Light-weight deep learning framework for 6-D pose estimation

SYNOPSIS FOR

RCO507- Report 1

Supervisor:

Prof. Anil Singh Parihar

Submitted By:

Sanjay Singh Gurjar

25/RCO/02

Department of Computer science and engineering

Delhi Technological University

Delhi, India

Title: Lightweight deep learning framework for 6-D pose estimation

1. Abstract

The 6D object pose estimation problem has been a central topic in computer vision. It involves estimating an object's 3D rotation and 3D translation—collectively known as its 6D pose—from a single RGB image. Traditional approaches such as PVNet, Uni6D, and other deep learning-based models have shown strong performance but often require large computational resources and object-specific training, limiting their scalability and real-time applicability.

This research focuses on developing category-agnostic and model-free pose estimation methods that can generalize across unseen object categories without relying on pre-defined 3D models. We propose a lightweight deep learning framework designed for efficient and effective 6D pose estimation, maintaining high accuracy while significantly reducing computational complexity.

The proposed framework enables robust performance in various robotics, augmented reality, and computer graphics applications, where real-time inference and generalization to novel objects are crucial. We evaluate our method on benchmark datasets including LINEMOD, Toyota Light, HO3D, YCB-INEOAT, and LINEMOD-O, demonstrating its strong generalization ability and potential for deployment in real-world scenarios

2. Introduction

Object pose estimation is a fundamental problem in computer vision and robotics, concerned with determining the rigid 6D transformation—comprising a 3D rotation and a 3D translation—that defines an object's position and orientation in space. Despite extensive research, accurate and efficient 6D pose estimation remains a challenging task due to factors such as occlusion, clutter, lighting variation, and intra-class diversity.

In industrial and automation domains, pose estimation plays a vital role in enabling precise robotic manipulation and environmental understanding. For example, in automated assembly lines such as car manufacturing, robotic arms must identify and localize components accurately to perform tasks like picking, placing, and fitting. Precise 6D pose estimation ensures that these operations are performed reliably and efficiently, minimizing human intervention and production errors.

Similarly, in autonomous driving, estimating the 3D pose of surrounding objects—such as vehicles, pedestrians, and road elements—is essential for safe navigation and collision avoidance. The system must understand not just the presence of an object but also its orientation and spatial configuration relative to the vehicle.

In robotic grasping and manipulation, the pose of an object provides the critical spatial information needed for grasp planning. A robotic gripper can only interact effectively with an object when its position and orientation are precisely known. Without accurate pose estimation, the robot cannot determine how to approach, align, and manipulate the object.

Therefore, developing robust, efficient, and generalizable methods for 6D object pose estimation is key to advancing automation in robotics, augmented reality, and intelligent perception systems.

The 6D object pose of an object is defined by its 3D rotation and 3D translation with respect to a given coordinate frame, typically the camera coordinate frame. Together, these six parameters describe the full spatial configuration of an object — its orientation (rotation) and position (translation) in 3D space. In simple terms, estimating the 6D pose means determining where an object is and how it is oriented relative to the camera or robot.

Accurate 6D pose estimation forms the foundation for numerous tasks in computer vision, robotics, and augmented reality (AR).

In robotics, it enables robots to localize and manipulate objects precisely. For instance, a robotic arm performing an assembly task or grasping an object must know the object's exact pose to align the gripper with the object's geometry and surface orientation.

In autonomous driving, pose estimation assists in understanding the 3D configuration of surrounding vehicles, pedestrians, and obstacles, supporting safe decision-making and navigation.

In augmented and mixed reality, pose estimation ensures that virtual objects are rendered in alignment with the physical environment, maintaining spatial coherence and realism.

A simple example can be seen in an AR-based interactive game, where a player virtually lifts or moves physical objects. The system must estimate the precise 6D pose of the object to correctly render and track its virtual counterpart. Without accurate pose information, both robotic manipulation and virtual alignment become unreliable.

Conventional deep learning methods for 6D pose estimation—such as PoseCNN, PVNet, DenseFusion, CosyPose, and Uni6D—achieve strong performance but rely on object-specific 3D models and category supervision. These methods typically assume the availability of accurate CAD models or synthetic renderings during training and inference. As a result, they perform well on known object categories but fail to generalize to unseen objects or new environments.

Furthermore, these models are computationally intensive, requiring large-scale datasets and GPU resources, which limit their deployment in resource-constrained settings such as mobile robotics, edge devices, and real-time AR systems. This dependency on pre-defined models and high computational cost constrains scalability and practical usability.

To overcome these limitations, the research community has shifted toward model-free and category-agnostic pose estimation frameworks.

Model-free methods eliminate the need for explicit 3D CAD models of objects during training or inference. Instead, they learn geometric representations directly from visual features, enabling pose prediction even for objects that were not seen during training. Such approaches leverage implicit shape understanding through cues like object contours, keypoints, or dense correspondences without requiring any prior mesh information.

Category-agnostic frameworks are designed to generalize across object categories. Instead of learning category-specific pose priors, these methods capture shared geometric patterns and spatial structures across diverse objects, allowing the network to estimate poses for unseen categories. This makes them highly suitable for real-world robotics and AR applications, where encountering unknown objects is inevitable.

A robust category-agnostic and model-free system can be trained on a diverse dataset of generic objects and then applied directly to new domains without retraining. Such generalization is a key step toward building scalable, adaptable, and deployment-ready pose estimation systems.

This research aims to design a lightweight, model-free, and category-agnostic deep learning framework for 6D object pose estimation that can operate efficiently on limited computational resources while maintaining strong generalization and accuracy. The proposed framework minimizes dependence on pre-defined 3D models, instead learning robust geometric priors that can adapt to unseen objects, unstructured environments, and occluded scenes. Such a system has wide applicability in domains like robotic grasping, AR-based interaction, and autonomous perception, where scalability, efficiency, and generalization are as crucial as precision.

Would you like me to extend this section into a complete paper-style introduction (with references to existing works, challenges, and contributions)? That would make it publication-ready.

3. Datasets

3.1 LINEMOD Dataset

The LINEMOD dataset is one of the most widely used benchmarks for evaluating 6D object pose estimation methods. It consists of RGB and depth image sequences of multiple household and industrial objects captured under varying lighting and background conditions. Each object is associated with a precisely aligned 3D CAD model, allowing accurate ground-truth pose annotations.

The dataset contains 15 distinct texture-less objects such as cans, drills, and boxes, recorded from

different viewpoints. The main challenge in LINEMOD lies in handling texture-less surfaces, cluttered backgrounds, and viewpoint variations, making it an essential benchmark for assessing the robustness and accuracy of pose estimation algorithms.

3.2 YCB-Video Dataset

The YCB-Video dataset extends the popular YCB (Yale-CMU-Berkeley) Object and Model Set, which includes a variety of everyday objects with accurate 3D mesh models. The dataset provides RGB-D video sequences of these objects under different lighting, occlusion, and motion conditions, captured using a moving camera. It contains 92 video sequences covering 21 object categories from the YCB set, along with ground-truth 6D poses for each frame. The YCB-Video dataset is considered more diverse and realistic than LINEMOD, making it a crucial benchmark for evaluating a model's generalization, temporal consistency, and real-world performance.

3.3 Occlusion LINEMOD Dataset:

The Occlusion LINEMOD dataset is an extension of the original LINEMOD benchmark, specifically designed to evaluate pose estimation performance under severe occlusion. In this dataset, multiple LINEMOD objects appear simultaneously within the same scene, partially covering each other. The dataset includes RGB-D images of eight objects that are significantly occluded in most frames. This poses a challenging scenario for algorithms that rely on full object visibility or contour information. The Occlusion LINEMOD dataset is therefore widely used to test the robustness and occlusion-handling capability of pose estimation models.

Together, these datasets provide a comprehensive evaluation framework — from clean, single-object scenes (LINEMOD) to complex, cluttered, and occluded real-world environments (Occlusion LINEMOD and YCB-Video) — enabling thorough assessment of pose estimation methods across varying levels of difficulty and generalization

4. Evaluation metric:

4.1 ADD Metric:

The ADD metric measures the mean Euclidean distance between the 3D model points transformed by the predicted pose and the ground-truth pose.

4.2 ADD(S) metric:

For **symmetric objects**, the standard ADD metric can produce misleading errors because multiple poses can yield visually identical object appearances. To handle this, the **ADD-S** (or **ADD-Symmetric**) metric is used.

Instead of measuring the distance between corresponding model points, ADD-S computes the **average distance from each transformed model point to its nearest neighbor** on the model under the ground-truth pose.

5. Work Plan

The objective of this research is to develop a lightweight, modelfree, and category agnostic deep learning framework for 6D object pose estimation that is both computationally efficient and robust to occlusions, clutter, and unseen object categories. The work plan is structured into systematic phases to ensure a thorough exploration, implementation, and validation of the proposed approach.

5.1. Literature Review and Background Study

- Conduct an extensive survey of existing 6D pose estimation methods, including modelbased, modelfree, and categoryagnostic approaches.
- Analyze the architectures and methodologies of state of the art models such as PVNet, PoseCNN, DenseFusion, GDRNet, and Uni6D.
- Study relevant topics such as geometric deep learning, keypoint regression, correspondence learning, and transformer-based perception.
- Identify the strengths, weaknesses, and computational limitations of current frameworks to define the design requirements for a lightweight model.

5.2. Problem Definition and Gap Identification

- Clearly define the research problem within the context of modelfree, categoryagnostic pose estimation.
- Identify specific research gaps such as:
  - High computational cost in existing architectures.
  - Limited generalization to unseen object categories.
  - Sensitivity to occlusion and viewpoint variation.
  - Formulate hypotheses and establish measurable objectives to address these limitations.

5.3. Model Design and Implementation

- Design a lightweight neural network architecture that can estimate 6D poses directly from RGB or RGBD inputs without relying on objectspecific CAD models.
- Explore feature extraction backbones optimized for speed (e.g., MobileNet, EfficientNet, or lightweight transformers).
- Integrate geometric reasoning components such as 2D–3D correspondences, implicit shape representation, or keypoint localization.
- Implement training pipelines using benchmark datasets (e.g., LINEMOD, Occlusion LINEMOD, YCBVideo, HO3D, Toyota Light).

5.4. Experimental Analysis and Ablation Studies

- Conduct detailed ablation studies to evaluate the impact of architectural choices, loss functions, and feature representations on performance.
- Compare the proposed framework with existing baselines on standard evaluation metrics such as ADD and ADDS.
- Analyze model performance under different conditions — occlusion, lighting changes, background clutter, and unseen object categories.
- Optimize the network for inference speed, memory footprint, and scalability.

5.5. Evaluation and Validation

- Evaluate the trained models on multiple datasets to ensure generalization across different environments and object types.
- Validate the proposed approach in realworld scenarios, such as robotic grasping or ARbased object interaction tasks.
- Use visualization and quantitative metrics to demonstrate pose accuracy, robustness, and computational efficiency.

5.6. Documentation and Future Work

- Summarize findings, insights, and experimental results.
- Identify potential extensions such as temporal pose tracking, multiobject scene understanding, or domain adaptation for crossdataset generalization.
- Prepare research manuscripts for publication in relevant computer vision and robotics conferences.

References

[1.] Negar Nejatishahidin, Pooya Fayyazsanavi, Review on 6D Object pose estimation with the focus on Indoor scene understanding, Advances in Artificial Intelligence and Machine Learning

[2.] Sida Peng, PVNet, CVPR 2019

[3.] Xiaoke Jiang, Donghai Li, Uni6D: A Unified CNN Framework without Projection Breakdown for 6D Pose Estimation

[5.] Taeyeop Lee, Bowen Wen, Any6D: Model-free 6D Pose Estimation of Novel Objects

[6.] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In ICCV, 2017

[7.] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In CVPR, 2019.

[8.] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield.

Deep object pose estimation for semantic robotic grasping of household objects. In CoRL, 2018

[9.] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. In ICRA, 2022

[10.] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In ECCV, 2020

[11.] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In CVPR, 2019 [12.] Evin Pınar Ornek, Yann Labb ¨ e, Bugra Tekin, Lingni Ma, ´ Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In ECCV, 2024.

[13.] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In CVPR, 2024.

[14.]Review on 6D Object Pose Estimation With the Focus on Indoor Scene Understanding https://www.oajaiml.com/uploads/archivepdf/24821141.pdf