Light field Image Compression Using Depth-based CNN in Intra Prediction

Tingting Zhong, Xin Jin, *Senior Member*, *IEEE*, Lingjun Li and Qionghai Dai, *Senior Member, IEEE*

***Abstract*—Recently, light field images have received extensive attention due to their potential applications. Since they take up a huge memory because of its super-high resolution, efficient compression methods are fundamentally required. In this paper, we propose a novel intra prediction mode by using depth-adaptive convolutional neuro network (DCNN). Light field projection finds the imaging response distribution for each object point using the depth estimated from each macropixel in the light field image. The highly correlated imaging responses are used to select the neural network structure. The network structure also adapts to the to-be-encoded block size. Adding the proposed DCNN-based prediction mode into the rate-distortion optimization loop with other 35 intra prediction modes of HEVC, the proposed encoding scheme achieves a significant bit-rate saving compared to representative compression approaches with limited computational complexity increment. Statistical data are also provided and analyzed to demonstrate the efficiency of the proposed method.**

***Index Terms*—Light field image compression, intra prediction, depth-adaptive convolutional neuro network.**

I. INTRODUCTION

Recently, plenoptic cameras [1], like *Lytro* [2] and *Raytrix* [3], have attracted great attentions both from academy and industry. Unlike traditional cameras, the plenoptic cameras capture spatial and angular intensity of light using a single exposure for the three-dimensional (3D) scene. There are many applications that the captured plenoptic images can be applied to, such as digital refocusing [4], synthesizing viewpoints [5], depth estimation [6]. Since a plenoptic image is a super-high-definition image, it requires efficient compression methods for storage and transmission.

Existing compression methods for plenoptic images can be mainly divided into two categories: approaches that compress plenoptic images directly and approaches that compress pseudo-video sequences generated from plenoptic images. The approaches in the first category improve the compression efficiency by utilizing the spatial correlations [7]. The spatial best match is retrieved by neighborhood macropixels displacement prediction [8] or spatial motion search [9], which greatly improves compression efficiency with the introduction of overhead in encoding the displacement vector or mode information. The approaches in the second category utilize the correlation among the subapertures to improve the compression efficiency. They compress the pseudo-videos generated from the subaperture images directly [10][11], or compress them using multiview coding structure [12], or compress the key image group and the residue between the key images and other images [13][14]. However, all of them discover the spatial and inter-view correlation only from the pixel intensities in the lenslet images. We proposed an intra prediction work in [15] to exploit correlations among the images under the microlens considering that the image responses on the sensor are directly affected by the object distance and the optical parameters during acquisition. While, since the exact optical parameters are generally not available, such image response correlation cannot be extracted accurately. So, we are considering whether it is possible to introduce machine learning into compression to find the inner correlation among the imaging responses.

Some researchers have tried to apply machine learning to video/image compression. They can also be categorized into applying machine learning to the existing encoding platforms or encoding directly by machine learning. The approaches applying machine learning to the existing encoding platforms are used to replace or enhance the coding tools in encoding platforms, such us the learning-based intra prediction mode decision [16], the CNN-based residual prediction of each CTU [17], the fully connected network-based Intra prediction [18]. The approaches that realize encoding directly by machine learning generates novel encoding architecture. G. Toderici *et al.*[19] proposed an architecture consists of a recurrent neural network-based encoder and decoder, a binarizer, and a neural network for entropy coding. Some others implement automatic loop coding by using multi-stage residual encoder in [20][21]. These methods can improve the coding efficiency to some extent, but none of them is designed for light field image compression.

Consequently, in this paper, we propose a novel light field image compression method by designing depth-adaptive convolutional neuro network (DCNN). The correlations among the imaging responses are exploited according to light field projection and CNNs are designed to generate the accurate prediction for the current block using the depth-adaptive reference blocks containing the highly correlated imaging responses. The network structure also adapts to the to-be-encoded block size. Adding the proposed CNN prediction mode into the rate-distortion optimization (RDO) loop with other 35 intra prediction modes of HEVC, experimental results and analyses demonstrate obvious

bit-rate saving compared to representative compression approaches, like HEVC.

The remainder of this paper is organized as follows. Our method is described in Section II. Experimental results are provided in Section III and Section IV concludes the paper.

II. Proposed Compression Algorithm

A. *Proposed system architecture*

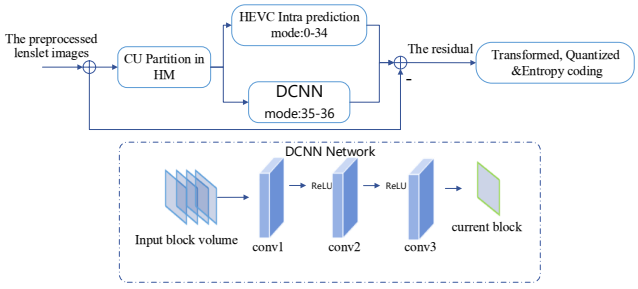


Fig.1. The proposed encoding architecture.

The system architecture of the proposed method is shown in Fig.1. As shown in the figure, the preprocessed lenslet image is taken as the input, which is generated by light field decoding [22] and image reshaping [23]. Light field decoding contains devignetting, demosaicing and rotation and scaling to obtain a colored lenslet image with regular macropixel spacing. Image reshaping is to align the macropixels with the coding unit grid and to maximize the intensity continuity among the adjacent macropixels. After preprocessing, the lenslet image presents regular and well-aligned macropixel structure, which improves the horizontal/vertical correlations obviously and is friendly to block based hybrid encoding.

The preprocessed lenslet images are encoded in HM platform. After the quad-tree based CU partition in HM, the intra prediction is performed. For each CU Block size, we proposed two new DCNN-based intra prediction modes, which select different reference blocks according to the depth of current block. One mode benefits the blocks imaging the focused objects and the other benefits the blocks imaging the defocused objects. The references used as the inputs of the CNN network are designed using the depth of the current block. The structure of the network is three-layer convolutional neural network. RDO is performed for the proposed intra modes and the existing 35 intra modes in HEVC to determine the best prediction mode. After intra prediction, the residual will be transformed, quantized and entropy coded according to HEVC. Also, the mode information related to the new modes will be encoded.

B. *Proposed DCNN-based intra prediction*

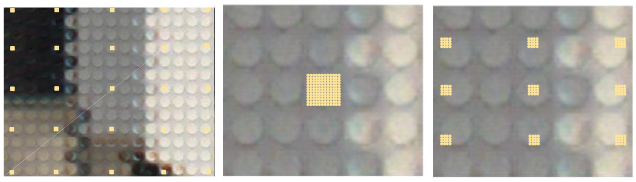
The proposed DCNN-based intra prediction tries to exploit the inner correlations among the image responses. Since the spatial distribution the image responses is determined by the separation between the object distance and the focus distance of the camera, its variation with the depth is first analyzed to define the best references for the network.

Using two-plane parameterization, a pixel in the lenslet image can be denoted by $L(x, y, u, v)$,where (x, y) is the spatial coordinates, and (u, v) is the angular coordinates signaling the ray direction. For an object point in the 3D space, i.e. for a given object distance, the distribution of its image responses varies with the focus distance of the camera by [25]

$$L_\alpha(x', y', u, v) = L(x + u(1 - \frac{1}{\alpha}), y + v(1 - \frac{1}{\alpha}), u, v), \quad (1)$$

where α is the relative depth between the new focus distance and the focus distance of $L(x, y, u, v)$; and L_α is the image response at α . Specifically, $\alpha = 1$ represents that the current focus plane coincides with the focus plane of $L(x, y, u, v)$. Thus, as α is smaller than or larger than 1, the image response can be derived and the samples are shown in Fig. 2. It can be found that as $\alpha = 1$, the image responses of the same object point are gathered in one macropixel. As α is smaller than or larger than 1, the image responses of the same object point scatter in the macropixels away from the current micropixel, while, they may scatter discretely in a much wider region as $\alpha < 1$.

It is noted that such distribution feature is invertible. If the focus distance is fixed and the object distance varies, which is very general when capture an image for a real scene, the inverse variation in the distribution happens. As the object distance is larger/smaller than the focus distance, the distribution is identical to $\alpha < 1 / \alpha > 1$. Since the image responses correspond to the same object point in the 3D space, they present higher correlation among each other than other pixels.



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.5$

Fig.2. Ray contribution to pixels with the variation in focus plane.

Considering the range and the sparsity of image responses distribution is determined by the depth of the object, the depth ranges of the natural scenes are analyzed by applying a representative depth estimation algorithm proposed by Jeons *et al*. [26] to several natural scenes captured by Lytro Illum camera. Setting the range of α from 0.02 to 2 and the default focused layer to $\alpha = 1$, the number of pixels on each α is accumulated. It is found that the depth range [0.5, 1.5] is the range with the highest number of pixels, which corresponds to the objects are placed a bit closer or farther from the focused plane, i.e. a bit defocused, while not too far away from it, i.e. not too defocused. Using Eq. (1), the coordinates (x', y', u, v) of the image responses of the object point can be derived. It is found that as $\alpha = 1.5$, they are distributed in the region that is two to three macropixel-distance around the

current block. However, when $\alpha = 0.5$, the responses are distributed seven to eight macropixel-distance away from the current block, which requires much bigger on-chip memory to buffer the reference pixels. Considering the computational complexity and storage complexity overhead introduced by selecting the reference blocks over a wider range in intra prediction jointly, we select the reference blocks located two to three macropixel-distance around the current block for the defocused blocks.

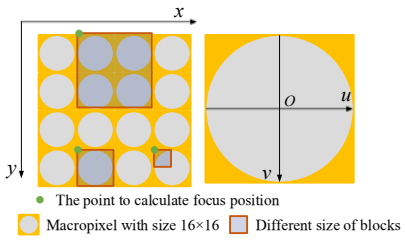


Fig. 3. The top-left point of different size of blocks. The size is from big to small: 32×32; 16×16; 8×8.

Thus, treating the depth value α of the top-left point (the green points in Fig.3) in the current block to be the depth of the whole block, the reference blocks derived according to Eq. (1) are shown in Fig. 4. For the focused blocks, as shown in Fig.4(a), four collocated blocks in the left, the top-left, the top, and the top-right marcopixels are selected as the reference blocks for CNN. For the defocused blocks, as shown in Fig.4(b), the maximum ten collocated reference blocks in the neighboring macropixels are selected as the input of CNN.

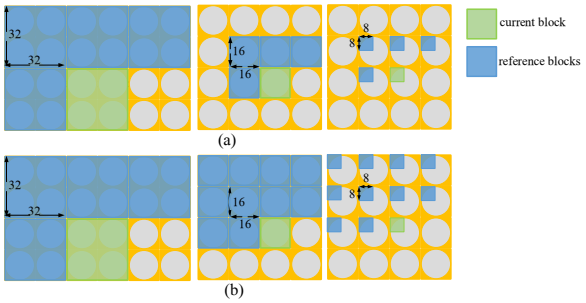


Fig. 4. Reference block selection for (a) focused blocks; (b) defocused blocks.

For the network structure, we have compared the complexity and compression efficiency among the fully connected network, the convolutional network [27] and the ResNet [28]. We adopted the convolutional network as a network choice in our design because of its low complexity and relative high efficiency. Thus, the network structure of the proposed DCNN-based intra prediction, as shown in Fig. 1, contains a sequence of 3 2D-Convolutional layers, each of which has a ReLU activation except the last layer. The first convolutional layer takes the volume of reference blocks with the dimension $m \times m \times n$, where m is the size of coding block, n is the quantity of reference blocks. The amount of filters is 256 with the size 5×5, and a stride of one with padding two. In second convolutional layer , the amount of filters is 128 with the size 1×1, and a stride of one with no padding. In last convolutional layer, the amount of filters is 1 with the size 3×3, and a stride of one with padding one. Note that in all the convolution layers the input is padded such that the activation map of each filter has the same size as the input. In last, we obtain the current prediction blocks with the size of $m \times m \times 1$.

For each CU Block size, except size 64×64, we increase two new intra prediction modes, a network forward operation will be conducted when calculate the cost of each new intra prediction mode in each CU size, it will increase the complexity of the program sharply when encoding, For decoding, it also brings complexity but better than encoding since no CU partition in decoding.

III. Experimental Results and Analysis

In this section, we have conducted experiments to evaluate the proposed depth-adaptive convolutional neuro network (DCNN) in HM encoding platform.

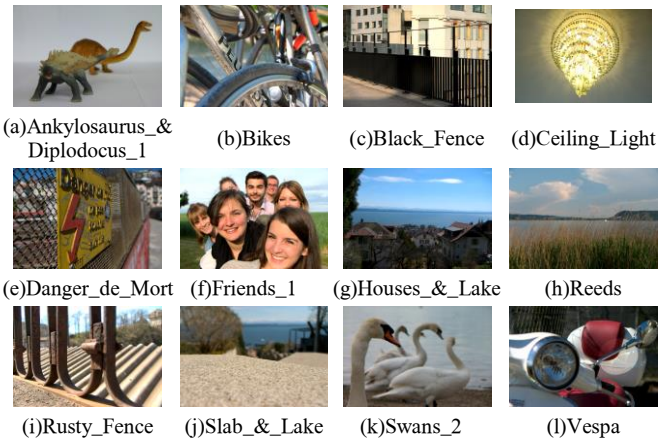
The experimental evaluation is done on the EPFL Light Field dataset [29], which is available online [30] and contains 118 lenslet images divided into 10 categories. The plenoptic images of this dataset were entirely captured by Lytro Illum cameras. We select 106 lenslet images of dataset as training samples for DCNN, and the other 12 lenslet images are selected as testing samples, which are shown in Fig.5.

At each coding size 32×32, 16×16 and 8×8, the focused and defocused situations were treated as two intra prediction modes, so we should train six networks. Since it has the same reference blocks in the focused and defocused case at the coding size 32×32 according to Fig.4, we treat these two cases as one, so we only need to train five networks. In addition, we selected the luminance component data for training network because our eyes are more sensitive to luma, and the prediction of chorma format utilized the same network with the luma prediction in HM platform. Also, we only train the dataset at QP=22, and it just uses the same network at different QP in HM platform. The training work was conducted in a PC with Intel® CPU E5-2620 V3 @ 2.40 GHz with 160 RAM and 64-bits Windows Server 2012 R2 standard operating system, the final training results were all convergent.

The HEVC reference software HM-16.9_SCM8.0 [31] was used with the configure file of main-RExt defined in [32] at four QP values 22,27, 32 and 37, and we use BD-Bitrate [33] to measure the compression efficiency, which is measured by the subaperture images rendered from the original lenslet image and the reconstructed lenslet image.

To demonstrate the effectiveness of the proposed algorithm, three testing cases as TPVA [34], two proposed situations are tested. TPVA is a pseudo video coding approach which generates the pseudo video by tiling the lenslet image into 464 by 320 size image sequences. *Proposed*1(Focused Mode Only) is our algorithm which was considered all situations as a focused situation, and *Proposed*2(Focused + Defocused Mode) is the proposed

algorithm which was considered the focused and defocused situation separately. Compression efficiency comparison results are shown in Table I. The proposed method always outperforms HEVC with considerable gain. *Proposed*1 can achieve a maximum of 31.37%/34.99% bitrate reduction relative to HEVC/ TPVA, and it reduces bitrate by an average of 10.76%/1.11% relative to HEVC/ TPVA. *Proposed*2 can achieve a maximum of 34.04%/37.45% bitrate reduction relative to HEVC/ TPVA, and it reduces bitrate by an average of 11.42%/1.79% relative to HEVC/ TPVA. It shows that TPVA has 11.36%/17.36% bitrate increment relative to HEVC for image(d)/(g) but also has huge bitrate reduction relative to HEVC for image(i)/(j) in Table I. Our method has bitrate reduction for almost all test images and it seems more stable than TPVA. The number of bits signaling the intra coding mode is increased from 5 to 6 bits. When the overhead bit increment is larger than the prediction efficiency improvement, a small performance loss relative to HEVC is introduced like that for image(d)in Fig. 5.



(a)Ankylosaurus_& Diplodocus_1 (b)Bikes (c)Black_Fence (d)Ceiling_Light

(e)Danger_de_Mort (f)Friends_1 (g)Houses_&_Lake (h)Reeds

(i)Rusty_Fence (j)Slab_&_Lake (k)Swans_2 (l)Vespa

Fig.5. Testing lenslet images.

To evaluate the performance of our proposed method, the mode selection proportion was counted at three cases, HEVC, *Proposed*1 and *Proposed*2. The statistic results of average intra prediction mode proportion in 12 test samples are shown in Table II. The proposed mode occupies the most proportion in all modes, which shows that the proposed can achieve more accurate prediction. Obviously, compare *Proposed*1 with *Proposed*2, the first five columns mode proportion is nearly never change, and the sum proportion of the mode 35 and 36 in *Proposed*2 is nearly equal with the proportion of the mode 35 in *Proposed*1, it shows that adding the defocused case make some defocused blocks achieve more accurate prediction.

TABLE I BD-BITRATE: COMPRESSION EFFICIENCY COMPRASION.

Image	TPVA vs.HEVC	*Proposed*1 vs. HEVC	*Proposed*2 vs. HEVC	*Proposed*1 vs. TPVA	*Proposed*2 vs. TPVA
(a)	3.09%	-23.22%	-24.09%	-25.60%	-26.46%
(b)	-22.69%	-10.60%	-11.38%	16.03%	15.02%
(c)	-5.99%	-7.14%	-7.74%	-1.35%	-1.98%
(d)	11.36%	0.39%	0.42%	-9.87%	-9.85%
(e)	-8.99%	-7.25%	-7.67%	1.82%	1.36%
(f)	-16.43%	-2.95%	-3.23%	16.14%	15.81%
(g)	17.36%	-20.45%	-21.14%	-32.51%	-33.09%
(h)	-7.22%	-1.47%	-1.74%	6.17%	5.87%
(i)	-23.60%	-4.16%	-4.41%	25.63%	25.31%
(j)	-33.46%	-6.30%	-6.59%	41.43%	40.95%
(k)	4.37%	-31.37%	-34.04%	-34.99%	-37.45%
(l)	1.50%	-14.61%	-15.43%	-16.19%	-16.95%
Average	-6.73%	-10.76%	-11.42%	-1.11%	-1.79%

IV. Conclusion

This paper proposed a novel intra prediction algorithm by using depth-adaptive convolutional neuro network. We have integrated DCNN to HM encoding platform, and combine the relative between the reference and the current blocks when object point at different depth to predict the current block. the coding results demonstrated the plenoptic images were compressed efficiently by using proposed method. It outperforms existing compression methods such as HEVC/TPVA by an average of 11.42%/1.79% bitrate reduction. In the future, we may use the depth maps to train the neural networks, which can exploit depth information for accurate prediction.

V. Acknowledgements

This work was supported in part by Shenzhen project JCYJ20170307153135771 and Foundation of Science and Technology Department of Sichuan Province 2017JZ0032c, China.

TABLE II PROPORTION OF DIFFERENT INTRA PREDICTION MODES

QP	Method	Planar: 0	DC: 1	Horizontal: 10	Vertical: 26	Others	Focused: 35	Defocused: 36
22	HEVC	18.11%	40.74%	4.69%	5.74%	30.72%		
	*Proposed*1	12.58%	30.10%	2.48%	4.09%	17.51%	33.24%	
	*Proposed*2	12.23%	29.46%	2.43%	3.93%	17.29%	26.08%	8.57%
27	HEVC	21.00%	36.85%	10.06%	11.76%	20.32%		
	*Proposed*1	13.87%	24.76%	4.88%	7.97%	12.38%	36.14%	
	*Proposed*2	13.61%	24.22%	4.84%	7.79%	12.33%	29.41%	7.81%
32	HEVC	22.09%	28.67%	14.97%	20.85%	13.42%		
	*Proposed*1	15.94%	20.90%	6.84%	12.10%	9.70%	34.52%	
	*Proposed*2	15.47%	20.53%	7.05%	12.03%	9.81%	30.97%	4.14%
37	HEVC	21.73%	21.14%	18.14%	26.80%	12.20%		
	*Proposed*1	17.90%	16.91%	9.10%	17.90%	9.88%	28.32%	
	*Proposed*2	17.82%	16.80%	9.14%	17.63%	9.93%	27.36%	1.33%

VI. REFERENCES

[1] R. Ng, M. Levoy, M. Brédif. G. Duval, M. Horowitz, "Light field photography with a hand-held plenoptic camera." Stanford University Cstr, 2005.

[2] Lytro. [Online]. Available: https://www.lytro.com/.

[3] Raytrix. [Online]. Available: https://raytrix.de/

[4] C. Chen, Y. C. Lu, and M. S. Su. "Light field based digital refocusing using a DSLR camera with a pinhole array mask." IEEE International Conference on Acoustics Speech and Signal Processing, 2010:754-757.

[5] Taguchi, Yuichi, and T. Naemura. "View-Dependent Coding of Light Fields Based on Free-Viewpoint Image Synthesis." IEEE International Conference on Image Processing (ICIP), 2006:509-512.

[6] Tao, W. Michael, et al. "Depth from Combining Defocus and Correspondence Using Light-Field Cameras." IEEE International Conference on Computer Vision IEEE Computer Society (ICCV), 2013:673-680.

[7] R. J. Monteiro, P. Nunes, N. Rodrigues, et al. "Light Field Image Coding using High Order Intra Block Prediction." IEEE Journal of Selected Topics in Signal Processing, 2017, PP(99):1-1.

[8] Y. Li, M. Sjöström, R. Olsson, &U. Jennehag, et al. "Coding of Focused Plenoptic Contents by Displacement Intra Prediction." IEEE Transactions on Circuits & Systems for Video Technology, 2016, 26(7):1308-1319.

[9] D. Liu, P. An, R. Ma, L. Shen, et al. "Disparity compensation based 3D holoscopic image coding using HEVC". IEEE China Summit and International Conference on Signal and Information Processing. IEEE, 2015:201-205.

[10] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, W. Zeng, et al. "Pseudo-sequence-based light field image compression." IEEE International Conference on Multimedia & Expo Workshops. IEEE, 2016:1-4.

[11] L. Li, Z. Li, B. Li, D. Liu, H. Li, "Pseudo-Sequence-Based 2-D Hierarchical Coding Structure for Light-field Image Compression." IEEE JSTSP, Vol. 11, No. 7, p1107-1119, Oct. 2017.

[12] W. Ahmad, R. Olsson, & M. Sjöström. "Interpreting plenoptic images as multi-view sequences for improved compression." IEEE International Conference on Image Processing. IEEE, 2018:4557-4561.

[13] S. Zhao, Z. Chen. "Light field image coding via linear approximation prior." IEEE International Conference on Image Processing. IEEE, 2018:4562-4566.

[14] J. Hou, J. Chen, L. Chau. "Light Field Image Compression Based on Bi-Level View Compensat ion with Rate-Distortion Optimization." IEEE Transactions on Circuits & Systems for Video Technology, 2018, PP(99):1-1.

[15] X. Jin, H. Han, Q. Dai. "Plenoptic Image Coding Using Macropixel-Based Intra Prediction." IEEE Transactions on Image Processing, 2018, 27(8):3954-3968.

[16] T. Laude, J. Ostermann. "Deep learning-based intra prediction mode decision for HEVC." Picture Coding Symposium. IEEE, 2017.

[17] Z. Zhang, C. Yeh, L. Kang, et al. "Efficient CTU-based intra frame coding for HEVC based on deep learning." Asia-Pacific Signal and Information Processing Association Summit and Conference. IEEE, 2018:661-664.

[18] J. Li, B. Li, J. Xu, et al. "Fully Connected Network-Based Intra Prediction for Image Coding." IEEE Trans Image Process, 2018, PP(99):3236-3247.

[19] G. Toderici, D. Vincent, N. Johnston, et al. "Full Resolution Image Compression with Recurrent Neural Networks." 2016:5435-5443.

[20] D. Minnen, G. Toderici, M. Covell, et al. "Spatially adaptive image compression using a tiled deep network." IEEE International Conference on Image Processing. IEEE, 2018:2796-2800.

[21] M. Baig, V. Koltun, L. Torresani. "Learning to Inpaint for Image Compression." Neural Information Processing Systems (NIPS) 2017.

[22] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenslet-based plenoptic cameras," in Proc. 2013 IEEE Conf. Comput. Vis. Pattern Recog., Portland, OR, USA, 2013, pp. 1027–1034.

[23] X. Jin, H. Han, Q. Dai. "Image Reshaping for Efficient Compression of Plenoptic Content." IEEE Journal of Selected Topics in Signal Processing, 2017, PP(99):1-1.

[24] E. Y. Lam, "Computational photography with Plenoptic camera and light field capture: tutorial," Journal of the Optical Society of America A, 2015, 32(11):2021-2032.

[25] R. Ng, "Digital light field photography." Phd Thesis Stanford University, 2006, 115(3):38-39.

[26] H. G. Jeon, J. Park, G. Choe, J. Park, Y. Bok, Y. W. Tai, et al. "Accurate depth map estimation from a lenslet light field camera." 2015:1547-1555.

[27] C. Dong, C.L. Chen, K. He, X. Tang. "Learning a Deep Convolutional Network for Image Super-Resolution." Computer Vision – ECCV 2014. Springer International Publishing.

[28] K. He, X. Zhang, S. Ren, et al. "Deep Residual Learning for Image Recognition." 2015:770-778.

[29] M. Rerabek and T. Ebrahimi, "New Light Field Image Dataset," in Proc. Quality of Multimedia Experience, 2016.

[30] " JPEG Pleno Database: EPFL light-field data set," https: //jpeg.org/plenodb/lf/epfl/.

[31] [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/ HM-16.9 + SCM-8.0/. Accessed on: 2017.

[32] O.C. Au, X. Zhang, C. Pang, and X. Wen, "Suggested Common test conditions and software reference configurations for Screen Content Coding," Joint Collaborative Team on Video Coding (JCT-VC), Torino, JCTVC-F696, July, 2011.

[33] G. Bjontegaard, Calculation of Average PSNR Difference Between RDCurves, ITU-T VCEG-M33, 2001.

[34] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW), Jul. 2016, pp. 1–4.