Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects

Presenter: Kunjal Shah

8th November, 2022

Motivation

- **Estimating 6DOF in Robot Manipulation** - an important consideration
- **Difficulties in collecting and labelling data** - for 3D objects
- **Issues with real life data** - cannot efficiently generalize to diverse test data
- **Synthetic data** - bridging the reality gap

Key Insights

The authors present:

- One-shot deep neural network for 3D pose estimation (DOPE) without requiring post-alignment to well estimate poses.
- The authors prove through experiments that a combination of domain randomization and photorealistic data can well generalize to estimating 3D poses.
- Robotic system showing estimated poses to solve real-world tasks

Problem Setting

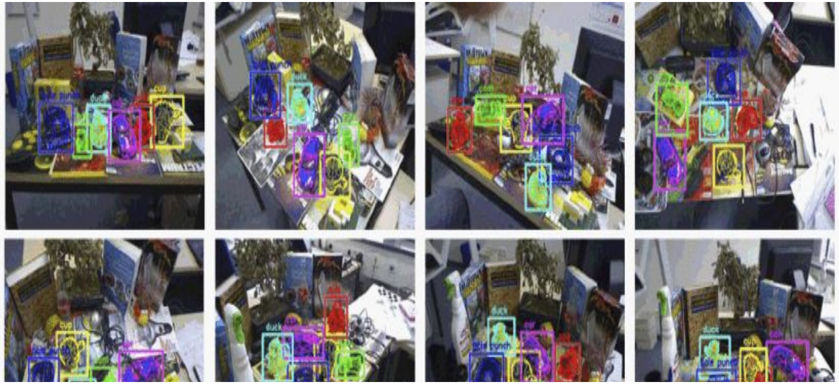
- For 6D object pose estimation, real data is difficult to gather and label.
- Solution-synthetic data with domain randomization and photorealistic data
- Domain Randomization- involves randomizing training data in non-realistic ways
- Photorealistic data- involves combining object models and backgrounds

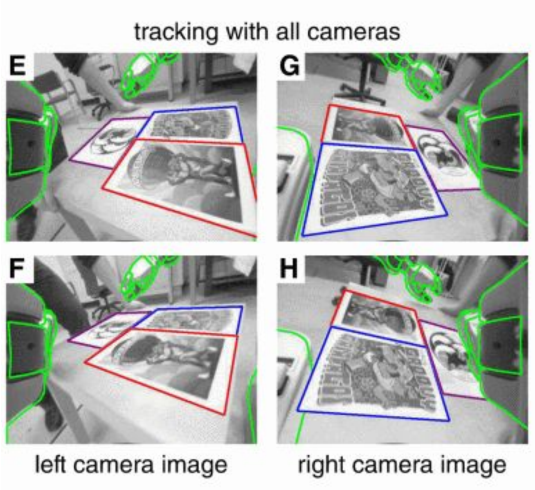
Context / Related Work / Limitations of Prior Work

- Object Detection using 6DOF
    - Traditional CV methods

      -Gradient Response Maps for Real-Time Detection of Textureless Objects

      -SimTrack: A simulation-based framework for scalable real-time object pose detection and tracking



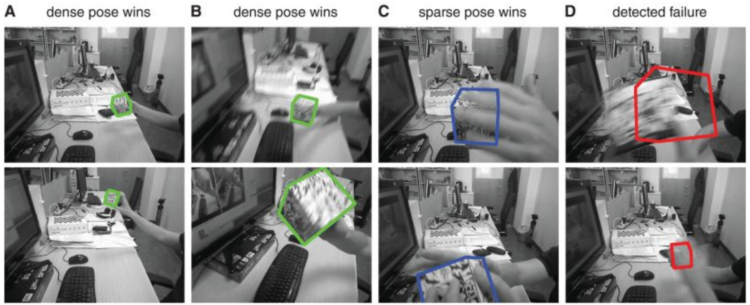


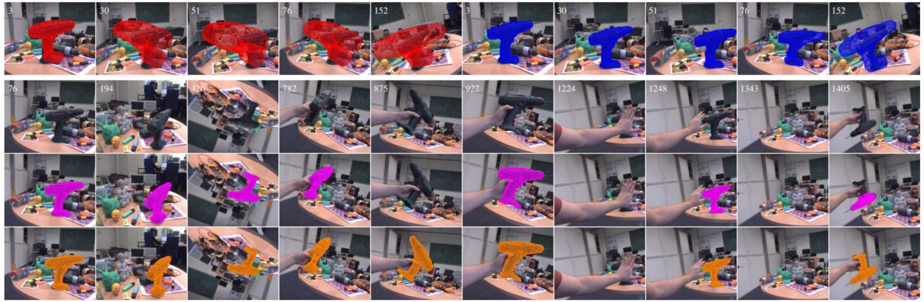
Context / Related Work / Limitations of Prior Work

- Object Detection using 6DOF
  - Traditional CV methods

    -Real-Time Pose Detection and Tracking of Hundreds of Objects

    -Real-Time Monocular Pose Estimation of 3D Objects using Temporally Consistent Local Color Histograms



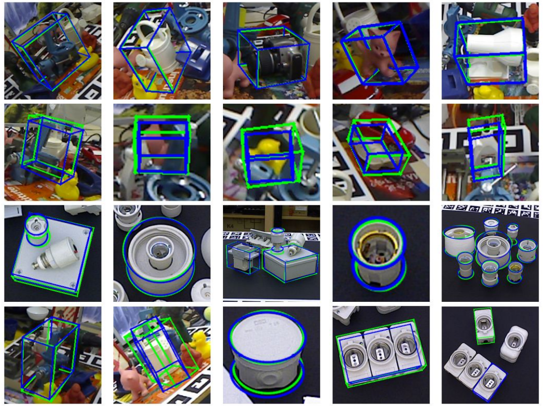


Context / Related Work / Limitations of Prior Work

- Object Detection using 6DOF
  - Deep Learning Methods

    -BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth

    -PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes.



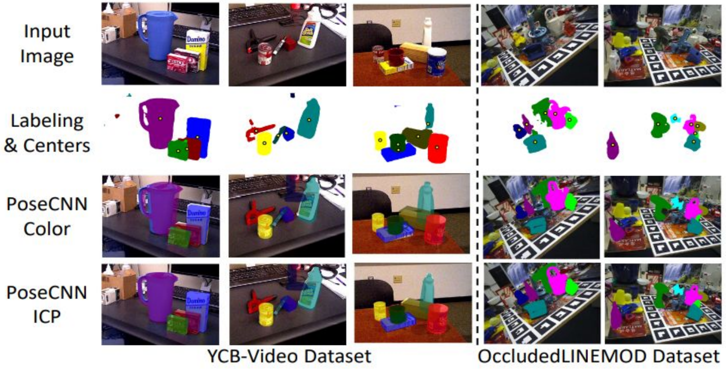


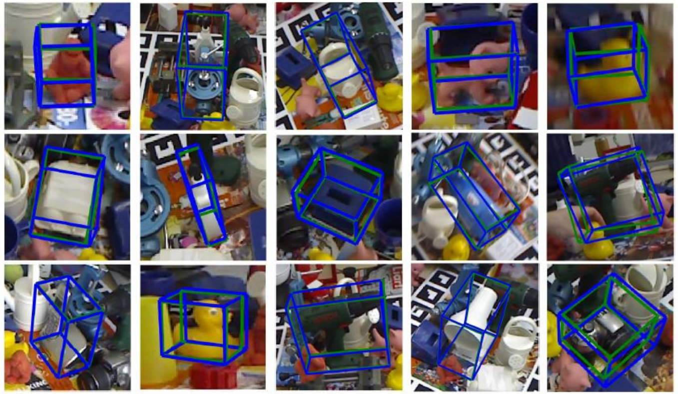
Fig. 9. Examples of 6D object pose estimation results on the YCB-Video dataset from PoseCNN.

Context / Related Work / Limitations of Prior Work

- Object Detection using 6DOF
  - Deep Learning Methods

    - Real-time seamless single shot 6D object pose prediction (YOLO)

    -SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again.



(a) 2D Detections (b) Unrefined (c) RGB refinement (d) RGB-D refinement

Context / Related Work / Limitations of Prior Work

- Synthetic datasets
  - Photorealistic ones require modelling effort

    -Sim4CV: A photo-realistic simulator for computer vision applications

    -Falling things: A synthetic dataset for 3D object detection and pose estimation

  - Domain randomization ones cannot beat state of the art

    -Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World

    -Training deep networks with synthetic data: Bridging the reality gap by domain randomization.

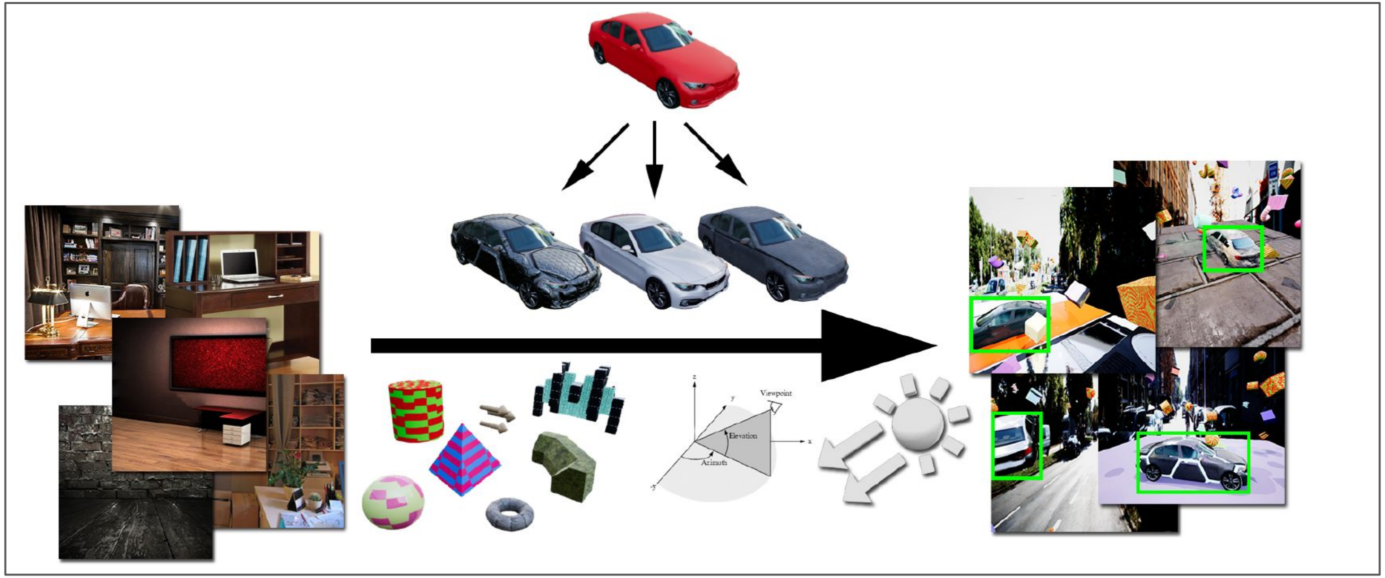
Proposed Approach / Algorithm / Method

- Two-step solution to estimating the 6-DOF pose
- Deep neural network estimates belief maps of 2D object keypoints
- Belief map peaks input to the perspective-n-point (PnP) algorithm for 6-DoF pose estimation.

Generating Data

- The use of synthetic data and YCB objects
- Domain Randomization
    - Foreground objects in virtual environments with randomizations
    - Number & types of distractors, set of 3D models, texture, background, photograph, pose, lights, distractor visibility
- Photorealistic Data
    - Foreground objects in 3D scenes with physical constraints
    - Interactions in physically possible ways

Domain Randomization



Training deep networks with synthetic data: Bridging the reality gap by domain randomization

Our Synthetic Data

Neural Network Architecture

- One shot neural network with multistage architecture to detect keypoints.
- RGB image of size w*h*3 is taken as input, producing 9 belief maps and 8 vector fields.
- Image features are computed by first 10 VGG-19 layers and later on convolutional layers for dimensionality reduction.
- 128-dimensional features input to the 1st stage with three 3*3*128 layers and one 1*1*512 layer, with either a 1*1*9 or a 1*1*16 layer.
- Rest stages have a 153-dimensional input (128+16+9 = 153) and comprise 5 7*7*128 layers and 1 1*1*128 layer before the 1*1*9 or 1*1*16 layer.
- ReLU

Pose Detection and Estimation

- Local peaks in the belief maps are searched, with a greedy assignment algorithm associating projected vertices to detected centroids.
- For all vertices, comparison is done between the vector field evaluated at the vertex with the direction from the vertex to each centroid, assigning the vertex to the closest centroid.
- PnP algorithm used for pose estimation.
- Projected vertices, camera intrinsics, object dimensions used to recover final translation

Experimental Setup

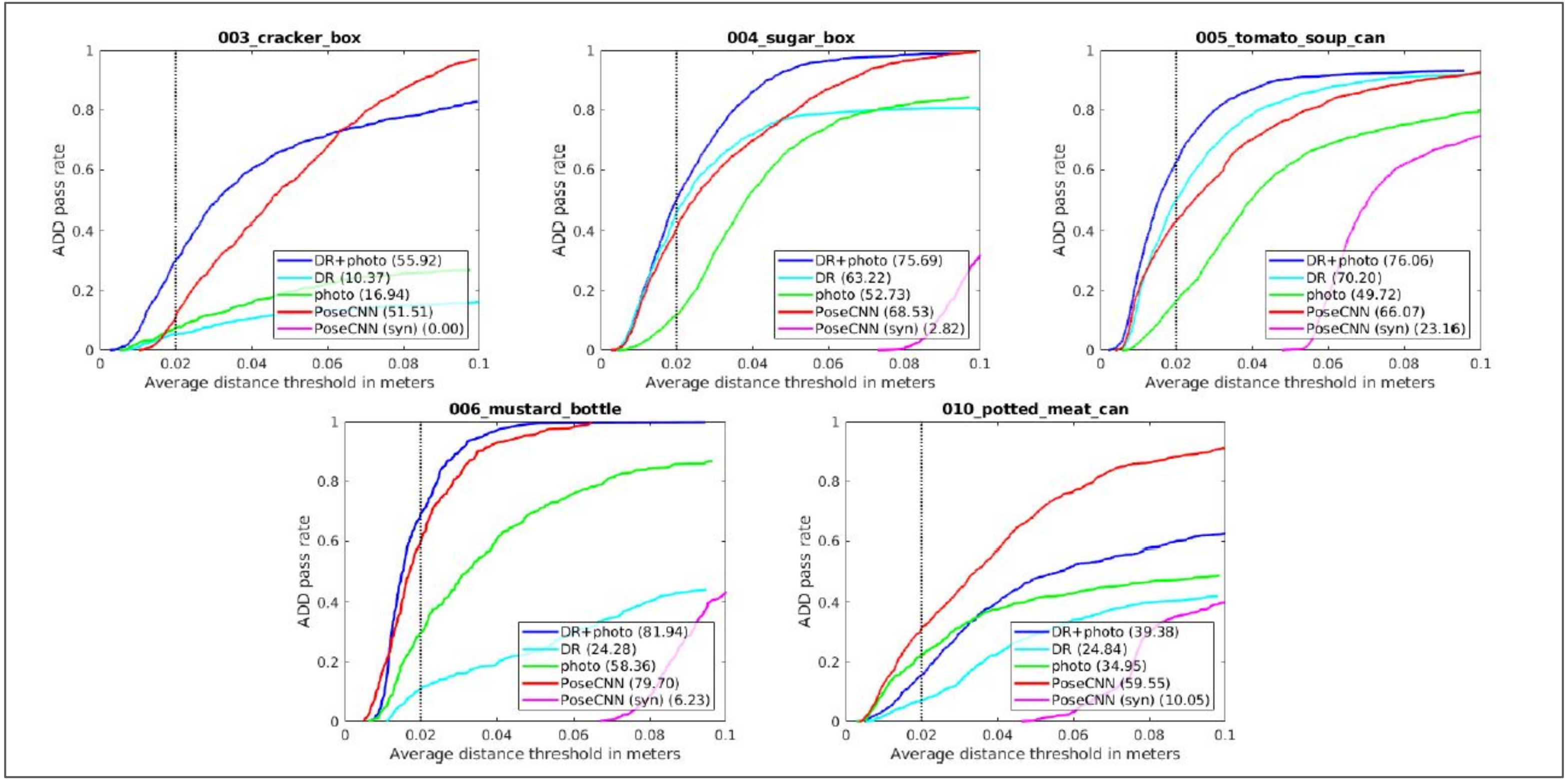
- Databases used:
    - Objects from YCB video dataset (21 objects)
    - self-created dataset: 4 videos with lighting conditions, 5 objects (cracker box, sugar box, tomato soup can, mustard bottle, and potted meat)
- Metric used - Average Distance (ADD) metric - average 3D Euclidean distance of all model between ground truth and estimated pose.

$$\text{ADD} = \frac{1}{m} \sum_{\mathbf{x} \in \mathcal{M}} \|(\mathbf{R}\mathbf{x} + \mathbf{T}) - (\tilde{\mathbf{R}}\mathbf{x} + \tilde{\mathbf{T}})\|,$$

- Training = 60k domain-randomized images, 60k photorealistic
- Baseline is PoseCNN

Objects from the YCB-Video Dataset

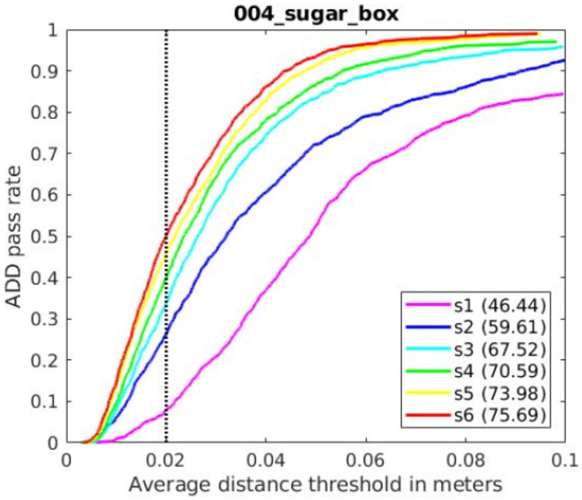
Results



003_cracker_box
DR+photo (55.92)
DR (10.37)
photo (16.94)
PoseCNN (51.51)
PoseCNN (syn) (0.00)
004_sugar_box
DR+photo (75.69)
DR (63.22)
photo (52.73)
PoseCNN (68.53)
PoseCNN (syn) (2.82)
005_tomato_soup_can
DR+photo (76.06)
DR (70.20)
photo (49.72)
PoseCNN (66.07)
PoseCNN (syn) (23.16)
006_mustard_bottle
DR+photo (81.94)
DR (24.28)
photo (58.36)
PoseCNN (79.70)
PoseCNN (syn) (6.23)
010_potted_meat_can
DR+photo (39.38)
DR (24.84)
photo (34.95)
PoseCNN (59.55)
PoseCNN (syn) (10.05)
ADD pass rate
Average distance threshold in meters

Experimental Setup

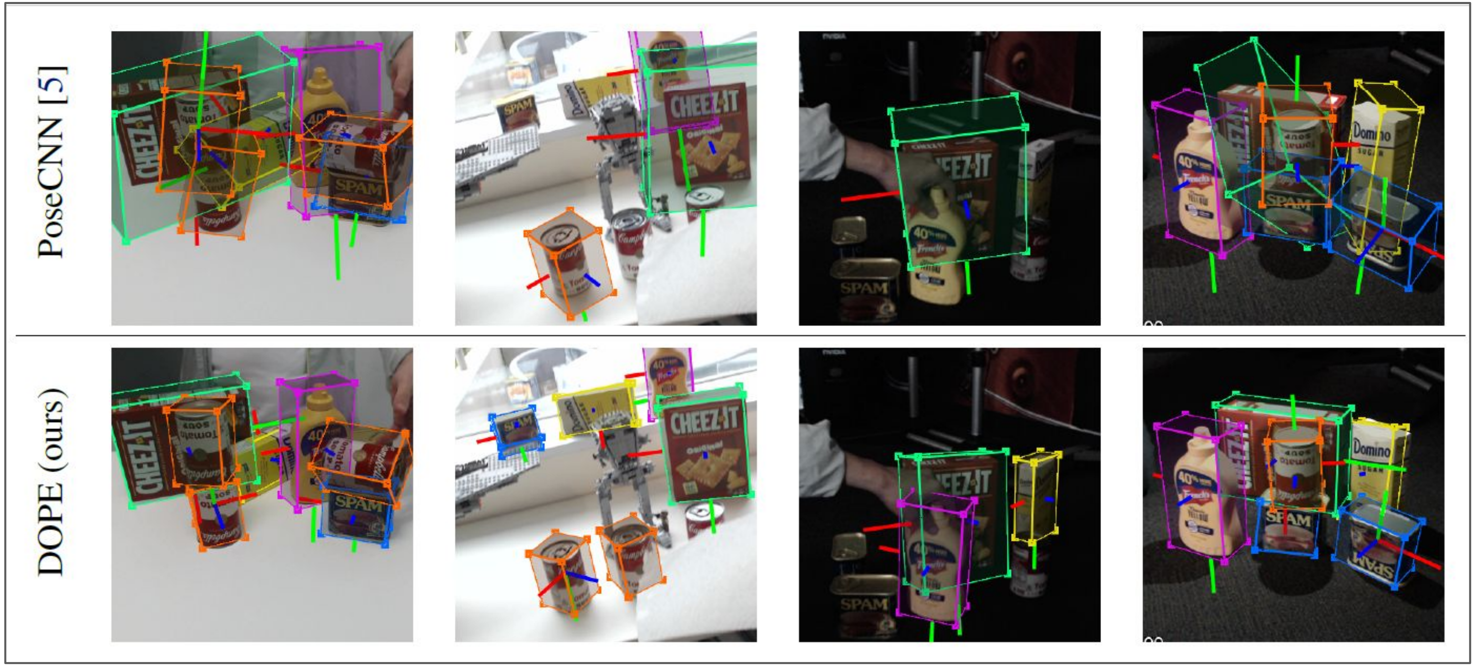
❖ Effect of dataset size and mixing percentages on data: individually on datasets and DR+Photorealistic data



	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

Experimental Setup

- Extreme lighting conditions - comparison between PoseCNN and DOPE

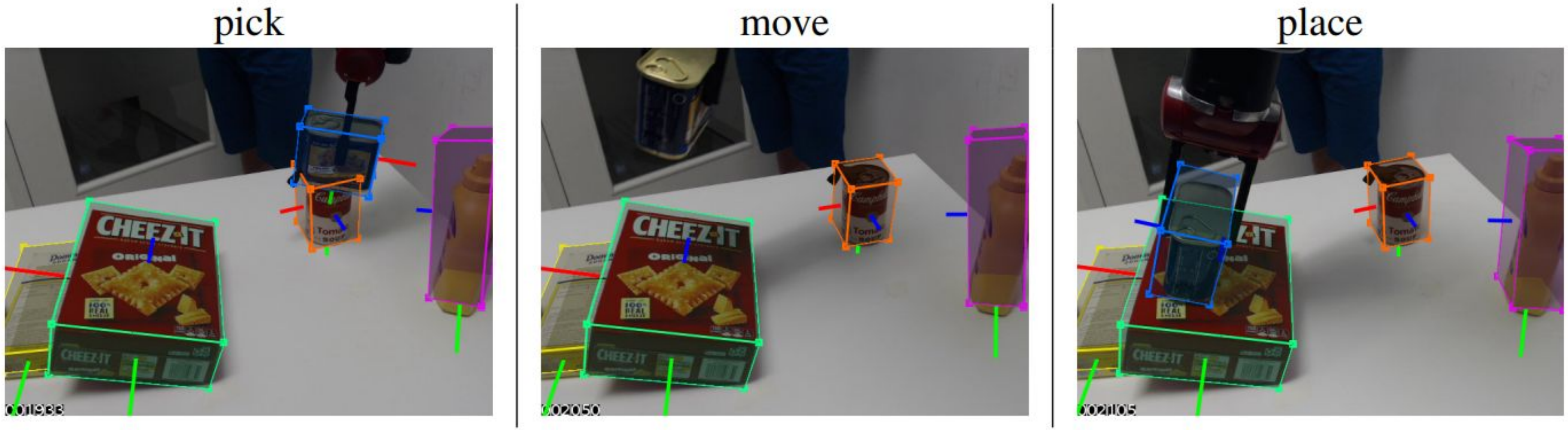


Experimental Results

- Robotic manipulation, Logitech C960 Baxter robot
- Demonstrate that it is robust in real-world conditions.
- 5 objects at different positions
- Robot successfully grasped the 5 objects, results of 12 trials:

Object	Successful Attempts
Cracker	10
Meat	10
Mustard	11
Sugar	11
Soup	7

Experimental Results



Discussion of Results

Conclusions drawn by the authors:

- DOPE trained on synthetic data achieves results similar to PoseCNN on synthetic+real data.
- Use of synthetic data achieves better generalization.
- Mixing domain randomization and photosynthetic data achieves better results than either.
- DOPE is robust and not limited to top-down grasps

- The conclusions are well supported by the results and experiments performed.

Critique / Limitations / Open Issues

- Unreliable in low-resolution or low-texture images, or texture mismatch

  -Self-supervised 6D Object Pose Estimation for Robot Manipulation

Critique / Limitations / Open Issues

- Synthetic data not all-inclusive (does not properly model the highly reflective metallic material)

Critique / Limitations / Open Issues

- Errors in grasping, pose estimation algorithm, miscalibration of devices, imprecise control

  -Comparing model predictive control and input shaping for improved response of low-impedance robots

Critique / Limitations / Open Issues

- Incorporating more objects and symmetry

  -EPOS: Estimating 6D Pose of Objects with Symmetries

Critique / Limitations / Open Issues

- Difficult when location of different objects is not clear

  -An Annotation Saved is an Annotation Earned: Using Fully Synthetic Training for Object Detection

Future Work for Paper / Reading

What interesting questions does it raise for future work?

- ❖ Interactive perception can be used to enhance perception- https://dl.acm.org/doi/pdf/10.5555/3546258.3546288
- ❖ Parameters in domain randomization http://openaccess.thecvf.com/content_CVPR_2020/html/Hodan_EPOS_Estimating_6D_Pose_of_Objects_With_Symmetries_CVPR_2020_paper.html
- ❖ Making synthetic data more inclusive
- ❖ Incorporating closed loop refinement to increase grasp success.

Extended Readings

1. A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms
   (to learn in detail about manipulation, its challenges and various works)
2. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion
   (a novel way for 6D pose estimation fusing complementary data sources)

Summary

- 3D Object grasping
- Important part of robot manipulation, hard due to data collection and labelling
- Prior work either uses real or synthetic data with either domain randomization or photorealistic data
- The authors propose a mixture of both types passed through a neural network and use PnP for pose estimation
- Demonstrated that the techniques help in ease of data collection, generalization and robustness