Disentangling Local and Global Information for Light Field Depth Estimation

Xueting Yang[1] Junli Deng[*1] Rongshan Chen[2,3] Ruixuan Cong[2,3] Wei Ke[4] Hao Sheng[*2,3,4]

[1]School of Information and Communication Engineering,
Communication University of China, Beijing 100024, P.R.China

[2]State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, P.R.China

[3]Beihang Hangzhou Innovation Institute Yuhang,Xixi Octagon City,
Yuhang District,Hangzhou 310023,PR.China

[4]Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR 999078, PR.China

{yangxueting, dengjunliok}@cuc.edu.cn

{rongshan, congrx, shenghao}@buaa.edu.cn wke@mpu.edu.mo

Abstract

Accurate depth estimation from light field images is essential for various applications. Deep learning-based techniques have shown great potential in addressing this problem while still face challenges such as sensitivity to occlusions and difficulties in handling untextured areas. To overcome these limitations, we propose a novel approach that utilizes both local and global features in the cost volume for depth estimation. Specifically, our hybrid cost volume network consists of two complementary sub-modules: a 2D ContextNet for global context information and a matching cost volume for local feature information. We also introduce an occlusion-aware loss that accounts for occlusion areas to improve depth estimation quality. We demonstrate the effectiveness of our approach on the UrbanLF and HCInew datasets, showing significant improvements over existing methods, especially in occluded and untextured regions. Our method disentangles local feature and global semantic information explicitly, reducing the occlusion and untextured area reconstruction error and improving the accuracy of depth estimation.

1. Introduction

Light field imaging has emerged as a potent tool for capturing and analyzing intricate three-dimensional (3D) scenes. Through the capture of multiple views of a scene, light field cameras permit the reconstruction of depth information, thereby enabling various applications such as augmented reality [12, 28], 3D modeling [7], and autonomous driving [9, 26]. However, accurate depth estimation from light field images remains a challenging problem due to the high dimensionality of the data and the large number of possible views.

In recent years, deep learning-based techniques [1, 15, 19, 22, 25] have exhibited considerable potential for addressing the problems mentioned. These techniques generally entail constructing a cost volume, which quantifies the similarity between each pixel in the reference view and its corresponding pixels in other views. The cost volume is subsequently utilized to estimate the depth map by identifying the disparity that minimizes the cost. Despite the achievements of these methods, they are associated with a few limitations, such as sensitivity to occlusions, high computational costs, and difficulties in handling untextured areas. To surmount these obstacles, we propose a novel approach for depth estimation that employs cost volume consisting of both local and global features using light field images. Our approach employs a hybrid cost volume network that learns both local matching information and global contextual information about the reference view. By focusing more on global contextual information, our method mitigates untextured area reconstruction errors and enhances depth estimation accuracy. Furthermore, we introduce an occlusion-aware loss that accounts for occlusion boundaries between views, facilitating more robust depth estimation in the presence of occlusions.

Leading light field depth estimation models employ a single fully 3D convolutional network for cost regularization to learn both local feature matching information and global context information [5, 11, 22, 24], which are crucial for achieving accurate depth estimation. While local feature information is necessary for matching texture-rich regions,

[*]represents the corresponding author.

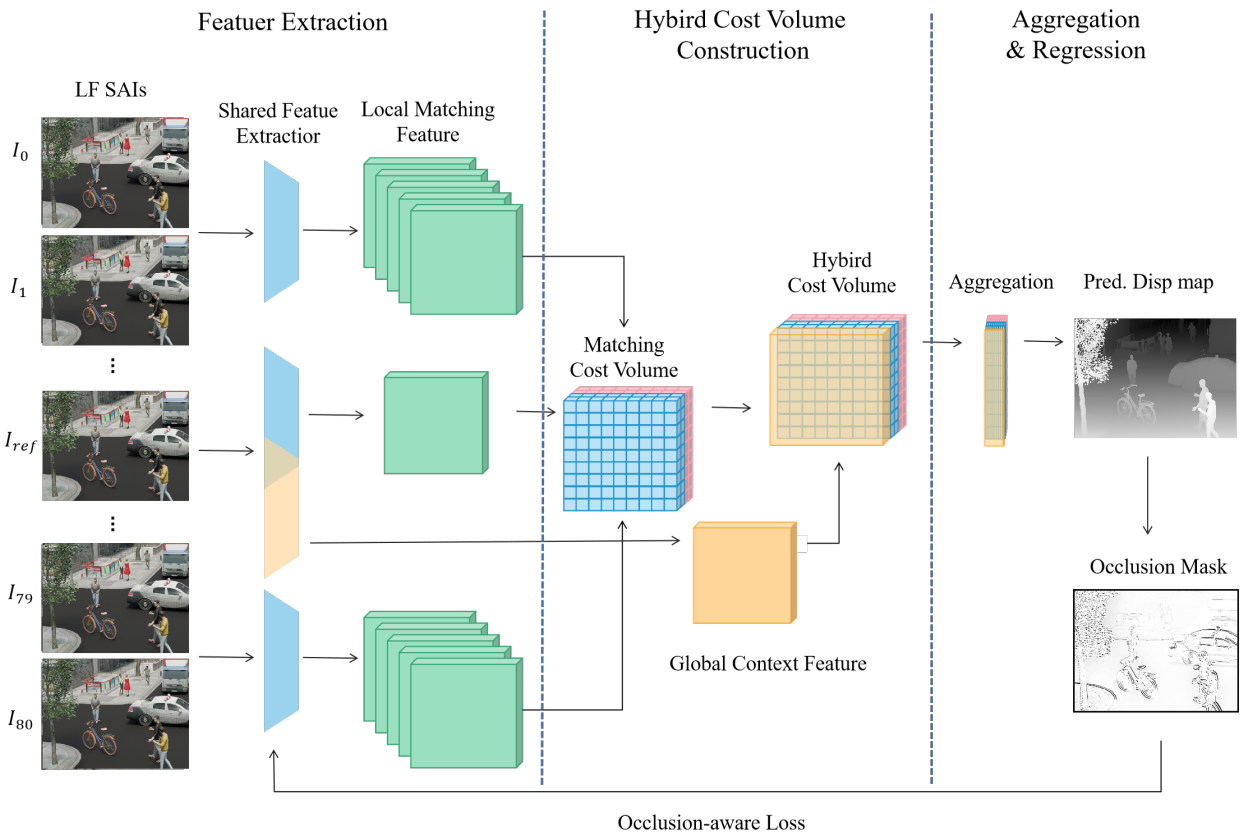

Figure 1. The overview of our network. Features extracted from each SAI are utilized to construct the matching cost volume. Notably, we propose a hybrid cost volume network with two sub-modules that separately focus on 3D local matching information and 2D global context information. Lastly, a disparity aggregation and regression module predicts the disparity map with the supervision of the proposed occlusion-aware loss

global context information is vital for scenes with textureless regions. However, existing networks tend to overlook the importance of global context features and rely heavily on deep neural networks or attention mechanisms to learn local matching information. Consequently, such networks face challenges in handling textureless regions. In this regard, we propose a novel approach that explicitly disentangles local feature information and global semantic information by utilizing two separate networks for depth estimation. Our hybrid cost volume network consists of two complementary expert sub-modules: a 2D ContextNet, which focuses on 2D global context information, and a matching cost volume network, which concentrates on 3D local information. Our method significantly reduces the untextured area reconstruction error, thereby improving the accuracy of depth estimation.

Furthermore, image-based depth estimation of the light field often faces challenges due to degraded geometry in the occluded regions. To address this problem, we first infer the occluded mask in the image domain to construct an occlusion-aware loss. Through extensive experiments on UrbanLF [17] and HCInew [4] datasets, we demonstrate the effectiveness of our approach, especially in occlusion and untextured areas.

Our contributions can be summarized as follows:

1. We present a hybrid cost volume that focuses on both local feature information and global semantic information to improve the accuracy of reconstruction results in textureless regions.

2. We propose a occlusion-aware loss for depth estimation of Light Field images to maintain geometry in the occluded regions.

3. Extensive experimental results demonstrate the effectiveness and universality of our method using the challenging test datasets UrbanLF and HCInew.

2. Related Work

In this section, we review traditional and deep learning based methods in Light Field depth estimation.

2.1. Traditional Methods

Depth estimation from light field (LF) images has been a prominent area of research within the field of computer vision for a number of years. Traditional methods [10, 13, 16, 18, 21, 29] for estimating depth can be broadly categorized into three distinct classes based on the properties of LF images.

Epipolar plane image (EPI)-based methods leverage the epipolar geometry of light field (LF) images for predicting depth. One such approach, proposed by Wanner et al. [27], involves utilizing a structure tensor to estimate line slopes in horizontal and vertical EPIs, which is further refined through global optimization. Another method, introduced by Zhang et al. [29], employs a spinning parallelogram operator (SPO) to estimate line slopes for depth estimation. Multi-orientation EPIs were introduced by Sheng et al. [18] for slope estimation, which resulted in improved results over SPO. Schilling et al. [16] proposed an inline occlusion handling scheme that operates on EPIs. On the other hand, multiview stereo matching (MVS)-based methods utilize the multi-view information of LF images for depth estimation. These methods establish correspondences between multiple views of the scene and estimate depth from triangulation. For example, Jeon et al. [6] developed a phase-based multi-view stereo matching technique that enabled Fourier domain depth estimation. Tao et al. presented a shading-based refinement strategy to augment the precision of depth estimation [21]. Williem et al. [13] proposed an approach that employed angular entropy cost and adaptive defocus cost. In contrast, defocus-based methods measure the consistency at different focal stacks to calculate the depth of a pixel. Tao et al. [20] presented a method that combines defocus and correspondence depth cues from LF images to obtain dense depth estimation. Additionally, Wang et al. [23] introduced a new occlusion-aware cost function to estimate the depth map.

While traditional methods have exhibited optimistic outcomes, they encounter restrictions such as non-linear optimization and manually crafted characteristics that demand significant computation and are susceptible to occlusions, weak textures, and highlights. Additionally, these methods may exhibit subpar performance in intricate settings with multiple occlusions and varying lighting conditions.

2.2. Deep Learning-based methods

The utilization of deep networks for depth estimation has gained significant popularity in recent years [1, 19, 30], exhibiting remarkable outcomes compared to traditional methods. Heber et al. [2] were pioneers in introducing an end-to-end network that learns the mapping between a 4D light field and its corresponding depths and uses high-order regularization to refine depth estimation. Subsequently, Heber, Yu, and Pock [3] developed an efficient U-shaped encoder-decoder architecture that extracts geometric information from light field images to generate high-quality depth maps. Shin et al. [19] presented a multi-stream network and a range of data augmentation strategies that facilitate fast and precise light field depth estimation. Tsai et al. [22] introduced an attention-based view selection network that dynamically incorporates all angular views for depth estimation. Furthermore, Peng et al. [14] proposed an unsupervised technique for light field depth estimation that eliminates the requirement for ground-truth depth maps during training. They later introduced a zero-shot learning-based approach that performs unsupervised depth estimation without relying on external datasets [15]. More recently, Chen et al. [1] devised an attention-based multi-level fusion network to address the occlusion issue in depth estimation, while Huang et al. [5] employed a multi-disparity-scale cost aggregation method to accelerate light field depth estimation. DistgDisp [25] disentangles light fields into view-specific high-resolution images and a low-resolution disparity map, enabling super-resolution and accurate depth estimation.

Prior research has evidenced the efficacy of deep neural networks in Light Field (LF) depth estimation. Nevertheless, there has been inadequate attention devoted to the relevance of global context features and occlusions in LF images, leading to challenges in dealing with textureless and occluded regions. To address these limitations, in this paper, we introduce a new approach that explicitly exploits global context features and incorporates an occlusion-aware loss function to enhance depth estimation accuracy.

3. Method

This paper introduces a novel approach for Light Field (LF) depth estimation utilizing sub-aperture images (SAIs) based on cost volume. Our proposed method comprises multiple modules that collaborate to achieve accurate and robust disparity estimation, as depicted in Fig. 1. The first module is the shared feature extractor, which extracts features from each SAI (Sec. 3.1). These features are then utilized to construct the cost volume through pixel view shifting in the second module. Notably, we propose a hybrid cost-volume network with two sub-modules that separately focus on 3D local matching information and 2D global context information (Sec. 3.2). The third step uses the cost aggregation module to synthesize the hybrid cost volume data. Lastly, a disparity regression module predicts the disparity map under the supervision of our proposed occlusion-aware loss (Sec. 3.3). The subsequent subsections provide an elaborate description of each module.

3.1. Feature Extraction

To extract features from sub-aperture images (SAIs), we employ the Spatial Pyramid Pooling (SPP) module, which

has demonstrated effectiveness in multiple computer vision tasks for multi-scale feature extraction [22]. We adopt the SPP module to create a feature map F, which contains significant information necessary for estimating disparity. This feature map serves as an input component to our hybrid cost volume network.

3.2. Hybrid Cost Volume

This paper employs a parallel plane parameterization to represent the four-dimensional light field image. The light traveling through space that passes through a point (u, v) on the primary lens plane and a point (x, y) on the microlens plane can be expressed as a four-dimensional light field image $L(x, y, u, v)$. Specifically, (u, v) denotes the angular coordinate, representing the plane coordinates of the camera, while (x, y) represents the spatial coordinate, indicating the plane coordinates of the image. The formulation of LF disparity structure is given by:

$$L\left(u_c, v_c, x, y\right) = L\left(u, v, x + (u - u_c) * d(x, y), \\ y + (v - v_c) * d(x, y)\right). \quad (1)$$

To find the corresponding point (X_s, Y_s) from another viewpoint s given a point (x_c, y_c) on the central viewpoint, we consider that each viewpoint is situated on a regular grid and their internal camera parameters can be regarded as equal. This allows us to formulate the corresponding point as:

$$(X_s, Y_s) = (du + x_c, dv + y_c), \quad (2)$$

The parameter d represents the disparity between adjacent views. In our approach, we uniformly sample virtual disparities over a range of values that cover the observed ranges in datasets, as previously done in relevant literature [5, 22, 24].

A raw cost volume for the center image is constructed by shifting the source feature map F_s at the virtual disparities. Specifically, we warp the feature map of each SAI into all virtual disparities to generate a feature volume with dimensions $R^{C \times D \times H \times W}$, which is termed the Matching Cost Volume, which serves the purpose of exclusively learning local features for matching.

In order to enhance depth estimation efficiency, we utilize another network, namely ContextNet, to learn global context information that complements the Matching Cost Volume. The ContextNet produces a learned feature volume with dimensions $R^{C_1 \times H \times W}$, where C_1 denotes the number of feature channels, which coincides with the number of virtual disparities D. To combine the 3D matching volume and the 2D context volume, we expand the dimensions of the context volume to $R^{1 \times D \times H \times W}$. Finally, we concatenate the regularized matching cost volume with the expanded context volume to obtain a hybrid cost volume, with dimensions $R^{(C+1) \times D \times H \times W}$. We repeat this hybrid cost volume generation procedure for each SAI.

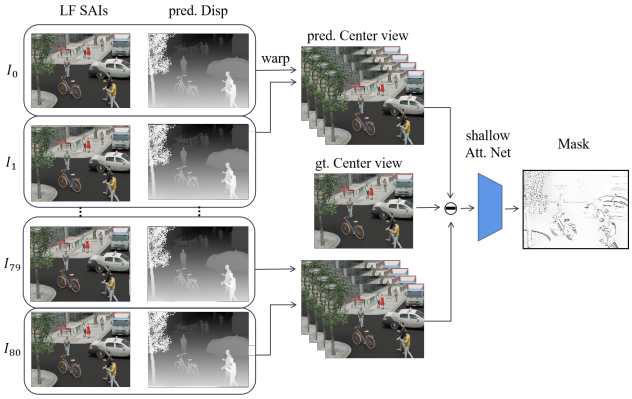

Figure 2. Based on photometric consistency, the projected view $I_{k \to c}$ should possess identical values to the center view I_c at non-occluded regions. Therefore, the occlusion mask $M_{k \to c}$ can be determined by calculating the absolute residuals between the projected view $I_{k \to c}$ and the center view I_c. Then a generalized occlusion mask is predicted by a shallow view-attention network from all SAIs masks.

3.3. Cost Aggregation and Disparity Regression

The hybrid cost volume has dimensions $R^{D \times H \times W \times (C+1)}$, and we use a 3D CNN for cost aggregation. Consistent with LFattNet [22], our cost aggregation approach incorporates eight $3 \times 3 \times 3$ convolutional layers and two residual blocks. The final cost volume C_f of size $D \times H \times W$ is obtained after processing through these 3D convolution layers. To produce the probability of the disparity distribution, Y_{dist}, we normalize C_f using softmax operation along dimension D. Finally, the output disparity $\hat{d}$ can be computed as:

$$\hat{d} = \sum_{d_k = D_{\min}}^{D_{\max}} d_k \times \mathrm{softmax}\left(-C_{d_k}\right) \quad (3)$$

where $\hat{d}$ denotes the estimated center-view disparity, $D_{\min}$ and $D_{\max}$ represent the minimum and maximum disparity values, respectively, and d_k represents the sampling interval between $D_{\min}$ and $D_{\max}$.

Occlusion-aware loss. Prior studies have typically treated all pixels equally, without considering the challenges of occlusion regions. In this paper, we introduce an occlusion-aware loss to supervise the network more effectively in regards to occluded regions by aggregating occlusion masks from source views, thereby achieving robustness to occlusions. Specifically, rather than deriving the occlusion mask of each view during cost volume reconstruction, we compute the occlusion mask of the center view by warping other source views using the predicted disparity. In areas where occlusions occur, a scene point that is visible in the center view may not be present in surrounding views,

Methods	hybrid Cost Vol.	Occlusion Loss	UrbanLF				HCInew			
			MSE×100↓	BP(0.01)↓	BP(0.03)↓	BP(0.07)↓	MSE×100↓	BP(0.01)↓	BP(0.03)↓	BP(0.07)↓
$\text{Ours}_{\text{w/o}}$	×	×	7.692	94.842	84.218	79.470	1.890	49.046	14.370	5.681
$\text{Ours}_{\text{w/o Hyb. cv}}$	×	✓	6.371	92.203	81.623	67.476	1.782	42.236	10.049	4.017
$\text{Ours}_{\text{w/o occ.loss}}$	✓	×	6.147	86.168	78.512	64.914	1.763	47.360	9.817	4.055
Ours	✓	✓	5.989	90.850	77.411	59.738	1.752	36.574	8.813	3.507

Table 1. Ablation studies on hybrid cost volume and occlusion-aware loss.

	UrbanLF			
Method	MSE×100↓	BP(0.01)↓	BP(0.03)↓	BP(0.07↓)
Ours	5.989	90.850	77.411	59.738
OCV	10.217	97.171	91.563	80.928
Multi-Net	10.206	98.678	95.812	88.972
UOAC-T1	6.035	71.226	41.489	26.788
pixelplus	5.145	94.764	84.754	66.294
ASO	3.891	94.303	83.294	63.295
MultiBranch	2.776	86.35	64.915	43.402
SF-Net	0.936	26.69	15.331	9.869
HRDE	0.798	32.846	14.637	7.534
EPI-Cost	0.785	67.222	30.662	14.155
SF-Net-M	0.712	28.007	14.595	8.311
MS3D	0.61	47.261	18.957	9.254
HRDE-aug	0.395	33.504	15.322	7.312

Table 2. Quantitative evaluation: compared to the existing state-of-the-art methods evaluating on UrbanLF. Our method is in the top ten of all submitted methods.

resulting in occluded pixels in these views that do not correspond to their equivalent pixels in the center view. However, based on photometric consistency, the projected view $I_{k\to c}$ should possess identical values to the center view I_c at non-occluded regions. Therefore, the occlusion mask $M_{k\to c}$ can be determined by calculating the absolute residuals between the projected view $I_{k\to c}$ and the center view I_c:

$$M_{k\to c} = |\mathcal{I}_{k\to c} - \mathcal{I}_c| \tag{4}$$

The previous calculation assumes that the depth predicted by the center view is entirely accurate, and utilizing $M_{k\to c}(k = c)$ directly would inevitably result in errors. To account for this uncertainty, we present a shallow 2D view-attention network that predicts a generalized occlusion mask M_c from occlusion masks of all SAIs, as shown in Fig. 2. Our insight is that the occlusion areas of the central view are somewhat related to those of surrounding views, and the occlusion areas of surrounding views can also cause corresponding pixels to be unmatched. Firstly, the $M * M$ raw occlusion masks are concatenated into $R^{1\times M*M\times H\times W}$ and then processed by two 2D convolution layers, reducing their dimensions to $R^{1\times 1\times H\times W}$. Next, a view average pooling operation aggregates information across different SAIs, resulting in a single aggregated occlusion mask for the central view, namely M_c. Finally, the occlusion-aware loss L_{occ} supervises the predicted disparity as follows:

$$L_{occ} = \|M_c\|^{\beta} \odot \|d - \hat{d}\|_1, \tag{5}$$

where $\hat{d}$ represents the predicted disparity and d denotes the ground truth disparity. The symbol $\odot$ signifies the element-wise multiplication operation, and β serves as the coefficient factor regulating the dynamic weight assignment ratio. When $\beta = 0$, the occlusion loss is equivalent to the standard L_1 loss.

4. Experiment

In this section, we provide an overview of the datasets we used and implementation details, followed by experiments conducted to evaluate our models. Finally, we compare our approach with several state-of-the-art LF depth estimation methods.

4.1. Datasets

We use two datasets in our experiments, a novel and challenging dataset released by [17] and the 4D Light Field Dataset [4].

The UrbanLF-Syn dataset is generated through the use of Blender software, specifically utilizing the Cycles and Eevee renderer. The synthetic urban environment is carefully designed with various elements added to capture images under diverse lighting conditions. This is accomplished by equipping the environment with multiple light sources to simulate different situations. The dataset is collected using a camera array made up of 81 virtual cameras with identical configurations, and disparity can be controlled via adjustment of the distance between adjacent cameras. The UrbanLF-Syn subset comprises 250 synthetic LF samples, including 172 training, 28 validation, and 50 test samples. Each sample consists of 81 sub-aperture images with a spatial resolution of 480×640 and angular resolution of 9×9, along with disparity maps for all views.

The 4D Light Field Dataset, created by [4], is widely regarded as the benchmark for evaluating disparity estimation techniques for light field images. The dataset is comprised of 28 synthetic light field scenes, which have been divided into four subsets: "Stratified", "Test", "Training", and "Additional". These scenes are composed of various materials, lighting conditions, and complex occlusions. The

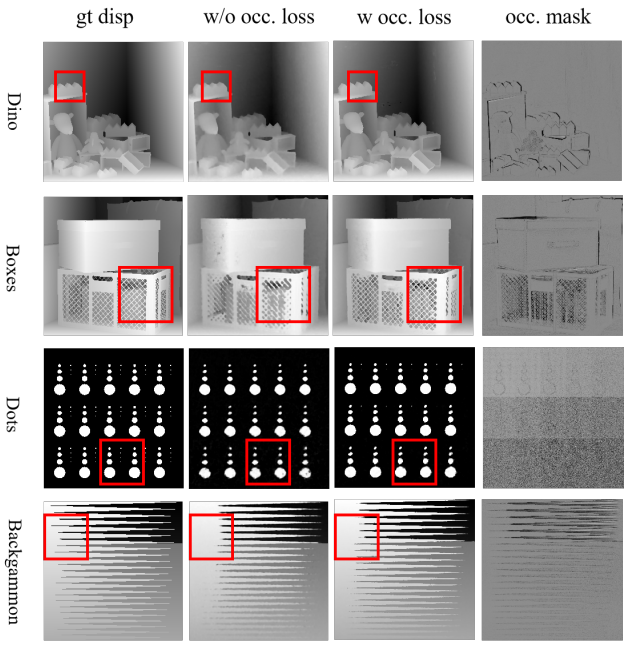

Figure 3. Disparity maps of the predicted scenes with and without occlusion-aware loss.

resolution of the images is 512×512, and they were rendered using the Blender renderer, with 9×9 sub-aperture views. Since the scenes are synthetic, obtaining ground truth depth is straightforward. For our experiments, we utilized 16 scenes from the "Additional" subset for training, 8 scenes each from both "Stratified" and "Training" for validation, and 4 scenes from "Test" for testing purposes. During training, we randomly sampled 32×32 grayscale patches from the training dataset, whereas we employed full resolution 512×512 images for validation.

4.2. Implementation Details

For our implementation, we employ patch-wise training by randomly selecting grayscale patches measuring 32×32 from the light field images present in the training set. Our approach was trained using a supervised methodology equipped with the proposed occlusion-aware loss function and optimized through use of the Adam method [8] with $\beta = 0.2$. The batch size is set to 12 and the learning rate was assigned a value of 1×10^{-3}. The range of disparity sampling varies according to the disparity range of the datasets. Training was executed using PyTorch and completed on a PC equipped with a single Nvidia RTX 3090Ti GPU, requiring approximately one week for completion.

4.3. Metrics

Regarding quantitative evaluation, we adopt standard metrics, including mean square errors (MSE$\times$100) and bad pixel ratios (BadPix(σ)). MSE$\times$100 represents the mean square error of all pixels within a given mask multiplied by 100. Meanwhile, BadPix(σ) specifies the percentage of pixels for which the absolute difference between the actual label at a given mask and the algorithm's predicted outcome exceeds a threshold σ. Typically, σ is selected as 0.01, 0.03, and 0.07.

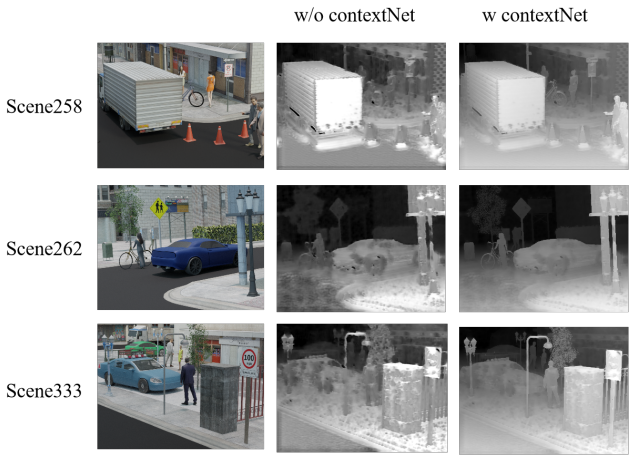

Figure 4. Disparity maps of the predicted scenes with and without ContextNet.

4.4. Compared to state-of-the-art

We compare our proposed method with other approaches submitted on the UrbanLF-Syn dataset. Table 2 presents a summary of the results. In comparison, our proposed method achieves a MSE$\times$100 of 5.989 on the UrbanLF-Syn dataset, outperforming most of the above-mentioned methods. Our method utilizes a novel hybird cost volume mechanism and occlusion-aware loss to better capture scene geometry in textureless and occlusion regions.

4.5. Ablation Study

In this section, we present an ablation study to analyze the individual contributions of various components in our proposed method. We perform this study on UrbanLF-Syn and HCInew datasets and report the results in Table 1.

Firstly, we compare our full model against a baseline method that does not include the occlusion-aware loss or any other additional components. The baseline obtains an MSE$\times$100 of 3.692 and a BadPix(0.01) of 84.842 on the test set. Then, we evaluate the contribution of the occlusion-aware loss by removing it from the full model. This results in a decrease in both the MSE$\times$100 and BadPixs, demonstrating the effectiveness of this loss in addressing occlusion challenges. The visualized results are shown in Fig. 3. We also examine the effect of our hybird cost volume network by removing it from the full model. This results in a slight

decrease in the MSE$\times$100, the BadPix(0.01) remains relatively unchanged, and more qualitative results are shown in Fig. 4. Overall, the ablation study confirms the efficacy of our proposed method, demonstrating that each component provides a valuable contribution towards achieving better performance on both UrbanLF-Syn and HCInew datasets.

5. Limitations

Despite achieving better reconstruction results, our proposed method still has some limitations, which we discuss in this section. Firstly, our method heavily relies on the availability of high-quality light field data. The quality of the estimated depth maps is directly influenced by the resolution, angular and spatial sampling density of the input light field images. Therefore, the performance of our method may degrade when applied to low-quality or sparsely sampled light field data. Secondly, our method assumes a static scene and does not consider moving objects in the scene. This is a particular limitation for dynamic scenes such as autonomous driving scenarios, where objects may move at different speeds and directions. Future work could explore incorporating motion estimation into our framework to handle dynamic scenes. Lastly, our proposed method is computationally expensive due to the use of hybrid cost volume and the multi-scale architecture. This limits its real-time applicability and could hinder its adoption in resource-constrained settings. Future work should focus on developing more efficient architectures that can achieve comparable performance while reducing computational costs.

6. Conclusion

In this paper, we propose a novel method for light field depth estimation by leveraging hybrid cost volume and an occlusion mechanism. Our method achieves better performance on the challenging UrbanLF and HCI datasets, outperforming most existing approaches in terms of accuracy and robustness. We introduce a hybrid cost-volume network that focuses on both local matching information and global context features while preserving details in textureless regions. Additionally, we proposed a new occlusion-aware loss function that encourages the network to focus more on occlusion areas, which leads to more accurate depth estimation. In conclusion, our proposed method shows promising results for various datasets and will stimulate further research in this area.

Acknowledgement

This study is partially supported by the National Key R&D Program of China (No.2022YFC3803600), the National Natural Science Foundation of China (No.61872025), and the Open Fund of the State Key Laboratory of Software Development Environment (No.SKLSDE-2021ZX-03). Thank you for the support from HAWKEYE Group.

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*, volume 35, pages 1009–1017, 2021. 1, 3

[2] 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. 3

[3] 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. 3

[4] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pages 19–34. Springer, 2017. 2, 5

[5] Zhicong Huang, Xuemei Hu, Zhou Xue, Weizhu Xu, and Tao Yue. Fast light-field disparity estimation with multi-disparity-scale cost aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6320–6329, 2021. 1, 3, 4

[6] 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. 3

[7] Changil Kim, Henning Zimmer, Yael Pritch, Alexander Sorkine-Hornung, and Markus H Gross. Scene reconstruction from high spatio-angular resolution light fields. *ACM Trans. Graph.*, 32(4):73–1, 2013. 1

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

[9] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE intelligent vehicles symposium (IV)*, pages 163–168. IEEE, 2011. 1

[10] Yaoxiang Luo, Wenhui Zhou, Junpeng Fang, Linkai Liang, Hua Zhang, and Guojun Dai. Epi-patch based convolutional neural network for depth estimation on 4d light field. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part III 24*, pages 642–652. Springer, 2017. 3

[11] Kazu Mishiba. Fast depth estimation for light field cameras. *IEEE Transactions on Image Processing*, 29:4232–4242, 2020. 1

[12] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 1

[13] 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. 3

[14] Jiayong Peng, Zhiwei Xiong, Dong Liu, and Xuejin Chen. Unsupervised depth estimation from light field using a convolutional neural network. In *2018 International Conference on 3D Vision (3DV)*, pages 295–303. IEEE, 2018. 3

[15] Jiayong Peng, Zhiwei Xiong, Yicheng Wang, Yueyi Zhang, and Dong Liu. Zero-shot depth estimation from light field using a convolutional neural network. *IEEE Transactions on Computational Imaging*, 6:682–696, 2020. 1, 3

[16] Hendrik Schilling, Maximilian Diebold, Carsten Rother, and Bernd Jähne. Trust your model: Light field depth estimation with inline occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4530–4538, 2018. 3

[17] Hao Sheng, Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, and Zhenglong Cui. Urbanlf: a comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7880–7893, 2022. 2, 5

[18] Hao Sheng, Pan Zhao, Shuo Zhang, Jun Zhang, and Da Yang. Occlusion-aware depth estimation for light field using multi-orientation epis. *Pattern Recognition*, 74:587–599, 2018. 3

[19] 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, 3

[20] Michael W Tao, Sunil Hadap, Jitendra Malik, and Ravi Ramamoorthi. Depth from combining defocus and correspondence using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 673–680, 2013. 3

[21] Michael W Tao, Jong-Chyi Su, Ting-Chun Wang, Jitendra Malik, and Ravi Ramamoorthi. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1155–1169, 2015. 3

[22] 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, 3, 4

[23] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 3487–3495, 2015. 3

[24] 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, 4

[25] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 1, 3

[26] Zian Wang, Wenzheng Chen, David Acuna, Jan Kautz, and

Sanja Fidler. Neural light field estimation for street scenes with differentiable virtual object insertion. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 380–397. Springer, 2022. 1

[27] Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013. 3

[28] Jingyi Yu. A light-field journey to virtual reality. *IEEE MultiMedia*, 24(2):104–112, 2017. 1

[29] 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. 3

[30] Wenhui Zhou, Enci Zhou, Gaomin Liu, Lili Lin, and Andrew Lumsdaine. Unsupervised monocular depth estimation from light field image. *IEEE Transactions on Image Processing*, 29:1606–1617, 2019. 3