计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900005

基于编码-解码器架构的光场深度估计方法

晏 旭[1,2,3] 马 帅[1,2,3] 曾凤娇[1,2,3] 郭正华[1,2,3] 伍俊龙[1,2,3] 杨 平[1,2] 许 冰[1,2]
1 中国科学院光电技术研究所自适应光学重点实验室 成都 610209
2 中国科学院光电技术研究所 成都 610209
3 中国科学院大学 北京 100049
(18123087889@163.com)

摘 要 针对现有光场深度估计方法存在的计算时间长和精度低的问题,提出了一种融合光场结构特征的基于编码-解码器架构的光场深度估计方法。该方法基于卷积神经网络,采用端到端的方式进行计算,一次输入光场图像就可获得场景视差信息,计算量远低于传统方法,大大缩短了计算时间。为提高计算精确度,网络模型以光场图像的多方向极平面图堆叠体(Epipolar Plane Image Volume,EPI-volume)为输入,先利用多路编码模块对输入的光场图像进行特征提取,再使用带跳跃连接的编码-解码器架构进行特征聚合,使网络在逐像素视差估计时能够融合目标像素点邻域的上下文信息。此外,模型采取不同深度的卷积块从中心视角图中提取场景的结构特征,并将该结构特征引入对应的跳跃连接中,为视差图预测提供了额外的边缘特征参考,进一步提高了计算精确度。对 HCI-4D 光场基准测试集的实验结果表明,所提方法的坏像素率(BadPix)指标比对比方法降低了 31.2%,均方误差(MSE)指标比对比方法降低了 54.6%。对于基准测试集中的光场图像,深度估计的平均计算时间为1.2 s,计算速度远超对比方法。

关键词:光场;深度估计;极平面图;编码-解码器结构;上下文信息
中图法分类号 TP391

Light Field Depth Estimation Method Based on Encoder-decoder Architecture

YAN Xu[1,2,3],MA Shuai[1,2,3],ZENG Feng-jiao[1,2,3],GUO Zheng-hua[1,2,3],WU Jun-long[1,2,3],YANG Ping[1,2] and XU Bing[1,2]
1 Key Laboratory on Adaptive Optics,Institute of Optics and Electronics,Chinese Academy of Sciences,Chengdu 610209,China
2 Institute of Optics and Electronics,Chinese Academy of Sciences,Chengdu 610209,China
3 University of Chinese Academy of Sciences,Beijing 100049,China

Abstract Aiming at the solution to the time-consuming and low-precision disadvantage of present methodologies,the light field depth estimation method combining context information of the scene is proposed. This method is based on an end-to-end convolutional neural network,with the advantage of obtaining depth map from a single light field image. On merit of the reduced computational cost from this method,the time consumption is consequently decreased. For improvement in calculation accuracy,multi orientation epipolar plane image volumes of the light field images are input to network,from which feature can be extracted by the multi-stream encoding module,and then aggregated by the encoding-decoding architecture with skip connection,resulting in fuse the context information of the neighborhood of the target pixel in the process of per-pixel disparity estimation. Furthermore,the model uses convolutional blocks of different depths to extract the structural features of the scene from the central viewpoint image,by introducing these structural features into the corresponding skip connection,additional references for edge features are obtained and the calculation accuracy is further improved. Experiments in the HCI 4D Light Field Benchmark show that the BadPix index and MSE index of the proposed method are respectively 31.2% and 54.6% lower than those of the comparison method,and the average calculation time of depth estimation is 1.2 seconds,which is much faster than comparison method.

Keywords Light field,Depth estimation,Epipolar plane image,Encoder-decoder,Context information

到稿日期:2020-09-01 返修日期:2021-02-02
基金项目:国家自然科学基金(J19K004)
This work was supported by the National Natural Science Foundation of China(J19K004).
通信作者:许冰((bing_xu_ioe@163.com)

1　引言

通过对图像进行分析而获得拍摄场景中物体深度信息的被动式三维测量技术，是近年来计算机视觉领域的研究热点之一。光场相机作为被动式三维测量技术的一种，通过单次曝光实现对空间中光线方向及强度信息的捕捉。利用这些捕捉到的信息实现对拍摄场景的深度信息的提取，是光场高层视觉应用中至关重要的环节，受到了国内外学者的广泛关注。

自 1939 年 Gershun[1] 首次提出光场概念以来，传统光场深度估计算法经过了 80 多年的发展，目前主要分为两大类：基于极平面图的算法和基于匹配的算法。

极平面图(Epipolar Plane Image，EPI)是光场图像的二维切片，能直观反映空间中光线的空间及角度信息。Wanner 等[2]提出利用 EPI 水平和垂直方向的结构张量来计算局部深度信息，并利用全局可见性约束优化，获得了在所有视角图中一致的深度图。Tosic 等[3]利用高斯核将 EPI 像纹理映射到光场深度-尺度空间，实现了光线探测及密集光场深度估计。Zhang 等[4]利用平行四边形算子在 EPI 中进行区域划分，通过最大化区域的分布距离来计算极图中蕴含深度信息的直线斜率，用于解决深度估计中的遮挡问题。基于 EPI 的算法在镜面反射区域、遮挡场景中能够取得较好的效果，但是，算法的计算量大，计算时间长，上述算法中速度最快的[4]计算时间超过 10 min。并且基于 EPI 的算法的精度受噪声的影响较大，在复杂场景中与真值图的误差普遍较大。

根据光场多视角成像特点，利用光场各视角图像间颜色的一致性，可将光场图像深度估计问题转化为匹配问题，通过构造成本量来获取光场图像的深度信息。Jeon 等[5]针对光场视角图之间基线窄的问题，使用相位移理论来获取光场子孔径图像间亚像素的位移，然后将光场图像中心视角图像与其他视角图进行立体匹配，并利用图割优化算法进行多标签匹配，最后通过局部优化迭代获得连续深度图。Chen 等[6]提出使用表面相机模型来建立光场子孔径图像立体匹配，并引入双边滤波的度量思想来解决光场深度估计中的遮挡问题。基于多视角立体匹配的算法能够很好地应对光场混叠和视角串扰的问题，并且在遮挡场景中也能计算得到较精确的深度图，但是基于匹配的深度估计方法得到的深度图与真值图仍存在较大的误差。而且此类算法需要对所有视角图像分别构建代价函数，计算量相对更大，时间复杂度更高，算法的计算时间通常超过 10 min。

综上，传统算法能通过特定方法计算得到对应场景的深度图，但是深度图与真值图存在较大误差。并且，传统方法在大部分复杂场景中存在性能严重退化的问题。同时，传统光场深度估计算法的计算量大，计算需要耗费大量的时间。

近年来，深度学习技术在光场领域得到了广泛应用，如视角合成[7]、光场图像超分辨率[8]、目标检测[9]、由单个图像合成光场图像[10]、光场图像压缩[11]等。

而在解决光场深度估计问题上，Heber 等[12]利用卷积神经网络来实现光场图像及其视差图的端到端的映射，并利用高阶正则化进一步优化预测结果。在此基础上，Heber 等[13]运用 U 型编码-解码器结构实现了对四维光场深度图的快速计算。Heber 等的方法大大减小了光场深度估计的计算时间，在数据集上的平均计算时间为 0.8 s。但是，其网络的设计只考虑了单一方向的极图，从而导致深度估计计算精度较低。Zhou 等[14]根据多尺度和多方向的 EPI 像素块提供的不同特征，提出了一种基于多方向和尺度感知的 EPI 像素块学习模型，实现了对光场图像准确的深度估计。但是，由于该方法需要运用两种类型的网络对 EPI 像素块的尺度和方向特征进行融合，会耗费大量的时间，因此计算速度相对较慢。

为了解决上述方法表现出的计算时间长、精度低的问题，本文提出了一种融合光场结构特征的基于编码-解码器架构的光场深度估计方法。为了有效利用目标像素点邻域信息，即上下文信息，辅助目标像素点的视差值预测，本文提出的网络模型在特征聚合模块采用一种带跳跃连接的编码-解码器结构，将光场图像的浅层较小感知域的局部特征与深层较大感知域的抽象特征通过跳跃连接的方式在通道维度串联在一起。编码器中聚合上下文信息而得到的抽象特征对于单像素点的视差值预测很有帮助，但是下采样操作却丢失了图形中物体的轮廓及像素点的位置信息。通过解码器部分的跳跃连接操作，特征在传递过程中就能够既具有强语义信息的抽象特征，又具有包含丰富位置信息的高分辨率局部特征。这样就保证了网络不仅能够对视差图进行精准预测，而且具有保边性能。同时，我们还将中心视角经过不同深度的卷积块后得到的结构特征引入对应各级跳跃连接中。实验结果表明，该结构特征能够极大程度地降低预测视差图的全局误差。对于 HCI-4D 光场基准测试集中多个场景的相同尺寸的光场图像，本文方法的坏像素率(BadPix)指标在多个场景中均优于对比方法，除一个场景外，均方误差(MSE)指标在其余场景都优于对比方法，在运算速度上更是远超对比方法。

2　本文方法

为了实现准确而又快速的光场深度估计，我们设计了一个基于卷积神经网络的端到端的光场深度估计网络。本文方法受到图像语义分割思想的启发。语义分割旨在将图像的每个像素分类为一个实例，其中每个实例对应一个类[15]，属于稠密预测的一种，这与光场图像深度估计需要进行逐像素点视差值预测不谋而合。因此，本文参考语义分割领域的多种上下文信息聚合架构，选择效果优异的框架——带跳跃连接的编码-解码器架构作为特征聚合网络的主体架构。图 1 给出了本文方法的网络结构及超参数。网络输入为光场图像 4 个方向的极平面图堆叠体(Epipolar Plane Image Vo-lume，EPI-volume)，输出为光场图像中心视角图的视差图。网络共分为 3 个模块：多路编码模块、特征聚合模块和视差回归模块。首先多路编码模块实现对多方向 EPI-volume 的特征提取，得到光场图像多方向的浅层特征，接着将这些特征串联后传入特征聚合模块。如图 2 所示，特征聚合模块采用带有跳跃连接的编码-解码器结构，保证特征在传递过程中能够将浅层较小感知域的局部特征与深层较大感知域的抽象特征融合，使网络在对单像素点进行视差估计时能够有效利用抽象特征提供的上下文信息辅助像素点进行视差预测的同时，补

充下采样过程中丢失的物体轮廓及像素点的位置信息。此外，将中心视图经不同深度的卷积块后得到的不同分辨率及精细度的结构特征映射连接到各级跳跃连接部分，提高了预测视差图的整体精确度。最后，视差回归模块实现对光场图像中心视图逐像素点的视差值预测。下面将对上述几个模块进行详细介绍。

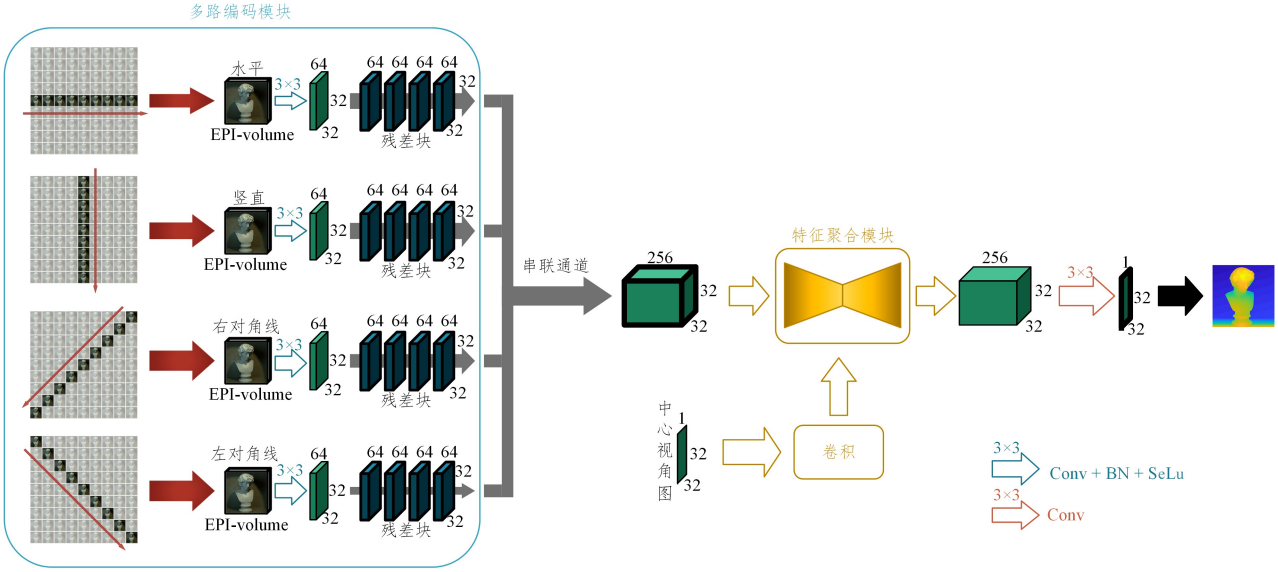


图 1 本文方法的网络结构

Fig. 1 Network architecture of the method

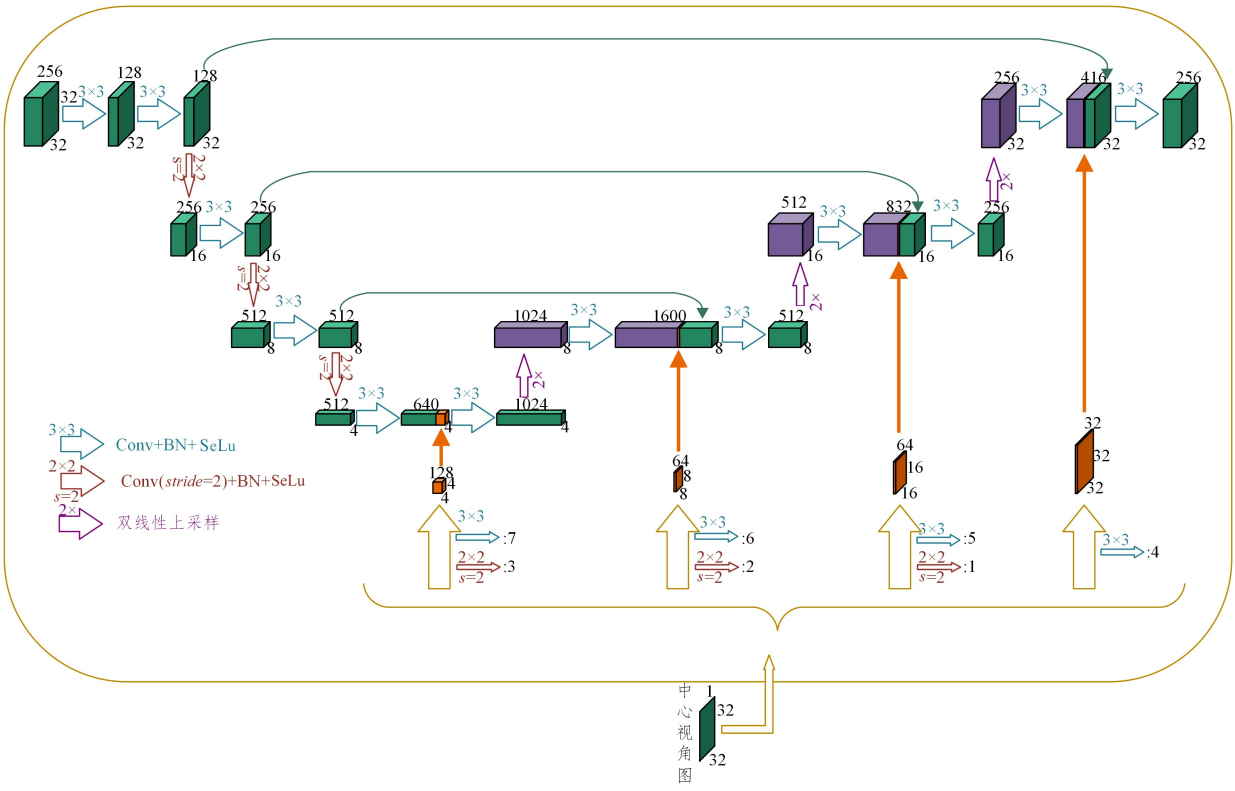


图 2 特征聚合模块的网络结构

Fig. 2 Network architecture of the feature aggregation network

2.1 多路编码模块

光场图像具有表示相机阵列平面上采样密集程度的角度分辨率，以及表示子孔径图像分辨率的空间分辨率，其数据量和深度估计所需的计算量远远大于立体匹配问题[16-18]。为了解决这个问题，文献[4,19-21]仅选用水平或垂直单一方向的 EPI-volume。但是，单一方向的 EPI-volume 无法有效地利用光场图像多视角的冗余信息，并且对遮挡和噪声的鲁棒性较差。

针对这个问题，本文选用光场图像水平、竖直、左对角线、右对角线 4 个方向的 EPI-volume 作为网络的输入。多方向 EPI-volume 能够在尽量减少计算量的情况下，充分利用光场多视角的有效信息，并且有实验表明，当 EPI 方向与遮挡边

界方向相同时，该方向 EPI 计算得到的深度图不受遮挡的影响[22]。

如图 3 所示，遮挡物的边缘与右对角线 EPI-volume 方向平行，此时在右对角线 EPI-volume 中只包含被遮挡点的信息而不包含遮挡物的信息。

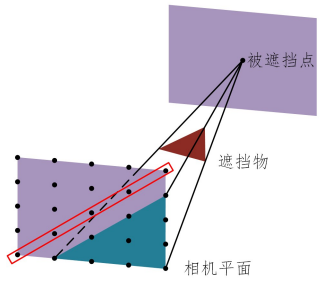


图 3　光场遮挡场景

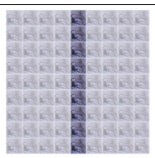
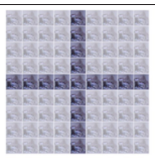
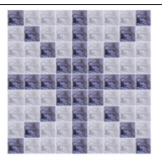
Fig. 3　Occlusion scene of light field

因此，本文采用的多路编码网络选择四路并联且完全相同的网络，每路网络分别输入一个方向的 EPI-volume。网络包含 1 个“Conv-BN-SeLu”卷积块和 4 个基本残差块，且在最后一个残差块加入空洞卷积，在不增加网络运算成本的情况下，使网络有更大的感受野。多路编码模块实现了对 4 个方向 EPI-volume 的像素块匹配特征的提取。

为了更清晰地展示多路编码网络相比单/双路编码的优势，我们设计了除多路编码模块外其余部分完全相同的 3 个网络架构，它们在多路编码部分分别使用了 1，2，4 个方向的 EPI-volume。表 1 比较了 3 个网络架构在测试集上的性能差异，可以看出，选择 4 个方向 EPI-volume 的网络架构的性能最佳。

表 1　视角数量对性能的影响

Table 1　Effect of the number of viewpoints on performance

	单路	双路	四路
输入视角			
BadPix(0.07)	11.353	8.140	4.785
MSE	7.092	1.870	1.677

经过多路编码网络提取 4 个视角方向的 EPI-volume 的特征后，生成了 4 组 32 通道的像素块匹配特征。将 4 组特征串联在一起，特征数量拓展为了原来的 4 倍。此时，像素块匹配特征属于浅层的局部粗略特征，需要传输到特征聚合模块中得到融合浅层特征与深层特征的区域级特征来实现视差预测。

2.2　特征聚合模块

对于光场图像的视差预测来说，从图像中提取有效的特征非常重要。然而，在各式各样的复杂场景中，在对目标像素点进行视差预测时，该像素点被当作孤立像素点看待，若仅仅只考虑局部区域的信息，则无法得到准确的视差预测值。因此，有效获取全局上下文信息，将特征传递到更大的空间区域，使网络有效结合全局和局部特征线索，是获得可靠的视差预测的关键。

通常，下采样操作是融合上下文信息的有效操作，通过缩小特征图的尺寸来扩大网络的感受野，这样能有效地融合像素点邻域的上下文信息，从而对目标像素点进行视差预测。但是，下采样操作降低了图像的分辨率，会丢失图像的细节信息，即图像中物体的轮廓、位置信息以及像素点与物体的对应关系，降低了网络的分割能力。在语义分割领域，带跳跃连接的编码-解码器结构是聚合上下文信息同时保留图像细节信息的一种有效方式[17]，该方法利用跳跃连接将浅层较小感知域的局部特征与深层较大感知域的抽象特征进行整合来实现上下文信息的聚合。

如图 2 所示，我们设计了如下带跳跃连接的编码-解码器模块。对于多路编码模块传递来的像素块匹配特征，首先进行卷积操作以减少通道数，提高其数值稳定性，编码器部分的每一级是一个“Conv-BN-SeLu-Conv-BN-SeLu”卷积块，并采用卷积核大小为 2×2、步长为 2 的卷积操作来实现下采样。相比池化操作，利用卷积实现的下采样操作能通过控制步长更好地实现目标像素点的上下文信息的有效融合，减少因池化操作而导致图像中姿态和空间位置等对图像分割具有重要影响的信息的丢失。解码器部分的每一级由一个“Conv-BN-SeLu-Conv-BN-SeLu”卷积块和一次双线性上采样构成，同时在上采样操作后进行一次“Conv-BN-SeLu”操作来增加其数值稳定性。图 2 对特征聚合模块的网络结构以及超参数进行了详细的描述。

四维光场函数利用两个平行平面与光线的交点坐标来参数化表示光场，其表达式为：

$$L(u,v,s,t) \tag{1}$$

其中，(u,v)，(s,t)是光线与两个平面的交点坐标。

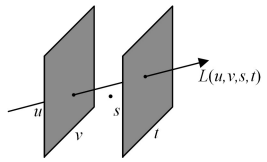


图 4　“双平面”模型参数化四维光场

Fig. 4　4D light field parameterized by two-plain model

在光场深度估计领域，无论是传统方法还是基于深度学习的方法，计算的都是以中心视角图为基准的视差图，且周围视角图与中心视角图之间存在着如下对应关系：

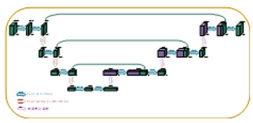
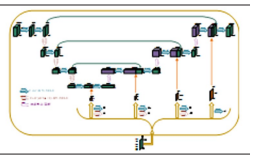
$$L(0,0,s,t)=L(u,v,s+d(s,t)*u,t+d(s,t)*v) \tag{2}$$

其中，$d(s,t)$是中心视角图中像素点(s,t)与相邻视角图中对应点的视差值，即其余视角图像与中心视角图像之间的视差呈线性关系。因此，我们将光场图像中心视角图作为参考图像，并将该参考图像经过不同深度的卷积块后得到的不同分辨率和精细度的结构特征映射连接到每级跳跃连接中，如图 2 所示。对中心视图进行不同深度的卷积操作，是为了获得对应精细度的边缘特征，这些特征反映了光场图像所拍摄场景的结构信息，将这些结构特征映射连接到对应深度的跳跃连接部分，能够提供传递特征图在边缘处的变化信息，也就是因获取上下文信息而进行的下采样操作丢失的物体轮廓信息。我们设计了相同架构的网络模型，该网络模型取消了将中心视图的结构特征在各级跳跃连接过程的映射链接。同

时，我们将该模型在测试集上的测试结果与本文方法的结果进行了对比，如表 2 所列。从对比结果可以看出，中心视图不同精细度的结构特征的引入有效提高了算法的精确度。

表 2 中心视角图的结构特征对性能的影响

Table 2 Effect of the structural features of center-view on performance

	无	有
中心视角图参考		
BadPix(0.07)	4.907	4.785
MSE	2.349	1.677

2.3 视差回归模块与损失

在视差回归模块，采用一个“Conv-BN-SeLu-Conv-BN-SeLu-Conv”卷积块，将特征聚合网络中聚合得到的融合浅层特征与深层特征的区域级特征体转换为单通道视差图。

对于整个模型，采用的是以随机初始化的方式进行端对端的训练，并利用真值视差图进行有监督的学习来训练本文模型。所使用的训练集中的视差标签是稀疏的，因此，对于图像的像素点，采用平均绝对误差(MAE)，即像素点(i,j)的真值视差与模型预测视差之间的绝对误差来训练本文模型。这种有监督的回归损失的定义如下：

$$loss=\frac{1}{N}\sum|disp(i,j)-disp_{gt}(i,j)| \tag{3}$$

其中，$disp(i,j)$表示输入图片中像素点(i,j)的模型预测视差值；$disp_{gt}(i,j)$表示输入图片中像素点(i,j)的真实视差值；N表示输入图像的像素点总数。

3 数据集与图像预处理

本文实验中选用的数据集是 Honauer 等[23-24]在 HCI-4D 光场基准测试中发布的 4D 光场数据集。

3.1 数据集介绍

4D 光场数据集是近两年来最常用的光场深度估计测试基准，分为“additional”“stratified”“training”3 类，共计 24 组场景。每个场景的光场图像的空间分辨率为 9×9，子孔径图像分辨率为 512×512，具有 R,G,B 3 个颜色通道值。所有光场图像均由 blender 渲染器渲染后获得，每个场景都包含不同的场景物体和纹理，设置不同的光照条件，并且场景物体间存在着符合场景客观条件的复杂遮挡情况的精细结构，是极其符合真实场景条件的光场图像。因为数据集是人工合成的，所以我们可以很轻松地获得所有场景的真值视差图。在网络训练过程及测试实验中，我们选择其中“additional”类的 16 组光场场景作为训练集，“stratified”“training”类的 8 组场景作为测试集。在对从训练集中随机选取 32×32 的灰度像素块进行训练后，使用图像的全分辨率即 512×512 进行验证。

3.2 数据增强

对于深度学习而言，网络模型的训练需要大量的训练数据做支撑，训练数据越多，对应的真值图就越精准，模型的效果就越好。但是，对于数据集中作为训练集的 16 组场景的数据量来说，无法满足网络训练至收敛的要求。因此，需要使用符合光场图像特殊几何结构的数据增强方法，通过保持相应输出标签的输入变换来增强训练数据集的大小和多样性，以解决网络训练过程中的过拟合问题[25]。

目前，计算机视觉领域的单样本数据增强主要有翻转、旋转、裁剪、变形、缩放以及颜色变换等。光场成像技术与经典多视图场景的关键区别在于密集而规则的采样[24]，故光场子孔径图像间存在着固定的几何关系，单纯采用上述数据增强方法而不对真值视差图做相应调整会破坏光场图像的几何结构，使网络模型的预测值的准确性受到极大的影响。因此，本文结合文献[26]中提出的数据增强方法，考虑了光场图像特殊的几何特性，使用了以下几种数据增强方法，并对对应的真值视差图做了相应的调整，最终有效扩充了训练数据量，达到了网络训练的基本要求。

(1)图像旋转(90°，180°，270°)。旋转是计算机视觉领域最常见的图像增强技术。然而，鉴于传统图像旋转方式没有考虑到子孔径图像间的方向特性，光场图像的图像增强不能单纯地用传统图像旋转方式进行数据扩充。在多路编码网络部分，每路网络提取的是对应方向的视角图集的极线特征。如果对图像进行旋转操作，则需要将其输入旋转后的极线特征对应方向的编码网络。如图 5 所示，以垂直编码网络为例，子孔径图像中黄线代表垂直方向的点集，它们构成的极平面图反映的是垂直方向上的点的极线特征，如果运用旋转操作将子孔径图像旋转 90°，那么此时黄线代表的是水平方向的点集，它们构成的极平面图反映的是水平方向的极线特征，因此需要将旋转 90°的垂直方向的视角图集组成的 EPI-volume 输入到水平编码网络中。

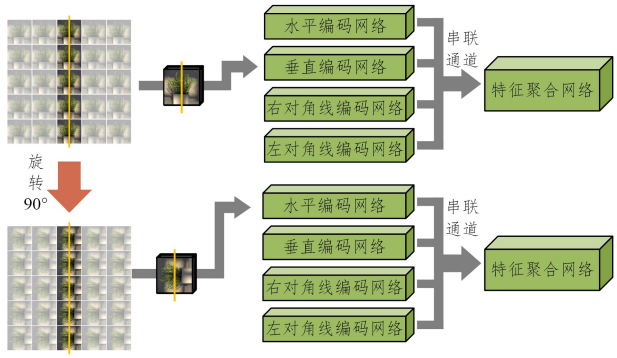


图 5 图像旋转(电子版为彩色)

Fig. 5 Image rotation

(2)图像缩放。图像缩放按照原图像大小的 1，1/2，1/3 的比例进行缩放，采用的是按照 1，2，3 大小的步长进行像素点的选取。在图像进行缩放后，视差图也需要进行相应的缩放。

(3)图像翻转。对图像进行翻转后，视差图也需要进行翻转。

(4)随机灰度变换。将子孔径图像按照式(4)进行随机灰度变换。

$$gray=w_R*R+w_G*G+w_B*B \tag{4}$$

其中，w_R,w_G,w_B 满足以下关系：

$$\begin{cases} w_R + w_G + w_B = 1 \\ w_R > 0 \\ w_G > 0 \\ w_B > 0 \end{cases} \tag{5}$$

(5)随机 gamma 变换。将图像处理为随机灰度变换后的光场图像，*gamma* 值为[0.8，1.2]的随机数，目的在于改变图片的整体灰度值，增强对比度，其计算式为：

$$s = r^{\gamma} \tag{6}$$

图 6　gamma 变换

Fig. 6　Gamma transform

值得一提的是，本文采取的数据增强方式是在训练过程中对输入数据进行处理，无需在训练之前对输入数据进行任何操作。经过上述数据增强操作后，用于训练的光场图像数量得到了极大的提升，为网络训练提供了可靠且充实的数据来源。

4　实验与结果分析

本文实验使用的电脑配置如下：CPU 为 Intel Core i7-4770，内存为 16GB，GPU 为 NVIDIA GTX 1080Ti。我们的网络模型以 Tensorflow 为后端，将 Keras 作为应用程序接口。本文方法花费大约 3 天的时间就能训练得到较好实验效果的网络模型。

4.1　网络参数及训练细节

网络训练采用的是 patch-wise 训练策略，在子孔径图像中随机裁剪 32×32 大小的像素块进行训练。并采用小样本随机梯度下降的方法，batch-size 大小设定为 32，学习率设定为 10^{-3}，使用 Adam 优化器对网络权重进行迭代优化。作为参考图像的中心视角图，按照其他视角图的参数进行随机灰度转换后，经过不同深度的卷积层得到对应各级分辨率和精细度的语义信息。在训练时，由于训练集中存在镜面及反射区域(如玻璃表面、金属表面等)，它们的值会在训练过程中产生错误的对应关系，我们采取将这些部分的像素块从训练中排除的策略。同时，我们在训练时还手动屏蔽了中心像素值与周围像素值绝对差小于 0.02 的无文本区域。

4.2　评价指标

在对模型的定量评价上，选用坏像素率和均方差作为评价指标。

(1)坏像素率(Bad Pixel Ratio，BadPix)。当模型对待测光场图像中某一像素点的视差预测值与给定真值视差图中对应像素点值的绝对误差大于设定的阈值时，则认为该像素点是一个坏像素点。坏像素点的数目占总像素数的比例被称为坏像素率，坏像素率反映的是模型计算的准确度，坏像素率越低，模型的准确度就越高。其计算式为：

$$BadPix(\varepsilon) = \frac{\{(i,j) \in N : |disp(i,j) - disp_{gt}(i,j)| > \varepsilon\}}{N} \tag{7}$$

其中，ε 表示设定的阈值。根据 HCI-4D 光场基准测试中的测试指标，将 ε 设定为 0.07。

(2)均方差(Mean Square Error，MSE)。均方差用于检测模型的视差预测值和真值之间的偏差，用对应像素点的预测数据和真实数据误差的平方和的均值来表示。MSE 常被用于评价数据的偏差程度，MSE 的值越小，说明模型对光场图像视差预测的精确度就越高。其计算式如下：

$$MSE = \frac{1}{N} \sum (disp(i,j) - disp_{gt}(i,j))^2 \tag{8}$$

4.3　实验结果

我们将本文方法与 4D 光场基准测试中排名靠前的算法在测试集上进行了对比。其中，PS_RF[27]，SPO[4] 是基于传统光场深度估计的方法，EPN+OS+GC[28]，SOA-EPN[14] 是将 EPI-volume 与卷积神经网络相结合进行深度估计的方法。并且，我们在表 3 中详细列出了本文方法和对比方法在 4D 光场数据集的测试集上的定量指标。从表中可以看出，本文方法在反映算法准确度的 BadPix(0.07)指标上取得了较好的结果，在反映精确度的 MSE 指标上更是普遍优于对比方法，证明了中心视图的结构特征能够对网络的视差图预测起到很好的辅助作用，保证网络的整体预测结果和真值不会产生较大的偏差。

表 3　实验结果对比

Table 3　Comparison of experimental results

Algorithm	AVG	*BadPix*(0.07)							
		Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard
PS_RF[27]	6.961	7.142	7.975	**0.107**	2.964	18.946	2.425	4.379	11.752
SPO[4]	8.233	3.781	16.274	0.861	14.987	15.889	2.594	2.184	9.297
EPN+OS+GC[28]	11.200	**3.328**	39.248	0.242	18.545	15.304	2.060	2.877	7.997
SOA-EPN[14]	7.166	3.452	31.005	0.243	**2.835**	**11.515**	0.854	1.707	5.722
Ours	**4.784**	4.163	**6.223**	0.508	5.362	15.034	**0.727**	**1.551**	**4.710**
Algorithm	AVG	*MSE*							
		Backgammon	Dots	Pyramids	Stripes	Boxes	Cotton	Dino	Sideboard
PS_RF[27]	3.694	6.892	8.338	0.043	1.382	9.043	1.161	0.751	1.945
SPO[4]	3.572	4.587	5.238	0.043	6.955	9.107	1.313	0.310	1.024
EPN+OS+GC[28]	5.980	3.699	22.369	0.018	8.731	9.314	1.406	0.565	1.744
SOA-EPN[14]	4.894	4.016	24.004	0.020	1.296	7.698	**0.484**	0.471	1.167
Ours	**1.677**	**3.204**	**1.774**	**0.013**	**0.951**	**6.157**	0.511	**0.141**	**0.666**

同时，图 7 给出了本文方法以及对比方法在“训练”类 4 个测试场景中预测的视差图的可视化结果。从左往右第一张图展示的是场景的中心视角图，紧接着展示了该场景的真值视差图。其中，Cotton 场景由简单平滑的表面结构组成，模型的预测难度较低。Dino 中通过在背景中加入墙角来增加场景深度，并加入了丰富的边缘信息及复杂的遮挡场景，提高了模型的预测难度。Sideboard 场景使用了纹理复杂的墙体背景，并在其中设置参差不齐的书本、电线等细小物体，大大提高了模型的预测难度。Boxes 场景中盒子表面由复杂的纹理且非连续深度的镂空框架组成，与框内的书本组成极为复杂的遮挡场景，对网络的要求极高。从图中可以看出，本文方法计算得到了准确的视差图，并且在遮挡场景中也能得到清晰的遮挡边界特征，证明了本文方法在聚合上下文信息的同时能够保留图中物体的轮廓、大小信息。同时，中心视图结构特征的引入为视差图的预测提供了特征图在边缘处的变化信息，极大地降低了预测视差图的全局误差。

此外，表 4 列出了本文方法与对比方法在测试集上的运算时间。从表中可以得到，本文方法的平均计算时间为 1.217 s，计算效率远远高于对比方法，能够在最短时间内得到准确的计算结果。

表 4 运算时间对比

Table 4 Comparison of calculating time

（单位：s）

Algorithm	Boxes	Cotton	Dino	Sideboard
PS_RF[27]	1189.644	1227.095	1297.555	1031.817
SPO[4]	2128.000	2025.000	2045.000	2073.000
EPN+OS+GC[28]	331.375	200.031	239.734	276.313
SOA-EPN[14]	72.257	74.267	72.259	75.598
Ours	**1.307**	**1.116**	**1.143**	**1.301**

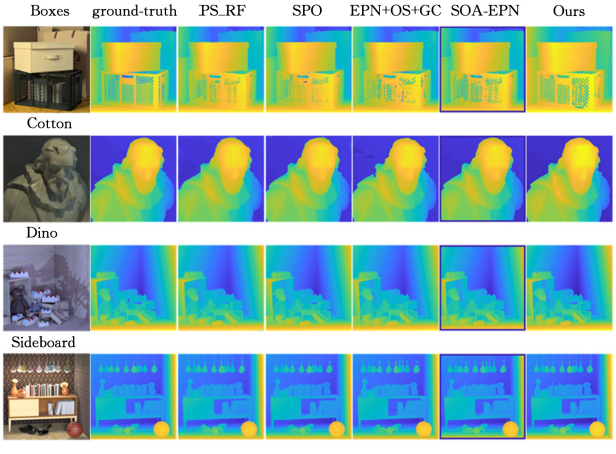


图 7 本文方法与对比方法的估计视差图的可视化结果

Fig. 7 Visualization results of estimated disparity maps of our method and the comparison method

最后，我们将合成光场图像训练集中训练的模型对“EPFL Light-Field Image Dataset[29]”和“INRIA Lytro Light Field Dataset[30]”提供的真实光场图像进行了测试，测试结果如图 8 所示，其中，前 4 个场景来自 EPFL 数据集，第 5 个场景来自 INRIA 数据集，选取的场景中都包含复杂的遮挡情况。从预测视差图中可以看到，本文方法在真实光场场景中也能够得到准确的结果，即使是在复杂遮挡场景下，也能计算得到清晰且锐利的遮挡边缘。这证明了本文提出的网络具有较强的泛化性能。

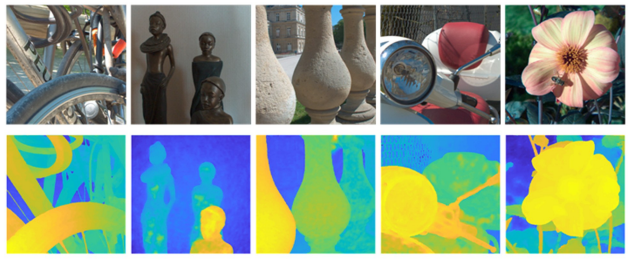


图 8 真实光场场景的估计视差图

Fig. 8 Estimated disparity maps of real light filed scene

结束语 本文提出了融合光场结构特征的基于编码-解码器架构的光场深度估计方法。本文方法以多方向 EPI-volume 为网络的输入，利用带跳跃连接的编码-解码器模块，在实现聚合图像上下文信息的同时补充了在下采样过程中丢失的物体轮廓、位置等信息，并在跳跃连接过程中加入作为参考图像的中心视图的不同精细度的结构特征，进一步补充了丢失的轮廓信息，极大地降低了预测视差图的全局误差，最终实现了对输入光场图像快速准确的深度估计。同时还使用符合光场结构特性的数据增强方式，使网络有效收敛并具有良好的泛化能力。但是，本文方法无法在复杂纹理及深度不连续区域取得准确的结果。下一步研究将考虑加入注意力机制，使网络能够充分学习传输特征的空间信息和通道信息，提高复杂纹理区域和深度不连续区域分割的准确性，从而提高网络在这些区域的深度估计精度。

参 考 文 献

[1] GERSHUN A. The Light Field[J]. Studies in Applied Mathematics, 1939, 18(1/2/3/4): 51-151.

[2] WANNER S, GOLDLUECKE B. Globally consistent depth labeling of 4D light fields[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 41-48.

[3] TOSIC I, BERKNER K. Light Field Scale-Depth Space Transform for Dense Depth Estimation[C]// Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition Workshops. 2014: 435-442.

[4] 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.

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

[6] CHEN C, LIN H, YU Z, et al. Light Field Stereo Matching Using Bilateral Statistics of Surface Cameras[C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1518-1525.

[7] KALANTARI N K, WANG T C, RAMAMOORTHI R. Learning-based view synthesis for light field cameras[J]. ACM Transactions on Graphics (TOG), 2016, 35(6): 1-10.

[8] YOON Y，JEON H G，YOO D，et al. Light-field image super-resolution using convolutional neural network[J]. IEEE Signal Processing Letters，2017，24(6)：848-852.

[9] WANG T C，ZHU J Y，HIROAKI E，et al. A 4d light-field dataset and cnn architectures for material recognition[C]// Proceedings of the European Conference on Computer Vision. 2016：121-138.

[10] SRINIVASAN P P，WANG T，SREELAL A，et al. Learning to synthesize a 4d rgbd light field from a single image[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：2243-2251.

[11] ZHONG T，JIN X，LI L，et al. Light field image compression using depth-based CNN in intra prediction[C]// Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). 2019：8564-8567.

[12] HEBER S，POCK T. Convolutional networks for shape from light field[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：3746-3754.

[13] 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.

[14] ZHOU W，LIANG L，ZHANG H，et al. Scale and Orientation Aware EPI-Patch Learning for Light Field Depth Estimation [C]// Proceedings of the International Conference on Pattern Recognition. 2018：2362-2367.

[15] TAGHANAKI S A，ABHISHEK K，COHEN J P，et al. Deep Semantic Segmentation of Natural and Medical Images：A Review[J]. Artificial Intelligence Review，2021，54(1)：137-178.

[16] KENDALL A，MARTIROSYAN H，DASGUPTA S，et al. End-to-end learning of geometry and context for deep stereo regression[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017：66-75.

[17] CHANG J R，CHEN Y S. Pyramid stereo matching network [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：5410-5418.

[18] HUANG P H，MATZEN K，KOPF J，et al. Deepmvs：Learning multi-view stereopsis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：2821-2830.

[19] WANNER S，GOLDLUECKE B. Variational Light Field Analysis for Disparity Estimation and Super-Resolution [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2014，36(3)：606-619.

[20] JOHANNSEN O，SULC A，GOLDLUECKE B. What Sparse Light Field Coding Reveals about Scene Structure[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016：3262-3270.

[21] STRECKE M，ALPEROVICH A，GOLDLUECKE B. Accurate Depth and Normal Maps from Occlusion-Aware Focal Stack Symmetry [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：2814-2822.

[22] SHENG H，ZHANG S，CAO X，et al. Geometric Occlusion Analysis in Depth Estimation Using Integral Guided Filter for Light-Field Image[J]. IEEE Transactions on Image Processing，2017，26(12)：5758-5771.

[23] HONAUER K，JOHANNSEN O，KONDERMANN D，et al. A Dataset and Evaluation Methodology for Depth Estimation on 4D Light Fields[C]// Proceedings of the Asian Conference on Computer Vision. 2016：19-34.

[24] JOHANNSEN O，HONAUER K，GOLDLUECKE B，et al. A Taxonomy and Evaluation of Dense Light Field Depth Estimation Algorithms[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017：82-99.

[25] BUSLAEV A，IGLOVIKOV V I，KHVEDCHENYA E，et al. Albumentations：fast and flexible image augmentations[J]. Information，2020，11(2)：125.

[26] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems，2012，25：1097-1105.

[27] JEON H，PARK J，CHOE G，et al. Depth from a Light Field Image with Learning-based Matching Costs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，41(2)：297-310.

[28] 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 International Conference on Neural Information Processing. 2017：642-652.

[29] RERABEK M，EBRAHIMI T. New Light Field Image Dataset [C]// 8th International Conference on Quality of Multimedia Experience (QoMEX). 2016.

[30] PENDU M L，JIANG X，GUILLEMOT C. Light Field Inpainting Propagation via Low Rank Matrix Completion[J]. IEEE Transactions on Image Processing，2018，27(4)：1981-1993.

YAN Xu，born in 1995，postgraduate. His main research interests include computer vision and deep learning.

XU Bing，born in 1960，senior research scientist，Ph.D supervisor. His research interests include application of adaptive optics in improving laser beam quality，wavefront detector development，and application of light field cameras.