Computer Science and Application 计算机科学与应用, 2022, 12(1), 221-232
Published Online January 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.121023

Hans汉斯

通过像素级XYZ坐标映射的实时6DoF姿态估计

吴　勇，程良伦

广东工业大学计算机学院，广东　广州

收稿日期：2021年12月25日；录用日期：2022年1月21日；发布日期：2022年1月28日

摘　要

为了解决在严重遮挡和存在无纹理物体情况下，从单一RGB图像中进行6DoF姿态估计的挑战，本文提出了一种通过像素级XYZ坐标映射的实时6DoF姿态估计方法。我们引入了联合的坐标－置信度损失函数来直接回归三维模型的空间坐标，以有效地处理无纹理物体和遮挡的杂乱场景。同时，我们还考虑了2D目标检测误差导致的问题，引入了一种动态缩放策略来提高算法的性能。实验表明，我们的方法在Occlusion LINEMOD和T-LESS数据集下的评估指标优于现有的基线方法。

关键词

6DoF姿态估计，遮挡，无纹理，像素级

Real-Time 6DoF Pose Estimation via Pixel-Level XYZ Coordinates Mapping

Yong Wu, Lianglun Cheng

School of Computer Science and Technology, Guangdong University of Technology, Guangzhou Guangdong

Received: Dec. 25th, 2021; accepted: Jan. 21st, 2022; published: Jan. 28th, 2022

Abstract

To address the challenge of 6DoF pose estimation from a single RGB image in the presence of severe occlusion and texture-less objects, this paper proposes a real-time 6DoF pose estimation approach via pixel-level XYZ coordinates mapping. We introduce a joint coordinates-confidence loss function to directly regress the spatial coordinates of the 3D model to effectively handle texture-less objects and occluded in cluttered scenes. Meanwhile, we consider the problems caused by 2D object detection errors and introduce a dynamic scaling strategy to improve the perfor-

文章引用: 吴勇, 程良伦. 通过像素级 XYZ 坐标映射的实时 6DoF 姿态估计[J]. 计算机科学与应用, 2022, 12(1): 221-232.
DOI: 10.12677/csa.2022.121023

mance of the algorithm. Experiments show that our method outperforms the existing baseline methods in terms of evaluation metrics under Occlusion LINEMOD and T-LESS datasets.

Keywords

6DoF Pose Estimation, Occlusion, Texture-Less, Pixel-Level

1. 引言

六自由度(6DoF)姿态估计，是估计相机坐标系下物体的 6DoF 姿态，即三维位置和三维旋转。6DoF 姿态估计的意义在于能够获得物体的准确姿态，支撑对物体的精细化操作，主要应用于机器人抓取领域和增强现实领域[1]。在机器人抓取领域中，主流的方法是估计已知物体的 6DoF 姿态，进而获得末端执行器的目标 6DoF 抓取姿态。在增强现实领域中，可以在物体上叠加虚拟元素，随着相机的移动，物体的相对姿态保持不变。随着 SLAM 等技术的成熟，机器人已经能够在空间中进行很好的定位，但如果想要和环境中的物体进行交互操作，物体的 6DoF 姿态估计是必不可少的技术。然而，对于复杂环境下的场景下，往往会存在无纹理物体或遮挡的情况，算法的鲁棒性不足，极大地限制了机器人的灵活性。因此，本文提出了一种通过像素级 XYZ 坐标映射的实时 6DoF 姿态估计方法，有效地处理无纹理物体和遮挡的杂乱场景，在不影响精度的基础下，还保证了算法的鲁棒性和实时性。

2. 相关工作

与传统的目标检测任务不同，6DoF 姿态估计需要估计出物体的三维位置和三维旋转。现有基于特征点[2]或模板[3]的方法，通过匹配点对特征或检索类似的模板图像来解决物体 6DoF 姿态的问题。诸如此类的传统方法实现了在固定场景下物体的 6DoF 姿态估计，广泛应用于工业中，但在处理杂乱的场景时，它们暴露了精度不足、鲁棒性差等问题。

随着深度学习和大数据的兴起，学术界出现了一些新的基于深度学习的方法。Xiang 等人[4]首次提出了通过卷积神经网络估计 6DoF 姿态的 PoseCNN。受 YOLO 网络的启发，Tekin 等人[5]直接预测物体的三维边界框的投影顶点，然后通过 Perspective-n-Point (PnP)算法进行 6DoF 姿态估计。然而，对于杂乱场景中的无纹理和被遮挡的物体，这些方法的效果并不好。Peng 等人[6]引入了有向单元向量来解决遮挡物体的 6DoF 姿态问题，提出了 PVNet 网络架构，但在 T-LESS 数据集上仍然显示出精度不足的缺点。为了解决无纹理物体的姿态估计问题，出现了基于坐标和自动编码器的方法。Zakharov 等人[7]通过 PnP 和 RANSAC 计算，估计了输入图像和现有三维模型之间密集的多类 2D-3D 对应图。Sundermeyer 等人[8]提出了由潜在空间中的样本定义物体姿态的隐式表示，他们在合成数据上训练了该模型，并将其泛化到现实图像上。然而，这些方法在遇到姿态歧义时也暴露出鲁棒性不足的问题。

另一个基本问题是，6DoF 姿态的有效性依赖于其预置的 2D 目标检测器的准确性。许多方法[4] [5] [6]将 2D 目标检测器和 6DoF 姿态估计模块作为一个整体来训练网络，但由于任务不同，姿态估计网络通常不能很好地收敛。现有的方法[7] [8]将其解耦成两个独立的模块，训练时互不干扰，取得了很好的性能。然而，由于 2D 目标检测器存在检测误差，6DoF 姿态网络对目标检测结果依旧非常敏感。

为了解决上述问题，我们提出了一个全新的网络框架，用于物体，6DoF 姿态估计，在单一 RGB 图像中进行像素级 XYZ 坐标映射。本文的主要贡献包括以下三个方面：

- 我们提出了一个简洁的网络结构和联合的坐标－置信度损失(Coordinates-Confidence Loss, CC Loss)函数，基于高质量逼真的合成数据进行训练，以有效地处理杂乱场景中的无纹理物体和被遮挡的情况。
- 考虑到 2D 目标检测器的误差，我们引入了一种动态缩放(Dynamic Scaling, DS)策略，对检测结果进行参数化的调整。
- 与现有的方法基线相比，我们的方法在 Occlusion LINEMOD 和 T-LESS 数据集的准确性和实时性方面表现得更好。

本文的其余部分组织如下。第 3 节描述了所提议的方法。第 4 节展示了实验结果。第 5 节给出了本文的结论。

3. 所提议的方法

3.1. 数据获取

深度学习是一种数据驱动的技术，因此大量具有准确注释的数据是非常重要的。6DoF 数据集与传统视觉任务不同，6DoF 姿态直观表示为 3D 矩形框的关键点，这在真实世界中不容易被注释。此外，采集数据时，物体的姿态往往不能覆盖整个球体，这会导致数据不全、部分姿态难以估计的问题。随着 Blender 和虚幻 4 引擎的广泛应用，许多方法开始使用合成图像，这些图像通常是通过在随机背景上渲染三维物体模型获得的。因此，我们也在 BlenderProc [9] (三维轻量级渲染工具)中构建了我们的数据集，以获得合成训练样本(见图 1)。我们提供了 5 万张带有 6DoF 姿态注释的高逼真度合成图像，具体统计信息见表 1。

Figure 1. Sample synthetic images for 6DoF pose estimation
图 1. 用于 6DoF 姿势估计的合成图像样本

Table 1. Statistical information on the Occlusion LINEMOD (LM-O) dataset and the T-LESS dataset
表 1. 关于 Occlusion LINEMOD (LM-O)数据集和 T-LESS 数据集的统计信息

数据集	物体种类	训练样本数		测试样本数	样本实例数
		真实	合成		
LM-O	8	-	50,000	1214	9038
T-LESS	30	37,584	50,000	10,080	67,308

3.2. 网络架构

在本文中，我们提出了一种用于物体 6DoF 姿势估计的全新 CNN 网络架构，如图 2 所示。具体来说，该网络由 2D 目标检测模块和 6DoF 姿态估计模块组成。2D 目标检测模块用于定位物体的 2D 位置，而 6DoF 姿态估计模块逐像素学习 XYZ 的坐标映射，最后通过 RANSAC/PnP 算法得出 6DoF 姿态。它主要分为三个阶段：1) 给定一张 RGB 图像，我们采用 FCOS 与 VoVNet 高效骨干网络提取特征图，然后将检测到的 2D 边界框动态缩放为 256 × 256 大小的统一裁切图像，2)裁切后的图像流入 ResNet 网络，通过预训练模型获得高级特征图。物体的 XYZ 坐标图和置信度图由编码器－解码器结构预测，该结构是一个完全卷积网络。

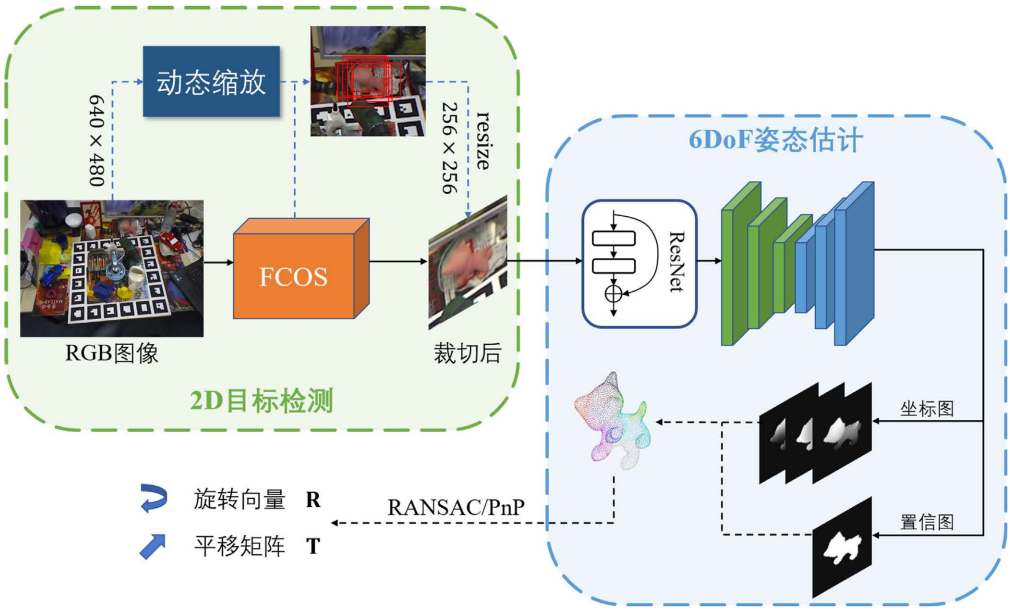

Figure 2. Overview of 6DoF pose estimation network architecture
图 2. 6DoF 姿态估计网络架构的概述

3.3. 坐标映射

为了获得物体的 XYZ 坐标图，我们的方法执行了两项任务：2D 目标检测和 6DoF 姿态估计。对于 2D 目标检测，给定一张输入 RGB 图像，我们通过所提议网络进行前向传播，获得分类分数 $p_{x,y}$ 和对每个位置的特征图 F_i 进行回归预测 $t_{x,y}$。我们定义训练损失函数如公式(1)：

$$\ell_{\text{det}} = \frac{1}{N_{pos}} \sum_{x,y} \ell_{\text{cls}} \left(p_{x,y}, c^*_{x,y} \right) + \frac{\lambda}{N_{pos}} \sum_{x,y} \mathbb{I}_{\left\{ c^*_{x,y} > 0 \right\}} \ell_{\text{reg}} \left(t_{x,y}, t^*_{x,y} \right) \tag{1}$$

其中，ℓ_{cls} 是 Focal 损失函数，ℓ_{reg} 是 UnitBox 中的 IoU 损失函数，N_{pos} 为正样本的数量，λ 是 ℓ_{reg} 的权重参数，$\mathbb{I}$ 为指示函数。

对于 6DoF 姿态估计，我们直接预测每个物体像素的空间坐标。此外，该网络还给出了置信度预测，以表明该像素是否属于该物体。具体来说，我们以 ResNet 为骨干网络，从目标区域中提取特征。然后，引入编码器－解码器结构，对特征进行处理，并将其统一比例为坐标－置信图，其中包含三通道的坐标图 M_{coor} 和单通道的置信图 M_{conf}。它们共享网络权重，M_{coor} 中的每个像素分别代表物体 3D 模型的 XYZ 空间坐标(如图 3 所示)。

Figure 3. XYZ spatial coordinates of the 3D model
图 3. 3D 模型的 XYZ 空间坐标

当我们从 3D 模型中估计空间坐标时，由于背景中的未知空间坐标，这会导致物体边缘处的坐标图出现明显的误差。为了解决这个问题，我们提出了一个联合的坐标－置信度损失(Coordinates-Confidence Loss, CC Loss)函数，具体如公式(2)：

$$\ell_{\mathrm{CC}} = \alpha \cdot \left\| M_{\mathrm{coor}} - \widehat{M}_{\mathrm{coor}} \right\|_1 + \beta \cdot \left\| \sum_{i=1}^{n_c} \left(M_{\mathrm{conf}} \otimes \left(M_{\mathrm{coor}_i} - \widehat{M}_{\mathrm{coor}_i} \right) \right) \right\|_1 \tag{2}$$

其中，$n_c = 3$ 是坐标图的通道数，M_* 和 $\widehat{M}_*$ 分别代表真值坐标图和估计坐标图，α 、β 是权重系数，$\otimes$ 是矩阵外积。

更具体地说，我们只关注一个物体的坐标图，而对于置信图，我们计算的是裁切图像的损失而不是整个区域。这种设计避免了来自非感兴趣区域(如背景、遮挡)的干扰，使网络能够准确预测空间坐标。

3.4. 动态缩放

2D 目标检测的性能会影响 6DoF 姿态估计的结果，因此使用定制的 2D 目标检测器模型是常见的做法，但不能保证每个检测器在不同的场景下都具有良好的表现。因此，我们引入动态缩放策略，以提高 6DoF 姿态估计的鲁棒性。

2D 目标检测的数学表示为物体的 2D 边界框 (x, y, w, h)，其中 x 、y 是边界框的中心点，w 、h 为尺寸大小。我们引入截断的正态分布进行随机抖动采样，如下公式(3)、(4)、(5)：

$$\tilde{x} \sim f_x = \frac{1}{\sigma_x} \cdot \frac{\phi\left(\frac{\tilde{x} - x}{\sigma_x} \right)}{\Phi\left(\frac{\alpha w}{\sigma_x} \right) - \Phi\left(\frac{-\alpha w}{\sigma_x} \right)} \tag{3}$$

$$\tilde{y} \sim f_y = \frac{1}{\sigma_y} \cdot \frac{\phi\left(\frac{\tilde{y}-y}{\sigma_y}\right)}{\Phi\left(\frac{\beta h}{\sigma_y}\right) - \Phi\left(\frac{-\beta h}{\sigma_y}\right)} \tag{4}$$

$$\tilde{s} \sim f_s = \frac{1}{\sigma_{\min(w,h)}} \cdot \frac{\phi\left(\frac{\tilde{s}-s}{\sigma_s}\right)}{\Phi\left(\frac{\gamma \min(w,h)}{\sigma_s}\right) - \Phi\left(\frac{-\min(w,h)}{\sigma_s}\right)} \tag{5}$$

其中，$\tilde{x} \in [-\alpha w, \alpha w]$，$\tilde{y} \in [-\beta h, \beta h]$，$\tilde{s} \in [-\gamma \min(w,h), \gamma \min(w,h)]$，$\phi(\bullet)$为其累计分布函数，$\alpha$、$\beta$、$\gamma$是超参数。

由于图像缩的原因，RGB 图像上的像素点位置与坐标图不同。为了构建 3D-2D 对应关系，需要将预测的坐标映射到RGB图像上(见图4)。对于每个像素点(i, j)，我们可以通过公式(6)得到3D-2D点(p_x, p_y)：

$$\begin{cases} p_x = c_x + \frac{w}{\tilde{w}} \cdot (i - \text{coor}_x) \\ p_y = c_y + \frac{h}{\tilde{h}} \cdot (j - \text{coor}_y) \end{cases} \tag{6}$$

其中，c_x、c_y、w、h分别为 RGB 图像中物体的中心点和尺寸大小，coor_x、coor_y、$\tilde{w}$、$\tilde{h}$分别为坐标图中物体的中心点和尺寸大小。

为了减少点对关系的异常值，我们引入了 RANSAC，使得估计的 6DoF 姿态更加稳健。

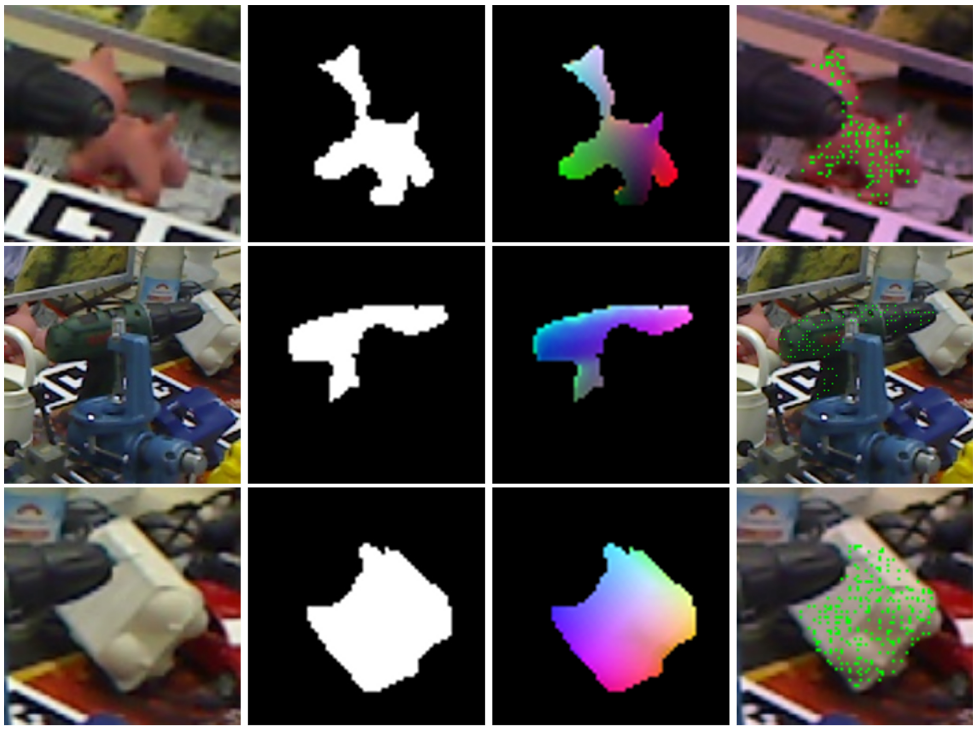

Figure 4. 3D-2D correspondence points
图 4. 3D-2D 对应点

4. 实验结果

4.1. 数据集

Occlusion LINEMOD [10]数据集是 LINEMOD 数据集的一个子集，它主要被用来比较在不同程度的遮挡情况下 6DoF 物体姿态估计方法的性能。

T-LESS [11]数据集主要被广泛用于评估无纹理物体的 6DoF 姿态估计性能。它包含了大量的行业相关的物体和测试图像，这些图像具有较大的视角变化，物体处于多个实例中，受到杂波和遮挡的影响。

4.2. 评估指标

物体的 6DoF 姿态数学化抽象表示为 4 × 4 矩阵：

$$P=\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \tag{7}$$

其中，R 为 3 × 3 的旋转矩阵，T 为 3 × 1 的平移向量。

对于 Occlusion LINEMOD 数据集，一般的评估指标为：非对称物体的 3D 模型点平均距离(ADD) [3]和对称物体的平均最邻近点距离(ADD-S) [4]。给定真值的旋转矩阵 R 和平移向量 T，以及估计值的旋转矩阵 R 和平移向量 T，评估指标如公式(8)、(9)：

$$e_{\text{ADD}}=\operatorname*{avg}_{x\in M}\left\|\left(Rx+T\right)-\left(\hat{R}x+\hat{T}\right)\right\| \tag{8}$$

$$e_{\text{ADD-S}}=\operatorname*{avg}_{x_1\in M}\min_{x_2\in M}\left\|\left(Rx_1+T\right)-\left(\hat{R}x_2+\hat{T}\right)\right\| \tag{9}$$

其中，M 表示为 3D 模型点的集合。如果估计姿态和真值姿态之间的距离 $<10\%\cdot d$ (d 为物体 3D 模型的直径)，则可以认为估计的姿态是正确的。

为了与 PVNet 进行比较，我们还采用了 2D 投影度量指标，即使用真值姿态和估计姿态将物体 3D 模型投影至图像中。如果整个模型顶点的平均投影误差<5 px，则可以认为估计的姿态是正确的。

我们为 T-LESS 数据集引入了 BOP Challenge [12]的评估指标。第一个姿态误差函数为可见表面误差(VSD)，它只考虑物体的可见部分，将无法区分的姿态视为等同姿态。第二个姿态误差函数为最大对称性感知表面距离(MSSD)，由于它评估了物体模型中的表面偏差，因此该函数与机器人操纵强相关。第三个姿态误差函数为最大对称性投影距离(MSPD)，它考虑了全局物体的对称性，并以最大距离取代平均值，以提高对网格顶点采样的鲁棒性。由于 MSPD 不评估沿相机光轴(Z 轴)的对齐情况，只测量可感知的差异，因此它与增强现实应用有关，适合用于评估基于 RGB 图像的方法。

结合上述三个姿态误差函数，T-LESS 数据集的评价指标是由平均召回率(AR)来衡量的：

$$\text{AR}=\frac{\text{AR}_{\text{VSD}}+\text{AR}_{\text{MSSD}}+\text{AR}_{\text{MSPD}}}{3} \tag{10}$$

4.3. 实现细节

为了提高方法的鲁棒性，我们引入了不同的图像对比度、亮度、高斯模糊和颜色失真。此外，我们使用带有黑色方块的随机物体掩码来模拟遮挡情况。在训练过程中，初始学习率为1×10^{-4}，批量大小为 4。我们采用 RMSProp 函数($\alpha=0.99$，$\sigma=1\times10^{-8}$)进行优化。该模型总共训练了 200 个迭代，每 50 个迭代的学习率将被除以 10。坐标标签通过前向投影与 Z-Buffer 计算的。

4.4. 定性结果

我们与基于RGB图像的6DoF姿态估计方法进行了比较，如表2、表3所示。对于Occlusion LINEMOD数据集，我们的方法在所有基线方法中取得了最好的性能。使用动态缩放(DS)策略的结果比使用真值边界框(BBs)的结果更为接近。

Table 2. Comparison of 2D projection metrics between our method and the baseline method on the Occlusion LINEMOD dataset (Objects annotated with * possess symmetric pose ambiguity)
表2. 我们的方法和基线方法在Occlusion LINEMOD数据集上的2D投影指标比较(带*号的物体具有对称的姿态歧义性)

方法 类别	Tekin [5]	PoseCNN [4]	Oberweger [13]	PVNet [6]	Ours		
					w/o DS	w/DS	w/BBs
ape	7.01	34.6	69.6	69.14	71.45	73.29	**75.43**
can	11.20	15.1	82.6	86.09	84.88	85.96	**88.89**
cat	3.62	10.4	65.1	65.12	69.36	71.25	**72.42**
duck	5.07	31.8	61.4	61.44	64.14	65.12	**67.50**
driller	1.40	7.4	73.8	73.06	75.44	77.21	**79.67**
eggbos*	-	1.9	**13.1**	8.43	9.33	9.96	10.89
glue*	4.70	13.8	54.9	**55.37**	46.29	48.67	51.29
holepuncher	8.26	23.1	66.4	69.84	80.01	82.11	**84.50**
mean	6.16	17.2	60.9	61.06	62.61	64.20	**66.32**

Table 3. Comparison of ADD(-S) metrics between our method and the baseline method on the Occlusion LINEMOD dataset (Ob-jects annotated with * possess symmetric pose ambiguity)
表3. 我们的方法和基线方法在Occlusion LINEMOD数据集上的ADD(-S)指标比较(带*号的物体具有对称的姿态歧义性)

方法 类别	Tekin [5]	PoseCNN [4]	Oberweger [13]	PVNet [6]	Ours		
					w/o DS	w/DS	w/BBs
ape	2.48	9.6	17.6	15.81	20.57	23.87	**25.71**
can	17.48	45.2	53.9	63.30	62.81	62.64	**64.32**
cat	0.67	0.93	3.31	16.68	23.98	24.45	**25.15**
duck	1.14	19.6	19.2	25.24	60.00	61.98	**64.05**
driller	7.66	41.4	62.4	**65.65**	42.22	42.56	42.78
eggbos*		22	25.9	**50.17**	37.78	41.55	45.56
glue*	10.08	38.5	39.6	49.62	65.71	66.03	**66.43**
holepuncher	5.45	22.1	21.3	39.67	41.00	42.22	**44.50**
mean	6.42	24.9	30.4	40.77	44.26	45.66	**47.31**

此外，我们将定性结果与 PVNet 进行比较，在不同程度的遮挡情况下，我们的方法产生了令人满意的姿态精度(见图 5)。对于有严重遮挡的物体，PVNet 存在漏检和较低的识别率，而我们的方法可以解决这种情况。

Figure 5. Visualization of results on the Occlusion LINEMOD dataset (The upper left corner of the bounding box is the object's label and confidence respectively)
图 5. Occlusion LINEMOD 数据集上的结果可视化(边界框的左上角分别是物体的标签和置信度)

对于 T-LESS 数据集，我们在表 4 中把我们的方法与现有基线方法进行了比较。值得一提的是，在平均召回率指标方面，我们基于 RGB 的方法比基于 RGB-D 的方法要高出 4.43%。我们将定性结果与 EPOS 进行比较，如图 6 所示，从比较结果来看，我们的方法具有更高的鲁棒性和准确性。

在性能方面，我们使用了两个训练数据集的版本：合成数据集和真实数据集。如表 5 和图 7 所示，少量的真实数据对所有指标的性能都有约 10%的提升。在推理速度方面，二维检测器每张图片只需要约 8 ms，而 6DoF 姿态估计网络需要 20 ms。我们的方法在 NVIDIA RTX 2060 显卡上运行约 35 ms，基本满足了实时的性能。

Table 4. System resulting data of standard experiment
表 4. 标准试验系统结果数据

方法 / 指标	DPOD [7]	Sundermeyer [8]		EPOS [14]	Ours		
		RGB	RGB-D		w/o DS	w/DS	w/BBs
AR_{VSD}	-	-	-	-	42.34	42.72	**43.87**
AR_{MSSD}	-	-	-	-	46.89	47.67	**48.12**
AR_{MSPD}	13.9	50.4	51.4	63.5	68.32	69.13	**70.52**
AR	8.1	30.4	48.7	47.6	52.52	53.13	**54.17**

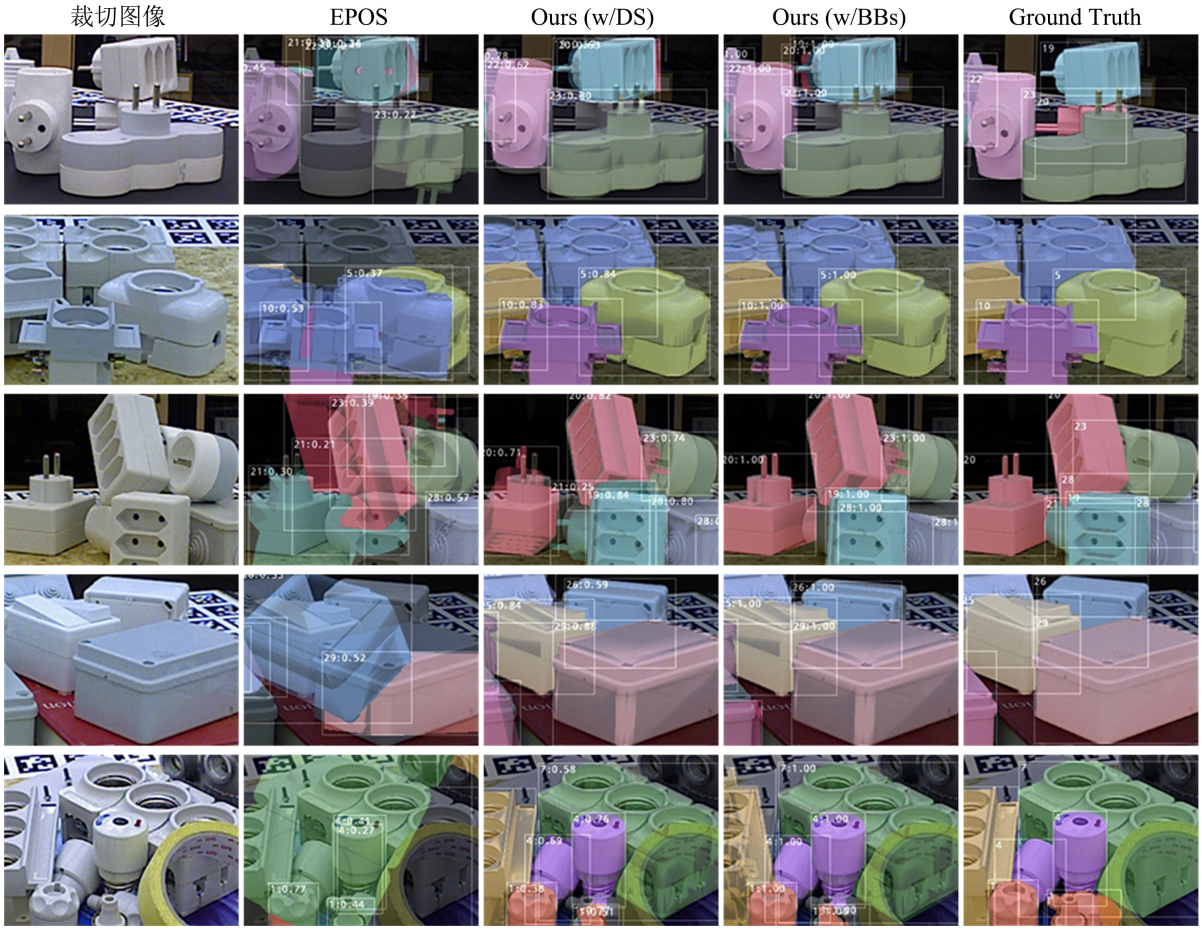

Figure 6. Visualization of results on the T-LESS dataset (The upper left corner of the bounding box is the object's label and confidence respectively)
图 6. T-LESS 数据集上的结果可视化(边界框的左上角分别是物体的标签和置信度)

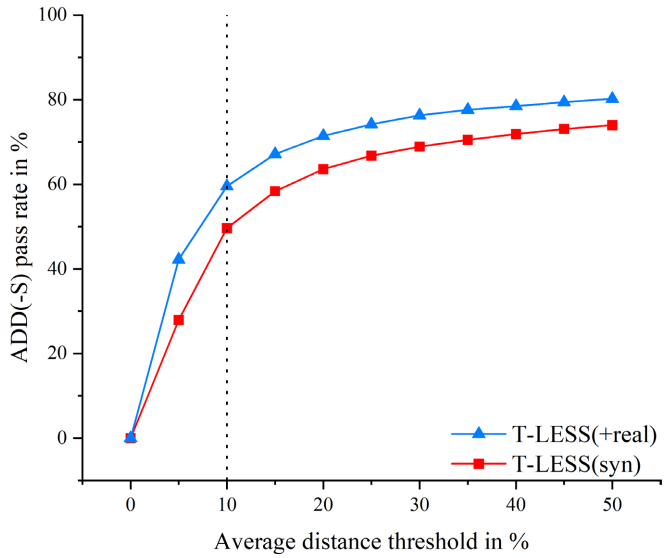

Figure 7. Accuracy-threshold curves under different training samples
图 7. 在不同训练样本下准确度－阈值曲线

Table 5. 6DoF pose estimation metrics under different training samples
表 5. 在不同训练样本下的 6DoF 姿态估计指标

指标	纯合成图像	+真实图像
2D 投影	36.98	46.33
ADD(-S)	49.63	59.54
AR	44.42	52.52

我们根据估计的 6DoF 姿势将物体的 3D 模型渲染到 RGB 图像上(见图 8)，已能够满足机器人抓取领域和增强现实领域。

Figure 8. Example results of our method on Occlusion LINEMOD (first row) and T-LESS (second row)
图 8. 我们的方法在 Occlusion LINEMOD (第一行)和 T-LESS (第二行)上的结果示例

5. 结论

在本文中，我们提出了一个全新的 6DoF 姿态估计网络框架，通过像素级的 XYZ 坐标映射进行 6DoF 姿态估计，这可以有效地处理遮挡情况和无纹理物体。同时，考虑了到 2D 目标检测对 6DoF 姿势估计的影响，我们的方法引入了动态缩放策略来提高鲁棒性。我们通过与现有的基线方法进行定量比较来评估我们的方法。实验结果表明，对于杂乱场景中的无纹理和遮挡物体，我们的方法优于基线方法。在未来的工作中，我们将通过多任务学习和注意力机制来优化网络结构，以提高准确性和稳健性。

参考文献

[1] Du, G.G., Wang, K. and Lian, S.G. (2020) Vision-Based Robotic Grasping from Object Localization, Pose Estimation, Grasp Detection to Motion Planning: A Review. *Artificial Intelligence Review*, **54**, 1677-1734. https://doi.org/10.1007/s10462-020-09888-5

[2] Mur-Artal, R., Montiel, J.M.M., Tardos, J.D. (2015) ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Transactions on Robotics*, **31**, 1147-1163. https://doi.org/10.1109/TRO.2015.2463671

[3] Hinterstoisser, S., Lepetit, V., Ilic, S., *et al.* (2012) Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. *Proceedings of the Asian Conference on Computer Vision*, Daejeon, 5-9 November 2012, 593-596. https://doi.org/10.1007/978-3-642-37331-2_42

[4] 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

[5] Tekin, B., Sinha, S.N. and Fua, P. (2018) Real-Time Seamless Single Shot 6D Object Pose Prediction. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 292-301. https://doi.org/10.1109/CVPR.2018.00038

[6] Peng. S., Liu. Y., Huang, Q., *et al.* (2019) PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4556-4565. https://doi.org/10.1109/CVPR.2019.00469

[7] Zakharov, S., Shugurov, I. and Ilic, S. (2019) DPOD: 6D pose Object Detector and Refiner. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 1941-1950. https://doi.org/10.1109/ICCV.2019.00203

[8] Sundermeyer, M., Marton, Z.-C., Durner, M., *et al.* (2018) Implicit 3D Orientation Learning for 6D Object Detection from RGB Images. *European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 712-729. https://doi.org/10.1007/978-3-030-01231-1_43

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

[10] Brachmann, E., Krull, A., Michel, F., *et al.* (2014) Learning 6D Object Pose Estimation Using 3D Object Coordinates. *European Conference on Computer Vision* (*ECCV*), Zurich, 6-12 September 2014, 536-551. https://doi.org/10.1007/978-3-319-10605-2_35

[11] Hodan, T., Haluza, P., Obdržálek, Š., *et al.* (2017) T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. 2017 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Santa Rosa, 24-31 March 2017, 880-888. https://doi.org/10.1109/WACV.2017.103

[12] Hodaň, T., *et al.* (2020) BOP Challenge 2020 on 6D Object Localization. *European Conference on Computer Vision* (*ECCV*), Glasgow, 23-28 August 2020, 577-594. https://doi.org/10.1007/978-3-030-66096-3_39

[13] Oberweger, M., Rad, M. and Lepetit, V. (2018) Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation. *European Conference on Computer Vision*, Munich, 8-14 September 2018, 125-141. https://doi.org/10.1007/978-3-030-01267-0_8

[14] Hodan, T., Barath, D. and Matas, J. (2020) EPOS: Estimating 6D Pose of Objects with Symmetries. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11700-11709. https://doi.org/10.1109/CVPR42600.2020.01172