DOI: 10.13382/j.jemi.B2306424

基于混合通道注意力的类别级物体六自由度位姿估计*

刘崇沛 孙 炜 刘 剑 杨 慧 张 星 范诗萌
(湖南大学电气与信息工程学院 长沙 410082)

摘 要:针对有光照变化、距离变化、背景杂乱、遮挡等干扰的场景下物体六自由度位姿估计精度低的问题,提出了一种结合多尺度特征融合和注意力机制的混合通道注意力模块(mixed channel attention,MCA)。在 MCA 基础上进一步构建了类别级物体六自由度位姿估计方法(MCA6D),其关键步骤包括物体的实例分割,特征提取与基于 MCA 的特征优化,基于先验形状的物体模型重建,及基于点云配准的位姿估计。本文方法在公共数据集 CAMERA 和 REAL 分别取得 86.3%(5°2 cm)、73.4%(5°5 cm)和 39.2%(5°2 cm)、43.3%(5°5 cm)的均值平均精度,领先于 NOCS、SPD、SGPA 等主流方法;同时实物实验表明本文方法在存在光照变化、距离变化、背景杂乱、遮挡等干扰的场景可准确估计物体六自由度位姿。

关键词: 物体六自由度位姿估计;类别级;注意力机制;通道注意力
中图分类号: TP391.4;TN0 **文献标识码**: A **国家标准学科分类代码**: 520.20

Category-level 6D object pose estimation based on mixed channel attention

Liu Chongpei Sun Wei Liu Jian Yang Hui Zhang Xing Fan Shimeng
(College of Electrical and Information Engineering, Hunan University, Changsha 410082, China)

Abstract: Aiming at the low accuracy of object six-degree-of-freedom (6D) pose estimation in scenes with interferences such as illumination changes, distance changes, background clutter, and occlusions, a mixed channel attention module (MCA) is proposed, which combines multi-scale feature fusion and attention mechanisms. Based on MCA, a category-level object 6D pose estimation method (MCA6D) is further constructed. The key steps include object instance segmentation, feature extraction and optimization based on MCA, object model reconstruction based on prior shape, and pose estimation based on point cloud registration. Relevant experiments show that our method achieves 86.3% (5°2 cm), 73.4% (5°5 cm) and 39.2% (5°2 cm), 43.3% (5°5 cm) mean average precision in the public datasets CAMERA and REAL, respectively, which is ahead of mainstream methods such as NOCS, SPD, and SGPA. At the same time, the practical experiment shows that the proposed method can accurately estimate the 6D pose of the object in scenes with interference, such as illumination changes, distance changes, background clutter, and occlusions.

Keywords: 6D object pose estimation; category-level; attention mechanism; channel attention

0 引 言

物体六自由度位姿估计可同时获得物体的类别和六自由度位姿,即在三维空间中物体相对于相机的 3D 平移和 3D 旋转关系。物体六自由度位姿估计应用广泛,包括增强现实[1]、机器人操作[2-4]、无人驾驶[5];在增强现实领域,可以利用物体位姿在物体上叠加虚拟元素,随着物体的移动而保持和物体相对位姿不变;在机器人领域,随着 SLAM 等技术的成熟,机器人已经能够在空间中进行很好的定位,同时需要物体六自由度位姿估计技术定位物体,帮助机器人与物体的交互;在无人驾驶领域,利用物体位姿估计技术可以感知其他交通参与者与障碍物,提供决策所需信息。位姿估计的结果会影响后续的操作,

收稿日期: 2023-04-11 Received Date: 2023-04-11
*基金项目:国家自然科学基金(U22A2059)、深圳科技计划项目(2021Szvup035)、湖南大学汽车车身先进设计制造国家重点实验室自主研究项目、电子制造业智能机器人技术湖南省重点实验室开放课题项目资助

低精度的估计结果会导致后期操作与规划任务的失败。一方面六自由度位姿估计任务复杂,需要识别物体的类别、确定物体在图像中的区域、估计物体的六自由度位姿;另一方面,为了该技术能真正用于实际任务,需要应对光照变化、距离变化、背景杂乱、遮挡等干扰;现有方法还难以满足实际应用的要求,仍需要持续不断的研究。

现有的方法根据泛化性可分为实例级和类别级物体六自由度位姿估计方法。实例级方法[6-17]旨在估计已经在训练集中出现或 3D 模型已给定的物体的位姿。它们可大致分为 3 类基于对应、基于模板、直接回归的方法。基于对应的方法又可分为 2D-3D 对应[13-16]与 3D-3D 对应[11-12]。2D-3D 对应方法首先在物体的 3D 模型上定义好关键点,然后在 RGB 图像中估计预定义关键点,最后通过 PnP(perspective-n-point)算法求解物体的六自由度位姿。3D-3D 对应方法又可分为局部方法和全局方法,局部方法与 2D-3D 对应方法类似,在 3D 点云中估计预定义关键点,然后通过最小二乘法求解物体六自由度位姿;全局方法通过观测点云与物体 3D 模型的全局对应获取物体六自由度位姿。总的来说,基于对应的方法依赖于丰富的纹理特征或突出的形状特征,因此不适用于形状特征不明显或弱纹理的物体。基于模板的方法[9-10]首先离线地从不同角度投影物体 3D 模型,生成标记有真实六自由度位姿的模板,然后在线搜索与当前图像最相似的模板,以其位姿作为输出。这些方法可以应对无纹理物体,但难以应对遮挡且计算成本较高。基于回归的方法[6-8]利用神经网络的学习能力,从大量数据中学习直接回归六自由度位姿。实例级方法具有较高的精度,但只能用于已知物体,泛化性较差。

类别级物体六自由度位姿估计可以估计已知类别的新物体的位姿,泛化性得到极大的增强。NOCS[18]提出了一种可用于建立同类别物体间 3D-3D 对应的归一化物体坐标空间,然后利用 Umeyama 算法进行点云配准解出六自由度位姿。一些方法[19-21]尝试直接回归六自由度位姿。FS-Net[19]提出了一种新的三维图卷积网络,用于增强六自由度位姿估计的形状特征提取。CenterSnap[20]提出了一种单阶段方法来估计六自由度位姿。DualPoseNet[21]引入了一种新的两分支位姿估计网络,改进位姿一致性学习,提高了位姿估计的精度。也存在一些间接的方法[22-26]。iCaps[23]开发了一个自动编码器网络来编码同类别物体的位姿,并在粒子滤波框架中使用该网络来估计和跟踪物体的六自由度位姿。SPD[26]引入了类别级先验形状,通过学习同类别物体的几何特征获得,并提出了一个中间步骤,通过调整先验形状来重建物体实例的形状,最后通过点云配准获得六自由度位姿。CR-Net[22]提出了一种循环重建网络,以迭代的方式对 NOCS 形状的重建结果进行细化。由于类别级先验形状是固定的,它不能很好地适应具体物体。SGPA[25]利用先验形状与观测物体实例的结构相似性来构造动态先验形状,改善物体重建结果。SAR-Net[24]探索每个物体实例的形状与其对应的类别级模板形状的对齐,以及每个物体类别的对称对应。现有的方法大多数关注于在空间的维度提升方法的性能,如发掘多尺度信息、全局局部信息,且存在有光照变化、距离变化、背景变化、遮挡等干扰时会出现漏检或位姿估计精度低的问题。

为了提高类别级物体六自由度位姿估计方法的精度,增强对干扰的鲁棒性,本文提出混合通道注意力,一方面引入通道注意力,建模通道间的相互关系,为重要的通道赋予更大的权重,从特征的通道维度出发提高网络性能,另一方面考虑到全局特征与局部特征所携带的信息不同,提出先分别施加通道注意力后级联的特征融合结构,称为混合通道注意力,以优化类别级物体六自由度位姿估计方法,提高鲁棒性与精度。

1　MCA6D

给定包含待检测物体的 RGB-D 图像 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 4}$、与待检测物体同类别的先验形状 $\boldsymbol{P}_P \in \mathbb{R}^{N_P\times 3}$,其中 H 和 W 分别为 RGB-D 图像的高和宽,通道数包含 RGB 与深度为 4, N_P 为点云数量,本文方法以它们为输入,实现类别级物体六自由度位姿估计。

1.1　方法概述

本文提出的 MCA6D 框架结构如图 1 所示,可大致分为 4 个步骤。首先输入 RGB-D 图像 $\boldsymbol{I}$,通过实例分割方法得到目标物体的点云 $\boldsymbol{P}_I \in \mathbb{R}^{N_I\times 3}$、图像块 $\boldsymbol{I}_I \in \mathbb{R}^{h\times w\times 3}$、同类别的先验形状 $\boldsymbol{P}_P$。然后,通过特征提取网络获得先验几何特征 $\boldsymbol{F}_{PG} \in \mathbb{R}^{N_P\times d}$、实例颜色特征 $\boldsymbol{F}_{IC} \in \mathbb{R}^{N_I\times d}$、实例几何特征 $\boldsymbol{F}_{IG} \in \mathbb{R}^{N_I\times d}$,通过基于结构引导的先验调整模块获得先验颜色特征 $\boldsymbol{F}_{PC} \in \mathbb{R}^{N_P\times d}$;接着, $\boldsymbol{F}_{PG}$ 与 $\boldsymbol{F}_{PC}$ 融合得到先验特征 $\boldsymbol{F}_P \in \mathbb{R}^{N_P\times 2d}$, $\boldsymbol{F}_{IG}$ 与 $\boldsymbol{F}_{IC}$ 融合得到实例特征 $\boldsymbol{F}_I \in \mathbb{R}^{N_I\times 2d}$,先验特征和实例特征都被输入 MCA 模块。本文提出的 MCA 模块,采用了通道注意力机制,建模通道的重要性,从通道维度优化网络结构,同时为充分利用局部与全局特征,且考虑到两者各自的优势,分别为其设置通道注意力。最后,通过先验形状变形重建物体模型,通过对应网络获取重建模型与物体点云的对应,最终通过点云配准方法得到目标物体的六自由度位姿与尺寸。为了获得先验形状,本文参考 SPD[26]中描述的方法,该方法训练一个编码器-解码器网络,编码器提取输入形状的嵌入特征,解码器通过解码特征恢复形状。通过解码同类别已知形状的平均嵌入特征,得到该类别的

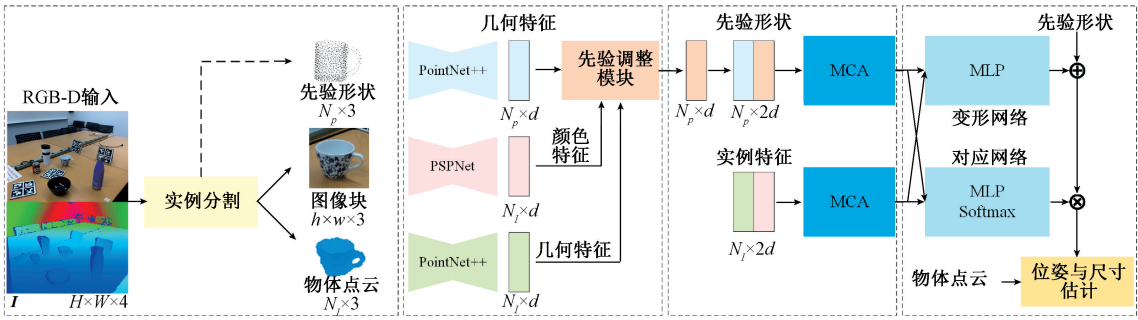

图 1 MCA6D 模型结构

Fig. 1 MCA6D model structure

先验形状。

1.2 实例分割

本文使用现成的实例分割方法 Mask R-CNN[27] 从 RGB 图像中生成目标物体的掩模,利用掩模对深度图进行处理得到物体点云 $\boldsymbol{P}_I$,当然,实例分割也可以通过其他现有方法[28]实现。

1.3 实例与先验特征提取

为了获得高质量的 $\boldsymbol{F}_{IC}$ 与 $\boldsymbol{F}_{IG}$,$\boldsymbol{I}_I$ 与 $\boldsymbol{P}_I$ 分别被输入到 PSPnet[29] 与 PointNet++[30]。PSPNet 通过金字塔池模块与金字塔场景解析网络来进行不同区域的上下文聚合以利用全局上下文信息。PSPNet 中的全局先验表示有利于在场景解析任务中产生高质量结果,非常适合像素级预测。Pointnet 是点云深度学习的先驱,PointNet++在 Pointnet 的基础上引入一个分层神经网络,对分块的点云都应用 Pointnet,通过下采样也可以学习不同上下文尺度的局部特征。这样解决了 Pointnet 无法提取度量空间内点云的局部特征的问题,从而提升了对细粒度模式的识别能力与对复杂场景的泛化能力。通过 PSPnet 和 Pointnet++获得 $\boldsymbol{F}_{IC}$ 与 $\boldsymbol{F}_{IG}$。类似地,将 $\boldsymbol{P}_P$ 输入到 PointNet++获得 $\boldsymbol{F}_{PG}$。为了完善先验信息,这里采用补充先验颜色信息的方式,参考 SGPA[25] 将 $\boldsymbol{P}_P$ 与 $\boldsymbol{P}_I$ 输入一个基于结构引导的先验调整模块。该模块利用一个基于 transformer 的结构关联 $\boldsymbol{F}_{IG}$ 与 $\boldsymbol{F}_{PG}$,建立 $\boldsymbol{P}_I$ 与 $\boldsymbol{P}_P$ 的整体结构相似度模型,通过该模型,$\boldsymbol{F}_{IC}$ 沿着结构相似度推广获得 $\boldsymbol{F}_{PC}$。为减少计算量,没有在所有位置点上密集地关联 $\boldsymbol{F}_{IC}$ 与 $\boldsymbol{F}_{PG}$,而是利用一个辅助网络,从 $\boldsymbol{P}_I$ 的 N_I 点中提取 n 个关键点。然后通过结构相似性获得 $\boldsymbol{F}_{PC}$:

$$\boldsymbol{F}_{PC}=F_A(F_S(\boldsymbol{F}_{IC},\boldsymbol{F}_{IG})\cdot\boldsymbol{F}_{IG},\boldsymbol{F}_{PG})\cdot\boldsymbol{F}_{IC}=F_A(\boldsymbol{E}\cdot\boldsymbol{F}_{IG},\boldsymbol{F}_{PG})\cdot\boldsymbol{F}_{IC} \tag{1}$$

其中,F_S 表示辅助网络,F_A 表示基于 transformer 的结构,$\boldsymbol{P}_K=\boldsymbol{E}\boldsymbol{P}_I$ 表示提取的关键点。实例颜色特征 $\boldsymbol{F}_{IC}$ 与实例几何特征 $\boldsymbol{F}_{IG}$ 通过逐点地级联获得实例特征 $\boldsymbol{F}_I$,类似地获得先验特征 $\boldsymbol{F}_P$。

1.4 MCA

1)注意力机制

注意力机制是人工神经网络领域模仿人类认知注意力而产生的技术。这一技术可以增强输入数据的某些部分,而削弱其他部分,其目的是使网络将更多的注意力放到小但重要的部分数据上。通过上下文信息学习决策部分数据比其他部分数据更重要,这一决策能力可通过优化算法进行训练。注意力机制可以追溯至 1990 年代的乘法模块、超网络等,其灵活性来源于其“软权重”的特性,即它与运行时保持固定的标准权重不同,它可以根据输入数据实时变化。注意力的用途包括神经图灵机中的记忆功能、可微分神经计算机中的推理任务、翻译器中的语言处理和 LSTMs 以及感知器中的多感官数据处理(声音、图像、视频和文本等)。基于注意力机制的方法成为研究热点,未来也将会有持续不断的相关研究。

2)通道注意力

通道注意模块是卷积神经网络中从通道维度应用注意力机制的模块,其通过建模特征通道间的关系生成通道注意力图。由于特征图的每个通道都可被视为特征检测器,通道注意力集中于给定输入图像中什么是重要的。为了高效地计算通道注意力,这里首先压缩输入特征图的空间维度,通过平均池化聚合特征图的空间信息,生成一个空间上下文描述符 $\boldsymbol{F}_{avg}^{C}$,称为平均池化特征。接着,$\boldsymbol{F}_{avg}^{C}$ 输入到一个简单的网络生成通道注意力图 $\boldsymbol{A}\in\mathbb{R}^{C\times1\times1}$,其中 C 是通道数。这里的简单网络是包含一个隐藏层的多层感知机(multilayer perceptrons, MLP)。为了减少参数量,隐藏层的神经元数为 C/r 其中 r 是衰减系数。通道注意力的计算过程,可表示如下:

$$\boldsymbol{A}=\sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F})))=\sigma(\mathrm{MLP}(\boldsymbol{F}_{avg}^{C})) \tag{2}$$

其中,σ 表示 sigmoid 激活函数。

3)混合通道注意力

$\boldsymbol{F}_I$ 与 $\boldsymbol{F}_P$ 表征的是每个点的几何与颜色信息,属于局部特征,在图像中较丰富,反映细节,不易受遮挡的等影

响。全局特征是对图像信息一个整体性的表示,从大尺度表征图像内容。局部和全局特征各有优势。为了高效的利用局部、全局特征,结合两者优势,同时引入通道注意力机制,从通道维度建模特征信息的重要性,本文提出 MCA 模块。

具体结构如图 2 所示,从图中可以看出 MCA 的主体结构包含两条支路:局部特征通道注意力支路、全局特征通道注意力支路。两条支路的输入均是局部特征 $\boldsymbol{F} \in \mathbb{R}^{N\times 2d}$。看向上方的局部支路,首先,$\boldsymbol{F}$ 通过全局平均池化,得到一个全局聚合特征 $\boldsymbol{F}_A \in \mathbb{R}^{1\times 2d}$,该操作从空间维度进行了压缩,将每个特征通道变成一个实数,该实数具有全局感受野;接着,$\boldsymbol{F}_G$ 被输入进一个包含 3 个全连接层神经网络,生成逐通道的权重,这里的神经网络建模了通道间的相关性,权重可以视为对每个特征通道的注意力的度量;最后,$\boldsymbol{F}$ 的每个通道与相应的权重相乘,生成加权局部特征 $\boldsymbol{F}_{LW} \in \mathbb{R}^{N\times 2d}$。再看向下方的全局支路,整体结构与局部支路类似,不同的是这里首先将 $\boldsymbol{F}$ 输入一个多层感知机扩充特征的维度得到 $\boldsymbol{F}_{ex} \in \mathbb{R}^{N\times D}$,这样的操作可以保证有足够的参数描述特征;接着 $\boldsymbol{F}_{ex}$ 经过全局平均池化得到全局特征 $\boldsymbol{F}_G \in \mathbb{R}^{1\times D}$;后续的网络结构与局部支路类似,最终得到加权全局特征 $\boldsymbol{F}_{GW} \in \mathbb{R}^{1\times D}$。为充分利用局部与全局特征优势,对 $\boldsymbol{F}_{LW}$ 与 $\boldsymbol{F}_{GW}$ 进行融合;由于 $\boldsymbol{F}_{GW}$ 与 $\boldsymbol{F}_{LW}$ 空间尺寸不同,对 $\boldsymbol{F}_{GW}$ 其进行重复操作获得 $\boldsymbol{F}_{GW} \in \mathbb{R}^{N\times D}$,$\boldsymbol{F}_{GW}$ 与 $\boldsymbol{F}_{LW}$ 进行逐点融合,获得混合特征 $\boldsymbol{F}_M$。

综上所述,混合通道注意力以如下公式总结:

$$\boldsymbol{F}_M = \mathrm{CON}(\mathrm{FC}_G(\mathrm{AvaPool}(\mathrm{MLP}(\boldsymbol{F}))) \cdot \boldsymbol{F}, \mathrm{FC}_L(\mathrm{AvaPool}(\boldsymbol{F})) \cdot \boldsymbol{F}) = \mathrm{CON}(\boldsymbol{F}_{GW}, \boldsymbol{F}_{LW}) \tag{3}$$

其中,FC_G,FC_L 分别为全局支路和局部支路的全连接神经网络,CON 为级联操作。

1.5　六自由度位姿估计

$\boldsymbol{F}_I$ 与 $\boldsymbol{F}_P$ 均输入 MCA 模块,获得混合实例特征 $\boldsymbol{F}_{IM}$ 与混合先验特征 $\boldsymbol{F}_{PM}$,两者交换特征,通过两个网络获得物体六自由度位姿。第 1 个是变形网络,由 3 层神经网络的 MLP 实现,通过预测一个逐点的变形场 $\boldsymbol{T} \in \mathbb{R}^{N_P\times 3}$,使得先验形状 $\boldsymbol{P}_P$ 变形从而重建物体模型 $\boldsymbol{M} \in \mathbb{R}^{N_P\times 3}$:

$$\boldsymbol{M} = \boldsymbol{P}_P + \boldsymbol{T} = \boldsymbol{P}_P + F_T(\boldsymbol{F}_{IM}, \boldsymbol{F}_{PM}) \tag{4}$$

其中,F_T 表示变形网络。第 2 个是对应网络,由 3 层神经网络的 MLP 加上 Softmax 实现,通过预测一个 $\boldsymbol{M}$ 到 $\boldsymbol{P}_I$ 的对应矩阵 $\boldsymbol{C}$ 建立起两者间的软对应:

$$\boldsymbol{P}_I' = \boldsymbol{C} \cdot \boldsymbol{M} = \boldsymbol{F}_C(\boldsymbol{F}_{IM}, \boldsymbol{F}_{PM}) \cdot \boldsymbol{M} \tag{5}$$

其中,$\boldsymbol{F}_C$ 表示对应网络。$\boldsymbol{C}$ 计算 N_I 个对应点,用于估计位姿。$\boldsymbol{P}_I'$ 表示预测的对应点,与 $\boldsymbol{P}_I$ 存在点对点的对应关系。最终,利用 $\boldsymbol{P}_I'$ 与 $\boldsymbol{P}_I$,通过 Umeyama 算法求解物体的六自由度位姿与尺寸。

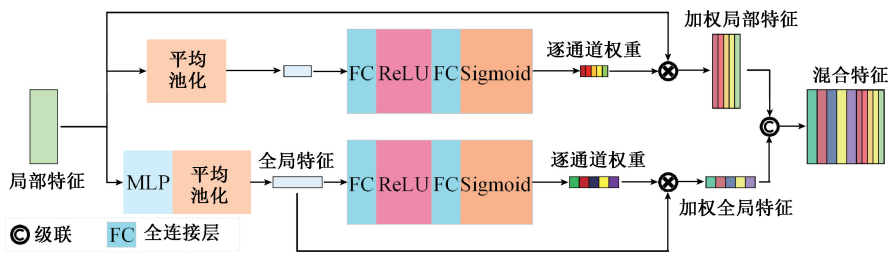

图 2　MCA 模块

Fig. 2　MCA module

1.6　损失函数

MCA6D 有 3 个估计目标用于估计物体六自由度位姿,包括关键点提取矩阵 $\boldsymbol{E}$、逐点的变形场 $\boldsymbol{T}$、对应矩阵 $\boldsymbol{C}$。为了训练 MCA6D,这里参考 SGPA 采用如下的损失函数:

$$L = \lambda_1 L_{pose} + \lambda_2 L_K \tag{6}$$

L_{pose} 用于估计 $\boldsymbol{T}$ 和 $\boldsymbol{C}$,它包含 4 项,其中两项通过真实的物体模型与六自由度位姿监督预测的 $\boldsymbol{T}$ 和 $\boldsymbol{C}$,其他两项用于正则化 $\boldsymbol{T}$ 和 $\boldsymbol{C}$ 的值。L_K 通过 $\boldsymbol{P}_I$ 与 $\boldsymbol{P}_K$ 间的倒角距离(chamfer distance, CD)构建,使得 $\boldsymbol{E}$ 可以通过 n 个点代表物体点云,其中 $n \ll N_I$。λ_1 与 λ_2 表示损失权重,在本文的实验中均设为 1.0。

2　实验与讨论

在本节中,本文方法在两个公开数据集上进行详细的测试,并通过真实场景的实验进一步检验本文方法。

2.1　数据集

本文采用类别级物体六自由度位姿估计领域权威的公开数据集 CAMERA 和 REAL[18] 用于训练和测试。数据集中包含瓶子、碗、相机、易拉罐、笔记本电脑和马克杯 6 类物体。CAMERA 是一个合成数据集,包含 300 K 合成图像;其中 275 K 为训练集,包含 1 085 个不同的物体实例;剩余的 25 K 用作测试集,包含 184 个不同的物体实例。REAL 是从真实场景中采集的,是 CAMERA 的补

充;7 个场景中的 4 300 张图片用于训练;6 个场景中的 2 750 张图像用于测试;训练集和测试集中每个类别包含有 3 个不同的物体实例。CAMERA 和 REAL 的测试集分别称为 CAMERA25 和 REAL275。

2.2 评估指标

本文采用用于物体六自由度姿态估计和 3D 物体检测的指标来评价类别级物体六自由度位姿估计的效果。对于六自由度位姿估计,本文报告 $n°$ $m$cm 的均值平均精度(mean average precision, mAP)。$n°$ $m$cm 表示当估计的位姿的旋转误差小于 $n°$ 且平移误差小于 $m$cm 时,该位姿被认为是正确的。注意,对于对称物体(如瓶子、碗和罐子),任何沿对称轴的旋转预测都被认为是正确的。此外,当马克杯的把手不可见,认为其是对称物体,反之为非对称物体。对于 3D 目标检测,本文报告 3D IoU (intersection over union)的 mAP,其精度阈值为 75%,简写为 $3D_{75}$。

2.3 实验细节

输入的 RGB-D 图像的分辨率为 640×480。为了减少计算量,通过实例分割获取的 RGB 图像块被统一缩放至分辨率 192×192。在权衡了计算量和信息的完整度后,先验形状的点云数量 N_p 设为 1 024,为了统一不同的物体点云的点云数量 N_l,通过下采样或重复转换至 1 024。先验几何特征、先验颜色特征、实例几何特征、实例颜色特征的维度 d 设为 64。MCA 模块中全局特征通道注意力支路中扩充后的特征维度为 1 024,衰减系数设为 16,即第 1 个全连接层将全局特征的维度从 1 024 降为 64,接着第 2 个全连接层将特征维度恢复至 1 024;局部特征通道注意力支路中为了保证足够的信息量,相对全局支路衰减系数减半,设为 8。本文使用 Adam 优化器训练 MCA6D,将训练步数设置为 120 K,epoch 设为 15,每步的批大小设置为 24。学习率按 epoch 动态调整,分别在 epoch 大于 5、10 时将学习率缩减为初始学习率的 0.6 倍、0.3 倍。初始学习率设置为 0.000 1。本文实验在一台带有 NVIDIA RTX 3090 GPU、Intel Xeon Gold 6138 CPU 和 128 G 内存的主机上进行。

2.4 实验结果

1)方法对比

在 CAMERA25 上的测试:本文方法只在合成数据集 CAMERA 上进行训练,并与 2 个基线方法[18,26]和 4 个前沿方法[20-21,24-25]进行对比。为了公平起见,NOCS[18],SPD[26],DualPoseNet[21],SGPA[25],CenterSnap[20],SAR-Net[24]和本文方法一样均只采用数据集 CAMERA 进行训练,具体结果如表 1 所示,其中 SGPA[25]的结果来自本实验平台按照原论文所提供细节训练和测试的结果,其他方法的结果来自公开论文。从表中第 3 行可以看出本文方法 $3D_{75}$ 的 mAP 达到了 86.3%,超过了基线方法 NOCS[18],SPD[26]分别达 16.8% 和 3.2%,超过了 SAR-Net[24]达 7.3%。本文方法在 10°2 cm 和 10°5 cm 的 mAP 也高达 81.8% 和 87.3%,超过 SAR-Net[24]分别达 6.5% 和 7.0%。在更严苛的评价标准 5° 2 cm 和 5°5 cm,本文方法的 mAP 仍然具有领先优势,为 69.3% 和 73.4%,超过 SAR-Net[24]达 2.6% 和 2.5%。

在 REAL275 上的测试:由于真实数据集 REAL 的数量较少,因此加上 CAMERA 的数据集作为补充进行训练,并保持来自 REAL 和 CAMERA 的数据量为 1∶3,对照的方法有 2 个基线方法[18,26]和 7 个前沿方法[19-25],为了公平比较,它们与本文方法采用同样的数据进行训练,具体实验结果如表 1 所示。同样,SGPA[25]的结果来自本实验平台,其他方法的结果来自公开论文。从表的第 4 行可以看出,本文 $3D_{75}$ 的 mAP 超过基线方法 NOCS[18],SPD[26]分别达 35.3% 和 12.2%,为 65.4%。本文方法在 10°2 cm 和 10°5 cm 的 mAP 达 61.7% 和 71.6% 超越 SAR-Net[24]达 11.4% 和 3.3%。在更严格的标准 5° 2 cm 和 5°5 cm 上同样能体现本文方法的优越性,超过 SAR-Net[24]达 7.6% 和 1.0%,超过 SGPA[25]达 3.0% 和 3.4%。上述结果说明本文方法具有较强的通用性,能准确地估计同类别新物体的位姿,非常符合实际应用的需求,同时,与前沿方法的量化对比说明了本文方法具有先进的性能。

表 1 各方法在数据集 CAMERA25 和 REAL275 上的性能对比

Table 1 Comparative results on CAMERA25 and REAL275 datasets

数据集	方法	mAP				
		$3D_{75}$	5°2 cm	5°5 cm	10°2 cm	10°5 cm
CAMERA25	NOCS[18]	69.5	32.3	40.9	48.2	64.6
	SPD[26]	83.1	54.3	59.0	73.3	81.5
	DualPoseNet[21]	**86.4**	64.7	70.7	77.2	84.7
	SGPA[25]	85.6	68.3	72.3	81.2	86.8
	CenterSnap[20]	-	-	66.2	-	81.3
	SAR-Net[24]	79.0	66.7	70.9	75.3	80.3
	本文方法	86.3	**69.3**	**73.4**	**81.8**	**87.3**
REAL275	NOCS[18]	30.1	7.2	10.0	13.8	25.2
	SPD[26]	53.2	19.3	21.4	43.2	54.1
	iCaps[23]	-	-	22.3	-	-
	FS-Net[19]	63.5	-	28.2	-	60.8
	CR-Net[22]	55.9	27.8	34.3	47.2	60.8
	DualPoseNet[21]	62.2	29.3	35.9	50.0	66.8
	SGPA[25]	64.8	36.2	39.9	61.5	70.7
	CenterSnap[20]	-	-	29.1	-	64.3
	SAR-Net[24]	62.4	31.6	42.3	50.3	68.3
	本文方法	**65.4**	**39.2**	**43.3**	**61.7**	**71.6**

图 3 较详细地展示了本文方法与 2 个基线方法 NOCS,SPD 的性能曲线,图 3(a)是在 CAMERA25 上的曲线,图 3(b)是在 REAL275 上的曲线。从图中可以看出,本文方法在 3D IOU 阈值大于 75%时具有较明显得领先优势。同时,本文方法在旋转的估计上明显较基线方法有明显的提升,在平移的估计方面同样展现了优势。

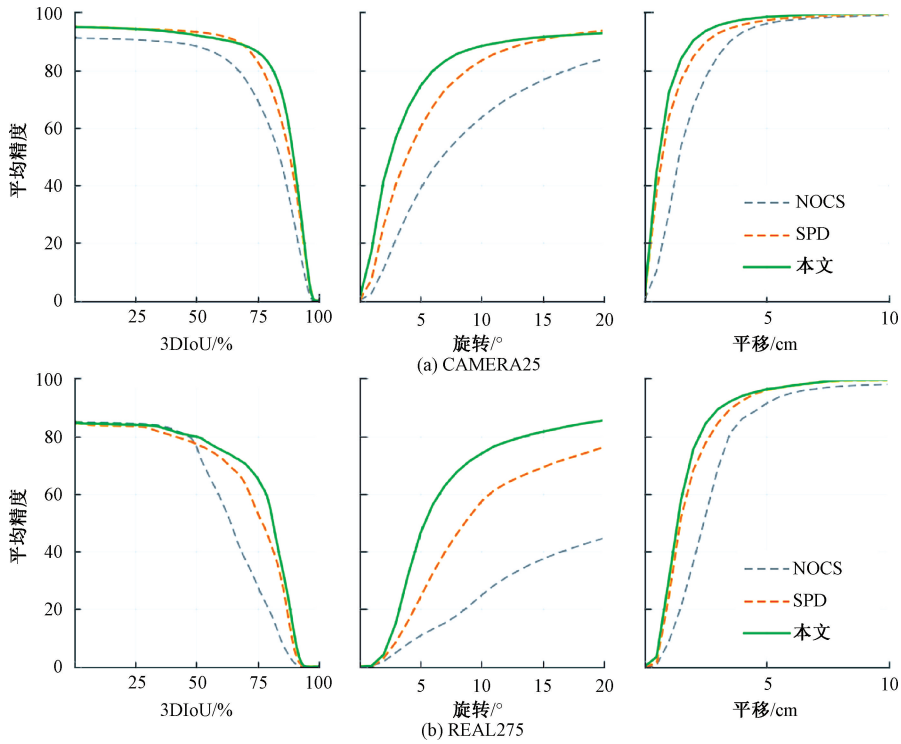

图 3　不同方法在 CAMERA25 和 REAL275 的性能曲线对比

Fig. 3　Performance curves comparison on CAMERA25 and REAL275 under different thresholds on 3D IoU, rotation, and translation

图 4 详细展示了本文方法对于估计每类物体六自由度位姿的性能,同样,图 4(a)展示的是在 CAMERA25 的结果,图 4(b)展示的是在 REAL275 的结果。从图中可以看出本文方法在各类物体上都有较优良的结果。比较而言,由于同类物体的外观变化,准确的旋转估计颇具挑战性,成为限制性能提升的关键,这一点在相机这一类上体现得尤为明显,因为不同的物体实例其外观变化最为明显,如图 5 所示。

图 6 展现了一些本文方法在 CAMERA25 和 REAL275 的六自由度位姿估计的可视化结果,从图中可以看出本文方法能较准确地估计物体的位姿。相较于基线方法 NOCS,本文方法能更准确地估计物体的旋转与尺寸。

2)消融实验

为了说明本文方法的创新,本文进行了针对 LCA 和 GCA 的消融实验,并使用第 1 节描述的方法作为基线。本文进行消融实验来说明 LCA 模块和 GCA 模块对 MCA6D 最终性能的贡献。首先,移除 GCA 模块只采用 LCA 模块。量化结果如表 2 第 3 行所示,与基线相比 $3D_{75}$ 从 65.4%下降至 64.4%,5° 2 cm 和 5°5 cm 也分别从 39.2%和 43.3%下降至了 34.4%和 38.7%。然后,移除 LCA 模块只采用 GCA 模块,同样的,从表 2 的第 4 行可以看到,与基线相比,$3D_{75}$,5° 2 cm 和 5°5 cm 也分别从 65.4%,39.2%,43.3%下降至 63.9%,33.1%,36.3%。实验结果表明 GCA 和 LCA 模块都有助于提升精度。具体来说,GCA 和 LCA 模块使用了通道注意机制根据通道的重要性为全局特征和局部特征的不同通道赋予不同的权重,使得这些特征更加高效和具有代表性。

表 2　在 REAL275 上本文方法不同模块,局部通道注意力(LCA)和全局通道注意力(GCA)的消融实验

Table 2　Ablation study of different modules, local channel attention (LCA) and global channel attention (GCA), in our method on REAL275

方法	LCA	GCA	mAP				
			$3D_{75}$	5°2 cm	5°5 cm	10°2 cm	10°5 cm
1	√		64.4	34.4	38.7	60.4	70.5
2		√	63.9	33.1	36.3	57.8	68.8
3	√	√	**65.4**	**39.2**	**43.3**	**61.7**	**71.6**

3)实物实验

为了进一步说明本文方法处理同类新物体的性能,针对 4 类存在干扰的场景,5 类物体,6 个物体实例对本方法进行了测试。这里通过微软 RealSense L515 相机采

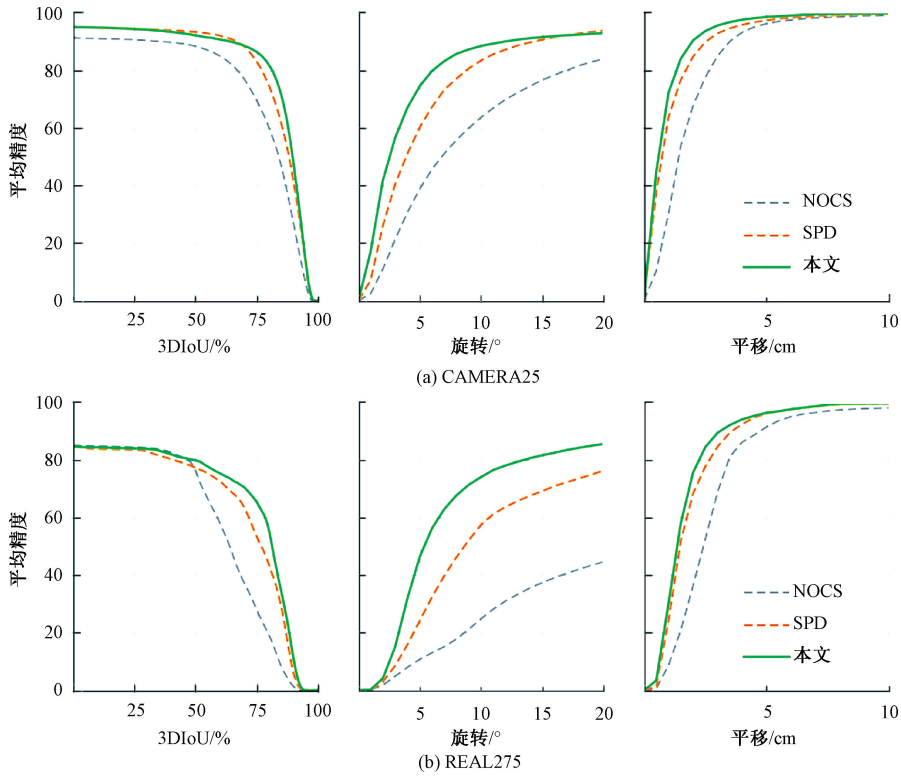

图 4 本文方法在 CAMERA25 和 REAL275 上对 6 类不同物体的性能曲线

Fig. 4 Performance curves of our method on CAMERA25 and REAL275, in terms of per-category 3D IoU, rotation, and translation

图 5 不同的相机实例

Fig. 5 Different camera instances

集 RGB-D 图像。4 类干扰包括光照变化、距离变化、背景杂乱和遮挡;5 类物体包括相机、瓶子、马克杯、碗和易拉罐。图 7 中展示了一些在真实场景下的测试结果,从图中可以看出本文方法在存在干扰的场景下对新物体的检测结果,展现了鲁棒且优良的性能,能准确地估计各物体的六自由度位姿。

3 结 论

为实现精准的物体六自由度位姿估计,解决干扰场景下位姿估计难题,本文提出了 MCA6D 方法。通过全局与局部特征级联实现多尺度特征融合,有效聚合细节与整体信息;利用通道注意力机制从通道维度优化特征,突出重要信息;通过先验形状变形重建物体模型,解决类别级位姿估计中的同类别物体形状变化难题。在公共数据集 CAMERA 和 REAL 上的实验,本文方法取得了

图 6 在 CAMERA25 和 REAL275 上的感性结果

Fig. 6 Some qualitative results of 6D object pose estimation on CAMERA25 and REAL275

86.3%(5°2 cm)、73.4%(5°5 cm)和 39.2%(5°2 cm)、43.3%(5°5 cm)的均值平均精度,优于 NOCS、SPD、

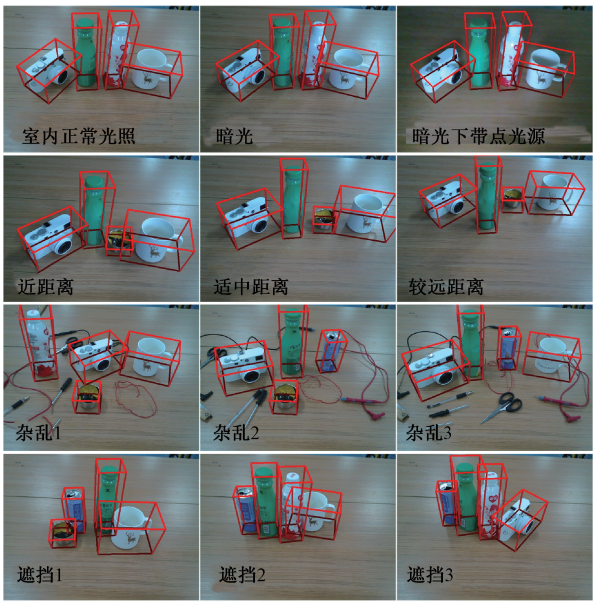

图 7　本文方法在真实场景中测试的感性结果

Fig. 7　Qualitative results of our method in the real-world scenes

SGPA 等主流类别级物体六自由度位姿估计方法。进一步的实物实验表明本文方法在存在光照变化、距离变化、背景杂乱、遮挡等干扰的场景中仍能准确估计物体六自由度位姿。

未来将进一步改进本文方法适用于透明物体位姿估计,透明物体在家庭场景中较常见,具有实用价值;将继续探索注意力机制用于改善场景信息提取,增强在存在干扰的场景的鲁棒性。

参考文献

[1] 吴国新,左云波,秦文丽,等. 工业室内环境中建立增强现实系统模型研究[J]. 电子测量与仪器学报, 2021, 35(5):196-201.
WU G X, ZUO Y B, QIN W L, et al. Research on an augmented reality system model in the industrial indoor environment[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(5): 196-201.

[2] 葛俊彦,史金龙,周志强,等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报, 2021, 41(8): 146-153.
GE J Y, SHI J L, ZHOU ZH Q, et al. A robotic grasping method based on three-dimensional detection network[J]. Chinese Journal of Scientific Instrument, 2021, 41(8): 146-153.

[3] 张自杰,张国良,曾静,等. 基于双视角点云拼接的机械手抓取方法[J]. 国外电子测量技术, 2022, 41(11):102-108.
ZHANG Z J, ZHANG G L, ZENG J, et al. Grasping method of manipulator based on two registered point clouds[J]. Foreign Electronic Measurement Technology, 2022, 41(11): 102-108.

[4] LI S Q, ZHANG X F. Research on hand-eye calibration technology of visual service robot grasping based on ROS[J]. Instrumentation, 2022, 9(1): 23-30.

[5] 陈特,蔡英凤,陈龙,等. 面向无人驾驶的三轴应急救援车辆并行控制方法[J]. 电子测量与仪器学报, 2022, 36(9):72-79.
CHEN T, CAI Y F, CHEN L, et al. Parallel control method for unmanned driving of three-axis emergency rescue vehicle[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(9): 72-79.

[6] WANG G, MANHARDT F, TOMBARI F, et al. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 11661-16606.

[7] LIU C, SUN W, ZHANG K, et al. Prior geometry guided direct regression network for monocular 6D object pose estimation [C]. 2022 41st Chinese Control Conference (CCC), 2022: 6241-6246.

[8] WANG C, XU D, ZHU Y, et al. Densefusion: 6D object pose estimation by iterative dense fusion[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 3338-3347.

[9] KEHL W, MANHARDT F, TOMBARI F, et al. SSD-6D: making RGB-based 3D detection and 6D pose estimation great again [C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 1530-1538.

[10] LI C, BAI J, HAGER G D. A unified framework for multi-view multiclass object pose estimation[C]. 2018 European Conference on Computer Vision (ECCV), CHAM: Springer International Publishing, 2018: 263-281.

[11] ZENG A, SONG S, NIEßNER M, et al. 3DMatch: Learning local geometric descriptors from RGB-D reconstructions[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 199-208.

[12] GOJCIC Z, ZHOU C, WEGNER J D, et al. The perfect match: 3D point cloud matching with smoothed densities[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 5540-5549.

[13] HUANG W, HUNG C, LIN I. Confidence-based 6D object pose estimation [J]. IEEE Transactions on Multimedia, 2022, 24(1): 3025-3035.

[14] RAD M, OBERWEGER M, LEPETIT V. Feature mapping for learning fast and accurate 3D pose inference from synthetic images[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2018: 4663-4672.
[15] PENG S, ZHOU X, LIU Y, et al. Pvnet: Pixel-wise voting network for 6DoF object pose estimation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3212-3223.
[16] 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]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 3848-3856.
[17] LIU J, SUN W, LIU C, et al. HFF6D: Hierarchical feature fusion network for robust 6D object pose tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 7719-7731.
[18] WANG H, SRIDHAR S, HUANG J, et al. Normalized object coordinate space for category-level 6D object pose and size estimation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 2637-2646.
[19] CHEN W, JIA X, CHANG H J, et al. FS-Net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 1581-1590.
[20] IRSHAD M Z, KOLLAR T, LASKEY M, et al. Centersnap: Single-shot multi-object 3D shape reconstruction and categorical 6D pose and size estimation[C]. 2022 International Conference on Robotics and Automation (ICRA), 2022: 10632-11064.
[21] LIN J, WEI Z, LI Z, et al. Dualposenet: Category-level 6D object pose and size estimation using dual pose network with refined learning of pose consistency[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 3540-3549.
[22] WANG J, CHEN K, DOU Q. Category-level 6D object pose estimation via cascaded relation and recurrent Reconstruction networks [C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: 4807-4814.
[23] DENG X, GENG J, BRETL T, et al. Icaps: Iterative Category-level object pose and shape estimation[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 1784-1791.
[24] LIN H, LIU Z, CHEANG C, et al. SAR-Net: Shape alignment and recovery network for category-level 6D object pose and size estimation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 6697-6707.
[25] CHEN K, DOU Q. Sgpa: Structure-guided prior adaptation for category-level 6D object pose estimation[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 2753-2762.
[26] TIAN M, ANG M H, LEE G H. Shape prior deformation for categorical 6D object pose and size estimation[C]. 2020 European Conference on Computer Vision (ECCV), CHAM: Springer International Publishing, 2020: 530-546.
[27] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. 2017 IEEE International Conference on Computer Vision (ICCV), 2017: 2980-2988.
[28] 黄聪,杨珺,刘毅,等. 基于改进 DeeplabV3+的遥感图像分割算法[J]. 电子测量技术, 2022, 45(21): 148-155.
HUANG C, YANG J, LIU Y, et al. Remote sensing image segmentation algorithm based on improved DeeplabV3+[J]. Electronic Measurement Technology, 2022, 45(21): 148-155.
[29] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6230-6239.
[30] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point SETS in a metric space[C]. Advances in Neural Information Processing Systems (NIPS), 2017.

作者简介

刘崇沛,目前于湖南大学电气与信息工程学院机器人视觉感知与控制技术国家工程研究中心攻读博士学位,主要研究方向为三维计算机视觉、深度学习、物体六自由度位姿估计和机器人操作。
E-mail: 2018088374@qq.com

Liu Chongpei is now a Ph. D. candidate at the National Engineering Research Center of Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha, China. His main research interests include 3D computer vision, deep learning, 6D object pose estimation, and robotic manipulation.

孙炜(通信作者),于 1997 年、1999 年和 2003 年获得湖南大学自动化工程系学士、硕士和博士学位,现为湖南大学电气与信息工程学院教授,主要研究方向为计算机视觉、机器人、神经网络和智能控制。
E-mail: david-sun@126.com

Sun Wei (Corresponding author) received the B. Sc., M. Sc., and Ph. D. degrees from the Department of Automation Engineering, Hunan University, Changsha, China, in 1997, 1999, and 2003, respectively. He is now a professor with the College of Electrical and Information Engineering, Hunan University. His main research interests include computer vision and robotics, neural networks, and intelligent control.