Benchmark Data Set and Method for Depth Estimation From Light Field Images

Mingtao Feng, Yaonan Wang, Jian Liu, Liang Zhang, Hasan F. M. Zaki, and Ajmal Mian

***Abstract*— Convolutional neural networks (CNNs) have performed extremely well for many image analysis tasks. However, supervised training of deep CNN architectures requires huge amounts of labeled data, which is unavailable for light field images. In this paper, we leverage on synthetic light field images and propose a two-stream CNN network that learns to estimate the disparities of multiple correlated neighborhood pixels from their epipolar plane images (EPIs). Since the EPIs are unrelated except at their intersection, a two-stream network is proposed to learn convolution weights individually for the EPIs and then combine the outputs of the two streams for disparity estimation. The CNN estimated disparity map is then refined using the central RGB light field image as a prior in a variational technique. We also propose a new real world data set comprising light field images of 19 objects captured with the Lytro Illum camera in outdoor scenes and their corresponding 3D pointclouds, as ground truth, captured with the 3dMD scanner. This data set will be made public to allow more precise 3D pointcloud level comparison of algorithms in the future which is currently not possible. Experiments on the synthetic and real world data sets show that our algorithm outperforms existing state of the art for depth estimation from light field images.**

***Index Terms*— Depth estimation, disparity, light field, two stream CNN, deep learning, plenoptic camera, Lytro camera.**

I. Introduction

LIGHT field imaging, introduced by Adelson and Bergen [1], not only captures the color intensities at each pixel but also the directions of incoming light rays. There are many devices for capturing light field images such as camera arrays [2] or a gantry consisting of a single moving camera [3]. Camera arrays are hardware-intensive and need a sophisticated calibration process. They are also expensive and difficult to build and are, therefore, not widely used. The gantry is less expensive and simpler but limited to static scenes. Recently, specially designed plenoptic cameras such as the Lytro [4] and Raytrix [5] have become commercially available. Plenoptic cameras use an array of microlenses to capture many sub-aperture images arranged in an equally spaced rectangular grid. These cameras make it possible to acquire a large number of light field images of indoor or outdoor scenes in a single photographic exposure. The sub-aperture images have been exploited to improve the performance of many applications, such as digital refocusing [6], image segmentation [7], material classification [8], saliency detection [9], view synthesis [10] and in particular depth estimation.

Depth estimation from a light field image has been a challenging and active research problem for the last few years. Since a plenoptic camera provides angular as well as spatial information, the angular and spatial resolutions of the captured images are limited by the hardware. The limited angular resolution combined with large depth range, large number of views and different types of noises in the real environment [11] make it difficult to estimate a dense and accurate depth map from a light field image.

Light field image based depth estimation methods can be broadly divided into those which employ Epipolar Plane Images (EPI) and those which do not employ EPI and use the sub-aperture images [12]. Classical methods do not use EPIs and extend multiview [13] and stereo algorithms [14] on the sub-aperture images. However, treating the sub-aperture images as multiple stereo pairs is not optimal because the sub-aperture images have an extremely short baseline and contain optical distortions. Moreover, stereo and multiview methods do not exploit the pattern obtained by placing the micro-lenses in an equally spaced rectangular grid. This pattern is exploited by the methods that use EPIs and e.g. [11], [12].

Convolutional Neural Networks (CNN) have obtained state-of-the-art results for object recognition [15]–[18], human action recognition [19], image segmentation [20] and stereo depth estimation [21]. However, CNNs have not been fully explored for depth estimation from light field images because of the complexity of directly feeding the 4D RGB light field images into CNNs and the lack of sufficient training data. Kalantari *et al.* [10] used two sequential CNNs to synthesize novel views from only a few of the light field sub-aperture images; generating a large number of combinations as training data. The first CNN provides rough disparity estimates to the second which performs color refinement. Heber and

Manuscript received July 11, 2017; revised January 4, 2018; accepted February 13, 2018. Date of publication March 9, 2018; date of current version April 20, 2018. This work was supported in part by the China Scholarship Council and National Natural Science Foundation of China under Grant 61573134, Grant 61401046, and Grant 61573135 and in part by the Australian Research Council under Grant DP160101458. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yonggang Shi. *(Corresponding author: Ajmal Mian.)*

M. Feng and Y. Wang are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: mintfeng@hnu.edu.cn; yaonan@hnu.edu.cn).

J. Liu, H. F. M. Zaki, and A. Mian are with the School of Computer Science and Software Engineering, The University of Western Australia, Perth, Crawley, WA 6009, Australia (e-mail: 21884024@uwa.edu.au; hasan.mohdzaki@uwa.edu.au; ajmal.mian@uwa.edu.au).

L. Zhang is with the School of Software, Xidian University, Xi'an 710071, China (e-mail: liangzhang@xidian.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes figures, tables, and datasets relevant to the main paper. The total size of the file is 21.2 MB. Contact ajmal.mian@uwa.edu.au for further questions about this work.

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

Digital Object Identifier 10.1109/TIP.2018.2814217

Pock [22] use a CNN to estimate the disparity at a single pixel from the stacked EPIs corresponding to the pixel.

In this paper, we propose a deep learning method that extends the ideas of [22] and [23]. We propose a two stream CNN that learns to estimate the disparity values of a small group of local pixels from their corresponding horizontal and vertical EPIs. Estimating the disparity of a small group of pixels, rather than a single one, exploits the correlation between the pixels to get multiple accurate and robust estimates. A two stream CNN architecture is preferred so that the convolution weights are learned individually for the horizontal and vertical EPIs since they are not related except at their point of intersection which is a single pixel only. The CNN is trained using synthetic light field images for which the ground truth disparities are available [24].

Once the CNN disparity estimates are obtained, we use variational techniques to eliminate inaccurate estimations in textureless regions. For this, we extend the depth reconstruction method of Liu *et al.* [25] which uses the sparseness of depth maps as a prior to refine depth images obtained from low resolution depth sensors as opposed to light field cameras. The main difference is that we use the light field central RGB image as a prior to guide the refinement process. In particular, we use the RGB edge based weight matrix as confidence measures for a new fidelity term to gauge the reliability of the estimation. The proposed method also uses a nonlocal mean prior with the RGB image structure and a flattening regularization to make the final disparity more smooth. The final disparity maps are converted into metric depth maps using the camera calibration parameters and subsequently to pointclouds by mapping the metric depth maps to the 3D space for comparison with ground truth 3D pointclouds.

A major bottleneck of depth estimation research is the unavailability of real data with ground truth i.e. light field images of real objects and the their corresponding 3D pointclouds. We fill this gap and propose a new dataset of 19 objects which are scanned by the 3dMD scanner [26] to get their dense 180° 3D pointclouds. We also provide Lytro Illum camera [4] images of these objects in outdoor scenes and lighting. Our dataset will be made public allowing researchers to compare the pointclouds estimated by various methods from light field images to the pointclouds of the objects obtained from the 3D scanner which can be considered as ground truth.

To summarize, our contributions: (1) We propose a two stream CNN architecture for disparity estimation from EPIs and train it from scratch using synthetic data. We show with extensive experiments that this network generalizes well to real data. (2) We propose a post refinement method based on the central RGB image and its edge map. (3) We propose the first light field image dataset that is acquired in outdoor environments and is accompanied with ground truth 3D pointclouds. We show that our algorithm outperforms existing state-of-the-art methods quantitatively and qualitatively on synthetic and real data. More importantly, we compare the pointclouds obtained from different methods on our real world dataset and show that this approach is good for precise evaluation.

II. Related Work

Multiview stereo methods for depth estimation have a longer history and [13], [14], [21], [27]–[35]. Some multi-view stereo methods, like Goldlücke *et al.* [28], use one of the multiview images as reference and compute the disparity of all remaining views from it. Liu *et al.* [29] integrated silhouette information and epipolar constraint into the variational method for continuous depth map estimation. Mayer *et al.* [33] synthesized three stereo video datasets to train a network for real time disparity estimation. Zbontar and LeCun [34] used CNN to compute the stereo matching cost, followed by a refinement process and a left-right consistency check. Zagoruyko and Komodakis [35] proposed a CNN-based model to represent a general similarity function. A central-surround two stream network was proposed to utilize multi-resolution information for image patches comparison. Luo *et al.* [21] employed a siamese network by treating stereo matching problem as multi-class classification. The proposed siamese network classifies all pixels to their possible disparity labels.

Since the introduction of light field cameras, many methods have been proposed for light field image based depth estimation. Early methods extended multi-view and stereo algorithms to estimate depth from light field images. For example, Jeon *et al.* [14] corrected the sub-aperture images and estimated the multiview stereo correspondences with sub-pixel accuracy using cost volume calculated from the sum of absolute color and gradient differences. Navarro and Buades [13] proposed a method that exploits multiple two-view stereo pairs in light filed images and estimates multiple sparse disparity maps which are interpolated to obtain a single robust disparity map.

Treating the sub-aperture images of a light field image as multiple stereo pairs or multi-view images, has its limitations. Firstly, compared with conventional multi-view images, light field images have an extremely short baseline. Secondly, the images from light field cameras contain optical distortions that are caused by both the main lens and micro-lenses. Consequently, conventional stereo and multiview approaches based on correspondence matching do not typically work well as the sub-pixel shift in the spatial domain is subtle and the sub-aperture image resolution is low.

It is well known that depth can also be estimated from a stack of images focused at different distances to the camera [36], [37]. Tao *et al.* [38] sheared the light field image to perform refocusing and combined depth cues from both defocus and correspondence into a final global depth estimate. Tao *et al.* [39] extended their work in [38] by adding the shading cues and used it to refine fine details of the estimated shape. Wang *et al.* [40] proposed an occlusion prediction framework using a modified angular photo-consistency. The recent method by Williem and Kyu Park [41] proposed novel data costs based on the angular entropy metric and adaptive defocus response to improve the occlusion and noise handling capabilities. Strecke *et al.* [37] constructed cost volumes based on focal stack symmetry, followed by a joint regularization of depth and normals as post-processing step. All these methods discretize the depth values which has

the advantages of removing noise and better handling of occlusions. However, the downside is that they sacrifice the depth resolution resulting in discontinuous depth values.

The sub-aperture light field images are arranged in a more structured way compared to multiview images. This structure is visually apparent in the EPI (Epipolar Plane Image). Some methods exploit this structure and estimate the disparity image from EPI. For example, Wanner and Goldluecke [42] used the 2D structure tensor to measure the EPI texture direction corresponding to each pixel in the vertical and horizontal EPIs. They combined the two direction estimates to measure a reliability map for the final optimization framework. Zhang *et al.* [11] proposed a spinning parallelogram operator (SPO) dividing the regions in an EPI, the lines that indicate depth information are located by maximizing the distribution distances of the regions. Zhang *et al.* [12] minimized a matching cost that aggregates intensity pixel value, gradient pixel value, spatial consistency as well as reliability measures to select the optimal line slope from a redefined set of directions. However for real world data, the outputs of [11] and [12] are noisy because low quality sub-aperture images affect the accuracy of line slope estimation. Li *et al.* [43] proposed an optimization method based on iteratively solving a sparse linear system with a conjugate gradient method starting from an initial disparity estimate to give a final disparity estimate.

Recently, deep learning techniques have been used for light field image processing. Heber and Pock [22] proposed a method for estimating shape from light field images that applies a CNN on the horizontal and vertical EPIs corresponding to a single pixel. They take local patches from the EPIs and stack them to make a 6 channel (RGBRGB) epipolar image. A CNN is then learned using labelled training data to regress the 6 channel image on to the single disparity value at the center pixel of the EPI patches. Regression allows the estimation of continuous depth values as opposed to discrete values in [14] and [40]. The learned CNN is used to estimate the pixel-wise disparities of new data and the disparity image is refined using higher order regularization. The limitations of this method [22] are that it stacks the *unrelated* horizontal and vertical EPIs to learn common weights for them in the single stream CNN, and it does not exploit correlations between neighboring pixels. Heber *et al.* [44] proposed a u-shaped networks to predict the depth information from EPI volumes. The u-shaped network uses 3D convolution layers to propagate information from a group of RGB sub-views to disparity volumes. However, training the u-shaped network requires disparity maps of all the sub-views as label, which is unrealistic for existing light field datasets.

In contrast to [22] and [23], we propose a two stream CNN where each stream learns features individually from the vertical or horizontal EPIs. The features are then combined at the fully connected layers for regression over the disparity values of a neighbourhood if pixels rather than a single pixel. Sliding a cross shaped window over the image, each disparity value is estimated multiple times. This allows the network to implicitly learn neighbourhood correlations in disparities and facilitates better disparity estimates through averaging. To account for textureless regions, the central RGB image and its edge map are used as priors in a variational framework to refine the network output disparity image.

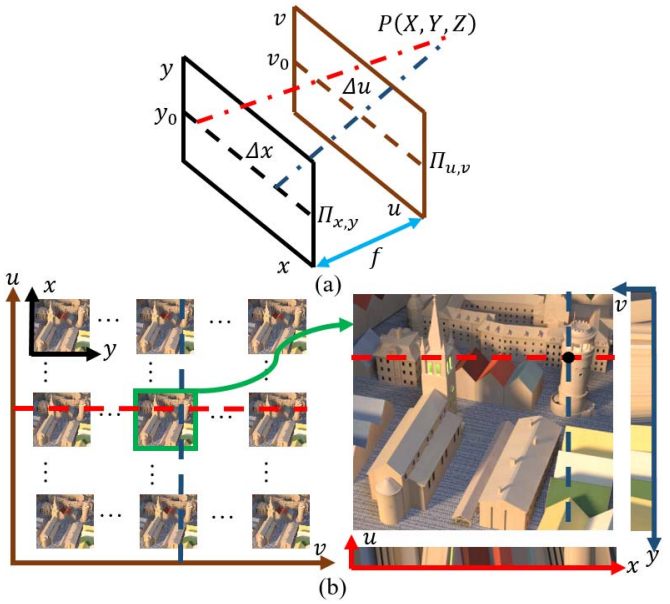


Fig. 1. (a) Two-plane parametrization of 4D light field. (b) 2D slices through 4D light field : the light field sub-aperture images are assumed to be arranged in an equally spaced rectangular grid. Horizontal and vertical EPIs are obtained between the central view and those that are in the same row and column, i.e. the red and blue lines intersect at one pixel (black dot) in the central view. In theory, the slope of black dot in two EPIs is the same.

III. 4D Light Field Structure

We use the two-plane parametrization introduced in the seminal work by [45] to represent light field images as shown in Figure 1-(a). The 2D plane $\Pi_{x,y}$ is the camera plane and $\Pi_{u,v}$ is the image plane. It is enlightening to think of light field as a collection of images of the same scene, taken by several cameras at different positions parallel to $\Pi_{x,y}$. In a discretely sampled light field, (u, v) can be regarded as a pixel in an image and (x, y) can be regarded as the position of the camera in the grid of cameras. Specifically, each ray is decoded by its intersections with these two parallel planes, that is $L(x, y, u, v)$. To visualize the position changes in $\Pi_{u,v}$ due to changing camera position, we draw out the horizontal line of constant camera coordinate y_0 and a constant v_0 in the image plane, resulting in a map called a horizontal Epipolar Plane Image or EPI. Note that we can get a vertical EPI in a similar way by keeping x_0 and u_0 constant. Thus, at each pixel, we get a horizontal EPI and a vertical EPI as shown in Figure 1-(b). These two EPIs have only one common pixel in the central image. If we vary x in Figure 1-(a), the coordinate u changes as

$$\Delta u = -\frac{f}{z} \cdot \Delta x, \tag{1}$$

where Δu is the distance between the scene points moved in the image plane $\Pi_{u,v}$, Δx is the distance between the two cameras located at the camera plane $\Pi_{x,y}$ along the line and f is the distance between the two parallel planes. Equation (1)

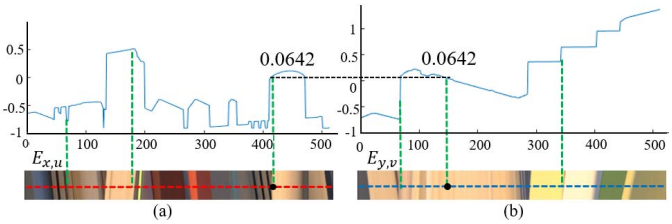


Fig. 2. Visualization of the relationship between disparity values and slopes in EPIs from Fig. 1. Our two stream network learns this relationship from the EPI patches. A central image pixel (143, 423) has the same slope/disparity value at horizontal and vertical EPIs.

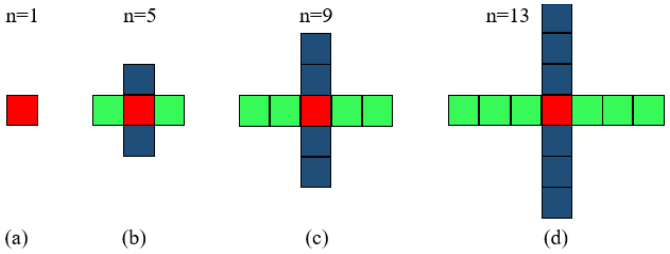


Fig. 3. Different pixel combinations used for regression. The blue pixels are in horizontal direction EPI and the green pixels are in vertical direction EPI. The red pixel is where the two EPIs intersect.

can be rearranged as

$$z = -f \cdot \frac{\Delta x}{\Delta u}, \tag{2}$$

where $\frac{\Delta x}{\Delta u}$ is the slope of the line in EPI indicating that the real metric depth value z is inversely proportional to the slope of its corresponding line in the EPI. In other words, the slope relates to the inverse disparity of the corresponding point in 3D space, which is termed as the depth-slope relationship [46]. Thus, predicting depth from light field images is essentially equivalent to computing the slope of lines in the EPIs.

Interestingly, even though the actual images contain quite complex shapes, the EPIs consist of simple linear structures which are projections of the corresponding 3D space points as shown in Figure 2.

IV. Proposed Method

The proposed method has two parts: (1) Learning a two stream CNN to predict disparity values of a small group of local pixels and (2) using the central RGB image and its edge map to refine the disparity image estimated by our network.

A. Learning a Two Stream CNN for Disparity Estimation

1) Data for CNN Training: CNNs require large amounts of labelled training data but current light field image datasets are small and inherently devoid of ground truth disparity/depth maps. Current real light field datasets include the New Light Field Image Dataset [47], [48] which were captured by the Lytro plenoptic camera, the Stanford Light Field Archive [2] captured with a camera array and the High-Resolution Disney Dataset [3] captured by gantry. Since these datasets do not have ground truth disparity/depth images, they cannot be used for supervised learning and quantitative analysis of the results.

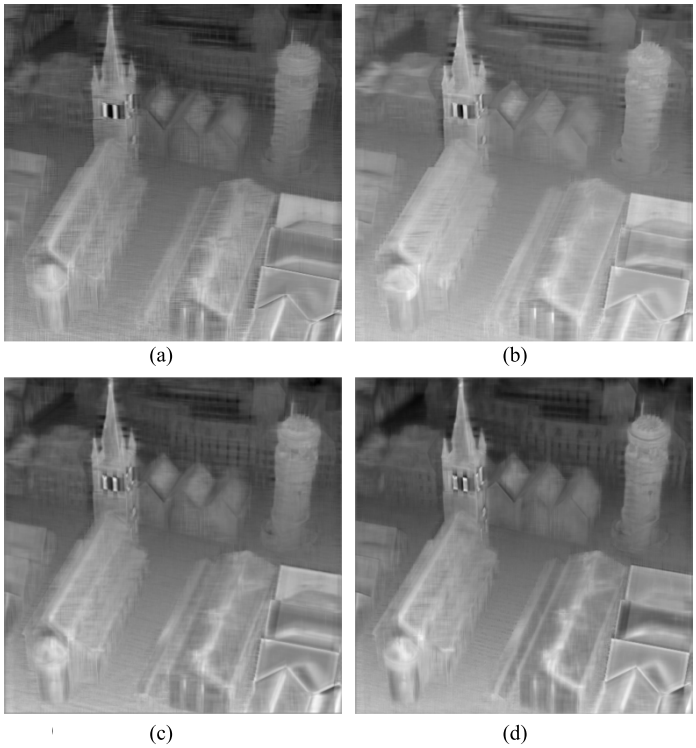


Fig. 4. Comparison of the proposed network outputs when $n = 1, 5, 9, 13$ for the Town image of the synthetic dataset [24]. Best performance is achieved at $n = 5$ indicating that multiple regression is a better than single i.e. $n = 1$. (a) $n = 1$, RMSE $= 0.265$. (b) $n = 5$, RMSE $= 0.2589$. (c) $n = 9$, RMSE $= 0.2899$. (d) $n = 13$, RMSE $= 0.297$.

Synthetic datasets have two advantages: they have ground truth disparity images and the ground truth is perfectly aligned with the central light field image. Only two synthetic datasets are publicly available, the HCI Light Field Benchmark by Wanner *et al.* [49] and the new HCI Light Field Benchmark by Honauer *et al.* [24]. We use the latter for training our network as it is more suitable [24]. This dataset includes light field images with 512×512 spatial resolution and 9×9 angular resolution. The dataset includes 24 carefully designed scenes with ground truth disparity images as shown in Figure 5. We use 19 light field images for training and the remaining 5 light field images for testing and quantitative comparison with existing algorithms.

Since the depth of a point is related to the line orientation in its corresponding EPIs in the 4D light field data, our proposed network learns this orientation from a predefined neighborhood of a given point $(x, y) \in \Pi_{x,y}$. More precisely, we extract two EPI patches of size 31×9 centered at (x, y) from the horizontal and vertical EPIs, as shown in Figure 6.

2) Network Architecture: Wanner and Goldluecke [23] proposed a method that estimates the slopes of the lines visible in the two direction EPIs to predict two rough disparity values for each pixel in the central image. It also returns a confidence map that measures the reliability of each depth estimate from the two EPIs. The disparity of high confidence pixels is then propagated to the low confidence ones in an objective function. Luo *et al.* [21] used a siamese network that takes patches from the left and right images of a stereo pair as input and learns a

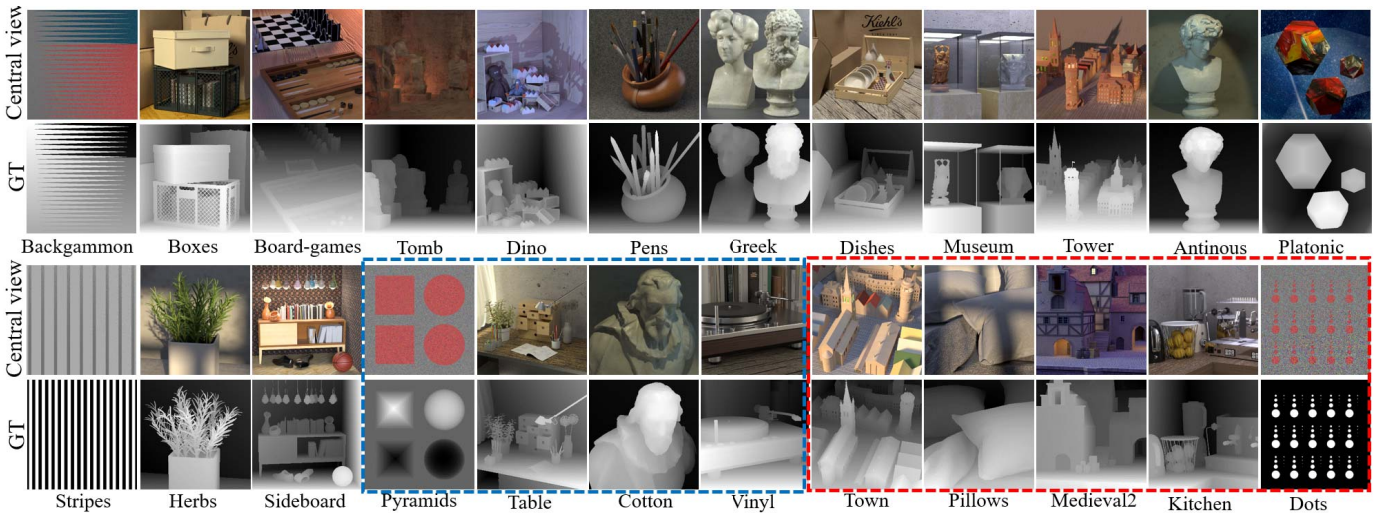


Fig. 5. Synthetic light field images [24] used for this work. The images in the red box are used for testing and comparison with other algorithms. The images in the blue box are used for validation (and to monitor overfitting) and the remaining images are used to train the network.

probability distribution over all disparity values. We propose a two stream CNN architecture, where the streams process the horizontal and vertical EPI patches separately and combine them at the fully connected layer.

We design and train our network from scratch. Not relying on pre-trained networks allows us to better adapt the network structure to the light field disparity estimation problem. An overview of the network architecture is given in Figure 6, where the two streams are highlighted in purple and green. Each branch takes an EPI image patch and passes it through a set of four convolution layers denoted as L_i. The outputs of the convolution layers are concatenated and then followed by two fully-connected layers. All convolution layers are followed by ReLU, howevever, we do not use pooling or normalization layers in our network. The two fully-connected layers fuse the horizontal and vertical EPI features from the two streams to estimate the final disparity values. The first and third convolution layers are padded so that the input and output have the same size.

The input to our network are two $31 \times 9 \times 3$ EPI patches which are quite small. However, to effectively learn the EPI slopes we need bigger kernel sizes which will shrink the input EPI patches quite rapidly as they progress through the network. Therefore, we do not include pooling layers. In fact, to account for rapid shrinking of the EPI patches, we use padding. For example our first and second kernel are 7×7 which already reduces the EPI patch to 25×3. The remaining kernels are 5×5 and 3×3 respectively. We perform a padding of 3 at the first layer and 2 at the third layer. With pooling, the four convolution layers will reduce to just two layers making the network quite shallow. The ResNet [18] and DenseNet [50] blocks which do not contain any pooling have already shown to be an effective strategy. As the layers go deeper, the number of feature maps increase (32, 64, 128, and 128). The size of the two fully-connected layers are 4096 and n (where $n = 1, 5, 9,$ or 13).

3) Neighbourhood Size Analysis: We treat the disparity estimation as a multi label regression problem and show that compared to single pixel estimation, estimating multiple pixel values gives more accurate results since the network learns correlation between the pixels i.e. neighboring pixels in the disparity map have a strong correlation. Moreover, with multiple regression using a cross shaped sliding window, the disparity at each pixel is estimated n times (where n is the number of labels/pixels in the cross) which can be combined to increase the accuracy and robustness.

Figure 3 shows the different size ($n = 1, 5, 9,$ or 13) cross shaped pixel neighborhoods we tested. The blue and green pixels correspond to the horizontal and vertical direction EPIs respectively and the red pixel indicates the intersection of the EPIs. Note that we do not use a rectangular shaped neighbourhood window since pixels outside the cross are not represented by the EPIs corresponding to the center pixel. We train four networks to compare the different output settings. Figure 4 shows the estimated disparity map of the scene Town. We can observe that $n = 5$ gives the smoothest output and with the least root mean squared error. Thus, we use $n = 5$ for the last fully connected layer in the remaining experiments. Since every pixel is predicted five times, we average the values.

4) Training: Although, we have only 19 light field images from the HCI dataset for training and validation, we extract five million patches from them. To avoid overfitting, we augment the training data by changing the color and brightness of the EPI images increasing the number of patches to 10 million. Out of these, eight million are used for training and two million for validation. We subtract the mean patch and train the proposed two stream network using the Caffe library [51] with Euclidean loss. We use back propagation with 0.001 learning rate and set the weight decay rate to 0.004 and batch size to 128. We stopped the training process after 78K iterations (2.5 epochs) as the network had already converged.

B. CNN Output Refinement

The proposed two stream CNN gives a disparity map aligned with the central light field image where the disparity values are reliable around edges but noisy and unreliable at

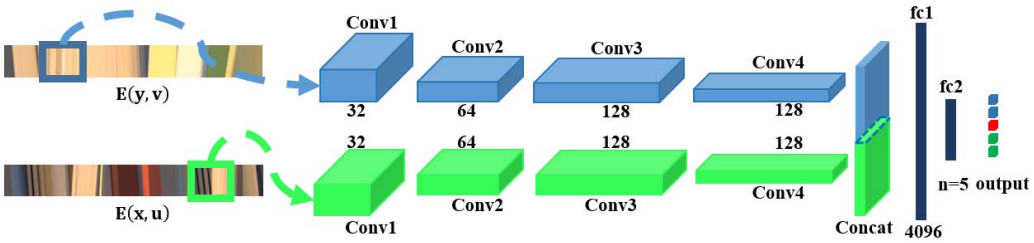


Fig. 6. Proposed two stream CNN for disparity estimation. The CNN inputs are an EPI pair of size $31 \times 9 \times 3$ each. Each CNN stream (blue, green) consists of four convolutional layers, each followed by a ReLU layer. The output of the two streams are concatenated and then followed by two fully-connected layers to estimate $n = 5$ values.

textureless image regions which have no dominant orientation in the EPI patches, see Figure 7-(c) and (d). This is similar to the aperture problem in optical flow where only the optical flow component in the direction of the gradient can be measured. Note that the disparity map output by our raw network is still competitive with modern stereo algorithms which exploit different forms of cost aggregation and post processing, see Table I. As shown in Figure 7-(b) and (c), we assume that there exists a joint occurrence between depth discontinuities and image edges. To refine the predicted orientations with a global optimization criterion, we use the central RGB image to guide the disparity map refinement.

The key idea is to select a reliable subset from the CNN output disparity map and use an optimization method to reconstruct the final dense map which is smoother and noise free. As shown in Figure 7-(a) and (b), if we compare natural images to depth maps, the latter would show a much sparse structure than the former. Similar to other dictionary-based disparity refinement methods, we assume that the final disparity map $\hat{D}$ can be represented by a sparse linear combination of some basis functions:

$$\hat{D} = A\alpha, \tag{3}$$

where $A = [a_1, \ldots, a_m]^T$ is a dictionary of basis vectors and $\alpha = [\alpha_1, \ldots, \alpha_m]^T$ are the sparse coefficients.

The disparity refinement problem can be posed as an optimization problem in which the goal is to recover the final disparity map $\hat{D}$ by seeking a sparse vector $\alpha \in R^m$ such that the observed disparity samples D are best approximated. Mathematically, we consider the problem

$$\min_{\alpha} \frac{1}{2} \left\| M_e \odot (\hat{D} - D) \right\|_2^2 + \lambda_1 \|\alpha\|_1, \tag{4}$$

where M_e is the binary confidence matrix that selects only the reliable pixels, $\|\alpha\|_1$ is the L_1-norm of α, and λ_1 is a parameter to control the sparsity of coefficients. We apply Canny edge detection on the central view image followed by dilation of the edges to get the the confidence matrix M_e.

The dictionary A is overcomplete and $AA^T = I$. Therefore, from (3) we have $\alpha = A^T\hat{D}$ and substitute it in (4) to get

$$\min_{\hat{D}} \frac{1}{2} \left\| M_e \odot (\hat{D} - D) \right\|_2^2 + \lambda_1 \left\| A^T \hat{D} \right\|_1. \tag{5}$$

The regularization in (5) is not sufficient because the ℓ_1 norm only enforces sparsity in the coefficients α. However, it is desirable to additionally enforce smoothness in the reconstructed disparity map. Therefore, we introduce a non-local mean regularization and a flatness penalty in (5) to get

$$\min_{\hat{D}} \frac{1}{2} \left\| M_e \odot (\hat{D} - D) \right\|_2^2 + \lambda_1 \left\| A^T \hat{D} \right\|_1 + \lambda_2 R_1(\hat{D}) + \lambda_3 R_2(\hat{D}). \tag{6}$$

Where $R_1(\hat{D})$ denotes the nonlocal-mean (NLM) smoothness regularizer, $R_2(\hat{D})$ denotes the Laplacian constraint for flatness of the output disparity map and λ_2 and λ_3 control the weight between smoothness and flatness respectively.

The NLM regularizer has been successfully applied to image restoration and denoising [52] and has proven effective for the refinement of depth maps [53], [54]. We use the NLM properties of the central image C to regularize the function and define it as

$$R_1(\hat{D}) = \sum_{p} \sum_{q \in N(p)} S_{pq} \left(\hat{D}(p) - \hat{D}(q) \right)^2, \tag{7}$$

where N is the window size around pixel p to search for a similar pixel q. We use an 11×11 window. The similarity weight S_{pq} between the pixels p and q is calculated as

$$S_{pq} = exp(- \sum_{\substack{p' \in N'(p) \\ q' \in N'(q)}} \frac{[C(p') - C(q')]^2}{\rho_{Color}^2} + \frac{[\nabla_{p,p'} C - \nabla_{q,q'} C]^2}{\rho_{Grad}^2}) \tag{8}$$

where ρ_{Color}^2 and ρ_{Grad}^2 are the variances in the color and gradient values of the image, $\nabla_{p,p'} C = C(p) - C(p')$ is the image gradient at pixel position p and $N'(p)$ is a small window around p. As shown in Figure 3-(b), we use $n = 5$ regression network where the neighbourhood pixels around the center are similar to a 3×3 window but without the corner pixels. Therefore, we select a 3×3 square window for N' in all experiments. This NLM regularization maintains edges in the disparity map at locations where edges exist in the light field center RGB image but smoothens out the remaining regions. The flattening constrain $R_2(\hat{D})$ is defined as

$$R_2(\hat{D}) = \sum_{(x,y)} \left(\left\| \frac{\partial \hat{D}}{\partial x} \right\|_{(x,y)} + \left\| \frac{\partial \hat{D}}{\partial y} \right\|_{(x,y)} \right), \tag{9}$$

where $\frac{\partial \hat{D}}{\partial x}$ and $\frac{\partial \hat{D}}{\partial y}$ are the first order finite difference operators in the horizontal and vertical directions.

Fig. 7. Refinement of the network output disparity image. We use the central RGB image and its edge map as the confidence matrix to reconstruct the final disparity image.

TABLE I

QUANTITATIVE COMPARISON OF DEPTH ESTIMATION METHODS ON THE TEST IMAGES (SEE FIG.8 AND 9) OF THE SYNTHETIC HCI DATASET [24]. THE VALUES (UNIT: PIXEL) SHOW THE AVERAGE ROOT MEAN SQUARED ERROR (RMSE), THE MEAN ABSOLUTE ERROR (MAE) AND THE PEAK SIGNAL TO NOISE RATIO (PSNR) BETWEEN THE ESTIMATED AND GROUND TRUTH DISPARITY MAPS

	Town	Pillows	Medieval2	Kitchen	Dot
Tao et al. [38]					
RMSE	0.4489	0.4297	0.3987	0.4222	0.5859
MAE	0.3948	0.2455	0.3649	0.369	0.5525
PSNR	55.0878	55.4675	56.1179	55.6204	52.7743
Jeon et al. [14]					
RMSE	0.5411	0.6634	0.3664	0.5703	0.2955
MAE	0.488	0.6067	0.2526	0.4943	0.1738
PSNR	53.4653	51.6953	56.8517	53.0087	58.7197
Wang et al. [40]					
RMSE	0.4493	0.5078	0.1507	0.5877	**0.1652**
MAE	0.3912	0.4527	0.0876	0.472	**0.0661**
PSNR	55.0801	54.0169	64.5685	52.7477	**63.7706**
Heber and Pock's net output [22]					
RMSE	0.6917	0.8205	0.158	0.7846	0.5805
MAE	0.6121	0.7396	0.067	0.6668	0.5761
PSNR	51.3324	49.8492	64.1577	50.2378	52.8548
Proposed net output					
RMSE	0.2589	0.165	0.1432	0.2739	0.2809
MAE	0.1649	0.1019	0.0506	0.1713	0.1761
PSNR	59.8682	63.7811	65.0119	59.3790	59.1598
Proposed final output					
RMSE	**0.1782**	**0.1403**	**0.101**	**0.2673**	0.2127
MAE	**0.1047**	**0.0971**	**0.0474**	**0.1605**	0.1282
PSNR	**63.1126**	**65.1897**	**68.0444**	**59.5908**	61.5755

Liu *et al.* [25] show that a refined Wavelets based dictionary encodes disparity maps more optimally compared to natural images. Therefore, we use Wavelets as the dictionary A in (3). To solve (6), we use the modified Alternating Direction Method of Multipliers (ADMM) proposed by Liu *et al.* [25]. ADMM can be traced back to the proximal operators proposed by Eckstein and Bertsekas [55] and comprehensive tutorials are available [56], [57]. Grid search is used to select the lambda parameters of the objective function (6).

V. REAL WORLD DATASET WITH GROUND TRUTH

Existing real world datasets [2], [3], [48], [47] lack ground truth 3D data. Therefore, they cannot be used for supervised learning or detailed comparisons. We cover this gap and propose a new dataset. Since it is difficult to obtain ground truth 3D images of real world scenes that are perfectly aligned with the central image of light field cameras, we provide a dataset that contains unaligned 3D scans of real objects and their corresponding light field images. The 3D scans are obtained with the 3dMD scanner [26] which scans objects to the millimeter resolution simultaneously from two viewpoints to cover 180° of their surface. We first scan 19 objects multiple times to get their dense 3D pointclouds and then collect light field images, using the Lytro Illum camera [4], of real scenes where the 19 objects are placed one by one. Most 3D scanners work only indoors since they use active near-IR illumination. An advantage of light field cameras is that they are passive and work equally well outdoors. Therefore, we capture the light field images in outdoor environments which makes our dataset unique.

We calibrate the Lytro Illum camera following the method and code of Bok *et al.* [58]. The calibration parameters were kept constant for all imaging sessions and are included in the dataset. The dataset containing the light field images, camera calibration parameters and ground truth 3D scans (with texture) can be downloaded from http://staffhome.ecm.uwa.edu.au/~00053650/databases.html. Note that the 3D scans can be rotated and re-rendered to get depth maps aligned with the light field central images. Similarly, the disparity maps estimated from light field images can be transferred to metric depth maps, using the camera calibration parameters, and to subsequently reconstruct the 3D pointclouds from the pixels. This will facilitate more detailed comparison and benchmarking of future algorithms.

VI. EXPERIMENTAL RESULTS

We conducted experiments on the recently proposed synthetic HCI light field benchmark [24] and our proposed real world dataset. The test partition of [24] containing five images is used for quantitative comparison with state-of-the-art methods [14], [22], [38], [40]. To extract the 4D real light field image, we use the toolbox provided by Dansereau *et al.* [59]. We use the authors' provided codes with default parameters for Tao *et al.* [38], Jeon *et al.* [14], and Wang *et al.* [40] and set the depth label range to 75. Since Heber and Pock's [22] code is not publicly available, we implemented their network and trained it on exactly the same training data as we trained our network on. For the refinement, we set the parameters as $\lambda_1 = 0.001$, $\lambda_2 = 0.003$, $\lambda_3 = 0.002$.

Fig. 8. Qualitative comparison on the Town, Pillows and Medieval2 images of the HCI dataset [24]. Table I provides quantitative comparisons.

A. Evaluation on Synthetic Data

Table I shows quantitative comparison of our method with others on the five test images of the HCI dataset [24]. The comparison is done using three metrics: the average Root Mean Squared Error (RMSE) for all pixels, the Mean Absolute Error (MAE) for all pixels and the Peak Signal to Noise Ratio (PSNR) for all pixels. Table I shows results for the network output only of [22] (using our implementation) without refinement to highlight that our network only (without refinement) achieves better accuracy. Both networks are trained on exactly the same training data. Hence the comparison of [22] in Table I is only with our proposed network output and not with [14], [38], and [40]. Our method obtains the lowest error rates for all test images except for the Dots image where it achieves the second lowest. The light field image of Dots exhibits significant noise in the lower right regions and has only two ground truth depth layers (see Figure 9). The method by Wang *et al.* [40] is able to achieve better results on this image since it uses discrete labels for depth. However, such an approach does not generalize well to other scenes where the depth values are not discrete. For the remaining four images,

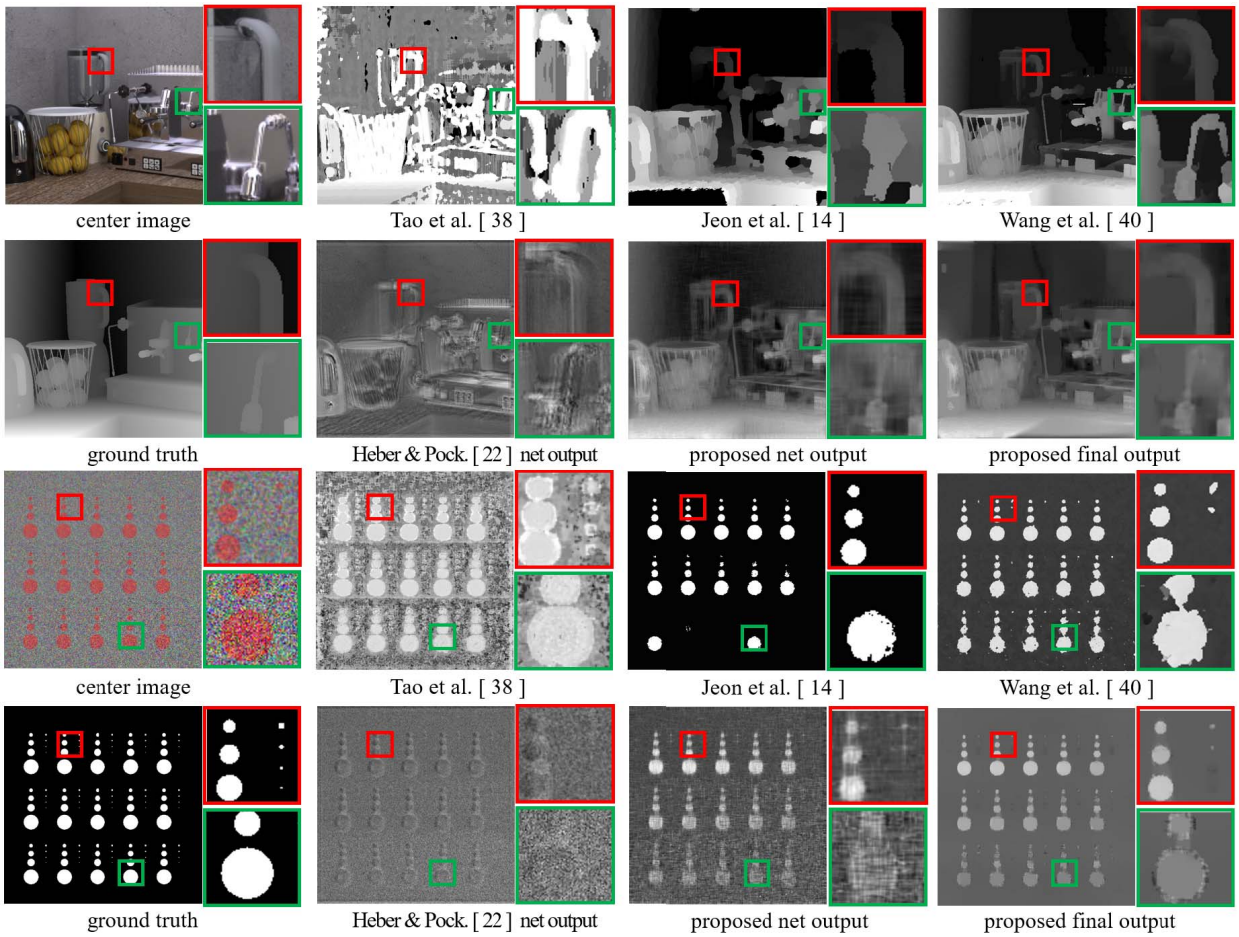


Fig. 9. Qualitative comparison on the Kitchen and Dots images of the HCI dataset [24]. Table I provides quantitative comparisons.

even the raw network output of our CNN outperforms all others by significant margins.

Figure 8 and Figure 9 qualitatively compares the raw network output and final refined output of our method with the others. We can see that our network gives a reasonable disparity map that looks more accurate than Heber and Pock's network [22]. The network output is robust to depth discontinuities and accurate in well textured and edge regions, but degrades otherwise. Our refinement process further reduces the errors in textureless regions while simultaneously preserving sharp depth discontinuities.

Figure 8 shows the results on the Town, Pillows, and Medieval2 scenes. In the Town scene, our network obtains more regular shape of the lighthouse roof than Heber and Pock's network. Our final disparity map preserves sharper edges and gives more accurate disparities (see highlighted areas) while the other methods merge the raised blocks or estimate inaccurate disparity values of the roof. In the Pillows scene, the contours of the Pillows are more accurately estimated by our network compared to Heber and Pock's network. Our final refined output gives a smoother disparity map following the shape of the Pillows whereas other methods produce irregularities, non-smooth results or artefacts near the discontinuities. In the Medieval2 scene, our network outputs a smoother and less noisy disparity map compared to Heber and Pock's network. The result of Wang *et al.* [40] is close to ours, however, our method is more regular in planar areas. Our final result also obtains smoother transitions in the disparity of the window and the ground compared to other approaches.

Figure 9 shows qualitative results for the Kitchen and Dots scenes. In the Kitchen scene, the kettle and stove are metal whereas the juicer is made of glass. The specularities and transparency make disparity estimation a real challenge. Our network output is again more regular than that of Heber and Pock's network for the overall scene. The result by Wang *et al.* [40] is close to ours, however, our method gives a smoother and accurate output as shown in the highlighted regions. The Dots image is noisy and has only two ground truth disparities. Hence, discrete depth estimation methods like Wang *et al.* [40] and Jeon et al. [14] achieve better visual results compared to our method. However, note that the raw disparity estimate of our network is still better than Heber and Pock's network.

B. Evaluation on Real Data

We compare the final output of our method with state of the art methods [14], [38], [40] on our real dataset captured in challenging outdoor environments. We present results for nine images here and provide the rest as supplementary material. To match the input of our trained CNN model, we use only the inner 9×9 sub-aperture images of the Lytro Illum.[1]

[1]The Lytro Illum camera provides 15×15 images.

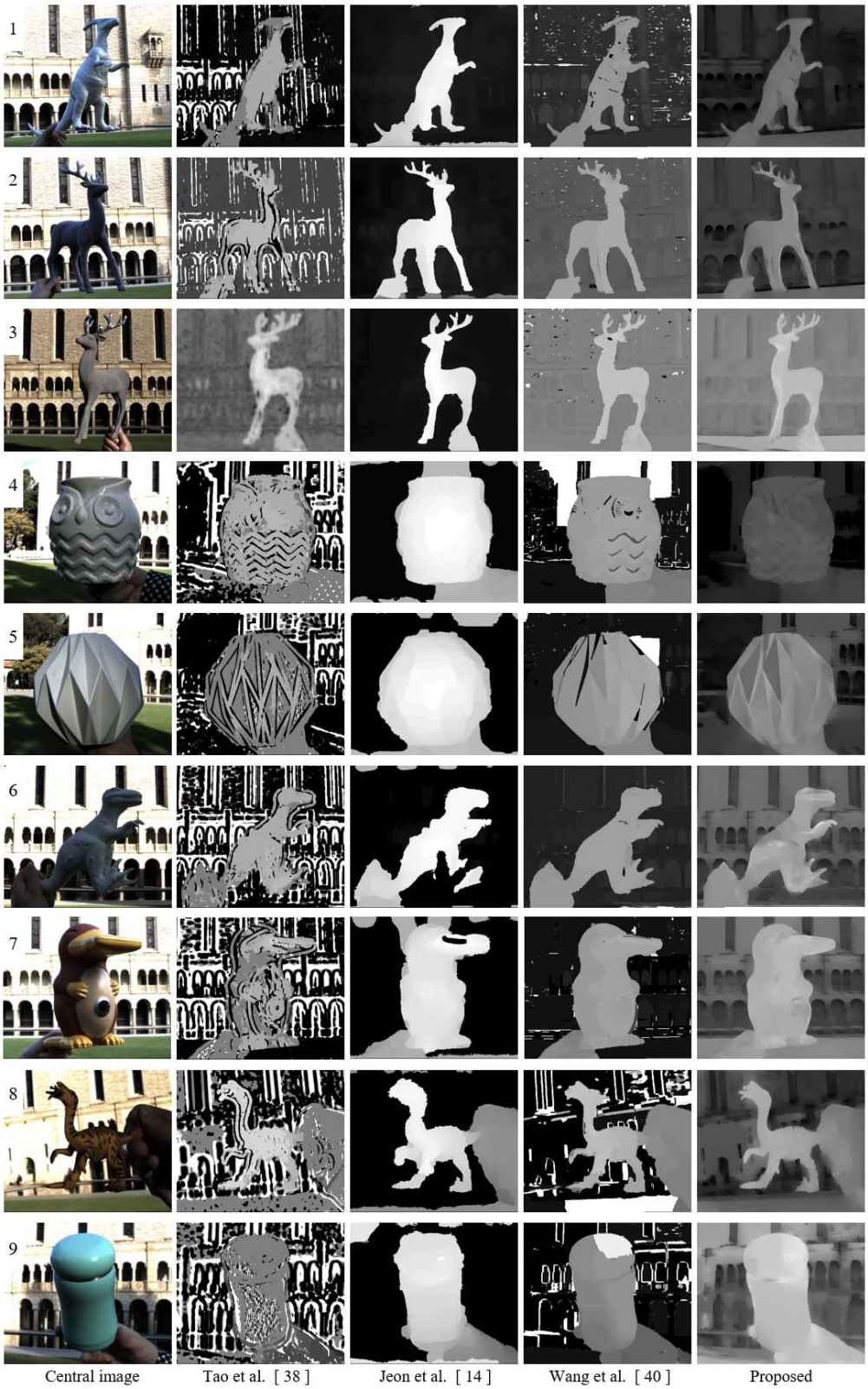


Fig. 10. Evaluation on the proposed real dataset. The first column is central light field image and the following columns are the outputs from [14], [38], and [40] and our proposed method respectively.

Figure 10 compares the disparities computed by the four methods. We can see that Tao *et al.* [38] method is adversely affected by illumination changes. Jeon *et al.* [14] method only estimates a few discrete disparity values on the object and completely misses the background. Wang *et al.* [40] method correctly estimates the silhouettes of the objects,

TABLE II

QUANTITATIVE COMPARISON USING RMSE (UNIT: MM) BETWEEN NEAREST NEIGHBOUR POINTS OF THE ESTIMATED POINTCLOUD AND GROUND TRUTH POINTCLOUD OF REAL OBJECTS FROM FIG. 10 AFTER RIGID REGISTRATION OF THE TWO POINTCLOUDS

	1	2	3	4	5	6	7	8	9
Jeon et al. [14]	19.059	14.647	15.021	20.718	24.514	21.48	25.671	17.438	17.32
Wang et al. [40]	13.218	12.312	11.493	18.115	20.47	16.034	18.69	19.152	15.83
Proposed method	**10.97**	**8.416**	**8.714**	**15.379**	**13.552**	**14.61**	**16.427**	**8.152**	**9.071**

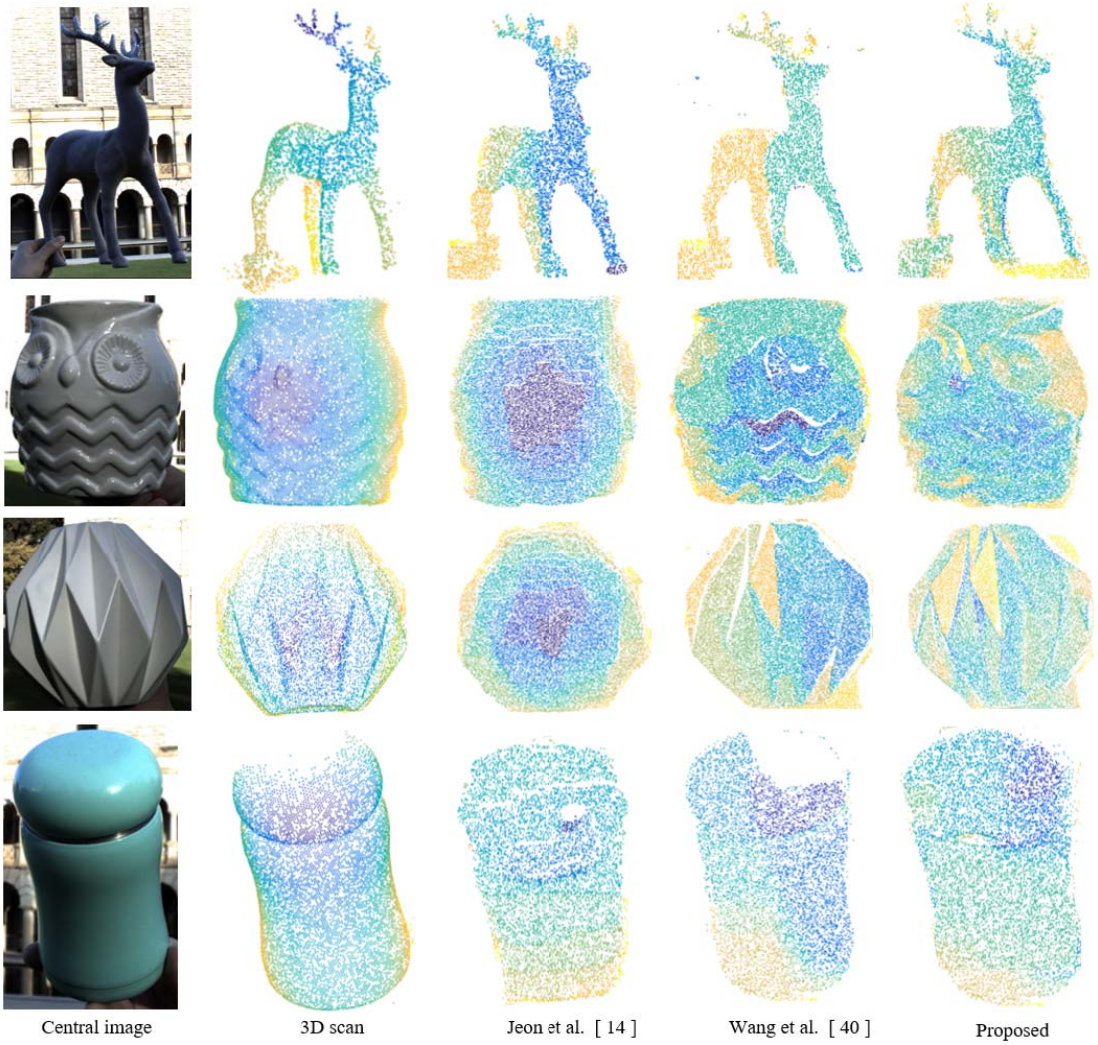


Fig. 11. Pointcloud evaluation of real data. The first column is the central light field image and the second column is the respective ground truth pointcloud from 3D scanner. Column 3 to 5 show pointclouds reconstructed by [14] and [40] and our proposed method respectively from the light filed image (column 1).

however, the disparity maps are contaminated by noise and artefacts. Moreover, their method does not accurately estimate the background.

Our proposed algorithm is able to accurately estimate the object silhouettes and smoother disparities on the objects as well as the background. In fact, our method is the only one to correctly estimate the background. In the disparity map of image 3 and 8, only our method correctly estimates the sharp boundaries of the deer antlers and the dinosaur head and in image 5, only our method correctly captures the shape of the object contours. Recall that our CNN model was neither trained nor fine tuned on this dataset.

For comparison with ground truth 3D pointclouds, we calibrated the Lytro Illum using the method proposed by Bok *et al.* [58]. We transform the disparity map to depth map using the conversion formula given in the supplementary material of [24]. Using the camera parameters, we convert the depth maps estimated by Jeon *et al.* [14] and Wang *et al.* [40] and our method then compare them to the ground truth 3D pointclouds of the objects only.[2]

Table II shows quantitative comparison of our method with Jeon *et al.* [14] and Wang *et al.* [40] on the real

[2]We do not have ground truth of the background.

world light field images given in Fig. 10. We calculated the RMSE (Unit:mm) between nearest neighbour points of the estimated pointcloud and ground truth pointcloud of real object from Fig.10 (top to bottom) after rigid registration of the two pointclouds. Note that our proposed method achieves significantly lower errors for all objects compared to the other two methods. Fig. 11 shows qualitative comparison (2,4,5,9 from Fig. 10) where we can see that the pointclouds of our method more accurately resemble the ground truth pointclouds obtained from the 3dMD scanner.

VII. Conclusion

We proposed a data-driven method for depth estimation from light field images. We trained a two stream network to learn the relationship between disparity values and line slope in the EPI domain. To refine the network output, we exploit prior information in the central image to optimize the variational model. We also proposed a new real world dataset that consists of light field images of 19 objects in outdoor scenes and their ground truth 3D scans. The trained network model[3] and the real dataset will be released publicly to facilitate better comparison of future methods. Experiments on synthetic and real data show that our method quantitatively and qualitatively outperforms existing state-of-the-art in depth from light field images.

Acknowledgments

The Tesla K-40 GPU was donated by NVIDIA corporation.

References

[1] E. H. Adelson and J. R. Bergen, "The plenoptic function and the elements of early vision," in *Computational Models of Visual Processing*. Cambridge, MA, USA: MIT Press, 1991, pp. 3–20.

[2] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.

[3] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–73, Jul. 2013.

[4] *Lytro*. Accessed: Mar. 14, 2018. [Online]. Available: https://www.lytro.com/

[5] *Raytrix*. Accessed: Mar. 14, 2018. [Online]. Available: https://raytrix.de

[6] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep. CSTR*, vol. 2, no. 11, pp. 1–11, 2005.

[7] S. Wanner, C. Straehle, and B. Goldluecke, "Globally consistent multi-label assignment on the ray space of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1011–1018.

[8] T.-C. Wang, J.-Y. Zhu, E. Hiroaki, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 121–138.

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

[10] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, p. 193, 2016.

[11] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.

[12] Y. Zhang *et al.*, "Light-field depth estimation via epipolar plane image analysis and locally linear embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 739–747, Apr. 2017.

[13] J. Navarro and A. Buades, "Robust and dense depth estimation for light field images," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1873–1886, Apr. 2017.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[17] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[19] H. Rahmani and A. Mian, "3D action recognition from novel viewpoints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1506–1515.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[21] W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5695–5703.

[22] S. Heber and T. Pock, "Convolutional networks for shape from light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3746–3754.

[23] 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.

[24] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.

[25] L.-K. Liu, S. H. Chan, and T. Q. Nguyen, "Depth reconstruction from sparse samples: Representation, algorithm, and sampling," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1983–1996, Jun. 2015.

[26] *3dMD Laser Scanner*. Accessed: Mar. 14, 2018. [Online]. Available: http://www.3dmd.com

[27] R. Szeliski, "A multi-view approach to motion and stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 1999, p. 163.

[28] B. Goldlücke, M. A. Magnor, and B. Wilburn, "Hardware-accelerated dynamic light field rendering," in *Proc. VMV*, 2002, pp. 455–462.

[29] Y. Liu, X. Cao, Q. Dai, and W. Xu, "Continuous depth estimation for multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR*, Jun. 2009, pp. 2121–2128.

[30] M. Bleyer, C. Rother, and P. Kohli, "Surface stereo with soft segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1570–1577.

[31] T. E. Bishop and P. Favaro, "Full-resolution depth map estimation from an aliased plenoptic light field," in *Proc. Comput. Vis.–ACCV*, 2011, pp. 186–200.

[32] Y. Wei and L. Quan, "Asymmetrical occlusion handling using graph cut for multi-view stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun. 2005, pp. 902–909.

[33] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.

[34] J. Žbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1592–1599.

[35] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.

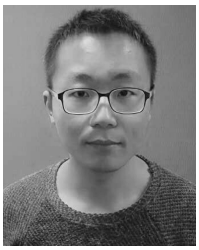
[36] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 8, pp. 824–831, Aug. 1994.

[37] M. Strecke, A. Alperovich, and B. Goldluecke, "Accurate depth and normal maps from occlusion-aware focal stack symmetry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2529–2537.

[38] 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.*, Dec. 2013, pp. 673–680.

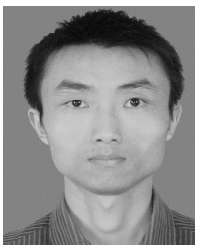
[3]Note that Heber and Pock's trained network [22] is not publicly available.

[39] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, and R. Ramamoorthi, "Shape estimation from shading, defocus, and correspondence using light-field angular coherence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 546–560, Mar. 2017.
[40] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Occlusion-aware depth estimation using light-field cameras," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3487–3495.
[41] W. Williem and I. Kyu Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4396–4404.
[42] S. Wanner and B. Goldluecke, "Globally consistent depth labeling of 4D light fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 41–48.
[43] J. Li, M. Lu, and Z.-N. Li, "Continuous depth map reconstruction from light fields," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3257–3265, Nov. 2015.
[44] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2271–2279.
[45] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.
[46] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Comput. Vis.*, vol. 1, no. 1, pp. 7–55, 1987.
[47] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, 2016.
[48] A. Mousnier, E. Vural, and C. Guillemot. (2015). "Partial light field tomographic reconstruction from a fixed-camera focal stack." [Online]. Available: https://arxiv.org/abs/1503.01903
[49] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4d light fields," in *Proc. VMV*, 2013, pp. 225–226.
[50] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[51] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[52] A. Buades, B. Coll, and J.-M. Morel, "Nonlocal image and movie denoising," *Int. J. Comput. Vis.*, vol. 76, no. 2, pp. 123–139, Jul. 2007.
[53] P. Favaro, "Recovering thin structures via nonlocal-means regularization with application to depth from defocus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1133–1140.
[54] H. Kwon, Y.-W. Tai, and S. Lin, "Data-driven depth map refinement via multi-scale sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 159–167.
[55] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 3, pp. 293–318, 1992.
[56] D. Han and X. Yuan, "A note on the alternating direction method of multipliers," *J. Optim. Theory Appl.*, vol. 155, no. 1, pp. 227–238, 2012.
[57] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.
[58] Y. Bok, H.-G. Jeon, and I. S. Kweon, "Geometric calibration of micro-lens-based light field cameras using line features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 287–300, Feb. 2017.
[59] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1027–1034.



Mingtao Feng is currently pursuing the Ph.D. degree with the College of Electrical and Information Engineering, Hunan University. He is currently a visiting Ph.D. student with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include image processing, computer vision, and machine learning.

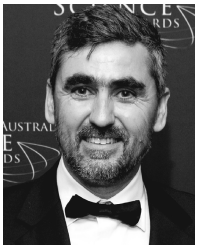
Yaonan Wang received the Ph.D. degree in electrical engineering from Hunan University, China, in 1994. From 1994 to 1995, he was a Post-Doctoral Research Fellow with the National University of Defence Technology, China. From 1998 to 2000, he was a Senior Humboldt Fellow in Germany. From 2001 to 2004, he was a visiting Professor with the University of Bremen, Bremen, Germany. Since 1995, he has been a Professor with Hunan University. His research interests include robot control, intelligent control and information processing, industrial process control, and image processing.



Jian Liu is currently pursuing the Ph.D. degree with the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include computer vision, deep learning, and human pose estimation.

Liang Zhang received the Ph.D. degree in instrument science and technology from Zhejiang University in 2009. He is currently an Associate Professor and the Director of the Embedded Technology and Vision Processing Research Center, Xidian University. His research interests focus on the areas of big data processing, multicore embedded systems, computer vision, deep learning, simultaneous localization and mapping, human–robot interaction, and image processing.

Hasan F. M. Zaki received the B.E. degree in mechatronics from the International Islamic University of Malaysia, Malaysia, in 2010, and the M.E. degree in mechatronics from the University of Malaya, Malaysia, in 2013. He is currently pursuing the Ph.D. degree in computer science with The University of Western Australia. His research interests include robotic vision, RGB-Depth object and scene recognition, machine learning, 3D shape analysis, and action recognition.



Ajmal Mian is currently an Associate Professor of computer science with The University of Western Australia. His research interests include computer vision, image processing, machine learning, 3D shape analysis, human action recognition, and hyperspectral image analysis. He has received several awards, including the West Australian Early Career Scientist of the Year Award, the Aspire Professional Development Award, the Vice-chancellors Mid-career Research Award, the Outstanding Young Investigator Award, IAPR Best Scientific Paper Award, and the EH Thompson Award. He has received two prestigious Fellowships from the Australian Research Council, the Australian Research Fellowship, and the Australian Post-Doctoral Fellowship. He has received seven major research grants from the Australian Research Council and the National Health and Medical Research Council of Australia with a total funding of over $3.0 million.