

# Attention-based View Selection Networks for Light-field Disparity Estimation

Yu-Ju Tsai,<sup>1</sup> Yu-Lun Liu,<sup>1,2</sup> Ming Ouhyoung,<sup>1</sup> Yung-Yu Chuang<sup>1</sup>

<sup>1</sup>National Taiwan University, <sup>2</sup>MediaTek

{r06922009, yulunliu}@cmlab.csie.ntu.edu.tw, {ming, cyy}@csie.ntu.edu.tw

## Abstract

This paper introduces a novel deep network for estimating depth maps from a light field image. For utilizing the views more effectively and reducing redundancy within views, we propose a view selection module that generates an attention map indicating the importance of each view and its potential for contributing to accurate depth estimation. By exploring the symmetric property of light field views, we enforce symmetry in the attention map and further improve accuracy. With the attention map, our architecture utilizes all views more effectively and efficiently. Experiments show that the proposed method achieves state-of-the-art performance in terms of accuracy and ranks the first on a popular benchmark for disparity estimation for light field images.

## Introduction

Light field cameras collect and record light from different directions in the scene. With the light field images captured by light field cameras, users are empowered with the capability to change the focal plane or viewpoint even after image shooting. The modern hand-held light field camera is often equipped with a micro-lens array which is placed one focal length away from the image plane of the sensor. With this structure, the measurements captured by the 2D sensor can be converted into a multi-view image with different viewpoints. The multi-view image offers several advantages over a conventional image captured by a regular camera. First, we can change the viewpoint to the scene and refocus on the object we want to see clearly in the scene for creating the effect of “Depth-of-Field.” Second, light field cameras have faster shooting speeds than conventional cameras because there is less need to focus before taking a picture. Third, the use of a larger aperture enables us to take better photographs under low-light environments. Finally, light field cameras also implicitly record the depth information which enables many interesting applications.

Although light field cameras record depth information implicitly, extracting depth information from light field images could be challenging because the baseline between sub-aperture images is very narrow, and the spatial and angu-

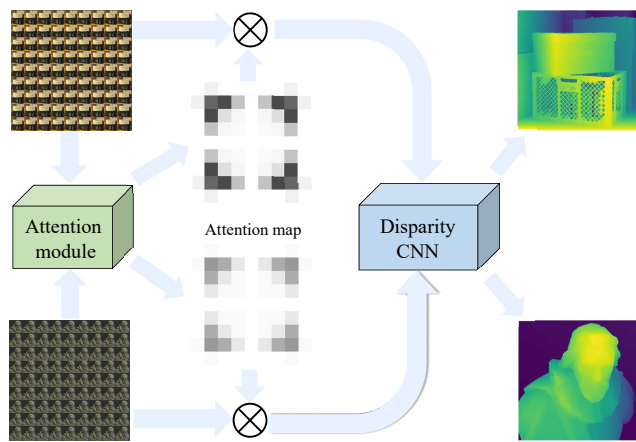


Figure 1: The light field image contains repetitive and redundant information among views. For utilizing views more effectively and efficiently for depth estimation, we design an attention-based view selection module to help the disparity estimation network determine how to weigh views’ contributions according to the properties of the input scene. With the attention map, the disparity CNN can predict the disparity map using the input views more adaptively. This figure shows that different attention maps could be generated for different scenes for better adapting to their characteristics.

lar resolutions within the image sensor are restricted by the hardware design. Several methods have been proposed to address these challenges in extracting accurate depth information from light field images. These methods often have to make a balance between computation overhead and accuracy. Conventional methods such as stereo matching can obtain depth maps of sufficient quality while suffering from heavy computation costs. At the same time, due to the narrow baseline, the resultant depth map could contain noise which causes problems in applications. Recently, several deep neural networks have been proposed for striking a better balance between accuracy and computation overhead. However, they often use only a subset of images for reducing computation and do not fully utilize the information within the light field.

This paper proposes an attention-based view selection network for estimating depth maps from light field images. Our method makes a good trade-off between accuracy and computation by exploring the following specific properties of light field images.

- *The repetitive structure of light field images.* Because of the design of light field cameras, there are correlations among views. To utilize the correlations, several methods use the epipolar geometry of light field images and only use the views at the horizontal, vertical, crosshair or diagonal directions for depth estimation. In our network, we use all views but utilize them more effectively with the help of the attention map.
- *The redundancy among sub-aperture views.* There is great redundancy among views. Using all views leads to heavy computation and does not necessarily lead to better accuracy. We propose an attention-based view selection module which can determine the importance of each view so that they can be utilized more effectively and efficiently for depth estimation. Figure 1 shows that each sub-aperture view provides different contribution and the contribution pattern among views often depends on the characteristics of the input scene.

The attention map not only reduces redundancy but also effectively indicates how important each view is in the following disparity estimation step. Different views could have different contributions as they have different spatial and angular distances from the target view. Also, some views could provide redundant information and their contribution needs to be discounted. By adding the attention map, the estimation module can focus on more important views, leading to better accuracy. Experiments show that the proposed method achieves the most accurate disparity estimation on a popular benchmark to date.

## Related Work

This section reviews disparity estimation methods of light field images in two categories: conventional methods and deep learning methods.

### Conventional Methods

Some depth estimation methods of light field (LF) images use the special structure called the epipolar plane images (EPIs), which contain the spatial and angular information of 2D slices of the light field images (Gortler et al. 1996; Levoy and Hanrahan 1996). To the best of our knowledge, the first paper which uses the EPIs for depth estimation is for depth estimation in the structure from motion (SfM) setting (Bolles, Baker, and Marimont 1987). The similarity between LF and SfM is that they both have a dense sequence of images. The EPIs contain lines with different slopes, which are formed by the projections of the same point from different viewpoints. By calculating the slope of such a line in the EPIs, we can obtain the disparity of the pixels in the images. (Wanner and Goldluecke 2012; 2014) compute the slopes in EPIs by using the structure tensor and get a high-quality depth map from light field images.

(Zhang et al. 2016) propose a spinning parallelogram operator (SPO) for estimating the depth value from EPIs, and their method is insensitive to occlusions, noise, and spatial aliasing, detrimental factors causing undesirable results of depth estimation for the light field images. (Zhang et al. 2017) also uses the EPIs and introduces the locally linear embedding (LLE) for depth estimation, which enhances the quality of depth map with faster computational time without the need for global optimization.

Some approaches do not utilize EPIs. (Yu et al. 2013) apply Constrained Delaunay Triangulation (CDT) and encode 3D line constraints by using the line-assisted graph-cut (LAGC) algorithm for light field stereo matching. (Chen et al. 2014) introduces a method to tackle occlusions in the light field depth estimation by applying a bilateral consistency metric (BCM) on the surface camera (SCam) introduced by (Jingyi Yu, McMillan, and Gortler 2002). (Tao et al. 2013) presents a method that combines both defocus and correspondence depth cues from the light field images for obtaining dense depth estimation.

Conventional methods share the inevitable problem on the trade-off between accuracy and computational cost. Our method utilizes a convolutional neural network (CNN) to achieve both better accuracy and faster computational time.

## Deep Learning Methods

In the past few years, deep learning techniques have been used in many applications of light field images such as view synthesis (Kalantari, Wang, and Ramamoorthi 2016), image compression (Zhong et al. 2019), material recognition (Wang et al. 2016), super-resolution (Yoon et al. 2017), synthesis of light field images from a single image (Srinivasan et al. 2017), and depth estimation (Shin et al. 2018).

For the problem of depth estimation, (Heber and Pock 2016) proposes a network to learn the end-to-end mapping between 4D light field images and apply the high-order regularization to refine the network. (Heber, Yu, and Pock 2017) build a U-shaped encoder and decoder to extract geometric information from light field images and produce a high-quality result at a low computational cost. (Alperovich et al. 2018) presents a fully convolutional autoencoder to encode light field images into low-dimensional representation and decode it for the depth estimation and the separation of diffuse and specular intrinsic components of the light field images. (Shin et al. 2018) introduces a fully convolutional neural network with fast and accurate performance in the depth estimation and proposes a data augmentation method to address the issue with the lack of training data.

For the trade-off between accuracy and computation, these methods often only use a sub-set of views by considering some directions in the epipolar geometry of light field images, such as the horizontal, vertical or diagonal directions. Thus, they do not fully utilize the information within light field images. We address this issue by taking all the sub-aperture views of light field images as input and design an attention-based view selection module to find out the more important sub-aperture views for estimating the depth information more efficiently and effectively.

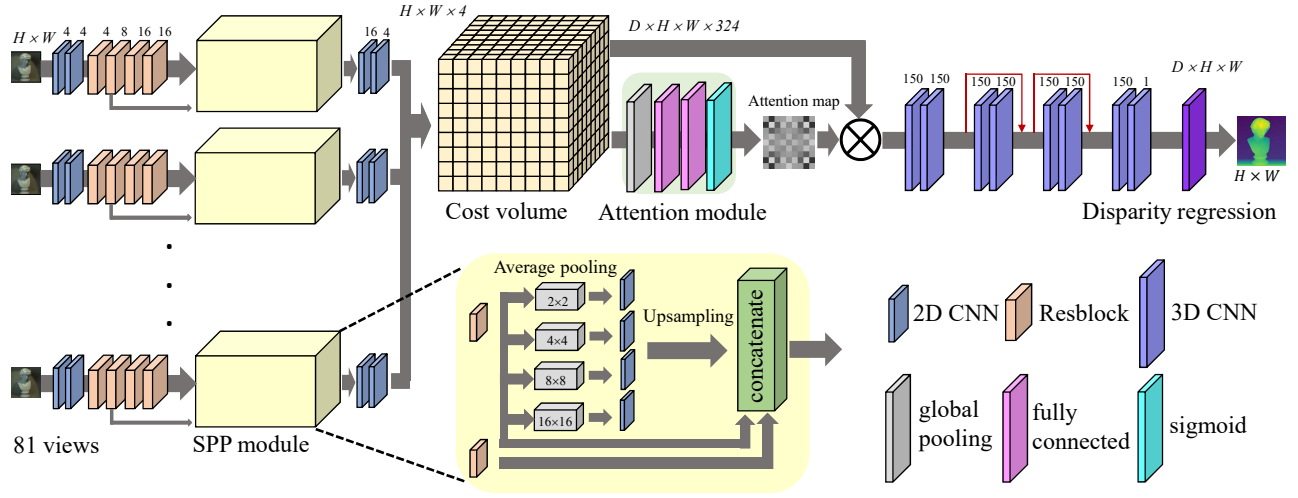


Figure 2: The architecture of the proposed method. Each sub-aperture view of the light field image passes four basic residual blocks for the unary feature extraction. After obtaining the feature maps, we apply a spatial pyramid pooling (SPP) module to extract the context information of the scene and obtain more effective feature maps. We then concatenate all the feature maps of the sub-aperture views from the SPP module into a 5D cost volume. Before sending the cost volume for disparity regression, we apply the attention-based view selection module to obtain an attention map which indicates the importance of each view. Finally, the cost volume is combined with the attention map and then sent to the disparity regression module for calculating the disparity map of the center view in the light field image.

## Method

Figure 2 depicts the architecture of the proposed network. The inputs are images of 81 views and the output is the depth map for the center view. Each sub-aperture view of the input light field image passes through four basic residual blocks (He et al. 2016). In the third and fourth residual blocks, we add the dilation convolution so that the network has a larger receptive field. The obtained feature map for each view is then fed into a spatial pyramid pooling (SPP) module (He et al. 2015) to extract the context information of the scene. Inspired by (Kendall et al. 2017) and (Chang and Chen 2018), we concatenate all the feature maps of the sub-aperture views from the SPP module into a cost volume. Before sending the cost volume into the disparity regression module, we apply the attention-based sub-aperture view selection module for learning the importance of each view. Finally, the cost volume combined with the attention map is fed into the disparity regression module for estimating the disparity map of the center view in the light field image.

### Feature Extraction and the SPP Module

For estimating disparity, it is necessary to extract effective features from images. For difficult regions such as texture-less regions or specular areas, it is challenging to have effective features. The context information is important to such regions so that their disparity values can still be estimated reliably by utilizing information of nearby regions. The SPP module in the proposed network can provide meaningful features by utilizing hierarchical context information or the relationship from nearby regions.

As explored by (He et al. 2015; Zhao et al. 2017; Chang and Chen 2018), the goal of the SPP module is to extract fea-

tures from different scales and sub-regions and provide the hierarchical context information about the region. As shown in Figure 2, we design our SPP module as follows. First, we apply four average pooling operations at different scales to compress the features. The sizes of the average pooling blocks are  $2 \times 2$ ,  $4 \times 4$ ,  $8 \times 8$ , and  $16 \times 16$ . After pooling, a  $1 \times 1$  convolution layer is used for reducing the feature dimension for each scale. We then use the bilinear interpolation to upsample these low-dimensional feature maps to the same size. Finally, we concatenate the feature maps of all levels as the output feature map of the SPP module.

### Cost Volume Construction

After passing the feature map of each sub-aperture view through the SPP module, we obtain the feature map for each view. The characteristic of CNNs makes it difficult to directly estimate the displacement by concatenating feature maps due to the finite receptive field. If the displacement is larger than the receptive field, it is impossible for CNNs to predict the correct disparity. For well utilizing these feature maps, we adopt the approach called cost volume introduced by (Zbontar and LeCun 2016; Kendall et al. 2017; Chang and Chen 2018). Given the feature maps from the SPP module, we manually shift the input images along the  $u$  or  $v$  direction with different disparity levels, so that the later part of the network can directly see pixel information at different spatial positions by using a relatively small receptive field. In our setting, we have 9 disparity levels ranging from -4 to 4. After shifting the feature maps, we concatenate these feature maps into a 5D cost volume whose size is equal to  $\text{Batch\_size} \times \text{\#Disparity} \times \text{Height} \times \text{Width} \times \text{Feature\_dimension}$ .

## Attention-based View Selection Module

Different from the stereo matching problem (Zbontar and LeCun 2016; Kendall et al. 2017; Chang and Chen 2018), there are many more views in the light field image. As mentioned previously, sub-aperture views often provide abundant but potentially redundant information for the estimation of disparity. Because the structure of the light field image is highly symmetric, we would like to have a module that can utilize this property and indicate the importance of individual views. Inspired by SENet (Hu, Shen, and Sun 2018), we propose the attention-based view selection module to find meaningful views with more importance as they have a higher potential to contribute to the accurate estimation of the disparity map in the center view.

The attention map is essentially a  $9 \times 9$  map whose entries indicate the importance of corresponding views. By exploring the structure of the light field images, we have tried three types of attention maps as shown in Figure 3. The first type is the free attention map in which each view has its own importance value. There are 81 weights to learn in this type. The second type is the symmetric attention map, in which we enforce the map is symmetric along the  $u$  and  $v$  axes. Thus, we only have to estimate the map at a quarter with 25 learnable weights. The full map can be constructed by mirroring along the  $u$  axis and then  $v$  axis. The third type is radial in which we assume the map is symmetric along the  $u$ ,  $v$  and two diagonal axes. This way, we only need to estimate 1/8 of entries (15 weights) and then construct the full attention map by mirroring along the diagonal,  $v$  and  $u$  axes. By imposing constraints on the structure of the attention map, we reduce the number of learnable weights and effectively perform regularization via domain knowledge of light field cameras. It helps with the training of the view selection network. Given the cost volume as input, the view-selection module generates the attention map by a global pooling layer, followed by two fully connected layers and ended with a sigmoid layer. We then multiply the features from the cost volume with the corresponding attention scores using the element-wise product to form the attended features. Thus, the attention map works as a scaler for each view.

## 3D CNN and Disparity Regression

For disparity regression on the attended cost volume, we employ a 3D CNN architecture. Following (Chang and Chen 2018; Kendall et al. 2017), our architecture consists of eight  $3 \times 3 \times 3$  convolutional layers, with two residual blocks from the third to the sixth 3D convolutional layers.

After passing through these 3D convolutional layers, we obtain the output cost volume from the final 3D convolutional layers and convert it from 5D to 4D,  $\text{Batch\_size} \times \text{\#Disparity} \times \text{Height} \times \text{Width}$ , in order to apply the disparity regression. Inspired by (Chang and Chen 2018; Kendall et al. 2017), we apply *soft argmin* for better and more robust performance for the stereo matching problem. It is a differentiable version of the winner-takes-all algorithm. The traditional winner-takes-all algorithm takes the argmin operation along the disparity dimension on the cost volume. However, this operation is not differentiable and cannot be

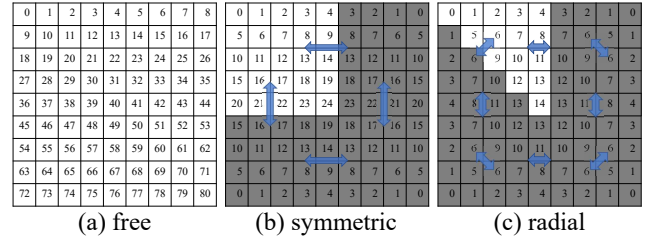


Figure 3: Different types of attention maps. Without imposing any constraint on the attention map, the free type has 81 learnable weights. By imposing symmetric constraints, the symmetric type has 25 learnable weights while the radial type has only 15, further regularizing the training of the view selection network. The gray cells indicate that their importance values can be automatically inferred by mirroring the learned weights from other views.

optimized through backpropagation. By taking a weighted sum, we can approximate the argmin operator. We modify this equation to fit with our problem and to estimate the continuous disparity maps. To calculate the normalized probability of each disparity  $d$ , we need to take the negative of each value in the predicted cost  $c_d$  from the cost volume (for the disparity values with higher costs, they will have lower probability) and normalize these values by the softmax operation  $\sigma(\cdot)$ . After we obtain the normalized probability of each disparity value, we can calculate the final predicted disparity  $\hat{d}$  by the weighted sum of each disparity  $d$  with its normalized probability as the weight:

$$\hat{d} = \sum_{d=-D_{max}}^{D_{max}} d \times \sigma(-c_d), \quad (1)$$

where  $c_d$  is the cost for the disparity value  $d$ .

## Experiments

In this section, we first introduce the datasets we used for training and evaluation. We then describe the implementation details. Finally, both quantitative and qualitative results are reported and compared with the state-of-the-art methods, along with the ablation study, discussions and limitations.

### Datasets

We use two datasets in our experiments, the 4D Light Field Dataset (Honauer et al. 2016) and a dataset released by (Alperovich et al. 2018).

**4D Light Field Dataset (Honauer et al. 2016).** This dataset is often used as the benchmark for evaluating disparity estimation methods for light field images. It contains 28 light field scenes that are partitioned into four sub-sets: “Stratified”, “Test”, “Training” and “Additional”. The light field images are rendered by the Blender renderer. The scenes in this dataset are composed of different materials, lighting conditions, and fine structures with complex occlusions. The resolution of the images is  $512 \times 512$  and the number of sub-aperture views is  $9 \times 9$ . Since the scenes are synthetic, the

	Stratified															
	Backgammon				Dots				Pyramids				Stripes			
	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE
Epinet-fcn	3.580	6.289	20.89	<b>3.629</b>	3.183	12.73	41.05	1.635	0.192	0.913	11.87	0.008	2.462	3.115	15.67	0.950
Epinet-fcn-m	3.501	5.563	19.43	3.705	2.490	9.117	35.61	1.475	0.159	0.874	11.42	0.007	2.457	<b>2.711</b>	<b>11.77</b>	0.932
Epinet-fcn9x9	3.287	4.482	15.39	3.909	4.030	18.70	44.64	1.980	0.147	0.604	8.913	0.007	<b>2.413</b>	2.876	14.75	0.915
PS_RF	7.142	13.93	74.65	6.892	7.975	17.54	78.80	8.338	<b>0.107</b>	6.235	83.23	0.043	2.964	5.790	41.64	1.382
EPN+OS+GC	3.328	10.56	55.97	3.699	39.24	82.74	84.90	22.36	0.242	3.169	28.55	0.018	18.54	19.59	28.16	8.731
SPO	3.781	8.639	49.94	4.587	16.27	35.06	58.07	5.238	0.861	6.263	79.20	0.043	14.97	15.46	21.87	6.955
Ours	<b>3.126</b>	<b>3.985</b>	<b>11.58</b>	3.648	<b>1.432</b>	<b>3.012</b>	<b>15.05</b>	<b>1.425</b>	0.195	<b>0.488</b>	<b>2.063</b>	<b>0.004</b>	2.933	5.417	18.21	<b>0.892</b>
	Training															
	Boxes				Cotton				Dino				Sideboard			
	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE
Epinet-fcn	12.84	19.76	49.04	6.240	0.508	2.310	28.06	<b>0.191</b>	1.286	3.452	22.40	0.167	4.801	12.08	41.88	0.827
Epinet-fcn-m	12.34	<b>18.11</b>	46.09	5.968	0.447	2.076	25.72	0.197	1.207	3.105	19.39	0.157	4.462	10.86	36.49	0.798
Epinet-fcn9x9	12.25	18.66	45.73	6.036	0.464	2.217	25.27	0.223	1.263	3.221	23.44	0.151	4.783	11.82	40.49	0.806
PS_RF	18.94	35.23	76.39	9.043	2.425	14.98	70.40	1.161	4.379	16.44	75.96	0.751	11.75	36.28	79.97	1.945
EPN+OS+GC	15.30	29.01	67.35	9.314	2.060	9.767	54.84	1.406	2.877	12.79	58.79	0.565	7.997	23.87	66.34	1.744
SPO	15.89	29.52	73.23	9.107	2.594	13.71	69.05	1.313	2.184	16.36	69.87	0.310	9.297	28.81	73.36	1.024
Ours	<b>11.04</b>	18.97	<b>37.04</b>	<b>3.996</b>	<b>0.271</b>	<b>0.697</b>	<b>3.644</b>	0.209	<b>0.848</b>	<b>2.339</b>	<b>12.22</b>	<b>0.093</b>	<b>2.869</b>	<b>7.243</b>	<b>20.73</b>	<b>0.530</b>
	Test															
	Bedroom				Bicycle				Herbs				Origami			
	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE	0.07	0.03	0.01	MSE
Epinet-fcn	2.403	6.921	33.99	0.213	9.896	18.05	46.37	4.682	12.10	28.95	62.67	9.700	5.918	14.37	45.93	<b>1.466</b>
Epinet-fcn-m	2.299	6.345	31.82	<b>0.204</b>	9.614	16.83	42.83	4.603	10.96	25.85	59.93	9.491	5.807	13.00	42.21	1.478
Epinet-fcn9x9	<b>2.287</b>	6.291	31.23	0.231	9.853	17.19	43.85	4.929	17.75	34.54	59.86	9.423	6.339	13.92	42.17	1.646
PS_RF	6.015	22.45	80.67	0.288	17.17	32.32	79.79	7.926	10.48	21.89	66.47	15.24	13.57	36.45	80.32	2.393
EPN+OS+GC	7.543	16.76	58.93	1.188	11.59	24.85	64.10	6.411	9.190	25.72	67.13	11.58	10.75	27.08	67.35	10.09
SPO	4.864	23.53	72.37	0.209	10.91	26.90	71.13	5.570	8.260	30.62	86.62	11.23	11.69	32.71	75.58	2.032
Ours	2.792	<b>5.318</b>	<b>13.33</b>	0.366	<b>9.511</b>	<b>15.99</b>	<b>31.35</b>	<b>3.350</b>	<b>5.219</b>	<b>9.483</b>	<b>19.27</b>	<b>6.605</b>	<b>4.824</b>	<b>8.925</b>	<b>22.19</b>	1.733

Table 1: Comparisons of our method and the compared methods on the ‘‘Stratified’’, ‘‘Training’’ and ‘‘Test’’ sets of the 4D Light Field Dataset in terms of Badpix 0.07, 0.03, 0.01 and MSE\*100.

ground truth depth can be obtained with ease. In our experiment setting, we use 16 scenes in ‘‘Additional’’ for training, 8 scenes from ‘‘Stratified’’ and ‘‘Training’’ for validating and 4 scenes from ‘‘Test’’ for testing. While we randomly sample  $32 \times 32$  gray-scale patches from the training dataset for training, we use the full resolution  $512 \times 512$  for validation. **Dataset released by (Alperovich et al. 2018).** This dataset is also rendered using Blender with the same resolution and number of views as the 4D Light Field Dataset. The scenes contain up to five objects of different scales and complexity in geometry. To prevent overfitting to certain types of scenes, the positions and orientations of objects are randomly adjusted and the environment light is also rotated randomly. There are 36 pre-built scenes with 321 textures and 109 environment maps. The dataset provides 175 scenes with different conditions. In our experiment setting, we choose 100 scenes for training and 21 scenes for validation and testing. The other settings are the same as the ones in the 4D Light Field Dataset.

## Implementation Details

In our implementation, we use patch-wise training by randomly choosing gray-scale patches of size  $32 \times 32$  from the light field images in the training set. To avoid incorrect correspondences, when training on the 4D Light Field Dataset, we exclude patches from the areas containing objects with non-diffuse reflection and refraction, such as glass, metal and textureless regions. We manually mask out the non-diffuse reflection and refraction areas and remove the tex-

tureless regions where the mean absolute difference of the patch is less than 0.02 between the center pixel and other pixels. For the dataset of (Alperovich et al. 2018), we utilize whole scenes without any exclusion instead.

For training the network, given the predicted disparity map  $\hat{d}$ , the ground-truth disparity map  $d$ , and corresponding exclusion mask  $M$ , we use Adam optimizer (Kingma and Ba 2014) to minimize the following  $L1$  loss

$$\mathcal{L} = \sum_{x \in X} M(x) \cdot \|\hat{d}(x) - d(x)\|_1, \quad (2)$$

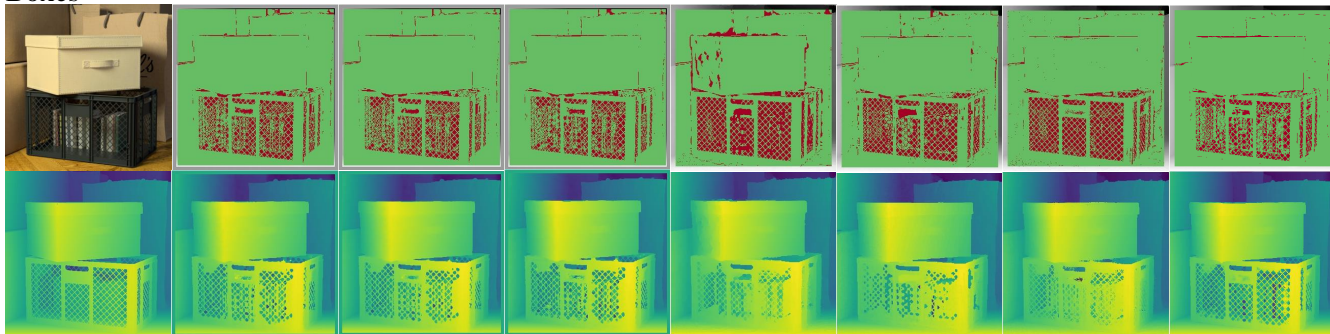
where  $x \in X$  denotes pixels in the image, and  $M(x) = 0$  if  $x$  is in the excluded regions; otherwise  $M(x) = 1$ .

The following parameters are set for training: the batch size is 12 and the learning rate is  $1e-3$ . The method is implemented using Keras with TensorFlow as the backend. Training took about one week on an NVIDIA GTX 1080Ti GPU.

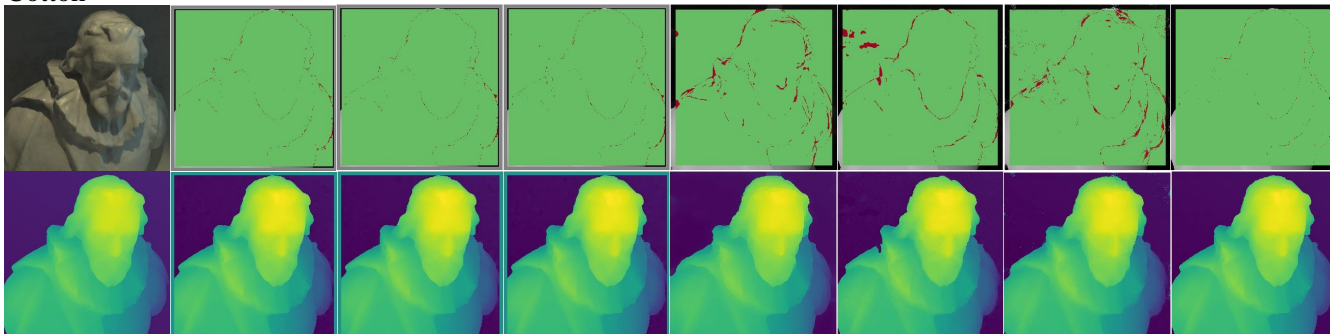
## Evaluation

For the quantitative evaluation, we mainly use the three test sets in the 4D Light Field Benchmark, which are named ‘‘Stratified’’, ‘‘Training’’ and ‘‘Test’’. Among the three test sets, the ground-truth depth maps of the ‘‘Stratified’’ and ‘‘Training’’ sets are available to the public while the ground truth of the ‘‘Test’’ set is not released. The ‘‘Test’’ set is often used as the benchmark for evaluating methods and, for obtaining the performance on this set, one has to submit the results to the benchmark website. There are several popular metrics for evaluation, including mean square errors (MSE)

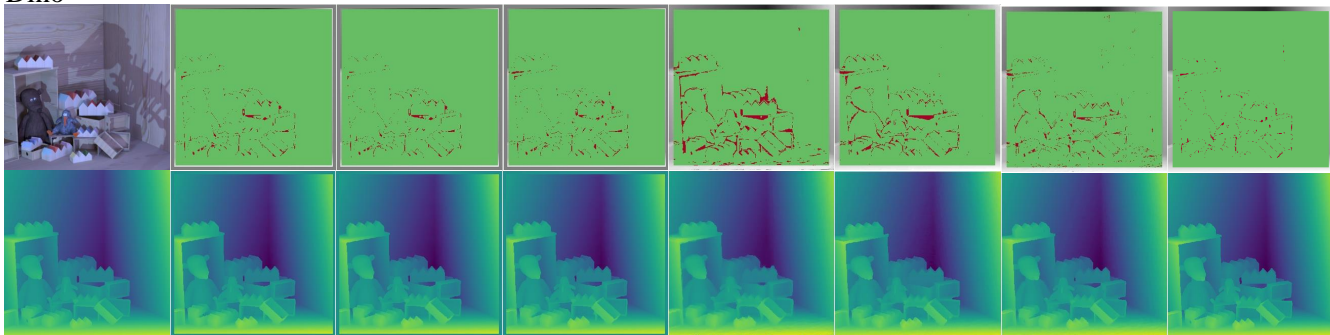
## Boxes



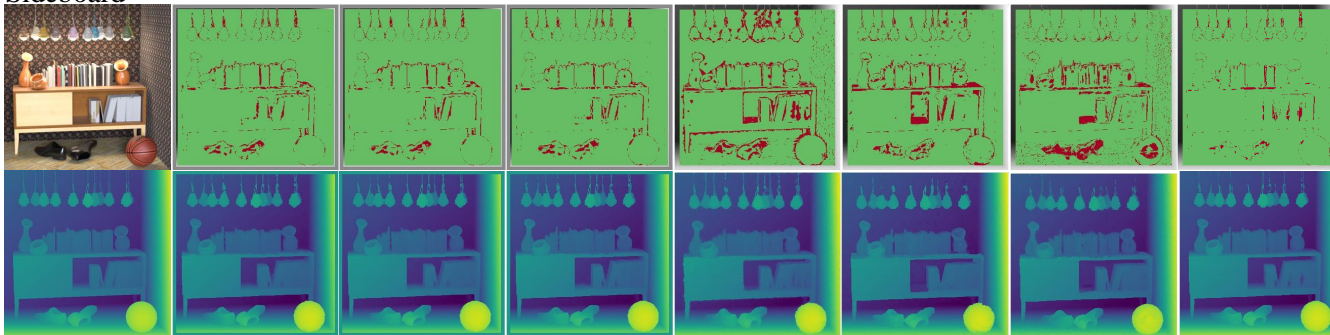
## Cotton



## Dino



## Sideboard



Ground truth

(a)

(b)

(c)

(d)

(e)

(f)

Ours

Figure 4: The estimated disparity maps of our method and compared methods for the four scenes in the “Training” set. For each scene, the first image of the top row is the center-view image, whose ground-truth disparity map is shown underneath the image. We then show the error map for Badpix 0.07 at the top and the disparity map at the bottom for each method. The compared methods include: (a) Epinet-fcn (b) Epinet-fcn-m (c) Epinet-fcn-9x9 (d) PS\_RF (e) EPN+OS+GC and (f) SPO. In the error map, red pixels indicate where the error of the estimated depth exceeds the threshold, 0.07 in this case, while the green pixels denote the ones with more accurate depth estimations. It is clear that our method has much fewer bad pixels than others.

MSE*100			BadPix 0.07		
Algorithm	MEDIAN No preview	AVG No preview	Algorithm	MEDIAN No preview	AVG No preview
LFatNet	1.159	1.904	LFatNet	2.901	3.756
CAPNet	0.973	1.974	OBER-cross+ANP	3.371	4.594
PDE-Net-e	1.315	2.097	Epinet-fcn-m	2.996	4.646
VommaNet_full	1.034	2.151	FusionNet	3.912	4.674
VommaNet9	1.096	2.218	Epinet-fcn	3.381	4.931
CrossResLFNet	1.504	2.341	CAPNet	2.655	4.953

BadPix 0.03			BadPix 0.01		
Algorithm	MEDIAN No preview	AVG No preview	Algorithm	MEDIAN No preview	AVG No preview
LFatNet	5.367	6.823	LFatNet	16.635	17.226
FusionNet	6.664	8.343	CAPNet	24.686	26.693
FocalStackNet	7.335	9.270	FusionNet	28.683	27.935
Epinet-fcn-m	7.731	9.537	CrossResLFNet	30.152	29.968
CAPNet	6.236	9.539	Epinet-fcn-m	33.721	31.898
RefocusedNet	8.501	9.846	Epinet-fcn9x9	35.863	32.982

Figure 5: The snapshot of the benchmark website (<https://lightfield-analysis.uni-konstanz.de/>). We submitted our results to the benchmark website. Our method is named “LFatNet”. It is ranked the first in the four popular error metrics as highlighted by red outlines.

	MSE*100	Badpix 0.01	Badpix 0.03	Badpix 0.07
Alperovich	18.559	59.635	56.654	45.578
Ours	2.866	19.806	7.123	3.413

Table 2: Comparison with Alperovich et al.’s method (Alperovich et al. 2018). The numbers reported here are the average errors over 21 scenes in the test set of the dataset (Alperovich et al. 2018).

and bad pixel ratios. The definition of the bad pixel ratio (Badpix) is the percentage of pixels whose absolute errors exceed the specified threshold, i.e.,  $|\hat{d}(x) - d(x)| > \epsilon$ , where  $\epsilon$  is the threshold. Three thresholds are often used for calculating the bad pixel ratios: 0.01, 0.03 and 0.07.

**Comparisons with state-of-the-art methods.** We compare our method with several top-ranked methods with publications on the 4D Light Field Benchmark, Epinet-fcn (Shin et al. 2018), Epinet-fcn-m (Shin et al. 2018), Epinet-fcn9x9 (Shin et al. 2018), PS\_RF (Jeon et al. 2017), EPN+OS+GC (Luo et al. 2017), and SPO (Zhang et al. 2016). Table 1 reports performance of our method and the compared methods on the “Stratified”, “Training” and “Test” sets of the 4D Light Field Dataset. Our method achieves the best performance in most scenes of the three sets. We have submitted our results to the benchmark website. Our method (LFatNet) ranks the first as shown in the snapshot of the benchmark website as shown in Figure 5 as of November 2019. Our method outperforms all methods in terms of MSE\*100, Badpix 0.07, Badpix 0.03 and Badpix 0.01. Figure 4 shows the visual results of our method and the compared methods on the four scenes of the “Training” set’. For each method, we show its depth map and error map in which red pixels indicate bad pixels. It is clear that our method has the lowest number of bad pixels in all scenes.

For the dataset of (Alperovich et al. 2018), Table 2 compares the performance of our proposed method and Alperovich et al.’s method (Alperovich et al. 2018). Our method outperforms their method significantly in all metrics. Fig-

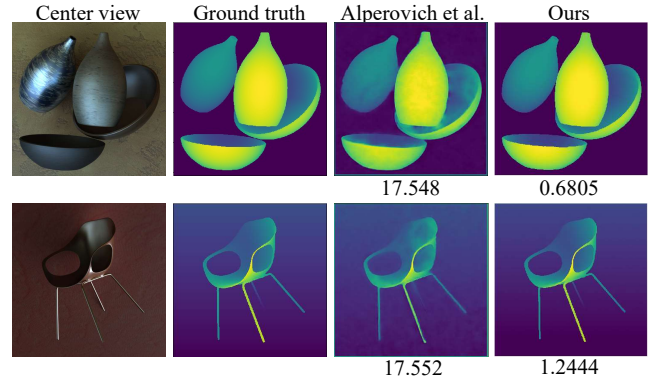


Figure 6: Comparison with (Alperovich et al. 2018) on their dataset. This figure shows the results of two examples using both our method and Alperovich et al.’s method. The numbers under the resultant depth maps are their MSE\*100 errors. Our method outperforms Alperovich et al.’s method significantly both quantitatively and qualitatively.

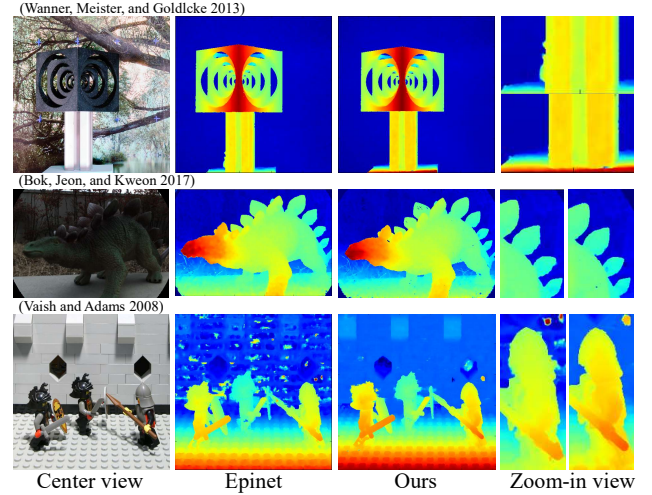


Figure 7: Evaluation of real-world light field images. Three light field images from previous papers are used. We compare our method with the Epinet (Shin et al. 2018). In the zoom-in views on the rightmost column, we compare our results (the right or bottom inset) with Epinet’s (the left or top inset) in detail. Our results generally contain fewer artifacts than Epinet’s results.

ure 6 shows the result depth maps of both our and their methods for two scenes. From this example, our method can handle the textureless and glossy regions better than Alperovich et al.’s method.

We have also tested our method on real-world light field images from previous work (Wanner, Meister, and Goldlücke 2013; Bok, Jeon, and Kweon 2017; Vaish and Adams 2008). Our model is trained using the synthetic images in the 4D light field dataset. We compare our results with the Epinet (Shin et al. 2018) in Figure 7. Our results generally exhibit fewer artifacts than the Epinet’s results.

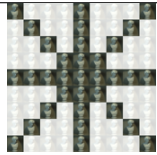
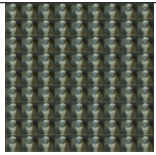
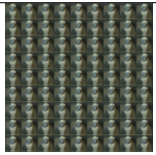
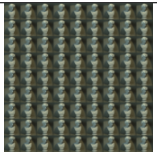
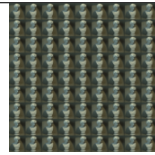
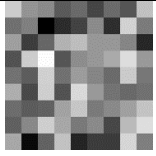
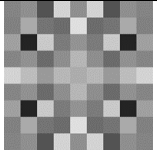
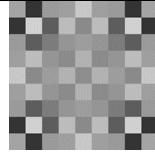
	Epinet (Shin et al. 2018)	Ours w/o attention	Ours w/ free attention	Ours w/ symmetric attention	Ours w/ radial attention
Input views					
#Parameters	5.12M	<b>5.00M</b>	5.06M	5.06M	5.06M
MSE*100	1.461	1.438	1.284	1.174	<b>0.982</b>
Badpix 0.07	3.91	3.44	3.08	2.73	<b>2.45</b>
Attention map	N/A	N/A			

Table 3: Comparisons of different types of attention maps. By imposing constraints on the structure of the attention map, the performance of our method can be significantly improved. We also compare our method with the Epinet (Shin et al. 2018) which has almost the same number of network parameters as our model. It uses a set of pre-selected views and has worse performance than our model.

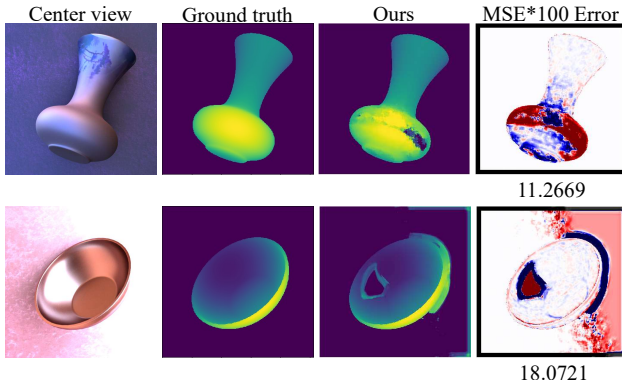


Figure 8: Failure cases. When the scene contains glossy materials or large textureless regions, our method could predict wrong disparity values, as shown in the two examples from the dataset (Alperovich et al. 2018). The numbers under the error maps are the MSE\*100 errors of the corresponding depth maps.

**Ablation study.** We have experimented with the three different types of attention maps listed in Figure 3. Table 3 compares their performance and the last row shows the learned attention maps of the three different types. It shows that the imposed constraints improve performance significantly. Table 3 also compares our models with the Epinet (Shin et al. 2018) which ranks high on the benchmark. It uses only pre-selected views and can be considered as a method with the pre-defined attention map. It has almost the same number of network parameters as our model, but with worse MSE and bad pixel ratio than ours.

**Discussions.** Although our method demonstrates good performance and provides a great improvement over previous methods on the benchmark and other datasets, it still has several limitations. As shown in Figure 8, when the scene

	Stratified			
	Backgammon	Dots	Pyramids	Stripes
Epinet-fcn	1.973	1.969	1.965	1.968
Epinet-fcn-m	10.67	10.64	10.70	10.60
Epinet-fcn9x9	1.886	2.034	2.027	2.032
PS_RF	979.8	969.9	929.7	1093
EPN+OS+GC	249.2	381.0	386.7	172.2
SPO	2195	2138	2256	1945
Ours	5.542	5.817	5.704	5.918

Table 4: Computation time (seconds) as reported by the authors on the 4D light field benchmark.

contains shining materials or has large textureless regions, our method would fail to estimate the accurate disparity values for those areas. As for the speed, Table 4 reports computation time for our method and several methods. Our method is reasonably fast.

## Conclusion

This paper proposes an attention-based view selection network for disparity estimation for light field images. By exploring the repetitive structure of light field cameras and the inherited redundancy within views, our method can utilize all views for estimating disparity maps both effectively and efficiently. Experiments demonstrate that our method achieves the best performance on a popular benchmark and other datasets. In the future, we would like to improve the proposed network so that it is more robust to glossy materials and textureless regions.

## Acknowledgements

This work was supported in part by Ministry of Science and Technology (MOST) and MOST Joint Research Center for AI Technology and All Vista Healthcare, under grants 107-2221-E-002-147-MY3 and 108-2634-F-002-004.

## References

- Alperovich, A.; Johannsen, O.; Strecke, M.; and Goldluecke, B. 2018. Light field intrinsics with a deep encoder-decoder network. In *Proceedings of IEEE CVPR*.
- Bok, Y.; Jeon, H.; and Kweon, I. S. 2017. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Bolles, R. C.; Baker, H. H.; and Marimont, D. H. 1987. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*.
- Chang, J.-R., and Chen, Y.-S. 2018. Pyramid stereo matching network. In *Proceedings of IEEE CVPR*.
- Chen, C.; Lin, H.; Yu, Z.; Kang, S. B.; and Yu, J. 2014. Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of IEEE CVPR*.
- Gortler, S. J.; Grzeszczuk, R.; Szeliski, R.; and Cohen, M. F. 1996. The lumigraph. In *Proceedings of ACM SIGGRAPH*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE CVPR*.
- Heber, S., and Pock, T. 2016. Convolutional networks for shape from light field. In *Proceedings of IEEE CVPR*.
- Heber, S.; Yu, W.; and Pock, T. 2017. Neural EPI-volume networks for shape from light field. In *Proceedings of IEEE ICCV*.
- Honauer, K.; Johannsen, O.; Kondermann, D.; and Goldluecke, B. 2016. A dataset and evaluation methodology for depth estimation on 4D light fields. In *Proceedings of ACCV*.
- Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of IEEE CVPR*.
- Jeon, H.-G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.-W.; and Kweon, I. S. 2017. Depth from a light field image with learning-based matching costs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jingyi Yu; McMillan, L.; and Gortler, S. 2002. Scam light field rendering. In *Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*.
- Kalantari, N. K.; Wang, T.-C.; and Ramamoorthi, R. 2016. Learning-based view synthesis for light field cameras. *ACM Transactions on Graphics*.
- Kendall, A.; Martirosyan, H.; Dasgupta, S.; and Henry, P. 2017. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of IEEE ICCV*.
- Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Levoy, M., and Hanrahan, P. 1996. Light field rendering. In *Proceedings of ACM SIGGRAPH*.
- Luo, Y.; Zhou, W.; Fang, J.; Liang, L.; Zhang, H.; and Dai, G. 2017. EPI-patch based convolutional neural network for depth estimation on 4D light field. In *Proceedings of the 24th International Conference on Neural Information Processing*.
- Shin, C.; Jeon, H.; Yoon, Y.; Kweon, I. S.; and Kim, S. J. 2018. EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of IEEE CVPR*.
- Srinivasan, P. P.; Wang, T.; Sreelal, A.; Ramamoorthi, R.; and Ng, R. 2017. Learning to synthesize a 4D RGBD light field from a single image. In *Proceedings of IEEE ICCV*.
- Tao, M. W.; Hadap, S.; Malik, J.; and Ramamoorthi, R. 2013. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of IEEE ICCV*.
- Vaish, V., and Adams, A. 2008. The (new) stanford light field archive. <http://lightfield.stanford.edu/index.html>.
- Wang, T.-C.; Zhu, J.-Y.; Hiroaki, E.; Chandraker, M.; Efros, A.; and Ramamoorthi, R. 2016. A 4D light-field dataset and cnn architectures for material recognition. In *Proceedings of ECCV*.
- Wanner, S., and Goldluecke, B. 2012. Globally consistent depth labeling of 4D light fields. In *Proceedings of IEEE CVPR*.
- Wanner, S., and Goldluecke, B. 2014. Variational light field analysis for disparity estimation and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Wanner, S.; Meister, S.; and Goldlücke, B. 2013. Datasets and benchmarks for densely sampled 4D light fields. In *Vision, Modeling & Visualization*.
- Yoon, Y.; Jeon, H.; Yoo, D.; Lee, J.; and Kweon, I. S. 2017. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*.
- Yu, Z.; Guo, X.; Ling, H.; Lumsdaine, A.; and Yu, J. 2013. Line assisted light field triangulation and stereo matching. In *Proceedings of IEEE ICCV*.
- Zbontar, J., and LeCun, Y. 2016. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*.
- Zhang, S.; Sheng, H.; Li, C.; Zhang, J.; and Xiong, Z. 2016. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*.
- Zhang, Y.; Lv, H.; Liu, Y.; Wang, H.; Wang, X.; Huang, Q.; Xiang, X.; and Dai, Q. 2017. Light-field depth estimation via epipolar plane image analysis and locally linear embedding. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of IEEE CVPR*.
- Zhong, T.; Jin, X.; Li, L.; and Dai, Q. 2019. Light field image compression using depth-based CNN in intra prediction. In *Proceedings of IEEE ICASSP*.