Learning Sub-Pixel Disparity Distribution for Light Field Depth Estimation

Wentao Chao, Xuechun Wang, Yingqian Wang, Member, IEEE, Guanghui Wang, Senior Member, IEEE, and Fuqing Duan, Member, IEEE

Abstract—Light field (LF) depth estimation plays a crucial role in many LF-based applications. Existing LF depth estimation methods consider depth estimation as a regression problem, where a pixel-wise L1 loss is employed to supervise the training process. However, the disparity map is only a sub-space projection (i.e., an expectation) of the disparity distribution, which is essential for models to learn. In this paper, we propose a simple yet effective method to learn the sub-pixel disparity distribution by fully utilizing the power of deep networks, especially for LF of narrow baselines. We construct the cost volume at the sub-pixel level to produce a finer disparity distribution and design an uncertainty-aware focal loss to supervise the predicted disparity distribution toward the ground truth. Extensive experimental results demonstrate the effectiveness of our method. Our method significantly outperforms recent state-of-the-art LF depth algorithms on the HCI 4D LF Benchmark in terms of all four accuracy metrics (i.e., BadPix 0.01, BadPix 0.03, BadPix 0.07, and MSE ×100). The code and model of the proposed method are available at https://github.com/chaowentao/SubFocal.

Index Terms—Light field, depth estimation, disparity distribution, sub-pixel cost volume.

I. INTRODUCTION

Light Field (LF) [1], [2] depth estimation is a fundamental task in computer vision with many subsequent applications, such as refocusing [1], [3], super-resolution [4], [5], [6], [7], [8], [9], view synthesis [10], [11], [12], 3D reconstruction [13], and virtual reality [14]. The 4D LF records abundant spatial and angular information in the scene, from which the scene depth can be estimated. Based on the characteristics of light field images, traditional methods [15], [16], [17], [18] and learning-based methods [19], [20], [21], [22], [23], [24] have been proposed for light field image depth estimation.

Early traditional LF depth estimation algorithms can be divided into three categories based on the properties of LF images. Multi-view stereo matching (MVS) based methods [16] exploit the multi-view information of the LFs for depth estimation. Epipolar plane image (EPI) based methods [25] calculate the slopes using epipolar geometric properties of the LFs to predict the depth. Defocus-based methods [15] calculate the depth of a pixel by measuring the consistency at different focal stacks. However, conventional methods involve nonlinear optimization and manually designed features, which are computationally demanding and sensitive to occlusions, weak textures, and highlights.

Deep learning has been applied to LF depth estimation in recent years [19], [20], [21], [22], [23]. The mainstream deep learning-based methods typically involve four key steps: feature extraction, cost construction, cost aggregation, and depth (disparity) regression.¹ Cost construction provides initial similarity measures of possible matched pixels among different views. Existing techniques construct cost volumes by shifting the features of each view with specific integer values, and regress the final disparity values (in sub-pixel accuracy) by performing a weighted average on each candidate disparity. However, the regressed disparity is only a sub-space projection (more precisely, an expectation) of the predicted disparity distribution, which can be considered as a set of probability values that a pixel belongs to different candidate disparities. Although these methods have achieved remarkable performance by minimizing the pixel-wise loss between the predicted disparity and the ground truth, the disparity distribution issue has been ignored. Since there are unlimited disparity distributions that may produce the same expectation, simply minimizing the loss in the disparity domain can prevent the deep networks from learning the core information and yield sub-optimal depth estimation results.

In this paper, we aim to learn the accurate disparity distribution for LF depth estimation, which requires a sub-pixel cost volume to produce a finer disparity distribution. In the previous method [16], the sub-pixel cost volume is constructed at the image level by displacing sub-aperture images (SAI) based on the phase shift theorem, which is computationally intensive and

¹Depth $\gamma$ and disparity $d$ can be transformed directly according to $\gamma = \frac{f B}{d}$, where $B$ and $f$ stand for the baseline length and the focal length of the LF camera, respectively. The paper does not make a distinction between them.
time-consuming. In our method, we propose an interpolation-based sub-pixel cost volume construction approach at the feature level to achieve sub-pixel level sampling of disparity distribution. In addition, we design an uncertainty-aware focal loss (UAFL) to reduce the difference between the estimated and ground-truth disparity distributions. Since the ground-truth disparity distribution is unavailable in public LF datasets, we follow the common settings in [26] to model the ground-truth disparity distribution as a Dirac delta distribution (i.e., the possibility of the ground-truth disparity is 1 while the others are 0), and further discretize the Dirac delta distribution into two adjacent sampling points to facilitate the training of our method. We conduct extensive experiments to validate the effectiveness of our method.

In summary, the contributions of this paper are as follows:
• We reframe LF depth estimation as a pixel-wise distribution alignment problem and make the first attempt to learn the sub-pixel level LF disparity distribution using deep neural networks.
• We propose a simple yet effective method for sub-pixel disparity distribution learning, in which a sub-pixel level cost volume construction approach and an uncertainty-aware focal loss are developed to supervise the learning of disparity distribution.
• Extensive experimental evaluations demonstrate the effectiveness of our approach. As shown in Fig. 1 and Table II, our model outperforms the state-of-the-art LF depth estimation method [22] surpasses by 21.7% on BadPix 0.01, 28.5% on BadPix 0.03, 23.9% on BadPix 0.07, and 8.1% on MSE ×100, respectively.

The rest of this paper is organized as follows. Section II overviews the related work on LF depth map estimation. Section III provides the network components and loss functions of our method in detail. Section IV reports the experimental results on both synthetic and real LF datasets. Section V describes the failure cases and limitations of our method. Finally, the paper is concluded in Section VI.

II. RELATED WORK

A. Light Field Visualization

There is spatial and angular information in the LF image, which is often parametrically represented as \( L(u, v, x, y) \) by a two-plane [27]. As shown in Fig. 2(a), where \((u, v)\) is the first plane, representing the angular resolution, and \((x, y)\) is the second plane, representing the spatial resolution. Based on the characteristics of the LF image, the LF can be represented as EPI or SAI as shown in Fig. 2(b). We can obtain scene depth based on different representations of LF images.

B. Traditional Methods in LF Depth Estimation

Traditional depth estimation algorithms can be divided into the following three categories based on the properties of LF images:

Multi-view stereo matching (MVS) based methods: The LF image contains multi-view information from the scene, some works employ disparity cues by performing stereo matching from the LF image for depth estimation. Jeon et al. [16] used phase shift theory to represent the sub-pixel translation between sub-aperture images. The central sub-aperture image was matched with other sub-aperture images to perform stereo matching.

Epipolar plane image (EPI) based methods: Other methods calculate the slope of lines in EPI for depth estimation. Wanner et al. [28] proposed a structure tensor to estimate the slope of lines in horizontal and vertical EPIs, and refined the initial results by global optimization. Zhang et al. [25] proposed a Spinning Parallelogram Operator (SPO) to compute the slope of straight lines in EPI, which is insensitive to occlusion, noise, and spatial blending. Chen et al. [29] performed superpixel regularization over partially occluded regions based on the characterization of EPI. Zhang et al. [30] exploited the special linear structure of an EPI and locally linear embedding (LLE).

Defocus-based methods: The defocused-based method performs depth estimation by calculating the color consistency...
| TABLE I |
| QUANTITATIVE COMPARISON RESULTS WITH THE TOP-RANKED METHODS ON HCI 4D BENCHMARK, CONTAINING THE SPO [25], CAE [43], EPIFNET-CN-M [20], EPI ORM [36], LFATTNet [21], AttMLFNNet [22], FastLFNet [44], DistGmDist [45], OACC-Net [23] AND SubFOcal(Ours) |

| Backgmon | Dots | Pyramids | Stripes |
|----------|------|----------|---------|
| 0.07 | 0.03 | 0.01 | MSE | 0.07 | 0.03 | 0.01 | 0.003 | MSE |
| SPO [25] | 3.781 | 8.639 | 94.94 | 4.587 | 16.27 | 35.06 | 58.07 | 5.238 | 0.861 | 6.263 | 79.20 | 0.043 | 14.99 | 15.46 | 21.87 | 6.955 |
| CAE [43] | 3.924 | 4.313 | 17.32 | 6.074 | 12.40 | 42.50 | 83.70 | 5.082 | 1.681 | 7.162 | 27.54 | 0.048 | 7.872 | 16.90 | 39.95 | 3.556 |
| PS_RF [46] | 7.142 | 13.94 | 74.66 | 6.892 | 7.975 | 17.54 | 78.80 | 8.338 | 1.073 | 6.235 | 63.23 | 0.382 | 2.964 | 7.990 | 41.65 | 1.382 |
| OBER-cross-ANP [47] | 3.413 | 4.956 | 13.66 | 4.700 | 0.974 | 37.66 | 73.19 | 1.757 | 0.364 | 1.130 | 8.171 | 0.008 | 3.065 | 9.352 | 44.72 | 1.435 |
| OADC [31] | 1.321 | 5.117 | 49.05 | 3.832 | 60.91 | 35.80 | 12.33 | 6.084 | 0.781 | 0.498 | 2.350 | 0.010 | 2.709 | 6.858 | 19.28 | 0.932 |
| Epinet-fcn-in [20] | 5.301 | 5.563 | 39.94 | 3.705 | 2.490 | 9.117 | 55.61 | 1.475 | 0.159 | 0.874 | 11.42 | 0.007 | 2.457 | 2.711 | 11.77 | 0.932 |
| EPI ORM [36] | 5.984 | 5.373 | 34.34 | 3.411 | 2.588 | 27.48 | 55.71 | 1.448 | 0.354 | 1.301 | 19.06 | 0.016 | 6.871 | 13.94 | 59.14 | 1.744 |
| LFattNet [21] | 3.367 | 25.33 | 39.88 | 4.189 | 1.554 | 13.02 | 55.06 | 1.425 | 0.195 | 0.489 | 2.063 | 0.004 | 2.933 | 5.417 | 18.21 | 0.892 |
| AttMLFNNet [22] | 5.224 | 5.442 | 13.79 | 3.963 | 1.606 | 2.021 | 55.61 | 1.061 | 0.174 | 0.429 | 1.767 | 0.003 | 3.923 | 7.432 | 55.44 | 1.814 |
| FastLFNet [44] | 5.138 | 11.41 | 39.84 | 3.986 | 21.17 | 41.11 | 68.15 | 3.407 | 0.620 | 2.193 | 22.19 | 0.018 | 9.442 | 32.60 | 63.04 | 0.984 |
| DistGmDist [45] | 5.824 | 10.24 | 26.17 | 4.712 | 1.826 | 46.42 | 35.71 | 1.367 | 0.108 | 0.539 | 4.939 | 0.004 | 3.913 | 6.885 | 19.25 | 0.917 |
| OACC-Net [23] | 3.931 | 6.640 | 21.61 | 3.938 | 1.510 | 3.040 | 55.01 | 1.418 | 0.157 | 0.536 | 3.852 | 0.004 | 2.920 | 6.448 | 15.24 | 0.845 |
| SubFOcal(Ours) | 3.194 | 4.281 | 12.47 | 3.667 | 0.899 | 1.524 | 55.15 | 1.310 | 0.220 | 0.411 | 1.867 | 0.005 | 2.464 | 5.368 | 9.386 | 0.821 |
| SubFOcal(L-Ours) | 3.079 | 3.651 | 7.821 | 3.868 | 0.935 | 1.153 | 8.355 | 1.279 | 0.253 | 0.543 | 2.017 | 0.005 | 2.122 | 2.219 | 3.992 | 0.874 |

| TABLE II |
| AVERAGE BADPIX 0.07, BADPIX 0.03, BADPIX 0.01 AND MSE > 100 ON THE HCI 4D LF BENCHMARK |

| Average | SPO [25] | CAE [43] | PS_RF [46] | OBER-cross-ANP [47] | OADC [31] | Epinet-fcn-in [20] | EPI ORM [36] |
|----------|----------|----------|------------|-------------------------|----------|-------------------|---------|
| BP 0.07  | 8.466    | 8.211    | 8.578      | 4.594                  | 12.45    | 4.646             | 8.642   |
| BP 0.03  | 22.30    | 22.95    | 21.63      | 13.10                   | 27.13    | 9.537             | 17.36   |
| BP 0.01  | 66.70    | 55.84    | 74.03      | 35.23                   | 62.14    | 31.90             | 49.84   |
| MSE > 100 | 3.986    | 3.730    | 4.617      | 2.584                   | 3.968    | 2.418             | 2.944   |

| Average | LFattNet [21] | AttMLFNNet [22] | FastLFNet [44] | DistGmDist [45] | OACC-Net [23] | SubFOcal(Ours) | SubFOcal(L-Ours) |
|----------|----------------|-----------------|----------------|----------------|----------------|----------------|------------------|
| BP 0.07  | 3.756          | 3.596           | 9.071          | 4.521          | 3.734          | 2.956          | 2.735            |
| BP 0.03  | 6.823          | 6.668           | 22.32          | 8.699          | 6.502          | 5.602          | 6.077            |
| BP 0.01  | 17.23          | 16.42           | 56.64          | 23.32          | 15.05          | 12.85          | 15.05             |
| MSE > 100 | 1.904          | 1.720           | 2.456          | 1.896          | 1.618          | 1.581          | 1.581             |

The best results are in bold faces and the second-best results are underlined.
of the views at different focal lengths. Tao et al. [15] fused scattering and matching cues to obtain a local depth map using Markov random field for global optimization. William et al. [17] employed the information entropy between different angles and adaptive scattering to improve robustness to occlusion and noise. An occlusion-aware vote cost (OAVC) method was presented by Han et al. [31] to preserve depth map edges. However, these methods rely on manually designed features and subsequent optimization, which are time-consuming with limited accuracy. In comparison, the proposed approach can fully utilize the power of deep networks to achieve higher accuracy and faster speed by learning the sub-pixel disparity distribution.

C. Deep Learning-Based Methods in LF Depth Estimation

In recent years, deep learning has been gradually used for various tasks [32], [33], [34], e.g., LF depth estimation [19], [20], [21], [22], [23]. Some learning-based methods estimate the slopes of the lines from the EPI. Luo et al. [35] used pairs of EPI image patches (horizontal and vertical) to train a CNN network. Li et al. [36] designed a network to learn the relationship between the slopes of straight lines in horizontal and vertical EPIs. Leistner et al. [37] introduced the idea of EPI-Shift to increase the receptive field of the network by virtually shifting the LF stack. These EPI-based methods only consider the features in the horizontal and vertical directions of angular views and require complex subsequent refinement.

Other methods design networks by directly exploring the correspondence among views in LF images. Shin et al. [20] introduced a multi-stream input structure to concatenate views from four directions for depth estimation. Guo et al. [38] proposed an occlusion-aware network that learns occlusion maps from features from multiple views, capable of accurately estimating the depth of sharp edges. Tsai et al. [21] proposed an attention-based viewpoint selection network that generates attention maps to represent the importance of each viewpoint. Chen et al. [22] proposed an attention-based multi-level fusion network, using an intra-branch and inter-branch fusion strategy to perform a hierarchical fusion of effective features from different perspectives. Wang et al. [23] constructed an occlusion-aware cost volume based on dilated convolution [39], which achieved a better trade-off between accuracy and speed.

The state-of-art methods [21], [22], [23] treat depth estimation as a regression problem supervised by a pixel-wise L1 loss and lack explicit supervision of disparity distribution. Different from previous methods, we directly learn sub-pixel disparity distribution for LF depth estimation, which can generate a more accurate depth map. To the best of our knowledge, we are the first to apply this methodology to LF depth estimation.

D. Loss Functions in Depth Estimation

The loss functions commonly used in supervised depth estimation methods include L1 Loss (MAE Loss), L2 Loss (MSE Loss), and Smooth L1 Loss. L2 Loss calculates the mean squared error between the predicted depth map and the actual depth label. The L1 Loss calculates the mean absolute error and has better robustness than L2 Loss to outliers. Smooth L1 Loss is a smoothed version of L1 Loss, which exploits the advantages of L1 and L2. Recently, Li et al. [40] proposed the BinsFormer for the classification-regression-based monocular depth estimation, which can generate proper adaptive bins and ensure sufficient interaction between probability distribution and bin predictions. However, BinFormer’s loss functions only consist of classification loss and regression loss, without considering depth distribution differences. Peng et al. [41] proposed a unified depth representation and a unified focal loss, which can be more uniform and reasonable to alleviate sample imbalance. However, a unified focal loss is based on a cross-entropy classification loss and lacks explicit supervision of disparity distribution. Our UAFL is inspired by Focal Loss [26], [42]. Different from previous studies, we introduced UAFL to address sample imbalance and adopted the dynamic weight adjustment strategy to make the model focus more on the hard points with significant differences in disparity distributions.

III. METHOD

The architecture of our proposed approach is shown in Fig. 3. First, the features of each SAI are extracted using a shared feature extractor. Second, the sub-pixel view shift is performed to construct the sub-pixel cost volume. Third, the cost aggregation module is used to aggregate the cost volume information. The predicted disparity distribution and disparity map are then produced by attaching a disparity regression module, where the Jensen-Shannon (JS) divergence is employed to calculate the uncertainty map. Finally, the proposed uncertainty-aware focal loss is used to simultaneously supervise the disparity distribution and the disparity map. We will describe each module in detail below.

A. Network

1) Feature Extraction: First, two 3 × 3 convolutions (i.e., Conv2D_1 and Conv2D_2) are employed to extract the initial feature with a channel of 4. In order to extract multi-scale features of SAI, the widely used Spatial Pyramid Pooling (SPP) module [21], [22] is adopted as the feature extraction module in our network. Four average pooling operations at different scales are used to compress the features. The sizes of the average pooling blocks are 2 × 2, 4 × 4, 8 × 8, and 16 × 16, respectively. A 1 × 1 convolution layer is used for reducing the feature dimension for each scale. Bilinear interpolation is adopted to upsample these low-dimensional feature maps to the same size. The features of the SPP module contain hierarchical context information and are concatenated to form the feature map F. The hierarchical feature extraction strategy incorporates additional useful features from neighboring regions for challenging scenarios, such as texture-less and reflection areas.

2) Sub-Pixel Cost Volume: The LF can be represented using the two-plane model, i.e., \( L(u, v, x, y) \), where \((x, y)\) stands for the spatial coordinate and \((u, v)\) for the angular coordinate. The LF disparity structure can be formulated as:

\[
L(u_c, v_c, x, y) = L(u, v, x + (u - u_c) \cdot d(x, y),
\]

\[
y + (v - v_c) \cdot d(x, y),
\]

(1)
where \( d(x, y) \) denotes the disparity between the center view and its adjacent view at \((x, y)\).

Cost volume is constructed by shifting the feature according to (1) within a predefined disparity range (such as from -4 to 4). It can be transformed into the disparity distribution by subsequent cost aggregation and disparity regression modules. To obtain a finer disparity distribution, a sub-pixel level disparity sampling interval is required. Different from the previous method [16] using phase shift theorem to construct image level cost volume, we construct a sub-pixel feature level cost volume based on bilinear interpolation, which can save memory-consuming, as shown in Fig 4. Suppose the shape of the feature map \( F \) is \( H \times W \times C \). We want to shift the point \( P_{11} \) to point \( S \), the shift distance is sub-pixel. The feature \( F_S \) of the point \( S \) can be calculated by bilinear interpolation:

\[
F_S = \frac{(x_2 - x)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} F_{P_{11}} + \frac{(x - x_1)(y_2 - y)}{(x_2 - x_1)(y_2 - y_1)} F_{P_{21}} + \frac{(x_2 - x)(y_2 - y_1)}{(x_2 - x_1)(y_2 - y_1)} F_{P_{12}} + \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)(y_2 - y_1)} F_{P_{22}}
\]

where the coordinates of \( S \) are \((x, y)\). \( P_{11}, P_{12}, P_{21} \) and \( P_{22} \) are the four adjacent points, whose coordinates are \((x_1, y_1), (x_1, y_2), (x_2, y_1), \) and \((x_2, y_2)\), respectively.

Experimental results in Section IV-C1 demonstrate that bilinear interpolation can achieve improved accuracy with higher efficiency. In our default setting, we adopt 17 disparity levels ranging from -4 to 4, where the sub-pixel interval is 0.5. After shifting the feature maps, we concatenate these feature maps into a 4D cost volume \( D \times H \times W \times C \).

It is worth noting that a smaller sampling interval can generate a finer disparity distribution but will increase computation time and slow down inference. Therefore, we conduct ablation experiments to investigate the trade-off between accuracy and speed with respect to the sampling interval, which is described in detail in Section IV-C2.

3) Cost Aggregation and Disparity Regression: The shape of the sub-pixel cost volume is \( D \times H \times W \times C \), and we employ 3D CNN to aggregate the sub-pixel cost volume. Following [21], our cost aggregation consists of eight \( 3 \times 3 \times 3 \) convolutional layers and two residual blocks. After passing through these 3D convolutional layers, we obtain the final cost volume \( C_f \in D \times H \times W \). We normalize \( C_f \) by using the softmax operation along dimension \( D \) to produce the probability of the disparity distribution \( \hat{Y}_{dist} = \text{softmax}(-C_f) \). Finally, the output disparity \( \hat{d} \) can be calculated according to

\[
\hat{d} = \sum_{d_k = D_{\text{min}}}^{D_{\text{max}}} d_k \times \text{softmax}(-C_{d_k}), \tag{3}
\]

where \( \hat{d} \) denotes the estimated center view disparity, \( D_{\text{min}} \) and \( D_{\text{max}} \) stand for the predefined minimum and maximum disparity values, respectively, \( d_k \) is the disparity value between \( D_{\text{min}} \) and \( D_{\text{max}} \) according to the predefined disparity interval, and \( C_{d_k} \) is the cost for the disparity value \( d_k \).
To address this issue, we design an uncertainty-aware focal loss to supervise the disparity distribution. Similar to Focal Loss [26], [42], we adopt the dynamic weight adjustment strategy to make the model focus more on the hard points with significant differences in disparity distributions. Fig. 6 compares the disparity distribution with and without using UAFL. By using UAFL, our model can predict a reasonable disparity distribution and accurate disparity in challenge regions. The process of UAFL is detailed below.

In order to quantitatively evaluate the difference between the predicted and the ground-truth disparity distributions as presented in (4).

\[ U = JS(\hat{d}_{\text{dist}}|\hat{Y}_{\text{dist}}), \]  

where \( U \) is the uncertainty map, \( \hat{Y}_{\text{dist}} \) represents the predicted disparity distribution, and \( Y_{\text{dist}} \) is the ground-truth disparity distribution. We follow [26] to consider the ground-truth disparity distribution as a Dirac delta distribution and discretize it to the two adjacent sampling points \( d_l \) and \( d_r \), according to:

\[ Y_{\text{dist}} = p_d d_l + p_{d_r} d_r, \]  

where \( d_l \) and \( d_r \) are the left and right adjacent points respectively, and \( d \) denotes the ground-truth disparity. \( p_d = \frac{d - d_l}{d_r - d_l} \) and \( p_{d_r} = \frac{d - d_l}{d_r - d_l} \) represent their corresponding probability values. We employ discretization for two reasons. Firstly, it approximates the actual disparity distribution (Dirac delta) more closely. Secondly, it ensures that the expected value of the distribution remains constant before and after discretization.

Finally, we design an uncertainty-aware focal loss to supervise both disparity distribution and disparity map:

\[ U \text{AFL} = \left\| U \right\|^\beta \odot \left\| d - \hat{d} \right\|_1, \]  

where \( U \) is the uncertainty map, \( \hat{d} \) is the predicted disparity, and \( d \) is the ground-truth disparity. \( \odot \) stands for the element-wise multiplication operation, and \( \beta \) is the coefficient factor that controls the ratio of dynamic weight assignment. When \( \beta = 0 \), UAFL degenerates to the standard L1 loss. In ablation experiments of Section IV-C4, we analyze the effect of different \( \beta \) values.

IV. EXPERIMENTS

In this section, we first describe the dataset, evaluation metrics, and implementation details, and then compare our method with state-of-the-art methods. Finally, we present ablation experiments to analyze the effectiveness of the proposed method.

A. Datasets, Evaluation Metrics, and Implementation Details

The 4D LF Benchmark (HCI 4D) [49] is the current mainstream benchmark dataset for the evaluation of LF disparity estimation methods. This dataset is rendered by Blender and provides 28 densely arranged 4D LFs, including 16 scenes in the Additional category, 4 scenes in the Training category, 4 scenes in the Test category, and 4 scenes in the Stratified category. Scenes in the Training, Additional, and Stratified have ground-truth disparity maps while scenes in the Test category

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
do not provide ground-truth disparity annotations. The spatial resolution of each scene is 512 × 512, and the angular resolution is 9 × 9. UrbanLF-Syn dataset [48] contains 230 synthetic LF samples, with 170 training, 30 validation, and 30 test samples for LFNAT LF Depth Estimation Challenge at the CVPR 2023 Workshop. Each sample consists of 81 SAIIs with a spatial resolution of 480 × 640 and an angular resolution of 9 × 9.

For the quantitative evaluation, we adopt the standard metrics, including MSE × 100 and BadPix(ε). Specifically, MSE × 100 represents the mean square errors of all pixels at a given mask multiplied by 100. BadPix(ε) denotes the percentage of pixels whose absolute value between the actual label at a given mask and the algorithm’s predicted result surpasses a threshold ε, which is usually chosen as 0.01, 0.03, and 0.07.

We follow the previous methods [20, 21, 22, 23] and use 16 scenes from Additional as the training set, 8 scenes from Stratified and Training as the validation set, and 4 scenes from Test as the test set. We employ LFattNet [21] as the baseline model, change disparity interval, and loss functions. During training, LF images were randomly cropped into 32 × 32 grayscale patches. The same data augmentation strategy [20, 21] was used to improve the model performance. We set the batchsize to 32 and used the Adam [50] optimizer. The disparity range was set to [-4, 4], and the sampling interval was set to 0.5 by default (i.e., SubFocal). To speed up the model training, a two-step strategy was adopted. In the first stage, the model was trained for 50 epochs using L1 Loss, and the learning rate was set to 1 × 10⁻³. In the second stage, the L1 loss was replaced by our UAIL. The model was finetuned for 10 epochs, and the learning rate was set to 1 × 10⁻⁴. Our network is implemented using TensorFlow [51]. The training process took roughly 65 hours on an NVIDIA RTX 3090 GPU. We also introduced an enhanced model called SubFocal-L by setting the sampling interval to 0.1. SubFocal-L was trained for 165 hours on an NVIDIA A40 GPU, which was approximately equivalent to 330 hours on an NVIDIA RTX 3090 GPU. As for the Urban-Syn dataset, we follow the official splitting with 170 scenes for training and randomly select 8 scenes for testing. The disparity range was set to [-0.5, 1.6], and the sampling interval was set to 0.1. It took about 158 hours on an NVIDIA RTX 3090 GPU to train the model.

B. Comparison With the State-of-the-Art Methods

1) Quantitative Comparison: We performed quantitative comparison with the state-of-the-art methods, among which PS_RF [46] is the MVS-based method, SPO [25] and EPI_ORM [36] are the EPI-based methods, CAE [43], OBER-cross-ANP [47] and OAVC [31] are the Defocus-based methods, Epinet-fcn-m [20], LFattNet [21], AttMLFNet [22], FastLFNet [44], DistgDisp [45] and OACC-Net [23] are the DL-based methods. As demonstrated in Fig. 7 our method ranks first place on all the four commonly used metrics (BadPix 0.01, BadPix 0.03, BadPix 0.07, and MSE × 100) on the HCI 4D LF benchmark. Table I shows the quantitative comparison results on the validation and test sets. Our method achieves the best or second-best performance in most scenes, especially on the unlabeled test sets, and also achieves an overall best performance, as shown in Table II. SubFocal surpasses the state-of-the-art method (i.e., AttMLFNet [22]) by 8.3% (from 16.42 to 15.06) on BadPix 0.01, 14.7% (from 6.568 to 5.602) on BadPix 0.03, 17.7% (from 3.596 to 2.956) on BadPix 0.07, and 5.9% (from 1.720 to 1.618) on MSE × 100, respectively. Furthermore, SubFocal-L (the enhanced version) achieves a 21.7% (from 16.42 to 12.85) gain on BadPix 0.01, 28.5% (from 6.568 to 4.697) gain on BadPix 0.03, 23.9% (from 3.596 to 2.735) gain on BadPix 0.07, and 8.1% (from 1.720 to 1.581) gain on MSE × 100, respectively. We also have compared our method with three state-of-the-art methods on the UrbanLF-Syn dataset [48], including LFattNet [21], DistgDisp [45], OACC-Net [23]. For fair comparisons, we retrained these methods based on released codes and adjusted disparity ranges, respectively. The ranges of LFattNet and DistgDisp were set to [-1, 2], and OACC-Net was [-2, 2]. The disparity interval was set to 1. Table III shows the comparison results on the UrbanLF-Syn dataset. Our method ranks first in most scenes and achieves the top average result in terms of BadPix 0.07, outperforming current state-of-the-art methods.

2) Qualitative Comparison: Fig. 8 shows the qualitative results achieved by different methods on four scenes of the HCI 4D benchmark: Backgammon, Dots, Stripes, and Boxes. Scenes Backgammon and Stripes are stratified and contain multiple lines. Scene Dots is composed of regular dots with clear edges. Scene Boxes consists of multiple objects and contains severe occlusions. The BadPix 0.07 error maps show that our approach has substantially less error as compared with previous methods, especially in the occlusion and edge regions. Fig. 9 shows qualitative comparison results on Img27, Img54, Img69, and Img70 scenes of Urban-Syn dataset. From this example, our method can handle the edge regions better than other methods. We attribute
Fig. 8. Visual comparisons of disparity and error maps on validation scenes Backgammon, Dots, Boxes, and Stripes [52]. Corresponding quantitative scores (BadPix 0.07, BadPix 0.03, and MSE × 100) are reported on the top-left corner of each error map. Red pixels on the error map represent areas where the error is higher than the threshold, i.e., 0.07. It is clear that our method has less error as compared to previous methods, especially in the occlusion and edge regions.
TABLE III
QUANTITATIVE COMPARISON RESULTS WITH STATE-OF-THE-ART METHODS ON THE SCENES OF URBANLF-SYN DATASET [48] IN TERMS OF BADPIX 0.07 AND RUNTIME

| Method     | Params. | Img11 | Img27 | Img34 | Img50 | Img54 | Img68 | Img69 | Img70 | BP 0.07 | Runtime |
|------------|---------|-------|-------|-------|-------|-------|-------|-------|-------|----------|---------|
| LFattNet   | 5.06M   | 10.179| 23.648| 22.637| 9.668 | 13.533| 9.819 | 12.075| 7.467 | 13.629   | 0.743s  |
| DistgDisp  | 4.89M   | 6.825 | 20.240| 24.522| 8.304 | 13.830| 9.367 | 10.118| 7.691 | 12.612   | 0.444s  |
| OACC-Net   | 5.02M   | 6.845 | 24.058| 27.792| 10.390| 13.922| 11.799| 13.234| 8.750 | 14.599   | 0.439s  |
| Ours       | 5.06M   | 9.194 | 22.223| 22.619| 6.008 | 11.476| 8.385 | 9.357 | 6.542 | 11.976   | 1.839s  |

Runtime is the average result of running each model 50 times with an input LF image of 9×9×256×256. The best results are shown in boldface.

Fig. 9. Visual comparisons between our method and state-of-the-art methods on the scenes of UrbanLF-Syn dataset [48], i.e., Img27, Img54, Img69 and Img70, including LFattNet [21], DistgDisp [45], OACC-Net [23], with the corresponding BadPix 0.07 error maps. Lower is better. The best results are shown in boldface. Please zoom in for a better comparison.

our improved performance to two factors. Firstly, our sub-pixel cost volume produces a more refined disparity distribution compared to previous methods. Secondly, our uncertainty-aware focal loss assigns higher weights to difficult points in the edge regions by computing the disparity distribution difference, leading to even better results. Therefore, our approach is able to produce a more reasonable disparity distribution and accurate disparity map.

3) Generalization Comparison: In order to validate the generalization ability of our method, we compare it with the mainstream methods in various scenes. We employ the model trained on the HCI dataset directly for inference. First, we compare the visual results achieved by SPO [25], EPINET [20], and our method on the old HCI LF dataset [53]. As shown in Fig. 10, our results exhibit sharp edges and less noise in comparison with other methods, such as the scene Papilon and monasRoom. We also evaluate our method on the real-world LF images taken with a Lytro Illum camera [54] and a moving camera mounted on a gantry [53]. Fig. 11 illustrates that our method can produce more accurate disparity maps than SPO [25] and EPINet [20], such as the boom of the scene Truck and the background of the scene Dinosaur. It clearly demonstrates the superior generalization capability of our approach.

C. Ablation Study

We conduct extensive ablation experiments to analyze the effectiveness of our method. Our ablation studies include the interpolation approaches, disparity intervals, combinations of loss functions, hyper-parameters of UAFL, and visualization of uncertainty maps.

1) Interpolation Methods: We compare three representative interpolation methods: phase [16], nearest neighbor, and bilinear. As shown in the upper part of Table IV, the method of phase shift requires more training and testing time, while the result is inferior to the nearest neighbor and bilinear approaches. The bilinear approach improves the results over the nearest approach (0.02 in terms of MSE) while taking a little more time in training and testing. Therefore, we finally choose the bilinear approach to construct a sub-pixel cost volume.
2) Disparity Interval: The selection of the disparity interval determines the sampling density and further affects the inference speed of the model. We need to choose the proper disparity interval to achieve a good trade-off in accuracy and speed. We set the disparity intervals to 1, 0.5, 0.25, and 0.1, respectively, and then train our model for 50 epochs on the HCI 4D LF benchmark and then evaluate their performance on the validation set. The lower part of Table IV shows the results of the model with different disparity intervals. As the disparity interval decreases and the number of samples increases, the accuracy of the model is improved, while the training time increases as well. Specifically, reducing the disparity interval from 1 to 0.5 improves the BadPix 0.07 and MSE values.

3) Combinations of Loss Functions: Selecting appropriate loss functions is a crucial aspect of our method. We test some combinations of common losses, including L1 loss, MSE loss, KL divergence loss, and JS divergence loss. We use the model with a disparity interval of 0.5 as the pre-trained baseline, and its network structure adopts LFattNet [21]. Then the model is finetuned for 10 epochs with the learning rate being set to $1 \times 10^{-4}$. Table V shows the results of different loss functions. Our UAFL achieves the lowest BadPix 0.07 metrics.

4) Hyper-Parameter of UAFL: The function of UAFL is to dynamically assign weights based on an uncertainty map to supervise the training of disparity distribution and disparity simultaneously. We introduce the hyper-parameter $\beta$ to adjust the ratio of the assigned weights. Based on the disparity interval of 0.5, we choose different $\beta$ to investigate its influence, and the results are shown in Fig. 13(a). We observe that as $\beta$ decreases...
from 1 to 0.1, the BadPix 0.07 metric is improved. Fig. 13(b) also shows that the BadPix 0.07 error map with $\beta = 0.1$ is significantly better than that with $\beta = 1$. Finally, we set $\beta = 0.1$ in our model.

5) **Visualization of Uncertainty Map**: Fig. 14 shows the per-pixel qualitative result on the Dino scene, which contains several hard regions, e.g., heavy occlusions and sharp edges. We can observe from the uncertainty map that our method assigns higher weights to the hard regions and lower weights to easy regions. Therefore, our approach can improve the overall accuracy of disparity estimation.

V. LIMITATIONS AND DISCUSSION

Although our method performs well in most scenes, we do not explicitly consider occlusions and texture-less regions. As shown in Fig. 15, relatively large errors occur in occlusion regions (see Dino scene error map), and artifacts may appear in texture-less regions (see the black hole in Knight scene). We also verified the effect of the sub-pixel cost volume. As shown in Tables IV and III, a small sampling interval results in increased computational cost and longer inference time, making it unfeasible for practical applications. Therefore, it is wise to make a trade-off between accuracy and speed based on specific requirements. In future work, we plan to enhance the robustness of our method to handle challenging scenarios such as non-Lambertian surfaces, weak textures, and highlight regions. Additionally, inspired by OACC-Net [23], we plan to explore the construction of cost volume based on dilated convolution to reduce inference time.

VI. CONCLUSION

In this paper, we have presented a straightforward and efficient approach to learning the disparity distribution for LF depth estimation. On one hand, we have constructed an interpolation-based cost volume at the sub-pixel level for a finer disparity distribution. On the other hand, we have designed an uncertainty-aware focal loss based on JS divergence to supervise the disparity distribution. Extensive experiments validate the effectiveness of our method. Our method outperforms state-of-the-art methods, as demonstrated by our experimental results. Our method holds great potential for advancing the field of LF-based depth estimation research.

REFERENCES

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photoACM trans. graphraphy with a hand-held plenoptic camera,” Comput. Sci. Tech. Rep., vol. 2, no. 11, pp. 1–11, Apr. 2005.
[2] J. Peng, Z. Xiong, Y. Wang, Y. Zhang, and D. Liu, “Zero-shot depth estimation from light field using a convolutional neural network,” IEEE Trans. Comput. Imag., vol. 6, pp. 682–696, 2020.
[3] Y. Wang, J. Yang, Y. Guo, C. Xiao, and W. An, “Selective light field refocusing for camera arrays using bokeh rendering and superresolution,” IEEE Signal Process. Lett., vol. 26, no. 1, pp. 204–208, Jan. 2019.
[4] S. Zhang, Y. Lin, and H. Sheng, “Residual networks for light field image super-resolution,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 11046–11055.
[5] J. Jin, J. Hou, J. Chen, and S. Kwong, “Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2260–2269.
[6] Z. Cheng, Z. Xiong, C. Chen, D. Liu, and Z.-J. Zha, “Light field super-resolution with zero-shot learning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10010–10019.
[7] Y. Chen, S. Zhang, S. Chang, and Y. Lin, “Light field reconstruction using efficient pseudo 4D epipolar-aware structure,” IEEE Trans. Comput. Imag., vol. 8, pp. 397–410, 2022.
[8] Z. Cheng, Y. Liu, and Z. Xiong, “Spatial-angular versatile convolution for light field reconstruction,” IEEE Trans. Comput. Imag., vol. 8, pp. 1131–1144, 2022.
[9] V. Van Duong, T. N. Huu, J. Yim, and B. Jeon, “Light field image super-resolution network via joint spatial-angular and epipolar information,” IEEE Trans. Comput. Imag., vol. 9, pp. 350–366, 2023.
[10] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, “Light field reconstruction using convolutional network on EPI and extended applications,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1661–1694, Jul. 2019.
[11] N. Meng, H.-K.-H. So, X. Sun, and E. Y. Lam, “High-dimensional dense residual convolutional neural network for light field reconstruction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 3, pp. 873–886, Mar. 2021.
[12] J. Jin, J. Hou, J. Chen, H. Zeng, S. Kwong, and J. Yu, “Deep coarse-to-fine light field reconstruction with flexible sampling and geometry-aware fusion,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 1819–1836, Apr. 2022.
[13] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. H. Gross, “Scene reconstruction from high spatio-angular resolution light fields,” ACM Trans. Graph., vol. 32, no. 4, pp. 73–1, Jul. 2013.
[14] J. Yu, “A light-field journey to virtual reality,” IEEE MultiMedia, vol. 24, no. 2, pp. 104–112, Apr.-Jun. 2017.
[15] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, “Depth from combining defocus and correspondence using light-field cameras,” in Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 673–680.
[16] H.-G. Jeon et al., “Accurate depth map estimation from a lenslet light field camera,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1547–1555.
[17] W.威廉 and J. K. Park, “Robust light field depth estimation for noisy scene with occlusion,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4396–4404.
[18] H. Zhu, Q. Wang, and J. Yu, “Occulsion-model guided antiocclusion depth estimation in light field,” IEEE J. Sel. Top. Signal Process., vol. 11, no. 7, pp. 965–978, Oct.-Dec. 2017.
[19] S. Heber and T. Pock, “Convolutional networks for shape from light field,” in Proc. IEEE Int. Conf. Comput. Vis., 2016, pp. 3746–3754.
[20] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, “Einet: A fully-convolutional neural network using epipolar geometry for depth from light field images,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4748–4757.
[21] Y.-J. Tsai, Y.-L. Liu, M. Ouyoung, and Y.-Y. Chiang, “Attention-based view selection networks for light-field disparity estimation,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 12095–12103.
[22] J. Chen, S. Zhang, and Y. Lin, “Attention-based multi-level fusion network for light field depth estimation,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 10099–10107.
[23] W. Wang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, “Occlusion-aware cost constructor for light field depth estimation,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2022, pp. 19809–19818.
[24] H. Sheng et al., “LFNAT 2023 challenge on light field depth estimation: Methods and results,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops, 2023, pp. 3472–3484.
[25] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, “Robust depth estimation for light field via spinnaling parallelogram operator,” Comput. Vis. Image Underst., vol. 145, pp. 148–159, Apr. 2016.
[26] X. Li et al., “Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection,” in Proc. Adv. Neural Inform. Process. Syst., 2020, pp. 21082–21092.
[27] M. Levoy and P. Hanrahan, “Light field rendering,” in SIGGRAPH, 1996, pp. 31–42.
[28] S. Wanner and B. Goldluecke, “Variational light field analysis for disparity estimation and super-resolution,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 3, pp. 606–619, Mar. 2013.
[29] J. Chen, J. Hou, Y. Ni, and L.-P. Chau, “Accurate light field depth estimation with superpixel regularization over partially occluded regions,” IEEE Trans. Image Process., vol. 27, no. 10, pp. 4889–4900, Oct. 2018.
[30] Y. Zhang et al., “Light-field depth estimation via epipolar plane image analysis and locally linear embedding,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 4, pp. 739–747, Apr. 2017.
[31] K. Han, W. Xiang, E. Wang, and T. Huang, “A novel occlusion-aware vote cost for light field depth estimation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 11, pp. 8022–8035, Nov. 2022.
[32] L. He, G. Wang, and Z. Hu, “Learning depth from single images with deep neural network embedding focal length,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4676–4689, Sep. 2018.
[33] L. He, M. Yu, and G. Wang, “Spindle-Net: CNNs for monocular depth inference with dilated kernel method,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2504–2509.
[34] K. Li, Z. Zhang, C. Zhong, and G. Wang, “Robust structured declarative classifiers for 3D point clouds: Defending adversarial attacks with implicit gradients,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 15294–15304.
[35] Y. Xian, W. Zhou, J. Fang, L. Liang, H. Zhang, and G. Dai, “EPF-based convolutional neural network for depth estimation on 4D light field,” in Proc. 24th Int. Conf. Neural Inf. Process., 2017, pp. 642–652.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.

Wentao Chao received the B.E. degree from the College of Automation Engineering, Shanghai University of Electric Power, Shanghai, China, in 2017, and the M.S. degree from the Department of Automation, North China Electric Power University, Baoding, China, in 2020. He is currently working toward the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. His research interests include generative models, image processing, light field imaging, and 3D reconstruction.
Xuechun Wang received the B.S. degree in information and computing science from Yunnan University, Kunming, China, in 2017, and the M.S. degree in computing mathematics from Yunnan University, Kunming, China, in 2020. She is currently working toward the Ph.D. degree with the School of Artificial Intelligence, Beijing Normal University, Beijing, China. Her research interests include camera calibration, light field imaging, image processing, and 3D reconstruction.

Yingqian Wang (Member, IEEE) received the B.E. degree in electrical engineering from Shandong University, Jinan, China, in 2016, the master’s and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018 and 2023, respectively. He is currently an Assistant Professor with the College of Electronic Science and Technology, NUDT. His research interests include optical imaging and detection, particularly on light field imaging, image super-resolution and infrared small target detection.

Guanghui Wang (Senior Member, IEEE) received the Ph.D. degree in engineering from the University of Waterloo, Waterloo, ON, Canada, in 2014. He is currently an Associate Professor with the Department of Computer Science, Toronto Metropolitan University, Toronto, ON. From 2014 to 2020, he was an Assistant Professor with the University of Kansas, Lawrence, KS, USA, and then an Associate Professor. He has authored one book Guide to Three Dimensional Structure and Motion Factorization, Springer-Verlag. He has authored or coauthored more than 180 papers in peer reviewed journals and conferences. His research interests include computer vision, image analysis, machine learning, and intelligent systems.

Yingqian Wang (Member, IEEE) received the B.E. degree in electrical engineering from Shandong University, Jinan, China, in 2016, the master’s and Ph.D. degrees in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018 and 2023, respectively. He is currently an Assistant Professor with the College of Electronic Science and Technology, NUDT. His research interests include optical imaging and detection, particularly on light field imaging, image super-resolution and infrared small target detection.

Fuqing Duan (Member, IEEE) received the B.S. and M.S. degrees in mathematics from Northwest University, Xi’an, China, in 1995 and 1998, respectively, and the Ph.D. degree in pattern recognition from the National Laboratory of Pattern Recognition, Beijing, China, in 2006. He is currently a Professor with the School of Artificial Intelligence, Beijing Normal University, Beijing. He has authored more than 80 conference and journal articles on related topics. His research interests include 3D reconstruction, skull identification, and machine learning and applications.