计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(3):187−194 [doi: 10.15888/j.cnki.csa.009444]

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

融合残差注意力和标准偏差的 6D 姿态细化网络①

邓　江, 陈姚节, 张梦杰

(武汉科技大学 计算机科学与技术学院, 武汉 430081)
通信作者: 陈姚节, E-mail: chenyaojie@wust.edu.cn

摘　要: 在 6D 物体姿态估计领域中, 现有算法往往难以实现对目标物体精准且鲁棒的姿态估计. 为解决该问题, 提出了一种结合残差注意力、混合空洞卷积和标准差信息的物体 6D 姿态细化网络. 首先, 在 Gen6D 图片特征提取网络中, 采用混合空洞卷积模块替换传统卷积模块, 以此扩大感受野、加强全局特征捕获能力. 接着, 在 3D 卷积神经网络中, 加入残差注意力模块, 这有助于区分特征通道的重要程度, 进而在提取关键特征的同时, 减少浅层特征的丢失. 最后, 在平均距离损失函数中, 引入了标准差信息, 从而使模型能够区分物体的更多姿态信息. 实验结果显示, 所提出的网络在 LINEMOD 数据集和 GenMOP 数据集上的 *ADD* 指标分别达到了 68.79% 和 56.03%. 与 Gen6D 网络相比, *ADD* 指标分别提升了 1.78 个百分点和 5.64 个百分点, 这一结果验证了所提出的网络能够显著提升 6D 姿态估计的准确性.

关键词: 6D 姿态估计; 混合空洞卷积; 残差注意力; 标准差

引用格式: 邓江,陈姚节,张梦杰.融合残差注意力和标准偏差的 6D 姿态细化网络.计算机系统应用,2024,33(3):187–194. http://www.c-s-a.org.cn/1003-3254/9444.html

6D Pose Refiner Network Combining Residual Attention and Standard Deviation

DENG Jiang, CHEN Yao-Jie, ZHANG Meng-Jie

(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China)

Abstract: In the domain of 6D object pose estimation, existing algorithms often struggle to achieve precise and robust pose estimation of the target objects. To address this challenge, this study introduces an object 6D pose refinement network that incorporates residual attention, hybrid dilated convolution, and standard deviation information. Firstly, in the Gen6D image feature extraction network, traditional convolutional modules are replaced with hybrid dilated convolution modules to expand the receptive field and enhance the capability to capture global features. Subsequently, within the 3D convolutional neural network, a residual attention module is integrated. This assists in distinguishing the importance of feature channels, hence extracting key features while minimizing the loss of shallow-layer features. Finally, the study introduces standard deviation information into the average distance loss function, enabling the model to discern more pose information of the object. Experimental results demonstrate that the proposed network achieves *ADD* scores of 68.79% and 56.03% on the LINEMOD dataset and GenMOP dataset, respectively. Compared to the Gen6D network, there is an improvement of 1.78% and 5.64% in the *ADD* scores, validating the significant enhancement in the accuracy of 6D pose estimation brought about by the proposed network.

Key words: 6D pose estimation; hybrid dilated convolution; residual attention; standard deviation

① 基金项目: 装备发展部“慧眼行动”项目 (62602010214)
收稿时间: 2023-09-21; 修改时间: 2023-10-25; 采用时间: 2023-11-09; csa 在线出版时间: 2024-01-19
CNKI 网络首发时间: 2024-01-22

姿态估计的目的是估计物体在三维空间中的位置和方向, 随着人工智能的不断发展, 机器人操作、自动驾驶、VR 对姿态估计的要求越来越高. 早期的方法如 SIFT[1]和 ORB[2]通过提取具有丰富纹理的物体特征点, 并利用这些 2D 点与对应的 3D 点使用 PnP 算法来估计物体的姿态. 但是这种方法对于无纹理的物体并不能很好地处理, 无法用在工业上. 随着深度学习在计算机视觉取得了巨大的成功, 提出了 PoET[3]、single-stage 6D[4]、PVNet[5]等多种方法. PoET 中可以仅使用 RGB 图片达到精准的估计结果, 但是无法对未经训练的物体进行姿态估计. DeepIM[6]、Latentfusion[7]方法可以对未经训练的物体进行姿态估计, 但是需要高质量的 3D 模型或者深度图片. Objectron[8]虽然不需要 3D 模型但是只能对训练的同类对象进行姿态估计. OnePose[9]提出了一种无模型可推广未训练物体的方法, 但该方法对于弱纹理物体并不能很好地处理. 最近, Gen6D[10]提出一种无模型并且可以估计未经训练物体的方法并且对于弱纹理物体也能取得较好的姿态估计结果. Gen6D 采用的是姿态细化的方法, 也就是获得图像物体中的初始检测结果后对估计的姿态进行细化. Gen6D 通过 2D 卷积网络提取参考图片和查询图片特征然后投影到三维空间中用深层的 3D 卷积网络来估计物体的姿态. 由于参考图片和查询图像极度相似, 为了让网络更好地从参考图片的特征中学习到全局特征信息, 本文在 2D 卷积网络中将 3×3 的卷积模块替换为混合空洞卷积模块, 增大感受野更好地提取全局特征. 同时在 3D 卷积网络中把特征拼接后的卷积模块替换成残差注意力模块, 使网络更加关注空间和通道中的重要特征, 防止了特征的丢失. Gen6D 采用的训练损失函数是根据网络回归得到的信息对预测物体进行相似变换, 通过预测物体的三维点和实际物体相似变换后的三维点的欧几里得距离求平均值来对网络进行反向传播. 但是这样会造成一个问题, 当预测物体是实际物体绕某个顶点旋转一定角度时的损失跟预测物体是实际物体平移一定距离时的损失相同, 但是在旋转情况下 *ADD* 值可能更大. 为了让网络区别上述两种情况的不同, 同时给予旋转情况更大的损失, 本文将每个点的距离标准差加入到损失函数中, 并通过参数控制损失函数对标准差的注重程度.

本文做出如下贡献: (1) 在 2D 卷积网络中将 3×3 的卷积模块替换为混合空洞卷积模块, 这有助于增大感受野并更好地提取图片的全局特征. (2) 在 3D 卷积网络中特征拼接后的卷积模块替换为残差注意力模块, 以便网络更加关注空间和通道中的重要特征, 从而避免特征丢失. (3) 在损失函数中加入了预测点和实际点距离差的标准差信息, 让模型可以区别更多情况, 增加模型鲁棒性.

1 相关工作

1.1 依赖 3D 先验模型的姿态估计

许多方法利用已知的物体 3D 模型进行姿态估计. 例如 CDPN[11]通过 PnP 从坐标间接求解旋转, 从图像中估计平移. DPOD[12]通过构建 2D-3D 对应关系, 然后使用 PnP 求解姿态. 但是这些方法需要为每个对象训练一个单独的网络. 还有一些方法可以学习类别内共享的形状先验, 这样在测试时可以不需要同一类别中的 3D 模型[13–15], 但这类方法在处理未知类别的物体时显示出其局限性. 近期的一些研究尝试利用 2D 特征匹配的泛化能力对未经训练的物体进行姿态估计, 使用 3D 模型渲染出参考图像, 然后通过稀疏关键点匹配[16]或是 2D 到 2D 的像素对应关系[17]来与查询图像匹配. 然而上述方法都高度依赖于高质量的有纹理的 3D 模型进行训练或渲染, 这在实际应用场景中是很难获取的. 本文的方法是使用检测器和选择器来初始化查询图像的位姿, 然后通过回归位姿残差对其进行细化, 不需要 3D 模型.

1.2 不依赖 3D 先验模型的姿态估计

最近提出了一些方法不需要 3D 模型, RLLG[18]使用图像对之间的对应关系作为训练的监督, 使用 PnP 和 RANSAC 来确定 6D 姿态. 虽然不需要已知的对象模型, 但它只能对实例对象进行处理, 并且需要高度精确的实例掩码来分割前景像素. Ahmadyan 等人[8]提出了一种数据驱动的方法, 该方法使用大量带注释的训练数据学习回归每个类别的投影框角的像素坐标. 这种方法由于学习的模型是特定于类别的, 因此只能局限于少数类别. 但是这些方法不能推广到未训练的对象. OnePose[9]从所有支持视点的 RGB 序列中重建稀疏点云, 并将目标视图与稀疏点云进行匹配来求解姿态, 可以推广到未经训练的物体, 但当处理弱纹理对象时由于局部纹理特征不明显, 无法进行准确的姿态估计. 本文方法对于不同强弱纹理的物体都具有良好的泛化性.

1.3 Gen6D

Gen6D[10]是一种可推广的无模型 6D 物体姿态估计器. Gen6D 只需要未经训练的物体的一些姿态图像, 并且能够准确地预测任意环境下物体的姿态. Gen6D 由一个对象检测器、一个视点选择器和一个姿态细化器组成, 所有这些都不需要 3D 对象模型, 并且可以泛化到看不见的对象. 其中检测器的原理是在查询图像的不同位置、不同尺度上与查询图像进行相关性比较, 通过将相关性由分数表示出来, 分数最大的地方就是检测出来的物体位置与尺度. 选择器的主要作用是进行图像匹配, 选择器将查询图像与每个参考图像进行比较, 计算相似度分数. 选择器首先对参考图像和查询图像应用 VGG-11[19]提取特征映射. 然后对每个参考图像的特征图计算其与查询图像的特征图的元素积, 以生成相关评分图. 最后通过相似度网络处理相关分数图, 生成相似度分数和相对平面内旋转, 使查询图像与参考图像对齐, 这样可以得到一个初始的目标姿态, 然后通过细化器来获得更准确的姿态. 细化器首先选择 6 张靠近输入姿态的参考图像, 通过 2D 卷积在这些选定的参考图像上提取特征映射, 提取多尺度二维特征后进行特种融合. 然后将这些融合后的特征映射通过线性插值到三维体积中, 并计算所有参考图像之间的特征均值和方差作为三维体积的特征. 对于查询图像也使用相同的 2D 卷积提取其特征再通过线性插值映射到三维体积中, 并将插值后的查询特征与参考图像特征的均值和方差进行拼接. 最后对拼接特征应用 3D 卷积网络模型来预测姿态, 并通过相似变换更新物体预测的姿态, 继续对更新后的姿态进行优化, Gen6D 中默认应用迭代器 3 次.

2 改进的 Gen6D 细化器网络

本文的网络结构图如图 1 所示, 本文对于图片的特征提取部分将常规的卷积模块替换为 HDC (hybrid dilated convolution) 模块, 增大了感受野, 强化了全局特征. 考虑到在深层网络中可能造成信息丢失以及让模型在空间和通道中更加关注重要的信息. 在对将查询图片特征和参考图片均值、参考图片方差拼接后的回归网络中加入 RA (residual attention) 模块. 在 RA 模块中的第 1 个卷积层后加入通道注意力, 第 2 个卷积层加入空间注意力, 然后将该部分的结果与输入该部分的特征相加. 最后将该网络的输出通过线性层回归到四元数旋转 (q), 二维平移 (t) 和尺度因子 (s).

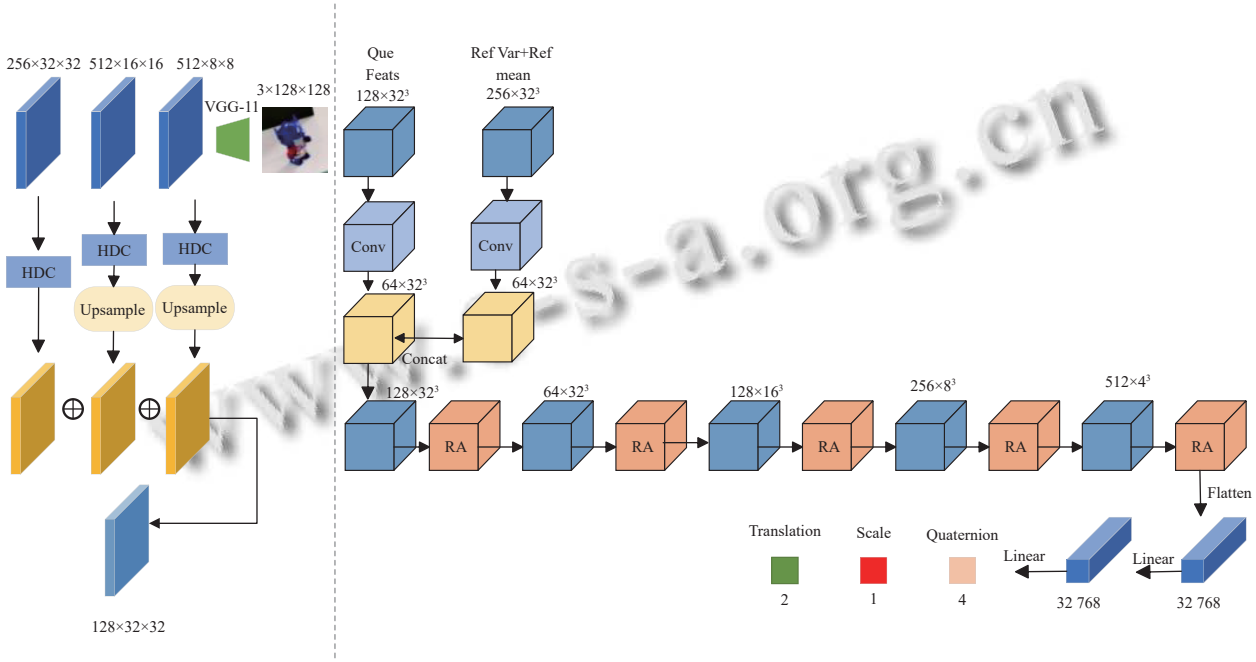


图 1 6D 姿态细化器网络架构

2.1 混合空洞卷积

空洞卷积是一种独特的卷积方法, 通过在卷积核中引入“空洞”来增大其感受野. 这种操作的扩张率可以由参数 a 来控制. 如图 2 所示, 当 a=1 时, 空洞卷积与常规卷积无异. 但是, 当 a 的值超过 1 时, 会有一些中间像素不参与计算, 这可能会降低网络的表现. 因此,

本文引入了混合空洞卷积 (HDC) 来替代传统空洞卷积, 并将其嵌入到二维卷积神经网络的原始卷积模块中. 这种混合空洞卷积增大了感受野, 可以网络捕获更广泛的上下文信息. 通过不同的扩张率, 实现了多尺度特征的捕获, 增强了全局和局部特征之间的协同作用. 特别是在估计弱纹理物体的姿态的时候, 这类物体主要依赖于其整体结构而非局部特征. 由于经过特征提取后的特征图尺寸较小, 空洞率过大反而会造成卷积核超过特征图边界从而导致无法捕捉边缘的信息. 因此在本文设计的 HDC 模块中, 混合卷积的扩张率为 1 和 2, 并通过 1×1 的卷积在最后进行通道数调整, 这样不仅能够有效地增大网络的感受野, 捕捉到更多的上下文信息, 而且不会导致边缘信息损失, 确保模型能够充分地利用特征图上的所有信息.

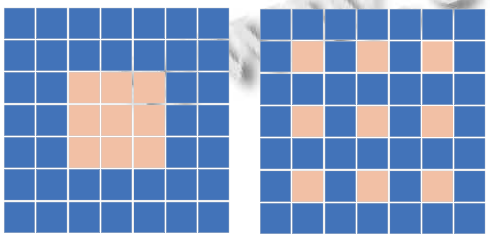


图 2 空洞卷积示意图

2.2 RA 模块

实践表明网络层数不断加深可能会导致网络退化等问题. 针对上述问题, He 等人[20]加入跳跃连接, 即残差学习框架 (ResNet), 将浅层特征与深层特征相加共同构成后续操作的输入, 有效缓解了输入特征的丢失, 提高了特征恢复的性能. 本文在 3D 卷积网络中加入残差连接, 减少浅层特征的损失. 注意力机制在 2D 卷积网络中得到了广泛的应用[21], 本文将其引入到 3D 卷积网络中, 在两个卷积层后分别加入通道注意力 (CA) 和空间注意力 (SA). 这样可以更高效地利用两种注意力机制分别对空间关系和通道关系进行建模. 空间注意力机制可以使模型专注于复杂纹理的物体上的重要细节, 为这些细节赋予更高的权重, 从而确保在姿态预测时充分考虑这些关键信息. 同时, 考虑到局部信息可能不足以描述物体的所有纹理特点, 通道注意力允许模型对于各个通道的信息进行动态加权, 从而更加聚焦于那些对当前任务更具区分性的特征通道. 这样模型可以在一个更广的上下文中, 区分出纹理间的微妙差异, 从而提高姿态估计的准确性和鲁棒性. 本文设计的 RA 模块结构图如图 3 所示, 一个残差注意力结构中包含两个 3×3×3 卷积层和一个跳跃连接的 1×1×1 卷积, 最后将两个分支相加实现特征融合, 减少特征损失的同时进一步解决网络退化问题.

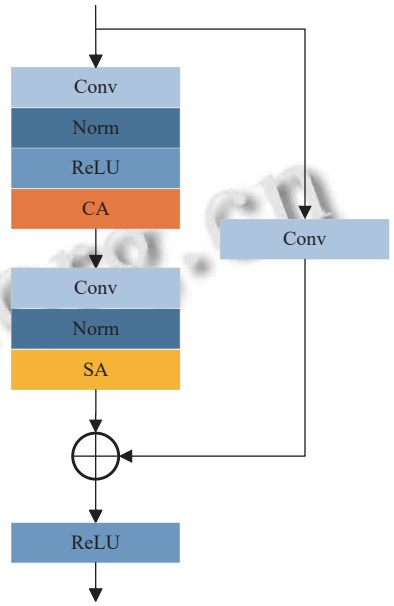


图 3 RA 模块结构图

通道空间注意力模块中首要步骤是通过自适应平均池化和最大池化对输入的空间信息进行全局压缩. 输入一个空间特征图 $F \in R^{C\times D\times H\times W}$, 使用平均池化和最大池化进行空间信息的聚合, 得到平均池化后的空间特征图 F_{avg} 和最大池化后的 F_{max}. 将得到的空间特征图输入到共享网络, 该共享网络由一个隐含层的感知机构成, 输出层的大小为 1×1×1×C. 将 2 个新得到的空间特征图相加经过 Sigmoid 激活函数得到权重系数 $M \in R^{1\times 1\times 1\times C}$. 最后, 将得到的通道注意力权重与原始的输入 x 做逐元素乘法操作, 使输入 x 的每个通道都乘以对应的通道注意力权重. 计算方式如下:

$$M_c(F)=\sigma(MLP(AvgPool(F)))+MLP(MaxPool(F)) \tag{1}$$

其中, σ 为 Sigmoid 激活函数.

在空间注意力模块中, 首要步骤是计算输入 x 在通道维度上的最大值和平均值. 通过最大池化和平均池化, 得到了每个空间位置在所有通道上的最大和平均特征 $F^s_{\text{max}} \in R^{1\times D\times H\times W}$ 和 $F^s_{\text{avg}} \in R^{1\times D\times H\times W}$, 接着将最大响应和平均响应在通道维度上进行拼接, 形成一个新的特征图. 然后使用一个卷积核为 3×3×3 的卷积核对特征图进行卷积操作. 这个卷积操作可以看作是对

最大响应和平均响应进行了一个线性组合. 接下来通过一个 Sigmoid 激活函数将卷积结果映射到 0–1 之间, 这样得到的值可以看作是三维空间特征图 $M^s \in R^{H\times W\times D}$. 最后将得到的空间注意力权重与原始输入 x 做逐元素乘法, 得到基于空间注意力的新特征. 计算方式如下:

$$M_s(F)=\sigma(Conv_{3\times3\times3}([AvgPool(F);MaxPool(F)])) \tag{2}$$

其中, σ 为 Sigmoid 激活函数, $Conv_{3\times3\times3}$ 为卷积核为 3×3×3 的标准卷积层.

2.3 标准差损失函数

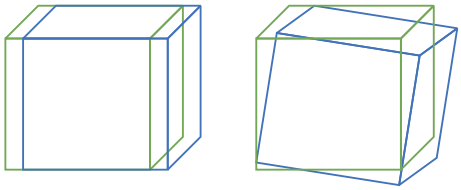
Gen6D 通过将输入姿态下对象坐标系中的 32^3 个体素点转换至相机坐标系来进行工作. 接着, 它计算了预测姿态与实际姿态下体素点的平均距离差异. 但这种方法在某些特定情况下可能会失真, 导致预测准确度的下降.

如图 4 所示, 绿色框表示真实的姿态外接矩形, 而蓝色框表示模型预测的姿态外接矩形. 在图 4(a) 情况, 预测框相对于实际框有一定的平移; 而在图 4(b) 情况, 预测框相对于实际框绕某一顶点旋转了特定角度. 尽管这两种情境下的损失相同, 但其 *ADD* 可能会有所不同. 特别是当图 4(a) 情况的平移距离恰好在物体直径的 10% 以内时. 由于使用的损失函数是基于均匀采样的体素点, 而 *ADD* 指标计算的点云分布并非均匀, 这可能导致在图 4(b) 情况中部分密集的点云差值超过 10%, 使得总体评估指标下降. 为了解决这一问题, 本文对原有的损失函数进行了修改, 引入了体素点距离差的标准差作为一个指标, 并通过参数来控制其影响力度. 这一改进帮助模型在训练过程中避免上述特殊情况, 从而增强了网络的鲁棒性, 并提高了物体姿态预测的准确性. 具体的改进公式如下:

$$D=\frac{1}{32^3}\sum_{k}^{32^3}\left\|s_pR_p\left(p_k+t'_p\right)-s_gR_g\left(p_k+t'_g\right)\right\|_2 \tag{3}$$

$$L=D+\alpha\sqrt{\frac{\sum_{k}^{32^3}(D_k-D)^2}{32^3}} \tag{4}$$

其中, p_k 为输入相机坐标系中样本点的坐标, s_p 和 s_g 分别为预测尺度和真实尺度, R_p 和 R_g 分别为预测的旋转和真实的旋转, t'_p 和 t'_g 分别代表预测的平移和真实的平移. D 为体素点的平均距离, D_k 为体素点的距离.



(a) 情况 1 (b) 情况 2

图 4 特殊情况示意图

3 实验结果与分析

本实验环境为 Ubuntu 22.04 操作系统, Nvidia RTX4080 显卡、32 GB 内存. 输入图片的大小为 128×128. 设置 300 000 次迭代, 使用 Adam 优化器进行优化, 初始学习率为 0.000 01, 每训练 100 000 次学习率下降 0.5. 实验中损失函数的参数 a=0.2. 本文实验结果迭代细化器 3 次.

3.1 数据集

本文在 LINEMOD[22]、GenMOP[10]、ShapeNet[23] 模型渲染的图片以及 Google 扫描对象数据集由 IBRNet[24]渲染的图片对细化器进行训练. 并在 LINEMOD 和 GenMOP 数据集上对模型进行评估. LINEMOD 数据集由 15 783 幅图像组成, 其中包含 13 个弱纹理对象, 每个对象包含约 1 400 幅图像. GenMOP 数据集由 10 个物体组成, 从平坦的物体 (如“剪刀”) 到薄结构物体 (如“椅子”). 对于每个物体, 在不同的背景、光照条件等环境下采集同一物体的两个视频序列. 每个视频序列被分成 200 个图像.

3.2 评估指标

为了评价模型的性能, 本文使用 2D 投影指标 Proj-2d[25]和点平均距离指标 *ADD*[6]分别对模型进行评估.

Proj-2d 指标计算的是预测的姿态投影的点与真实标注姿态投影的点之间的平均距离. 当平均距离小于 5 个像素的时候, 模型估计的姿态被认为是正确的. 相关定义如下:

$$\text{Proj-2d}=\frac{1}{m}\sum_{x\in M}\left\|K(Rx+T)-K\left(R_px-T_p\right)\right\| \tag{5}$$

其中, K 为摄像机的参数矩阵; M 为目标物体的顶点合集; m 代表顶点的个数; R、T 分别代表真实的旋转和平移; R_p 和 T_p 分别代表预测的旋转和平移.

ADD 指标指当预测的点云与实际的点云差值小于物体直径的 10% 时, 该指标认为估计的旋转矩阵是正确的. 相关定义如下:

$$ADD=\frac{1}{m}\sum_{x\in M}\left\|(Rx+T)-\left(R_px+T_p\right)\right\| \tag{6}$$

3.3 实验结果分析

在 GenMOP 数据集实验中本文选择了基于特定示例的估计器 PVNet 和可泛化模型 Gen6D 和 OnePose 与本文修改后的细化器在 GenMOP 通过 *ADD* 和 Proj-2d 评价指标做实验对比, 其中对于可泛化的模型 OnePose, 使用 YOLOv5 作为检测器, 实验结果如表 1 所示. 本文从实验结果中选择部分物体进行结果可视化展示如图 5 所示, 其中绿色为真实姿态外界矩形, 蓝色姿态为模型预测姿态外界矩阵. 从表 1 中可以看出本文改进后的网络性能是优于 Gen6D 的, Proj-2d 指标平均提升了 1.31 个百分点, *ADD* 指标平均提高了 5.64 个百分点. Tformer 和 Piggy 都拥有丰富的纹理 *ADD* 分别提升了 5.95 个百分点和 3.01 个百分点, 这是因为通过残差注意力机制让细化器网络保留了浅层特征更加注重细节特征, 提高了模型的预测准确度. 其中 Scissors 相较于 Gen6D 的提升最大, 提升了 16.8 个百分点. 这是由于改进前的损失函数会导致模型预测这种长方形物体更易产生误差. 从图 5 中可以看到 Gen6D 中的 Scissors 实际预测框相当于绕某一点旋转了一定角度, 这正是本文提到的特殊情况, 而改进后的方法很好地解决了这个问题. OnePose 虽然对这些未经训练的物体取得了不错的姿态估计结果, 但是对于弱纹理物体如 PlugEN、Scissors 与本文方法有较大差距. 本文的网络也在所有物体上优于 PVNet, 这是由于对于 PVNet 来说训练数据过少并不能满足网络训练到最好的程度.

表 1 在 GenMOP 数据集上各方法对比 (%)

Metrics	Method	Chair	PlugEN	Piggy	Scissors	Tformer	Avg.
ADD	PVNet[5]	49.50	2.33	77.89	44.40	19.84	38.79
	OnePose[11]	52.00	4.67	**79.89**	26.29	57.53	44.08
	Gen6D[12]	61.50	19.63	75.38	32.76	62.7	50.39
	Ours	**63.00**	**20.56**	78.39	**49.56**	**68.65**	**56.03**
Proj-2d	PVNet[5]	15.00	30.37	83.42	96.55	59.52	56.97
	OnePose[11]	**58.00**	68.22	93.97	77.58	86.50	76.85
	Gen6D[12]	55.00	72.90	92.96	93.53	98.81	82.64
	Ours	53.50	**73.83**	**94.97**	**97.84**	**99.60**	**83.95**

在 LINEMOD 数据集实验中本文选择特定实例姿态估计器 PVNet、OnePose、Gen6D 和本文修改后的模型做实验对比, 实验结果如表 2 所示, 其中 PVNet 在真实数据上进行训练. 同时选择部分物体进行可视化展示如图 6 所示, 可以看到本文改进后的方法在这些纹理较弱的物体上相比 Gen6D 更加接近真实结果. 从表 2 可以看出本文优化后的网络相比 Gen6D 在 *ADD* 指标上平均提高了 1.78 个百分点, Proj-2d 指标平均提高了 2.33 个百分点. 在 LINEMOD 数据集中, 绝大多数物体呈现出较弱的纹理特点. 是由于空洞卷积增大了感受野, 增强了物体的全局特征, 有助于模型从参考姿态中学习到查询图片的姿态信息. 其中 cat 和 lamp 提升较高是由于改进后的网络对于这种长方体物体效果提升较为显著. driller 由于本身颜色很暗, 在部分弱光环境下本文改进后的网络仍然无法准确地识别他们的姿态. OnePose 平均 *ADD* 指标与本文改进后的模型相接近, 但是对于弱纹理物体 cat、duck 并不能很好的处理. 本文模型相比 PVNet 仍有很大的差距, 这是因为 PVNet 是用 LINEMOD 中充足的真实数据进行训练的, 而本文模型在训练的时候并没有训练过这些数据所以对这些物体的深度估计并不是很准确.

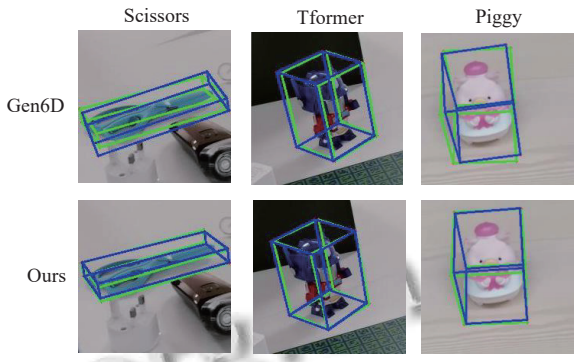


图 5 GenMOP 可视化示意图

表 2 在 LINEMOD 数据集上各方法对比 (%)

Metrics	Method	Cat	Duck	Bvise	Cam	Driller	Lamp	Avg.
ADD	PVNet[5]	**79.34**	**52.58**	**99.90**	**86.86**	**96.43**	**99.33**	**85.74**
	OnePose[11]	45.81	33.99	88.95	85.09	71.26	85.32	68.40
	Gen6D[12]	60.68	40.47	77.03	66.67	67.39	89.83	67.01
	Ours	63.27	41.97	78.49	67.94	68.38	92.71	68.79
Proj-2d	PVNet[5]	99.30	98.02	99.81	99.21	96.43	99.33	98.68
	OnePose[11]	75.64	71.17	92.63	94.61	75.02	87.52	82.77
	Gen6D[12]	95.81	80.85	81.89	90.88	73.34	91.17	85.66
	Ours	97.90	82.72	83.13	93.23	75.22	95.68	87.98

3.4 运行时间

处理尺寸为 540×960 的图像, 本文改进后的网络总共花费约 0.12 s, 其中目标检测器花费约 0.05 s, 视点选择器花费约 0.01 s, 优化 1 次的细化器花费约 0.07 s.

3.5 损失函数参数对比实验

为了探究在损失函数中引入标准差信息后不同参

数的影响程度, 本文以 Gen6D 为基线模型设计了一系列对比实验, 旨在明确不同参数值如何影响模型的性能, 实验结果如表 3 所示.

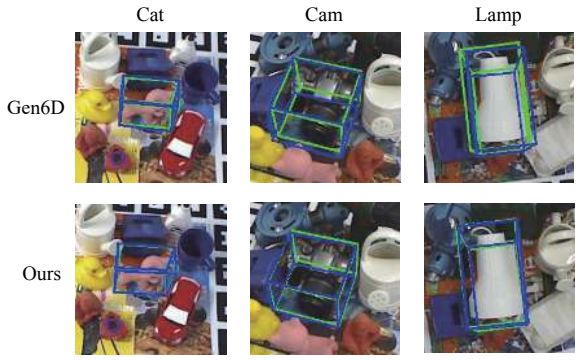


图 6 LINEMOD 可视化示意图

表 3 损失函数参数对比结果

a	Proj-2d (Avg.) (%)		*ADD* (Avg.) (%)	
	LINEMOD	GenMOP	LINEMOD	GenMOP
0	85.67	82.64	50.39	67.01
0.1	85.85	82.84	52.12	67.51
0.2	86.72	83.08	53.86	67.94
0.3	86.13	82.96	51.38	67.29
0.4	85.78	82.42	50.12	67.12
0.5	85.19	80.97	48.12	66.38

从表 3 中可观察到, 当参数 a=0.2 时, 模型的性能达到了最佳表现. 在$a \leqslant 0.3$的范围内, 两个不同数据集上 *ADD* 指标和 Proj-2d 指标相较于原本的损失函数都有一定提升, 证明引入的标准差信息确实提高了模型估计姿态的准确性, 在此参数范围内, 模型能够在训练过程中充分考虑到标准差信息, 从而避免了之前提到的特定情况. 然而, 当 a>0.4 时, 损失函数对标准差的重视程度过高, 使模型对标准差的关注超出了合理范围. 这进一步导致模型对平均距离的关注度下降, 从而影响了模型的整体准确度.

3.6 消融实验

为了证明对细化器各个改进模块的有效性在 LINEMOD 和 GenMOP 数据集上设计了消融实验, 实验验证如表 4 所示. 第 1 行是原网络的性能结果, √代表加入该改进模块后的网络模型.

从表 4 可以看到加入空洞卷积后 LINEMOD 的提升相对 GenMOP 的提升更显著, 这是因为空洞卷积能够扩大特征图的感受野, 从而捕获更丰富的上下文信息. 对于 LINEMOD 中的弱纹理物体, 上下文信息尤其重要, 因为这些物体更加依赖于全局特征来进行区分. 而加入 RA 模块后 GenMOP 的提升更为明显, 这是因为 GenMOP 中的大部分物体纹理较多, 而 RA 模块中的注意力机制可以帮助模型更加注重细节特征和复杂纹理的相关特征. 加入修改后的损失函数后, 由于通过在原有的损失函数中加入标准差信息使模型解决了之前无法区分特殊情况下两种不同姿态相同损失的问题, 两种数据集所有指标有了较为明显的提升, 证明了本文改进的有效性.

表 4 消融实验结果 (%)

RA	HDC	Standard deviation loss	Proj-2d (Avg.)		*ADD* (Avg.)	
			LINEMOD	GenMOP	LINEMOD	GenMOP
—	—	—	85.67	82.64	67.01	50.39
√	—	—	85.88	82.94	67.65	52.69
—	√	—	85.95	82.72	67.82	51.86
—	—	√	86.72	83.08	67.94	53.86
√	√	—	86.54	83.16	68.19	53.56
√	√	√	87.98	83.95	68.79	56.03

4 结论与展望

针对目前算法难以实现对物体进行精准且鲁棒的姿态估计的问题, 在 Gen6D 细化器的基础上本文提出了一种融合残差注意力的标准偏差的姿态优化模型方法, 该方法在对图片特征提取部分融合了混合空洞卷积, 在 3DCNN 网络中加入了残差注意力模块. 这一设计有助于减少网络中浅层特征的丢失, 从而使得提取的特征更加全面和丰富. 同时, 为了确保模型能够准确区分特殊情况下的姿态信息, 本文在损失函数中加入了标准差信息. 这一改进有助于提高模型对于不同姿态变化的判别能力. 实验结果证明本文改进后的网络在弱纹理物体和丰富纹理物体上相较原网络估计的姿态均有提升, 本文改进后的细化器与 Gen6D 中初始的细化器相比在 LINEMOD 数据集上 *ADD* 指标提高了 1.78 个百分点, GenMOP 数据集中 *ADD* 指标提高了 5.64 个百分点, 显著提升了估计物体 6D 姿态的准确性. 但是在研究过程中还存在一些问题, 例如在弱光线环境下不能精确地估计物体姿态. 因此, 后续研究的重点将会针对图片特征提取阶段加入针对弱光环境下更为鲁棒的算法.

参考文献

1 Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 2003, 31(13): 3812–3814. [doi: 10.1093/nar/gkg509]

2 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.
3 Jantos T, Hamdad MA, Granig W, *et al*. PoET: Pose estimation Transformer for single-view, multi-object 6D pose estimation. Proceedings of the 6th Conference on Robot Learning. Auckland: PMLR, 2023. 1060–1070.
4 Hu YL, Fua P, Wang W, *et al*. Single-stage 6D object pose estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 2927–2936.
5 Peng SD, Liu Y, Huang QX, *et al*. PVNet: Pixel-wise voting network for 6DoF pose estimation. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4556–4565.
6 Li Y, Wang G, Ji XY, *et al*. DeepIM: Deep iterative matching for 6D pose estimation. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 695–711.
7 Park K, Mousavian A, Xiang Y, *et al*. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10707–10716.
8 Ahmadyan A, Zhang LK, Ablavatski A, *et al*. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7818–7827.
9 Sun JM, Wang ZH, Zhang SY, *et al*. OnePose: One-shot object pose estimation without CAD models. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 6815–6824.
10 Liu Y, Wen YL, Peng SD, *et al*. Gen6D: Generalizable model-free 6-DoF object pose estimation from RGB images. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 298–315.
11 Li ZG, Wang G, Ji XY. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7677–7686.
12 Zakharov S, Shugurov I, Ilic S. DPOD: 6D pose object detector and refiner. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 1941–1950.
13 Chen X, Dong ZJ, Song J, *et al*. Category level object pose estimation via neural analysis-by-synthesis. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 139–156.
14 Lee T, Lee BU, Kim M, *et al*. Category-level metric scale object shape and pose estimation. IEEE Robotics and Automation Letters, 2021, 6(4): 8575–8582. [doi: 10.1109/LRA.2021.3110538]
15 Wang JZ, Chen K, Dou Q. Category-level 6D object pose estimation via cascaded relation and recurrent reconstruction networks. Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021. 4807–4814.
16 Zhong CL, Yang C, Sun FC, *et al*. Sim2Real object-centric keypoint detection and description. Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI, 2022. 5440–5449.
17 Shugurov I, Li F, Busam B, *et al*. OSOP: A multi-stage one shot object pose estimation framework. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 6825–6834.
18 Cai M, Reid I. Reconstruct locally, localize globally: A model free method for object pose estimation. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 3150–3160.
19 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2014.
20 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.
21 王昕, 董琴, 杨国宇. 基于优化 CBAM 改进 YOLOv5 的农作物病虫害识别. 计算机系统应用, 2023, 32(7): 261–268. [doi: 10.15888/j.cnki.csa.009175]
22 Hinterstoisser S, Lepetit V, Ilic S, *et al*. Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. Proceedings of the 11th Asian Conference on Computer Vision. Daejeon: Springer, 2013. 548–562.
23 Chang AX, Funkhouser T, Guibas L, *et al*. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012, 2015.
24 Wang QQ, Wang ZC, Genova K, *et al*. IBRNet: Learning multi-view image-based rendering. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 4688–4697.
25 Brachmann E, Michel F, Krull A, *et al*. Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 3364–3372.

(校对责编: 牛欣悦)