Supplementary Material for “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

1. Comparison Results of Efficiency

In this section, we compare our method with several available disparity estimation algorithms for light field, and evaluate the advantages of our method in GPU memory consumption and model parameters.

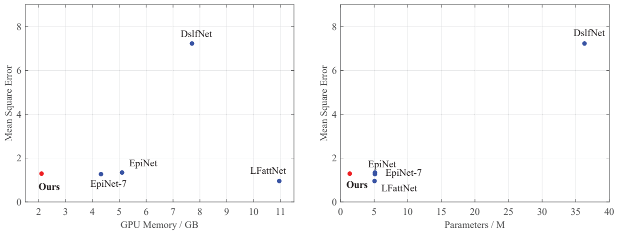

Figure 1. Comparison in performance and efficiency of light field disparity estimation algorithms.

It can be seen from Fig. 1, compared to EpiNet-7 and EpiNet [9], our proposed FastLFnet greatly reduces GPU memory consumption and network parameters while maintaining competitive performance. DslfNet [8] requires a large number of model parameters because *FlowNet 2.0* [4] is used as the backbone, and LFattNet [10] demands excessive GPU memory owing to its huge 3D CNN layers.

2. Network Details

We report a detailed description of the feature extraction and EFE module of FastLFnet. We first introduce the structure of Basic Block (or Resblock in the text). As shown in Fig. 2, the basic structure we used follows ResNet [1], with the exception that it does not apply ReLU after summation.

The parameters of the feature extraction module of our FastLFnet are detailed in Tab. 1, while the structure of the BAM module is shown in Tab. 2. For the center view of the light fields, after feature extraction, the feature maps (*first_conv*, *layer1*, *layer2*, *layer3_up*, *layer4_up*, *layer5_up*) are sent to the edge guidance sub-network for edge feature extraction.

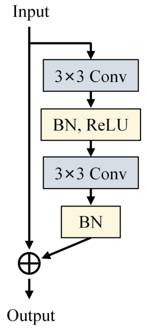

Figure 2. The structure of Basic Block

Name	Convolution layers	Output dimension
input		H × W × 1
first_conv	(3 × 3 conv, 16) × 2	H × W × 16
layer1	(basicblock, 32) × 6	H × W × 32
layer2	(basicblock, 64) × 2	H × W × 64
layer3	(basicblock, 64) × 2, stride 2	H/2 × W/2 × 64
layer4	(basicblock, 64) × 2, stride 2	H/4 × W/4 × 64
layer5	(basicblock, 64) × 2, stride 2	H/8 × W/8 × 64
layer3_up	3 × 3 conv, 16 bilinear interpolation	H × W × 16
layer4_up	3 × 3 conv, 8 bilinear interpolation	H × W × 8
layer5_up	3 × 3 conv, 8 bilinear interpolation	H × W × 8
concat[first_conv, layer1, layer2, layer3_up, layer4_up, layer5_up]		H × W × 144
last_conv	3 × 3 conv, 64 3 × 3 conv, 32	H × W × 32
output	bam module	H × W × 32

Table 1. Parameters of the feature extraction module of our proposed FastLFnet. H and W denote the height and width of the input image. After each convolution, batch normalization and ReLU are followed, except for the last convolution of last_conv.

The edge feature extraction (EFE) module of the edge guidance sub-network is illustrated in Fig. 3 and the parameters are detailed in Tab. 4.

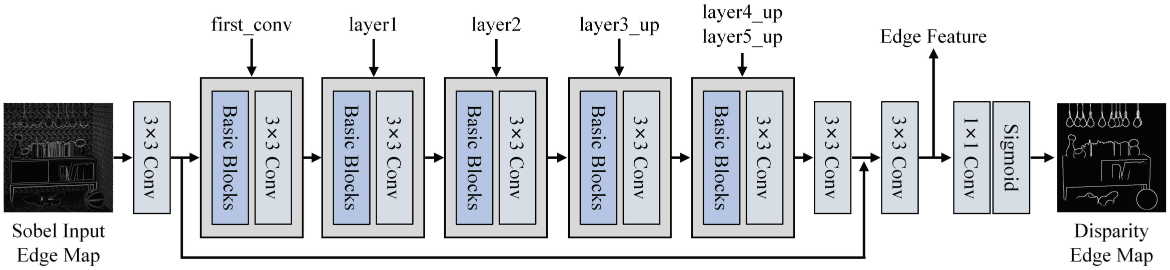

Figure 3. Overview of the edge feature extraction (EFE) module

Name	Convolution layers	Output dimension
input		H × W × 32
gate_0	3 × 3 conv, 16 BN + ReLU	H × W × 16
gate_1	3 × 3 conv, 16 BN + ReLU	H × W × 16
gate_2	3 × 3 conv, 16 BN + ReLU	H × W × 16
gate_out	1 × 1 conv, 1	H × W × 1
output	input ⊙ (1 + sigmoid(gate_out))	H × W × 32

Table 2. Parameters of the BAM module. BN denotes batch normalization and ⊙ denotes element-wise multiply operation. The dilated rate of *gate_2* is set to 2.

Methods	CAE [7]	PS_RF [6]	RPRF-5 [3]	EpiNet-7 [9]	EpiNet [9]	LFattNet [10]	w/o Edge	Ours
Boxes	22.11	23.49	27.55	16.41	15.79	10.57	15.25	11.44
Cotton	17.52	13.60	7.84	2.30	2.71	2.64	3.63	3.83
Dino	1.96	4.50	2.35	1.02	0.90	0.56	2.10	0.98
Sideboard	2.89	7.57	4.24	3.29	3.18	2.04	4.16	2.75
Average	11.12	12.29	10.49	5.75	5.65	3.95	6.28	4.75
Boxes	34.92	46.49	49.94	34.33	33.01	30.91	49.84	45.05
Cotton	22.70	16.07	15.68	7.82	7.39	5.49	18.20	9.90
Dino	13.05	23.34	27.90	9.02	8.49	6.68	24.22	14.83
Sideboard	19.04	32.98	21.10	15.81	14.70	9.66	31.29	19.18
Average	22.43	29.72	28.65	16.75	15.90	13.19	30.89	22.24

Table 3. Quantitative comparison on the metrics of Discon_MSE (row 2 - 6) and Discon_BadPix (row 7 - 11).

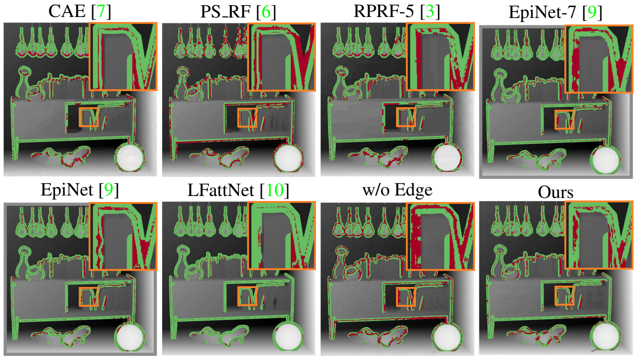

Figure 4. Error maps of Discon_BadPix for the scene *Sideboard*.

3. More Evaluations on Discontinuity Regions

We make more quantitative and visual evaluations with different metrics at discontinuity and occlusion regions. The Discon_MSE and Discon_BadPix metrics on the 4D Light Field Dataset [2], i.e., the mean square error and the percentage of bad pixels (the absolute error greater than 0.07) at discontinuity regions, are given in Tab. 3. The error maps of Discon_BadPix for *Sideboard* scene are shown in Fig. 4. As shown, compared with the state-of-the-art methods, our method achieves better Discon_MSE scores and comparable Discon_BadPix with high computational efficiency. Besides, we also add the results of our method without edge guidance, and we can see that by introducing the light-weight edge guidance module (refer to the effec-

Index	Input	Convolution layers	Output	Output dimension
1	sobel_edge	(3 × 3 conv, 16) × 2	pre_conv	H × W × 16
2	pre_conv *first_conv*	concat[pre_conv, first_conv]	cat1_0	H × W × 32
3	cat1_0	(basicblock, 32) × 3	cat1_1	H × W × 32
4	cat1_1	3 × 3 conv, 16	cat1_2	H × W × 16
5	*layer1*	3 × 3 conv, 16	layer1_down	H × W × 16
6	cat1_2 layer1_down	concat[cat1_2, layer1_down]	cat2_0	H × W × 32
7	cat2_0	(basicblock, 32) × 3	cat2_1	H × W × 32
8	cat2_1	3 × 3 conv, 16	cat2_2	H × W × 16
9	*layer2*	3 × 3 conv, 32 3 × 3 conv, 16	layer2_down	H × W × 16
10	cat2_2 layer2_down	concat[cat2_2, layer2_down]	cat3_0	H × W × 32
11	cat3_0	(basicblock, 32) × 3	cat3_1	H × W × 32
12	cat3_1	3 × 3 conv, 16	cat3_2	H × W × 16
13	cat3_2 *layer3_up*	concat[cat3_2, layer3_up]	cat4_0	H × W × 32
14	cat4_0	(basicblock, 32) × 3	cat4_1	H × W × 32
15	cat4_1	3 × 3 conv, 16	cat4_2	H × W × 16
16	cat4_2 *layer4_up* *layer5_up*	concat[cat4_2, layer4_up, layer5_up]	cat5_0	H × W × 32
17	cat5_0	(basicblock, 32) × 3	cat5_1	H × W × 32
18	cat5_1	3 × 3 conv, 16	cat5_2	H × W × 16
19	cat5_2	3 × 3 conv, 16	res_conv	H × W × 16
20	res_conv pre_conv	ReLU(pre_conv + res_conv)	out_feature	H × W × 16
21	out_feature	3 × 3 conv, 16	edge_feature	H × W × 16
22	edge_feature	1 × 1 conv, 1 sigmoid	disparity_edge	H × W × 1

Table 4. The architecture of the EFE module. The input sobel_edge is obtained from the input image with the *Sobel* edge detection operation [5]. The features indicated by italic text are obtained from the previous feature extraction module. Batch normalization and ReLU are employed in all convolution layers, apart from the layers of index 19 and 22.

tiveness ablation results in Tab. 2 in the paper), the performance of our method on edge regions is largely improved on average in an efficient way.

References

[1] 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. 1

[2] 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. 2

[3] 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. 2

[4] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 1

[5] FG Irwin et al. An isotropic 3x3 image gradient operator. *Presentation at Stanford AI Project*, 2014(02), 1968. 2

[6] 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. 2

[7] 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. 2

[8] 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

[9] 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. 1, 2

[10] 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