# 三维模板跟踪的基准合成数据集构建及算法评估

何弦 [1),2)] 李佳宸 [1),2)] 金立 [1),2)] 刘力 [3)] 钟凡 [4)] 秦学英 [1),2)]

[1)](山东大学 软件学院 济南 250101)

[2)](数字媒体技术教育部工程研究中心 济南 250101)

[3)](视辰信息科技(上海)有限公司 上海 201203)

[4)](山东大学 计算机科学与计算学院 青岛 山东 266237)

**摘 要** 三维模板跟踪旨在将预先构建的三维 CAD 模型与输入图像中的相应目标进行精确配准，在增强现实、机器人等领域具有重要的应用，也是计算机视觉领域的关键问题之一。近年来，三维模板跟踪的准确率和稳定性都得到了持续提升，但仅有少量的工作关注三维模板跟踪数据集的构建。随着深度学习的普及，各领域中大规模数据集的构建越来越被重视，为算法的训练、测试和评估奠定了基础，极大地推动了相关领域的发展。以往的三维模板跟踪数据集大多存在规模有限，画面不够自然、真实，多样性不足等问题。基于此，本文创建了一个大规模的基于真实感渲染的三维模板跟踪数据集(Render Dataset for Object Tracking，简称 RDOT)，其包含多种不同结构和材质的物体、复杂的运动模式，并且在场景、光照、噪声、运动模糊和遮挡等方面有丰富细致的设置，是目前三维模板跟踪领域最大的数据集，满足三维模板跟踪算法评估的各种需求。针对现有三维模板跟踪算法测评时使用的数据集不统一，测评结果难以客观全面地反映算法性能的问题，本文基于所构建的数据集，利用平均边缘距离、平均表面距离和重初始化率三种度量标准全面评估了目前主流的三维模板跟踪算法，并对评测结果进行了深入的分析讨论，给出了全面的分析报告和技术展望。此外，基于所构建的数据集，本文提出了对跟踪结果建立误差分析模型，并对结果进行校正的方法，有效改善了三维模型跟踪的准确率。

**关键词** 三维模板跟踪；数据集构建；算法测评；增强现实；真实感渲染

**中图法分类号** TP391

# A Synthetic Dataset and Performance Evaluation for 3D Template Tracking

HE Xian[1),2)] LI Jia-Chen[1),2)] JIN Li[1),2)] LIU Li[3)] ZHONG Fan[4)] QIN Xue-Ying[1),2)]

[1)](Department of Software, Shandong University，Ji'nan 250101)

[2)](Engineering Research Center of Digital Media Technology, Ministry of Education, Shandong University, Ji'nan 250101)

[3)](Shichen Information Technology (Shanghai) Co., Ltd, Shanghai 201203)

[4)](Department of Computer Science and Technology, Shandong University, Qingdao 266237)

**Abstract** 3D template tracking aims to accurately align pre-constructed 3D CAD models with the corresponding targets in the input images, and has important applications in augmented reality and robotics. It is also one of the key problems in the field of computer vision. In recent years, various approaches have been proposed to improve the accuracy and robustness of 3D template tracking, but only a small amount of work has contributed to the construction of 3D template tracking datasets. With the development and wide applications of deep learning, the construction of large-scale datasets in various fields has been paid more and more attention, laying the foundation for the training, testing and evaluation of algorithms, which has greatly promoted the development of related fields. Previous datasets for 3D template tracking are acquired by either video capture or computer rendering. Video-captured datasets are realistic, but since the pose is computed based on hand-crafted

本课题得到工信部2019年工业互联网创新发展工程项目、之江实验室项目(2020NB0AB02)、国家自然科学基金项目(61907026)、山东省高等学校科学技术计划项目(J18KA392)资助.**何 弦**, 硕士研究生，主要研究领域为增强现实、计算机视觉.E-mail: hexian_18@163.com.**李佳宸**，博士研究生，主要研究领域为增强现实.**金 立**，硕士研究生，主要研究领域为增强现实.**刘 力**，硕士，主要研究领域为计算机视觉.**钟 凡**，博士，副教授，主要研究领域为图像视频处理、计算机视觉.E-mail: zhongfan@sdu.edu.cn.**秦学英**(通讯作者)，博士，教授，主要研究领域为图像视频处理、计算机视觉.E-mail: qxy@sdu.edu.cn.

markers, the accuracy of the ground-truth pose is not guaranteed and the size of these datasets are also limited due to the time-consuming labelling process. Computer-rendered datasets could be synthesized massively, but the quality of rendered image sequences is limited by the adopted render techniques. Altogether, previous datasets suffer from problems such as limited scale, inaccurate ground-truth poses, unrealistic images and insufficient diversity of model settings, therefore it is meaningful and challenging to construct a high-quality and large-scale dataset for 3D template tracking. In this paper, we propose to construct a large-scale 3D template tracking dataset RDOT (Render Dataset for Object Tracking) based on photorealistic rendering. RDOT is rendered with photorealistic rendering method. The model set contains tens of objects with different physical structures and realistic materials, it also allows the camera and objects to move in pre-defined complex motion modes. Moreover, compared with previous datasets, RDOT takes more accurate control of settings of rendering scenes, it offers various detailed settings of lighting, noise, motion blur and occlusion in different degrees of difficulty. To the best of our knowledge, RDOT is currently the largest 3D template tracking dataset which meets the demands of performance evaluation. Based on RDOT, we evaluated previous 3D template tracking methods in an objective and fair way. Previous approaches have been evaluated on different datasets that suffer the aforementioned problems. In our evaluation, the tracking methods are evaluated with three precision metrics, including ADE (Average Edge Distance), ASD (Average Surface Distance) and RR (Reinitialization Rate). We analyze the evaluation results from multiple aspects considering structures of objects, materials of objects and different settings of rendering scenes. In addition, since RGB-based 3D tracking method usually produce significant errors in the depth direction due to the missing of depth constraint, we propose a statistical model of tracking errors that can be computed based on the accurate ground-truth pose of RDOT. By applying the error model to compensate the resulting object pose parameters, the tracking accuracy can be improved significantly. Finally, we discuss the disadvantages of different tracking approaches, and give an overall conclusion and perspective for future 3D template tracking approaches.

**Key words** 3D template tracking; dataset construction; algorithm evaluation; augmented reality; photorealistic rendering

## 1 引言

基于视觉的三维模板跟踪算法是在视频数据中通过拟合预先建模的实体三维模型与图像特征，跟踪估计相机与物体间的位姿参数，实现实体与相机三维空间注册的技术。三维模板跟踪技术广泛地用于增强现实、机器人技术以及自然的人机交互技术等。

由于三维空间注册技术在很多领域具有实时性和高精度的要求，在无实体三维模型作为先验的情况下直接采用视觉分析技术进行实体与相机之间的空间注册很难获得鲁棒的结果，因此将物体的三维模型作为先验信息能够产生有力的约束条件，提高三维跟踪算法的精度，由此产生了三维模板跟踪技术。

三维模板跟踪技术作为重要的视觉问题之一，已有大量的相关工作。根据算法在实际应用场景下处理数据的类型，主要分为两类：基于深度数据的三维模板跟踪算法[1-3]和基于 RGB 数据的三维模板跟踪算法[4-10]。深度数据由深度相机获取，往往受限于深度相机的使用条件，不具通用性；有纹理的三维模板跟踪算法已较为成熟，而弱纹理或无纹理的模板跟踪算法在速度和精度上有待提高，且具有更广泛的应用场景。因此我们主要关注基于弱纹理或无纹理的 RGB 数据三维模板跟踪算法。

三维模板跟踪算法在性能上不断提高，但在评估算法性能时也存在不可忽视的问题：部分算法[6,7,9]在测评性能时使用各自实拍数据，此类数据规模小且不统一，这导致比较算法间的精确度和鲁棒性非常困难；而且目前常用的实拍数据集[1,11]的基准位姿利用标定算法计算获得，不可避免存在一定误差，且其画面大多包含用于标定的二维标志，画面不自然，容易对算法构成干扰；部分算法[12]测评使用渲染数据集，此类数据集存在图像画面不真实、模型材质较少等问题。因此，为了客观全面地评估三维模板跟踪算法，构建一个大型的高真实感且基准位姿准确的三维模板跟踪数据集是非常有意义且兼具挑战性的工作。

本文建立了一个大规模的高真实感、基准位姿

无误差的 RGB 图像数据集，在一定程度上解决了现有数据集数据量小、基准位姿不准确、渲染画面不真实等问题，可用于弱纹理或无纹理三维模板跟踪算法的性能评测。表 1 对比了本数据集与其他数据集的基本信息。本数据集利用基于图像的光照技术（image based lighting，简称 IBL）渲染得到的虚实融合数据集，是目前最真实的三维模板跟踪渲染数据集之一。本数据集模拟了多样性、随机性的运动模式，增加了光照变化、噪声、运动模糊、遮挡等场景设置，生成超过 160 万帧图像，本数据集的图像示例图如图 1、图 2 所示。

本文的主要贡献总结如下：

（1）建立了目前最大规模的高质量三维模板跟踪合成数据集。与现有实拍数据集[1,11]相比，本数据集的画面自然、无标志物侵入，基准位姿准确；与现有渲染数据集[11,13]相比，本数据集的数据量增加了 20 余倍，画面真实感高，数据构成更为丰富；本数据集公开在 https://github.com/Xian-He/RDOT。

（2）增加了多样化的物体材质和复杂的场景设置，解决了现有合成数据集不具备半透明、高光材质和运动模糊等，实现了更全面的数据集构建。

（3）提出了误差分析模型，建立与本数据集类似的训练数据集，根据算法在训练数据集中反映的统计规律，建立误差修正模型，能提升测试数据集中算法性能。

（4）提出了多维度的三维模板跟踪算法评估方式，并对主流算法进行测评分析，得出了最全面的测评结果。与其它数据集测评工作[12,13]相比，本文评测的三维模板跟踪算法种类更多，同时根据物体的属性和场景设置进行了多维度的评估和分析。

**表 1 三维模板跟踪数据集基本信息**

| | 场景 | 运动模式 | 物体属性 | | | 场景设置 | | | | 测试算法 | 数据规模 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 物体数量 | 物体结构 | 物体材质 | 光照变化 | 噪声 | 运动模糊 | 遮挡 | | |
| 文献[12] | 1 个 | 4 种 | 4 个 | 2 种 | 1 种 | × | × | × | × | 2 种 | 4000 帧 |
| 文献[11] | 1 个 | 5 种 | 6 个 | 1 种 | 1 种 | √ | × | √ | × | 2 种 | 3324 帧 |
| 文献[13] | 1 个 | 1 种 | 18 个 | 4 种 | 1 种 | √ | √ | × | √ | 2 种 | 72072 帧 |
| OURS | 6 个 | 12 种 | 27 个 | 4 种 | 4 种 | √ | √ | √ | √ | 6 种 | 1603800 帧 |

图 1 本文建立的 RDOT 数据集基础图像示例图（该图展示了不同物体在不同场景的图像示例图，其中每一列图像来自同一场景）

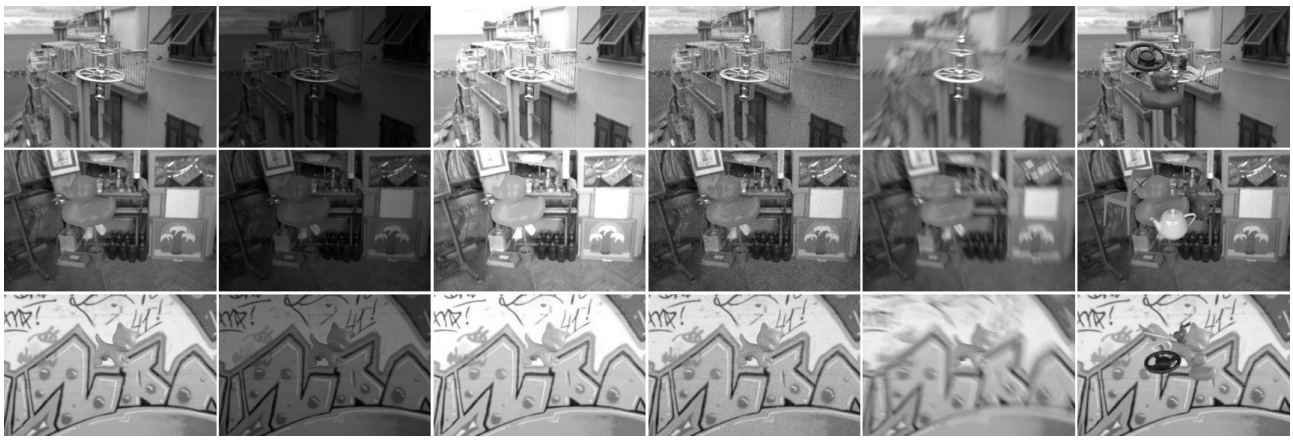



图 2 本文建立的 RDOT 数据集增强图像示例图（每一行图像展示物体在同一运动路径下的不同场景设置，从左到右场景设置依次为：基础图像，光照 1，光照 2，噪声，运动模糊，遮挡）

## 2 相关工作

### 2.1 三维模板跟踪概述

三维模板跟踪技术实现了动态物体的空间注册，即计算三维模型所在的物体坐标系与相机坐标系的连续相对位姿关系[14]，实现虚实融合。位姿即位置与姿态，由平移和空间旋转构成，共 6 个自由度。

三维模板跟踪算法输出的位姿可用于增强现实领域，将三维虚拟物体准确地放置在实体定义的现实空间中的指定位置上，从而呈现虚拟与现实在几何和空间上一致性融合；也可用于机器人抓取物体的过程，根据算法得出的位姿信息规划机器臂的行为路径，实现对实物的准确抓取。三维模板跟踪算法主要分为基于深度数据[1-3]和基于 RGB 数据[4-10]的三维模板跟踪算法。

基于深度数据的三维模板跟踪算法中，一类是将位姿估计建模为三维点云配准问题，利用迭代最近点算法[15]（Iterative Closest Point，简称 ICP 算法）进行能量最小化求解最优位姿，此类算法求解复杂度高；另一类是通过基于学习的方法提取深度图特征，将其映射到物体位姿[1-3]。虽然当前深度传感器（Kinect 等）较为普及，但是由于传感器成像原理的限制，目前基于深度数据的三维跟踪算法局限于室内环境。

基于 RGB 数据的三维模板跟踪算法通常需要通过图像特征建立图像平面点与三维物体表面点的关联，然后利用迭代优化算法求解最优位姿[4,8]。图像特征在光照、尺度和几何形变等情形下的稳定性决定了算法的性能。因此，这类算法重点关注如何通过图像特征准确建立图像平面点与三维物体表面点的关联。根据图像平面上可利用的特征类型大致分为基于边缘[4-7]和区域[8-10]的三维模板跟踪算法。

基于边缘的三维模板跟踪算法[4-7]的基本思想是先将已知的三维模型初始位姿投影到二维图像，获得其轮廓线，然后在模型投影点的法线方向上寻找对应图像边缘点，通过最小化二者之间的误差迭代求得跟踪物体的最佳位姿。Harris 等人[5]提出了第一个实时的基于边缘的三维模板跟踪系统（Real-time Attitude and Position Determination，简称 RAPiD）。Wang 等人[16]提出的全局最优搜索算法（Global Optimal Searching，简称 GOS）采用动态规划进行了对应边缘点的全局寻优，并且增加了几何约束。该算法在复杂场景和镂空物体中表现较好。Wang 等人[14]提出的基于边缘距离场算法（Edge Distance Field，简称 EDF）最小化跟踪物体的投影轮廓与图像边缘的距离,利用边缘方向一致性来验证跟踪物体姿态。基于边缘的方法依赖于图像边缘线提取效果，对光照的变化有较强的适应性，而当物体所处环境背景复杂或存在运动模糊时，边缘特征提取质量较低，算法容易跟踪失败。

基于区域的三维模板跟踪算法[8-10]通常利用物体表面的颜色区域信息，使用水平集函数对物体进行隐式分割，寻找到能将图像进行前景与背景最大化分割的最优位姿。Prisacariu 等人[9]提出了具有代表性的基于区域的三维跟踪方法（Pixel-Wise Posterior，简称 PWP）。PWP 算法是逐像素计算图

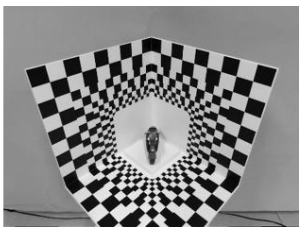
像颜色属于前景和背景的概率，计算所有像素的联合后验概率，最大化前景和背景像素概率，利用一阶最速梯度下降算法优化，求解跟踪位姿。 Tjaden 等人[17]将上述优化算法改进为高斯牛顿算法，更好地处理快速旋转和尺度变化，降低了整体运行时间。Tjaden 等人[13]采用局部颜色直方图计算前背景概率，保持物体在跟踪过程中局部颜色直方图的时间一致性，提高了在动态遮挡情况下姿态跟踪的鲁棒性。基于区域的方法在复杂背景、运动模糊、遮挡等条件下具有更好的鲁棒性，但当前景与背景颜色相近或场景光照变化剧烈，算法不能准确分割前背景时，容易跟踪失败。另外，其涉及的采样点为轮廓周围的环形区域，比基于边缘的方法更加耗时。

在算法开始或跟踪失败时，可以借助三维检测与姿态估计算法[18, 19]来初始化或重启跟踪。三维检测与姿态估计算法对于部分遮挡和复杂背景有一定容忍性，但此类方法的速度较慢，难以完全替代三维模板跟踪算法。

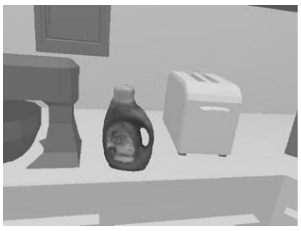
### 2.2 三维模板跟踪数据集概述

在计算机视觉领域，随着深度学习的普及，大规模数据集[20, 21]对算法研究具有极大的推动作用，不仅用于测试和评估，也在很大程度上为算法提供学习和训练的基础，因此广受注目。目前只有少量数据集（如图 3 所示）是为三维模板跟踪任务而创建，三维模板跟踪领域也亟待统一的大规模基准数据集。

（1）文献[14]测试数据示例图



（2） 文献[11]数据集图像示例图



（3）文献[12]数据集图像示例图

（4）文献[13]数据集图像示例图

图 3 不同三维模板跟踪数据集的图像示例图

目前三维跟踪数据集的构建主要包括实拍真实三维实体所在场景[11, 22]和渲染[12, 13]两种方式。实拍数据的采集一般是将三维物体放置在场景中，利用相机拍摄获得的一系列连续画面，其中三维物体的姿态通过基于标志物的标定算法计算获得。基于实拍真实场景建立的数据集画面真实，但基准位姿一般通过标定算法计算得到，涉及二维标志以及二维标志与物体之间的多次标定，容易产生误差，难以得到准确的位姿信息，图 4 所示现有相关工作[1, 11]均存在基准姿态不准确的问题; 并且二维标志一般是黑白分明的人工标志，严重影响画面的自然性。渲染数据集的采集一般是由渲染程序将三维模型合成到实拍视频中或直接在虚拟场景中渲染得到。渲染数据集基准姿态准确，而且可以短时间生成大量数据，但现有渲染数据集的图像缺乏真实性，前景和背景融合效果较差，如图 2（3）（4）。因此如何平衡准确位姿与真实感画面的问题尤为重要。

（1）文献[1]位姿示例图

（2）文献[11]位姿示例图

图 4 实拍数据集存在基准位姿不准确的问题（其中，虚线表示数据集中提供的基准位姿下物体轮廓，实线表示手工提取的物体实际轮廓，两者间存在一定误差）

在统一的数据集出现之前，三维模板跟踪算法通常使用各自拍摄的视频序列作为测试数据（如图 3（1）），这导致不同算法的评估结果难以统一。而且这类数据规模较小，往往是为了说明算法性能而拍摄，难以全面的反映算法的问题。

为了解决数据集规模小且不统一的问题，Wu 等人[11]提出的物体姿态跟踪数据集（Object Pose Tracking，简称 OPT）（如图 3（2））是目前最复杂的实拍三维模板跟踪数据集。该数据集利用机械臂手持相机以不同运动路径、不同速度、在不同光照环境下拍摄。尽管该数据集的模式较为复杂，但拍摄的背景环境较为简单，运动路径受限于机械臂的活动范围。Tejani 等人[22]建立的多物体数据集将相同的物体放置在同一环境中，物体间存在部分遮挡。虽然该数据集提供了部分连续图像测试序列，但主要用于三维检测与姿态估计技术。

实拍数据集获得的基准位姿存在一定误差，而

且构建大型实拍数据集需要大量的时间成本和硬件资源。基于渲染的方式能有效解决上述问题。Choi 等人[12]建立的三维跟踪数据集（如图 3（3））包含 4 个全渲染图像序列共 4000 帧图像，但存在渲染场景的纹理较为简单等问题。Pauwels 等人[23]建立的虚实融合数据集增加了噪声和遮挡，但该数据的渲染环境光照单一，渲染物体不真实。Tjaden 等人[13]提出的基于区域物体跟踪数据集（Region-based Object Tracking，简称 RBOT）（如图 3（4））改进了前者的问题，模拟物体和相机同时运动的情况，共生成 7 万多帧有效图像。但该数据集组织方式较为单一，虽然画面具备了一定真实感，但与真实场景还存在一定差距。

为了解决目前数据集存在的问题，本文利用基于图像的光照的渲染方式，在保证基准位姿准确的前提下建立了具有高真实感的三维模板跟踪数据集，生成的图像序列的光照条件能与周围的环境光照一致，而且能真实的表现镜面、半透明等特殊材质，表 1 给出了本数据集的基本信息。为了增加数据集的多样性，我们设计了多种运动模式并对场景设置进行了控制，定量增加了光照、噪声、运动模糊、遮挡等变量。本文对主流的三维模板跟踪算法进行了测评，提供目前最全面的三维模板跟踪算法测试结果。

## 3 数据集构建方法

本节首先介绍本数据集的组成信息，包括基础数据集和增强数据集的基础信息；然后详细说明数据集的具体建立方法，包括场景与三维物体模型、物体材质、运动模式和基准姿态、场景设置等设置。

### 3.1 RDOT数据集总览

本文建立的 RDOT 数据集由两部分数据组成：基础数据集和增强数据集。基础数据集共 1944 个基础图像序列，每个序列 300 帧，共生成 583200 帧；增强数据集增加了光照、噪声、运动模糊和遮挡等不同的设置，共生成 3402 个增强图像序列，共 1020600 帧。RDOT 数据集是目前最大的三维跟踪数据集，具体组成信息如表 2，基础数据集渲染效果如图 1，增强数据集渲染效果如图 2。

表 2 RDOT 数据集组成信息

| | |
|---|---|
| 图像分辨率 | 720x480 |
| 基础数据集 | 由 6 个场景、27 个物体、12 种运动模式组合生成 1944 个基础数据图像序列，每个序列 300 帧 |
| 增强数据集 | 在上述四分之一基础图像序列中进行 4 种不同的场景设置，即在 3 个场景、27 个物体、6 种运动模式基础上增加了 4 种场景设置（包括 2 种光照变化、2 种不同程度的噪声、2 种不同程度的运动模糊以及 1 种遮挡）组合生成 3402 个增强数据图像序列，每个序列 300 帧 |

本数据集构建流程为：准备高质量的实拍世界高动态光照渲染（High-Dynamic Range，简称 HDR）环境贴图和物体模型，设计渲染环境；自动化生成不同模式下相机和物体的运动路径，结合相机和物体的运动模式，计算得到物体相对相机的基准位姿；根据该基准位姿变换相机位姿，并逐帧渲染得到基础图像序列，即基础数据集；在基础图像序列中增加不同的场景设置得到增强图像序列，即增强数据集。

### 3.2 场景与三维物体模型

RDOT 数据集共使用 6 个室内和室外的真实世界 HDR 环境贴图和 27 个无纹理或弱纹理的三维物体模型，真实世界 HDR 环境贴图和渲染物体模型示例图分别如图 5、图 6 所示。

图 5 RDOT 数据集渲染世界 HDR 环境贴图示例图（前三个为室外场景，后三个为室内场景）

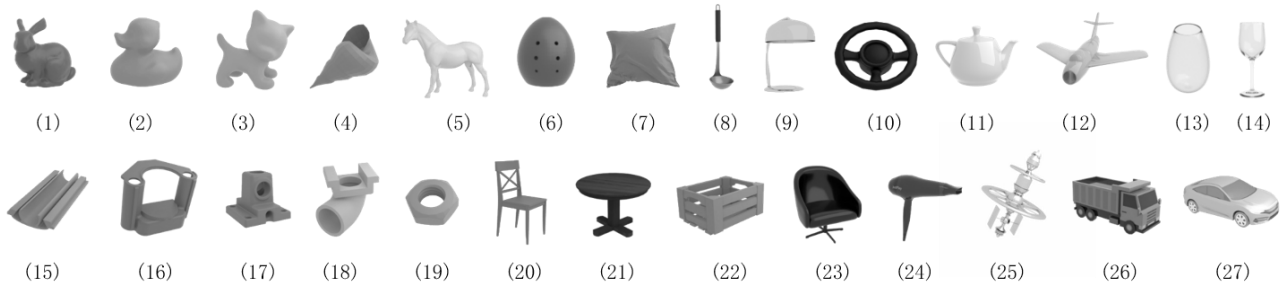



图 6 RDOT 数据集中三维物体图像示例图((1)bunny ,(2)cat ,(3)duck ,(4)seashell ,(5)horse ,(6)meteorite ,(7)pillow ,(8)spoon ,(9)lamp ,(10)wheel ,(11)teapot ,(12)airplane ,(13)bottle ,(14)glass ,(15)bottom ,(16)lock ,(17)clamp ,(18)tube ,(19)parts ,(20)chair，(21)table，(22)crate，(23)leather chair，(24)hair dryer，(25)space station，(26)truck，(27)car)

当前数据集[11][13]多为室内场景，故本数据集选择了室内和室外两种不同的真实世界环境贴图，共 6 个，来自于开源库[1]，可满足多样化的测试需求，如图 5。本文利用基于图像的光照渲染方式，将虚拟物体放置到真实世界光照环境 HDR 贴图中，该方式能有效模拟真实的物体同环境真实的光线追踪和真实世界光照信息，具有正确的投影和明暗效果，真实感强。

三维物体模型来自于开源模型库[2]或手工建模，共 27 个，如图 6。模型共有四种结构，分别为对称物体、非对称物体、存在孔洞的物体、不存在孔洞的物体；模型的材质共有四种，包括没有镜面效果的哑光材质、有镜面高光效果的金属和瓷器材质、半透明玻璃材质和弱纹理的木纹材质模型等。表 3 给出了具体物体模型的属性信息，图 7 给出特殊材质在本数据集中渲染效果图像示例图。

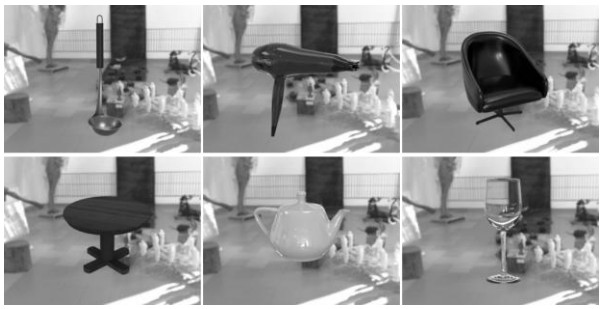


图 7 不同物体材质在本数据集中渲染效果图像示例图（材质依次为：哑光材质、金属材质、皮质材质、木纹材质、瓷器材质和玻璃材质）

**表 3 RDOT 数据集三维物体模型属性信息**

| 物体属性 | | 物体名称 |
|---|---|---|
| 物体结构 | 非对称 | bunny，cat，duck，seashell，horse，tube； |
| | 对称 | meteorite，pillow，spoon，lamp，wheel，teapot，airplane，bottle，glass，bottom，lock， clamp，parts，chair，table，crate，leather chair，hair dryer，space station，truck，car； |
| | 无孔洞 | bunny，cat，duck，horse，pillow，spoon，lamp，airplane，bottle，glass，bottom，table，leather chair，hair dryer，truck，car； |
| | 有孔洞 | seashell，meteorite，wheel，teapot， lock，clamp，tube，parts，chair，crate，space station。 |
| 物体材质 | 非镜面 | bunny，cat，duck，seashell， meteorite，pillow，wheel，bottom，clamp，tube，parts，chair，crate，table, truck； |
| | 镜面 | horse，spoon，lamp，teapot，airplane，lock，leather chair，hair dryer，space station，car，bottle，glass； |
| | 半透明 | bottle，glass； |
| | 弱纹理 | table， truck，hairdryer。 |

### 3.3 运动模式与基准位姿

为了增加运动路径的多样性和不可预测性，本文为相机和物体设计了不同的运动模式，组合二者的运动模式模拟相机和物体同时运动，最后计算得到基准位姿。

本文为相机设计了 6 种运动模式，分别为缩放、平移、平面内旋转、平面外沿经线和沿纬线旋转、螺旋式运动。

**缩放。**相机沿着视线方向靠近或远离三维物体，图像中物体被放大或缩小。

**平移。**相机在视平面(与视线方向垂直并通过相机坐标系原点）内自由移动。

**平面内旋转。**相机保持位置不变，以视线方向

[1] https://hdrihaven.com/hdris/?c=all

[2] http://benedikt-bitterli.me/resources/
https://assetstore.unity.com/

为轴进行旋转。

**平面外沿经线和沿纬线旋转。**相机注视物体，并在以物体为球心的球面上沿球面上的经线或纬线移动。

**螺旋式移动。**相机沿着视线方向以螺旋线方式进行移动。

本文为物体设计了 2 种运动模式，分别为旋转和自由移动。

**旋转。**物体依次沿模型局部坐标系的 X、Y、Z 轴自转，生成的物体路径在一个固定点上旋转。

**自由移动。**物体依次在模型局部坐标系的 XOY、YOZ、ZOX 平面内以随机加速度沿固定的局部坐标原点自由移动。物体运动到每一时刻的位置由迭代公式（1）计算得到.

$$\begin{cases}\mathbf{P}_{k+1}=\mathbf{P}_k+\mathbf{V}_k\cdot\nabla t\\ \mathbf{V}_{k+1}=\mathbf{V}_k+\mathbf{A}_k\cdot\nabla t\\ \mathbf{A}_{k+1}=\lambda_a\cdot(\mathbf{P}_k-\mathbf{T})+\lambda_n\cdot\mathbf{A}_k^n\\ \mathbf{A}_k^n\backsim \mathrm{N}(X;U,\Sigma)\end{cases}\tag{1}$$

其中，$\mathbf{P}_k$ 表示第 k 帧物体的位置，$\mathbf{V}_k$ 表示第 k 帧物体的速度，$\nabla t$ 表示帧间时间间隔，$\mathbf{A}_k$ 表示第 k 帧物体加速度，$\mathbf{A}_k^n$ 表示随机加速度，其满足均值为$U$，方差为$\Sigma$的高斯分布，$\lambda_a$、$\lambda_n$用来控制物体加速度随机性的尺度因子，以增加路径的不可预测性。

上述运动模式在全局世界坐标系下生成，本文通过式（2）方程复合相机和物体变换矩阵计算得到物体相对于相机的基准位姿：

$$\begin{cases}\mathbf{R}_{CO}=\mathbf{R}_{WC}^{-1}\mathbf{R}_{WO}\\ \mathbf{t}_{CO}=\mathbf{R}_{WC}^{-1}(\mathbf{t}_{WO}-\mathbf{t}_{WC})\end{cases}\tag{2}$$

其中，$\mathbf{R}_{CO}$、$\mathbf{t}_{CO}$分别为物体在相机坐标下旋转矩阵与平移向量，$\mathbf{R}_{WC}$、$\mathbf{t}_{WC}$分别为相机在世界坐标系下旋转矩阵和平移向量，$\mathbf{R}_{WO}$、$\mathbf{t}_{WO}$分别为物体在世界坐标系下旋转矩阵和平移向量。本文的基准位姿均表示物体在相机坐标系下的位姿。

本文设计了丰富的运动模式，共 12 种复合运动模式，在不同场景中不同物体运动路径均不相同，尽可能模拟物体和相机运动路径，增加的噪声具有不可预测性，且允许修改运动模式的采样率模拟不同运动速度。图 8 展示了本数据集部分运动模式的轨迹示意图。

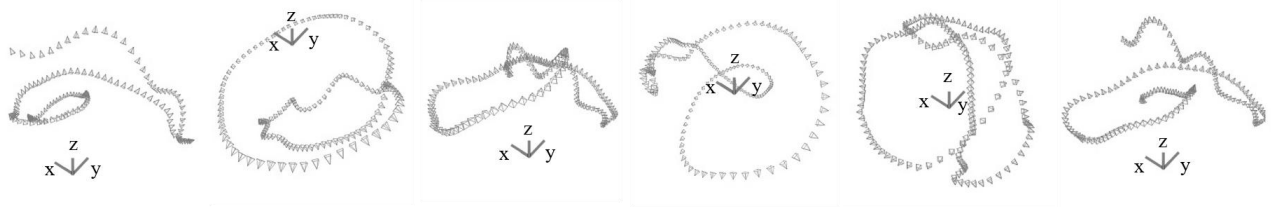


图 8　本数据集中复合运动模式轨迹示意图。从左到右依次为：自由移动缩放曲线、自由移动平移曲线、自由移动平面内旋转曲线、自由移动平面外沿经线旋转曲线、自由移动平面外沿纬线旋转曲线、自由移动螺旋线曲线。

### 3.4　场景设置

为了增加数据集的挑战性，本文对场景的复杂度进行了设置，在基础序列的基础上，增加了光照变化、噪声、运动模糊和遮挡，并对其进行定量控制。不同场景设置效果图如图 2 所示。

**光照。**光照变化会导致物体的颜色变化，对于图像特征提取具有一定干扰，特别是基于区域的算法。因而为了全面测试算法对光照变化的鲁棒性，本文设计了两种强度变化的闪光灯来改变场景渲染画面的明暗。

**噪声。**噪声是在实拍数据中常有的变量，光线较暗的时候噪声明显，但基于渲染的数据集前背景往往比较干净，难以模拟真实数据的效果。因而本文为了更好的模拟实拍数据的效果，在场景的渲染画面中合成两种强度不同的彩色高斯噪声来增加场景复杂度，用来测试算法对噪声的稳定性。

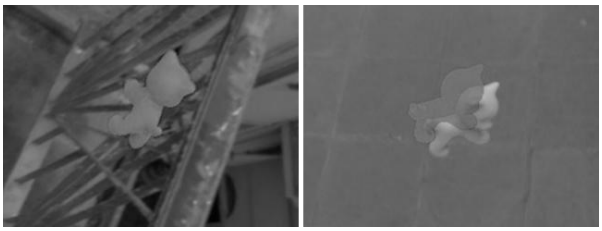
**运动模糊。**实拍数据中一旦物体发生模糊就难以获得准确的基准位姿，而普通的虚实融合渲染方式[21]将渲染物体与实拍视频进行 alpha 通道合成，难以生成真实的模糊效果。本文基于物理的渲染方法，能较好的模拟运动模糊的效果，得到高真实感的运动模糊。本文设计两种不同程度的运动模糊，分别将 25%和 50%的渲染帧进行模糊，用来测试算法对运动模糊的稳定性。

**遮挡。**实拍数据中，被遮挡的物体难以获得较为准确的基准位姿，渲染数据则不存在此问题。为

了模拟真实世界中的遮挡效果，本文在渲染世界中对跟踪物体设置其它遮挡物，使得物体在渲染图像序列中存在一定的遮挡，生成了遮挡数据集，用来测试算法对遮挡的稳定性。

## 4 误差分析建模

在单视角画面中，三维模板跟踪算法获得的物体位姿在该图像上观察具有良好的吻合度（如图 9（1）所示）；然而，当从另外的视点观察时，却存在较大的误差（如图 9（2）所示）。经过观察发现，物体在相机视线方向的平移不会导致物体图像投影图的显著变化，但位姿却存在误差，这是由于算法的平移分量沿相机视线方向缺少约束，会导致物体的抖动甚至晃动。



（1） （2）

图 9 文献[13]算法在本数据集中的位姿投影图（（1）本数据集相机视线下投影图，（2）该姿态在另一相机视线下投影图）

本文构建的数据集具有准确的基准位姿，可对算法进行精确的误差分析。因为该误差模型与物体表面的形状、距离相机的远近有关，因此本文建立了与测试数据集物体、距离类似的训练子集，对跟踪算法进行误差分析。通过分析训练数据集上的平移误差分布得出统计规律，建立经验模型，对物体在相机视线方向上进行修正，可用于三维模板跟踪算法的后处理，以去除测试数据集位姿轨迹的抖动。

基于上述方法，本文对训练数据进行了统计分析，所建模型如公式（3）所示。因为误差与物体在图像上的大小有关，采用该方式能有效平衡距离上的误差。

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mathbf{t}_i^g|-|\mathbf{t}_i|}{|\mathbf{t}_i^g|^2} \qquad (3)$$

其中，$\mathbf{t}_i^g$、$\mathbf{t}_i$ 分别表示训练数据集中第 i 帧基准位姿与跟踪算法位姿的平移分量，n 表示图像帧数，$\mu$ 为误差因子。

对于每一个训练数据集可利用上述公式计算得到误差因子 $\mu$。因为训练集与测试集的误差分布类似，故在测试集中，可利用训练集中的误差因子 $\mu$ 对测试集中算法跟踪结果进行相机视线方向上的修正，利用公式（4）可计算出修正后算法平移分量的模长 $|\mathbf{t}_i^r|$，结合已知的相机视线方向，利用公式（5）可求解得出算法修正后位姿的平移分量。

$$\frac{|\mathbf{t}_i^r|-|\mathbf{t}_i|}{|\mathbf{t}_i^r|^2} = \mu \qquad (4)$$

$$\mathbf{t}_i^r = |\mathbf{t}_i^r| \cdot \frac{\mathbf{t}_i}{|\mathbf{t}_i|} \qquad (5)$$

其中，$\mathbf{t}_i^r$、$\mathbf{t}_i$ 分别表示测试数据集中第 i 帧修正后算法位姿与跟踪算法位姿的平移分量，$\mu$ 为误差因子。

## 5 实验结果与分析

本节首先介绍评测的主流的 6 种三维模板跟踪算法，以及所用的三种性能度量准则，从整体、不同物体属性和不同场景设置三个维度测评算法并全面分析实验结果，然后对算法进行误差分析建模，修正算法的姿态，提高算法准确率，最后分析了本数据集与现有数据集的区别和联系。

### 5.1 参与测评的算法及性能度量准则

本文从主流的基于 RGB 图像的三维模板跟踪算法中选择了 6 种算法进行测评。表 4 列举了评测算法的主要特征。

表 4 参与评测的三维模板跟踪算法

| 算法 | 代码 | 算法特征 |
|---|---|---|
| PWP3D[9] | C++ | 基于区域的算法 |
| GOS[15] | C++ | 基于边缘的算法 |
| 文献[17] | C++ | 基于区域的算法 |
| 文献[24] | C++ | 基于区域的算法 |
| EDF[7] | C++ | 基于边缘的算法 |
| 文献[13] | C++ | 基于区域的算法 |

为了全面评估算法精度，本文使用二维度量方法平均边缘距离[14]（Average Edge Distance，简称 AED）、三维度量方法平均表面距离[1]（Average

Surface Distance，简称 ASD）和重初始化率（Reinitialization Rate, 简称 RIR)度量算法的误差，使用每秒处理图像帧数（Frames Per Second，简称 FPS）度量算法的时间性能。

5.1.1 平均边缘距离

平均边缘距离是度量三维物体模型估计位姿到基准位姿下边缘轮廓之间距离的平均值。第 $i$ 帧图像， AED 的计算方式为：

$$AED=\frac{1}{C}\sum_{k=0}^{C-1}\|\pi\left(\mathbf{K}\left(\mathbf{R}_i\mathbf{X}_k+\mathbf{t}_i\right)\right)-\pi\left(\mathbf{K}\left(\mathbf{R}_i^g\mathbf{X}_h+\mathbf{t}_i^g\right)\right)\|_2 \quad (6)$$

其中，$\mathbf{R}_i$、$\mathbf{t}_i$ 分别是第 $i$ 帧算法估计的旋转矩阵和平移向量为物体在相机坐标下旋转矩阵与平移向量，$\mathbf{R}_i^g$、$\mathbf{t}_i^g$ 分别为相机在世界坐标系下旋转矩阵和平移向量，$\mathbf{X}_k$ 是当前位姿 $\mathbf{R}_i$ 和 $\mathbf{t}_i$ 下的三维轮廓点，$\mathbf{X}_h$ 是基准位姿 $\mathbf{R}_i^g$ 和 $\mathbf{t}_i^g$ 下到 $\mathbf{X}_k$ 二维投影点最近的三维轮廓点，C 是当前位姿 $\mathbf{R}_i$ 和 $\mathbf{t}_i$ 下的三维轮廓点数量，**K** 是相机内参矩阵，$\pi$ 将齐次坐标映射为非齐次坐标。

当 AED 的值小于设定阈值 $\alpha$ 时，该帧图像的位姿被认为正确地估计。本数据集上的算法平均边缘距离准确率为正确帧数占总帧数的比例，误差范围为 0~4 个像素。

5.1.2 平均表面距离

平均表面距离是度量三维物体模型估计位姿到基准位姿下三维模型对应点之间距离的平均值。对于第 $i$ 帧图像，ASD 的计算方式为：

$$ASD=\frac{1}{V}\sum_{k=0}^{V-1}\|\left(\mathbf{R}_i\mathbf{X}_k+\mathbf{t}_i\right)-\left(\mathbf{R}_i^g\mathbf{X}_k+\mathbf{t}_i^g\right)\|_2 \quad (7)$$

其中，$\mathbf{R}_i$、$\mathbf{t}_i$ 分别是第 $i$ 帧算法估计的旋转矩阵和平移向量为物体在相机坐标下旋转矩阵与平移向量，$\mathbf{R}_i^g$、$\mathbf{t}_i^g$ 分别为相机在世界坐标系下旋转矩阵和平移向量，$\mathbf{X}_k$ 是三维物体模型上的第 k 个三维点，V 是三维物体模型上所有顶点的数量。

当 ASD 的值小于设定 $\beta$ D 时，该帧图像的位姿被认为正确地估计，其中 $\beta$ 是设定阈值，D 是物体最大直径（物体包围盒边长中的最大值）。本数据集上的算法平均表面距离准确率为正确帧数占总帧数的比例,误差范围为物体最大直径的 0~40%。

5.1.3 重初始化率

在三维跟踪算法跟踪的过程中，可能会出现跟踪失败的情况，为了保证跟踪算法完整测试每个序列，本文设置每种算法在跟踪失败时使用相应帧的基准位姿进行重初始化。同时，重初始化率也反映了算法的鲁棒性和数据集的挑战性，重初始化率越高说明算法鲁棒性越差或数据集挑战性越强。本文中每个序列上的重初始化率计算方式如下：

$$RIR=\frac{R_n}{F_n} \quad (8)$$

其中，$R_n$、$F_n$ 分别为重初始化次数和序列帧数。

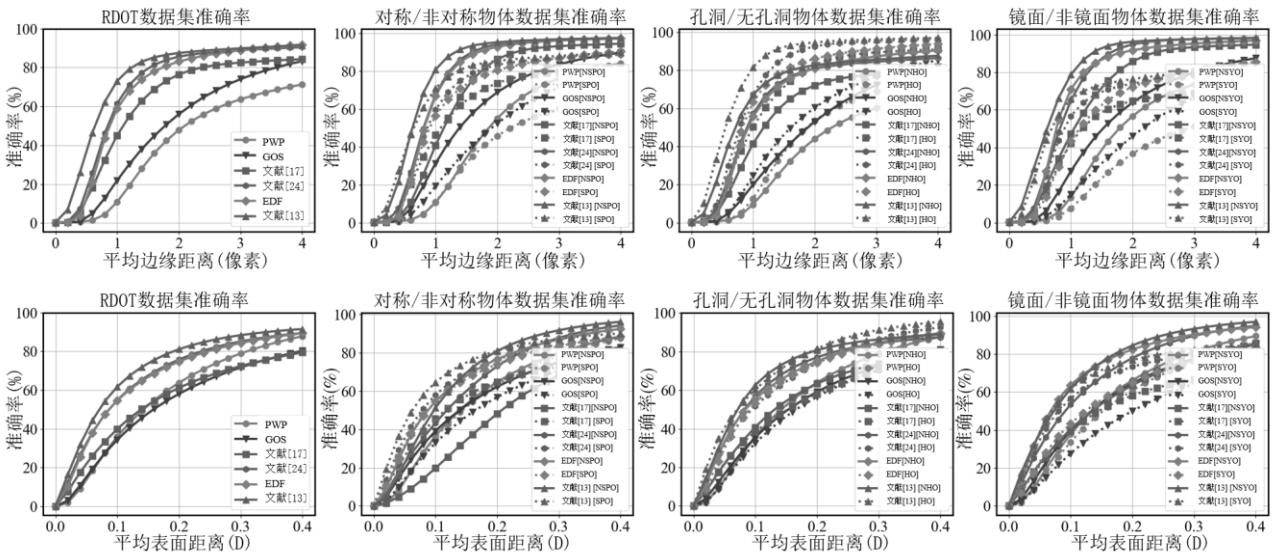
本文设置当算法跟踪失败并启动重初始化时的条件为：算法的预测姿态与基准姿态的旋转角度误差超过 10 度或平移位置误差超过三维模型最大直径的 50%。

### 5.2 算法评测与分析

参与测评的算法均用 C++语言实现，运行环境为 Intel（R）Core（TM）i7-8700CPU（3.2GHz）、8 GB RAM、Nvidia GT 1060 显卡以及 64 位 Windows 10 操作系统，图像的分辨率为 720x480。

5.2.1 整体性能

三维模板跟踪算法在本数据集上的整体性能的评测结果如图 10（1）所示。整体上，不同算法在 AED 误差和 ASD 误差准确率的性能表现相似。文献[13]算法在 AED 和 ASD 误差评估中性能均为最优，误差在 1 像素内的 AED 误差准确率为 73.05%，误差在 0.1 倍物体直径的 ASD 准确率为 61.69%。文献[13]算法使用物体边缘附近区域的局部颜色直方图作为跟踪物体位姿的特征，提取特征的区域较多，提取特征效果较好，因此跟踪性能较优。虽然文献[13]算法已是目前最优算法，但该算法在三维跟踪领域还有很大的提升空间。EDF 算法和文献[24]算法性能次之，表现相当，AED 误差准确率分别为 58.85%和 61.36%，ASD 误差准确率分别为 54.60%和 54.79%。GOS 算法、文献[17]算法和 PWP3D 算法性能表现较差，这三种算法在两种评估方式下准确率均低于 45%。分别来看，文献[17]算法在二维上的性能优于其余两种算法，而 PWP3D 算法在三维上的性能要略优于其余两种算法。




（1） （2） （3） （4）

图 10 三维模板跟踪算法在本数据集上整体性能和不同物体属性性能评测结果。其中，（1）表示算法在基础数据集中的整体准确率；（2）表示对称和非对称物体数据集准确率；（3）表示孔洞和无孔洞物体数据集准确率；（4）表示镜面和非镜面物体数据集准确率；D 表示物体最大直径；NSPO、SPO、NHO、HO、NSYO、SYO 分别表示非对称、对称、无孔洞、有孔洞、非镜面、镜面物体数据集。

6 种三维模板跟踪算法在本数据集中的重初始化率见表 5。整体上看，PWP 算法的重初始化率最低，为 19.3%，说明 PWP 算法的稳定性较好，但是根据上两种精度分析，PWP 算法虽然跟踪目标不易丢失，但跟踪效果较差，精度略低。EDF 算法和文献[13]的重初始化率次之，分别为 20.9%和 22.17%，结合上述两种度量方式可以看出，这两种算法在物体跟踪的稳定性和跟踪精度之间平衡较好，整体性能较优。GOS 算法效果最差，可能时由于该算法未考虑场景出现噪声、遮挡等情况。

**表 5 三维模板跟踪算法在本数据集中的重初始化率**

（BD、NSPO、SPO、NHO、HO、NSYO、SYO、HFL、HNO、HMB、OCC 分别表示非对称、对称、无孔洞、有孔洞、非镜面、镜面物体、严重光照变化、严重噪声、严重运动模糊和遮挡数据集） （单位：%）

| | PWP | GOS | 文献[17] | 文献[24] | EDF | 文献[13] |
|---|---|---|---|---|---|---|
| BD | 16.05 | 32.96 | 24.86 | 17.04 | 14.41 | **13.62** |
| NSPO | 15.48 | 27.85 | 17.71 | 9.63 | 11.10 | **7.17** |
| SPO | 16.22 | 34.42 | 26.88 | 19.16 | **15.36** | 15.46 |
| NHO | 17.57 | 37.70 | 30.90 | 22.57 | **14.67** | 19.15 |
| HO | 17.88 | 38.84 | 20.45 | 11.76 | 16.81 | **8.42** |
| NSYO | 14.00 | 26.82 | 14.85 | 8.45 | 9.94 | **5.77** |
| SYO | 18.63 | 40.63 | 37.32 | 27.79 | **20.00** | 23.43 |
| HFL | 23.63 | 38.78 | 31.29 | 27.48 | **19.78** | 23.39 |
| HNO | 19.06 | 45.53 | 31.34 | 23.46 | **18.17** | 19.90 |
| HMB | **18.02** | 38.72 | 30.46 | 22.22 | 19.74 | 18.43 |
| OCC | **19.74** | 53.41 | 81.59 | 42.84 | 32.44 | 35.50 |
| Avg. | **19.3** | 41.88 | 39.91 | 26.60 | 20.90 | 22.17 |

表 6 给出了三维模板跟踪算法在本数据集中的平均速度对比。其中，EDF 算法时间性能最优，平均速度为 35.80FPS，文献[13]算法次之。这是因为 EDF 算法使用边缘特征提取较快，而文献[13]算法提取特征的区域较多，位姿优化的时间也相对较长。文献[17] 速度最慢，因其使用全局颜色直方图，特征计算时间较长。该 6 种三维模板跟踪算法的处理速度与目前移动端增强现实应用所需的强实时性（平均速度达到 60FPS）仍存在一定差距，后续新的算法应关注在不损失跟踪精度的前提下提高处理速度的方法。

**表 6 三维模板跟踪算法在本数据集中的平均速度**

（单位：FPS）

| | PWP | GOS | 文献[17] | 文献[24] | EDF | 文献[13] |
|---|---|---|---|---|---|---|
| BD | 28.74 | 23.82 | 17.69 | 28.28 | **35.97** | 28.4 |
| HFL | 28.61 | 24.23 | 17.32 | 27.32 | **36.29** | 28.01 |
| HNO | 29.36 | 19.42 | 16.9 | 27.38 | **32.88** | 27.19 |
| HMB | 29.65 | 24.27 | 16.78 | 27.52 | **36.92** | 27.63 |
| OCC | 28.74 | 22.8 | 16.56 | 27.24 | **35.83** | 27.32 |
| Avg. | 29.15 | 23.23 | 17.01 | 27.47 | **35.80** | 27.56 |

### 5.2.2 不同物体属性下的性能

不同的三维物体的几何结构和表面材质能渲

染出不同的成像效果，因此是影响三维模板跟踪算法进行图像特征提取的关键因素，进而影响算法跟踪结果。为了评估 6 种三维模板跟踪算法对不同物体属性的物体的跟踪结果，本文按照表 3 的信息对数据集进行了划分，不同物体属性下三维模板跟踪算法的测评结果如图 9（2-4）所示，重初始化率性能如表 5 所示。

整体上，6 种三维模板跟踪算法在对称物体数据集、非孔洞物体数据集、镜面物体数据集和半透明物体数据集的准确率低于其在非对称物体数据集、孔洞物体数据集、非镜面物体数据集结果。RBOT 算法在不同物体属性数据集中的性能均是最优的，EDF 算法和文献[24]次之。

对称物体的边缘在图像上展现出对称性，会造成算法求解出歧义的位姿，算法在对称物体数据集中的准确率较非对称物体准确率下降约 10 个百分点；非孔洞物体相比于孔洞物体缺乏内部边缘，因此可供算法跟踪的特征较少，故算法在非孔洞物体中表现较差，准确率比孔洞物体低 10-20%。镜面物体相比于非镜面物体存在高光区域，对于算法提取边缘特征或颜色直方图特征造成干扰较大，对基于边缘的 EDF 算法准确率影响超过 20%，文献[13]重初始化率相差近 20 个百分点。半透明物体的颜色根据其所在背景颜色而变化，对于基于区域颜色信息的算法特征影响较大，难以提取到有效的特征信息，具体性能如表 7 所示。基于区域的 PWP 算法、文献[13][17][24]性能远低于平均性能，文献[17]的重初始化率高达 90%。后续算法应该重点研究镜面、半透明和对称物体的稳定特征提取方法。

**表 7　三维模板跟踪算法在半透明物体中的性能**

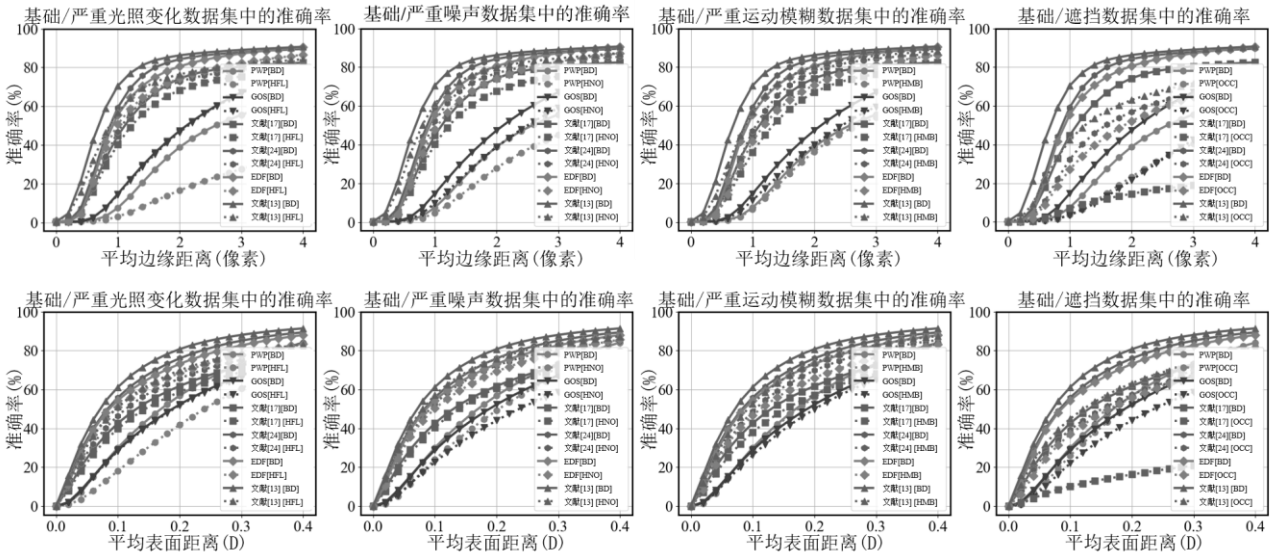
**（1 至 3 行分别表示误差在 1 像素的 AED 准确率、误差在 0.1 倍物体直径的 ASD 准确率和重初始化率**　　（单位：%）

| | PWP | GOS | 文献[17] | 文献[24] | EDF | 文献[13] |
|---|---|---|---|---|---|---|
| AED | 2.46 | 6.22 | 3.46 | 6.56 | **30.46** | 9.44 |
| ASD | 16.92 | 17.92 | 6.40 | 9.54 | **33.62** | 10.14 |
| RIR | 31.29 | 67.17 | 90.14 | 83.62 | **39.75** | 80.88 |

### 5.2.3　不同场景设置下的性能

不同的场景设置直接影响生成的数据集的图像质量，本数据集模拟了现实世界中可能出现的情况，如运动模糊、噪声等，可在不同程度上分析影响三维模板跟踪算法的原因。

为了评估不同场景设置对三维模板跟踪算法的跟踪结果的影响，图 11 展示了不同场景设置下 6 种三维模板跟踪算法的测评结果，表 5 展示了算法的重初始化率性能。为了更好展现测评结果，本文按照场景设置选择了基础数据集和严重难度的场景设置数据集进行展示。相比于基础数据集上的准确率，不同设置的数据集对算法都有一定的影响，准确率在一定程度降低，其中遮挡和严重光照变化对算法的准确率影响较大。




（1）　（2）　（3）　（4）

图 11 三维模板跟踪算法在本数据集不同场景设置中的评测结果。其中，（1）表示算法在基础和严重光照变化数据集准确率；（2）表示基础和严重噪声数据集准确率；（3）表示基础和严重运动模糊数据集准确率；（4）表示基础和遮挡数据集准确率；D 表示物体最大直径；BD 表示 RDOT 基准数据子集；HFL、HNO、HMB、OCC 分别表示严重光照变化、严重

噪声、严重运动模糊、遮挡数据集。

文献[13]在不同场景设置中 AED 和 ASD 误差性能表现最好，EDF 算法与文献[24]算法次之；除 PWP 算法外，EDF 算法的平均重初始化率最低，为 22.53%。本数据集中噪声对算法影响较小，原因是本数据集是高质量序列，噪声强度有限。光照和运动模糊会影响算法特征提取的准确性，在遮挡数据集中，由于物体在图像序列中始终存在遮挡，算法难以提取完整的图像特征，因而准确率最低。

在实际场景中，当光照变化、噪声、运动模糊和遮挡出现时，三维跟踪算法往往会跟踪失败，当场景设置恢复正常，三维跟踪算法应能够快速从失败中恢复，建议后续算法可以进一步探索三维跟踪算法失败后的错误恢复方法，保证算法能够完整地跟踪。

### 5.3 误差修正

本文将目前两个最优算法（EDF 算法和文献[13]算法）的位姿误差进行分析，统计出算法位姿与基准位姿的误差分布情况，如表 8 所示。对于算法的误差数据结果统计发现，EDF 算法和文献[13]算法在三个位移上的位姿误差存在较大差异，其中 z 轴方向上的误差超过 x 轴、y 轴方向的 5 倍。利用公式（3）-（5）可对算法进行相机视线方向的修正，修正结果如表 9 所示，修正结果可视化效果如图 12 所示。

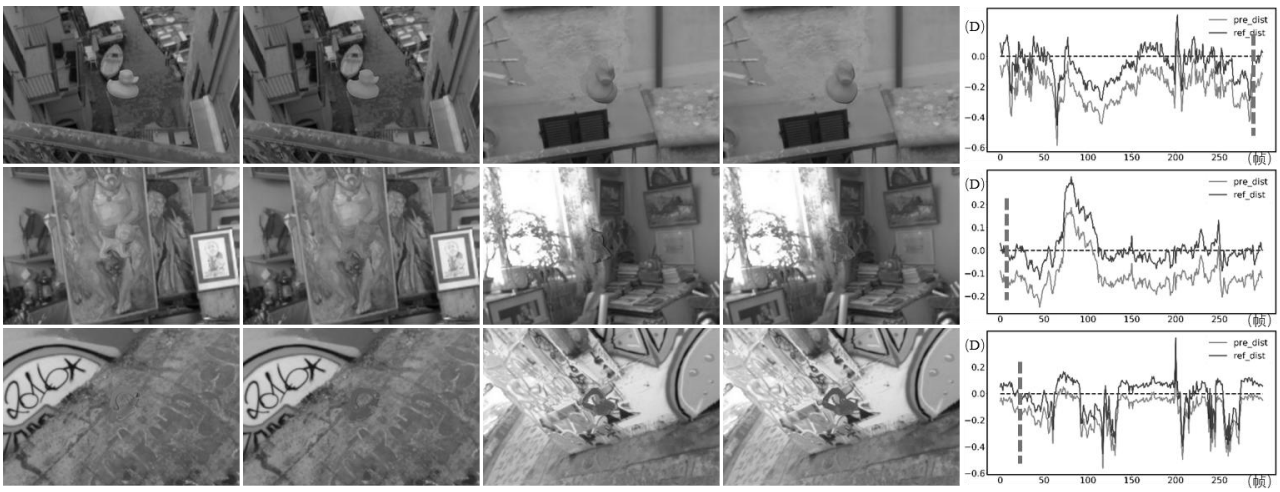
**表 8 EDF 算法和文献[13]算法的误差分布** （单位：$10^{-2}$）

| | μx | μy | μz | σx | σy | σz |
|---|---|---|---|---|---|---|
| EDF | 1.76 | 1.69 | 18.8 | 3.78 | 3.71 | 38.33 |
| 文献[13] | 6.58 | 7.34 | 45.58 | 347.79 | 470.52 | 1579.31 |

本文建立不同物体训练集，对不同算法进行了误差因子计算。实验结果显示，利用该修正算法，EDF 算法与文献[13]算法在平均表面距离误差上有了一定程度减小，其中利用 bunny 进行误差因子计算修正后的 EDF 算法平均误差减小了接近 10 个百分点。该修正算法对于物体的距离、物体表面形状、运动路径等较为敏感，对于部分模型优化效果较好，如 bunny，平均误差均减小了 10 多个百分点，但是对于部分模型来说效果欠佳，如 hairdryer，故不用对所有物体进行修正。但从整体上来说，修正补偿能在一定范围内对算法姿态进行优化，具体原因有待进一步研究。

**表 9 EDF 算法和文献[13]算法修正后误差在 0.1 倍物体直径的 ASD 准确率变化** （单位：%）

| 算法 | 计算误差因子所用物体 | airplane | bottom | bunny | cat | duck | horse | lock | hairdryer | seashell | truck | Agv. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EDF | bunny | 8.74 | 3.49 | 15.44 | 9.65 | 3.61 | 8.78 | 12.21 | 20.98 | 9.30 | 6.71 | 9.89 |
| | cat | 4.30 | -6.76 | 3.42 | -10.95 | -16.26 | 7.95 | -1.76 | 30.24 | -4.55 | 2.47 | 0.81 |
| | teapot | 8.81 | 3.07 | 15.44 | 8.57 | 2.49 | 9.11 | 11.64 | 22.16 | 8.46 | 6.78 | 9.65 |
| 文献[13] | bunny | -14.54 | 17.99 | 21.14 | 33.87 | 24.38 | 8.16 | 9.14 | -12.04 | 20.93 | 1.71 | 11.07 |
| | cat | -5.25 | 17.08 | 22.26 | 36.51 | 27.39 | 8.25 | 22.65 | -1.58 | 20.41 | 4.86 | 15.26 |
| | teapot | -0.56 | 14.12 | 19.17 | 28.91 | 22.34 | 7.38 | 23.13 | 2.06 | 16.31 | 5.17 | 13.80 |




（1） （2） （3） （4） （5）

图 12 文献[13]算法在本数据集中修正前后位姿投影图可视化。其中，（1）表示本数据集图像序列，（2）表示文献[13]算法跟踪位姿投影图，（3）表示文献[13]算法跟踪位姿在另一视角投影图，（4）表示文献[13]算法修正后位姿投影图，算法在该序列中时序上与基准位姿的距离变化，浅灰表示修正前的位姿与基准位姿的距离，深灰表示修正后的位姿与基准位姿的距离，竖线表示投影图所在帧数，D 表示物体的最大直径）

### 5.4 与现有数据集的区别和联系

本文数据集是用于三维模板跟踪算法测评的合成数据集，与现有数据集有以下三个方面的区别。

首先，本文采用的基于图像的光照渲染技术，是一种基于物理的渲染方法，其得到的渲染图片能有效的反映真实世界光照条件，而且能更好的表现特殊材质（如半透明、镜面材质等）效果和运动模糊效果。文献[13]建立的渲染数据集是将渲染的前景与实拍背景视频合成而得到的，前背景光照不一致，真实感较差。图 13 对比展示了两者渲染图像的效果。

图 13 RBOT 数据集与本数据集渲染效果对比（左：文献[13];右:本数据集）

其次，本数据集的构成更为丰富。在数据集设置上，本文选择了非镜面、镜面、半透明和弱纹理四类物体材质，还建立了运动模糊数据集，而文献[13]仅考虑了非镜面和弱纹理物体，镜面、半透明物体和运动模糊是文献[13]未考虑的，恰恰也是现实世界中普遍存在的，而且目前三维跟踪算法不能鲁棒处理的；在场景属性设置上，本文数据集中每个序列都具备完全不同的背景、光照参数和运动路径参数，如图 1，而文献[13]的序列除了前景物体不同，均使用相同的实拍背景视频、运动路径和光照条件，构成较为单一。总体上，本数据集的规模是文献[13]数据集的 20 余倍，其它设置如物体数量、运动路径数量等数据量均更为丰富。

最后，本文根据物体的属性和场景设置不同对现有 6 种三维跟踪算法进行了多维度评估，全面分析了利用不同图像特征进行跟踪的局限性和当前三维模板跟踪算法的优劣，而文献[11]和文献[13]仅评测了两种不同的三维模板跟踪算法，同时未考虑特殊物体属性和场景设置对三维模板跟踪算法鲁棒性的影响。

为了验证本数据集的挑战性和有效性，本文利用文献[13]算法在现有最大数据集 RBOT 数据集与本数据集进行了评估，表 10 对比展示了本数据集和 RBOT 数据集上性能测评。

**表 10 文献[13]算法在 RBOT 数据集和本增强数据集中的重初始化率对比（FL、NO、MB、OCC 分别表示光照变化、噪声、运动模糊和遮挡数据集）** （单位：%）

| | FL | NO | MB | OCC |
|---|---|---|---|---|
| RBOT | 14.45 | **20.07** | -- | 21.11 |
| Ours | **23.39** | 19.90 | **18.43** | **35.50** |

在增强数据集中，文献[13]算法在本数据集上的平均重初始化率比其在RBOT数据集上的平均重初始化率高，反映了本数据集对测评算法更有挑战性。具体来看，文献[13]算法在本数据集中的光照变化和遮挡数据集上均超过 RBOT 数据集，高出约 10 个百分点，说明本数据集中光照变化和遮挡对测评算法的难度更大，还建立了运动模糊数据集，有效地填补了 RBOT 数据集缺失的部分。

本文还将不同算法在RBOT数据集和本数据集中的时间性能进行对比，如表 11 所示。时间性能受图像分辨率、物体的大小、距离相机的远近等多个因素相关。EDF 算法在两个数据集中性能表现均最优，最快可达 44FPS。GOS 算法、文献[13] [17][24]算法在本数据集中的平均速度略快于 RBOT 数据集，而 PWP 算法和 EDF 算法则相反。

**表 11 三维模板跟踪算法在本数据集和 RBOT 数据集中的平均速度对比** （单位：FPS）

| | PWP | GOS | 文献[17] | 文献[24] | EDF | 文献[13] |
|---|---|---|---|---|---|---|
| RBOT | 35.16 | 17.61 | 16.63 | 26.84 | **44.14** | 25.68 |
| Ours | 29.15 | 23.23 | 17.01 | 27.47 | **35.80** | 27.56 |

## 6 结语

本文针对三维模板跟踪领域建立了一个大型高真实感的数据集，并对当前主流的三维模板跟踪算法进行测评。本数据集对场景的真实性和复杂程度进行了定量合理的控制，共生成了具有真实位姿参数的 5342 个图像序列，1603800 帧图像，是目前最大数据集的 20 多倍。本数据集为理解三维模板跟踪算法的局限性提供帮助，也为有效量化评估三维模板跟踪算法提供数据支持。本文还建立了误差分

析模型能提高算法跟踪结果的精度。但本数据集仍然有改进的空间。采用真实感渲染技术生成的数据集与实拍数据相比，图像真实感仍存在一定差距；尚未生成有纹理物体的相关数据；本数据目前仅针对三维模板跟踪算法，帧间位姿具有时序上的连续性，可进一步扩展用于三维物体检测和位姿估计算法。未来工作将对上述局限性进行改进，最终得到一个更为丰富的数据集。

## 参 考 文 献


[1] Hinterstoisser S, Lepetit V, Ilic S, et al. 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, Korea, 2012: 548–562

[2] Huang C H, Boyer E, Navab N, et al. Human Shape and Pose Tracking Using Keyframes//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 3446–3453

[3] Newcombe R A, Izadi S, Hilliges O, et al. KinectFusion: Real-Time Dense Surface Mapping and Tracking//Proceedings of the 10th IEEE/ACM International Symposium on Mixed and Augmented Reality. Basel, Switzerland, 2011: 127–136

[4] Lowe D G. Robust model-based motion tracking through the integration of search and estimation. International Journal of Computer Vision, 1992, 8(2): 113-122

[5] Harris C, Stennett C. RAPID-a video rate object tracker //Proceedings of British Machine Vision Conference. Oxford, UK, 1990: 1-6

[6] Huang Hong, Zhong Fan, Qin Xueying. Textureless 3D Object Tracking Based on Adaptive Feature Fusio. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(5): 833-841) (in Chinese)
(黄鸿, 钟凡, 秦学英, 基于自适应特征融合的无纹理 3D 目标跟踪. 计算机辅助设计与图形学学报, 2018, 30(5): 833-841)

[7] Wang B, Zhong F, Qin X. Pose optimization in edge distance field for textureless 3D object tracking//Proceedings of the Computer Graphics International Conference. Yokohama, Japan, 2017: 32:1-32:6

[8] Kass M, Witkin A, Terzopoulos D. Snakes: Active contour models. International Journal of Computer Vision, 1988, 1(4): 321–331

[9] Prisacariu V A, Reid I D. PWP3D: Real-time segmentation and tracking of 3D objects. International Journal of Computer Vision, 2012, 98(3):335–354

[10] Hexner J, Hagege R R. 2D-3D pose estimation of heterogene-ous objects using a region based approach. International Journal of Computer Vision, 2016, 118(1): 95-112

[11] Wu P C, Lee Y Y, Tseng H Y, et al. A Benchmark Dataset for 6DoF Object Pose Tracking//Proceedings of the IEEE/ACM International Symposium on Mixed and Augment-ed Reality (ISMAR-Adjunct). Los Alamitos,USA: IEEE Computer Society Press, 2017: 186-191

[12] Changhyun C, Henrik I. C. RGB-D object tracking: A particle filter approach on GPU[C] //Proceedings of the IEEE Confer-ence on International Conference on Intelligent Robots and Systems. Nantes, France, 2013: 1084-1091

[13] Tjaden H, Schwanecke U, Schömer E, et al. A region-based gauss-newton approach to real-time monocular multiple object tracking. IEEE Transactions on Pattern Analysis and Ma-chine Intelligence, 2019, 41(8): 1797-1812

[14] Wang Bin. Research on Monocular Textureless 3D Object Tracking[Ph.D. thesis]. Shandong: Shandong University
(王斌. 单目无纹理三维物体跟踪研究[博士论文]. 山东大学计算机科学与技术系, 济南，2019)

[15] Besl P J, Mckay N D. A method for registration of 3-D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2):239-256

[16] Wang G F, Wang B, Zhong F, et al. Global optimal searching for textureless 3D object tracking. The Visual Computer, 2015, 31(6-8): 979-988

[17] Tjaden H, Schwanecke U, Schmer E. Real-Time Monocular Segmentation and Pose Tracking of Multiple Objects// Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 423-438

[18] Kehl W, Manhardt F, Tombari F, et al. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1521-1529

[19] Sida P, Yuan L, Qixing H, et al. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4561-4570

[20] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248-255

[21] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2411-2418

[22] Tejani A, Tang D , Kouskouridas R , et al. Latent-Class Hough Forests for 3D Object Detection and Pose Estimation//Proceedings of European Conference on Computer Vision. Zurich, Switzerland：Springer,2014: 462-477

[23] Pauwels K , Rubio L , Diaz J , et al. Real-Time Model-Based Rigid Object Pose Estimation and Tracking Combining Dense and Sparse Visual Cues//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2347-2354

[24] Tjaden H, Schwanecke U, Schomer E. Real-time monocular pose estimation of 3D objects using temporally consistent local color histograms//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 124-132.

**HE Xian**，M. S. candidate. Her research interests include augmented reality and computer vision.

**LI Jia-Chen**，Ph. D. candidate. His research interests include augmented reality.

**JIN Li**，M. S. candidate. His research interests include augmented reality.

**LIU Li**，M. S. His research interests include computer vision.

**ZHONG Fan**，Ph. D. , associate professor. His research interests include video processing and computer vision.

**QIN Xue-Ying**，Ph. D. , professor. Her research interests include video processing and computer vision.

**Background**

3D template tracking is one of the most important tasks in computer vision. As one of the fundamental technique, 3D template tracking has a wide application in augmented reality, the human-computer interaction and robotics, etc. In recent years, many 3D template tracking methods are proposed, with better accuracy and robustness evaluated on different datasets. However, previous datasets are of limited scale and low quality, and it is difficult to comprehensively and objectively evaluate the tracking approaches due to the lacking of uniform dataset and standard metrics.

Currently, several works have constructed 3D template tracking datasets, including both video-captured datasets and computer-rendered datasets. However, the ground-truth pose of the video-captured datasets is inaccurate, since the pose are obtained by the calibration method of hand-crafted markers. The scale of these datasets is also limited due to the time-consuming labelling process. There are also some problems in the computer-rendered datasets, such as unreal pictures and insufficient settings of rendering scenes.

In this work, we construct a large-scale 3D template tracking benchmark RDOT based on photorealistic rendering, which effectively addresses the problems of inaccurate ground-truth and unrealistic images in previous datasets. It includes tens of objects with different structures and materials, and allows cameras and objects to move in complex motion modes, which increases the difficulty of the benchmark. RDOT also considers different settings of noise, motion blur, illumination change and occlusion. RDOT simulates the real lightings in order to evaluate the performance of the leading 3d tracking approaches under various conditions. Three metrics are used to evaluate the current tracking approaches, resulting in a comprehensive analysis and perspective of 3d template tracking.

This work is supported by Industrial Internet Innovation and Development Project in 2019 of China，Zhejiang Lab（No. 2020NB0AB02），the National Natural Science Foundation of China（No. 61907026）and a Project of Shandong Province Higher Educational Science and Technology Program（No. J18KA392）。