# Sim2Real Transfer for Object Pose Estimation

Jacob Hoffman, Ethan Li, Antonio Tan-Torres
{jacobmh, ethanli, tantonio}@stanford.edu - March 2018

## Problem Statement

- Estimate 6D pose of query object in cluttered scene
  - Inputs: RGB-D image of scene, RGB-D image of object
  - Output: X,Y, Z, Roll, Pitch, Yaw of object
- Train model on simulated dataset and evaluate on real-world dataset

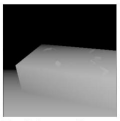
## Motivation

- Real-world data for deep learning approaches to robotic object manipulation: difficult & expensive to obtain
- Deep neural networks trained on simulated robotics tasks may not transfer directly to real-world tasks
- Domain randomization in simulation improves transferability (1)

## Simulated Data

- 4000 RGB-D scenes of objects on a table simulated in Unity3D
- 58800 examples of an object's pose in randomized scene
- 55 objects and their RGB-D images


RGB image of scene


Depth image of scene


RGB image of object


Depth image of object

## Real Data

- Shelf Dataset from Amazon Picking Challenge (2)
- 2130 examples of an object's pose in a cluttered scene of a shelf
- 27 objects and their RGB-D images
- Shelf Dataset provides object-to-world transformation matrix, and camera extrinsic matrix, which we use to extract the 6D pose of object from the camera's perspective (same as Sim Dataset)


RGB image of scene


Depth image of scene

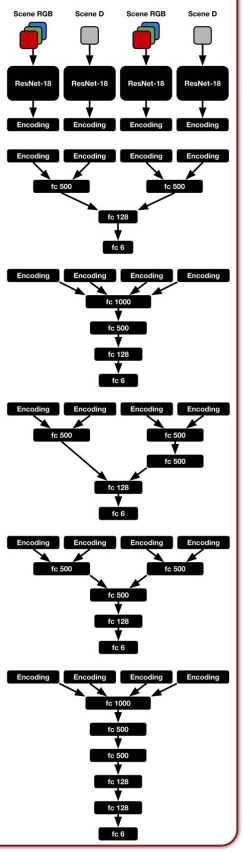

RGB image of object


Depth image of object

## Models

### ResNet Encodings

- Transfer learning with fixed CNN as fixed feature extractor
- ResNet-18 pre-trained on ImageNet
  - Last hidden layer for encodings
  - Depth images triplicated as input




### Model A

- Baseline: hierarchical combination
- Architecture selection MSE losses: train 4.83, train-dev: 5.56

### Model B

- Immediate combination, pass through extra layer to output
- Architecture selection MSE losses: train 1.76, train-dev: 6.3

### Model C

- Hierarchical combination, but pass scene through extra layer
- Architecture selection MSE losses: train 4.91, train-dev: 5.52

### Model D

- Hierarchical combination, pass through extra layer to output
- Architecture selection MSE losses: train 4.15, train-dev: 5.74

### Model E

- Immediate combination, pass through three extra layers to output
- Architecture selection MSE losses: train 2.15, train-dev: 6.10

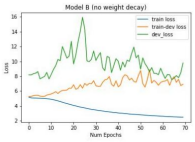
## Experiments

Architecture Selection:

- Trained for 100 epochs on 3.5% of dataset, minibatch size 10
- MSE loss as sum of position and orientation losses
- **Model B** achieved lowest bias

Train Set Size Analysis:

- Trained model B for 70 epochs
- Changed minibatch size to 40
- Losses with 7% of dataset: train 0.44, train-dev 6.56
- **Losses with full dataset: train 2.48, train-dev 6.91, dev 9.78**
- Moderate variance reduction



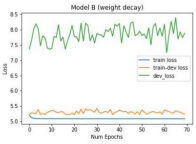
Training with full dataset, λ = 0

L2 Penalty Hyperparameter Selection:

- Trained for 50 epochs on 10% of dataset
- $\lambda = 10^{-1}$: train 5.14, train-dev 5.18
- **$\lambda = 10^{-2}$**: train 4.22, train-dev 6.97
- $\lambda = 10^{-3}$: train 1.74, train-dev 6.97

Full Dataset Training:

- Trained model B for 70 epochs
- **Losses: train 4.94, train-dev 5.27, dev 8.33, test 8.13**
- Position loss: train-dev 0.85 (RMSE 0.92 m), test 2.2 (RMSE 1.47 m)
- Orientation loss: train-dev 4.4 (RMSE 120 deg) test 6.0 (RMSE 140 deg)
- L2 penalty reduces sim2real mismatch



Full training, λ = 0.01

## Challenges and Next Steps

- L2 penalty search with full dataset training needed
- Sim2real transfer may require further domain randomization
- Final layer of ResNet may not provide appropriate encoding for pose estimation task
- Providing single viewpoint of query object may be too difficult, especially for orientation estimation
- End-to-end pose estimation may require more complex architecture or multi-stage approach, as in (3)

**References**

1. J Tobin, R Fong, A Ray, J Schneider, W Zaremba, P Abbeel. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. IROS 2017.
2. A Zeng, KT Yu, S Song, D Suo, E Walker Jr., A Rodriguez, J Xiao. Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge. ICRA 2017.
3. Y Xiang, T Schmidt, V Narayanan, D Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes