文章编号　1004-924X(2026)16-2595-17

不确定性感知的RGB图像6D物体位姿估计方法

李露帆[1]，党建武[1,2,3]，雍　玖[1,2,3*]
(1. 兰州交通大学 电子与信息工程学院，甘肃 兰州 730070；
2. 兰州交通大学 轨道交通信息与控制国家级虚拟仿真实验教学中心，甘肃 兰州 730070；
3. 兰州大成科技股份有限公司，甘肃 兰州 730070)

摘要：针对复杂背景与弱纹理条件下6D物体位姿估计精度受限的问题，本文提出了一种不确定性感知的位姿估计框架。首先，在特征提取阶段于编码器-解码器结构中引入多尺度感受野增强模块，通过融合不同卷积尺度与空洞率的上下文信息，提升网络对弱纹理目标的多尺度表征能力。其次，设计像素级置信度感知投票机制，对关键点密集单位向量场预测结果进行置信度建模，并在关键点投票过程中对像素贡献进行自适应加权，以提高关键点定位的鲁棒性。最后，将关键点定位结果的不确定性显式引入姿态求解过程，构建不确定性感知的PnP位姿估计框架，以减弱噪声关键点对姿态估计的影响。在LineMOD，Occlusion LineMOD及YCB-Video数据集上的实验结果表明，与BB8，DenseFusion，PoseCNN等多种方法相比，所提方法在位姿估计精度方面取得显著提升，平均距离(ADD)指标平均提高约7%，并在复杂背景与弱纹理场景下表现出良好的鲁棒性。实验结果充分验证了该方法的有效性与复杂场景适应能力。

关　键　词：物体位姿估计；关键点定位；像素级置信度投票；不确定性感知；PnP位姿求解
中图分类号：TP391　　**文献标识码**：A
doi：10.37188/OPE.20263416.2595　　**CSTR**：32169.14.OPE.20263416.2595

Uncertainty-aware monocular RGB 6D object pose estimation method

LI Lufan[1], DANG Jianwu[1,2,3], YONG Jiu[1,2,3*]

(1. *School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*;
2. *National Virtual Simulation Experimental Teaching Center for Rail Transit Information and Control, Lanzhou Jiaotong University, Lanzhou* 730070, *China*;
3. *Lanzhou Dacheng Technology Co., Ltd., Lanzhou* 730070, *China*)
* *Corresponding author, E-mail: yongjiu@mail.lzjtu.cn*

Abstract: To address the limited accuracy of 6D object pose estimation under complex backgrounds and weak-texture conditions, an uncertainty-aware 6D object pose estimation framework was proposed. Firstly, a multi-scale receptive field enhancement module was introduced into the encoder - decoder architec-

收稿日期：2026-04-29；**修订日期**：2026-06-30.
基金项目：国家自然科学基金项目(No. 62367005)；铁路青年人才托举工程项目(第七届)；中国高等教育学会高校数字思政精品项目(No. GJXHSZSZZY047)；兰州交通大学“科研反哺教学典型案例”教学改革研究项目(No. JGZ202503)

ture at the feature extraction stage. By integrating contextual information with different convolutional scales and dilation rates, the network was able to effectively improve the multi-scale feature representation capability for weak-texture objects. Secondly, a pixel-wise confidence-aware voting mechanism was designed to model the reliability of the predicted dense keypoint unit vector fields, and the contribution of each pixel was adaptively weighted during the voting process, thereby enhancing the robustness and accuracy of keypoint localization. Finally, the uncertainty of keypoint localization was explicitly incorporated into the pose estimation stage by constructing an uncertainty-aware PnP framework, which effectively reduced the influence of noisy keypoints on pose estimation results. Experimental results on the LineMOD, Occlusion LineMOD, and YCB-Video datasets demonstrate that, in comparison with several representative methods such as BB8, DenseFusion, and PoseCNN, the proposed method achieves significant improvements in pose estimation accuracy, with an average increase of approximately 7% in the Average Distance (ADD) metric, and exhibits strong robustness under complex background and weak-texture scenarios. These results demonstrate the effectiveness, robustness, and adaptability of the proposed method under complex background and weak-texture conditions.

Key words: 6D object pose estimation; keypoint localization; pixel-wise confidence-aware voting; uncertainty-aware modeling; PnP pose estimation

1 引　言

6D物体位姿估计是具身智能领域的核心任务,其目标是在二维图像中同时恢复物体在三维空间中的旋转与平移参数,该技术在机器人抓取[1-3]、自动驾驶[4]、工业装配[5-6]以及增强现实[7-8]等领域具有重要应用价值。随着实际应用场景对复杂环境感知需求的不断提升,亟须提升6D物体位姿估计的精度与可靠性。然而,在复杂背景、弱纹理以及遮挡等非理想条件下,目标区域的判别性特征往往不足,导致位姿估计结果易受到干扰,从而显著降低系统性能。因此,如何在复杂环境中实现稳定且准确的6D位姿估计,已成为当前研究的关键问题。

现有6D物体位姿估计方法主要包括基于传统几何的方法和基于深度学习的方法。传统几何方法通常依赖关键点检测、特征匹配以及投影几何模型,通过构建二维图像与三维模型之间的几何约束恢复物体姿态。在光照稳定、背景简单的理想条件下,该类方法能够获得较高精度,但在光照变化、遮挡及复杂背景干扰等情况下,其特征匹配稳定性显著下降,鲁棒性与环境适应能力受到限制。随着深度学习技术的发展,研究者开始利用神经网络对图像特征进行端到端建模,从而提升复杂场景下的位姿估计性能。Han[9]等人提出跨模态多尺度特征融合网络,通过融合不同模态与尺度的特征信息增强模型对目标的表征能力,从而提高位姿估计性能;Li[10]等人引入不确定性建模策略,通过刻画关键点预测结果的可靠性并将其引入姿态求解过程,从而缓解误差传播问题并提升位姿估计精度。尽管上述方法在特征表达与模型建模方面取得了一定进展,但在弱纹理、遮挡及复杂背景条件下,仍存在特征表达能力不足与预测结果不稳定的问题。

在基于深度学习的6D位姿估计方法中,关键点建模、特征融合以及姿态求解等方面已成为研究热点。近年来,Donad[11]等人提出基于关键点投票与可微RANSAC的位姿估计方法,有效提升了复杂场景下姿态求解的鲁棒性。随着多模态感知技术的发展,Hong[12]等提出基于Transformer的多模态特征融合方法,通过自注意力机制实现RGB与点云等多源信息的深度交互,有效提升了模型在复杂环境下的位姿估计性能。此外,Song[13]等针对遮挡场景下关键点易失效的问题,提出基于RGB-D特征融合的位姿估计方法,通过引入深度信息增强关键点表达能力,从而提升遮挡情况下的估计精度。进一步地,An[14]等利用层次化特征Transformer结构对多模态信息进行建模,实现跨层级特征交互与融合,提升了模型的整体表达能力与位姿估计精度。尽管

上述方法取得了显著进展，但在复杂背景与弱纹理条件下，关键点预测仍易受到噪声与遮挡的影响，不同关键点定位精度差异较大，从而限制了整体位姿估计性能的进一步提升。

综上所述，现有6D物体位姿估计方法虽然在建模、关键点定位及姿态求解等方面取得了一定进展，但在复杂背景与弱纹理条件下仍存在特征表达能力不足、关键点预测稳定性较差以及姿态求解易受噪声干扰等问题。针对上述问题，所提方法的主要工作有：(1)在特征提取阶段引入RFB_a(多尺度感受野增强模块)，通过融合不同尺度与不同空洞率的上下文信息，有效提升网络对弱纹理目标的多尺度表征能力；(2)构建像素级置信度感知投票机制，通过对关键点密集单位向量场进行置信度建模，实现像素贡献的自适应加权，从而提高关键点定位的准确性；(3)提出不确定性感知PnP位姿求解方法，将关键点定位结果的不确定性引入姿态优化过程，对不可靠关键点进行自适应抑制，从而提升复杂场景下6D位姿估计的整体精度。

2　本文方法

针对6D位姿估计过程中存在的特征表达能力不足、关键点预测不稳定以及姿态求解对噪声敏感等问题，所提方法由：RFB_a特征增强模块、像素级置信度感知投票模块以及不确定性感知PnP模块三部分组成。整体模型结构，如图1所示。

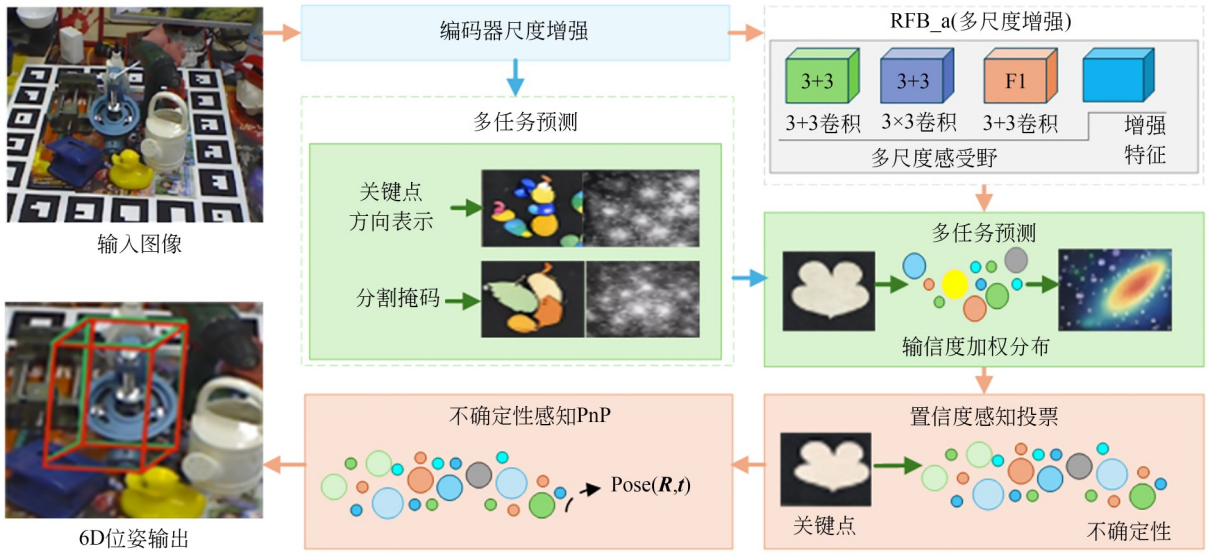


图1　网络整体结构

Fig. 1　Overall network architecture diagram

图1给出了所提方法的整体流程。首先，输入RGB图像经过编码器—解码器网络进行特征提取，并在编码器与解码器之间引入RFB_a多尺度感受野增强模块，以提升复杂背景与弱纹理场景下的特征表达能力。基于增强后的共享特征，网络分别通过关键点方向向量场分支、目标语义分割分支以及像素级置信度分支进行多任务预测。接着，利用分割掩码筛选前景像素，并结合像素级置信度信息对关键点方向约束进行加权投票，获得关键点二维位置及其对应的不确定性信息。最后，将关键点定位结果及其不确定性引入PnP优化框架，实现物体6D位姿的鲁棒估计。在上述整体流程的基础上，本文将在后续各小节中分别对RFB_a特征增强模块、像素级置信度感知投票模块以及不确定性感知PnP模块进行详细阐述。

2.1　RFB_a特征增强模块

在弱纹理且背景复杂的工业场景中，目标物体表面往往缺乏显著的局部纹理特征，导致基于单一尺度卷积的特征提取方式难以有效刻画目

标结构信息，从而引起特征判别性不足及边界模糊等问题。尤其在存在遮挡、反光及尺度变化的情况下，局部感受野难以覆盖目标整体区域，进一步影响后续语义分割与关键点方向预测的准确性。为此，本文在PVNet编码器—解码器结构之间引入RFB_a（Receptive Field Block-a）多尺度感受野增强模块，对中高层特征进行上下文增强，该模块结构如图2所示。

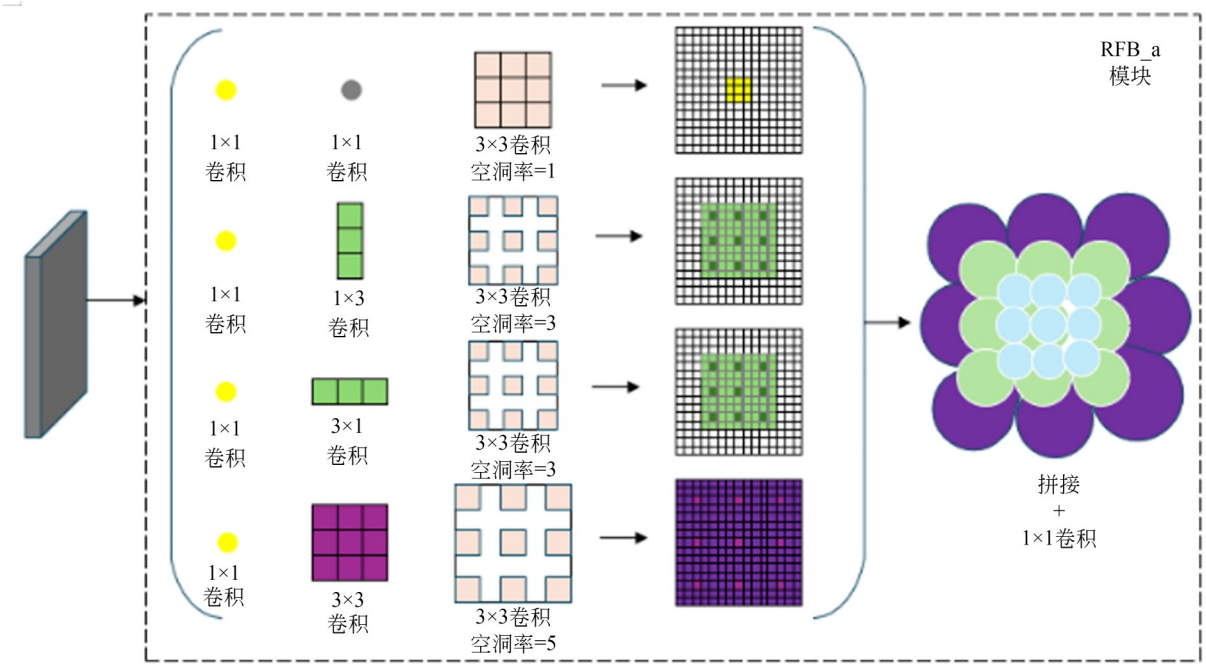


图2　RFB_a模块

Fig. 2　RFB_a module

RFB_a模块通过多分支并行卷积结构对输入特征进行多尺度建模，在保持特征图空间分辨率不变的前提下扩大有效感受野。设输入特征映射为F，具体为：

$$F=\boldsymbol{R}^{H\times W\times C}. \tag{1}$$

RFB_a首先将输入特征F分别送入多个并行卷积分支，每个分支通过不同卷积组合对特征进行变换，以捕获不同尺度与不同方向的上下文信息。该设计使网络能够在单一特征层中同时建模局部细节与全局结构，有效缓解弱纹理条件下特征表达不足的问题。

在图2中，RFB_a模块由四个相互独立的卷积分支组成。各分支首先通过1×1卷积对输入特征进行通道压缩与线性映射，以降低后续计算复杂度并提升特征变换效率。其中，两条分支进一步采用1×3与3×1的分离卷积结构，以增强对方向性结构特征的表达能力，其余分支则采用标准1×1或3×3卷积，用于补充局部空间信息。

在完成基础卷积操作后，各分支分别引入不同空洞率的3×3空洞卷积（dilated convolution），以构建多尺度上下文感受野。本文设置空洞率为r，具体为：

$$r=\{1,3,3,5\}. \tag{2}$$

从而使不同分支在不降低特征图分辨率的前提下感知不同尺度范围的上下文信息。对于卷积核大小为k的空洞卷积，其等效感受野可表示为k_{eff}，具体为：

$$k_{\mathrm{eff}}=k+(k-1)(r-1). \tag{3}$$

由此可见，在相同核尺寸条件下，增大空洞率能够显著扩展感受野范围，使网络在局部纹理缺失或被遮挡时仍可利用更大范围的语义信息进行判别。

各分支输出的多尺度特征在通道维度进行拼接，得到融合特征表示F_{cat}。随后，通过1×1卷积对拼接后的特征进行通道融合与维度修复，生成RFB_a的输出特征F_{rfb}。为稳定网络训练并保留主干网络的原始语义信息，RFB_a模块采用残

差连接方式将输入特征与增强后的特征进行融合，其形式为F_{out}，具体为：

$$F_{\text{out}}=F_{\text{rfb}}+F. \tag{4}$$

该残差结构不仅有助于缓解深层网络中的梯度消失问题，还能够保留输入特征中的原始语义信息，使增强后的多尺度特征与原始特征形成互补表达。同时，该设计使RFB_a模块能够无缝嵌入现有编码器-解码器框架，在保证网络收敛稳定性的同时提升特征融合效果。

通过多尺度感受野增强，RFB_a模块能够在中高层特征空间中同时捕获局部细节信息与大范围上下文语义，使网络在弱纹理与复杂背景条件下获得更加稳定且具有判别性的特征表示。增强后的特征有助于提升语义分割分支对目标区域的完整性与边界一致性，同时增强向量场分支在遮挡区域和纹理缺失区域的方向预测稳定性。作为关键点投票与不确定性感知PnP位姿求解的前端特征基础，RFB_a模块为后续像素级置信度感知投票过程提供了更可靠的前景特征支持，从而在整体上提升了6D位姿估计的精度与鲁棒性。

2.2　像素级置信度感知投票模块

在获得编码器-解码器网络输出的共享特征表示后，网络采用多分支预测结构对前景像素进行联合建模，分别输出关键点的密集单位向量场、目标语义分割结果以及像素级置信度信息。该多任务建模方式在统一特征空间中同时刻画几何约束、前景区域及预测可靠性，为关键点定位与姿态求解提供互补信息支撑。

其中，语义分割分支用于生成目标物体的像素级掩码，以限定后续投票过程中的有效前景区域，避免背景像素对几何投票过程的干扰；向量场分支用于预测每个前景像素指向各关键点的单位方向向量，从几何角度为关键点位置提供稠密约束；置信度分支则用于刻画对应像素预测结果的可靠程度，使网络能够区分高质量预测与潜在噪声预测。通过三者协同建模，构建像素级“方向预测—置信度评估—加权投票”的统一建模流程，从而提高关键点定位在复杂场景下的鲁棒性。像素级置信度感知投票模块整体框架如图3所示。

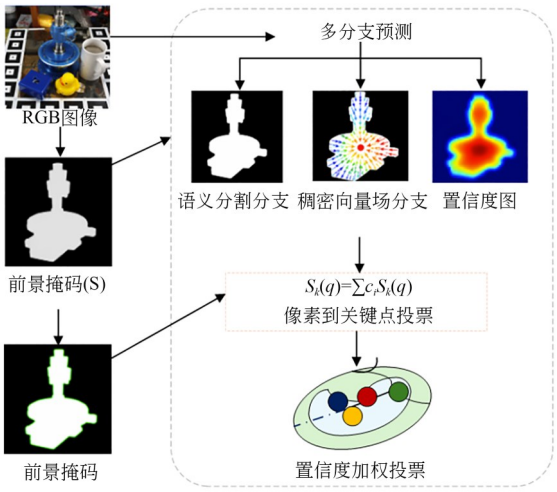


图3　像素级置信度感知投票模块

Fig. 3　Pixel-wise confidence-aware voting module

2.2.1　关键点定义与表示

在本文中，关键点是指在物体三维CAD模型表面预先选取的一组具有代表性的几何点，通常分布于边缘或结构特征显著的位置。作为连接二维图像观测与三维模型的中间表示，关键点用于建立稳定的2D-3D对应关系，从而为位姿求解提供几何约束。相较于直接回归旋转和平移参数，基于关键点的表示方式具有更好的可解释性和几何约束能力。该方式将位姿估计问题转化为几何对应关系求解，在遮挡、弱纹理及复杂背景条件下能够有效利用局部结构信息，提高估计稳定性。因此，关键点建模已成为RGB图像6D位姿估计的重要表示形式。

对于每一类目标物体，首先在其三维CAD模型表面预先选取一组具有代表性的三维关键点，具体为：

$$\kappa=\left\{\boldsymbol{X}_k\in\boldsymbol{R}^3 \mid k=1,2,\cdots,N\right\}, \tag{5}$$

其中：$\boldsymbol{X}_k$表示物体模型坐标系下的第k个三维关键点，N为关键点数量。该关键点集合在训练与推理阶段保持一致，以保证几何约束的一致性。

本文采用与PVNet相同的关键点定义方式，在CAD模型表面通过最远点采样（Farthest Point Sampling，FPS）自动选取若干具有代表性的三维关键点，以保证关键点在目标表面的空间分布均匀。训练阶段，根据已知CAD模型、真实位姿标注及相机内参，将三维关键点投影至图像平面生成对应的二维监督标签。在实际应用中，

关键点无需人工标注或直接观测，而是由训练完成的网络根据输入RGB图像自动预测获得，随后结合关键点投票与PnP求解实现目标位姿估计。

在给定真实位姿与相机内参条件下，三维关键点可投影至图像平面得到对应二维位置。不同于直接回归关键点坐标，本文采用像素级方向建模策略，通过预测前景像素指向关键点的单位方向向量场，并结合投票机制对关键点位置进行估计。该表示方式能够利用稠密前景像素提供几何约束，从而在遮挡及局部特征不稳定情况下保持较高的定位鲁棒性，并为后续置信度感知投票与PnP位姿求解提供基础。

本文所研究的复杂物体主要指在6D位姿估计过程中存在弱纹理、复杂背景干扰、部分遮挡、局部特征不明显或结构相似等情况的目标物体。这类目标难以仅依赖纹理信息完成稳定匹配，更需要利用目标的几何结构信息建立可靠的关键点约束。本文所采用的LineMOD，Occlusion LineMOD及YCB-Video数据集均包含上述类型目标，因此可用于验证所提方法在复杂场景下的位姿估计性能。

2.2.2　像素级关键点投票建模

设输入图像中第i个前景像素的二维坐标为u_i，具体为：

$$\boldsymbol{u}_i=\left[x_i,y_i\right]^{\mathrm{T}}\in\boldsymbol{R}^2. \tag{6}$$

网络预测该像素指向第k个关键点的单位方向向量，具体为：

$$\boldsymbol{v}_i^{(k)}=\frac{\boldsymbol{y}_k-\boldsymbol{u}_i}{\left\|\boldsymbol{y}_k-\boldsymbol{u}_i\right\|_2}, \tag{7}$$

$$\hat{\boldsymbol{v}}_i^{(k)}\in\boldsymbol{R}^2,\left\|\hat{\boldsymbol{v}}_i^{(k)}\right\|_2=1. \tag{8}$$

并同时预测其对应的像素级置信度权重，具体为：

$$c_i\in(0,1]. \tag{9}$$

该置信度权重用于刻画像素预测结果在几何意义上的可靠程度，其数值大小反映了当前像素在关键点定位过程中应被赋予的约束强度。

在语义分割掩码约束下，每个前景像素u_i可视为对第k个关键点位置提供一条二维几何约束，其形式为一条经过$\boldsymbol{u}_i$、方向为$\boldsymbol{v}_i^{(k)}$的射线。该建模方式将关键点定位由直接坐标回归转化为方向约束的聚合过程，使单个像素仅提供方向信息，从而降低局部预测误差对整体定位结果的影响。

对于任意关键点候选位置$\boldsymbol{q}\in\boldsymbol{R}^2$，其与像素约束之间的一致性通过垂直距离进行度量，具体为：

$$d\left(\boldsymbol{q},\boldsymbol{u}_i,\hat{\boldsymbol{v}}_i^{(k)}\right)=\left\|\left(\boldsymbol{q}-\boldsymbol{u}_i\right)\times\hat{\boldsymbol{v}}_i^{(k)}\right\|_2, \tag{10}$$

其中：$\times$表示二维向量叉积运算。该距离量化了候选位置$\boldsymbol{q}$与像素预测几何约束之间的偏离程度，距离越小表示候选位置越符合该像素所提供的方向信息。

基于上述一致性度量，第k个关键点的投票响应函数定义为$S_k(\boldsymbol{q})$，具体为：

$$S_k(\boldsymbol{q})=\sum_{i\in\Omega}c_i^{(k)}\exp\left(-\frac{d\left(\boldsymbol{q},\boldsymbol{u}_i,\hat{\boldsymbol{v}}_i^{(k)}\right)^2}{2\sigma^2}\right), \tag{11}$$

其中：Ω表示前景像素集合，σ为带宽参数，用于控制投票响应函数对距离误差的敏感程度。需要说明的是，式(11)中的$c_i^{(k)}$与式(9)中的定义一致，均表示第i个前景像素针对第k个关键点预测结果的置信度。因此，置信度同时与像素索引i和关键点索引k相关，并作为加权系数参与关键点投票过程。

与传统等权投票(如PVNet)策略不同，本文引入像素级置信度权重c_i，使不同像素对投票结果的贡献能够根据其预测可靠性进行自适应调节，从而在聚合过程中强化高质量约束并抑制噪声干扰。

最终，第k个关键点在图像平面中的二维位置通过最大化投票响应函数获得，具体为：

$$\hat{y}_k=\arg\max_{\boldsymbol{q}}S_k(\boldsymbol{q}). \tag{12}$$

上述过程构建了像素级置信度感知的关键点投票机制，通过对前景像素方向约束进行加权聚合，在图像平面内搜索几何一致性最优的位置作为关键点估计结果。为进一步展示关键点投票过程及其定位结果，本文在第2.3节给出了关键点投票样本分布、二维高斯拟合结果以及对应的位置坐标统计参数。相比直接回归关键点坐标或采用等权投票方法，该方法能够有效降低遮挡区域、边界不确定区域及噪声像素对关键点定位的影响，从而提升复杂场景下的定位精度与稳定性。

2.2.3　姿态监督损失与置信度加权学习

为约束像素级方向预测结果在整体几何空间中的一致性，并引导网络学习合理的置信度分布，本文引入基于三维模型点距离的姿态监督损失，对像素级预测结果进行间接优化。该损失通过比较预测姿态与真实姿态在三维空间中的作用效果，从全局几何层面评估预测精度，从而避免仅依赖二维误差所带来的不稳定性。

设从物体三维模型点集 $\boldsymbol{x}$ 中随机采样 M 个模型点，具体为：

$$\boldsymbol{x}_j \in \boldsymbol{R}^3, j = 1, 2, \cdots, M. \tag{13}$$

物体在相机坐标系下的真实姿态表示为旋转矩阵 $\boldsymbol{R}$ 与平移向量 $\boldsymbol{t}$ 的组合形式，具体为：

$$\boldsymbol{p} = [\boldsymbol{R}|\boldsymbol{t}], \boldsymbol{R} \in SO(3), \boldsymbol{t} \in \boldsymbol{R}^3. \tag{14}$$

在像素级建模框架下，不同前景像素对关键点定位及最终位姿估计的贡献存在明显差异。为刻画像素预测的可靠性，本文利用姿态一致性误差构造像素级监督信号，并据此对像素置信度进行加权学习。当某一像素的方向预测与真实姿态诱导的几何关系更一致时，该像素应被赋予更高置信度；反之，当其预测误差较大时，则在训练过程中被自适应降低权重。具体地，局部姿态估计 $\hat{\boldsymbol{P}}_i$ 由当前像素预测的关键点方向约束与对应三维关键点构建局部2D-3D对应关系，并通过PnP求解获得；随后依据该局部姿态与真实姿态之间的姿态一致性误差生成对应的监督置信度 c_i。在此过程中，记由当前像素方向预测所诱导的局部姿态估计为 $\hat{\boldsymbol{P}}_i$，具体为：

$$\hat{\boldsymbol{P}}_i = \left[\hat{\boldsymbol{R}}_i|\hat{\boldsymbol{t}}_i\right]. \tag{15}$$

该姿态估计表示由当前像素预测所对应的局部几何关系诱导得到的姿态估计，并非直接由网络回归得到，而是由像素级关键点方向预测结果所隐含的几何约束间接确定，用于衡量当前像素预测与整体真实姿态之间的几何一致性。

对于非对称物体，其姿态具有唯一解，因此采用模型点在三维空间中的欧氏距离作为姿态监督误差度量，像素级姿态损失定义为 $L_p(i)$，具体为：

$$L_p(i) = \frac{1}{M}\sum_{j=1}^{M}\left\| \boldsymbol{R}x_j + t - \left(\hat{\boldsymbol{R}}_i + \hat{\boldsymbol{t}}_i\right)\right\|_2. \tag{16}$$

该损失刻画了真实姿态与预测姿态作用下模型点在三维空间中的平均偏差程度，从整体几何层面反映预测结果的一致性。

对于对称物体，由于其在某些旋转条件下存在多解姿态，仅使用欧氏距离会引入错误的监督信号。为消除对称性带来的姿态歧义，本文采用最近点距离形式定义姿态损失函数，具体为：

$$L_p(i) = \frac{1}{M}\sum_{M}^{1}\min_{x^k \in \chi}\left\| \boldsymbol{R}x_j + t - \left(\hat{\boldsymbol{R}}_i + \hat{\boldsymbol{t}}_i\right)\right\|_2. \tag{17}$$

该损失通过为每个预测点匹配模型点集中距离最近的对应点，有效消除对称结构引起的姿态歧义。

在此基础上，为进一步引导网络区分不同像素预测结果的可靠性，引入像素级置信度加权机制，整体训练损失函数定义为 L，具体为：

$$L = \frac{1}{N}\sum_{i=1}^{N}\left(c_i L_p(i) - \lambda \log c_i\right), \tag{18}$$

其中：N 表示参与训练的前景像素数量，c_i 表示像素 i 的预测置信度，λ 为平衡系数，用于调节姿态误差项与置信度正则项之间的权重关系。

与传统方法仅利用方向预测进行投票不同，本文将像素级置信度显式引入训练过程，使网络能够根据姿态一致性误差自适应调整像素权重，从而在优化过程中强化高质量约束并抑制不可靠预测。

该损失设计使模型不仅能够学习符合几何约束的方向预测结果，同时能够在像素级别对预测可靠性进行建模，从而为后续关键点投票与PnP位姿求解提供稳定且可信的先验信息。

2.3　不确定性感知模块

在本文中，不确定性主要来源于关键点方向预测误差与像素级置信度分布。根据不确定性在位姿估计流程中的作用，本文主要从关键点定位与像素预测两个层面对其进行建模：关键点不确定性用于刻画关键点在二维图像中的定位稳定性，像素级不确定性用于描述前景像素预测方向与真实方向一致性的可信度。推理阶段，关键点不确定性通过加权投票过程进行传播，并进一步引入PnP位姿求解过程，从而影响最终6D位姿估计结果。上述分析明确了不确定性的来源及其在位姿估计流程中的作用，为后续像素级置信度感知投票机制与不确定性感知PnP位姿求解提供了理论依据。

通过像素级置信度感知投票机制，本文为每个关键点在图像平面中获得其二维空间分布。

具体而言，根据投票阶段得到的关键点候选位置样本，可统计其均值与协方差，并将第k个关键点的投票结果近似建模为二维高斯分布，其均值μ_k表示关键点的期望位置，协方差矩阵Σ_k用于刻画关键点定位的不确定性。该建模不仅提供关键点位置估计，还能够量化其空间分布的离散程度，为后续位姿求解提供可靠的权重依据。

为进一步验证所提不确定性建模机制对关键点定位结果的表征能力，选取LineMOD数据集中典型目标进行可视化分析。通过像素级置信度感知投票机制获得关键点候选位置样本，并采用二维高斯分布对其空间分布进行建模。关键点投票结果的空间分布特征能够通过均值位置与协方差矩阵进行有效描述，其中协方差矩阵反映了关键点定位结果的不确定性大小。图4给出了关键点定位结果的不确定性可视化分析过程，表1给出了对应的不确定性统计参数。

如图4所示，通过像素级置信度感知投票机制获得关键点候选位置样本，图4(b)展示了关键点投票样本在图像平面中的空间分布情况，大量投票结果集中于目标关键区域附近。进一步地，采用二维高斯分布对投票结果进行建模，图4(c)给出了对应的高斯拟合结果及协方差椭圆。其中，均值位置可作为关键点最终估计结果。表1给出了对应的关键点定位统计参数，其中关键点估计坐标为(310.42，186.57)pixels。上述结果验证了所提投票机制能够有效完成关键点位置搜索与不确定性建模，为后续不确定性感知PnP位姿求解提供可靠的二维观测信息。

在此基础上，将关键点的空间不确定性显式引入姿态求解过程，构建不确定性感知的PnP(Perspective-n-Point)位姿估计框架。不确定性感知PnP位姿估计流程如图5所示：

设物体模型中第k个三维关键点在模型坐标系下的坐标为$\boldsymbol{X}_k\in\boldsymbol{R}^3$，其在相机坐标系下经旋转平移变换后的投影位姿表示为$\tilde{x}_k$，具体为：

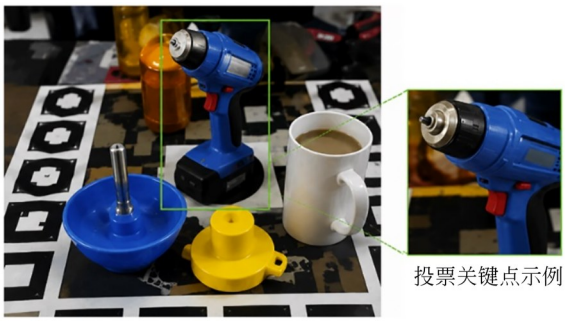
$$\tilde{x}_k=\pi(RX_k+t), \tag{19}$$

其中：$\boldsymbol{R}\in SO(3)$和$\boldsymbol{t}\in\boldsymbol{R}^3$分别表示物体的旋转矩阵和平移向量，$X_k$是关键点的3D坐标，$\tilde{x}_k$是$X_k$的2D投影，$\pi(\cdot)$为相机的透视投影函数。

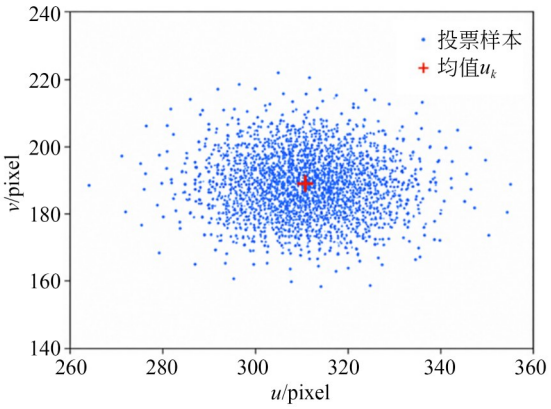
为刻画不同关键点定位精度的差异，本文采用马氏距离形式对关键点的重投影误差进行加权建模，最终的位姿求解被表示为如下优化问题，具体为：

$$\min_{R,t}\sum_{k=1}^{K}(\tilde{x}_k-u_k)^{\mathrm{T}}\sum_k^{-1}(\tilde{x}_k-u_k). \tag{20}$$

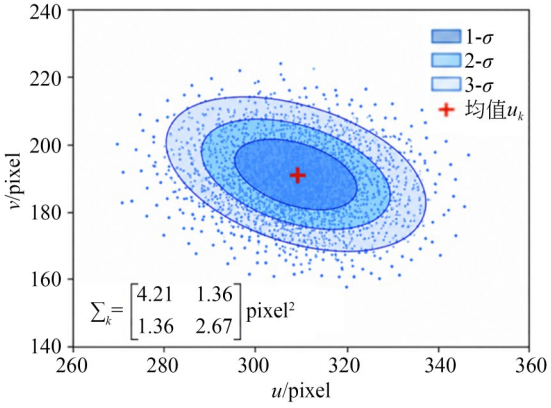
通过协方差矩阵Σ_k对不同关键点的约束进行自适应加权，使得定位不确定性较大的关键点



(a) 原始RGB图像与目标区域(LineMOD-Driller)
(a) Original RGB image and target region（LineMOD-Driller）



(b) 关键点投票样本分布(图像平面坐标)
(b) Keypoint voting sample distribution (image-plane coordinates)



(c) 二维高斯拟合结果(协方差椭圆)
(c) Two-dimensional Gaussian fitting result (covariance ellipse)

图4　关键点定位结果的不确定性可视化分析
Fig. 4　Visualization of keypoint localization uncertainty

表 1　关键点定位结果不确定性统计参数

Tab. 1　Statistical parameters of keypoint localization uncertainty

参数名称	数值
均值位置	(310.42,186.57)pixels
横向标准差	2.05 pixels
纵向标准差	1.63 pixels
协方差行列式	9.29
不确定性水平	中等

在优化过程中被赋予较小权重，而高置信度关键点对位姿估计结果起主导作用。相比传统PnP方法中对所有关键点等权处理的策略，该不确定性感知建模能够有效抑制噪声关键点及局部错误匹配对姿态求解结果的干扰。

在实际求解过程中，首先利用标准EPnP方法对位姿参数进行初始化[15]，随后采用非线性最小二乘优化算法对上述目标函数进行迭代求解，从而获得最终的物体6D位姿估计结果。

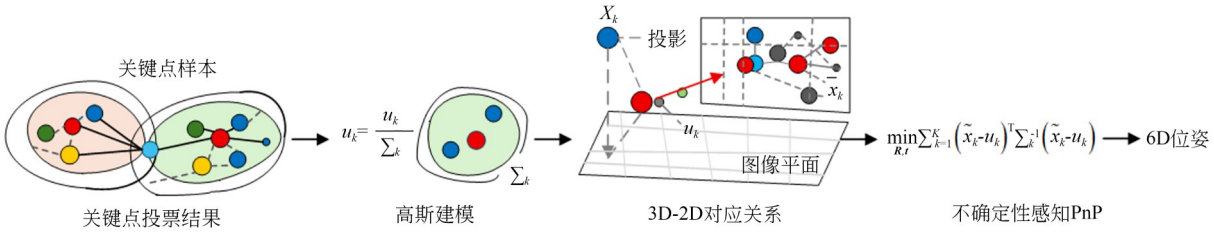


图 5　不确定性感知模块

Fig. 5　Uncertainty-aware module

在引入关键点不确定性建模的基础上，有必要对不确定性感知PnP的计算复杂度与收敛特性进行讨论。设每个目标物体选取K个三维关键点，则优化变量为旋转$R\in SO(3)$与平移$t\in R^3$，采用最小参数化后变量维度为6。每个关键点对应二维重投影误差，因此残差向量维度为$2K$，雅可比矩阵规模为$2K\times 6$。采用Levenberg-Marquardt(LM)进行非线性最小二乘迭代时，主要开销来源于雅可比构建与解线性方程组，单次迭代的时间复杂度随K近似线性增长，即$O(K)$。由于实际应用中K通常取8～9，且LM迭代次数一般小于10次，因此该模块不会对整体推理速度造成显著负担。

在收敛性方面，本文首先使用EPnP提供位姿初值，使优化从合理解域启动；随后通过协方差矩阵Σ_k对重投影误差进行自适应加权，等价于加权最小二乘形式，可对高不确定性关键点进行“软抑制”，从而降低异常观测对目标函数的干扰。为保证数值稳定性，实践中可在Σ_k上加入微小对角正则项以确保正定性，使目标函数连续可导并满足LM收敛条件。实验表明，该策略在复杂背景与遮挡场景下能够显著降低优化震荡与发散风险，获得更稳定一致的位姿解。

2.4　损失函数

为实现语义分割、关键点方向预测与置信度学习的协同优化，本文采用多任务联合训练策略，将各分支损失进行加权组合。整体损失函数具体为：

$$\mathcal{L}=\alpha\mathcal{L}_{\text{seg}}+\beta\mathcal{L}_{\text{vec}}+\gamma\mathcal{L}_{\text{conf}},\tag{21}$$

其中：$\mathcal{L}_{\text{seg}}$，$\mathcal{L}_{\text{vec}}$和$\mathcal{L}_{\text{conf}}$分别表示语义分割损失、关键点方向向量场损失以及置信度损失，α,β,γ为权重系数。语义分割损失用于约束网络生成目标前景区域，为后续关键点投票提供有效的前景像素约束；关键点方向向量场损失用于监督像素到关键点的单位方向预测，以保证关键点几何约束的准确性；置信度损失则依据像素预测结果与真实姿态之间的一致性误差进行构建，引导网络学习不同像素预测结果的可靠程度，从而提高关键点定位与位姿求解的稳定性。

在训练阶段，上述三个分支所需监督信息均由已知CAD模型及对应位姿标注自动生成。其中，语义分割标签通过将目标CAD模型按照真实位姿投影至图像平面获得；关键点方向向量标签由CAD模型关键点投影位置与前景像素坐标计算得到；置信度标签则依据像素局部预测所对应的姿态一致性误差构建，用于指导网络学习像

素级预测可靠性。通过上述方式，可以实现语义分割、关键点方向预测以及置信度学习的联合监督训练。

需要说明的是，本文方法依赖于目标物体的CAD模型及对应位姿标注，因此主要适用于具有CAD模型先验的已知物体位姿估计任务。当目标CAD模型不可获取时，相关监督信息难以构建，从而限制了方法在未知类别物体场景中的直接应用。然而，在工业装配、机器人抓取以及增强现实等典型应用场景中，目标CAD模型通常是可获取的，因此该方法仍具有较好的工程应用价值。

由于语义分割、方向向量场回归与置信度学习三项任务对网络优化的重要程度存在差异，因此本文参考PVNet及相关关键点回归方法的参数设置，并结合预实验结果对损失权重进行调节。最终设置损失权重系数$\alpha=1$，$\beta=1$，$\gamma=0.5$，同时进一步说明各权重系数的作用及设置原因，使损失函数设计更加完整。其中，分割损失与向量场回归损失采用相同权重，以保证目标区域提取与关键点方向预测的协同优化；置信度损失作为辅助监督项，赋予较小权重，以避免其对主任务优化过程产生过度影响。预实验结果表明，该参数配置能够兼顾语义分割、关键点回归及置信度学习效果，从而获得较优的综合性能。

对于语义分割分支，采用像素级交叉熵或Dice损失以增强前景区域的完整性与边界一致性；对于向量场分支，采用方向一致性损失对单位向量进行监督，通过最小化预测方向与真实方向之间的角度误差或余弦距离，保证像素级几何指向的准确性。

置信度分支的学习与姿态监督紧密耦合：基于式(18)，将像素级置信度引入加权损失，使网络能够根据姿态一致性误差自适应调整像素权重：误差较大的预测对应较低置信度，从而在训练过程中抑制不可靠像素的影响。

3 实验结果分析

为验证所提方法的有效性，本文在LineMOD，Occlusion LineMOD和YCB-Video三个公共数据集上开展实验。

3.1 实验数据与实验环境

LineMOD数据集包含13类刚性物体，每类物体约1 200张图像，其中eggbox和glue为对称物体，总计15 783张图像。该数据集涵盖多视角变化、尺度变化以及不同复杂程度的背景与光照条件，具有较强的多样性与挑战性。由于提供精确的位姿标注，LineMOD数据集被广泛用于评估单目标场景下6D物体位姿估计算法的性能。

Occlusion LineMOD是LineMOD数据集的子集，专门用于评估算法在严重遮挡场景下的性能表现。该数据集包含8类目标物体，共1435张图像。在该数据集中，目标物体通常仅部分可见，显著增加了关键点定位与姿态求解的难度，因此被广泛用于检验算法在复杂遮挡条件下的鲁棒性。

YCB-Video数据集是基于YCB物体集构建的6D物体位姿估计数据集，由真实室内场景中的视频序列组成。该数据集涵盖了从92段视频中采集得到的21种不同物体，总计包含133 827张图像。数据集中的场景具有复杂背景、光照变化以及多物体同时存在等特点，对算法在真实环境下的位姿估计精度与鲁棒性提出了较高要求，被广泛用于评估多目标及复杂场景下6D位姿估计方法的综合性能。

此外，上述数据集均提供目标物体对应的CAD模型、相机内参以及标准6D位姿标注信息。本文采用数据集官方提供的位姿标注作为真值(Ground Truth)，用于训练监督及不同算法之间的定量性能评估。其中，位姿真值由目标物体相对于相机坐标系的旋转矩阵和平移向量组成，可用于建立二维图像观测与三维模型之间的对应关系，从而为ADD和ADD-S等评价指标的计算提供依据。

实验在Linux操作系统下进行，硬件平台为NVIDIA RTX 2080Ti(11 GB)GPU，系统内存为80 GB。模型基于Python 3.7.9环境，采用PyTorch深度学习框架实现。在训练阶段，采用Adam优化器对网络参数进行优化，初始学习率设置为1×10^{-4}，权重衰减系数为5×10^{-4}，学习率采用分段衰减策略，在第80与120个epoch分别

衰减至原来的0.5倍，总训练轮数为140个epoch，batch size设置为4。为提升模型的泛化能力，在训练过程中引入了多种数据增强策略，包括随机旋转、尺度缩放、裁剪、随机高斯噪声以及运动模糊等。此外，在训练后期引入迭代细化机制，当模型收敛至稳定阶段后自动进入细化优化过程，以进一步提升位姿估计精度。

3.2　评价指标

为客观评估6D位姿估计性能，使用平均距离（ADD）和平均最近点距离（ADD-S）两种常见指标进行客观评价。

$$D_{\mathrm{ADD}}=\frac{1}{m}\sum_{x\in M}\left\|(\boldsymbol{R}x+\boldsymbol{t})-\left(\hat{\boldsymbol{R}}x+\hat{\boldsymbol{t}}\right)\right\|, \tag{22}$$

$$D_{\mathrm{ADD\text{-}S}}=\frac{1}{m}\sum_{x_1\in M}\min_{x_2\in M}\left\|(\boldsymbol{R}x_1+\boldsymbol{t})-\left(\hat{\boldsymbol{R}}x_2+\hat{\boldsymbol{t}}\right)\right\|, \tag{23}$$

其中：M表示模型点集；m为模型点数量；x_1和x_2均表示模型点集M中的三维点，其中x_1表示真实模型点，x_2表示用于最近邻匹配的模型点；$\boldsymbol{R}$和$\boldsymbol{t}$分别表示真实的旋转矩阵和平移向量；$\hat{\boldsymbol{R}}$和$\hat{\boldsymbol{t}}$分别表示预测姿态的旋转矩阵和平移向量。式（22）和式（23）分别给出ADD与ADD-S评价指标的计算公式。

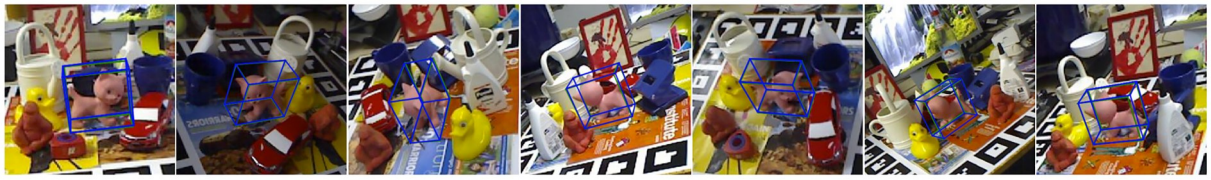
ADD指标衡量的是变换后的模型点的平均偏差是否小于物体直径的10%。当距离小于模型直径的10%时，则认为位姿估计正确。对于对称目标，使用ADD-S指标进行衡量，该指标衡量的是到最近模型点的平均距离误差，其中平均距离是根据最近的点距离计算。在YCB-Video数据集上进行评估时，使用ADD-S和AUC评估目标。其中AUC是精度阈值曲线下的面积，通过改变评定中的距离阈值得到。

3.3　准确性分析

3.3.1　LineMOD数据集的实验结果

图6是在LineMOD数据集上的可视化结果（彩图见期刊电子版）。其中，蓝色3D边框表示真实姿态，绿色3D边框表示预测姿态。蓝色和绿色的3D边框重合度越高，表示估计的位姿越准确。通过可视化结果可见，蓝色3D边框和绿色3D边框基本贴合，说明所提方法位姿估计表现良好，表明所提方法在该场景下具有较好的鲁棒性。

(a) Ape



(b) Cat

(c) Duck

图6　LineMOD数据集上的可视化结果

Fig. 6　Visualization results on the LineMOD dataset

表 2 是在 LineMOD 数据集上的实验结果。将本文算法的实验结果与 BB8[16]，DenseFusion[17]，PoseCNN[18]，TexPose[19]，Self6D＋＋[20]，Dual-Stream[21]，PVNet[22]和 OK-POSE＋＋[23]这 8 种具有代表性的算法的实验结果对比，均采用 ADD 指标进行评价，其中，以*标注的物体为对称物体，采用 ADD-S 指标进行评价。表 2 中加粗结果表示各类别中的最优性能，本文方法在多数物体类别上均取得最优或次优结果，在 bench vise，camera，can，driller，eggbox*，glue*，iron 及 lamp 等物体上表现尤为突出[24]，体现出良好的整体性能优势。由表 2 可以看出，本文算法对比其他 8 种算法平均精度分别提升了 7.4%，2.4%，8.1%，5.0%，11.1%，1.9%，10.4% 和 10.3%。

上述结果表明，所提方法在整体性能上具有明显优势。其主要原因在于，本文方法在特征表达、关键点定位及姿态求解等关键环节实现了协同优化，使模型能够在复杂背景与弱纹理条件下获得更加稳定且具有判别性的特征表示[25]，并有效提升关键点定位的可靠性，从而在整体上提高位姿估计的精度与鲁棒性。对于 ape，cat，duck 及 holepuncher 等物体，由于其尺寸较小或有效可见区域受限，导致判别性特征不足；同时，在 ADD 指标采用相对直径阈值的情况下，小尺度目标对应更严格的误差容忍范围，因此其性能提升相对有限。对于 camera，can 和 lamp 等典型弱纹理目标，其表面缺乏稳定纹理信息，关键点定位主要依赖局部结构与上下文特征。总体而言，所提方法在上述目标上均取得了较高的位姿估计精度，说明 RFB_a 模块提取的多尺度上下文信息能够有效弥补纹理特征不足的问题，提高弱纹理场景下的特征表达能力，从而增强模型对目标结构信息的感知能力，进一步提升位姿估计的准确性与稳定性。

表 2　LineMOD 数据集上精度结果对比

Tab. 2　Comparison of accuracy results on LineMOD dataset　(%)

Object	BB8	DenseFusion	PoseCNN	TexPose	Self6D＋＋	Dual-Stream	PVNet	OK-POSE＋＋	Ours
Ape	96.6	92.3	77.0	80.9	76.0	91.3	43.6	86.6	**96.7**
Bench vise	90.1	93.2	97.5	99.0	91.6	93.5	**99.9**	84.7	99.2
Camera	86.0	94.4	93.5	94.8	**97.1**	94.0	86.9	90.4	95.8
Can	91.2	93.1	96.5	99.7	**99.8**	94.3	95.5	87.0	**99.8**
Cat	**98.8**	96.5	82.1	92.6	85.6	95.8	79.3	90.6	95.4
Driller	80.9	87.0	95.0	97.4	**98.8**	92.9	96.4	92.6	97.3
Duck	92.2	92.3	77.7	83.4	56.5	**94.7**	52.6	83.8	88.6
Eggbox*	91.0	99.8	97.1	94.9	91.0	**99.9**	99.2	85.0	99.4
Glue*	92.3	**100.0**	99.4	93.4	92.2	99.9	95.7	88.7	99.5
Holepuncher	**95.3**	92.1	52.8	79.3	35.4	92.8	82.0	84.0	89.7
Iron	84.8	97.0	98.3	**99.8**	99.5	95.1	98.9	89.2	**99.8**
Lamp	75.8	95.3	97.5	98.3	97.4	94.6	**99.3**	83.3	**99.3**
Phone	85.3	92.8	87.7	78.9	91.8	94.0	92.4	77.9	**96.6**
Mean	89.3	94.3	88.6	91.7	85.6	94.8	86.3	86.4	**96.7**

3.3.2　Occlusion LineMOD 数据集的实验结果

图 7 是在 Occlusion LineMOD 数据集上的可视化结果(彩图见期刊电子版)。蓝色 3D 边框表示真实姿态，绿色 3D 边框表示预测姿态。可以看出，在存在明显遮挡的情况下，预测姿态与真实姿态仍保持较高一致性，三维边框在图像中的投影基本重合，表明所提方法在遮挡场景下能够准确恢复目标位姿。

表 3 是在 Occlusion LineMOD 数据集上的实验结果。将本文算法的实验结果与 SSPE[26]，Self6D＋＋，HybridPose[27]，NVR-Net[28]，OberWeger[29]，OK-POSE＋＋，RNNPose[30]和 PVNet 这 8 种算法的实验结果对比，采用与测试 LineMOD 数据集一样的指标。表 3 中加粗结果

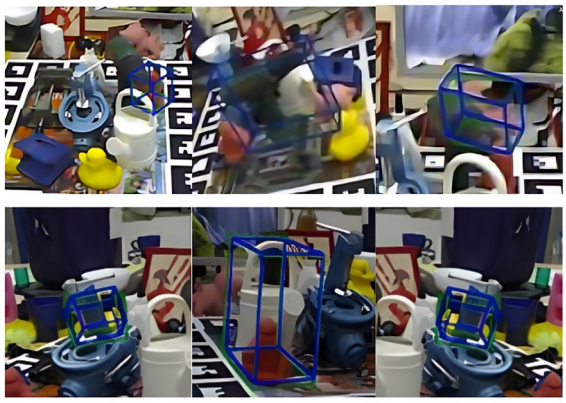


图7　Occlusion LineMOD数据集上的结果

Fig. 7　Visualization results on the Occlusion LineMOD dataset

表示各类别中的最优性能。由表3可以看出，本文方法在多数物体类别上取得了最优或次优结果，尤其在driller，eggbox及glue等物体上优势较为明显。从平均精度（mean）来看，本文方法达到66.6%，明显优于大多数对比方法，体现出较好的整体性能。

通过表3的实验数据可知，该算法对比这8种算法平均精度分别提升了23.3%，1.9%，19.1%，10.0%，36.2%，5.2%，5.9%和25.8%，在整体性能上具有明显优势。从具体类别来看，所提方法在can，driller，eggbox，glue及holepuncher等物体上均取得最优结果，说明该方法在严重遮挡条件下仍能够保持稳定的关键点定位能力。对于ape、cat和duck等物体，其性能相对较低，主要由于这些目标尺寸较小且在遮挡条件下可见区域进一步减小，导致有效特征信息不足，从而影响位姿估计精度。

结合图7可视化结果进一步分析，以driller目标为例，其钻头区域存在明显遮挡，同时部分关键结构仅保留有限可见区域。尽管如此，本文方法在driller类别上仍取得了当前最优结果，并能够获得较准确的位姿预测结果，说明像素级置信度感知投票机制能够有效降低遮挡区域不可靠像素对关键点定位的干扰。同时，不确定性感知PnP模块进一步增强了位姿求解过程的鲁棒性，从而提高了复杂遮挡场景下的位姿估计性能。

表3　Occlusion LineMOD数据集上精度结果对比

Tab. 3　Comparison of accuracy results on Occlusion LineMOD dataset　　(%)

Object	SSPE	Self6D++	HybridPose	NVR-Net	OberWeger	OK-POSE++	RNNPose	PVNet	Ours
Ape	19.2	59.4	20.9	43.1	17.6	**60.9**	37.2	15.8	46.8
Can	65.1	**96.5**	75.3	82.9	53.9	59.0	88.1	63.3	79.9
Cat	18.9	60.8	24.9	27.2	3.3	**64.8**	29.2	16.7	49.1
Duck	25.3	**92.0**	27.9	69.7	19.2	78.2	88.1	25.2	52.0
Driller	69.0	30.6	70.2	44.2	62.4	62.6	49.2	65.7	**83.8**
Eggbox	52.0	51.1	52.4	49.7	25.9	57.4	67.0	50.2	**74.3**
Glue	51.4	**88.6**	53.8	74.3	39.6	73.4	63.8	49.6	75.9
Holepuncher	45.6	38.3	54.2	61.7	21.3	34.9	62.8	39.7	**70.7**
Mean	43.3	64.7	47.5	56.6	30.4	61.4	60.7	40.8	**66.6**

3.3.3　YCB-Video数据集的实验结果

表4是在YCB-Video数据集上的实验结果，加粗结果表示各类别中的最优性能。将本文算法的实验结果与PoseCNN，DenseFusion对比，本文方法在ADD-S和ADD(S)指标上均取得更优性能，平均精度分别达到92.9%和88.7%，整体表现优于对比方法。从具体类别来看，所提方法在多数物体上均取得较高精度，尤其在弱纹理及形状复杂物体（如banana，scissors，clamp等）上表现出更稳定的估计效果，表明该方法具有良好的特征表达能力与几何建模能力。总体而言，所提方法不仅适用于单目标弱纹理场景，在复杂背景和多目标共存条件下同样具有较好的鲁棒性和复杂场景适应能力。

表 4　YCB-Video 数据集上精度结果对比

Tab. 4　Comparison of accuracy results on YCB-Video dataset　(%)

Object	PoseCNN		DenseFusion		Ours	
	ADD-S	ADD(S)	ADD-S	ADD(S)	ADD-S	ADD(S)
02 Master chef can	83.9	50.2	**95.3**	70.7	95.2	**77.4**
03 Cracker box	76.9	53.1	92.5	86.9	**93.8**	**91.3**
04 Sugar box	84.2	68.4	95.1	90.8	**97.0**	**94.1**
05 Tomato soup can	81.0	66.2	93.8	84.7	**95.7**	**89.1**
06 Mustard bottle	90.4	81.0	95.8	90.9	**96.7**	**94.2**
07 Tuna fish can	88.0	70.7	95.7	79.6	**96.1**	**89.8**
08 Pudding box	79.1	62.7	**94.3**	**89.3**	94.1	88.7
09 Gelatin box	87.2	75.2	**97.2**	**95.8**	97.0	95.2
10 Potted meat can	78.5	59.5	89.3	79.6	**91.2**	**83.0**
11 Banana	86.0	72.3	90.0	76.7	**93.2**	**92.0**
19 Pitcher base	77.0	53.3	**93.6**	**87.1**	92.2	87.0
21 Bleach cleanser	71.6	50.3	94.4	**87.5**	**95.5**	86.9
24 Bowl*	69.6	69.6	86.0	86.0	**87.8**	**87.8**
25 Mug	78.2	58.5	95.3	83.8	**96.9**	**92.5**
35 Power drill*	72.7	55.3	92.1	83.7	**95.5**	**92.3**
36 Wood block	64.3	64.3	**89.5**	**89.5**	89.3	89.3
37 Scissors	56.9	35.8	90.1	77.4	**93.1**	**89.1**
40 Large marker	71.7	58.3	**95.1**	**89.1**	94.2	85.0
51 Large clamp*	50.2	50.2	71.5	71.5	**78.3**	**78.3**
52 Extra large clamp*	44.1	44.1	70.2	70.2	**87.0**	**87.0**
61 Foam brick*	88.0	88.0	**92.2**	**92.2**	92.1	92.1
Mean	75.9	59.9	91.2	82.9	**92.9**	**88.7**

注：AUC 是将对称物体和非对称物体集成到同一评估中的指标，均采用 ADD-S 指标计算；其中以*标注的物体为对称物体，其他物体则是非对称物体；ADD(S)是对对称物体采用 ADD-S 指标，其他物体采用 ADD 指标的结果。

3.4　消融实验

为验证所提方法中各模块的有效性，在 LineMOD 数据集上进行消融实验，以 PVNet 为基线模型，在仅使用 RGB 输入的条件下，逐步引入 RFB_a 特征增强模块、像素级置信度感知投票模块以及不确定性感知 PnP 模块，分别用 B_1，B_2 和 B_3 表示。实验结果如表 5 所示。

表 5　所提模型消融实验比较

Tab. 5　Comparative analysis of model ablation experiments

Input	Module				ADD/%
	B_1	B_2	B_3	Baseline model	
√				√	88.1
√	√			√	92.7
√	√	√		√	95.2
√	√	√	√	√	96.7

由表 5 可见，基线模型的 ADD 准确率为 88.1%。在此基础上，引入 RFB_a 特征增强模块后，模型 ADD 准确率由 88.1% 提升至 92.7%，提高了 4.6%。说明 RFB_a 模块通过融合不同卷积尺度与空洞率的上下文信息，能够有效扩大网络感受野，增强目标局部细节与全局语义特征表达能力，从而提高复杂背景及弱纹理场景下的关键点定位精度和位姿估计性能。实验结果验证了多尺度上下文建模策略的有效性。进一步引入像素级置信度感知投票模块后，通过对像素预测结果进行可靠性建模，实现了对关键点定位过程中不稳定预测的自适应抑制，使 ADD 指标提升至 95.2%，说明该模块在提升关键点定位鲁

棒性方面具有显著作用。在此基础上，加入不确定性感知PnP模块，将关键点定位的不确定性显式引入位姿求解过程，通过对不同关键点进行加权优化，有效降低了噪声关键点对结果的影响，使最终精度进一步提升至96.7%。总体来看，各模块均对性能提升具有稳定贡献，且随着模块逐步叠加，模型性能呈现持续提升趋势，表明所提方法在特征增强、关键点定位及姿态求解三个层面形成了有效协同，从而显著提升了整体位姿估计精度与鲁棒性。

3.5　模型复杂度与推理效率分析

为进一步评估所提方法的计算开销与运行效率，本文从模型参数量（Params）、推理速度（FPS）以及单张图像平均推理时间（Runtime）三个方面对模型复杂度进行分析。选取经典RGB 6D位姿估计方法PVNet，PoseCNN以及RGB-D方法DenseFusion作为对比对象，在相同实验环境下统计各方法的参数量和推理速度，模型复杂度与推理效率实验结果如表6所示。

表6　模型复杂度与推理效率对比

Tab. 6　Comparison of model complexity and efficiency

Method	Params/M	FPS	Runtime/ms	ADD/%
PVNet	21.5	28.3	35.3	86.3
DenseFusion	42.1	16.4	61.0	94.3
PoseCNN	67.2	12.8	78.1	88.6
Ours	23.1	24.1	41.5	96.7

从表6可以看出，PVNet由于网络结构轻量，推理速度较快且运行时间较短；DenseFusion和PoseCNN由于网络结构较复杂，参数量和运行时间均明显增加。相比之下，本文方法在保持较低计算开销的同时，推理速度仍然较高，并在位姿估计精度（ADD指标）上取得明显提升。相较于PVNet，本文方法参数量仅增加1.6 M，单张图像平均推理时间增加6.2 ms，但ADD指标提升10.4%，表明所提方法能够以较小额外计算开销获得显著性能增益。综合来看，所提方法在精度与效率之间实现了良好的平衡，能够满足实际工业装配、机器人抓取及增强现实等应用场景对实时性的要求。

4　结　论

针对弱纹理条件下RGB图像6D物体位姿估计中关键点定位不稳定及姿态求解精度不足的问题，所提方法通过引入RFB_a特征增强模块、像素级置信度感知投票机制以及不确定性感知PnP模块，实现了特征表达、关键点定位与姿态求解过程的协同优化。在LineMOD，Occlusion LineMOD及YCB-Video数据集上的实验结果表明，所提方法在位姿估计精度和鲁棒性方面均优于PVNet等对比方法，尤其在弱纹理与遮挡场景下表现出更好的适应能力。消融实验进一步验证了各组成模块的有效性，说明联合利用多尺度特征增强、预测置信度建模与不确定性感知姿态求解能够有效提高关键点定位可靠性和位姿估计精度。后续工作将重点研究弱纹理场景下的高效特征建模与轻量化位姿估计方法。

作者贡献说明：

李露帆：研究思路的提出与方法框架设计，完成模型结构构建与算法实现，参与实验方案设计、论文撰写及整体修改工作，并负责数据预处理与模型参数调优；

党建武：实验结果分析与可视化展示，参与理论推导与模型性能验证工作；

雍玖：论文写作指导，提供论文整体框架设计，并参与论文审阅与修改工作。

参考文献：

[1] SONG X, LI Y Y, ZHANG Y F, *et al*. An overview of learning-based dexterous grasping: recent advances and future directions[J]. *Artificial Intelligence Review*, 2025, 58(10): 300.

[2] AI B, TIAN S, SHI H C, *et al*. A review of learning-based dynamics models for robotic manipulation [J]. *Science Robotics*, 2025, 10(106): eadt1497.

[3] LIU J, XIE J L, XIAO L B, *et al*. Hierarchical multi-modal fusion for language-conditioned robotic grasping detection in clutter[J]. *IEEE Robotics and Automation Letters*, 2024, 9(10): 8762-8769.

[4] 许凌志，符钦伟，陶卫，等. 基于三维模型的单目

车辆位姿估计[J]. 光学 精密工程, 2021, 29(6): 1346-1355.
XU L Z, FU Q W, TAO W, *et al*. Monocular vehicle pose estimation based on 3D model[J]. *Opt. Precision Eng.*, 2021, 29(6): 1346-1355. (in Chinese)

[5] GENG J H, WANG Y T, CHENG Y, *et al*. Systematic AR-based assembly guidance for small-scale, high-density industrial components[J]. *Journal of Manufacturing Systems*, 2025, 79: 86-100.

[6] 王立辉, 苏余足威, 韩华春, 等. 智能换电站电池包锁止机构位姿视觉估计[J]. 光学 精密工程, 2023, 31(21): 3135-3144.
WANG L H, SU Y, HAN H C, *et al*. Visual 6D pose estimation of battery package locking mechanism in intelligent battery swapping station [J]. *Opt. Precision Eng.*, 2023, 31(21): 3135-3144. (in Chinese)

[7] 王怡雯, 赵玺. 输入输出双视角下的增强现实人机交互技术综述[J]. 中国图象图形学报, 2026, 31(2): 349-373.
WANG Y W, ZHAO X. Input-output modalities in augmented reality: a two-dimensional survey of human-computer interaction[J]. *Journal of Image and Graphics*, 2026, 31(2): 349-373. (in Chinese)

[8] YANG H, PAVONE M. Object pose estimation with statistical guarantees: conformal keypoint detection and geometric uncertainty propagation[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 8947-8958.

[9] HAN Z W, CHEN L, WU S Q. CMFF6D: Cross-modality multiscale feature fusion network for 6D pose estimation[J]. *Neurocomputing*, 2025, 623: 129416.

[10] LI Y, HU H, XIAO J, *et al*. Uncertainty-aware keypoint-based 6D object pose estimation [J]. *IEEE Transactions on Image Processing*, 2024, 33: 2150-2164.

[11] DONADI I, PRETTO A. KVN: keypoints voting network with differentiable RANSAC for stereo pose estimation[J]. *IEEE Robotics and Automation Letters*, 2024, 9(4): 3498-3505.

[12] HONG J X, ZHANG H B, LIU J H, *et al*. A Transformer-based multi-modal fusion network for 6D pose estimation [J]. *Information Fusion*, 2024, 105: 102227.

[13] SONG Y W, TANG C H. A RGB-D feature fusion network for occluded object 6D pose estimation [J]. *Signal, Image and Video Processing*, 2024, 18(8): 6309-6319.

[14] AN Y N, YANG D D, SONG M Y. HFT6D: Multimodal 6D object pose estimation based on hierarchical feature transformer [J]. *Measurement*, 2024, 224: 113848.

[15] 蔡旭东, 王永才, 白雪薇, 等. 基于先验地图的视觉重定位方法综述[J]. 软件学报, 2024, 35(2): 975-1009.
CAI X D, WANG Y C, BAI X W, *et al*. Survey on visual relocalization in prior map[J]. *Journal of Software*, 2024, 35(2): 975-1009. (in Chinese)

[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*). *October* 22-29, 2017, *Venice*, *Italy*. IEEE, 2017: 3848-3856.

[17] LI H Q, WAN G Y, LI X N, *et al*. DenseFusion-DA2: end-to-end pose-estimation network based on RGB-D sensors and multi-channel attention mechanisms[J]. *Sensors*, 2024, 24(20): 6643.

[18] XIANG Y, SCHMIDT T, NARAYANAN V, *et al*. PoseCNN: a convolutional neural network for 6d object pose estimation in cluttered scenes[C]. *Proceedings of Robotics: Science and Systems*, 2018.

[19] CHEN H Z, MANHARDT F, NAVAB N, *et al*. Te$_x$Pose: neural texture learning for self-supervised 6D object pose estimation[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 4841-4852.

[20] WANG G, MANHARDT F, LIU X Y, *et al*. Occlusion-aware self-supervised monocular 6D object pose estimation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(3): 1788-1803.

[21] LI Q N, HU R M, XIAO J, *et al*. Learning latent geometric consistency for 6D object pose estimation in heavily cluttered scenes[J]. *Journal of Visual Communication and Image Representation*, 2020, 70: 102790.

[22] PENG S D, LIU Y, HUANG Q X, *et al*.

PVNet: pixel-wise voting network for 6DoF pose estimation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 15-20, 2019, *Long Beach, CA, USA*. IEEE, 2019: 4556-4565.

[23] ZHANG S B, ZHAO W Q, GUAN Z Y, *et al*. Learning cross-view consistent 3D keypoints for object 6D pose estimation[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, 35(7): 6816-6831.

[24] 王金聪，杨海峰，宋文龙，等. 基于改进DenseFusion的卫星6D位姿估计方法[J]. 激光杂志，2025, 46(3): 161-168.
WANG J C, YANG H F, SONG W L, *et al*. Satellite 6D position estimation method based on improved DenseFusion[J]. *Laser Journal*, 2025, 46(3): 161-168. (in Chinese)

[25] 白一凡，党选举. 多层特征融合与混合注意力的物体位姿估计[J]. 组合机床与自动化加工技术，2024(6): 32-36, 41.
BAI Y F, DANG X J. Object pose estimation based on multi-layer feature fusion and hybrid attention [J]. *Modular Machine Tool & Automatic Manufacturing Technique*, 2024(6): 32-36, 41. (in Chinese)

[26] HU Y L, FUA P, WANG W, *et al*. Single-stage 6D object pose estimation[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 13-19, 2020. *Seattle, WA, USA*. IEEE, 2020: 2927-2936.

[27] SONG C, SONG J R, HUANG Q X. HybridPose: 6D object pose estimation under hybrid representations[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 13-19, 2020. *Seattle, WA, USA*. IEEE, 2020: 428-437.

[28] FENG G K, XU T B, LIU F L, *et al*. NVR-net: normal vector guided regression network for disentangled 6D pose estimation[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(2): 1098-1113.

[29] OBERWEGER M, RAD M, LEPETIT V. Making deep heatmaps robust to partial occlusions for 3D object pose estimation[C]. *Computer Vision-ECCV* 2018. *Cham*: Springer, 2018: 125-141.

[30] XU Y, LIN K Y, ZHANG G F, *et al*. RNNPose: 6-DoF object pose estimation *via* recurrent correspondence field estimation and pose optimization [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(7): 4669-4683.

作者简介：

李露帆(2002—)，女，河南周口人，硕士研究生，主要从事6D位姿估计方法的研究。E-mail: 1520934457@qq.com

通讯作者：

雍　玖(1993—)，男，博士，副教授，甘肃临夏人，主要从事智能推荐算法研究。E-mail: yongjiu@mail.lzjtu.cn