Computer Science and Application 计算机科学与应用, 2022, 12(1), 252-261
Published Online January 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.121025

Hans汉斯

基于碎片模型的弱纹理物体位姿估计

李　耀，程良伦，王　涛

广东工业大学计算机学院，广东　广州

收稿日期：2021年12月26日；录用日期：2022年1月21日；发布日期：2022年1月28日

摘　要

位姿估计已广泛应用于智能机器人，无人驾驶以及增强现实等多种应用场景，现有基于神经网路的点对匹配方法大多未能处理好前景遮挡，弱纹理以及对称。本文提出了一种基于碎片模型的弱纹理物体位姿估计方法，该方法利用现有的3D碎片模型从一幅RGB输入图像中估计出刚体物体的6D姿态；物体由紧凑表面碎片表示，能够系统地处理物体对称；使用编码器－解码器网络预测密集采样的像素和片段之间的对应关系；还提出了一种通用的数据合成方案，创建了高相似度的弱纹理物体数据集；最后，设计并改进了一种PNP-RANSAC算法稳健而有效地估计可能多个对象实例的姿态，并在3个弱纹理数据集上进行对比实验并验证了该方法的有效性。

关键词

位姿估计，弱纹理物体，碎片模型，编码－解码，PNP

Pose Estimation of Weak Textured Objects Based on Fragment Model

Yao Li, Lianglun Cheng, Tao Wang

School of Computer, Guangdong University of Technology, Guangzhou Guangdong

Received: Dec. 26th, 2021; accepted: Jan. 21st, 2022; published: Jan. 28th, 2022

Abstract

Pose estimation has been widely used in intelligent robot, unmanned driving, augmented reality and other application scenarios. Most of the existing point pair matching methods based on neural network failed to deal with foreground occlusion, weak texture and symmetry. In this paper, a pose estimation method for weakly textured objects based on the fragment model is proposed. This method uses the existing 3D fragment model to estimate the 6D pose of rigid bodies from an RGB input image. Objects are represented by compact surface fragments, which can systematically

文章引用: 李耀, 程良伦, 王涛. 基于碎片模型的弱纹理物体位姿估计[J]. 计算机科学与应用, 2022, 12(1): 252-261.
DOI: 10.12677/csa.2022.121025

deal with object symmetry. The encoder-decoder network is used to predict the correspondence between pixels and fragments of dense sampling. A general data synthesis scheme is proposed to create a weak textured object dataset with high similarity. Finally, a PNP-RANSAC algorithm is designed and improved for robust and efficient attitude estimation of multiple object instances, and the effectiveness of the method is verified by comparison experiments on three weakly textured datasets.

Keywords

Pose Estimation, Texture-Less Object, Fragment Model, Encoder-Decoder, PNP

1. 引言

刚体 6D 位姿估计在智能机器人，无人驾驶以及增强现实等多种应用场景有重要的研究意义。基于模型的 6D 位姿估计指的是物体在相机坐标系下的 3D 平移以及 3D 旋转[1]，可由齐次变换矩阵 P 表示，齐次变换矩阵 P 又可表示为旋转矩阵 R 和平移向量 T。其中，旋转矩阵表征物体的 3D 姿态，即物体相对于相机坐标系的 x、y、z 轴的旋转角度；平移向量 P 表征物体的 3D 位置，即物体相对于相机坐标系的三维坐标。

随着 1963 年第一个关于位姿估计的方法被 Roberts [2]提出以来，涌现了许多解决位姿估计的方法。该问题的一种常见方法是在输入图像和对象模型之间建立一组 2D-3D 对应关系，并通过 PnP-RANSAC 算法[3] [4]稳健地估计姿态。

2000 年以来一些局部特征描述子[5]的提出，使得具有颜色纹理信息的物体位姿估计难度大大下降，但在实际场景中往往很多物体缺少或没有纹理信息，不同于一般目标物体的位姿识别，工业中的目标物体通常具有表面低纹理的特性，即为弱纹理物体，没有明显的纹理，目标物体表面颜色、明暗变化不明显，与背景区分度不高，强光下往往还会伴随反光，在形状或大小上具有很高相似性，难以从中提取出鲁棒的特征点，并且工业场景往往还伴随着杂乱、堆叠等复杂操作环境，这对目标物体的识别以及位姿估计带来极大的挑战[6]。对于这些弱纹理物体，传统上基于纹理局部特征描述子匹配的方法则无法适用，因此复杂环境下弱纹理物体六自由度位姿估计也成为了近几年的研究热点。

近年来，深度学习在 2D 视觉领域进展火热，研究人员很自然的会将深度学习引入到物体 6D 位姿估计，而且是全方位的，无论是基于纯 RGB 图像、RGB 和 Depth 图像、还是只基于 3D 点云，无论是寻找对应、寻找模板匹配、亦或是进行投票，都展现了极好的性能[7]。但是根据对弱纹理物体六自由度位姿估计现有方法的调研，目前主要存在以下问题：在场景同时存在背景杂乱(Background Clutter)、弱纹理(Texture Less)以及前景遮挡(Foreground Occlusions)的情况下仍然存在挑战，大部分方法无法解决前景遮挡的问题[6]。

针对上诉问题，本文提出了一种基于碎片模型的弱纹理物体位姿估计(Pose Estimation of Weak Textured Objects Based on Fragment Model)方法，该算法基于以下假设，每个物体可以分解为若干碎片模型替代，算法对多碎片模型进行编码－解码操作，可从局部推理全局信息关系，解决物体弱纹理以及前景遮挡，也能很好地处理物体对称。本文的主要工作如下：

1) 提出一种物体表面碎片表示，允许以系统的方式处理物体对称性，并确保任何物体上候选 3D 位置的数量和覆盖范围一致。

2) 提出一种基于单 RGB 的弱纹理物体位姿估计方法框架，适用于众多物体实例，包括具有对称性，表面弱纹理的物体，在合成以及真实数据集上实现了最好的效果。

3) 提出一种多场景弱纹理物体数据集渲染方法，并创建泛化能力较好的弱纹理物体数据集，达到工业检测需求。

2. 数据集获取

目前公开的位姿估计数据集普遍都是针对日常家用场景下的物体设计的，比如 Ycb-Vedio [8]，Linemod [9]，针对工业场景下的数据集非常少，只有 T-less 数据集，但是其局限性很大，训练集复杂性不高，泛化能力也随之减弱，针对此数据集问题，本文提出一种数据生成的方法。

本文数据集采集方法是一个模块化的程序管道，可以生成照片级场景图像，并提供完美的分割掩码深度图像以及 RGB 图像，应用开源项目 Blender 的 python API [10]渲染出逼真的场景效果，虽然在渲染速度上比不上 OpenGL，但是在渲染效果以及泛化性能上能够达到真实数据级别。

2.1. 模型创建

本节描述了一个快速且方便的方法来构造物体数据集。获得相应的三维模型并不困难，因为工件在制造之前有自己的三维模型，现今也有许多可以扫描出物体精确模型的相机工具。与此同时，我们在每个场景空间中随机设置物体的堆叠方式以及虚拟摄像机从不同视角拍摄方式，并且通过随机调整自定义参数(比例、位置和方向)自动生成信息样本，可以很好的获取模型的完整信息，包含更多的信息。场景初始化如图 1 所示。

Figure 1. Scenario initialization
图 1. 场景初始化

2.2. 数据集制作

在完成每个场景的模型创建之后，我们需要通过 blender 管道运行 python API，这个管道由几个模块组成并可以轻松使用脚本进行配置。这些模块可以通过加载新对象或新相机位置采样来更改 blender 中的场景状态。

每次管道运行都包含摄像机位置的定义，这些位置被保存为相机对象的关键点。每个关键点对应特定时间内特定对象的位置和旋转，我们使用它来渲染每个场景的图像，然后每个关键点生成相应的图像。

最后一步是照片级渲染，颜色渲染器和深度渲染器会依次执行，这个过程能够轻松地生成图像，将所需标签和相应的数据存储在同一个文件中。

3. 碎片模型

对于弱纹理对象，往往物体之间形状和纹理十分相似或具有全局或部分对称，为这种物体建立 2D-3D 对应是一个非常大的挑战。这些物体的可见部分，由自遮挡和其他物体遮挡决定，可能对物体模型有多次拟合。因此，相应的二维和三维位置形成多对多的关系，即一个二维图像位置可能对应模型表面的多个三维位置，反之亦然。另外，基于局部图像特征的方法对无纹理目标的性能较差，这是因为特征检测器往往不能提供足够数量的可靠位置，描述符也不再具有足够的判别性。

为了解决这个问题，本文提出了一种从单一 RGB 输入图像中利用可用的 3D 模型估计可能多个刚体实例的 6D 位姿的方法。该方法适用范围很广，可以处理无纹理物体和具有全局或部分对称的物体，比如碗和杯子。

其关键思想是用数量可控的紧凑表面碎片来表示一个对象，如图 2 所示。这种表示允许以系统的方式处理对称性，并确保在任何类型的物体上的候选 3D 位置的一致性。利用编码器－解码器卷积神经网络预测密集采样像素与表面碎片之间的对应关系。在每个像素处，网络预测：1) 每个物体存在的概率，2) 给定物体碎片存在的概率，3) 每个碎片上的精确 3D 位置。通过有条件地建模碎片的概率，将物体对称引起的不确定性与物体存在的不确定性解耦，并用于预测出在每个像素处选择数据依赖的 3D 位置数量。我们将物体 i 的碎片模型定义为：

$$S_{ij}=\left\{x\middle|x\in S_i\wedge d\left(x,g_{ij}\right)<d\left(x,g_{ik}\right),\forall k\in J,k\neq j\right\}$$

其中这里 d 表示两个 3D 点的欧几里德距离，J 表示碎片集合，g 表示预选的碎片中心，碎片中心通过最远点采样算法(FPS)得到。S_{ij} 表示物体 i 中碎片 j 的点集，x 代表入选该碎片的点，k 为该物体其他碎片的点。

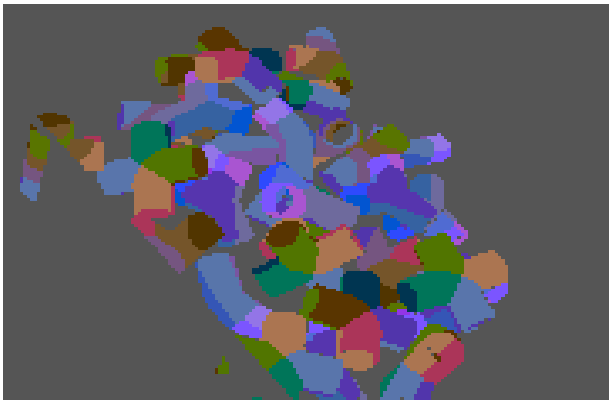

Figure 2. Object fragment model representation
图 2. 物体碎片模型表示

4. 网络结构

本文方法的主要网络架构如图 3 所示，提出了一种新的六自由度目标位姿估计框架。给定一幅 RGB 图像，在此图像中检测目标并估计其方向和平移。6D 位姿由物体坐标系到摄像机坐标系的刚性变换(R; T)表示，其中 R 表示三维旋转，T 表示三维平移。

受到最近方法的启发，我们使用两阶段的编码－解码管道来估计目标的位姿：首先使用 CNN 检测二维目标关键点，然后使用 PnP 算法计算 6D 位姿参数。我们的创新之处在于改进了一种新的编码架构，在轻量化的同时达到精度要求，以及一种改进的 PnP 位姿估计算法建立多对多的 2D-3D 对应关系。具体来说，我们的方法使用以类似 Ransac 的方式检测 2D 关键点，它能够很好地处理弱纹理和对称以及遮挡的对象。基于 PnP-RANSAC 预测每个像素上的精确 3D 位置的数据依赖数量，然后预测多对多 2D-3D 对应关系，估计出可能的多个目标实例的位姿。

我们网络的输入是一张包含多个类别实例的 RGB 图像，从一幅 RGB 输入图像中利用可用的 3D 模型估计可能多个刚体对象的可能多个实例。在训练期间，以对象标签、片段标签和 3D 片段坐标的形式向编码器－解码器网络提供逐像素的标注。在推断过程中，在每个像素预测可能多个片段上的 3D 位置，这很大程度地提升了对物体对称性处理的能力。然后通过将像素与预测的 3D 位置相关联来建立许多 2D-3D 对应关系，并且使用 PNP-RANSAC 算法的稳健而高效的变体来估计 6D 姿态。更加细节的介绍见 51，5.2 小节。

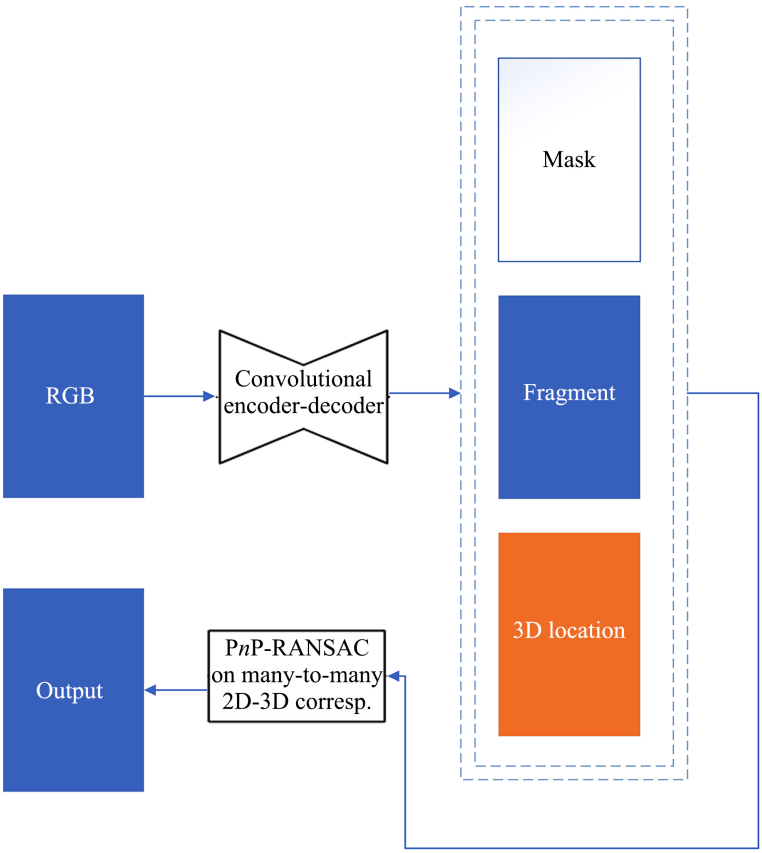

Figure 3. Network architecture
图 3. 网络架构图

4.1. 编码－解码网络

我们的网络采用较稳定的编码－解码结构，为了更好地提取出弱纹理物体的特征并使得模型更加地轻量化，我们创新性地提出了一种新的编码结构，采用 swin transform [11]，先经过一层卷积进行 patch 映射，具体是将图像先分割成 4 × 4 的小块，然后将每一个小块通过映射成一个像素点，进行了通道上的扩充。为了减少计算量，swin 的做法是将输入图片划分成不重合的 windows，然后在不同的 window 内

进行 self-attention 计算。这种新的结构可以在很大提升识别准确率的同时，对于计算速率的提升以及模型压缩效果比较明显。解码结构是将得到的关键点热力图进行上采样，通过反向回归预测每个像素的掩码，三维碎片坐标类别和三维坐标点，其中保留背景类。对于每个对象，每个对象由 n 个表面片段表示，网络具有 4 mn + mM + 1 个输出通道(对象的分类和背景，用于表示表面片段概率以及 3D 片段坐标)。

4.2. 物体位姿对应

传统位姿估计方法大多处理单实例的物体，如 LineMod，利用 RANSAC 算法从一组包含“局外点”的观测数据集中，通过迭代方式估计数学模型的参数，进而利用 PNP 算法进行 3D-2D 对应关系求解物体位姿。但对于本文的对象来说，要同时预测出每张图片中同一物体的多个实例，尤其是每个物体还重建成了多个碎片，要预测这种碎片模型的 3D-2D 对应关系，传统方法就不太适用。

本文提出一种几何多模型拟合位姿估计方法，通过重复假设提议、快速拒绝和通过标记最小化将新假设整合到一个实例集，从而对当前数据解释进行采样和维护。通过逐步迭代数据，采用 RANSAC 的概率保证参数，并采用了一种基于集合重叠的通用度量进行有效估计，为了实现多个模型实例的估计，修改了 RANSAC 的模型质量函数。

多实例拟合主要由 PnP-RANSAC 变体算法实现，首先按顺序提出姿态假设，通过利用对应关系的空间连贯性优化添加到一组维护的假设中，然后通过描述由 2D 和 3D 坐标组成的 5D 向量来构建邻域图。如果它们的欧几里德距离低于 inlier-outlier 阈值，则链接两个 5D 描述符，inlier-outlier 阈值在重投影误差上手动设置并定义。具体步骤如下：

4.2.1. 位姿提议生成

提议策略的主要目标是提议出看不见的实例，即那些可能不在集合中的实例。实现该目标的直接方法是优先选则具有合理数量的不与复合实例共享点的实例。提出了一个新的质量函数，用于度量一个模型实例的得分：

$$Q=\sum_{p\in P}\left[\phi\left(h,P\right)<\epsilon\wedge\phi\left(h_{\cup},P\right)\geq\epsilon\right]$$

Q 表示质量函数，P 表示空间的一个点，这个点属于提议点集 p，这里 ϕ 表示点 P 到物体实例 h 的距离，$h_{\cup}$ 代表复合实例。

4.2.2. 提议验证

验证用于确定是否应该将实例优化。要做到这一点，必须定义一个实例到实例的距离，以衡量提议实例和复合实例的相似性。如果距离很小，则建议很可能是已经可以确定的实例，因此没有必要进行优化。一般来说，表现形式对结果有很大影响，有一个用点集表示实例的简单解决方案，模型通过偏好点集来描述，两个实例的相似度通过它们的 Jaccard 得分来定义。实例的偏好集为 $P_h\in\{0,1\}\left|\mathrm{P}\right|$，如果其中第 j 个点是实例的 inlier，则其 P_h 值为 1，否则为 0。提议的接受标准是：

$$J\left(h,h_{\cup}\right)=\frac{\left|P_h\cap P_{h\cup}\right|}{P_h\cap P_{h\cup}}>\varepsilon_S$$

其中，h 表示目标物体实例，$h_{\cup}$ 代表复合实例，P 代表每个实例的集合，J 表示 Jaccard 得分，如果两个实例的 Jaccard 相似性高于手动设置的阈值 $\varepsilon_S\in[0,1]$，则 J 成立，否则 J 为 FALSE。

4.2.3. 多实例优化

多实例可能是一个合理的选择。本文使用简化的 PEARE 算法，因为它的主要贡献是在标签空间中移动，要用相应的密度模式替换标签集。如果只有几个标签，则不需要执行此操作。因此，使用 PEARE

算法作为优化程序：

$$E(L)=\sum_{p}\left\|p-L_{p}\right\|+\lambda\cdot\sum_{(p,q)\in N}w_{pq}\cdot\delta\left(L_{p}\neq L_{q}\right)$$

其中 L 表示模型，p，q 分别表示一个点集，L_p 表示标签点集，w 表示权重，δ 代表一个判断，如果括号内的指定条件成立为 1，否则为 0。PEARL 方法通过上诉优化来估计模型及其空间支持度。

5. 实验及结果分析

本节介绍了本文方法的实验环境，损失函数及评估标准，并比较了本文方法与其他基于模型的 6D 目标姿态估计方法的性能优势。

5.1. 实验条件

本文中使用 Intel i5 处理器，6 核 12 线程，GPU 采用了 Nvidia RTX 2080Ti，以及 16 GB 内存空间。为了方便 GPU 加速，配置了 CUDA 平台，包含多个加速库，包含用于深度学习、计算机视觉、计算机图形和多媒体处理的加速库，可以加快开发者、企业对于程序上的应用开发，还包含有 CUDA、CUDNN、完整版桌面 Linux 操作系统。

5.2. 评估准则

我们遵循 2019 年 BOP Challenge [12] (简称 BOP19)的评估协议。该任务是估计单一图像中不同数量物体的不同实例的 6D 姿态，每个图像提供的实例数量。

用三个位姿误差函数计算了估计位姿 $\hat{P}$ 和地面真实位姿 $\bar{P}$ 的误差。第一个，可见表面差异，只考虑可见物体部分，将不可区分的姿势视为等价的：

$$e_{VSD}=\underset{p\in\hat{V}\cap\bar{V}}{avg}\begin{cases}0, & \text{if } p\in\hat{V}\cap\bar{V}\wedge\left|\hat{D}(p)-\bar{D}(p)\right|<\tau\\ 1, & \text{otherwise}\end{cases}$$

其中，$\hat{D}$ 和 $\bar{D}$ 是分别通过估计姿态和地面真实姿态渲染对象模型而获得的距离图。将距离图与测试图像 I 的距离图 D_I 进行比较，以获得可见性掩模 $\hat{V}$ 和 $\bar{V}$，即对象模型在图像 I 中可见的像素组。距离图 D_I 可用于 BOP 中的所有图像。参数 τ 是未对准公差。

第二个位姿误差函数(Maximum Symmetry Aware Surface Distance)测量 3D 中的曲面偏差，因此与机器人应用相关：

$$e_{MSSD}=\min_{T\in T_I}\max_{x\in V_I}\left\|\hat{P}_X-\bar{P}T_X\right\|_2$$

其中 T_I 是对象 I 的一组对称变换(在 BOP19 中提供)，V_I 是一组模型顶点。

第三个姿态误差函数，最大投影距离，测量可感知的偏差。它与增强现实应用相关，并且适用于 RGB 方法的评估，对于 RGB 方法，估计 Z 平移分量更具挑战性：

$$e_{MSPD}=\min_{T\in T_I}\max_{x\in V_I}\left\|proj\left(\hat{P}_X\right)-proj\left(\bar{P}T_X\right)\right\|_2$$

其中 $proj$ 表示 2D 投影操作，其他符号的含义与 e_{MSPD} 相同。

当位姿误差函数 $e<\theta_e$ 时，估计的姿态被认为是正确的，其中 $e\in\{e_{VSD},e_{MSSD},e_{MSPD}\}$ 和 θ_e 是正确阈值。为其估计正确姿态的带注释的对象实例的分数称为召回率。平均召回率函数 $e(AR_e)$ 被定义为针对阈值 θ_e 的多个设置以及在 e_{VSD} 的情况下针对未对准容差 τ 的多个设置计算的召回率的平均值。一种方法的整体

性能是通过平均召回率来衡量的： $AR=\left(AR_{VSD}+AR_{MSSD}+AR_{MSPD}\right)/3$ 。由于本文方法只使用 RGB，所以除了 AR，我们还统计 AR_{MSPD} 分数。

5.3. 实验结果

为了证明我们方法的泛化性，我们在 T-less 数据集、LineMod 数据集以及自建数据集中进行了实验，该网络针对几种类型的合成图像进行训练。对于 T-less，我们使用基于物理的渲染(PBR)图像。对于 LM-O，我们使用了场景 1 和场景 2 的 PBR 图像，以及用 OpenGL 渲染的对象随机照片图像，没有用真实图像进行训练。我们在 T-less 以及 LineMod 两个弱纹理数据集表现较好的三种方法 CDPN [13]、CosyPose [14]、Epos [15]进行了比较，主要实验结果如下：

Table 1. AR (average recall rate) results
表 1. AR (平均召回率)结果

Approaches	LM-O	T-less	Pipes
	AR	AR	AR
CDPN [13]	0.624	0.407	0.413
CosyPose [14]	**0.633**	0.640	0.530
Epos-resnet [15]	0.332	0.230	0.214
Epos-xc65	0.547	0.467	0.441
Ours	0.612	**0.651**	**0.579**

Table 2. AR_{MSPD} results
表 2. AR_{MSPD} 结果

Approaches	LM-O	T-less	Pipes
	AR_{MSPD}	AR_{MSPD}	AR_{MSPD}
CDPN	0.815	0.579	0.591
CosyPose	0.812	**0.761**	0.769
Epos-resnet	0.514	0.453	0.468
Epos-xc65	0.750	0.619	0.637
Ours	**0.832**	0.753	**0.782**

1) 精确度

表 1 和表 2 将我们方法的表现与目前表现较好的方法进行了比较。在 AR 得分上，我们的方法在 T-less 和自建数据集上的表现都好于所有 RGB 方法。在 T-less 和自建数据集上，我们的方法得到了最好的 AR_{MSPD} 得分。

下表 3 是推断速率结果：

Table 3. Inferred time results
表 3. 推断时间结果

Approaches	LM-O	T-less	Pipes
	Time	Time	Time
CDPN	**0.489**	1.849	2.031
CosyPose	1.550	1.693	1.769
Epos-resnet	0.771	**0.937**	**0.768**
Epos-xc65	1.268	1.992	2.237
Ours	0.546	1.253	0.930

2) 速度

在未经优化的实施中，我们的方法平均每张图像耗时约 0.9 秒(采用 6 核 Intel i5-10500 CPU、16 GB RAM 和 NVIDIA 2080TiGPU)。与其他基于卷积神经网络的 RGB 方法一样，EPOS 的速度明显快于 RGB-D 和 D 方法，因为 ICP 后处理步骤而较慢。在有着弱纹理特征与极度相似性的数据集 T-less 与自建数据集中，Epos-resnet 方法比我们的方法快，但精确度明显低于我们方法(见图 4)。

根据应用需求，我们的方法精度和速度之间的平衡达到了很高的标准，并且可以通过例如表面碎片的数量、主干网络大小、图像分辨率、预测对应的像素密度来控制。

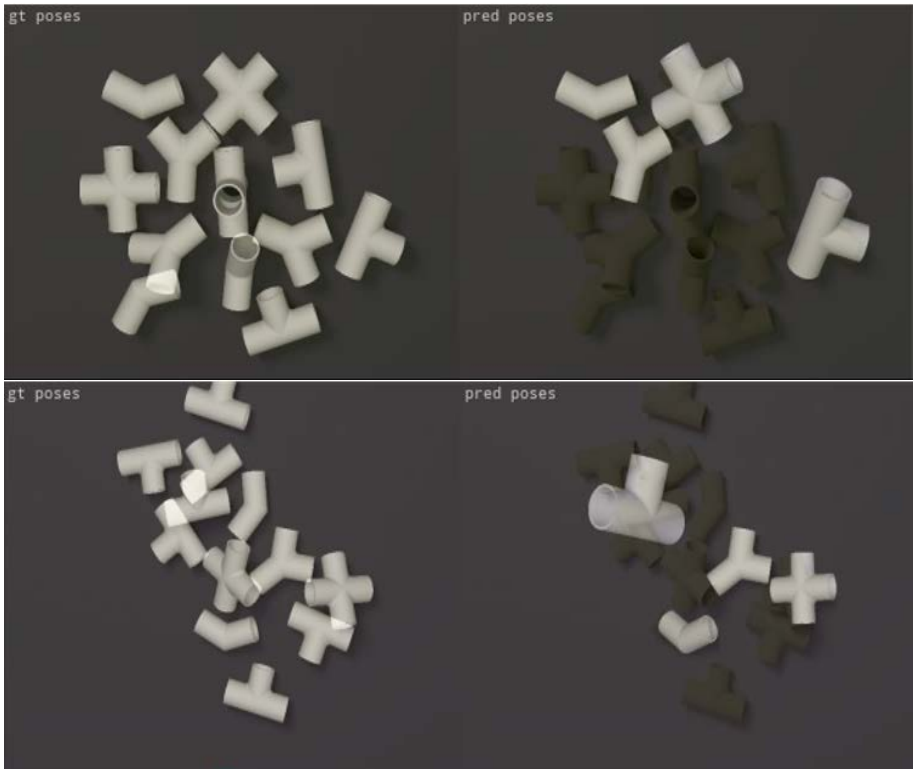

Figure 4. Effect comparison diagram (top: our method; bottom: Epos-resnet method)
图 4. 效果对比图(上：本文方法；下：Epos-resnet 方法)

6. 结论

本提出了一种新的基于模型的单 RGB 图像 6D 目标姿态估计方法。其核心思想是通过紧凑的表面碎片来表示对象，在每个像素上预测可能的多个对应的 3D 位置，并使用 PNP-RANSAC 算法的健壮而高效的变体来求解姿态。实验评估表明，该方法适用于弱纹理，表面相似性极高的物体，包括具有对称性的挑战性目标，很好地平衡了姿态估计精度与实时性。本文还提出了一种通用的数据集创建方案，能够达到数据集的高泛化性与扩展性。另外，本文对特定于物体的碎片数量的研究还未很好地深入，这取决于物体的大小、形状或物体到相机的距离范围等因素，这将留到未来的工作中去做。

参考文献

[1] 涂文哲. 基于合成样本的弱纹理物体 6D 位姿估计[D]: [硕士学位论文]. 成都: 电子科技大学, 2020.

[2] Roberts, L.G. (1963) Machine Perception of Three-Dimensional Solids. Massachusetts Institute of Technology, Cambridge.

[3] Fischler, M.A. and Bolles, R.C. (1981) Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, **24**, 381-395.

[4] Lepetit, V., Moreno-Noguer, F. and Fua, P. (2009) Ep*n*p: An Accurate O(*n*) Solution to the P*n*P Problem. *International Journal of Computer Vision*, **81**, Article No. 155. https://doi.org/10.1007/s11263-008-0152-6

[5] Lowe, D.G. (1999) Object Recognition from Local Scale-Invariant Features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, Kerkyra, 20-27 September 1999, 1150-1157. https://doi.org/10.1109/ICCV.1999.790410

[6] 张昊若. 面向机器人抓取的弱纹理物体六自由度位姿估计方法研究[D]: [博士学位论文]. 上海: 上海交通大学, 2019.

[7] Du, G.G., Wang, K., Lian, S.G., *et al*. (2020) Vision-Based Robotic Grasping from Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review. *Artificial Intelligence Review*, **54**, 1677-1734. https://doi.org/10.1007/s10462-020-09888-5

[8] Xiang, Y., Schmidt, T., Narayanan, V., *et al*. (2017) PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. https://doi.org/10.15607/RSS.2018.XIV.019

[9] Hinterstoisser, S., Holzer, S., Cagniart, C., *et al*. (2011) Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 858-865. https://doi.org/10.1109/ICCV.2011.6126326

[10] Denninger, M., Sundermeyer, M., Winkelbauer, D., *et al*. (2019) BlenderProc. https://arxiv.org/abs/1911.01911

[11] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. https://arxiv.org/abs/2103.14030

[12] Hodan, T., Michel, F., Brachmann, E., *et al*. (2018) BOP: Benchmark for 6D Object Pose Estimation. *Computer Vision—ECCV* 2018, Munich, 8-14 September 2018, 19-35. https://doi.org/10.1007/978-3-030-01249-6_2

[13] Li, Z.G., Wang, G. and Ji, X.Y. (2019) CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-Dof Object Pose Estimation. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 7678-7687. https://doi.org/10.1109/ICCV.2019.00777

[14] Labbé, Y., Carpentier, J., Aubry, M., *et al*. (2020) Cosypose: Consistent Multi-View Multi-Object 6D Pose Estimation. *European Conference on Computer Vision*, Glasgow, 23-28 August 2020, 574-591. https://doi.org/10.1007/978-3-030-58520-4_34

[15] Hodan, T., Barath, D. and Matas, J. (2020) EPOS: Estimating 6D Pose of Objects with Symmetries. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11703-11712. https://doi.org/10.1109/CVPR42600.2020.01172