LFNAT 2023 Challenge on Light Field Depth Estimation: Methods and Results

Hao Sheng, Yebin Liu, Jingyi Yu, Gaochang Wu, Wei Xiong, Ruixuan Cong*,
Rongshan Chen*, Longzhao Guo, Yanlin Xie, Shuo Zhang, Song Chang, Youfang Lin,
Wentao Chao, Xuechun Wang, Guanghui Wang, Fuqing Duan, Tun Wang,
Da Yang, Zhenglong Cui, Sizhe Wang, Mingyuan Zhao, Qiong Wang, Qianyu Chen,
Zhengyu Liang, Yingqian Wang, Jungang Yang, Xueting Yang, Junli Deng,

Abstract

This paper reviews the 1st LFNAT challenge on light field depth estimation, which aims at predicting disparity information of central view image in a light field (i.e., pixel offset between central view image and adjacent view image). Compared to multi-view stereo matching, light field depth estimation emphasizes efficient utilization of the 2D angular information from multiple regularly varying views. This challenge specifies UrbanLF [20] light field dataset as the sole data source. There are two phases in total: submission phase and final evaluation phase, in which 75 registered participants successfully submit their predicted results in the first phase and 7 eligible teams compete in the second phase. The performance of all submissions is carefully reviewed and shown in this paper as a new standard for the current state-of-the-art in light field depth estimation. Moreover, the implementation details of these methods are also provided to stimulate related advanced research.

1. Introduction

Over the past few years, light field (LF) has gradually developed into one of the mainstream research areas of computer vision. Benefiting from the potential capability of additional directional information, LF simultaneously records both spatial information and angular information of all light rays. In addition, with the advent of plenoptic cameras, LF acquisition is greatly simplified and can be obtained in a single shot. Therefore, a large variety of research fields try to use LF instead of single image or video sequence for superior performance by mining internal structural information.

As a fundamental and crucial step, LF depth estimation provides geometric information of the scene and is considered as a basis for various researches, such as super resolution [31], view synthesis [11] and image segmentation [35]. Compared to the classical multi-view stereo matching problem, LF contains dense and regular sampled views, making it possible to design novel and accurate methods tailored for depth reconstruction.

However, it is challenging to extract the depth information recorded in LF images. Conventional methods mainly concentrate on exploiting photo-consistency among sub-aperture images (SAIs) [19, 21], linear structures in epipolar plane images (EPIs) [39] as well as focusness in focal stacks [17]. They are difficult to be deployed in applications due to the heavy computation costs. With the emergence of deep learning technologies, a series of methods [22, 26, 30] based on convolutional neural network (CNN) have been proposed to solve this problem. They suffer from occlusion, texture-less or other regions that do harm to LF structure.

In order to stimulate scientific progress and inspire new solution for this problem, the 1st light field depth estimation challenge on LFNAT 2023 workshop is held on schedule. This challenge focuses on predicting disparity information for central view image, discards the commonly used LF datasets (*i.e.,* HCIold [34] and HCInew [9]) and chooses large-scale challenging UrbanLF [20] for training and evaluating models. Moreover, this challenge supports detailed comparisons among different methods and takes a step forward in benchmarking LF depth estimation.

2. Related Work

2.1. Conventional Methods

Conventional methods adopt different attributes of LF to obtain the depth information of scene, and have different characteristics and applicable ranges. For instance, EPI-based methods project the same points into different views

Hao Sheng, Yebin Liu, Jingyi Yu, Gaochang Wu and Wei Xiong are the organizers of LFNAT 2023 challenge, Ruixuan Cong and Rongshan Chen (State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, and Beihang Hangzhou Innovation Institute Yuhang. email: {congrx, rongshan}@buaa. edu.cn) are corresponding authors, while other authors are the participants of LFNAT 2023 challenge on light field depth estimation.

Section 7 provides affliations of each organizer and participant.

LFNAT CVPR workshop webpage: http://lfchallenge.com/LFNAT@CVPR2023/home/

of LF, forming lines in EPI whose slope has a linear relationship with disparity. Multi-view based methods regard LF as multi-view image array, using color consistency to estimate depth based on multi-view stereo. Focus stack-based methods apply refocusing technology to generate multiple refocusing images and obtain depth through the focal length that makes image clearest.

2.1.1 EPI-Based Methods

Bolles et al. [1] first propose the concept of EPI, and detecte edge, crest and trough through linear fitting in EPI. Wanner et al. [33] propose a globally consistent labeling algorithm, which uses structure tensor to extract direction of lines in EPI and obtains depth through global optimization. Inspired by this, Li et al. [14] use iterative conjugate gradient method to build a sparse linear system, getting depth information in EPI. Kim et al. [12] carry out sparse representation of LF, and utilize a fine-to-coarse depth propagation strategy. Zhang et al. [39] propose a spinning parallelogram operator to calculate the slope of line in EPI by distance. Zhang et al. [40] select optimal slope of line from EPI, and divide pixels into reliable and unreliable category, in which the latter is filled via disparity propagation of the former.

2.1.2 Multi-View-Based Methods

Yu et al. [38] explore geometric structure of 3D lines in ray space to improve the triangulation and stereo matching of LF. Heber et al. [8] construct a matching item of principal component analysis for multi-view stereo reconstruction, assuming that the matrix is low-rank if each SAI is projected to the same center and each projected image is regarded as a row. Chen et al. [4] model 3D point angular radiosity distribution, and introduce bilateral consistency metric. Jeon et al. [10] propose a phase shift theory based on Fourier transform to solve small disparity situation. They innovate subpixel-level offset into phase shift in frequency domain. Liu et al. [18] acquire LF video and depth map through Fourier phase shift and graph cutting.

2.1.3 Focus Stack-Based Methods

Tao et al. [23] integrate defocusing clues and consistency clues to get depth maps for the first time. They [24] further use color, depth and shadow consistency to correct results. Wang et al. [28] propose a occlusion perception method by ensuring imaging consistency restricted in occluded view regions. Williem et al. [36] use angular entropy and adaptive defocusing to improve the robustness and noise sensitivity based on occlusion model. Zhu et al. [41] extend the above method to the case of multiple occlusions. Tian et al. [25] propose three-clue fusion methods including defocusing, correspondence and propagation, thus eliminating the influence of light backscattering and attenuation.

2.2. Learining Methods

With the wide application of deep learning in advanced computer vision tasks such as image classification, segmentation and recognition, LF depth estimation method based on deep learning comes into being. Different from conventional methods, deep learning methods are mainly divided into two mainstreams: EPI-based methods and cost volume-based methods.

2.2.1 EPI-Based Methods

Shin et al. [22] propose EPINet, the first end-to-end network using CNN to extract EPI geometry disparity. They also come up with data augmentation methods tailored for LF. Leistner et al. [13] design EPI-shift to virtually shift LF stack which enables to retain a small receptive field to be effective in the case of wide-baseline. Li et al. [15] construct an oriented relation module to extract oriented relation features between center pixel and its neighborhood from EPI patches. Hassan et al. [6] improve a light-weight epinet architecture that can quickly calculate disparity. It greatly improves the speed and reduces calculation consumption. Li et al. [16] introduce the transformer [27] into LF depth estimation. They combine EPI feature extraction with transformer to establish global features.

2.2.2 Cost Volume-Based Methods

Tsai et al. [26] first implement cost volume with CNN by sequentially shifting each SAI with a series of predefined disparity. Chen et al. [5] develop a new solution with four branches and use attention mechanism to establish relationship among cost volumes. Wang et al. [32] design a class of domain-specific convolutions to disentangle LF from different dimensions, and then leverage these features to construct cost volume. They [30] further propose an occlusion-aware cost constructor to handle occlusions by modulating pixels from different views. Chao et al. [3] construct a more refined cost volume at the sub-pixel level and propose an elaborate loss function. Wang et al. [29] divide the LF into four regions to build a four-branch cost volume to reduce computational consumption and get a more accurate result.

3. LFNAT Depth Estimation Challenge

The objectives of LFNAT 2023 challenge on LF depth estimation are: (1) measure and compare the state-of-the-art methods in related research field. (2) push new solutions with high efficiency as well as accurate performance. (3) promote a novel benchmark to replace the widely used one[1]

[1] http://www.lightfield-analysis.net

Figure 1. Example central view images and corresponding ground truth disparity images from UrbanLF-Syn.

for further advanced research.

3.1. Dataset

This challenge takes UrbanLF [20] as sole data source. As a high-quality and challenging urban scene dataset, UrbanLF aims at understanding complex urban scenes through the rich information in 4D LF to improve the practical system performance and reliability. Since there is no available ground truth disparity information for real-world sample, all experiments are only performed on UrbanLF-Syn subset created by Blender[2] software. Specifically, there are 170 synthetic samples for training, 30 samples for validation and 30 samples for test. Each sample is composed of 81 sub-aperture images with an angular resolution of 9×9 and a spatial resolution of 640×480, as well as ground truth depth image and disparity image of all views. In addition, UrbanLF-Syn subset contains densely sampled LF with a minimum disparity range [-0.47,1.55] pixels between adjacent views. Fig. 1 shows some representative samples in UrbanLF-Syn. For detailed description, please refer to the official repository[3].

3.2. Timeline

This challenge starts at January 24, 2023, ends at April 6, 2023, lasting for 73 days in total. It is divided into the following two phases:

1. **Submission phase** ($1.24 \sim 3.28$): The participants can download UrbanLF-Syn subset to develop models for generating predict disparity results. Specifically, training and validation samples are fully publicly available to participants while ground truth depth and disparity information for 30 test samples remains confidential to equally evaluate submitted works. It is worth noting that disparity range of each test sample is also released. The participants should organize their prediction files according to the official instruction for submission and evaluation. A performance leaderboard is available online to compare the submitted methods developed by different participants. In order to have access to submit prediction files to the challenge evaluation server, each participant should register an account on http://www.lfchallenge.com/main/.

2. **Final evaluation phase** ($3.28 \sim 4.6$): All methods in the submission phase are first compared with several baseline models [3, 22, 26, 30, 32] (provided by challenge organizer), and only methods with higher performance than baseline models can enter the final evaluation phase. All participants of the methods that meet the condition are required to resubmit their prediction files (**must be the same as the prediction files in the submission phase**), and a 2-page extended abstract which should include a short description of their methods and resulting insights about performance. Participants that do not submit the method description before deadline will not be counted in final ranking.

3.3. Evaluation Metrics

The evaluation method only needs to predict disparity results of central view rather than all views. The mean square disparity error (MSE) and the bad pixel ratio (BP) with three thresholds (0.01, 0.03 and 0.07 pixels) are the quantitative measures. The calculation formulas are as follows:

$$MSE = \sum_{i}^{n} (pre(i) - gt(i))^2 \tag{1}$$

$$BP(\sigma) = \frac{1}{n} \sum_{i}^{n} (|pre(i) - gt(i)| \geq \sigma) \tag{2}$$

[2]https://www.blender.org/
[3]https://github.com/HAWKEYE-Group/UrbanLF

Rank	Team	Method	MSE	BP			Params.
				BP(0.01)	BP(0.03)	BP(0.07)	
1	$INSIS_1$	CBPP	<u>0.394 / 2</u>	<u>27.385 / 2</u>	**12.628 / 1**	**5.907 / 1**	5.00M
2	$INSIS_2$	LRDE	**0.368 / 1**	27.802 / 3	12.825 / 3	<u>6.205 / 2</u>	0.16M
3	BNU-AI-TRY	SF-Net	0.416 / 3	**24.681 / 1**	<u>12.649 / 2</u>	6.750 / 3	5.06M
4	HawkeyeGroup	EPI-Cost	0.738 / 4	57.946 / 6	27.041 / 5	14.327 / 5	5.51M
5	eker	MTLF	1.156 / 6	46.933 / 4	22.852 / 4	13.518 / 4	0.32M
6	AnsLab301	ConvCC	0.953 / 5	53.926 / 5	28.211 / 6	15.582 / 6	5.02M
7	CUC001team	Hybrid CV	5.989 / 7	90.850 / 7	77.411 / 7	59.738 / 7	5.11M
-	baseline	LFattNet [26]	1.723	86.592	63.847	39.320	5.06M
-	baseline	EPINet [22]	1.948	90.809	73.352	34.004	5.12M
-	baseline	SPO [39]	8.617	68.952	42.276	30.123	-

Table 1. LFNAT 2023 Light Field Depth Estimation Challenge results, final rankings and network parameters on UrbanLF-Syn test set. Note that column MSE and column BP are displayed in a form like value / ranking. The best results are in bold and the second results are underlined. The comprehensive ranking is determined by averaging MSE ranking and BP ranking, in which BP ranking result is calculated by averaging BP(0.01), BP(0.03) and BP(0.07) ranking result.

where i is a single pixel. n is the total number of pixels. σ is the threshold value. pre and gt denote the predict result and ground truth, respectively. For these metrics, they are computed by averaging over all test samples and small value signifies good performance. The submitted results are finally ranked by averaging MSE ranking result and BP ranking result, in which BP ranking result is calculated by averaging BP(0.01), BP(0.03) and BP(0.07) ranking result.

4. Challenge Results

From 75 registered participants, 7 teams successfully enter in the final evaluation phase and submit results, codes and extended abstract. Tab. 1 reports the final ranking of the challenge, as well as single evaluation metric results (MSE and BP) and network parameters. The methods are briefly described in Sec. 5 and the corresponding teams and affiliations are listed in Sec. 7.

Architectures and main ideas. All competition methods apply deep learning techniques and extend the stereo-matching framework based on cost volume. Among them, 3 participants explore the computational element of matching cost, and enhance its robustness by introducing the unique LF features; 3 methods focuse on the way of cost construction, such as faster convolutional constructor and finer sub-pixel constructor; 1 solution is to study the post-processing technology for cost volume refinement. And by better handling the texture-less areas, the INSIS team wins the championship and runner-up in this challenge.

Forecast Accuracy. From Tab. 1, it can be observed that the INSIS team achieve the 1st and 2nd ranking by well handling the non-texture problem in UrbanLF, and the proposed LRDE exceed the 3rd solution 0.048 in terms of MSE with only about 3% network parameters, which is competitively competitive in all participation methods. Additionally, the top of 6 solutions surpass the state-of-the-art method LFattNet, and 4 of them double the accuracy of MSE compared to previous works on UrbanLF, which significantly boosts the performance on light field depth estimation.

Data Preprocessing and Augmentation. Similar to previous learning-based methods, all participating teams adopt numerous data augmentation strategies, such as scale enhancement, transposing, rotating, and color transformation. However, the view augmentation approach is not used here and most methods take all 9×9 LF as input. Additionally, 3 teams explicitly remove texture-less regions to reduce interference; 3 teams adjust the predefined disparity range according to the characteristics of UrbanLF; and 1 participant rearrange the training and validation subset for more trainable data.

Conclusions. By analyzing the experimental settings, the proposed methods and their results, we can conclude that:

- The competition methods significantly boost the performance in LF depth estimation.
- The stereo matching theory of cost volume is still popular and useful in LF depth estimation, and there still exists a large room for exploration, such as cost constructor, post-processing.
- It seems that more attention needs to be paid to the handling of texture-less regions in LF, only 2 participants

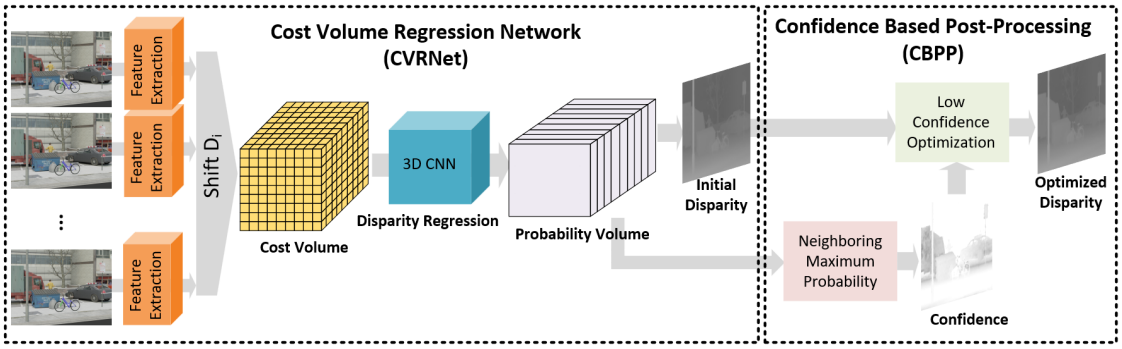

Figure 2. INSIS$_1$ Team: The network architecture of the proposed CBPP.

give some solutions for it except completely removing.

5. Teams and Methods

In this section, all methods are described in order of final ranking from top to bottom. For simplicity, we uniformly specify U×V and H×W as the angular resolution and spatial resolution of the input LF, respectively.

5.1. INSIS$_1$: CBPP

Most recent deep learning methods for LF depth estimation generate results by constructing 3D cost volume. Due to memory and time limitation, they use small patches for training instead of entire images. However, such strategy poses difficulty to generate final depth map using information from entire image. As a result, the model may perform poorly in some areas, especially large texture-less regions.

To address this issue, this team proposes a Confidence-Based Post-Processing (CBPP) method. They derive confidence from probability volume generated by cost volume and optimize depth values in low confidence regions. Their method effectively improves results in texture-less regions without introducing any additional parameters.

Theoretically speaking, their post processing method is applicable to all networks that build 3D cost volume, such as [2, 26]. Here, they adopt a basic network shared by these methods, named Cost Volume Regression Network (CVRNet). The overall network architecture is shown in 2. They use the same feature extraction structure and 3D disparity regression strategy as [2].

Neighboring Maximum Probability. Methods that generate depth by constructing 3D cost volume usually perform 3D convolution operations on cost volume to obtain probability volume $P \in R^{D\times H\times W}$, where D represents disparity levels. Given a certain pixel $p(h, w)$, the probability distribution $P(:, h, w)$ of the pixel p must be centralized at one disparity level when the estimated depth value of p is correct. Conversely, if the probability of one pixel p is dispersed, it suggests an inaccurate estimated depth value. Therefore, they define the confidence $C \in R^{H\times W}$:

$$C(h,w) = max(\sigma(\sum_{i=0}^{1} P(d_i,h,w), ..., \sum_{i=D-2}^{D-1} P(d_i,h,w))) \tag{3}$$

where σ denotes the softmax operation.

Algorithm 1: Confidence Based Post-Processing

Input: Confidence map C, initial depth map D^{init}, center image I, window W_i, three thresholds σ_l, σ_h, σ_c
Output: Optimized depth map D^{opt}
for p *in* W_i **do**
  if $C_p < \sigma_l$ **then**
    for q *in* W_i **do**
      if $C_q > \sigma_h \&\& |I_q - I_p| < \sigma_c$ **then**
        $List_p$.append(D_q^{init})
      end
    end
    $avg(List_p) \rightarrow D_p^{opt}$
  else
    $D_p^{init} \rightarrow D_p^{opt}$
  end
end

Low Confidence Optimization. After obtaining the confidence of each point, they seek to improve results in low confidence regions (mainly texture-less regions). Texture-less areas typically exhibit similar color and continuous surface. Thus they define a window in which the depth with low confidence is approximately consistent, then refine it with depth values from high confidence regions. Moreover, to improve depth results on entire map, the window must be global, not limited to a small patch. Specifically, they divide entire map of size 480×640 into several windows of

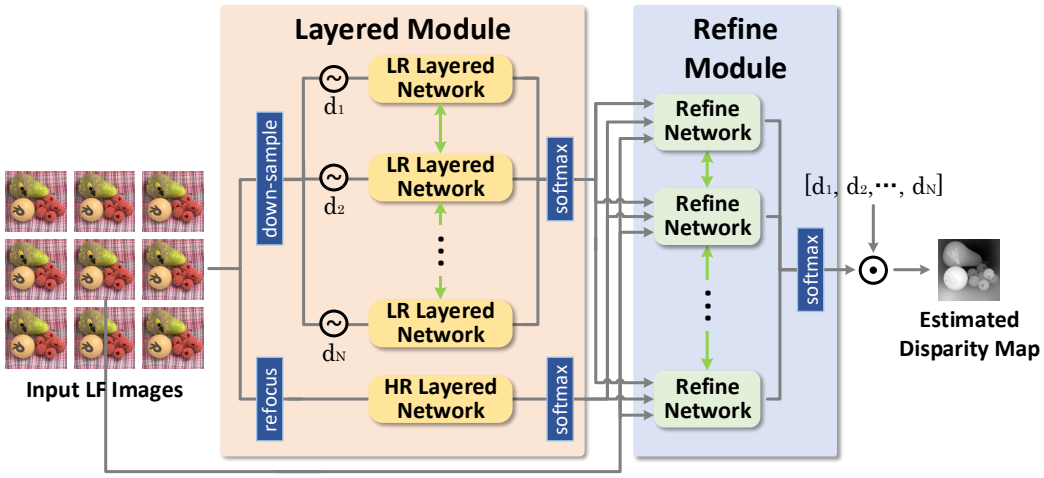

Figure 3. INSIS$_2$ Team: The network architecture of the proposed LRDE.

size 1×640. Also, to distinguish between high confidence and low confidence, they set two thresholds (σ_l and σ_h). For each window W_i, they compute average depth d_i of pixels with high confidence, and assign d_i to pixels with low confidence in W_i. The processing is shown in Tab. 1. During training, thresholds $\sigma_l, \sigma_h, \sigma_c$ are set as 0.227, 1.231, 16.

5.2. INSIS$_2$: LRDE

This team proposes a novel Layered Refined Depth Estimation (LRDE) method for LF depth estimation, following their previous method [2] for LF image super-resolution with the hybrid lenses LF images. For weak texture regions in the depth estimation, the results may be wrong due to the lack of rich texture information for matching. Therefore, they downsample the original LF to reduce the influence of weak texture regions. At the same time, the high-resolution (HR) LF images are used to supplement the high-frequency information to refine rich texture regions for better details.

The overall network architecture is shown in Fig. 3. It includes two modules: Layered Module and Refine Module. In Layered Module, they firstly use the downsampled LF images and HR focal stack to generate two sets of coarse alpha maps of multi-plane images (MPI). Then, in Refine Module, two sets of alpha maps are fused and combined with HR central view to generate refined counterparts. Finally, the refined alpha maps are used to generate the disparity map of central view.

Layered Module. The original LFs are firstly downsampled or refocused to obtain low-resolusion (LR) LF images and focal stack. For LR LF images, they shift every view toward central view according to a set of disparity values and obtain view stacks, in which every stack contains U×V views. Then, they design Layered Network to calculate the coarse alpha maps. Specifically, for every view stack, they first concatenate all views in channel dimension, then feed them into LR Layered Network, which has multiple cascaded 2D CNN residual blocks and a 2D deconvolutional layer with one output channel in the last. To ensure that the most appropriate alpha map is selected, they apply a softmax normalization to generate coarse alpha maps. Note, for every view stack, the parameters of Layered Network are shared. For focal stack, they feed it into an HR Layered Network, which is similar to LR Layered Network, just replacing 2D CNN with 3D CNN and removing deconvolutional layer to obtain coarse alpha maps from focal stack.

Refined Module. The two sets of coarse alpha maps and HR central view are fed into Refine Network, which follows U-net structure, to interact with high and low-frequency information for refinement. Especially, they introduce HR central view with the corresponding scale only in the encoding stage of U-Net. It can not only adequately explore and provide the high-frequency information in HR central view, but also not interfere the final output during decoding stage. Besides, the average pooling layer is replaced by convolutional layer and the refined alpha maps are obtained after a softmax normalization. Similar to Layered Network, the parameters of Refine Network are also shared in every alpha map. Finally, they use the refined alpha maps and a set of disparity values to synthesize disparity map.

5.3. BNU-AI-TRY: SF-Net

This team proposes a novel yet effective method called SF-Net based on SubFocal [3] to learn multi-scale features of extensive texture-less regions for LF depth estimation. The overall network architecture is shown in Fig. 4. First, the features of each SAI are extracted using a shared fea-

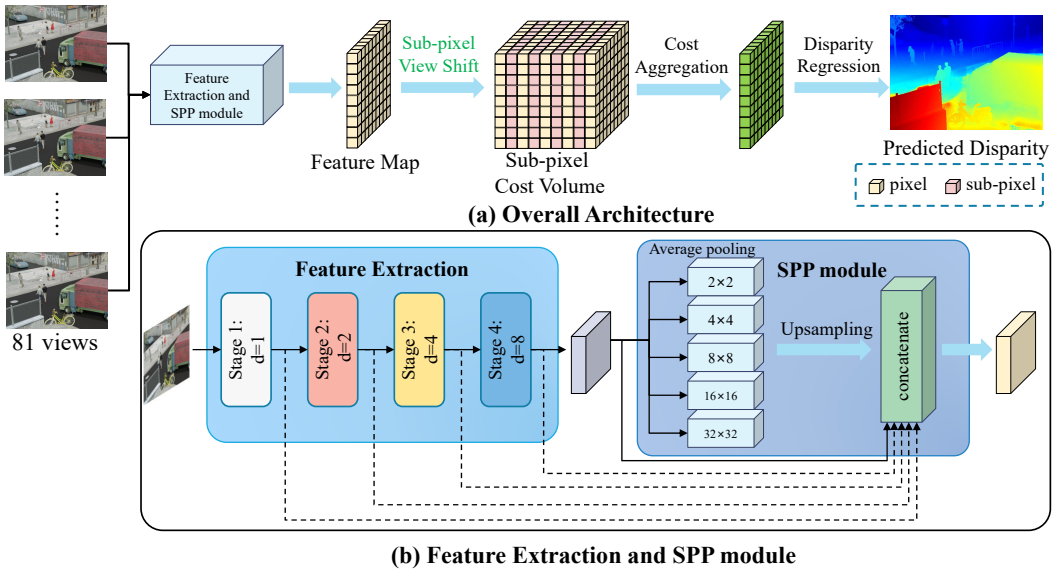

Figure 4. BNU-AI-TRY Team: The network architecture of the proposed SF-Net.

ture extraction based on dilated convolution [37] and spatial pyramid pooling (SPP) module [7]. Second, the sub-pixel view shift is performed to construct sub-pixel cost volume. Third, the cost aggregation module is used to aggregate cost volume information. The predicted disparity map is produced by attaching a disparity regression module.

Feature Extraction and SPP Module. After getting initial feature, they use a feature extraction module based on dilated convolution and SPP to extract features of SAI. Concretely, it contains four stages with 1, 2, 4, and 8 dilation rate. Different from previous methods [3, 26], an extra average pooling with 32×32 size is inserted into SPP, resulting in five pooling operations at different scales to compress the features. Bilinear interpolation is adopted to upsample low-dimensional features. Finally, features of different stages are concatenated to improve discrimination by skip connection, and the channel dimension is reduced via convolution. In this way, The output feature contains multi-scale discriminative context information and incorporates additional information from neighboring regions for challenging scenarios, such as texture-less and reflection areas.

Sub-pixel Cost Volume. In order to handle narrow baseline of LF, different from previous method [10] using phase shift theorem to construct image level cost volume, they follow [3] to form a sub-pixel feature level cost volume within predefined disparity via bilinear interpolation. After shifting features, they concatenate them into a 4D cost volume $R^{D\times H\times W\times C}$. Considering that a smaller sampling interval can generate a finer sub-pixel cost volume but increase computation and slow down inference, they adopt 22 disparity levels ranging from -0.5 to 1.6, where the sub-pixel interval is 0.1 to trade off accuracy and speed.

Cost Aggregation and Disparity Regression. Following [3, 26], eight 3×3×3 3D convolutional layers and two residual blocks are used to cost aggregation for final cost volume $C_f \in R^{D\times H\times W}$. Then, they normalize C_f with softmax operation and calculate output disparity of central view $\hat{d}$:

$$\hat{d} = \sum_{d_k=D_{min}}^{D_{max}} d_k \times \mathrm{softmax}(-C_{d_k}), \tag{4}$$

where D_{min} and D_{max} stand for disparity range. d_k is the specific sampling value.

5.4. HawkeyeGroup: EPI-Cost

This team uses EPI lines to guide cost volume construction, which not only takes advantage of the relationship between EPI but also overcomes the problem of noise and occlusion during matching to a certain extent. The overall network architecture is shown in Fig. 5. Firstly, they design an efficient two-branch EPI extraction module to extract horizontal and vertical line features. Secondly, guided by these characteristics, they construct two special cost volumes in a specific order and match them. Thirdly, the horizontal and vertical volumes are merged using the intervolume attention mechanism. Finally, they use the 3D CNN to aggregate and regress it to get the final disparity map.

EPI Extraction Module. The input is horizontal and vertical center images of the SAI. They design a two-branch

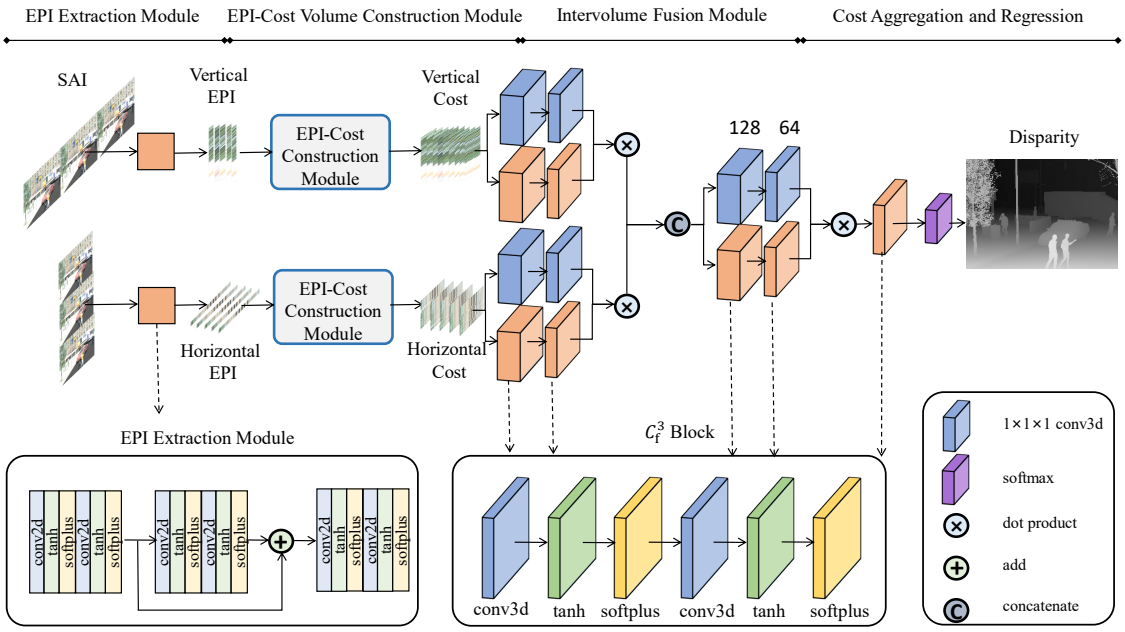

Figure 5. HawkeyeGroup Team: The network architecture of the proposed EPI-Cost.

feature extraction structure, and the horizontal and vertical images are convolved separately. Then, the residual training method is used to thoroughly combine the initial, intermediate, and final features.

EPI-Cost Volume Construction Module. Firstly, they select nine equally spaced disparity values from the range. Secondly, they manually shift the horizontal EPI lines along the vertical direction according to their offset and disparity ranges. Similarly, vertical EPI lines are carried out. Then, they concatenate the shifted EPI lines in the feature dimension into 5D cost volume, which includes color consistency, EPI line characteristics, and disparity information.

Intervolume Fusion Module. Firstly, they extract local attention weights by using 3D $conv + bn$, and global attention weights by using the adaptive average pooling layer for two cost volumes. Secondly, they concatenate the cost map of two branches and use a 2D convolution to fuse information. Finally, the two EPI-Cost volumes are multiplied by attention maps and concatenated.

5.5. eker: MTLF

In order to be implemented on resource-constrained devices, memory consumption is an important aspect of deploying efficient depth estimation models for LF images. This team is motivated to design a memory-efficient network. The overall pipeline is illustrated in Fig. 6. The proposed MTLF consists of several prominent components, i.e.

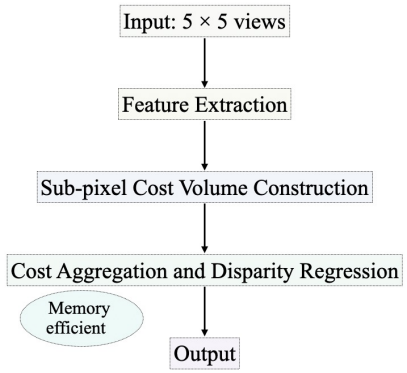

Figure 6. eker Team: The network pipeline of the proposed MTLF.

"Feature Extraction", "Sub-pixel Cost Volume Construction" and "Cost Aggregation and Disparity Regression".

The proposed MTLF is implemented based on [3]. Compared with the baseline, the center 5×5 SAIs are extracted from the given LF image, and then fed into the network. In "Feature Extraction", the Spatial Pyramid Pooling module adopts 2d convolutional layer to compress the features. In "Cost Aggregation", the input sizes of "Conv3D_2", "ResBlock3D_2", "Conv3D_3" "Cost", and the output sizes of "Conv3D_1", "Conv3D_2", "ResBlock3D_2" "Conv3D_3" are set as $R^{D\times H\times W\times 32}$, where D represents disparity levels. Note that the texture-less regions are excluded from the training data.

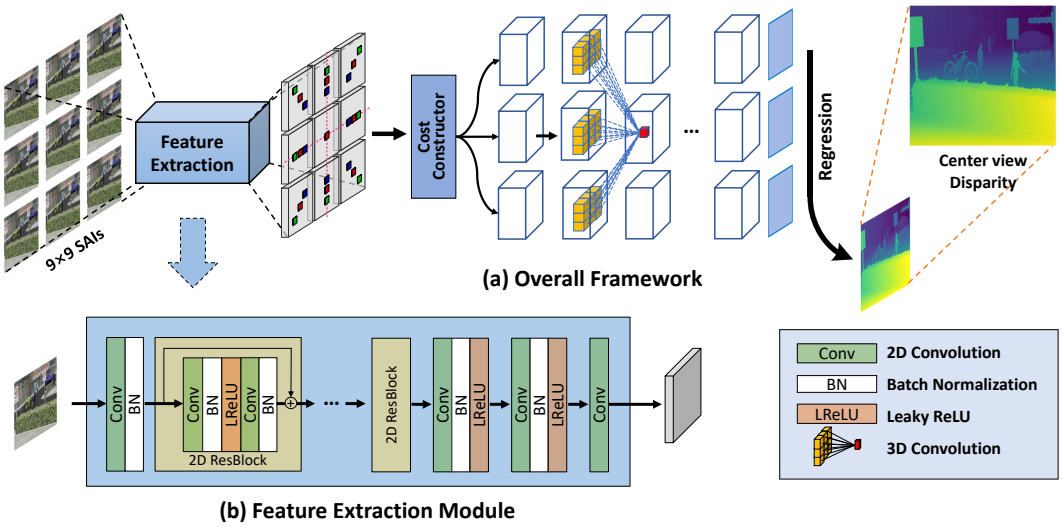

Figure 7. AnsLab301 Team: The network architecture of the proposed ConvCC.

5.6. AnsLab301: ConvCC

The method of ConvCC is inspired by recent OACC-Net [30] for LF depth estimation, in which the predefined disparities can be integrated without performing the shifting operation. The overall network architecture is shown in Fig. 7. Following the pipeline of OACC-Net, they design a Convolutional Cost Constructor cascading a series of dilated convolutions according to the disparity range, and construct an improved model with better convergence.

Feature Extraction. The proposed ConvCC takes $U \times V$ SAIs as inputs and adopts a spatial convolution for initial feature extraction. The feature is then fed into deep feature extraction cascading 8 residual blocks. After 3 spatial convolutions, it generates final features for cost construction.

Cost Construction. Their cost constructor is designed as a sequence of convolutions whose kernel size is consistent with angular dimension (*i.e.,* $U \times V$), and dilation rate is dependent on spatial dimension (*i.e.,* $H \times W$) and predefined disparity (*i.e.,* d). The dilation rate is formulated as:

$$dila(d) = [H - d, W - d] \tag{5}$$

This formula shows that object with larger disparity corresponds to a smaller atrous rate, which is consistent with the shift-and-concatenate situation.

To reduce the matching aliasing, a zero-padding strategy is introduced to each SAI, whose vertical padding value η_h and horizontal padding value η_w are defined as:

$$\eta_h = \frac{U-1}{2} \times \tilde{d}, \eta_w = \frac{V-1}{2} \times \tilde{d} \tag{6}$$

where $\tilde{d} = max\{|d_{max}|, |d_{min}|\}$. The dilation rates are recalculated on padded SAIs and the output size of cost constructor is $(H + (U-1)(d+\tilde{d}), W + (V-1)(d+\tilde{d}))$. Therefore, cropping is required to adjust spatial resolution to original (H, W). The vertical and horizontal cropping values $c_h(d)$ and $c_w(d)$ are defined as

$$c_h(d) = \frac{U-1}{2} \times (d+\tilde{d}), c_w(d) = \frac{V-1}{2} \times (d+\tilde{d}) \tag{7}$$

Cost Aggregation and Regression. After getting matching cost, the channel number is first reduced through a 3D convolution. Several 3D convolutions and channel attention mechanism are then applied to generate final pixel-wise disparity cost volume $F_{final} \in R^{D \times H \times W}$ on center view.

$$\hat{D}_c = \sum_{d_k=d_{min}}^{d_{max}} d_k \times Softmax(F_{final}) \tag{8}$$

$\hat{D}_c$ refers to estimated disparity of center view.

5.7. CUC001team: Hybrid CV

This team utilizes SAIs based on cost volume for LF depth estimation. The overall network architecture is shown in Fig. 8. Their method comprises four modules that collaborate to achieve accurate and robust disparity estimation. The first module is a shared feature extractor, which extracts features from each SAI. These features are then utilized to construct cost volume through pixel view shift in the second module. Notably, they propose a hybrid cost volume network with two sub-modules that separately focus on 3D local matching information and 2D global context information. The third step uses cost aggregation module to synthesize hybrid cost volume data. Lastly, a disparity regression

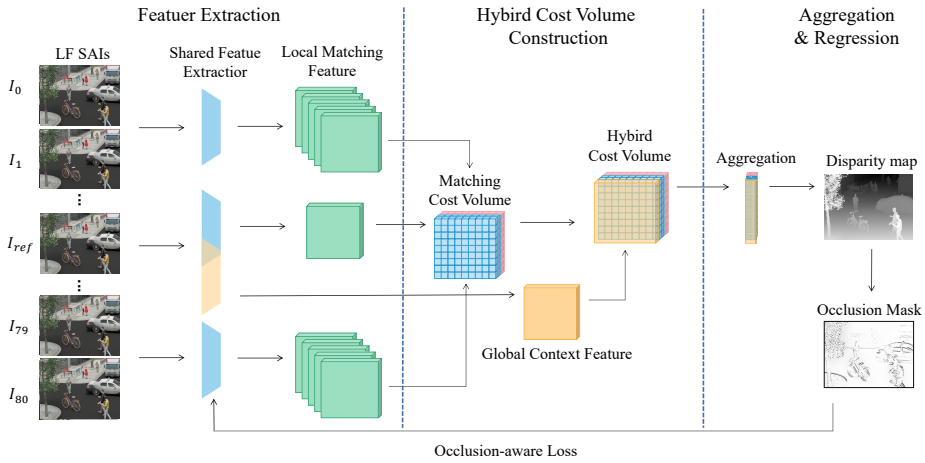

Figure 8. CUC001team Team: The network architecture of the proposed Hybrid CV.

module predicts disparity map with the supervision of the proposed occlusion-aware loss.

Hybrid Cost Volume. Firstly, they 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 as Matching Cost Volume that serves the purpose of exclusively learning local features for matching. Then, they utilize ContextNet to learn global context information and produces a volume with dimensions $R^{D\times H\times W}$. Finally, they concatenate the regularized matching cost volume with the expanded context volume to obtain a hybrid cost volume, with dimensions $R^{(C+1)\times D\times H\times W}$. They repeat this hybrid cost volume generation procedure for each SAI.

Occlusion-aware loss. They propose an occlusion-aware loss in regards to occluded regions by aggregating occlusion masks from source views. Specifically, they compute occlusion mask of center view by warping other source views using predicted disparity. Based on photometric consistency, the projected view $I_{k\to c}$ possesses identical values to center view I_c at non-occluded regions. Therefore, occlusion mask $M_{k\to c}$ is determined by calculating absolute residuals between projected view $I_{k\to c}$ and center view I_c:

$$M_{k\to c} = |\mathcal{I}_{k\to c} - \mathcal{I}_c| \tag{9}$$

They also present a view-attention network to predict a occlusion mask M_c from occlusion masks of all SAIs. The insight is that occlusion areas of central view are also related to surrounding views by causing corresponding pixels to be unmatched. Firstly, raw occlusion masks are concatenated into $R^{1\times M*M\times H\times W}$ and processed by two 2D convolution layers, reducing dimensions to $R^{1\times 1\times H\times W}$. Next, a view average pooling operation aggregates information across SAIs, resulting in a single aggregated occlusion mask for central view, namely M_c. Finally, occlusion-aware loss L_{occ} supervises predicted disparity as follows:

$$L_{occ} = \|M_c\|^{\beta} \odot \|d - \hat{d}\|_1, \tag{10}$$

where $\hat{d}$ and d represents predicted disparity and ground truth. β serves as coefficient factor regulating the dynamic weight assignment ratio. When $\beta = 0$, the occlusion loss is equivalent to the standard L_1 loss.

6. Acknowledgements

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.

We thank the LFNAT 2023 sponsors: Beijing MEET YUAN Co.,Ltd and Macao Polytechnic University.

7. Teams and Affiliations

Challenge Organizers

Members:
Hao Sheng[1,2,3] (shenghao@buaa.edu.cn)
Yebin Liu[4] (liuyebin@mail.tsinghua.edu.cn)
Jingyi Yu[5] (yujingyi@shanghaitech.edu.cn)
Gaochang Wu[6] (wugc@mail.neu.edu.cn)
Wei Xiong[7] (weixiong@lusterinc.com)

Affiliations:
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University

[2]Beihang Hangzhou Innovation Institute Yuhang
[3]Faculty of Applied Sciences, Macao Polytechnic University
[4]Tsinghua University
[5]Shanghaitech University
[6]State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University
[7]Beijing MEET YUAN Co.,Ltd

INSIS$_1$

Members: *Longzhao Guo[1] (guolongzhao@bjtu.edu.cn)*
Yanlin Xie[1], Shuo Zhang[1]
Affiliations:
[1]Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University

INSIS$_2$

Members: *Song Chang[1] (changsong@bjtu.edu.cn)*
Youfang Lin[1]
Affiliations:
[1]Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University

BNU-AI-TRY

Members: *Wentao Chao[1] (chaowentao@mail.bnu.edu.cn)*
Xuechun Wang[1], Guanghui Wang[2], Fuqing Duan[1]
Affiliations:
[1]Beijing Normal University
[2]Toronto Metropolitan University

HawkeyeGroup

Members: *Tun Wang[1] (wangtun@buaa.edu.cn)*
Da Yang[1,2], Zhenglong Cui[1,2], Sizhe Wang[1,2], Mingyuan Zhao[1]
Affiliations:
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University
[2]Beihang Hangzhou Innovation Institute Yuhang

eker

Members: *Qiong Wang[1] (wangqiong@zjut.edu.cn)*
Affiliations:
[1]College of Computer Science and Technology, Zhejiang University of Technology

AnsLab301

Members: *Qianyu Chen[1] (chenqianyu18@nudt.edu.cn)*
Zhengyu Liang[1], Yingqian Wang[1], Jungang Yang[1]
Affiliations:
[1]National University of Defense Technology

CUC001team

Members: *Xueting Yang[1] (yangxueting@cuc.edu.cn)*
Junli Deng[1]
Affiliations:
[1]School of Information and Communication Engineering, Communication University of China

References

[1] Robert C Bolles, H Harlyn Baker, and David H Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International journal of computer vision*, 1(1):7–55, 1987. 2

[2] Song Chang, Youfang Lin, and Shuo Zhang. Flexible hybrid lenses light field super-resolution using layered refinement. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5584–5592, 2022. 5, 6

[3] Wentao Chao, Xuechun Wang, Yingqian Wang, Liang Chang, and Fuqing Duan. Learning sub-pixel disparity distribution for light field depth estimation. *arXiv preprint arXiv:2208.09688*, 2022. 2, 3, 6, 7, 8

[4] Can Chen, Haiting Lin, Zhan Yu, Sing Bing Kang, and Jingyi Yu. Light field stereo matching using bilateral statistics of surface cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1518–1525, 2014. 2

[5] 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. 2

[6] Ali Hassan, Mårten Sjöström, Tingting Zhang, and Karen Egiazarian. Light-weight epinet architecture for fast light field disparity estimation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2022. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 7

[8] Stefan Heber and Thomas Pock. Shape from light field meets robust pca. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 751–767. Springer, 2014. 2

[9] 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. 1

[10] 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. 2, 7

[11] Jing Jin, Junhui Hou, Hui Yuan, and Sam Kwong. Learning light field angular super-resolution via a geometry-aware network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11141–11148, 2020. 1

[12] 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. 2

[13] Titus Leistner, Hendrik Schilling, Radek Mackowiak, Stefan Gumhold, and Carsten Rother. Learning to think outside the box: Wide-baseline light field depth estimation with epi-shift. In *2019 International Conference on 3D Vision (3DV)*, pages 249–257. IEEE, 2019. 2

[14] Jianqiao Li, Minlong Lu, and Ze-Nian Li. Continuous depth map reconstruction from light fields. *IEEE Transactions on Image Processing*, 24(11):3257–3265, 2015. 2

[15] Kunyuan Li, Jun Zhang, Rui Sun, Xudong Zhang, and Jun Gao. Epi-based oriented relation networks for light field depth estimation. In *British Machine Vision Conference (BMVC)*, 2020. 2

[16] Peng Li, Jiayin Zhao, Jingyao Wu, Chao Deng, Haoqian Wang, and Tao Yu. Opal: Occlusion pattern aware loss for unsupervised light field disparity estimation. *arXiv preprint arXiv:2203.02231*, 2022. 2

[17] Haiting Lin, Can Chen, Sing Bing Kang, and Jingyi Yu. Depth recovery from light field using focal stack symmetry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3451–3459, 2015. 1

[18] Fei Liu, Guangqi Hou, Zhenan Sun, and Tieniu Tan. High quality depth map estimation of object surface from light-field images. *Neurocomputing*, 252:3–16, 2017. 2

[19] 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

[20] 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*, 2022. 1, 3

[21] Hao Sheng, Shuo Zhang, Xiaochun Cao, Yajun Fang, and Zhang Xiong. Geometric occlusion analysis in depth estimation using integral guided filter for light-field image. *IEEE Transactions on Image Processing*, 26(12):5758–5771, 2017. 1

[22] 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, 3, 4

[23] 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. 2

[24] Michael W Tao, Pratul P Srinivasan, Jitendra Malik, Szymon Rusinkiewicz, and Ravi Ramamoorthi. Depth from shading, defocus, and correspondence using light-field angular coherence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1940–1948, 2015. 2

[25] Jiandong Tian, Zachary Murez, Tong Cui, Zhen Zhang, David Kriegman, and Ravi Ramamoorthi. Depth and image restoration from light field in a scattering medium. In *Proceedings of the IEEE international conference on computer vision*, pages 2401–2410, 2017. 2

[26] 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, 3, 4, 5, 7

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[28] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Depth estimation with occlusion modeling using light-field cameras. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2170–2181, 2016. 2

[29] Xucheng Wang, Chenning Tao, and Zhenrong Zheng. Occlusion-aware light field depth estimation with view attention. *Optics and Lasers in Engineering*, 160:107299, 2023. 2

[30] 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

[31] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. Ntire 2023 challenge on light field image super-resolution: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 1

[32] 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*, 2022. 2, 3

[33] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48. IEEE, 2012. 2

[34] Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pages 225–226. Citeseer, 2013. 1

[35] 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

[36] W Williem and In Kyu Park. Robust light field depth estimation for noisy scene with occlusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4396–4404, 2016. 2

[37] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 7

[38] Zhan Yu, Xinqing Guo, Haibing Lin, Andrew Lumsdaine, and Jingyi Yu. Line assisted light field triangulation and stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2792–2799, 2013. 2

[39] 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, 4

[40] Yongbing Zhang, Huijin Lv, Yebin Liu, Haoqian Wang, Xingzheng Wang, Qian Huang, Xinguang Xiang, and Qionghai Dai. Light-field depth estimation via epipolar plane image analysis and locally linear embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):739–747, 2016. 2

[41] Hao Zhu, Qing Wang, and Jingyi Yu. Occlusion-model guided antiocclusion depth estimation in light field. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):965–978, 2017. 2