Automated Pallet Pick-and-Place Manipulation Using Physics-Based Simulation in NVIDIA Isaac Sim

Joungbin Choi Instructed by Prof. YoungKeun-Kim
School of Mechanical and Control Engineering, Handong Global University
Personal Contact: (Mobile) 010-5932-0865 (e-mail) jb4310@handong.ac.kr

1. Introduction

1.1 Background

Automating pallet loading and unloading requires **accurate pose estimation** and reliable manipulation, but collecting large-scale real-world training data is costly, time-consuming, and difficult to manage safely. **High-fidelity** and **physics-based simulation** offers an efficient alternative by generating large-scale pose data and enabling the development of pose estimation and manipulation algorithms that reliably reflect real-world robotic interactions.

1.2 Isaac Sim

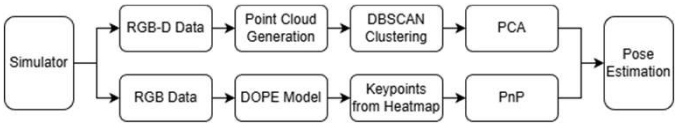
NVIDIA Isaac Sim is a high-fidelity robotics simulation platform that provides **photorealistic rendering**, **accurate physics**, and **scalable synthetic RGB-D data generation** with precise ground-truth annotations. Its flexible scene configuration, built-in robot models, motion controllers, and sensor simulations make it well-suited for developing and evaluating perception and manipulation algorithms quickly and safely without **the cost and risk of physical experimentation**.

Figure 1. Human Stacking a Pallet

2. Pose Estimation

2.1 Pipeline

This project presents two approaches for pallet pose estimation: an RGB-D–based method and a model-based method. Both approaches are evaluated within a physics-enabled simulation environment. The overall pipeline consists of the following stages.



3. Method 1

3.1 RGB-D Geometry-based Pose Estimation

1. RGB-D Acquisition

RGB images and depth maps are captured from a virtual RGB-D camera in Isaac Sim. YOLO is applied to detect the pallet and define the region of interest (ROI).

2. Point Cloud Generation

Depth values and camera intrinsics (K) are used to project each pixel into a 3D point in the camera coordinate frame, which is then transformed into the world coordinate frame using the camera pose.

3. Clustering (DBSCAN)

DBSCAN is applied to the point cloud inside the bounding box to isolate the pallet surface. The largest cluster is selected to extract the side structure.

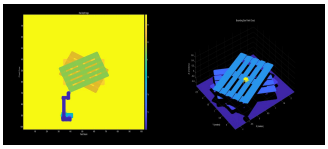


Figure 2. Raw Point Cloud(right) and Point Cloud in ROI(left)

4. Surface Extraction & PCA

PCA is performed on the clustered points to estimate the dominant axis and compute the surface normal, determining the orientation of the side picking surface.

Equations of PCA

Covariance Matrix $C = \frac{1}{n}(X-\mu)^T(X-\mu)$

Eigenvalue Decomposition $Cv_i = \lambda_i v_i$

Projection onto Principal Components $Z = (X-\mu)V$

5. Pose Generation

The surface center and derived normal vector are used to construct the rotation matrix, which is then converted to an Isaac Sim-compatible quaternion to compute the final 6D pick pose.

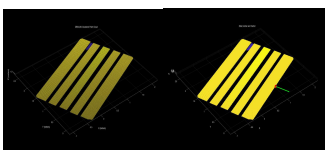


Figure 3. Clustering with DBSCAN(right) and Normal Vector(left)

4. Method 2

4.1 DOPE Model Description

Deep Object Pose Estimation (DOPE) is a deep-learning framework that infers an object's full 6D pose from a single RGB image through a multi-stage architecture that progressively refines keypoint belief maps and affinity fields.

4.2 Advantages

1. Single-RGB 6D Pose Estimation
2. Synthetic-Data Friendly
3. Robust in Multi-Object Scenes

4.3 Network Architecture

1. Feature Extractor (VGG19)
2. Belief Map Heads
3. Vector Field Heads (Affinity Fields)
4. Keypoint Grouping + PnP

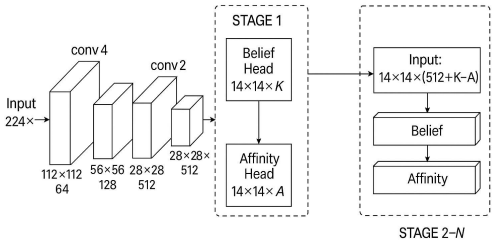


Figure 4. Predicted Heat map and Keypoints Result

4.4 Training Setup

Training data were **generated using Isaac Sim**, where pallet objects were placed in randomized environments with **automatically labeled keypoints** and ground-truth pose files (JSON).

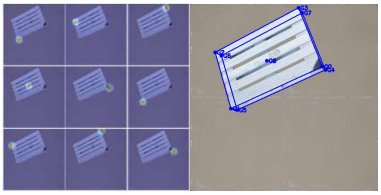


Figure 5. Predicted Heat map and Keypoints Result

5. Result and Discussion

5.1 The Performance

To test RGB-D geometry-based and DOPE model-based pose estimation, 100 pallets were generated with random orientations between −80° and 80° in simulation.

DOPE achieves lower rotation error and greater robustness, whereas the RGB-D method is more affected by depth noise and surface variation.

Table 1. Mean Error Angle

Method	Mean Rotation Error (°)
RGB-D Geometry	±2.308°
DOPE	±0.973°

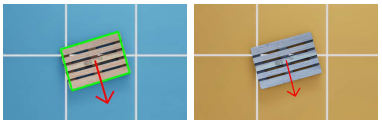


Figure 6. The Result of DOPE(right) and RGB-D(Left)

5.2 Actual Implementation

Both the RGB-D–based pose estimation pipeline and the DOPE model demonstrated successful pallet pick-and-place performance on both of **Isaac Sim simulation** and **a real PIPER robot** in a physical environment.

Figure 7. Demonstration of Pallet Pick-and-Place Execution

6. Conclusion

6.1 Conclusion and Improvements

Simulation enables efficient collection of high-quality training data, and the use of a physics-based engine reduces the **sim-to-real gap**, making experiments safer and more cost-effective.

However, applying the system to full-scale pallets and developing a complementary fusion of the two pose estimation methods remain important future steps.