Unsupervised Light Field Depth Estimation via Multi-View Feature Matching With Occlusion Prediction

Shansi Zhang, *Graduate Student Member, IEEE*, Nan Meng, *Member, IEEE*, and Edmund Y. Lam, *Fellow, IEEE*

***Abstract*— Depth estimation from light field (LF) images is a fundamental step for numerous applications. Recently, learning-based methods have achieved higher accuracy and efficiency than the traditional methods. However, it is costly to obtain sufficient depth labels for supervised training. In this paper, we propose an unsupervised framework to estimate depth from LF images. First, we design a disparity estimation network (DispNet) with a coarse-to-fine structure to predict disparity maps from different view combinations. It explicitly performs multi-view feature matching to learn the correspondences effectively. As occlusions may cause the violation of photo-consistency, we introduce an occlusion prediction network (OccNet) to predict the occlusion maps, which are used as the element-wise weights of photometric loss to solve the occlusion issue and assist the disparity learning. With the disparity maps estimated by multiple input combinations, we then propose a disparity fusion strategy based on the estimated errors with effective occlusion handling to obtain the final disparity map with higher accuracy. Experimental results demonstrate that our method achieves superior performance on both the dense and sparse LF images, and also shows better robustness and generalization on the real-world LF images compared to the other methods.**

***Index Terms*— Light field, unsupervised depth estimation, feature matching, occlusion prediction.**

I. Introduction

A LIGHT field (LF) camera can capture both the intensities and directions of the light rays [1], [2] to obtain LF images, each of which consists of an array of sub-aperture images (SAIs) to record the scene from multiple viewpoints [3], [4]. As the LF images contain rich geometric information, useful clues are available for depth (disparity) estimation. Usually, depth estimation is an essential task for scene understanding and also a crucial step for various LF applications and researches, such as auto refocusing [5], scene reconstruction [6], novel view synthesis [7], [8], compressed sensing [9], and semantic segmentation [10].

Traditional methods for LF depth estimation mainly adopt two approaches. The first approach is to explore the structure of epipolar-plane images (EPIs) [11], [12], [13], [14], where the pixels corresponding to the same scene point in different views form a line with a slope proportional to the disparity value. Another approach leverages the classical stereo matching to find the corresponding pixels among different views [15], [16], [17], [18]. However, EPI-based methods are mainly applicable to the densely sampled LF images, and the traditional stereo matching-based methods usually suffer from heavy computational costs. Recently, many learning-based methods [19], [20], [21], [22], [23], [24] have been proposed for LF depth estimation with improved accuracy and efficiency. They use a deep neural network to represent the estimator, which learns from the depth labels. However, it is costly to acquire sufficient, accurate depth annotations, especially for the real-world LF images, and the lack of training data usually leads to limited generalization ability. To alleviate the reliance on a large number of labeled data, some unsupervised learning-based methods [25], [26], [27], [28], [29] are developed, which implicitly learn the correspondences by a plain network with the photo-consistency constraint [30] to minimize the warping errors. However, these methods are mainly applicable to the dense LF images but are not effective enough for the sparse LF images with large disparity values.

Our target is to develop an unsupervised LF depth framework applicable to both the dense and sparse LF images. We first design a disparity estimation network (DispNet) that explicitly performs multi-view feature matching by constructing memory-efficient cost volumes to learn the correspondences among the input views with a variety of disparity ranges. Our DispNet leverages a coarse-to-fine structure, with a coarse branch to estimate a initial disparity map and a refinement branch to further improve the disparity accuracy. Moreover, occlusion is a challenging issue in many LF tasks [8], [31], [32], and it leads to the violation of photo-consistency in LF depth estimation. To tackle this issue, we introduce an occlusion prediction network (OccNet) during training to predict the occlusion maps, which are used as the element-wise weights of the photometric loss to mitigate the effect of occlusions on the disparity learning. In order

Manuscript received 21 March 2023; revised 2 July 2023; accepted 13 August 2023. Date of publication 17 August 2023; date of current version 5 April 2024. This work was supported in part by the Research Grants Council of Hong Kong under Grant GRF 17201822; and in part by ACCESS—Artificial Intelligence (AI) Chip Center for Emerging Smart Systems, Hong Kong SAR. This article was recommended by Associate Editor S. Berretti. *(Corresponding author: Edmund Y. Lam.)*

Shansi Zhang and Edmund Y. Lam are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, SAR, China (e-mail: sszhang@eee.hku.hk; elam@eee.hku.hk).

Nan Meng is with the Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, SAR, China (e-mail: nanmeng@hku.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3305978.

Digital Object Identifier 10.1109/TCSVT.2023.3305978

to fully utilize the views of each LF image, multiple view combinations are input to the DispNet to obtain multiple estimated disparity maps. We then propose a disparity fusion strategy with effective occlusion handling to obtain the final disparity map with higher accuracy. The main contributions of our work are summarized as follows:

- We develop a DispNet, which employs a coarse-to-fine structure and performs multi-view feature matching to estimate disparity maps from different input combinations.
- To tackle the occlusion issue, we introduce an OccNet for occlusion prediction, which aims to eliminate the adverse impact of occlusion on the training of DispNet.
- With the multiple estimated disparity maps, we propose a disparity fusion strategy with occlusion handling to obtain the final disparity map.
- Experimental results demonstrate that our method achieves superior performance on both the dense and sparse LF images with better robustness and generalization compared to the other methods.

II. Related Work

The existing methods for LF depth estimation are reviewed in terms of the traditional methods and the learning-based methods.

A. Traditional Methods

Traditional methods for LF depth estimation mainly focus on the EPI structure and stereo matching. For the EPI-based methods, Wanner and Goldluecke [11] estimated the disparity maps locally by using EPI analysis, which works fast without the expensive matching cost minimization. Zhang et al. [12] proposed a spinning parallelogram operator (SPO) to locate the lines in EPI and calculate their slopes to acquire the depth information. Zhang et al. [13] exploited the line structure of EPI and the locally linear embedding (LLE) to estimate the local depth by minimizing the matching cost. Sheng et al. [14] developed a strategy to extract EPIs in multiple directions besides the horizontal and vertical EPIs to calculate the local depth, which is combined with the predicted occlusion boundaries to obtain the final depth map.

For the stereo matching-based methods, Jeon et al. [15] constructed a cost volume to estimate the multi-view stereo correspondences, which was used to optimize the depths in weak texture regions. Then, the local depth map was refined iteratively by fitting the local quadratic function. Tao et al. [16] leveraged the shading information to improve the local shape estimation from defocus and correspondence, and developed a framework that exploits LF angular coherence for depth and shading optimization. Lee et al. [17] computed binary maps through foreground–background separation to obtain the disparity maps. Huang et al. [18] proposed a stereo matching algorithm using an empirical Bayesian framework, which employs the pseudo-random field to explore the statistical cues of LF. Zhang et al. [33] leveraged graph spectral analysis to exploit the angular and spatial structure information for depth estimation.

Occlusion issue is often encountered in LF depth estimation since the photo-consistency assumption does not hold in the occlusion regions. There are some methods focusing on tackling this issue. Wang et al. [34] found that the photo-consistency still holds in about half of the views when the occlusions exist, and they predicted the occlusions, which was used as a regularizer to improve depth estimation. Williem et al. [35] focused on the robust depth estimation from noisy LF with occlusions. They introduced two data costs with angular entropy metric and adaptive defocus response to handle the occlusions and noises. Chen et al. [36] proposed a method with partially occluded region detection through super-pixel regularization and showed that even a simple least square model can achieve superior depth estimation after manipulating the label confidence and edge strength.

These traditional methods usually involve complex optimization process and long execution time, and cannot achieve a good balance between the accuracy and efficiency.

B. Learning-Based Methods

The recent work on LF depth estimation mainly focuses on the learning-based methods, which usually achieves higher accuracy and inference efficiency. Most existing learning-based methods adopt supervised training using depth labels. Sun et al. [19] developed a convolutional neural network (CNN) to estimate LF disparity by extracting enhanced EPI features. Heber et al. [20] proposed a U-shaped network to extract LF geometric information, with 3D convolutional layers to examine the EPI volumes for robust depth prediction. Shin et al. [21] developed a CNN framework by taking as input the views from different angular dimensions. They also proposed some data augmentation methods for LF to overcome the deficiency of training data. Shi et al. [22] proposed a framework to learn depth from the dense and sparse LF images with three steps, including initial depth estimation by a fine-tuned network, occlusion-aware depth fusion and refinement by an additional network. Tsai et al. [23] proposed a view selection network by learning an attention map to estimate the contribution of each view on depth. The attention map was constrained to be symmetric in accordance with the LF views. Chen et al. [24] developed a multi-level fusion network, which contains four branches to perform intra-branch and inter-branch fusion, and incorporates attention to select the features that can provide more useful information for depth. Wang et al. [37] proposed a fast approach to construct matching cost for LF depth estimation, which does not require any shifting operation and can also handle the occlusions. These supervised methods heavily rely on the labeled data, which results in poor generalization ability when the labeled data are not sufficient.

Unsupervised methods can overcome the reliance on the labeled data. Peng et al. [25] proposed an unsupervised CNN framework by designing a combined loss with compliance and divergence constraints to estimate LF disparity. Zhou et al. [26] developed an unsupervised monocular LF depth network, which was trained by the improved photometric losses and takes only one view as input. Jin et al. [27] proposed an unsupervised occlusion-aware framework by exploring the angular coherence among different LF subsets. Iwatsuki et al. [28] developed an unsupervised learning framework with pixel-wise weights to evaluate the

warping errors and an edge loss to enforce edge alignment between the image and the disparity map. Lin et al. [29] proposed to integrate the traditional LF constraints into an unsupervised framework with an adaptive spatial-angular consistency loss. These methods directly output the disparity values from the last convolution layer without explicitly learning the correspondences among different views, which leads to poor performance when applied to the sparse LF images with large disparity values.

III. Proposed Method

We first describe our overall framework for LF disparity estimation in Sec. III-A. Then, we introduce the architecture of our DispNet in Sec. III-B, the occlusion prediction method in Sec. III-C, and the loss functions for training in Sec. III-D. Finally, we introduce our disparity fusion strategy in Sec. III-E.

A. Overall Framework

A 4D LF image is represented as $\mathbf{L} \in \mathbb{R}^{U\times V\times X\times Y}$, with angular index (u, v) and spatial index (x, y). It consists of $U \times V$ SAIs, each of which records the scene from one viewpoint with a spatial resolution of $X \times Y$. Our target is to estimate the disparity map of the central view relative to its adjacent views. According to the LF geometry, the relationship between the central view and any other view in terms of the central disparity $\mathbf{d}(x, y)$ is expressed as

$$\mathbf{L}(u_c, v_c, x, y) = \mathbf{L}\big(u, v, x + \mathbf{d}(x, y) \times (u_c - u), y + \mathbf{d}(x, y) \times (v_c - v)\big), \quad (1)$$

where (u_c, v_c) is the angular position of the central view.

The overall framework of our method is depicted in Fig. 1. Three views, including the central view $\mathbf{I}_c$, the left source view $\mathbf{I}_l$ and the right source $\mathbf{I}_r$ view ("source" means that they are warped according to the disparity to reconstruct the central view), are fed to the DispNet. They are in the same row or column, and the two source views framed by the same color are located symmetrically to the central view. This input strategy can avoid the matching ambiguity caused by the occlusion without incorporating any redundant inputs, since each scene point is usually visible in at least two views of the three input views. Fig. 2 gives an intuitive illustration. The left view and right view are warped to the central view using the ground-truth disparity map of the central view, and their warping error maps are presented. The bright regions with large errors in the warping error maps correspond to the occlusion regions, which are visible in the central view but invisible in the left or right views. It can be seen that the occlusion regions for the left and right views usually near the object boundaries and locate in the opposite positions, which indicates that the pixel in the central view has correspondence in at least one view of the left and right views to enable accurate disparity estimation.

For a 7×7 LF image (Fig. 1(a)), there are totally 6 input combinations. The views from the same column should be rotated by 90° (counterclockwise) before being input to the network in order to convert the vertical disparity to be horizontal, and the corresponding output disparity map needs to be rotated by −90° (clockwise) to recover the orientation. In this way, only horizontal disparity estimation is involved to make the learning easier. Multiple disparity maps can be obtained from different input combinations by the shared DispNet, and they need to be scaled by the distance between the source views and the central view, as shown in Fig. 1(b). Using nonadjacent views to estimate disparity helps to alleviate the inaccurate estimation caused by the narrow LF baseline. Then, these disparity maps are fused based on their estimated errors using the auxiliary views (in the diagonal direction) to obtain the final disparity map (Fig. 1(c)). In what follows, we will introduce each part in detail.

B. Disparity Estimation

The architecture of our DispNet is shown in Fig. 3. The central view $\mathbf{I}_c$ and the two source views $\mathbf{I}_l$ and $\mathbf{I}_r$ are fed to a shared feature extractor. The feature extractor (Fig. 3(b)) consists of several residual blocks, each of which has two 3×3 convolution layers with leaky Rectified Linear Unit (ReLU) activation, and an atrous spatial pyramid pooling (ASPP) [38] block to encode multi-scale features, which contains three dilated convolutions with rates 3, 6 and 8 and a global average pooling (GAP) operation for global receptive field.

The extracted features of the three views are used to construct the cost volumes through variance-based feature matching. The detailed procedures of constructing the coarse cost volume is illustrated in Fig. 4. First, we prescribe D equally spaced disparity samples with the minimum value $s_{\min}$ and the maximum value $s_{\max}$, expressed by a vector $\mathbf{s} = [s_{\min}, \cdots, s_i, \cdots, s_{\max}]^T$. Given the features of the three views with a size $C \times X \times Y$ (C is the channel number), the left and right source features are warped to match the reference (central) feature by a disparity sample s_i from $\mathbf{s}$. Then, the element-wise variance of the warped features and reference feature is calculated to measure the difference, which is treated as the matching cost at disparity sample s_i (accurate disparity at a pixel should lead to a small variance). The cost volume is obtained by concatenating the matching costs at all the disparity samples, with a size $C \times D \times X \times Y$. Compared to the common approach used in stereo matching [39], which builds the cost volume by concatenating the reference feature and warped source features, our variance-based feature matching is much more memory-efficient and can also adapt to any number of input views without increasing the size of cost volume.

The coarse cost volume is further processed by the cost filters (Fig. 3(c)), each of which consists of several 3D residual blocks with skip connections. The blocks with stride $= 2$ aim to downsample the cost volume to reduce memory consumption. Then, a coarse disparity regression module (Fig. 3(d)) is employed, with 3D convolution layers to yield the coarse cost $\mathbf{c}_{\text{coa}} \in \mathbb{R}^{D\times X\times Y}$, and the coarse disparity map $\tilde{\mathbf{d}}_{\text{coa}}$ is obtained by

$$\tilde{\mathbf{d}}_{\text{coa}}(x, y) = \mathbf{s} \cdot \text{softmax}\big(\mathbf{c}_{\text{coa}}(x, y)\big), \quad (2)$$

where softmax$(\cdot)$ is the softmax function to obtain the probability of each disparity sample, and $\cdot$ denotes the inner product.

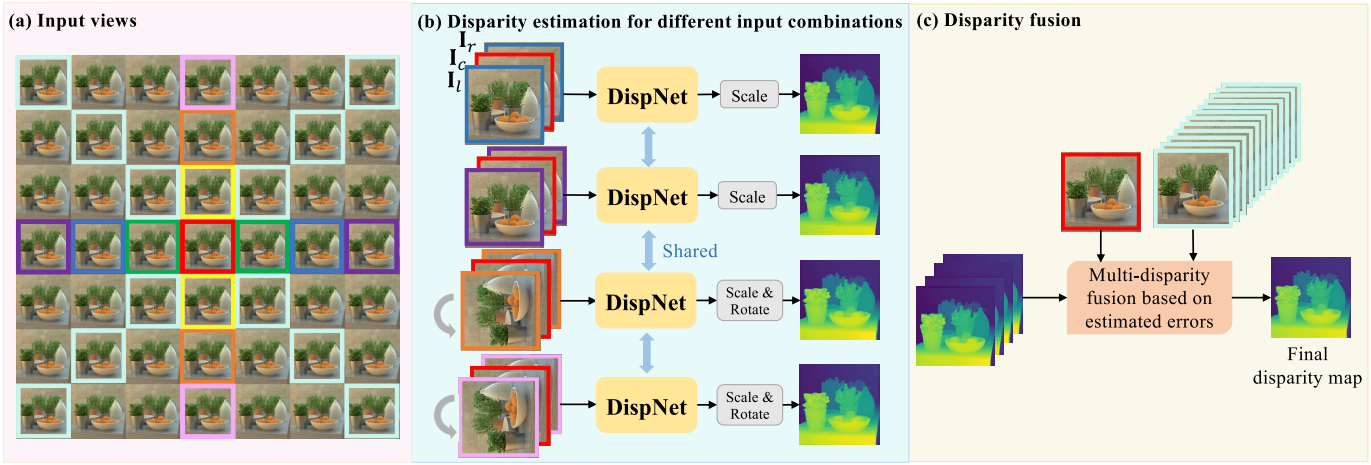


Fig. 1. (a) The views input to the DispNet and used for determining the errors during fusion. (b) The DispNet takes as inputs the central view $\mathbf{I}_c$, the left source view $\mathbf{I}_l$ and the right source view $\mathbf{I}_r$. The views from the same column are rotated by 90°. Multiple disparity maps (after scaling and rotation) are obtained from different input combinations. (c) The estimated disparity maps are fused according to their estimated errors using the auxiliary views to obtain the final disparity map.

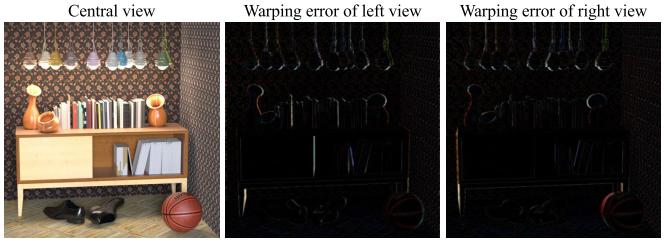


Fig. 2. The central view and the warping error maps of the left and right views. The occlusion regions for the left and right views usually near the object boundaries and locate in the opposite positions.

To further refine the disparity map, a residual cost volume is constructed by using the same extracted features. Similarly, we set a residual sample vector $\mathbf{s}' = [s'_{\min}, \cdots, s'_i, \cdots, s'_{\max}]^T$ with the minimum value $s'_{\min}$ and the maximum value $s'_{\max}$. The left and right source features are warped according to the coarse disparity map plus a residual sample s'_i from $\mathbf{s}'$, with $\tilde{\mathbf{d}}_{\text{coa}}(x, y) + s'_i$ at each position, to match the reference feature. The value and interval of residual sampling are much smaller than that of coarse sampling in order to improve the disparity accuracy. The matching cost at each residual sample is derived by calculating the variance of the warped and reference features, and the residual cost volume is obtained by concatenating the matching costs at all the residual samples. Then, the residual cost volume is processed by the shared cost filters and the residual disparity regression module to derive the residual map $\tilde{\mathbf{d}}_{\text{res}}$, with

$$\tilde{\mathbf{d}}_{\text{res}}(x, y) = \mathbf{s}' \cdot \text{softmax}\big(\mathbf{c}_{\text{res}}(x, y)\big), \tag{3}$$

where $\mathbf{c}_{\text{res}} \in \mathbb{R}^{D \times X \times Y}$ is the residual cost.

Thus, the refined disparity map $\tilde{\mathbf{d}}$ is obtained by

$$\tilde{\mathbf{d}} = \tilde{\mathbf{d}}_{\text{coa}} + \tilde{\mathbf{d}}_{\text{res}}. \tag{4}$$

When the input views are not adjacent, the output disparity maps from DispNet need to be scaled according to the view distance. In addition, the disparity maps predicted by the views from the same column need to be rotated by $-90°$ to recover the orientation. Thus, the estimated central disparity map $\hat{\mathbf{d}}$ is obtained by

$$\hat{\mathbf{d}} = \begin{cases} \dfrac{\tilde{\mathbf{d}}}{u_c - u_l}, & \text{from the same row} \\ \text{rot}_{-90°}\left(\dfrac{\tilde{\mathbf{d}}}{v_c - v_l}\right), & \text{from the same column} \end{cases} \tag{5}$$

where u_l and v_l are the angular coordinates of the left source view, and $\text{rot}_{-90°}$ means rotating by $-90°$.

C. Occlusion Prediction

During training, $\tilde{\mathbf{d}}$ is used to warp the left and right source views to the central view, yielding $\mathbf{I}_{l\to c}$ and $\mathbf{I}_{r\to c}$, with

$$\mathbf{I}_{l\to c}(x, y) = \mathbf{I}_l\big(x + \tilde{\mathbf{d}}(x, y), y\big), \tag{6}$$

$$\mathbf{I}_{r\to c}(x, y) = \mathbf{I}_r\big(x - \tilde{\mathbf{d}}(x, y), y\big). \tag{7}$$

To predict the occlusion regions for the two source views, we introduce an OccNet that takes the concatenation of $\mathbf{I}_{l\to c}$, $\mathbf{I}_{r\to c}$ and $\tilde{\mathbf{d}}$ as inputs, as shown in Fig. 5. The OccNet adopts a U-shape structure with residual blocks and skip connections. A softmax activation is used in the last convolution layer to output the confidence maps, $\mathbf{O}_l$ and $\mathbf{O}_r$ for $\mathbf{I}_{l\to c}$ and $\mathbf{I}_{r\to c}$, respectively, which are used to reconstruct the central view, with,

$$\mathbf{I}_{\text{rec}} = \mathbf{O}_l \odot \mathbf{I}_{l\to c} + \mathbf{O}_r \odot \mathbf{I}_{r\to c}, \tag{8}$$

where $\mathbf{I}_{\text{rec}}$ is the reconstructed central view, $\odot$ represents the element-wise multiplication, and $\mathbf{O}_l(x, y) + \mathbf{O}_r(x, y) = 1$. The OccNet is trained by the reconstruction loss, with

$$\ell_{\text{rec}} = \|\mathbf{I}_{\text{rec}} - \mathbf{I}_c\|_1, \tag{9}$$

where $\|\cdot\|_1$ is the ℓ_1-norm operator.

The photometric loss for unsupervised disparity training is based on the photo-consistency assumption, which does not hold in the occlusion regions. Therefore, the pixel-wise weights of the photometric loss in the occlusion regions are expected to be small. The confidence maps $\mathbf{O}_l$ and $\mathbf{O}_r$ can be treated as the occlusion maps of the two warped views, since the occlusion regions usually lead to relatively larger warping errors, and therefore less confidence for the reconstruction.

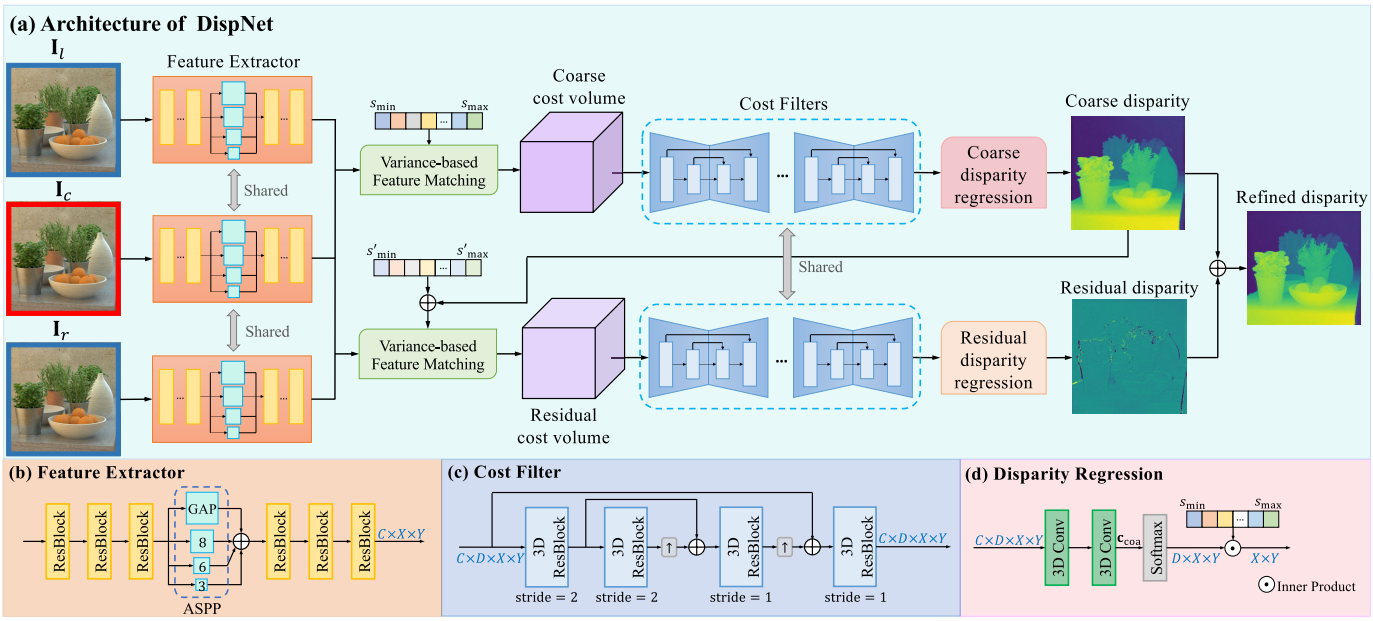


Fig. 3. (a) Architecture of DispNet. It consists of two branches with shared feature extractor and cost filters to estimate the coarse disparity map and residual map by constructing the coarse and residual cost volumes with variance-based feature matching. (b) Feature extractor. (c) Cost filter. (d) Disparity regression.

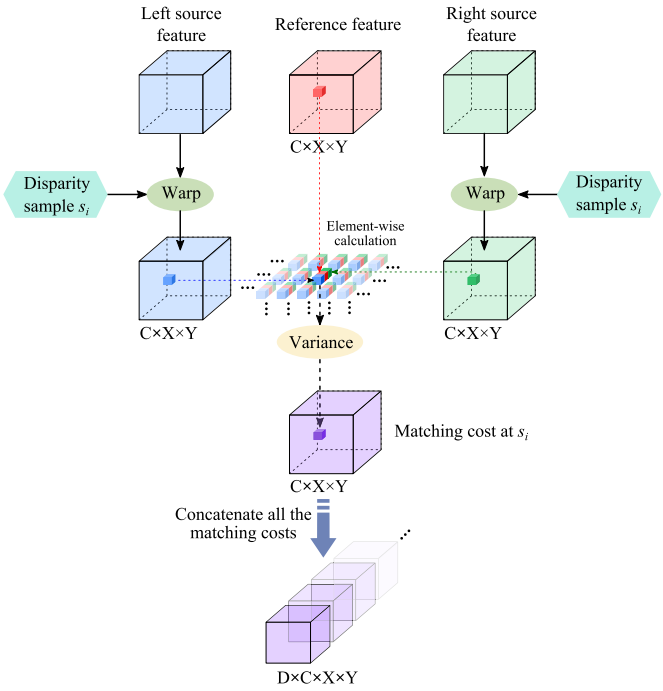


Fig. 4. Variance-based feature matching. The left and right source features are warped according to a disparity sample s_i to match the reference feature. The element-wise variance of the warped features and reference feature is calculated to obtain the matching cost at the disparity sample s_i. The final cost volume is obtained by concatenating the matching costs at all the disparity samples.

Thus, we propose the weighted photometric loss with the occlusion maps, expressed as

$$\ell_{\text{wpm}} = \frac{1}{XY}\sum_{x,y}\mathbf{O}_l^{(x,y)} \odot |\mathbf{I}_{l\to c}^{(x,y)} - \mathbf{I}_c^{(x,y)}| + \frac{1}{XY}\sum_{x,y}\mathbf{O}_r^{(x,y)} \odot |\mathbf{I}_{r\to c}^{(x,y)} - \mathbf{I}_c^{(x,y)}|. \quad (10)$$

The OccNet plays an important role during the training of DispNet. First, it helps to enforce the similarity between the warped source views and the central view for disparity learning through the reconstruction loss, as it is jointly trained with the DispNet. Second, with the improvement of estimated disparity map, large warping errors mainly lie in the occlusion regions, which can be identified by the OccNet to alleviate the adverse impact on further improvement. Note that the OccNet is only employed during training to address the occlusion issue and assist the disparity learning without influencing the inference efficiency.

D. Loss Function

In addition to the weighted photometric loss ℓ_{wpm} and the reconstruction loss ℓ_{rec}, we apply a structural similarity (SSIM) loss [40] to further enforce the similarity, with

$$\ell_{\text{SSIM}} = 1 - \frac{\text{SSIM}(\mathbf{I}_{l\to c}, \mathbf{I}_c) + \text{SSIM}(\mathbf{I}_{r\to c}, \mathbf{I}_c)}{2}. \quad (11)$$

To improve the smoothness of the estimated disparity map while preserving the boundary structures of the objects, we leverage the structure-aware smoothness loss [41], [42], expressed as

$$\ell_{\text{smd}} = \frac{1}{XY}\sum_{x,y}|\nabla\tilde{\mathbf{d}}^{(x,y)}| \odot \exp\left(-\eta|\nabla\mathbf{I}_c^{(x,y)}|\right), \quad (12)$$

where ∇ denotes the gradients along both the horizontal and vertical directions, and η is a hyperparameter for structure preservation.

Moreover, we apply a similar smoothness loss to the occlusion map, with

$$\ell_{\text{smo}} = \frac{1}{XY}\sum_{x,y}|\nabla\mathbf{O}_l^{(x,y)}| \odot \exp\left(-\eta|\nabla\mathbf{I}_c^{(x,y)}|\right), \quad (13)$$

which is only applied to $\mathbf{O}_l$ since $\mathbf{O}_r = \mathbf{1} - \mathbf{O}_l$.

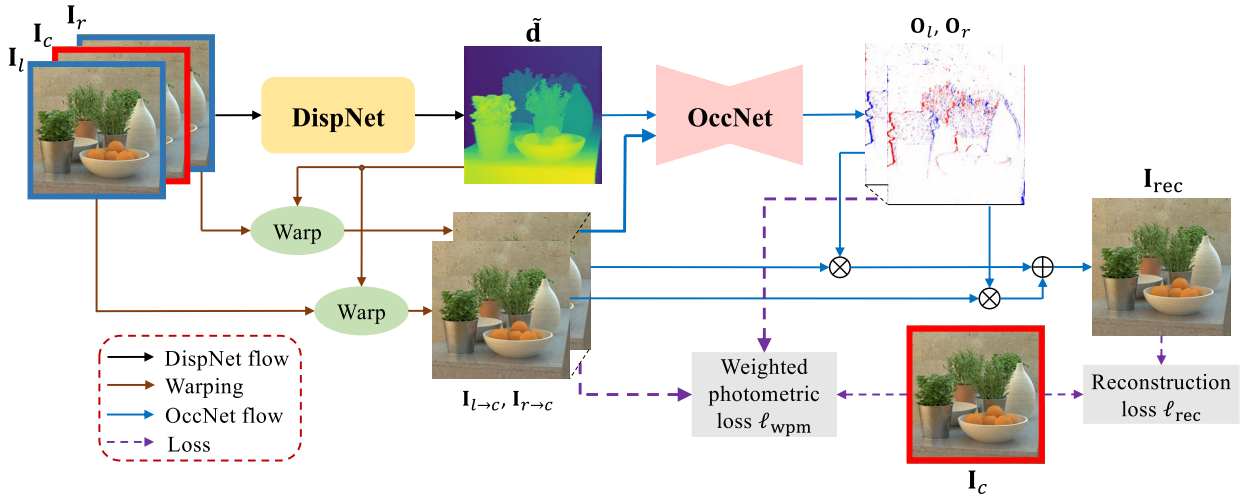


Fig. 5. Training with predicted occlusion maps. The estimated disparity map from the DispNet is used to warp the two source views to the central view. Then, the concatenation of the warped views and the disparity map is input to the OccNet to predict the confidence maps for reconstructing the central view, which are treated as occlusion maps and used as the pixel-wise weights of the photometric loss ℓ_{wpm}. The OccNet is trained by the reconstruction loss ℓ_{rec}. The arrows for DispNet flow, warping, OccNet flow and loss are distinguished by different colors.

The DispNet and OccNet are trained simultaneously by the following full loss,

$$\ell_{\text{full}} = \ell_{\text{wpm}} + \ell_{\text{rec}} + \alpha_1 \ell_{\text{SSIM}} + \alpha_2 \ell_{\text{smd}} + \alpha_3 \ell_{\text{smo}}, \tag{14}$$

where ℓ_{wpm} and ℓ_{rec} are the necessary losses, and the others are the auxiliary losses with coefficients $\alpha_1 \sim \alpha_3$. Note that these loss terms are also applied to the coarse disparity map $\tilde{\mathbf{d}}_{\text{coa}}$, but we omit the procedures for simplicity.

E. Multi-Disparity Fusion Based on Estimated Errors

Multiple disparity maps can be obtained by the DispNet with different input combinations. To obtain the final disparity map, we propose a disparity fusion strategy to merge these disparity maps.

Suppose that we have n estimated disparity maps $\{\hat{\mathbf{d}}_j\}_{j=1}^{n}$ and Z auxiliary views to evaluate their accuracy. With each disparity map, the auxiliary views are warped to the central view, and the warping errors are calculated. However, some of the warping errors are not accurate due to the occlusions. If occlusions exist in some of the auxiliary views, the warping errors at the corresponding positions would be large, resulting in large variances among the Z warping errors. Thus, we calculate the standard deviation of the warping errors to judge if occlusion exists at each position, and obtain a binary mask. Here, we define $\boldsymbol{\epsilon}_j \in \mathbb{R}^{X\times Y\times Z}$, indexed by (x, y, z), as the warping error maps of the Z auxiliary views using the disparity map $\hat{\mathbf{d}}_j$, and $\sigma_z(\cdot)$ as the standard deviation along Z dimension. Then, a binary mask $\mathbf{M}_j$ is formulated as

$$\mathbf{M}_j^{(x,y)} = \begin{cases} 1, & \sigma_z(\boldsymbol{\epsilon}_j^{(x,y,z)}) > \theta(q) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

where $\theta(\cdot)$ is to determine the threshold for occlusion using the quantile q of $\sigma_z(\boldsymbol{\epsilon}_j^{(x,y,z)})$. The effect of q is analyzed in Sec. IV-D.4.

For a pixel with occlusion, we use the median value of the warping errors to represent its estimated error, which can eliminate the influence of large warping errors due to occlusion. Otherwise, the estimated error is represented by the mean of all the warping errors. Then, the estimated error map $\mathbf{e}_j$ for $\hat{\mathbf{d}}_j$ is obtained by

$$\mathbf{e}_j^{(x,y)} = \text{median}_z\big(\boldsymbol{\epsilon}_j^{(x,y,z)}\big) \odot \mathbf{M}_j + \text{mean}_z\big(\boldsymbol{\epsilon}_j^{(x,y,z)}\big) \odot (\mathbf{1} - \mathbf{M}_j), \tag{16}$$

where $\text{median}_z(\cdot)$ and $\text{mean}_z(\cdot)$ denote the median and mean along Z dimension.

With the above procedures, we can derive the error maps $\{\mathbf{e}_j\}_{j=1}^{n}$ for all the disparity maps. Next, we need to merge these disparity maps according to their estimated errors. Here, we consider several different fusion approaches. The first one is the minimum error fusion by choosing the pixels with the minimum errors from each disparity map, and the final disparity map $\hat{\mathbf{d}}_{\text{final}}$ is derived by

$$j' = \arg\min_j \big(\mathbf{e}_j(x, y)\big), \tag{17}$$

$$\hat{\mathbf{d}}_{\text{final}}(x, y) = \hat{\mathbf{d}}_{j'}(x, y). \tag{18}$$

The second approach is the weighted fusion, where the weights are obtained by the softmax function and negatively correlated with the errors. Moreover, we can choose different numbers of disparity pixels for weighted fusion at each position. If n' $(n' \le n)$ disparity pixels with the smallest errors are used at each position, the final disparity map is obtained by

$$\mathbf{W}(x, y) = \text{softmax}\big(-\mathbf{E}(x, y)\big), \tag{19}$$

$$\hat{\mathbf{d}}_{\text{final}}(x, y) = \sum_{j=1}^{n'} \mathbf{w}_j(x, y) \times \hat{\mathbf{d}}_j(x, y), \tag{20}$$

where $\mathbf{E}(x, y) = \{\mathbf{e}_j(x, y)\}_{j=1}^{n'}$ denotes the smallest n' errors at position (x, y), and $\mathbf{W}(x, y) = \{\mathbf{w}_j(x, y)\}_{j=1}^{n'}$ denotes their corresponding weights.

IV. Experiments

A. Datasets

To evaluate the model performance comprehensively, we used both the synthetic and real-world LF datasets. The

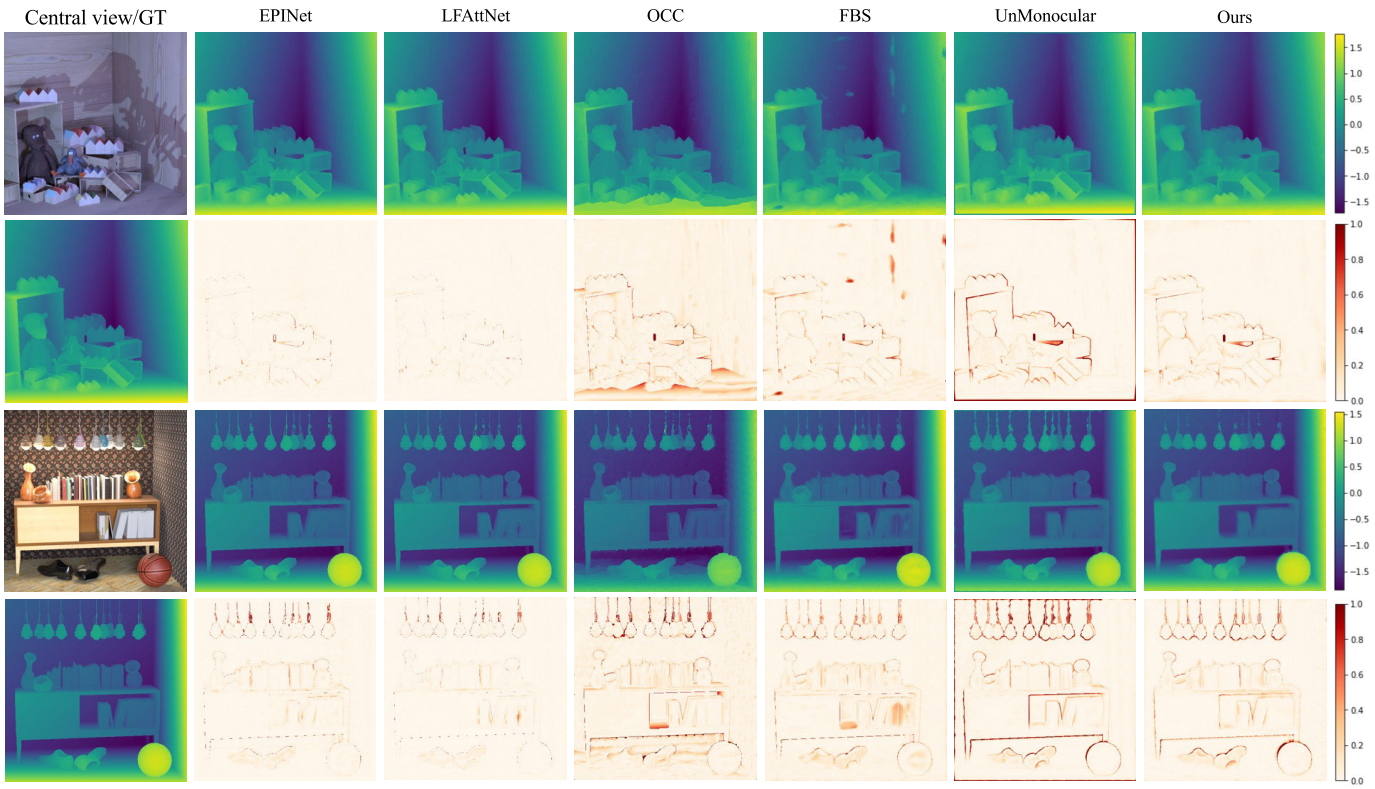


Fig. 6. Visual results of different methods on the scenes from the HCI dataset [43], with the error maps relative to the ground truths.

TABLE I

EVALUATION ON THE SCENES FROM HCI DATASET. **BOLD**: BEST AMONG THE UNSUPERVISED METHODS

Methods	Dino				Sideboard				Backgammon				Pyramids			
	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
Supervised																
EPINet [21]	0.167	1.286	3.452	22.401	0.742	4.277	10.824	37.999	3.629	3.580	6.289	20.899	0.008	0.192	0.913	11.876
LFAttNet [23]	0.093	0.848	2.340	12.224	0.531	2.870	7.243	20.739	3.648	3.126	3.984	11.582	0.004	0.195	0.489	2.063
Non-learning																
OCC [34]	0.944	15.366	50.167	88.810	2.073	17.910	50.550	84.653	22.782	13.522	44.899	91.402	0.077	1.450	25.574	92.860
FBS [17]	0.664	8.427	23.533	65.390	1.072	13.296	32.516	70.042	5.805	10.162	22.181	65.407	0.029	0.549	5.705	78.243
Unsupervised																
UnCNN [25]	1.807	23.660	47.876	78.724	3.149	26.173	45.384	82.924	11.034	31.783	65.583	87.987	0.191	10.849	43.972	79.113
UnMonocular [26]	1.031	**5.402**	**14.757**	**43.258**	2.770	**10.947**	**23.646**	61.406	11.833	12.311	28.524	68.312	0.027	**0.262**	8.725	35.594
UnPlug [28]	0.788	6.178	-	-	1.999	12.766	-	-	9.399	14.200	-	-	**0.022**	0.658	-	-
UnOcc [27]	**0.63**	8.25	-	-	1.79	14.20	-	-	6.684	14.371	-	-	0.213	7.348	-	-
Ours	0.650	6.586	16.722	46.380	**1.738**	12.013	25.848	**58.744**	**5.740**	**10.710**	**18.452**	**51.066**	0.023	0.670	**4.720**	**24.314**

TABLE II

EVALUATION ON THE SCENES FROM DLF DATASET. **BOLD**: BEST AMONG THE UNSUPERVISED METHODS

Methods	Toys				Antiques				Pinenuts white				Smiling crowd roses			
	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
Supervised																
EPINet [21]	0.431	15.540	42.438	75.800	1.265	6.992	32.292	72.562	0.509	15.232	35.826	69.557	3.148	14.997	41.297	77.024
LFAttNet [23]	0.405	10.231	35.711	74.166	0.827	4.206	21.862	67.076	0.406	10.698	27.042	65.995	2.025	12.454	33.356	66.702
Unsupervised																
UnCNN [25]	0.960	20.031	53.104	82.405	4.876	21.944	56.391	84.684	2.099	35.955	65.235	89.028	5.143	32.653	62.062	86.544
UnMonocular [26]	0.859	8.783	47.252	82.033	4.551	7.073	21.986	63.595	0.803	14.146	39.468	74.769	6.257	13.522	28.691	73.238
UnCon [29]	0.886	12.238	57.656	84.407	3.322	10.859	43.703	80.262	0.540	14.402	54.013	84.173	3.916	14.595	31.483	72.385
UnOcc [27]	1.021	18.497	44.714	76.093	4.034	7.018	22.748	61.508	0.706	21.065	49.413	79.825	3.678	15.403	32.678	71.964
Ours	**0.643**	**6.745**	**24.210**	**60.513**	**2.336**	**5.514**	**12.937**	**42.088**	**0.434**	**8.106**	**25.164**	**61.817**	**3.243**	**11.325**	**23.254**	**50.507**

synthetic LF datasets include both the densely and sparsely sampled LF images. The synthetic dense LF images are from the HCI dataset [43] and the DLF dataset [22], with a disparity range $[-4, 4]$ pixels between the adjacent views. The synthetic sparse LF images are from the SLF dataset [22], with a much larger disparity range $[-20, 20]$ pixels. The real-world LF images are from the Stanford Lytro LF Archive [44], Kalantari [45] and EPFL LF [46] datasets. They can also be

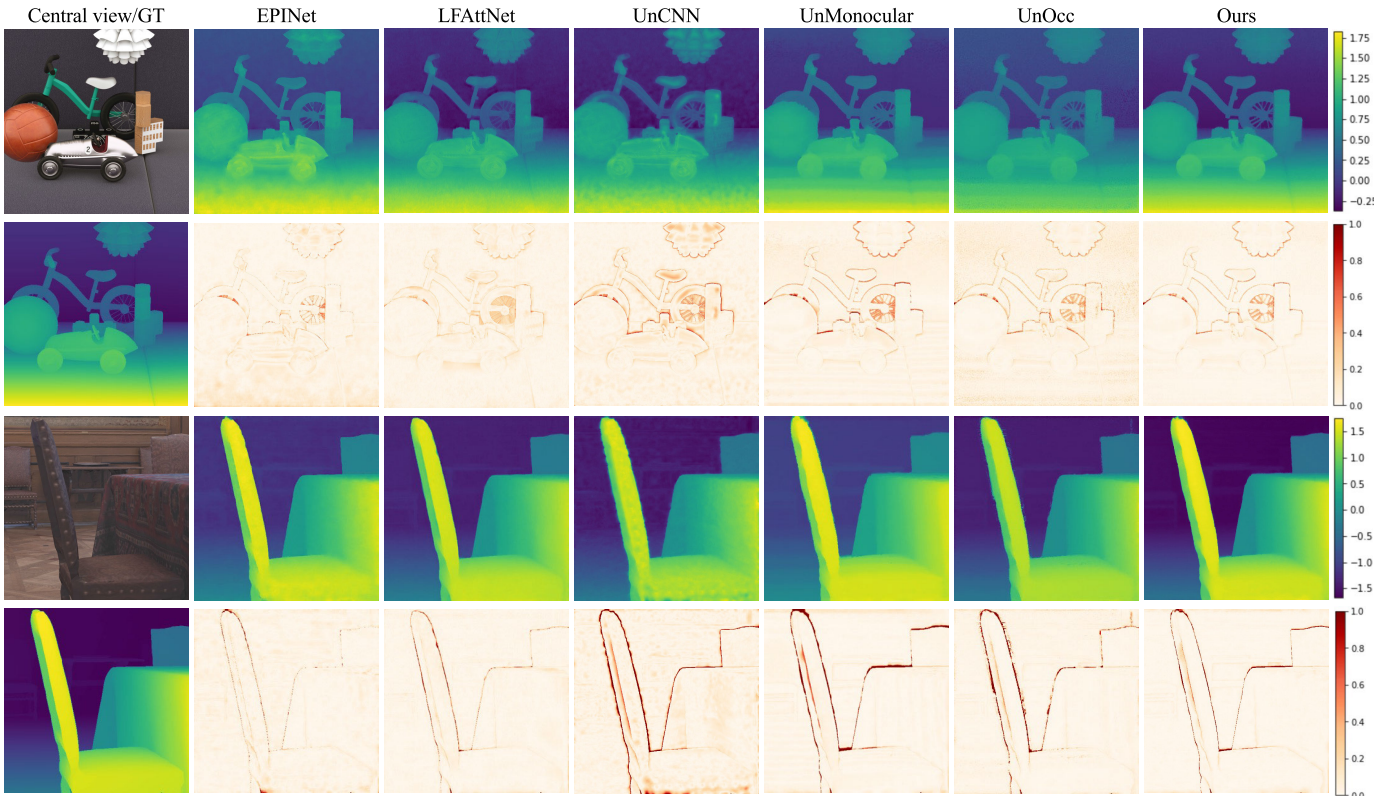


Fig. 7. Visual results of different methods on the scenes from the DLF dataset [22], with the error maps relative to the ground truths.

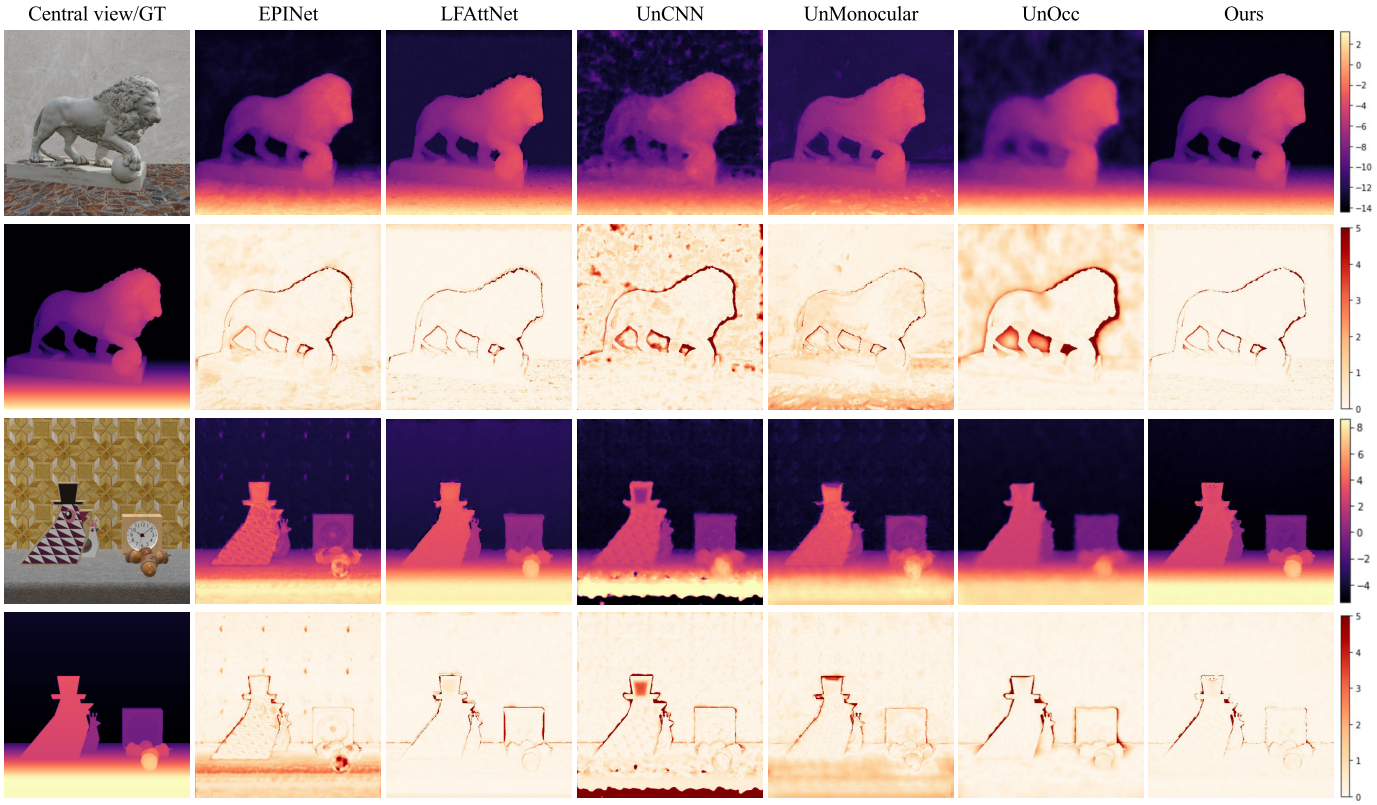


Fig. 8. Visual results of different methods on the sparse LF images from the SLF dataset [22], with the error maps relative to the ground truths.

treated as the dense LFs since the disparity values between the adjacent views are very small. For all the LF images, the central 7×7 SAIs were used to estimate the disparity maps for the central views.

B. Implementation Details

Regarding the DispNet, the maximum channel number within the feature extractor is 128, and the number of cost

TABLE III
EVALUATION ON THE SCENES FROM SLF DATASET. **BOLD**: BEST AMONG THE UNSUPERVISED METHODS

Methods	Lion				Rooster clock				Toy bricks				Electro devices			
	MSE↓	BPR↓ (0.3)	BPR↓ (0.1)	BPR↓ (0.05)	MSE↓	BPR↓ (0.3)	BPR↓ (0.1)	BPR↓ (0.05)	MSE↓	BPR↓ (0.3)	BPR↓ (0.1)	BPR↓ (0.05)	MSE↓	BPR↓ (0.3)	BPR↓ (0.1)	BPR↓ (0.05)
Supervised																
EPINet [21]	0.476	31.457	68.138	82.954	0.740	36.868	71.562	85.147	1.057	34.658	61.957	77.491	1.112	39.809	64.455	81.793
LFAttNet [23]	0.372	12.994	29.594	49.327	0.278	6.831	25.813	50.920	0.738	11.607	29.018	52.576	0.703	13.819	38.731	60.993
Unsupervised																
UnCNN [25]	1.442	58.221	84.867	92.373	9.213	34.586	75.021	87.383	2.254	53.922	82.172	90.950	2.181	41.163	72.970	85.866
UnMonocular [26]	0.760	48.305	87.371	94.673	0.664	39.931	73.601	86.370	1.226	47.594	71.098	84.982	2.736	45.612	75.238	87.332
UnCon [29]	1.192	35.493	75.392	87.891	0.491	16.928	53.889	75.597	1.741	50.471	79.470	89.474	3.604	42.443	75.654	87.271
UnOcc [27]	1.454	53.460	80.650	90.212	0.421	11.148	42.627	68.151	1.906	61.457	84.587	91.570	2.720	40.834	71.622	84.935
Ours	**0.360**	**8.766**	**24.420**	**47.911**	**0.261**	**5.796**	**25.303**	**48.668**	**0.772**	**10.048**	**30.427**	**55.732**	**0.842**	**13.631**	**36.038**	**57.880**

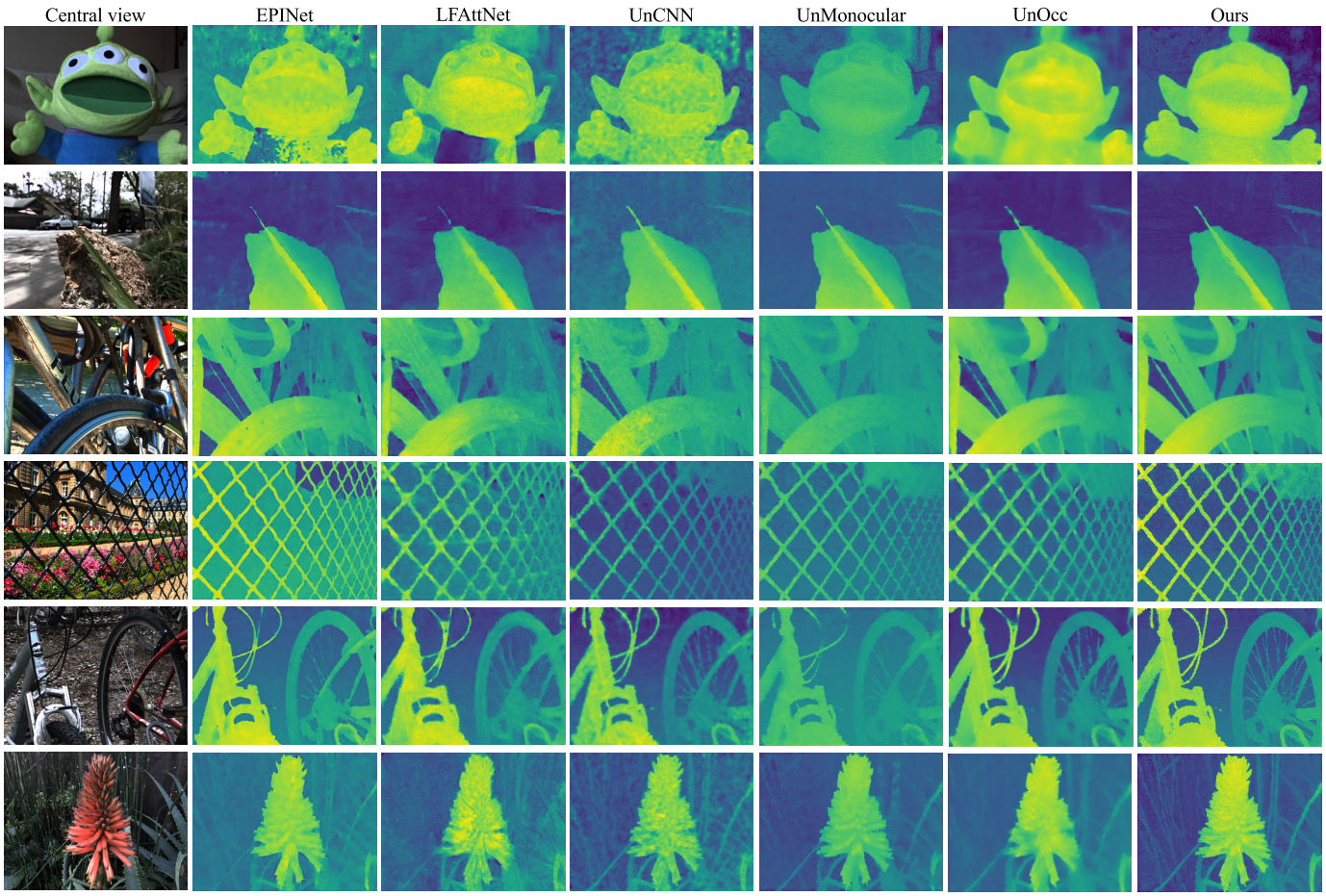


Fig. 9. Visual results of different methods on the real-world LF images from the datasets [44], [45], [46]. Zoom in for best view.

filters is 2. For the dense LFs, The range of disparity samples was set to $[-12, 12]$ (three times the disparity range $[-4, 4]$) with a coarse sampling interval 1, and the range of residual samples was set to $[-1, 1]$ with a finer sampling interval 0.1. For the sparse LFs, only the central view and its adjacent views were used since a large disparity range would lead to high memory consumption. Thus, the range of disparity samples was set to $[-20, 20]$ with a coarse sampling interval 1.2, and the range of residual samples was set to $[-2, 2]$ with a finer sampling interval 0.12. For the OccNet, the maximum channel number is 64. Therefore, it is very lightweight with only 0.113 M parameters.

During training, all the views were cropped to 256×256 randomly. The learning rate was set to 1×10^{-3} initially, and multiplied by 0.8 every 50 epochs. Moreover, we empirically set $\alpha_1 = 1$ for the SSIM loss, and $\eta = 100$, $\alpha_2 = 0.1$, $\alpha_3 = 0.05$ for the smoothness losses. The DispNet and OccNet were jointly trained using the loss in Eq. 14 with the Adam optimizer for about 500 epochs.

During inference, multiple view combinations were input to the DispNet to obtain multiple disparity maps. For the dense LFs, the input combinations (expressed by the angular coordinates) are $[(u_c - 3, v_c), (u_c, v_c), (u_c + 3, v_c)]$, $[(u_c - 2, v_c), (u_c, v_c), (u_c + 2, v_c)]$, $[(u_c, v_c - 3), (u_c, v_c), (u_c, v_c + 3)]$, $[(u_c, v_c - 2), (u_c, v_c), (u_c, v_c + 2)]$, and the weighted fusion with $n' = 2$ was used to obtain the final disparity map. For the sparse LFs, the input combinations are $[(u_c - 1, v_c), (u_c, v_c), (u_c + 1, v_c)]$, $[(u_c, v_c - 1), (u_c, v_c), (u_c, v_c +$

1)], and the minimum error fusion was used since there are only two estimated disparity maps. The quantile q in Eq. 15 was set to 0.95.

C. Comparison

We compared our method with several state-of-the-art LF depth estimation methods, including both the supervised and unsupervised methods, on a variety of LF datasets.

1) Evaluation on Synthetic Dense LF Images: First, we list the quantitative results of different methods on several scenes from the HCI dataset in Table I, including two supervised methods, EPINet [21] and LFAttNet [23], two non-learning-based methods, OCC [34] and FBS [17], and four unsupervised methods, UnCNN [25], UnMonocular [26], UnPlug [28], and UnOcc [27]. The evaluation metrics are the mean square error (MSE $\times 100$) and bad pixel ratios (BPR) [43] with thresholds 0.07, 0.03 and 0.01 (lower is better for all). The results of the other methods were obtained from the benchmark of the HCI dataset [43] or the original papers.[1] It can be seen that our method still has some gaps with the supervised methods but can generally outperform the other unsupervised methods. Some of the visual results (Dino and Sideboard), with the error maps relative to the ground truths, are presented in Fig. 6, where we can find that our estimated disparity maps have better visual qualities with smaller errors than the other non-learning-based and unsupervised methods.

Then, we compared the performance of different methods on several scenes from the DLF dataset [22], as recorded in Table II. The methods for comparison include EPINet [21], LFAttNet [23], UnCNN [25], UnMonocular [26], UnCon [29], and UnOcc [27]. We re-built these models and trained them using the same synthetic dense LF datasets since there are no off-the-shelf models for evaluation. From our experiments, we can see that our method obtains better quantitative results than the other unsupervised methods. Compared to the supervised methods, our method still has some gaps in terms of the MSE, but with comparable or lower BPRs. Fig. 7 presents the visual results on the scenes, Toys and Antiques, which reflects that our estimated disparity maps are more visually compelling with proper smoothness and clear object details.

2) Evaluation on Synthetic Sparse LF Images: To evaluate the performance of different methods on the sparse LF images, we retrained them using the SLF dataset [22]. The quantitative results of several scenes in terms of the MSE and BPR with thresholds 0.3, 0.1 and 0.05 are listed in Table III. It can be seen that our method achieves comparable quantitative results with the supervised method LFAttNet, and has better results than the others. Both our method and LFAttNet predict the probability distribution of the disparity samples by constructing cost volumes, while the other methods directly output the disparity values from the last convolution layer, which increases the difficulty to learn large disparities for the networks. The visual results on the scenes, Lion and Rooster clock, are shown in Fig. 8, with a different color map from the dense LFs for distinction. It can be seen that our disparity maps have better visual qualities with smaller errors compared to the other methods.

[1]The results of UnOcc [27] and UnPlug [28] are from their published papers, and they provide only the MSE and BPR with threshold 0.07.

TABLE IV
NUMBER OF PARAMETERS AND RUN TIME OF DIFFERENT METHODS

Method	Param. (M)	Run time (s)
EPINet [21]	2.466	2.765
LFAttNet [23]	5.540	6.479
UnCNN [25]	1.315	5.768
UnMonocular [26]	1.878	1.434
UnOcc [27]	1.891	1.840
UnCon [29]	2.185	2.102
UnPlug [28]	2.466	2.765
Ours	1.802	2.688

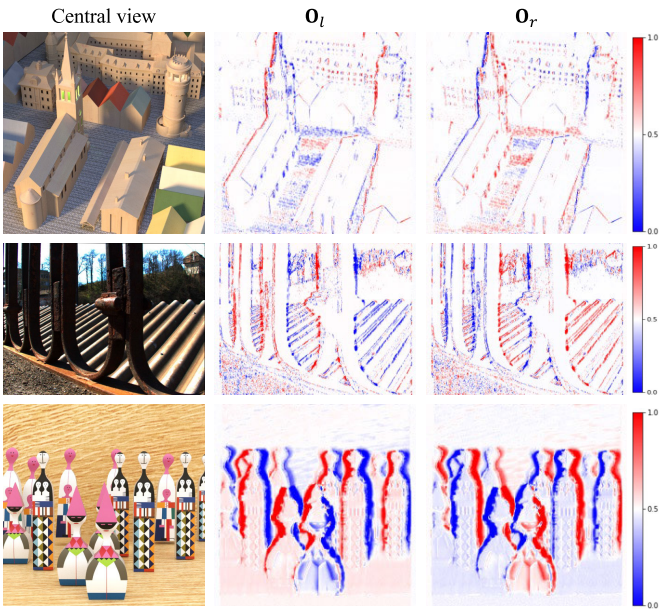


Fig. 10. The predicted occlusion maps of a synthetic dense LF, a real-world LF and a synthetic sparse LF.

3) Evaluation on Real-World LF Images: We further evaluated different methods on the real-world LF images from the datasets [44], [45], [46]. All the methods were trained only on the synthetic LF images. Fig. 9 presents several predicted disparity maps of some methods. As the real-world scenes do not have ground truths, we can only compare their visual qualities. It can be seen that our method achieves more smooth predictions while preserving better object details, and therefore demonstrates better robustness and generalization compared to the other methods.

4) Efficiency: We list the number of parameters and run time (the average time for outputting the disparity map for one LF image) of different methods, shown in Table IV.[2] All the methods were implemented on a NVIDIA Tesla P100 GPU. It can be seen that our network is lightweight, and achieves a good balance between the accuracy and efficiency.

D. Ablation Studies

1) Design of DispNet: We first validated the coarse-to-fine structure of DispNet by training an additional model without the branch of residual estimation. We chose several scenes

[2]UnPlug [28] employs the architecture of EPINet [21], so their number of parameters and run time are the same.

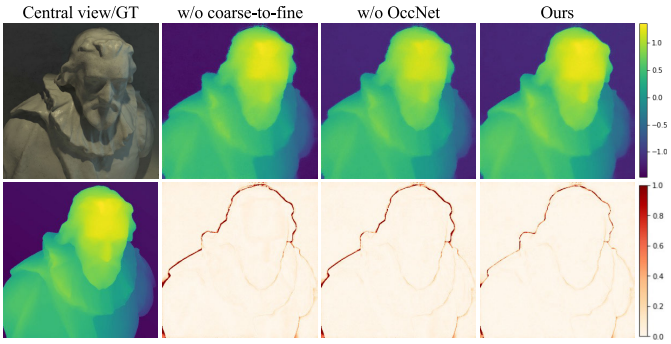


Fig. 11. Visual comparison of different network configurations. Zoom in for best view.

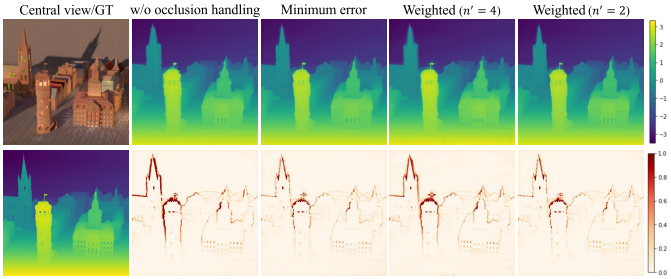


Fig. 12. Visual comparison of different disparity fusion strategies. Zoom in for best view.

TABLE V
ABLATION STUDY ON THE FRAMEWORK. BOLD: BEST

Configuration	Param. (M)	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
w/o coarse-to-fine	1.799	3.057	11.487	27.302	62.725
w/o shared weights	2.230	2.282	9.365	21.882	54.137
w/o OccNet	1.802	2.701	10.806	24.466	58.244
Default	1.802	**2.266**	**9.238**	**21.683**	**54.042**

TABLE VI
ABLATION STUDY ON THE LOSS TERMS. BOLD: BEST

Loss terms	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
w/o ℓ_{SSIM}	2.393	9.535	22.027	55.250
w/o ℓ_{smd}	2.457	9.749	22.136	55.856
w/o ℓ_{smo}	2.276	9.884	22.304	54.275
Full loss	**2.266**	**9.238**	**21.683**	**54.042**

TABLE VII
ABLATION STUDY ON DISPARITY FUSION APPROACHES. BOLD: BEST

Fusion strategy	MSE↓ (×100)	BPR↓ (0.07)	BPR↓ (0.03)	BPR↓ (0.01)
w/o occlusion handling	2.837	10.371	23.441	56.778
Minimum error fusion	2.411	10.210	23.743	57.842
Weighted fusion ($n' = 4$)	2.564	10.022	22.513	54.462
Weighted fusion ($n' = 2$)	**2.266**	**9.238**	**21.683**	**54.042**

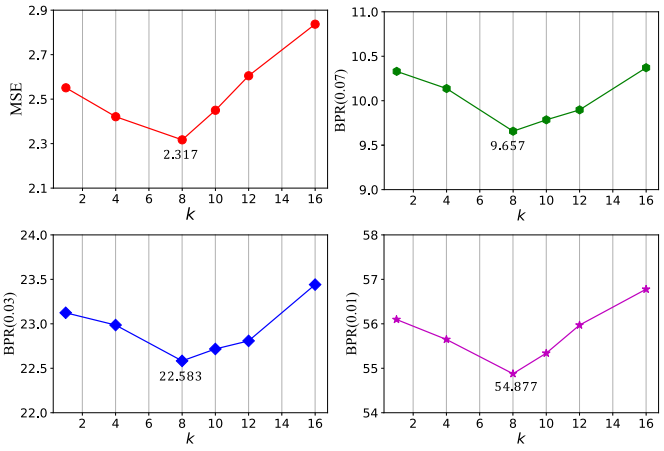


Fig. 13. The effect of k on the quantitative results.

from the HCI and DLF datasets for evaluation. Table V lists the number of parameters (during inference) and quantitative results of different configurations. It can be seen that the quantitative results have obvious decline if the coarse-to-fine structure is not employed, which suggests that the refinement branch with finer samplings helps to improve the disparity accuracy. Moreover, the refinement branch only introduces extra 0.003 M parameters since the feature extractor and the cost filters are all shared by the two branches. The visual comparison in Fig. 11 suggests that the DispNet with a coarse-to-fine structure achieves better disparity estimation with smaller errors.

We also trained an additional model without shared weights for the cost filters. From Table V, we can see that separate cost filters for the two branches lead to more parameters but no improvement on the performance. Therefore, we chose to use shared cost filters in our DispNet.

2) Occlusion Prediction: To verify the effectiveness of our OccNet for occlusion prediction, we trained an additional model by removing the OccNet and the loss terms ℓ_{rec} and ℓ_{smo} in Eq. 14. Thus, the pixel-wise weight of the photometric loss in Eq. 10 was reduced to 0.5. From Table V, we observe that the performance is degraded if the OccNet is not employed. The visual comparison in Fig. 11 reflects that the disparity map with occlusion prediction has smaller errors.

The predicted occlusion maps for a synthetic dense LF, a real-world LF and a synthetic sparse LF are presented in Fig. 10. The occlusion regions are usually near the object boundaries [8], which leads to larger or smaller values within these regions in the occlusion maps. The red pixels with values approximate to 1 denote larger contributions for the reconstruction of the central view and also larger weights for the photometric loss. The case is opposite for the blue pixels with values approximate to 0. Moreover, due to the larger disparity range, the sparse LFs usually have larger occlusion regions, resulting in much thicker red and blue regions near the object boundaries compared to the dense LFs.

3) Loss Terms: We trained additional models by excluding the SSIM loss ℓ_{SSIM}, the smoothness loss for disparity map ℓ_{smd} and the smoothness loss for occlusion map ℓ_{smo}, since ℓ_{wpm} and ℓ_{rec} are indispensable to the training of DispNet and OccNet. Table VI lists the corresponding results, which suggests that the MSE and BPRs slightly increase after removing each of them. Therefore, these loss terms contributes to a better performance.

4) Disparity Fusion Strategy: We first evaluated different disparity fusion strategies, and our default strategy is the weighted fusion by using two disparities ($n' = 2$) with the smallest errors at each position. The other fusion approaches include the weighted fusion by using all the disparities

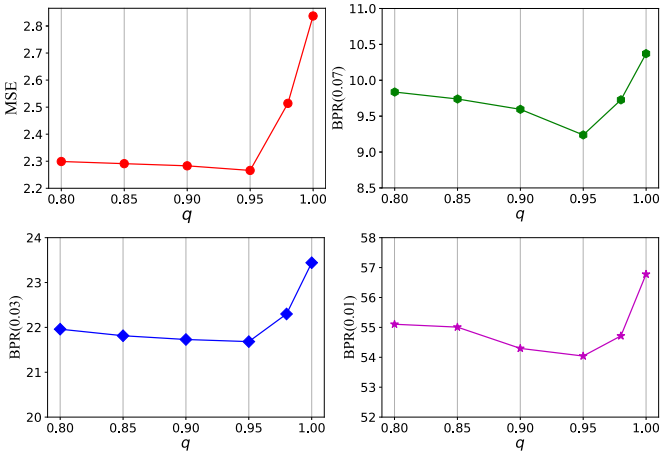


Fig. 14. The effect of q on the quantitative results.

($n' = 4$), the minimum error fusion that selects the disparity with the minimum error at each position. We also list the results of no occlusion handling when deriving the error maps (the estimated error of each position is represented by the mean of all the warping errors) with weighted fusion ($n' = 2$). Their quantitative results are given in Table VII and the visual comparison is shown in Fig. 12, where we observe that the weighted fusion with $n' = 2$ obtains better results than the other strategies, and the performance has obvious decline if occlusion is not handled during fusion.

Another approach to address occlusion is to use the average value of the k smallest warping errors. We experimented with $k = 1, 4, 8, 10, 12, 16$ (totally 16 auxiliary views) using weighted fusion with $n' = 2$. Fig. 13 shows the quantitative results at different k values. It can be seen that the results are best at $k = 8$, but they are still worse than those of using median value as listed in Table VII. As large warping errors are mainly caused by the occlusion and inaccurate disparity estimation, large k value may lead to incomplete occlusion handling while small k value may lead to inaccurate error estimation at the pixels with inaccurate disparity values. Therefore, using median value is more effective to address these issues.

We then investigated the effect of quantile q for determining the threshold of occlusion. We experimented with $q = 0.8, 0.85, 0.9, 0.95, 0.98, 1$. $q = 0.95$ means that 5% pixels with the largest standard deviations use the median value of the warping errors to address the occlusion issue, and $q = 1$ means that all the pixels use the mean value of the warping errors, which is equivalent to the strategy without occlusion handling in Table VII. Fig. 14 shows the effect of q on the quantitative results, where we can see that $q = 0.95$ obtains better results than the other values, and the MSE increases significantly at $q = 0.98$ and $q = 1$, verifying the effectiveness of occlusion handling.

V. Conclusion

This paper presents an unsupervised framework for LF depth estimation. We first develop a DispNet that takes as inputs different view combinations to predict multiple disparity maps. It adopts a coarse-to-fine structure with two branches to estimate the coarse and residual disparity maps, respectively, through multi-view feature matching. Photo-consistency is the main supervisory signal for the unsupervised training. However, it does not hold in the occlusion regions. To tackle this issue, we introduce an OccNet to predict the occlusion maps, which are used as the element-wise weights of the photometric loss to alleviate the impact of occlusions on the disparity learning. The OccNet is trained by the reconstruction loss with no ground truth required. The multiple estimated disparity maps after rotating and scaling are fused based on their estimated errors using the auxiliary views with effective occlusion handling to obtain the final disparity map. Our framework achieves superior performance on both the dense and sparse LF images, and also demonstrates better generalization ability on the real-world LF images.

References

[1] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, 2005.

[2] E. Y. Lam, "Computational photography with plenoptic camera and light field capture: Tutorial," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 32, no. 11, pp. 2021–2032, Nov. 2015.

[3] S. Zhang and E. Y. Lam, "A deep retinex framework for light field restoration under low-light conditions," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 2042–2048.

[4] S. Zhang, N. Meng, and E. Y. Lam, "LRT: An efficient low-light restoration transformer for dark light field images," *IEEE Trans. Image Process.*, vol. 32, pp. 4314–4326, 2023.

[5] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," in *Proc. IEEE Int. Conf. Comput. Photography (ICCP)*, May 2014, pp. 1–9.

[6] C. Kim, H. Zimmer, Y. Pritch, A. Sorkine-Hornung, and M. Gross, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–12, Jul. 2013.

[7] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11141–11148.

[8] N. Meng, K. Li, J. Liu, and E. Y. Lam, "Light field view synthesis via aperture disparity and warping confidence map," *IEEE Trans. Image Process.*, vol. 30, pp. 3908–3921, 2021.

[9] J. Chen and L.-P. Chau, "Light field compressed sensing over a disparity-aware dictionary," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 855–865, Apr. 2017.

[10] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, "UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, Nov. 2022.

[11] 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. 2014.

[12] S. Zhang, H. Sheng, C. Li, J. Zhang, and Z. Xiong, "Robust depth estimation for light field via spinning parallelogram operator," *Comput. Vis. Image Understand.*, vol. 145, pp. 148–159, Apr. 2016.

[13] 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.

[14] H. Sheng, P. Zhao, S. Zhang, J. Zhang, and D. Yang, "Occlusion-aware depth estimation for light field using multi-orientation EPIs," *Pattern Recognit.*, vol. 74, pp. 587–599, Feb. 2018.

[15] H.-G. Jeon et al., "Accurate depth map estimation from a lenslet light field camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1547–1555.

[16] M. W. Tao, P. P. Srinivasan, J. Malik, S. Rusinkiewicz, and R. Ramamoorthi, "Depth from shading, defocus, and correspondence using light-field angular coherence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1940–1948.

[17] J. Y. Lee and R.-H. Park, "Depth estimation from light field by accumulating binary maps based on foreground–background separation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 955–964, Oct. 2017.

[18] C.-T. Huang, "Empirical Bayesian light-field stereo matching by robust pseudo random field modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 552–565, Mar. 2019.

[19] X. Sun, Z. Xu, N. Meng, E. Y. Lam, and H. K.-H. So, "Data-driven light field depth estimation using deep convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 367–374.

[20] S. Heber, W. Yu, and T. Pock, "Neural EPI-volume networks for shape from light field," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2271–2279.

[21] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4748–4757.

[22] J. Shi, X. Jiang, and C. Guillemot, "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.

[23] Y.-J. Tsai, Y.-L. Liu, M. Ouhyoung, and Y.-Y. Chuang, "Attention-based view selection networks for light-field disparity estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12095–12103.

[24] 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.

[25] 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. (3DV)*, Sep. 2018, pp. 295–303.

[26] 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.

[27] 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.

[28] T. Iwatsuki, K. Takahashi, and T. Fujii, "Unsupervised disparity estimation from light field using plug-and-play weighted warping loss," *Signal Process., Image Commun.*, vol. 107, Sep. 2022, Art. no. 116764.

[29] L. Lin, Q. Li, B. Gao, Y. Yan, W. Zhou, and E. E. Kuruoglu, "Unsupervised learning of light field depth estimation with spatial and angular consistencies," *Neurocomputing*, vol. 501, pp. 113–122, Aug. 2022.

[30] T.-H. Tran, G. Mammadov, and S. Simon, "GVLD: A fast and accurate GPU-based variational light-field disparity estimation approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2562–2574, Jul. 2021.

[31] 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.

[32] X. Wang, J. Liu, S. Chen, and G. Wei, "Effective light field de-occlusion network based on Swin transformer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2590–2599, 2023.

[33] Y. Zhang, W. Dai, M. Xu, J. Zou, X. Zhang, and H. Xiong, "Depth estimation from light field using graph-based structure-aware analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4269–4283, Nov. 2020.

[34] 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.

[35] W. Williem and I. K. Park, "Robust light field depth estimation for noisy scene with occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4396–4404.

[36] 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.

[37] Y. Wang, 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. (CVPR)*, Jun. 2022, pp. 19809–19818.

[38] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[39] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[40] S. Zhang and E. Y. Lam, "Learning to restore light fields under low-light imaging," *Neurocomputing*, vol. 456, pp. 76–87, Oct. 2021.

[41] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6602–6611.

[42] S. Zhang and E. Y. Lam, "An effective decomposition-enhancement method to restore light field images captured in the dark," *Signal Process.*, vol. 189, Dec. 2021, Art. no. 108279.

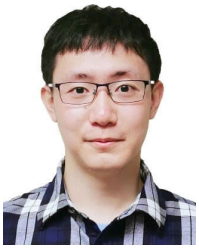
[43] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldluecke, "A dataset and evaluation methodology for depth estimation on 4D light fields," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 19–34.

[44] R. Shah, G. Wetzstein, A. S. Raj, and M. Lowney. (2016). *Stanford Lytro Light Field Archive*. [Online]. Available: http://lightfields.stanford.edu/LF2016.html

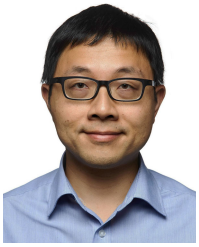
[45] N. K. Kalantari, T.-C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–10, Nov. 2016.

[46] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *Proc. Int. Conf. Quality Multimedia Exper.*, 2016. [Online]. Available: https://infoscience.epfl.ch/record/218363

Shansi Zhang (Graduate Student Member, IEEE) received the B.S. degree in mechanical engineering from the Beijing Institute of Technology in 2016 and the M.S. degree in electrical and electronic engineering from Nanyang Technological University in 2019. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, The University of Hong Kong. Her research interests include image processing, computer vision, and computational imaging.



Nan Meng (Member, IEEE) received the B.S. degree in computer science from the University of Electronic Science and Technology of China in 2015 and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong. Currently, he is a Post-Doctoral Researcher with the Department of Orthopaedics and Traumatology, The University of Hong Kong. His research interests include machine learning, light field reconstruction, light field rendering, and medical imaging, intelligent orthopaedics.



Edmund Y. Lam (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA. He was a Visiting Associate Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor of electrical and electronic engineering with The University of Hong Kong, Hong Kong. He is also a Computer Engineering Program Director and a Research Program Coordinator with the AI Chip Center for Emerging Smart Systems. His research interests include computational imaging algorithms, systems, and applications. He is a fellow of OPTICA, SPIE, IS&T, and HKIE and a Founding Member of the Hong Kong Young Academy of Sciences.