文章编号：2095-2163(2022)06-0013-08 中图分类号：TP391 文献标志码：A

PVANet：针对弱纹理工业零件的像素级 6DoF 位姿估计方法

杨 纯，陈 权，王 涛
（广东工业大学 计算机学院，广州 510006）

摘 要： 位姿估计在工业场景进行零件的拣选抓取时扮演着非常重要的角色。但是，目前针对杂乱场景下工业弱纹理零件的 6DoF（6 Degrees Of Freedom）位姿估计的研究还较少。特别是当这些工件使用相同的材质，且形状相近时，对位姿估计提出了更大的挑战。本文针对杂乱场景下工业弱纹理零件的 6DoF 位姿估计方法进行了研究。在杂乱场景下，首先使用了针对该场景下工业弱纹理零件位姿数据集的获取方法。接着，提出了一种基于 PAVNet 网络与注意力机制的学习框架。在该框架中，选用多阶段处理的方法从像素级对目标对象进行特征提取，再通过基于 RANSAC 的投票算法选出阈值范围内的关键点，最后通过求解这些关键点位置与工件的旋转和平移的关系，由此来估计其位姿。本文通过在公用数据集以及真实数据集上的实验验证了本文提出算法的精度，并且满足了工业应用要求。

关键词： 金属工件；弱纹理；位姿估计；像素级

PVANet：Pixel-wise 6DoF pose estimation method for weakly textured industrial parts

YANG Chun，CHEN Quan，WANG Tao
（School of Computer Science and Technology，Guangdong University of Technology，Guangzhou 510006，China）

【Abstract】 Pose estimation plays a very important role in picking and grabbing parts in industrial scenes. However，there are few studies on 6DoF（6 Degrees Of Freedom）pose estimation for industrial weakly textured parts in cluttered scenes. Especially when these workpieces use the same material and have similar shapes，it poses a greater challenge for pose estimation. This paper studies the 6DoF pose estimation method for industrial weakly textured parts in cluttered scenes.In the cluttered scene，this paper uses the acquisition method for the pose dataset of industrial weakly textured parts in this scene，and proposes a learning framework based on PAVNet network and attention mechanism. In this framework，this paper first uses a multi-stage processing method to extract features in the target objects from the pixel level，then selects key points within the threshold range through a RANSAC-based voting algorithm，finally solves the relationship between the positions of these key points and rotation and translation of the workpiece to estimate its pose. Based on the above，this paper verifies the accuracy of the proposed algorithm through experiments on public datasets and real datasets，and meets the requirements of industrial applications.

【Key words】 metal workpiece；weakly texture；pose estimation；pixel-wise

0 引 言

位姿估计在机器视觉领域扮演着十分重要的角色，尤其是一些应用场景通过使用视觉传感器进行导航，增强现实等操作，需要找到现实世界和图像投影之间的对应点。比如，在工业作业场景的抓取任务中，经常会遇到几种工件堆放散乱、待抓取物体的表面纹理信息不够丰富的场景，由于材质相同，光线在金属介质表面的传播性质，以及在光线不足的情况下，甚至会因为彼此间遮挡产生阴影，导致工件边缘的重要信息较为模糊，特征提取不够突出，从而严重影响到指定任务的抓取执行。

现如今的位姿估计方法大都在公用数据集上具有很好的鲁棒性，由于场景的改变存在诸多不确定性问题。如 Tless 等数据集往往体量过于庞大，He 等人[1]提出的方法在这些数据集上表现良好，但是受硬件因素的约束导致训练困难，虽然网络效果很好，却因网络设计复杂而难以快速部署到机器人系统，从而影响实际作业效率。而其他的一些数据集、如 linemod 等存在遮挡或截断等特点，且过于生活化，表面纹理色彩都很丰富，无法满足一些特殊的场景需求、如本文探讨的金属工件抓取问题，为此本文从数据集的制作开始，结合其他网络的优点进行弱纹理金属工件的 6DoF 位姿估计的实用型研究。

本文针对上述问题，拟从单个 RGB 图像的角度，结合注意力机制，将像素级上效果良好的一种方

作者简介：杨 纯（1998-），女，硕士研究生，主要研究方向：6 自由度位姿估计；陈 权（1989-），男，博士，讲师，硕士生导师，主要研究方向：物联网、无线网络、分布式算法设计和分析等；王 涛（1983-），男，博士，副教授，硕士生导师，主要研究方向：制造物联网、工业过程大数据、智能装备与机器人系统等。

通讯作者：陈 权 Email：quan.c@ gdut.edu.cn

收稿日期：2022-01-05

法[2]扩展到一个新的位姿估计分支 PVANet，使其用于工件的精确抓取任务。本文的主要贡献是将工件的小型数据集成功拟合进这个网络，对网络模型的部分结构做出重要调整，优化精确度。

将注意力机制结合深度学习网络进行训练的方法主要是通过掩码来实现，通过不断地学习，使深度神经网络学习到数据集中每一张图片中感兴趣的区域。一些网络[3]挖掘到了通道注意力机制的优点，指出不同通道的特征图的作用权重不同会严重影响结果，Jaderberg 等人[4]提出的空间注意力机制，发现包含对象的检测区域相较于其他背景信息的重要性要大很多。鉴于这些优点，很多研究提出了结合通道注意力和空间注意力的方法[5]，充分发挥两者的重要性，并将其功能进行结构化设计，本次研究合理利用了这一优点来提取了局部信息。

分析可知，对于这种距离相机视点较远的情况，深度信息已经不太可靠，相较于一些使用 3D 定位和旋转的方法，本文从 pixel-wise 或者 patch-wise 上进行投票选出 2D 关键点的方法，如图 1 所示，这在 Yu 等人[6]的方法中也有体现。但是实验中忽略像素点和关键点之间的距离对假设偏差影响不大的情况，此后将采用(Effective Perspective-n-Point，EPnP)根据 2D-3D 对应的方法进行位姿估计，在原工作基础上提出一些改进，结合目标检测和位姿估计的端到端通道，通过二维 RGB 图像和相关的 3D 模型建立对应关系，回归位姿参数 R 和 T。本文主要贡献如下：

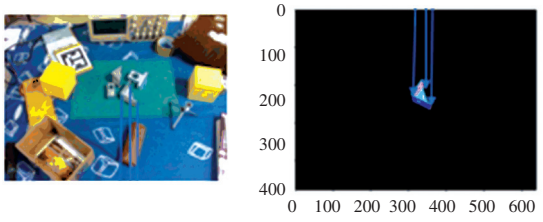
(1) 使用较少的数据模态预测弱纹理工件位姿，弥补了位姿估计数据集在工业零件方面的空缺。

(2) 分析网络深度和数据集规模的关系，将注意力机制融入像素级投票网络，并进行一些重要的调整使其能够更好地进行迁移使用。

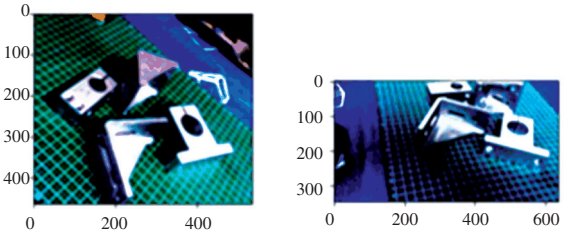
(3)改进后的方法在自定义数据集和 Linemod 上的 *ADD* 评估精度在 0.9 以上，达到工业应用要求，且可视化效果更好。

1 相关工作

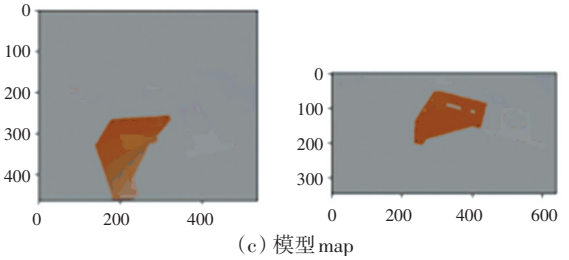
目前比较成熟的位姿估计方法包括但不限于基于对应、基于模板、基于投票这三种，并且具有比较完整的实现过程。其中，基于对应的方法[7-8]通过隐式地回归 3D 点在 2D 图像上的若干投影点，再使用 PnP 进行位姿细化。基于模板的方法[9-11]将模型的 RGB 图像结合精心设计的 CNN 取得很好的位姿估计的结果。使用投票策略的方法中，最重要的是充分利用像素信息，Brachmann 等人[12]充分利用每一个像素来产生一个 3D 坐标轴，Peng 等人[2]通过像素投票生成 2D 关键点，另一部分则是使用霍夫投票获得很好的结果。这些方法通过直接或间接地从 RGB 图像中恢复 6D 位姿。另一个大的分支是在卷积网络中结合深度信息，自从 PointNet 系列[13]的重大创新后，直接通过点云信息进行位姿估计的方法被提出[1,14]，而 Wada 等人[15]使用该方法在处理弱纹理目标时甚至都能获得很好的效果，但有关的研究一般是在大型公用数据集上不断提升算法的精确度，这在应用于实际场景时就会出现如下类似问题的探索。



(a) 2D 和对应的 3D 结果



(b) 数据增强



(c) 模型 map

图 1 投票后选出 2D 关键点

Fig. 1 The selected 2D keypoints after voting

对于工业场景中常见的无纹理金属工件，在光照等因素的干扰下，RGB 图像中可用的信息很少，目前主要的解决办法是利用图像中边缘像素的底层特征进行计算，如 Zhang 等人[16]提出使用多阶段细化的方法实现简单的抓取任务。由于金属在不同光照角度导致粗糙表面反光使得 RGB 不可忽视，充分考虑这些 RGB 图像和模型本身携带的信息能够在一定程度上降低成本，仅从 RGB 图像检测 6D 位姿对于其他类别的机器人应用也是同样重要。例如从单目稀疏视角考虑直线轮廓之间的相互关系作为描述金属零件的高级几何特征，放弃利用像素这一重要元素[17]，或是在有限样本数的单 RGB 研究上给

出了很好的示例[18]，但目前仍是在具有丰富纹理的常见生活物品对象上做进一步的提升。本文的研究对象是弱纹理的金属工件，从像素级进行探索，并使用投票对遮挡物体进行位姿预测。

投票预测局部不可见点的位置时，先根据 3D 模型点中的关键点投影到 2D 像素平面，目前已有方法[19]提供了一些 3D 特征描述子的实验效果，表明都能检测一定数量的特征点，但是如果限制特征点的数量进行投影，用于表面信息本就不丰富的工件则情况不一定很好。一些基于点对特征的方法试图通过使用点云上的少量点对构成描述子进行位姿估计，如 Drost 等人[20]、Papazov 等人[21]提出的全局建模、局部匹配，Hinterstoisser 等人[22]优化前者也使用到的 PPF 描述子来达到最佳效果，但这些方法对于场景简单、成本低的数据来说很有可能导致过拟合，且需要使用点云扫描仪器进行额外的数据采集。另外一些使用随机森林[12,23]的方法，通过霍夫投票逐像素投票，或者使用深度学习的方法提取特征[11,18]，甚至结合深度信息[24-25]，这些密集的 2D-3D 对应虽然对遮挡场景具有鲁棒性，但是网络体量大，鉴于此，本文采用 FPS 随机选择 8 个点作为候选关键点的方法，保证每次的点都不一样，减少人为因素的影响，将 RANSAC 方法重新定义为投票方法，通过逐像素迭代淘汰假设关键点的方法对 2D 关键点进行投票，结合了密集融合的方法和基于关键点的方法的优点，针对特征提取不够全面的问题，有效融合了注意力机制，进行网络效益的提升。

2　工业弱纹理零件位姿数据集的获取方法

2.1　方法概述

在进行位姿估计之前，首先要构建符合场景并带有位姿标签的数据集。目前很多先进的方法都是在公用数据集上进行精度提升，这些通用措施导致的一些局限性无法扩展到其他特殊场景。

本文方法使用 Glocker 等人提出的主要步骤进行多边形模型的 3D 重建，并将其用于单个物体的检测[26]。相较于其他流行的模型重建方法，这是为数不多的利用物体表面信息进行重建的手段，在小场景的重建上相较于其他算法取得更好的效果，和 Weise 等人[27]的研究类似，这使得一些操作虽然枯燥、但容易着手，具体的标注流程如图 2 所示。由图 2 可知，获取视频流序列中间的 100 s，通过对这些序列进行场景稠密重建，截取包含工件的一定范围场景后，导入工件的 CAD 模型进行粗略关键点匹配，再利用 ICP 进行细化后，手动调整工件模型位姿，并根据获取的位姿对工件模型进行投影获取标签。研究可知，初始场景为包含 4 个形状不同、纹理和材质相同的工件随意摆放在背景杂乱的工作台，数据的采集过程是将相机安装在机械臂末端，通过机械臂的运动来采集数据。

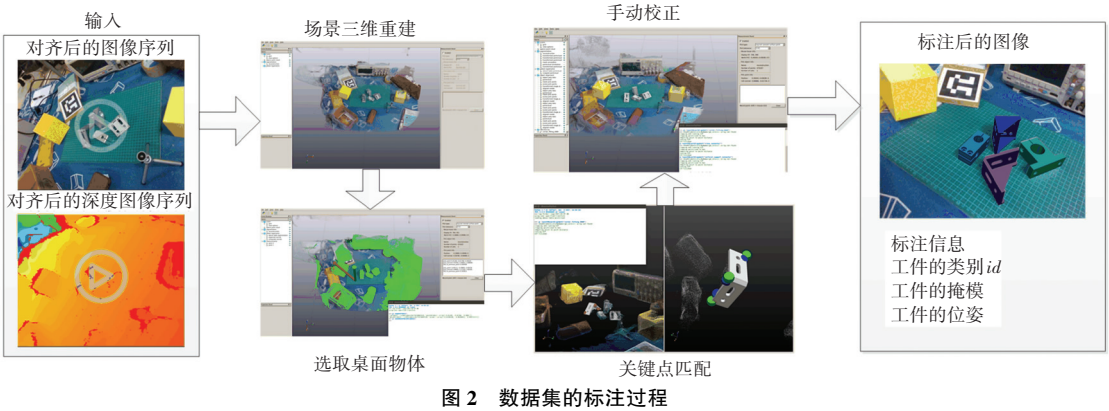


图 2　数据集的标注过程

Fig. 2　Annotation process of data set

实验使用的都是真实数据集，经测试按照随机 3:1 的比例分别抽取数据制作训练集和测试集进行训练和测试时效果最好，在无合成数据的情况下，能够尽量维持不同帧之间标签的语义相关性。为了得到 mask 这一重要因素，一些算法[28-29]通过实例分割把对象从场景中分离出来，但是目前的分割网络为了得到精度更高的结果，模型体量都比较大，这在工业应用上将显著影响作业效率，本文在实际使用中利用标注的位姿，通过模型投影可直接获得。

2.2　位姿描述

对姿态的描述是机器人进行位姿估计的基础，包括欧拉角、旋转+平移，以及四元数表示。

对于 3D 空间的任一参考系，任何其他的坐标系都可以用 3 个欧拉角表示，即通过绕着 x，y，z 这 3 个轴旋转的 3 个角度进行组合表示，由于参数的显式意义，这种表示是直观的，并且旋转向量与旋转矩

阵的相互转换可以用罗德里格斯公式来解决,但是在一些情况下却不能实现平滑插值,甚至还会产生万向节死锁问题,通常在有关旋转的应用场景中基本不使用欧拉角来旋转,而是使用上述后 2 种进行表示,相互之间也可进行转换。

相较于旋转矩阵需要满足单位正交的限制,如何在训练目标中加入该限制条件是难点之一,在这项工作中,本文使用的是四元数(1)这种计算量偏小的位姿表示:

$$Q = t + x\,i_1 + y\,i_2 + z\,i_3 \quad x, y, z \in \mathbb{R} \tag{1}$$

其中,$i_1^2 = i_2^2 = i_3^2 = -1$,形式上和复合函数的表示一样;实部 t 表示目标对象的平移,且 $t \in \mathbb{R}^3$;3 个虚部 x,y,z 表示 3D 空间的旋转 R,且 $R \in SO(3)$。

在进行网络训练前,本文对真实位姿进行预处理,在数据处理的过程中,尤其要注意实部与虚部的相对位置关系,否则回归研究后的结果就会出现如图 3 所示的由于旋转矩阵转换为四元数时使用了函数的默认顺序导致的位姿偏差过大的问题。

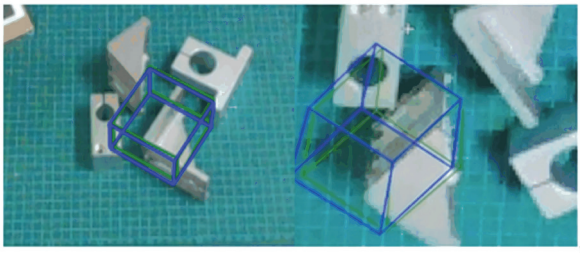


图 3 可能出现的偏差过大的问题

Fig. 3 Possible problems of excessive deviation

3 本文方法

3.1 结合注意力机制的 PVNet 网络架构

图 4 为经过调整后的模型。图 4 中,以 ResNet-18 为主干网络,增加注意力机制强化特征提取性能,网络的输入为自定义数据集,输出为掩膜分割和向量,然后用 RANSAC 投票出关键点,最后使用 PnP 回归位姿。和 Peng 等人[2]的相关研究类似,使用预训练的 ResNet-18 为主线,重点在预测像素的方向、而不是从图像中直接回归关键点的位置,即网络的主要作用是预测向量场和生成对象标签,通过重视目标的局部特征,减轻了杂乱背景的影响。对于图像中的任意一个像素点 p,坐标表示为 (u,v),将其到目标对象的 2D 关键点 x_k 的方向定义为向量 $\boldsymbol{v}_k$,即:

$$\boldsymbol{v}_k(p) = \frac{x_k - p}{\|x_k - p\|^2} \tag{2}$$

其中,x_k 是通过最远点采样方法获取的模型 3D 点通过投影矩阵获得,p 的坐标是根据式(1)所得姿态,结合相机内参通过向投影矩阵公式(3)带入 K 计算得到,即:

$$\boldsymbol{Z}\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} \triangleq \boldsymbol{K}\boldsymbol{P} \tag{3}$$

其中,(X,Y,Z) 表示世界坐标系下点 P 的位置,$\boldsymbol{K}$ 为本实验中 D435 系列相机对应的内参矩阵。

给定语义标签和单位向量,物体的所有像素都对通过基于投票的 RANSAC 机制生成关键点假设进行投票,这些投票中会有置信度分数较高的一些假设(大于设定的阈值),通过 RANSAC 策略,使用循环迭代计算出来的最好模型再一次生成假设坐标并进行关键点的投票,用这些假设表示图像中关键点的空间概率分布是很可靠的,因为这样与更多的预测方向重合,局部不合适的点的投票只占少量。

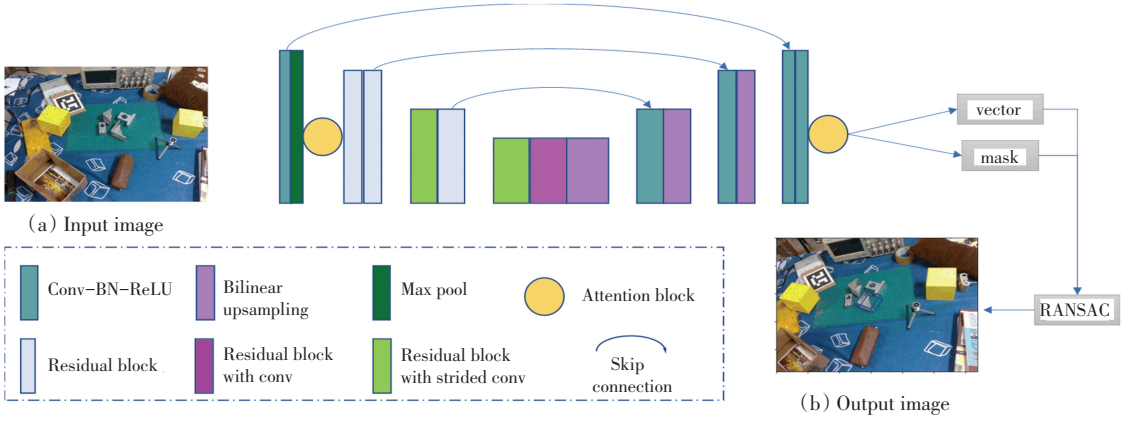


图 4 本文网络管道说明

Fig. 4 The network pipeline in the paper

本文在对原始网络进行迁移使用的时候，发现其本身效果已经具有一定的准确性，可视化结果详见实验部分，但是网络庞大，本文在保持原来方法的主要步骤的情况下，对局部结构进行了调整。对此可给出研究分述如下。

（1）对一些效果不明显的设计进行了更改，具体但不仅仅包括将原来在 skip connection 中间部分的 Residual block with dilated conv，改为普通的 3×3 卷积，该卷积结构试图通过空洞卷积的参与来增加感受野范围，学习到更多的特征，然而对于本文这种远离视角的小物体分割，特别是尺寸都差不多的工件来说有弊无利，dla 可能导致局部信息缺失，颜色纹理相近的工件特征相关性匮乏，从而影响最终的分类结果。

（2）将注意力机制模块结合进 ResNet-18 进行特征提取，其有效性已经在某些工作[5]中体现得很充分，本文的工作是在网络的第一层，即使用最大值池化前、最后一层，即使用均值池化前加入注意力机制模块，而不是放在残差块中，并且是用 ImageNet 的预训练权重字典，以充分提取局部特征，不忽略每一层特征图在训练时的不同作用比率。值得注意的是，注意力用在位姿估计的场景还不是很广泛。

（3）PVNet 工作在对 ResNet-18 进行 fine tuning 时，是将最后的 1×1 之前的所有的 FC 改为 Conv，这么做是考虑到 FC 如果过多，且形状都不小，容易导致内存消耗严重。但是一些研究中[30]表明适当的 FC 设计可在模型表示能力迁移过程中充当防火墙的作用，不含 FC 的网络微调后的结果要差于含 FC 的网络，事实确实如此，特别是本文的自定义数据集和原始结构使用的公用 linemod 数据集的对象完全不一样的情况下，FC 可保持较大的模型容忍度，从而保证模型表示能力的迁移，因此本文又一次强调在合适的情况下 ResNet-18 的最后一层 FC 设计的重要性，以及允许部分 FC 存在。另外，ResNet 在很多应用场景中都占有很重要的一部分比重，但是相对于其他很多领域的数据集，所有位姿估计的数据集体量都非常地大，如何减少内存浪费是很重要的事情。

当然网络的大部分还是值得本文借鉴的，输入图像大小为 $H \times W \times 3$，当网络的特征图的大小为 $H/8 \times W/8$ 时，不再为了提高分辨率而对特征图进行下采样，丢弃后续的池化层，这在一定程度上阻止了后续无意义的操作导致的消耗。

3.2　损失函数设计

为了训练网络，本文使用了比较稳妥的损失函数来联合训练包围框位置、分割、投票、框内的姿态。形式上，损失计算包含 2 部分，投票的损失计算使用 Smooth L_1 损失函数[1,18,31]，$L_{EntropyCross}$、即 *softmax* 交叉熵损失，可用于训练语义标签，实验中使用的损失函数定义如式（4）所示：

$$L = L_{vote} + L_{seg} = L_1 + L_{EntropyCross} \tag{4}$$

其中，Smooth L_1 是 Smooth L_1 损失，是 L_1 与 L_2 损失的结合，$L_{EntropyCross}$ 是该分类问题中的常见解决办法。

4　实验与分析

本文研究的对象是金属工件，为了提高分拣抓取等操作的精确度，针对这些弱纹理工件进行 6DoF 位姿估计。实验使用的数据集经过格式转换，以适配一些算法的数据读取接口。实验中涉及的高性能计算的网络训练部分均在 2080Ti 上进行。

研究可知，PVNet 中，输入 RGB 图像，通过基于 RANSAC 的投票方法给所有向量、即像素指向每个关键点的方向进行打分，由此得到分数高于一定阈值的关键点的空间分布，详细的介绍参见文献[2]。

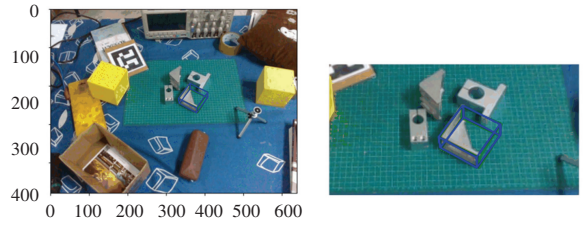
本文的方法步骤也是如此，但是基于本文对数据集更改的考虑、即目标对象完全不一样，以及新数据集中检测对象的视点较远的情况，对网络进行了一些改动，使其更好地适用于本文的工作。

由于数据集中的数个初始研究对象为 3D 对称物体，本文实验中使用了由 Xiang 等人[10]提出的 ADD-s 指标，用来评估网络输出位姿和真实位姿转换后的 2 个模型对应点之间的平均距离，即当这个距离小于模型直径的 10%时，就认为估计出来的位姿是正确的，对于这种立体几何形状的对象直径则根据模型最远对角点的距离进行计算。为此，这里将给出剖析阐释如下。

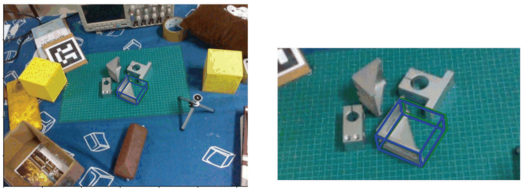
（1）Linemod 数据集上的性能。由于本文的方法大多集中在场景不同于 BOP 等数据集的弱纹理场景进行位姿估计，而且主要是使用 RGB 进行这项工作，因此本文对比了使用 Depth 后的先进网络效果、原网络进行较大改动后的效果、以及使用本文的方法进行微调后的更好的结果，优化后的算法在 Linemod 数据集上的性能表现见表 1。由表 1 可知，相较于 PVNet，增加了注意力机制后的效果有所提升。

（2）真实数据集上的性能。对比网络深度的实

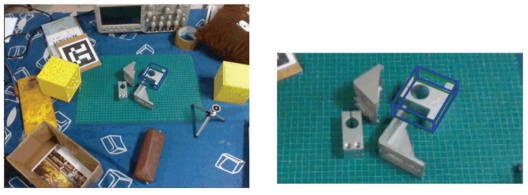
验效果如图 5 所示。图 5 中,绿色框表示 ground truth,蓝色框表示网络输出结果。图 5 从(a)~(d)依次为 ResNet50(工件一)、ResNet34(工件一)、ResNet-18(工件二)、ResNet-18(工件一),其中 ResNet-18 为调整后的网络,层数变动不大。仍需指出的是,图 5(a)~(d)中,左侧图为经过网络调整后的可视化结果,右侧图为意在方便比较进行的相同比例放大。图 5 的结果表明随着网络的加深,效果并没有较大的改进,但是使用较少的残差块,对网络适当地剪枝,得到的效果更好。调整后的算法对比其他网络使用本文的数据集的 *ADD* 结果见表 2。由表 2 可知,本文的实验效果更好,但是目前比较好的网络都已经能达到这样的效果。本文对场景中的其他数个物体也进行了相同的步骤,但是实验结果相近就不在文中加以赘述了。



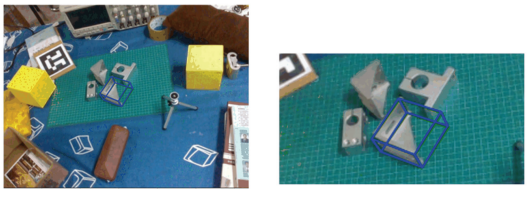
(a) ResNet50(工件一)



(b) ResNet34(工件一)



(c) ResNet-18(工件二)



(d) ResNet-18(工件一)

图 5 对比网络深度的实验效果

Fig. 5 Comparing the experimental effect of network depth

表 1 和其他算法在 Linemod 数据集上的表现相比较

Tab. 1 Comparison with the performance of other algorithms on Linemod data set

	DenseFusion	PVNet	PVANet(The proposed)
ape	79.5	43.6	75.6
bench	84.2	99.9	99.8
camera	76.5	86.9	87.1
can	86.6	95.5	95.5
cat	88.8	79.3	89.0
driller	77.7	96.4	96.0
duck	76.3	52.6	71.0
eggbox	99.9	99.2	99.8
glue	99.4	95.7	99.8
hole	79.0	81.9	79.0
iron	92.1	98.9	92.0
lamp	92.3	99.3	91.0
phone	88.0	92.4	90.2
MEAN	86.2	86.3	89.2

表 2 调整后的算法对比其他网络使用本文的数据集的 *ADD* 结果

Tab. 2 The comparison of *ADD* results using the dataset in this paper between the adjusted network and other networks

methods	DenseFusion	PVNet	PVANet(The proposed)
ADD(-*s*)*AUC*	90.35	93.34	94.51

5 结束语

由于不同的抓取场景所针对的研究对象的自身属性、诸如金属工件反光等因素会导致不同的位姿估计问题,本文从数据集制作、方法实现等角度探讨了输入 RGB 进行位姿估计的框架,并做出一些重要的改进以便执行实际场景下的任务,如抓取、拣选等。但这些方法都是基于一定的使用条件下,并且研究可知一个正确且精确度高的对象模型对于 3D 目标检测、位姿估计任务极具重要性。但是为了更好地服务于工业发展的需要,仍会有很多当模型不存在时进行精确操作的情况、如类级别位姿估计。本文虽然对常见的工件进行了探索,但是零件间不同的遮挡情况会导致零件外形在孔的位置、形状等地方有些许的不一样,因此后续工作可以在此基础上进行拓展,以应对更多的特殊场景。

此外,本文在数据集上的规模上还有一些不足,一方面受制于没有掌握合成包含符合场景的数据集制作方法,另一方面真实数据集的标注需要耗费较大的人力,因此后期在数据集的扩充上也要再做更进一步的探索。

参考文献

[1] HE Yisheng , SUN Wei , HUANG Haibin , et al. PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2020:1-14.

[2] PENG Sida, LIU Yuan, HUANG Qixing, et al. PVNet: Pixel-wise voting network for 6DoF pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Long Beach, CA, USA:IEEE,2019:4556-4565

[3] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(8):2011-2023.

[4] JADERBERG M, SIMONYAN K, ZISSERMAN A. Spatial transformer networks [C]//Advances in Neural Information Processing Systems.Montreal, Quebec, Canada:NIPS Foundation, 2015: 2017-2025.

[5] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module [M]//FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision - ECCV 2018. ECCV 2018. Lecture Notes in Computer Science. Cham:Springer, 2018,11211:3-19.

[6] YU Xin, ZHUANG Zheyu, KONIUSZ P, et al. 6DoF object pose estimation via differentiable proxy voting loss[J]. arXiv preprint arXiv:2002.03923, 2020.

[7] HU Yinlin, HUGONOT J, FUA P , et al. Segmentation-driven 6D object pose estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Long Beach, CA, USA: IEEE, 2019:3380-3389.

[8] ZAKHAROV S, SHUGUROV I, ILIC S. DPOD: 6D pose object detector and refiner [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV).Seoul, Korea (South): IEEE, 2019:1941-1950.

[9] SUNDERMEYER M, MARTON Z C, DURNER M, et al. Implicit 3D orientation learning for 6D object detection from RGB images[J]. International Journal of Computer Vision,2020, 128: 714-729. International Journal of Computer Vision volume 128, pages714-729 (2020)

[10] XIANG Yu, SCHMIDT T, NARAYANAN V, et al. PoseCNN: A Convolutional Neural Network for 6D object pose estimation in cluttered scenes[J]. arXiv preprint arXiv:1711.00199,2017.

[11] LI Zhigang, WANG Gu, JI Xiangyang. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation[C]//IEEE/CVF International Conference on Computer Vision (ICCV).Seoul, Korea:IEEE,2019:7677-7686.

[12] BRACHMANN E, KRULL A, MICHEL F, et al. Learning 6D object pose estimation using 3D object coordinates[M]//FLEET D, PAJDLA T, SCHIELE B, et al. Computer Vision - ECCV 2014. ECCV 2014. Lecture Notes in Computer Science. Cham: Springer, 2014,8690:536-551.

[13] QI C R, SU Hao, MM Kaichun, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017:77-85.

[14] SONG Chen, SONG Jiaru, HUANG Qixing. HybridPose: 6D object pose estimation under hybrid representations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA:IEEE, 2020:428-437.

[15] WADA K, SUCAR E, JAMES S, et al. MoreFusion: Multi-object reasoning for 6D pose estimation from volumetric fusion [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Seattle, WA, USA: IEEE, 2020: 14528-14537.

[16] ZHANG Haoruo, CAO Qixin. Detect in RGB, optimize in edge: Accurate 6D pose estimation for texture-less industrial parts[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal, QC, Canada:IEEE,2019:3486-3492.

[17] 江智伟. 基于单目稀疏视角的无纹理零件位姿估计技术研究[D]. 杭州:浙江大学,2019.

[18] DO T T, CAI Ming, PHAM T, et al. Deep-6DPose: Recovering 6D object pose from a single RGB image[J]. arXiv preprint arXiv:1802.10367, 2018.

[19] ZHOU Ruqin, LI Xixing, JIANG Wanshou. 3D surface matching by a voxel-based buffer-weighted binary descriptor[J]. IEEE Access, 2019, 7:86635 - 86650.

[20] DROST B, ULRICH M, NAVAB N, et al. Model globally, match locally: Efficient and robust 3D object recognition[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA:IEEE, 2010:998-1005.

[21] PAPAZOV C, BURSCHKA D. An efficient RANSAC for 3D object recognition in noisy and occluded scenes[M]// KiMMEL R, KLETTE R, SUGIMOTO A. Computer Vision-ACCV 2010. ACCV 2010. Lecture Notes in Computer Science. Heidelberg/Berlin: Springer, 2010,6492:135-148.

[22] HINTERSTOISSER S, LEPETIT V, RAJKUMAR N, et al. Going further with point pair features[M]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision - ECCV 2016. ECCV 2016. Lecture Notes in Computer Science. Cham:Springer,2016,9907: 834-848.

[23] MICHEL F, KIRILLOV A, BRACHMANN E, et al. Global hypothesis generation for 6D object pose estimation[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE, 2017:115-124.

[24] KEHL W, MILLETARI F, TOMBARI F, et al. Deep learning of local RGB - D patches for 3D object detection and 6D pose estimation [C]//European Conference on Computer Vision. Cham:Springer, 2016:205-220.

[25] WANG He, SRIDHAR S, HUANG Jingwei, et al. Normalized object coordinate space for category-Level 6D object pose and size estimation[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA:IEEE, 2019: 2637-2646.

[26] WHELAN T, SALAS-MORENO R F, GLOCKER B, et al. ElasticFusion: Real-time dense SLAM and light source estimation [J]. International Journal of Robotics Research, 2016, 35(14): 1697-1716.

[27] WEISE T, WISMER T, LEIBE B, et al . In-hand scanning with online loop closure [C]//IEEE International Conference on Computer Vision Workshops. Kyoto, Japan:IEEE, 2009:1630-1637.

(下转第 27 页)