**研究方向科研报告**

姓名： 王富新

拟选取方向：基于RGB-D的高反光物体6D位姿估计算法研究与应用

## 研究背景与意义

### 1.1 研究背景

物体的6D位姿估计是计算机视觉领域的一个关键任务，它和许多快速发展的技术领域密切相关，如自动驾驶[1]、机器人技术[2]和增强现实[3]。6D位姿估计是物体坐标系和与视觉传感器（相机）坐标系之间的刚体坐标变换，由3-DoF平移变换（3-DoF Translation）与3-DoF旋转变换（3-DoF Rotation）组成，简称6-DoF（Degree of Freedom）位姿[4]。

物体的6D位姿估计的发展历程中视觉传感器起到了决定性的作用。在研究初期，视觉传感器捕获得到的数据以RGB数据为主，大多数工作在RGB图像和CAD模型之间（2D-3D）的特征点或者模板匹配，建立RGB数据和CAD模型的对应关系，基于对应关系预测出物体位姿。但是由于基于RGB的方法高度依赖于RGB图像中纹理丰富程度且受光照、噪声等因素影响，导致基于RGB数据的方法精度有限。随着深度相机的发展，获取RGB-D数据更加容易且成本低，基于RGB-D的6D位姿估计的深度学习模型逐渐出现，由于RGB-D数据包含额外的深度信息，使得模型在低纹理、光照变化较大的情况下仍然具有较高的精度。

尽管近年来6D位姿估计相关的研究和应用已经取得了较大的进展，但在实际应用中，特别是在工业领域的抓取、拣选等应用上，仍然存在一些亟待解决的具有挑战性的问题，例如当应用场景中出现光照变化较大、物体之间相互遮挡或者截断、场景中的目标对象纹理较少、目标对象对称等情况时，预测结果的准确性和鲁棒性会降低；另外由于深度学习模型高度依赖于数据集，在相应场景的数据集匮乏时，模型的准确度也会极大的降低。

### 1.2 研究意义

## 国内外研究现状与存在的问题

### 2.1 研究现状

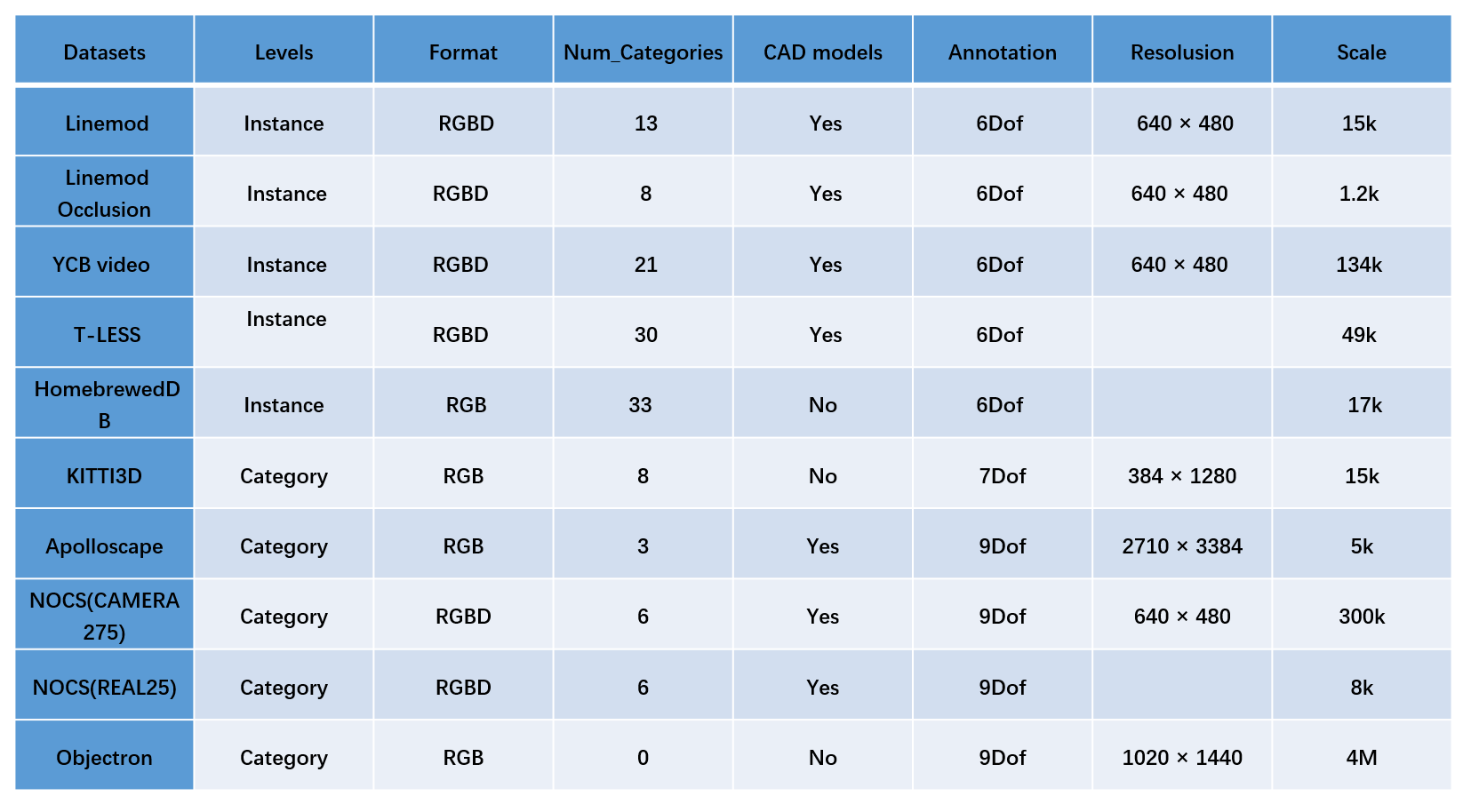
目前物体的姿态估计任务可以分为两大类，分别是实例级和类别级的物体姿态估计[5]。实例级的物体姿态估计是给定图像（RGB/RGBD）和CAD模型，恢复图像中目标物体在相机坐标系下的3D平移和3D旋转信息，目标物体由已知物体的集合构成。类别级的物体位姿估计是给定图像（RGB/RGBD）恢复物体在相机坐标系下的3D平移、3D旋转和物体的3D尺寸信息（物体的长宽高），目标物体由已知类别的多个物体实例组成。

**2.1.1物体的6D位姿估计数据集**

6D位姿估计任务中，可用于训练神经网络的数据可以分为真实数据和合成数据。在计算机视觉领域中，真实数据使用视觉传感器捕获真实场景，然后对图像进行人工标注；合成数据通过专业的三维图形软件，使用虚拟相机渲染对象或场景，且无需进行手动标注。此外，合成数据还可以进一步分为具有高质量照片级物理渲染

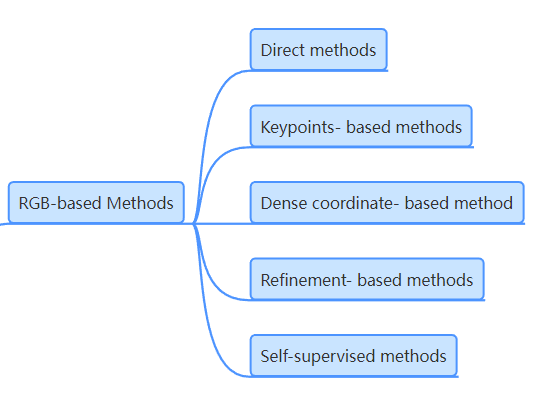
（Physically Based Rendering，PBR）数据和域随机化（Domain Randomization，DR）[6][7]数据。域随机化的核心思想是通过对随机物体姿势进行抽样并将模型放在随机背景图像上产生图像。这样一来，真实世界的领域只是生成领域的一个子集，这样模型就可以通过利用这些合成图像来学习尽可能多的6D姿势相关特征。

目前用于6D位姿估计的常见的数据集有Linedmod[9]、YCB video[10]、T-LESS[11]等。根据物体的CAD模型是否可用以及每个类别是否存在多实例，将它们的分为实例级的物体姿态估计和类别级的物体姿态估计。表xx是各个常用数据集之间的比较。



**2.1.2基于RGB的实例级物体位姿估计**

基于RGB的深度学习的方法可以分为五大类，分别是直接方法、基于关键点的方法、基于密集对应的方法、基于细化的方法和自我监督的方法。



直接方法是指将物体姿势估计作为一项回归或分类任务，并直接从输入图像中预测物体姿势相关的参数（例如，欧拉角或旋转的四元数）。

|  |  |  |
| --- | --- | --- |
|  | 主要阅读文献 | 创新点 |
| 1 | Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes.[12] （Arxiv 2017） | 利用图像坐标来推测实际3D坐标），并且通过hough投票来确定物体位置中心；提出新的姿态估计损失函数ShapeMatch-Loss：解决旋转对称物体姿态估计问题；提出了YCB-Video数据集。 |
| 2 | Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again.[13]（IEEE 2017） | 分解模型姿态空间， 便于对称的训练和处理；扩展SSD， 产生2D检测并推断合适的6D位姿。 |
| 3 | EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. [14]（Arxiv 2020） | 网络由EfficientDet扩展而来，具有高度准确、高效和可扩展性。提出了一种用于6D位姿估计数据集的增强方法，称为6D增强，可以显著提高网络的性能和泛化能力。 |

基于关键点的方法使用CNN来检测图像中的二维关键点，然后解决Perspective-n-Point（PnP）[15]问题进行姿势估计。

|  |  |  |
| --- | --- | --- |
|  | 主要阅读文献 | 创新点 |
| 1 | YOLO6D Real-Time Seamless Single Shot 6D Object Pose Prediction.[16]（CVPR2018） | 提出了一个新的CNN结构，直接预测目标的射影顶点的2D图像位置，然后用PnP算法估计目标的3D位姿。（不需要多阶段、假设检验、不需要精确目标纹理模型、精度足够不用后期处理） |
| 2 | Pvnet: Pixel-wise voting network for 6dof pose estimation.[17]（CVPR 2019） | 使用了一种逐像素投票网络(PVNet)来回归逐像素的指向关键点的向量，并且用这些向量给关键点位置投票，能更好地定位遮挡和截断关键点。创造了关键点位置的不确定性机制，能更好的利用PnP。 |

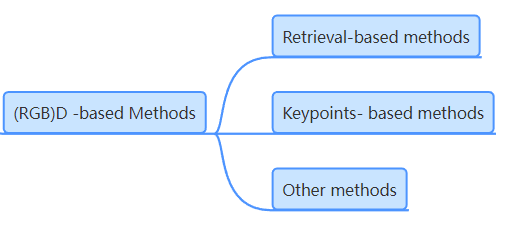
基于密集坐标的方法是将6D的物体位置估计任务表述为建立密集的2D-3D对应关系，然后用PnP解算器恢复物体位置。密集的2D-3D对应关系是通过预测每个物体像素的3D物体坐标或预测密集的UV图获得的。基于密集坐标的方法对严重的遮挡和对称有很强的鲁棒性，然而，由于连续搜索空间较大，回归物体坐标比预测稀疏关键点更难。

基于细化的方法指的是使用一个步骤来提高模型的性能，一般是通过最小化当前位姿下的观察图像和渲染图像之间的差异来反复优化位姿，并将优化后的位姿作为下一次迭代的初始位姿。基于细化的方法的速度在很大程度上取决于迭代的数量和所使用的渲染器，这也成了此种方法的瓶颈。

自监督的方法，在6D的物体姿势检测任务中，目前的深度学习模型高度依赖于对有注释的真实世界数据的训练，而这些数据是很难得到的。一个直观的方法是使用无成本的合成数据进行训练。然而，许多工作已经证明，由于合成数据和自然数据之间存在巨大的领域差距，单纯在合成数据上训练的模型对真实世界场景的概括能力很差。基于自监督的方式可以产生新的数据用于自我监督学习，并以终身学习的方式改进其姿势估计结果。

**（3）基于RGB-D的实例级物体位姿估计**

基于RGB-D的方法将RGB-D图像或深度掩码作为输入，并通过充分利用点云来预测物体的姿势。一般来说，基于RGB-D的方法可以分为基于检索的方法、基于关键点的方法和其他基于深度学习的方法。



基于检索的方法首先使用CAD模型生成一个涵盖所有可能的物体姿势的RGB-D图像数据集，然后我们可以从数据集中检索目标物体的最相似的图像，以确定给定观察到的RGB-D图像或深度掩码的物体姿势，检索可以通过比较图像描述符或匹配模板来实现，基于检索的方法在大多数情况下可以达到稳健的性能，但是这种方法需要离散旋转空间来定义编码本，当离散区间较大时，会造成粗略的预测，而当离散区间较小时，则会造成计算效率的下降。

基于关键点的方法与基于RGB的方法类似，在基于RGB-D的方法中，也有一系列试图通过提取物体关键点来预测物体姿势的工作，这种方法总是构建3D-3D对应关系，利用预测的关键点来解决物体姿势。

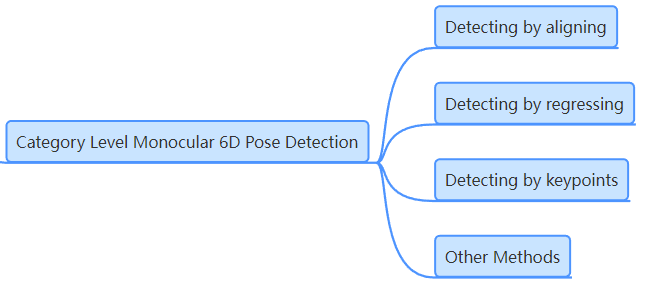
|  |  |  |
| --- | --- | --- |
|  | 主要阅读文献 | 创新点 |
| 1 | Ffb6d: A full flow bidirectional fusion network for 6d pose estimation.[18]（CVPR 2021） | 提出了一种新的关键点选择算法SIFT-FPS，取代了FPS，充分利用纹理和几何信息。 |

除了上面提到的方法外，还有一类其他方法，这类方法不追求精度，而是解决一些特定的问题，比如由复杂场景引起的遮挡和杂波、实时性问题、计算成本问题等等。

|  |  |  |
| --- | --- | --- |
|  | 主要阅读文献 | 创新点 |
| 1 | Densefusion: 6d object pose estimation by iterative dense fusion.[19]（CVPR 2019） | 提出了能结合颜色和深度信息的方法；不需要ICP的端到端的微调网络。 |

**（4）类别级物体位姿估计**

类别级6D位姿估计任务中，我们需要预测9个自由度（9Dof），包括3D平移、3D旋转和物体的3D尺寸信息，严格来说，应该将这个任务命名为9D物体姿态估计。解决类别级位姿估计任务的方法可以分为四类，分别是基于对齐、基于回归、基于关键点和其他方法。



基于对齐的方式。Wang等人为一个类别中所有可能的物体情况引入了一个共享的典型表示，名为归一化物体坐标空间（NOCS），严格来说，这是第一个用于类别级单眼6D位姿估计的深度学习方法。

|  |  |  |
| --- | --- | --- |
|  | 主要阅读文献 | 创新点 |
| 1 | Normalized object coordinate space for category-level 6d object pose and size estimation.[20]（CVPR 2019） | 提出了一个共享的典型表示NOCS（归一化物体坐标空间）；提出了了一种新的上下文感知技术来生成大量的完整的混合现实数据，发布了NOCS数据集 |

### 2.2 存在的问题

在数据集方面，标记物体的6D姿态数据是非常困难的，目前可以使用Blender、Unity等生成合成数据集，但是在合成数据集上训练的模型在真实世界的图像上通常表现不佳，所以提高在合成数据集上训练的模型的概括能力成了一大难题。对于类别级的物体姿态估计任务，目前主要使用的是NOCS[8]数据集，该数据集由合成的CAMERA275数据集和真实世界的REAL25数据集组成。目前，NOCS数据集中同一类别的实例之间的类内变化仍然十分有限，即实例不够丰富，这可能会导致训练后的模型难以估计同一类别外形变化很大的物体的位姿。

在实时性方面，尽管现有的方法经常使用大型的骨架如ResNet-101来提取特征，以确保高精确度和有效性，但它们同时也降低了效率，再加上比较耗时的检测过程，6D位姿估计架构很难保证其实时性。

## 研究路线

主要有3个研究点，依次针对于：合成数据集、引入注意力机制、搭建轻量化和高实时性网络，最后综合以上三点实现基于MR的位姿估计系统。

### 3.1 研究内容一：高反光物体数据集制作方法

目前的6D位姿估计深度学习模型高度依赖于有注释的真实世界的数据，但是这些数据往往是难以得到的，这个问题的解决方法是使用自动化工具生成合成数据对模型进行训练，但是由于合成数据和自然数据之间存在巨大的域间隙，单纯在合成数据上训练的模型对真实世界场景的概括能力很差。

真实+虚拟。

本文将研究现有用于6D位姿估计的数据采集和制作方法，归纳各种方法的优劣之处，从6D位姿估计相关公开数据集中结合实际需求总结数据集的标准，结合Unity合成数据的开源工具Perception 1.0（自带深度标注器，使用深度信息估计物体位姿精度更高），快速构建高质量合成数据集，同时研究数据增强对实验的影响，以选用合适的数据增强方法增加模型的鲁棒性。

### 3.2 研究内容二：基于xxx抗遮挡

由于RGB图像缺乏深度信息，所以使用RGB图像估计物体的姿态精度往往不够高，这是因为模型需要具有深度感知的能力。RGB-D深度相机的出现推动了物体的6D位姿估计任务的发展，基于RGB-D的6D位姿估计方法将RGB-D图像或深度掩码作为输入，并充分利用点云来物体的姿势。

本文将研究在网络中引入注意力机制，有效利用隐藏在图像中的非局部特征，高效地提取和融合颜色信息和几何信息，提高模型的深度感知能力，进而提高预测的精度。

引入注意力机制能够在严重遮挡的情况下恢复物体的姿态，且鲁棒性很高。

### 3.3 研究内容三：基于改进的EfficientPose的高实时性位姿估计网络

在实时性方面，尽管现有的方法经常使用大型的骨架如ResNet-101来提取特征，以确保高精确度和有效性，但它们同时也降低了效率，再加上比较耗时的检测过程，6D位姿估计架构很难保证其实时性。

EfficientPose基于EfficientDet进行了改进，是一种高度可扩展的端到端6D物体姿态估计方法，它保持了后者的优势并降低了计算开销，最终使得位姿估计的时间和物体的数量几乎无关。本文将研究并改进该网络，提出一个新的轻量化且高实时性的网络，使得新的网络能够适用于MR

### 3.4 研究内容四：高实时性位姿估计MR应用设计与实现

在MR应用中，常常需要获取视野中的平面（如桌面、墙面等）、物体的空间信息，目前已经成熟的技术是空间感知技术，但是空间感知技术需要预先扫描场景，将场景的空间信息存储到本地，如果场景中的物体位置发生改变，应用将无法察觉到空间已经发生改变。将6D位姿应用到MR应用中能够顺利解决这个问题。

本文将设计并实现一个MR应用，利用Socket通信结合xxx，测试位姿估计的实时性。

## 预期成果

使用Unity Precepton制作高质量合成数据集，并使用数据增强方法提升数据集的概括能力，最终得到数据集并产生论文；

设计并实现基于Preception实现的6D位姿估计数据集自动生成系统，产生相关软件著作权或专利；

调整EfficientPose参数与结构优化部分，将完成神经网络模型并产生论文；

将研究的位姿估计算法应用到MR端的位姿估计中，形成软件著作权或者专利。

## 短期计划

**表x 文献特点表**

|  |  |  |
| --- | --- | --- |
| 日期 | 工作内容 | 预期成果 |
| 2022.01-2022.02 | 学习Unity Perception，阅读数据增强相关论文 | 数据增强方法Demo，基于Preception的制作合成数据集的Demo |
| 2022.03-2022.04 | 学习注意力机制，并尝试将注意力机制融入到特征提取网络中 | 含有注意力机制的特征提取网络 |
| 2022.05-2022.06 | 精读EfficientPose等轻量化、高实时性论文并复现，提出改进想法 | 复现Demo，相关论文的总结 |
| 2022.07-2022.08 | 进行初步改进EfficientPose的实验 | 改进的位姿估计模型 |

## 附录：引用文献

1. Grigorescu S, Trasnea B, Cocias T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of Field Robotics, 2020, 37(3): 362-386.
2. James S, Wohlhart P, Kalakrishnan M, et al. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12627-12637.
3. Gattullo M, Scurati G W, Fiorentino M, et al. Towards augmented reality manuals for industry 4.0: A methodology[J]. Robotics and Computer-Integrated Manufacturing, 2019, 56: 276-286.
4. Du G, Wang K, Lian S, et al. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review[J]. Artificial Intelligence Review, 2021, 54(3): 1677-1734.
5. Fan Z, Zhu Y, He Y, et al. Deep learning on monocular object pose detection and tracking: A comprehensive overview[J]. ACM Computing Surveys, 2022, 55(4): 1-40.
6. Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world[C]//2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, 2017: 23-30.
7. Sundermeyer M, Marton Z C, Durner M, et al. Implicit 3d orientation learning for 6d object detection from rgb images[C]//Proceedings of the european conference on computer vision (ECCV). 2018: 699-715.
8. Wang H, Sridhar S, Huang J, et al. Normalized object coordinate space for category-level 6d object pose and size estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2642-2651.
9. Hinterstoisser S, Lepetit V, Ilic S, et al. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes[C]//Asian conference on computer vision. Springer, Berlin, Heidelberg, 2013: 548-562.
10. Xiang Y, 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. Hodan T, Haluza P, Obdržálek Š, et al. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects[C]//2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2017: 880-888.
12. Xiang Y, 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.
13. Kehl W, Manhardt F, Tombari F, et al. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1521-1529.
14. Bukschat Y, Vetter M. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach[J]. arXiv preprint arXiv:2011.04307, 2020.
15. Lepetit V, Moreno-Noguer F, Fua P. Epnp: An accurate o (n) solution to the pnp problem[J]. International journal of computer vision, 2009, 81(2): 155-166.
16. Tekin B, Sinha S N, Fua P. Real-time seamless single shot 6d object pose prediction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 292-301.
17. Peng S, Liu Y, Huang Q, et al. Pvnet: Pixel-wise voting network for 6dof pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4561-4570.
18. He Y, Huang H, Fan H, et al. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3003-3013.
19. Wang C, Xu D, Zhu Y, et al. Densefusion: 6d object pose estimation by iterative dense fusion[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3343-3352.
20. Wang H, Sridhar S, Huang J, et al. Normalized object coordinate space for category-level 6d object pose and size estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2642-2651.

## 附录：项目地址

1. https://github.com/Unity-Technologies/com.unity.perception