

文章编号: 1002-1582(2025)05-0610-07

基于多尺度特征融合的双目位姿估计

朱家奇¹, 李彬^{2*}, 赵新华¹

(1. 天津理工大学 计算机科学与工程学院, 天津 300384)

(2. 天津理工大学 机械工程学院 天津市先进机电系统设计与智能控制重点实验室, 天津 300384)

摘 要: 针对病态区域位姿估计精度不足的问题, 提出基于多尺度特征融合的双目位姿估计算法。算法通过构建双目视觉-位姿估计端到端联合优化框架, 突破传统级联式架构的信息传递瓶颈。首先, 设计多尺度成本积模块, 通过融合不同感受野下的上下文特征, 提升弱纹理与遮挡区域的立体匹配精度; 其次, 提出金字塔特征筛选融合模块, 利用跨层注意力机制动态聚合低层几何细节与高层语义信息, 以增强位姿估计的几何一致性; 最后, 以Linemod数据集为基础, 搭建基于双目的Linemod-Stereo数据集, 并验证本文算法的有效性。实验结果表明, 所提算法相比传统基于RGB-D的位姿估计方法具有竞争性优势。

关 键 词: 立体匹配; 双目视觉; 位姿估计; 多模态融合

中图分类号: TP394.1

文献标识码: A

Stereo vision pose estimation based on multi-scale features

ZHU Jiaqi¹, LI Bin², ZHAO Xinhua¹

(1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China)

(2. Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, School of Mechanical Engineering, Tianjin University of Technology, Tianjin 300384, China)

Abstract: To address the insufficient pose estimation accuracy in degenerate regions, a stereo pose estimation algorithm based on multi-scale feature fusion is proposed. An end-to-end joint optimization framework for stereo vision and pose estimation is established, overcoming the information transmission bottleneck in traditional cascaded architectures. Firstly, a multi-scale cost volume module is designed to enhance stereo matching accuracy in weakly-textured and occluded regions through context feature fusion across varying receptive fields. Additionally, a pyramid screening-feature fusion module is proposed. This module dynamically aggregates low-level geometric details and high-level semantic information. It uses a cross-layer attention mechanism to achieve this aggregation to significantly improve the geometric consistency of pose estimation. Finally, a new binocular vision-based Linemod-Stereo dataset is constructed to validate the algorithm's effectiveness. Experimental results demonstrate that the proposed method outperforms conventional RGB-D-based approaches.

Keywords: stereo matching; stereo vision; pose estimation; multi-modal fusion

0 引 言

基于双目的位姿估计是利用两个相机同时获取被测物体不同视角的图像, 根据相机参数和对应点的

位置偏差, 在三维空间中描述物体坐标系与相机坐标系间的转换关系。随着计算机视觉的快速发展, 双目立体视觉技术被广泛应用于摄影测量、无人自动驾驶、机器人智能装配与导航、虚拟现实等领域^[1-4]。

收稿日期: 2025-04-23; 收到修改稿日期: 2025-06-13

基金项目: 国家自然科学基金项目 (52375026)

作者简介: 朱家奇 (1995—), 男, 博士研究生, 从事双目视觉、深度学习、位姿估计研究。13110090339@163.com

李彬 (1981—), 男, 教授, 博士, 从事视觉感知、机器人技术研究。

* 通讯作者: stereovision@stud.tjut.edu.cn

目前,依据于其输入数据,可分为基于RGB的位姿估计算法和基于RGB-D的位姿估计算法。基于RGB的方法^[5-9]采用数据驱动的方式,利用卷积神经网络(CNN)或Transformer架构,通过预测二维关键点建立2D-3D对应关系,或通过定位物体边界框来提升估计的鲁棒性。然而,由于2D-3D投影固有的不稳定性,基于RGB的位姿估计方法难以有效解决逆投影导致的病态问题,从而限制了6D位姿预测的精度和可靠性。

为了克服这些难题,深度传感器技术的最新进展推动了向RGB-D数据融合的转向,旨在结合几何深度信息以弥补传统方法的不足。根据训练方法不同,将基于RGB-D的位姿估计方法分为直接回归法和两步法。其中,直接回归法^[10-12]采用端到端的架构,以RGB-D作为输入直接预测目标的旋转和平移。Jiang等^[11]将UV信息与RGB-D图像共同作为网络输入,以缓解投影失效问题。该方法采用单一CNN主干网络进行特征提取,不仅简化了网络架构,还显著提升了计算效率。为进一步解决对称目标导致位姿歧义问题,Mo等^[12]提出基于2D卷积的全新网络架构,统一处理深度信息与RGB数据,实现了特征提取的一体化,并且设计了新的误差度量函数,有效解决一对多的问题。

与直接法不同,两阶段法^[13-16]首先在物体表面预设关键点,提取并融合这些关键点的多模态特征;其次预测关键点的实际位置,建立2D-3D关键点对应关系,最终通过PnP-RANSAC算法或最小二乘拟合算法精确计算目标的6D位姿。He等^[13]基于偏移量与语义标签的迭代投票机制,自适应筛选

最优预测结果,实现了三维关键点检测。在此基础上,He等^[14]进一步优化了特征融合策略,通过深度耦合特征提取与融合模块,在保持计算效率的同时显著增强了特征表示的判别性。Zhou等^[16]通过深度挖掘跨多模态相关性构建全局语义相似性模型,并提出非迭代加权向量投票方案以减少计算复杂度。

尽管基于RGB-D的方法在几何信息融合和位姿估计精度上取得了显著进展,其在处理病态区域时仍面临特征检测困难、融合不充分等问题,导致估计精度下降。为此,本文提出基于多尺度特征融合的双目位姿估计算法(Stereo vision pose estimation based on multi-scale features, MFpose)。核心创新在于构建了多层次特征融合机制与双目视觉的深度信息协同框架。首先,通过模拟人类双眼视差原理,利用两台相机从不同视角捕获同一场景的图像,计算像素点的深度信息。其次,通过跨层次特征整合融合浅层的细节特征与深层的语义特征,解决了单层次特征在几何细节与全局上下文表达上的局限性,并结合自适应注意力机制动态调整融合权重。本文算法充分利用双目系统的立体视觉优势,将深度学习与几何约束结合,在端到端训练中联合优化深度估计与位姿预测任务,既提升匹配鲁棒性,又避免尺度模糊性问题。

1 基于双目的位姿估计算法

1.1 网络结构

本文提出的算法MFpose主要包括两个核心模块:双目立体匹配和位姿估计模块,其算法结构如图1所示。

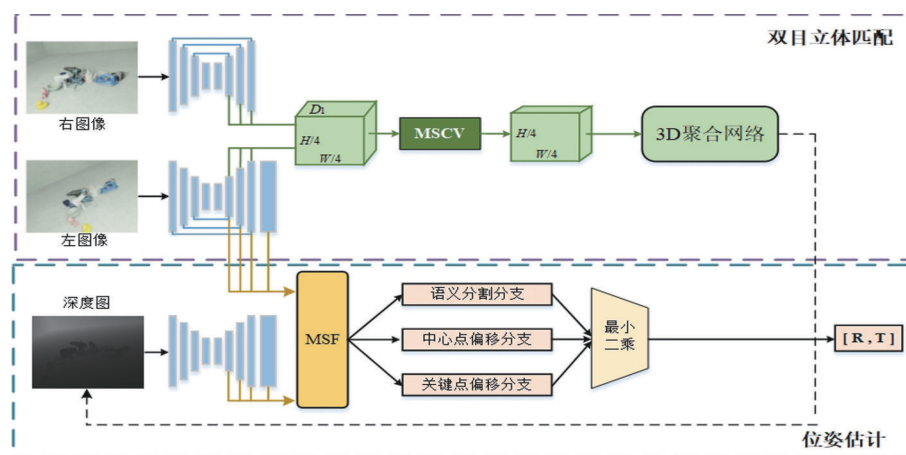


图1 网络结构图

首先,通过双目相机捕获左右视角的图像对,并利用立体匹配算法生成高精度的深度图;其次,结合RGB图像与深度信息,利用共享主干的多模态特征提取网络,充分挖掘目标物体的几何与纹理特征,并基于关键点坐标结合最小二乘法,回归6D位姿。

1.2 双目立体匹配

1.2.1 特征提取模块

特征提取模块通过提取输入图像对中的可辨识特征,进而构建多尺度成本积,实现精确的视差估计。由于编码器-解码器架构网络能够充分学

习输入和输出之间的复杂映射关系, 因此特征提取网络同样采用此结构提取双目图像的信息。在编码器阶段, 采用预训练后的 MobileNetV2^[17]模型将输入左右图像 F_l 和 F_r 下采样到 1/32 分辨率。然后, 将得到的特征输入到类似于 U-Net 的解码器模块, 以提取多尺度特征 $f_{(l,i)}^{ms}(f_{(r,i)}^{ms}) \in C_i \times \frac{H}{i} \times \frac{W}{i}$ 。其中, H 和 W 表示输入图像的高和宽; i 表示不同层级特征的分辨率。

1.2.2 多尺度成本积构建和聚合模块

为了显著减轻成本积的计算压力和提高视差预测的可靠性, 设计基于高斯上下文激励的多尺度成本积构建方法 (Multi-scale Cost Volume, MSCV)。该方法利用高斯注意力机制自适应调整不同层级的权重, 通过激励策略匹配不同层级特征的视差, 进而实现更加准确的相似性度量。MSCV 模块网络结构如图 2 所示。

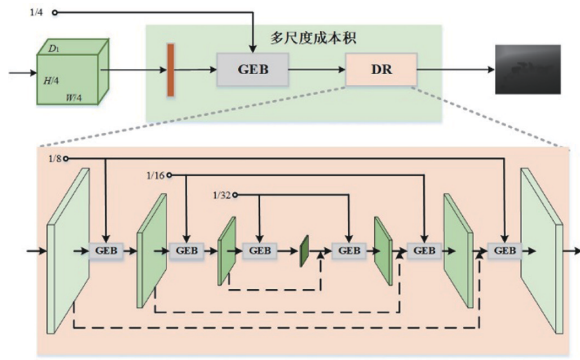


图 2 多尺度成本积示意图

给定左右视图 1/4 分辨率特征 $f_{(l,4)}^{ms}$ 和 $f_{(r,4)}^{ms}$ 后, 沿着通道维度将特征分为 N_g^{ms} 组, 然后计算分组相关卷 C_{corr}^{ms} 如下

$$C_{corr}^{ms}(j, d, x, y) = \frac{1}{N_c^{ms}/N_j^{ms}} \langle f_{(l,4)}^{ms}(x, y), f_{(r,4)}^{ms}(x - d, y) \rangle \quad (1)$$

式中, N_c^{ms} 表示特征通道的数量; $\langle \cdot, \cdot \rangle$ 是内积; d 是差异指数; (x, y) 表示像素坐标; j 表示第 j 个特征组。在得到初始的成本积 C_{corr}^{ms} 后, 将其输入一个 $3 \times 3 \times 3$ 步长为 1 的 3D 卷积层和一个高斯激励块 (Gaussian Excitation Block, GEB), 筛选后得到维度为 $D_1' \times H/4 \times W/4$ 的成本积。

为了构建左右特征图与多尺度成本积之间的关系, 利用不同尺度的特征图捕捉几何信息, 并通过高斯激励函数获取高权重特征, 自适应学习映射关系, 以过滤初始成本积, GEB 基本结构如图 3 所示。

首先对不同尺度的输入特征图通过 2 个 $1 \times 1 \times 1$ 的 3D 卷积将通道数量与同尺度的成本积调整一致, 然后利用高斯激活函数自适应学习各尺度的

权重并与对应的尺度成本积相乘, 进而得到不同尺度下的成本积, 具体过程表示如下

$$f_i' = \text{conv}(\text{conv}(f_i)) \quad (2)$$

$$g(f_i) = e^{-\frac{(\text{norm}(f_i'))^2}{2c^2}} \quad (3)$$

$$\text{GEB}_i^{ms}(f_i) = g(f_i) \odot C_i^{ms} \quad (4)$$

式中, conv 表示 3D 卷积层; norm 表示归一化函数; g 代表高斯函数; $\odot$ 代表哈达玛乘积; c 代表超参数; f_i 表示分辨率为 i 的串联代价体, f_i' 表示调整通道数后的代价体, $i = 4, 8, 16, 32$, $\text{GEB}_i^{ms}(\cdot)$ 表示高斯激励模块过滤后分辨率为 i 的多尺度代价体。

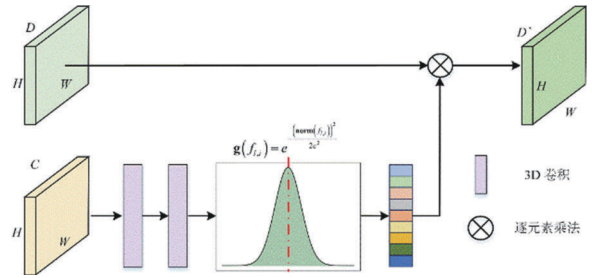


图 3 GEB 模块结构示意图

仅依靠高斯激活函数对 4D 成本积进行筛选虽然可以简化成本积的计算量, 但是该操作缺乏捕捉全局上下文的能力。许多算法通过堆叠基于 3D 卷积的沙漏网络聚合成本积, 以获取更强的空间特征的表达, 但是 3D 卷积模块的大量使用必将带来计算复杂度高, 模型参数多等问题。目前主流的网络通过减少 3D 卷积层的数量, 逐步细化成本积和视差图以及压缩 4D 代价立方体等方法提高算法效率。为平衡网络对于精度和效率的需求, MFpose 网络采用基于 GEB 的轻量化成本聚合网络。该网络利用编码器-解码器架构, 通过由粗到细融合不同尺度的输入图像, 结合 GEB 监督并筛选成本积, 以此对图像的上下文信息进行整合, 最终获得多尺度成本积 C^{ms}

$$C^{ms} = \text{DR}(\text{GEB}_i^{ms}(f_{(l,i)})) \quad (5)$$

式中, DR 为轻量化聚合网络^[18]。

1.2.3 匹配损失

MFpose 采用 L_1 损失函数监督双目预测值与真实值差距, 其损失定义为

$$L_{\text{ASD}} = \frac{1}{M} \sum_{i=1}^M \text{Smooth}_{L_1}(d_i - \hat{d}_i) \quad (6)$$

式中, d_i 表示像素点 i 的预测视差值; $\hat{d}_i$ 表示像素点 i 的真实视差值; M 表示待预测深度值像素点的数目。

1.3 位姿估计

1.3.1 金字塔筛选融合模块

在获取深度图后,首先通过相机内参矩阵将深度值转换为三维点云信息。不同于图像分支使用预提取的颜色特征,点云分支采用共享主干的特征提取网络捕获几何特征。这种设计确保了跨模态特征在潜在空间的一致性表征,同时避免了冗余计算。为优化特征融合过程,提出金字塔筛选融合模块(Pyramid Screening-feature Fusion Module, PSF)。该模块通过注意力过滤模块(Attention Filtering Module, AF),考虑不同层级特征显著特征的关联与重要性,实现对多层级信息潜在表征解码。其具体结构如图4所示。

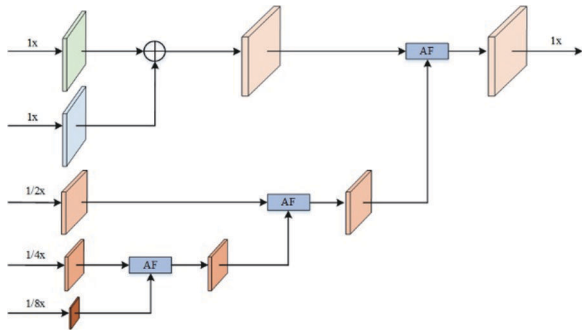


图4 金字塔筛选融合模块示意图

以1/8分辨率与1/4分辨率特征图融合为例,给定高级特征序列 $F'_{(pw,8)}$ 和低级特征序列 $F'_{(pw,4)}$,通过AF模块筛选 $F'_{(pw,8)}$ 中的显著特征,并将其通道数压缩到与 $F'_{(pw,4)}$ 相同的通道,得到融合后的 $F'_{(pw,4)}$ 。

为了统一输入特征的尺度,AF模块采用转置卷积和双线性插值相结合的方法来恢复高级特征的尺度信息,基本结构如图5所示。AF模块包含 3×3 转置卷积层、双线性插值块和SE(Squeeze-and-Excitation)激活模块。首先使用转置卷积和双线性插值块恢复高级特征的尺度;然后利用SE模块将高级特征转换为相应的注意力权重,以过滤低级特征;最后利用得到的权重信息与低层特征相乘后相加,即可得到融合后的特征。

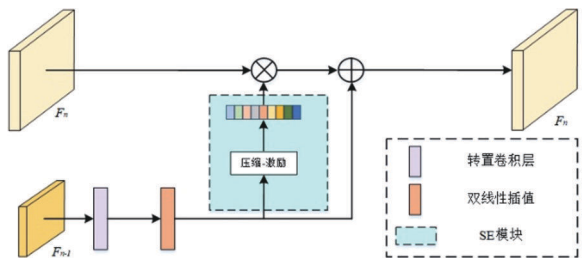


图5 注意力过滤模块示意图

给定低层特征 F_n 和高层特征 F_{n+1} ,通过两个上采样层将 F_n 和 F_{n+1} 的通道数统一,计算过程如下

$$F_{att} = BL(TC(F_{n+1})) \quad (7)$$

式中, BL 表示双线性插值块; TC 表示转置卷积。然后,利用SE模块将特征转换为相应的注意力权重,计算过程如下

$$F_{out} = F_n \cdot SE(F_{att}) + F_{att} \quad (8)$$

式中, SE 表示注意力激活模块。

1.3.2 6D位姿估计

MFpose算法通过搭建语义分割分支、中心点偏移检测分支和关键点偏移检测分支三个互相约束的子任务预测关键点坐标,进而通过最小二乘法拟合位姿数据。

(1) 中心点偏移分支

给定金字塔筛选融合后的特征序列,利用多层感知器(Multilayer Perceptron, MLP)预测物体中心点的偏移量。为监督对该分支的训练,本文采用 L_{center} 损失函数,其具体计算过程表示为

$$L_{center} = \frac{1}{N} \sum_{n=1}^N \|O_n - \hat{O}_n\| \quad (9)$$

式中, N 表示待估计目标的数目; K 表示预定义目标关键点的数目; O_n 表示第 n 个物体中心点预测值, $\hat{O}_n$ 表示第 n 个物体中心点真实值。

(2) 语义分割分支

为了区分图像中的交互物体,语义分割分支通过全局几何特征和局部像素特征,预测同一类物体的边界,进而辅助定位不同物体的中心点位置。语义分割分支同样采用MLP层预测的边界,其输出通道数为待估计物体中心点的数目。本章算法采用Focal Loss损失函数监督该分支,具体计算过程表示为

$$L_{seg} = -a(1-b)^s \log(b) \quad (10)$$

式中, a 表示平衡正负样本的超参数; s 表示调节难易样本的超参数; b 表示对目标的预测概率。

(3) 关键点偏移分支

关键点偏移分支是基于物体表面可见点与预定义关键点的估计偏移量,通过投票的方式选出集群中点作为预测的坐标值。预测关键点的三维坐标可以表示为

$$P_n = \text{Vot}(p_n + O) \quad (11)$$

式中, Vot 表示投票模块; P_n 表示投票后置信度最高第 n 个关键点的坐标; p_n 表示第 n 个关键点的估计坐标; O 表示物体中心点估计坐标。为了监督关键点的学习,关键点偏移分支采用了 L_1 损失函数进行监督

$$L_{keypoints} = \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K \|P_n^k - \hat{P}_n^k\| \quad (12)$$

式中, N 表示待估计目标的数目, K 表示预定义目

标关键点的数目, P_n^k 表示第 n 个物体中第 k 个关键点的预测偏移量; $\hat{P}_n^k$ 表示第 n 个物体中第 k 个关键点的真实偏移量。

1.4 多任务损失函数

MFpose 采用分阶段多任务联合训练, 其损失函数包含视差损失函数, 中心点偏移损失函数、语义分割损失以及关键点偏移损失三部分。整体任务损失函数计算表示为

$$L_{\text{all}} = \lambda_1 L_{\text{ASD}} + \lambda_2 L_{\text{center}} + \lambda_3 L_{\text{seg}} + \lambda_4 L_{\text{keypoints}} \quad (13)$$

式中, 各子任务损失权重为 $\lambda_1 = 2\lambda_2 = 2\lambda_3 = \lambda_4 = 2$ 。

2 实 验

2.1 数据集及评估指标

为公平验证 MFpose 的性能, 以 Linemod 数据集^[19]原始高精度 3D 物体模型为输入, 通过 BlenderProc 配置虚拟双目相机参数, 在可控光照与多样化材质属性下渲染生成多视角双目图像对, 同步输出亚像素级精度的左右图像对及对应的深度和 6D 位姿标签, 以获取改进后 Linemod-Stereo 数据集。该数据集包括不同光照、不同背景和不同遮挡条件等多种场景下的图像构成, 如图 6 所示。



图 6 Linemod-Stereo 数据集示意图

为验证本文算法的有效性, 采用以下指标评估:

(1) 端点误差 (EPE)

端点误差为预测视差图与真实视差图之间所有有效像素点的平均绝对误差。计算公式为

$$\text{EPE} = \frac{\sum_{i=1}^U |d_i - \hat{d}_i|}{U} \quad (14)$$

式中, U 表示除噪点外的有效像素; d_i 表示像素点 i 的预测视差值; $\hat{d}_i$ 表示像素点 i 的真实视差值。

(2) s 像素误差百分比 (>s-px)

s 像素误差百分比是指视差图中误差大于 s 的百分比, 其公式表示为

$$\text{s-px} = \frac{\sum_{i=1}^U (|d_i - \hat{d}_i|) \cdot K(s)}{U} \quad (15)$$

$$K(s) = \begin{cases} 1, & s \geq d_i \\ 0, & s < d_i \end{cases} \quad (16)$$

式中, d_i 为预设定的视差阈值。

(3) 平均距离度量 (ADD, ADD-S)

平均距离度量通过衡量预测位姿与真实位姿对应点间的欧式距离的平均值, 得到位姿估计误差。其计算过程表示为

$$\text{ADD} = \frac{1}{O} \sum_{p \in O} \| (Rp + t) - (R^*p + t^*) \| \quad (17)$$

式中, O 表示物体 S 中所有顶点 p 数目; R, t 分别表示预测的旋转矩阵和平移向量, R^*, t^* 分别表示预测的旋转矩阵和平移向量。对于对称物体, 通过计算每个顶点到目标模型最近顶点的距离 ADD-S 为

$$\text{ADD-S} = \frac{1}{O} \sum_{p \in O} \min_{p_2 \in O} \| (Rp_1 + t) - (R^*p_2 + t^*) \| \quad (18)$$

式中, p_1 和 p_2 表示预测点和模型点。

由于 Linemod-Stereo 数据集 21 类物体中含有 2 项对称物体, 因此, 将评价指标 ADD 与 ADD-S 简写为 ADD (-S)。对于 Linemod-Stereo 数据集, 遵循 Linemod 数据集^[10, 13-14]惯例将 ADD (-S) 阈值设定为目标物体最大直径的 10%。该阈值设定方式充分考虑了物体尺度差异。ADD (-S) 的数值越大, 表示位姿预测越准确, 最佳预测值为 100%。

2.2 实验环境及参数设置

本文所提算法 MFpose 使用 Pytorch 在硬件环境为 Intel i7-13700KF CPU 和 RTX4080 GPU 的计算机进行训练和评估。在 Linemod-Stereo 数据集, 将训练集的批次大小设为 8, 并进行 50 个周期的训练。优化过程采用 AdamW 优化器, 设置初始学习率为 2×10^{-4} 并实施单周期学习率调度策略, 同时将梯度范数裁剪至 $[-1, 1]$ 区间以增强训练稳定性。

2.3 消融实验

为验证共享权重策略、MSVC 模块、MSF 模块的有效性, 设计消融实验, 如表 1 所示。其中, 第一组基础模型仅采用共享权重的特征提取网络处理 RGB 和深度图像。

由表 1 可知, 引入 MSVC 模块后, EPE 指标降低至 0.43px, 1 像素误差率 ($>1\text{-px}$) 下降至 3.92%, 而模型参数量仅增加 0.6M。这验证了 MSVC 模块能够有效提升深度感知能力。在此基础上, 进一步引入 MSF 模块, ADD (-S) 指标提升 1.53%, 表明该模块通过注意力机制实现了判别性特征增强与冗余信息抑制。值得注意的是, 单独引入 MSVC 模块的模型在 EPE 指标上优于单独使用 MSF 模块的方案。该对比实验证实, 在多模态位姿估计框架中, 原始数据源的几何一致性保障比后期多模态特征融合和优化更重要。

表 1 消融实验

共享权重	MSVC	MSF	EPE/px	>1-px/%	ADD (-S)/%	参数量/M
✓			0.61	5.36	81.91	17.3
✓	✓		0.43	3.92	84.88	17.9
✓		✓	0.61	5.36	84.21	27.7
✓	✓	✓	0.43	3.92	86.32	28.3

2.4 对比实验

2.4.1 立体匹配模块对比实验

为全面评估所提出的双目立体匹配模块的性能，在 Linemod-Stereo 数据集上与当前主流的 PSMNet^[20]、GWCNet^[21]和 RAFT-Stereo^[22]算法进行了对比实验。实验结果如表 2 所示，其中加粗数值表示最优性能指标。实验数据分析表明，MFpose 算法相较于基于单尺度成本积的 RAFT-Stereo 算法，在 EPE 和 1 像素误差率 (>1-px) 两个关键指标上分别提升了 0.02px 和 0.46%，充分验证了多尺度特征融合策略的有效性。尽管与同样采用多尺度特征的 GWCNet^[21]算法相比，MFpose 在推理速度上略有不足，但在 EPE 和 1 像素误差率评估指标上均取得了显著提升，其中 EPE 指标达到 0.18px，充分体现了所提出的金字塔特征筛选融合模块的优越性。

表 2 立体匹配模块对比实验

模型	EPE/px	>1-px/%	时间/s
PSMNet ^[20]	0.84	6.24	0.19
GWCNet ^[21]	0.61	4.52	0.16
RAFT-Stereo ^[22]	0.45	4.38	0.20
MFpose	0.43	3.92	0.24

图 7 显示了 MFpose 的立体匹配模块与 RAFT-Stereo^[22]算法在 Linemod-Stereo 数据集可视化分析结果。通过对比红色矩形标注的关键区域可知，在鸡蛋盒以及打孔器等弱纹理对象，本文所提算法能够更精准地捕捉鸡蛋盒开口位置以及打孔器的端面空洞轮廓，其视差图呈现出更细腻的几何层次。针对受遮挡的手电钻以及高光照的鸭子区域，MFpose 的视差预测在手电钻按钮处以及鸭子头部的深度估计保持了更好的空间连续性。

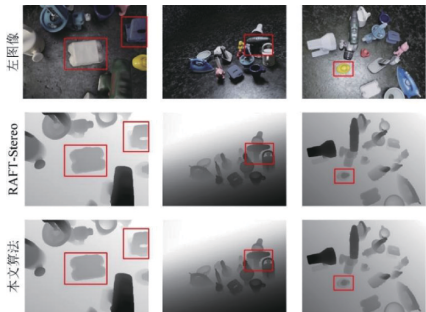


图 7 Linemod-Stereo 数据集立体匹配可视化结果

2.4.2 整体网络对比实验

为公平对比 MFpose 与其它经典方法的性能，本实验中所有算法均采用双目立体匹配模块生成的深度图和左图像作为输入，在 Linemod-Stereo 数据集进行训练和验证。由表 3 可知，本文算法在 ADD (-S) 均值指标上优于大部分算法。与 FFB6D^[14]相比，虽然本文算法 ADD (-S) 均值略低，但是在猿、猫和鸭子等小目标检测效果更优，证明通过多尺度特征与立体几何信息的互补，显著提升微小目标的位姿估计精度。为直观对比本文算法的检测效果，本文在 Linemod-Stereo 数据集抽取猿、鸭、猫、电话等目标进行对比。如图 8 所示，本文所提算法 MFpose 对场景中不同类别物体准确的位姿估计。

表 3 位姿估计模块对比实验

类别	DF ^[10]	PVN3D ^[13]	FFB6D ^[14]	MFpose
猿	43.2	61.8	67.2	71.8
台钳	63.3	98.1	97.5	95.2
相机	36.7	98.5	99.3	98.3
罐子	59.2	98.3	98.7	96.5
猫	31.1	61.8	61.2	63.7
电钻	49.9	83.2	88.4	85.7
鸭子	33.4	59.7	61.5	62.9
鸡蛋盒	69.7	76.7	84.9	85.3
胶水	53.8	79.2	90.7	89.4
打孔器	38.9	91.1	95.6	93.5
熨斗	50.4	90.4	90.8	89.9
灯	43.8	91.7	92.6	91.3
电话	69.4	98.9	99.1	98.6
均值	49.45	83.80	86.73	86.32

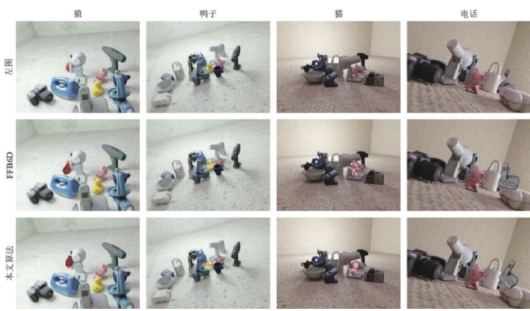


图 8 Linemod-Stereo 数据集位姿估计可视化结果

3 结论

本文提出了基于多尺度特征融合的双目位姿估计算法 MFpose，通过构建端到端联合优化框架，实现了立体匹配与位姿估计的双向信息交互，解决了传统级联式架构中的特征传递损失问题。首先，设计多尺度成本积模块提升病态区域的立体匹配精度；其次，提出金字塔特征筛选融合模

块, 利用注意力机制动态调整不同尺度特征的权重, 以实现几何和语义信息的融合; 最后, 搭建基于双目的 Linemod-Stereo 数据集为算法性能评估提供了可靠的基准测试平台。实验结果表明, 与现有方法相比, 本算法在保持较高精度的同时, 突破了对 RGB-D 硬件的依赖, 能够灵活适配多种双目视觉系统实现目标位姿估计。

未来将进一步结合动态场景分割与多模态传感器 (如 LiDAR/IMU) 融合策略, 并探索轻量化模型压缩技术, 以提升算法在移动端设备上的实时性及动态目标场景下的泛化能力。

参考文献:

- [1] 庄宏伟, 吴东岷, 张宝顺. 基于深度语义模型的双目位姿估计[J]. 光学技术, 2023, 49 (02): 205—210.
Zhuang Hongwei, Wu Dongmin, Zhang Baoshun. Binocular pose estimation based on depth semantic model[J]. Optical Technique, 2023, 49 (02): 205—210.
- [2] 邓鹏, 唐文涛, 黄开明. 基于级联融合网络的密集型抓取位姿检测[J]. 现代制造工程, 2024, (07): 126—134+60.
Deng Peng, Tang Wentao, Huang Kaiming. Intensive grasp pose detection based on cascaded fusion network[J]. Modern Manufacturing Engineering, 2024, (07): 126—134+60.
- [3] Papazov C, Haddadin S, Parusel S, et al. Rigid 3D geometry matching for grasping of known objects in cluttered scenes[J]. International Journal of Robotics Research, 2012, 31(4): 538—553.
- [4] Alvaro C, Dmitry B, Siddhartha S S, et al. Object recognition and full pose registration from a single image for robotic manipulation[C]//2009 IEEE International Conference on Robotics and Automation. Kobe, Japan: IEEE, 2009: 48—55.
- [5] Peng S, Liu Y, Huang Q, et al. PVNet: pixel-wise voting network for 6 DoF pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 4561—4570.
- [6] Park K, Patten T, Vincze M. Pix2pose: pixel-wise coordinate regression of objects for 6D pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 7668—7677.
- [7] Nguyen V N, Groueix T, Ponimatkin G, et al. Nope: novel object pose estimation from a single image[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle WA, USA: IEEE, 2024: 17923—17932.
- [8] Li Y, Mao Y, Bala R, et al. Mrc-net: 6-dof pose estimation with multiscale residual correlation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle WA, USA: IEEE, 2024: 10476—10486.
- [9] 王神龙, 雍宇, 吴晨睿. 基于伪孪生神经网络的低纹理工业零件 6D 位姿估计[J]. 电子学报, 2023, 51(01): 192—201.
Wang Shenlong, Yong Yu, Wu Chenrui. 6D Pose estimation of low texture industrial parts based on pseudo-siamese neural network[J]. Acta Electronica Sinica, 2023, 51(01): 192—201.
- [10] 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. Long Beach, USA: IEEE, 2019: 3343—3352.
- [11] Jiang X, Li D, Chen H, et al. Uni6d: a unified cnn framework without projection breakdown for 6d pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 11174—11184.
- [12] Mo N, Gan W, Yokoya N, et al. Es6d: a computation efficient and symmetry-aware 6d pose regression framework[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 6718—6727.
- [13] He Y, Sun W, Huang H, et al. PVN3D: a deep point-wise 3d keypoints voting network for 6dof pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 11632—11641.
- [14] He Y, Huang H, Fan H, et al. Ffb6d: a full flow bidirectional fusion network for 6d pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021: 3003—3013.
- [15] Wu J, Chen L, Wang S, et al. Geometric-aware dense matching network for 6D pose estimation of objects from RGB-D images[J]. Pattern Recognition, 2023, 137: 109293.
- [16] Zhou J, Chen K, Xu L, et al. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6d object pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Vancouver, Canada: IEEE, 2023: 13967—13977.
- [17] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: inverted residuals and linear bottlenecks[C]//Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 4510—4520.
- [18] Xu G, Wang X, Ding X, et al. Iterative geometry encoding volume for stereo matching[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 21919—21928.
- [19] Hinterstoisser S, Lepetit V, Ilie S, et al. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes[C]//Computer Vision-ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea: IEEE, 2013: 548—562.
- [20] Chang J R, Chen Y S. Pyramid stereo matching network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City, USA: IEEE, 2018: 5410—5418.
- [21] Guo X, Yang K, Yang W, et al. Group-wise correlation stereo network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA: IEEE, 2019: 3273—3282.
- [22] Lipson L, Teed Z, Deng J. Raft-stereo: multilevel recurrent field transforms for stereo matching[C]//2021 International Conference on 3D Vision (3DV). Prague, Czech Republic: IEEE, 2021: 218—227.