Supplementary Materials: Epipolar Consistent Attention Aggregation Network for Unsupervised Light Field Disparity Estimation

Chen Gao, Shuo Zhang *, Youfang Lin
Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence,
School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China.
{gaochen, zhangshuo, yflin}@bjtu.edu.cn

In this supplementary material, we first present more discussion of our parallex attention maps. Then, we provide more results of different models on the HCI and HCI old dataset. Finally, more intermediate visual results of our model are shown.

1. Discuss about ECAAN

1.1. The key point of our methods.

Unlike PAM in binocular vision, the offset scales and occlusion patterns in the attention maps between views in LFs are different. Visualization of more attention maps are present in Fig. A. As shown, the attention maps exhibit varying scales and the corresponding matching points are hard to find in occluded areas (labeled with red arrows). Our approach is the first to focus on aggregating these attention maps for LF disparity estimation. On one hand, the designed ECSU successfully unify scale of the attention maps. On the other hand, we design COFA to detect occluded areas and aggregate disparity information from non-occluded areas, producing clearer aggregated attention maps.

1.2. The superiority of our method

The ability of handling large disparities: Our model excels in handling both **large disparity scenes** and **real-world scenes** (Tab.1 and Fig.6). Moreover, our approach yields smoother results across varying disparity inputs (Fig.5). Both cost-volume based methods and our method require a predefined disparity range. The difference is that for a large disparity range, the cost volume based methods have to change the way they construct the cost volume accordingly by using large granularity, where the receptive field may be insufficient to capture correspondence over the large disparities. Therefore, their disparity maps exhibit noticeable disruptions in depth smoothness as in Fig.5. By contrast, our model is able to capture the long-range correspondence using the attention mechanism, which achieves better performance in large disparity scenes and is robust to different disparity variations.

*Corresponding author: zhangshuo@bjtu.edu.cn

2. Experiments

2.1. Comparison of Efficiency

We compare the efficiency of different methods measured by the running time (in seconds) for inferring the disparity map from a 4-D LF image in Tab. A. All the non-learning methods are implemented on the CPU and the learning-based methods are tested on the NVIDIA A4000. Tab. A shows that traditional non-learning methods [2, 4, 9] cause a sizeable computational burden. With the GPU acceleration, the estimation time for learning-based methods is reduced. Compared with the other learning methods that need cost volume construction [3, 5–8], our model achieves comparable running times using self-attention for disparity estimation.

2.2. Comparison on HCI Benchmark

We submitted our results to the HCI online benchmark evaluation labeled as "LF_depth_2" and Tab. B presents the results of different methods on the HCI 'test' dataset. Compared with UnOcc [3], our epipolar consistent aggregation perceives occlusion information and fully extracts and fuses disparity information from different views by assigning weights to views. Our model shows significant improvement on the Bad Pixel metric over unsupervised methods [3, 5] and achieves comparable results with some supervised methods [8] in some scenarios, indicating that our model significantly reduces the performance gap between supervised and unsupervised methods.

The estimated disparity maps and the corresponding error maps are shown in Fig. B. Our results have more accurate edges than the current best-unsupervised method UnOcc [3] with obviously fewer errors in all scenes. UnOcc [3] processes the input image by dividing it into 4 sub-images and discards all the information in the possible occluded

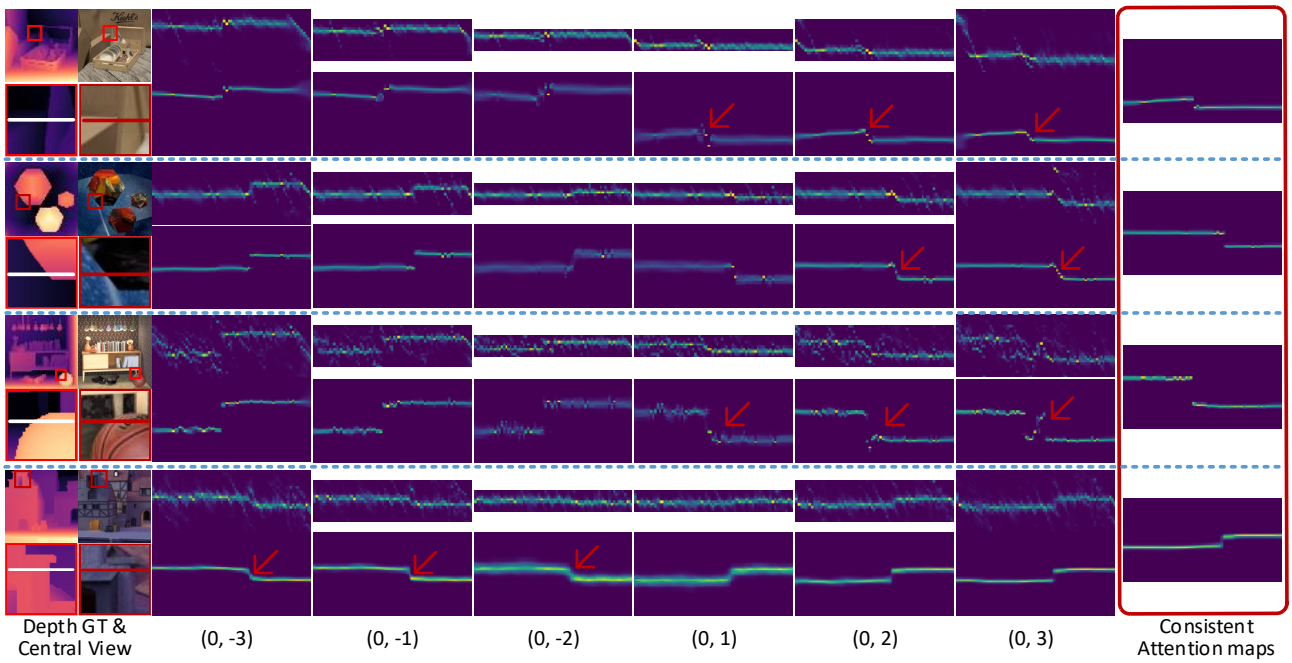

Figure A. Addition examples of the attention maps.

Table A. Comparison of the running time (seconds) of different methods for disparity map estimation from a $512 \times 512 \times 7 \times 7$ LF.

Non-Learning			Supervised			Unsupervised		
SPO [9]	CAE [4]	OAVC [2]	EPINet [6]	LFattNet [7]	OACC-Net [8]	Unsup [5]	UnOcc [3]	Ours
65.00	229.3	68.63	1.35	7.04	21.7	5.57	0.16	1.02

input sub-images. In contrast, our occlusion-aware aggregation assigns different fusion weights to filter the ambiguous information and accurately fuse the disparities obtained from different views.

We also visualize the performance of our model on the HCI old dataset in Fig. C. Our results have more accurate disparity information than other methods, even compared to those supervised methods [1, 6–8]. For example, the supervised methods make significant errors in "papillon" and "stilllife". This shows that our model has strong generalization ability and achieves good prediction results on different datasets.

3. Visualization

3.1. Visualization of Occlusion Masks

In this part, we verify the effectiveness of the occlusion mask calculation. Specifically, we show the original generated four occlusion masks $O_{(0,-3)}$, $O_{(0,3)}$, $O_{(-3,0)}$, $O_{(3,0)}$ as well as the extended four occlusion masks $O_{(-3,-3)}$, $O_{(-3,3)}$, $O_{(3,-3)}$, $O_{(3,3)}$ using Equ (5) in section 4.4, as in Fig. D. We also show the ground truth occlusion masks for comparison, which are calculated by warping the related views to the central view according to ground truth disparity and thresholding the differences.

As in Fig. D, our occlusion masks are able to cover the occluded regions in input LF images. Taking the ball in the scene "Sideboard" as an example, the occlusion area in $O_{(3,0)}$ mostly appears above the object, while the perceived occlusion in $O_{(0,3)}$ mostly appears on the left side of the object. The view $I_{(3,3)}$ is located on the upper right side of the central views. The occlusion appears on the left and above the object in $O_{(3,3)}$, the same as in ground truth. The results fully prove the effectiveness of our occlusion mask generation method and ensure that the model pays more attention to the learning of non-occluded regions.

3.2. Visualization of Intermediate Results

We also show the $D_{(u,0)}$ and $D_{(0,v)}$ regressed from parallax attention maps computed from different views in Fig. E to show the differences of these disparity maps.

Compare with the disparity map $D_{(u,0)}$ and $D_{(0,v)}$, the errors caused by occlusions in $D_{(u,0)}$ disappeared in $D_{(0,v)}$. Specifically, in the red region, our model perceives occlusion in the vertical moving views while perceiving the correct depth information in the horizontal moving views.

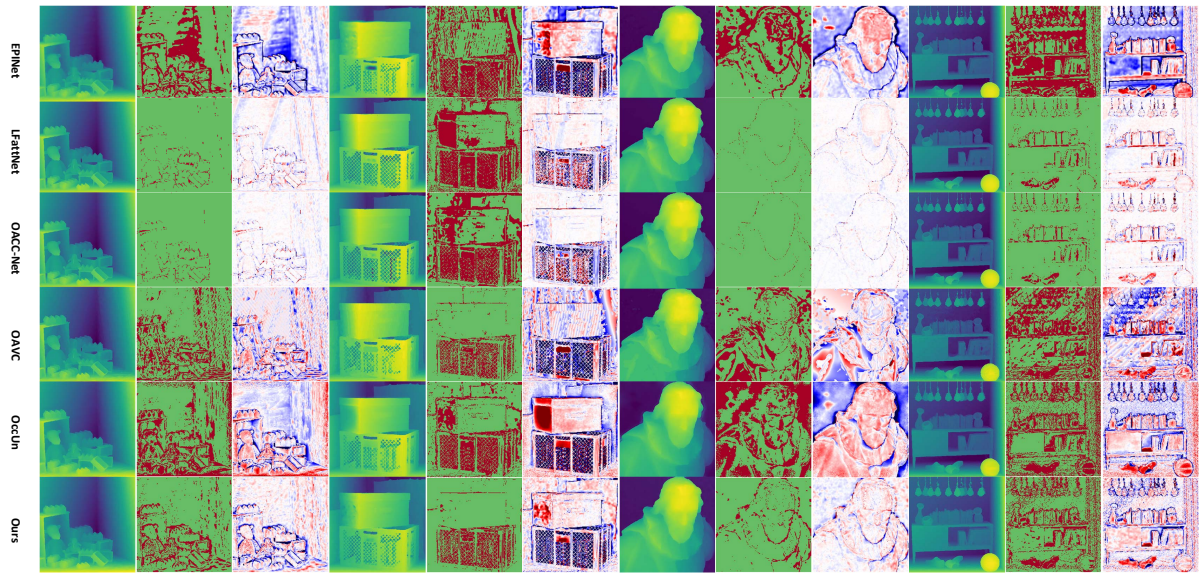

Figure B. Visual comparison of scenes "Boxes", "Cotton", "Dino", "Sideboard" with state-of-the-art methods. The BadPix (0.03) error maps, the MSE error map, and the disparity maps are shown.

Table B. Quantitative comparison of Bad Pixel 0.07, 0.03, 0.01 and MSE*100 on HCI 4D LF synthetic scenes.

		Bed				Bic				Herbs				Ori			
		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
Non-L	CAE[4]	5.79	25.36	68.59	0.24	11.23	23.62	59.65	5.14	9.55	23.16	59.24	11.67	10.03	28.36	64.16	1.78
	SPO[9]	4.87	23.53	72.38	0.21	10.91	26.91	71.14	5.57	8.26	30.63	86.63	11.24	11.70	32.71	75.58	2.04
	OAVC[2]	4.92	19.09	64.76	0.22	12.22	25.46	64.74	4.89	8.73	29.65	74.76	10.37	12.56	30.59	69.36	1.48
Sup	EPINet [6]	2.40	6.93	33.99	0.22	9.90	18.05	46.37	4.69	12.10	28.95	62.67	9.70	5.92	14.37	45.94	1.47
	LFattNet [7]	2.79	5.32	13.33	0.37	9.51	16.00	31.36	3.35	5.22	9.49	19.27	6.61	4.82	8.93	22.19	1.74
	AttMLFNet [1]	2.07	5.28	16.19	0.13	8.84	16.06	32.71	3.09	5.43	9.47	18.84	6.38	4.40	9.04	22.46	1.00
	OACC-Net [8]	2.31	5.71	21.98	0.15	8.08	14.40	32.75	2.91	6.52	46.79	86.42	6.57	4.07	9.72	32.25	0.88
Unsup	Unsup[5]	21.61	43.62	75.44	0.93	30.24	50.31	78.42	11.74	63.94	79.66	92.55	145.56	53.41	72.74	89.04	8.82
	UnOcc[3]	12.69	37.82	74.56	0.39	21.65	45.66	77.93	6.24	16.96	57.12	85.78	13.95	19.82	62.55	87.19	1.93
	Ours	6.40	18.42	50.84	0.27	17.12	30.46	59.66	7.21	10.81	24.17	57.20	14.81	11.04	28.10	61.33	1.99

Therefore, the disparity information $D_{(u,0)}$ produces errors, while $D_{(0,v)}$ contains the correct disparity information. The yellow area is the opposite. There is an occlusion in this area in $I_{(0,v)}$ while the original information is retained in $I_{(u,0)}$. At the same time, we show the occlusion perceived by different views in these two block regions. Our module perceives the correct occlusion relation in the region and uses occlusion information to perform occlusion-aware aggregation.

Moreover, since views $I_{(0,-3)}$ and $I_{(0,3)}$ are far from the central view, the position offsets are $3\times$ larger than the adjacent views $I_{(0,-1)}$ and $I_{(0,1)}$ due to the larger baseline. On one hand, the errors caused by the occlusion areas are more serious in $D_{(0,-3)}$ and $D_{(0,3)}$ than in $D_{(0,-1)}$ and $D_{(0,1)}$. On the other hand, since the baseline is larger, it is easier to estimate a more precise disparity in sub-pixel level in $D_{(0,-3)}$ and $D_{(0,3)}$. Therefore, the background plane is smoother in $D_{(0,-3)}$ and $D_{(0,3)}$ than in $D_{(0,-1)}$ and $D_{(0,1)}$. Therefore, in this paper, we propose to fully exploit the complementary information between the disparity information calculated from different views.

References

[1] Jiaxin Chen, Shuo Zhang, and Youfang Lin. Attention-based multi-level fusion network for light field depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1009–1017, 2021. 2, 3

[2] Kang Han, Wei Xiang, Eric Wang, and Tao Huang. A novel occlusion-aware vote cost for light field depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8022–8035, 2021. 1, 2, 3

[3] Jing Jin and Junhui Hou. Occlusion-aware unsupervised learning of depth from 4-d light fields. *IEEE Transactions on Image Processing*, 31:2216–2228, 2022. 1, 2, 3

[4] 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. 1, 2, 3

[5] Jiayong Peng, Zhiwei Xiong, Dong Liu, and Xuejin Chen. Unsupervised depth estimation from light field using a convolutional neural network. In *Proceedings of the IEEE Interna-*

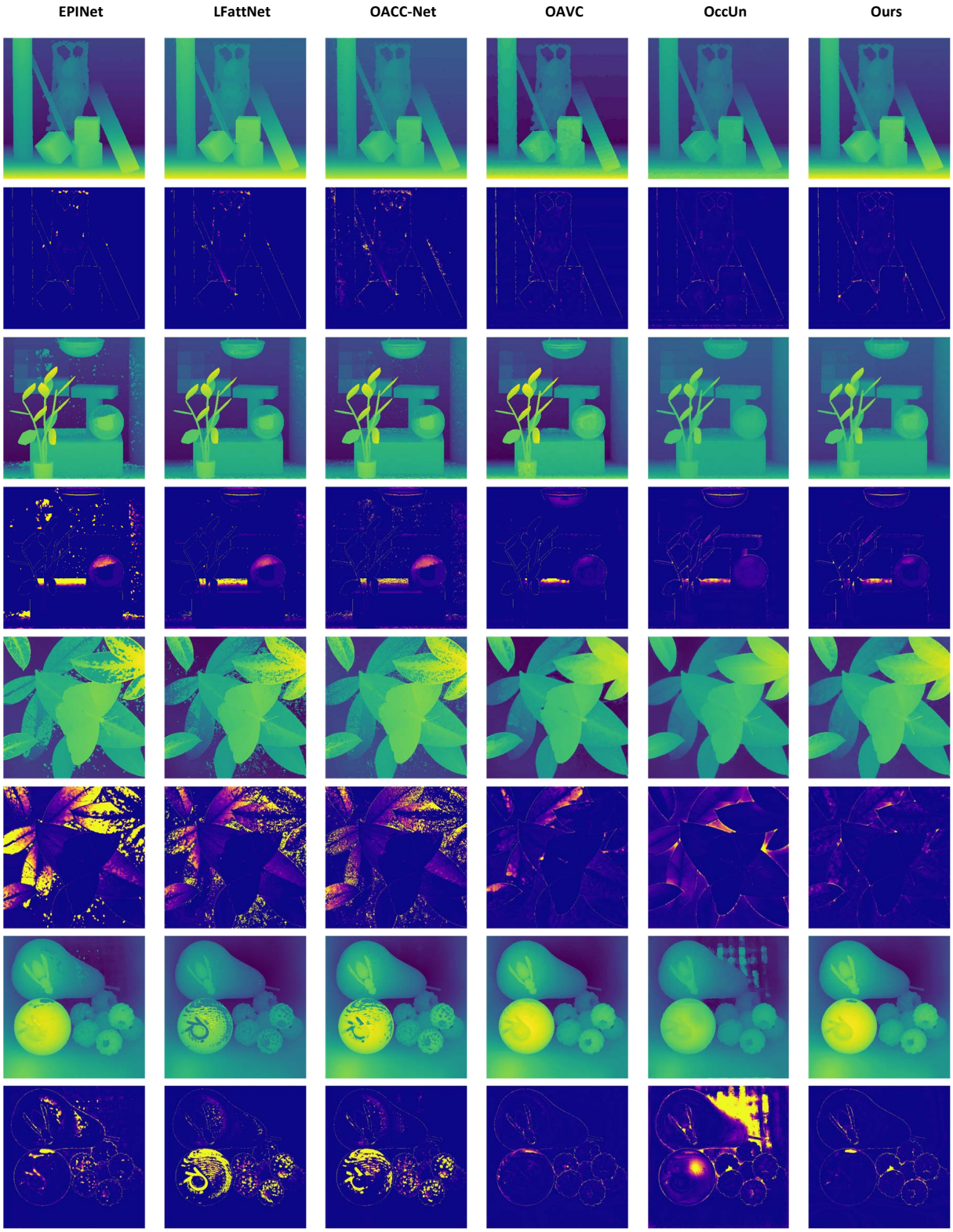

Figure C. Visual comparison with state-of-the-art methods on the HCI old dataset. The MSE error map and the disparity maps are shown.

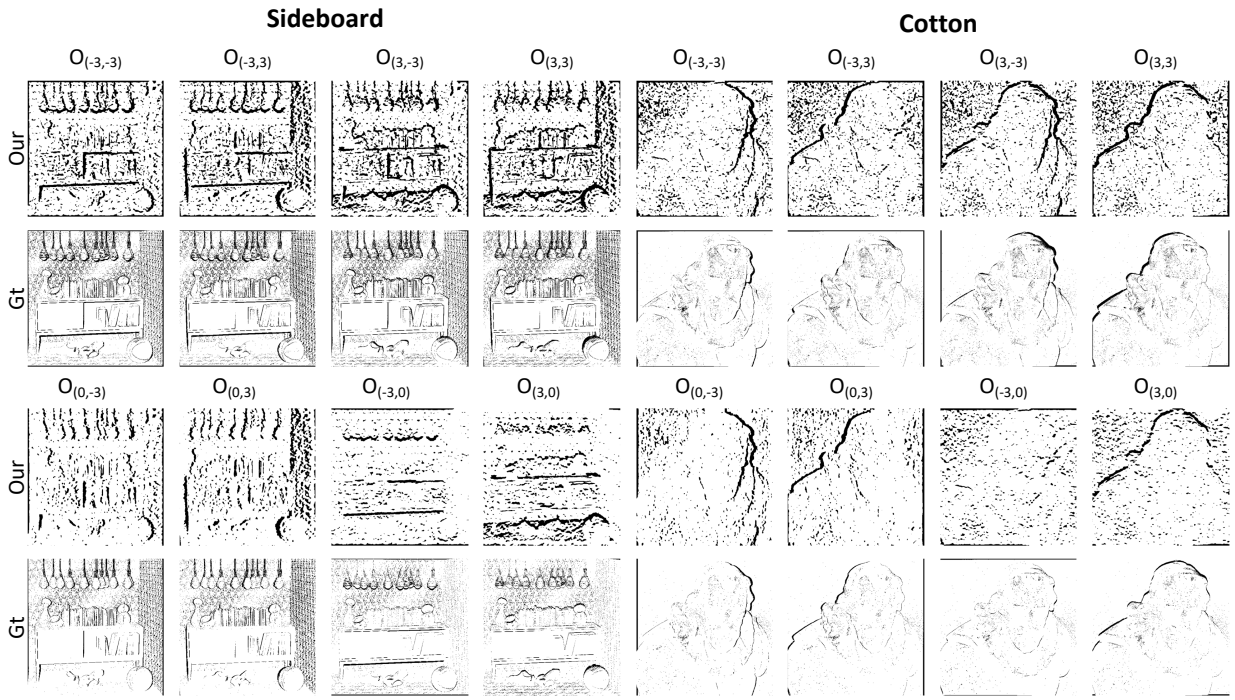

Figure D. Visualization of the generated occlusion masks.

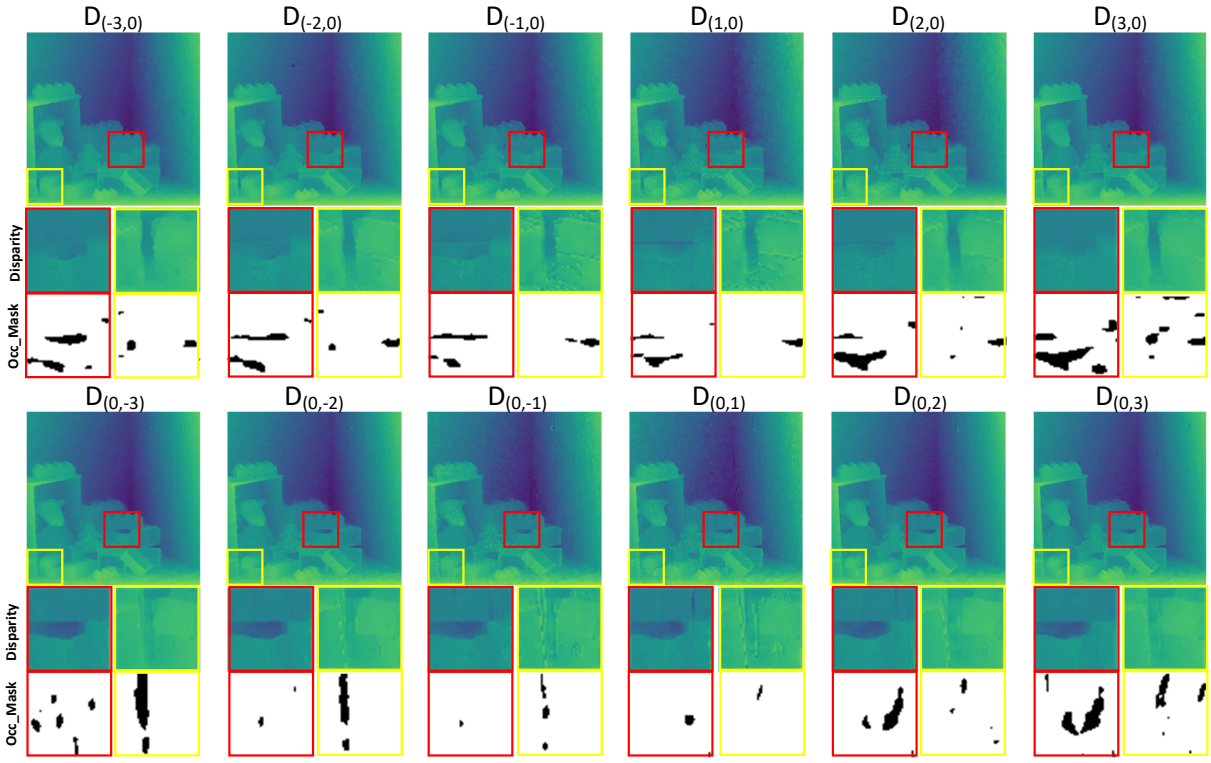

Figure E. Visualization of disparity maps regressed from parallax attention maps computed from different views.

tional Conference on 3D Vision, pages 295–303. IEEE, 2018. 1, 2, 3

[6] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 4748–4757, 2018. 2, 3

[7] 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*, pages 12095–12103, 2020. 2, 3

[8] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Wei An, and Yulan Guo. Occlusion-aware cost constructor for light field depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2022. 1, 2, 3

[9] Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145:148–159, 2016. 1, 2, 3