Fast Disparity Estimation from a Single Compressed Light Field Measurement

1st Emmanuel Martínez
Department of Computer Science
Universidad Industrial de Santander
Bucaramanga, Colombia
emmanuel2162134@correo.uis.edu.co

2nd Edwin Vargas
Department of Computer Science
Universidad Industrial de Santander
Bucaramanga, Colombia
edwin.vargas4@correo.uis.edu.co

3rd Henry Arguello
Department of Computer Science
Universidad Industrial de Santander
Bucaramanga, Colombia
henarfu@uis.edu.co

***Abstract*—The abundant spatial and angular information from light fields has allowed the development of multiple disparity estimation approaches. However, the acquisition of light fields requires high storage and processing cost, limiting the use of this technology in practical applications. To overcome these drawbacks, the compressive sensing (CS) theory has allowed the development of optical architectures to acquire a single coded light field measurement. This measurement is decoded using an optimization algorithm or deep neural network that requires high computational costs. The traditional approach for disparity estimation from compressed light fields requires first recovering the entire light field and then a post-processing step, thus requiring long times. In contrast, this work proposes a fast disparity estimation from a single compressed measurement by omitting the recovery step required in traditional approaches. Specifically, we propose to jointly optimize an optical architecture for acquiring a single coded light field snapshot and a convolutional neural network (CNN) for estimating the disparity maps. Experimentally, the proposed method estimates disparity maps comparable with those obtained from light fields reconstructed using deep learning approaches. Furthermore, the proposed method is 20 times faster in training and inference than the best method that estimates the disparity from reconstructed light fields.**

***Index Terms*—Disparity estimation, light field, compressive sensing, convolutional neural networks.**

I. Introduction

Disparity estimation has been well studied in recent years, since it allows to infer the scene geometry from 2D images. Some of its most relevant applications include computational 3D modeling of real-life scenarios, face recognition, robotics and medicine [1]–[5]. Disparity estimation can be achieved taking advantage of the angular information of the light fields [6]–[10]. Light fields collect the amount of light coming from all directions at each spatial point of a physical scene, nevertheless, the huge captured amount of spatial and angular information produces large computational storage and processing costs. Additionally, the multiplexing in microlenses of some of the optical systems also imposes a trade-off between spatial and angular resolution [10]. To overcome these limitations, different optical architectures [11]–[13] that leverage the CS theory [14] have been proposed for efficient acquisition. Specifically, CS exploits the fact that if the light fields are sparse in some representation basis, they can be compressed [14].

Once acquired the compressed measurements, the recovery of the light field must be carried out in a digital processing step. Light field reconstruction can be performed using traditional CS algorithms [11]–[13], or more recently, using deep learning approaches [10], [15]. Most advanced methods for light field reconstruction proposed to design coded masks in an end-to-end (E2E) approach [16]–[19]. This allows the coded aperture be directly adapted from the training data domain and achieves better performance than traditional approaches. However, the training stage requires long processing times and entails high computational costs. Furthermore, to achieve high quality reconstructions, multiple compressed light field captures are often required.

Traditionally, disparity estimation from compressed measurements requires first an algorithm to recover the full light field and then a disparity estimation algorithm to obtain the desired disparity. This two-step methodology degrades the quality of the final estimated disparity maps since the reconstruction errors in the first step are propagated to the second step. Furthermore, each step takes long processing times, and performing them sequentially increase the overall time for disparity estimation. Although, some approaches estimate disparity maps in an unsupervised manner as an implicit step in the reconstruction process of light fields from compressed random measurements [15], [20], the performance is lower than a supervised method and still requires long processing times also related to their multi-step methodology.

In this work, we propose a fast algorithm to directly estimate the disparity maps from compressed light fields based on an E2E approach, omitting the reconstruction process. Specifically, we jointly optimize the attenuation values of the coded mask in the optical architecture proposed by [11] and the parameters of a CNN model that decode the sensor measurement to directly obtain the disparity estimation from the compressed measurements. In addition, we experimentally demonstrate that the proposed method is quantitative and qualitatively comparable to the state-of-the-art methods that use high-quality reconstructed light fields based on deep learning methods. Besides, our method is 20 times faster in

This work was supported by the project Sistema general de regalías-Colombia BPIN 2020000100415 with UIS code 8933.

both training and inference stages than the best comparison method based on deep learning as demonstrated in subsection IV-E.

II. Light Field Compressed Acquisition

To acquire the compressed light field, we consider the optical architecture developed by [11]. The optical configuration consists of a camera with an objective lens at a distance d_a from the sensor and an attenuation coded mask (ACM) located at a short distance d_m in front of the sensor. Specifically, a light field $F(x,y,u,v)$ is modulated by an ACM $\phi(x,y)$, where (x,y) corresponds to spatial dimension and (u,v) corresponds to the angular dimension.The acquired projections after the light field passes through the ACM can be modeled as

$$I(x,y)=\iint \phi(x+\tau(u-x),y+\tau(v-y))F(x,y,u,v)\,du\,dv, \tag{1}$$

where $\tau = d_m/d_a$ is the shear of the ACM pattern associated to the input of the scene. Eq. (1) can be expressed in vector form as

$$\hat{\mathbf{I}} = \mathbf{H}_\Phi \mathbf{f} + \epsilon, \tag{2}$$

where $\hat{\mathbf{I}} \in \mathbb{R}^m$ is the compressed measurement, $\mathbf{f} \in \mathbb{R}^n$ is the vector form of the light field, $\mathbf{H}_\Phi \in \mathbb{R}^{m\times n} = \left[\mathbf{H}_\Phi^{(1)}, \ \mathbf{H}_\Phi^{(2)}, \ \cdots, \ \mathbf{H}_\Phi^{(S)}\right]$ represents the sensing matrix of the compressive light field photography architecture [11], where $\mathbf{H}_\Phi^{(s)} \in \mathbb{R}^{m\times m}$ is a diagonal matrix that represents the modulation for the s-th angular view, and its non-zero values depend on the discrete representation of the ACM $\Phi \in \mathbb{R}^m$. Finally, the compressed measurement of the entire scene can also be viewed as a weighted sum of each modulation of the angular views of the scene as

$$\hat{\mathbf{I}} = \sum_{s=1}^{S} \mathbf{H}_\Phi^{(s)} \mathbf{f}^{(s)} + \epsilon. \tag{3}$$

where each $\mathbf{f}^{(s)}$ contains the s-th angular view and S is the total number of angular views.

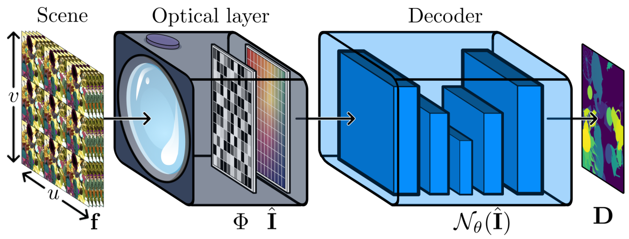

Fig. 1. Proposed E2E method. It is divided into two layers: The optical layer generates the compressed measurements from the spatial and angular modulation of the scenes through the ACM. The decoder is focused on estimating the disparity maps from the compressed measurements.

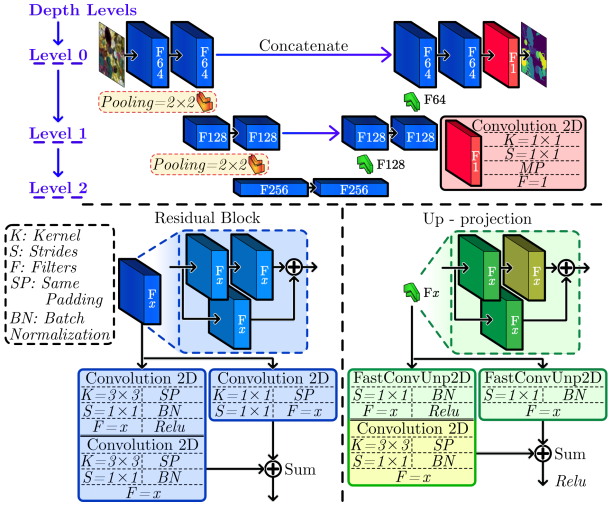

Fig. 2. CNN model for the decoder of the proposed method. Based on wide U-net architecture where each layer of the network consists of residual blocks, MaxPooling, and up-projections.

III. Disparity Estimation from Compressed Light Field

The main idea of the proposed method is to jointly optimize the optical system represented by the ACM $\mathbf{\Phi}$ and a electronic decoder represented by a CNN for a fast disparity estimation. The proposed method is divided in two stages: a training stage that leverages a dataset of light fields and corresponding disparity maps to optimize the ACM and the CNN in a E2E manner; and an inference stage where compressed measurements of testing light fields acquired by the optimized optical system are fed to the trained CNN to estimate the disparity maps (Fig. 1).

A. Training Stage

For the E2E optimization of the optical encoding and electronic decoder, consider a dataset of light fields $\{\mathbf{f}_i\}_{i=1}^{M}$ that are modulated by the ACM to generate compressed measurements $\hat{\mathbf{I}}_i = \mathbf{H}_\Phi \mathbf{f}_i$ according to (3). These measurements are fed to a CNN $\boldsymbol{\mathcal{N}}_\theta$ with parameters θ to generate estimated disparity maps $\hat{\mathbf{D}}_i = \boldsymbol{\mathcal{N}}_\theta(\mathbf{H}_\Phi \mathbf{f}_i)$. Then, to learn the ACM Φ and the CNN parameters θ we optimize a cost function $\mathcal{L}$ to minimize the error between the estimated $\hat{\mathbf{D}}_i$ and ground truth $\mathbf{D}_i$ disparity maps

$$\begin{aligned}\{\Phi^*,\theta^*\} = \quad & \underset{\Phi,\theta}{\text{argmin}} \qquad \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}(\boldsymbol{\mathcal{N}}_\theta(\mathbf{H}_\Phi \mathbf{f}_i),\mathbf{D}_i) \\ & \text{subject to} \quad \Phi_i \in [0,1], \quad 1 \leq i \leq m.\end{aligned} \tag{4}$$

where Φ^* is the optimal ACM and θ^* are the optimal parameters of $\mathcal{N}_\theta$. It is important to highlight that when using a CNN to estimate the disparity, the error given by the loss function $\mathcal{L}$ can be propagated back to the optical system, allowing a joint optimization of the entire model.

B. Inference Stage

Once the training stage is carried out, the optimized optical parameters are employed to simulate compressed measurements of a light field sampled from a given test set $\hat{\mathbf{I}}^* = \mathbf{H}_{\Phi^*}\mathbf{f}$, where Φ^* represents the optimized ACM. Then the estimated disparity map $\hat{\mathbf{D}}$ is computed from the trained CNN $\hat{\mathbf{D}} = \boldsymbol{\mathcal{N}}_{\theta^*}(\hat{\mathbf{I}}^*) = \boldsymbol{\mathcal{N}}_{\theta^*}(\mathbf{H}_{\Phi^*}\mathbf{f})$. We emphasize here that the trained optical parameter can be translated to a physical device and employ the trained CNN to infer disparity maps from real captures [17], [21].

IV. Simulation Results

A. Dataset

The dataset employed in this work was built in [22]. It consists of 500 synthetic scenes of RGB light fields with their respective central disparity map. The dataset is divided into 3 parts: 400 samples for training, 50 samples for validation, and 50 samples for testing. To reduce the computational cost, all samples were angularly and spatially reduced to 7×7 and 256×256, respectively. Each sample was cropped into patches with the same angular resolution and spatial resolution of 32×32, for a total of $32,000$ patches for the entire dataset.

B. Decoder

Inspired by the CNNs [23], [24] based on the U-net architecture [25] with residual blocks [26] and ascending projections [23], we propose the CNN $\mathcal{N}_\theta$ shown in Fig. 2 to estimate the disparity maps from the optimized projections. Furthermore, considering that deep CNNs can suffer the feature reuse problem [27], the proposed CNN is a shallow network with 3 layers and we increase the number of filters. We experimentally find that the proposed network works better for the decoding of light fields than deeper CNNs.

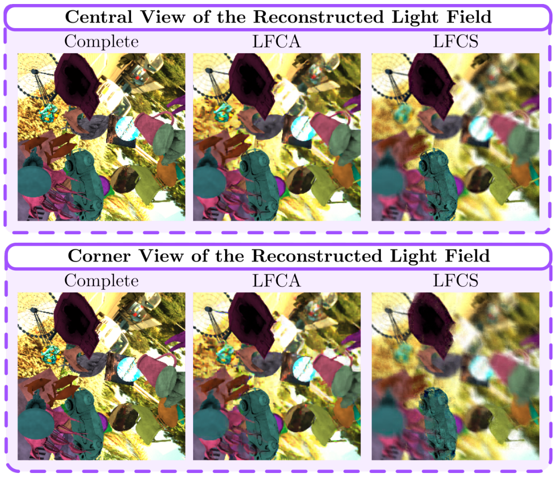
Central View of the Reconstructed Light Field
Complete
LFCA
LFCS
Corner View of the Reconstructed Light Field
Complete
LFCA
LFCS

Fig. 3. Light field reconstructions. Center view sample and view sample located in a corner of the light field with the respective LFCS and LFCA reconstructions where complete is the synthetic sample.

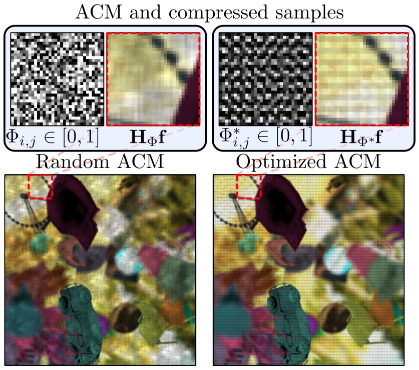
ACM and compressed samples
$\Phi_{i,j} \in [0,1]$
$\mathbf{H}_\Phi \mathbf{f}$
$\Phi^*_{i,j} \in [0,1]$
$\mathbf{H}_{\Phi^*} \mathbf{f}$
Random ACM
Optimized ACM

Fig. 4. ACMs and corresponding compressed measurements. Left: Random ACM employed to run the CNN configuration. Right: E2E optimized ACM.

C. Loss Function and Metrics

To quantitatively evaluate the results obtained, various metrics were used to accurately quantify the pixel-to-pixel differences between the true disparity maps and the predictions. To train and evaluate the proposed method, we employ as loss function $\mathcal{L}$ in (4) the pseudo-Huber loss, allowing to control the outliers between the true disparity maps and their estimations [28]. The other metrics used to evaluate the quality of the predictions are the mean square error (MSE) and mean absolute error (MAE), total variation (TV) and the bad pixel ratio (BadPix) [29].

D. Experimental setup

For the proposed method, we consider two configurations: an E2E configuration where the ACM and the digital decoder weights are jointly learned, and a CNN configuration where the ACM is fixed and random. The random ACM was elaborated using a normal distribution function. It is worth noting that the CNN requires as input the single coded snapshot measurement of the light fields, thus the required processing would be 7×7 times smaller compared to the processing of the full light fields with 7×7 angular resolution and 256×256 spatial resolution. An augmentation was applied to both configurations with the dataset during the training, which consist of random horizontal flip, rotation and gamma stretch applied to the channels of the light fields. For the training of the proposed method (both configurations), 100 epochs were performed, with mini-batches of 16 samples. The root mean square propagation (RMSprop) optimizer with a fixed learning rate of $5e{-}4$ was used to solve the optimization problem in (4).

We compare the proposed method with the strategy that reconstructs the light field and then estimates the disparity map. Two different types of light field reconstructions were considered: a reconstruction based on traditional iterative algorithm (LFCS) [11] and a reconstruction based on deep learning (LFCA) [16], as shown in Fig. 3. We note that the LFCA method employs a different acquisition system and we adapt this method to employ the architecture presented in section II. For disparity estimation, we employ two state-of-the-art

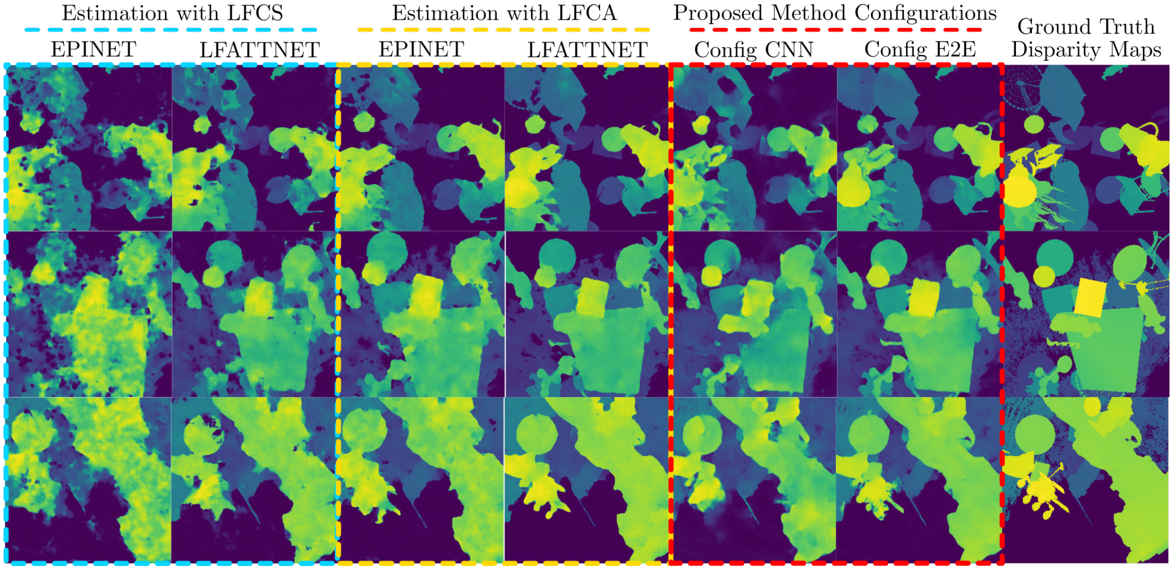

Fig. 5. Qualitative results of the estimation of disparity maps. Dashed cyan and yellow blocks are the estimations by comparison models, and the dashed red block are the estimations by proposed method.

TABLE I
QUANTITATIVE RESULTS OF THE ESTIMATION OF DISPARITY MAPS.

Configurations and Comparison Models	Pseudo Huber	MAE	MSE	BadPix01	BadPix03	BadPix07	Time	
							Training (hours)	Inference (minutes)
Reference	0.00035	0.00488	0.00055	4.6237	3.0980	2.0553	18.5	1.1
Epinet + LFCS	0.00410	0.04541	0.00680	61.1211	44.5674	23.3791	*20.944*	30.16
Epinet + LFCA	*0.00172*	*0.02240*	*0.00291*	*41.0658*	*16.5027*	*7.3672*	104.457	6.78
Lfattnet + LFCS	0.00333	0.03632	0.00580	50.3294	27.7080	14.6844	26.638	30.45
Lfattnet + LFCA	0.00134	**0.01729**	0.00229	**29.4294**	**11.0752**	5.8249	110.512	*7.61*
Configuration CNN	0.00367	0.04580	0.00571	73.8832	43.4794	19.1005	**2.639**	**0.33**
Configuration E2E	**0.00127**	0.01891	**0.00204**	38.3157	13.9681	**5.6851**	5.416	

CNN models, Epinet [30] and Lfattnet [31]. We train these networks from scratch using the reconstructed light fields. The experiments and comparison models were trained with Tensorflow [32] using an NVIDIA GeForce RTX 3090 video card. The LFCS method was trained in MATLAB without GPU usage.

E. Simulation Results

The ACMs of the proposed two configurations are shown in Fig. 4. Following the same practice in [11], the ACM has the same image size but with a periodic structure built with a mask of 32 × 32. It is observed that the optimized ACM learns a semi-uniform grid pattern that results in compressed measurements whit less distortion than the measurements acquired with the random ACM. In this way, the proposed CNN can extract richer features for the disparity estimation task from the designed compressed light field measurements than the non-designed ones.

The quantitative evaluation and qualitative performance of the estimated disparity maps with the proposed method and the comparison methods are reported in the table I and in the Fig. 5, respectively. The table results correspond to the average of the best performance. The "Reference" in the table refers to the highest performance of the **Lfattnet model** trained using the full light field samples. Therefore, it will not be taken into account for comparison. Table I shows that the performance of the proposed method using the E2E configuration is comparable to the Lfattnet model having as input the reconstructed light field from LFCA. The lower performance was achieved with the approaches that employ LFCS due to the low quality of its reconstructions.

The most notable gain of the proposed method is found in the training time, where the E2E configuration takes 195 seconds by epoch. Therefore, for 100 epochs it takes a total time of 5.416 hours. Lfattnet trained with reconstructions from LFCA takes a total of 18.5 hours being 3.4 times slower than the proposed E2E configuration. Furthermore, considering that the light field reconstruction using LFCA took approximately 4 days, the proposed method is 20 times faster in training than the best competitive result obtained using Lfattnet and LFCA. Applying the same analysis to Epinet, the E2E configuration is 2.3 times faster and taking into account the reconstruction using LFCA, it would be 19 times faster.

It can be seen that training with the CNN configuration is the fastest to be performed as expected since it does not require the optimization of the ACM. When analyzing the inference times, it is observed that both configurations of the proposed method are faster in this process than all the comparison methods. Specifically, the E2E configuration manages to be 23 times faster than the best comparison method.

V. Conclusions

A fast method was proposed to estimate disparity from a single compressed measurement of light fields using an E2E approach. The optical coding of the light field photography system is optimized jointly with a CNN for direct disparity estimation from the compressed measurement. The results obtained suggest that the proposed E2E method is comparable with state-of-the-art models for disparity estimation that employ as input reconstructed light fields by compressive sensing or deep learning algorithms. Finally, the proposed method with E2E configuration proved to be 20 times faster in training and inference than the best comparison method that separately learns two deep networks for recovering the light fields and estimating the disparity maps.

References

[1] H. Huang, A. Kuhn, M. Michelini, M. Schmitz, and H. Mayer, "3d urban scene reconstruction and interpretation from multisensor imagery," in *Multimodal Scene Understanding*. Elsevier, 2019, pp. 307–340.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European conference on computer vision*. Springer, 2020, pp. 405–421.

[3] C.-H. Chen and R. Chellappa, "Face recognition using an outdoor camera network," in *Human Recognition in Unconstrained Environments*. Elsevier, 2017, pp. 31–54.

[4] G. De Cubber and D. Doroftei, "Human victim detection and stereo-based terrain traversability analysis for behaviour-based robot navigation," in *Using Robots in Hazardous Environments*. Elsevier, 2011, pp. 476–498.

[5] L. Wu, A. Jaiprakash, A. K. Pandey, D. Fontanarosa, Y. Jonmohamadi, M. Antico, M. Strydom, A. Razjigaev, F. Sasazawa, J. Roberts *et al.*, "Robotic and image-guided knee arthroscopy," in *Handbook of Robotic and Image-Guided Surgery*. Elsevier, 2020, pp. 493–514.

[6] E. H. Adelson, J. R. Bergen *et al.*, *The plenoptic function and the elements of early vision*. Vision and Modeling Group, Media Laboratory, Massachusetts Institute of . . . , 1991, vol. 2.

[7] M. Levoy and P. Hanrahan, "Light field rendering," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 31–42.

[8] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 43–54.

[9] W.-C. Chen, "Light field mapping: Efficient representation of surface light fields," *Energy, simulation-training, ocean engineering, and instrumentation: research papers of the Link Foundation fellows*, p. 89, 2003.

[10] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, 2017.

[11] K. Marwah, G. Wetzstein, Y. Bando, and R. Raskar, "Compressive light field photography using overcomplete dictionaries and optimized projections," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, pp. 1–12, 2013.

[12] Y. Inagaki, Y. Kobayashi, K. Takahashi, T. Fujii, and H. Nagahara, "Learning to capture light fields through a coded aperture camera," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 418–434.

[13] S. Hajisharif, E. Miandji, C. Guillemot, and J. Unger, "Single sensor compressive light field video camera," in *Computer Graphics Forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 463–474.

[14] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, 2008.

[15] A. K. Vadathya, S. Girish, and K. Mitra, "A unified learning-based framework for light field reconstruction from coded projections," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 304–316, 2019.

[16] M. Guo, J. Hou, J. Jin, J. Chen, and L.-P. Chau, "Deep spatial-angular regularization for compressive light field reconstruction over coded apertures," in *European Conference on Computer Vision*. Springer, 2020, pp. 278–294.

[17] E. Vargas, J. N. Martel, G. Wetzstein, and H. Arguello, "Time-multiplexed coded aperture imaging: Learned coded aperture and pixel exposures for compressive imaging systems," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2692–2702.

[18] G. Le Guludec, E. Miandji, and C. Guillemot, "Deep light field acquisition using learned coded mask distributions for color filter array sensors," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 475–488, 2021.

[19] K. Tateishi, K. Sakai, C. Tsutake, K. Takahashi, and T. Fujii, "Factorized modulation for singleshot lightfield acquisition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3253–3257.

[20] A. K. Vadathya, S. Cholleti, G. Ramajayam, V. Kanchana, and K. Mitra, "Learning light field reconstruction from a single coded image," in *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 2017, pp. 328–333.

[21] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.

[22] M. Schambach and M. Heizmann, "A multispectral light field dataset and framework for light field deep learning," *IEEE access*, vol. 8, pp. 193 492–193 502, 2020.

[23] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[24] K. Harsányi, A. Kiss, A. Majdik, and T. Szirányi, "A hybrid cnn approach for single image depth estimation: A case study," in *International Conference on Multimedia and Network Information System*. Springer, 2018, pp. 372–381.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[28] K. Gokcesu and H. Gokcesu, "Generalized huber loss for robust learning and its efficient minimization for a robust statistics," *arXiv preprint arXiv:2108.12627*, 2021.

[29] 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.

[30] 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.

[31] 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, no. 07, 2020, pp. 12 095–12 103.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.