DOI: 10.13382/j.jemi.B2407259

基于光场解耦的 6D 位姿估计方法*

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

摘 要:光场成像技术能同时捕捉场景中光线的空间和角度信息,被广泛应用到许多计算机视觉任务中。针对基于 RGB 图像位姿估计方法在严重遮挡和截断、光照变化、物体和背景相似等复杂场景下难以有效预估出位姿的问题,提出一种光场解耦特征融合的两阶段 6D 位姿估计方法。该方法采用多种特征提取器解耦光场宏像素图像并将其映射到特征空间,并引入注意力机制融合空间角度及 EPI 信息,为下游位姿估计网络提供有效可靠的关键特征。同时,将反投影应用到关键点预测网络以减少特征传递过程中信息的损耗。在光场位姿估计数据集 LF-6Dpose 上的实验表明,该方法在平均最近点三维距离(ADD-S)和二维投影(2D Project)两个指标下的结果分别为 91.37%和 70.12%,在三维距离指标上相比现有先进方法提升 12.5%,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: 位姿估计;光场解耦;空间角度特征提取;特征融合
中图分类号: TP391.41;TN91 **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 decoupling

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

Abstract: Light field imaging technology can capture both the spatial and angular information of light in a scene simultaneously. It is commonly used in various computer vision tasks. A two-stage 6D pose estimation method leveraging light field decoupled feature fusion is proposed. The aim of this method is to overcome the limitations of RGB image pose estimation methods when predicting pose in complex scenes with severe occlusion and truncation, illumination changes, and similarity between objects and backgrounds. Various feature extractors are utilised to decouple the light field macro-pixel image and map it to the feature space. An attention mechanism is then introduced to fuse the spatial, angular and EPI information to provide effective and reliable features for the downstream pose estimation network. Additionally, the back-projection is applied to the keypoints prediction network to minimise information loss during feature transfer. Experiments on the LF-6Dpose light field pose estimation dataset demonstrate that this method achieves 91.37% and 70.12% for the average closest point 3D distance for symmetric objects (ADD-S) and 2D Project metrics, respectively. This represents a 12.5% improvement compared to existing state-of-the-art methods in 3D distance metrics and more effectively solves the problem of estimating the 6D pose of objects in complex scenes.

Keywords: pose estimation; light field decoupling; spatial and angle feature extraction; feature fusion

0 引 言

物体的 6 自由度(6 degrees of freedom, 6DoF)位姿包括三维旋转和三维平移,物体的位姿估计旨在检测目标物体并估计其相对于世界坐标系的位姿。准确且稳定的位姿估计在机器人导航与操作[1]、自动驾驶[2]、自动化机械臂抓取[3]等领域起着至关重要的作用。在严重遮挡和截断、目标物体与周围环境的视觉特征相似、光照变化等复杂场景下,位姿估计极具挑战。

收稿日期: 2024-02-03 Received Date: 2024-02-03
*基金项目:安徽省科技重大专项(202103a06020010)、安徽省自然科学基金(2208085MF158)项目资助

传统的位姿估计方法[4-5]从已知的物体三维模型生成多个视角的包含纹理和形状信息模板,然后将场景图像特征与模板进行匹配计算位姿。然而传统方法通常手动设计和提取特征,对图像变化的鲁棒性较差,容易受到背景杂波的干扰,在复杂场景下效果受到明显的限制。随着深度学习的快速发展,一些直接从 RGB 图像估计物体位姿的端到端卷积神经网络(convolutional neural network, CNN)被提出。Xiang 等[6]通过训练端到端的神经网络直接预测目标物体的平移量和旋转,并使用迭代最近点方法进行细化。Trabelsi 等[7]提出一种基于 CNN 的多解码器网络用以解耦平移和旋转,并使用多注意机制校正位姿。为了解决由于旋转空间的非线性导致端到端直接位姿估计方法学习难度大、耗时长的问题,Tekin 等[8]提出一个两阶段方法 YOLO-6D,首先检测目标物体 3D 边界框角点的 2D 投影,然后通过 n 点透视(perspective-n-point, PnP)求解位姿。但是当目标物体处于场景边缘被截断时,网络难以有效预测到物体的边界框从而导致位姿求解失效。于是 Peng 等[9]提出一种基于随机采样一致(random sample consensus, RANSAC)投票的关键点定位方法,先训练深度网络回归出指向关键点的逐像素单位向量,然后用属于目标物体像素的向量投票选出关键点求解位姿,这种矢量场的表示在一定程度上可以从可见部分恢复出被遮挡或截断的关键点。由于基于关键点的两阶段方法不能应用到许多可微位姿任务中,Wang 等[10]将直接和间接方法结合,提出了端到端训练的 GDR-Net 网络,使用残差网络提取几何特征构建 2D-3D 密集对应关系,并用 CNN 和全连接层模拟 PnP 估计位姿。Hu 等[11]提出一种名为 PFA 的非迭代细化方法,通过查询一组离散位姿并估计它们相对于目标物体的密集 2D 位移场,从而使初始位姿渲染的图像和真实图像之间密集对应,组合成一组 2D-3D 对应关系校正位姿。

基于 RGB 的深度学习方法显著提高了位姿估计的精度和泛化性。然而单目图像只能从一个固有的视角和尺度获取场景中的空间信息,难以完全表示出目标物体的几何特征,导致在严重遮挡和背景干扰等场景下位姿估计的性能欠佳。为此,最近有工作使用光场(light field, LF)图像获取多视角信息来增强位姿估计的性能。Zhou 等[12]提出一种使用光场感知的两阶段位姿估计器,输入光场子孔径堆栈,经过双流网络估计目标物体的分割掩码和中心点,然后采用深度似然来回归位姿。Huo 等[13]提出一种线性方法估计不同光场点对之间的投影变换并提取相对位姿的方法。李扬等[14]提出基于极平面图像(epipolar plane image, EPI)栈的位姿估计方法 EPI-Pose,用 EPI 卷积算子显式建模光场,获取空间和角度信息之间的关系用于回归位姿。但是仅仅使用子孔径图像堆栈或者 EPI 堆栈作为输入,都只在单方面解耦光场图像,其丰富的角度信息没有被充分利用。

受 Wang 等[15]光场解耦机制的启发,为了充分利用光场多视角图像的角度信息,本文设计了一种光场解耦特征融合的两阶段预测网络以解决复杂场景下的位姿估计问题。其关键思想是用特定的卷积算子解耦宏像素图像(macro-pixel image, MacPI)并使用注意力机制进行特征级别的交互融合,提取空间-角度融合特征以及空间-角度交互特征,为下游关键点预测任务提供可靠有效的图像特征。此外,在关键点预测网络中,将反投影机制应用到降采样的过程,把深层特征反投影到浅层弥补降采样的信息缺失。最后根据 RANSAC 投票选出二维图像上关键点的几何位置,通过 PnP 的变体方法[16]求解位姿。在光场位姿估计数据集 LF-6Dpose[14]上评估所提方法,ADD-S 和 2D Project 指标平均检测精度分别为 91.37%和 70.12%,与最优方法相比,ADD-S 提升了 12.5%。本文的贡献如下:

1)提出一种高性能的光场解耦特征融合模块,该模块解耦光场宏像素图像以简化模型学习,同时提取光场空间-角度融合特征以及空间-角度交互特征从而增强在严重遮挡和截断、光照变化和背景干扰等复杂场景下的位姿估计性能。

2)对光场关键点预测网络进行优化改进,将反投影机制应用到特征降采样过程中,减少降采样中特征信息的丢失,增加特征提取的有效性。

1　基于光场解耦的 6D 位姿估计方法

本文利用光场成像解决遮挡等复杂场景下的位姿估计问题,关键是如何有效获取光场多视角图像的场景信息。为此本文提出了一种基于光场解耦的两阶段位姿估计方法,如图 1 所示。整体网络结构由 3 个模块组成,分别是光场解耦特征融合模块、关键点预测模块和位姿回归模块。为高效提取光场图像中包含的结构特征,设计了空间-角度融合特征提取器(spatial-angular fusion features extractor, SAFFE)和空间角度交互特征提取器(spatial-angular interaction feature extractor, SAIFE)用于提取场景的全局结构和局部细节信息。通过多种算子解耦宏像素图像,在特征空间交互融合以有效挖掘光场的多视角信息。光场解耦特征融合模块的输出前向传播到关键点预测模块,采用带跳跃连接的深层网络对特征进行上下文聚合,将反投影应用到降采样的过程中减少特征传递过程中信息的衰减。最后通过位姿回归模块求解出旋转矩阵 $\boldsymbol{R}$ 和平移量 t。

1.1　光场图像输入

基于目前主流的双平面模型对光场进行参数化表征,即 $L(u,v,h,w) \in R^{U\times V\times H\times W}$,$U$ 和 V 表示光场图像的角

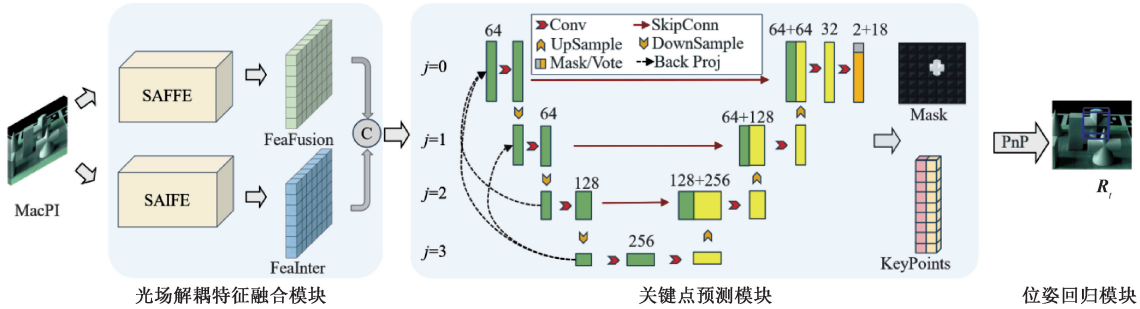
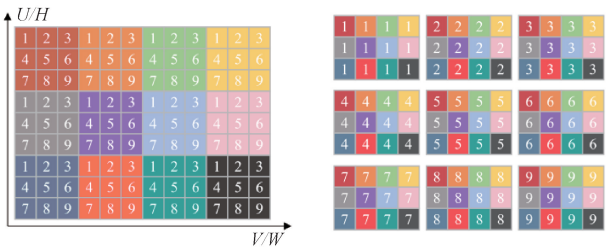


图 1 光场解耦位姿估计网络结构

Fig. 1 Light field decoupled pose estimation network structure

度分辨率，H 和 W 表示空间分辨率。在图 2 光场解码的示意图中，$U=3$，$V=3$；$H=3$，$W=4$。图 2(a)表示宏像素图像，各种颜色方块表示不同的宏像素，图中的数字表示多个不同的视角。每个宏像素是由不同视角下的像素组合而成的，在空间上具有相同的位置，每个宏像素代表了某个特定位置上的所有视角变化。图 2(b)是子孔径图像，是多个特定视角下的图像集，不同颜色表示不同的像素。宏像素图像虽然不太适合人类视觉感知，但其空间和角度特征是均匀分布的，对于模型而言更加适合提取位姿估计所需要的空间角度特征，本文采用宏像素图像作为输入。



(a) 宏像素图像
(a) Macropixel image

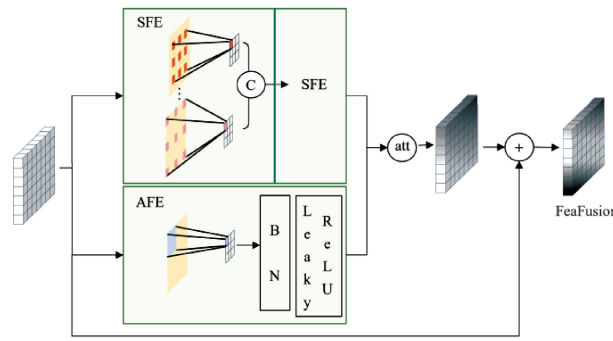
(b) 子孔径图像
(b) Subaperture image

图 2 光场图像解码示意图

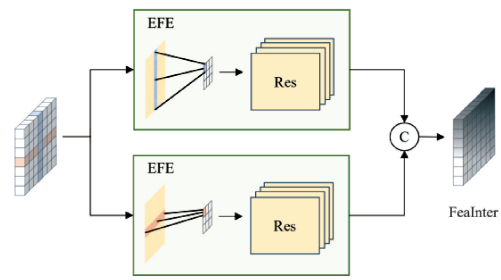
Fig. 2 Schematic diagram of light field image decoding

1.2 光场解耦特征融合模块

如图 3(a)所示，分别用空间特征提取器(spatial feature extractor, SFE)和角度特征提取器(angular feature extractor, AFE)解耦光场宏像素图像，提取空间特征 FeaSpa 和角度特征 FeaAng。然后用注意力机制进行融合，与原始图像浅层特征残差连接后输出空间-角度融合特征 FeaFusion。对整张宏像素图像而言，FeaSpa 表征了整张图片的全局信息，涉及了整个图像的内容与结构，有利于理解场景，估计目标物体的大致方向和位置；FeaAng 代表了场景的局部信息，更加关注图像中较小的区域和多视角信息，提取目标物体的细节特征，能够有效减小场景中光照变化的影响。综合利用空间和角度特征的注意力融合策略不仅能够提高对场景上下文的感知能力，还能增强模型在不同环境条件下的适应性，使其更有效地应对复杂的现实场景。在图 3(b)中，用极平面特征提取器(EPI features extractor, EFF)提取光场垂直和水平方向的 EPI 特征，并通过堆叠的残差块 Res 学习图像特征的不同抽象层次，最后将两个方向的 EPI 特征拼接得到空间-角度交互特征 FeaInter。现实场景中不同深度的 3D 对象有不一样的视差，EPI 图像的极线的分布和变化体现了场景中不同深度区域的位置和关系，反映出光场图像空间和角度之间的交互信息。



(a) 空间-角度融合特征提取器
(a) Space-Angle Fusion Feature Extractor(SAFFE)



(b) 空间-角度交互特征提取器
(b) Space-Angle Interaction Feature Extractor(SAIFE)

图 3 光场空间-角度特征融合、交互模块

Fig. 3 Light field space-angle feature fusion and interaction module

为了扩大 SFE 的感受野，获取目标物体更加丰富的

上下文信息,使用空洞率不一致的卷积提取多个层次的结构特征。针对维度为 $(U \times H, V \times W)$ 的宏像素图像,采用卷积核大小为 3,步长为 1,扩张率为 $(N \times U, N \times V)$,$N \in (1,3,5)$ 的卷积处理,将多层次特征在特征维度拼接输出 $\mathrm{FeaSpa} \in R^{H \times W \times spaChannel}$。在 AFE 中,使用多层卷积核大小为 3,步长和扩张率都为 (U,V) 的卷积层处理宏像素图像,经过批量归一化 BN 和激活层 LeakyReLU 后输出 $\mathrm{FeaAng} \in R^{H \times W \times angChannel}$。AFE 本质上是在宏像素内做卷积,将某个空间位置不同视角的特征进行组合。如果某个视角下目标物体部分关键点被遮挡,但是在另一个视角下相同的关键点位置是可见的,多视角信息有利于处理此类遮挡情况。然后采用注意力融合空间和角度特征,将二者拼接后分别映射到不同的特征空间,再计算两种特征映射之间的相似度得到注意力权重,最后用注意力引导特征融合得到 $\mathrm{FeaFusion} \in R^{H \times W \times FChannel}$。

EPI 特征能够同时建模光场的空间结构和角度变化,解决不同视角图像的视差问题。用 EFE 提取宏像素图像垂直和水平两个方向的 EPI 特征,两个方向的 EFE 共享权重。以水平方向为例,采用两层卷积核大小为 $1 \times V$,步长和扩张率均为 $(1,V)$ 的卷积层将宏像素图像映射到特征空间,然后经过 4 个堆叠的残差块 Res 学习深层次的特征,得到水平 EPI 特征 $\mathrm{FeaEpi-H} \in R^{U \times H \times W \times epiChannel}$,同理垂直 EPI 特征为 $\mathrm{FeaEpi-V} \in R^{H \times V \times W \times epiChannel}$。将水平和垂直的 EPI 特征拼接后 reshape 为 $\mathrm{FeaInter} \in R^{H \times W \times IChannel}$。最后,将 FeaFusion 和 FeaInter 在特征维度上进行拼接,和原始图像浅层特征残差连接,前向传播到关键点预测网络。

1.3　关键点预测网络和位姿回归

关键点预测网络是以 U-net[17] 为基础,如图 1 所示,将光场特征提取模块的输出特征经过上下采样和跳跃连接在深层次的特征空间进行交互。最后一层卷积输出一个 $(H,W,20)$ 的张量,由 2 位掩码和 18 位指向 9 个关键点的单位向量组成,用 RANSAC 投票求解出关键点的二维坐标。传统的 U-net 结构的网络在降采样时会有信息丢失,非相邻层次之间的特征缺乏足够的连接。受到文献[18]启发,将反投影应用到特征传递过程的降采样中,如图 1 关键点预测模块中的虚线箭头所示,在 4 个不同层次的特征之间将深层特征反投影到非相邻的浅层,该过程可以用式(1)表示。

$$f_{j+1} = \sum_i \varphi_{j+1}[f_j - g_j(f_i)],\ j = 0,1,\cdots;i = j+1, j+2,\cdots \tag{1}$$

其中,j 表示在降采样过程中不同层次的特征,f_j 表示在第 j 层的特征;$g_*(\,)$ 表示上采样,$g_j(f_i)$ 的特征图尺寸以及通道数都和 f_j 一致;$\varphi_{j+1}[\,]$ 表示第 j 层的特征向第 $j+1$ 层的映射,实现方式是“Conv2d-BN-Conv2d-BN-LeakyReLU”。通过将深层次的特征向上反投影,加深了非相邻层次之间的特征交互,弥补前向过程中的信息衰减。图 1 中关键点预测模块中的数字表示特征通道数,当特征图的尺寸缩小到原来的 1/16 时停止降采样。转而对特征进行上采样,并将降采样和上采样过程中同一层次的特征经过跳跃连接在特征维度上拼接后卷积输出,最后用 RANSAC 求解出在目标物体表面的 9 个关键点和物体掩码。

位姿回归模块采用 PVNet[9] 的方式,考虑了不同关键点具有不一样的置信度,通过最小化重投影误差结合透视投影法 PnP 求解三维旋转矩阵 $\boldsymbol{R}$ 和平移量 t 作为目标物体相对于相机坐标系的位姿。

2　实　验

本文整体流程遵循两阶段的方法回归位姿。首先通过神经网络在光场图像上找出目标物体上的 9 个关键点,然后用考虑置信度的透视投影法求解位姿。

2.1　实验设置

数据集。本文方法在 LF-6Dpose 数据集上训练模型并验证方法的可行性。数据集中的物体采用的是无纹理的具有朗伯特性的石膏体,该类物体相对于有色彩和纹理的物体提取特征和应对光照变化处理的难度更大。数据集中包含的子孔径图像空间分辨率是 416×608,空间分辨率是 9×9。为方便训练模型,对数据集进行预处理,将包含中心视角的多张子孔径图像转化成不同角度分辨率的宏像素图像,本文采用的宏像素图像的空间分辨率是 2 080×3 040(角度分辨率为 5×5 的光场图像)。防止模型过拟合,对宏像素图像采用了裁剪、旋转(−30° ~ 30°)等数据增强的手段。

评价指标。使用位姿估计领域两个常用且具有代表性的指标进行评估:2D 投影[19] 和模型点的平均最近点 3D 距离[20]。给定神经网络的输出位姿 $[\bar{R} \mid \bar{t}]$ 和真值 $[R \mid t]$,2D 投影指标计算 $[\bar{R} \mid \bar{t}]$ 和 $[R \mid t]$ 的 3D 模型投影到 2D 平面之间的平均距离,距离小于 5 pixels 判定位姿估计正确。平均最近点三维距离(average closest point 3D distance for symmetric objects, ADD-S),考虑旋转对称的物体可以有多个估计的位姿。给定估计的位姿 $[\bar{R} \mid \bar{t}]$ 和真值 $[R \mid t]$,ADD-S 计算从 $[\bar{R} \mid \bar{t}]$ 变换的每个 3D 模型点到由 $[R \mid t]$ 变换的最近点的平均距离,当 ADD-S 小于模型直径的 20%时表示位姿估计正确。

训练细节。从特征提取到关键点的预测采用 smoothl_1 损失[21] 进行训练,掩码预测使用交叉熵损失,训练过程包含由 ADAM 优化器进行优化训练 200 个 epoch,其中 $\beta_1 = 0.9, \beta_2 = 0.999$,batchsize 为 4。初始学习率设置为 1×10^{-3},$\gamma = 0.5$,经过 20 个 epoch 调节学习率。卷积

层采用 He 初始化,批归一化层采用标准初始化。所有实验在 NVIDIA RTX A6000 GPU 上进行训练,训练时长为 24 h 左右。

2.2 定量实验

将本文方法和 YOLO6D[8]、PVNet[9]、GDR-Net[10]、PFA[11]、EPI-Pose[14]在 LF-6Dpose 上进行比较,基于 RGB 图像的方法输入光场的中心子孔径图像。为公平比较,在数据集 LF-6Dpose 上对所有方法重新训练和评估。表 1 和 2 分别表示了上述方法在 2D 投影和 ADD-S 指标上的比较结果。在本节的所有表格中,加粗表示最优,下划线表示次优。

从表 1 的数据可以看出本文方法在 2D 投影上达到了次优或最优的水平。从表 2 的数据显示在 ADD-S 指标上每个类都达到最优,相比之前最优的 EPI-Pose 平均提升了 12%。2D 投影指标缺少一个维度的信息,指标在部分 3D 场景中不能完全反映位姿估计精度,三维指标 ADD-S 更能评价位姿估计的准确性。实验表明,预测角点的方法 YOLO6D 不太适用于遮挡和截断场景较多的数据集,所以在指标上分数较低;PVNet 使用的投票策略在一定程度上解决了部分截断问题,相比角点预测位姿估计的准确率更高;GDR-Net 用中间几何特征引导位姿估计从而使准确率有较大提升;PFA 采用的细化方法有效提高了 RGB 方法性能,但单图像方法受限于固有的视角,缺乏场景的角度信息;EPI-Pose 用光场 EPI 图像栈估计位姿引入了光场的角度特征,相比较单图像的方法有较大的提升,但是 EPI-Pose 没有充分利用光场图像的特性,对角度和空间特征的提取不充分;本文方法采用的光场解耦特征融合机制提取了 FeaFusion 和 FeaInter,在面对遮挡、光强变化等复杂场景性能更高,更加适合处理现实复杂场景的位姿估计,在 ADD-S 上能够很好体现本文模型的优越性。

表 1 不同方法的 2D Project 定量比较

Table 1 Quantitative comparison of 2D Project using different methods

方法	2D Project/%					
	Average	Cone	Crosscone	Crosscuboid	Cube	Cuboid
YOLO6D	81.82	81.13	83.44	76.27	87.35	80.89
PVNet	87.06	86.89	87.09	85.62	88.91	86.77
GDR-Net	90.37	89.54	90.24	89.02	92.21	90.83
PFA	90.64	**92.18**	89.65	88.96	90.88	91.54
EPI-Pose	**91.97**	91.73	90.38	90.51	**94.86**	**92.38**
本文方法	91.37	90.31	**92.47**	**90.51**	93.14	90.40

表 2 不同方法的 ADD-S 定量比较

Table 2 Quantitative comparison of ADD-S using different methods

方法	ADD-S/%					
	Average	Cone	Crosscone	Crosscuboid	Cube	Cuboid
YOLO6D	40.50	26.02	32.43	30.41	60.38	53.24
PVNet	47.27	39.35	36.17	32.56	65.51	62.74
GDR-Net	53.61	42.70	47.55	43.63	69.55	64.61
PFA	54.38	43.62	48.21	45.92	69.63	64.52
EPI-Pose	57.61	42.56	53.97	48.14	74.51	68.87
本文方法	**70.12**	**56.75**	**71.55**	**65.42**	**81.70**	**75.17**

2.3 消融实验

本节将验证所提模块的有效性,除非另有说明,所有实验输入尺寸为 2 080×3 040(空间分辨率为 5×5),采用相同的数据增强。在本小节的所有表格中,加粗表示最优,下划线表示次优。为验证光场空间-角度特征混合特征提取模块(SAFFE)、空间-角度交互特征提取模块(SAIFE)以及反投影模块(BackProj)的有效性,设计了下列的实验。实验 1:只含有 SAIF 和 BackProj;实验 2:只含有 SAFF 和 BackProj;实验 3:只含有 SAFF 和 SAIF;实验 4:所有模块均包含。实验的 2D 投影指标和 ADD-S 指标分别如表 3、4 所示,表中的向下箭头表示相比较实验 4 指标的平均下降值。综合两个表格分析得出,缺少任何一个模块都会对 6D 位姿估计精度产生较大的影响。特别是缺少空间-角度交互模块时,2D-project 和 ADD-S 分别下降了 8.17%和 16.68%,这说明空间角度交互特征同时建模场景的结构和角度信息,对于处理复杂场景下目标物体的位姿估计更加有效。

此外,对输入不同角度分辨率的宏像素图像进行了比较,结果如表 5 所示,向下箭头表示和 5×5 比较指标的平均降低值。3×3 的角度分辨率相比 5×5 的在 2D-

project 上并没有特别明显的变化,但是在三维指标 ADD-S 上降低了 8.50%。6D 位姿估计的准确度和角度分辨率时呈正相关,这也说明了光场解耦特征融合能够有效提取光场图像的角度特征,并能有效地提升 6D 位姿估计精度。

表 3　不同模块的 2D Project 定量比较

Table 3　Quantitative comparsion of 2D Project of different modules

实验	2D Project/%					
	Average	Cone	Crosscone	Crosscuboid	Cube	Cuboid
SAIF+BackProj	89.52(↓1.58)	89.66	90.68	87.80	90.07	89.39
SAFF+BackProj	83.2(↓8.17)	78.01	88.47	84.92	83.53	84.17
SAFF+SAIF	90.02(↓1.35)	90.04	91.63	89.16	90.20	89.07
SAFF+SAIF+BackProj	91.37	90.31	92.47	90.51	93.14	90.40

表 4　不同模块的 ADD-S 定量比较

Table 4　Quantitative comparison of ADD-S of different modules

实验	ADD-S/%					
	Average	Cone	Crosscone	Crosscuboid	Cube	Cuboid
SAIF+BackProj	89.52(↓1.58)	89.66	90.68	87.80	90.07	89.39
SAFF+BackProj	83.2(↓8.17)	78.01	88.47	84.92	83.53	84.17
SAFF+SAIF	90.02(↓1.35)	90.04	91.63	89.16	90.20	89.07
SAFF+SAIF+BackProj	91.37	90.31	92.47	90.51	93.14	90.40

表 5　不同角度分辨率位姿回归的定量比较

Table 5　Quantitative comparsion of pose regression with different angle resolutions

尺寸	指标	Average	Cone	Crosscone	Crosscuboid	Cube	Cuboid
3×3	2D Project/%	91.05(↓0.32)	90.31	90.58	88.81	90.52	94.04
	ADD-S/%	61.62(↓8.50)	46.02	59.41	58.64	74.18	69.87
5×5	2D Project/%	91.37	90.31	92.47	90.51	93.14	90.40
	ADD-S/%	70.12	56.75	71.55	65.42	81.70	75.17

2.4　定性实验

将本文方法与 PVNet[9] 和 EPI-Pose[14] 进行定性分析,图 4~6 为不同场景下的实验结果。在图 4~6 中,(a)表示使用 RGB 图像的方法 PVNet;(b)是使用 EPI 图像栈的方法 EPI-Pose;(c)是本文方法。使用相机的内参矩阵和网络估计得到的物体位姿,通过 3D 边界框(3D bounding box)在 2D 图像上可视化位姿。定性试验结果图中虚线 3D 边界框表示数据集中标注的真值,实线 3D 边界框表示不同方法的预测值。

从图 4 中可以观察到,在第 1 列、第 3 列场景中,目标物体的截断比较明显,PVNet 和 EPI-Pose 预测框和真实框之间的差距比较明显;在第 4 列的场景中,目标物体是交叉锥体,被前面的长方体严重遮挡,PVNet 预测的旋转和平移量都存在一定的差距,EPI-Pose 的平移量由较大偏差。本文方法有效减小了严重遮挡和截断场景下预测和真值之间的误差,能提高位姿估计的性能。

图 5 是目标物体和背景板相似的几个场景下的位姿估计框,从图中可以看到,单图像的方法在这种场景下的旋转预测值和真值之间存在较大的偏差;使用 EPI 图像栈的方法估计的旋转大致和真值相符合,但是平移量在部分场景中有较大的偏移;本文方法在旋转和平移量上都大致符合位姿的真值。这是因为光场解耦融合机制能很好的利用光场的角度信息,从不同的视角建模场景结构。

图 6 是光照变化场景的位姿估计的定性分析,在光照强度过大或者过小的场景中,PVNet 和 EPI-Pose 预计的位姿框都会有比较都会有相对较大的偏差,在成像距离比较远的情况下更加明显。本文方法通过深度挖掘光场图像中记录的光线的强度和角度信息,能够有效缩小预测框和真实框之间的差距。

3　结　论

基于 RGB 的位姿估计方法很难有效处理存在遮挡严重、光照变化、背景干扰等复杂场景的位姿估计。为此本文提出一种光场解耦特征融合的两阶段方法估计位姿。使用不同的特征提取器解耦光场图像,充分利用光场图像中包含的空间和多视角信息处理遮挡光照等问

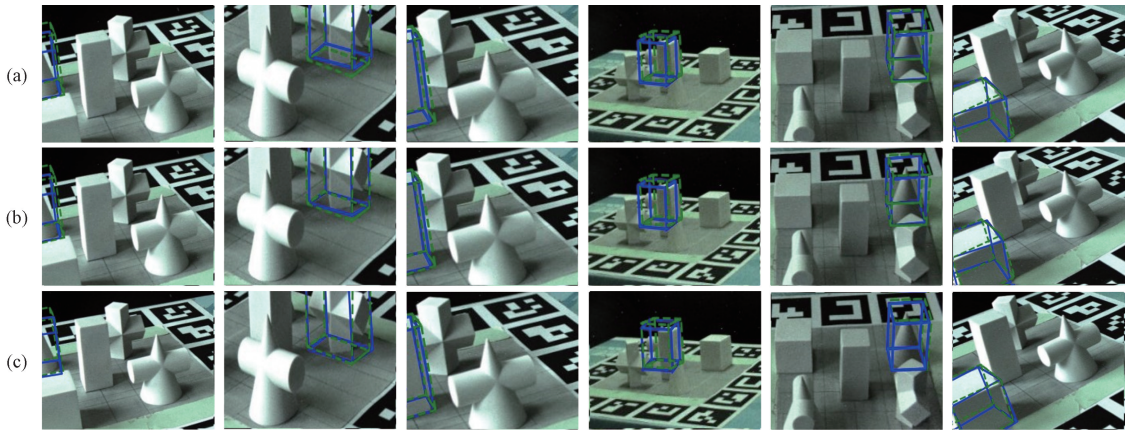


图 4 遮挡或截断场景定性实验

Fig. 4 Qualitative experiments with masked or truncated scenes

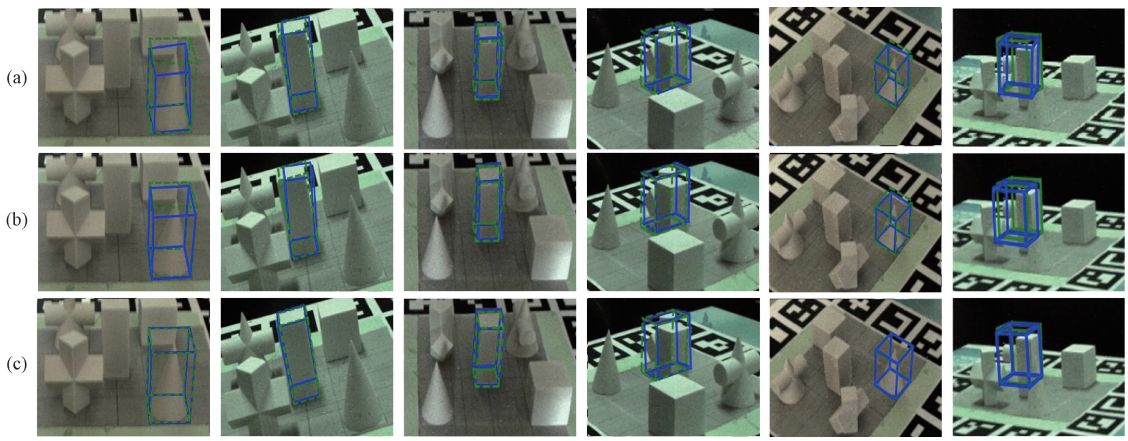


图 5 物体与背景相似场景定性实验

Fig. 5 Qualitative experiments on object and background similarity scenes

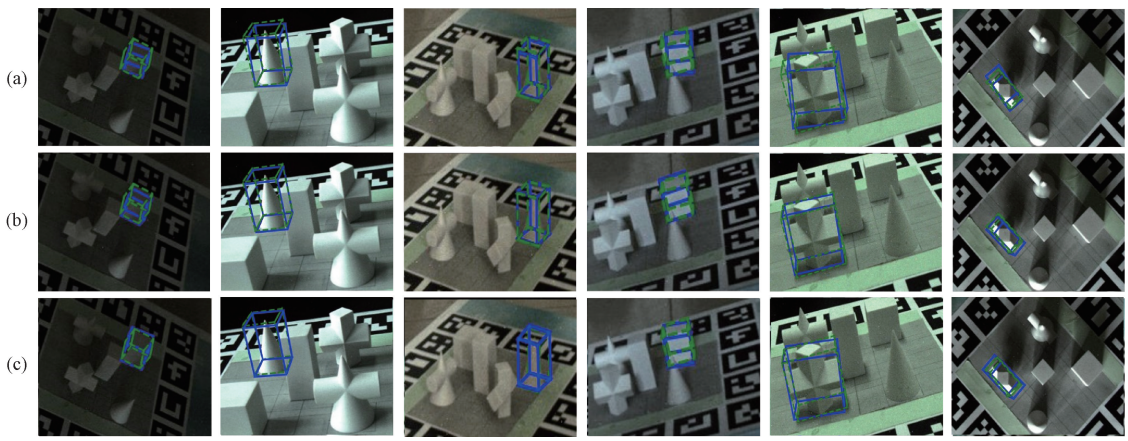


图 6 光照变化场景定性实验

Fig. 6 Qualitative experiment of illumination variation scenes

题,并使用注意力机制进行特征融合,为下游任务提供有效的信息。在关键点预测网络中将反投影应用到特征前向传播的过程以减少信息损耗。实验表明本文方法能够有较地解决复场景下的目标 6D 位姿估计问题。下一步

工作将致力于研究更加高效的特征提取网络,在保证精度的前提下更快速地处理光场图像用于估计位姿。

参考文献

[1] 杨雪梅, 李帅永. 移动机器人视觉 SLAM 回环检测原理, 现状及趋势[J]. 电子测量与仪器学报, 2022, 36(8): 1-14.
YANG X M, LI SH Y. Principle, current situation and trend of visual SLAM loop closure detection for mobile robot [J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(8): 1-14.

[2] HU J, XU Z. CMTR6D: Cross-modal transformer for 6D pose estimation [C]. Proceedings of the 2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE), 2023: 577-580.

[3] 林麒光, 刘宇, 李杰, 等. 基于轨迹测量与人机映射的六自由度机械臂运动追踪模型[J]. 电子测量与仪器学报, 2023, 37(3): 102-110.
LIN Q G, LIU Y, LI J, et al. Motion tracking model of 6-DOF manipulator based on trajectory measurement and human-machine mapping [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(3): 102-110.

[4] 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]. Proceedings of the Computer Vision - ACCV 2012: 11th Asian Conference on Computer Vision, 2013: 548-562.

[5] PAVLAKOS G, ZHOU X, CHAN A, et al. 6-dof object pose from semantic keypoints[C]. Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017: 2011-2018.

[6] 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:171100199, 2017.

[7] TRABELSI A, CHAABANE M, BLANCHARD N, et al. A pose proposal and refinement network for better 6D object pose estimation[C]. Proceedings of the the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 2382-2391.

[8] TEKIN B, SINHA S N, FUA P. Real-time seamless single shot 6D object pose prediction[C]. Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 292-301.

[9] PENG S, LIU Y, HUANG Q, et al. Pvnet: Pixel-wise voting network for 6D of pose estimation [C]. Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4561-4570.

[10] WANG G, MANHARDT F, TOMBARI F, et al. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation[C]. Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 16611-16621.

[11] HU Y, FUA P, SALZMANN M. Perspective flow aggregation for data-limited 6d object pose estimation[C]. Proceedings of the European Conference on Computer Vision, 2022: 89-106.

[12] ZHOU Z, CHEN X, JENKINS O C J I R, et al. Lit: Light-field inference of transparency for refractive object localization [J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4548-4555.

[13] HUO X, JIN D, ZHANG S, et al. Light field rectification based on relative pose estimation [J]. IEEE Transactions on Instrumentation, 2024, 73: 1-18.

[14] 李扬, 张旭东, 孙锐, 等. 基于光场 EPI 图像栈的 6D 位姿估计方法[J]. 电子测量与仪器学报, 2023, 37(4): 122-130.
LI Y, ZHNAG X D, SUN R, et al. 6D pose estimation method based on lighted field EPI image stack [J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(4): 122-130.

[15] WANG Y, WANG L, WU G, et al. Disentangling light fields for super-resolution and disparity estimation [J]. IEEE Transactions on Pattern Analysis, 2022, 45(1): 425-443.

[16] FERRAZ L, BINEFA X, MORENO-NOGUER F. Very fast solution to the PnP problem with algebraic outlier rejection[C]. Proceedings of the the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 501-508.

[17] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]. Proceedings of the Medical Image Computing and Computer-Assisted Intervention - MICCAI, 2015: 234-241.

[18] DONG H, PAN J, XIANG L, et al. Multi-scale boosted dehazing network with dense feature fusion [C]. Proceedings of the the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2157-2167.

[19] BRACHMANN E, MICHEL F, KRULL A, et al. Uncertainty-driven 6D pose estimation of objects and scenes from a single rgb image[C]. Proceedings of the the IEEE Conference on Computer Vision and Pattern

Recognition, 2016: 3364-3372.
[20] HODAN T, MATAS J, OBDRŽáLEK Š. On evaluation of 6D object pose estimation [C]. Proceedings of the Computer Vision-ECCV 2016 Workshops, 2016: 606-619.
[21] GIRSHICK R. Fast R-CNN[C]. Proceedings of the the IEEE International Conference on Computer Vision, 2015: 1440-1448.

作者简介

丁潇,2022 年于湘潭大学获得学士学位,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: 2022111063@ mail. hfut. edu. cn

Ding Xiao received his B. Sc. degree from Xiangtan University in 2022. Now he is a M. Sc. candidate in the School of Computer and Information of 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 the B. Sc. degree from Hefei University of Technology in 1989, the M. Sc. degree from Hefei University of Technology in 1992, and the Ph. D. degree from University of Science and Technology of China in 2005. Now he is a professor at Hefei University of Technology. His main research interests include intelligent information processing and machine vision.