DOI:10.19652/j.cnki.femt.2204430

混合坐标注意力与改进空间金字塔池化融合的物体位姿估计*

党选举[1,2] 李启煌[1,2]
(1.桂林电子科技大学电子工程与自动化学院 桂林 541004;2.广西智能综合自动化高校重点实验室 桂林 541004)

摘 要:在物体杂乱放置非遮挡和遮挡构成的复杂场景下,针对位姿实时、准确和稳定地估计的问题,提出了混合坐标注意力与改进空间金字塔池化融合的目标位姿估计算法。搭建了由坐标特征、通道特征和空间特征组成的混合坐标注意力残差模块,有效提高了关键点估计的准确率。改进了空间金字塔池化网络,并通过颈部位置的多尺度特征细化方法,获得边缘姿态及空间位置的高精确估计。将所制作的遮挡数据集,进一步验证所提出算法性能和泛化能力。在公开 LineMod 及 Partial Occlusion 遮挡数据集上,所提算法与基于组特征注意力(SA)算法相比 ADD 指标分别提高 2.26%和 2.57%,5cm5°指标分别提高 5.16%和 4.1%,达到了 30 fps 实时处理速度,为遮挡等复杂场景下的物体位姿估计提供一个有效的方法。

关键词:遮挡;混合坐标注意力;空间金字塔池化;位姿估计
中图分类号:TP391.4 **文献标识码**:A **国家标准学科分类代码**:520.60

Pose estimation of objects combining shuffle coordinate attention and improved spatial pyramid pooling

Dang Xuanju[1,2] Li Qihuang[1,2]
(1. School of Electronic Engineering and Automation, Guilin University of Electronic Technology, Guilin 541004, China;
2. Key Laboratory of Guangxi College Intelligent-Comprehensive Automation, Guilin 541004, China)

Abstract: In the complex scene composed of non-occlusion and occlusion of objects placed in disorder, aiming at the problem of real-time, accurate and stable pose estimation, a target pose estimation algorithm combining shuffle coordinate attention and improved spatial pyramid pooling is proposed. A shuffle coordinate attention residual module consisting of coordinate features, channel features and spatial features has been built to effectively improve the accuracy of key point estimation. The spatial pyramid pooling network is improved, and the multi-scale feature thinning method of neck position is used to obtain highly accurate estimation of edge pose and spatial position. The produced occlusive dataset is used to further validate the performance and generalization capability of the proposed algorithm. On the public LineMod and Partial Occlusion occlusive datasets, the proposed algorithm improves ADD metrics by 2.26% and 2.57% respectively, and 5cm5° metrics by 5.16% and 4.1%, respectively, compared to the shuffle attention (SA)-based algorithm, reaching a real-time processing speed of 30 fps, providing an effective method for object pose estimation in complex scenes such as occlusion.

Keywords: occlusion; shuffle coordinate attention; spatial pyramid pooling; pose estimation

0 引 言

近年来,大量研究致力于复杂场景下估计目标物体 6 自由度的平移及旋转位姿信息,这一直是工业机器人抓取发展的重要研究课题[1]。针对在遮挡和非遮挡复杂场景中杂乱放置情况下,估计目标物体的 6 D 位姿主要有 3 种方式。1)基于投票方式。文献[2-4]分别通过引入像素级向量投票技术、考虑图片颜色输入信息和增加距离影响因

收稿日期:2022-10-20
* **基金项目**:国家自然科学基金(62263004,61863008)项目资助

素的方法，降低遮挡对目标位姿估计的影响，提高目标位姿估计的准确度，但运算量大。2）直接回归方式。文献[5]对图像中心进行定位，预测其与相机距离，估计对象位姿。文献[6]扩展 SSD 算法产生的 2 D 检测框信息，推断物体位姿。文献[7]采用了融合通道空间注意力网络 CSA6D，直接输出物体 6 D 位姿。此类方式减少了运算处理量，但网络估计的位姿精度有待提高。3）关键点方式。文献[8]先预测多个小块的热图，再综合热图信息获得位姿。文献[9]从细化的感兴趣目标区域，得到 3D 检测框的角点在图像上的投影点坐标，采用多阶段的方法回归关键点，提高了位姿估计的精度，但结构较为复杂。为了简化模型结构，文献[10]采用 YOLO 框架，加快了网络定位关键点速度，但模型训练效果和鲁棒性有待加强。文献[11]引入通道注意力机制，提高了位姿估计精度。以上 3 种方式，在遮挡和非遮挡场景下，都存在位姿估计准确度、速度和稳定性三者之间有效平衡问题。

围绕位姿估计准确度、速度及稳定性问题，本文采用 YOLO 框架，设计了一种混合坐标注意力（shuffle coordinate attention，SCA）机制，并融合改进的空间金字塔池化（spatial pyramid pooling，SPP）网络，构造了单阶段端到端 6 D 位姿估计模型（YOLO-SS）。该模型无需细化处理，具备轻量化特点。为了适应特定的遮挡环境，制作了遮挡数据集，丰富了训练数据，验证了模型的鲁棒性和泛化能力。

1 YOLO-SS 的网络结构

本文提出了一种 YOLO-SS 的网络结构，精简了颈部网络结果，以提高推理速度；提出一种改进的 SPP 网络，增强整体与细节信息的融合；构造了混合坐标注意力（SCA）残差模块，在网络结构中添加位置特征信息；选择合适的激励函数，增强网络的鲁棒性。

1.1 YOLO-SS 的改进结构

本文所提出的 YOLO-SS 网络如图 1 所示，以 Darknet53 作为主干网路[12-13]，提出了 SCA 注意力，具体结构如虚线框中 SCA 残差模块（SCA-RESX）所示。通过该模块抑制部分与目标无关的背景和其他弱相关物体带来干扰的特征信息，如数据集图片中所包含的标定板和周围杂乱物体，突出目标区域的重要信息。

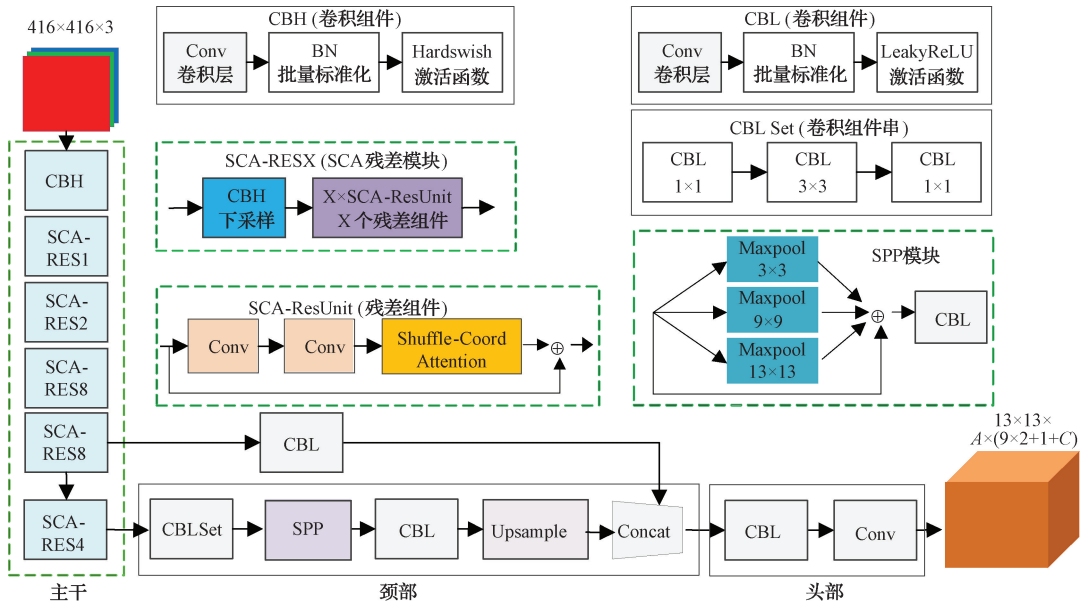


图 1 整体框架（YOLO-SS 的网络结构）

由于原始颈部采用多尺度处理时，位姿估计实时性差。YOLO-SS 网络在原始 YOLO v3 的颈部加入了图 1 中改进的 SPP 模块，对局部目标区域的位置细节信息和整体的语义轮廓信息进行提取和融合，得到更高维度的特征信息。在 SPP 输出引入 1/16 倍原分辨率的特征图，实现通道维度上的拼接，减少了特征图多次下采样带来的位置信息丢失。

将网络多尺度输出删减为单尺度（分辨率 13×13）输出，通道数为 $A\times(9\times2+1+C)$，其中 A 为锚框（Anchor）个数，每个锚框包含 9 对关键点坐标、1 个置信度值和 C 个类别概率信息。

1.2 混合坐标注意力残差模块

在遮挡目标位姿估计中，缺失估计目标特征信息，产生特征提取偏差现象及降低目标检测精度。此外，主干网络层数变多将进一步弱化物体的坐标特征信息，组特征注意力（shuffle attention，SA）[14]仅通过参数化的卷积形式，细化空间位置信息的能力不足，降低了关键点位置的定位精度。

针对以上问题，本文在 SA 注意力机制上，提出了混合坐标注意力残差模块如图 2 所示。采用空间注意力模

块，根据特征信息的重要程度生成不同比值的空间区域，突出关注目标区域特征，减弱其他物体的干扰。采取具有跳跃连接的残差连接分支，缓解注意力和层数迭代带来的位置信息丢失和模型退化的问题。添加通道坐标注意力新分支，融合通道信息和坐标系信息，提高网络对特定通道上位置信息的感知能力，增加网络丢失的位置信息。

由描述通道坐标信息和空间信息的注意力模块和聚合模块构成混合坐标注意力残差模块。

1)注意力模块

将输入维度为[C，W，H]的特征图分为 G 组，每组通道输入 $X=[X_i, \cdots, X_G]$，通道维度变为 C/G。将每组输入 X_i 划分为图 2 的上下两分支，分别为通道坐标注意力分支和空间注意力分支，通道坐标注意力由坐标信息模块和通道信息模块构成。两分支输入为 X_{i1} 和 X_{i2}，通道维度数为 $C/2G$，输出为特征图 X'''_{i1} 和特征图 X'''_{i2}。

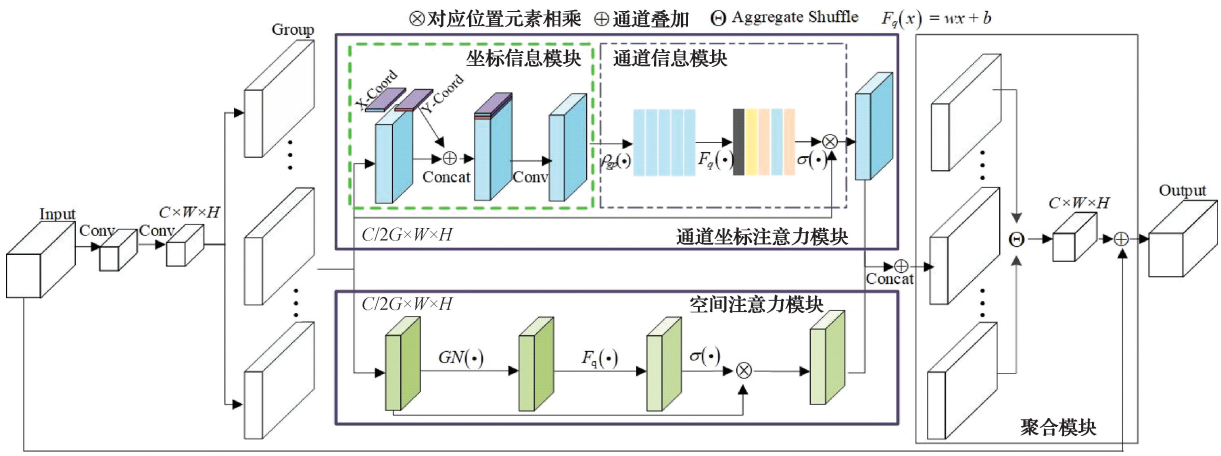


图 2 混合坐标注意力残差模块

坐标信息模块，为了弥补残差模块中坐标信息丢失，在通道信息模块前加入坐标信息模块[15]（图 2 绿色虚线框）。将输入特征图 $X_{i1}(i,j)$ 分别在通道上赋予 X 和 Y 轴信息，由于每次提取特征时张量数值不一致，需要将 X 和 Y 信息数值固定缩放至 $-1\sim1$。通道上叠加了两层坐标信息的特征图 $X'_{i1}(i,j)$，为保持后续网络回归，通过卷积方式，转化成 $X_{i1}(i,j)$ 的通道大小，同时增强了通道上的坐标特征。转化公式如下：

$$X'_{i1}(i,j)=Concat(X_{i1}(i,j),x_{coord},y_{coord}) \tag{1}$$

$$X''_{i1}(i,j)=Conv(X'_{i1}(i,j)) \tag{2}$$

式中：Concat 表现通道级联；Conv 表示卷积操作；$X_{i1}(i,j)$ 为输入特征图；x_{coord}、y_{coord} 为 X、Y 坐标信息。

通道信息模块，为了提取坐标信息进行通道信息整合，坐标信息模块后串联了通道信息模块（图 2 通道信息模块）。将添加坐标信息的特征图全局池化，获得各个通道的权重。将输入权重进行平移，调整成适合的 Sigmoid 函数输入，使输入权重对应到 0～1 的范围，获取特征图各通道的重要程度信息。将调整后的权重与输入特征图通道相乘，实现网络对通道的感知能力，以减少受无关通道的干扰，如式(3)、(4)所示。

$$m=\rho_{gp}(X''_{i1})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X''_{i1}(i,j) \tag{3}$$

$$X'''_{i1}=\sigma(F_q(m))X_{i1}=\sigma(w_1m+b_1)X_{i1} \tag{4}$$

式中：W 和 H 为特征图的宽度和高度；$X''_{i1}(i,j)$ 坐标信息模块输出特征图；$\rho_{gp}(\cdot)$ 为全局平均池化；X_{i1} 为通道坐标注意力分支输入特征图；$w_1,b_1\in R^{C/G\times1\times1}$ 为权重向量；$\sigma(\cdot)$ 为 Sigmoid 激活函数。

空间注意力模块，通过图 2 的空间注意力模块，实现特征图空间信息整合。对输入特征图进行组归一化，并对输入特征进行空间参数化。通过权重向量进行值域的平移，使其满足 Sigmoid 函数输入要求，得到特征图的各个区域的比值，筛选出重要空间区域，达到补充通道坐标注意力模块中空间信息的目的，如式(5)所示。

$$X'''_{i2}=\sigma(w_2GN(X_{i2})+b_2)X_{i2} \tag{5}$$

式中：X_{i2} 为通道注意力分支输入特征图；$GN(\cdot)$ 为组归一化；$w_2,b_2\in R^{C/G\times1\times1}$ 为权重向量；$\sigma(\cdot)$ 为 Sigmoid 激活函数。

2)聚合模块

图 2 的聚合模块，将注意力双分支进行通道拼接，以保持网络整体维度一致。为了有效融合目标特征在不同组之间的信息，基于 ShuffleNet[16] 网络中 Channel Shuffle 结构，进行数据重组与整合，得到了跨组交流的融合图。将输出与初始输入进行数值相加，实现有效特征信息的叠加。

相比 SENet[17] 注意力，SCA 注意力通过卷积替代多参数复杂运算的全连接层，减少了模型的训练参数，加快网络训练和推理速度。混合坐标注意力残差模块增加了空间和坐标信息融合操作，各分支训练参数只涉及局部变化，在分组机制下，参数的数量已倍数级减少，加快了响应速度，有效地获取了更多的关键点定位信息，提高模型的

位姿估计精度。

1.3 改进的空间金字塔池化

在目标检测中，增大感受野，有利于提取目标细节特征信息[18]。本文在加入注意力后，在整体框架中的颈部区域提出改进空间金字塔池化如图 3 所示。改进部分为图 3 虚线框，将传统最大池化层中 5×5 尺度的池化核改成 3×3 尺度，添加的卷积组件(CBL)融合 3 个尺度的池化特征。特征图池化后所包含的信息熵计算如式(6)[18]所示。

$$H(p,k)=(S-k)^2\left(-\frac{1}{S^2}\log\frac{1}{S^2}\right)+2(S-k)\left(-\frac{k}{S^2}\log\frac{k}{S^2}\right)+\left(-\frac{k^2}{S^2}\log\frac{k^2}{S^2}\right) \tag{6}$$

式中：S 是颈部区域中的特征图对应尺寸固定为 13；k 是池化核尺度的大小；p 为随机事件的概率。

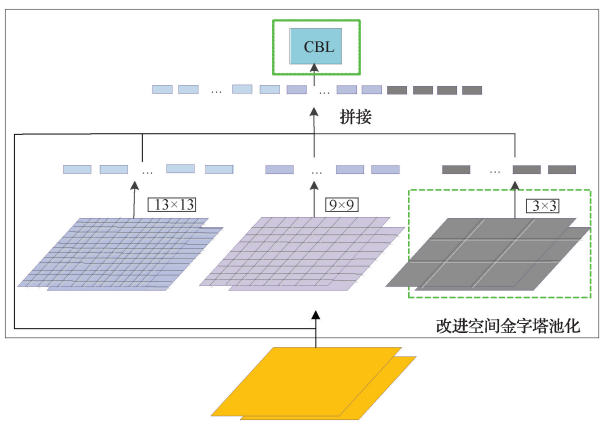


图 3 组件框架图

由式(6)可知，为增大特征图的信息熵 $H(p,k)$，丰富特征图含有的信息特征，选择较小的 3×3 尺度的池化核。网络将获得更多的颗粒度信息，有利于挖掘边缘特征，对小尺寸目标更敏感。对于不同尺寸的目标物体，改进 SPP 通过结合 3 种尺度的池化核，增大网络感受野，增强特征图中目标的特征信息，补充目标位置的细节信息，细化上下文信息，达到与主干信息有效融合目的，提高网络位姿估计的准确度。

1.4 激活函数

激活函数可对特征数据进行非线性映射，增强网络的表达能力。本文改用 HardSwish 函数作为主干部分的激活函数，改用计算简单的 LeakyReLU 函数在颈部部分。整体达到增强网络非线性拟合效果、改善特征提取能力和提高检测速度。HardSwish 和 LeakyReLU 表达式分别如下：

$$HardSwish=\frac{x\ ReLU6(x+3)}{6} \tag{7}$$

$$LeakyReLU=\max(0,x)+leak\cdot\min(0,x) \tag{8}$$

式中：x 为输入特征图；max 为取最大值；$leak$ 为一个很小的常数；$ReLU6(\cdot)$ 为非线性激活函数。

相比 Swish 函数，HardSwish 函数输出具有下边界及正值无穷大的特性，有助于消除网络的饱和问题改善网络正则化，提高检测模型对不同数据集的性能。HardSwish 具有封装好的调用接口，运算速度比 Swish 更快，减短了网络训练的周期。HardSwish 求导数值在边缘区域更加优异 Swish，在主干中有助于学习更有表现力的特征，增强系统的鲁棒性。

随着网络层数加深，特征图的分辨率逐渐减半，对于非线性部分的拟合逐渐降低。对于网络深层的部分，LeakyReLU 满足网络拟合要求，计算量小，输出梯度恒定，能缓解梯度消失情况，解决网络梯度饱和问题。

2 改进的遮挡数据集制作方法

遮挡数据集制作存在如下 3 个问题：1)3D 数据集依靠手动标注难以实现；2)直接使用公开 Occlusion Line-Mod 数据集无法检验遮挡目标实际效果；3)传统数据集制作方法无法实现遮挡情况下数据集制作。为了解决上述问题，本文提出一种用于复杂环境中遮挡目标的数据集制作方法。

实验平台包括 Intel Realsense D435i 相机、三脚架、ArUco 标识码和目标物体，如图 4 所示，其中 ArUco 标识码由不同的内部 ID 和外部边框组成，标识码中心设为世界坐标系原点，不同的 ID 用于标记特定的场景位置，目标物体的中心点作为物体坐标系原点。

图 4 实验平台图

目标物体建模时，缺失目标任意角度的真实标签数据，无法得到目标模型。传统制作数据集方法，采集遮挡物体第一帧的位置数据作为初始点，按照一定时间间隔获取每帧的变换矩阵数据，出现遮挡目标标签数据缺失的问题。实验中，固定相机的初始角度和位置，分别进行两次同一场景不同角度的数据采集，具有获取遮挡物体的各个角度的模型信息和遮挡物体数据集信息的能力，改进的数据集采集部分如图 5 所示(绿色虚线框)。

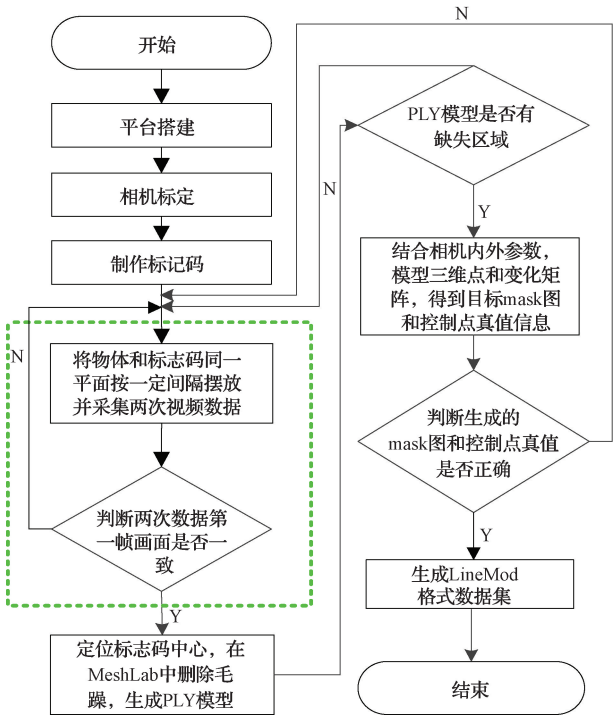
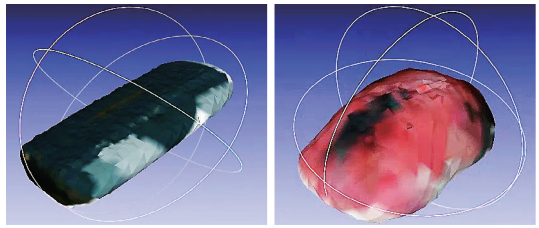


图 5　数据集制作流程

数据集制作整体流程如下：1)将目标置于复杂场景，调节相机角度，实时采集目标各角度的数据，建立目标模型；2)采集存有清晰标识码的每帧数据，检测和定位每一帧数据中的标识码中心，得到每帧数据中物体相对标识码的位姿信息；3)通过计算每一帧相对第一帧实时位置数据，得到每一帧的位姿变换矩阵；4)通过像素值白化处理，将 PLY 模型的 3D 位置投射到平面区域，删除目标区域不符要求的图片，得到 Mask 文件；5)通过 Mask 文件信息，获取 2 D 投影关键点坐标和宽高数值，写入 Label 文件中。物体 PLY 模型文件如图 6 所示，掩码 Mask 文件和关键点坐标 Label 文件组成 Partial Occlusion 遮挡数据集。



(a) 眼镜盒模型　　(b) 鼠标模型

图 6　PLY 模型

3　位姿估计

物体 6 D 位姿中，结合 PLY 模型的三维关键点坐标和 YOLO-SS 网络输出的 2D 关键点映射坐标，通过 PnP (perspective-n-point)算法[19]，得到相机坐标系到物体坐标系的变换矩阵。

为减少遮挡环境对目标定位影响，采用 Anchor 机制进行处理。将网络输出特征图划分成 13 × 13 个网格区域，每块区域预设不同大小的 Anchor。通过 Anchor 与真实标签进行 IOU (intersection-over-union) 计算，选择计算中最大的 Anchor。通过网络训练，使 Anchor 贴近目标真实关键点位置，解决遮挡物体特征缺失的问题，提高网络对遮挡物体的位姿估计能力。为了缩小预测范围，减少运算推理时间，通过掩码矩阵，将目标预测区域对应的矩阵位置设置为非零值，以过滤背景区域。

模型训练时，将坐标损失 L_{coord} 和置信度损失 L_{conf} 改用 Smooth L1 损失求和函数，减少网络对离群点的敏感性，使整体梯度趋于稳定，降低梯度爆炸的影响。分类损失 L_{cls} 使用交叉熵损失函数。考虑训练初期置信度较低和误差大，前十五个批次不增加置信度损失。损失函数整体采用权重求和方式，其中分类损失和坐标损失权重 λ_1、λ_3 为 1，置信度权重 λ_2 为 5。损失函数如下：

$$L = \lambda_1 L_{cls} + \lambda_2 L_{conf} + \lambda_3 L_{coord} \tag{9}$$

求解置信度损失函数时，涉及置信度计算。原 YOLO 置信度计算仅是 4 个角点数据，相对位姿估计的 9 个关键点数据而言运算难度大且时间长。针对此问题，本文引入了直接求解置信度函数[10]，如式(10)所示。

$$c(x) = \begin{cases} \dfrac{e^{\mu\left(1-\frac{D_T(x)}{d_{th}}\right)}-1}{e^{\mu}-1}, & D_T(x) < d_{th} \\ 0, & \text{其他} \end{cases} \tag{10}$$

式中：d_{th} 为置信度函数的像素阈值设为 80；μ 为指数函数的锐度设为 2；$D_T(x)$ 为网络预测坐标和真实标注坐标之间的欧氏距离。

模型测试时，通过网络输出的类概率与置信度分数相乘得到每个类别的类置信度，采取阈值机制过滤类置信度较低的 Anchor。过滤中类置信度值最大的 Anchor 作为所需预测类别的整体信息，包含目标关键点位置信息、所属类别信息和该类别置信度大小，结合 PnP 算法，得到估计 6 D 位姿。

4　实验结果与分析

实验平台中，硬件包括：Intel Xeon Silver 4210R，RAM 为 24 G，GPU 为 RTX 3090；软件包括：操作系统为 64 位 Ubuntu，版本号 18.04LTS，Python 3.6，Cuda 11.0，Cudnn 8.1.1，Anaconda 3 及 Pytorch 1.7.1。

4.1　数据集

为了验证本文方法的有效性，分别在 LineMod 标准 6D 位姿估计数据集和自制 Partial Occlusion 遮挡数据集，进行了实验验证。

LineMod 数据集，公开标准 6 D 位姿估计数据集，主要在背景杂乱，光照不足的复杂环境中，以 13 个物体对象为中心，每个非遮挡对象包含约 1 200 张图片实例，真实标注 6 D 位姿文件、对应位置掩码文件和 PLY 模型文件。

训练与测试中，直接采用数据集提供的15%作为训练集，85%作为测试集。

Partial Occlusion遮挡数据集，自制的实际复杂环境下遮挡且杂乱放置的对象数据集中包含9个常见物体的图片实例。每个对象都有对应的PLY模型文件、对应的掩码图和真实标注位姿信息。平均每个对象有1 100张图片实例，按照常用数据集分配方法，将其中20%作为训练集，80%作为测试集。

4.2 评价指标

由于PnP算法的输入标签和网络预测出的2 D关键点不一致，导致旋转和平移矩阵不同，不可避免位姿估计误差。在复杂背景下，为了更加准确说明位姿估计的效果，涉及到衡量算法的精确性评价指标如下。

5cm5°[9]，网络预测和真实标签分别求出的旋转矩阵误差在5°之内和平移矩阵误差低于5 cm，求出的位姿正确。

ADD[10]，模型关键点通过真实位姿和预测位姿转换后的3D平均距离误差在模型直径允许的10%数值之内，位姿估计正确。

旋转矩阵误差为：

$$e_T = \sqrt{\sum (\boldsymbol{T} - \boldsymbol{T}_{pr})^2} \tag{11}$$

平移矩阵误差为：

$$e_{ang} = \arccos\left[(tr(\boldsymbol{R}\boldsymbol{R}_{pr}^{-1}) - 1)/2\right] \tag{12}$$

3D平均距离误差为：

$$e_{ADD} = \frac{1}{|M|}\sum_{x\in M} \| (\boldsymbol{R}x + \boldsymbol{T}) - (\boldsymbol{R}_{pr}x + \boldsymbol{T}_{pr}) \| \tag{13}$$

式中：$\boldsymbol{T}$ 和 $\boldsymbol{T}_{pr}$ 分别为真实和预测的平移矩阵；tr 代表求矩阵的迹；M 是3D模型顶点坐标；x 是三维关键点；$\boldsymbol{R}$ 和 $\boldsymbol{R}_{pr}$ 分别为真实和预测的旋转矩阵。

4.3 实验结果分析

为了与文献算法进行精确度比较，分别基于LineMod标准数据集和Partial Occlusion遮挡数据集，使用ADD和5cm5°指标来衡量。另外，基于SPP模块，进行注意力的消融实验。

1)不同算法实验分析

为检测本文提出的方法在特定复杂环境中实际效果，与文献[9]提出的基于感兴趣区域进行位姿估计的方法、文献[10]基于YOLO v2的方法和文献[20]优化YOLO v2网络框架的方法进行了6 D位姿估计的实验对比分析。

(1)LineMod公开的标准数据集实验分析

本文进行了SCA注意力与改进SPP网络融合，改进了网络结构，提升了位姿估计的精度。对于ADD指标，实验结果如表1所示。与文献[9,20]方法相比较，分别提高了17.59%和8.81%。相比文献[10]的方法，在部分物体也取得了较大的精度提高。例如在复杂背景下的小尺寸猿，大尺寸的手机等目标提高了10%以上。

表1 LineMod数据集上ADD指标对比 (%)

目标模型	猿	台虎钳	浇水壶	钻工	鸭子	鸡蛋盒	胶水瓶	熨斗	台灯	手机	平均值
文献[9]	27.90	62.00	48.10	58.60	**32.80**	40.00	27.00	67.00	39.90	35.20	43.85
文献[10]	21.62	81.80	68.80	63.51	27.23	**69.58**	**80.02**	74.97	71.11	47.74	60.64
文献[20]	23.24	85.47	73.03	64.72	24.04	34.55	41.22	73.24	66.89	39.87	52.63
本文	**32.29**	**88.01**	**77.66**	**67.99**	24.65	66.32	52.54	**78.65**	**76.20**	**50.05**	**61.44**

注：加粗部分为该列最优值

对于5cm5°指标，实验结果如表2所示。本文相比文献[10,20]的方法，在可允许的旋转和平移误差范围内取得了不错的效果，分别提高了16.31%和7.6%，并相比文献[9]的结果平均值得到一定的提高。例如，在较大尺寸浇水壶和台灯目标上的准确率都存在较大提高。

表2 LineMod数据集上5cm5°指标对比 (%)

目标模型	猿	台虎钳	浇水壶	钻工	鸭子	鸡蛋盒	胶水瓶	熨斗	台灯	手机	平均值
文献[9]	**80.20**	81.50	76.80	69.60	**53.20**	81.30	54.00	**61.10**	67.50	58.60	68.38
文献[10]	41.81	77.91	83.17	60.05	38.59	39.62	52.22	30.34	51.54	46.78	52.20
文献[20]	55.81	**89.05**	88.78	59.46	35.40	59.34	57.92	43.11	79.75	40.44	60.91
本文	57.52	84.72	**89.87**	**72.75**	39.58	**81.61**	**66.92**	51.07	**81.67**	**59.37**	**68.51**

注：加粗部分为该列最优值

整体而言，对于多尺度物体都有一定提升，但对于对称物体，例如小尺寸的胶水瓶和鸡蛋盒，位姿估计的精度提升都较低。加入注意力和SPP的网络，更加关注目标的细节信息，提高了多尺度目标的检测精度，但网络对于对称信息学习不敏感，降低了位姿估计精度的提升幅度。

(2)Partial Occlusion 遮挡数据集实验分析

对于遮挡且杂乱放置物体,本文改进了文献[10]提出的 YOLO v2 方法,进行的对比实验结果如表 3 和 4 所示。通过引入混合坐标注意力,增强目标信息,抑制干扰信息。在平衡网络计算时间和性能情况下,增加了网络坐标信息。对于 ADD 指标,小尺寸电容提高了 13.41%,较大目标的相机盒和羽毛球桶提高了 15% 以上,平均值从 22.47%提高至 28.45%。对于 5cm5°指标,本文算法平均提高了 15.95%,在大部分的目标物体上都得到显著的提高,例如小目标电容和中等目标黑板擦等提高了 15%。实验结果验证,本文所提出的方法,在遮挡环境中具有较强鲁棒性,适合于复杂背景下进行稳定的位姿估计。

表 3 Partial Occlusion 遮挡数据集上 ADD 指标对比 (%)

目标模型	黑板擦	鼠标	相机盒	橙子	电容	魔方	牛奶盒	固体胶	羽毛球桶	平均值
文献[10]	21.19	14.89	28.97	5.62	2.70	7.58	**40.04**	**15.54**	65.66	22.47
本文	**30.16**	**18.44**	**52.20**	**12.39**	**16.11**	**8.67**	23.56	11.02	**83.53**	**28.45**

注:加粗部分为该列最优值

表 4 Partial Occlusion 遮挡数据集上 5cm5°指标对比 (%)

目标模型	黑板擦	鼠标	相机盒	橙子	电容	魔方	牛奶盒	固体胶	羽毛球桶	平均值
文献[10]	7.06	30.78	10.15	26.44	4.75	22.16	**78.10**	10.36	65.66	28.38
本文	**25.22**	**35.00**	**50.73**	**38.06**	**70.59**	**28.30**	34.23	**28.62**	**88.25**	**44.33**

注:加粗部分为该列最优值

2)消融实验分析

针对特定非遮挡和遮挡环境中目标位姿估计任务,未加 SPP 模块的网络无法满足预期位姿估计的精度要求。融合 SPP 模型后,注意力机制对各项性能指标进行了对比。如表 5 所示,其中 SPP 为改进后的空间金字塔池化模块、SA 为组特征注意力残差模块和 SCA 为设计的混合坐标注意力残差模块。由表 5 可得,在 SPP 模块下引入 SA 注意力机制,进行空间和通道信息处理,提升了目标位姿估计的精确度。在 LineMod 数据集上,加入 SCA 注意力相比加入 SA 注意力,ADD 和 5cm5°指标分别提高了 2.26%和 5.16%,在 Partial Occlusion 遮挡数据集上 ADD 和 5cm5°指标分别提高了 2.57%和 4.1%。

表 5 消融实验结果 (%)

主干	SPP	SA	SCA	LineMod		Partial Occlusion	
				ADD	5cm5°	ADD	5cm5°
Darknet53	√			55.73	59.16	23.27	35.38
	√	√		59.18	63.35	25.88	40.23
	√		√	**61.44**	**68.51**	**28.45**	**44.33**

注:加粗部分为该列最优值

实验结果表明,与融合空间和通道信息的 SA 注意力相比,本文增加了像素维度坐标信息的 SCA 注意力,使网络在精度和效率上达到平衡,改善了位姿估计的结果。

3)运行时间实验

本文所提出的位姿估计模型,整体时间与输入图片尺寸、网络大小和后期处理有关。对于 640×480 的图片,使用 RTX 3090 时,测试结果可以达到 30 fps,其中数据载入耗时 0.5 ms,网络前向处理耗时 24.6 ms,关键点处理和 PnP 操作耗时 8 ms。

4.4 实验结果可视化

对 LineMod 公开数据集和 Partial Occlusion 遮挡数据集的实验结果随机可视化,如图 7 所示。绿色框是基准位姿信息投影的 3D 框,蓝色框是网络预测位姿信息投影的 3D 框。

由图 7 可知,本文的效果图相比文献[10]的效果图,实验结果更加接近基准 3D 框,未出现较大的关键点偏差。实验表明通过混合坐标注意力残差块进行目标聚焦,实验结果更加稳定和准确。对于 LineMod 数据集中小尺度的猿,与背景颜色重合的鸡蛋盒,在复杂光照情况下,算法也实现了精准的关键点定位。对于 Partial Occlusion 数据中遮挡的电容、固体胶和鼠标目标,在电容大小仅有鼠标 1/3 和固体胶可见区域极低情况下,算法均取得了精确的位姿估计可视化结果。综合分析可知,通过加入 SPP 网络充分利用特征图中多尺度位置细节信息,增强对小目标的感知能力,结合 SCA 注意力减少遮挡因素带来的干扰,增强网络稳定输出贴近基准关键点的能力,使得目标位姿检测精度整体提高。通过在不同数据集上实现特定复杂情景下的实验结果可视化,说明了本文算法具有的鲁棒性和精确性。

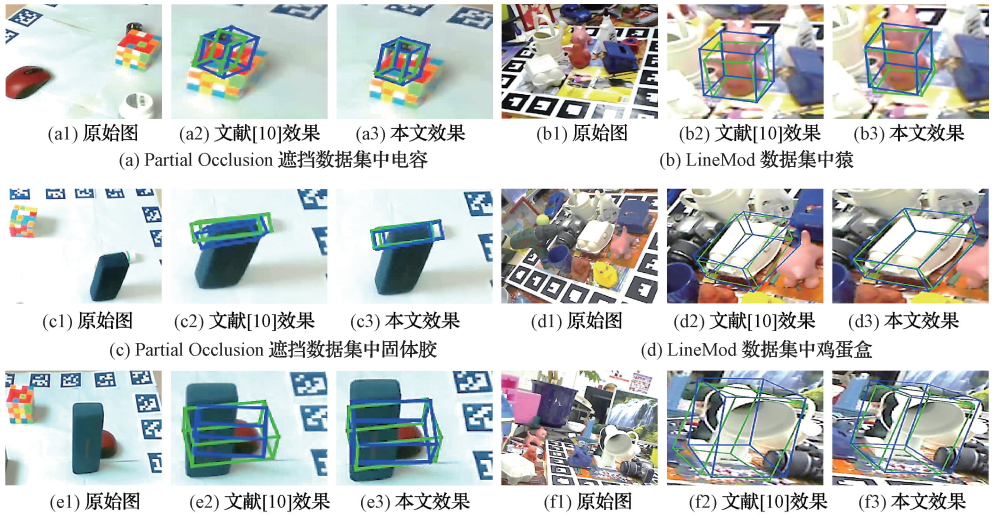


图 7　实验对比

5　结　论

面向复杂场景下杂乱放置的遮挡和非遮挡物体，本文基于 YOLO 网络构架，提出了混合坐标注意力与改进空间金字塔池化融合的物体位姿估计算法。采用改进 SPP 网络细化颈部位置的多尺度特征信息，解决了底层位置信息与高层语义信息断层问题。利用特征堆叠的方式有效降低了网络变深导致位置信息丢失的影响，进一步融合高效率的 SCA 注意力机制，实现了遮挡目标的精确位姿估计。所改进的位姿估计方法在推理时间上可达到 30 fps 的实时效果。相比 SA 注意力，消融实验也进一步验证了所提出的 SCA 注意力机制对提高位姿估计精度的有效性。本文算法与基于 YOLO v2 经典算法相比，在公开数据集 LineMod 中，ADD 和 5cm5°指标分别提高了 0.8%和 16.31%，在自制 Partial Occlusion 遮挡数集下，分别提高了 5.98%和 15.95%。通过在公开和自制遮挡的数据集下的结果，验证了所提出的算法在遮挡等复杂情况下进行目标位姿估计的准确性和鲁棒性。

参 考 文 献

[1] 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.

[2] PENG S，ZHOU X，LIU Y，et al. PVNet：Pixel-wise voting network for 6DOF object pose estimation [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2022：3212-3223.

[3] WANG C，XU D，ZHU Y，et al. Densefusion：6D object pose estimation by iterative dense fusion[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：3343-3352.

[4] YU X，ZHUANG Z，KONIUSZ P，et al. 6dof object pose estimation via differentiable proxy voting loss[J]. Computer Science，2020，arXiv：2002.03923.

[5] XIANG Y，SCHMIDT T，NARAYANAN V，et al. PoseCNN：A convolutional neural network for 6D object pose estimation in cluttered scenes [J]. Computer Science，2017，arXiv：1711.00199.

[6] KEHL W，MANHARDT F，TOMBARI F，et al. SSD-6d：Making rgb-based 3D detection and 6D pose estimation great again[C]. Proceedings of the IEEE International Conference on Computer Vision，2017：1521-1529.

[7] CHEN T，GU D. CSA6D：Channel-spatial attention networks for 6D object pose estimation[J]. Cognitive Computation，2022，14(2)：702-713.

[8] OBERWEGER M，RAD M，LEPETIT V. Making deep heatmaps robust to partial occlusions for 3D object pose estimation [C]. Proceedings of the European Conference on Computer Vision (ECCV)，2018：119-134.

[9] 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]. Proceedings of the IEEE International Conference on Computer Vision，2017：3828-3836.

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

[11] 黄晨，高岩. 结合通道注意力的特征融合多人位姿估

计算法[J]. 小型微型计算机系统，2021，42(1)：142-146.

[12] 刘素行，吴媛，张军军. 基于 YOLO v3 的交通场景目标检测方法[J]. 国外电子测量技术，2021，40(2)：116-120.

[13] 韩玉洁，曹杰，刘琨，等. 基于改进 YOLO 的无人机对地多目标检测[J]. 电子测量技术，2020，43(21)：19-24.

[14] ZHANG Q L，YANG Y B. Sa-net：Shuffle attention for deep convolutional neural networks[C]. IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP)，2021：2235-2239.

[15] LEE B，KU B，KIM W，et al. Feature sparse coding with CoordConv for side scan sonar image enhancement[J]. IEEE Geoscience and Remote Sensing Letters，2020，99：1-5.

[16] GOMES R，ROZARIO P，ADHIKARI N. Deep Learning optimization in remote sensing image segmentation using dilated convolutions and ShuffleNet[C]. IEEE International Conference on Electro Information Technology (EIT)，2021：244-249.

[17] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2018：7132-7141.

[18] 徐印赟，江明，李云飞，等. 基于改进 YOLO 及 NMS 的水果目标检测[J]. 电子测量与仪器学报，2022，36(4)：114-123.

[19] 王平，周雪峰，安爱民，等. 一种鲁棒且线性的 PnP 问题求解方法[J]. 仪器仪表学报，2020，41(9)：271-280.

[20] 包志强，邢瑜，吕少卿，等. 改进 YOLO V2 的 6D 目标姿态估计算法[J]. 计算机工程与应用，2021，57(9)：148-153.

作 者 简 介

党选举，博士，教授，博士生导师，主要研究方向为机器视觉、机器人智能控制等。

E-mail：xjd69@163.com

李启煌，硕士研究生，主要研究方向为机器视觉。

E-mail：lqh646942825@163.com