文章编号:1674-2869(2024)03-0304-06

基于大核注意力改进的工业管件位姿估计

蒋珺阳[1,2],吴晶华*[2],赵娜娜[2]
1. 安徽建筑大学机械与电气工程学院,安徽 合肥 230009;
2. 中国科学院合肥物质科学研究院智能机械研究所,安徽 合肥 230031

摘 要:在物体六自由度位姿估计任务中,现有算法在真实场景下对具有弱纹理性且摆放存在遮挡的工件难以实现准确的识别。为提高工件识别精度,提出一种基于深度学习改进的位姿估计算法。该算法采用编码器-解码器架构,引入大核注意力组成视觉注意力网络,聚焦不确定性关键点,增强特征提取能力。根据关键点对应构建密集点对关系,求解出候选位姿。实验结果表明,该算法在公共数据集和自建工业管件数据集上识别准确率分别达到了57.4%和62.1%。与高密度表面编码(Surfemb)算法相比准确率分别提升了5.5%和1.9%。这验证了该算法在遮挡场景下有更高的精准度和鲁棒性。

关键词:位姿估计;密集对应;深度学习;大核注意力
中图分类号:TP391 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202310018

Pose estimation of industrial pipe fittings based on large kernel attention improvement

JIANG Junyang[1], *WU Jinghua*[*2], *ZHAO Nana*[2]
1. School of Mechanical and Electrical Engineering, Anhui Jianzhu University, Hefei 230009, China;
2. Institute of Intelligent Machines, Hefei Institute of Physical Sciences, Chinese Academy of Sciences, Hefei 230031, China

Abstract: In the task of six-degree-of-freedom pose estimation, the existing algorithms cannot accurately recognize the workpieces with weak texture and occluded placement in real settings. To improve the accuracy of workpiece recognition, an improved pose estimation algorithm based on depth learning was proposed. A large kernel attention was first added to the encoder-decoder architecture to construct a visual attention network so that the network could focus on uncertain key points and enhance feature extraction capability. Then, the candidate pose was obtained according to the key points corresponding to building a dense point-to-point relationship. The experimental results show that the recognition accuracy of the algorithm is 57.4% on the public dataset and 62.1% on the self-built industrial pipe fittings dataset, respectively. Compared with the surface embeddings (Surfemb) algorithm, the accuracy is improved by 5.5% and 1.9%, respectively. This proves that the proposed algorithm has a higher accuracy and robustness in occluded scenes.

Keywords: pose estimation; dense correspondence; deep learning; large kernel attention

刚体的六自由度位姿估计是机器视觉领域的关键技术之一[1],在机械臂抓取[2]、自动驾驶、工业产线等领域都有广泛的应用。通过建立世界坐标系和物体像素坐标系之间的对应关系,清楚描述估计的对象在图像中的位置和方向。针对复杂场景下刚体的位姿估计受多种因素干扰,导致特

收稿日期:2023-10-24
基金项目:安徽省重点实验室基金(IRKL2022KF04);江苏省重点研发计划项目(BE2017001-1)
作者简介:蒋珺阳,硕士研究生。Email:jiangjunyang7@163.com
***通信作者**:吴晶华,博士,副研究员。Email:wjh@iamt.ac.cn
引文格式:蒋珺阳,吴晶华,赵娜娜. 基于大核注意力改进的工业管件位姿估计[J]. 武汉工程大学学报,2024,46(3):304-309,316.

征提取不准确、识别精度低等问题。现阶段六自由度位姿估计方法不断发展，利用深度学习的强大自主学习能力[3-4]，逐步演化出稀疏对应方法，此类方法虽简单易行，但当特征点的可见部分受到遮挡时，精确定位特征点将受到挑战，从而使识别产生较大误差。相比之下，密集对应方法展现优势，增加关键特征点，进一步发展为像素级二维（two dimensional，2D）到三维（three dimensional，3D）的特征点配对。2017年Hu等[5]提出PoseCNN算法，该算法以目标检测网络为主干，使用不同方法分开求取位移和旋转向量，不直接回归关键点而是输出密集对应，在遮挡情况下也有很强的鲁棒性。迭代匹配深度神经网络[6]（deep iterative matching，DeepIM）在PoseCNN算法密集对应的基础上进一步改进。2019年Peng等[7]提出逐像素投票网络（pixel voting network，PVNet），设计2D到3D的密集对应回归到特征向量场，再使用多点透视成像[8]（perspective-n-point，PnP）算法求取位移和旋转，对预测位姿准确度有较大提升。Park等[9]在2019年提出Pix2Pose算法，利用编码器-解码器结构，由输入图像预测每个目标像素对应的3D模型坐标以及置信度。李少飞等[10]使用特征融合网络将3D点云映射到2D平面生成深度特征图和法线特征图。Fan等[11-13]基于关键点密集对应对遮挡情况下的位姿估计方法进一步拓展。

弱纹理物体具有均匀的颜色和质地，缺乏明显的纹理特征，目前位姿估计算法研究主要依赖公共数据集进行训练和验证，而这些数据集多数是丰富纹理的物体或生活化物体，不符合工业中弱纹理工件的实际情况，这使得验证工件位姿估计算法识别的有效性变得困难。在自制数据集过程中，标注成本高昂，流程繁琐，需要耗费大量的时间，真实场景设置也存在不确定性和干扰因素，使工业领域弱纹理工件位姿估计存在困难。

针对上述问题，本文基于Surfemb算法，结合密集对应可以聚焦不确定性关键点的优势，进一步提升位姿估计精准度。本文算法设计融合新颖的视觉注意力机制，组成具有强大特征提取能力的视觉注意力网络，将2D图像与3D点云映射为高维特征表示，构建稠密的对应关系，有效应对弱纹理物体识别和遮挡场景，优化识别精确度，获得更好的可视化识别效果。同时为解决工作数据集标注问题，利用计算机图形学技术，调用物理引擎渲染合成弱纹理管件自建数据集，渲染虚拟背景，随机增加不同遮挡和光照影响，更适用于工业场景识别，验证本文算法具有泛化性和鲁棒性。

1　位姿估计网络结构优化改进

受当前先进的密集对应位姿估计方法Surfemb[14]算法启发，设置目标检测识别后，将裁剪完成的图像作为位姿估计网络的起始输入，通过改进主干特征提取网络，拟合成具有更强特征提取能力的视觉注意力网络，获得密集对应的像素级特征图和掩码图。通过学习特征点对应，将三原色（red green blue，RGB）图像的特征映射到有3层隐藏层的多层感知机，组成关键点对应模块，将3D模型信息输入此模块中，学习表征物体表面特征点与掩码的对应关系。使用对比损失优化特征点对应值，通过PnP算法[8]获得物体精确位姿。位姿估计结构框架由关键点对应模块、特征提取模块、损失优化模块和PnP算法模块组成，如图1所示。

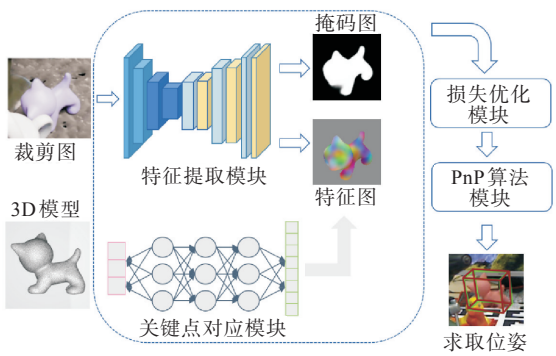


图1　位姿估计整体框架图

Fig. 1　Overall framework diagram of pose estimation

1.1　大核注意力

将注意力机制融合进神经网络等同于自适应选择的过程，根据输入特征聚焦关键特征，并自动忽略噪声响应。根据不同特征之间的关系，采用不同的优化方式。

为避免改进后的网络结构过于繁杂，增加运算时间，选取大核注意力[15]（large kernel attention，LKA）嵌入网络主干。LKA为结合卷积和注意力机制优点的线性视觉注意力。将具有大尺寸卷积核的卷积分解为3个部分，分别是深度卷积、深度膨胀卷积和点卷积。

在深度卷积的操作中，其每个卷积核只负责1个通道，区别于普通卷积的卷积核需要负责多个通道，其聚焦图像的局部特征信息使分类性能提升0.5%。深度膨胀卷积在LKA中具有长程依赖性作用，与普通卷积相比扩大了感受野。卷积结合注意力机制可以增加通道维度的适应性，LKA

模块对于输入特征的操作如式(1)所示：

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{C}_{1\times1}\left\{\boldsymbol{W}-\boldsymbol{D}-\boldsymbol{C}\left[\boldsymbol{W}-\boldsymbol{C}(\boldsymbol{F})\right]\right\} \\ \boldsymbol{O} &= \boldsymbol{A}\otimes\boldsymbol{F} \end{aligned} \tag{1}$$

式中，$\boldsymbol{A}$为注意力映射，$\otimes$为元素乘积，$\boldsymbol{D}$为深度卷积，$\boldsymbol{W}$为深度膨胀卷积，$\boldsymbol{C}$为点卷积，$\boldsymbol{F}$为输入特征，$\boldsymbol{O}$为输出特征结果。注意力映射中的值表示每个特征的重要性。设计LKA模块由视觉注意力模块和前馈网络组成，其中，视觉注意力模块结构为1×1卷积、GELU激活函数、LKA和另一个1×1卷积按顺序依次叠加，前馈网络由1×1卷积、3×3卷积、GELU激活函数和1×1卷积组成。在每个模块前添加归一化层，共同构成大核注意力模块网络，其组成结构如图2所示。

1.2 改进特征提取网络

在特征提取主干网络部分，采用经典的编码器-解码器结构。主干网络在U型网络(unity network，U-Net)架构[16]的基础上联合加载具有预训练权重的残差网络，选取残差网络ResNet34加深网络深度，避免梯度消失。

在引用LKA模块的基础上结合原有网络结构提出改进后的特征提取网络，用于处理输入特征图中的复杂局部结构，从而提取更加准确和细化的局部特征。改进后的特征提取网络结构如图3所示。

1.3 位姿生成

PnP算法核心思想为通过最小化重投影误差来估计相机的旋转矩阵和平移向量。其基本步骤为提取特征点，将特征点与物体在3D空间中的对应点进行匹配，根据匹配的2D和3D点对，求解出相机姿态。

通过计算特征提取模块提取到的2D特征图和关键点对应模块提取到的3D特征之间的相似

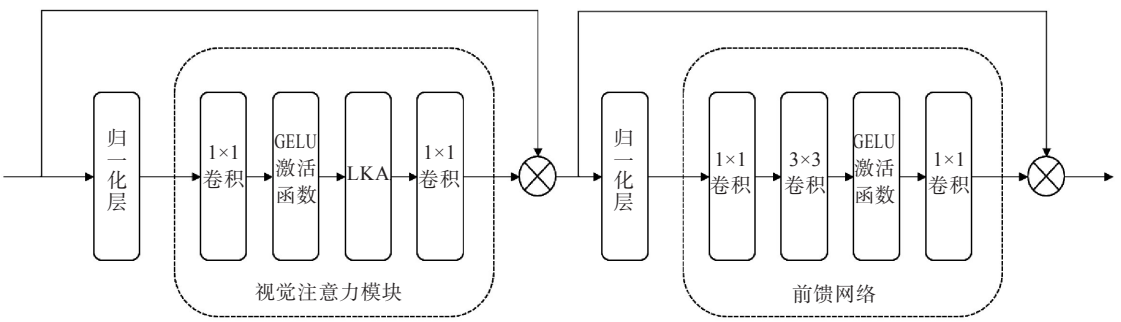


图2 LKA模块网络结构

Fig. 2 Network structure of LKA module

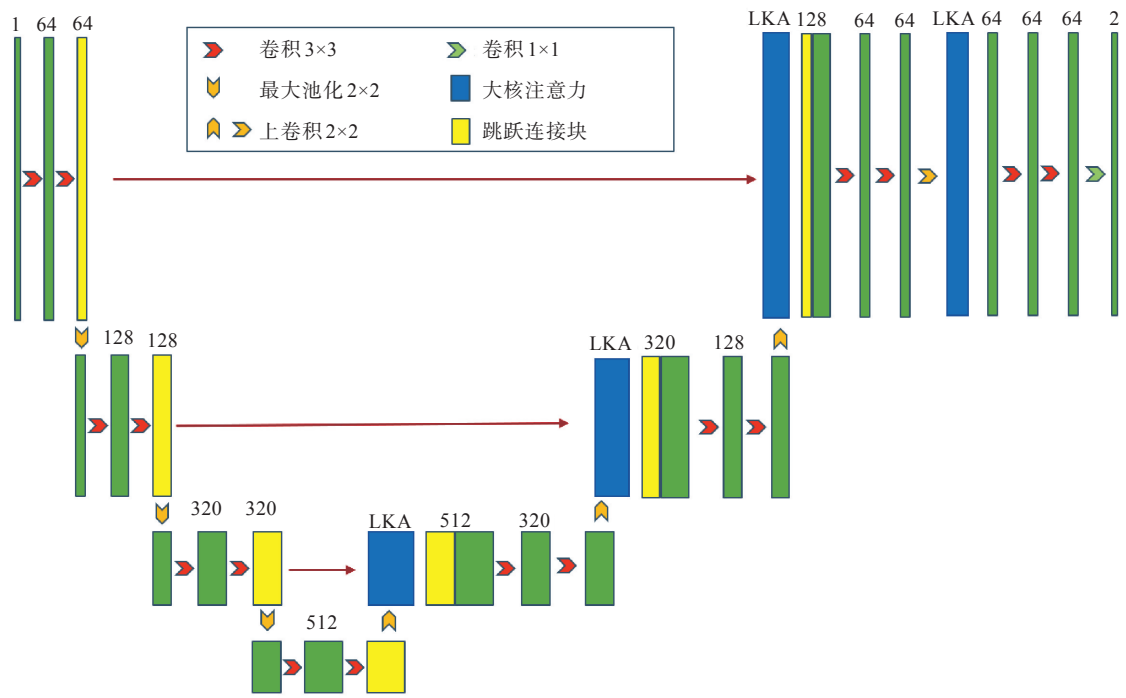


图3 改进后的特征提取网络结构

Fig. 3 Network structure of improved feature extraction

矩阵，将预测的掩码与相似矩阵进行加权可得到网络最终的相关矩阵，从而建立起密集的对应点关系，通过多次随机采样获得2D到3D的点对，结合损失优化模块，使用PnP算法[8]求解出候选位姿。

2　数据集选取与制作

2.1　公共数据集

实验采用BOP格式下的公共Occlusion LINEMOD数据集进行训练。其中50 000张不同类别图像用作训练集，1 517张图像用作验证集。

Occlusion LINEMOD数据集引入不同的光照条件和散乱遮挡情况，包含8种弱纹理物体。数据集中每一幅图像包含其对应类别、3D CAD模型和掩码图。其中光照条件包含室内以及室外环境中的自然光照和人工灯光照射下的光照变化，通过在不同光照条件下拍摄物体，使数据集具有综合随机性，提供对光照变化的鲁棒性评估。数据集中设置不同程度的随机遮挡，在遮挡环境下识别物体，使物体的姿态估计更具挑战性，符合本实验设计的针对性问题，充分体现算法整体性能。

2.2　合成数据集

为提高算法落地效率，解决标注数据繁琐、成本高昂的问题，调用物理引擎模块化BlenderProc生成大量合成数据。BlenderProc是一种基于Python开发的模块化物理渲染工具，通过参数设置调用相机模块、照明模块、材质模块和场景模块进行渲染，引入随机光照变化，模拟真实场景材质，设置符合工业场景应用中的随机遮挡。本实验模块化渲染4种工业管道连接件，生成合成数据，并按照BOP标准数据集格式输出。数据集格式包含RGB目录、实例物体的掩膜目录和实例物体深度目录。数据集中含有50个不同虚拟渲染场景，每个场景包含1 000张RGB图像和不同物体对应掩码图。选择数据集中2种不同的虚拟场景渲染图进行示例表示，如图4所示。

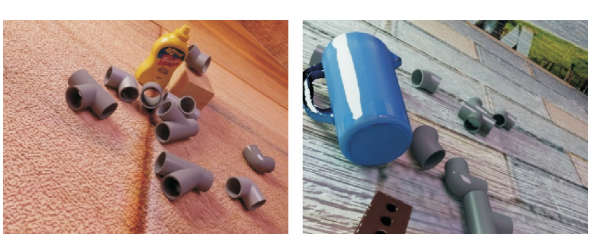


图4　管件自建数据集中不同虚拟场景渲染图示例

Fig. 4　Example of rendering different virtual scenes in self-built dataset for pipe fittings

3　实验设计与结果分析

实验环境采用Ubuntu 18.04系统，主机CPU采用32 GB内存的Xeon W-2245，显卡配置为搭载11 GB显存的2080Ti。

3.1　评价指标

数据通过训练后使用量化指标判断实验准确率，采用2D投影指标与模型点平均距离（average distance of model points，ADD）作为评价指标。

2D投影指标方法是预测位姿和真实位姿的投影之间的2D距离平均值。其计算公式如式（2）所示：

$$L_{2D}=\frac{1}{N}\sum_{i=0}^{N}\left\|K(\boldsymbol{R}x_i+\boldsymbol{t})-K(\bar{\boldsymbol{R}}x_i+\bar{\boldsymbol{t}})\right\| \quad (2)$$

式中：L_{2D}表示2D距离平均值，N为特征点集合的总数，$\boldsymbol{R}$为预测姿态的旋转矩阵，$\boldsymbol{t}$为预测平移矩阵，$\bar{\boldsymbol{R}}$为真实标注的旋转矩阵，$\bar{\boldsymbol{t}}$为真实标注的平移矩阵，K为相机参数，x_i为特征点集合中第i个点。

ADD方法为计算真实旋转矩阵和平移向量与预测的位姿进行变换后所有对应点之间距离的平均值。对于对称物体，选用平均最近点距离（average closet point distance，ADD-S）进行准确度评估。当计算差距低于物体直径的10%，则认为该位姿估计结果正确，统计估计正确的准确率。ADD，ADD-S计算公式如式（3–4）所示：

$$L_{ADD}=\frac{1}{N}\sum_{i=0}^{N}\left\|(\boldsymbol{R}x_i+\boldsymbol{t})-(\bar{\boldsymbol{R}}x_i+\bar{\boldsymbol{t}})\right\| \quad (3)$$

$$L_{ADD\text{-}S}=\frac{1}{N}\sum_{i=0}^{N}\min_{x_j}\left\|(\boldsymbol{R}x_i+\boldsymbol{t})-(\bar{\boldsymbol{R}}x_j+\bar{\boldsymbol{t}})\right\| \quad (4)$$

式中L_{ADD}表示平均距离，$L_{ADD\text{-}S}$表示平均最近点距离，x_j为特征点集合中第j个点。

3.2　对比实验

实验基于检测网络的图像裁剪，输入分辨率为224的正方形图像。使用Adam优化器优化，特征提取模块设置学习率为3×10^{-4}，关键点对应模块设置学习率为3×10^{-5}，批处理大小设置为16，最大迭代步数设置为10 000，实验训练时长为20个周期。

表1和表2分别表示在公共数据集上的2D投影指标评价结果和ADD、ADD-S评价结果。

由表1可知，在具有遮挡情况的数据集中，本文算法在2D投影指标下具有突出的鲁棒性和较高的检测精度。与Surfemb算法相比准确度平均值提升1.3%，特别是钻孔机、鸡蛋盒和打孔器类别

分别提升了3.6%、3.9%和5.1%。与PVNet和PoseCNN算法相比准确度分别提升6.4%和50.3%。

由表2可知，在ADD、ADD-S指标下，本文算法在遮挡情况下识别精度显著提高。与Surfemb、PVNet、Pix2Pose和PoseCNN算法相比，识别准确度分别提升了5.5%、16.6%、25.4%和32.5%。其中猿类别和钻孔机类别分别获得了14.8%和12.1%的大幅度提升。

挑选3种类别将位姿识别结果进行可视化展示，定性结果如图5所示。其中3D边框是由位于物体坐标系的物体包围框通过6D位姿变换渲染到图中，绿色边框代表真实的位姿结果，红色边框代表算法估计的位姿结果，因此红色边框和绿色边框重合度越高则证明算法性能越好，可以看出本文算法估计的位姿与真实位姿更为符合。

为进一步验证本文算法在工业场景下的泛化性，在自建管件数据集上进行实验和验证，实验结果如表3所示。

表1 Occlusion LINEMOD数据集2D投影指标准确度对比

Tab. 1 Comparison of 2D projection metrics accuracy for Occlusion LINEMOD dataset

算法	2D投影准确度/%								
	猿	水壶	猫	钻孔机	鸭子	鸡蛋盒	胶棒	打孔器	平均值
本文算法	70.6	93.3	75.9	84.2	74.6	15.5	48.0	78.2	67.5
Surfemb[14]	71.4	93.5	75.6	80.6	74.4	11.6	49.5	73.1	66.2
PVNet[7]	69.1	85.9	65.2	73.1	61.4	8.4	55.4	69.8	61.1
PoseCNN[5]	34.6	15.1	10.4	7.4	31.8	1.9	13.8	23.1	17.2

表2 Occlusion LINEMOD数据集ADD(ADD-S)指标准确度对比

Tab. 2 Comparison of ADD(ADD-S) metrics accuracy for Occlusion LINEMOD dataset

算法	ADD(ADD-S)准确度/%								
	猿	水壶	猫	钻孔机	鸭子	鸡蛋盒*	胶棒*	打孔器	平均值
本文算法	50.9	88.9	50.8	90.3	18.1	64.0	60.0	36.4	57.4
Surfemb[14]	36.1	82.6	47.3	78.2	17.2	64.2	55.1	34.4	51.9
PVNet[7]	15.8	63.3	16.7	65.7	25.2	50.2	49.6	39.7	40.8
Pix2Pose[9]	22.0	44.7	22.7	44.7	15.0	25.2	32.4	49.5	32.0
PoseCNN[5]	9.6	45.2	0.9	41.4	19.6	22.0	38.5	22.1	24.9

注：*为对称物体，使用ADD-S指标。

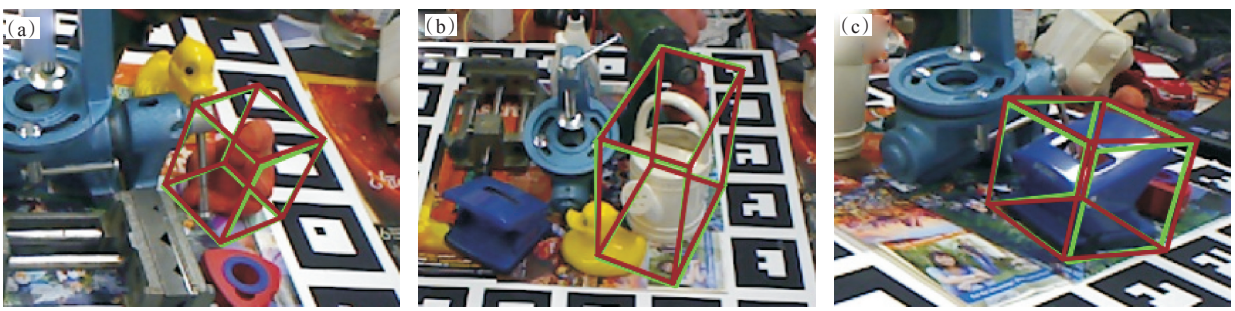


图5 部分实验结果可视化：(a)猩猩，(b)水壶，(c)打孔器

Fig. 5 Visualization of some experimental results: (a)ape, (b)watering can, (c)hole puncher

表3 自建数据集ADD指标准确度对比

Tab. 3 Comparison of ADD metrics accuracy for pipe fitting self-build dataset

算法	ADD准确度/%				
	三通道管	直角弯头管	斜三通道管	斜弯管	平均值
Surfemb[14]	55.7	56.1	65.6	63.4	60.2
本文算法	57.2	58.5	66.9	65.8	62.1

数据集中被识别管件有着不同程度的遮挡，Surfemb算法识别结果中单个管件的预测位姿与真实的位姿相差较大，而本文算法的估计结果与真实位姿更加符合，准确度提高1.9%，符合工业应用要求。直角弯头管、斜三通道管和斜弯管部分的可视化结果如图6所示，将上述3个部分分别在Surfemb算法和本文算法下的识别效果进行对比，并展示识别管件在当前图中的对应掩码图。

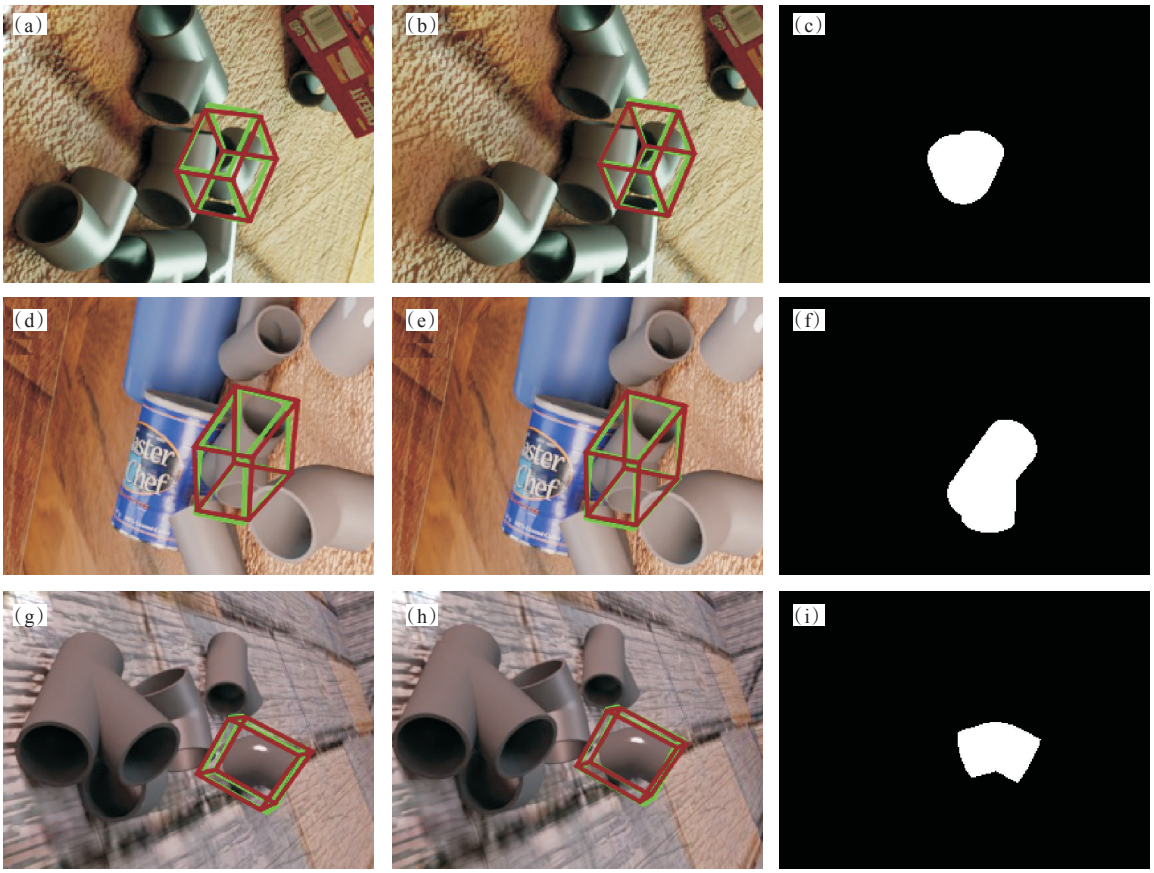


图6　自建数据集部分实验结果可视化：(a、d、g)Surfemb识别，(b、e、h)本文算法识别，(c、f、i)对应掩码

Fig. 6　Visualization of experimental results for part of self-built dataset: (a, d, g) Surfemb recognition, (b, e, h) our algorithms recognition, (c, f, i) corresponding masks

4　结　论

本文提出了一种基于深度学习的工业管件位姿估计算法，有效利用密集对应可以聚焦不确定性关键点的优势，融合具有卷积分解结构的大核注意力机制，实现对遮挡场景中物体六自由度位姿的高精度估计。实验结果表明，该算法不仅有效提高了公共数据集中存在遮挡情况的物体位姿的识别精度，还能克服纹理特征的限制，在自建工件数据集中也表现出更为精准的识别效果，验证了算法的有效性，可推广应用于其他需要识别的工件中。在未来的研究中，计划通过添加深度信息，进一步提高算法在复杂遮挡场景下的位姿估计准确度。

参考文献

[1] 郭楠，李婧源，任曦. 基于深度学习的刚体位姿估计方法综述[J]. 计算机科学，2023，50(2)：178-189.

[2] 陈立挺，聂晓根. 基于双目视觉的机械手识别、定位、抓取系统研究[J]. 机电工程，2019，36(8)：862-866，872.

[3] 张苏沛，刘军，肖澳文，等. 基于卷积神经网络的验证码识别[J]. 武汉工程大学学报，2019，41(1)：89-92.

[4] 陈希彤，卢涛. 基于全局深度分离卷积残差网络的高效人脸识别算法[J]. 武汉工程大学学报，2019，41(3)：276-282.

[5] HU Y L, HUGONOT J, FUA P, et al. Segmentation-driven 6D object pose estimation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3385-3394.

[6] LI Y, WANG G, JI X Y, et al. DeepIM: deep iterative matching for 6D pose estimation[J]. International Journal of Computer Vision, 2020, 128(7): 657-678.

[7] PENG S, LIU Y, HUANG Q X, et al. PVNet: pixel-wise voting network for 6DoF pose estimation[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 4561-4570.

[8] LEPETIT V, MORENO-NOGUER F, FUA P. EP*n*P: an accurate o(*n*) solution to the P*n*P problem[J]. International Journal of Computer Vision, 2009, 81(2): 155-166.

（下转第316页）