文章编号:1671-4598(2025)11-0236-08 DOI:10.16526/j.cnki.11-4762/tp.2025.11.028 中图分类号:TP391 文献标识码:A

基于 YOLO 模型的目标位姿估计

董伟嗣[1], 毛 钪[2], 付东翔[2]
(1. 华为技术有限公司 苏州研究所, 苏州 215128;
2. 上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘要: 文章针对共轴双旋翼无人机抓取应用中的目标位姿估计方法展开研究, 提出基于 CDPN 的无人机目标位姿估计模型 EPRO-CDPN, 其前置目标检测器采用改进的 YOLO 网络算法, 提高其目标检测能力; 引入注意力机制, 使模型关注关键特征信息, 加强网络训练过程中通道间的特征融合; 引入 EPRO-PnP 替换原来的传统 PnP 方法, 将传统求解转化为对位姿概率分布的预测; 整个位姿估计网络实现为一个端到端的网络; 位姿估计网络模型在公开数据集 LineMod、自制数据集上进行了算法性能测试, 并以共轴双旋翼无人机为物体目标进行抓取实验, 验证姿态估计算法的可行性和有效性; 检测精度在 95%以上, 检测速度快, FPS 达到 35.2 帧/秒, 实现了实时目标姿态估计, 为视觉引导的机械臂自动抓取回收共轴无人机奠定了研究和实验方案。

关键词: 深度学习; 目标检测; 位姿估计; 共轴双旋翼无人机; CDPN; 机械臂抓取

Pose Estimation for Targets Based on Yolo Model

DONG Weisi[1], MAO Kang[2], FU Dongxiang[2]
(1. Huawei Technologies Co., Ltd., Suzhou Research Institute, Suzhou 215128, China;
2. School of Optical-Electrical and Computer Engineering University of Shanghai
for Science and Technology, Shanghai 200093, China)

Abstract: Research on target pose estimation methods for co-axial dual-rotor drone grasping applications is conducted, a CDPN-based drone target pose estimation model, EPRO-CDPN, is proposed. The front-end target detector employs an improved YOLO network algorithm to enhance its target detection capability. An attention mechanism is introduced to enable the model to focus on critical feature information, strengthening the feature fusion between channels during the network training process. The traditional PnP method is replaced with the EPRO-PnP, transforming the conventional solving process into the prediction of pose probability distribution. The entire pose estimation network is implemented as an end-to-end network. The performance testing of pose estimation network model is carried out on the public dataset LineMod and a self-created dataset, and grasping experiments with the co-axial dual-rotor drone as the target object verify the feasibility and effectiveness of the pose estimation algorithm. The detection accuracy reaches over 95%, with a fast detection speed, achieving 35.2 frames per second, achieving a real-time target pose estimation, and laying research and experimental schemes for visually guided robotic arms to automatically grasp and recover co-axial drones.

Keywords: deep learning, object detection, pose estimation, coaxial dual-rotor drone, CDPN, robotic arm grasping

0 引言

物体姿态估计是计算机视觉领域的一项关键任务[1], 其目标是准确获得真实场景中物体姿态的 6DoF[2] (6 个自由度: 3 个自由度的旋转和 3 个自由度的平移)。目前其在自动驾驶、机器人操作等领域应用广泛。随着无人机技术的不断发展和成熟, 在军事、民用等各领域应用广泛。在军事领域, 从当前的俄乌战场应用反馈来看, 小型无人机在作战中表现优异, 从战场情报收集、对敌攻击等战果显著。在民用领域, 民用无人机在快递运输、农业喷洒等领域已经开始应用并快速增长。随着低空经济的不断发展, 无人机应用规模的不断扩大, 数量会快速上升, 蜂群式应用会是未来的发展方向之一, 即大量的无人机集群应用, 这些小型无人机

收稿日期:2025-05-06; **修回日期**:2025-06-16。
作者简介:董伟嗣(1974-),男,硕士,高级工程师。
通讯作者:付东翔(1971-),男,博士,副教授。
引用格式:董伟嗣,毛 钪,付东翔. 基于 YOLO 模型的目标位姿估计[J]. 计算机测量与控制,2025,33(11):236-243,251.

的起飞、载荷装填、回收降落等如果还是按照目前人工干预、手工操作的模式，将导致效率低下，难以满足无人机大规模的应用需求。视觉引导的机械臂系统智能抓取系统可以高效地完成上述无人机应用中的自动抓取回收降落、载荷装填等任务，极大提升工作效率。目前还未见有视觉引导的智能回收系统应用报道。

在上述视觉引导的机械臂智能抓取目标系统中，目标物体实时位姿估计是关键技术之一。6D 位姿估计是给出相机坐标系与目标物体坐标系之间的转换关系[3]。

随着视觉相机硬件技术的不断发展，视觉测量采集数据有 RGB 图像、RGB-D 深度图、点云等形式[4]。根据视觉采集输出数据类型不同，相应位姿估计方法也不同[5]。传统的物体姿态估计方法多依赖于几何推理和特征匹配等技术[6]，具有较强的数学基础和可解释性，但在面对复杂场景下，如遮挡、低纹理区域或多样化的物体形状时，其检测结果表现较差。随着深度学习技术的快速发展，现代的姿态估计方法已逐渐转向基于 CNN 和其它深度模型的端到端的学习方法，这些方法在复杂和多变的环境中有更强的适应性。常见基于深度网络的位姿估计方法如下。

基于 RGB 图像的位姿估计：由于缺少直接的三维几何信息与深度信息，只依靠二维视觉数据来学习并补充，因此该领域的研究极具挑战性，清华大学的 CDPN 深度模型方法给出基于单幅 RGB 图像的位姿求解[7]，创新性提出间接法与直接法对结合地对目标物体进行位姿估计。Lepetit 等人提出了 BB8[8]位姿估计算法，是一个基于深度学习的两阶段算法，专注于单一的 RGB 图像，通过检测和分割 RGB 图像中的物体获得感兴趣的目标，进一步预测其 3D 边界框在 2D 图像中的角点，再结合 PNP 算法来估计物体的 6D 姿态。

基于 RGB-D 深度图的位姿估计：深度图像会提供场景的几何距离信息，每个像素值代表相机到场景中各个点的距离。Chen 提出了 G2L-Net，网络从深度图获取粗糙点云，根据提取到的点云信息估计目标物体的位姿[9]。来自港科大、深大与旷视研究院合作的工作 PVN3D 将基于 2D 关键点的方法引入到 3D 位姿估计中，以深度图像为输入，估计物体的 6-DoF 信息[10]。

基于点云的位姿估计：点云数据包含三维空间信息，适用于精确的目标建模和识别，Gualtieri 等研究通过点云数据，在复杂环境中识别高精度抓握位姿[11]。基于点云的 PointNet 模型[12]及其后续工作[13]则最早将 CNN 卷积应用在点云检测领域，是 Point-base 的经典模型，该方法直接对点云进行处理以减少位置信息损失，但计算量庞大，难以实现实时检测。后续提出的 Voxel-base 的模型在推理速度上能有所提升，该模型使用 Voxelization（体素化）来处理点云，将三维空间划分成无数个相同的小立方体，每个立方体即为 voxel（体素）。通过将这些体素组织成 3 维数组的形式来进行处理，再使用 3D 以及 2D 卷积网络进行分析处理，PointPillars[14]是 Voxel-base 的经典模型之一，将点云分为柱体 pillars，再对 pillars 进行特征提取，最后映射到二维图像，在速度与精度上有所改善，但模型依旧使用了三维卷积，算法检测速度不高。

在共轴双旋翼抓取姿态估计应用中，上述各类模型方法在具体应用场景中往往无法直接采用，需要根据目标的检测精度和速度等要求，设计相应的深度模型算法，通过数据集训练与测试验证模型方法的可行性和有效性。

本文针对共轴双旋翼无人机抓取应用中的目标位姿估计方法展开研究，提出基于 CDPN 的无人机目标位姿估计模型 EPRO-CDPN，其前置目标检测器采用改进的 YOLO 网络算法，提高其目标检测能力；引入注意力机制，使模型关注关键特征信息，加强网络训练过程中通道间的特征融合；引入 EPRO-PnP 替换原来的传统 PnP 方法，将传统求解转化为对位姿概率分布的预测。整个位姿估计网络实现为一个端到端的网络。位姿估计网络模型在 LineMod 公开数据集上进行训练和测试，并在自制数据集上测试和验证模型的可行性。最后将姿态估计模型部署与边缘计算平台进行共轴双旋翼无人机的抓取，验证模型算法的可行性和有效性。

1　位姿估计原理与模型

1.1　位姿估计定义

在计算机视觉中，6D 位姿估计是求取视觉相机与目标物体坐标系之间的转换关系，包含 3 个平移自由度（沿 X、Y、Z 轴的位移）和 3 个旋转自由度（绕 X、Y、Z 轴的欧拉角或旋转角），6 个信息量表示物体的位置和姿态[15]，如图 1 所示。

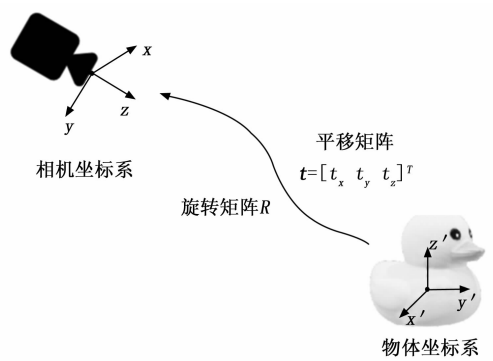

图 1　相机坐标系与物体坐标系之间的 6D 位姿关系

有了该转换关系，可将视觉中的目标物体位置转换到目标坐标系或其它参考坐标系下，在机械臂抓取应用中，根据位姿估计关系，可将目标坐标转换为机械臂末

端坐标，实现机械臂末端爪手的抓取。

1.2 CDPN 模型原理

CDPN（Coordinates-based Disentangled Pose Network）[16]是一种基于单张 RGB 图像的六自由度（6DoF）位姿预测框架。网络采用直接法、间接法结合的方式，运用直接法进行物体平移量的估计，结合间接法进行旋转量的估计。直接法通过对 RGB 图像进行算法处理，直接输出物体的三维旋转和位移参数，一般是基于回归的方法；间接法则先建立 2D 图像点与 3D 模型点之间的特征点对应关系，然后利用 PnP（Perspective-n-Point）算法[17]求解位姿。两者有各自的局限性，例如对称性物体，其不同姿态在 RGB 图像上可能呈现出相同外观，导致直接法在估计旋转参数精度波动。间接法由于依赖于特征点，对尺度变化较大或出现遮挡时，平移估计可能误差较大。CDPN 网络使用直接法来得到平移量，间接法来得到旋转量，使用两个平行的检测架构来分别预测，实现对无纹理和被遮挡物体的高精度实时位姿估计，CDPN 的网络架构图如图 2 所示。

1.3 EPRO-CDPN 位姿估计模型

CDPN 原模型中目标检测器采用的是 Faster R-CNN 检测器[18]，其在识别精度和对不同尺度物体的检测能力方面存在局限性，同时在边缘计算场景下，其计算资源占用较大，相比于现今的 YOLOv8 等[19]算法有着较大差距。为提升模型的目标检测算法性能，这里提出基于 YOLO v8 的轻量检测网络 ECASC _ YOLO，将其作为位姿估计模型的前置目标检测器。

引入注意力机制模块 ECBAM，使位姿估计网络着眼于关键特征信息，并加强网络训练过程中通道间的特征融合，以此进一步提升网络性能。

EPRO-PnP 替换原来的传统 PnP 方法，将位姿估计网络转换成端到端的训练网络，并通过将将传统的确定性求解转化为对位姿概率分布的预测，提高 PnP 的性能，使得网络可以输出更精确的旋转姿态信息。

1.3.1 目标检测器 ECASC _ YOLO

目标检测器设计基于 YOLOv8 模型，ECASC _ YOLO 主要工作包括：改进其 C2f 模块，设计了一个模

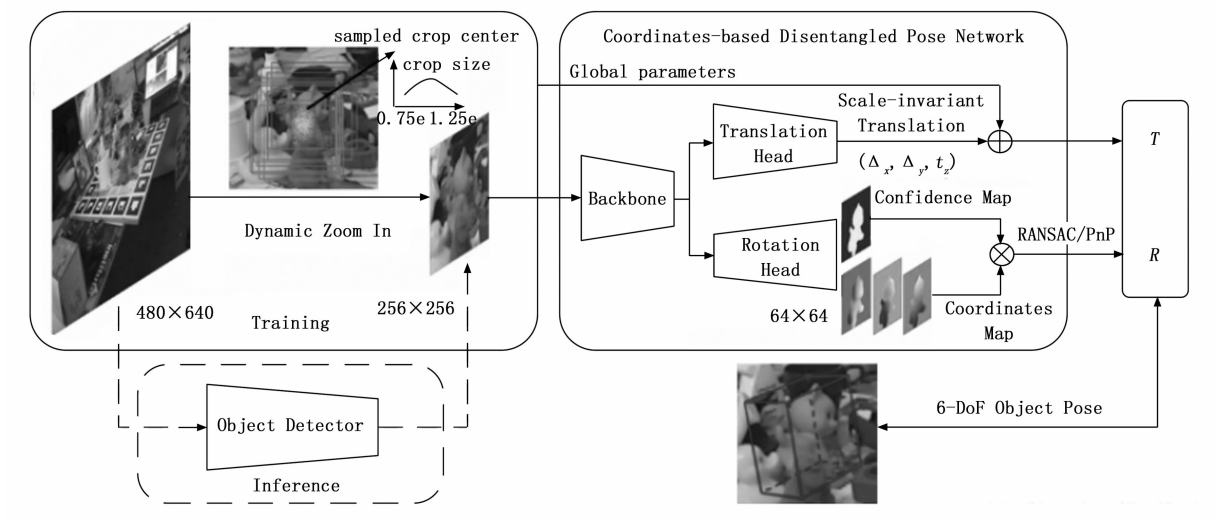

图 2 CDPN 网络结构图

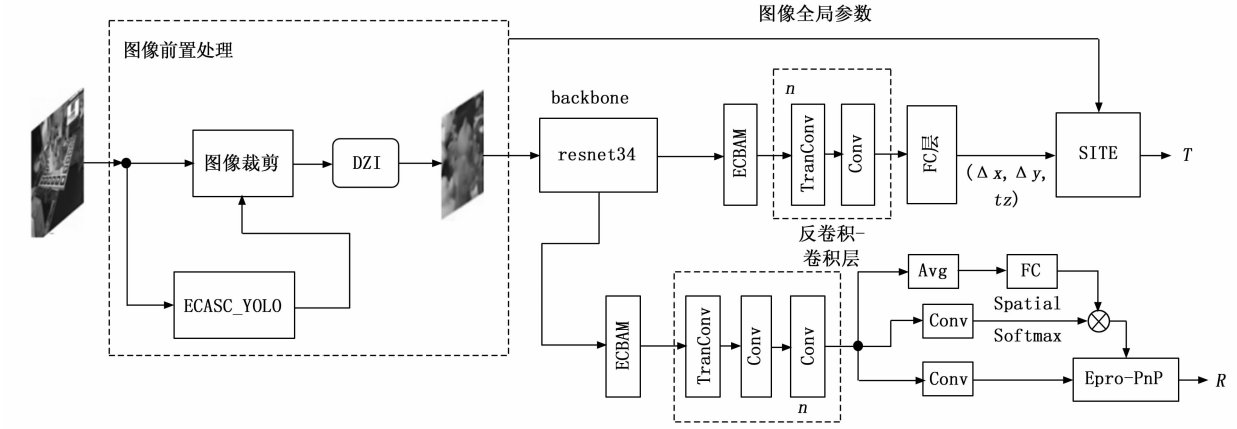

图 3 EPRO-CDPN 网络模型

块 C2f _ mk。引入 ECA 通道注意力机制（ECA，efficient channel attention）[20]、位置注意力机制改进常规 Concat，并结合跳跃链接和空间和通道重建卷积（SCconv，spatial and channel reconstruction convolution）[21]改进特征融合网络（FPN，feature pyramid networks）。网络结构如图 4 所示。

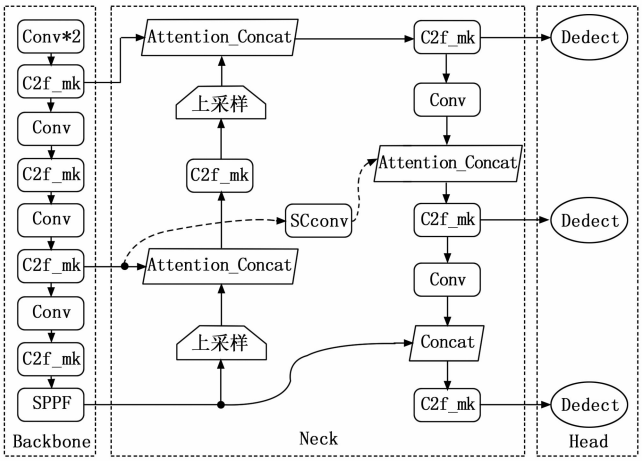

图 4　ECASC _ YOLO 网络结构

1）C2fmk 模块结构如图 5 所示，模块引入 ECA 注意力机制来减少模型计算量，增强特征通道之间的信息交互，通过 1×1 卷积改变通道数，使用 Split 将通道数分割为四份，增加模块分支数。前两部分进行直接拼接，第三个部分进行 ECA 注意力机制和常规卷积的 Add 操作，第四部分进行常规卷积 Conv 的重复特征提取。相比原模块 C2f，模块将参与进行重复卷积的特征通道减半，然后使用 ECA 注意力机制与普通卷积进行输出融合。

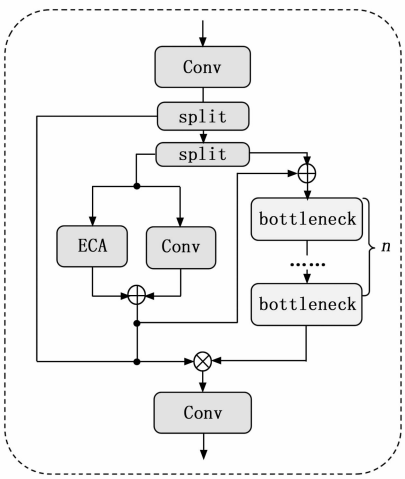

图 5　C2fmk 结构图

2）Attention _ Concat 模块，在网络的 FPN 特征融合部分，将位置注意力机制 PA（position attention，PA）的输出和 ECA 注意力机制的输入作拼接，进行不同输入的特征融合，构成 Attention _ Concat 模块，如图 6 所示。PA 可以增加网络的空间位置信息的获取能力，拼接后输入到 ECA 通道注意力机制可以进一步滤除特征通道冗余，且 ECA 和位置注意力机制 PA 都是轻量型注意力机制，将其与模型结合不会带来太大的参数量。由于其结合了不同尺度的特征层输入，将两个特征输入进行关键信息获取，再将获取的信息继续输入到其他模块进行特征信息进一步提取。这种操作可以有效增加特征融合网络的性能，提高网络的检测性能。

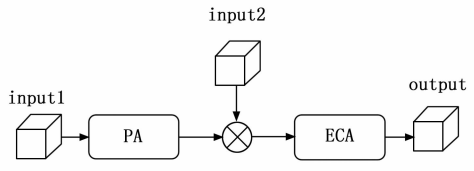

图 6　Attention _ Concat 结构图

3）SCconv 模块和跨越连接，SCconv 模块由两个单元组成：空间重构单元（SRU，spatial reconstruction unit，）和信道重构单元（CRU，channel reconstruction unit），首先通过 SRU 操作得到空间细化特征，然后利用 CRU 操作得到通道细化特征。基于其良好的性能和参数量，将其放到 Backbone 和 FPN 的跳跃连接中，增强模型在不同尺度上的特征融合能力，提高模型的检测性能。模块结构如图 7 所示。

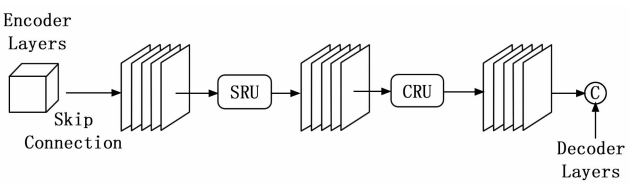

图 7　SCconv 结构图

1.3.2　空间通道注意力 ECBAM

为提升姿态估计网络对各通道特征信息的提取效率并改善整体性能，模型引入 ECBAM 注意力机制，其基于 CBAM 注意力机制原理，通过更改空间注意力机制和通道注意力机制的组合，来增加特征通道的融合和特征空间的信息提取能力。针对注意力机制带来的模型复杂度增加问题，提出轻量化注意力重构方案：采用环境认知自适应模块 ECA 实现通道维度特征校准，结合空间特征聚焦模块 SAM[22]构建双路轻注意力模块。该方案在保持原网络模块化集成特性的前提下，通过特征融合自适应增强机制提升模型判别能力，并有效平衡计算效率与特征表征力的矛盾。这种基于 CBAM 注意力机制已经在多个计算机视觉任务得到有效证明。例如在多移动车辆跟踪算法中，集成了 LCBAM 的模型在检测阶段能够使检测器在有限计算资源下更多地关注前景特征，提高跟踪性能[23]。ECBAM

结构如图 8 所示。

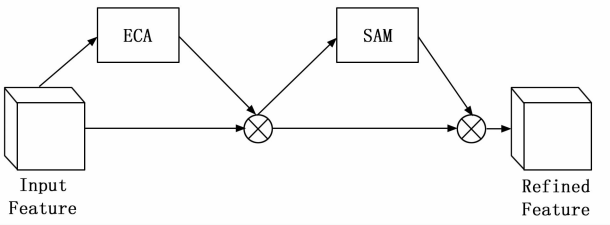

图 8 ECBAM 模块结构图

ECA 是在 SE 注意力机制（Squeeze-and-Excitation Attention Module）的基础上优化改进的，采用自适应一维卷积来捕捉通道间的局部交互信息，避免 SE 模块中可能导致信息损失的降维操作。ECA 模块首先对输入特征图进行全局平均池化（GAP，global average pooling）[24]，以获取每个通道的全局信息，利用大小为 k 的一维卷积核对池化后的特征进行卷积操作，模拟相邻通道之间的交互。为有效捕捉不同层次的通道间依赖关系，卷积核的尺寸 k 被设计为根据输入特征的通道数 C 自适应生成，这不仅增强通道注意力机制对特征表达的建模能力，同时大幅降低计算开销。给定通道数 C，卷积核大小 k 可通过式（1）计算得到：

$$k = \psi(C) = \left| \frac{\log_2(c)}{\gamma} + \frac{b}{\gamma} \right|_{\text{odd}} \tag{1}$$

其中：γ 和 b 是预设的超参数，通常设置为 1 和 0。$|t|_{\text{odd}}$ 表示最接近 t 的奇数。该公式通过将通道数 C 映射到合适的卷积核大小 k，以平衡模型性能和计算复杂度。通过上述设计，ECA 模块能够有效捕捉通道间依赖关系，提升模型的表达能力，同时保持轻量化。ECA 结构如图 9 所示。

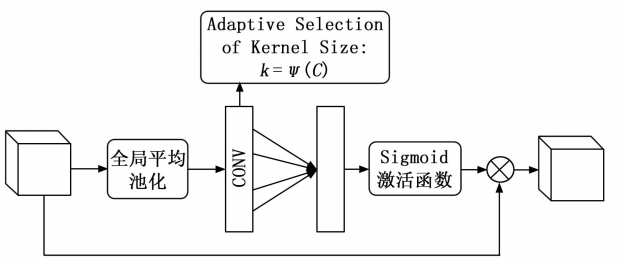

图 9 ECA 注意力机制结构图

空间注意力模块 SAM 通过学习空间维度的信息来提升网络性能，其先使用全局平均和最大池化来进一步获取全局信息，然后使用 $k\times k$ 的卷积核去进行特征融合捕捉空间特征。最终通过激活函数 Sigmoid 的计算得到空间注意力机制权重图。由此将其与原始输入特征相乘来使模型加强对关键空间特征的关注，提升网络性能。在本文应用场景中，SAM 可以细化特征映射的特征分布，增强前景特征，抑制背景特征，进一步减小背景干扰对精度的影响，SAM 结构如图 10 所示。

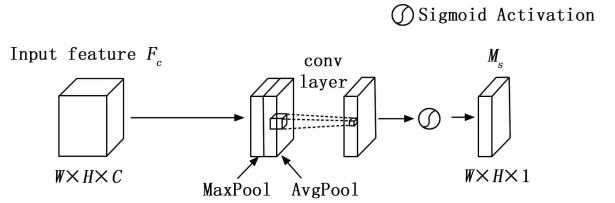

图 10 SAM 结构图

1.3.3 EPro-PnP

EPro-PnP 是近些年提出的，其将传统的确定性求解转化为对位姿概率分布的预测，实现整个过程的可微分性。旋转分量求取是将物体各个像素的置信度 map 和三维空间坐标 map 来形成 2D-3D 点对，使用间接法通过 PnP 来求取物体姿态。核心思想：基于重投影误差构造似然函数，利用贝叶斯公式得到关于旋转矩阵 R 和平移向量 t 的后验分布。

PnP 算法是相机位姿估计经典方法，原理是已知多个三维空间点及其成像点图像坐标，通过建立 3D-2D 点对之间的映射关系，求解相机坐标坐标系与目标坐标系之间的转换，即旋转矩阵 R 与平移分量 t。PnP 方法虽然常用且在理论上可以获得较优解，但实际应用中，受噪声、错误匹配和模型非线性等影响，其性能难以保障。此外 PnP 求解不能成为一个网络模块，使得网络实现端到端训练。原因在 argmin 运算并不能完全可微，且参与 PnP 运算的关键点不连续使得无法使用损失函数来通过反向传播学习所有的 2D-3D 点。本文这里采用 EPro-PnP 方法来进行位姿求解，同时其精度较高。

具体实现：EPro-PnP 首先定义了一个基于重投影误差的似然函数，令每个特征点的重投影误差为 $\| x_i - \pi(RX_i + t) \|$，则似然函数可写为式（2）：

$$p(\{x_i\} \mid R,t) \propto \exp\left(-\sum_{i=1}^{N} \| x_i - \pi(RX_i + t) \| 2\right) \tag{2}$$

在此基础上，结合一个先验 $p(R,t)$（此时对所有 R，t 赋予相同权重），通过归一化操作得到后验概率分布式（3）：

$$p(R,t \mid \{x_i\}) = \frac{p(\{x_i\} \mid R,t)\,p(R,t)}{\int p(\{x_i\} \mid R,t)\,p(R,t)\,\mathrm{d}R\mathrm{d}t} \tag{3}$$

由此将传统最优化问题转化为概率密度的预测，消除 argmin 运算不连续性。后续训练时则采用 KL 散度（KLD，kullback-leibler divergence）作为损失函数[25]，使得预测的分布能够与由真实数据隐式定义的目标分布接近。

上述过程使模块能够通过网络逆向传播来学习，使 EPro-PnP 能够作为模块嵌入到深度网络中。同时，EPro-PnP 内部已经考虑了数据中的不确定性和异常值

问题。因此不需要 PnP 与 RANSAC 结合运算，可直接输出位姿估计结果。

2　实验与结果分析

2.1　数据集

实验所用数据集采用公共数据集 LineMod 和自制的无人机身圆柱体状数据集。公共数据集用于进行模型的对比测试，检测模型有效性。自制数据集由于数据较少且是单目标数据集，无法进行各模型性能对比，故该数据只用于验证位姿估计模型性能可行性，为后续的抓取实验作准备。

1）公共数据集 LineMod：由 Stefan Hinterstoisser 等学者提出的 RGB-D 公开数据集[26]，广泛用于评估物体识别与位姿估计算法的性能。该数据集涵盖了 15 类常见的日常物体，如橡胶鸭、猫雕像和胶水瓶等，每类物体均配有约 1 100 张在复杂环境中采集的 RGB-D 图像序列。图像拍摄过程中引入了多种现实因素，如背景干扰与光照变化，以模拟真实应用场景下的挑战性条件。此外，LineMod 数据集还为每个物体提供了精确的三维模型，并配套标注了其在每帧图像中的平移向量、旋转姿态以及类别标签等信息，成为当前位姿估计研究中的标准评测基准之一。如图 11 所示为数据集样例图。

图 11　LineMod 数据集图片

2）自制数据集：要训练和测试位姿估计网络对无人机抓取圆柱体检测，需要进行共轴双旋翼无人圆柱体机身筒状物数据集的制作，自制数据集的格式为 Linemod 格式，共提取 1 156 帧图像参与制作。在经过数据采集、点云获取与三维表面重建，以及生成 Mask 和标签信息等步骤后，得到完整的自制数据集，图 12 为无人机机身圆柱体图及其 Mask。数据集主要有以下文件：

（1）depth 文件夹：存放目标物体的深度图像，通常以特定的深度图格式（如 .dpt）保存。

（2）JPEImgs 文件夹：对应的 RGB 彩色图像，以 JPEG 格式存储。

（3）mask 文件夹：包含每张图像的掩膜，用于指示目标物体在图像中的位置和轮廓。

（4）labels 文件夹：存储每张图像对应的 3D 包围盒标签信息，包括物体的位姿、类别等详细标注。

（5）intrinsics.json 文件：保存相机的内参信息，如焦距、主点坐标等，便于后续的图像处理和分析。

（6）registeredScene.ply 文件：包含场景的 3D 模型，以 PLY 格式存储，用于三维重建和可视化。

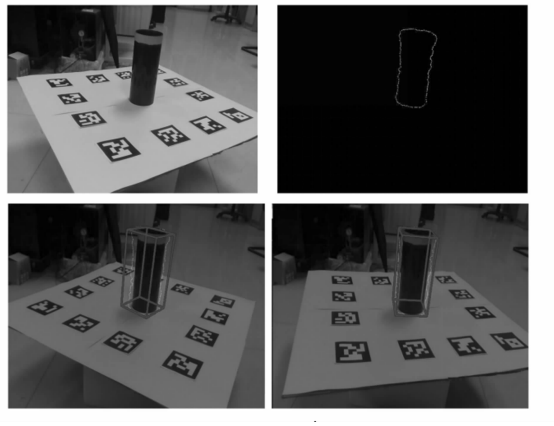

图 12　目标物体及其 Mask

2.2　模型训练

模型在两个数据集上的训练基于 Nvidia RTX 3060 GPU 完成。在训练过程中，为避免参数过拟合，采用数据增强来进行数据预处理。由于位姿估计网络有前置目标检测处理和双分支网络，训练参数设置 batch size 为 4，epoch 为 120，学习率初始化为 0.001，从第 60 个 epoch 后将学习率设置为 0.000 1。置信度图的阈值门限设置为 0.5。采用了多层交替训练策略：首先单独训练带有旋转 head 的 backbone，重点学习物体旋转参数的信息，此阶段有助于模型更好地捕捉旋转特征；在旋转 head 训练较为稳定后，将平移 head 整合到网络中进行训练，以专门学习物体的平移信息；再对整个网络进行端到端调优，使旋转和平移两个分支能够协同优化，提升整体位姿估计的准确性。前置目标检测网络的训练则单独使用 RGB 图像和标签进行训练。训练过程中部分参数的配置如表 1 所示。

表 1　模型训练参数

序号	参数	值
1	batch size	4
2	置信度阈值	0.5
3	总训练轮次	120
4	初始学习率	0.001
5	优化器	RMSProp
6	threads_num	12

评价标准：姿态估计常见的评价指标有：2D projection、ADD（－S）和 5 cm 5°等[27]。

2D 重投影误差[28]：衡量的是关键点的 2D 投影之间的平均像素。将模型的关键点通过预测得到的旋转矩阵 R_p、平移矩阵 T_p 进行 2D 重投影，同时将目标真实的 R、T 也进行 2D 重投影，然后将两者作差，若平均误差小于等于 5 个像素，则认为模型的位姿估计准确，否则模型失效：

$$p.2D = \frac{1}{m}\sum_{x\in M} \| K(Rx+T) - K(R_p x + T_p) \| \quad (2)$$

式中，m 表示关键点的个数，M 表示模型关键点的集合，K 是相机的内参矩阵，x 表示关键点的投影坐标。

ADD（－S）[29]：衡量的是模型的关键点之间的平均距离。如果物体是对称的，通常使用 ADD 指标来评价位姿估计的准确性，即将模型顶点的坐标与预测的坐标之间的平均距离，如式（3）所示，如果平均距离小于目标直径的 10％（也可以使用 2％和 5％作为评判标准），那么预测是正确的：

$$\mathrm{ADD}(-\mathrm{S}) = \frac{1}{m}\sum_{x\in M} \| (Rx+T) - (R_p x + T_p) \| \quad (3)$$

其中：m 是目标模型上的点个数，M 是模型关键点的集合，R 、T 是真实的位姿，R_p 、T_p 是估计的位姿，$R \cdot x + T$ 表示模型姿态转换后的关键点坐标。

本文这里模型评价指标采用 2D projection、ADD（－S）两种指标。

2.3 实验结果

1）LineMod 数据集实验。在该数据集上对包括本文方法在内的几种常见位姿估计网络模型进行性能对比，所有模型均在 GPU RTX3060 平台上进行训练和测试，其 2D projection、ADD（－S）、FPS 等精度和速度指标如表 2 所示。

表 2 Linemod 数据集上不同算法的性能对比

算法	2D projection/％	ADD(－S)/％	FPS/(帧/s)
PoseCNN[57]	70.2	62.7	27.4
BB8[58]	83.9	62.7	34.3
YOLOv5-6D[59]	99.4	96.8	41.9
PVNet[60]	99.0	86.3	30.9
CDPN	94.3	89.8	45.7
Ours	98.2	95.4	59.2

实验数据表明，本文模型在检测精度和检测速度上高于原模型。与 PoseCNN 等经典位姿估计模型比较，性能有提升。与最近提出的 YOLOv5-6D 相比，本文模型在检测精度上略有不足，但实时性更好。图 13 为 LineMod 数据集位姿估计部分结果可视化。

2）自制数据集实验。在自制的无人机身圆柱体状数据集实验为进一步验证模型对无人机身圆柱体的位姿估计能力，避免因泛化能力不足导致的精度下降，该数据集上的 EPRO-CDPN 模型与原 CDPN 模型性能对比结果如表 3 所示。

表 3 基于自制数据集的算法性能对比

算法	2D projection/％	ADD(－S)/％
CDPN	98.8	94.8
Ours	99.1	95.3

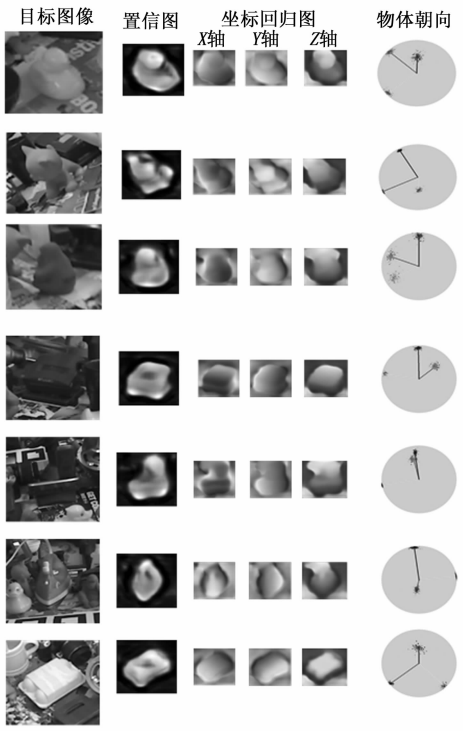

图 13 目标图像与对应的置信图和坐标回归可视化

从表 2 数据看出，EPRO-CDPN 模型在两指标上均略优于原始 CDPN 模型，表明其对于圆柱体目标物体姿态估计性能良好。表 2 中两者性能相差不大的原因是自制数据集为单类目标且目标为简单圆柱状，模型性能差距无法体现。图 14 为自制数据集训练结果可视化图。

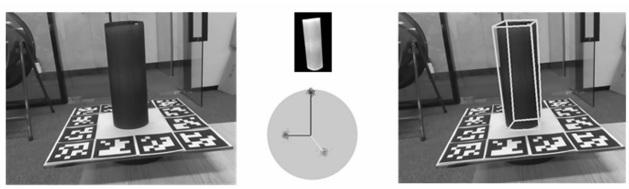

图 14 自制数据集训练结果可视化

3）共轴双旋翼无人机抓取实验。实验硬件环境：双目相机 ZED2i、NVIDIA® Jetson Xavier 为核心的边缘计算平台 TW-T506S、AUBO-i10 机械臂。将算法模型部署于 TW-T506S 上，在外场空旷环境下，风速低于 2 m/s，对悬停状态的共轴双旋翼无人机开展基于姿态估计的抓取实验，评判标准为抓取成功率。同时分析单次抓取中模型估计值与真实值的误差。无人机抓取实验如图 15 所示。

实验进行了 30 次抓取，成功抓取次数为 25 次，成功率为 83.3％，模型检测速度为 35.2 frame/s，由于是在边缘计算平台上运行，相比公开数据集的测试硬件环境，FPS 有所下降。但满足实时性要求。算法估计的目标位姿与实际抓取中机械臂夹爪的末端位姿其中部分数

图 15　共轴双旋翼无人机抓取实验

据见表 4。表 4 中目标位置信息的误差采用 X、Y、Z 的平均相对误差来衡量，姿态信息的误差采用 RX、RY、RZ 的绝对误差去分析，5 组数据中位置信息的平移相对误差在 4.26％以内，X 轴、Y 轴、Z 轴的旋转绝对误差在 1.77°以内。这些误差在机械臂末端爪手允许误差范围内。

表 4　目标位姿与末端位姿对比

		位置信息/cm			姿态信息/(°)		
		X	Y	Z	RX	RY	RZ
1	目标姿态	−11.3	57.8	−107.9	89.9	0.18	−0.21
	末端姿态	−11.5	58.4	−106.7	90.2	0.28	−0.86
	误差	2.60％			−0.3	−0.10	0.65
2	目标姿态	−13.3	50.1	−102.3	92.34	0.86	−4.57
	末端姿态	−12.8	48.9	−103.8	91.59	1.05	−3.89
	误差	3.34％			0.75	−0.19	−0.68
3	目标姿态	−14.7	56.9	−109.7	92.9	−0.63	0.98
	末端姿态	−15.6	55.5	−110.3	91.8	0.32	0.19
	误差	4.29％			1.10	−0.95	0.79
4	目标姿态	−15.4	52.1	−107.9	90.13	−1.25	−3.05
	末端姿态	−16.6	52.7	−106.8	91.6	−0.57	−1.91
	误差	3.32％			−1.47	−0.68	−1.14
5	目标姿态	−15.43	53.25	−105.4	91.62	0.53	−4.07
	末端姿态	−14.32	54.42	−105.7	90.57	0.98	−5.28
	误差	3.50％			1.17	0.24	1.59

3　结束语

本文基于 RGB 图像的位姿估计方法展开研究与实验，提出 EPRO-CDPN 姿态估计算法，该算法将基于 YOLO 的目标检测网络作为位姿估计的前置检测器，模型引入 ECBAM 注意力机制，其由 CBAM 注意力机制改进，通过更改空间注意力机制和通道注意力机制的组合，来增加特征通道的融合和特征空间的信息提取能力。采用基于位姿概率分布预测，整个过程可微的 EPro-PnP 方法进行位姿求解，其作为模块嵌入到深度网络，实现一个端对端的位姿估计模型。

在公开数据集 LineMod、自制数据集上进行了算法性能测试，并以共轴双旋翼无人机为物体目标进行抓取实验，验证姿态估计算法的可行性和有效性。检测精度在 95％以上，检测速度快，FPS 达到 35.2 帧/秒，实现了实时目标姿态估计，为视觉引导的机械臂自动抓取回收共轴无人机奠定了研究和实验方案。

参考文献：

[1] MORRISON D, CORKE P, LEITNER J. Learning robust, real-time, reactive robotic grasping [J]. International Journal of Robotics Research, 2020, 39 (2 - 3): 183 - 201.

[2] PARK D, SEO Y, SHIN D, et al. A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection [J]. 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020: 7300 - 7306.

[3] ZHAO B, ZHANG H, LAN X, et al. REGNet: REgion-based grasp network for single-shot grasp detection in point clouds [J]. Arxiv-CS-Robotics.

[4] PATTEN T, PARK K, VINCZE M. DGCM-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping [J]. Frontiers in Robotics and AI, 2020, 7: 120.

[5] ZENG L, LV W J, ZHANG X Y, et al. ParametricNet: 6DoF pose estimation network for parametric shapes in stacked scenarios [C] //2021 IEEE International Conference on Robotics and Automation (ICRA), 2021: 772 - 778.

[6] HAMZA M. Histogram of oriented gradients (HOG) in computer vision [Z]. Towards Data Science, 2024.

[7] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [J]. ArXiv, 2014.

[8] GIRSHICK R. Fast R-CNN [C] //2015 IEEE International Conference on Computer Vision (ICCV), 2015: 1440 - 1448.

[9] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [M]. ArXiv, 2016.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [M]. ArXiv, 2016.

[11] 张　聪. 基于卷积神经网络的旋转目标检测方法研究 [D]. 成都：电子科技大学，2024.

[12] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [C] //LEIBE B, MATAS J, SEBE N, et al. Computer Vision-ECCV 2016, Cham: Springer International Publishing, 2016: 21 - 37.

（下转第 251 页）