计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.010115] [CSTR: 32024.14.csa.010115]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 RTMPose-BRNM 的盒式关键点识别及位姿计算①

李浩萌[1,2], 王少威[1,3]

[1](武汉科技大学 计算机科学与技术学院, 武汉 430065)
[2](武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430065)
[3](武汉科技大学 机器人与智能系统研究院, 武汉 430065)
通信作者: 王少威, E-mail: wangshaowei@wust.edu.cn

摘　要: 针对在遮挡和日常环境中, 盒式物体位姿计算精度差等问题, 提出采用基于 RTMPose-BRNM 的盒式物体 2D 关键点检测和点云深度信息相结合的位姿计算方法. 首先, 引入 RFAConv 替换普通 Conv 卷积, 提高遮挡 2D 关键点坐标识别精确度; 使用 NATTEN 模块, 提高模型对盒式物体边缘轮廓点抽取能力; 设计混合感受野卷积 (mixed-perception convolution, MPC) 结构, 增强不同尺寸盒式物体识别适应性. 实验结果表明, RTMPose-BRNM 关键点识别算法平均像素距离误差为 0.98, 相比原 RTMPose 模型, 降低 1.19 像素误差; 改进后平移误差和旋转误差为 1.32% 和 0.96°左右.

关键词: 关键点识别; 混合感受野卷积模块; RTMPose-BRNM; 位姿计算

引用格式: 李浩萌,王少威.基于 RTMPose-BRNM 的盒式关键点识别及位姿计算.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/10115.html

RTMPose-BRNM-based Box Keypoint Recognition and Pose Computation

LI Hao-Meng[1,2], WANG Shao-Wei[1,3]

[1](School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)
[2](Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China)
[3](Institute of Robotics & Intelligent System, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: To address the low pose-estimation accuracy of box-shaped objects in occluded and general environments, a method is proposed that integrates 2D keypoint detection based on RTMPose-BRNM with point-cloud depth information. First, standard convolutions are replaced with RFAConv to improve the localization accuracy of occluded 2D keypoints. Subsequently, the NATTEN module is employed to enhance the model's capability in extracting edge contour points of box-shaped objects. Furthermore, a mixed-perception convolution (MPC) structure is designed to increase the model's adaptability to objects of varying sizes. Experimental results show that the proposed RTMPose-BRNM keypoint detection algorithm achieves an *MPDE* of 0.98, which is 1.19 pixels lower than that of the original RTMPose model. The improved framework yields translation and rotation errors of approximately 1.32% and 0.96°, respectively.

Key words: keypoint detection; mixed-perception convolution module; RTMPose-BRNM; pose estimation

位姿计算是机器人感知体系中的关键技术, 尤其在物流自动化领域具有广泛的应用前景. 在高效分拣、自动化堆码以及仓储盘点等典型场景中, 精准的 6 自由度 (6-DOF)[1]位姿解算是实现机械臂高效抓取的

① 基金项目: 国家重点研发计划 (2022YFB4700400); 国家自然科学基金 (62073249)
收稿时间: 2025-07-27; 修改时间: 2025-09-22, 2025-10-28; 采用时间: 2025-11-07; csa 在线出版时间: 2026-02-06

关键. 针对盒式物体关键点定位问题, 当前主流方法多采用多模态感知机制: 首先通过彩色图像提取物体表面的二维特征点, 继而结合深度相机或 TOF 传感器所获取的三维信息, 构建关键点三维定位与姿态估计模型. 现有技术路线归纳为两类: 基于传统几何特征的方法[2]和基于深度学习端到端方法.

在传统方法中, 盒式物体关键点检测基于三维点云数据, 融合形状和颜色信息迭代优化关键点位置. 然而该方法存在局限性: 首先该方法对干扰敏感, 易受深度相机噪声、盒式包装遮挡以及其他混乱背景等因素干扰[3]; 其次盒式包装关键点存在对称性特征, 易造成左右配对错误[4]; 由于传统方法运用穷举法, 导致关键点计算的时间复杂度呈指数上升, 关键点识别实时性能不高[5], 无法满足实际场景下的需求.

针对上述问题, 本文提出了一种基于深度学习的方法 RTMPose-BRNM (box-oriented RFAConv with NATTEN-enhanced mixed-perception context aggregation), 可以识别出盒式物体 2D 关键点, 并在 3D 关键点计算阶段结合深度信息, 利用高效多点透视位姿求解算法 (efficient perspective-n-point, EPnP) 与最小二乘优化 (levenberg-marquardt optimization, LM) 算法计算盒式物体的位姿信息.

1 相关工作

基于深度学习的盒式物体关键点识别主要有自下而上和自上而下两种方法.

在自下而上方法中, OpenPose[6]方法构建双分支卷积神经网络, 分别预测关键点热图与部件亲和场. 在推理阶段, OpenPose 通过贪婪图匹配算法将检测到的关键点组合为独立个体, 摆脱对外部检测器的依赖, 在遮挡、交互密集等复杂场景中展现出关键点关联鲁棒性. PifPaf[7]模型基于复合场预测机制, 构建关键点热图 (part intensity field, PIF) 与连接向量场 (part association field, PAF), 用于精细建模关键点位置及其拓扑结构. 为缓解低分辨率输入下的定位误差, 模型引入基于拉普拉斯分布的自适应回归损失函数. HigherHRNet[8]将高分辨率网络架构引入自下而上姿态估计任务, 构建特征金字塔, 增强多尺度关键点的建模能力. 该方法在保持高分辨率特征的同时, 结合逐层解码策略融合不同尺度表示, 并通过嵌入引导的聚类机制完成关键点的分组.

而在自上而下方法中, AlphaPose[9]提出对称积分回归 (symmetric integral regression), 以提升关键点的回归精度. 同时为解决多检测框冗余问题, 设计参数化姿态非极大抑制 (parametric pose NMS) 机制. HRNet (high-xresolution network)[10]构建并行多分辨率子网络架构, 在保持高分辨率特征流的同时, 引入跨尺度特征融合机制, 有效增强关键点定位的精度与鲁棒性, 但由于多分支结构设计导致模型参数量大. 为解决姿态估计中遮挡问题, 级联金字塔网络 (cascaded pyramid network, CPN)[11]采用双阶段架构: GlobalNet 利用特征金字塔提取多尺度语义信息, 初步定位关键点; RefineNet 进一步整合不同层级特征, 并引入在线困难关键点挖掘机制. TransPose[12]首次将 Transformer 架构引入关键点检测任务, 构建轻量卷积特征提取器与空间交互 Transformer 编码器的双模块结构, 该结构利用自注意力机制显式建模关键点间的全局依赖关系, 增强遮挡和远距离关节定位能力. YOLOv8-pose[13]设计并行的目标检测与关键点检测, 实现多任务学习, 该模型轻量高效, 通过任务头分离, 实现检测与姿态估计的解耦优化.

MMPose 团队提出的 RTMPose[14]构建了一种面向实时应用的轻量级姿态估计框架, 采用深度可分离卷积与结构重参数化策略, 有效压缩计算开销并提升推理效率. 模型引入自适应热图解码器, 根据下游任务需求动态调整关键点输出分辨率, 增强在不同硬件环境下的适应性.

但 RTMPose 在盒式物体关键点检测上仍面临一系列不足: 首先盒式物体相互遮挡对关键点定位造成难度, 部分关键点可能在图像中不可见, 影响关键点估计的准确度; 其次由于 RTMPose 作为人体姿态估计模型, 偏向识别内部关键点, 但盒式物体的 6 个关键点为轮廓边缘点, 造成精度下降; 除此之外, 盒式物体与深度相机之间距离远近不一, 并且不同盒式物体尺寸存在差异性, 导致不同尺度的关键点在图像中直观呈现为所占的像素值不一致.

本文提出的 RTMPose-BRNM 模型在现有的 RTMPose 上有如下改进.

(1) 针对同类物品遮挡引发的关键点误识别问题, 采用感受野增强卷积 (receptive field augmented convolution, RFAConv) 替代 Conv 标准卷积模块, 增强模型对盒式物体局部遮挡区域的语义推断能力.

(2) 将 GAU 改进为 NATTEN 模块, 使用其滑动窗口局部自注意力机制, 在邻域内计算位置敏感的特征

响应, 从而使盒式物体的轮廓关键点定位更贴合物体, 盒式物体边缘锯齿状区域表现出更强的几何一致性.

(3) 为缓解盒式关键点平均像素距离误差 (mean pixel distance error, *MPDE*) 下降, 提出混合感受野卷积 (mixed-perception convolution, MPC) 模块, 取代原单一 7×7 卷积层.

2 盒式物体关键点的 2D 检测

2.1 盒式关键点的描述

由于盒式模型自我遮挡, 将存在的 8 个关键点分类为未遮挡关键点和遮挡关键点, 根据具体视角捕捉角度, 如图 1 所示, 将遮挡分为如下 3 种情况: (1) 视角捕获 1 个平面, 设置 4 个未遮挡关键点和 4 个遮挡关键点; (2) 视角捕获 2 个平面, 设置 6 个未遮挡关键点和 2 个遮挡关键点; (3) 视角捕获 3 个平面, 设置 7 个未遮挡关键点和 1 个遮挡关键点. 实际场景下多个同类盒式物体重叠时, 在上述之一情况下继续多重遮挡.

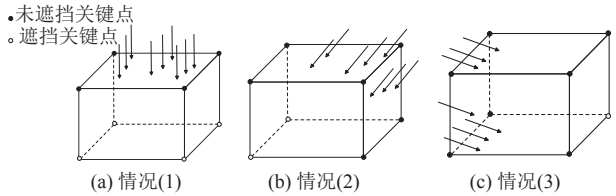


图 1 盒式关键点自我遮挡的 3 种情况

2.2 针对关键点检测 RTMPose 改进模型

2.2.1 重塑感受野注意力卷积

在多个盒式物体摆放的场景中, 由于相机角度不同, 盒式物体遮挡无法避免. 而 RTMPose 标准卷积采用均匀采样策略, 对遮挡区域产生过度平滑效应, 出现盒式关键点特征响应峰值偏移. 如图 2 所示, 模型易将盒式物体的实际关键点 (黑色点) 误识别至白色点位置, 导致 *MPDE* 值降低.

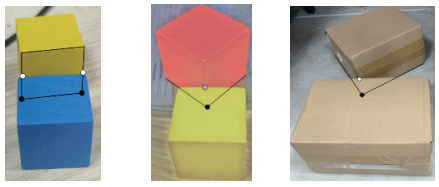


图 2 遮挡关键点的拟合错误现象

为此采用动态感受野注意力卷积[15]的方法, 建立关键点感知的权重分配机制, 解决卷积核参数共享问题, 对盒式各关键点采用不同权重, 增强模型对盒式物体遮挡关键点的提取能力, 增强对遮挡点的响应. 式 (1) 为感受野注意力卷积计算公式:

$$\begin{cases} A_{\mathrm{rf}} = Softmax(\boldsymbol{g}^{1\times1}(AvgPool(X))) \\ F_{\mathrm{rf}} = ReLU(Norm(\boldsymbol{g}^{k\times k}(X))) \\ F = A_{\mathrm{rf}} \times F_{\mathrm{rf}} \end{cases} \tag{1}$$

其中, $AvgPool(\cdot)$为平均池化操作, $\boldsymbol{g}^{k\times k}$表示大小为 $k\times k$ 的分组卷积, k 表示卷积核的大小, *Norm* 表示归一化, X 表示输入的盒式物体特征图, F 通过将盒式物体注意力图A_{rf}与变换后的盒式关键点的感受野空间特征 F_{rf} 相乘获得.

根据式 (1), 以下为感受野注意力卷积的具体步骤.

(1) 输入盒式物体特征图 X 经过$AvgPool(X)$, 并使用$\boldsymbol{g}^{1\times1}$交互信息, 利用 *Softmax* 函数处理平均池化后的结果, 生成盒式物体的A_{rf}.

(2) 将盒式物体特征图 X 通过$\boldsymbol{g}^{k\times k}$操作进行盒式关键点特征提取, 对$\boldsymbol{g}^{k\times k}(X)$的结果应用 *Norm* 和 *ReLU* 激活函数, 生成盒式关键点的F_{rf}.

(3) 将A_{rf}与F_{rf}进行乘法操作, 生成最后结果.

上述卷积的计算过程称为 RFA, 并将 RTMPose 的 Backbone 结构中常规卷积替换为 RFAConv, 在处理复杂的遮挡情况时, 降低同类盒式物体遮挡导致关键点特征混淆的错误率.

2.2.2 轮廓注意力改进

在 RTMPose 基准模型中, 全连接层 (FC) 与 SimCC 回归头之间引入 GAU, 充分挖掘人体骨骼结构的全局依赖关系, 该策略对人体姿态估计任务表现优异.

然而人体骨架的判别信息集中于关节内部, 而盒式物体的关键点则更偏向于物体边缘轮廓区域, 二者在空间分布存在差异, 直接沿用 GAU, 导致盒式物体 2D 关键点坐标的精度降低. 为缓解上述差异, 本文将 GAU 替换为基于邻域注意力机制的 NATTEN[16]模块. NATTEN 仅在 $k\times k$ 局部窗口内计算自注意力, 显著降低计算开销, 并实现逐像素滑动窗口的细粒度特征聚合.

具体地, 记输入特征图的尺寸为 $B\times H\times W\times C$, 其中 B 为 batch size, H 和 W 分别为高和宽, C 为通道数. 输入张量首先通过 3 个独立的 1×1 卷积, 分别投影生成 Query (Q)、Key (K) 和 Value (V). 随后, neighborhood attention 使用每个像素 Q 与其最近邻的 K 之间的点积, 加上相对位置偏置 (relative positional bias, RPB), 计算注意力权重. 具体计算形式如式 (2) 所示:

$$A_i^k = \begin{bmatrix} Q_i K_{\alpha_1(i)}^{\mathrm{T}} + B_{(i,\alpha_1(i))} \\ Q_i K_{\alpha_2(i)}^{\mathrm{T}} + B_{(i,\alpha_2(i))} \\ \vdots \\ Q_i K_{\alpha_k(i)}^{\mathrm{T}} + B_{(i,\alpha_k(i))} \end{bmatrix} \tag{2}$$

其中, $\alpha_i(j)$表示 j 的第 i 个近邻, B 为可学习的相对位置偏置.

并对每个像素的所有邻居权重A_i^k进行 *Softmax* 归一化, 将其与对应的邻域 V 向量相乘并加权求和, 以[B, H, W, C]的形式返回, 直接送入残差连接和多层感知器 (MLP) 单元, 构成 NATTEN 单元模块.

NATTEN 结构采用邻居注意力, 分为如下重要模块, 其结构如图 3 所示.

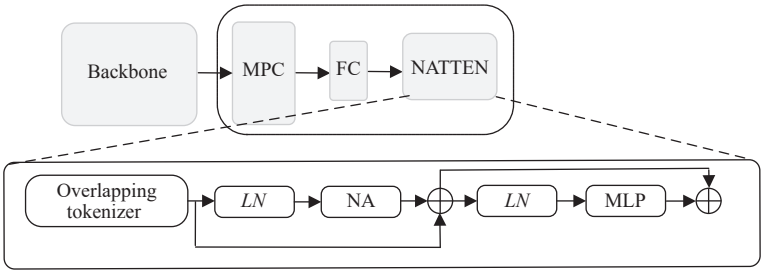


图 3 NATTEN 结构图

(1) *LN* (layer normalization), 对其输入图像进行归一化处理, 包括对通道方向归一化, 加快学习的速度并使训练过程平稳, 提高模型的抗干扰能力, 具体的计算如式 (3) 所示:

$$\begin{cases} LN(x_{i,c}) = \dfrac{x_{i,c} - \mu_i}{\sigma_i}\gamma + \beta \\ \mu_i = \dfrac{1}{C}\displaystyle\sum_{c=1}^{C} x_{i,c} \\ \sigma_i = \sqrt{\dfrac{1}{C}\displaystyle\sum_{c=1}^{C}(x_{i,c} - \mu_i) + \varepsilon} \end{cases} \tag{3}$$

其中, $x_{i,c}$为第 i 个像素在第 c 个通道上的激活值, μ和σ是均值和标准差, γ和β是可学习的缩放和平移参数.

(2) NA 中采用局部自适应视窗注意力机制, 根据各个像素和周围像素的相关程度对每一个像素进行衡量, 并采用可学习的卷积核灵活地设置局部关注的区域大小, 根据边界曲线的轮廓变化自动地重构点的关注权重.

(3) MLP 由两个线性层和 GELU 激活函数组成, 对通道维度执行非线性变换, 增强特征表达能力. 该模块与残差连接共同构成 NATTEN 的第 2 个子层.

2.2.3 混合感受野卷积

在 RTMPose 盒式关键点检测的任务中, 不同尺寸盒式包装与其成像距离存在较大差异, 导致关键点在特征图上像素覆盖范围变化显著: 1 m 内拍摄时, 纸箱关键点覆盖范围为 5×5 像素, 而桌面级木块关键点仅占 2×2 像素区域, 同时根据相机与盒式物体的距离不同, 相同的盒式物体在特征图中呈现的关键点覆盖范围也不同. 使用单一 7×7 Conv 模块, 在细粒度定位时难以同时兼顾大尺寸盒式包装的上下文信息和小尺寸目标的精确细节, 从而出现 *MPDE* 值偏高的问题.

为了充分兼顾不同尺度关键点的定位需求, 解决 *MPDE* 值偏高问题, 本文在 MPC 模块中设计 3 条并行分支, 如图 4 所示, 分别命名为局部细节分支 (fine-grained branch, FG-B)、上下文关联分支 (contextual branch, CT-B) 和全局语义分支 (global context branch, GC-B). MPC 将输入特征划分为 3 条并行分支, 每条分支依次经过卷积、归一化与激活, 再在通道维度汇聚.

(1) 对于覆盖 2×2 像素的小范围关键点, 若使用 2×2 或 4×4 偶数尺寸的可变形卷积 (deformable conv), 会因缺乏唯一中心像素, 卷积响应在方向的定位会产生偏置, 导致方向与位置估计不稳定; 而 3×3 可变形卷积可完整覆盖 2×2 区域并提供单一中心像素, 在微小位移下保持平移不变性与稳定性. 故 FG-B 采用 3×3 可变形卷积, 其计算如式 (4):

$$\begin{cases} F_{\mathrm{out}}(p) = \displaystyle\sum_{k=1}^{K} W_k F_{\mathrm{in}}(p + p_k + \Delta p_k) \\ \Delta p_k = \Omega_k(F_{\mathrm{in}}) \end{cases} \tag{4}$$

其中, F 为特征图, K 为核大小的采样点数, W_k为卷积权重, p 为当前输出位置, p_k为规则卷积采样格点, Δp_k为可学习偏移, Ω_k为 3×3 普通卷积, 预测所有Δp_k.

配合 BN+GELU, 侧重捕获 2×2 像素级微小关键点的边缘与角点信息, 保证高分辨率细粒度表达.

(2) 对于更大尺度的上下文, 6×6 可变形卷积无中心像素; 7×7 可变形卷积计算开销陡增, 且过大的感受野易引入背景干扰与过度平滑, 均会带来训练不稳定. 基于以上权衡, CT-B 采用 5×5 Deformable Conv, 在计

算量可控的同时, 其感受野扩展至 9×9 像素, 与 FG-B 感受野范围形成互补.

(3) GC-B 由 Channel-SE 利用全局平均池化生成通道权重; 为了与 FG-B 和 CT-B 形成轻量且稳定的语义门控配合, 使用 1×1 卷积在不改变空间感受野的前提下进行逐像素通道混合, Spatial-SE 通过 1×1 卷积生成空间注意力图, 为局部分支提供长程语义先验, 提升复杂场景下的鲁棒性. 本文将并行的 Channel-SE 与 Spatial-SE 记为 SCSEBolck 模块.

FG-B、CT-B 和 GC-B 在消融实验中的表现如表 1 所示. FG-B 分支中 3×3 卷积在定位精度与稳定性间表现最佳; 在 CT-B 分支中, 5×5 卷积在精度与计算量间达到最佳平衡; 保留 GC-B 并同时启用 Channel-SE 与 Spatial-SE 时性能最优.

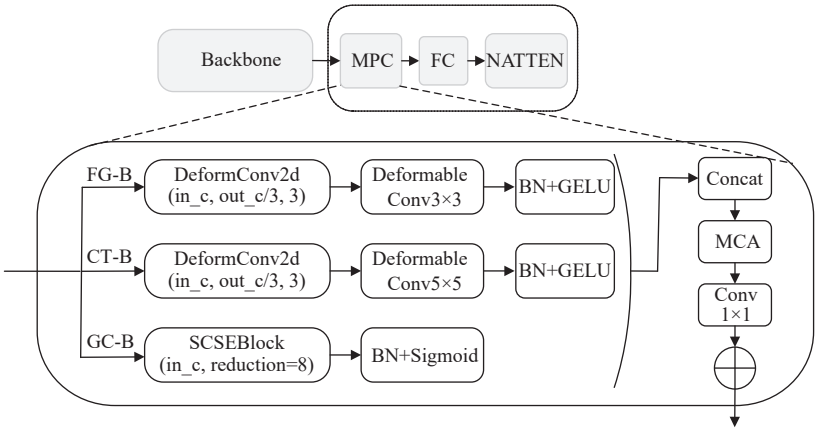


图 4 MPC 结构图

表 1 MPC 模块组件消融实验

目标	实验配置	*MPDE*	FPS (f/s)
FG-B	2×2 DConv	1.83	74.6
	3×3 DConv	0.98	65.8
	4×4 DConv	1.99	61.2
CT-B	5×5 DConv	0.98	65.8
	6×6 DConv	1.92	61.4
	7×7 DConv	1.87	57.3
GC-B	保留GC-B	0.98	65.8
	移除GC-B	2.13	72.4
	仅保留Channel-SE	1.96	68.9
	仅保留Spatial-SE	1.73	69.5

通过混合感受野覆盖关键点区域的多尺度空间上下文, 最后采用门控融合机制动态整合多分支特征, 融合后特征经 1×1 卷积实施跨通道重组, 嵌入运动感知通道注意力 (motion-aware channel attention, MCA), 其权重系数根据相邻帧光流幅值动态调整, 增强时序连续性, MCA 的计算如式 (5):

$$\widetilde{F} = (\sigma(W_2\varphi(W_1 GAP(F \odot M)))) \odot F \tag{5}$$

其中, ⊙表示按通道广播的逐元素乘, $GAP(\cdot)$为全局平均池化, W_1和W_2为两层全连接权重, $\varphi(\cdot)$为 GELU, $\sigma(\cdot)$为 Sigmoid 函数.

综上, 改进后的 RTMPose 结构如下, 将其命名为 RTMPose-BRNM, 如图 5 所示.

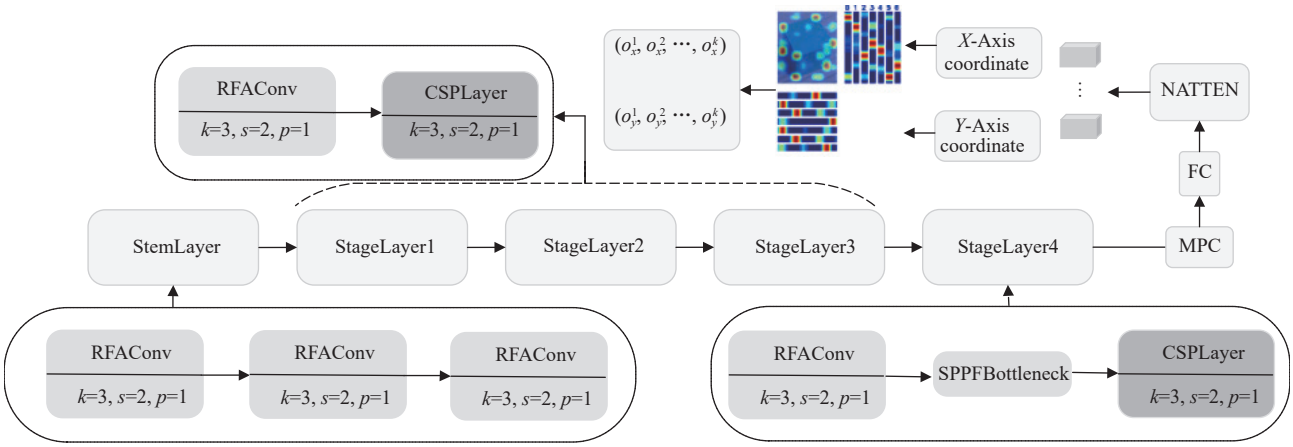


图 5 RTMPose-BRNM 结构图

3 盒式物体 3D 位姿计算

3.1 盒式物体 2D 关键点转换 3D 关键点

3.1.1 2D 关键点转换 3D 关键点计算公式

通过上述过程已获取 2D 关键点, 但使用 EPnP 算法计算深度相机相对于盒式物体的位姿, 需要将其转换为对应 3D 关键点. 给定一个像素点的 2D 图像坐标 (u, v) 和对应点的深度信息 $D(u, v)$, 将其转换到深度相机坐标系中的三维坐标 (X, Y, Z), 该算法的公式如式 (6) 所示:

$$\begin{cases} X = \dfrac{(u-c_x)\cdot Z}{f_x} \\ Y = \dfrac{(v-c_y)\cdot Z}{f_y} \\ Z = D(u,v) \end{cases} \tag{6}$$

其中, f_x 与 f_y 为像素坐标系下的焦距, c_x 与 c_y 为图像中心的位置.

3.1.2 深度异常点预处理

由于盒式物体受反光与低纹理影响, 深度图中部分像素的深度值 Z 出现异常波动和缺失. 为保证位姿计算稳定性, 本文在 2D 关键点至 3D 关键点转换前, 对每个盒式关键点局部邻域实施自适应筛选.

为避免对全帧 640×480 点云的高开销处理, 将全局点云对象转化为局部点云, 以(u_0, v_0)为中心取 31×31 邻域, 使用式 (6) 计算盒式关键点邻域内点云, 并对有效深度像素的局部点云, 使用 RANSAC 拟合局部平面, 其中, 最小采样集 s=3, 置信度 p= 0.99, RANSAC 的残差定义式 (7) 所示:

$$r_i = \left| Z_i - (\bar{a}X_i + \bar{b}Y_i + \bar{c}) \right| \tag{7}$$

其中, $\bar{a}$, $\bar{b}$, $\bar{c}$为拟合平面参数.

RANSAC 算法阈值 T 采用动态计算, 以适应不同盒式场景的深度, 其定义如式 (8) 所示:

$$T = a + b\bar{Z}^2 \tag{8}$$

其中, $\bar{Z}$为邻域有效点的深度均值; 根据经验取 a=6 mm, b=2 mm/m^2. 满足 r_i<T 的像素组成内点集 K, 若$|K|$<5, 判定该关键点邻域深度为不可靠估计, 直接剔除该关键点, 不进入后续位姿计算.

3.2 基于 EPnP 算法和 LM 算法的盒式物体位姿测量

EPnP 算法用于求解深度相机相对于盒式物体的旋转矩阵 R 和平移矩阵 t, 算法输入 3D 关键点坐标、对应 2D 关键点坐标、深度相机内参矩阵、畸变系数、初始化位姿参数, 算法输出位姿参数 R 与 t; 但 EPnP 为基于线性模型的近似算法, 存在误差; 还需进一步优化 R 和 t, 使用 LM 算法对 EPnP 输出的位姿参数进行非线性最小二乘优化, LM 算法迭代优化矩阵 R 和平移矩阵 t, 使重投影误差最小, 式 (9) 为该优化过程:

$$\begin{cases} p_i = K(RP_i + t) \\ M_{\mathrm{RT}} = \arg\min\limits_{R,t} \sum\limits_{i=1}^{n} \left| p_i - p_i' \right|^2 \end{cases} \tag{9}$$

其中, p_i是 2D 图像点, P_i为 3D 点的世界坐标系坐标, K 为内参矩阵.

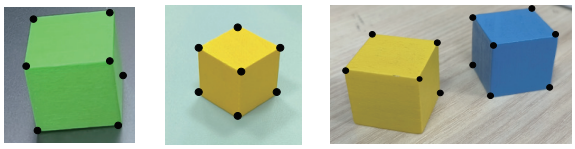
4 实验结果与分析

4.1 图像采集与环境搭建

实验平台运行在 Windows 10 64 位操作系统上, 使用 Python 3.9 作为开发语言. 硬件配置为 NVIDIA GeForce RTX 4060Ti GPU (16 GB 显存) 搭配 Intel i5-12600KF CPU 和 CUDA 12.4. 深度信息传感器选用 Intel RealSense D435 相机, 支持 30 f/s 帧率, 分辨率为 640×480 像素.

4.2 盒式物体数据集与评价指标

本文制作数据集由下述类别构成, 小尺寸桌面级精细检测采用 DOD (desktop object dataset), 由相机采集构成; 中尺寸盒式包装采用 SCD (stacked carton dataset)[17]公开数据集; 大尺寸盒式物体采用 CD (container dataset), 整合网络开源数据, 包含 300 组类型集装箱图像. DOD 和 SCD 数据集的部分示例如图 6 和图 7 所示.



(a) 绿色情况 (b) 黄色情况 (c) 混合情况

图 6 DOD 数据集部分示例

(a) 不规则包装 (b) 规则包装 (c) 干扰包装

图 7 SCD 数据集部分示例

在数据预处理流程中, 通过预训练 YOLOv8 检测, 将样本数据中重叠比例大于 40% 的过滤, 获得 16 248

张数据集, 19 457 个实例分配, 其中 DOD、SCD、CD 分别有 5 815、5 416、5 017 张, 并按照 6:2:2 比例划分训练集、验证集和测试集.

为验证遮挡情况下模型性能, 使用图片处理技术将盒式包装进行人为遮挡, 以模拟遮挡发生, 设计 3 组遮挡情况, 10% 比例的轻微遮挡, 30% 比例的中度遮挡和 50% 比例的严重遮挡, 如图 8 所示.

(a) 人为10%遮挡 (b) 人为30%遮挡 (c) 人为50%遮挡

图 8 盒式包装人为遮挡情况

盒式物体关键点定位采用 *MPDE* 来评估, 其计算公式如式 (10):

$$MPDE = \frac{1}{K}\sum_{k=1}^{K}\|\widetilde{p}_i - p_i\| \tag{10}$$

其中, *MPDE* 为平均像素距离误差, K 为图像内的关键点总数, $\widetilde{p}_i$和p_i为第 i 个关键点的预测与真值像素坐标.

MPDE 用于评估 RTMPose-BRNM 的模型关键点预测性能, 旋转误差和平移误差 (式 (11)) 用于评估位姿计算. 旋转误差和平移误差需要实时的深度数据, 作为实时检测时的衡量指标.

$$\begin{cases} e_t = \dfrac{\left\|t_{\text{true}} - t_{\text{pred}}\right\|}{\|t_{\text{true}}\|} \times 100 \\ e_{\text{rot}} = \max_{i=1}^{3} \cos^{-1}\left[\text{dot}(r_{\text{true}}^{k}, r_{\text{pred}}^{k})\right] \times \dfrac{180}{\pi} \end{cases} \tag{11}$$

其中, e_t为平移矩阵误差, e_{rot}为旋转矩阵误差.

4.3 RTMPose 关键点模型改进实验

将上述数据集在 YOLOv8n 和 RTMPose-BRNM 模型中训练, 将训练得到的权重文件部署后结合 Realsense D435 深度相机, 实时获取盒式物体的彩色图与深度图数据, 将实时数据分别输入模型中, 将 YOLOv8n 识别的目标框权重传递给 RTMPose-BRNM, 其盒式物体关键点预测结果如表 2、表 3 所示, 可视化如图 9 所示, RTMPose-BRNM 的盒式物体关键点识别中像素距离误差为 0.85, 表明盒式关键点坐标精度识别稳定. 最后测量盒式物体 3D 位姿, 平移误差和旋转误差为 1.13% 和 0.93°左右, 误差在合理区间内. 综上实验结果表明, 改进后的测量方法具有可行性.

为了更直观地对比改进后的模型在盒式物体关键点定位任务中的效果, 选取数据集中的部分图像进行比较, 结果如图 10 所示.

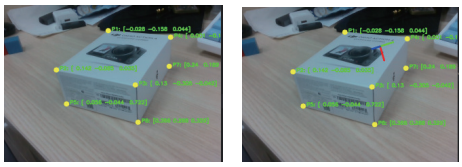
表 2 盒式关键点数据 (m)

编号	*X*	*Y*	*Z*
1	−0.028 434 175	−0.304 831 2	0.722 000 062 46
2	−0.158 032 183	−0.041 884 9	0.366 000 026 46
3	0.044 432 694 9	0.000 947 81	0.268 000 006 67
4	0.142 299 954 2	−0.120 341 7	0.329 000 025 98
5	−0.085 161 803	0.036 549 44	0.240 000 009 53
6	0.035 212 827 9	0.055 927 49	0.189 000 010 49
7	0.129 634 996 1	−0.043 793 9	0.307 000 011 20

注: *X*, *Y*, *Z*表示一个语义关键点的3D坐标

表 3 位姿计算结果

Rotation (rad)	Translation (m)
(−1.929 359 5, −1.492 962 0, 0.399 768 9)	(0.178 343 24, −0.264 560 2, 1.379 395)



(a) 关键点识别效果 (b) 位姿识别效果

图 9 盒式物体关键点和位姿可视化

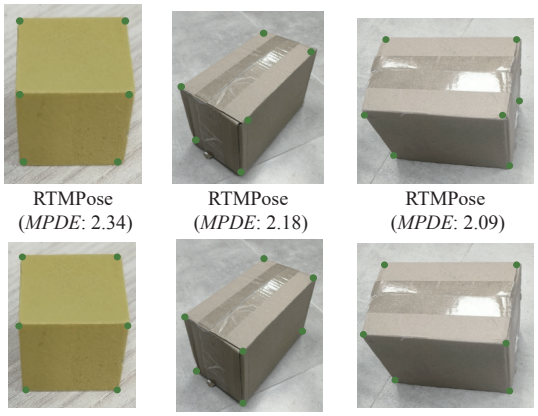


图 10 数据集上关键点定位效果

4.3.1 盒式物体关键点对比实验

为验证改进 RTMPose-BRNM 的有效性, 选取了 HRNet、OpenPose、YOLO-pose[18]、AlphaPose 等人体姿态估计模型及物体位姿估计模型 CenterSnap[19]、GDR-Net[20]进行对比. 如图 11 所示, 实验结果显示: 基线 RTMPose 以 83.1 f/s 的速度居首, *MPDE* 为 2.17; 传统模型 HRNet 与 AlphaPose 在速度与精度上均不占优. 改进后的 RTMPose-BRNM 的 *MPDE* 降至 0.98, 相

对 RTMPose 下降 54.8%, 较 YOLO-pose 下降 59.5%, 较 OpenPose 下降 58.8%, 较 HRNet 下降 79.1%, 在复杂背景和遮挡场景下关键点定位更稳健; 同时保持 65.8 f/s 的实时性, 较 RTMPose 速度仅下降 20.8%. GDR-Net 物体位姿估计模型的 FPS 为 74.4 f/s, 但其 *MPDE* 为 2.78, 高于 RTMPose-BRNM.

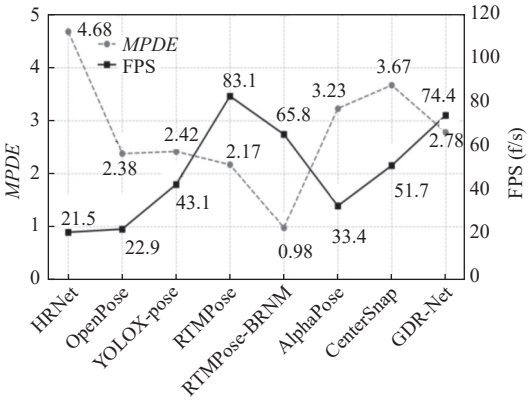


图 11 关键点识别模型性能比较

4.3.2 改进 RTMPose 模型消融实验

为验证本文模型所提方法中 RFAConv、NATTEN、MPC 的有效性 (表 4 分别以 R、N、M 替换), 设计以 RTMPose 网络为基准的 T1、T2、T3、T4 这 4 组消融实验, “√”代表使用该结构, 结果如表 4 所示.

表 4 消融实验结果

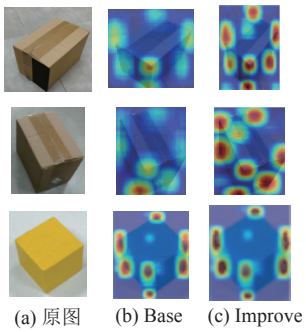
模型	R	N	M	*MPDE*	Params	FPS (f/s)
Base	—	—	—	2.17	5.42	83.1
T1	√	—	—	1.42	6.59	74.3
T2	—	√	—	1.58	6.75	75.7
T3	—	—	√	1.21	9.91	68.9
T4	√	√	√	0.98	10.5	65.8

在消融实验结果中, 实验组 T1 在主干网络中引入 RFAConv, 相较于 Baseline 的 *MPDE* 降低了 0.75; 实验组 T2 采用 NATTEN, 增强了对轮廓边缘点的识别率, *MPDE* 降低了 0.59; 实验组 T3 使用 MPC 结构, *MPDE* 降低 0.96; 实验组 T4 进行三者的综合改进, *MPDE* 降低 1.19.

如图 12 所示, 输入相同原图展示消融结果, 图 12(b) 为 Base 模型, 图 12(c) 为 Improve 模型, 在 Base 上分别接入 RFAConv、NATTEN、MPC 的输出. RFAConv 关键点感知的权重分配机制, 补偿被遮挡区域的特征缺失; NATTEN 利用邻域注意力对边缘邻近像素进行自适应加权; MPC 通过 3×3、5×5 可变形卷积与全局上下文模块覆盖从局部细节到大范围上下文的特征, 各类尺度与全局场景的 *MPDE* 均有降低. 与 Base 相比, 3 种模块均带来热度聚焦.

4.3.3 不同卷积模块比较

为验证 RFAConv 在 RTMPose 中的表现, 在相同网络结构和参数情况下引入不同的变体卷积, 包括部分卷积 PConv[21]、线性可变形卷积 LDConv[22]、全维动态卷积 ODConv[23]与可切换空洞卷积 SAConv[24], 对比其对模型指标影响, 结果如表 5 所示.



(a) 原图 (b) Base (c) Improve

图 12 不同模块下可视化效果

表 5 不同卷积模块在 RTMPose 的效果

模型	*MPDE*	FPS (f/s)
Conv	2.17	83.1
PConv	2.02	81.7
LDConv	1.95	85.3
ODConv	1.76	88.9
RFAConv	1.42	74.3
SAConv	1.64	74.4

PConv 减少冗余计算和内存访问次数, 提高特征提取的效率, 但 PConv 只对部分特征通道进行卷积操作, 导致对盒式包装模型的可解释性降低. LDConv 引入任意参数数量和采样形状, 提供在网络性能与开销之间进行权衡的更多选择, 但对本模型提升有限. ODConv 采用多维注意力并行策略, 对卷积核 4 个维度加权调整, 以增强特征提取能力. SAConv 使用不同膨胀率的卷积核, 捕获不同尺度的特征, 并且其全局上下文模块可以提取图像的全局信息, 并将其用于指导卷积核的选择. RFAConv 引入网络后, 解决同类盒式包装物体遮蔽点位置偏移问题, 盒式包装模型的 *MPDE* 值显著降低.

4.3.4 遮挡情况下模型比较

为验证盒式包装遮挡情况下 RFAConv 模块对 *MPDE* 值的影响, 进行了 8 组实验. A1、B1、C1、D1 分别为模型 RTM-POSE、YOLO-pose、HRNet-W32、

RSN-50. A2 组在 FPN 的 P3、P4、P5 这 3 个特征层中, 将原本的 3×3 卷积统一替换为 RFAConv; B2 组在 YOLO-pose 颈部的 C3 模块内, 将所有 3×3 卷积替换为 RFAConv; C2 组在 HRNet-W32 的 Stage 3 和 Stage 4 中, 将每个 Bottleneck 块的第 2 个 3×3 卷积整体替换为 RFAConv; D2 组在 RSN-50 骨干网络中, 把所有 Bottleneck 块的中间 3×3 卷积替换为 RFAConv.

采用 10%、30% 和 50% 比例的盒式遮挡, 计算盒式包装在每种情况下的 *MPDE* 值, 其结果如表 6 所示.

表 6　遮挡情况下的盒式包装 *MPDE* 实验结果

组别	模型	遮挡比例		
		10%	30%	50%
A1	RTMPose	2.89	4.05	5.37
A2	RTMPose-RFA	1.91	2.75	3.44
B1	YOLO-pose	3.18	4.36	6.46
B2	YOLO-pose-RFA	2.68	3.89	5.12
C1	HRNet-W32	3.61	4.86	6.11
C2	HRNet-W32-RFA	2.62	3.89	4.93
D1	RSN-50	3.12	3.96	10.17
D2	RSN-50-RFA	2.16	2.81	8.42

4.4 盒式物体表面点位姿测量效果比较

4.4.1 未遮挡情况下的位姿测量

本文选取 HRNet、RSN[25]、YOLO-pose、RTMPose、RTMPose-BRNM、AlphaPose、SPM[26]、PifPaf、SAR-Net[27]和 RBP-pose[28]模型, 在实际场景下, 计算不同模型下平移误差和旋转误差 (结果如表 7 和图 13 所示), 用于衡量盒式物体关键点准确性.

表 7　不同模型实时位姿检测结果

模型	e_t (%)	e_{rot} (°)
HRNet	3.82	2.69
RSN	2.51	1.54
YOLO-pose	2.35	1.41
RTMPose	1.71	1.26
RTMPose-BRNM	1.32	0.96
AlphaPose	2.53	1.54
SPM	2.67	1.89
PifPaf	2.39	1.79
SAR-Net	4.45	2.98
RBP-pose	5.12	3.42

在盒式物体位姿检测实验中, RTMPose-BRNM 展现出最优性能. 传统模型 HRNet 因多分辨率结构对刚性物体约束不足, 误差最高; AlphaPose 则受限于级联优化机制, 精度失衡显著. 轻量化模型中, YOLO-pose 精度优于传统方法, 但仍落后于 RTMPose 系列. SPM 因沙漏结构的误差累积问题, 旋转误差突出, 而 PifPaf 的向量场方法在密集场景中易受局部混淆影响. SAR-Net 通过模板形状对齐和对称点云补全来编码形状先验, 仅用深度点云完成 6D 位姿与三维尺度估计. RBP-Pose 将形状先验融入位姿回归, 使用形状先导残差向量, 造成精度较差.

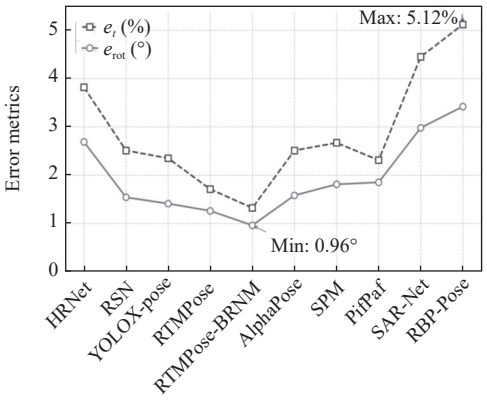


图 13　不同算法实时位姿检测结果

4.4.2 遮挡情况下的位姿测量

在 30% 中等遮挡条件下, 对 8 种主流关键点检测模型和 2 种物体位姿模型进行位姿测量实验. 结果如表 8 所示, 改进后的 RTMPose-BRNM 以 1.94% 的平移误差与 1.46°的旋转误差, 在全部模型中表现最佳.

表 8　不同模型遮挡时位姿检测结果

模型	e_t (%)	e_{rot} (°)
HRNet	7.14	5.65
RSN	6.35	4.17
YOLO-pose	3.47	2.15
RTMPose	3.34	1.88
RTMPose-BRNM	**1.94**	**1.46**
AlphaPose	3.72	2.32
SPM	3.49	2.43
PifPaf	3.43	2.27
SAR-Net	5.23	3.97
RBP-pose	6.92	4.52

整体来看, 具有显式上下文建模能力的模型 RTMPose-BRNM、RTMPose 和 PifPaf, 在 30% 中度遮挡场景下保持较高鲁棒性, 而单纯依赖多分辨率和深堆叠的 HRNet、RSN 精度退化明显, SAR-Net 与 RBP-Pose 在 30% 遮挡导致的可见点云轮廓不完整、关键边缘缺失与多目标近邻干扰的场景中, 两者的先验约束难以发挥.

5 结论

针对盒式物体位姿测量, 本文提出一种精准测量

的流程, 使用 RTMPose-BRNM 盒式关键点检测模型. 首先, 通过将 RTMPose 的 Backbone 网络中传统卷积替换为感受野注意力卷积, 对卷积核的重要性进行动态调节, 解决盒式物体类内遮蔽点拟合错误问题; 引入 NATTEN 模块, 让模型更好捕捉盒式物体轮廓边缘局部特征; 使用 MPC 结构, 加强对不同尺寸的盒式关键点学习. 实验结果表明, 本文的关键点检测算法测量具有盒式物体数量识别稳定和 *MPDE* 值误差较小的优势, 并能根据 EPnP 与 LM 算法精确测量出盒式物体的位姿. 但本文由于实地场景有限, 只能应用有限的盒式物体数据训练, 对于更多情况, 例如不规则的盒式物体, 还需要在未来的研究工作中, 继续增强模型泛化能力, 并在保持精度的同时进一步轻量化, 使模型在算力有限的场景下应用.

参考文献

1 Mousavian A, Eppner C, Fox D. 6-DOF GraspNet: Variational grasp generation for object manipulation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2901–2910.

2 Li XJ, Chen ZY, Chen JQ, *et al*. Automatic characterization of rock mass discontinuities using 3D point clouds. Engineering Geology, 2019, 259: 105131. [doi: 10.1016/j.enggeo.2019.05.008]

3 Lowe DG. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110. [doi: 10.1023/B:VISI.0000029664.99615.94]

4 Rublee E, Rabaud V, Konolige K, *et al*. ORB: An efficient alternative to SIFT or SURF. Proceedings of the 2011 International Conference on Computer Vision. Barcelona: IEEE, 2011. 2564–2571.

5 Bay H, Ess A, Tuytelaars T, *et al*. Speeded-up robust features (SURF). Computer Vision and Image Understanding, 2008, 110(3): 346–359. [doi: 10.1016/j.cviu.2007.09.014]

6 Cao Z, Hidalgo G, Simon T, *et al*. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172–186. [doi: 10.1109/TPAMI.2019.2929257]

7 Kreiss S, Bertoni L, Alahi A. PifPaf: Composite fields for human pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 11969–11978.

8 Cheng BW, Xiao B, Wang JD, *et al*. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 5385–5394.

9 Fang HS, Li JF, Tang HY, *et al*. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7157–7173. [doi: 10.1109/TPAMI.2022.3222784]

10 Wang JD, Sun K, Cheng TH, *et al*. Deep high-resolution representation learning for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(10): 3349–3364. [doi: 10.1109/TPAMI.2020.2983686]

11 Chen YL, Wang ZC, Peng YX, *et al*. Cascaded pyramid network for multi-person pose estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7103–7112.

12 Yang S, Quan ZB, Nie M, *et al*. TransPose: Keypoint localization via Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 11782–11792.

13 Maji D, Nagori S, Mathew M, *et al*. YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 2636–2645.

14 Jiang T, Lu P, Zhang L, *et al*. RTMPose: Real-time multi-person pose estimation based on MMPose. arXiv:2303.07399, 2023.

15 Zhang X, Liu C, Yang DG, *et al*. RFAConv: Innovating spatial attention and standard convolutional operation. arXiv:2304.03198, 2024.

16 Hassani A, Shi H. Dilated neighborhood attention Transformer. arXiv:2209.15001, 2023.

17 Yang JR, Wu SK, Gou LJ, *et al*. SCD: A stacked carton dataset for detection and segmentation. Sensors, 2022, 22(10): 3617. [doi: 10.3390/s22103617]

18 Debapriya M, Soyeb N, Manu M, Deepak P, *et al*. YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 2636–2645.

19 Irshad MZ, Kollar T, Laskey M, *et al*. CenterSnap: Single-shot multi-object 3D shape reconstruction and categorical 6D

pose and size estimation. Proceedings of the 2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022. 10632–10640.

20 Wang G, Manhardt F, Tombari F, *et al*. GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 16606–16616.

21 Chen JR, Kao SH, He H, *et al*. Run, don't walk: Chasing higher FLOPS for faster neural networks. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 12021–12031.

22 Zhang X, Song YZ, Song TT, *et al*. LDConv: Linear deformable convolution for improving convolutional neural networks. Image and Vision Computing, 2024, 149: 105190. [doi: 10.1016/j.imavis.2024.105190]

23 Tan H, Dong SJ. Pixel-level concrete crack segmentation using pyramidal residual network with omni-dimensional dynamic convolution. Processes, 2023, 11(2): 546. [doi: 10.3390/pr11020546]

24 Qiao SY, Chen LC, Yuille A. DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2020. 10208–10219.

25 Cai YH, Wang ZC, Luo ZX, *et al*. Learning delicate local representations for multi-person pose estimation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 455–472.

26 Nie X, Feng J, Zhang J, *et al*. Single-stage multi-person pose machines. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 6950–6959.

27 Lin HT, Liu ZC, Cheang C, *et al*. SAR-net: Shape alignment and recovery network for category-level 6D object pose and size estimation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 6697–6707.

28 Zhang RD, Di Y, Lou ZQ, *et al*. RBP-pose: Residual bounding box projection for category-level pose estimation. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 655–672.

(校对责编: 李慧鑫)