OPAL: Occlusion Pattern Aware Loss for Unsupervised Light Field Disparity Estimation

Peng Li, Jiayin Zhao, Jingyao Wu, Chao Deng, Yuqi Han, Haoqian Wang, Member, IEEE, and Tao Yu

Abstract—Light field disparity estimation is an essential task in computer vision. Currently, supervised learning-based methods have achieved better performance than both unsupervised and optimization-based methods. However, the generalization capacity of supervised methods on real-world data, where no ground truth is available for training, remains limited. In this paper, we argue that unsupervised methods can achieve not only much stronger generalization capacity on real-world data but also more accurate disparity estimation results on synthetic datasets. To fulfill this goal, we present the Occlusion Pattern Aware Loss, named OPAL, which successfully extracts and encodes general occlusion patterns inherent in the light field for calculating the disparity loss. OPAL enables i) accurate and robust disparity estimation by teaching the network how to handle occlusions effectively and ii) significantly reduced network parameters required for accurate and efficient estimation. We further propose an EPI transformer and a gradient-based refinement module for achieving more accurate and pixel-aligned disparity estimation results. Extensive experiments demonstrate our method not only significantly improves the accuracy compared with SOTA unsupervised methods, but also possesses stronger generalization capacity on real-world data compared with SOTA supervised methods. Last but not least, the network training and inference efficiency are much higher than existing learning-based methods. Our code will be made publicly available.

Index Terms—Light field disparity estimation, unsupervised disparity estimation, occlusion pattern aware loss.

I. INTRODUCTION

LIGHT field captures the light rays of a scene from multiple directions, thus simultaneously reserving the intensity and angular information of these light rays. Light field cameras [1], [2], [3] are compact sensors designed for single-shot and structured 4D light field capture. Such sensors are especially suitable for accurate depth sensing and volumetric imaging of complex scenes in a convenient and efficient manner which remains challenging for single camera setups (e.g., single image [4] or monocular video [5], [6]). Moreover, in contrast to stereo cameras [7], light field (LF) cameras provide dense sub-view images (SAIs), making disparity estimation more accurate and robust. Finally, compared with sensors like structured light [8] and TOF cameras [9], LF cameras do not require active lighting and are less susceptible to strong infrared light interference, which makes them also available in outdoor scenarios. As a result, light field cameras are not only used in general vision tasks, e.g., novel view synthesis and 3D defect inspection, but also widely used in microscopic imaging [10].

Light field disparity estimation is a long standing research topic in computer vision with strong potential for being used in various 3D applications [3], [11], [12]. According to training and estimation strategies, light field disparity estimation methods can be roughly classified into 3 categories: supervised learning [13], [14], [15], [16], [17], [18], unsupervised learning [19], [20], [21], [22] and optimization-based [18], [23], [24], [25], [26], [27], [28], [29], [30] methods. Although supervised learning provides a good balance between accuracy and efficiency, the heavy dependency on the ground-truth disparity from synthetic datasets for network training deteriorates the generalization capacity of such methods for real-world data. Unsupervised learning and optimization-based methods both eliminate the prerequisite of ground-truth disparity for training, thus having the potential for general, accurate, and efficient disparity estimation as demonstrated in recent works [19], [20], [21].

Existing optimization methods [23], [24], [25], [26], [27] estimate the disparity based on the photometric consistency assumption in angular patch. Photometric consistency assumption indicates the projection of light to each view has the same intensity, leading to the same color in the sub-view images. The typical two kinds of loss based on photometric consistency are defocus and correspondence [26]. Depth estimation from them have been studied extensively [30], [31], [32], [33] and achieved excellent performance. However, the photometric consistency no longer holds in the area where the observed target is occluded by other objects. As a result, the photometric consistency loss used in previous methods could not produce accurate disparity around occluded areas in which the warped pixel may come from occluders [25]. Some early work [25], [28], [29], [30] have been introduced to handle occlusion by excluding the occluded pixels. The most recent work [26] proposes an explicit voting strategy by first counting the number of refocused pixels in which absolute error to center view is below a predefined threshold and then selecting the disparity with the minimum number to achieve occlusion-aware disparity estimation.
Existing unsupervised methods [19], [20], [21] also suffer from occlusions and tend to produce even larger disparity estimation errors around those regions. Most recently, Jin [22] partition LFs into sub-LFs to separately predict multiple disparities and corresponding reliability maps to exclude the possible occluded sub-views, which improves the quality of disparity estimation to some extent and makes it possible to train the network with real-world data directly. Although [26] and [22] achieved accurate results on synthetic datasets, the ignorance of inherent occlusion patterns in the angular space leads to deteriorated performance on real-world data which may contain much more noise.

The key observation of our method is that, for every occluded pixel, the occlusion in the angular patch has strong patterns. Despite defocus and correspondence cue, the inherent occlusion patterns in light fields is another crucial cue for unsupervised learning. And if we can make the network "sense" the inherent occlusion patterns during the training process, we can exclude the occluded sub-views efficiently without redundant calculation in the inference process. To fulfill this goal, we propose OPAL: Occlusion Pattern Aware Loss, which incorporates occlusion patterns for calculating the photometric consistency loss directly. Specifically, for a pixel in the reference view, to calculate its OPAL, we first analyze and approximate its 2D occlusion pattern using a fixed number of 1D occlusion patterns in several directions. Note that the 1D occlusion pattern can further be simplified using a fixed number of pre-defined 1D patterns based on the observation that occlusions in light fields mainly occur from a single side in most cases. After getting the occlusion pattern, we can use it to mask out the occluded views when calculating the photometric consistency loss. Note that OPAL is only used for training, and the trained model successfully learned how to adaptively eliminate the occluded views for more accurate disparity estimation.

To fully release the potential of OPAL, we propose Occlusion Pattern Aware Network, named OPENet. Compared with the classical light field disparity estimation network, OPENet additionally introduces an EPI-transformer and a refinement module. The EPI-transformer is used to effectively integrate features of epipolar plane images (EPIs) with high inter-view similarity. We calculate the gradient maps of warped angular patches based on the initial disparity and take in the occlusion information to the refinement module for better aligning the final disparity to the reference center view. As shown in Table II, our network achieves the best overall accuracy in different datasets even when compared with the SOTA supervised methods, which demonstrates the strong performance of our method. We also conduct extensive comparisons on different real-world datasets to demonstrate the superior performance of our method. Finally, the proposed method significantly reduces the time required for network training (6 hours V.S. 1 week of LFatt [16]) and run-time inference (0.098 s V.S. 0.19 s of OAVC [26]).

The main contributions can be concluded as:

- We propose an unsupervised learning-based method with a novel training loss and a lightweight but effective network architecture for light field disparity estimation. Our method achieves more accurate and efficient disparity estimation performance compared to the existing disparity estimation solutions. The consistent performance on both synthetic and real-world data clearly demonstrates the strong potential of our method for practical usage.
- We propose OPAL, an Occlusion Pattern Aware Loss, which not only enables strong occlusion handling ability for accurate and robust light field disparity estimation but also significantly reduces the network parameters needed for accurate estimation.
- We propose OPENet, which is composed of a newly designed EPI-Transformer and gradient-based disparity refinement module, named GDRM, to further improve the overall performance by effective feature fusion and more accurate central-view alignment.

II. RELATED WORK

Currently, the light field disparity estimation algorithms can be categorized into the supervised method, the unsupervised method, and the optimization-based algorithm. Supervised learning-based methods have been widely used in LF disparity estimation due to their efficiency and excellent performance. The typical supervised methods are based on EPIs. For example, Shin et al. [15] presented a multi-branch CNN, which uses epipolar geometry information by extracting the features of EPIs from four directions and merges these features to improve accuracy. Tsai et al. [16] proposed to take all SAIs as input to build a cost volume for regularization and then regress the disparity by taking a weighted sum. Based on the angular attention, [34] further proposes a multi-level fusion network to perceive the angular occlusion and achieve the SOTA performance on the 4D light field benchmark. Unlike EPI-based methods, Shi et al. [14] adopted FlowNet2.0 for LF disparity estimation and refined the initial depth map according to warping errors. To reduce computational overhead and running time caused by 3D CNN, Huang et al. [17] proposed a lightweight network based on multi-disparity scale cost aggregation and introduced a boundary guidance sub-network to improve performance. Most recently, Wang et al. [18] proposed a simple cost constructor to construct matching costs and handle occlusion, which achieves the best performance in the supervised learning-based methods up to now.

The unsupervised methods and optimization-based methods eliminate the prerequisite of ground-truth disparity for training, thus having the potential for general, accurate and efficient disparity estimation as demonstrated in recent works [19], [20], [21], [22]. For the first time, Peng et al. [19] put forward an unsupervised learning-based method that uses the constraints between SAIs for training. Furthermore, Zhou et al. [21] designed three unsupervised loss functions (defocus loss, photometric loss, and symmetry loss) according to the inherent depth cues and geometry constraints of light field images. However, these two methods both performed poorly in occluded and textureless areas, which demonstrates that it is a tricky and crucial problem to handle occlusion without real disparity as supervision. Jin et al. [22] proposed to produce several initial depth maps, then fuse them based on reliability maps and an occlusion mask (generated
by a threshold) to obtain the final depth. Due to artificially set
the hard threshold, decision errors would occur at the edge of
occluded areas, resulting in massive estimation errors.

The occlusion problem is always the key to optimization-
based methods [23], [26], [27], [35], [36], [32]. Williem et
al. [27] designed an angle entropy measurement and an adaptive
defocus response function to construct data cost, but it cannot be
practicable due to time-consuming global optimization. Taking
advantage of GPU accelerating, Han et al. [26] proposed a
fast and occlusion-aware method that separates refocused pixels
by a threshold and utilizes the number of the separated pixels
to select the correct depth. Although it achieves competitive
performance to learning-based methods, the hyperparameters
may need to be adjusted carefully when applied to different
scenarios. More importantly, as shown in Figs. 8 and 9, it cannot
achieve consistently well performance for complex real-world
scenarios.

In summary, existing methods are either constrained by the
huge amount of synthesized training data or influenced by
photometric consistency loss as the “only cue” for disparity
estimation. Different from these methods, in this paper, we
consider the inherent patterns of the occlusion for disparity
estimation, thus refining the final estimation accuracy.

III. METHOD

In this section, we first introduce the basic photometric con-
sistency loss for disparity estimation, which handles disparity
ambiguities by removing the impact of occlusion. Then we elabo-
rate on the proposed Occlusion Pattern Aware Loss (OPAL) for
achieving occlusion-aware photometric consistency by utilizing
the inherent occlusion patterns of the light fields effectively.
Finally, we introduce the network architecture, named OPENet,
for accurate, efficient, and robust disparity estimation in an
unsupervised manner based on OPAL.

A. Analyzing Photometric Consistency Loss

The 4D light field systems usually use the intersection points
of the lines on two parallel planes to parameterize the light rays in
the 3D space [37], as known as the two-plane parameterization
of light field. The intersecting points can be denoted as 2-D
angular coordinate \((u, v)\) for the first plane and 2-D spatial
coordinate \((x, y)\) for the second plane, respectively. Without loss
of generality, we suppose the angular dimension of the 4D light
field is \(N \times N\), i.e., it has \(N \times N\) sub-views, and the spatial
dimension (or resolution) of each sub-view image is \(H \times W\).
In the following formulation, we use \(u\) to represent the 2-D angular
coordinate \((u, v)\) and \(x\) to represent the 2-D spatial coordinate
\((x, y)\) for simplicity. We define \(u_0\) as the angular coordinate of
the center view as the reference view for disparity estimation.
Under the Lambertian material assumption, sub-view images
warped to the center view according to the ground truth disparity
have high similarity to the center view image, which can be
formulated as

\[
I_{u_0}(x) = I_u(x + (u - u_0) \cdot D(x)), \quad (1)
\]

where \(I_u(x)\) is the intensity (or RGB) value of pixel \(x\) in view \(u,
D(x)\) is the ground truth disparity of pixel \(x\) in the center view.

According to (1), the disparity is estimated by finding the
pixel position interval of the consistent color of pixels between
the central view and the warped sub-views. However, the photom-
etric consistency loss no longer holds when occlusion happens.
As shown in Fig. 1, the rays from the background (blue) are oc-
cluded by the foreground (red), leading to the inconsistent color
description of the same point and further ambiguous disparity
estimation results. Here we present an example in Fig. 2 to better
demonstrate the benefits of occlusion-aware photometric loss
for disparity estimation. In Fig. 2, we sample two 4*4 patches,
one without occlusion and one with occlusion, from the scene
“Dino” of 4D Light Field Benchmark. Based on these patches,
we conduct a comparison between the basic photometric loss
and the occlusion-aware photometric loss (calculating the loss
by first masking out all the occluded views according to an
occlusion mask). The reported 3 losses are: photometric loss
calculated without occlusion mask (basic photometric loss, blue
line), with ground truth occlusion mask (orange line), and with
OPAL (our proposed occlusion-aware photometric loss, green
line), respectively. As for the non-occlusion case in Fig. 2(a),
all the 3 loss configurations reach the minimum value only when

![Fig. 1. Influence of photometric inconsistency in occluded regions (the black square is the refocused position). Under occlusion, the photometric consistency loss doesn’t hold no matter whether we refocused to the correct depth(left) or incorrect depth. The basic photometric consistency loss could not guide accurate depth estimation under such conditions.](image1)

![Fig. 2. Influence of occlusion mask on the photo-metric loss in non-occluded (a) and occluded (b) regions, respectively. Red Resolving disparity estimation ambiguity caused by occlusions using occlusion masks. The horizontal axis indicates the disparity estimation error \(dD\) and the vertical axis indicates the photometric loss calculated with (green and orange lines) or without (blue lines) occlusion mask. In the non-occluded regions (Fig. 2(a), the photometric loss reaches the minimal value when \(dD = 0.0\), which indicates we can estimate accurate disparity by optimizing the simple photometric loss without occlusion mask. However, in the occluded regions, without masking out the sub-views that obscure the center view using the ground truth or estimated occlusion masks, the photometric loss remains high even with ground truth disparity (e.g., when \(dD = 0.0\) as demonstrated by the blue line in (b), which demonstrate the disparity estimation ambiguity caused by occlusions. Only with proper occlusion masks, we can determine the smaller the photometric loss the more accurate the disparity estimation as demonstrated by the green and orange lines in (b).](image2)
the disparity error, denoted by $dD$, tends to be 0. However, for the occlusion case in Fig. 2(b), the disparity error cannot reach 0 (almost 0.2) by minimizing the loss without an occlusion mask. With the ground-truth occlusion mask, $dD$ is close to 0 when the loss reaches the minimum value. This demonstrates the ability of using occlusion masks for resolving disparity estimation ambiguity. Finally, the proposed OPAL Fig. 2(b) also approximates the performance using the ground truth occlusion mask, which demonstrates the strong occlusion handling performance of OPAL.

### B. OPAL: Occlusion Pattern Aware Loss

According to Fig. 1, occlusion-aware photometric consistency loss is necessary for handling occlusions. We first define the occlusion mask mentioned above, named $OP$, as a 2D bool matrix indicating whether the center view is occluded or not by the sub-views,

$$OP(u) = \begin{cases} 
1, & \text{center view is not occluded by sub-view } u \\
0, & \text{otherwise.}
\end{cases} \quad (2)$$

where $u$ is angular coordinate of the sub-view. The masked photometric cost at pixel $x$ can then be defined as

$$Cost(x) = \sum_u OP(u) \cdot |I_{u \rightarrow u_0}(x, \hat{D}) - I_{u_0}(x)|, \quad (3)$$

where $I_{u \rightarrow u_0}(x, \hat{D})$ represents the pixel value that warped from sub-view $u$ to central view $u_0$ using the estimated disparity $\hat{D}$. Since the light field is well structured, the candidate patterns of $OP$ are same for each of the $H \times W$ pixels. Based on the observation in Figs. 1 and 2(b), for each pixel, we can traverse all possible patterns of $OP$ and select the one with the lowest $Cost$, named $OP^*$, to construct the occlusion-aware photometric loss. However, the solution traversing of $2^N \times N$ patterns is computation-consuming. To further reduce the computation complexity, we present a pattern approximation as follows.

We first approximate the 2D $OP^*$ with two 1D occlusion masks, which only has $2^N$ possible patterns in horizontal and vertical direction, respectively. Thus, the possible 1D occlusion patterns can be defined as Occlusion Pattern in Line, named $OPL$. Another important observation is the candidate occlusion patterns of $OPL$ in any directions can be further simplified according to the structure of the light field. Specifically, when traversing $OPL$ for the optimal $OPL^*$ in one direction, we can assume in most cases: i) occlusion, if exist, usually starts from the boundary sub-views (e.g., the left most or the right most sub-view for the horizontal $OPL$) and ii) sub-views that obscure the center view are often next to each other. Based on the assumptions, we finally reduce the number of the candidate occlusion patterns of $OPL$ from $2^N$ to $N$, with only 9 patterns for a $9 \times 9$ light field in vertical Fig. 5(b) or horizontal Fig. 5(c) direction, thus significantly improving the training efficiency. Without loss of generalization, we only discuss $OPL$ in the vertical direction in the following part of the paper.

As illustrated in Fig. 5(b), for the vertical direction, we pre-defined $N$ candidate occlusion patterns of $OPL$. If $j \leq \left\lceil \frac{N-1}{2} \right\rceil$,

$$OPL_j(u) = \begin{cases} 
1, & u \leq j + \frac{N-1}{2} \\
0, & \text{otherwise.}
\end{cases} \quad (4)$$

Conversely, if $j > \left\lceil \frac{N-1}{2} \right\rceil$, the top views are not occluded, the $j$th $OPL$ can be defined as

$$OPL_j(u) = \begin{cases} 
1, & u \geq j - \frac{N-1}{2} \\
0, & \text{otherwise.}
\end{cases} \quad (5)$$

During training, we apply $j$th $OPL$ to obtain the masked photometric cost of pixel $x$ and revise (3) as

$$Cost_{opt}(j, x) = \sum_u OPL^*_j(u) |I_{u \rightarrow u_0}(x, \hat{D}) - I_{u_0}(x)|. \quad (6)$$

Then, we optimize the per-pixel $OPL^*_x$ in each direction by traversing all possible $OPLs$,

$$OPL^*_x = \arg \min_j Cost_{opt}(j, x). \quad (7)$$

Finally, the occlusion pattern awareness loss in each direction can be defined as

$$\ell_{optal}(\hat{D}) = \frac{1}{HW} \sum_x OPL^*_x \cdot |I_{u \rightarrow u_0}(x, \hat{D}) - I_{u_0}(x)|, \quad (8)$$

where $H \times W$ is the spatial resolution and the sum of $\ell_{optal}$ in all directions is used to penalize $\hat{D}$. As shown in Fig. 3, without OPAL, the disparity inferred by the network trained cannot generate a reasonable occlusion map, while OPAL can guide the network to sense the occlusion more accurately for light field disparity estimation.

### C. Network Architecture and Implementation Details

As illustrated in Fig. 4, the proposed OPENet contains a feature extractor, an EPI-Transformer, a disparity regression module, and a gradient-based disparity refinement module (GDRM). The input are SAIs from four angular directions (horizontal, vertical, left & right diagonal). We use an SPP module [39]
Fig. 4. Overview of the network architecture and training loss. Our network, named OPENet, contains four components: an image feature extractor, an EPI-Transformer, a disparity regression module and a gradient-based disparity refinement module (GDRM). The green part demonstrates the proposed training loss, OPAL, which is only used for the unsupervised training process.

with shared weights to extract global features of each sub-view, which could better exploit the hierarchical contextual information and the correlation between adjacent regions. After getting the feature maps of different views, an EPI-Transformer is introduced to construct cross-view correlations for effective feature fusion. The fused feature maps are then fed into the disparity regression module to estimate the raw disparity, which is denoted as $D_{raw}$. Finally, a gradient-based refinement module (GDRM) is introduced to better align the disparity map with the center view image. The final estimated disparity is denoted as $D_{final}$.

1) EPI-Transformer: Attention-based architectures have been applied in various computer vision tasks [40], [41], [42]. The Transformer [43] that relies on the attention mechanism can merge the information of all tokens and model the global dependencies of input and output effectively. Inspired by [40], [43], we propose an EPI-Transformer to enhance cross-view correlations for more effective feature fusion. Given a sequence of angular tokens of the SAIs in each direction, the information can be effectively fused by the EPI-Transformer for disparity regression. Specifically, our EPI-Transformer consists of four identical streams corresponding to the four directional inputs. Since EPI-based methods usually utilize the proportional relationship between disparity and slope of EPIs and most pixels outside the focal plane are not well-aligned, we concatenate the feature of 2*2 neighboring pixels to compensate for this “misalignment”. Thus, for each stream, we first concatenate $N$ feature maps from $N$ sub-views and rearrange the feature map from $N \times HW \times C$ to $N \times HW/4 \times 4C$, denoted as $E$, where $H \times W$ is the spatial resolution of SAIs and $4C$ is the embedding dimension. Then we calculate position encoding according to (9)

$$
\begin{align*}
PE(pos, 2i) &= \sin \left( \frac{pos}{10000} \right) \\
PE(pos, 2i + 1) &= \cos \left( \frac{pos}{10000} \right)
\end{align*}
$$

(9)

where $pos$ is the position of the SAI in the epipolar plane and $i$ is the channel index. We take the sum of $PE$ and $E$ as the input embedding and put it through layer normalization. Afterward, we multiply input embedding by $W_Q$, $W_K$ and $W_V$ to generate $Q$, $K$ and $V$ and apply multi-head self-attention ($MHSA$) to learn the dependence among different views

$$MHSA(E_A) = Concat(H_1, \ldots, H_h)W_O$$

(10)

$$H_i = \text{softmax} \left( \frac{Q_iK_i^T}{\sqrt{C/h}} \right)V_i$$

(11)

where $h$ is the number of head groups, $W_O$ is the output projection matrix. Then we feed the sum of $MHSA$ and $E$ into a residual structure composed of layer normalization and MLP to obtain $E_{fusion}$. We get the final fused feature maps after another reshaping operation and additional convolution layers (a 3D convolution layer with kernel size $N \times 3 \times 3$ and a 2D convolution layer). Finally, the features in four directions are concatenated to regress the raw disparity $D_{raw}$.

2) Gradient-Based Disparity Refinement Module (GDRM): Although our OPAL helps the model converge better and fast by handling most occlusion cases, we observe some dilation still appears in the object edges when foreground and background have similar intensity. To solve this issue, we introduce GDRM to align the estimated disparity with the center view. We first warp SAIs in each direction to the central view based on $D_{raw}$.
TABLE I
QUANTITATIVE (MSE × 100 AND BRP0.07) COMPARISONS WITH STATE-OF-THE-ART METHODS ON 4D LIGHT FIELD BENCHMARK [38] AND HCI BLENDER [35]

| Scene                | Optimization-based | Supervised | Unsupervised |
|----------------------|--------------------|------------|--------------|
|                      | SPO     | CAE     | OAVC | EPINet | LFattNet | OACC | OccUnNet | Ours-fast | Ours      |
| Buddha               | 0.54    | 0.64    | 0.51 | 0.39   | 0.33     | 0.476 | 1.27      | 0.34      | 0.31      | 0.32      |
| Buddha2              | 1.05    | 0.35    | 1.34 | 0.63   | 0.56     | 5.15  | 1.36      | N/A       | 1.53      | 1.06      |
| Horse                | 1.34    | 0.79    | 0.53 | 7.35   | 6.32     | 9.84  | 1.65      | 1.52      | 0.81      | 0.54      |
| Medieval             | 0.91    | 0.97    | 0.94 | 2.28   | 0.50     | 1.80  | 1.27      | 0.70      | 0.91      | 0.75      |
| MonaRoom             | 0.55    | 0.49    | 0.51 | 1.34   | 0.78     | 0.89  | 1.68      | 0.57      | 0.43      | 0.36      |
| Papilion             | 0.77    | 0.64    | 0.84 | 5.19   | 4.65     | 3.52  | 3.72      | 1.11      | 0.63      | 0.58      |
| StillLife            | 1.51    | 1.24    | 4.07 | 2.43   | 1.41     | 7.73  | 2.35      | 1.57      | 1.48      | 1.10      |
| Average              | 0.95    | 0.73    | 1.05 | 2.93   | 4.71     | 4.20  | 1.89      | 0.97      | 0.87      | 0.67      |

4D LF Benchmark (MSE×100)

| Scene                | Optimization-based | Supervised | Unsupervised |
|----------------------|--------------------|------------|--------------|
|                      |                    |            |              |
| Backgamon            | 4.58  | 6.07 | 3.84 | 3.71 | 3.65 | 3.94 | 34.71 | 6.68 | 5.42 | 5.17 |
| Dots                 | 5.23  | 5.08 | 1.68 | 1.48 | 1.42 | 1.42 | 59.91 | 6.56 | 12.09 | 7.64 |
| Pyramids             | 0.043 | 0.048 | 0.04 | 0.01 | 0.01 | 0.004 | 0.035 | 0.21 | 0.02 | 0.016 |
| Stripes              | 6.95  | 3.56 | 1.32 | 0.93 | 0.89 | 0.84 | 11.75 | 5.20 | 1.42 | 1.36 |
| Bedroom              | 0.21  | 0.23 | 0.21 | 0.20 | 0.36 | 0.15 | 0.92 | 0.38 | 0.33 | 0.31 |
| Bicycle              | 5.57  | 5.13 | 4.89 | 4.60 | 3.35 | 2.90 | 11.73 | 6.23 | 5.37 | 5.12 |
| Herbs                | 11.23 | 11.66 | 10.4 | 9.49 | 6.61 | 6.56 | 37.51 | 13.94 | 12.67 | 12.62 |
| Origami              | 2.03  | 1.78 | 1.48 | 1.48 | 1.73 | 0.88 | 8.81 | 1.92 | 1.62 | 1.41 |
| Boxes                | 9.11  | 8.42 | 6.99 | 6.05 | 4.09 | 2.59 | 11.35 | 7.45 | 6.34 | 5.76 |
| Cotton               | 1.31  | 1.51 | 0.60 | 0.23 | 0.21 | 0.16 | 6.46 | 0.80 | 0.55 | 0.43 |
| Dino                 | 0.31  | 0.38 | 0.27 | 0.18 | 0.08 | 0.083 | 1.89 | 0.63 | 0.43 | 0.34 |
| Sideboard            | 1.02  | 0.88 | 1.05 | 0.79 | 0.50 | 0.54 | 4.55 | 1.79 | 3.17 | 1.12 |
| Average              | 3.97  | 3.73 | 3.97 | 2.43 | 1.91 | 1.70 | 15.80 | 4.31 | 3.98 | 3.42 |

Overall average: 2.85 | 2.62 | 2.97 | 2.61 | 2.94 | 2.62 | 10.67 | 3.20 | 2.84 | 2.41 |
Excluding Stratified average: 2.50 | 2.34 | 2.30 | 2.90 | 3.33 | 2.90 | 6.43 | 2.78 | 2.33 | 2.12 |

4D LF Benchmark (BRP0.07)

| Scene                | Optimization-based | Supervised | Unsupervised |
|----------------------|--------------------|------------|--------------|
|                      |                    |            |              |
| Buddha               | 2.33  | 3.24 | 1.81 | 1.54 | 2.03 | 4.53 | 12.37 | 4.11 | 3.15 | 2.35 |
| Buddha2              | 14.19 | 5.05 | 12.2 | 34.77 | 34.23 | 35.08 | 30.89 | N/A | 14.91 | 12.44 |
| Horse                | 6.21  | 21.35 | 5.18 | 16.4 | 16.20 | 18.50 | 19.64 | 26.95 | 10.32 | 7.92 |
| Medieval             | 9.56  | 17.1 | 10.79 | 18.8 | 11.70 | 20.20 | 18.41 | 16.48 | 12.35 | 11.36 |
| MonaRoom             | 6.69  | 7.31 | 5.86 | 10.77 | 10.76 | 12.01 | 19.29 | 10.57 | 8.17 | 6.73 |
| Papilion             | 25.52 | 7.86 | 13.40 | 35.64 | 34.81 | 33.31 | 27.93 | 36.36 | 14.83 | 12.15 |
| StillLife            | 9.08  | 14.80 | 12.71 | 11.42 | 11.70 | 10.74 | 19.21 | 17.14 | 8.93 | 6.90 |
| Average              | 10.00 | 11.1 | 8.85 | 18.47 | 17.34 | 19.20 | 21.10 | 18.60 | 10.38 | 8.55 |

Overall average: 3.78 | 3.92 | 3.12 | 3.5 | 3.12 | 3.93 | 36.42 | 14.37 | 8.91 | 8.83 |
Excluding Stratified average: 3.62 | 4.53 | 69.13 | 2.49 | 3.43 | 1.51 | 56.10 | 45.34 | 36.21 | 48.31 |

OCIAL MODELS ON 4D LIGHT FIELD BENCHMARK [38] AND HCI BLENDER [35]

and then calculate the gradient maps between adjacent views for each pixel. Apparently, the average gradient indicates the probability of being miscalculated. As shown in Fig. 4, we concatenate the central view SAI, the raw disparity, and the average gradient map together as input and sequentially pass it through the network designed in [42] to obtain the refined disparity in different directions. The final disparity \( D_{\text{final}} \) is the average of the refined disparity maps from all directions.

3) OPENet-Fast: By eliminating the EPI-Transformer and GDRM and decreasing the directions (only using horizontal and vertical directions) as inputs, we can significantly improve the efficiency without sacrificing too much accuracy. We call this light-weight version OPENet-fast. In OPENet-fast, we only use the additional convolution layers mentioned in Section III-C1 for feature fusion. As shown in Table I, such a simple structure could also significantly improve the overall performance compared to existing methods, which also demonstrates the strong benefit of OPAL. We also compare parameters and the inference time in Table II, which demonstrates the high efficiency of OPENet-fast.

4) Implementation Details: We also apply the edge-aware smoothness loss [44] to further improve the smoothness of textureless areas and provide disparity with sharp edges, which is defined as

\[
\ell_{\text{smooth}}(\bar{D}, I_u 0) = |\partial_x \bar{D}| e^{-\gamma |\partial_x I_u 0|} + |\partial_y \bar{D}| e^{-\gamma |\partial_y I_u 0|},
\] (12)
where the edge weight $\gamma$ is set to 150. The total loss function is given by

$$\ell_{total} = \lambda_1 \ell_{raw} + (1 - \lambda_1) \ell_{final} + \lambda_2 \ell_{smooth},$$

(13)

where $\ell_{raw}$ and $\ell_{final}$ are OPAL (based on L1 norm) of $D_{raw}$ and $D_{final}$, respectively. $\lambda_1$ is set to 0.6 and 1.0 for OPENet and OPENet-fast, respectively. $\lambda_2$ is equal to 0.3.

We randomly crop $64 \times 64$ RGB patches from the LF images for training and the data augmentation strategy in EPINet [15] was used to improve robustness. We use Adam optimizer [46] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and the batch size is set to 16. The initial learning rate is 1e-3 and is decreased to 1e-6. It only takes 210 epochs for convergence, and the training time is 6 hours for OPENet and 4 hours for OPENet-fast on a single NVIDIA RTX 3080, respectively.

### IV. Experimental Results

In this section, we first introduce the datasets and metrics used for evaluation. Then we report the quantitative and qualitative comparisons with state-of-the-art methods. Finally, we perform ablation studies to analyze different components of our method.

#### A. Datasets and Evaluation Criteria

Our method is evaluated on three synthetic datasets (4D Light field Benchmark [38], the HCI Blender [35], and the dataset published by [45]) and four kinds of real-world datasets from Stanford Lytro LF Archive [47], Kalantari et al. [48], gantry camera dataset [49] and EPFL LF dataset [50].

**Synthetic LF Datasets:** The 4D Light field Benchmark [38] and HCI Blender [35] are the most widely used datasets for evaluating light field disparity estimation. Both of them have the angular resolution of $9 \times 9$. For all learning-based methods, we use 16 scenes in a subset (Additional) of the 4D Light field Benchmark for training and evaluating all the methods on other scenes of two datasets for better evaluating the generalization capacity. Note that we further verify the feasibility of our method on the textureless and non-uniform lighting scenes in [45] directly without any fine-tuning process.

**Real-World LF Datasets:** Stanford Lytro LF Archive dataset [47] consists of 251 real-world scenes captured by the Lytro Illum camera and the dataset published by Kalantari [48] has 97 different scenes, including 72 training scenes and 25 testing scenes. The two real-world datasets have the same angular resolution of $14 \times 14$. The EPFL [50] is another LF dataset captured by a Lytro camera, which has the angular resolution of $9 \times 9$. For the dataset [49] captured with a gantry camera, the angular resolution is $17 \times 17$. Due to the large disparity, we downsample the SAI’s by a factor of 2 to ensure all datasets have the same disparity range ($[-4, 4]$). Note that we only apply the central $9 \times 9$ SAI’s of all datasets for disparity estimation. Since the real-world datasets could not provide ground-truth disparity, we only retrain the unsupervised models using the training set in [48] and test all the methods (including optimization-based, supervised, and unsupervised methods) on the scenes of other datasets.

**Metrics:** We use Bad Pixel Ratio (BPR) defined in [38] and Mean Square Errors (MSE) for quantitative evaluation on synthetic datasets. Note that BPR measures the percentage of mistakenly estimated pixels whose errors exceed $\epsilon$, which is set to 0.07 for evaluation.

#### B. Comparisons With State-of-the-Art Methods

**Methods for Comparison:** We quantitatively and qualitatively compare our methods with other state-of-the-art methods, including optimization-based methods (CAE [27], SPO [24], OAVC [26]), supervised learning-based methods (EPINet [15], LFAttNet [16], OACC [18]) and existing unsupervised methods (Unsup [19], OccUnNet [22]). Noting that the code of OccUnNet [22] and AttMLFNet [34] is unavailable or incomplete, so we only show some results in their papers. Additionally, we conduct extensive experiments to evaluate the superiority of our methods compared with single image [4], monocular video [5], [6] and NeRF [52] based methods.

**Quantitative Comparison on Synthetic Data:** For synthetic datasets, we train the model with simplified occlusion patterns for better performance. Table I reports the results of all methods. Fig. 6 visualizes the disparity and error maps of different methods. Note that the Stratified scenes in 4D LF Benchmark is not exist in the real world, so we exclude the Stratified scenes
Fig. 6. Qualitative comparisons of the disparity and error maps on synthesis datasets 4D Light field Benchmark [38] and HCI Blender [35].

Fig. 7. Qualitative comparisons with two representative methods on dataset [45]. Even under non-uniform lighting conditions, our method outperforms the SOTA ones.

when calculating the average metrics to reflect the performance and generalization ability of the algorithms in real scenarios. For the dataset of [45], Fig. 7 shows the results of some representative methods in two scenes.

- OPENet-fast achieves comparable performance to the optimization-based methods on the accuracy, while OPENet surpasses them significantly. For areas with poor texture or severe occlusion, our method performs better.
- Compared to the previous unsupervised methods [19], [22], ours could handle most occlusion cases well and produce less noisy disparity, thus greatly surpassing them.
- Supervised methods achieve better performance on 4D LF Benchmark [38] than ours. However, significant performance degradation occurs when generalizing them to other datasets (right in Table I and Fig. 7). The performance of our method is much more robust across different datasets.

Qualitative Comparisons on Real-World Datasets: Fig. 8 shows the qualitative results on [47], [48]. The disparity maps generated by the existing unsupervised method [19] have obvious errors in occluded and texture-less areas. Due to severe noise and complex structures in real-world scenarios, even the SOTA optimization-based approach (OAVC) cannot maintain good performance with carefully adjusted parameters. Moreover, supervised methods inevitably suffer serious performance degradation on real-world datasets. While the proposed method maintains the sharp disparity in real-world LF images, showing obvious performance improvement compared to supervised baselines. Benefiting from the delicate design of OPAL, the ability to re-train on diverse real data, and the smoothness loss constrain, non-occlusions are also well regularized in our method. In Fig. 9, we convert the disparity into 3D meshes to verify the accuracy. The impressive results show our method tends to provide more accurate disparity and even could be used for further 3D applications.

To verify the generalization and robustness of our method in real scenes, we additionally compare the results of indoor scenes captured by the gantry camera [49] and another outdoor dataset [50]. As shown in Fig. 10, the indoor scene has only slight noise, and the supervised method will produce some artifacts.
The results of the optimization algorithm and the unsupervised algorithm produce obvious errors. In the outdoor scene with severe noise, like Fig. 11, affected by the lighting and complex scene structure, these methods degrade seriously. In comparison, our results maintain great integrity in both indoor and outdoor environments. We believe this is due to the fact that our algorithm is able to perform additional training without ground truth, which allows our network to adapt to varying levels of noise, while [19] does not take full advantage of the superiority of unsupervised learning, yielding less than satisfactory results.

Comprehensive Comparisons of Efficiency and Generalization Capacity: We comprehensively compare the overall performance with other state-of-the-art methods according to running time, network parameters, MSE $\times 100$ and BPR0.07 on synthetic datasets. As shown in Table II, the overall accuracy of our method outperforms SOTA methods. Although LFattNet and OACC achieve the best accuracy in 4D LF Benchmark [38] (Table I), the performance decreases significantly in HCI Blender [35] due to data distribution bias. Note that the performance of CAE also decreases in [38] although it
performs well in [35]. This obviously demonstrates the strong generalization capacity of our method. Moreover, we show the running time of different methods on CPU and GPU, which significantly illustrates the efficiency of our method. Note that OAVC runs faster than OPENet by fully taking advantage of GPU acceleration. In addition to high run-time efficiency, we would like to mention the high training efficiency (6 hours V.S. 1 week of LFattNet) of our method. Since OPAL is strong enough to guide the training to achieve fast convergence, we can use a lightweight model (1.047 Mb) for achieving high efficiency without sacrificing accuracy. To ensure fairness, all the methods are tested on the same PC equipped with an Intel 2.1 GHz CPU and an NVIDIA RTX 3080 GPU.

Comparisons With Different Kinds of LF Reconstruction and Depth Estimation Methods: To demonstrate the effectiveness of our method in a more comprehensive manner, we quantitatively and qualitatively compare with the fully-supervised single image based method MiDaS [4], self-supervised monocular video based method Monodepth2 [5] and NeRF-based method TensoRF [51] on the 4D LF Benchmark. Note that even though the light field data is dense and structured, light field depth estimation [3] is non-trivial due to i) the micro-baseline setup (the shorter the baseline the more estimation ambiguity as explained in [53]) and ii) the complex occlusions. Thus, it remains difficult for algorithms without carefully considering

the structured patterns of light field to produce comparable depth estimation results with light-field depth estimation algorithms as shown in Fig. 12 and Table III.

C. Ablation Study

We perform comprehensive ablation studies for network design, OPAL, smoothness loss, and different 1D pattern selections for OPAL on four representative scenes (Boxes, Cotton, Dino, and Sideboard) of 4D LF Benchmark [38]. First, we remove the different model components and OPAL to show their contributions to the estimation performance. As shown in Fig. 13, OPAL significantly improves the performance in occluded boundaries and the GDRM further helps OPENet promote the sharpness in subtle occlusion areas. Additionally, the introduced EPI-Transformer could optimize the overall performance. Meanwhile, OPENet trained on real-world datasets

| Methods | MSE×100 ↓ | BPR0.07 ↓ |
|---------|-----------|-----------|
| MiDaS [4] | 7.75 | 41.51 |
| Monodepth2 [5] | 20.64 | 58.56 |
| TensoRF [51] | 66.49 | 93.64 |
| Ours | 3.42 | 11.94 |
performs better than OPENet-fast obviously, which indicates the significance of the feature fusion and refinement strategy in our network. We think that severe noise and uneven lighting in the real world have a serious influence on occlusion pattern selection. The EPI-Transfomer and GDRM make our model more robust to handle these complicated scenes. The quantity ablation is also provided in Table IV.

Additionally, we also analyze the impact of the smoothness loss in (13). As in Table IV, the metric reported shows that it is effective for the overall performance. Since the pattern is pixel-wise, it is difficult for the pattern to maintain good continuity in textureless regions, resulting in slight noise in the estimated disparity map. The smoothness loss can ensure the local consistency of disparity maps, thus improving the overall quality of the disparity maps.

Finally, we evaluate different 1D approximations of 2D patterns for OPAL, including vertical (V), horizontal (H) and diagonal (D) directions. Obviously, the pattern in one direction cannot fully represent the 2D information, and additional diagonal patterns cause overall performance degradation due to warping errors. As a result, OPENet trained with the 1D pattern in vertical and horizontal directions performs best.
Fig. 13. Qualitative evaluation. We evaluate different components of our method, including OPAL, GDRM, and EPI-Transformer, qualitatively using disparity and error maps. (a) Baseline, (b) OPAL only, (c) OPAL+GDRM, (d) Ours.

Table IV
Ablation study on OPAL, Network Design, Smoothness Loss and Different 1D Pattern Selections on Four Scenes (Boxes, Cotton, Ding, and Sideboard) of 4D LF Benchmark [38]

| Methods          | MSE x 100 w/o OPAL | MSE x 100 w OPAL |
|------------------|---------------------|------------------|
| Fast             | 3.828               | 2.17             |
| Fast+GDRM        | 2.931               | 2.03             |
| Fast+Trans       | 3.137               | 2.07             |
| Fast+Trans+GDRM(Ours) | 2.504               | 1.91             |
| Ours w/o $\ell_{\text{smooth}}$ | -               | 2.13             |
| Ours, V          | -                   | 2.08             |
| Ours, II         | -                   | 2.10             |
| Ours, VHD        | -                   | 1.96             |
| Ours, VH          | -                   | 1.91             |

Trans: EPI-transformer, V: vertical, H: horizontal, VH: vertical and horizontal, VHD: vertical, horizontal and diagonal directions.

V. CONCLUSION

Limitations and Future Work: Despite the common issues caused by non-Lambertian or transparent materials on real-world data, the assumptions for 1D occlusion patterns can not cover some extreme cases like very fine grids in front of an object or very small but deep holes. Designing more occlusion patterns may resolve this problem and we leave this as future work. We would also apply OPAL for unstructured light fields. Finally, we would leverage GPU acceleration tools like TensorRT to further improve the overall efficiency of our method.

Conclusion: In this paper, we design the occlusion pattern aware loss (OPAL) and propose the unsupervised OPENet framework, achieving accurate, robust, and lightweight LF disparity estimation in general scenarios. Experimental results demonstrate that our method significantly reduced the computation overhead (for both training and inference) and achieve more robust performance even when compared with state-of-the-art supervised methods. More importantly, our method can effectively avoid the domain shift effect when generalizing to real-world scenarios. We believe OPAL and OPENet redefine the balance between accuracy, generalization, and efficiency for light field disparity estimation.

REFERENCES

[1] B. S. Wilburn, M. Smański, H.-H. K. Lee, and M. A. Horowitz, “Light field video camera,” in Media Processors 2002, vol. 4674, Paris: SPIE, 2001, pp. 29–36.
[2] Raytrix, “3D light field camera technology,” 2017. [Online]. Available: http://www.raytrix.de/
[3] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, “Light field photography with a hand-held plenoptic camera,” Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, 2005.
[4] B. Raeydt, K. Lasinger, D. Hafner, T. Bartow, and T. Koltun, “Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer,” IEEE Trans. Pattern Anal. Mach. Intel., vol. 44, no. 3, pp. 1623–1637, Mar. 2022.
[5] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into self-supervised monocular depth estimation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3828–3838.
[6] A. Gordon, H. Li, R. Joungchikowski, and A. Angelova, “Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 8977–8986.
[7] D. Lee and I. Kweon, “A novel stereo camera system by a biprism,” IEEE Trans. Robot. Autom., vol. 16, no. 5, pp. 528–541, Oct. 2000.
[8] Z. Xiong, Y. Zhang, F. Wang, and W. Zeng, “Computational depth sensing: Toward high-performance commodity depth cameras,” IEEE Signal Process. Mag., vol. 34, no. 3, pp. 55–58, May 2017.
[9] A. Bhandari and R. Raskar, “Signal processing for time-of-flight imaging sensors: An introduction to inverse problems in computational 3-D imaging,” IEEE Signal Process. Mag., vol. 33, no. 5, pp. 45–58, Sep. 2016.
[10] J. Wu et al., “Iterative tomography with digital adaptive optics permits hour-long intravitral observation of 3D cellular dynamics at millisecond scale,” Cell, vol. 184, no. 12, pp. 3318–3332, 2021.
[11] F.-C. Huang, K. Chen, and G. Wetzstein, “The light field stereo: Immerse graphics via factored near-eye light field displays with focus cues,” ACM Trans. Graph., vol. 34, no. 4, Jul. 2015, Art. no. 60, doi: 10.1145/2766922.
[12] M. Martinez-Corral and B. Javidi, “Fundamentals of 3D imaging and displays: A tutorial on integral imaging, light-field, and plenoptic systems,” Opt. Eng., vol. 45, no. 9, Sep. 2006.
[13] M. Tschopp and B. Pock, “Convolutional networks for shape from light field,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 3746–3754.
[14] J. Shi, X. Jiang, and C. Guilleminot, “A framework for learning depth from a flexible subset of dense and sparse light field views,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5867–5880, Dec. 2019.
[15] C. Shin, H.-G. Jeon, Y. Yoon, I. S. Kweon, and S. J. Kim, “EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 4748–4757.
[16] Y.-J. Tsai, Y.-L. Liu, M. Ouyoung, and Y.-Y. Chuang, “Attention-based view selection networks for light-field disparity estimation,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 12095–12103.
[17] Z. Huang, X. Hu, Z. Xue, W. Xu, and T. Yue, “Fast light-field disparity estimation with multi-disparity-scale cost aggregation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 6320–6329.
[18] W. Yang, L. Wang, Z. Liang, J. Yang, W. An, and Y. Guo, “Occlusion-aware cost constructor for light field depth estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 19809–19818, 2022.
[19] J. Peng, Z. Xiong, D. Liu, and X. Chen, “Unsupervised depth estimation from light field using a convolutional neural network,” in Proc. Int. Conf. 3D Vis., 2018, pp. 295–303.
[20] 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, doi: 10.1109/TCI.2020.2967148.
[21] W. Zhou, E. Zhou, G. Liu, L. Lin, and A. Lumsdaine, “Unsupervised monocular depth estimation from light field image,” IEEE Trans. Image Process., vol. 29, pp. 1606–1617, 2020, doi: 10.1109/TIP.2019.2944343.

[22] J. Jin and J. Hou, “Occlusion-aware unsupervised learning of depth from 4-D light fields,” IEEE Trans. Image Process., vol. 31, pp. 2216–2228, 2022, doi: 10.1109/TIP.2020.2967148.

[23] H.-G. Jeon et al., “Accurate depth map estimation from a lenslet light field camera,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1547–1555.

[24] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, “Robust depth estimation for light field via spinning parallelgram operator,” Comput. Vis. Image Understanding, vol. 145, pp. 148–159, 2016.

[25] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Occlusion-aware depth estimation using light-field cameras,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 3487–3495.

[26] 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. 3022–3035, Nov. 2022.

[27] W. William and I. K. Park, “Robust light field depth estimation for noisy scene with occlusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4396–4404.

[28] 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.

[29] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Depth estimation with occlusion modeling using light-field cameras,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2170–2181, Nov. 2016.

[30] I. K. Park et al., “Robust light field depth estimation using occlusion-noise aware data costs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 10, pp. 2484–2497, Oct. 2018.

[31] W. William and I. K. Park, “Robust light field depth estimation for noisy scene with occlusion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4396–4404.

[32] M. W. Tao, S. Hadap, J. Malik, and R. Ramamoorthi, “Depth from combining defocus and correspondance using light-field cameras,” in Proc. IEEE Conf. Comput. Vis., 2013, pp. 673–680.

[33] M. W. Tao, P. P. Srinivasan, S. Hadap, S. Rusinkiewicz, J. Malik, and R. Ramamoorthi, “Shape estimation from shading, defocus, and correspondance using light-field angular coherence,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 3, pp. 546–560, Mar. 2017.

[34] 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. 1009–1017.

[35] S. Wanner, S. Meister, and B. Goldlücke, “Datasets and benchmarks for densely sampled 4D light fields,” in Proc. Int. Symp. Vis. Model. Vis., 2013, pp. 225–226.

[36] R. C. Bolles, H. H. Baker, and D. H. Marrimont, “Epipolar-plane image analysis: An approach to determining structure from motion,” Int. J. Comput. Vis., vol. 1, no. 1, pp. 7–55, 1987.

[37] X. Gu, S. J. Gortler, and M. F. Cohen, “Polyhedral geometry and the two-plane parameterization,” in Proc. Eurographics Workshop Rendering Techn., Berlin, Heidelberg: Springer-Verlag, 1997, pp. 1–12.

[38] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldlücke, “A dataset and evaluation methodology for depth estimation on 4D light fields,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 19–34.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[40] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. 9th Int. Conf. Learn. Representations, May 3–7, 2021. [Online]. Available: https://openreview.net/forum?id=IcYfDUNT6

[41] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., Springer, 2020, pp. 213–229.

[42] Z. Li et al., “Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 6197–6206.

[43] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[44] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, “What matters in unsupervised optical flow,” in Proc. 16th Eur. Conf. Comput. Vis., Glasgow, U.K., Springer, 2020, pp. 557–572.

[45] A. Alperovich, O. Johannsen, M. Strecke, and B. Goldlücke, “Light field intrinsics with a deep encoder-decoder network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 9145–9154.

[46] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

[47] A. Sunder Raj, M. Lowney, R. Shah, and G. Wetzstein, “The (new) stanford light field archive,” 2016. [Online]. Available: http://lightfields.stanford.edu/LF2016.html

[48] N. K. Kalantari, T. C. Wang, and R. Ramamoorthi, “Learning-based view synthesis for light field cameras,” ACM Trans. Graph., vol. 35, no. 6, 2016, Art. no. 193.

[49] S. Sunder Raj and M. Lowney, R. Shah, and G. Wetzstein, “The (new) stanford light field archive,” 2016. [Online]. Available: http://lightfields.stanford.edu/LF2016.html

[50] M. Rerabek and T. Ebrahimi, “New light field image dataset,” in Proc. Int. Workshop Qual. Multimedia Experience, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:2409457

[51] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, “TensoRF: Tensorial radiance fields,” in Proc. 17th Eur. Conf. Comput. Vis., Tel Aviv, Israel, Springer, 2022, pp. 333–350.

[52] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” Commun. ACM, vol. 65, no. 1, pp. 99–106, 2021.

[53] M. Okutomi and T. Kanade, “A multiple-baseline stereo,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 1991, pp. 63–69.

Peng Li received the BS degree in communication engineering from Xidian University, China, in 2020, He is currently working toward the master’s degree with Shenzhen International Graduate School, Tsinghua University. His research interest includes depth estimation, 3D reconstruction, and neural rendering.

Jiayin Zhao received the BS degree in measurement and control from Beihang University, China, in 2020, He is currently working toward the master’s degree with Shenzhen International Graduate School, Tsinghua University. His research interest includes microscopy imaging and 3D reconstruction.

Chao Deng received the ME degree in signal and information processing from the High-Tech Institute of Xi’an, Xi’an, China, in 2014, and the PhD degree jointly from Tsinghua university and High-Tech Institute of Xi’an in 2019. He is currently an assistant researcher with the Department of Automation, Tsinghua University, Beijing, China. His research interests include computational imaging, computational vision, and vision navigation.
Yuqi Han received the BS degree in computer science and technology from Harbin Engineering University, Harbin, China, in 2015, and the PhD degree in computer science and technology from Tongji University, Shanghai, China, in 2021. She is a postdoctor with the Department of Automation, Tsinghua University. Her research interests include intelligent wireless network, mobile edge computing, and visual enhancement.

Haoqian Wang (Member, IEEE) received the PhD degree from the Harbin Institute of Technology, Harbin, in 2005. He was a post-doctoral fellow with Tsinghua University, Beijing, China, from 2005 to 2007. He has been a faculty member with the Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, since 2008, where he has also been an associate professor and a professor since 2011 and the director of the Shenzhen Institute of Future Media Technology. His current research interests include Stereo Vision and deep learning.

Tao Yu received the BS degree in measurement and control from the Hefei University of Technology, in 2012, and the PhD degree from Beihang University, in 2019. From 2020 to 2022, he worked as a postdoc with the Department of Automation, Tsinghua University, China. He is currently an assistant researcher of BNRist, Tsinghua University. His research area lies with the intersection between computer vision, deep learning, and computer graphics. His main research focus is digitizing humans using different kinds of sensors and algorithms.