文章编号：2095-2163(2024)02-0100-06 中图分类号：TP391.4 文献标志码：A

基于多级残差融合的复杂纹理光场图像深度估计

赵 以[1]，赵娟宁[2]，孙连山[1]
（1 陕西科技大学 电子信息与人工智能学院，西安 710021；2 陕西科技大学 物理与信息工程学院，西安 710021）

摘 要：光场的深度信息可以通过深度学习的光场深度估计算法计算，在图像视差、光场图像边缘以及光场图像的复杂纹理区域，获取高精度深度值仍然具有一定局限性。本文提出了一种用于光场图像深度估计的多级残差融合网络，通过组合残差模块提取多层次的残差特征，在保持网络深度的同时提升了网络对特征的表征能力。利用多级残差融合模块对多层次的残差特征进行融合，以获得包含浅层纹理信息和深层语义信息的融合特征。利用本文方法对 HCI4D 光场数据集进行处理，图像深度估计的均方误差指标达到 1.471，不良像素率指标达到 4.208，该实验结果表明本文方法在处理具有复杂遮挡的光场图像区域方面具有良好的处理效果。

关键词：光场图像；深度估计；组合残差模块；多级残差融合；复杂纹理

Depth estimation of complex textured light field image based on multi-level residual fusion

ZHAO Yi[1], ZHAO Juanning[2], SUN Lianshan[1]
(1 School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi′an 710021, China; 2 School of Physics and Information Engineering, Shaanxi University of Science and Technology, Xi′an 710021, China)

Abstract: The depth information of a light field can be computed using deep learning-based algorithms, yet it remains challenging to obtain high-precision depth values in areas with image parallax, edges in light field images, and complex textures. In this paper, we propose a multi-level residual fusion network for light field image depth estimation. This network leverages residual modules to extract multi-level residual features, enhancing the network′s ability to capture details while maintaining its depth. The multi-level residual fusion module is employed to combine these features, resulting in fusion features that encompass both shallow texture information and deep semantic details. We applied this method to process the HCI4D light field dataset, achieving a mean squared error index of 1.471 for image depth estimation and a bad pixel rate index of 4.208. Experimental results demonstrate that our approach effectively handles complex occlusion scenarios in light field image processing.

Key words: light field image; depth estimation; combined residual module; multi-stage residual fusion; complex textures

0 引 言

光场成像(Light Field Imaging，LFI)是一种与传统成像不同的计算光学成像技术[1]。利用光场相机获取初始图像，并通过计算处理获得光场图像。光场成像系统提供了额外的角度信息，使得图像不仅具有二维信息，还包含了场景中的角度信息。

深度估计是目标检测、三维重建等图像处理任务的基础，通过准确估计场景的深度，可以实现更精确的目标检测和三维重建，从而推动计算机视觉和图像处理领域的发展[2]。

根据光场成像原理及光图像成像的数据形式，将原始光场图像图进行数据重构，可获得不同对准深度的目标图像，不同视角的聚焦图像及相邻子孔径图像(Sub-Aperture Image，SAI)匹配处理的极平面图像(Epipolar Plane Image，EPI)，以上图像作为信号输入，利用视差法、灰度法、极平面法，可获得光场图像的深度信息。

由于光场成像系统对特定空间进行了重复采样，光场相机的空间分辨率和角度分辨率相互制约，

基金项目：陕西省自然科学基础研究计划资助项目(2023-JC-YB-581)。

作者简介：赵 以(2000-)，女，硕士研究生，主要研究方向：光场图像深度估计。
通讯作者：孙连山(1977-)，男，博士，副教授，硕士生导师，主要研究方向：数据起源及区块链技术、人工智能技术应用研究。Email：sunlianshan@sina.com

收稿日期：2023-10-24

所以光场图像存在冗余信息和空间分辨率的损失，但两者之间又具有特定的联系。为了获得空间分辨率较高的深度图像，需要建立更为精准的模型来对图像空间信息和深度信息进行恢复。Tao 等[3]利用光场相机的特性，通过分析图像中的焦散效果和对应关系，来推断场景中物体的深度信息；Jeon 等[4]提出了一种基于相位的多视点立体匹配方法，实现了傅里叶域深度估计；Wang 等[5]提出了一种基于部分角度一致性的遮挡感知算法；Zhang 等[6]通过 EPI 斜率估计局部深度信息，并计算其置信度，对于置信度较低的像素，采用局部线性嵌入方法计算视差，得到深度图；William 等[7]提出了角熵代价和自适应离焦代价来处理深度估计的噪声和遮挡问题，这些代价函数能够有效地降低噪声的影响，并提高存在遮挡情况下的深度估计的质量；Han 等[8]提出了一种遮挡感知投票成本，以保留深度图中的边缘。

近年来，卷积神经网络（CNN）深度学习模型从大规模的数据集中学习图像中纹理和深度之间的关系，能够捕捉到更复杂的纹理特征。Heber 等[9]使用 CNN 提取 EPI 的特征，进行深度估计，EPI 仅沿一个方向生成，因此深度估计结果的置信度在一定程度上受到限制。Luo 等[10]提出 EPI 十字网络输入，在图像预处理中减少了内存的占用；在此基础上，Shin 等[11]提出在光场的多视角图像阵列中选取水平、垂直、斜 45°等 4 个方向的图像，使用多层的全卷积网络进行深度估计；随后，Tsai 等[12]在网络中提出加入视图选择模块，根据不同场景生成注意力视图，来更好地适应场景的深度特征；Zhou 等[13]通过学习 EPI-patch（Epipolar Plane Images 的局部区域）来进行光场图像的深度估计。

本文针对纹理多样性和复杂性所带来的深度信息匹配不准确的问题，提出了一个多级残差融合的网络，该深度学习网络具有以下特点：

（1）设计组合残差模块（Composed Residual Block，CRB），提取多层次的残差特征，在保持网络深度的同时提升了网络对特征的表征能力，获取更丰富的图像上下文信息，更准确地进行深度估计；

（2）设计了多级残差融合模块（Multi-level Feature Fusion Block，MFFB），对多层次的残差特征进行融合，获得包含浅层纹理信息和深层语义信息的融合特征，通过增加图像上下文信息来扩展特征提取的范围，减少特征提取过程中有效信息的损失。

1　多级残差融合的光场图像深度估计

1.1　多级残差融合网络结构

视差是光场深度估计的核心，对于 4D 光场的子孔径图像，其中心子孔径图像 $L(x,y,0,0)$ 与相邻视图 $L(x,y,u,v)$ 之间的关系可以用式（1）表示：

$$L(x,y,0,0)=L(x+d(x,y)\times u,y+d(x,y)\times v,u,v) \tag{1}$$

其中，(x,y) 代表空间坐标；(u,v) 代表角度坐标；$d(x,y)$ 是中心视点像素与其相邻视点中相应像素的视差。

视差值的计算是基于中心视角像素点与其他视角像素点之间的差异，通过比较其位置偏移来确定像素之间的深度关系。这个偏移量可以表示为视差图，提供了场景中不同物体的深度信息。

由于光场相机的子孔径基线相对较窄，光场在不同角度上的投影图像之间存在更为广泛的信息覆盖，从而可以利用不同角度的光场图像的匹配信息进行深度估计。现有方法在不同角度的光场图像的纹理匹配、噪声去除以及计算复杂度等方面都面临各种挑战，降低了面向复杂纹理的光场深度估计的准确率。

为了解决光场图像复杂纹理信息的深度估计问题，本文提出了一个多级残差融合网络（Multi-level Residual Fusion Network，MRFNet）。首先，该网络在不同阶段利用组合残差模块提取不同层次的残差特征，获取更丰富的图像上下文信息，从而更加充分地对复杂纹理进行建模；其次，不同角度的光场图像还存在较多的冗余信息，在一定程度上降低了光场图像的信息独立性和有效性。

为了使网络在加深时仍然保留丰富的有效信息，本文进一步利用多级特征融合模块对不同层次的特征信息进行融合，以充分利用光场图像的浅层纹理细节信息和深层语义信息。本文提出的多级残差融合网络旨在改善复杂纹理区域的深度估计结果，其网络结构如图 1 所示。

网络结构以 $u\times v$ 的光场图像作为 MRFNet 输入，通过组合残差模块 CRB 进行特征提取，通过多级特征融合模块 MFFB 进行特征融合，最后通过代价构造和代价聚合进行深度图回归。

1.2　组合残差模块

为了提取光场图像的复杂纹理特征，需要建立更深层次的网络以提高网络的表征能力和性能。深层次的网络容易在训练过程中出现梯度消失或梯度爆炸的问题。为了克服这些问题，He 等[14]提出了

残差网络 ResNet,通过跳跃连接建立输入与输出的直接连接,允许梯度直接传递到更浅的层级,从而允许网络更容易地学习输入与期望输出之间的差异,即残差。

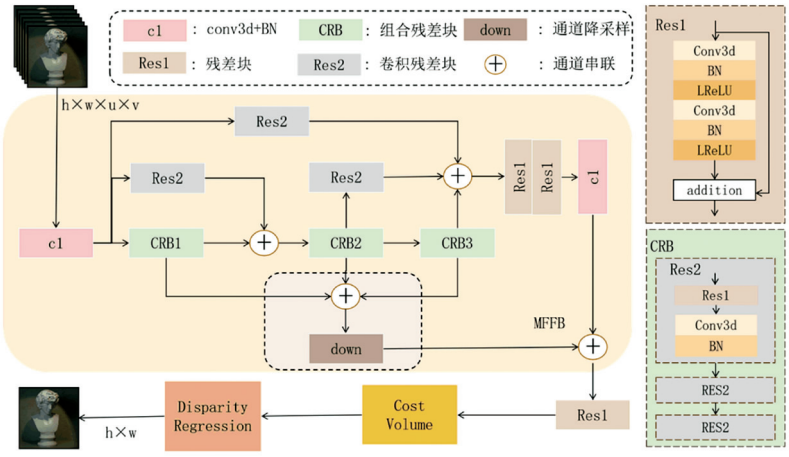

图 1 多级残差融合网络结构图

Fig. 1 Structure diagram of multi-level residual fusion network

基于残差网络 ResNet 的特点,本文提出了组合残差模块 CRB,通过将多个残差单元相互组合,增加网络的复杂性和表达能力。每个残差单元可以提取不同层级的特征,通过组合这些残差单元,网络可以更好地学习到复杂的特征表示,通过跨层级信息的传递来帮助优化网络的训练。

对于组合残差模块的输入 F_{in},经过 3 个级联的卷积残差块 ConvRes 得到输出 F_{out},其表示形式为式(2):

$$F_{out} = \mathrm{ConvRes}(\mathrm{ConvRes}(F_{in})) \tag{2}$$

其中,ConvRes 表示基于卷积模块的残差单元。

卷积残差块的表示形式,式(3)和式(4):

$$Y_{mid} = \mathrm{ConvBNR}(\mathrm{ConvBNR}(X)) \tag{3}$$

$$Y = \mathrm{ConvBN}(X + Y_{mid}) \tag{4}$$

其中,X 表示卷积残差块的输入;Y 表示卷积残差块的输出;ConvBNR 表示带有批量归一化 BN 和 LReLU 激活的 3D 卷积模块;ConvBN 表示仅带有 BN 的 3D 卷积模块。

1.3 多级特征融合模块

为了提高光场图像深度估计的精度、鲁棒性和质量,本文提出了多级特征融合模块 MFFB 来帮助网络捕捉多尺度信息、处理视角差异以及减轻噪声和伪影问题,从而提升深度估计的性能。首先,不同层级的特征可以提供不同细节和上下文信息,通过将这些特征进行融合,可以使模型更好地理解图像中的结构和纹理信息;其次,通过融合不同层级的特征,帮助提取光场图像不同视角下图像序列的一致性特征表示。多级特征融合模块还可以帮助减轻光场图像中的噪声和伪影。

假设光场图像经过卷积块 c_1 后的特征为 F,则经过 3 个组合残差模块 CRB 得到的特征分别表示为式(5)~式(7):

$$F_1 = \mathrm{CRB}(F) \tag{5}$$

$$F_2 = \mathrm{CRB}(F_1 + \mathrm{Res}(F)) \tag{6}$$

$$F_3 = \mathrm{CRB}(F_2) \tag{7}$$

为了融合不同层级的残差特征,本文通过通道串联的方式,在两个维度对多级残差特征进行融合,其表示形式分别为式(8)和式(9):

$$F_{down} = \mathrm{Concat}(F_1, F_2, F_3) \tag{8}$$

$$F_{mid} = \mathrm{Concat}(\mathrm{ConvRes}(F), \mathrm{ConvRes}(F_2), F_3) \tag{9}$$

其中,Concat 表示通道串联操作,ConvRes 表示卷积残差单元。

通过融合不同层次的特征,获得包含浅层纹理信息和深层语义信息的组合特征。最后,将两个维度的特征进行融合,用于后续的代价构造,代价构造的特征 Flast 形式如式(10)所示:

$$\mathrm{Flast} = \mathrm{Concat}(F_{down}, c_1(\mathrm{Res}(F_{mid}))) \tag{10}$$

其中,Res 表示残差单元。

1.4 代价构造模块

对融合后的特征进行匹配代价构造,用于光场图像的深度估计。代价构造模块的构造是指特征提取和融合后的特征被组织成子孔径图像数组。参考遮挡感知的代价体构造方式,通过一系列卷积核大

小为 $U*V$ 且具有不同膨胀率的卷积，融合不同视差下的图像块特征，作为代价体[15]。在模型推理阶段，首先生成一个初始的遮挡掩码模块，并使用该初始的掩码模块来进行深度估计；获得初始的视差图后，再用此视差图更新遮挡掩码，生成更准确的模块。以此类推，通过多次迭代操作得到最终结果。

1.5　代价聚合与回归

将输入的特征图应用 1×1 卷积，使通道深度由 512 降至 160。将 8 个三维卷积层级联在一起，每个卷积层用 3×3×3 的内核进行成本聚合；从第 3 到第 6 个 3D 卷积层，每个层都包含两个残差块，这些块的目的是在三维卷积后通过通道注意层来突出通道的贡献；最后，通过第 8 个三维卷积层生成一个三维张量 $F \in \mathbb{R}^{D\times H\times W}$，该张量包含了按照视差排列的信息。估计的中心视图视差按照式(11)进行回归：

$$D_c = \sum_{d_k=d_{\min}}^{d_{\max}} d_k \times \mathrm{Softmax}(F) \tag{11}$$

其中，d_k 为深度范围中的可能值，Softmax(·)表示沿视差轴进行的 Softmax 归一化。

2　实验及分析

2.1　训练细节

实验中采用了 4D 光场数据集 HCI，该数据集包含 28 组数据，选择 16 组数据作为训练样本。为了增加训练数据，执行多种数据增强，包括随机翻转和旋转、亮度和对比度调整、噪声注入、重新聚焦和下采样。在训练过程中，网络以 $L1$ 损失函数进行监督，并采用 Adam 方法进行优化，其中 β_1 设为 0.9，β_2 设为 0.999，批量大小设置为 1，学习率设置为 1×10^{-3}。实验的运行环境为 NVIDIA GTX3090Ti GPU，并使用 pytorch 框架作为后端，整个训练过程大约耗时 4 天。

实验结果以均方误差(Mean Square Error，*MSE*)，不良像素率(Bad Pixel，*BP*)作为评价指标，用于评价预测值与真实值之间的差异，式(12)和式(13)：

$$MSE = \frac{1}{m}\sum (y_i - \hat{y}_i)^2 \tag{12}$$

$$BP(t) = \left\{\frac{y_i \in m : |y_i - \hat{y}_i > t|}{m}\right\} \tag{13}$$

其中，取 $t = 0.07$，y_i 表示第 i 个像素的真实值；$\hat{y}_i$ 表示初始估计值；m 为视差像素点的总数。

2.2　实验分析

实验中效果评价的测试集采用了 4D 光场数据集 HCI 中包含 8 个场景的的测试集。

本文将方法 LF[16]、LF_OCC[5]、CAE[7]、SPO[6]、深度学习的方法 EPINET-fcn[11] 以及 MRFNet 在 4 个真实场景下的深度图处理结果进行分析。

在 4 个真实场景下的深度图如图 2 所示。在 cotton、dino、sideboard 3 个场景中，上述方法视觉效果几乎一致且误差较少，仅在 sideboard 场景下，出现少量明显误差，基于深度学习的方法在这类场景下表现比传统方法更好。

在 boxes 场景中，本文算法和其他算法都出现了明显误差，特别是在箱子的网格区域。由于箱子内部存在大量密集复杂的纹理和遮挡，基于深度学习的方法在 boxes 场景中复杂纹理的区域表现得比传统方法更好，本文算法在该区域的边缘处理结果更加清晰。

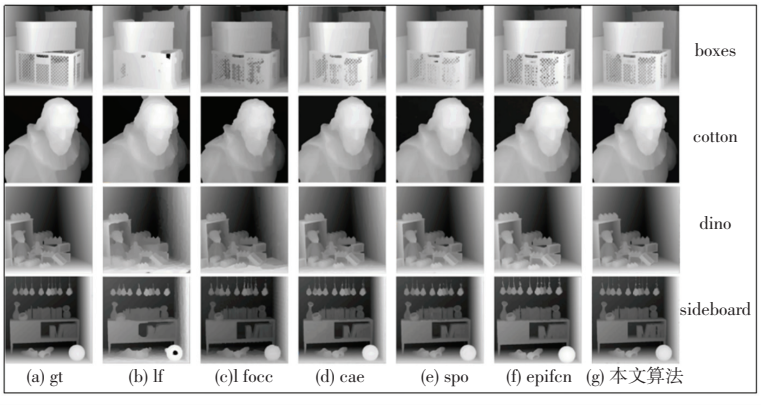

图 2　测试数据集下 4 个场景深度图

Fig. 2　Depth maps from four scenes in test dataset

不同方法场景下的坏点图如图 3 所示，$BP > 0.07$ 显示为红色，反之为绿色。从图 3 的坏点图可

得出,在遮挡较少的图像中,传统方法的处理较好,在较为复杂的遮挡区域,深度学习的方法能更准确的找出像素间深度差异,在处理复杂纹理时主观效果较好,但在边缘、以及复杂纹理区域边缘较多的位置仍存在模糊。

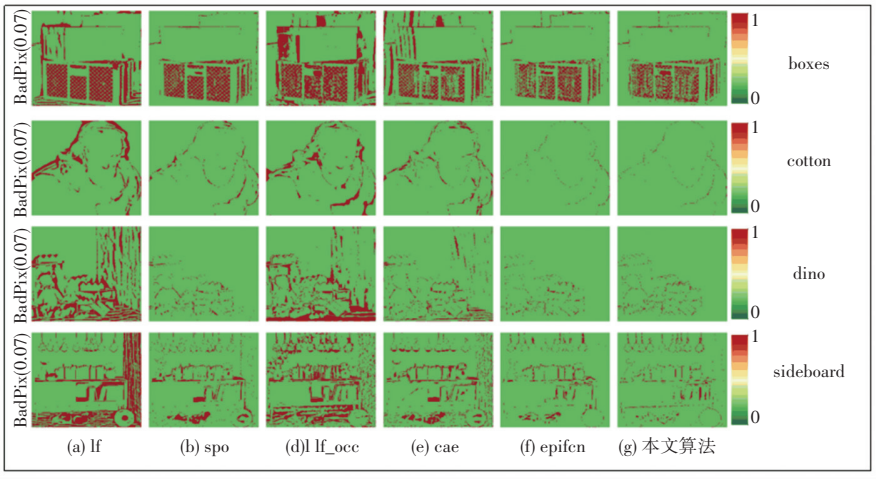

图 3 测试数据集下 4 个场景坏点图

Fig. 3 Dead pixel plots from four scenes in test dataset

不同方法的深度图 *mse* * 100 指标对比见表 1,不同方法的深度图 BP7%对比见表 2。表 1 和表 2 中最后一列(Avg)表示所有场景的平均值。黑色加粗为最佳值,下划线表示次佳。

由表 1 和表 2 可得出以下结论:本文方法在大部分场景下的 *MSE* 指标表现最优。在具有大量边缘和复杂纹理区域的场景(如 sideboard、backgammon、boxs),本文方法的坏点像素较少,主观图像效果明显优于其他方法。*MSE* 指标在大部分场景上达到最优,仅在少部分场景为次优。

表 1 不同方法的深度图 *MSE* * 100 指标对比

Table 1 Comparison of *MSE* * 100 indicators of depth maps with different methods

Method	boxes	cotton	dino	sideboard	backgammon	pyramids	stripes	dots	Avg
lf	17.43	9.168	1.163	5.071	13.01	0.273	17.45	5.676	8.655
spo	9.107	1.313	0.311	1.024	4.587	0.043	6.955	5.238	3.572
lf_occ	9.593	1.074	0.944	2.073	22.78	0.077	7.942	3.301	5.973
cae	8.427	1.506	0.382	0.876	6.074	0.048	3.556	5.082	3.244
epinetfcn	4.189	0.287	0.336	0.778	**3.411**	0.016	1.744	14.48	3.155
本文	**3.750**	**0.220**	**0.130**	**0.618**	4.480	**0.008**	**1.125**	**1.440**	**1.471**

表 2 不同方法的深度图 *BP7%*对比

Table 2 Depth plots *BP7%* comparison of different methods

Method	boxes	cotton	dino	sideboard	backgammon	pyramids	stripes	dots	Avg
lf	23.020	7.829	19.03	21.98	5.516	12.35	35.74	2.900	16.05
spo	15.890	2.594	2.184	9.297	3.781	0.861	14.98	16.27	8.233
lf_occ	26.520	6.218	14.91	18.49	19.07	3.172	18.41	5.822	14.08
cae	17.880	3.369	4.968	9.845	3.924	1.681	7.872	12.40	7.742
epinetfcn	**12.839**	**0.508**	**1.286**	4.801	**3.580**	**0.192**	**2.462**	3.183	**3.606**
本文	13.320	0.590	1.550	**4.157**	5.580	0.320	5.669	**2.480**	4.208

2.3 消融实验

消融实验采用了 4D HCI 光场数据集,并使用 *MSE* 作为评估指标对多个 CRB(M_CRB)和 MFFB 在深度估计中的影响进行了定量分析。消融实验结果见表 3,当使用 M_CRB 时,网络的均方误差指标显著降低,说明多级残差在光场特征提取方面对于

产生更好的边缘细节非常有帮助。当结合 MFFB 时，指标再次下降，证明两种方法融合所产生的特征提取方式更加有效，显著提高了深度估计的性能。

表 3　模块消融实验的定量比较

Table 3　Quantitative comparison of module ablation

Method	*MSE*
Base	2.26
M_CRB	1.67
MFFB	1.78
M_CRB+MFFB	1.47

3　结束语

针对光场深度估计问题，本文提出了一种用于光场图像深度估计的多级残差融合网络，其中包含多尺度残差模块和多层次信息融合模块。通过结合多级残差和多层次特征融合模块，获得更清晰的深度图边缘，并有效解决随着网络深度增加而导致的有效信息损失问题。多层次特征融合模块采用了多尺度特征的融合，将浅层特征和复杂信息特征进行融合，提高了深度估计的准确性。实验结果表明：相较于传统方法，本文提出的方法在处理复杂纹理区域时能够实现更准确的深度估计。未来的工作将聚焦于改进弱纹理区域的深度估计，以提高整体光场深度估计的精度。

参考文献

[1] DUN X, FU Q. Advances in computational imaging [J]. Chinese Journal of Image and Graphics, 2022, 27 (6): 1840-1876.

[2] DENG H P, SHENG Z C. Depth estimation of light field images based on semantic guidance [J]. Acta Electronica and Informatics, 2022,44 (8): 2940-2948.

[3] TAO M W, HADAP S, MALIK J, et al. Depth from combining defocus and correspondence using light-field cameras [C]// Proceedings of the IEEE International Conference on Computer Vision. 2013: 673-680.

[4] JEON H G, PARK J, CHOE G, et al. Accurate depth map estimation from a lenslet light field camera [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1547-1555.

[5] WANG T C, EFROS A A, RAMAMOORTHI R. Occlusion-aware depth estimation using light-field cameras [C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3487-3495.

[6] ZHANG S, SHENG H, LI C, et al. Robust depth estimation for light field via spinning parallelogram operator [J]. Computer Vision and Image Understanding, 2016, 145: 148-159.

[7] PARK I K, LEE K M. Robust light field depth estimation using occlusion-noise aware data costs [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(10): 2484-2497.

[8] HAN K, XIANG W, WANG E, et al. A novel occlusion-aware vote cost for light field depth estimation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(11): 8022-8035.

[9] HEBER S, YU W, POCK T. Neural epi-volume networks for shape from light field [C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2252-2260.

[10] LUO Y, ZHOU W, FANG J, et al. Epi-patch based convolutional neural network for depth estimation on 4d light field [C]//Proceedings of the 24th International Conference on Neural Information Processing. Guangzhou, China: Springer International Publishing, 2017: 642-652.

[11] SHIN C, JEON H G, YOON Y, et al. Epinet: A fully-convolutional neural network using epipolar geometry for depth from light field images [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4748-4757.

[12] TSAI Y J, LIU Y L, OUHYOUNG M, et al. Attention-based view selection networks for light-field disparity estimation [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 12095-12103.

[13] ZHOU W, LIANG L, ZHANG H, et al. Scale and orientation aware epi-patch learning for light field depth estimation [C]// Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018: 2362-2367.

[14] TARG S, ALMEIDA D, LYMAN K. Resnet in resnet: Generalizing residual architectures [J]. arXiv preprint arXiv:1603.08029, 2016.

[15] WANG Y, WANG L, LIANG Z, et al. Occlusion-aware cost constructor for light field depth estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 19809-19818.

[16] JEON H G, PARK J, CHOE G, et al. Accurate depth map estimation from a lenslet light field camera [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1547-1555.