基于多源数据融合的目标姿态估计方法研究与实现

答辩学生 王子航， 指导教师 张强

（江苏科技大学测控技术与仪器专业）

摘　要：在智能制造、智能抓取、自动驾驶等领域，物体六自由度（6-DoF）姿态估计是三维环境感知的关键技术，是机器对空间准确理解和操作的基础。物体6-DoF位姿估计易受到复杂环境如光照变化、物体表观信息缺失、遮挡等因素的影响，从而影响物体姿态估计的准确性。本文深入探索了复杂环境条件对物体位姿估计的影响机制，针对视觉传感器多源信息鲁棒特征提取与高精度视觉目标位姿估计展开研究。为了突破光照变化环境和弱纹理物体的姿态估计问题，本文提出了一种基于RGB-D数据的两阶段混合卷积神经网络CMCL6D，优化跨模态数据特征整合，提高了算法对光照变化和弱纹理情况下的物体姿态估计的准确性。为了降低颜色信息表征一致性差对物体姿态估计的影响，并进一步探索深度数据蕴含的丰富信息，本文还提出了一种基于深度图像的端到端姿态估计网络DMG6D，其通过扩展深度图像数据形成法向量角和空间结构表达，进而形成高鲁棒性目标特征表征和姿态估计算法。本文对提出的方法在大型基准数据集上进行了实验测试。实验表明，本文提出的CMCL6D方法在YCB-Video数据集上的AUC指标达94.5%，提出的DMG6D在LineMod数据集上的ADD(S)-0.1d指标达98.9%，证明了本文提出多源数据融合策略的有效性。本文涉及的CMCL6D算法和DMG6D算法代码已分别开源于https://github.com/wangzihanggg/CMCL6D 和 https://github.com/wangzihanggg/DMG6D。

关键词：6D姿态估计；语义分割；多源数据融合；图像特征提取

0　引　言

在《中国制造2025》[1]这一国家制造业战略蓝图的引领下，中国正极速向制造强国转型，力求在全球科技革命和产业变革的大潮中占据先机。在传统的工业生产线中，机器通过目标检测、语义分割等方法感知周围环境[2]，然而这些方法主要关注物体的二维信息。随着智能化技术要求的提升，如何实现复杂场景下对三维环境的感知已成为机器视觉研究的焦点。其中，物体6-DoF姿态估计旨在使机器能够精确感知并理解三维空间中物体的位置和姿态，是实现机器人灵活操作[3]、自动化装配[4]、自动驾驶[5]等核心技术的基础。

目前，基于深度学习的物体6-DoF姿态估计方法取得了长足的进展。基于深度学习的6-DoF姿态估计方法主要分为基于模板、基于投票和基于对应点等方法。基于投票的方法依赖事先标注好的6-DoF姿态模板，选取最匹配的模板来估计物体姿态。Kehl等人的SSD-6D[6]则结合监督与自监督学习，基于SSD框架扩展以涵盖完整的6D姿态估计，但其在处理不规则或被遮挡目标时具有一定的局限性。基于投票的方法分为直接与间接投票。Peng等人的PVNet[7]通过像素指向关键点的向量进行回归与投票定位，展现出对遮挡情况的良好适应性。旷视研究院提出PVN3D[8]在此基础上采用RGB-D输入，先检测关键点后拟合姿态，结合实例分割与关键点投票优化。该方法提升了在复杂条件下的位姿估计的可靠性，但牺牲了部分实时性。基于对应点的方法核心为在RGB图像中识别与3D模型点对应的2D像素点，如YOLO6D[9]在单阶段内通过卷积网络直接从输入图像获取包围盒顶点与中心点，继而借助PnP算法推算6D姿态。然而，该类型的算法虽然实现了高实时性估计，但姿态估计的精度较差。

因此，虽然目前基于深度学习的6-DoF姿态估计方法已经取得了很好的研究成果，但实际应用中场景的复杂性对目标位姿估计存在巨大影响。例如，遮挡会导致模型的可见部分减少，影响特征提取的可靠性，从而影响网络模型的识别能力；另外，待检测目标的弱纹理特性也会影响特征表达的鲁棒性，对 6-DoF姿态估计构成巨大挑战。

为解决遮挡、弱光照等情况对目标位姿估计的影响，本文提出了基于RGB-D数据的双模态数据融合的CMCL6D位姿估计算法，通过两种模态数据的有效特征提取与互补实现高精度物体6D姿态估计；为解决弱纹理及光照变化等因素导致目标色彩数据噪声高的问题，提出了基于深度图像的DMG6D目标位姿回归方法，充分挖掘深度信息，同时规避光照变化对表观颜色对目标姿态估计的影响。本文的贡献和创新点如下：

（1）提出了一种基于RGB-D数据输入的两阶段混合CNN网络架构CMCL6D算法，有效提升跨模态数据交互和特征融合性能，能够在弱光照、弱纹理情况下获取高可靠性场景实例分割图像并基于此进行目标物体的鲁棒姿态回归。

（2）提出了一种基于纯深度数据输入的端到端姿态估计DMG6D网络框架，探寻深度图像蕴含的丰富信息，从深度数据中挖掘潜在的角度信息和空间结构信息，形成深度、角度、点云三模态数据交叉融合的网络

框架，在无需彩色图像前提下实现目标物体的鲁棒特征提取和位姿回归。

（3）本文提出的CMCL6D方法在YCB-Video数据集上的AUC精度指标达94.5%，提出的DMG6D在LineMod数据集上的ADD(S)-0.1d精度指标达98.9%，形成了视觉目标6-DoF高精度检测结果。

1 基于CMCL6D跨模态跨层特征融合机制的物体姿态估计算法

本节探索遮挡、弱光照等情况对RGB-D成像的影响，着重研究物体6-DoF姿态估计中的RGB-D鲁棒特征提取方法，形成基于跨模态跨层机制的RGB-D数据融合与6-DoF物体姿态估计CMCL6D框架。

图1 CMCL6D算法结构图

图1所示为本方法框架的结构图。其流程为给定RGB-D标定图像，首先使用RGB-D融合分割网络得到场景中目标实例的掩膜，再根据掩膜的边界分别对RGB图、深度图进行分割和特征提取形成密集特征，最后采用密集特征进行姿态估计与调优。

1.1 跨模态特征交互算法

在低纹理和低光照条件下，RGB图像往往缺乏足够的特征信息，而深度图像可以稳定表达目标物体的三维结构特征。为充分利用两种数据的隐含信息，本节提出了跨模态特征引导模块(Cross-Modal guided Feature Extraction Module, CM)，建立RGB和深度特征信息流之间的有效信息交互机制。

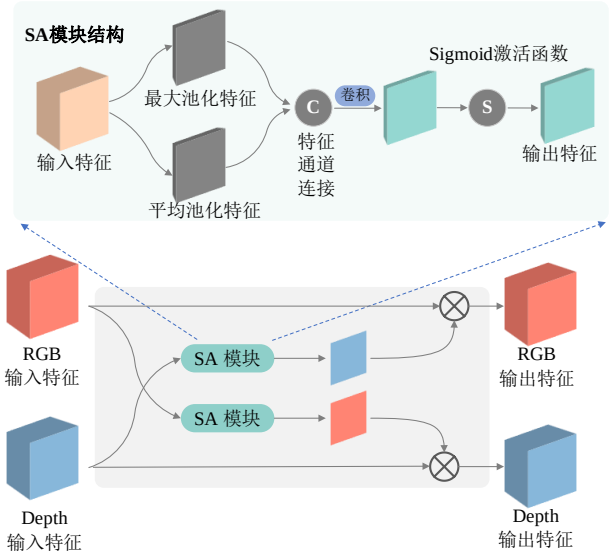

图2 CM模块结构图

图2为CM算法结构，实现RGB和D两种模态的优势特征交叉互补，该结构表达为：

$$F_{RSA} = \delta(MLP(AvgPool(F_R)) \oplus (MLP(MaxPool(F_R))) \quad (1)$$

$$F_{DSA} = \delta(MLP(AvgPool(F_D)) \oplus (MLP(MaxPool(F_D))) \quad (2)$$

$$F'_R = \delta(F_R \otimes F_{DSA}) \quad (3)$$

$$F'_D = \delta(F_D \otimes F_{RSA}) \quad (4)$$

式(1)-(4)中，F_R 和 F_D 分别为 RGB 和 D 分支的输入特征；$MaxPool$ 和 $AvgPool$ 分别为最大池化和平均池化；$\oplus$ 代表特征的通道级联；$\otimes$ 表示逐像素的乘法。MLP 表示多层感知机层，其包括两个卷积层和一个 ReLU 激活层δ。该结构采用的空间注意算法，实现了 RGB 特征和深度特征的有效融合。

在分割网络中，CM 模块通过 RGB 数据和深度数据之间的跨模态信息交互，使编码器的每一层都能对输入数据进行有效特征提取。该设计使 RGB 分支能够增强其特征提取能力，有效应对弱纹理和低光照复杂条件。同时，深度分支克服了缺乏颜色纹理信息的限制，实现了更全面的特征表示。然而，在复杂和多样化的场景中，仅依赖 CM 模块无法准确地提取和充分利用每个分支的特征信息。为解决该问题，提出了跨层特征提取引导模块，从而有效融合两个分支网络中的特征，进一步提高特征提取的鲁棒性。

1.2 跨层特征交互算法

本节提出了一种新的跨特征层特征提取引导模块(Cross-Layer guided Feature Extraction Module, CL)，旨在增强编码器的特征提取能力，其结构如图3所示。在基于深度学习的特征提取框架中，低层特征具备更多细节信息，高层次特征具备高级语义信息。CL模块实现不同层之间有价值信息的交互，从而使编码器有效利用低级特征层提供的丰富细节，增强整体特征的鲁棒性，进而提高物体姿态估计的准确性。公式(5)和(6)为CL算法的结构表达。

$$F_L = \delta(MLP(AvgPool(F_R)) \oplus (MLP(MaxPool(F_R))) \quad (5)$$

$$F_H = F_H \otimes \delta(MaxPool(F_L)) \quad (6)$$

式(5)-(6)中，F_L为空间注意模块提取的低层次空间注意力特征，F_H 为由低层次特征引导获得的高层次特征。通过将低级特征向高级特征的嵌入，有效地解决高级特征中整体特征拟合效率低的问题。

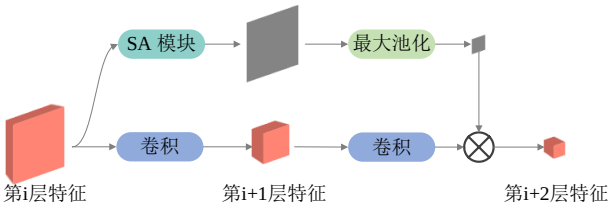

图3 CL模块结构图

1.3 基于改进自注意力机制的目标姿态检测头

受CBAM算法[11]的启发，本文关注如何从自注意力机制的角度更准确地提取目标物体的空间特征信息，形成目标物体边界框区域的鲁棒特征描述。因此，本文对姿态回归检测头网络进行改进，形成图4所示结构。该算法针对RGB与掩膜的鲁棒特征提取部分加入空间与通道自注意力机制，提高姿态矩阵回归的可靠性。

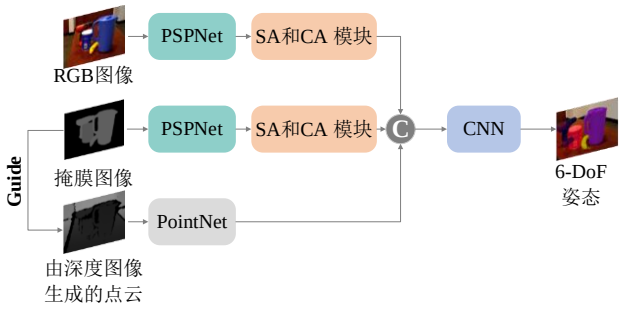

图4 6D姿态检测头模块结构图

2 基于DMG6D多流全局特征融合网络的物体姿态估计算法

针对光照变化使得物体RGB图像表征差异大的问题，本节探索仅基于深度图像的物体6-DoF检测方法，形成图5所示的基于多流全局特征融合DMG6D网络框架。

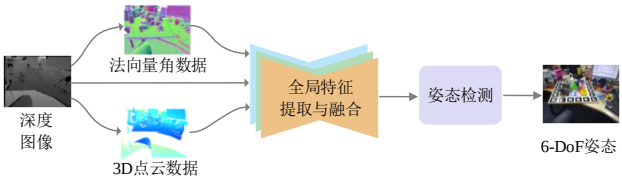

图5 DMG6D算法结构图

在DMG6D编码器-解码器的每个阶段进行全局特征融合，用于学习深度数据在各个模态下的特征表示并进行6D姿态估计。本文使用Swin Transformer[12]对原始深度图像和法向量角度图进行特征编码，使用RandLA-Net[13]对点云进行特征提取，并在此过程中对这三种模态的信息进行逐级特征互补融合，形成稠密特征嵌入。本文遵循[8]提出的3D关键点检测方法，采用得到的嵌入特征进行目标物体的3D关键点预测及6D姿态估计。

2.1 深度图像、法向量角度图像与点云的特征提取

以往的基于深度图的6D姿态估计算法往往直接将深度图输入进特征提取网络，但这类方法往往忽视了深度图隐含的目标物体的其他物理特征信息。为了从深度图像中获取方向信息和局部几何信息，本文遵循[14]的方法为每张深度图像中的每个像素点生成其与相机坐标系N与x,y,z轴对应的法向量角度，该过程描述为：

$$\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = \begin{bmatrix} \arccos(N \cdot x) \\ \arccos(N \cdot y) \\ \arccos(N \cdot z) \end{bmatrix} \tag{7}$$

之后，将得到的角度值(a_x, a_y, a_z)正则化到0~255区间，形成如图6所示的法向量角度数据。

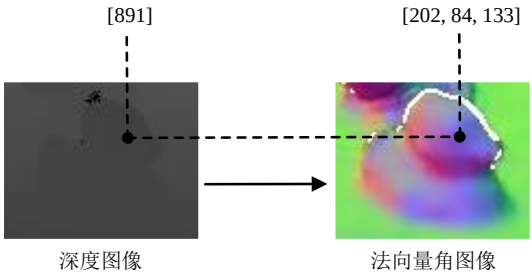

图6 法向量角度变换示例

为同时学习到深度特征和法向量角度图像的鲁棒特征，本节提出了基于编码器-解码器结构的主干网络。同时，考虑到两种数据的同源性，本文使用权重共享方式进行特征嵌入提取。网络采用Swin-T（微型Swin Transformer）作为编码器，UperNet[15]作为解码器，形成两种图像的多尺度特征表示。

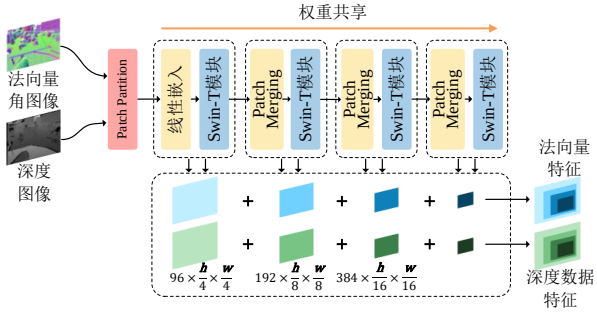

图7 Swin-T结构图

如图7所示，Swin-T编码器同时以深度图像和法向量角度图像作为输入，包含四个Swin Transformer模块，每个模块分别由Encoder、Bottleneck和Decoder组成。以深度图像为例，首先进行Patch Partition，将数据分为4x4个Patch；之后经过线性嵌入和两个自注意Swin Transformer模块进行特征提取；最后通过Patch Merging进行下采样，完成一个阶段的特征提取。同时，每次特征提取结果与其他两个信息流进行点级特征融合；通过迭代四次类似特征计算，完成对深度图像的特征提取。法向量角度图像特征提取过程与上述相同。

经特征提取后，本文使用UperNet进行上采样操作，恢复特征图尺寸。UperNet基于双线性插值进行上采样操作，并使用PSPNet[16]中的金字塔池化模块（PPM）作为特征金字塔的最后一层。上采样过程对三种模态的特征交互融合（见第2.2节）以及与编码器特征嵌入的拼接。通过最后一级的无监督上采样，得到与输入图

像尺寸相同的像素级特征嵌入结果。

对于点云模态，首先使用多尺度填充法补齐深度图像孔洞，再根据相机内参矩阵将深度图像转换为3D点云。为了提取点云的多尺度嵌入特征，点云特征提取网络包含一次特征预处理CNN模块、四次RandLA-Net特征提取下采样模块以及三次RandLANet特征提取上采样模块。在该上/下采样过程中，进行点云与深度图像和法向量角度模态的逐点特征融合。

2.2 基于全局特征融合的特征交互机制

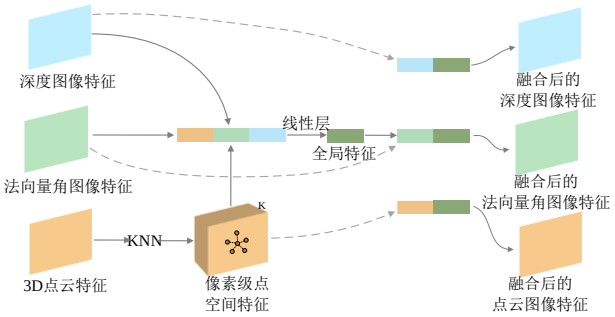

图8 基于全局特征的融合机制示意图

本节提出了一种基于全局特征的逐点特征融合机制，其结构如图8所示。由于法向量角度图像、点云两种模态都源自深度图像，各模态数据在空间上良好对齐，因此可以以像素-点的方式进行多模态融合。该过程为：深度图像和法向量角度图像上的每个像素在点云模态中都存在对应的点；采用kNN算法寻找目标点的k个近邻，并将这些邻近点整合输入至多层感知机来表征目标像素的空间结构特征。该过程表示为：

$$F_{p2g} = MLP(\{P_i | P_i \in \mathrm{kNN}(P, \{P_i\}, k)\}) \tag{8}$$

式(8)中，F_{p2g}指目标点的空间特征嵌入，P_i指目标点P最近邻的第i个点。为高效融合三种模态点对点的特征，采用如公式(9)所示的全局特征融合算法：

$$F_{global} = MLP(F_{p2g} \oplus F_{dpt} \oplus F_{nv}) \tag{9}$$

公式(9)中，F_{dpt} 和 F_{nv} 分别为深度图像和法向量角度图像经各自 Swin-T 算法后的特征嵌入。为防止过拟合，全局特征融合部分仅叠加三个模态的特征，之后输入至多层感知机。得到全局特征之后，即可将全局特征与每个模态的特征嵌入融合。首先，将全局特征与点云模态进行融合：

$$F_{fp} = MLP(Pool(F_{global}) \oplus F_{dpt}) \tag{10}$$

式(10)中，F_{dpt} 指点云模态数据经过点云网络处理后的空间特征嵌入。由于 F_{global} 的尺寸与 F_{dpt} 不一致，因此使用获得 F_{p2g} 时的点云索引对 F_{global} 进行下采样。之后，通过多层感知机进行特征点云模态空间特征嵌入与全局特征的融合，得到融合特征 F_{fp}。最后，进行全局特征与深度图像模态和法向量角度模态的特征融合，如公式(11)-(12)所示：

$$F_{fd} = MLP(F_{global} \oplus F_{dpt}) \tag{11}$$

$$F_{fn} = MLP(F_{global} \oplus F_{nv}) \tag{12}$$

上述算法完成了三个模态的特征提取与融合，形成全局特征表达，使得各个模态都具有了来自深度、角度和空间的特征嵌入。编码器-解码器结构的每个阶段都使用该特征融合机制，以提高深度、角度和空间特征的交互能力。

2.3 基于改进自注意力机制的3D关键点姿态检测算法

受He等人的FFB6D工作[17]的启发，本文通过基于最远点采样算法进行目标物体表面关键点采样。为了使密集特征嵌入对实例分割、位姿矩阵关键点回归具有更精确的指向性，受CBAM算法的启发，本节分别为平移矩阵回归、旋转矩阵回归以及图像实例分割三个任务的检测头进行了改进，在检测头模块头部加入基于空间和通道的自注意模块，实现了密集特征嵌入，提高了姿态回归的稳定性。

3 实验与结果分析

3.1 实验平台介绍

本文的实验执行环境为：LinuxMint 18.1操作系统，处理器为 Intel(R) Core(TM) i9-12900K CPU，内存64GB，显卡为 NVIDIA RTX3090Ti。编程语言为Python3.7，深度学习框架为Pytorch 1.8。本文实验在两个大型公开基准数据集上进行测试。两个数据集分别为：包含133827张真实标签的YCB-Video Dataset数据集[18]；包含22000张真实标签和100000张合成标签的LineMod Dataset数据集[19]。

3.2 评估指标

本文使用两种常用的 6-DoF 姿态估计评估指标来测量算法的性能：平均距离度量 ADD 和 ADD-S[20]。对于非对称目标，ADD 算法计算框架预测姿态与真值的目标顶点之间的平均距离：

$$\mathrm{ADD} = \frac{1}{m}\sum_{p\in M} \| (Rp + T) - (R^*p + T^*) \| \tag{13}$$

式(13)中，p表示目标物体的顶点，M表示目标顶点的集合，m表示顶点的个数，R、T表示模型预测的旋转矩阵和平移矩阵，R^*、T^*表示真值旋转矩阵和平移矩阵。对于对称物体，计算平均距离度量的算法 ADD-S 为：

$$\text{ADD-S} = \frac{1}{m}\sum_{p_1\in M} \min_{p_2\in M} \| (Rp_1 + T) - (R^*p_2 + T^*) \| \tag{14}$$

在 YCB-Video 数据集的实验中，使用 ADD(S)的精度阈值曲线积分（AUC）作为指标；在 LineMod 实验中，使用平均距离度量小于物体直径的 10%（ADD(S)-0.1d）作为指标。

3.3 基于RGB-D数据的姿态估计算法实验结果及分析

在 YCB-Video 数据集上评估了本文提出的CMCL6D 方法，并与经典方法 PointFusion[21]和基线方法 DenseFusion[10]方法进行了对比。各算法以 AUC 作为评估指标，结果如表 1 所示。

表1 各算法在YCB-Video数据集上测试结果

物体编号	PointFusion (%)	DenseFusion (%)	**CMCL6D** (%)
002	90.9	96.4	**97.5**
007	93.8	96.6	**98.2**
024	75.7	88.2	**88.5**
051	65.9	72.9	**75.0**
052	60.4	69.8	**72.7**
平均值	83.9	93.1	**94.5**

由表1可知，本文提出的方法比基线方法[10]准确率提升了1.4%，这表明本文所提方法对物体大小具有较强的判别能力，对纹理较弱的物体具有较好的特征提取能力并可以准确估计位姿。

另外，本文在YCB-Video数据集的典型数据上对CMCL6D算法各阶段结果进行了可视化，如图9所示。图9表明，本文提出的CMCL6D算法可以有效获得物体高精度位姿检测结果。

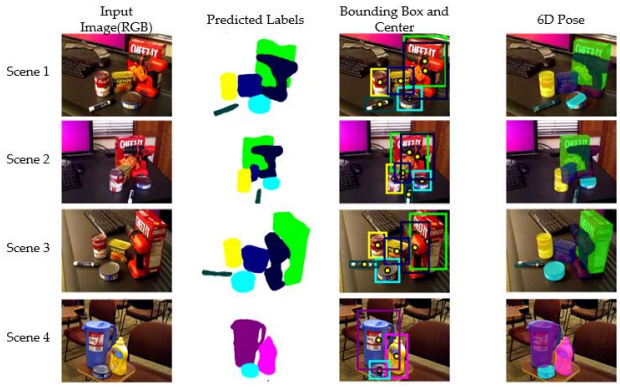

图9 CMCL6D方法姿态估计可视化

3.4 基于深度数据的姿态估计算法实验结果与分析

在LineMod数据集上定量评估了DMG6D方法，并与其他Depth-only的先进方法进行了比较。本节分别计算了非对称对象的ADD<0.1d和对称对象的ADD-S<0.1d。实验统计结果如表2所示。

表2 各算法在LineMod数据集上测试结果

物体	CATRE (%)	SwinDePose (%)	**DMG6D** (%)
ape	63.7	95.4	**96.6**
benchvise	98.6	98.2	**99.5**
camera	89.7	96.9	**99.1**
can	96.1	98.2	**99.3**
cat	84.3	98.6	**98.7**
平均值	90.9	97.5	**98.9**

表2表明，在所有以深度数据为输入的方法中，本文提出的DMG6D方法比目前最先进的方法[23]提升了1.4%。

如图6所示，在部分LineMod数据集中典型物体上对提出算法的效果进行了重投影可视化，可见DMG6D可以稳定准确的预测目标物体的位姿。

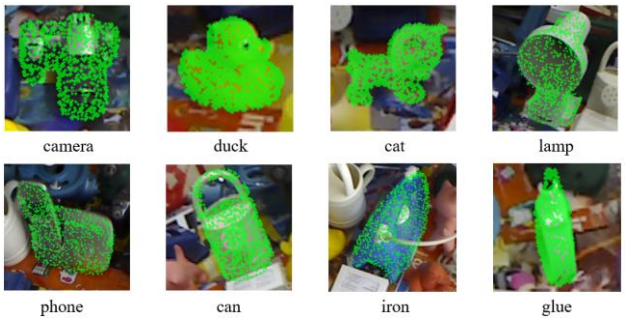

图10 DMG6D方法姿态估计可视化

4 结 论

本文致力于解决大范围光照变换场景下弱纹理目标物体的6-DoF姿态估计问题，通过提出的CMCL6D和DMG6D两种算法，有效提升了物体姿态估计的鲁棒性和精度。实验结果表明，本文提出的方法在理论和实践层面均取得了显著成效，具体结论如下：

（1）本文提出了CMCL6D方法通过跨模态特征交互模块与跨特征层特征交互模块，实现了RGB-D数据的深度融合，不仅提高了在低纹理和低光照条件下的姿态估计性能，还揭示了跨模态信息互补对于提升姿态估计鲁棒性的关键作用。

（2）本文提出了DMG6D算法，其利用深度数据和由深度数据获得法向量角度图像与点云数据，形成基于全局特征融合的端到端姿态估计框架，有效挖掘深度数据的潜力，展示了深度信息在6D姿态估计中的独特价值。

实验结果显示，这两种方法在YCB-Video和LineMod数据集上的性能超越了当前最优算法，验证了提出的多源数据融合策略对于提高姿态估计精度的有效性。这不仅解决了实际应用中因光照变化和弱纹理导致的估计困难，也为理论研究提供了新的视角和证据，即深度数据的综合运用可以显著增强姿态估计的稳健性。

尽管本研究在目标姿态估计领域取得了显著进展，但仍存在一些局限性，如对极端复杂环境下的姿态估计鲁棒性还有待提高。未来研究可探索更高效的特征提取和多源数据融合策略，以及轻量化网络设计，以适应更广泛的实时网络应用需求。

参考文献

[1] 国务院. 国务院关于印发《中国制造 2025》的通知[EB/OL]. https://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm. 2015.05.19

[2] JIANG J, CAO G, DENG J, et al. Robotic perception of transparent objects: A review[J]. IEEE Transactions on

Artificial Intelligence, 2023.

[3] COLLET A, MARTINEZ M, SRINIVASA S S. The moped framework: Object recognition and pose estimation for manipulation[J]. The international journal of robotics research, 2011, 30(10): 1284-1306.

[4] SKEIK O, ERDEN M S, KONG X. 6D Pose Estimation for Precision Assembly[C]//2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS). IEEE, 2022: 1-6.

[5] WU D, ZHUANG Z, XIANG C, et al. 6D-VNet: End-to-end 6-dof vehicle pose estimation from monocular rgb images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[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] PENG S, LIU Y, HUANG Q, et al. PVNet: Pixel-wise voting network for 6dof pose estimation[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4561-4570.

[8] 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. 2020: 11632-11641.

[9] KANG J, LIU W, Tu W, et al. Yolo-6d+: single shot 6d pose estimation using privileged silhouette information[C]//2020 International Conference on Image Processing and Robotics (ICIP). IEEE, 2020: 1-6.

[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. 2019: 3343-3352.

[11] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[12] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[13] HU Q, YANG B, XIE L, et al. Randla-net: Efficient semantic segmentation of large-scale point clouds[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11108-11117.

[14] BOUBOU S, NARIKIYO T, KAWANISHI M. Differential histogram of normal vectors for object recognition with depth sensors[C]//2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC). IEEE, 2016: 162-167.

[15] XIAO T, LIU Y, ZHOU B, et al. Unified perceptual parsing for scene understanding[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 418-434.

[16] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.

[17] 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. 2021: 3003-3013.

[18] CALLI B, SINGH A, WALSMAN A, et al. The YCB object and model set: Towards common benchmarks for manipulation research[C]//2015 international conference on advanced robotics (ICAR). IEEE, 2015: 510-517.

[19] HINTERSTOISSER S, HOLZER S, Cagniart C, et al. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes[C]//2011 international conference on computer vision. IEEE, 2011: 858-865.

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

[21] XU D, ANGUELOV D, JAIN A. PointFusion: Deep sensor fusion for 3d bounding box estimation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 244-253.

[22] LIU X, WANG G, LI Y, et al. CATRE: Iterative point clouds alignment for category-level object pose refinement[C]// European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 499-516.

[23] LI Z, STAMOS I. Depth-based 6dof object pose estimation using swin transformer[C]//2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2023: 1185-1191.