used the Heidelberg Collaboratory for Image Processing (HCI) LF dataset [19] which provides 14 synthetic LF images with 9×9 angular resolution and 768×768 spatial resolution or more. We extracted one million patches by randomly cropping from the 12 training examples. In order to monitor overfitting, we use a test set of 200000 patches from two images ("Buddha" and "Mona"). The reason for selecting these two test set is because they show various disparity ranges, texture types, material, and illumination conditions.

In Table I, we report peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) values to numerically compare the state-of-the-art methods to the proposed framework. We also added PSNR and SSIM values of results from multidimensional interpolations (4-D bilinear and 4-D bicubic).

Works in [13] and [14] are optimization-based approaches using disparity information on input images which model the synthesis of novel views directly. We observe that a significant number of occluded regions makes the problem difficult. As expected, multidimensional approaches lead to blur artifacts in regions with large parallax. This phenomenon can be observed in Section III-B. On the other hand, our learning-based approach produces high-quality spatial upsampled image and the shared network efficiently models novel view synthesis without ringing artifacts compared to our previous network [1].

B. Qualitative Evaluation

For the SR of real-world LF images, we captured 307 LF images utilizing the Lytro Illum camera whose spatial and angular resolution is 383×552 and 5×5, respectively.[1] Subaperture images from raw LF images were generated by using a geometric LF calibration toolbox [20]. We used 201 LF images as the training set, and ensured our training set contained a variety of different scenes including textiles, woods, flowers, and fruits in order to handle a diverse test set.

As shown in Fig. 3, we compare the proposed method against the methods of [14] and multidimensional bicubic interpolation. Different from [14] which assumes the images to be ideal, images captured from commercial LF cameras make it hard to estimate accurate disparities. The bicubic interpolation also fails to synthesize novel views between subaperture images. However, our method learns to handle these inaccuracies without producing artifacts. We note that our diverse training patches extracted from various subaperture views help handle spatially-variant color aberration of LF images. As an application, we estimate a depth map using the super-resolved LF image (first row of Fig. 3). We used a multiview stereo matching-based depth estimation algorithm [8] to find correspondences between subaperture images. As stated in [8], a LF image with high spatial and angular resolution is preferred to obtain accurate correspondences. The depth map from the super-resolved LF image preserves fine details and the highresolution image is more accurately discretized than the original image as shown in Fig. 4.

IV. Conclusion

We have presented a new method for 4-D LF image SR. To simultaneously upsample the spatial and angular resolution of a LF image, we proposed an end-to-end trainable architecture by cascading spatial and angular SR networks. By adopting weight sharing among the angular network modules, we improve the efficiency of network training and also generate consistent novel views without ringing artifacts. In addition, we provided more than 300 LF images captured from an off-the-shelf commercial LF camera and validated the practical performance of our method in real-world environments. Experimental results show that our method outperforms the state-of-the-art methods for LF image SR on synthetic and real-world datasets. In the future, we expect that the propose framework shows better results if we apply optical characteristics of LF imaging such as the wave diffraction model [21] into the learning framework.

[1]The actual angular resolution of the Lytro Illum cameras is 15×15. But, the five views from each side suffer from severe vignetting, and thus, we decided to use only 5×5 middle views.

REFERENCES

[1] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. So Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 57–65.

[2] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," *Comput. Models Vis. Process.*, vol. 1, no. 2, pp. 3–20, 1991.

[3] D. Cho, S. Kim, and Y.-W. Tai, "Consistent matting for light field images," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 90–104.

[4] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2806–2813.

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

[6] Z. Yu, X. Guo, H. Ling, A. Lumsdaine, and J. Yu, "Line assisted light field triangulation and stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2792–2799.

[7] S. Heber and T. Pock, "Shape from light field meets robust PCA," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 751–767.

[8] H.-G. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1547–1555.

[9] Raytrix, "3D light field camera technology." [Online]. Available. http://www.raytrix.de/

[10] Lytro, "The lytro camera." [Online]. Available. http://www.lytro.com/

[11] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[12] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen, "Deep network cascade for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 49–64.

[13] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 606–619, Mar. 2014.

[14] K. Mitra and A. Veeraraghavan, "Light field denoising, light field superresolution and stereo camera based refocussing using a GMM light field patch prior," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2012, pp. 22–28.

[15] T. E. Bishop and P. Favaro, "The light field camera: Extended depth of field, aliasing, and superresolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 972–986, May 2012.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[18] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems." 2015. [Online]. Available: tensorflow.org

[19] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," in *Proc. Vis., Model. Vis.*, 2013, pp. 225–226.

[20] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light-field cameras using line features," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 47–61.

[21] M. Broxton *et al.*, "Wave optics theory and 3-D deconvolution for the light field microscope," *Opt. Express*, vol. 21, no. 21, pp. 25 418–25 439, 2013.