第 43 卷 第 21 期/2023 年 11 月/光学学报 研究论文

光学学报

基于邻域像素注意力机制的光场深度估计方法

林曦，郭阳，赵永强*，姚乃夫

西北工业大学自动化学院，陕西 西安 710129

摘要 通过发掘深度信息与子孔径图像邻域像素间的高度相关性，提出了一种基于邻域像素注意力机制的光场深度估计方法。首先根据光场图像的数据特性提出了一种邻域像素注意力机制，该注意力机制考虑了不同子孔径图像在同一邻域间的极几何关系，能够增强网络对遮挡像素的感知能力。其次基于注意力机制设计了一个光场子孔径图像序列特征提取模块，该模块通过三维卷积将相邻序列图像上的特征编码到特征图上，并通过注意力机制增强网络对光场图像极几何特征的学习能力。最后联合邻域像素注意力机制和特征提取模块设计了一个多分支的全卷积神经网络，该网络使用部分光场子孔径图像序列即可估计图像的深度特征。实验结果表明，所提方法在均方误差(MSE)和平均坏像素率(BP)指标上总体表现优于其他先进方法，同时得益于高效注意力机制的加入，与其他先进方法相比所提方法运行速度最快。

关键词 光场图像；深度估计；邻域像素；注意力机制；神经网络

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS230786

1 引 言

光场(Light field)是一种空间中光线集合的完备表示，能够描述空间中任一条光线在不同时间下的位置、角度和波长信息，这为深度估计提供了大量的线索[1]。但是整个光场的数据难以采集且数量庞大，Levoy等[2]考虑到目前传感器的限制，将自由空间中的光场简化为四维，提出了光场双平面模型；基于该模型，Ren等[3]在常规相机的主透镜和感光元件之间放置一片微透镜阵列，实现以常规相机的体积来采集双平面光场图像，为光场成像技术的广泛应用奠定了基础。作为一种高效的被动深度估计方案，光场深度估计已经被成功应用于姿态估计[4]、显著性检测[5]、三维重建[6]、粒子图像测速[7]等领域。尽管光场数据隐含了深度信息，但是从光场图像中提取深度信息仍面临精度低、速度慢、遮挡场景误差大等问题[8-9]。

现有的光场深度估计方法大致可分为基于优化与基于学习两种：1)基于优化的方法主要依靠人为构建成本度量函数并计算出图像大致的深度图，再利用全局优化、局部平滑等算法对深度图进一步优化以提高得到的深度图的质量。Yu等[10]将三维双线性子空间映射到线性约束并使用LAGC(Line-assisted graph-cut)算法进行立体匹配。Tao等[11]提出使用MRF(Markov random fields)框架融合离焦线索和对应视差的线索来估计深度，并额外引入阴影线索用于改善物体的形状细节。Zhang等[12]借助在对极平面图像(Epipolar plane image，EPI)上构建旋转平行四边形算子(SPO)来测量极线的斜率。Johannsen等[13]采用了EPI块来构造EPI斜线字典，并使用各向异性平滑对结果进行优化。基于优化的方法在纹理明显且连续的情况下，效果较好，但在受遮挡或噪声影响时，该类方法难以获得准确的深度，同时还存在耗时长的问题。2)基于学习的方法是利用卷积神经网络强大的特征提取能力，放弃手动设计匹配模式而让网络学习出所需的拟合函数，实现端到端的深度图预测，这也是目前光场深度估计的主要研究方向。Heber等[14]建立了一个编码-解码与变分优化相结合的U型网络，该网络使用了3D卷积来提取空间信息和图像序列信息，但没有考虑到光场数据中存在的冗余，导致算法计算量大、运行速度慢；Shin等[15]在光场的子孔径图像中根据极几何线索使用了4个方向的图像，最后通过聚合网络来合并特征，但没能充分挖掘不同通道邻域像素间的相关性，导致深度估计的精度较低；Tsai等[16]考虑到光场图像中存在的冗余，引入注意力机制对子孔径图像赋予不同权重并构造了代价体(Cost volume)进行深度匹配，但是该结构需要大量的内存来存放代价计算的结果，后续的视差回归也导致算法运算缓慢；Wang等[17]引入膨胀卷积以构造视差匹配代价体，并使用了一种像素调制方法解决不同子孔径图像中存在的遮挡问题，但该调制方法对遮挡像素的建模仍不够充分。由

收稿日期：2023-04-07；**修回日期**：2023-05-18；**录用日期**：2023-06-26；**网络首发日期**：2023-07-14

通信作者：*zhaoyq@nwpu.edu.cn

上述文献可知，基于学习的方法仍存在以下挑战：场景中的遮挡导致了光场EPI极线不连续，使得遮挡处的深度预测结果不可信；光场深度估计需要进行亚像素层面的预测，现有方法的性能和运行速度仍有待提升。

针对上述挑战，通过分析光场图像特性，本文提出一种基于邻域像素注意力机制的光场深度估计方法：1)针对光场深度估计任务的特性以及子孔径图像序列的特点，利用光场中某一像素点的深度信息与该像素点周围有限邻域像素点的相关性，提出邻域像素注意力机制Mix Attention，联合空间与通道注意力高效地建立了特征图与深度的关系，提高了光场深度估计的精度，并使网络具有一定的抗遮挡能力；2)基于该注意力机制提出一种序列图像特征提取模块，利用三维卷积将子孔径图像序列包含的空间与角度信息编码到特征图中，并使用Mix Attention调整权重；3)提出一种多分支深度估计网络，使用光场部分子孔径图像作为输入，实现了任意尺寸光场图像快速的端到端的深度估计。运用所提方法对光场数据集New HCI[18]进行测试，验证了该方法在坏像素率(BP)、均方误差(MSE)和计算时间3个性能指标上的优越性。

2 基于邻域像素注意力机制的光场深度估计方法

2.1 光场子孔径图像分析

光场双平面模型用两个相互平行的平面来表示四维光场，通过双平面的架构来表述光场的空间分布和角度分布信息，即$L=L(u,v,s,t)$，式中u、v和s、t分别表示光线穿过两个平面的空间坐标。

子孔径阵列图像可以通过微透镜、相机阵列等方式获得，是双平面模型下光场图像数据可视化的表示形式。子孔径阵列在获取的过程中固定了相机平面的两个坐标，更侧重于反映光线的空间分布信息。通过固定相机平面位置(u,v)，令$u=u^*$，$v=v^*$，则$L(u^*,v^*,s,t)$可表示在视点位置(u^*,v^*)处对应的图像(s,t)，光场数据以子孔径图像阵列的形式实现可视化，如图1所示，其中右侧放大图为不同位置的子孔径图像。

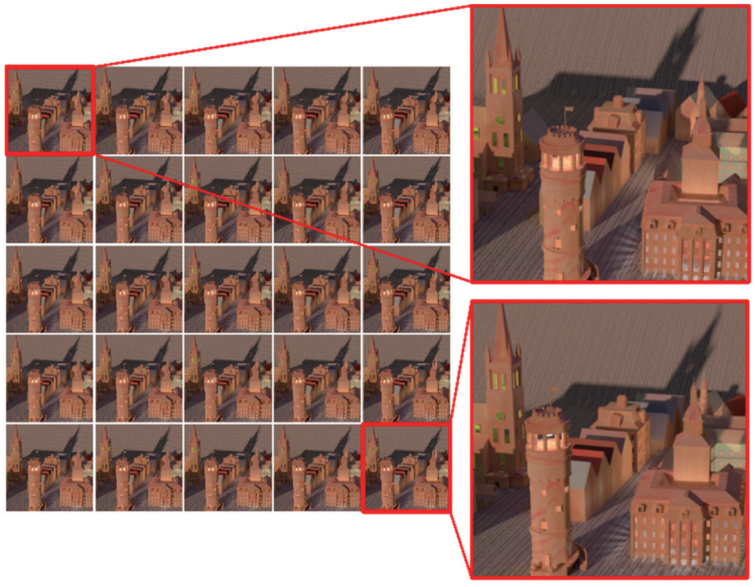

图1 光场子孔径图像阵列

Fig. 1 Subaperture image array of light field

现有的基于子孔径图像的深度估计方法，没有考虑子空间图像间的邻域像素相关性，这使得深度估计算法不能充分提取图像特征，并且难以排除场景中的遮挡对目标像素点的干扰[16]。图2通过圆点标出了场景中的某个特定点在不同子孔径图像中关于s和t轴产生的位移，由于光场图像的基线较窄，同一个像素点在相邻子孔径图像中的位移被局限在特定的邻域范围内，显示出邻域相关性。所提出的邻域像素注意力机制Mix Attention可以高效且有针对性地捕捉到这种相关性。

2.2 邻域像素注意力机制Mix Attention

目前，注意力机制已被应用于光场视觉，但通用的注意力机制得到的注意力权重都是逐通道或逐像素的，并没有关注到像素点在不同特征图上的邻域。由于光场子孔径间的基线较短，光场中心视角中像素点的深度信息几乎只与该像素点在其他光场视角上同一位置周围有限邻域的像素点有关。基于上述原因，提出了一种基于邻域像素的注意力机制Mix Attention，如图3所示。在图3中，FC为全连接层，LReLU为Leaky ReLU激活函数，C为特征图通道数，H为特征图的高，W为特征图的宽。

对于一组特征图上位置(x,y)的一组像素$F(x,y)\in\mathbb{R}^{C\times1\times1}$，取其3×3邻域做平均池化和最大池化操作。平均池化聚合了邻域包含的空间信息，当场景存在遮挡时，该操作有助于特征图像剔除无关像

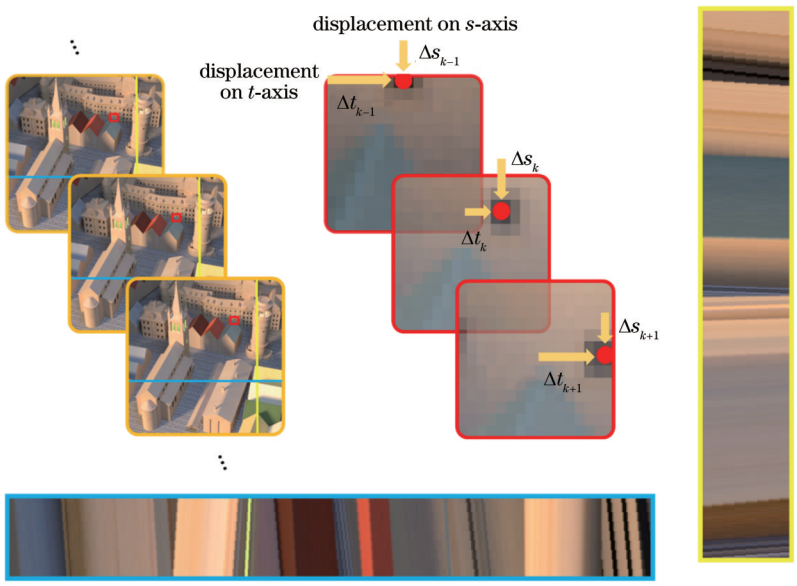

图 2　子孔径图像序列邻近像素示意图

Fig. 2　Schematic diagram of adjacent pixels of subaperture image sequence

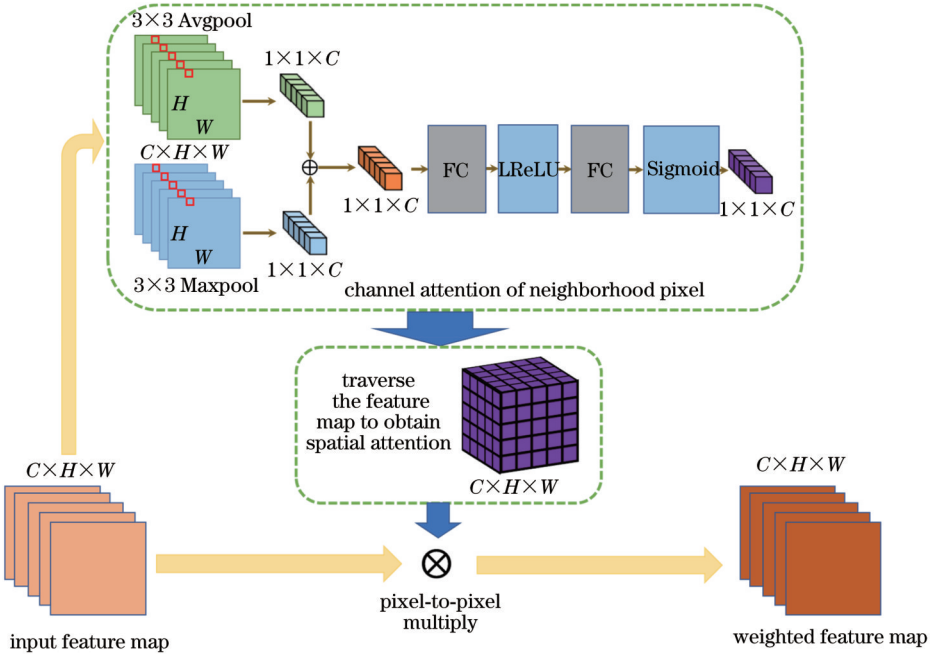

图 3　邻域像素注意力机制示意图

Fig. 3　Diagram of Mix Attention

素；最大池化能够捕捉到对应于卷积核的判别性特征，使网络能够判断不同特征图中极线的走向，

$$F_{\mathrm{Avg}}=P_{\mathrm{Avgpool}}\left[N_8(x,y)\right],F_{\mathrm{Max}}=P_{\mathrm{Maxpool}}\left[N_8(x,y)\right],\tag{1}$$

式中：F_{Avg}和F_{Max}分别表示平均池化和最大池化操作得到的结果；P_{Avgpool}和P_{Maxpool}分别表示平均池化和最大池化操作；$N_8(x,y)$表示像素点(x,y)的8邻域。将两种池化操作得到的结果相加再经过一个两层的多层感知机MLP（Multi-layer perceptron；M_{MLP}），得到该组像素的权重为

$$F'_{\mathrm{Attention}}(x,y)=M_{\mathrm{MLP}}\left(F_{\mathrm{Avg}}+F_{\mathrm{Max}}\right)。\tag{2}$$

使用上述步骤遍历整幅特征图就得到了全图注意力$F'_{\mathrm{Attention}}\in\mathbb{R}^{C\times H\times W}$。邻域像素注意力机制Mix Attention的计算过程可以表示为

$$F'_{\mathrm{Attention}}(x,y)=\sigma\left\{W_1\times\sigma\left\{W_0\left[f_{\mathrm{Avg}}(x,y)\right]\right\}+W_1\times\sigma\left\{W_0\left[f_{\mathrm{Max}}(x,y)\right]\right\}\right\},\tag{3}$$

式中：W_1，$W_0\in\mathbb{R}^{C\times C}$为多层感知机的权重；$f_{\mathrm{Avg}}$和$f_{\mathrm{Max}}$分别为对邻域进行的平均池化和最大池化操作；$\sigma$为Sigmoid激活函数。从式(3)可以看出，Mix Attention

在像素邻域层面上为通道注意力，遍历整幅特征图得到的是空间注意力图。

对于特征图边界上的像素点，采取零填充的方式进行池化操作，这在一定程度上会影响注意力模块的结果，所以训练时使用较大尺寸的图像块可以减少误差。由文献[19]可知，加一层用以降维的全连接层不能提升注意力模块的表征性能，反而会导致网络性能下降，因此没有使用 Bottleneck，而是利用所提方法设计的邻域像素注意力机制 Mix Attention 中的多层感知机进行降维。通过有效利用邻域信息，该邻域像素注意力机制 Mix Attention 所获得的注意力仅取决于目标点周边的像素，能够最大程度地挖掘光场的子孔径图像间的极几何信息。

2.3　序列图像特征提取模块

由于所提网络使用的是光场子孔径图像序列，为了让注意力模块能够感知到不同子孔径图像上的特征信息，提出一种利用三维卷积编码相邻子孔径图像序列的特征提取模块。首先使用“3D 卷积-批归一化-3D 卷积-批归一化-激活函数”结构将相邻图像的信息编码到特征图中，对于编码产生的多组序列特征图，采用加权融合的方式将每组序列特征图融合为单幅特征图，该权重由网络自适应学习得到。再对融合后的特征图进行 Mix Attention 注意力计算即可得到对应每组特征图的注意力，流程如图 4 所示。

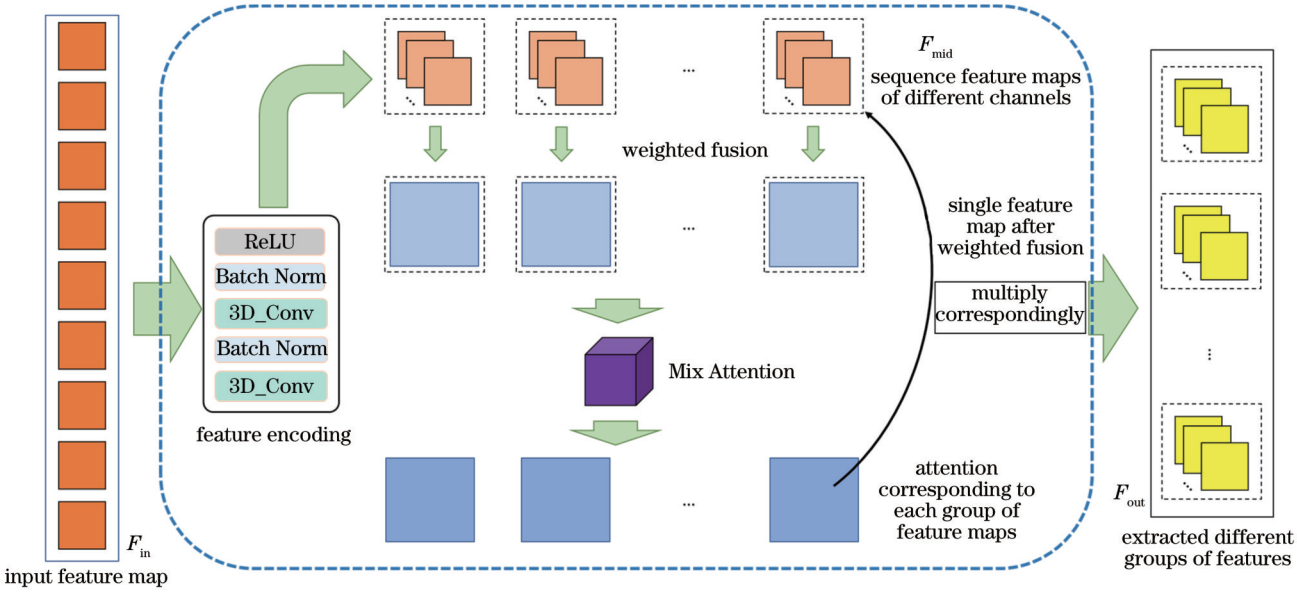

图 4　序列图像特征提取模块

Fig. 4　Sequence image feature extraction module

该模块的输入为一组或多组序列图像，提取出的特征为多组不同通道的序列图像，该模块可表示为

$$\begin{cases} F_{mid} = f_{in}(F_{in}, \theta_i) \\ F_{out} = F_{mid} \times F'_{Attention}(W_3 \times F_{mid}) \end{cases}, \tag{4}$$

式中：F_{in}、F_{mid}、F_{out} 分别表示输入特征图、中间编码特征图、模块输出特征图；f_{in} 表示特征编码模块；θ_i 表示特征编码模块中的可学习参数；W_3 表示每个通道加权融合的权重。

2.4　基于邻域注意力机制的多流光场深度估计网络

图 5 描述了所提网络的具体结构，网络主要分为两个部分：前半部分是用于提取序列图像低层次特征的多支流网络；后半部分则是用于整合提取到的极几何信息并抽象出深度特征的聚合网络。由于光场子孔径图像序列中存在大量空间和角度冗余信息，增加了计算的复杂度，同时使用所提出的注意力机制可以高效地提取出深度估计所需特征，因此所提网络仅选取了部分光场子孔径图像堆栈（方向分别为 0°、45°、90°、135°）作为输入，输出为输入图像中心视角的视差图。当相机参数已知时，可以从视差图直接算出深度。

对于 4 个支流中从不同方向输入的光场子孔径图像堆栈，首先使用序列图像特征提取模块提取光场子孔径图像间的极几何信息。其次，将 4 个支流的特征图直接进行拼接合并，再利用 Mix Attention 模块筛选支流中对深度估计最有效的信息。然后，将筛选后的特征图输入聚合网络中进行视差回归，聚合网络中使用了 6 个类 Inception 的并行模块同时提取稀疏和不稀疏的特征，但没有使用零填充边缘，而是通过直接减小经过每层卷积后的特征图的尺寸以免引入误差。最后，经过一个“二维卷积-激活函数-二维卷积”模块输出深度估计结果。该网络可表示为

$$F_{pred} = f_c\left[F_0, F_{45}, F_{90}, F_{135}, \theta_a, F'_{Attention}(x, y)\right], \tag{5}$$

式中：F_0、F_{45}、F_{90}、F_{135} 分别表示 4 个不同角度的输入序列图像；θ_a 表示可学习参数；f_c 表示网络中的卷积、归一化等操作。网络设计采用了全卷积结构，可以对

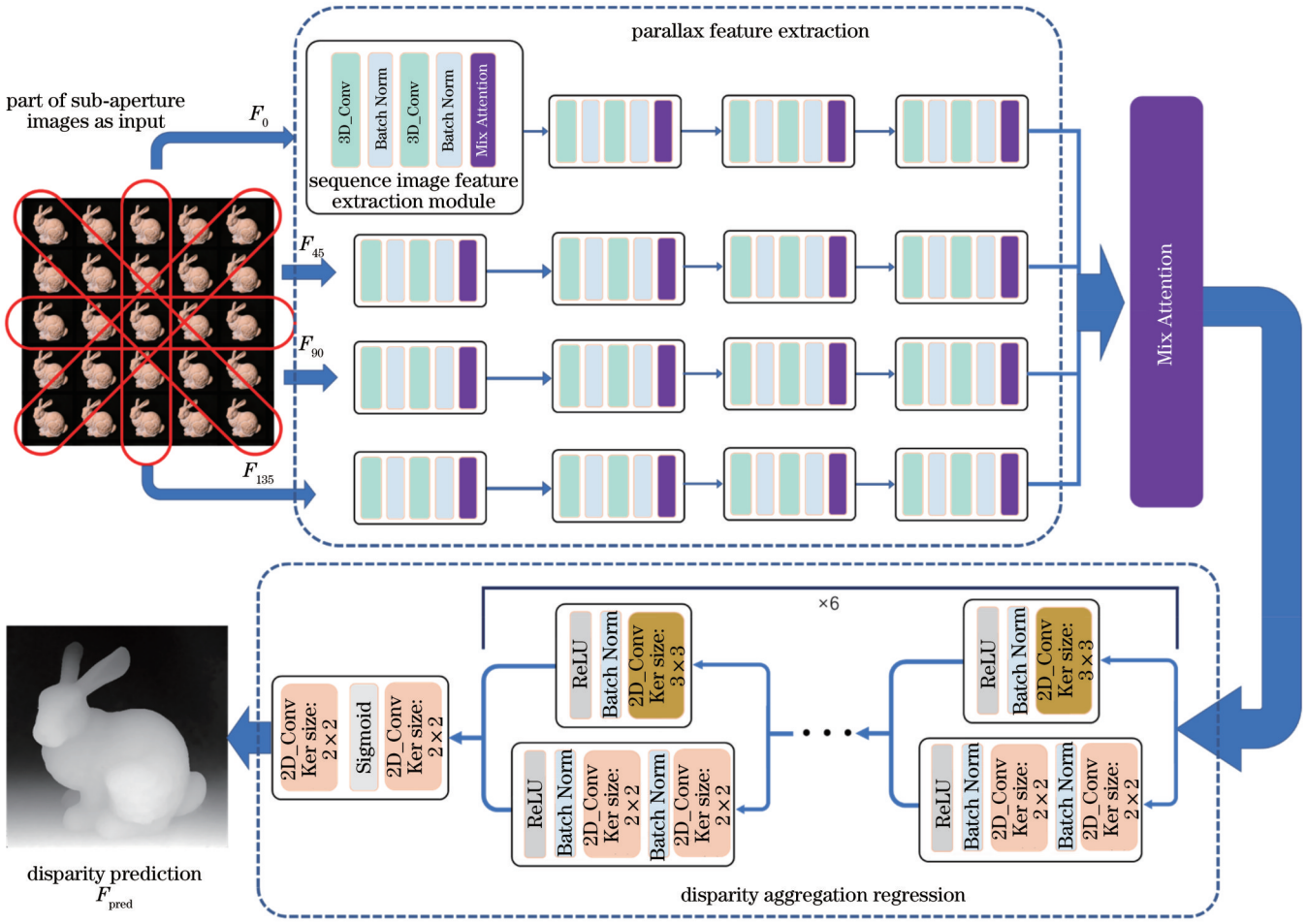

图 5　基于邻域注意力机制的多流光场深度估计网络

Fig. 5　Multi-stream depth estimation network based on Mix Attention

大于卷积深度(即长、宽均大于 23 pixel)的任意尺寸光场子孔径图片实现端到端的预测。为了网络的输出能与视差真值上的像素一一对应以计算 MSE,网络执行的数据降维(Data dimension reduction)的尺寸为偶数,以免产生插值从而引入额外误差。

3　分析与讨论

3.1　实验数据及实验环境

采用 Honauer 等[18]提供的 New HCI 光场数据集对网络进行训练和测试。该数据集中每幅光场图像包含空间分辨率为 512 pixel×512 pixel、角度分辨率为 9 pixel×9 pixel 的子孔径图像。选取其中 16 幅光场图像作为训练集来训练网络,并应用文献[15]的数据增强算法。本实验所用服务器搭载的型号为 Intel(R) Xeon(R) Gold 6240,CPU 为 18 核心 36 线程,内存为 314 G,显卡为 NVIDIA RTX2080Ti,并且实验采用 PyTorch 框架完成模型的搭建与训练。

3.2　网络训练过程

New HCI 光场数据集中将数据分为 4 类:Stratified 集、Training 集、Test 集以及 Additional 集。由于 Additional 集包含的场景类型丰富,因此将其作为训练数据,Training 集和 Stratified 集作为测试数据。训练在 Patch-wise 层面进行,这样能在增加训练图片数量的同时减少所需的显示内存,每个 patch 的尺寸为 50 pixel×50 pixel,采用 Adam 优化算法,学习率从 10^{-4} 衰减到 10^{-6},在 RTX2080TI 显卡上训练了一周收敛。

3.3　实验结果对比

为了评估算法的实验结果,定量分析的评价指标选择了均方误差 E_{MSE} 和坏像素率 R_{BP}。前者是计算预测深度和真实深度之间的 MSE,可以反映深度估计结果与真值的偏离程度;后者将预测出来的视差图和真实视差图逐像素值作差,如果绝对误差小于某个阈值 x 就认为该深度估计预测正确,在实验中,x 设置为 0.07。两种评价指标的计算表达式为

$$\begin{cases} E_{MSE}=\dfrac{1}{m}\sum\left(y_i-\hat{y}_i\right)^2\times 100\% \\ R_{BP}=\dfrac{\left|\left\{y_i\in m:\left|y_i-\hat{y}_i\right|>x\right\}\right|}{|m|} \end{cases}, \tag{6}$$

式中:m 表示图片像素点的总数量;y_i 表示单个预测的值;$\hat{y}_i$ 表示真值;x 表示选用的阈值。

图 6 分别展示了 BP 与 MSE 的迭代过程，其中 BP 阈值设置为 0.07。

为评估所提方法的性能，选取了如下算法进行比较：基于联合深度图正则化和焦点图像堆栈对称性的 OFSY_330[20]，基于约束角熵度量的 CAE[21]，融合 EPI 线索和重对焦线索的 EPI-refocus[22]、EPINET[15]，基于平滑双边滤波和给定三维点的方差度量的 OBER-cross+ANP[23]，基于旋转四边形算子的 SPO-MO[12]，基于注意力机制的视角选择卷积神经网络 LFattNET[16]，基于多视差尺度成本聚合的 Fast-LFnet[24]，基于遮挡感知代价的 OAVC[25]，基于像素调制代价的 OACC-Net[17]。

表 1、表 2 为所提方法与上述方法在 New HCI 测试数据集上 BP 与 MSE 的定量比较。

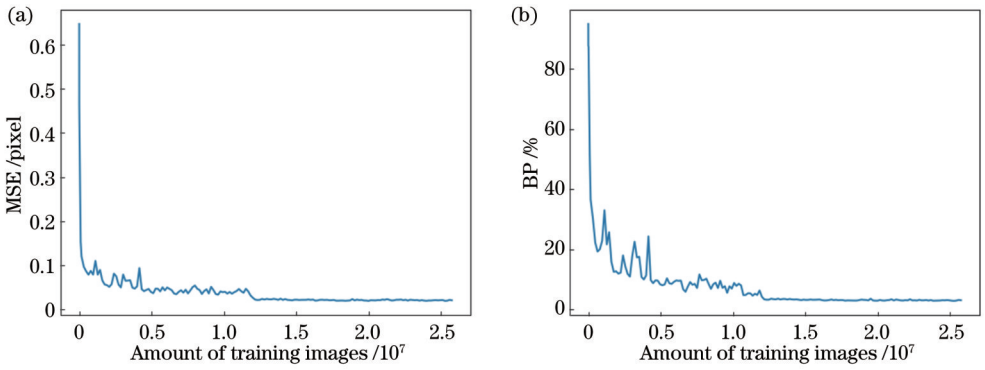

图 6 MSE 与 BP 迭代过程。(a)MSE 迭代过程；(b)BP 迭代过程

Fig. 6 MSE and BP iteration process. (a) MSE iteration process; (b) BP iteration process

表 1 所提方法与其他算法的 BP 对比

Table 1 Comparison of BP between proposed method and other algorithms unit: %

	Backgammon	Dots	Pyramids	Strips	Boxes	Cotton	Sideboard	Avg
OFSY_330	4.828	37.670	0.356	18.640	19.246	3.036	10.355	13.447
CAE	3.924	12.401	1.681	7.872	17.885	3.369	9.845	8.140
EPI-refocus	4.305	3.904	0.424	3.922	12.170	0.559	5.955	4.463
EPINET	3.501	2.490	0.159	2.457	12.341	0.447	4.462	3.694
OBER-cross+ANP	3.413	0.974	0.364	3.065	10.759	1.018	5.671	3.609
LFattNET	3.126	1.432	0.195	2.933	11.044	**0.272**	**2.870**	3.125
SPO-MO	3.450	2.781	**0.050**	4.118	15.494	2.161	7.515	5.081
Fast-LFnet	5.138	21.169	0.620	9.442	18.699	0.714	7.032	7.467
OAVC	3.120	69.100	0.830	2.900	16.100	2.550	12.400	15.286
OACC-Net	3.931	1.510	0.157	2.920	**10.700**	0.312	3.350	3.269
Ours	**2.179**	**0.772**	0.365	**1.779**	12.087	0.497	3.956	**3.091**

表 2 所提方法与其他算法的 MSE 对比

Table 2 Comparison of MSE between proposed method and other algorithms unit: pixel

	Backgammon	Dots	Pyramids	Strips	Boxes	Cotton	Sideboard	Avg
OFSY_330	7.549	14.756	0.008	7.269	9.561	2.653	2.478	6.325
CAE	6.074	5.082	0.048	3.556	8.424	1.506	0.876	3.652
EPI-refocus	5.553	3.063	0.041	1.870	7.552	0.573	1.609	2.894
EPINET	3.705	1.475	0.007	0.932	5.968	0.197	0.798	1.869
OBER-cross+ANP	4.799	1.757	0.008	1.435	4.750	0.555	0.941	2.035
LFattNET	3.648	1.425	**0.004**	0.892	3.996	0.209	**0.531**	1.529
SPO-MO	4.133	3.763	0.009	1.934	10.374	1.329	0.932	3.211
Fast-LFnet	**1.488**	3.070	0.018	**0.231**	4.260	0.339	0.742	1.450
OAVC	3.840	16.600	0.040	1.320	6.990	0.600	1.050	4.349
OACC-Net	3.938	1.418	**0.004**	0.845	**2.892**	**0.162**	0.542	1.400
Ours	2.056	**0.795**	0.010	0.339	3.771	0.260	0.649	**1.126**

显然，所提方法的性能更加稳定，平均坏像素率和误差分别为 3.091% 和 1.126 pixel，在大部分场景中取得了最优（加粗）或次优（下划线）的深度估计结果。

表 3 列出了光场深度估计算法计算深度图的平均用时，可以看到基于优化的算法（SPO-MO）在运行速度上大幅度慢于基于学习的算法，并且由于所提算法没有采用代价体等复杂设计，运行速度快于其他基于学习的对比算法，所需的内存也更少。

表 3 运行时间对比

Table 3 Comparison of operation time unit: s

Algorithms	Avg runtime	Algorithms	Avg runtime
SPO-MO	2115.417	LFattNET	5.862
EPI-refocus	72.742	OAVC	4.220
EPINET	2.041	Fast-LFnet	<u>0.624</u>
Ours	**0.387**		

3.4 消融实验

为验证所提出的邻域像素注意力机制 Mix Attention 的有效性，在 New HCI[18] 光场数据集上进行了消融实验。表 4 为消融实验结果，其中 Baseline 表示只使用 2D 卷积的多流网络结构，Baselines+3D_Conv 表示用 3D 卷积替换 2D 卷积的网络结构，Baseline+Mix_Att 表示只在 Baseline 多分支的最后加入 Mix Attention 模块。消融实验结果表明，所提出的邻域像素注意力机制 Mix Attention 有效地提高了深度估计的性能。

表 4 消融实验结果

Table 4 Results of ablation study

Model	MSE /pixel		BP /%	
	Training	Testing	Training	Testing
Baseline	1.932	2.279	12.087	14.610
Baselines+3D_Conv	1.475	2.012	9.424	11.919
Baseline+Mix_Att	1.352	1.656	5.406	9.825
Full	**0.728**	**1.184**	**2.838**	**4.820**

3.5 算法结果主观对比

图 7 和图 8 展示了所提方法与其他算法的预测结

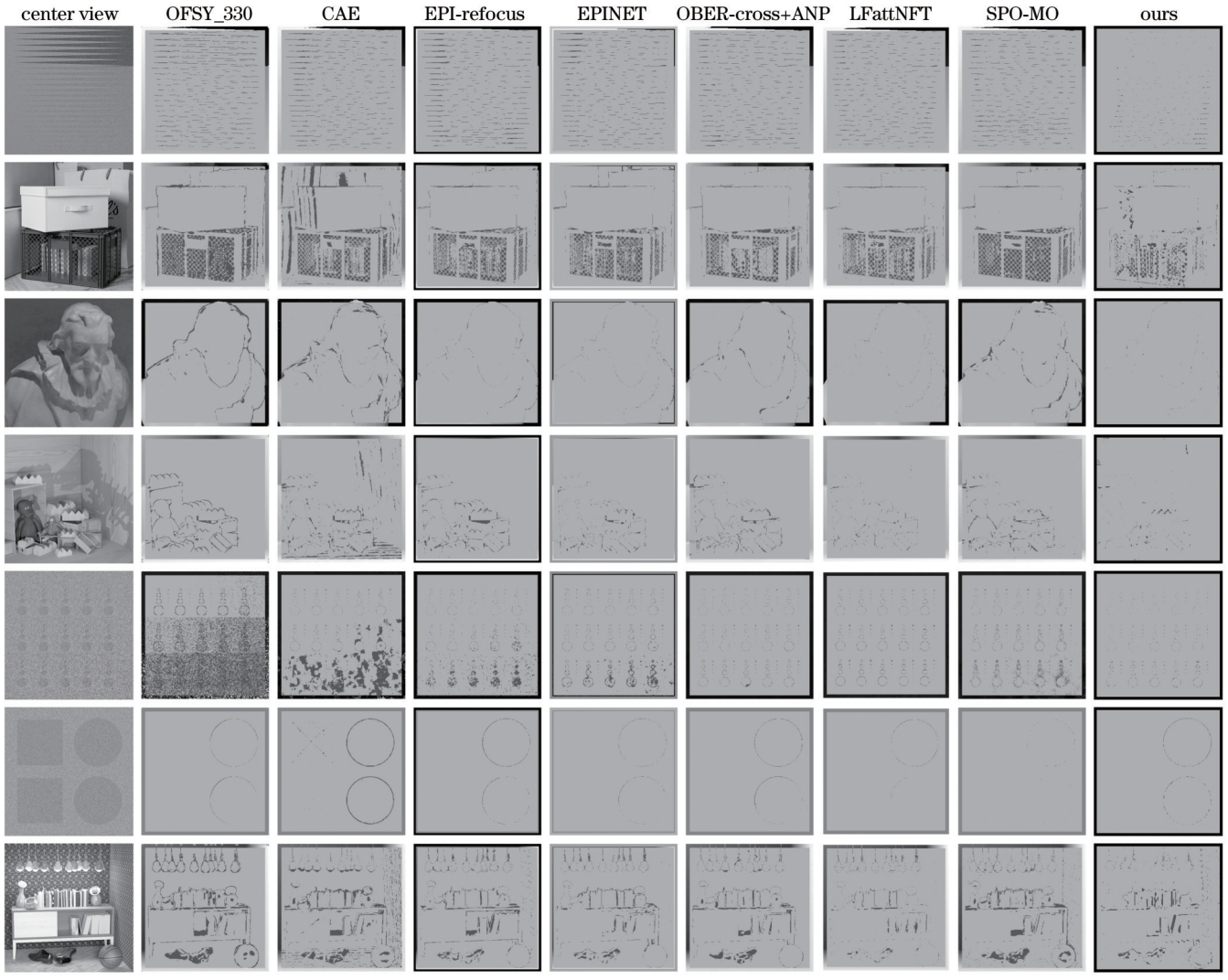

图 7 BP 可视化表示

Fig. 7 Visual representation of BP

果 BP 和 MES 的定性比较。在图 7 中，深色部分表示坏像素点，浅色部分表示视差与真值差距在 0.07 以内的点，各个场景的中心视角展示在图 7 的第一列中。

图 8 是逐个像素点计算深度图与真值图的 MES，差值越大颜色越深，若预测值小于真实值表示为红色（正值），反之为蓝色（负值）。从图 8 中可以看出，所提方法总体表现优于其他对比算法，在深度不连续区域（如 Sideboard 场景中悬挂的吊灯）有较强的鲁棒性，在纹理复杂的区域和深度连续区域（如 Cotton 场景及 Pyramids 场景）具有较高的精度，在反射高光区域（如 Sideboard 场景中地上的鞋）较其他对比算法减少了预测误差，在深度边缘区域（如 Backgammon 场景中的边缘）具有较高的平滑度，取得了更为理想的视差估计结果。

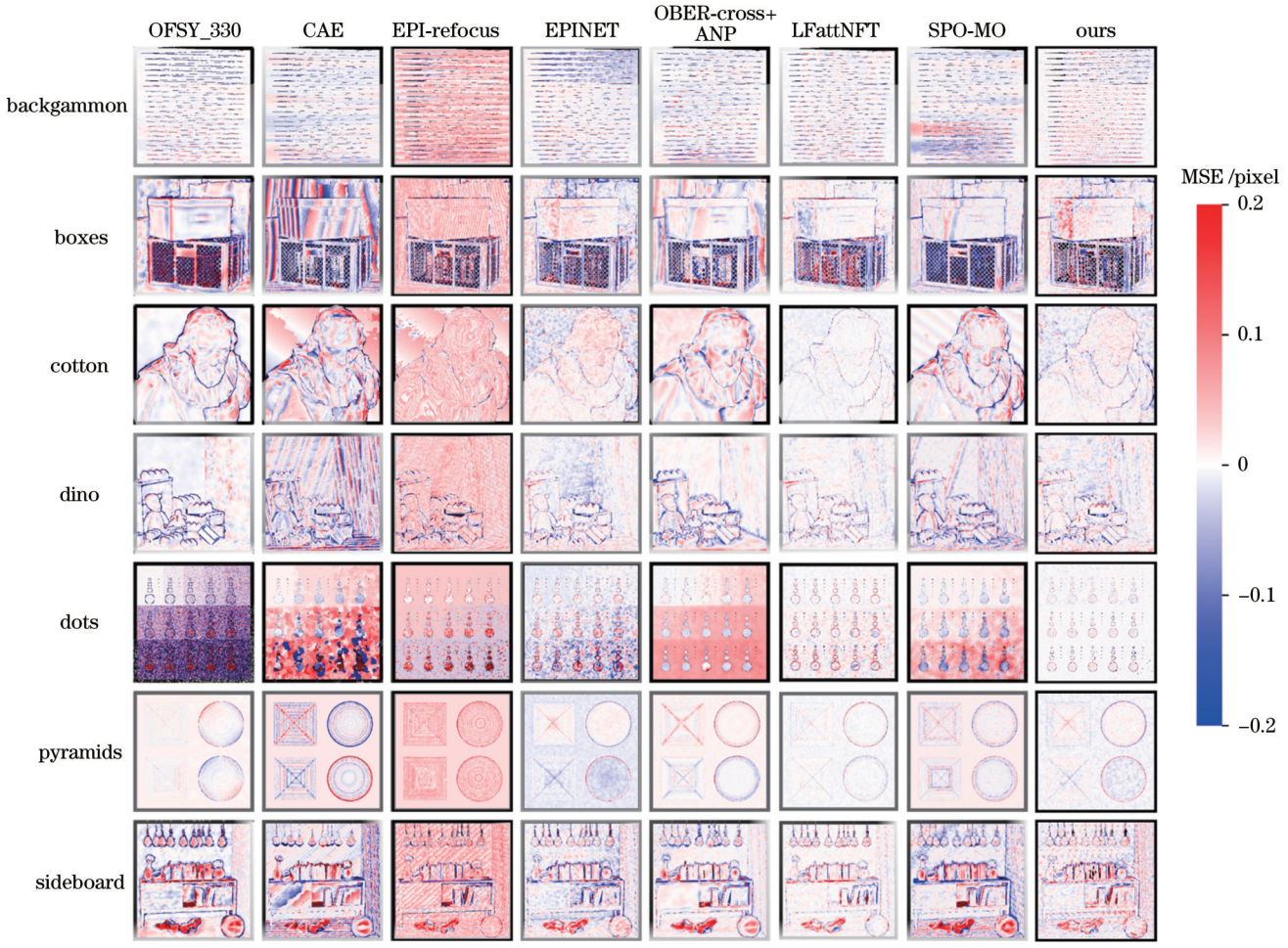

图 8　MSE 可视化表示

Fig. 8　Visual representation of MSE

图 9 为验证集中几何关系较为复杂的 Boxes 场景放大视图，网状的箱子使得后面的线条存在被遮挡的情况。针对这一场景，所提方法与目前光场深度估计算法中综合性能最佳的 LFAttNet、EPINET、SPO-MO、EPI-refocus 进行了对比。从图 7（第二行）和图 9 可以看出，现有算法在此场景的错误率较高，但是所提模型在此场景取得了最好的效果，证明了所提方法能够有效解决遮挡问题。

在真实数据集上用所提方法开展实验。与合成数据集相比，真实数据集会存在深度不连续、场景模糊和各类噪声等问题，从图 10 可以看出，所提出的方法具有很好的泛化性能，在真实数据集中得到了边缘更锐利、画面更干净的深度估计结果。

4　结　　论

针对光场深度估计任务的特性以及光场数据的特征，提出了一种邻域像素注意力机制 Mix Attention，该机制可以捕捉到光场中某一像素点周围有限邻域的像素点与深度特征之间的相关性，通过对邻域的特征图进行计算，对网络中不同的特征图进行选择，提高了光场数据的利用效率。同时通过分析不同光场子孔径图像间的像素位移情况，引入三维卷积核提取序列图像特征，提出了一种快速的端到端多流光场深度估计网络。在光场数据集 New HCI[18] 上的测试显示所提光场深度估计网络在 BP、MSE 和计算时间 3 个性能指标上优于现有光场深度估计方法，有效提高了网络预测

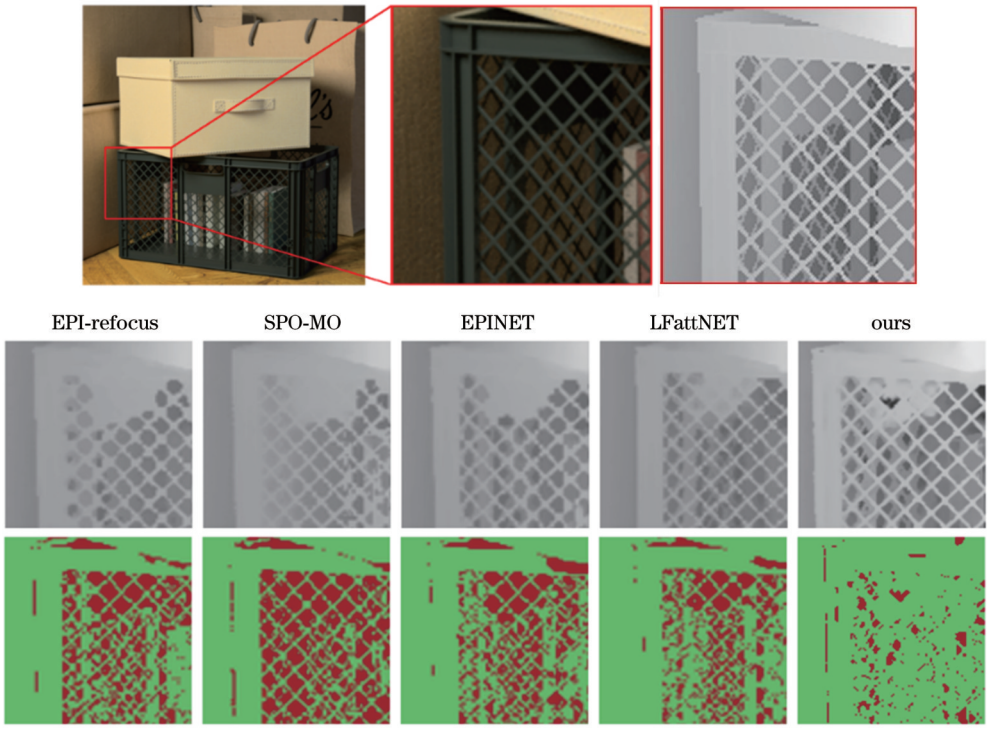

图 9　场景 Boxes 中本算法

Fig. 9　Our algorithm in the scene Boxes

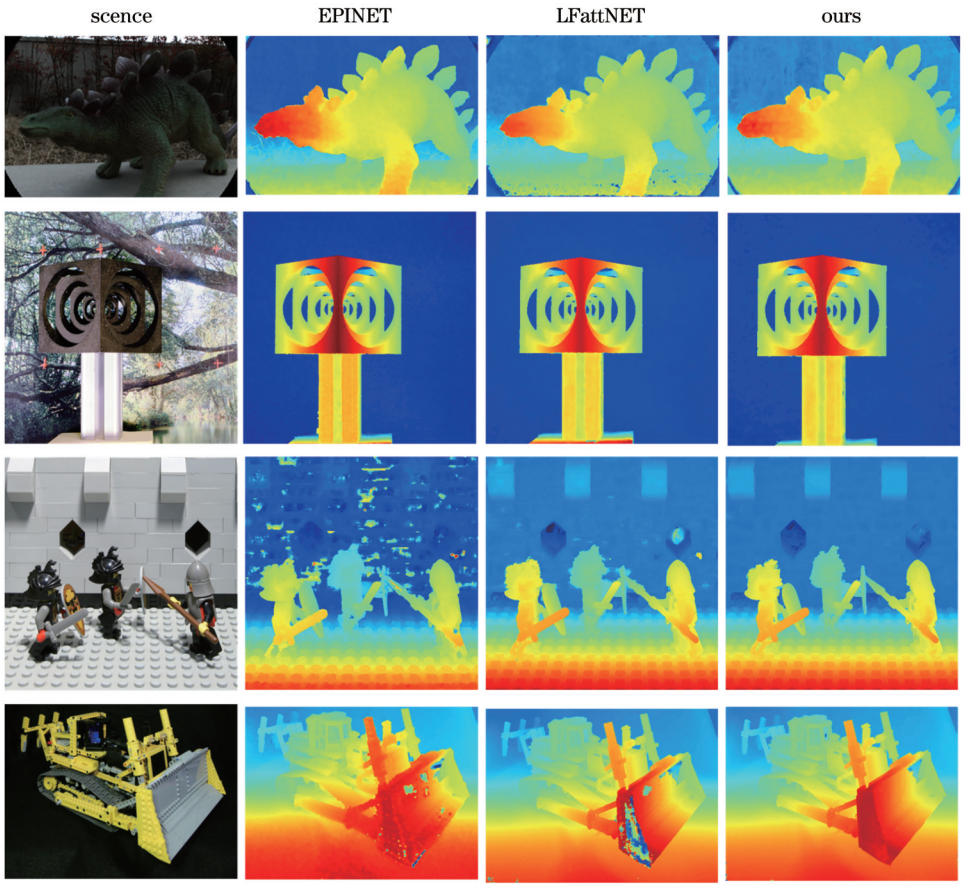

图 10　真实场景数据集实验结果

Fig. 10　Results in the real scene datasets

深度的性能，且对遮挡严重的Boxes等场景的测试也表明所提算法有较强的鲁棒性。消融实验表明，所提出的Mix Attention注意力机制充分挖掘了不同通道邻域像素间的相关性，有效提高了光场深度估计网络预测深度的性能。但是对于缺少纹理信息的区域，所提方法表现欠佳。下一步，将重点研究利用空间金字塔等结构增加网络对多尺度特征的提取能力，通过平滑对无纹理区域的深度结果进行优化，进一步提高深度估计可靠性。

参 考 文 献

[1] Adelson E H. The plenoptic function and the elements of early vision[J]. Computational Models of Visual Processing, 1991: 3-20.

[2] Levoy M, Hanrahan P. Light field rendering[C]//Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, August 4-9, 1996, Stanford University, CA, USA. New York: ACM Press, 1996: 31-42.

[3] Ren N, Levoy M, Bredif M, et al. Light field photography with a hand-held plenoptic camera[J]. Stanford University Computer Science Tech Report, 2005, 2(1): 1-11.

[4] Nousias S, Lourakis M, Keane P, et al. A linear approach to absolute pose estimation for light fields[C]//2020 International Conference on 3D Vision (3DV), November 25-28, 2020, Fukuoka, Japan. New York: IEEE Press, 2021: 672-681.

[5] Fu K R, Jiang Y, Ji G P, et al. Light field salient object detection: a review and benchmark[J]. Computational Visual Media, 2022, 8(4): 509-534.

[6] Suhail M, Esteves C, Sigal L, et al. Light field neural rendering [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8259-8269.

[7] 吴治安, 朱效宇, 李健, 等. 基于体标定追迹法的光场PIV权系数计算方法[J]. 光学学报, 2021, 41(20): 2010001.
Wu Z A, Zhu X Y, Li J, et al. Volumetric-calibration ray tracing-based calculation method of weight coefficient in light field PIV[J]. Acta Optica Sinica, 2021, 41(20): 2010001.

[8] 方璐, 戴琼海. 计算光场成像[J]. 光学学报, 2020, 40(1): 0111001.
Fang L, Dai Q H. Computational light field imaging[J]. Acta Optica Sinica, 2020, 40(1): 0111001.

[9] 殷永凯, 于锴, 于春展, 等. 几何光场三维成像综述[J]. 中国激光, 2021, 48(12): 1209001.
Yin Y K, Yu K, Yu C Z, et al. 3D imaging using geometric light field: a review[J]. Chinese Journal of Lasers, 2021, 48(12): 1209001.

[10] Yu Z, Guo X, Lin H, et al. Line assisted light field triangulation and stereo matching[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2013: 2792-2799.

[11] Tao M W, Hadap S, Malik J, et al. Depth from combining defocus and correspondence using light-field cameras[C]//2013 IEEE International Conference on Computer Vision, December 1-8, 2013, Sydney, NSW, Australia. New York: IEEE Press, 2014: 673-680.

[12] 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.

[13] Johannsen O, Sulc A, Goldluecke B. What sparse light field coding reveals about scene structure[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3262-3270.

[14] Heber S, Pock T. Convolutional networks for shape from light field[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 3746-3754.

[15] 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]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4748-4757.

[16] Tsai Y J, Liu Y L, Ouhyoung M, et al. Attention-based view selection networks for light-field disparity estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12095-12103.

[17] Wang Y Q, Wang L G, Liang Z Y, et al. Occlusion-aware cost constructor for light field depth estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 19777-19786.

[18] Honauer K, Johannsen O, Kondermann D, et al. A dataset and evaluation methodology for depth estimation on 4D light fields [M]//Lai S H, Lepetit V, Nishino K, et al. Computer vision - ACCV 2016. Lecture notes in computer science. Cham: Springer, 2017, 10113: 19-34.

[19] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 11531-11539.

[20] Strecke M, Alperovich A, Goldluecke B. Accurate depth and normal maps from occlusion-aware focal stack symmetry[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2529-2537.

[21] Williem, 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, 2018, 40(10): 2484-2497.

[22] Zhou W H, Liang L K, Zhang H, et al. Scale and orientation aware EPI-patch learning for light field depth estimation[C]//2018 24th International Conference on Pattern Recognition (ICPR), August 20-24, 2018, Beijing, China. New York: IEEE Press, 2018: 2362-2367.

[23] Schilling H, Diebold M, Rother C, et al. Trust your model: light field depth estimation with inline occlusion handling[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4530-4538.

[24] Huang Z C, Hu X M, Xue Z, et al. Fast light-field disparity estimation with multi-disparity-scale cost aggregation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 6300-6309.

[25] 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, 2022, 44(11): 8022-8035.

Depth Estimation Method of Light Field Based on Attention Mechanism of Neighborhood Pixel

Lin Xi, Guo Yang, Zhao Yongqiang*, Yao Naifu

School of Automation, Northwestern Polytechnical University, Xi'an 710129, *Shaanxi, China*

Abstract

Objective Accurate acquisition of depth information has always been a research hotspot in computer vision. Traditional cameras can only capture light intensity information within a certain time period, losing other information such as the incident light angle helpful for depth estimation. The emergence of light field cameras provides a new solution for depth estimation. Compared to traditional cameras, light field cameras can capture four-dimensional light field information. Micro-lens array light field cameras also solve the problems of large camera array size and impracticality to carry. Therefore, employing light field cameras to estimate the depth of a scene has broad research prospects. However, in the existing research, there are problems such as inaccurate depth estimation, high computational complexity, and occlusions in multi-view scenarios. Occlusions have always been challenging in tasks of light field depthestimation. For scenes without occlusions, most existing methods can yield good depth estimation results, but this requires the pixels to satisfy the color consistency principle. When occluded pixels exist in the scene, this principle among different views is no longer satisfied. In such cases, the accuracy of the depth map obtained using existing methods will significantly decrease, with more errors in the occluded areas and edges. Thus, we propose a method to estimate light field depth based on the attention mechanism of neighborhood pixel. By exploiting the high correlation between depth information and neighboring pixels in sub-aperture images, the network performance in estimating the depth of light field images is improved.

Methods First, after analyzing the characteristics of the sub-aperture image sequence, we utilize the correlation between the depth information of a pixel in the light field image and a limited neighborhood of surrounding pixels to propose a neighborhood pixel attention mechanism Mix Attention. This mechanism efficiently models the relationship between feature maps and depth by combining spatial and channel attention, thereby improving the estimation accuracy of light field depth and providing the network with a certain degree of occlusion robustness. Next, based on Mix Attention, a sequential image feature extraction module is proposed. It employs three-dimensional convolutions to encode the spatial and angular information contained in the sub-aperture image sequence into feature maps and adopts Mix Attention to adjust the weights. This module enhances the representation power of the network by incorporating both spatial and angular information. Finally, a multi-branch depth estimation network is proposed to take part of sub-aperture images of the light field as input and achieve fast end-to-end depth estimation for light field images of arbitrary input sizes. This network leverages the proposed attention mechanism and the sequential image feature extraction module to effectively estimate depth from the light field image. Overall, we propose a novel estimation approach for light field depth. By leveraging the correlation between neighboring pixels and incorporating attention mechanisms, this approach improves the depth estimation accuracy and enhances the network's ability to handle occlusions. The proposed network architecture enables efficient and robust depth estimation for light field images.

Results and Discussions In quantitative analysis, mean square error (MSE) and bad pixel rate are chosen as evaluation metrics. The proposed method demonstrates stable performance, with an average bad pixel rate and MSE of 3.091% and 1.126, respectively (Tables 1 and 2). In most scenarios, the method achieves optimal (bold) or suboptimal (underlined) depth estimation results. The effectiveness of the proposed attention mechanism (Mix Attention) is further demonstrated by ablation experiments (Table 3). Qualitative analysis (Figs. 7 and 8) reveals that the proposed method exhibits strong robustness in depth-discontinuous regions (hanging lamp in the Sideboard scene), high accuracy in texture-rich areas and depth-continuous regions (Cotton and Pyramids scenes), reduced prediction errors in areas with reflections (shoes on the floor in the Sideboard scene), and high smoothness at depth edges (edges in the Backgammon scene). Generally, the proposed method yields more desirable disparity estimation results. Experimental results indicate that the overall performance of the proposed network surpasses that of other algorithms. Therefore, the proposed method exhibits stable and superior performance in depth estimation, as indicated by the selected evaluation metrics, quantitative results, and qualitative analysis.

Conclusions Aiming at the estimation task characteristics of light field depth and the features of light field data, we propose an attention mechanism of neighborhood pixel called Mix Attention. This mechanism captures the correlation

between a pixel and its limited neighborhood pixel in the light field and depth features. By calculating the feature maps of the neighborhood, different feature maps in the network are selectively attended to improve the utilization efficiency of light field images. Additionally, by analyzing the pixel displacement between different sub-aperture images in the light field, a fast end-to-end multi-stream estimation network of light field depth is introduced to employ three-dimensional convolutional kernels to extract sequential image features. Tests on the New HCI light field dataset demonstrate that the proposed estimation network outperforms existing methods in three performance metrics, including 0.07 bad pixel rate, MSE, and computational time. It effectively enhances the depth prediction performance and exhibits robustness in occluded scenes such as Boxes. Ablation experiments show that the proposed mechanism fully exploits the correlation between neighboring pixels in different channels, improving the depth prediction performance in the estimation network of light field depth. However, the performance of the proposed method is unsatisfactory in regions lacking texture information. In the future, we will focus on techniques such as spatial pyramids to enhance the network′s ability to extract multi-scale features, smooth the depth results in textureless regions, and further improve the depth estimation reliability.

Key words light field image; depth estimation; neighborhood pixel; attention mechanism; neural network