DOI: 10.13382/j.jemi.B2306247

基于光场 EPI 图像栈的 6D 位姿估计方法*

李 扬 张旭东 孙 锐 范之国
(合肥工业大学计算机与信息学院 合肥 230601)

摘 要:光场相机单次拍摄可以同时记录光线的强度与方向信息,相较于 RGB 相机能够更好地揭示场景的三维结构和几何特征,在目标 6D 位姿估计领域具有独特优势。针对现有 RGB 位姿估计方法存在复杂场景下检测精度低、鲁棒性差的问题,本文首次提出了一种基于光场图像的端到端卷积神经网络目标位姿估计方法。该方法首先利用双路 EPI 编码模块实现高维光场数据的处理,通过重构出光场 EPI 图像栈和引入水平和垂直 EPI 卷积算子,提高对光场空间角度信息关联的建模能力,并由双分支孪生网络进行光场图像的浅层特征提取。其次,设计了带跳跃连接的特征聚合模块,对串联后的水平和垂直方向光场 EPI 浅层特征进行全局上下文聚合,使网络在逐像素关键点位置预测时有效结合全局和局部特征线索。针对光场数据不足问题,本文使用 Lytro Illum 光场相机采集真实场景,构建了一个丰富且场景复杂的光场位姿数据集——LF-6Dpose。在光场位姿数据集 LF-6Dpose 上的实验结果表明,该方法在 ADD-S 和 2D Projection 指标下平均位姿检测精度分别为 57.61%和 91.97%,超越了其他基于 RGB 的先进方法,能够更好地解决复杂场景下的目标 6D 位姿估计问题。

关键词: 光场;6D 位姿估计;光场位姿数据集;EPI 图像栈;特征聚合模块;关键点
中图分类号: TN91;TP391.41 **文献标识码**: A **国家标准学科分类代码**: 510.4050

6D pose estimation method based on light field EPI image stack

Li Yang Zhang Xudong Sun Rui Fan Zhiguo
(School of Computer and Information, Hefei University of Technology, Hefei 230601, China)

Abstract: A single shot of a light field camera can record the intensity and direction information of light at the same time. Compared with the RGB camera, it can better reveal the three-dimensional structure and geometric characteristics of the scene and has unique advantages in the field of object 6D pose estimation. Aiming at the problems of low detection accuracy and poor robustness in complex scenes in existing RGB pose estimation methods, this paper proposes an end-to-end convolutional neural network object pose estimation method based on light field images for the first time. In this method, the dual-channel EPI encoding module is used to process high-dimensional light field data. By reconstructing the light field EPI image stack and introducing horizontal and vertical EPI convolution operators, the modeling ability of the spatial angle information association of the light field is improved. Two-branch siamese network for shallow feature extraction of light field images. Secondly, a feature aggregation module with skip connection is designed to perform global context aggregation on the concatenated light field EPI shallow features in the horizontal and vertical directions, so that the network can effectively combine global and local feature clues when predicting pixel-by-pixel key point positions. To solve the problem of insufficient light field data, this paper uses the Lytro Illum light field camera to collect real scenes and constructs a rich and complex light field pose dataset—LF-6Dpose. The experimental results on the light field pose dataset LF-6Dpose show that the average pose detection accuracy of this method is 57.61% and 91.97% under the ADD-S and 2D Projection indicators, which surpasses other advanced methods based on RGB and can better solve the target 6D pose estimation problem in complex scenes.

Keywords: light field; 6D pose estimation; light field pose dataset; EPI image stack; feature aggregation module; key points

收稿日期: 2023-02-09 Received Date: 2023-02-09
*基金项目:国家自然科学基金(61876057)、安徽省科技重大专项(202103a06020010)、安徽省自然科学基金(2208085MF158)项目资助

0　引　言

物体六自由度(6-degree of freedom,6-DoF)位姿估计旨在确定场景中目标物体的三维空间位置和三维旋转姿态[1],精确的位姿估计对于增强现实[2]、自动驾驶[3]和机器人抓取操作[4]等领域具有关键意义。近些年,随着视觉传感器的快速发展,研究人员在基于 RGB 相机的 6D 位姿估计方面做了大量研究并取得显著的进展[5-7]。传统方法[8-10]首先从不同视角对 3D 物体模型进行投影,得到多个方向的投影模板图像,然后对目标图像和模板图像进行特征检测和匹配,间接寻找目标图像与 3D 物体模型的对应关系实现位姿求解,或直接以模板图像对应的位姿作为估计结果。但传统方法特征提取能力有限,无法有效应对光照变化和背景杂乱等场景,且计算步骤复杂,实时性较差。卷积神经网络凭借强大的特征提取能力,使得现今大多数位姿估计任务的研究都由其主导。Rad 等[11]提出 BB8 姿态估计网络检测图像中物体三维边界框顶点的 2D 投影位置,然后利用 n 点透视(perspective-n-point,PnP)算法求解 6D 位姿,但是该方法后期需要迭代最近点(iterative closest point,ICP)算法进行位姿迭代优化,严重影响网络整体运行速度。基于类似的想法,Tekin 等[12]在轻型目标检测器 YOLOv2[13]的基础上开发了 YOLO6D,可以快速地预测物体三维边界框顶点在二维图像上的投影位置实现位姿估计。但当物体处于图像边缘处时,边界框顶点对应的投影位置会出现在图像界外,导致检测失效。Peng 等[14]提出 PVNet 像素级投票网络,通过预先定义一组位于物体表面上的关键点,由图像中物体可见部位的每个像素生成一个指向关键点的单位方向向量,然后对所有向量进行随机抽样一致(random sample consensus,RANSAC)投票,其得到的 2D 关键点位置会更加鲁棒。Xiang 等[15]提出了一种基于回归的方法,直接从输入图像回归出六自由度位姿参数,但是卷积神经网络难以直接估计具有非线性的 3D 旋转参数。Wang 等[16]使用 ResNet[17]预测密集的 2D-3D 对应关系的几何特征,作为指导直接回归 6D 姿态的中间特征,然后使用简洁高效的 2D 卷积加全连接层拟合不可微分的 PnP 和 RANSAC,得到目标物体的位姿信息。但是当物体处于严重遮挡时,预测密集的 2D-3D 对应关系会存在很大的挑战。由于 RGB 图像只能从单一视角记录场景空间位置特征,不具备场景的多视角信息和深度信息,无法充分揭示场景的三维结构和几何特征,导致基于 RGB 的位姿估计方法在复杂场景下仍然无法精确稳定地预测物体位姿。

随着计算成像学的进步,光场成像技术在处理计算机视觉问题表现出显著的性能提升和良好的发展前景[18]。和传统 RGB 相机不同,光场相机[19]在主透镜和像素传感器之间插入微透镜阵列,单次拍摄就可以记录三维场景的空间和角度信息,获取的信息比传统相机多出两个自由度。其中空间信息反映了场景的位置特征,角度信息记录了场景的视角变化,因此光场数据能够更好地揭示场景的三维结构。正是由于光场的多视角信息蕴藏着场景的深度信息和几何特征,其在深度估计、显著性检测、三维重建等计算机视觉任务中得到了广泛的应用。Tao 等[20]利用角度一致性构造散焦和匹配线索提取场景的深度信息,并通过非线性最小二乘法优化初始深度图以获取高精度的深度图。Johannsen 等[21]提出第 1 个适用于光场相机的运动恢复结构算法—LF-SFM(light field-structure from motion)。Li 等[22]首次提出光场显著性检测算法,证明在相似前后景、尺度变化等复杂场景中,基于光场数据的显著性检测算法有着更高的检测精度。相比 RGB 图像,光场数据的多视角将场景的外观特征和深度特征融合在一起,能为物体位姿估计提供可靠的背景、深度等先验信息。本文首次将光场成像技术应用到目标 6D 位姿估计领域,提出一种基于光场极平面图像(epipolar plane image,EPI)图像栈的端到端卷积神经网络架构,有效提升复杂场景下物体位姿估计的效果。同时针对光场位姿数据不足问题,本文构建一个丰富且场景复杂的光场位姿数据集 LF-6Dpose(light field-6Dpose)作为数据支撑。本文的贡献总结如下:

1)首次提出一种基于光场 EPI 图像栈的端到端卷积神经网络架构预测目标物体的 6D 位姿信息。该框架通过双路 EPI 编码模块学习光场 EPI 图像栈的特征信息,并利用特征聚合模块进一步融合光场深层结构特征,构建端到端网络学习光场 EPI 图像栈与目标位姿信息之间的映射关系。

2)提出一种高效提取光场多视角信息的双路 EPI 编码模块。该模块首先重构光场 EPI 图像栈并设计水平和垂直 EPI 卷积算子,提高对光场空间角度信息关联的建模能力,同时适应特定维度输入的光场 EPI 图像栈,然后由双分支孪生网络提取光场浅层特征。

3)构建一个新的光场数据集 LF-6Dpose 用于物体 6D 位姿估计,该数据集包含 5 类物体,每类物体由约 1200 张光场图像、掩码图、物体三维模型和真值位姿标签组成,覆盖了目标纹理颜色相似、尺度变化、背景杂乱和遮挡等挑战性场景。

1　基于光场 EPI 图像栈的 6D 位姿估计方法

本文首次提出利用光场图像丰富的多视角信息解决复杂场景下的目标位姿估计问题,而如何高效挖掘光场的多视角信息是本文的关键。结合光场 EPI 图像的特点

和卷积神经网络的优势,本文设计了一种基于光场 EPI 图像栈的端到端卷积神经网络方法来预测场景中目标物体的 6D 位姿信息。网络整体结构如图 1 所示,主要分为 3 个部分,双路 EPI 编码模块,特征聚合模块、关键点回归和位姿解算。为了高效提取光场图像蕴藏的场景三维结构信息,本文设计的双路 EPI 编码模块首先选取中心行和列的子孔径堆叠图像,重构出水平和垂直 EPI 图像栈,然后再对其进行特征提取得到光场图像的浅层特征。在双路 EPI 编码模块中,本文引入水平和垂直 EPI 卷积算子,实现对光场空间角度信息的显式建模,并灵活地适应处理特定维度的 EPI 图像栈。同时采用双分支孪生网络分别对水平和垂直 EPI 图像栈的结构特征进行深层学习,最后将双路提取的 EPI 特征进行连接,传入到特征聚合模块。带跳跃连接的特征聚合模块可以对光场图像的局部特征和全局特征进行有效的上下文聚合,为像素级的关键点位置预测提供更加准确的全局和局部特征线索。最后通过关键点回归模块得到准确的 2D 关键点位置,并使用 PnP 法求出旋转矩阵 $\boldsymbol{R}$(Rotation)和平移量 $\boldsymbol{T}$(Translation),作为目标物体的 6D 位姿信息。下面对网络的 3 个模块进行详细介绍。

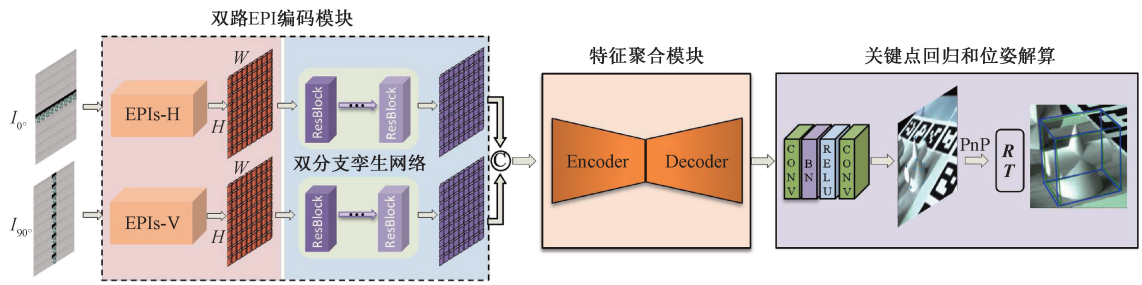

图 1　本文方法的网络结构

Fig. 1　The network structure of the method

1.1　双路 EPI 编码模块

对光场蕴藏的场景信息和几何特征进行高效提取是光场位姿估计的关键一步。相较于隐式包含 EPI 特征的子孔径堆叠图像,光场极平面图像栈可以显式地表达整个场景的 EPI 特征。因此,本文设计了能够高效提取光场特征信息的双路 EPI 编码模块。

1)EPI 图像栈重构和空间角度特征建模

EPI 图像是 4D 光场的二维切片,传统的 EPI 图像只包含了单一空间维度,而本文的任务是解决场景中目标物体的位姿信息估计,其需要整个场景的空间信息。因此将 EPI 图像在另一空间维度进行拼接,使得单一空间维度的 EPI 图像扩展为二维空间维度的 EPI 图像栈,从而得到整个场景的空间信息,同时 EPI 图像栈包含了该 EPI 方向的所有视角信息。本文对中心行和中心列子孔径堆叠图像的 EPI 切片进行了拼接,重构出水平 EPI 图像栈和垂直 EPI 图像栈,以此代替传统的子孔径堆叠图像作为网络的输入。极平面图像中的斜线能够很好地反映空间与角度之间的关联性,基于光场极平面图像这种结构先验,对于输入的 EPI 图像栈,本文设计了水平和垂直 EPI 卷积算子对 EPI 图像栈中空间角度信息进行显式建模,这样不仅可以对 EPI 结构特征进行显式学习,而且降低了卷积层的参数量和学习难度,提升网络性能。并且本文设计的水平和垂直 EPI 卷积算子可以灵活地处理特定维度的 EPI 图像栈,因此对不同维度大小的 EPI 图像栈,均输出与场景空间分辨率相同的特征图。水平 EPI 图像栈的处理过程如图 2(a)所示。

选取中心行输入的 N 张子孔径图像,对这 N 个视角形成的子孔径堆叠图像进行 transpose 和 reshape 操作,可以得到 EPI 图像栈尺寸大小为 $(N\times H,W)$,$N\in(3,5,7,9)$,本文中 $N=9$,$H\times W$ 为场景的空间分辨率。图中相同颜色的方格代表同一场景点在不同子孔径图像中的像素位置。当场景一点处于聚焦深度时,其在各个子孔径图像中的像素处于同一位置,在水平 EPI 图像中呈现为一条垂直线。当场景点处于非聚焦深度时,其在各个子孔径图像中的像素不再处于同一位置,而是产生了一些偏移,因此在水平 EPI 图像中表现为具有不同斜率的斜线,斜率的不同程度代表场景点处于不同的深度。本文构建的光场 EPI 图像栈可以对整个场景空间点进行 EPI 结构形式的显示,相较于单一空间维度的 EPI 图像,扩展到了二维空间。

针对输入维度为 $(N\times H,W)$ 的水平 EPI 图像栈,本文定义水平 EPI 卷积算子 EFE-H(horizontal EPI feature extractor)为一个 kernel size = (N,N), stride = $[N,1]$ 的“Conv-BN-ReLU”卷积块进行预处理。使用大小为 $N\times N$,水平空间维度步进为 1,垂直空间维度步进为 N 的卷积核,可以提取属于相同行视角的局部空间角度信息,实现对光场空间角度关联特征的显式建模,并得到与场景空间分辨率相同的输出特征图。通过水平 EPI 卷积算子预处理整个场景的 EPI 结构特征,能很好地学习到场景点的三维结构信息和几何特征,为位姿估计提供有效的场景信息。垂直 EPI 图像栈采用了类似的方法进行重构,可以得到垂直 EPI 图像栈的尺寸大小为 $(H,N\times$

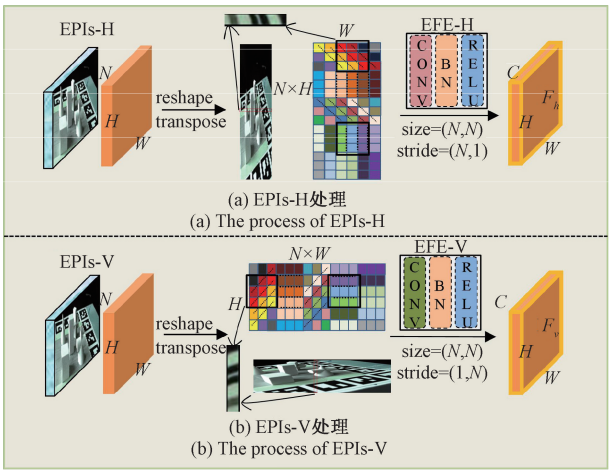

图 2　EPI 图像栈重构和空间角度特征建模

Fig. 2　EPI image stack reconstruction and spatial angle feature modeling

W)。同时对输入尺寸为(H,$N \times W$)的垂直 EPI 图像栈,本文定义垂直 EPI 卷积算子 EFE-V(vertical EPI feature extractor)为一个 kernel size = (N,N),stride = [1, N]的"Conv-BN-ReLU"卷积块。采用大小为 $N \times N$,水平空间维度步进为 N,垂直空间维度步进为 1 的卷积核,可以显式提取属于相同列视角的局部空间角度信息,同时得到与场景空间分辨率相同的输出特征图。其中垂直 EPI 图像栈的重构和建模过程如图 2(b)所示。

2)双分支孪生网络

在对 EPI 图像栈的空间角度信息进行显示建模后,本文设计了双分支孪生网络对场景的 EPI 结构特征进行深层学习,每支路网络处理一个方向的 EPI 图像信息,其中双分支孪生网络不采用权值共享。分支网络由 4 个基本残差块构成,每个残差块由"Conv-BN-ReLU-Conv-BN-ReLU"的卷积块组成,借助残差网络融合浅层的细节特征进一步提高网络性能。本文在最后 1 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。最后将水平和垂直方向提取的特征进行直接相连,传入到特征聚合模块。

1.2　特征聚合模块

在预测光场图像中目标物体的关键点位置时,虽然深层网络提取的高级特征包含抽象的语义信息,有助于目标关键点的定位和噪声去除,但是缺少低层特征包含的边缘、纹理等更详细的空间结构细节,仍然无法得到准确的关键点位置。因此,为了使网络有效结合全局和局部特征线索,引入带跳跃连接的编解码结构实现全局上下文信息的聚合。图 3 展示了特征聚合模块的网络结构。

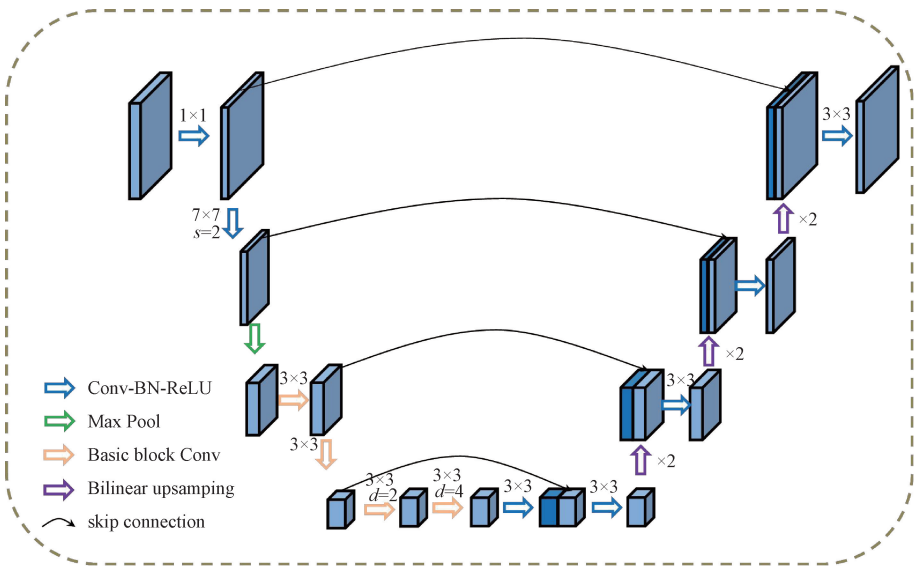

图 3　特征聚合模块

Fig. 3　Feature aggregation module

编码器部分由 ResNet-18 网络变体构成,为了适应双路 EPI 编码模块处理后的光场特征图,将原 ResNet-18 的第一个卷积替换为"Conv-BN-ReLU"卷积块,且当特征图的高宽降为原来输入的 1/8 时,不再进行下采样,同时将全连接层替换为卷积层。具体地,在对双分孪生网络提取的光场浅层特征进行串联后,首先使用 1×1 卷积块操作对水平和垂直 EPI 特征进行跨通道聚合,减少通道数,以提高数值稳定性。然后采用卷积核大小为 7×7、步长为 2 的"Conv-BN-ReLU"卷积块操作来实现下采样。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像素级预测具有重要影响的信息的丢失，最后交由后续网络继续处理。

解码器部分的每一级由一个“Conv-BN-ReLU”卷积块和一次双线性上采样构成，采用跳跃连接的架构将光场图像的浅层特征传递到解码器的每一级。在上采样操作完成后进行一次“Conv-BN-ReLU”操作来增加其数值稳定性。带跳跃连接的编解码结构是融合上下文信息同时保留图像细节信息的一种有效方式，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，从而为像素级的关键点位置预测提供更加准确的全局和局部特征线索。

1.3 关键点回归和位姿解算

本模块采用 PVNet[14] 的思想，网络通过学习代表关键点位置的方向矢量场，实现目标物体关键点在图像上的定位，从而间接预测目标物体的 6D 位姿信息。在关键点位置回归模块中，本文对特征聚合模块输出的最终 feature map 使用“Conv-BN-LeakyReLU-Conv”卷积块以得到方向矢量图，大小为 $H \times W \times 2 \times K$，$K$ 为关键点个数，$H \times W$ 为中心视角图像分辨率大小。方向矢量图代表图像中物体可见部位的每个像素，所生成一个指向每个关键点的单位向量。然后对这些单位向量使用 RANSAC 进行投票，得到关键点在中心视角图像上的 2D 像素位置。最后利用已知关键点在本体坐标系下的 3D 坐标，使用透视投影法求出旋转矩阵 $\boldsymbol{R}$ 和平移量 $\boldsymbol{T}$，作为目标物体 6D 位姿。其中，本文中提到的关键点是预先定义的，采取最远点采样法确定。网络模型采用光场图像中心视角的真值标签进行有监督的学习。

1.4 数据集构建

针对当前存在光场位姿估计数据不足问题，本文首次使用 Lytro Illum 光场相机构建了一个丰富且场景复杂的光场位姿估计数据集—LF-6Dpose。该数据集选择 5 类无纹理、颜色相似、具有朗伯特性的石膏体作为采集对象。将采集对象统一放置在科尔多瓦大学增强现实(Augmented Reality University of Cordoba，ArUco)标记板上，以便在采集图像时计算位姿真值。对每一个目标从不同的角度和距离采集光场数据，设置采集距离 30～180 cm，水平角 5°～85°，偏正角 5°～75°。本文为每类物体拍摄了约 1 200 张光场数据，覆盖多尺度、光照变化、背景杂乱和遮挡等复杂场景，并为每个光场数据中的目标物体提供了 6D 位姿真值标签和掩码图，同时使用 3D 软件为每类物体创建了三维模型。图 4 说明了 LF-6Dpose 数据集的构建过程。使用 Lytro Illum 光场相机(图 4(a))获取原始 4D 光场数据，并使用 MATLAB 光场工具箱将原始 4D 光场数据解码为 2D 子孔径图像阵列，其角度分辨率为 15×15，每个子孔径图像的空间分辨率为 416× 608。考虑到子孔径阵列的 4 个边角区域图像存在不同程度的颜色失真和模糊问题，选择中间的 9×9 视角作为采集的光场多视角图像，如图 4(b)所示，其中央视角图像如图 4(c)所示。结合物体的三维模型 4(图 4(d))和中央视角图像，通过 ArUco 标记板对目标物体进行了 6D 位姿真值标注和掩码图的分割。图 4(e)和(f)分别是十字圆锥(crosscone)的 6D 位姿标签可视化、掩码图。

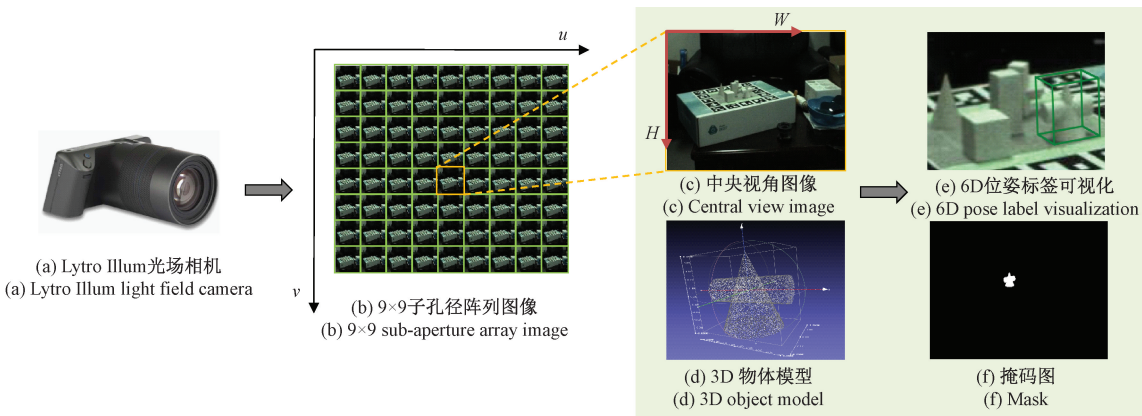

图 4 LF-6Dpose 数据集构建流程

Fig. 4 The process of LF-6Dpose dataset construction

2 实验结果与分析

本文实验使用的环境配置为 CPU E5-2620，RAM 32 GB，GTX TITAN X * 3，Ubuntu 18.04 操作系统，网络模型采用 Pytorch 实现。实验采用 LF-6Dpose 光场位姿数据集，为了防止过拟合，本文对该数据集进行了数据增强。网络初始学习率设置为 0.001，每 20 个 epoch 将其减半，

所有模型都经过 240 个 epoch 的训练。

2.1　评价指标

定量分析实验采用平均最近点三维距离 ADD-S (average closest point 3D distances for symmetric objects)和二维重投影误差 2D Projection 两个评价指标。ADD-S 是计算真实 3D 模型点到预测 3D 模型点的最近平均距离，适用于评估机器人抓取操作应用。

$$ADD-S=\frac{1}{m}\sum_{x_1\in M}\min_{x_2\in M}\|(\boldsymbol{R}x_1+\boldsymbol{T})-(\tilde{\boldsymbol{R}}x_2+\tilde{\boldsymbol{T}})\| \tag{1}$$

二维重投影误差，给定真值位姿和预测位姿，计算 3D 模型点在二维图像上的投影像素误差，适用于评估增强现实应用。

$$2D_{projection}=\frac{1}{m}\sum_{x_1\in M}\|K(\boldsymbol{R}x_1+\boldsymbol{T})-K(\tilde{\boldsymbol{R}}x_1+\tilde{\boldsymbol{T}})\| \tag{2}$$

式中：M 表示 3D 模型上点的集合，m 表示点的个数，x_1 和 x_2 表示 3D 模型上的点，K 表示相机内参，$\boldsymbol{R}$ 和 $\boldsymbol{T}$ 表示真实位姿，$\tilde{\boldsymbol{R}}$ 和 $\tilde{\boldsymbol{T}}$ 表示预测位姿。当 ADD-S 指标小于模型直径的 20%，判定当前位姿预测正确。当 2D Projection 小于 5 pixels，判定当前位姿预测正确。

2.2　消融实验

为了进一步验证本文方法的有效性，本文针对双路 EPI 编码模块进行了消融实验，网络其他部分保持不变。消融实验主要分为 3 个部分。

1)为了验证将子孔径堆叠图像重构成光场 EPI 图像栈进行处理的有效性，本文设计了由相同子孔径图像构成的子孔径堆叠图像为输入的处理模块，记作实验 1。

2)为了验证双分支网络的有效性，本文分别设计了水平和垂直两个单分支网络，分别记作实验 2 和实验 3。

表 1 给出了 ADD-S 和 2D Projection 两种指标下正确估计位姿所占的百分比。如表 1 所示，采用光场 EPI 图像栈作为输入，相比子孔径堆叠图像有更好的预测结果。这是因为子孔径堆叠图像隐式包含的 EPI 结构特征存在于更高维度，导致网络学习难度增加。而 EPI 图像栈显式构建了整个场景的空间角度信息，更好地揭示了场景的几何结构特征。通过设计高效的水平和垂直卷积算子，能够提取更有效的场景信息。而相较于两个独立的单分支处理网络，双路 EPI 编码模块可以同时提取水平和垂直两个方向的光场 EPI 特征，实验效果也有着明显提升，这验证了本文双路 EPI 编码模块结构的有效性。

表 1　消融实验结果

Table 1　Quantitative results of ablation experiment

指标	实验 1	实验 2	实验 3	本文
ADD-S	47.42	51.68	52.40	**57.61**
2D Projection	86.62	88.49	88.92	**91.97**

2.3　与其他方法对比

在 LF-6Dpose 光场位姿数据集上，将本文方法与当前先进的基于 RGB 图像的位姿估计方法进行了对比：YOLO6D[12]、PVNet[14]、GDR-Net[16] 和 PFA[23]。基于 RGB 图像的方法采用光场的中心视角图像作为输入。表 2 和 3 详细列出了本文方法与对比方法分别在 ADD-S 和 2D Projection 下正确估计位姿所占的百分比(加粗为最优结果，下划线为次优结果)。

从表 2 和 3 中可以看出，本文方法在 ADD-S 和 2D Projection 指标上的平均估计精度均优于其他方法。其中在 ADD-S 指标上，本文方法在 4 个目标物体上取得最优，1 个目标物体上取得次优。在 2D Projection 指标上，本文方法在 3 个目标物体上取得最优，2 个目标物体上取得次优。其中，相较于采用同种目标位姿估计策略的 PVNet，本文方法的性能得到了显著提升，在 ADD-S 上提升 23%，2D Projection 上提升 6%。两者投影的 3D bounding box 渲染在图像上的效果如图 5 所示。图中从左到右 4 列分别为 crosscone、cone、crosscuboid、cube 和 cuboid 5 类目标物体。每一列中分别为光场中央视角图像、PVNet 预测 3D 边界框的渲染结果、本文方法预测 3D 边界框的渲染结果。其中，浅色框(绿色)表示真值结果，深色框(蓝色)表示预测结果。

表 2　不同方法在 ADD-S 指标上的定量比较

Table 2　Quantitative comparison of different methods on ADD-S indicators

方法	AVG	ADD-S				
		Cone	Crosscone	Crosscuboid	Cube	Cuboid
YOLO6D	40.45	24.68	32.70	28.20	60.46	56.22
PVNet	46.89	38.75	36.82	32.58	64.64	61.67
GDR-Nett	52.87	40.82	46.44	42.63	<u>69.86</u>	64.61
PFA	<u>54.50</u>	**43.07**	<u>48.75</u>	<u>46.96</u>	68.64	<u>65.08</u>
本文	**57.61**	<u>42.56</u>	**53.97**	**48.14**	**74.51**	**68.87**

表3　不同方法在2D Projection指标上的定量比较

Table 3　Quantitative comparison of different methods on 2D projection indicators

方法	AVG	2D Projection				
		Cone	Crosscone	Crosscuboid	Cube	Cuboid
YOLO6D	80.53	81.85	79.12	72.27	86.82	82.61
PVNet	86.90	86.42	86.62	86.15	88.61	86.72
GDR-Net	89.61	89.96	88.93	87.02	91.88	90.24
PFA	90.78	**92.18**	89.01	**90.76**	91.32	90.62
本文	**91.97**	91.73	**90.38**	90.51	**94.86**	**92.38**

上述实验结果表明了本文提出的基于光场EPI图像栈的卷积神经网络方法在目标物体6D位姿预测上的优势,通过高效学习光场图像包含的丰富场景特征,可以保证网络的整体预测结果和真值不会产生较大的偏差,在复杂场景下有着更好的抗干扰能力。

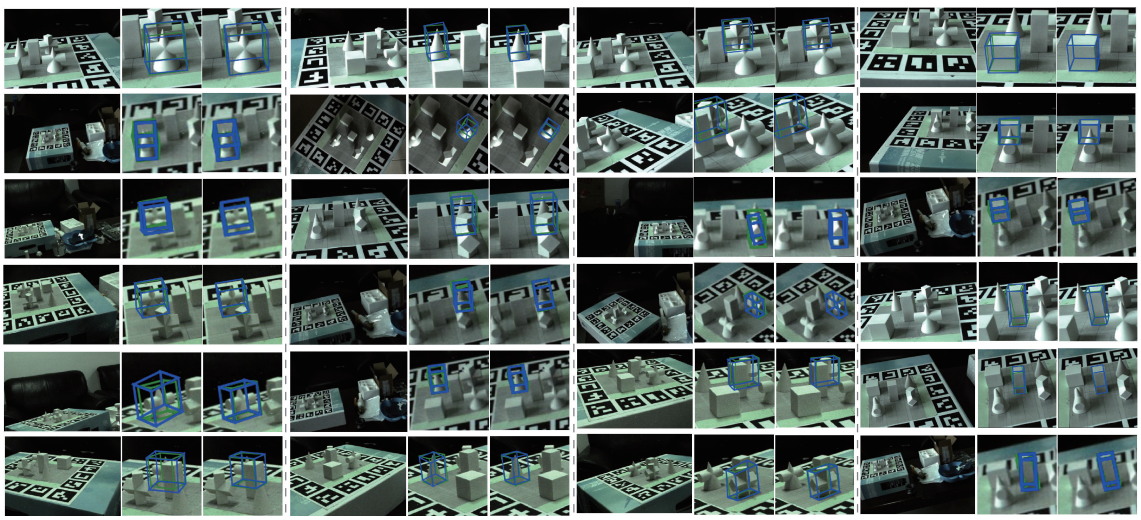

图5　3D bounding box可视化

Fig. 5　3D bounding box visualization

3　结　论

针对复杂场景下的目标位姿估计问题,本文提出了一种基于光场EPI图像栈的端到端卷积神经网络目标位姿估计方法,并使用Lytro Illum光场相机构建了一个具有挑战性的光场位姿数据集LF-6Dpose。本文设计的双路EPI编码模块首先构建水平和垂直光场EPI图像栈,然后使用EPI卷积算子和双分支孪生网络对光场空间角度信息进行显示建模和特征提取,高效挖掘光场图像的场景信息和几何特征。再使用带跳跃连接的编码-解码架构进行上下文特征聚合,提升光场结构特征利用效率。本文提出的方法可以更好地解决复杂场景下的目标6D位姿估计问题。下一步研究将考虑引入双重注意力机制,使网络能够充分学习水平和垂直EPI图像栈的空间信息和通道信息,进一步提高目标位姿估计精度。

参考文献

[1] XIANG Y, MOTTAGHI R, SAVARESE S. Beyond pascal: A benchmark for 3D object detection in the wild[C]. IEEE Winter Conference on Applications of Computer Vision. IEEE, 2014: 75-82.

[2] MARCHAND E, UCHIYAMA H, SPINDLER F. Pose estimation for augmented reality: A hands-on survey[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 22(12): 2633-2651.

[3] WU D, ZHUANG Z, XIANG C, et al. 6D-VNet: End-to-end 6-DoF vehicle pose estimation from monocular RGB images [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Workshops, 2019.

[4] TREMBLAY J, TO T, SUNDARALINGAM B, et al. Deep object pose estimation for semantic robotic grasping of household objects [J]. arXiv preprint arXiv: 1809.10790, 2018.

[5] SAHIN C, GARCIA-HERNANDO G, SOCK J, et al. A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators [J]. Image and Vision Computing, 2020, 96:103898.

[6] FAN Z, ZHU Y, HE Y, et al. Deep learning on monocular object pose detection and tracking: A comprehensive overview[J]. ACM Computing Surveys, 2022, 55(4): 1-40.

[7] DU G, WANG K, LIAN S, et al. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: A review[J]. Artificial Intelligence Review, 2021, 54 (3): 1677-1734.

[8] 崔建国,孙长库,李玉鹏,等. 基于 SURF 的快速图像匹配改进算法[J]. 仪器仪表学报,2022,43(8):47-53.
CUI J G, SUN CH K, LI Y P, et al. Improved fast image matching algorithm based on SURF[J]. Chinese Journal of Scientific Instrument,2022,43(8):47-53.

[9] HINTERSTOISSER S, LEPETIT V, ILIC S, et al. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes[C]. Computer Vision - ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. Springer Berlin Heidelberg, 2013: 548-562.

[10] 张艳,王宇. 基于视觉里程计的室内位姿测量技术研究[J]. 电子测量与仪器学报,2022,36(6):73-81.
ZHANG Y, WANG Y. Research on indoor pose measurement technology based on visual odometry[J]. Electronic Measurement Technology, 2022, 36 (6): 73-81.

[11] RAD M, LEPETIT V. BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 3828-3836.

[12] TEKIN B, SINHA S N, FUA P. Real-time seamless single shot 6D object pose prediction[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.

[13] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[14] PENG S D, LIU Y, HUANG Q X. PVNet: Pixel-wise voting network for 6DoF object pose estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 1-12.

[15] XIANG Y, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes[J]. arXiv preprint arXiv:1711.00199, 2017.

[16] WANG G, MANHARDT F, TOMBARI F, et al. GDR-Net: Geometry guided direct regression network for monocular 6D object pose estimation[J]. arXiv preprint arXiv:2102.12145, 2021.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[18] WU G, MASIA B, JARABO A, et al. Light field image processing: An overview[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(7):926-954.

[19] 周志良. 光场成像技术研究[D]. 合肥:中国科学技术大学, 2012.
ZHOU ZH L. Research on optical field imaging technology [D]. Hefei: University of Science and Technology of China, 2012.

[20] 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.

[21] JOHANNSEN O, SULC A, GOLDLUECKE B. On linear structure from motion for light field cameras[C]. 2015 IEEE International Conference on Computer Vision (ICCV). IEEE, 2015.

[22] LI N, YE J, YU J, et al. Saliency detection on light field[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2014, 2014.

[23] HU Y, FUA P, SALZMANN M. Perspective flow aggregation for data-limited 6d object pose estimation[C]. Computer Vision-ECCV 2022: 17th European

Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part II. Cham: Springer Nature Switzerland, 2022: 89-106.

作者简介

李扬,2018 年于西南科技大学获得学士学位,现为合肥工业大学计算机与信息学院硕士研究生,主要研究方向为机器视觉。
E-mail: leeyang@ mail. hfut. edu. cn

Li Yang received his B. Sc. degree from Southwest University of Science and Technology in 2018. Now he is a M. Sc. candidate in the School of Computer Science at Hefei University of Technology. His main research interest includes machine vision.

张旭东(通信作者),1989 年于合肥工业大学获得学士学位,1992 年于合肥工业大学获得硕士学位,2005 年于中国科学技术大学获得博士学位,现为合肥工业大学教授,主要研究方向为智能信息处理、机器视觉。
E-mail: xudong@ hfut. edu. cn

Zhang Xudong (Corresponding author) received his B. Sc. degree from Hefei University of Technology in 1989, M. Sc. degree from Hefei University of Technology in 1992 and Ph. D. degree from University of Science and Technology of China in 2005, respectively. Now he is a professor in Hefei University of Technology. His main research interests include intelligent information processing and machine vision.