同行专家业内评价意见书编号：20250854374

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名：汪丹竹

学号：22260007

申报工程师职称专业类别（领域）：电子信息

浙江工程师学院（浙江大学工程师学院）制

2025年04月03日

填表说明

一、本报告中相关的技术或数据如涉及知识产权保护、军工项目保密等内容，请作脱密处理。

二、请用宋体小四字号撰写本报告，可另行附页或增加页数，A4纸双面打印。

三、表中所涉及的签名都必须用蓝、黑色墨水笔，亲笔签名或签字章，不可以打印代替。

四、同行专家业内评价意见书编号由工程师学院填写，编号规则为：年份4位＋申报工程师职称专业类别(领域)4位+流水号3位，共11位。

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.对本专业基础理论知识和专业技术知识掌握情况(不少于200字)

我的专业为电子信息（控制工程），在杭州卫丰机器人有限公司的实践过程中，我通过全过程参与视觉驱动的六轴机械臂抓取系统，积累了丰富的实践经验。在工程建设的技能方面，我熟练运用先进的仪器设备、专业软件以及企业现场数据采集与算法分析等现代研究工具和方法。例如，我熟练使用Linux操作系统、Python数据处理和基于Pytorch的深度学习算法和优化。在技术应用与创新方面，我注重将理论与实践相结合处理困难问题，具备较强的技术创新能力和解决复杂工程问题的能力。例如针对位姿估计过程中物体对称性干扰的问题，我通过优化数据处理方法引入新的解决思路。此外，团队工作也培养了我良好的沟通合作能力，问题导向的工作方式也锻炼了我系统分析问题、不断反思优化的工程思维。

2.工程实践的经历(不少于200字)

本人于2023年11月至2024年五月在杭州卫丰机器人有限公司进行专业实践，主要实践内容为无纹理物的六自由度体位姿估计与机器人抓取，旨在利用计算机视觉技术和深度学习技术提高物体的定位精度，为机械臂抓取提高准确的位姿信息。针对无纹理物体特征提取困难和背景环境复杂的问题的问题，基于深度网络设计满足算法速度与精度平衡的快速准确的物体位姿估计方法：基于图网络融合物体CAD模型的先验知识，提高无纹理物体特征提取的准确度；对物体的对称性进行分析，避免位姿歧义对网络造成干扰。针对位姿估计训练数据获取困难，成本较高等问题，设计了仿真环境下的数据集渲染与生成方法。

3.在实际工作中综合运用所学知识解决复杂工程问题的案例（不少于1000字）

一、项目描述与难点

项目的主要内容为无纹理物的六自由度体位姿估计与机器人抓取，旨在利用计算机视觉技术和深度学习技术提高物体的定位精度，为机械臂抓取提供准确的位姿信息。在项目中，主要难点包括以下几个方面：

1、无纹理物体特征提取困难：无纹理物体表面通常非常平滑，缺乏可供匹配的特征点。而传统的基于特征点匹配的位姿估计方法，通常依赖于物体表面的纹理和特征点来进行匹配和定位，在面对无纹理物体时性能下降。

2、物体对称性的干扰：工业物体大多具有对称性，这进一步增加了六自由度位姿估计的难度。对称性会导致多个等价位姿的存在，从而影响位姿估计的精度和稳定性。

二、分析思路

针对上述难点，我们进行了系统性研究，重点关注了以下几个方面：

1、无纹理物体的位姿估计

针对无纹理物体，重点关注了结合深度信息的方法和基于先验知识的方法：结合深度信息的方法通过结合深度信息，可以在一定程度上弥补无纹理物体表面缺乏特征点的问题。深度信息可以提供物体的三维结构信息，从而辅助位姿估计；基于先验知识的方法则利用物体CAD模型、几何轮廓、比例约束等已知信息，辅助位姿估计。考虑到已有的硬件条件和应用场景，我们选择使用基于先验知识的方法。具体来说，我们利用CAD模型的先验知识来辅助特征提取。CAD模型提供了物体的精确几何信息，可以作为先验知识用于位姿估计。

2、对称物体的位姿估计：

针对对称物体，我们对物体的旋转对称性进行了系统性分析，识别物体的对称轴和对称阶数

。针对位姿歧义问题，我们研究了多种解决方案，最终选择数据集归一化的方式进行处理。归一化可以有效地减少位姿歧义性对网络的影响，提高位姿估计的精度。

三、解决方案

针对上述难点和分析思路，我们提出了以下解决方案：

1、无纹理物体特征提取：

基于RGB图像和CAD模型的先验知识位姿检测网络：我们设计了一个由两部分组成的网络。第一部分是基于改进的Yolov8的预估计网络，用于初步估计物体的位姿。第二部分是基于图神经网络（GNN）的关键点后修正网络，该网络将关键点的像素坐标和CAD模型的3D距离编码为GNN的节点（node）和边（edge）特征，利用先验知识辅助位姿估计任务。

2、对称物体位姿歧义性处理：

针对离散对称性，我们设计了特殊的损失函数，使得网络能够预测任意一个等价位姿。这种损失函数能够有效地处理多个等价位姿的问题，提高位姿估计的稳定性。针对连续对称性，我们使用归一化方案对自由度进行了限制。归一化可以有效地减少位姿歧义性对网络的影响，避免位姿歧义对网络造成干扰。

四、实验与结果

我们进行了大量的实验来验证所提出方法的有效性。具体包括：在LINEMOD、T-LESS等公开数据集上验证了算法的召回率，并设计了机械臂抓取分拣实验验证了算法在真实场景下的应用价值。实验结果表明，我们的方法在无纹理物体和对称物体的位姿估计任务中均取得了较好的效果。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】

1.
公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
Yolo-S: Texture-less object pose estimation with shape	会议论文	2024年05月17日	DDCLS’24	1/2	EI会议收录

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

参加中国高校智能机器人创意大赛，获得省级一等奖，排名第一。

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 85 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年（要求1年及以上） 考核成绩： 85 分
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：** 汪丹竹	

2226007

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀 ☐良好 ☐合格 ☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：　　年　月　日
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过 ☐不通过（具体原因：　　） 工程师学院教学管理办公室审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22260007	姓名：汪丹竹	性别：女	学院：工程师学院	专业：电子信息	学制：2.5年
毕业时最低应获：26.0学分	已获得：29.0学分			入学年月：2022-09	毕业年月：
学位证书号：		毕业证书号：			授予学位：

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2022-2023学年秋季学期	研究生英语能力提升		1.0	免修	跨专业课	2022-2023学年秋冬学期	工程伦理		2.0	78	专业学位课
2022-2023学年秋季学期	新时代中国特色社会主义理论与实践		2.0	90	专业学位课	2022-2023学年秋冬学期	高阶工程认知实践		3.0	82	专业学位课
2022-2023学年秋季学期	数值计算方法		2.0	92	专业选修课	2022-2023学年秋冬学期	智能工业机器人及其应用		3.0	87	专业选修课
2022-2023学年秋季学期	研究生英语基础技能		1.0	免修	公共学位课	2022-2023学年春季学期	科技创新案例探讨与实战		2.0	83	专业选修课
2022-2023学年秋季学期	研究生英语		2.0	免修	专业学位课	2022-2023学年春季学期	多相流检测技术		1.0	92	跨专业课
2022-2023学年秋季学期	工程技术创新前沿		1.5	90	专业学位课	2022-2023学年春季学期	自然辩证法概论		1.0	83	专业学位课
2022-2023学年秋冬学期	研究生论文写作指导		1.0	78	专业选修课	2022-2023学年春夏学期	人工智能制造技术		3.0	88	专业学位课
2022-2023学年冬季学期	产业技术发展前沿		1.5	86	专业学位课		硕士生读书报告		2.0	通过	

说明：1. 研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2. 备注中“*”表示重修课程。

学院成绩校核章：浙江大学研究生院（60）成绩校核章

成绩校核人：张梦依

打印日期：2025-03-20

2024 lEEE 13th Data Driven Control and Learning Systems Conference
May 17-19, 2024, Kaifeng, China

Yolo-S: Texture-less object pose estimation with shape prior

Danzhu Wang[1], Shan Liu[2]

1. Polytechnic Institute, Zhejiang University, Hangzhou 310015, China
E-mail: dzwang@zju.edu.cn

2. State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
E-mail: sliu@zju.edu.cn

Abstract: This paper aims to address the challenge of pose estimation for texture-less objects under conditions of stacking, occlusion, and complex backgrounds. The proposed method, called Yolo-S, consists of two parts: a pre-estimation network based on Yolov8 backbone and a key point post-refinement network based on Graph Neural Network. The first part uses off-the-shelf detectors result as inputs (e.g. MaskRCNN) and outputs the pre-estimate keypoints at the pixel coordinates. The second part encodes the roughly estimated keypoints and the priori knowledge of the object's CAD model, uses a graph neural network to intergrate feature and predict the final keypoints. PNP method is used to obtain the object's pose. The main advantage of Yolo-S is that is has a prior knowledge embedded GNN model for estimating the pose of texture-less objects, which greatly improves prediction quality and accuracy and achieves zero-shot sim-to-real model transfer. Experimental results on the LINEMOD dataset demonstrate the effectiveness of the proposed method and its significant competitiveness with other state-of-the-art methods.

Key Words: Pose Estimation, Yolov8, GNN, Texture-less object

2024 IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS) | 979-8-3503-6167-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/DDCLS61622.2024.10606564

1 Introduction

Pose estimation technology, which means recovering the rotation and translation of an object in the 3-D Euclidean space, is an active research area in computer vision with significant importance for many real-world applications, such as robotic grasping, autonomous driving and unit assembly.

Texture-less objects are very common in industrial products and components, like plastic pipes. Although many pose estimation methods have achieved promising results in recent years, the featureless nature of texture-less objects still poses a great challenge to them due to their heavy reliance on surface features [1].

The existing methods to handle this task can be classified as classical approaches and data-driven approaches [2]. Classical approaches first extract features from image data and establish the correspondence between an object image and the model to realize instance recognition and pose estimation. However, the reliance on handcrafted features and fixed matching procedures have limited their applicability in situations with complex backgrounds, occlusions, or texture-less objects that are difficult to extract features from.

Date-driven approaches have greatly improved the performance of pose estimation algorithms. On the basis of different data type, this method can be divided into RGB image based and RGB-D image-based methods. Although RGB-D image-based methods have achieved good performance in general scenarios with the assistance of depth information, the depth sensor has some limitations. Current consumer-level RGB-D cameras cannot handle non-Lambertian materials well, such as metal parts and glossy plastic, which often produce fragmented depth images. Therefore, it is desirable to rely on only RGB images for 6D pose estimation even if it is more challenging. Although many RGB based methods [3] [4] have work well on public benchmarks (e.g., LINEMOD [5]) of ordinary objects, they have limitations in predicting texture-less object due to the difficulty of feature extraction.

To address this problem, we extend Yolov8 to keypoints detection and use GNN to integrate a priori knowledge of object shape into the model. Thanks to the good performance of Yolov8 and the addition of shape priori, the proposed method, called Yolo-S (which stands for Yolov8 and shape prior), achieves significant accuracy on texture-free pipes using only RGB images. In summary, this work has the following contributions：

a) The Yolov8 network was extended to add a keypoints detection head to pre-estimate the projection of the 3D bounding box of a texture-less object on the 2D image plane. A GNN model is designed to encode the rough estimation of keypoints and priori knowledge of target object into node and edge and refine the keypoints prediction based on the integrated features.
b) A loss function based on the projection error of CAD model is designed to calculate the pixel distance deviation between the projected image in ground truth pose and predicted pose.
c) The proposed method obtains good results on a self-designed texture-less pipes dataset and achieves zero-shot sim-to-real model transfer by means of extensive data augmentation. In addition, the method also obtains good accuracy on the public dataset LINEMOD [5].

2 Related work

This section introduces methods for 6D pose estimation based on RGB images, which can be divided into the following four categories.

*This work is supported by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No.ICT2023B49).

Image-only Estimating Methods: These methods use only RGB images and perform feature extraction based on a CNN network, which in turn regresses the 6D position of the target object, or predicts the keypoints of the target object and calculates the target position using PNP or RANSAC methods. SSD-6D [6] extends the popular single-shot multibox detector (SSD [7]) object detection framework to cover the full 6-D pose space by adding a translation and orientation regression module. BB8 [3] uses a CNN as a keypoint detector to output the two-dimensional coordinates of the eight corners of the object's three-dimensional bounding box. These methods may be effective on objects with rich textures, but it is difficult to achieve ideal accuracy on texture-less target objects that are hard to extract features from.

Template Matching-Based Methods: Pose estimation methods using template matching discretize the cartesian space into a set of pre-defined templates. These templates, labeled with 6D poses, are created offline using the target object's CAD model. During online detection, the template most similar to the input RGB image is chosen as the estimated pose. A typical representative of this kind of method is the LINEMOD [5], which uses multimodal features (RGB image s and depth image) for template matching. This method relies on manually created features and is very time-consuming. AAE [8] proposes a 3D object orientation estimation method based on autoencoders. Instead of explicitly learning the mapping from input images to object poses, it provides an implicit representation of object orientation, defined as the orientation of samples in the latent space. Multipath-AEE [9] extends AAE to multi-object scenes using a single encoder-multi-decoder network structure. This approach achieves good results in pose estimation, but it is slow in computation and the process of creating the template library is also cumbersome.

Prior Knowledge-Based Methods: Despite the rich 2D features provided by RGB images, such as keypoint locations, determining the 6D pose of an object solely from 2D information presents significant challenges, particularly in scenarios involving textureless objects where keypoint extraction is difficult. Incorporating prior knowledge of the object model can enhance model robustness in these complex scenarios. ContourPose [1] uses the object contour as a geometric prior and designs an additional contour decoder implicitly constrains the prediction of keypoints, improving the accuracy of keypoint prediction. PSGMN [10] also uses GNN to integrate the prior knowledge of object CAD models to pose estimation network. The difference from our work lies in that PSGMN uses GNN as feature extractors for the CAD model, while our work uses GNN as an optimizer for the already extracted features. Compared with image-only method, pose estimation methods that incorporate a priori knowledge of the object are more suitable for industrial products with fixed geometric features.

3 Proposed approach

3.1 The Overall Structure of Yolo-S

The proposed pose estimation network Yolo-S includes a keypoints pre-estimation module based on Yolov8, and a key point post processing module based on GNN and shape prior. Yolo-S uses the result of arbitrary segmentation network as input, employing the Yolov8 architecture for backbone and neck for feature extraction and multi-scale feature fusion. Subsequently, through a keypoints head, it approximates the initial projection of the object's 3D bounding box onto the 2D image plane. In the post-refinement network based on the priori knowledge of the object, we encode the object model and the preliminary prediction results of the keypoints as node and edge, and use the GNN network to intergrate the node and edge features, and the fully-connected network is used to perform the post-refinement of the keypoints based on the intergrated features. The overall structure of model is shown in Fig. 1.

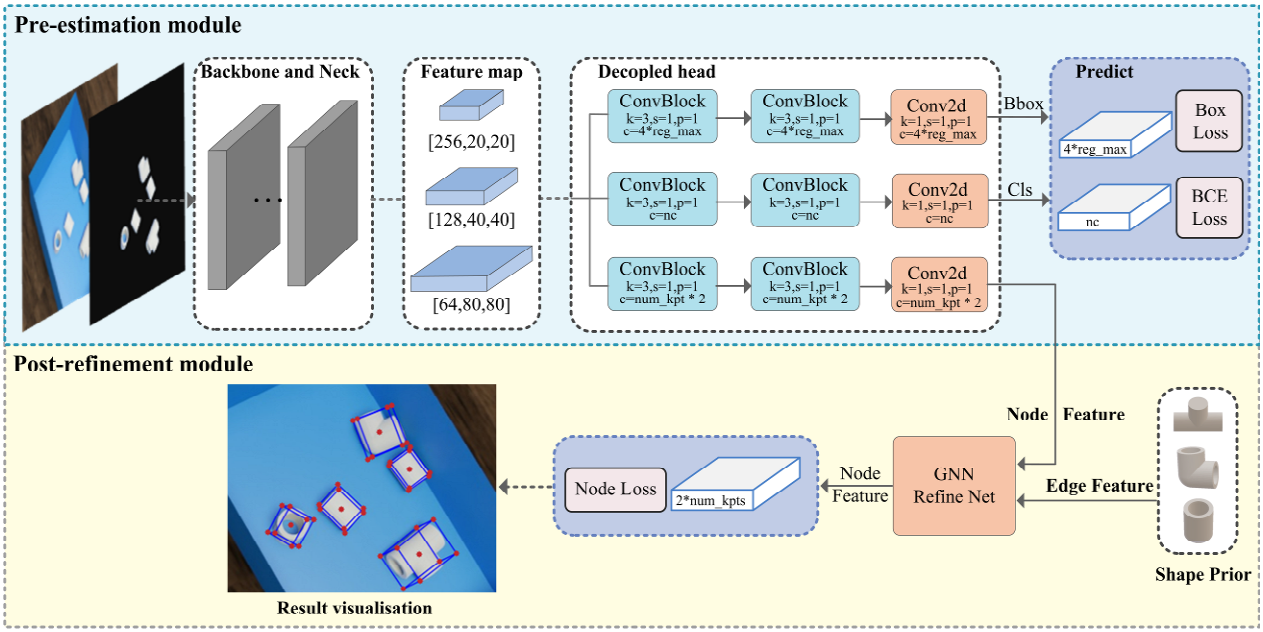

Fig. 1: The overall architecture of Yolo-S

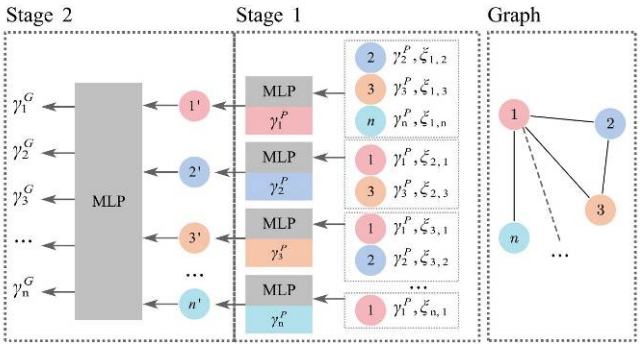

Fig. 2: The architecture of GNN Refine Net

3.2 Pre Estimation Module

The keypoints pre-estimation module is designed based on the extension of Yolov8 network. Yolov8 is a work in progress in 2D object detection but cannot be directly used in 6D bit-pose estimation scenarios, where more 2D points need to be predicted to compute the object's pose in Cartesian space. Therefore, our work enhances Yolov8 by integrating a decoupled keypoints prediction head, which accomplishes the task of predicting the projection of an object's 3D bounding box on the 2D image plane.

The pre-estimation module consists of backbone, neck, decoupled head and predict module. Backbone is designed with reference to the CSPDarknet53 architecture, where each layer down samples the feature maps using a 3×3 convolution with a stride of 2, and introduces cross-stage partial connections in the network using a CSP (cross-stage partial connections) structure. Neck uses FPN [11] (Feature Pyramid Network) and PAN [12] (Path Aggregation Network) for multiscale features fusing, where FPN fuses detailed features at lower levels with semantic features at higher levels through up-sampling and down-sampling operations to obtain a more comprehensive and enriched feature representation. PAN is used to aggregate these features across different layers of the network. Such structure can better utilize the multi-scale information, thus improving the accuracy and stability of detection. The input RGB image is normalized to a standard size of 640×640×3. After Backbone's down sample layer and CSP layer, three feature layers of different scales are obtained, which are future fused by Neck to finally obtain three feature layers for regression and classification tasks.

The prediction of bounding box, category and keypoints are performed by three decoupled heads consisting of two ConvBlock and one 1×1 convolution. The prediction results of bounding box and category are directly calculated through the NMS function. The keypoints prediction head outputs a tensor of size $[b, n, \sum_{i=1}^{3} w_i \times h_i]$, where n is the number of keypoints and (w_i, h_i) are the size of multi-level feature map.

3.3 GNN Refine Net

Graph Neural Network (GNN) are a type of neural network designed to process data that is represented as graphs. Graphs consist of nodes and edges and adjacency matrix. Each node represents an entity, and each edge represents a relationship between two nodes. Both nodes and edges can have features or attributes. Adjacency matrix is a way to describe whether there is an edge between any pair of these nodes. For a graph with N nodes, its adjacency matrix is an N×N matrix represented as $M_{adj\ N\times N}$ $M_{adj\ N\times N}(i,j)=1$ if node i connects to node j, otherwise $M_{adj\ N\times N}(i,j)=0$.

In the GNN post-processing network, the features of the n keypoints output from the previous stage are encoded as node γ:

$$\begin{aligned} \gamma^P &= \{\gamma^P{}_1, \gamma^P{}_2, \cdots \gamma^P{}_n\} \\ \gamma^P{}_i &= (x_i, y_i) \qquad i \in [1, n] \end{aligned} \tag{1}$$

where (x_i, y_i) is the initial estimated pixel coordinate of the projection of the object's 3D bounding box on the 2D image plane. The distance relationship between nodes is then encoded as an edge ξ based on the object model:

$$\begin{aligned} \xi &= \{\xi_{ij} \mid if\ M_{adj}(i,j)=1\} \\ \xi_{ij} &= d(\gamma_i, \gamma_j) \end{aligned} \tag{2}$$

where $d(\gamma_i, \gamma_j)$ represents the Euclidean distance between node γ_i and γ_j, then the task of GNN Refine Net can be summarised as the regression of point features, where γ^G represents the final output of node feature:

$$\gamma^G = Net(\gamma^P, \zeta) \tag{3}$$

The GNN Refine Net consists of two modules and the structure is shown in Fig. 2. The first module, the features of connected nodes and edges are merged. The first module consists of two fully connected layers that integrates the features of connected nodes and edges, and the second phase consists of three fully connected layers that integrates the information between different nodes and predict the final node features γ^G.

3.4 Loss Function

Node Loss: We combine the 3D projection error of the CAD model and the 2D pixel distances of the key points to train the GNN correction network. The 2D pixel distances is represented as:

$$L_{2D\ node} = \frac{1}{n}\sum_{i=1}^{n}(\gamma_i^G - \gamma_i^{gt})^2 \tag{4}$$

where γ^G represents the node feature predicted by GNN refine net, which means the projection of the 3D bounding box in the 2D image plane. γ^{gt} is the ground truth label. The 3D projection error is represented as:

$$L_{3D\ proj} = \frac{1}{m}\sum\nolimits_{P\in M}[\pi^G(P_i) - \pi^{gt}(P_i)]^2 \tag{5}$$

where M is the set of point clouds in the CAD model of object, $P_i = (x_i, y_i, z_i)$ i∈ (1,M) is a point in M, π is the camera projection matrix, which consists of the camera internal parameters and external parameters $[R,t]$, where the internal parameters is known in advance, and $[R,t]$ is obtained by the PNP algorithm based on the predicted keypoints.

Total Loss: The total loss of the network is the sum of node loss, classification loss and box loss.

$$
\begin{aligned}
L &= \lambda_1 L_{node} + \lambda_2 L_{cls} + \lambda_3 L_{box} \\
L_{node} &= \kappa L_{2D\ node} + (1-\kappa) L_{3D\ proj} \\
L_{cls} &= -\sum[(1-\hat{c})\log(1-c) - \hat{c}\log(c)] \\
L_{box} &= 1 - IoU + \frac{\rho(b, b^{gt})}{C^w} + \alpha v \\
v &= \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \\
\alpha &= \frac{v}{1 - IoU + v}
\end{aligned}
\tag{6}
$$

where $\lambda_1, \lambda_2, \lambda_3, \kappa, \alpha$ are the gain coefficients for different component. $\hat{c}$ and c are the ground truth and predicted values of the object categories, and IOU(Intersection over Union) denotes overlap ratio between the ground truth bounding box and the predicted bounding box. w^{gt}, h^{gt}, w, h denotes the width and height of ground truth bounding box and predicted bounding box, $\rho(b, b^{gt})$ denotes the Euclidean distance between the centroid of the predicted box and ground truth, and C represents the length of the diagonal of the smallest enclosing rectangle encompassing both the predicted box and the ground truth box.

4 Datasets and Model Training

4.1 Datasets

Texture-less pipe datasets: We use robot simulation software Coppeliasim to render the texture-less pipe dataset. 2000 initial images of different types of pipes are obtained by setting different camera positions in Cartesian space, and then the dataset size is expanded by ten times by data augmentation methods such as color shifting, adding random noise and occlusion.

LINEMOD datasets: To compare with other state-of-art pose estimation methods, we validated Yolo-S on the public dataset LINEMOD[4], which contains images of 13 texture-less objects in cluttered scenes. Each object's subset consists of approximately 1200 RGB-D images. We also expanded the dataset using data augmentation methods mentioned before.

GNN dataset: In order to speed up the convergence rate of the network, we provide and large amount of synthetic data for training the GNN refinement network to ensure that the network is able to learn the connection information between the key points of the objects. A 10k-sized ground truth value of the node and edge feature is obtained based on the rendering of a real object model, and then the training data is obtained by adding Gaussian noise to the ground truth.

4.2 Training Details

The training of the network is divided into two stages. In the first phase the GNN post-refinement network is trained using the GNN dataset, and in the second phase the overall network is trained end-to-end based on the pre-trained weights of GNN refinement network weights Yolov8 pre-estimation network using RGB image as inputs. Our algorithm is implemented using the Pytorch framework and in the first stage the learning rate is adaptively adjusted using the Adam (Adaptive Moment Estimation) optimizer with an initial learning rate of 0.005, a batch size of 128 and an overall training epoch of 50. The SGD (Stochastic Gradient Descent) optimizer was used in the second phase to train the network end-to-end holistically with an initial learning rate of 0.001, batch size of 8, and overall training epoch of 100. 85% of all the data was divided into training set and 15% was used as a test set, and the configurations of the software and hardware devices for the actual training and testing process are shown in the Table 1.

Table 1: Configuration of Software and Hardware

Item	Configuration
Operating system	Ubuntu 18.04
CPU	1 × Intel Core i7-12700KF
GPU	1 x NVIDIA GeForce GTX 3080Ti
Operating system	CUDA 11.1
Framework	Pytorch 1.9.1

5 Experiments and Results

In this section, we analyze the performance of Yolo-S using our own fittings dataset, and in order to demonstrate the necessity of the GNN post-refinement network, we compared the proposed Yolo-S with and without GNN module. Additionally, we transfer the network that trained on digitally rendered data to real-world environments without using any real samples, and have achieved good results.

In order to compare with other similar pose estimation methods, we also validated Yolo-S on the LINEMOD dataset.

5.1 Evaluation Metrics

We evaluate our method using a commonly used metric in pose estimation: the average 3D distance of the model point (ADD) metric. This metric computes the mean distance between two transformed model points using the estimated pose and the ground truth pose. It is claimed that the estimated pose is correct if the distance is less than 10% of the model diameter. The ADD metric is defined as follows:

$$
ADD = \frac{1}{m}\sum_{x \in M}\left\|(Rx + t) - (\hat{R}x + \hat{t})\right\| \tag{7}
$$

where m is the number of model points and x represents 3D points. The ground truth and estimated pose are represented as $[R \mid t]$ and $[\hat{R} \mid \hat{t}]$ respectively.

For objects with rotational symmetry whose pose is ambiguous, the ADD-S metric is used to calculate the average 3D distance of all nearest neighbor pairs in the two point sets after transformation. ADD-S is defined as follows:

$$
ADD - S = \frac{1}{m}\sum_{x_1 \in M}\min_{x_2 \in M}\left\|(Rx_1 + t) - (\hat{R}x_2 + \hat{t})\right\| \tag{8}
$$

5.2 Evaluation Results

Texture-less pipes dataset: The visualization results of the position estimation of Yolo-S on the texture-less pipes are shown in Fig. 3. In order to validate the generalization performance of the network, instead of using the

Coppeliasim synthetic data used to train the network, we used images rendered by BlenderProc for model validation, which were not learned by the network.

Fig. 3: Texture-less pipe estimation result based on BlenderProc rendered image.
Line 1: BlenderProc rendered image with proposed Yolo-S
Line 2: Yolo-S without GNN model.
Line 3: Real image with GNN model.

Line 1 of Fig. 3 shows the results of the proposed Yolo-S on BlenderProc rendered images, line 2 shows the prediction results based on Yolo-S without the GNN post-refinement module, and line 3 shows the results of the Yolo-S on real images with a complex background and stacked target objects. It can be clearly seen that the success rate of network prediction with the addition of the GNN module was 100%, while the prediction results without the GNN module are much unstable. The comparison of line 1 and line 2 shows that the addition of the GNN module greatly improves the accuracy of the prediction network and has an advantage in the case of a small amount of occlusion of the target object. Line 3 shows the prediction results of real images with difficulties such as occlusion, stacking, and complex background, and the difficulty of prediction is incremented from left to right. It is worth noting that the network without the GNN post-refinement module cannot be used in the real image.

Table 2 compares the network prediction results with and without the GNN module using ADD accuracy on BlenderProc's test set. From Table 2, it can be seen that the GNN module greatly improves the prediction performance of the model for texture-less pipes. These results strongly demonstrate the necessity of the GNN module.

Table 2: Comparison w and w/o GNN Module

Method	ADD metric
w	70.1
w/o	11.2

LINEMOD dataset: In order to compare more intuitively with other same type of pose estimation methods and to demonstrate the advantages of proposed Yolo-S, we also validate the proposed method on the public dataset LINEMOD. Fig. 4 shows the visualization results of Yolo-S on the LINEMOD dataset, where the predicted bounding box (blue) is very close to the ground truth bounding box (green), which implies that the estimation results are accurate. Table 3 shows the comparison of Yolo-S with other state-of-the-art methods, where BB8[2] uses CNN as a keypoints detector to output the 2D coordinates of the eight corner points of the 3D bounding box of the target object, and we adopt the optimization method of BB8, i.e., regressing the keypoints on the heatmap to make the comparison. AAE[7] uses an enhanced autoencoder to encode the object rotations into the latent space, given an image, AAE predicts only the rotation of the object, while the translation of the object is estimated using a 2-D bounding box. As can be seen from the table, the average ADD accuracy of Yolo-S exceeds that of the optimized BB8 method and the AEE method using depth information.

Fig. 4: LINEMOD data estimation result using Yolo-S

Table 3: Comparison in Terms of ADD Metric

Method/Object	BB8[2]	AAE[7]	OURS
Ape	40.4	24.4	**55.3**
Benchvise	91.8	89.1	**92.9**
Cam	55.7	82.1	**89.5**
Can	64.1	**70.8**	68.9
Cat	62.6	**72.2**	61.6
Driller	74.4	44.9	**93.9**
Duck	44.3	54.6	**60.8**
Eggbox	57.8	**96.6**	61.7
Glue	41.2	**94.2**	88.6
Hole puncher	**67.2**	51.3	47.3
Iron	84.7	77.9	**96.5**
Lamp	76.5	86.3	**90.8**
Phone	54.0	**86.2**	80.97
Average	62.7	71.6	**76.05**

6 Conclusion

In this paper, we propose an end-to-end position estimation network based on Yolov8 feature extractor and the priori knowledge of the object model called Yolo-S. GNN network is used to encode the pre-estimation of keypoints extracted by Yolov8 and the a priori knowledge of the object model as node and edge, and fuses the features of node the edge for keypoints post-refinement. The overall model is able to predict the projection of the 3D bounding box of target object on the 2D image plane and subsequently compute the object's 6D pose based on PNP algorithm. To improve the prediction accuracy, we combine the 3D projection error of the CAD model and the 2D pixel distance of the keypoints to design the loss function for keypoint prediction. Yolo-S demonstrates excellent pose estimation performance on texture-less pipes, and is able to handle the case of complex backgrounds, object stacking and occlusion, and achieves zero-shot sim-to-real model migration. What's more, Yolo-S shows superior performance over other methods with similar ideas on the LINEMOD dataset.

References

[1] Z. He et al., ContourPose: Monocular 6-D Pose Estimation Method for Reflective Textureless Metal Parts, *IEEE Transactions on Robotics,* 39(5): 4037-4050, 2023.

[2] C. Wang et al., DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019: 3338-3347.

[3] M. Rad and V. Lepetit, BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 3848-3856.

[4] B. Tekin, S. N. Sinha and P. Fua, Real-Time Seamless Single Shot 6D Object Pose Prediction, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 292-301.

[5] Hinterstoisser, Stefan, et al. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. in *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision*, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. Springer Berlin Heidelberg, 2013.

[6] W. Kehl, F. Manhardt, F. Tombari, S. Ilic and N. Navab, SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again, in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017: 1530-1538.

[7] Liu, Wei, et al. Ssd: Single shot multibox detector. in *Computer Vision–ECCV 2016*: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016.

[8] Sundermeyer, M., Marton, Z. C., Durner, M., Brucker, M., & Triebel, R., Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, 2018: 699-715.

[9] M. Sundermeyer et al., Multi-Path Learning for Object Pose Estimation Across Domains, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020: 13913-13922.

[10] C. Wu, L. Chen, Z. He and J. Jiang, Pseudo-Siamese Graph Matching Network for Textureless Objects' 6-D Pose Estimation, *IEEE Transactions on Industrial Electronics*, 69(3): 2718-2727, 2022.

[11] Kim, S. W., Kook, H. K., Sun, J. Y., Kang, M. C., & Ko, S. J., Parallel feature pyramid network for object detection. In *Proceedings of the European conference on computer vision (ECCV)* 2018: 234-250.

[12] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, Path Aggregation Network for Instance Segmentation, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 8759-8768.

经检索“Engineering Village”，下述论文被《Ei Compendex》收录。（检索时间：2024 年 12 月 16 日）。

<RECORD 1>
Accession number:20243516969118
Title:Yolo-S: Texture-Less Object Pose Estimation with Shape Prior
Authors:Wang, Danzhu (1); Liu, Shan (2)
Author affiliation:(1) Polytechnic Institute, Zhejiang University, Hangzhou; 310015, China; (2) College of Control Science and Engineering, Zhejiang University, State Key Laboratory of Industrial Control Technology, Hangzhou; 310027, China
Source title:Proceedings of 2024 IEEE 13th Data Driven Control and Learning Systems Conference, DDCLS 2024
Abbreviated source title:Proc. IEEE Data Driven Control Learn. Syst. Conf., DDCLS
Part number:1 of 1
Issue title:Proceedings of 2024 IEEE 13th Data Driven Control and Learning Systems Conference, DDCLS 2024
Issue date:2024
Publication year:2024
Pages:1889-1894
Language:English
ISBN-13:9798350361674
Document type:Conference article (CA)
Conference name:13th IEEE Data Driven Control and Learning Systems Conference, DDCLS 2024
Conference date:May 17, 2024 - May 19, 2024
Conference location:Kaifeng, China
Conference code:201823
Sponsor:Chinese Association of Automation (CAA); DCLOD; Henan University; IEEE; IEEE Beijing Section; Qingdao University
Publisher:Institute of Electrical and Electronics Engineers Inc.
Number of references:12
Main heading:Graph neural networks
Controlled terms:Computer vision - Object detection - Object recognition
Uncontrolled terms:GNN - Graph neural networks - Keypoints - Network-based - Object pose - Pose-estimation - S textures - Shape priors - Texture-less object - Yolov8
Classification code:1101 - 1106.3.1 - 1106.8
DOI:10.1109/DDCLS61622.2024.10606564
Funding details: Number: -, Acronym: -, Sponsor: State Key Laboratory of Industrial Control Technology;Number: -, Acronym: ZJU, Sponsor: Zhejiang University;
Funding text:This work is supported by the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No.ICT2023B49).
Database:Compendex

注：
1. 以上检索结果来自 CALIS 查收查引系统。
2. 以上检索结果均得到委托人及被检索作者的确认。

教育部科技查新工作站（Z09）
委托人（签章）：应湘艺
2024 年 12 月 16 日

获 奖 证 书

HONOR CERTIFICATES

证书编号
ELSFKZNJQR20230022

学校：浙江大学

学生：汪丹竹 潘力豪 萧力城

指导教师：刘山

参赛作品：超级拼接侠

在2023年第六届中国高校智能机器人创意大赛俄罗斯方块专项赛初赛中荣获

一等奖

特发此证，以资鼓励！

中国高校智能机器人创意大赛组委会

二〇二三年八月十二日