Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects

Presenter: Zhiyao Bao

2021/10/05

Main Problem – Overview

Difficulty of collecting sufficiently large amounts of labeled training data

Synthetic Data

Reality gap: training on synthetic data usually do not perform well on real data

DOPE

Main Problem – Motivation & Significance

General-purpose Robot Autonomy

- Bridge the reality gap of using synthetic data
- Generalize well to novel environments (extreme lighting conditions)

Application & Impact

- Detect and estimate the 6-DoF pose of instances of a set of known household objects from a single RGB image
- Household-objects-related robot manipulation tasks

Main Problem – Prior Approaches & Challenges & Targets

Prior Approaches

- The performance training with synthetic data not comparable to the one training with real data
- Need fine-tuning to achieve great performance

Challenge:
Reality Gap

Targets

trained on synthetic data

+

state-of-the-art performance (compare to real data ones)

+

without fine-tuning

+

real time

Main Problem – Key Insights

- Domain Randomized + Photorealistic Data
- Train only on synthetic data while achieve state-of-the-art performance compared with a network trained on a combination of real and synthetic data
- Infer the 3D pose of such objects, in clutter, from a single RGB image in real time for the purpose of enabling the robot to manipulate such objects

Problem Setting – Definitions & Formulation

Def: Pose Estimation

3D position and orientation of objects estimation in the scene

Formulation: Explore how to train a neural network for 6-DoF known household object pose estimation solely with synthetic data from a single RGB image.

Related Work & Limitations

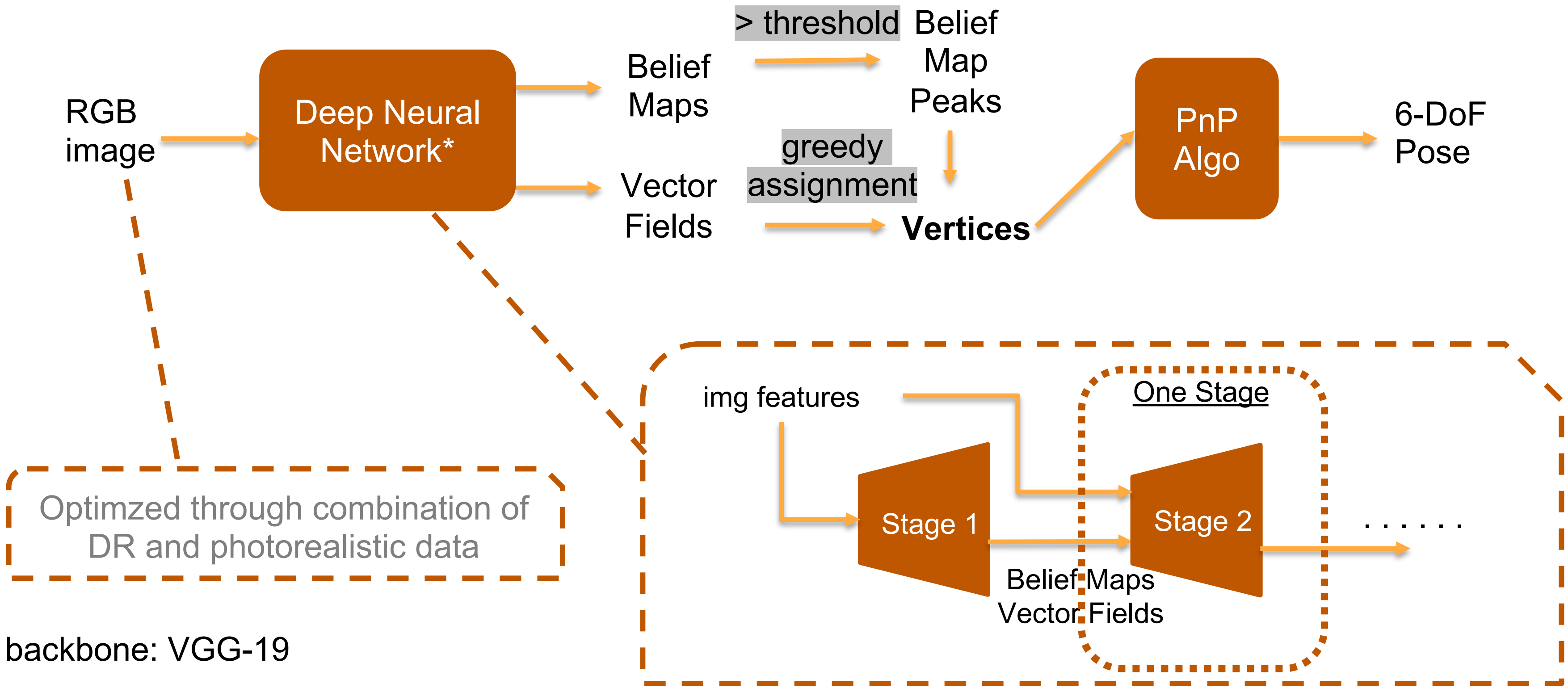
- Domain Randomization
    - Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPR Workshop on Autonomous Driving (WAD)*, 2018.

Require fine-tuning to achieve great performance

- Photorealistic Data
    - Falling things: A synthetic dataset for 3D object detection and pose estimation. In *CVPR Workshop on Real World Challenges and New Benchmarks for Deep Learning in Robotic Vision*, June 2018.

Only photorealistic dataset, not solving real-world problems

Proposed Approach – Network Overview

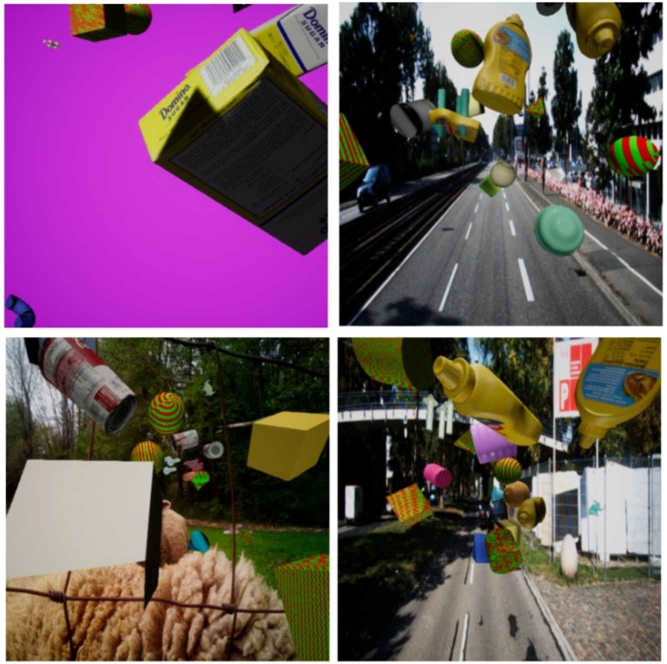


* backbone: VGG-19

Proposed Approach – Domain Randomization

Def: Place the foreground objects **within virtual environments** consisting of **various distractor objects** in front of a **random background**.

domain randomized



Proposed Approach – Photorealistic Images

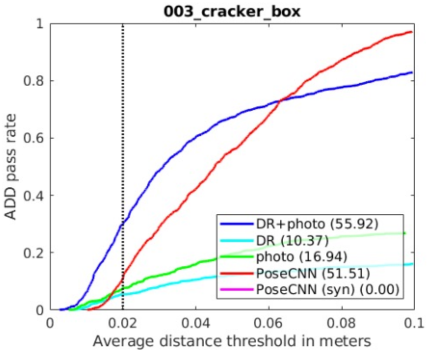
Def: Placing the foreground objects in 3D background scenes **with physical constraints**. Allowed to fall under the weight of gravity, and to collide with each other and with surfaces in the scene, these objects **interact in physically plausible ways**.

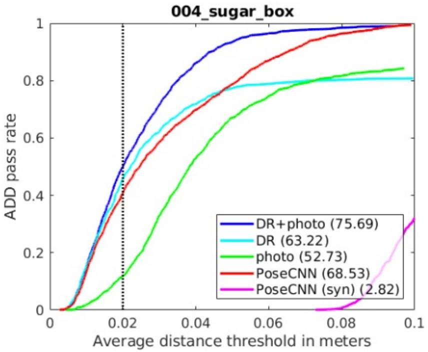
photorealistic

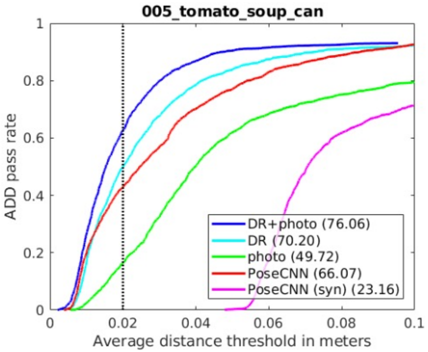
Experimental Setup

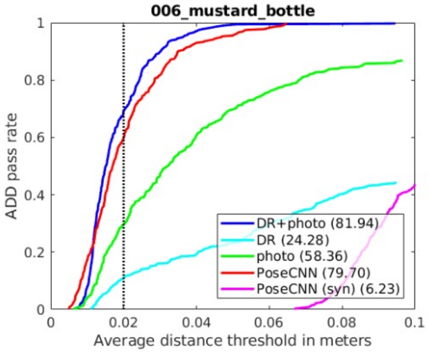
- **Dataset**: YCB-Video dataset + Extreme lighting dataset
- **Task**: Pose estimation
- **Hardware**: Logitech C960 camera & Baxter robot
- **Baseline**: Compare to PoseCNN (PoseCNN > Tekin, BB8 and SSD-6D on the standard LINEMOD, Occluded-LINEMOD datasets)
- **Evaluation Metric**: Average Distance (ADD) Metric (average 3D Euclidean distance of all model points between ground truth pose and estimated pose)
- **Goal**: detecting and estimating the 6-DoF pose of all instances of a set of known household objects from a single RGB image

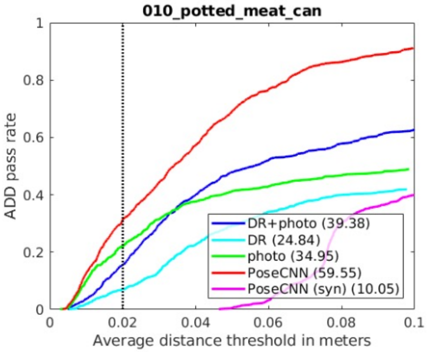
Experimental Results





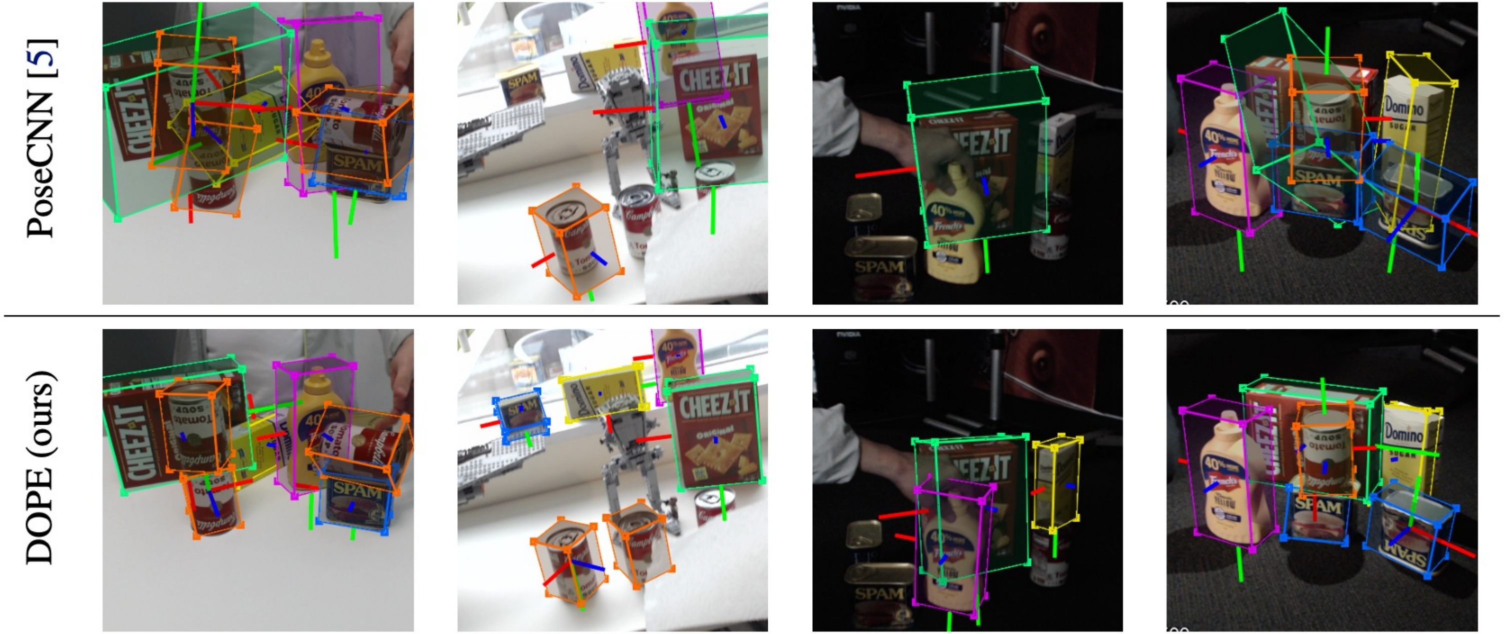






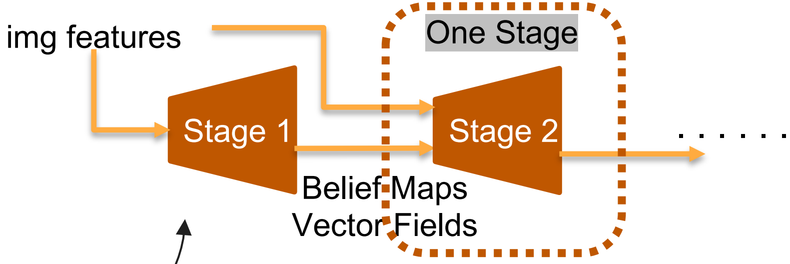
This paper's method (blue curve) yields the best results for 4 out of 5 objects

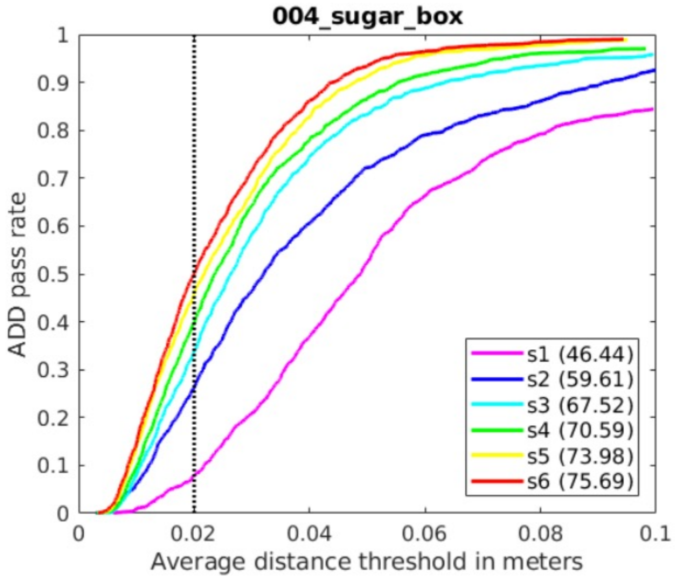
Experimental Results



This paper’s method **generalizes** better to **extreme real-world conditions.**

Experimental Results





	speed (ms)	AUC
1 stage	57	46.44
2 stages	88	59.61
3 stages	124	67.52
4 stages	165	70.59
5 stages	202	73.98
6 stages	232	75.69

Additional stages yield higher accuracy at the cost of greater computation.

Discussion of Results – Summary

- **Mixing** DR and photorealistic synthetic data achieves greater success at domain transfer than either DR or photorealistic images alone ✓ Figure 2
- The performance was comparable for all networks as long as **at least 40% of either dataset was included** ⊖ Section 3.5, but no experiment details
- **Generalizes** well to a variety of real-world scenarios, including extreme lighting conditions ✓ Figure 3

Discussion of Results – Strength & Weakness

Strength

- trained **only on synthetic data**
- Great **performance** on pose estimation
- the resulting poses are of **sufficient accuracy for robotic manipulation**
- **generalizes** better to novel environments including extreme lighting conditions

Weakness

- May not work perfect for severely occluded frames when the part visible is not properly modeled in synthetic data (E.g. potted meat can instances detection failures)

Critique & Limitations

- Medium novelty
- May not work well for severely occluded frames when the part visible is not properly modeled in synthetic data
- Limited to certain rigid and known household objects

Future Work

- Increasing the number of objects in the image
- Handling symmetry
- Incorporating closed-loop refinement to increase grasp success
- Extend to generalize on soft or unseen objects
- Investigate on the best ratio of domain randomization to photorealistic data

Extended Readings

- Bridging the Reality Gap for Pose Estimation Networks using Sensor-Based Domain Randomization
- Learning Object Localization and 6D Pose Estimation from Simulation and Weakly Labeled Real Images
- Deep ChArUco: Dark ChArUco Marker Pose Estimation
- Real-Time Object Pose Estimation with Pose Interpreter Networks

Summary

- The paper demonstrates a network trained only on synthetic data that can achieve great performance compared with a network trained on real data on 6-DoF object pose estimation.
- It uses only synthetic data, which shows the promising future for using synthetic data to generate sufficient data in training
- Prior work failed to solve the reality gap or at least need fine-tuning
- The proposed work uses a combination of domain randomized and photorealistic data, achieves great performance, and has high practicability
  (achieves state-of-the-art performance on object pose estimation)

Q & A

Thank you!

Backup Slides


Hand camera not used

Robotic Pick-and-place Experiment