Fast Light-field Disparity Estimation with Multi-disparity-scale Cost Aggregation

Zhicong Huang[1,2], Xuemei Hu[1], Zhou Xue[2], Weizhu Xu[1], Tao Yue[1]
[1]School of Electronic Science and Engineering, Nanjing University, Nanjing, China
[2]ByteDance Inc.
zcong17huang@smail.nju.edu.cn, xuemeihu@nju.edu.cn, xuezhou@bytedance.com
weizhuxunju@smail.nju.edu.cn, yuetao@nju.edu.cn

Abstract

Light field images contain both angular and spatial information of captured light rays. The rich information of light fields enables straightforward disparity recovery capability but demands high computational cost as well. In this paper, we design a lightweight disparity estimation model with physical-based multi-disparity-scale cost volume aggregation for fast disparity estimation. By introducing a sub-network of edge guidance, we significantly improve the recovery of geometric details near edges and improve the overall performance. We test the proposed model extensively on both synthetic and real-captured datasets, which provide both densely and sparsely sampled light fields. Finally, we significantly reduce computation cost and GPU memory consumption, while achieving comparable performance with state-of-the-art disparity estimation methods for light fields. Our source code is available at https://github.com/zcong17huang/FastLFnet.

1. Introduction

Disparity estimation from light fields has become a promising way to derive disparity information with the arise of consumer-level light field cameras [20, 24]. Many algorithms have been proposed to estimate disparity maps from the light field images [25, 12, 9, 29].

With the progress of the artificial neural network, the learning-based algorithms are proposed [14, 3, 33, 28, 29, 26] and greatly improve the performance of disparity estimation. Considering the high-dimensional essence of this problem, 3D CNN architecture is widely used to handle the space-disparity representation for higher accuracy [14, 3, 29]. However, the extremely high computational cost and huge GPU memory consumption bring lots of difficulties to train and deploy the model in practice. Although several fast disparity estimation methods [28, 8, 31] have been proposed, they suffer from the loss of accuracy.

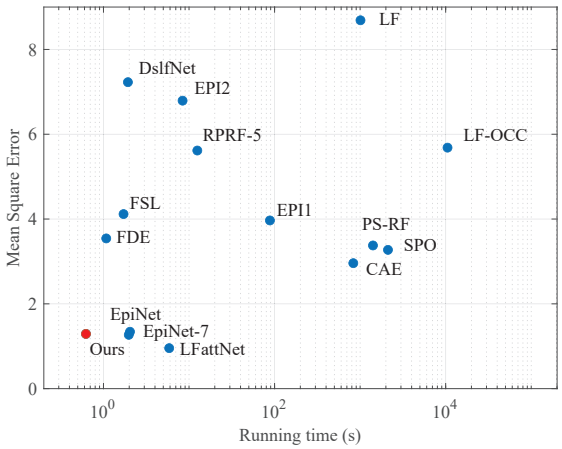


Figure 1. Comparison in performance and efficiency of light field disparity estimation algorithms.

In this work, we propose a fast and lightweight end-to-end deep architecture without using any 3D CNN modules for estimating disparity maps from light field images. Taking into account that different views of light fields have different disparity scales, we design a physical-based multi-disparity-scale cost aggregation module for efficient cost regularization. The proposed method can save computation and memory cost while providing pyramidal disparity information for better accuracy and robustness.

Abandoning the 3D CNN architecture may deteriorate the results in challenging regions with fine structures and detailed textures, thus an edge guidance sub-network is proposed to preserve subtle details by integrating edge information into the main network. The edge maps can highlight the regions where fine structures and detailed textures should be given more attention to, and guide the network handling these regions specially to achieve better results. Based on the edge-guided multi-disparity-scale cost aggregation, the proposed network could achieve competitive performance with the state-of-the-art methods with much faster computing speed and lower GPU memory consumption, as shown

in Fig. 1. In summary, the main contributions are as follows:

• We propose a fast and lightweight end-to-end network for light field disparity estimation.

• We present a physical-based multi-disparity-scale network for fast and high-performance cost volume regularization.

• We design an edge guidance sub-network to guide the disparity estimation with edge cues for better performance on challenging regions.

• We achieve competitive performance on par with state-of-the-art methods for both densely and sparsely sampled light fields while significantly reducing the computation cost and GPU memory consumption.

2. Related Work

Recently, with the development of neural networks, learning-based methods achieve state-of-the-art performances. Tsai *et al*. [29] propose to take all sub-aperture light field images as input to build a cost volume [14, 34] for regularization, which could get accurate disparity estimation. However, this method utilizes 3D CNN architecture for disparity regression, thus leading to heavy computational cost and huge GPU memory consumption.

Heber *et al*. [5, 6] used an artificial neural network to process the EPIs for the first time. They proposed an end-to-end deep network consisting of a U-shaped encoder and decoder to extract geometric and disparity information from light field images. Immediately afterward, Shin *et al*. [27] proposed a fully convolutional neural network [16] by considering the light field geometry for disparity estimation as well as a unique method to augment the light field images for training. However, these approaches are not robust enough to noise and cannot perform well in real-world data, and these EPI-based methods are not well suited for sparse light fields either.

Downsampling is useful to increase the receptive field while reducing computation. But at the same time, due to the loss of resolution, the performance in fine-grained details is sacrificed. Multi-scale aggregation has been proven to improve accuracy and reduce computation cost and GPU memory. GCNet [14] proposed an encoder-decoder architecture to get around the computational burden while preserving accuracy. Similarly, to learn more context information, PSMNet [3] used a stacked hourglass architecture in conjunction with intermediate supervision for cost volume regularization. SSPCV-Net [32] fused the cost volumes from the lowest level to the higher ones in a recursive way. AANet [33] constructed multi-scale cost volumes by correlating features at corresponding scales and proposed Intra-Scale Aggregation and Cross-Scale Aggregation modules of 3 pyramid levels for cost aggregation. However, these methods handled the cost volume as a 4D volume generally, downsampling both the spatial and disparity dimensions without distinction, which may greatly reduce the accuracy. Considering the physical structure of light field images, the cost volumes obtained from views with different lengths of baselines have different disparity scales. Therefore, for light field disparity estimation, the physical-based multi-disparity-scale cost aggregation can better adapt to the intrinsic structure of light fields and achieve more accurate estimation results with less computation cost.

Recently, edge information is proposed to effectively improve the performance of various computer vision tasks [17, 1, 18, 35]. Inspired by these methods, we introduce edge guidance to the multi-disparity-scale cost aggregation to guide the disparity estimation with edge cues and further improve the performance.

In all, in this paper, we propose a fast and light-weight end-to-end network based on edge-guided multi-disparity-scale cost volume aggregation to realize elegant performance on both estimation accuracy and computation cost.

3. Methodology

In order to estimate disparity maps for both densely and sparsely sampled light field images with a low computation cost and GPU memory consumption, we propose FastLFnet, a fast light-field disparity estimation network that can not only produce accurate estimations but also significantly speed up inference. The overview of the architecture of FastLFnet is illustrated in Fig. 2. Considering the redundancy of light field images, instead of using all sub-aperture images as input, we only use sub-aperture images along with two cross directions, i.e. horizontal and vertical of the center-view image, to estimate the disparity for reducing the computational cost as much as possible. Views at the same angular distance from the center view have the same disparity scale and are classified as a type of anchor, while different kinds of anchors are marked with different colors in Fig. 2. Details are discussed in this section.

3.1. Feature Extraction For Edge Guidance

As shown in Fig. 2, the input images are fed into the feature extraction module to produce an effective feature representation. Here we use basic residual blocks [4] for extracting reliable features and in the deep layers, we use convolutions with a stride of 2 for downsampling. Feature maps are downsampled to four scales followed by a bilinear interpolation to upsample these features of different scales to the original size. Then the features of different levels are concatenated and fed into a fusion layer for multi-level fusion. Before the final output of the feature extraction module, we pass the feature maps through a BAM module [22] to add attention to regions that are important for matching.

We propose to utilize an attention mechanism with the edge information extracted from the center view image to guide the network to focus more on fine structure and edge details. Specifically, we propose to extract edge features

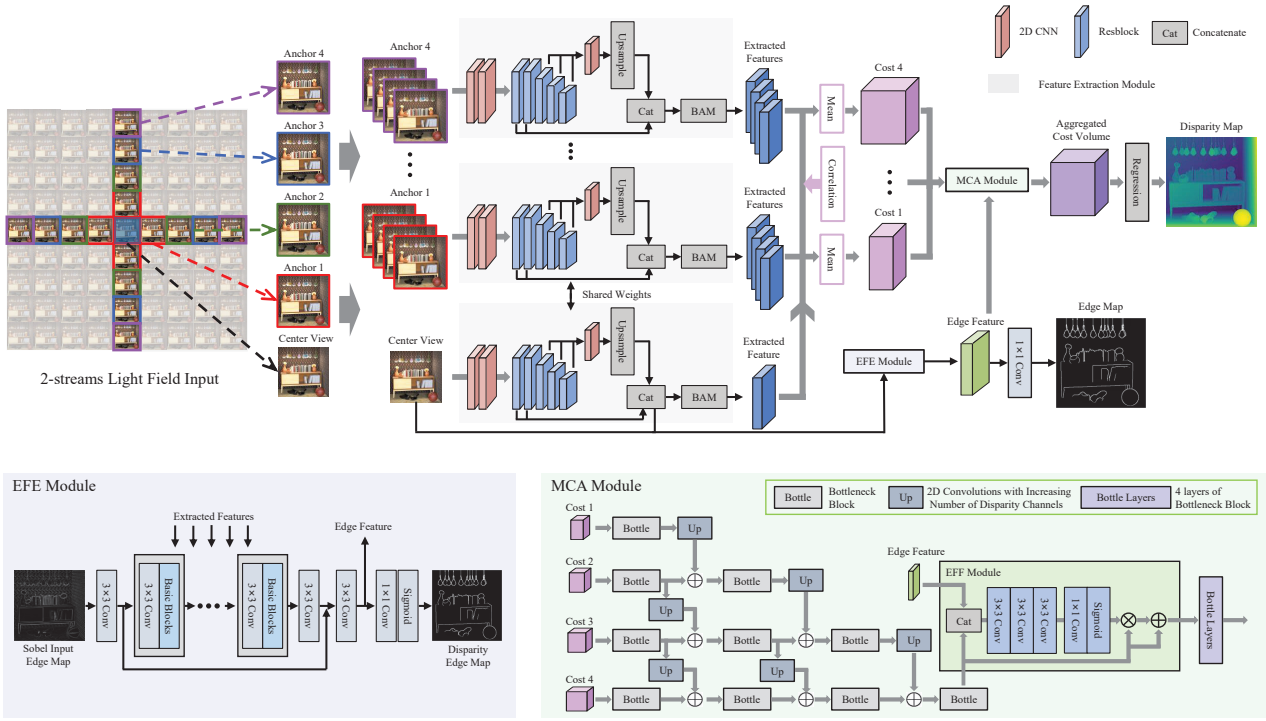


Figure 2. Overview of the proposed FastLFnet, the overall FastLFnet is at the top right of the figure. The 2-streams light field input representation is at the top left of the figure, while the EFE module and MCA module are at the bottom.

with the edge feature extraction (EFE) module from the center view image and the extracted edge feature maps are then integrated into a pixel-wise edge feature fusion (EFF) module to guide the disparity estimation (refer to Sec. 3.2 for details). As shown in Fig. 2, an initial edge map is obtained from the center image with the $Sobel$ edge detection operation [10]. Since the multi-level features from the feature extraction module carry rich structural information which is significant to the generation of disparity edge map, these representations with the same spatial resolution are fed into the EFE module (refer to the supplementary material for network details). We use the basic residual blocks to extract higher-level features and get the output edge features. The output edge map is produced by a 1×1 convolution layer and Sigmoid function.

The proposed edge guidance sub-network is effectively and efficiently combined with the main network of disparity estimation. On the one hand, the EFE module directly utilizes the feature maps from the feature extraction module as a prior, which can largely reduce the parameters and computation cost and can in turn impact the main network when training. On the other hand, we integrate the obtained edge feature maps into the EFF module to produce an adaptive weighted attention cost, which is fused with the aggregated cost volume to guide the disparity estimation.

3.2. Multi-disparity-scale Cost Aggregation

To aggregate the cost volume from different views, the extracted feature maps at different disparities are commonly concatenated to 4D $(height \times width \times disparity \times feature\_size)$ and 3D convolutions are required [14], which is computationally expensive and requires large GPU memory. To overcome these problems, we calculate the cost volume while eliminating the feature dimension as in [19], i.e.

$$C\left(d,h,w\right)=\frac{1}{N}\left\langle F_c\left(c,h,w\right),f_{\text{warp}}\left[F_s\left(c,h,w\right),d\right]\right\rangle, \quad (1)$$

where $f_{\text{warp}}\left[\cdot,\cdot\right]$ denotes the warp function to warp the surround feature F_s to the center feature F_c for a given disparity level d. $\langle\cdot,\cdot\rangle$ denotes the inner product along the feature dimension and N is the channel number of extracted features. $C(d,h,w)$ is the cost for aggregation at spatial location $(h,\ w)$ and disparity level d.

To take into consideration that the views of different anchors have different disparity scales, we propose a strategy of pyramid cost volumes to construct costs of different disparity scales for each kind of anchor. With views of the same anchor, the corresponding number of costs are first constructed, and then perform a mean operation, resulting in one output cost for each anchor. Specifically, defining the maximum disparity of the innermost views as $d_{\max}$, the disparity range of the innermost views is $[-d_{\max},\ d_{\max}]$, resulting in the disparity level of $2 * d_{\max} + 1$. As for the outermost views, the disparity range is $[-4d_{\max},\ 4d_{\max}]$, i.e., the disparity level of $2 * 4d_{\max} + 1$, leading to a more precise disparity shift. Here, the *maximum disparity* refers to the maximum absolute value of the disparity of the view, and the *disparity level* refers to the number of discrete dis-

	1-stream				2-streams				4-streams			
Channel Numbers	F4	F8	F16	F32	F4	F8	F16	F32	F4	F8	F16	F32
MSE_x100	1.902	1.705	2.068	1.955	1.756	1.546	1.218	1.653	1.815	1.523	1.437	1.476
Running time / s	0.162	0.212	0.357	0.804	0.256	0.354	0.593	1.438	0.416	0.592	1.065	2.591
GPU Memory / GB	1.605	1.745	1.911	2.773	1.695	1.877	2.107	3.305	1.821	2.127	2.649	4.349
Parameters / M	0.267	0.489	1.366	4.854	0.267	0.489	1.366	4.854	0.267	0.489	1.366	4.854

Table 1. Ablation results on the HCI benchmark for different sub-sets of views and channel numbers of features.

parities in the interval from the minimum disparity (negative) to the maximum disparity (positive).

Having obtained cost volumes of different disparity scales, we propose a layer-by-layer multi-disparity-scale cost aggregation architecture to integrate these pyramid cost volumes. Detailed structure is illustrated in the MCA module of Fig. 2. To fuse the disparity dimensions of different views, our proposed architecture integrates the cost volumes layer by layer, and finally obtains one output cost that aggregates cost volume information of different scales. From coarse to fine scale, our method aggregates the feature information along the disparity dimension as well as spatial dimension to improve precision and accuracy.

Furthermore, we propose a pixel-wise edge feature fusion (EFF) module to utilize the edge-attention-guided mechanism to guide each pixel in the aggregated cost volume to learn its own weight. After getting edge feature maps from the edge feature extraction (EFE) module, we concatenate these two features and then process them with three layers of 3×3 convolution and one layer of 1×1 convolution. The second layer of 3×3 convolution reduces the number of channels to the same as the aggregated cost. We adopt the sigmoid function to weight the cost with edge guidance. The guided cost volume is obtained through:

$$C_{d,h,w} = (1 + W_{d,h,w}) \odot C'_{d,h,w}, \tag{2}$$

where $C'_{d,h,w}$ is the aggregated cost volume and $C_{d,h,w}$ is the output cost that has been guided. $W_{d,h,w}$ represents the weighted attention map to guide the cost volume to focus more on edge details. $\odot$ denotes element-wise multiplication. At last, the output from the pixel-wise EFF module is regularized by 4 layers of bottleneck [4].

3.3. Disparity Regression and Loss

We utilize soft argmin operation as in [14] for disparity regression to estimate the continuous and precise disparity maps. First, we use the softmax operation $\sigma(\cdot)$ to calculate the probability for a probability volume. The final predicted disparity $\hat{d}$ is then calculated as the sum of each disparity d weighted by its normalized probability, i.e.

$$\hat{d} = \sum_{-D_{\max}}^{D_{\max}} d \times \sigma\left(C_d\right), \tag{3}$$

where $D_{\max}$ denotes the maximum disparity of the outermost views and C_d is the predicted cost of the disparity d. This regression is more robust than classification-based methods with sub-pixel precision.

For our method that can perform disparity estimation and edge guidance simultaneously, we propose a three-step training strategy. First, we train our FastLFnet without edge guidance to get coarse results. We adopt the smooth $L1$ loss function for the first step of training which has low sensitivity to outliers. The disparity loss L_{disp} is defined as

$$L_{\text{disp}}\left(d, \hat{d}\right) = \frac{1}{M} \sum_{(i,j)} \text{smooth}_{L_1}\left(d_{i,j}, \hat{d}_{i,j}\right), \tag{4}$$

where M is the number of pixels to be predicted, d is the ground-truth disparity, and $\hat{d}$ is the predicted disparity.

In the second step, we combine the edge guidance subnetwork with the feature extraction module while fixing the weights of other parts of the network. We only input the center view of the light field images and due to the lack of edge map labels of the corresponding ground-truth disparity, we manually mark out the edge maps on the dataset for training. We use weighted BCE loss L_{BCE} to supervise the predicted edge maps and the weights are defined as

$$\alpha = \lambda \cdot \frac{|Y^+|}{|Y^+| + |Y^-|}, \quad \beta = \frac{|Y^-|}{|Y^+| + |Y^-|}, \tag{5}$$

where α and β denote weights for negative and positive samples. Y^+ and Y^- denote positive sample set and negative sample set respectively. λ controls the weight of positive over negative samples.

Finally, the whole FastLFnet is jointly trained together. Because of the second step of training, our network has been able to predict the edge information of the disparity map. To get better performance in edge structures, we define an edge loss L_{edge}, which is effective guidance for disparity estimation:

$$L_{\text{edge}}\left(e, \hat{e}\right) = \frac{1}{M} \sum_{(i,j)} \text{smooth}_{L_1}\left(e_{i,j}, \hat{e}_{i,j}\right), \tag{6}$$

where e is the edge map of the ground-truth disparity, and $\hat{e}$ is the edge map of the predicted disparity. Edge maps are calculated by the $Sobel$ edge detection operation. Hence the overall loss at this step is defined as $L = L_{\text{disp}} + \lambda_b L_{\text{BCE}} + \lambda_e L_{\text{edge}}$, where λ_b and λ_e are the weights for balancing different loss terms.

4. Experiments

In this section, we first introduce the datasets and describe experimental settings. The ablation studies are conducted to evaluate the contribution of proposed modules. Finally, we demonstrate our method with both quantitative and qualitative results by comparing it with state-of-the-art methods on both synthetic and real-world light fields.

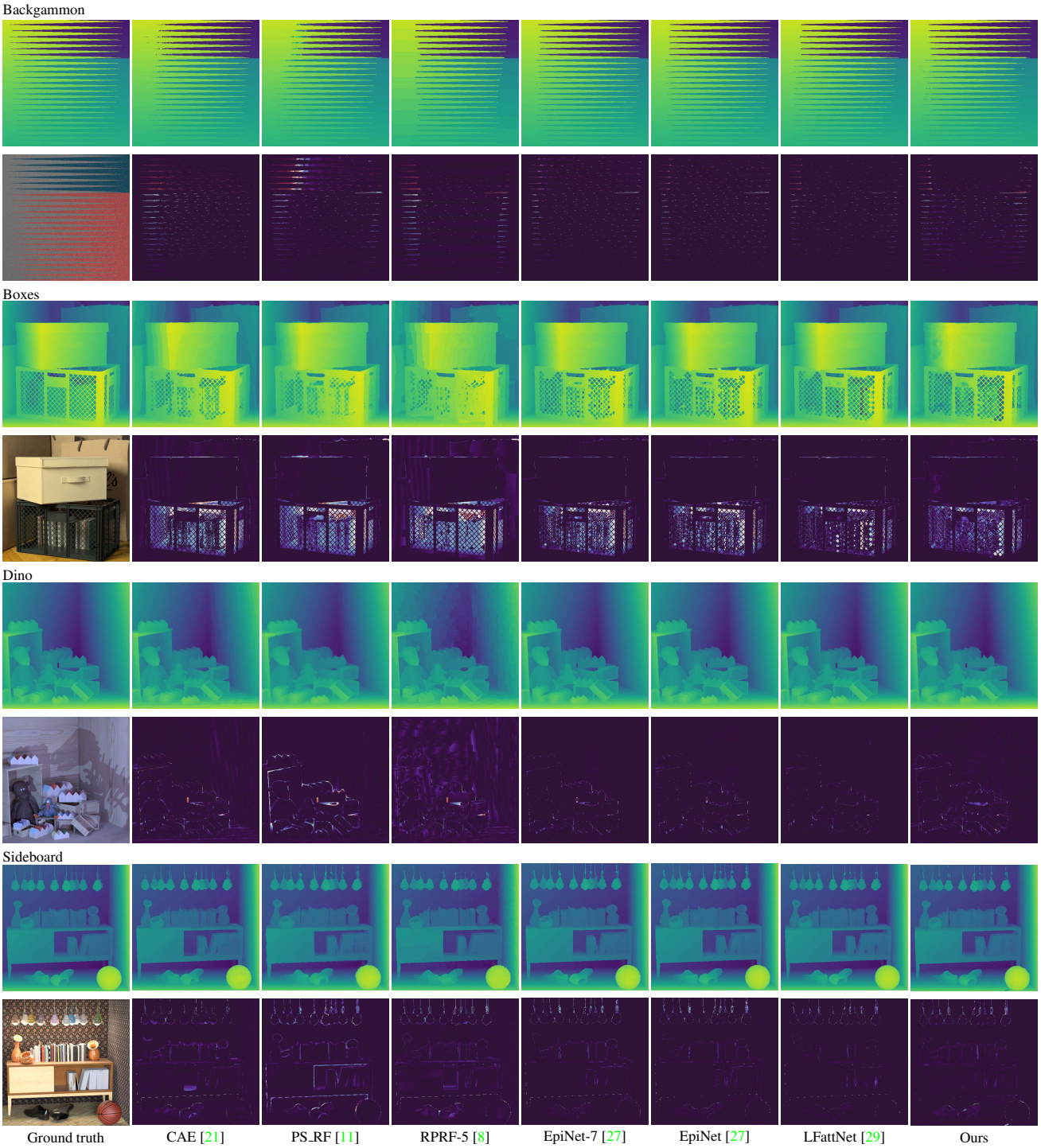


Figure 3. Qualitative results of our method and other compared methods. For each scene, the image in the lower-left corner represents the center-view image of input light fields. The first rows are the estimated disparity results and the second rows show the corresponding absolute error maps (bright color denotes large errors).

4.1. Dataset

4D light field dataset [7] A synthetic dataset with 28 carefully designed scenes. The scenes are composed of various challenging objects and structures, and partitioned into four subsets: *Stratified*, *Test*, *Training*, and *Additional*. Each light field has a spatial resolution of 512×512 and an angular resolution of 9×9 with a disparity range [-4, 4] pixels, while most disparities lie within the range of [-1.5, 1.5] pixels. In our experiment, we randomly sample 32×32 gray-scale patches for training, use the data augmentation

Processing methods	One Scale	w/o BAM	w/o Edge	FastLFnet
MSE_x100	1.650	1.492	1.844	1.218
Running time / s	0.725	0.582	0.576	0.593
GPU Memory / GB	2.323	2.103	2.189	2.107
Parameters / M	1.281	1.361	0.982	1.366

Table 2. Comparison of the contributions of each component we proposed.

strategy and exclude non-diffuse reflection and refraction regions as in [27]. We use the subset of *Additional* for training and the others for validation and testing.

Sparse light field dataset [26] A sparsely sampled synthetic light field dataset with large baselines between views. The dataset contains 53 scenes with a large disparity range, i.e. within the interval of [-20, 20] pixels, which is comparable to the real captured light fields with camera arrays. The scenes contain textureless background, specular reflection, diffusion and object occlusion. Each light field has the same spatial resolution (512×512), and angular resolution (9×9) as those in the 4D Light Field Dataset [7]. Because of the large disparity range, we crop the images to size $H = W = 128$ during training, and use four scenes (*Bear*, *Two_vases*, *Surfboard* and *Robots*) for validation, four scenes (*Furniture*, *Lion*, *Toy_bricks*, and *Electro_devices*) for test and the others for training.

4.2. Implementation Details

The proposed network is implemented with PyTorch platform [23] and Adam [15] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) is used as optimizer. We trained our model end-to-end for 40000 iterations with a batch size of 16 for the second step and a batch size of 8 for the others. The initial learning rate was set to 0.001 and was decayed by multiplying 0.2 for the second step and decayed by multiplying 0.5 for the others every other 10000 iterations. The loss weights in the third step of training are set to $\lambda_b = 100$ and $\lambda_e = 2.2$ respectively. The parameter λ in loss L_{BCE} is set to 1.1. The whole training process takes about 17 hours with one Nvidia 2080Ti GPU.

Methods	CAE [21]	PS_RF [11]	RPRF-5 [8]	EpiNet-7 [27]	EpiNet [27]	LFattNet [29]	w/o Edge	Fast-LFnet
Boxes	8.162	8.771	10.333	6.042	5.845	**3.869**	5.658	*4.260*
Cotton	1.704	1.227	0.949	**0.206**	0.235	*0.220*	0.318	0.339
Dino	0.376	0.730	0.603	0.162	*0.147*	**0.090**	0.350	0.184
Sideboard	0.860	1.899	1.224	0.814	0.794	**0.518**	1.070	*0.742*
Backgammon	4.762	5.559	3.024	*1.500*	1.893	1.762	2.658	**1.488**
Dots	4.589	7.881	20.114	*1.155*	1.549	**0.959**	4.508	3.070
Pyramids	0.047	0.043	0.042	0.008	*0.007*	**0.004**	0.010	0.018
Stripes	3.171	0.905	8.643	0.265	0.264	**0.220**	0.854	*0.231*
Average	2.959	3.377	5.616	*1.269*	1.342	**0.955**	1.928	1.291
Fattening	7.614	6.597	5.262	4.702	4.990	**3.810**	5.752	*4.300*
Thinning	**1.153**	2.237	2.568	1.548	*1.430*	2.230	3.499	2.427
Running time / s	832.081	1412.623	12.498	1.976	2.041	5.862	**0.611**	*0.624*
GPU Memory / GB	-	-	-	4.319	5.103	10.953	*2.189*	**2.107**
Parameters / M	-	-	-	5.116	5.118	5.058	**0.982**	*1.366*

Table 3. Quantitative comparison (i.e., MSE_x100) with other state-of-the-art methods on the 4D Light Field Dataset [7]. The best and secondary results are indicated by bold and italic text respectively.

4.3. 4D Light Field Dataset

Ablation studies First, we conduct experiments to evaluate the trade-off between performance and efficiency on the 4D Light Field Dataset. Here we use three different sub-sets of views of light field images as input, i.e. *1-stream* (horizontal), *2-streams* (horizontal and vertical), and *4-streams* (horizontal, vertical, and diagonal). Additionally, we evaluate the impact of the channel number of extracted features. For each input stream, we use four different channel numbers (4, 8, 16, and 32) for comparison. As shown in Tab. 1, incorporating more views or a larger feature number improves the performance while at the cost of much higher computational cost. Meanwhile, simply increasing the feature number may lead to overfitting problems. Using fewer views or smaller channel number reduce computation cost and GPU consumption, while the accuracy of results drops accordingly. By taking both the performance and efficiency into consideration, we choose the network with *2-streams* input and 16 channels, which performs elegantly well with relatively high efficiency.

Then, we conduct ablation studies to compare a number of different model variants for FastLFnet on the 4D Light Field Dataset, so that the importance of the two key contributions of the proposed method, i.e., edge guidance sub-network and multi-disparity-scale cost aggregation, can be evaluated. In addition, we also evaluate the effectiveness of the BAM module on the performance of the network. The comparison results are shown in Tab. 2 and we can clearly justify our design choices for FastLFnet. Here, MSE_x100 denotes 100×Mean Square Errors (MSE). The proposed edge guidance sub-network provides edge cues to refine object details for more accurate estimation results. Multi-disparity-scale cost volumes and the proposed cost aggregation architecture can bring more useful information

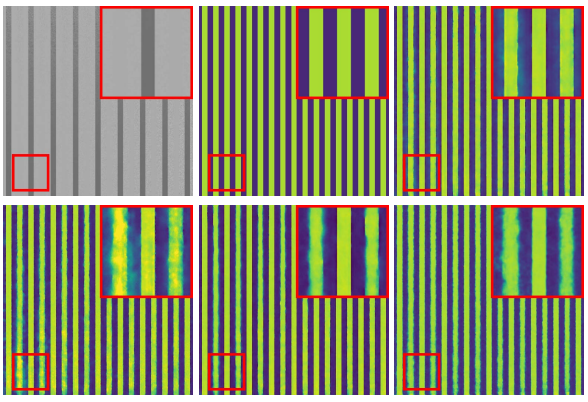


Figure 4. Visual comparisons of ablation study to show the contribution of each component of the proposed FastLFnet. From top to bottom and left to right: Center view, Ground truth, FastLFnet, One Scale, w/o BAM and w/o Edge. Our integrated FastLFnet produces sharper and better results in thin structures and edge details.

and greatly improve the performance of disparity estimation with relatively lower computational cost. Besides, the attention mechanism of the BAM module helps to improve the overall disparity estimation without much extra computation overhead.

To qualitatively analyze the importance of each component of our FastLFnet, we further show the results of different ablation setting upon the *Stripes* scene of the 4D Light Field Dataset in Fig. 4. As can be compared, utilizing cost aggregation with only one-disparity-scale leads to poor performance, and the disparity of the low contrast stripes is of bad effect, which further demonstrates the essential role of MCA in improving the estimation accuracy. Besides, through comparison, it could be found that the BAM module can help to improve the disparity accuracy in occlusion boundaries. It can be seen that with the integration of edge information, the disparity results of fine structures and textureless regions are recovered with much sharper details. The qualitative comparison further demonstrates the effectiveness of the proposed network structure.

Comparison with state-of-the-art methods We compare the performance of our FastLFnet with both traditional and learning-based state-of-the-art methods (CAE [21], PS_RF [11], RPRF-5 [8], EpiNet-7 [27], EpiNet [27], LFattNet [29]) for light field disparity estimation. We use two subsets (*Training* and *Stratified*) on the 4D Light Field Dataset and compare the performance by MSE. Since EpiNet [27] does not use zero-padding, their results lose 11 pixels at each border. Therefore, for a fair comparison, the margins of 11 pixels are cropped for all the methods in Tab. 3, which is also used in [26]. (This explains that the results of our method in Tab. 3 are slightly different from those in Tab. 1 and Tab. 2.).

As shown in Tab. 3, our method completely outperforms the first three methods for all the scenes in terms of MSE_x100 with a large margin and performs comparably with the other state-of-the-art methods [27, 29]. The comparison of MSE and computational cost of these methods are summarized in Tab. 3. For a fair comparison, the learning-based methods are all tested on an NVIDIA GTX 1080Ti GPU and here we use the average running time in all 12 scenes of 3 subsets (*Stratified*, *Test*, and *Training*) on the 4D Light Field Dataset. As shown, our method requires much less inference time and GPU memory, while at the same time with a comparable disparity estimation accuracy.

Light fields	MSE					
	EBSM [9]	OHLF [13]	SfIfNet [6]	EpiNet [27]	DslfNet [26]	FastLFnet
Furniture	0.37	1.94	9.18	1.73	0.42	**0.17**
Lion	0.10	0.87	1.59	3.41	0.09	**0.05**
Toy_bricks	0.22	1.10	3.70	0.36	0.57	**0.16**
Elec_dev	0.20	0.63	7.82	0.74	0.20	**0.09**
Average	0.22	1.14	5.57	1.56	0.32	**0.12**

Table 4. Results of the performance comparison on the Sparse Light Field Dataset in terms of MSE.

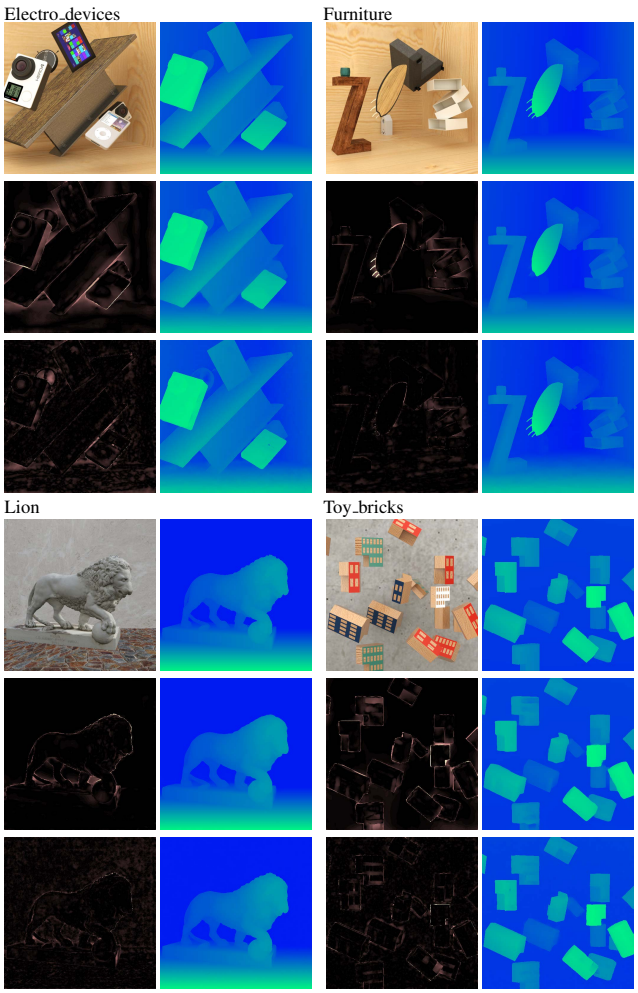
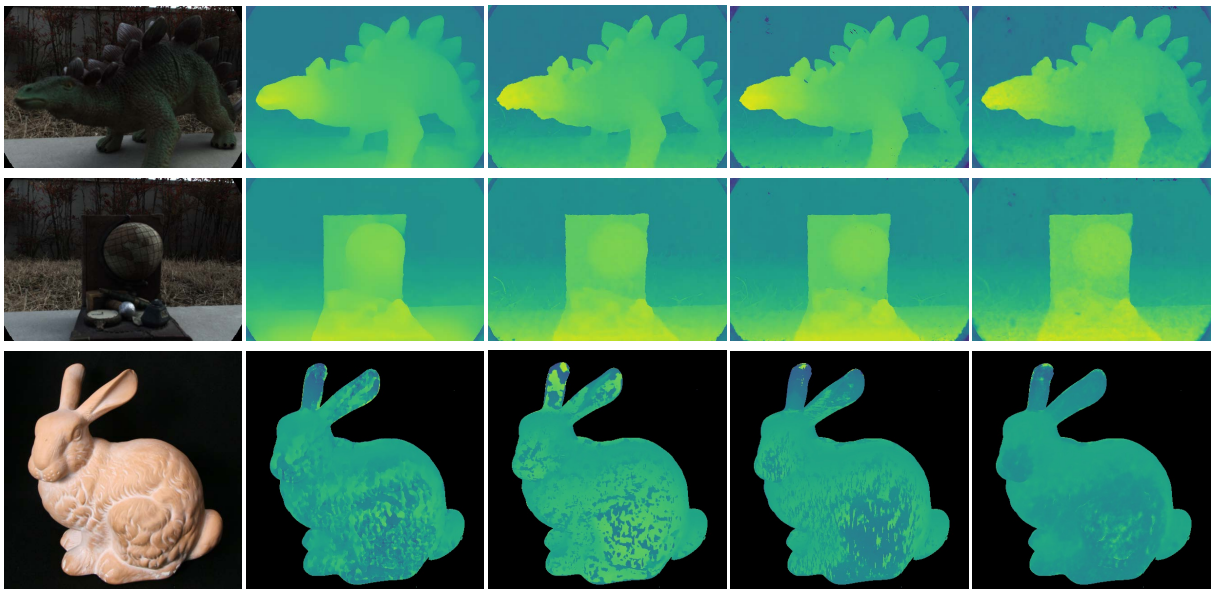


Figure 5. Qualitative results comparison with [26] on the Sparse Light Field Dataset. For each scene, the second row represents the results of [26] and the third row shows our performance.

In addition, for more evaluations on occlusion and edge regions, we apply Fattening and Thinning metrics, i.e., the fraction of false foreground/background pixels (see [7] for details), of scene *Backgammon* to evaluate the accuracy of results on occlusion boundaries between background and foreground. The proposed method also achieves comparable results with others. For more results and discussions of edge and discontinuity regions, please refer to the supplementary material.

For visual comparison, we show the disparity estimation results of different methods on four scenes of the 4D Light Field Dataset in Fig. 3. For each method, we show the estimated disparity maps and the corresponding absolute error maps. Combined with the error maps, we can see that in areas with fine structures and rich edge details, as the grids in *Boxes*, jagged foreground plane in *Backgammon*, and jagged edges of the table in *Sideboard*, our method performs comparably with the state-of-the-arts.



(a) Center view (b) DslfNet [26] (c) EpiNet [27] (d) LFattNet [29] (e) Ours

Figure 6. Qualitative results of real-world light field images. The dataset is provided by [2] and [30]

4.4. Sparse Light Field Dataset

Most of the existing light field depth estimation methods [27, 31, 6] mainly focus on the densely sampled light fields, like 4D Light Field Dataset [7], and few of them can handle the sparsely sampled ones. We hereby experimentally prove that our proposed method can not only obtain competitive results for densely sampled light fields but also achieve elegant performance in the sparse light fields dataset [26].

The large disparity range leads to huge 4D cost volumes, so the 3D-CNN-based methods [29] cannot be applied in the sparse cases directly for their excessive GPU memory demand. We compare our method with those without 3D CNNs, e.g., Bayesian-based EBSM [9], EPI-based EpiNet and SflfNet [27, 6], and optical-flow-based OHLF and DslfNet [13, 26]. Our model uses 9 sparsely sampled views of sparse light field images for comparison. The quantitative comparison results of 4 scenes are shown in Tab. 4, and we compute MSE for evaluation. We can see that the proposed method achieves much better results than others.

The qualitative results are shown in Fig. 5. For each scene, the top row represents the center view image and the disparity ground truth, and the second and third rows show absolute error maps (left) and disparity maps (right) of [26] and our method respectively. As shown in Fig. 5, the proposed method achieves more accurate results, especially in fine structures and edge details, demonstrating the superiority of our method on not only dense light field data, but also sparse light field data.

4.5. Real-world Results

Challenging real captured datasets usually suffer from deep discontinuities, blurred scenes and various noises problems. To evaluate the performance of the proposed network on real captured light fields, we directly use the model trained on the 4D Light Field Dataset and test our method on the real-world light field images captured by a Lytro Illum camera [2], and the (New) Stanford Light Field Archive [30]. We compare our results with [26], [27] and [29]. Some results are shown in Fig. 6. As shown, our method performs elegantly well and the estimated disparity is comparable or even better than the other methods, further demonstrating our method.

5. Discussions and Conclusions

To reduce the computational complexity, we abandon the 3D CNN architecture, leading to more errors at discontinuous regions. Although the proposed edge guidance mechanism greatly improves the MSE performance, the results on some other metrics like Thinning are still not particularly good. This could be improved by introducing specifically designed network modules and corresponding loss functions in the future.

In this paper, we propose a fast and lightweight end-to-end deep architecture for estimating disparity maps from light field images. A multi-disparity-scale cost aggregation module is proposed to regularize the cost volume efficiently, and an edge-based guidance sub-network is proposed to further improve the performance on the challenging regions with fine structures and detail textures. The method achieves competitive performance with state-of-the-art methods with much faster computing speed and lower GPU memory consumption.

Acknowledgement

This work was supported by NSFC Projects 61971465, and Fundamental Research Funds for the Central Universities, China (Grant No. 0210-14380184).

References

[1] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE, 2019. 2

[2] Yunsu Bok, Hae-Gon Jeon, and In So Kweon. Geometric calibration of micro-lens-based light field cameras using line features. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):287–300, 2016. 8

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[5] Stefan Heber and Thomas Pock. Convolutional networks for shape from light field. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3746–3754, 2016. 2

[6] Stefan Heber, Wei Yu, and Thomas Pock. Neural epi-volume networks for shape from light field. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2260, 2017. 2, 7, 8

[7] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016. 5, 6, 7, 8

[8] Chao-Tsung Huang. Robust pseudo random fields for light-field stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11–19, 2017. 1, 5, 6, 7

[9] Chao-Tsung Huang. Empirical bayesian light-field stereo matching by robust pseudo random field modeling. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):552–565, 2018. 1, 7, 8

[10] FG Irwin et al. An isotropic 3x3 image gradient operator. *Presentation at Stanford AI Project*, 2014(02), 1968. 3

[11] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Depth from a light field image with learning-based matching costs. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):297–310, 2018. 5, 6, 7

[12] Hae-Gon Jeon, Jaesik Park, Gyeongmin Choe, Jinsun Park, Yunsu Bok, Yu-Wing Tai, and In So Kweon. Accurate depth map estimation from a lenslet light field camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1547–1555, 2015. 1

[13] Xiaoran Jiang, Mikaël Le Pendu, and Christine Guillemot. Depth estimation with occlusion handling from a sparse set of light field views. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 634–638. IEEE, 2018. 7, 8

[14] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 1, 2, 3, 4

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[17] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998, 2017. 2

[18] Cheng Ma, Yongming Rao, Yean Cheng, Ce Chen, Jiwen Lu, and Jie Zhou. Structure-preserving super resolution with gradient guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7769–7778, 2020. 2

[19] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 3

[20] Ren Ng. Lytro redefines photography with light field cameras, 2018. 1

[21] In Kyu Park, Kyoung Mu Lee, et al. Robust light field depth estimation using occlusion-noise aware data costs. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2484–2497, 2017. 5, 6, 7

[22] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 2

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 6

[24] Christian Perwass and Lennart Wietzke. Single lens 3d-camera with extended depth-of-field. In *Human Vision and Electronic Imaging XVII*, volume 8291, page 829108. International Society for Optics and Photonics, 2012. 1

[25] Neus Sabater, Mozhdeh Seifi, Valter Drazic, Gustavo Sandri, and Patrick Pérez. Accurate disparity estimation for plenoptic images. In *European Conference on Computer Vision*, pages 548–560. Springer, 2014. 1

[26] Jinglei Shi, Xiaoran Jiang, and Christine Guillemot. A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing*, 28(12):5867–5880, 2019. 1, 6, 7, 8

[27] Changha Shin, Hae-Gon Jeon, Youngjin Yoon, In So Kweon, and Seon Joo Kim. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4748–4757, 2018. 2, 5, 6, 7, 8

[28] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–204, 2019. 1

[29] Yu-Ju Tsai, Yu-Lun Liu, Ming Ouhyoung, and Yung-Yu Chuang. Attention-based view selection networks for light-field disparity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12095–12103, 2020. 1, 2, 5, 6, 7, 8

[30] Vaibhav Vaish and Andrew Adams. The (new) stanford light field archive. *Computer Graphics Laboratory, Stanford University*, 6(7), 2008. 8

[31] Sven Wanner, Christoph Straehle, and Bastian Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1011–1018, 2013. 1, 8

[32] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7484–7493, 2019. 2

[33] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. 1, 2

[34] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. 2

[35] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13106–13113, 2020. 2