sensors

MDPI

*Article*

# Efficiency–Accuracy Trade-Off in Light Field Estimation with Cost Volume Construction and Aggregation

**Bo Xiao [1,2], Stuart Perry [2], Xiujing Gao [3,4] and Hongwu Huang [1,3,4,*]**

[1] State Key Laboratory of Advanced Design and Manufacturing for the Vehicle Body, Hunan University, Changsha 410012, China; xiaobo123@hnu.edu.cn
[2] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia; stuart.perry@uts.edu.au
[3] School of Smart Marine Science and Engineering, Fujian University of Technology, Fuzhou 350118, China; gaoxiujing@fjut.edu.cn
[4] Fujian Provincial Key Laboratory of Marine Smart Equipment, Fuzhou 350118, China
* Correspondence: hhwxm05@163.com

**Abstract:** The Rich spatial and angular information in light field images enables accurate depth estimation, which is a crucial aspect of environmental perception. However, the abundance of light field information also leads to high computational costs and memory pressure. Typically, selectively pruning some light field information can significantly improve computational efficiency but at the expense of reduced depth estimation accuracy in the pruned model, especially in low-texture regions and occluded areas where angular diversity is reduced. In this study, we propose a lightweight disparity estimation model that balances speed and accuracy and enhances depth estimation accuracy in textureless regions. We combined cost matching methods based on absolute difference and correlation to construct cost volumes, improving both accuracy and robustness. Additionally, we developed a multi-scale disparity cost fusion architecture, employing 3D convolutions and a UNet-like structure to handle matching costs at different depth scales. This method effectively integrates information across scales, utilizing the UNet structure for efficient fusion and completion of cost volumes, thus yielding more precise depth maps. Extensive testing shows that our method achieves computational efficiency on par with the most efficient existing methods, yet with double the accuracy. Moreover, our approach achieves comparable accuracy to the current highest-accuracy methods but with an order of magnitude improvement in computational performance.

**Keywords:** depth estimation; light field; convolution neural network

check for updates

**Citation:** Xiao, B.; Perry, S.; Gao, X.; Huang, H. Efficiency–Accuracy Trade-Off in Light Field Estimation with Cost Volume Construction and Aggregation. *Sensors* **2024**, *24*, 3583. https://doi.org/10.3390/s24113583

Academic Editor: Jacek Wojtas

Received: 2 May 2024
Revised: 26 May 2024
Accepted: 29 May 2024
Published: 1 June 2024

## 1. Introduction

Light field imaging can serve as a significant potential tool for constructing 3D environments. Unlike traditional imaging, light field imaging captures a richer array of information, describing the distribution of light rays in three-dimensional space. Furthermore, it has applications in virtual reality [1,2], view synthesis [3], 3D reconstruction [4], and autonomous driving [5,6]. Depth estimation is a fundamental and critical step in these important research areas. However, these applications require not only high accuracy but also rapid generation speeds, thus necessitating that depth estimation processes simultaneously meet the demands for both estimation accuracy and computational efficiency.

In recent years, deep learning-based methods [6–11] have achieved significant advancements and demonstrated considerable potential in the realm of light field depth estimation. These approaches typically utilize all available light field image information, effectively improving the accuracy of depth estimation. However, the presence of abundant light field information in light fields leads to a substantial increase in network memory consumption and computational load. Therefore, many researchers have tried to improve computational speed by pruning redundant light field images, which indeed achieves good results. However, reducing the input information inevitably lowers accuracy and the perception of occlusion and texture regions. To address this issue, state-of-the-art methods [10] have utilized full correlation to construct matching cost volumes, replacing 3D convolution operations with 2D convolutions to further reduce computational load while incorporating multi-scale aggregation to boost accuracy. Nonetheless, full correlation tends to lose critical depth features and performs poorly in textureless areas. There are also many in-depth studies on the perception of textureless regions [12–14]. In summary, although 2D convolutions reduce data volume, they lose significant spatial depth information compared with 3D convolutions, rendering the network less effective in handling complex spatial scenes.

To overcome these limitations, we propose a more effective method for cost volume construction and a cost aggregation architecture. Firstly, we replace full correlation with grouped correlation to enhance feature matching information. Secondly, we introduce feature dissimilarity operations to compensate for the shortcomings of feature correlation in textureless areas. Lastly, we present an architecture integrating Hourglass modules with 3D convolutions for multi-scale disparity fusion. This network structure more effectively captures multi-scale spatial features, and the use of 4D feature vectors further enhances the model's ability to detect texture details, thereby improving both performance and accuracy in depth estimation. Our performance is shown in Figure 1.

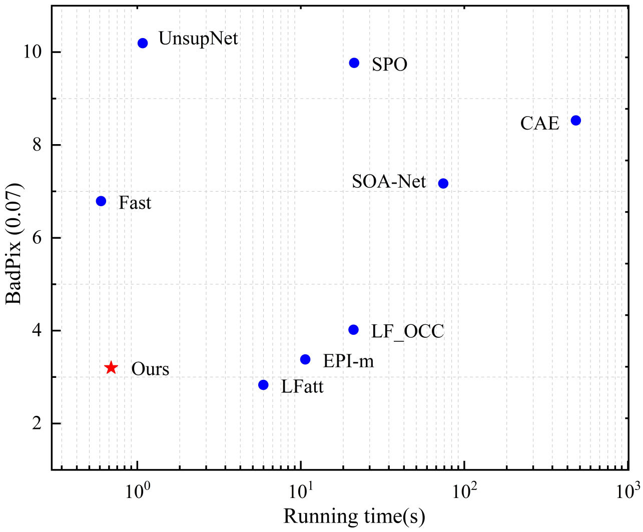

**Figure 1.** Comparison of efficiency and computation performance of light field disparity estimation algorithms.

## 2. Related Work

In this section, we review the main works in the direction of light field depth estimation based on traditional and deep learning methods.

### *2.1. Traditional Methods*

Previous work on light field depth estimation has leveraged various light field attributes to obtain scene depth information. Traditional depth estimation methods generally fall into three categories: methods based on epipolar-plane images (EPIs) [15–19], methods based on multi-view stereo matching (MVS) [20–23], and methods based on refocusing [7,24–27].

The concept of EPIs was first introduced by Bolles et al., employing light field (LF) epipolar geometry to calculate line slopes for depth prediction. Wanner et al. [28] then incorporated it into light field depth estimation, using only horizontal- and vertical-direction EPIs of light field images, optimizing the results with a global consistency labeling algorithm. The EPI method significantly accelerated disparity estimation speed. Zhang et al. [29]

further proposed the spinning parallelogram operator (SPO) to compute line slopes in EPIs, enhancing the accuracy of disparity estimation. Sheng et al. [30] used multi-directional EPIs to optimize slope estimation accuracy, achieving results surpassing the SPO. Heber et al. [21] developed a principal component analysis matching item for multi-view stereo reconstruction, combined with the projection of sub-aperture images (SAIs) for depth estimation. Jeon et al. [22] introduced a Fourier transform-based phase-shift theory to address small disparities between SAIs. In the refocusing approach, Tao et al. [24] combined defocus cues with consistency cues in light field images for depth map estimation, though performing poorly in occluded areas. Tao et al. [26] proposed a shadow-based refinement method to enhance the robustness of depth map estimation.

While these traditional methods have continuously progressed in accuracy and computational efficiency in light field disparity estimation, they are limited by nonlinear optimization and manually designed features. These features demand extensive computational resources and perform poorly in occluded and weakly textured areas, leaving substantial room for improvement in both accuracy and computational performance in disparity estimation.

### *2.2. Deep Learning Methods*

In recent years, the use of deep convolutional neural networks (CNNs) for light field depth estimation has achieved impressive results. Focusing on disparity estimation accuracy, Tsai et al. [31] introduced an attention-based visual selection module that integrates the importance of each view with depth estimation, significantly enhancing robustness against noise interference. Building on this, Chen et al. [8] combined attention mechanisms with multi-level fusion networks, using fusion between different angular branches to further enhance disparity estimation accuracy. Most recently, Yang et al. [32] integrated local and global features within view feature cost volumes to address the challenges of occlusions and textureless regions, further improving disparity estimation accuracy. However, these methods, due to the use of redundant information and extensive 3D convolution operations, tend to be slower in generating disparity maps.

In another direction, Heber et al. [33] proposed a U-shaped artificial neural network to extract geometric information from light fields for depth estimation, initially utilizing EPIs. Subsequently, Shin et al. [34] used CNNs to extract geometric disparities from EPIs and proposed a fully convolutional end-to-end network. Further, Huang [10] designed a disparity estimation model that replaced 3D convolutions with 2D convolutions, significantly reducing the learning parameters and enhancing computational performance. These methods, which generate disparity maps using a lower proportion of light field images as input, have good computational performance but are limited in disparity estimation accuracy and robustness against real-world noise, especially compared with methods that use inputs from all views.

Finally, previous research has already shown the advantages of deep neural networks in light field depth estimation. However, there has been insufficient focus on balancing accuracy and computational performance, often leading to a trade-off when generating disparity images. In this paper, we propose a lightweight convolutional neural network that employs multi-disparity cost aggregation. This network extracts richer depth information from fewer input data and achieves a balance between computational load and depth estimation accuracy.

## 3. Method

In this paper, we introduce a novel method that balances efficiency and accuracy in light field depth estimation. The overall architecture of the network is depicted in Figure 2. Initially, acknowledging the redundancy in light field images, we only use sub-aperture images (SAIs) from the horizontal and vertical directions as inputs to reduce computational costs as much as possible. A shared feature extraction module is then employed to extract SAI features (Section 3.1). Following this, we construct cost volumes based on the features of surrounding pixels after pixel shifting and central features. In this process, we propose a hybrid cost volume network to enhance detail perception (Section 3.2). Finally, a multi-scale disparity cost aggregation module is used to synthesize mixed cost depth information, which is then processed by a disparity regression module to predict the disparity map (Section 3.3).

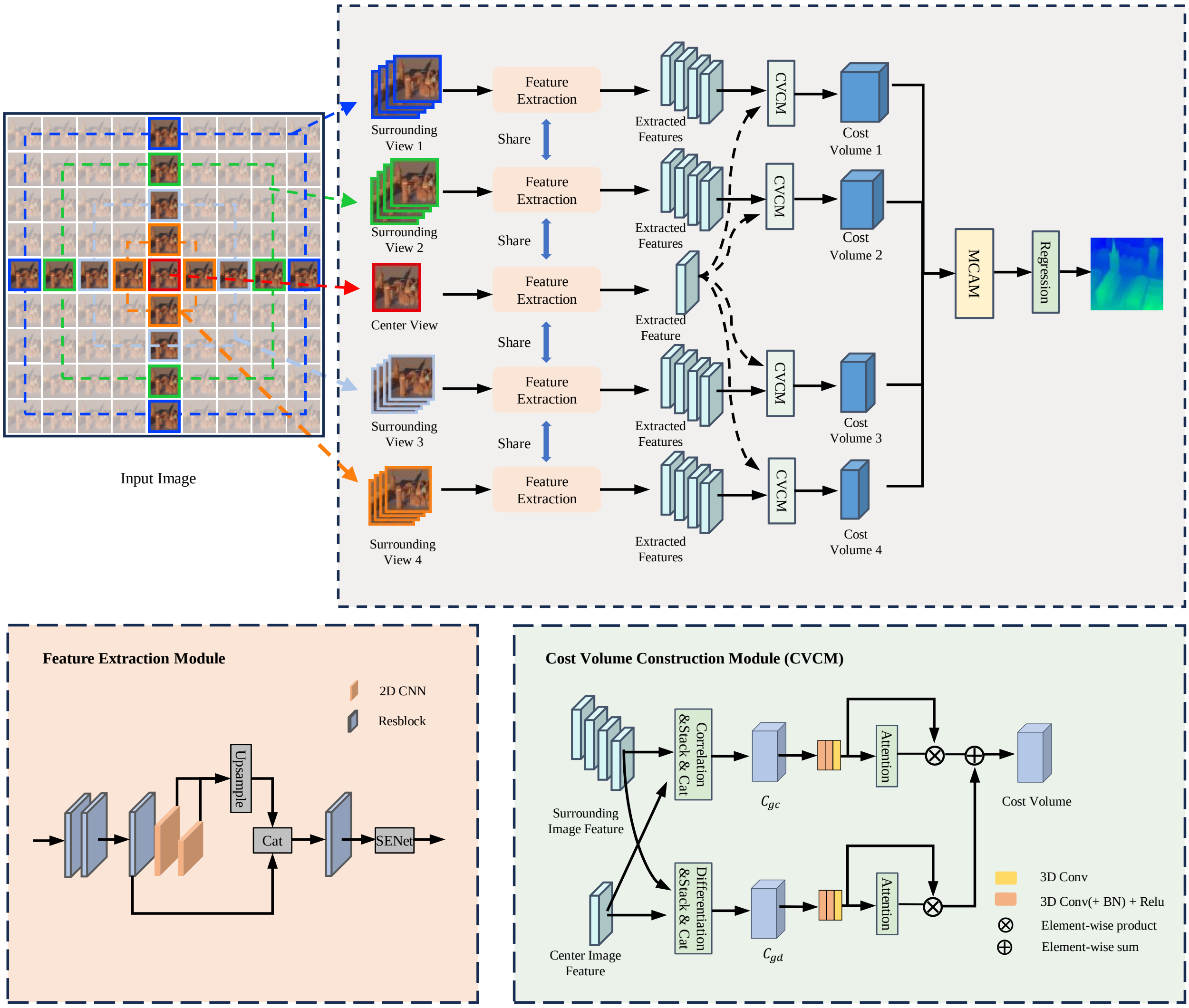

**Figure 2.** An overview of our network. The term "MCAM" denotes the multi-view cost volume aggregation module, and its specific structure is detailed in Figure 3.

### 3.1. Feature Extraction

The extraction of effective features is crucial to estimating the disparity map, particularly due to the small disparity range in light fields, which complicates accurate estimation in low-texture and occluded areas. To address this, we use multiple basic residual blocks for preliminary feature extraction, applying stride-2 convolutions for downsampling. The feature maps are downsampled at two different scales and subsequently restored to their original scale through bilinear interpolation. These features at different levels are concatenated and fused via convolution, and an SENet [35] module is added to enhance the weighting of key feature channels. The resultant feature map serves as the input to our hybrid cost volume network.

### 3.2. Texture-Aware Cost Volume

After extracting features of SAIs, a 4D cost volume is constructed to predict the disparity map by establishing a correspondence between shifted surrounding view features and the central view feature. We use parallel plane parameterization to represent the four-dimensional light field, $L(\mathbf{x}, \mathbf{u})$, where $\mathbf{x}$ and $\mathbf{u}$ are the spatial and angular coordinates, respectively. $I_c(\mathbf{x}, \mathbf{u}_c)$ is the central view. The disparity of the light field is denoted by $d(\mathbf{x}, \mathbf{u}_c)$, and $\mathbf{u}_c$ denotes the center view position. According to LF geometry, given a surrounding SAI $I_s(\mathbf{x}, \mathbf{u})$, the reconstructed central view, $\tilde{I}_c(\mathbf{x}, \mathbf{u})$, can be expressed as

$$\tilde{I}_c(\mathbf{x}, \mathbf{u} \to \mathbf{u}_c) = I_s(\mathbf{x} + (\mathbf{u} - \mathbf{u}_c) \cdot d(\mathbf{x}, \mathbf{u}_c), \mathbf{u}) \tag{1}$$

By using this equation, we can reconstruct the central view, $\tilde{I}_c(\mathbf{x}, \mathbf{u})$, from the surrounding views, $I_s(\mathbf{x}, \mathbf{u})$, where $\mathbf{x} + (\mathbf{u} - \mathbf{u}_c) \cdot d(\mathbf{x}, \mathbf{u}_c)$ calculates the displacement caused by the disparity.

Typically, the full correlation cost volume is obtained by using correlation operations [10,36] between the distorted features of surrounding views and the central view to regress the disparity map. However, relying solely on full correlation can result in the loss of significant information. To further reduce the computational load, we group features to compress the matching cost volume. The number of channels of a univariate feature is denoted by $N_c$, and the channels are uniformly divided into $N_g$ groups along the channel dimension. Therefore, each group feature has $N_c/N_g$ channels. Correlation is then computed for each group. The correlation between the surrounding view features, $F_s^g$, and the central view features, $F_c^g$, is represented as follows:

$$C_{gc}(d, \mathbf{x}, g) = \frac{N_g}{N_c} \left\langle F_c^g(\mathbf{x}, \mathbf{u}_c), F_s^g(\mathbf{x} + (\mathbf{u} - \mathbf{u}_c) \cdot d, \mathbf{u}) \right\rangle \tag{2}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of two features and $C_{gc}$ is the correlation cost volume for feature group $g$ and disparity $d$.

However, due to the low values of features in textureless areas, the multiplication operation in the correlation process results in a small variance between the feature costs at the correct and incorrect depths. This can easily lead to interference by noise and incorrect depth estimation. To address this issue, we introduce a new set of cost volumes. We construct cost volumes by using the sum of absolute differences between feature views, effectively increasing the variance range of depth-related feature costs and enhancing the network's perception of the correct depth in textureless areas. The differentiation cost volume between feature view pixels is represented as follows:

$$C_{gd}(d, \mathbf{x}, g) = \sum_{i=1}^{N_c/N_g} \left| F_c^g(\mathbf{x}, \mathbf{u}_c) - F_s^g(\mathbf{x} + (\mathbf{u} - \mathbf{u}_c) \cdot d, \mathbf{u}) \right| \tag{3}$$

Furthermore, the cost volumes are stacked in the depth direction as a 4D array $(G \times D \times H \times W)$ and then concatenated to connect volumes with the same disparity scale, forming the initial cost volume $(4G \times D \times H \times W)$. Here, $G$ represents the number of groups, $D$ the number of disparity layers, and $H$ and $W$ the dimensions of the input image. It is important to note that the relationship of disparity scales at different distance angles is proportional to the distance ($d$). Here, we define the maximum disparity for the innermost view as $d_{max}$, and considering that disparity estimation requires multiple downsampling operations, we set the total number of disparity layers to be even, with a disparity range of $[-d_{max}, 1 + d_{max}]$ and disparity levels set to $2 + 2d_{max}$. Similarly, the disparity for the outermost view is set to $[-4d_{max}, 1 + 4d_{max}]$, with disparity levels set to $2 + 4d_{max}$. The disparity level refers to the number of discrete disparities within the interval from the minimum to the maximum disparity.

Finally, through the network we propose, the correlation cost volume and the differentiation cost volume are fused, as illustrated in the CVCM module shown in the lower right corner of Figure 2. The correlation cost volume and the differentiation cost volume are processed separately through 3D convolution and a 3D channel attention mechanism to extract matching information. They are then combined to form the final cost volume.

During the construction of the cost volume matching process, we employed a grouping method to compress the most memory-intensive part, reducing computational load and memory consumption. Additionally, we proposed a correlation and dissimilarity fusion structure to enhance perception and accuracy in textureless regions.

### *3.3. Multi-View Cost Volume Aggregation and Disparity Regression*

To fuse cost volumes of different scales for disparity estimation and enhance the model's performance in occluded areas, we propose a multi-level fusion strategy. Given that spatially occluded areas in the central view are visible from other directional viewpoints, we employ a structure similar to U-Net, featuring upsampling and downsampling, along with Hourglass modules, to extract useful features from unoccluded areas. Furthermore, to achieve better accuracy in depth estimation, we designed a multi-level scale disparity fusion structure to enhance feature robustness. As shown in Figure 3, the first layer input is the maximum disparity cost volume, capturing the broadest spatial information to provide a comprehensive initial perspective for depth estimation. Information from different disparity scales is integrated from largest to smallest, offering a multi-level feature fusion strategy that transitions from local to global and back to local. Finally, each layer's disparity is upsampled to the same size and fused by using three-dimensional convolution.

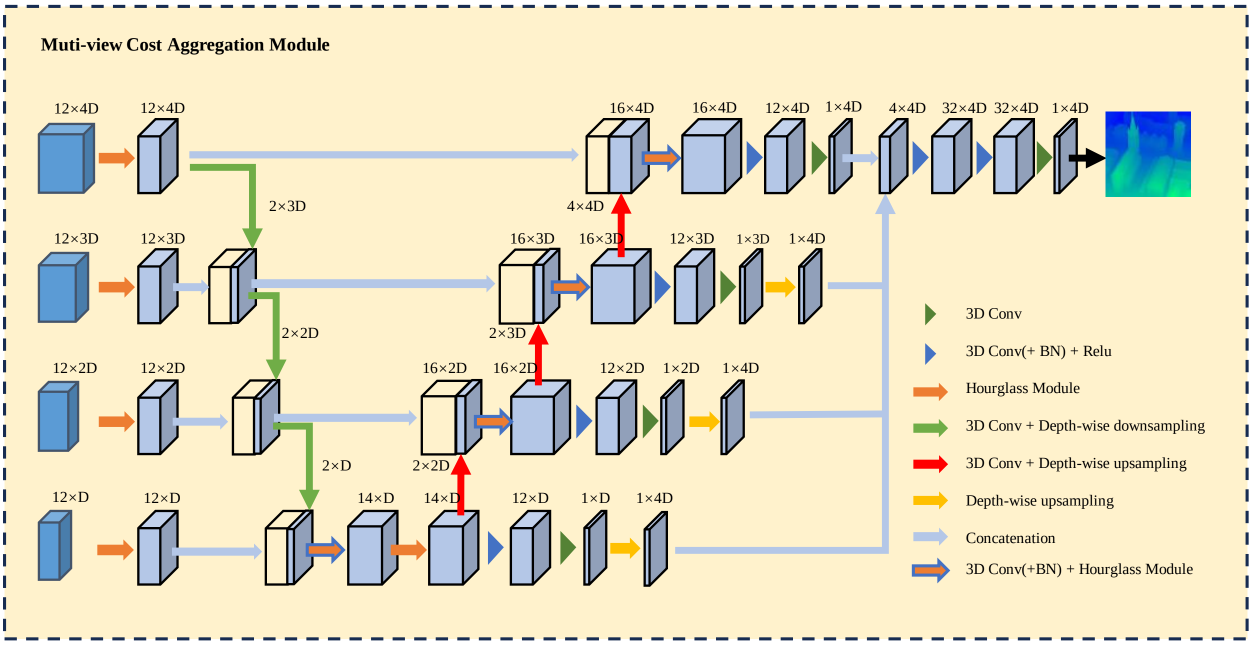

**Figure 3.** Multi-View Cost Aggregation Module architecture. Note that the feature size annotations in our diagram omit the dimensions $(H \times W)$ of the input light field images.

Furthermore, to more effectively capture global and local texture features at the current scale of each layer, we have embedded Hourglass modules [37,38] at every level. These modules not only recognize large-scale structures within the feature cost volumes but also process fine textures and edge details. The specific details of this process are illustrated in Figure 4. This approach ensures a more nuanced and comprehensive analysis of the light field data, significantly enhancing depth estimation accuracy and detail.

After obtaining the final cost volume, each pixel is represented by a vector of length $D_{max}$, containing the probabilities of all disparity levels. We use the softmax activation function introduced in [39] to generate a continuous distribution of disparity predictions. The predicted disparity value, $\hat{d}$, is defined as

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

where $\hat{d}$ represents the predicted disparity by the pixel; $D_{min}$ and $D_{max}$ denote the minimum and maximum disparities of the outermost view, respectively; and $C_f$ is the predicted cost at disparity $d_i$.

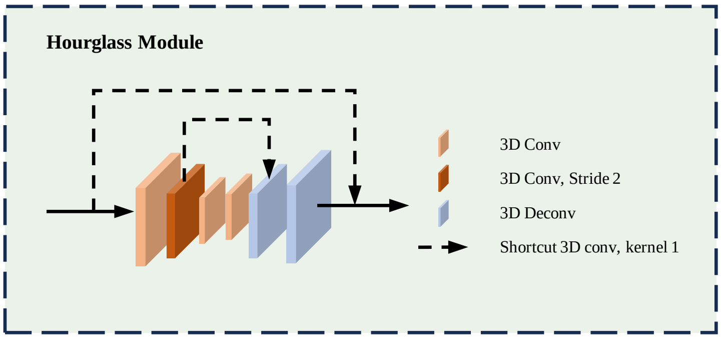

**Figure 4.** Hourglass module structure.

## 4. Experiments

In this section, we will introduce the details and results of our implementation. Finally, we will compare our method with several state-of-the-art light field depth estimation methods.

### *4.1. Dataset and Implementation Details*

The 4D light field dataset [40] is widely regarded as a benchmark for evaluating light field image disparity estimation methods. This dataset, rendered by using Blender, includes 28 densely arranged synthetic light field scenes and their corresponding ground-truth disparity maps divided into four subsets: "Stratified", "Test", "Training", and "Additional". These scenes incorporate a mix of various materials, lighting conditions, and complex spatial occlusions. All light field data possess a 9 × 9 angular resolution and a 512 × 512 image resolution.

We utilized the "Additional" category of the dataset for training our model. During training, we randomly cropped the SAIs into 48 × 48 grayscale patches and applied various image augmentation techniques, including random rotation, brightness and contrast adjustments, and noise injection. Inspired by Zhao [41], we adopted a joint L1 and SSIM [42] loss function for our training network, denoted by loss term $L$, and optimized it by using the Adam method [43]. The loss function is formulated as follows:

$$L = \frac{1}{M}\sum_{i,j}(\alpha\frac{1-\mathrm{SSIM}(d_{i,j},\tilde{d}_{i,j})}{2} + (1-\alpha)\left\|d_{i,j} - \tilde{d}_{i,j}\right\|) \tag{5}$$

where $\tilde{d}$ represents the predicted disparity map, $d$ is the true disparity map, and $M$ denotes the number of pixels, with $\alpha$ being set to 0.9. We tested weights $\alpha$ ranging from 0.1 to 0.9 and found that as the weight increases, accuracy in depth estimation also improves. The proposed method was implemented with PyTorch platform [44] and optimized by using the Adam [43] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) optimizer. The batch size was set to 16, and the initial learning rate was $10^{-3}$, decaying by 0.8 every 100 epochs. The total training comprised 1000 iterations. Our model was trained on a PC equipped with an Nvidia 6000× GPU, requiring approximately three days.

### *4.2. Comparison to State-of-the-Art Methods*

We compared our approach to various state-of-the-art methods, including traditional methods [25,29,45], unsupervised deep learning methods [46,47], and supervised deep learning methods [10,31,34,48].

To demonstrate the accuracy of our method, we compared its performance with other state-of-the-art methods on the "Stratified" and "Training" categories of 4D LF data in terms of bad pixel rate (BadPix) (0.07) and Mean Squared Error (MSE). The comparative results are reported in Table 1, showing that our method achieved good results overall compared with the other methods.

**Table 1.** Quantitative comparison with other state-of-the-art methods in terms of BadPix (0.07) and MSE on 4D light field benchmark dataset [40]. Lower scores represent better performance.

| **Methods** | **CAE [25]** | | **SPO [29]** | | **LFAtt [31]** | | **Fast [10]** | | **Distrib [46]** | | **EPI-m [34]** | | **Ours** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **BP07** | **MSE** | **BP07** | **MSE** | **BP07** | **MSE** | **BP07** | **MSE** | **BP07** | **MSE** | **BP07** | **MSE** | **BP07** | **MSE** |
| Backgammon | 2.967 | 5.170 | 2.608 | 3.607 | 3.126 | 3.648 | 3.323 | 1.818 | 19.22 | 13.68 | 2.229 | 2.579 | 2.342 | 1.830 |
| Boxes | 19.86 | 10.01 | 15.98 | 12.19 | 11.04 | 3.996 | 18.28 | 4.532 | 28.7 | 15.92 | 12.34 | 5.968 | 10.50 | 3.786 |
| Cotton | 3.562 | 1.844 | 2.343 | 2.009 | 0.272 | 0.209 | 0.984 | 0.341 | 21.22 | 12.74 | 0.55 | 0.287 | 0.537 | 0.265 |
| Dino | 5.752 | 0.407 | 2.469 | 0.407 | 0.848 | 0.093 | 3.122 | 0.208 | 13.41 | 8.775 | 1.207 | 0.157 | 1.245 | 0.125 |
| Dots | 15.50 | 8.127 | 35.29 | 16.68 | 1.432 | 1.425 | 16.36 | 3.524 | 35.70 | 6.663 | 2.490 | 1.475 | 3.939 | 2.322 |
| Pyramids | 1.822 | 0.053 | 0.271 | 0.02 | 0.195 | 0.004 | 0.407 | 0.017 | 8.992 | 2.029 | 0.159 | 0.008 | 0.382 | 0.008 |
| Sideboard | 11.05 | 0.876 | 7.670 | 1.027 | 2.870 | 0.530 | 7.472 | 0.823 | 19.81 | 11.31 | 4.462 | 0.798 | 3.865 | 0.584 |
| Stripes | 8.534 | 3.268 | 11.59 | 6.276 | 2.933 | 0.892 | 4.125 | 0.192 | 30.17 | 2.91 | 2.457 | 0.932 | 2.812 | 0.350 |

Bad pixel ratio of 0.07 (BP07) and MSE (multiplied with 100) are the metrics for accuracy evaluation, where lower scores represent better performance. The best result is shown in deep blue and the second best in orange.

For performance comparison, to ensure fairness, we executed these methods on the same platform and compared the average running time for these scenes. The results, as shown in Table 2, indicate that our method outperforms the other methods, except for the Fast method. Additionally, while our method's accuracy is second only to LFAtt [31], it computes faster. The traditional methods CAE [25] and SPO [29] were run on a CPU platform configured with an Intel i7-10850H.

**Table 2.** Quantitative comparison of the average performance and efficiency with state-of-the-art methods on the 4D LF Benchmark.

| **Method** | **Average BadPix (0.07)** | **Average Running Time/s** |
|---|---|---|
| CAE [25] | 8.530 | 481.0 |
| SPO [29] | 9.770 | 21.21 |
| LFAtt [31] | 2.836 | 5.913 |
| Fast [10] | 6.792 | 0.601 |
| Distrib [46] | 22.15 | 4.830 |
| EPI-m [34] | 3.383 | 10.65 |
| SOA-Net [48] | 7.170 | 74.30 |
| LF_OCC [45] | 4.021 | 21.00 |
| UnsuperNet [47] | 10.19 | 1.079 |
| Ours | 3.203 | 0.693 |

The data for methods SOA-Net [48] and UnsuperNet [47] are cited from reference [49]. The best result is shown in deep blue and the second best in orange.

In addition, we also subjectively compared the performance of our method with other methods in textureless and occluded areas. Figure 5 displays visual comparison results across four scenarios: "Dish", "Dots", "Rosemary", and "Origami". The depth estimation results and errors for the first three scenes show that our method performs well in spatial areas with occluded edges, comparable to the best methods available [31]. Additionally, in the "Origami" scene, as indicated, our method achieves better accuracy in the marked textureless areas. Overall, the results demonstrate that our method exhibits superior performance and robustness in handling the challenge of textureless and occluded areas.

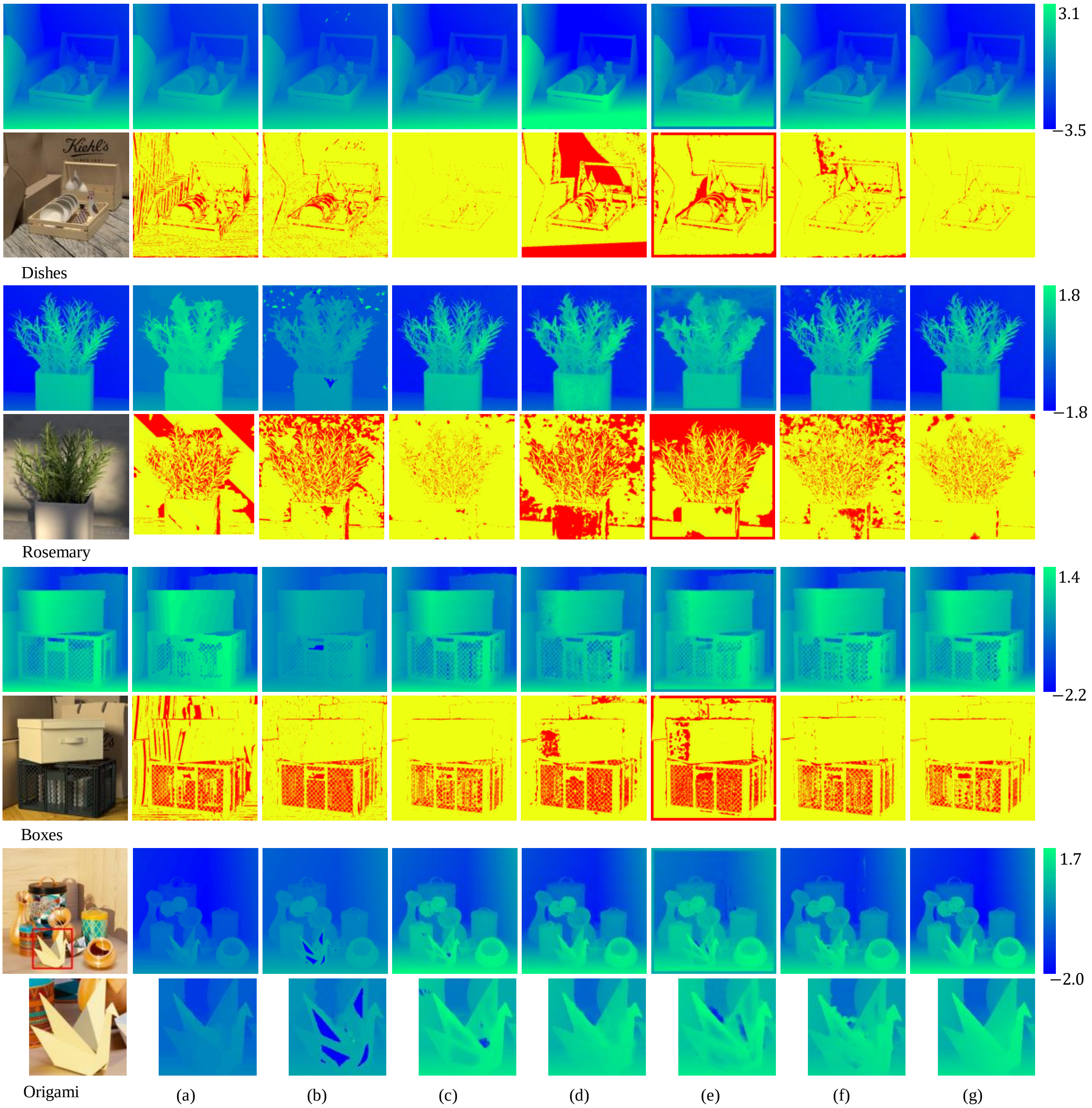

**Figure 5.** Quantitative comparison of the performance of different methods on the HCI light field benchmark. (**a**–**g**) show the results for CAE [25], SPO [29], LFAtt [31], Fast [10], Distrib [46], EPI-m [34], and our method, respectively. The first row in each scene represents the estimated disparity corresponding to the original image, and the second row displays the distribution of bad pixels, with red indicating areas where the bad pixel rate exceeds 0.07.

To comprehensively evaluate the performance of our method, we also used real-world datasets for testing and comparison with state-of-the-art methods. As illustrated in Figure 6, the depth maps generated by our method are more consistent and exhibit less noise. This indicates that our approach can be effectively generalized to real LF depth estimation. The scenes "Bench" and "Leaf" were captured by using our Lytro Illum camera, and "Knights" was captured by using the gantry setup from the Stanford Light Field Archive.

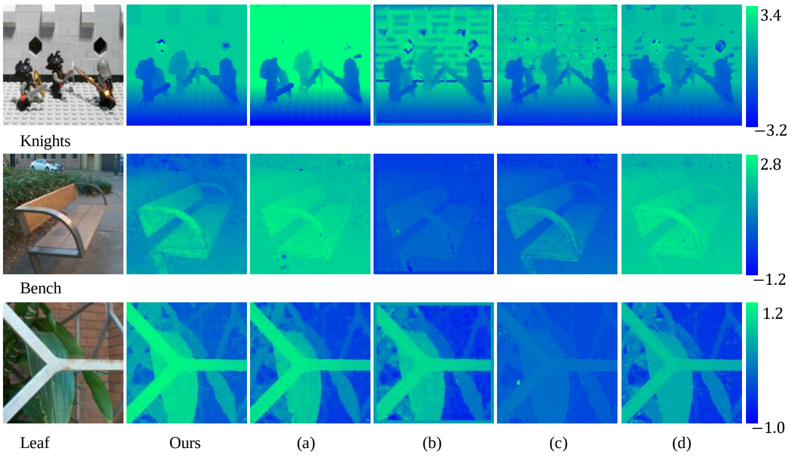

**Figure 6.** Qualitative results of real-world light field images. (**a**–**d**) represent depth maps generated by deep learning-based methods Fast [10], Distrib [46], EPI-m [34], and LFAtt [31], respectively.

### 4.3. Ablation Study

We conducted extensive ablation experiments to analyze the effectiveness of our method. Our ablation study includes the trade-off between performance and efficiency, the choice of disparity cost operations, and the combination of loss functions.

#### 4.3.1. Disparity Cost Calculation

The disparity cost has a significant impact on accuracy in depth estimation, so it is crucial to choose the appropriate cost generation operations. We conducted separate tests with and without the feature dissimilarity operation on the HCI 4D LF benchmark. When we removed the feature dissimilarity operation, the network's performance in terms of both MSE and BadPix (0.07) deteriorated. Figure 7 illustrates the influence of feature dissimilarity on depth estimation within the aggregated cost volume. It indicates that adding feature dissimilarity effectively improves the network's performance in weak-texture regions and enhances its robustness.

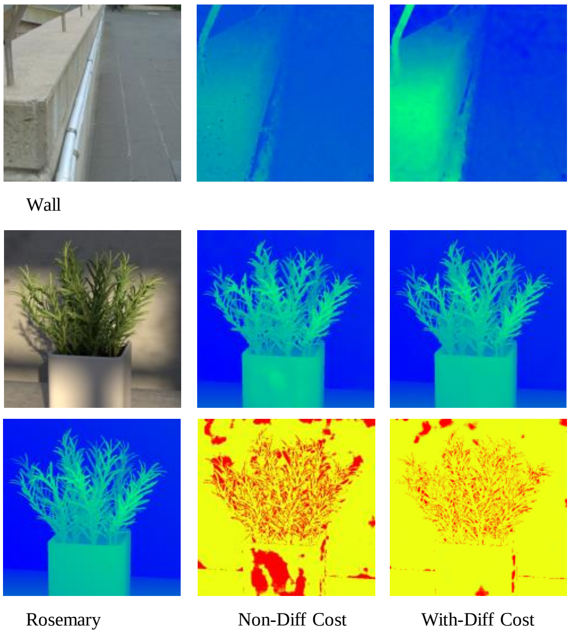

**Figure 7.** Disparity maps in synthetic and real scenes. "Wall" represents a light field image captured by us, and "Rosemary" is the synthetic light field image.

### 4.3.2. Computational Cost

To validate the impact of the light field input distribution and the number of grouped aggregation channels on the computational performance and effectiveness of our network, we used three different combinations of horizontal, vertical, and diagonal distributions as input variables for the network, as well as varying numbers of cost aggregation groups as intermediate variables. The results are shown in Table 3.

**Table 3.** Quantitative comparison of results with different inputs and varying numbers of aggregation group channels.

| | **Input** | | **Number within Group** | **Average MSE (×100)** | **Average BidPix (0.07)** | **Time (s)** |
|---|---|---|---|---|---|---|
| **Horizontal** | **Vertical** | **Diagonals** | | | | |
| ✓ | | | 4 | 2.552 | 8.540 | 0.410 |
| | | ✓ | 4 | 4.445 | 10.132 | 0.673 |
| ✓ | ✓ | ✓ | 4 | 1.205 | 3.102 | 1.612 |
| ✓ | ✓ | | **4** | **1.159** | **3.203** | **0.693** |
| ✓ | ✓ | | 8 | 1.275 | 3.560 | 0.740 |
| ✓ | ✓ | | 1 | 2.148 | 4.169 | 0.613 |

Horizontal, Vertical, and Diagonals represent inputs at 0°, 90°, and two diagonal directions through the central subspace image in the light field array, respectively.

The number of image inputs has a significant impact on network performance. Increasing the data based on the distribution of horizontal and vertical directions results in a small gain in accuracy but a significant performance drop. As indicated in the fourth column of the table, the optimal number of grouped aggregation channels is four, and the number of groups has little impact on network performance. Overall, our method achieves optimal performance in the network with the choice of input data distribution and the number of groups.

### 4.3.3. Effectiveness of Cost Aggregation Network

To verify the role of the cost aggregation network within the overall architecture, we constructed a cost aggregation network by using the Resnet structure [50], as depicted in Figure 4, as a benchmark module for comparison. Considering memory limitations, the input cost volumes were uniformly resampled to the same dimensions $B \times 12 \times 2D \times H \times W$ before aggregation, followed by two 3D convolutions and eight 3D convolutions within the Resnet structure, and then depth regression. We trained for a total of 500 epochs. The results on the HCI light field benchmark dataset are shown in Table 4.

**Table 4.** Depth estimation results of cost aggregation module and resnet benchmark module on HCI dataset.

| | **Average MSE** | **Average BP (0.07)** |
|---|---|---|
| ResNet-based | 3.201 | 7.422 |
| Ours | **1.456** | **4.232** |

The depth estimation accuracy achieved by our cost aggregation network structure is higher than that of the comparison experiment, indicating that our proposed multi-scale cost aggregation module plays a significant role in improving accuracy.

Overall, our ablation study validates the approach we proposed, demonstrating that each modular component of our model makes a valuable contribution to the final results.

## 5. Discussions and Conclusions

Despite achieving commendable results in both disparity estimation accuracy and computational performance, our method still has certain limitations. First, our approach heavily relies on high-quality light field data, and its robustness to distortion and noise in light field images is weak, especially when the number of input data is limited. Consequently, the performance of our method might decrease when applied to real-world light field data. In the future, we could explore integrating specially designed network modules to mitigate the impact of distortion.

Secondly, our ablation studies reveal that the total number of input data significantly impacts computational performance. While optimizing the cost aggregation can reduce the number of parameters and thus computation time, it does not substantially affect overall performance. Simultaneously, the quality of cost volume construction directly influences accuracy in depth estimation in challenging areas. Future work could, therefore, focus on exploring better input structures and cost construction methods to balance computational performance and accuracy.

In this paper, we propose an end-to-end network architecture that trades off computational performance and depth estimation accuracy. Our feature dissimilarity cost construction method effectively compensates for the shortcomings of feature correlation, enhancing network accuracy in textureless areas. Moreover, our multi-scale cost aggregation architecture significantly improves depth estimation accuracy while maintaining good computational performance. Overall, compared with state-of-the-art methods, our approach achieves the best trade-off between computational performance and accuracy, as demonstrated on a broad HCI benchmark set and on a real-world light field dataset.

**Author Contributions:** Conceptualization, B.X. and S.P.; X.G. also contributed. Methodology, B.X.; H.H. provided assistance. Software, B.X., assisted by X.G. Validation, B.X. and S.P.; assisted by H.H. Formal analysis, B.X., with support from X.G. Investigation, B.X., assisted by H.H. Resources, X.G. Data curation, H.H. Writing—original draft, B.X., with contributions from X.G. Writing—review and editing, S.P.; X.G. and H.H. provided feedback. Visualization, B.X., assisted by H.H. Supervision, S.P. Project administration, S.P., with assistance from X.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** The support provided by China Scholarship Council (CSC) during a visit of Bo Xiao to University of Technology Sydney is acknowledged. This research by authors Xiujing Gao and Hongwu Huang was supported by Fujian Provincial Department of Science and Technology Major Special Projects (2023HZ025003), Key Scientific and Technological Innovation Projects of Fujian Province (2022G02008), Education and Scientific Research Project of Fujian Provincial Department of Finance (GY-Z220233), and Fujian University of Technology (GY-Z23027).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to the protection of intellectual property.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Overbeck, R.S.; Erickson, D.; Evangelakos, D.; Pharr, M.; Debevec, P. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Trans. Graph.* **2018**, *37*, 197:1–197:15. [CrossRef]
2. Yu, J. A Light-Field Journey to Virtual Reality. *IEEE MultiMedia* **2017**, *24*, 104–112. [CrossRef]
3. Guo, M.; Jin, J.; Liu, H.; Hou, J. Learning Dynamic Interpolation for Extremely Sparse Light Fields with Wide Baselines. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 2450–2459.
4. Meng, N.; Ge, Z.; Zeng, T.; Lam, E.Y. LightGAN: A Deep Generative Model for Light Field Reconstruction. *IEEE Access* **2020**, *8*, 116052–116063. [CrossRef]
5. Levinson, J.; Askeland, J.; Becker, J.; Dolson, J.; Held, D.; Kammel, S.; Kolter, J.Z.; Langer, D.; Pink, O.; Pratt, V.; et al. Towards fully autonomous driving: Systems and algorithms. In Proceedings of the 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, 5–9 June 2011; pp. 163–168. [CrossRef]
6. Wang, Z.; Chen, W.; Acuna, D.; Kautz, J.; Fidler, S. Neural Light Field Estimation for Street Scenes with Differentiable Virtual Object Insertion. In *Proceedings of the Computer Vision—ECCV 2022*; Lecture Notes in Computer Science; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland, 2022; pp. 380–397. [CrossRef]

7. Wang, Y.; Wang, L.; Liang, Z.; Yang, J.; An, W.; Guo, Y. Occlusion-Aware Cost Constructor for Light Field Depth Estimation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19777–19786. [CrossRef]
8. Chen, J.; Zhang, S.; Lin, Y. Attention-based Multi-Level Fusion Network for Light Field Depth Estimation. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1009–1017. [CrossRef]
9. Wang, Y.; Wang, L.; Wu, G.; Yang, J.; An, W.; Yu, J.; Guo, Y. Disentangling Light Fields for Super-Resolution and Disparity Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 425–443. [CrossRef] [PubMed]
10. Huang, Z.; Hu, X.; Xue, Z.; Xu, W.; Yue, T. Fast Light-field Disparity Estimation with Multi-disparity-scale Cost Aggregation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 6300–6309. [CrossRef]
11. Zhang, S.; Meng, N.; Lam, E.Y. Unsupervised Light Field Depth Estimation via Multi-view Feature Matching with Occlusion Prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 2261–2273. [CrossRef]
12. Woodham, R.J. Photometric Method For Determining Surface Orientation From Multiple Images. *Opt. Eng.* **1980**, *19*, 191139. [CrossRef]
13. Shi, B.; Wu, Z.; Mo, Z.; Duan, D.; Yeung, S.K.; Tan, P. A Benchmark Dataset and Evaluation for Non-Lambertian and Uncalibrated Photometric Stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Ju, Y.; Lam, K.M.; Xie, W.; Zhou, H.; Dong, J.; Shi, B. Deep Learning Methods for Calibrated Photometric Stereo and Beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–19. [CrossRef] [PubMed]
15. Diebold, M.; Jähne, B.; Gatto, A. Heterogeneous Light Fields. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1745–1753. [CrossRef]
16. Kim, C.; Zimmer, H.; Pritch, Y.; Sorkine-Hornung, A.; Gross, M. Scene Reconstruction from High Spatio-Angular Resolution Light Fields. *ACM Trans. Graph.* **2013**, *32*, 1–12. [CrossRef]
17. Schilling, H.; Diebold, M.; Rother, C.; Jähne, B. Trust Your Model: Light Field Depth Estimation with Inline Occlusion Handling. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4530–4538.
18. Li, J.; Lu, M.; Li, Z.N. Continuous Depth Map Reconstruction From Light Fields. *IEEE Trans. Image Process.* **2015**, *24*, 3257–3265. [CrossRef]
19. Hou, G.; Li, J.; Wang, G.; Yang, H.; Huang, B.; Pan, Z. A novel dark channel prior guided variational framework for underwater image restoration. *J. Vis. Commun. Image Represent.* **2020**, *66*, 102732. [CrossRef]
20. Chen, C.; Lin, H.; Yu, Z.; Bing Kang, S.; Yu, J. Light Field Stereo Matching Using Bilateral Statistics of Surface Cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
21. Heber, S.; Pock, T. Shape from Light Field Meets Robust PCA. In *Proceedings of the Computer Vision—ECCV 2014*; Lecture Notes in Computer Science; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 751–767. [CrossRef]
22. Jeon, H.G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.W.; So Kweon, I. Accurate Depth Map Estimation from a Lenslet Light Field Camera. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1547–1555.
23. Yu, Z.; Guo, X.; Ling, H.; Lumsdaine, A.; Yu, J. Line Assisted Light Field Triangulation and Stereo Matching. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2792–2799. [CrossRef]
24. Tao, M.W.; Hadap, S.; Malik, J.; Ramamoorthi, R. Depth from Combining Defocus and Correspondence Using Light-Field Cameras. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
25. Williem.; Park, I.K.; Lee, K.M. Robust Light Field Depth Estimation Using Occlusion-Noise Aware Data Costs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2484–2497. [CrossRef] [PubMed]
26. Tao, M.W.; Srinivasan, P.P.; Malik, J.; Rusinkiewicz, S.; Ramamoorthi, R. Depth from shading, defocus, and correspondence using light-field angular coherence. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1940–1948. [CrossRef]
27. Williem, W.; Park, I.K. Robust Light Field Depth Estimation for Noisy Scene with Occlusion. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016 ; pp. 4396–4404. [CrossRef]
28. Wanner, S.; Goldluecke, B. Variational Light Field Analysis for Disparity Estimation and Super-Resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 606–619. [CrossRef]
29. Zhang, S.; Sheng, H.; Li, C.; Zhang, J.; Xiong, Z. Robust depth estimation for light field via spinning parallelogram operator. *Comput. Vis. Image Underst.* **2016**, *145*, 148–159. [CrossRef]
30. Sheng, H.; Zhao, P.; Zhang, S.; Zhang, J.; Yang, D. Occlusion-aware depth estimation for light field using multi-orientation EPIs. *Pattern Recognit.* **2018**, *74*, 587–599. [CrossRef]
31. Tsai, Y.J.; Liu, Y.L.; Ouhyoung, M.; Chuang, Y.Y. Attention-Based View Selection Networks for Light-Field Disparity Estimation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12095–12103, Number: 07. [CrossRef]

32. Yang, X.; Deng, J.; Chen, R.; Cong, R.; Ke, W.; Sheng, H. Disentangling Local and Global Information for Light Field Depth Estimation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 3419–3427. [CrossRef]
33. Heber, S.; Pock, T. Convolutional Networks for Shape from Light Field. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3746–3754. [CrossRef]
34. Shin, C.; Jeon, H.G.; Yoon, Y.; Kweon, I.S.; Kim, S.J. EPINET: A Fully-Convolutional Neural Network Using Epipolar Geometry for Depth From Light Field Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
36. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
37. Guo, X.; Yang, K.; Yang, W.; Wang, X.; Li, H. Group-Wise Correlation Stereo Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2019.
38. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
39. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 66–75. [CrossRef]
40. Honauer, K.; Johannsen, O.; Kondermann, D.; Goldluecke, B. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields. In *Proceedings of the Computer Vision—ACCV 2016*; Lecture Notes in Computer Science; Lai, S.H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer: Cham, Switzerland, 2017; pp. 19–34. [CrossRef]
41. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss Functions for Neural Networks for Image Processing. *arXiv* **2018**, arXiv:1511.08861.
42. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
43. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
44. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2019; Volume 32.
45. Wang, T.C.; Efros, A.A.; Ramamoorthi, R. Occlusion-Aware Depth Estimation Using Light-Field Cameras. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3487–3495. [CrossRef]
46. Iwatsuki, T.; Takahashi, K.; Fujii, T. Unsupervised disparity estimation from light field using plug-and-play weighted warping loss. *Signal Process. Image Commun.* **2022**, *107*, 116764. [CrossRef]
47. Zhou, W.; Zhou, E.; Liu, G.; Lin, L.; Lumsdaine, A. Unsupervised Monocular Depth Estimation From Light Field Image. *IEEE Trans. Image Process.* **2020**, *29*, 1606–1617. [CrossRef] [PubMed]
48. Zhou, W.; Liang, L.; Zhang, H.; Lumsdaine, A.; Lin, L. Scale and Orientation Aware EPI-Patch Learning for Light Field Depth Estimation. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 2362–2367. [CrossRef]
49. Lin, L.; Li, Q.; Gao, B.; Yan, Y.; Zhou, W.; Kuruoglu, E.E. Unsupervised learning of light field depth estimation with spatial and angular consistencies. *Neurocomputing* **2022**, *501*, 113–122. [CrossRef]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.