A learning based depth estimation framework for 4D densely and sparsely sampled light fields

Xiaoran Jiang, Jinglei Shi, Christine Guillemot

A LEARNING BASED DEPTH ESTIMATION FRAMEWORK FOR 4D DENSELY AND SPARSELY SAMPLED LIGHT FIELDS

Xiaoran Jiang, Jinglei Shi, Christine Guillemot

INRIA, Rennes, France

ABSTRACT

This paper proposes a learning based solution to disparity (depth) estimation for either densely or sparsely sampled light fields. Disparity between stereo pairs among a sparse subset of anchor views is first estimated by a fine-tuned FlowNet 2.0 network adapted to disparity prediction task. These coarse estimates are fused by exploiting the photo-consistency warping error, and refined by a Multi-view Stereo Refinement Network (MSRNet). The propagation of disparity from anchor viewpoints towards other viewpoints is performed by an occlusion-aware soft 3D reconstruction method. The experiments show that, both for dense and sparse light fields, our algorithm outperforms significantly the state-of-the-art algorithms, especially for subpixel accuracy.

Index Terms— depth estimation, light field, deep learning, multi-view stereo

1. INTRODUCTION

Unlike traditional cameras, light field (LF) devices capture ray intensities emitted by a 3D point along different orientations. Due to this rich description of the 3D scenes, great attention has been given to efficient and robust algorithms for depth estimation exploiting LFs. According to different kinds of images that the estimation depends on, existing algorithms can be classified into several main categories: methods based on Sub-Aperture Images (SAI) [1–3], on Epipolar Plane Images (EPI) [4–7] or on refocused images [8, 9]. However, most of these methods are designed for dense view sampling. Although some algorithms only use a subset of views, e.g. horizontal or crosshair viewpoints, disparity between the exploited views should remain small.

On the contrary, the authors in [10] employ an empirical Bayesian framework, which adapts parameters according to scene statistics. This algorithm is free of additional cues exploiting dense view sampling, e.g. phase shift [2], spinning parallelogram operator [5], defocus cue [8] and structured tensor [4], thus it is robust and relevant for both dense and sparse LFs. In [11], densely sampled depth (disparity) information is inferred from a subset of sparsely sampled light field views. Disparity between sparse views are first estimated and refined using optical flow estimator and edge-preserving filtering, then they are propagated to other viewpoints by exploiting angular correlation.

This work has been funded by the EU H2020 Research and Innovation Programme under grant agreement No 694122 (ERC advanced grant CLIM).

Meanwhile, deep learning has met great success in binocular vision, and similarly in optical flow estimation. As one of the pioneers, FlowNet [12] employs an end-to-end encoder-decoder architecture with additional skip-connections between contracting and expanding parts. Being a variant of FlowNet, DispNet [13] considers 1D correlation instead of 2D correlation to better adapt to disparity estimation task. By stacking several elementary networks, each of them being similar to FlowNet, FlowNet 2.0 [14] significantly improves the prediction accuracy. A cascade framework is proposed in [15], which corrects the disparity initialization by learning in a supervised fashion residual signals across multiple scales. Deep learning has been also successfully applied to light field depth estimation. Among them, EpiNet [7] achieves the state-of-the-art performance by using a multi-stream network, each stream exploiting one angular direction of light field views: horizontal, vertical, left or right diagonal directions. But this approach is well suited for dense light fields only.

In this paper, we focus on how to handle either a dense or a sparse light field for disparity/depth estimation. Similar to [11], our algorithm only exploits a sparse subset of anchor views (four corner views), and generates one disparity map for every viewpoint of the light field. Multi-view stereo (MVS) is implemented in a deep learning based cascaded framework. A pre-trained FlowNet 2.0 is fine-tuned by pairs of stereo images, and the obtained model is used to estimate disparity between pairs of anchor views, arranged horizontally or vertically. These coarse estimates are then fused at each anchor viewpoint by exploiting the warping error from other anchor viewpoints, and then refined by a second convolutional neural network (CNN), which we call Multi-view Stereo Refinement Network (MSRNet). For better subpixel accuracy of the disparity values, views are upsampled before being fed to CNNs. Correspondingly, the output disparity maps are rescaled. The propagation of disparity from anchor viewpoints towards other viewpoints is performed by an occlusion-aware soft 3D reconstruction method.

2. LEARNING-BASED DISPARITY ESTIMATION

2.1. Finetuned FlowNet 2.0 for stereo

We take the four corner views $I_{1,1}, I_{1,V}, I_{U,1}$ and $I_{U,V}$ (c.f. Fig. 1) as the set of anchor light field views. Arguably, these distinct views on the extreme corners of a densely sampled light field contain all color and geometric information, from which the whole light field can be reconstructed [16].

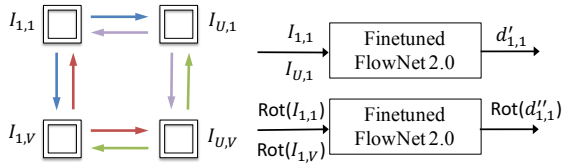

Fig. 1: Disparity estimation for LF anchor views by using the finetuned FlowNet 2.0. Vertical pairs are rotated 90 degrees to become horizontal pairs. An example is given on the right for the view $I_{1,1}$.

Disparity is estimated between pairs of anchor views. To work within the stereo vision framework, we assume that the light field is well rectified, i.e., a scene point moves only horizontally from position $(1,1)$ to $(U,1)$, and only vertically from position $(1,1)$ to $(1,V)$. In order to apply a single network to deal with both horizontal and vertical image pairs, vertical pairs are rotated 90 degrees. During the training process, their corresponding ground truth disparity maps are also rotated 90 degrees. For each anchor view I_r, two coarse estimates d'_r and d''_r are obtained, respectively using horizontal and vertical pairs.

2.2. Multi-view disparity refinement

In low-level vision tasks, such as segmentation, denoising and optical flow estimation, results can be generally improved with post-facto refinement. Similar to [15], our Multi-view Stereo Refinement Network (MSRNet) takes as input the coarse disparity estimates from a first stage network (finetuned FlowNet 2.0), and the refinement of these maps is implemented by a multi-scale encoder-decoder structure. Differently, our scheme generalizes the binocular vision refinement task to MVS scenario, and our network is flexible for any number of initialized disparity maps.

2.2.1. Disparity fusion

In order to fuse multiple coarse disparity estimates to a single one, we propose to leverage photo-consistency warping errors. $I^i_r(d'_r)$ denotes the warped image from position i to r (both i and r are corner positions) by using disparity d'_r. Traditionally, warping errors from different corner viewpoints are aggregated by computing the sum or average of them:

$$E_{\text{nocc}}(d'_r) = \text{mean}_i\Big(\mathcal{E}(I_r, I^i_r(d'_r))\Big), \tag{1}$$

with $\mathcal{E}(I, I')$ the pixel-wise sum of square errors for the three color components between image I and I':

$$\mathcal{E}(I, I') = \sum_{C\in\{R,G,B\}} (I_C - I'_C)^2. \tag{2}$$

The term E_{nocc} measures accurately the pixel warping error in occlusion-free areas, and therefore reflects well the disparity accuracy at these pixels, but fails in occlusion zones. In fact, a high error in occlusion zones is due to interpolation in large holes, rather than to disparity inaccuracy.

Compared to stereo vision, multi-view stereo should provide a better modeling for occlusion, since a pixel occluded from one viewpoint may be viewed from another viewpoint. Therefore, we define a second error map E_{occ}:

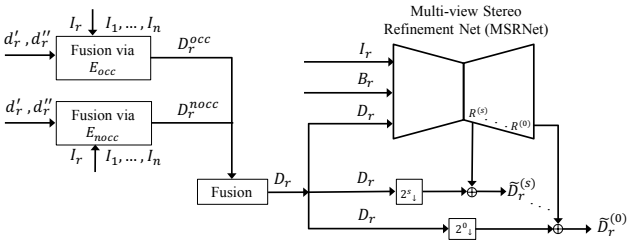

Fig. 2: Network structure for light field multi-view disparity refinement. Given the coarse disparity maps d'_r and d''_r computed using image pairs, the fusion of these maps is performed by exploiting the warping errors E_{occ} and E_{nocc}. The refinement of the fused disparity map D_r is learned with multi-scale residual learning.

$$E_{\text{occ}}(d'_r) = \min_i\Big(\mathcal{E}(I_r, I^i_r(d'_r))\Big). \tag{3}$$

In general, error due to interpolation is smaller for non-occluded pixels than occluded ones. At a pixel p that can be seen in the warped view $I^{i'}_r$, but not in I^i_r for any other corner viewpoint i, the error $E_{\text{occ}}(d'_r, p)$ may equals to the value $\mathcal{E}(I_r, I^{i'}_r(d'_r))$ at pixel p, which more faithfully reflects the disparity accuracy for this pixel.

Hence, disparity inaccuracy is better modeled by E_{nocc} in occlusion-free areas, and by E_{occ} in non-overlapped occlusion zones. Two fused disparity maps can be obtained: at each pixel p, one disparity value is selected among the candidate values $d'_r(p)$ and $d''_r(p)$ by minimizing the error term E_{nocc} or E_{occ}, respectively:

$$D^{\text{nocc}}_r(p) = \underset{d'_r(p), d''_r(p)}{\text{argmin}}\Big(E_{\text{nocc}}(d'_r, p), E_{\text{nocc}}(d''_r, p)\Big), \tag{4}$$

and

$$D^{\text{occ}}_r(p) = \underset{d'_r(p), d''_r(p)}{\text{argmin}}\Big(E_{\text{occ}}(d'_r, p), E_{\text{occ}}(d''_r, p)\Big). \tag{5}$$

One may also approximate the binary occlusion mask M by identifying the pixels that correspond to high values of E_{nocc}:

$$M(p) = \begin{cases} 1, & \text{if } \min\Big(E_{\text{nocc}}(d'_r, p), E_{\text{nocc}}(d''_r, p)\Big) > \theta \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

the threshold θ being chosen as the top 90% E_{nocc} value. Compared to [11] where the occlusion mask is explicitly computed based on disparity values, Eq. (6) shows to be a good approximation and above all more time efficient, which is compatible to the learning-based framework. Finally, the unique disparity value per pixel at corner viewpoint r is computed as:

$$D_r(p) = D^{\text{nocc}}_r(p) \times \Big(1 - M(p)\Big) + D^{\text{occ}}_r(p) \times M(p). \tag{7}$$

2.2.2. MSRNet

A cascade residual learning framework is proposed in [15] to refine disparity for binocular vision: residuals of disparity are learned explicitly by an encoder-decoder multi-scale network, which is able to correct the disparity initialized by a first stage CNN. Residual signals are supervised at each resolution scale.

The input of this refinement network is a set of 5 images: the left image I_L, the right image I_R, the initialized disparity map, warped image $\tilde{I}_L$ and warping error. In a MVS scenario, it is obvious that this scheme is no longer applicable: the multiplication of the number of stereo pairs, as well as the number of initialized disparity maps, will rapidly enlarge the size of the network and make the learning inefficient. Moreover, this scheme cannot easily adapt to the varying number of input images.

This is why we chose to fuse the horizontal and vertical disparity estimates (c.f. Section 2.2.1) before residual learning. Regardless of potentially different numbers of stereo pairs, only one fused disparity map D_r will be obtained at corner position r. Two other images are also fed to MSRNet: the color image I_r as guidance, and a binary map B indicating unreliable pixels. These pixels are supposed to be located near object contours which are determined using canny edge detectors. We detect the contours on the disparity map D_r that do not correspond to those on the color image I_r. The binary mask is therefore computed as $B = |\mathcal{D} \circ \mathcal{C}(D_r) - \mathcal{D} \circ \mathcal{C}(D_r) \odot \mathcal{D} \circ \mathcal{C}(I_r)|$. $\mathcal{D}$ and $\mathcal{C}$ are respectively dilation and canny edge detection operators. The symbol $\circ$ denotes function composition and $\odot$ is for Hadamard product.

The rest of the network structure is similar to the one of [15]. Compared to [15], we add a gradient term $\mathcal{L}_2$ in the loss function, which leads to smoother maps:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1(D, D_{gt}) + \lambda_2 \mathcal{L}_2(D, D_{gt}) \tag{8}$$

where $\mathcal{L}_1$ is sum of absolute differences (SAD), and the term $\mathcal{L}_2$ is defined as the sum of l^2-norm $\mathcal{L}_2(D, D_{gt}) = \sum_p \|G(p)\|_2$ with

$$G(p) = \Big(\nabla_x D(p) - \nabla_x D_{gt}(p), \nabla_y D(p) - \nabla_y D_{gt}(p)\Big)^\top. \tag{9}$$

3. DISPARITY PROPAGATION

The obtained disparity maps $\tilde{D}_{1,1}, \tilde{D}_{1,V}, \tilde{D}_{U,1}$ and $\tilde{D}_{U,V}$ at the 4 corners should be propagated to other viewpoints where we assume that the color information is absent. In [11], disparity values are projected to novel view positions, and low rank assumption of the matrix containing all warped maps is exploited to inpaint missing disparity values for occluded pixels. However, for light fields with wide baseline, this low rank assumption cannot be held anymore.

Authors in [17] have proposed a soft 3D reconstruction method for light field view synthesis. We apply this idea to disparity propagation. For each corner view, a consensus volume is constructed, each voxel encoding the consensus score of all voters (4 corner view disparity estimates) for the existence of a surface with a certain discretized disparity/depth value at a pixel p. To compute the consensus volume, a vote volume and a confidence volume are previously constructed, half of the voters with low confidence values abstain to vote. Occlusion/visibility is modeled by the soft visibility function: a pixel can be viewed at a depth z only if the consensus for a surface existing at any depth $z' < z$ is low (note that the disparity is inversely proportional to the depth). These soft visibility values at corner viewpoints are then used as weights for merging warped disparity maps in other viewpoints.

4. EXPERIMENTS

4.1. Datasets

In order to train our deep neural networks, two synthetic LF datasets with different disparity ranges have been created using Blender software [18]. 3D models, available under CC0 or CC license, have been downloaded from Chocofur [19] and Sketchfab websites [20]. The sparse light field (SLF) dataset contains scenes with disparity range $[-20, 20]$ between adjacent views, whereas the disparity range is $[-4, 4]$ for the dense light field (DLF) dataset. Each LF has the same resolution $512 \times 512 \times 9 \times 9$. We provide for each view of the LFs, a color view, a ground truth depth map and its corresponding disparity map.

Among the 53 scenes in the SLF dataset, 44 scenes are served as training data, and 9 others as valid data. For the DLF dataset, the training set contains 38 scenes and the valid set contains 5 scenes. To our knowledge, they are among the first synthetic light field datasets providing depth information for every viewpoint of the light fields (HCI 4D light field dataset [21] is only available with light fields with narrow baseline), and the size of the datasets is sufficient for training a deep neural network for light field disparity/depth estimation. Our datasets will be available online upon paper acceptance.

4.2. Training

Training stereo pairs are randomly selected among pairs of LF views in the same row or the same column. For dense LFs, the angular distance between these two stereo views is between 2 and 8 view indices. The same distance is chosen between 1 and 3 for sparse light fields. In order to avoid imbalanced data distribution, we have managed to keep equal the occurrence of training examples for each different distance. The model is first trained with data in SLF dataset. Then, for dense light fields, the model is further finetuned using the DLF dataset. We also include the 16 additional scenes (different from those evaluated in Table 1) from the HCI 4D light field dataset to enrich our DLF training set. Due to limited GPU memory, a batch size of 4 is adopted.

Data augmentation. Chromatic augmentation (variation of contrast, color and brightness) is applied with the same parameters as suggested in [12]. However, we do not apply any geometrical transformation, e.g., translation, rotation and scaling. In our experiments, it is observed that geometrical transformation implies interpolation errors in the transformed ground truth disparity maps, which harms the learning convergence.

Learning rate schedule. To finetune FlowNet 2.0, the initial learning rate is set to 0.0001 for the first 500 epochs, whereas this learning rate is applied for the first 1200 epochs to train the MSRNet. Then, for both training tasks, the learning rate is decreased by half after every 200 epochs for better convergence.

4.3. Dense light fields

We assess our algorithm against several state-of-the-art methods [2, 5, 10, 11] with 8 test dense LFs in the HCI dataset. The methods [2] and [5] make use of the full 4D LF. The method [10] considers a sparse subset of 3×3 views including the center view, whereas [11] and our method only exploit the 4 corner views.

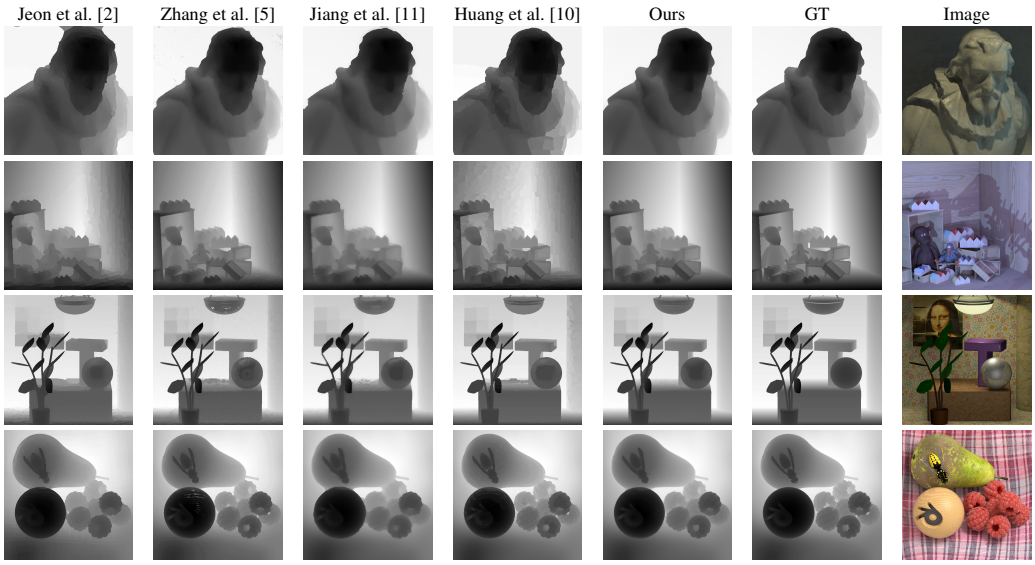

Fig. 3: Visual comparison of the estimated disparity maps on center view.

Table 1: Quality evaluation of the estimated disparity maps on center view for dense LFs. The best results are marked in **bold**.

	MSE*100					BadPix(0.01)					BadPix(0.03)					Q25				
Light fields	[2]	[5]	[11]	[10]	Ours	[2]	[5]	[11]	[10]	Ours	[2]	[5]	[11]	[10]	Ours	[2]	[5]	[11]	[10]	Ours
StillLife	2.02	1.72	2.56	1.16	**1.14**	81.2	76.2	**71.3**	74.4	71.5	51.0	32.1	25.0	37.1	**24.5**	1.36	1.02	0.87	**0.86**	0.88
Buddha	1.13	0.97	0.82	**0.40**	0.46	57.7	41.2	34.9	51.3	**25.8**	24.4	14.8	12.3	13.4	**6.6**	0.51	0.34	0.31	0.52	**0.28**
MonasRoom	0.76	0.58	0.53	0.56	**0.38**	46.0	42.5	38.6	45.5	**25.2**	22.1	17.8	18.6	17.8	**11.4**	0.38	0.34	0.33	0.35	**0.24**
Butterfly	4.79	0.74	1.84	0.70	**0.54**	82.5	78.9	70.8	82.4	**62.9**	49.1	48.5	36.0	50.8	**28.7**	1.47	1.22	0.85	1.28	**0.66**
Boxes	14.15	**8.23**	12.71	10.05	12.48	72.7	62.3	65.8	83.6	**60.5**	45.5	**28.1**	37.7	57.1	32.8	0.89	0.62	0.68	1.54	**0.55**
Cotton	9.98	1.44	1.18	1.23	**0.67**	60.5	41.7	42.6	72.1	**29.6**	23.3	11.1	10.7	33.7	**8.0**	0.59	0.36	0.42	0.89	**0.25**
Dino	1.23	**0.29**	0.88	0.53	0.50	76.6	57.5	49.1	80.9	**35.9**	48.4	17.9	20.0	48.0	**12.6**	1.08	0.55	1.32	0.42	**0.29**
Sideboard	4.16	**0.92**	10.31	1.31	1.60	67.8	64.3	61.7	79.8	**48.8**	39.3	31.0	37.5	46.4	**23.2**	0.74	0.66	1.26	0.51	**0.37**
Average	4.78	**1.86**	3.85	1.99	2.22	68.1	58.1	54.4	71.2	**45.0**	37.9	25.2	24.7	38.0	**18.5**	0.88	0.64	0.62	0.80	**0.44**

Table 2: Quality evaluation of the estimated disparity maps on center view for sparse LFs.

	MSE			BadPix(0.1)			Q25		
Light fields	[11]	[10]	Ours	[11]	[10]	Ours	[11]	[10]	Ours
Furniture	1.94	**0.38**	0.78	41.3	61.3	**22.0**	2.52	6.17	**1.10**
Lion	0.87	**0.08**	0.15	59.5	21.4	**8.0**	4.47	2.51	**0.61**
Toy_bricks	1.10	**0.18**	0.44	44.6	36.0	**16.6**	3.61	2.72	**0.94**
Electro_devices	0.63	**0.18**	0.23	43.4	55.5	**24.5**	2.71	4.93	**1.35**
Average	1.14	**0.21**	0.40	47.2	43.6	**17.8**	3.33	4.08	**1.00**

We use the same evaluation metrics defined in [21, 22]. MSE is mean-square-error, which penalizes large disparity errors on the object boundary, whereas BadPix(α) (the percentage of pixels having an error superior to α, α being set to small values) and Q25 (the error value *100 at the 25th percentile of the disparity estimates) measure the sub-pixel accuracy. Table 1 shows that in terms of MSE, our method is on par with [5] and [10], and better than other reference methods. In terms of BadPix(0.01), BadPix(0.03) and Q25, our method outperforms all the reference methods by a large margin. Moreover, unlike [2, 5, 10], the fact that our algorithm generates one disparity map per viewpoint, without exploiting the color information of the views other than the four corner views, is especially interesting for applications such as light field view synthesis.

4.4. Sparse light fields

Our algorithm has been also evaluated on 4 sparse LFs in our SLF dataset. An angular resolution of 3×3 is considered. In this case, methods [2] and [5] are no longer relevant for comparison, since they only rely on densely sampled views, and their performance drops drastically when the baseline increases. In Table 2, our algorithm performs significantly better than [11] and [10] in subpixel accuracy (BadPix and Q25), whereas the method [10] excels in MSE. Note that the loss function (Eq. (8)) computes SAD instead of MSE. For the test LFs, our algorithm obtains better SAD measures against [10].

5. CONCLUSIONS

In this paper, we have proposed a learning-based depth estimation solution both for densely and sparsely sampled light field data. The experiments show that our algorithm outperforms state-of-the-art algorithms by a large margin in most of the metrics. Our algorithm can be also naturally integrated into a light field view synthesis pipeline, since it is able to infer disparity information for a view that the color information is unknown.

6. REFERENCES

[1] Stefan Heber and Thomas Pock, "Shape from light field meets robust PCA," in *European Conference on Computer Vision (ECCV)*, 2014.

[2] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon, "Accurate depth map estimation from a lenslet light field camera," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Yang Chen, Martin Alain, and Aljosa Smolic, "Fast and accurate optical flow based depth map estimation from light fields," in *Irish Machine Vision and Image Processing Conference (IMVIP)*, 2017.

[4] Sven Wanner and Bastian Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions of Pattern analysis and machine intelligence*, vol. 36, no. 3, 2013.

[5] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Journal Computer Vision and Image Understanding*, vol. 145, pp. 148–159, 2016.

[6] Ole Johannsen, Antonin Sulc, and Bastian Goldluecke, "What sparse light field coding reveals about scene structure," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim, "Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] Michael W. Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi, "Depth from combining defocus and correspondence using light-field cameras," in *International Conference on Computer Vision (ICCV)*, 2013.

[9] Ting-Chun Wang, Alexei Efros, and Ravi Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *International Conference on Computer Vision (ICCV)*, 2015.

[10] Chao-Tsung Huang, "Empirical bayesian light-field stereo matching by robust pseudo random field modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–1, 2018.

[11] Xiaoran Jiang, Mikael Le Pendu, and Christine Guillemot, "Depth estimation with occlusion handling from a sparse set of light field views," in *IEEE International Conference on Image Processing (ICIP)*, 2018.

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, and Vladimir Golkov, "Flownet: Learning optical flow with convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Nikolaus Mayer, Eddy Ilg, Philip Hausser, and Philipp Fischer, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *ICCV Workshop on Geometry Meets Deep Learning*, Oct 2017.

[16] Xiaoran Jiang, Mikael Le Pendu, and Christine Guillemot, "Light field compression using depth image based view synthesis," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2017.

[17] Eric Penner and Li Zhang, "Soft 3d reconstruction for view synthesis," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2017.

[18] "Blender software," `https://www.blender.org/`.

[19] "Chocofur 3d models website," `http://www.chocofur.com/`.

[20] "Sketchfab 3d models website," `http://www.sketchfab.com/`.

[21] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision (ACCV)*, 2016.

[22] Ole Johannsen, Katrin Honauer, Bastian Goldluecke, Anna Alperovich, Federica Battisti, Yunsu Bok, Michele Brizzi, Marco Carli, Gyeongmin Choe, Maximilian Diebold, Marcel Gutsche, Hae-Gon Jeon, In So Kweon, Alessandro Neri, Jaesik Park, Jinsun Park, Hendrik Schilling, Hao Sheng, Lipeng Si, Michael Strecke, Antonin Sulc, Yu-Wing Tai, Qing Wang, Ting-Chun Wang, Sven Wanner, Zhang Xiong, Jingyi Yu, Shuo Zhang, and Hao Zhu, "A taxonomy and evaluation of dense light field depth estimation algorithms," in *Conference on Computer Vision and Pattern Recognition - LF4CV Workshop*, 2017.