Robust Pose Estimation of Boxes from Keypoints

Maggie Wang

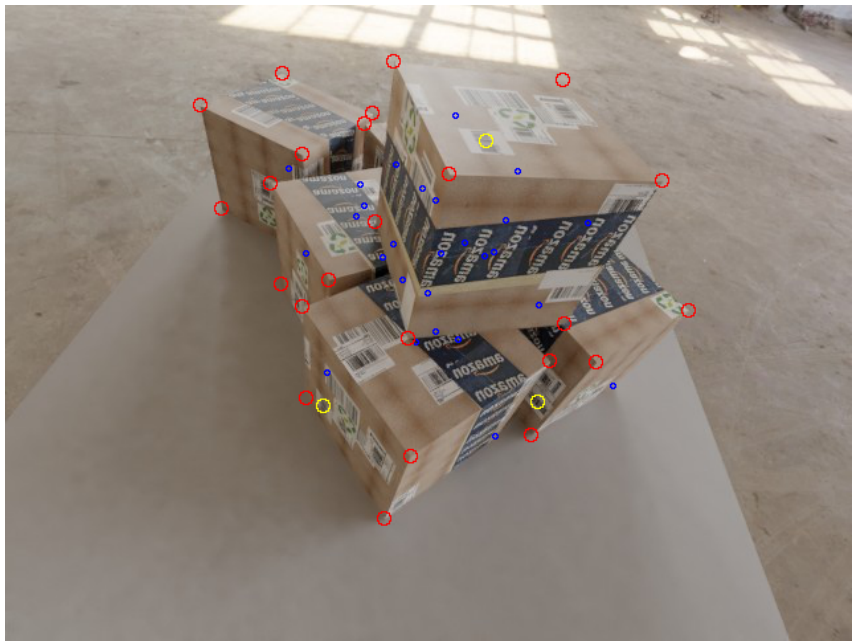
Spring 2021

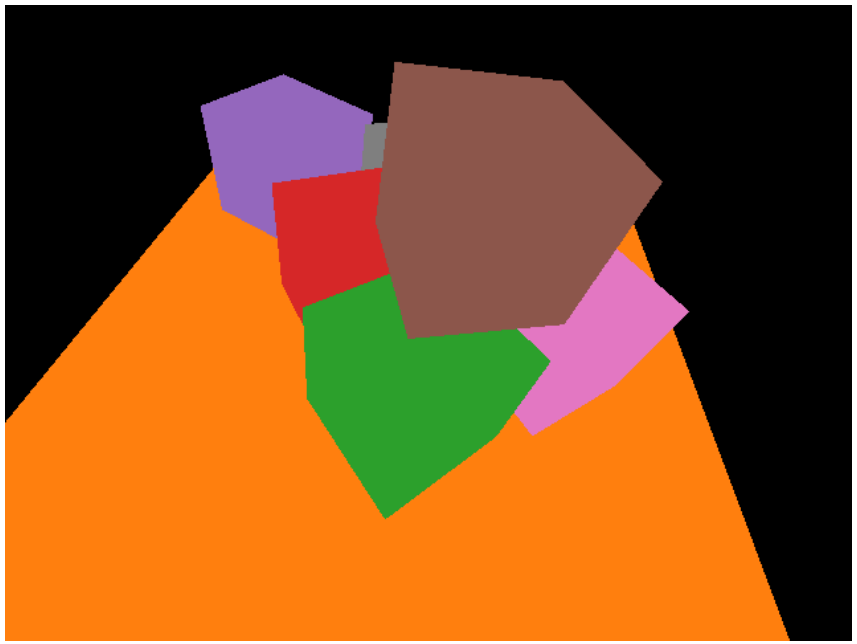
Motivation and Goals

- We would like **robust keypoints** for downstream manipulation tasks
  - Various object categories: **boxes**, mugs, shoes, etc.
  - Extract **pose and shape estimation** of boxes
- Search for **counterexamples**
  - Use distribution over scene parameters
  - Compare with black-box optimization methods
- Use counterexamples to **improve the perception pipeline**
  - Making interpretable improvements to the network and post-processing steps

Box Scene Simulation Pipeline

- Generated box scenes in Blender and Drake
- Mask R-CNN for instance segmentation

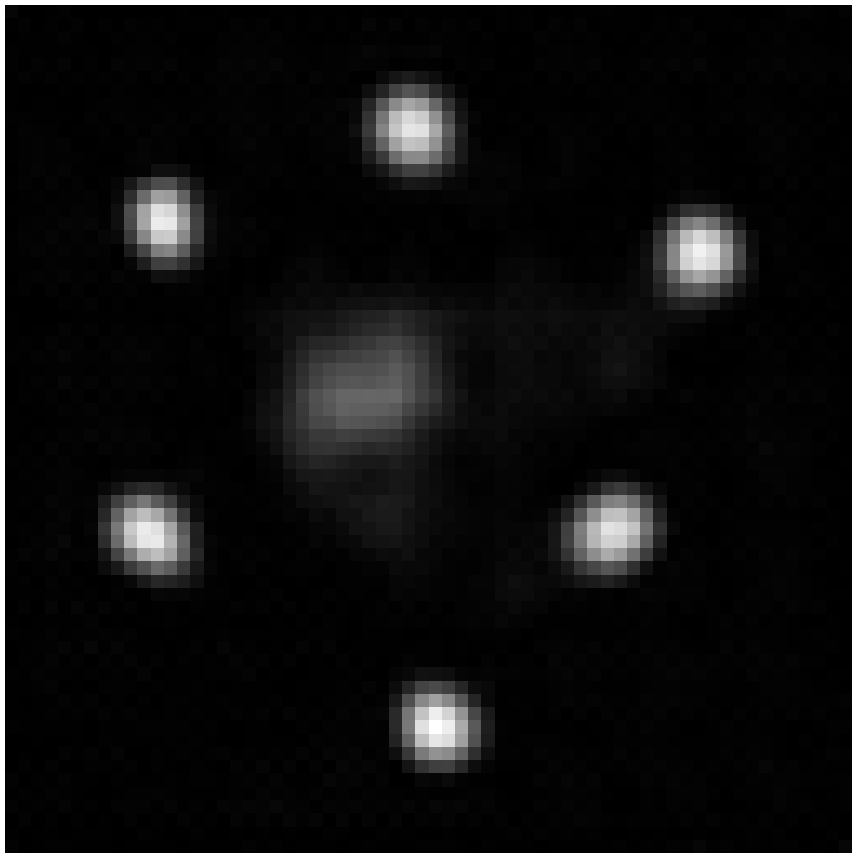




Box Scene Simulation Pipeline: Network

- ResNet 34-backbone and integral regression network
- Network input is an RGBD image, output is a xy-depth probability heatmap, where each pixel is the probability of a keypoint being at that point
- MSE loss between the target Gaussian kernel heatmap and the network prediction

Non-Maximum Suppression of 2D Points: Motivation to Jointly Estimate Corners



Jointly Estimating Box Pose and Shape from Heatmap

- Using 2D heatmap:
  - **Nonlinear optimization** - kind of worked, but had failure cases that were hard to resolve
  - **MIP** - non-convex constraints
- Using 3D heatmap:
  - **NMS and ICP** over all of the correspondences - works well

Joint Optimization to Detect Keypoints from 2D Heatmap

- Nonlinear optimization problem:

$$\begin{aligned} \min_{X} \quad & \left\| H - \sum_{i=0}^{7} k(PX \begin{bmatrix} p_i \\ 1 \end{bmatrix}) \right\|^2 \\ \text{s.t.} \quad & -0.3 \leq x, y \leq 0.3 \\ & 0.0 \leq z \leq 0.3 \end{aligned}$$

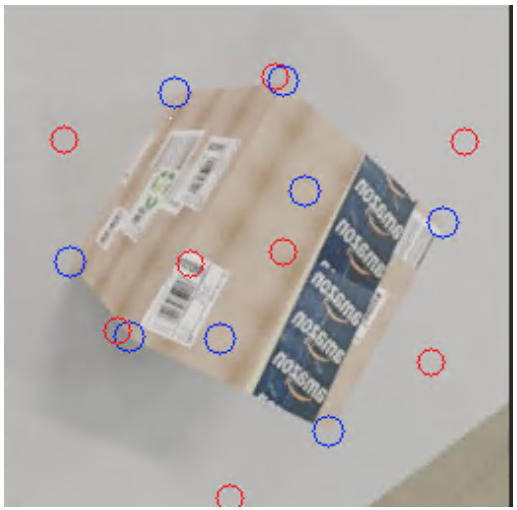
where

- $X \in \mathrm{Sim}(3) \subset \mathbb{R}^{4\times4}$: similarity transformation (translation (x, y, z), rotation (r, p, y), scale (a, b, c))
- $H \in \mathbb{R}^{64\times64}$: combined heatmap from the model output
- $k()$: returns a Gaussian heatmap given a kernel location
- $P \in \mathbb{R}^{4\times4}$: projects from the world to camera frame
- $p_i \in \mathbb{R}^3$: corner i of unit box

Using 2D Heatmap for Nonlinear Optimization to Jointly Estimate Box Pose

(a) Corners found with NMS

(b) Corners found with optimization



(c) Example of failure in joint optimization

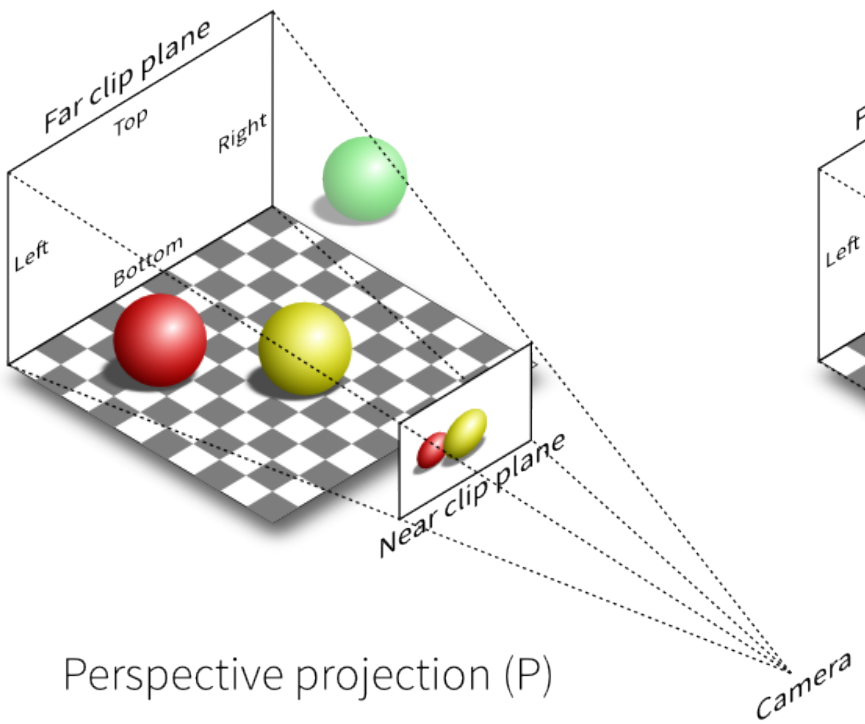
Red points are predicted box corners and blue points are ground truth box corners.

MIP: Perspective vs Orthogonal Projection

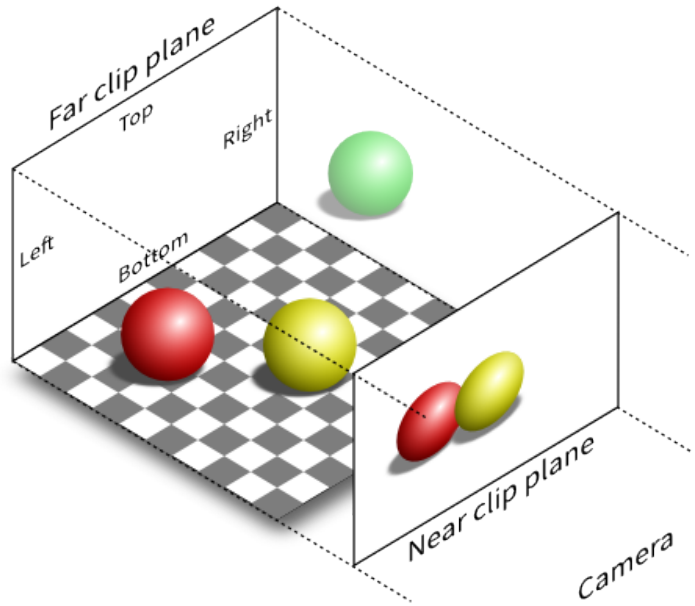
- Blocked by perspective projection step, which makes the pose estimation problem non-linear
- **Nonlinear in decision variables t and R**:
  - Let M be the 3×4 homogeneous matrix used to transform a 3D unit box corner into the 2D camera frame. If p is the homogeneous unit box coordinate (4×1), then Mp is the projected point in the 2D image. M is given by PO, where P is the camera matrix (3×4) and O is the box transform in camera frame (4×4).
  - Let $K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ be a matrix that removes the depth from Mp. So we have that KMp is a 2×1 point in the image frame. Since we have to divide by the depth in perspective projection, we divide the elements of KMp by the depth given by kMp, where $k = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$ extracts the depth.
  - The constraint would be $KMp \cdot (kMp)^{-1} - m_j$. Since $M = PO$ where O is composed of t and R through O, this is also nonlinear in our decision variables t and R.

Perspective vs Orthogonal Projection

- Using orthogonal projection instead would be an approximation (especially at shorter focal lengths)



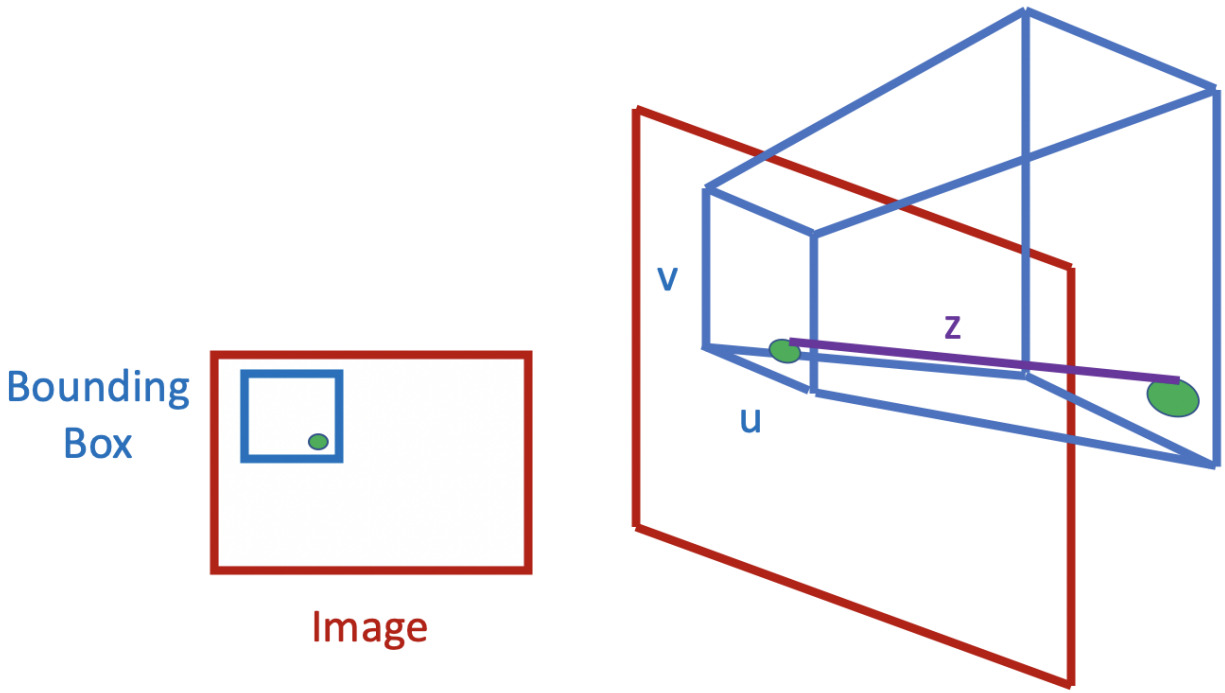
Perspective projection (P)



Orthographic projection (O)

Detection with 3D Heatmap

- Network takes in RGBD image and ground truth 3D gaussian heatmap (64x64x64) for training
- Non-maximum suppression (similar to 2D): find the maximum gaussian kernel and subtract it from the total heatmap to get eight total points

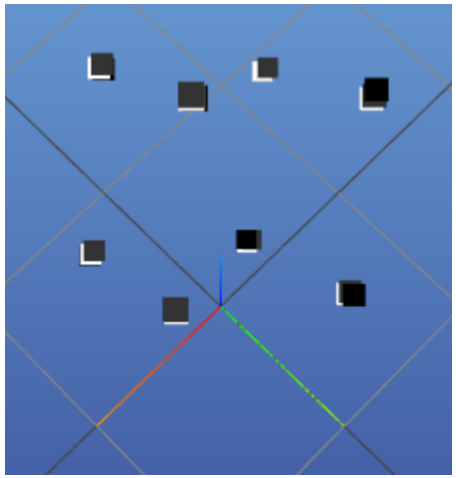


ICP for Box Pose Estimation

- Tried correspondences of all 8! permutations, used only five points for ICP. Parallelized in one batch pass using torch.
- Currently not finding scale (scaling ICP), not a shape estimate

$$\min_{R,t,c} \sum_i \| Rs_i + t - m_{c_i} \|_2^2$$

$$\text{s.t. } R^T = R^{-1}, \det\{R\} = 1$$



Ground truth (white), NMS output (black), ICP output (gray)

Error Metric

- Nearest-neighbor point-to-point distance between transformed unit box corner points and ground truth keypoints in camera frame

$$\text{error} = \sum_{i=0}^{8} \left\| R m_{c_i} + t - s_i \right\|_2^2$$

where

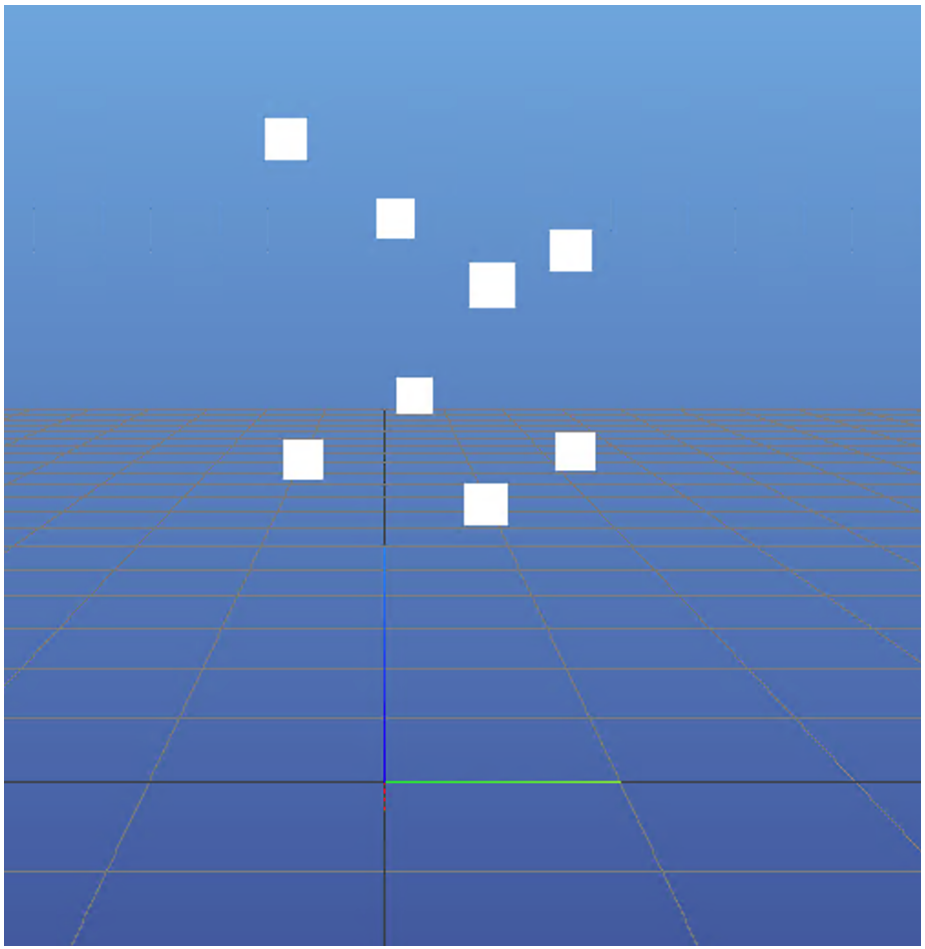
m_{c_i} is unit box corner, where c_i is the integer index correspondence

R is rotation matrix, t is transformation vector

s_i is corresponding nearest-neighbor ground truth keypoint

Failure Cases: Outliers

- Mostly resolved by only using 5 points in ICP
- Will have to be more careful when using scaling ICP

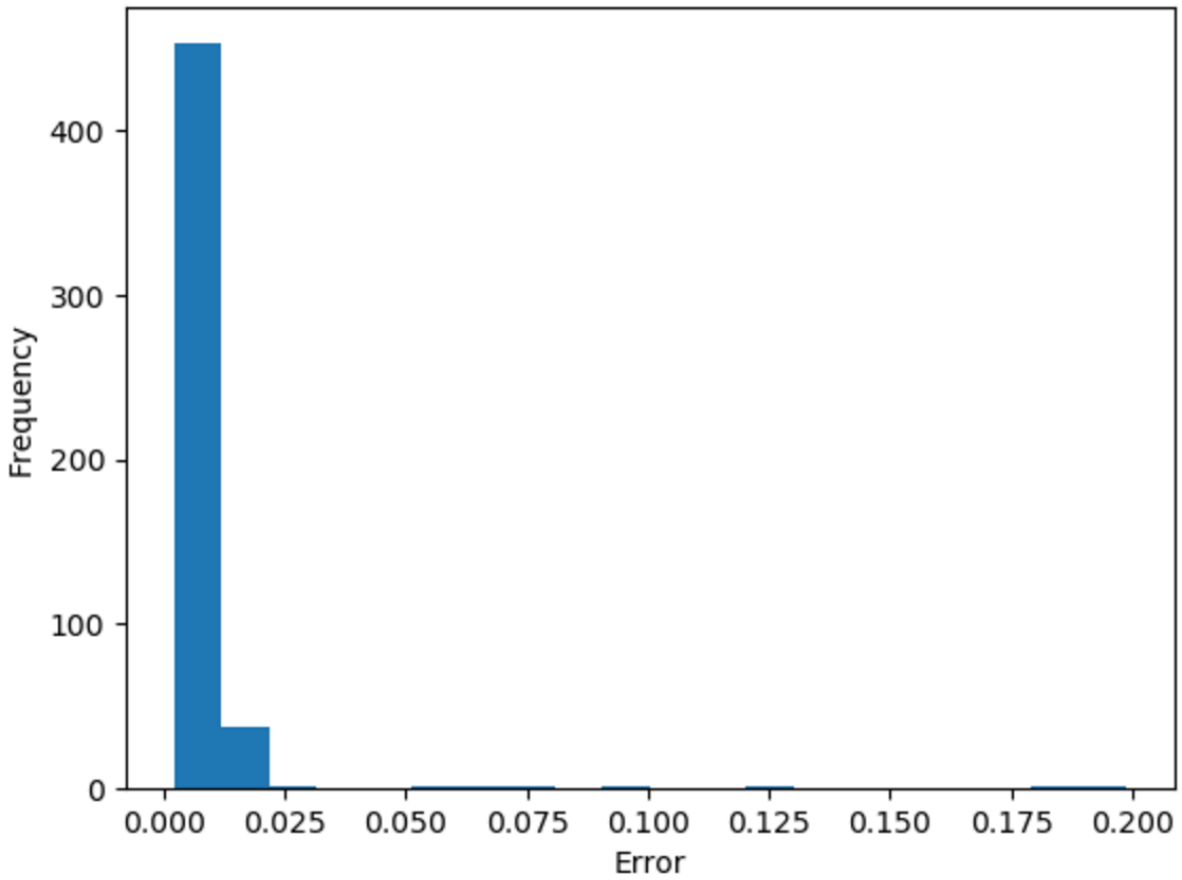


Counterexample Search

- Maximizing error, perturbing box poses ($\mathbb{R}^6$) relative to a fixed 3/4 camera
- Using Matt's Trustworthy AI Monte Carlo mode, where it returns a random sample over the uniform distribution

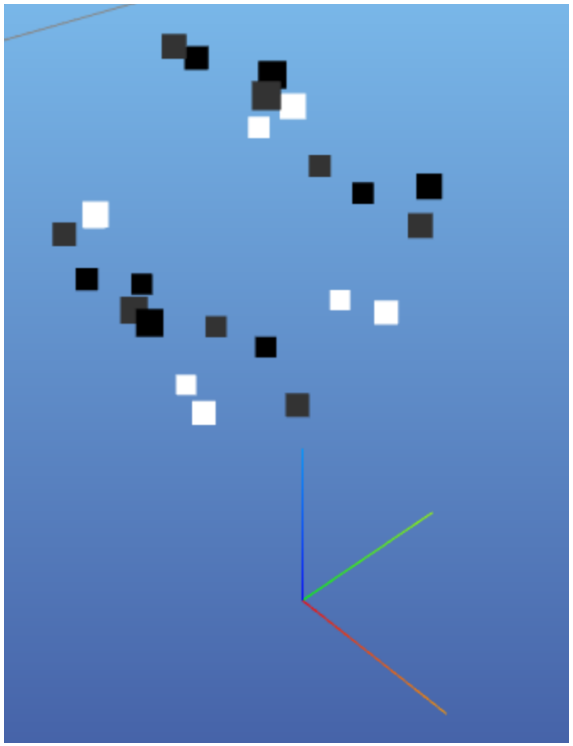
Failure Cases: Histogram

- 500 iterations (would add more but kept crashing every 100 iterations)



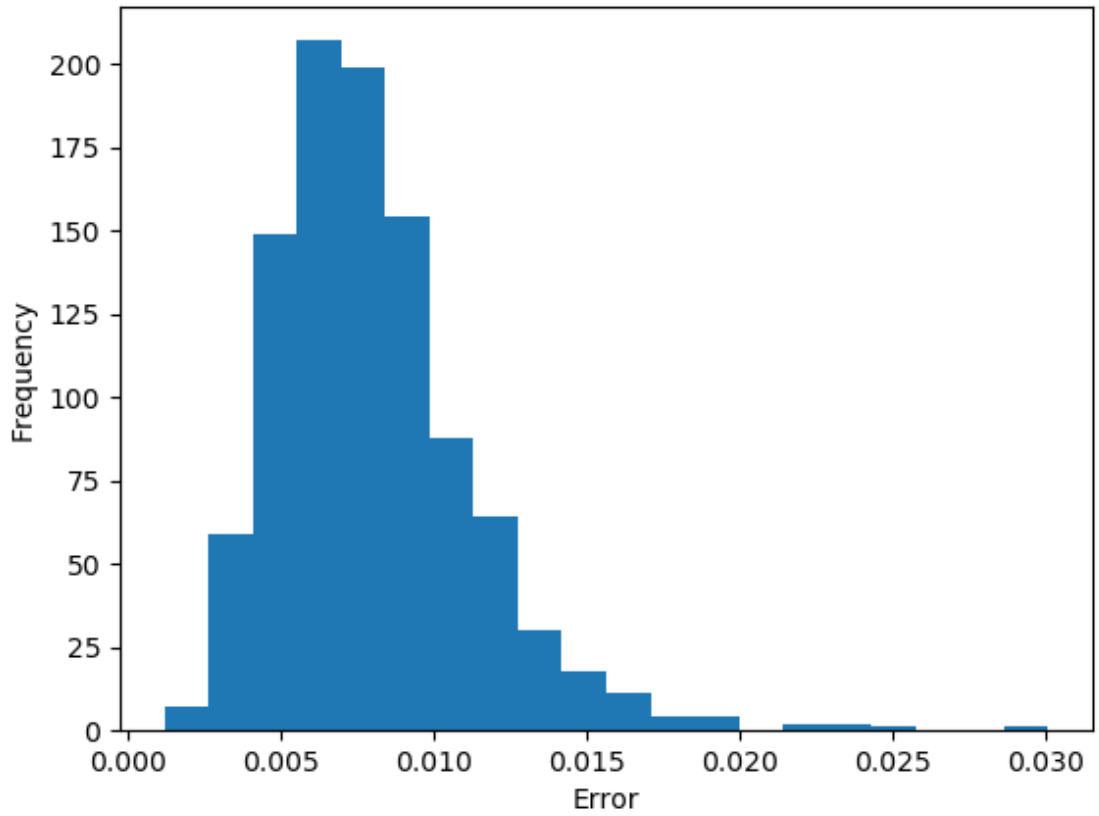
Highest Error Cases

- Counterexample search exposed simulation issues



Histogram of Errors

- 1000 iterations, no failure case yet
- Will try with order(s) of magnitude more samples



Keypoint Representation

- Larger error if camera is straight above box (which is the case for Amazon dataset)
- A different representation might be better in this case (or more shape information is needed)

Future Work

- Scaling ICP to get box shapes along with pose estimation
- Risk/adaptive Monte Carlo search for failure cases, which prioritizes coverage over simply finding a point that minimizes/maximizes a function
- Multiple boxes for more cluttered scenes
- Testing on real-world Amazon data
- Retraining with counterexamples (depending on failure type)
- Better visualization of similarity between failure cases
- Quantifying distributional robustness over scene graphs to detect out-of-distribution scenes