Physics-aware Simulation for Object Detection and Pose Estimation

Chaitanya Mitash, Kostas E. Bekris and Abdeslam Boularias
Department of Computer Science,
Rutgers University

Presented by: Zakary Littlefield (Rutgers University)

Motivation

- Object Detection and 6D pose estimation is an integral part of any Robotic Manipulation.

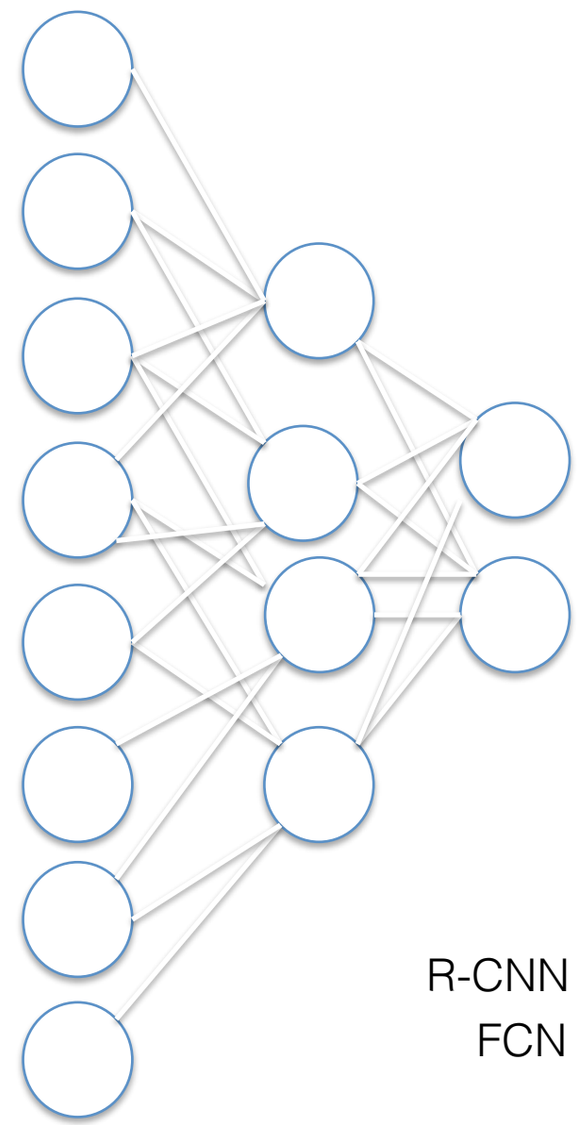
- State-of-the-art methods in pose estimation make use of Convolutional Neural Networks(CNNs) for object recognition.

- CNNs need access to large amount of labeled training data, which needs intensive human labor.

- The techniques used for generating synthetic data suffers from dataset bias.

Motivation

- Object Detection and 6D pose estimation is an integral part of any Robotic Manipulation.
- State-of-the-art methods in pose estimation make use of Convolutional Neural Networks(CNNs) for object recognition.
- CNNs need access to large amount of labeled training data, which needs intensive human labor
- The techniques used for generating synthetic data suffers from dataset bias.



R-CNN [1]
FCN [2]

Motivation

- Object Detection and 6D pose estimation is an integral part of any Robotic Manipulation.
- State-of-the-art methods in pose estimation make use of Convolutional Neural Networks(CNNs) for object recognition.
- CNNs need access to large amount of labeled training data, which needs intensive human labor
- The techniques used for generating synthetic data suffers from dataset bias.

ImageNet

MS-COCO

Rutgers APC RGB-D Dataset

Berkeley BigBIRD dataset

Princeton's "Shelf & Tote" Benchmark

Motivation

- Object Detection and 6D pose estimation is an integral part of any Robotic Manipulation.
- State-of-the-art methods in pose estimation make use of Convolutional Neural Networks(CNNs) for object recognition.
- CNNs need access to large amount of labeled training data, which needs intensive human labor.
- The techniques used for generating synthetic data suffers from dataset bias.

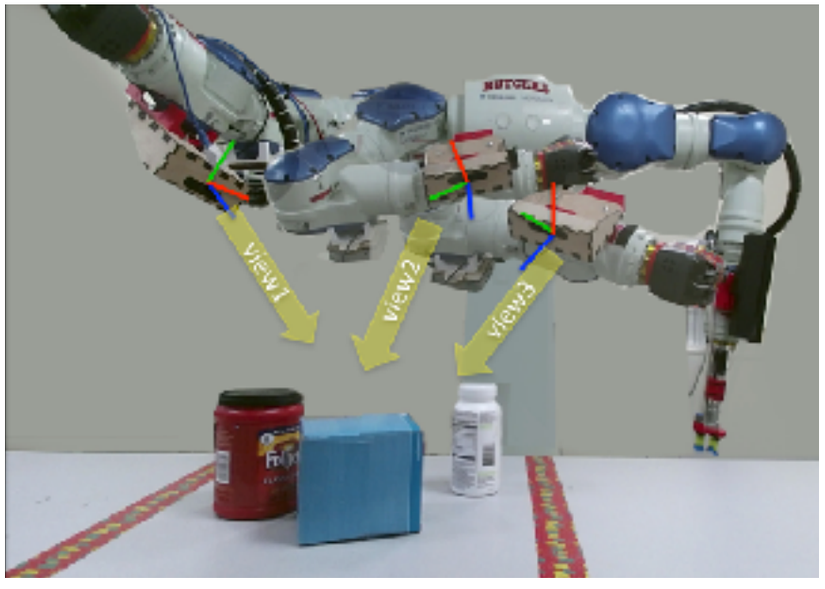
occlusions
texture
illumination
scale
…

Aim

To generate a labeled dataset for Object Detection in a clutter that mimics the data from the sensor.

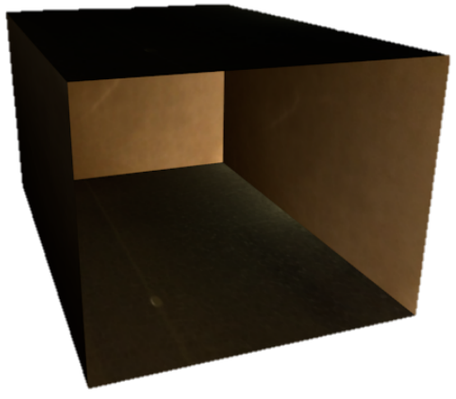
Known

camera configurations



objects to be detected

resting surfaces

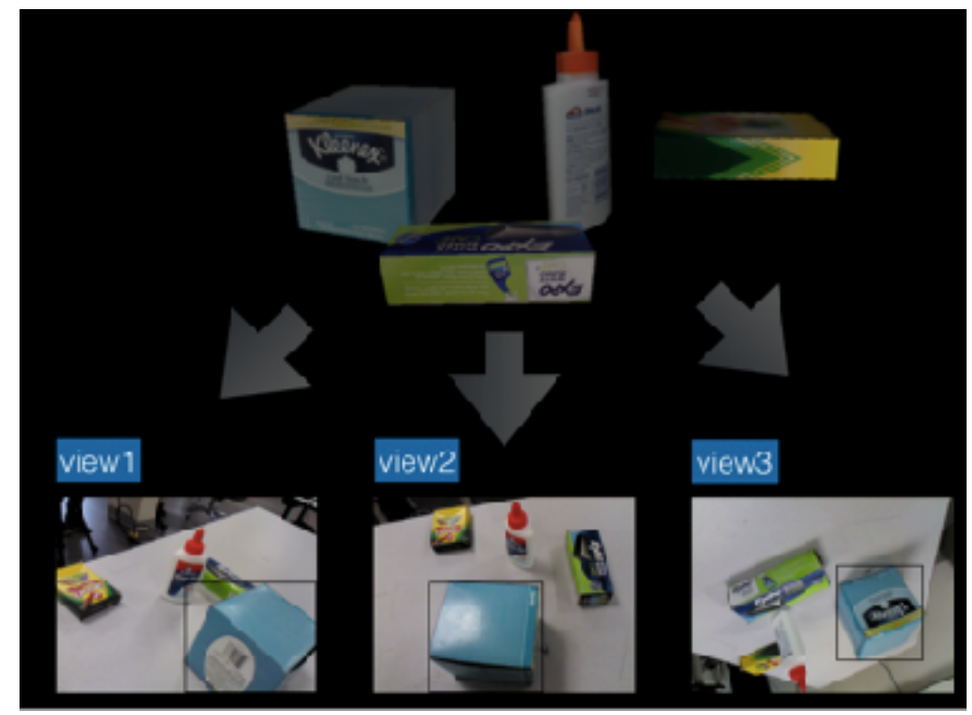


Approach

Physics-aware simulation

Simulation that generates realistic scenes.

Self-Learning



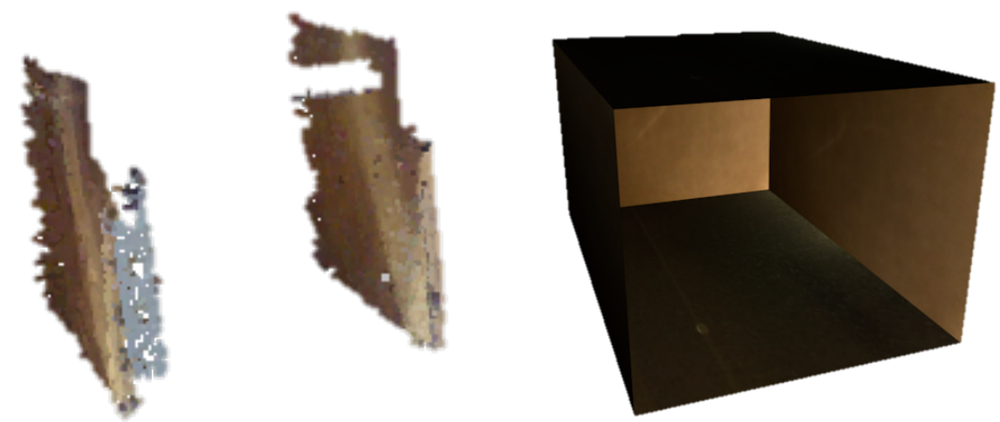
Lifelong learning process that autonomously labels multi-view real scenes

Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.

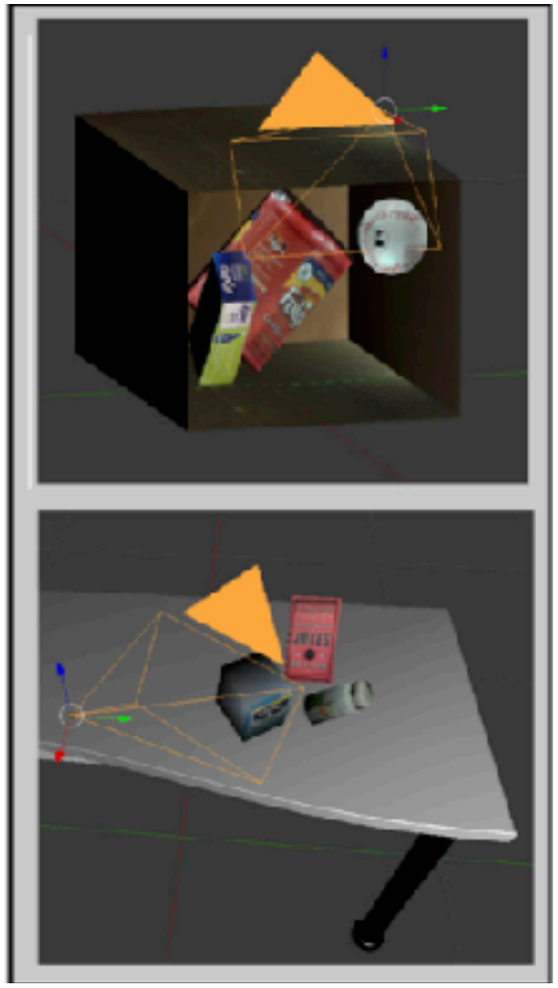
Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.



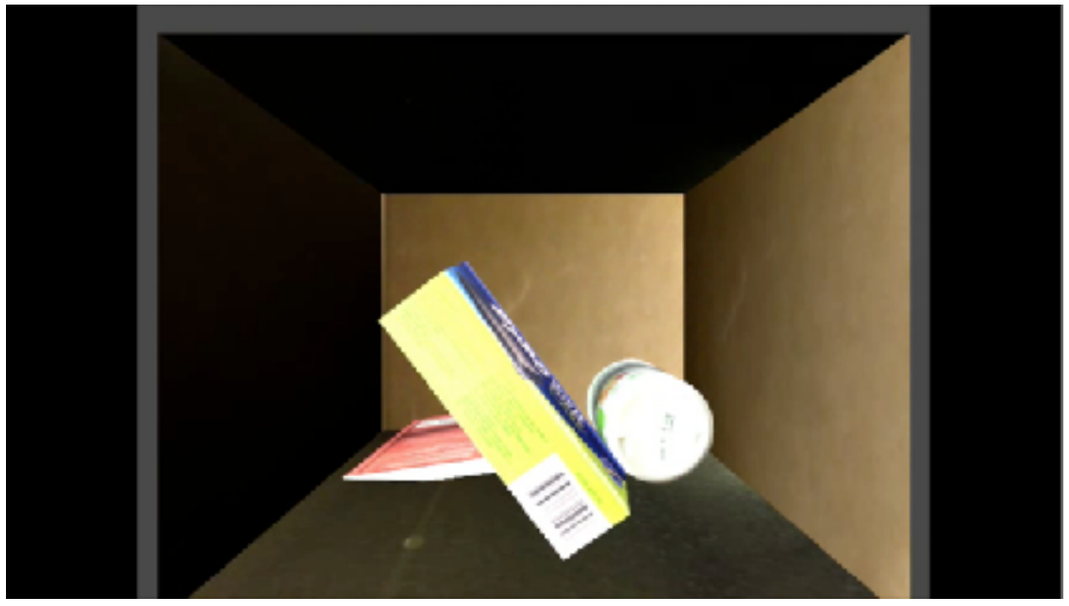
Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.



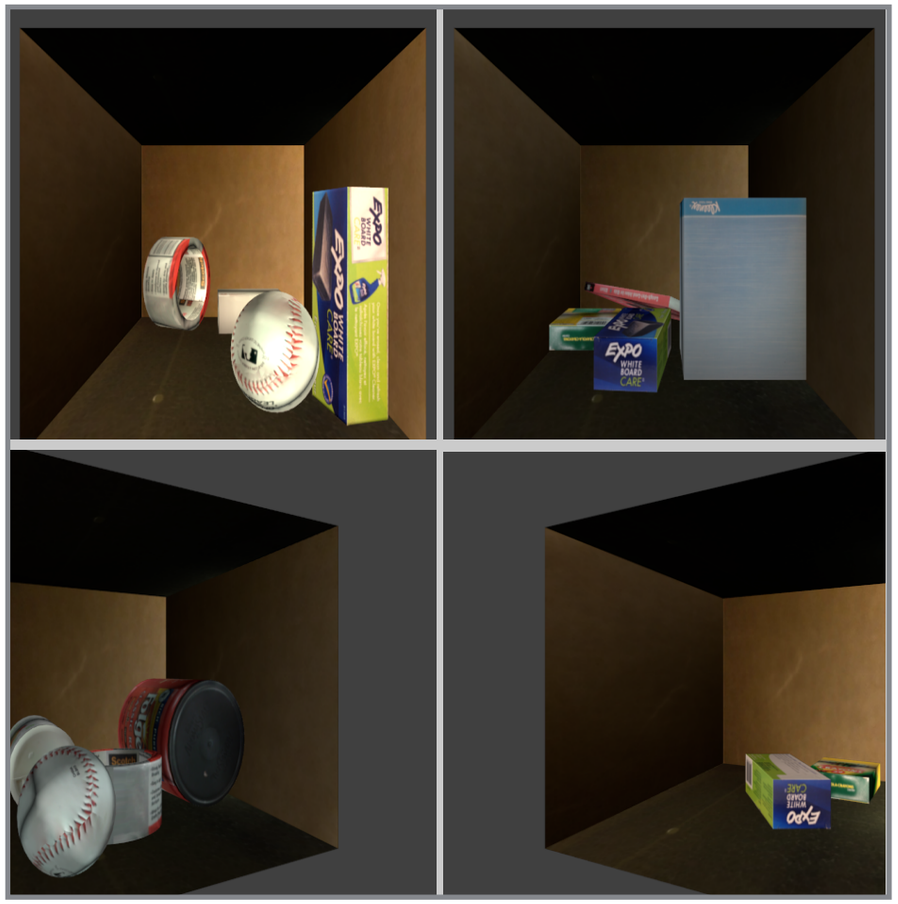
Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.



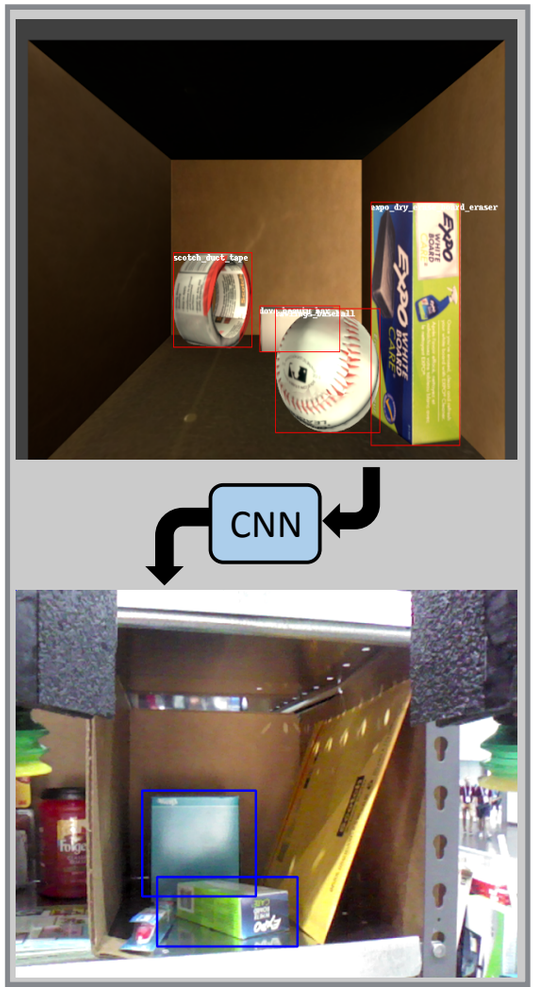
Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.



Physics-aware Simulation

- 3D CAD models are created for the objects and the resting surface.
- Resting surface is calibrated using RANSAC based technique.
- Scenes are generated by randomly choosing object combinations, and locations above the resting surface.
- Physics simulation is used to resolve collisions, unstable configurations and inter-penetrations.
- The final configurations of object represent natural object placements. The scenes are rendered from multiple known camera views.
- 2D Bounding box and class labels are assigned based on the known objects and projective geometry. This dataset is used to train R-CNN and test on real-world images.



Evaluation(Physics Simulation Object Detection)

Results are evaluated on the Princeton's APC "Shelf & Tote" dataset.

Training dataset	success (IoU > 50%)
synthetic data with uniform pose distribution	31%
synthetic data with known pose distribution	69%
physics-aware simulation (Our)	64%
physics-aware simulation, varying illumination (Our)	**70%**

- When using uniform distribution, a large dataset comprising different scales and views are required many of which do not represent natural scenes or views.
- In case the distribution of object poses is known in the test distribution, we can simulate these poses and the results indicate the upper limit of success with respect to varying object poses.
- For physics-aware simulation, training could be performed with small number of physically realistic simulated scenes (~2k in our results).
- The smaller number helps avoid overfitting with respect to texture and scene illumination which is the most significant cause of dataset bias.
- We further train by simulating different illumination conditions.

Evaluation(Physics Simulation Object Detection)

Results are evaluated on the Princeton's APC "Shelf & Tote" dataset.

Training dataset	success (IoU > 50%)
synthetic data with uniform pose distribution	31%
synthetic data with known pose distribution	69%
physics-aware simulation (Our)	64%
physics-aware simulation, varying illumination (Our)	**70%**

- When using uniform distribution, a large dataset comprising different scales and views are required many of which do not represent natural scenes or views.
- In case the distribution of object poses is known in the test distribution, we can simulate these poses and the results indicate the upper limit of success with respect to varying object poses.
- For physics-aware simulation, training could be performed with small number of physically realistic simulated scenes (~2k in our results).
- The smaller number helps avoid overfitting with respect to texture and scene illumination which is the most significant cause of dataset bias.
- We further train by simulating different illumination conditions.

Evaluation(Physics Simulation Object Detection)

Results are evaluated on the Princeton's APC "Shelf & Tote" dataset.

Training dataset	success (IoU > 50%)
synthetic data with uniform pose distribution	31%
synthetic data with known pose distribution	69%
physics-aware simulation (Our)	64%
physics-aware simulation, varying illumination (Our)	**70%**

- When using uniform distribution, a large dataset comprising different scales and views are required many of which do not represent natural scenes or views.
- In case the distribution of object poses is known in the test distribution, we can simulate these poses and the results indicate the upper limit of success with respect to varying object poses.
- For physics-aware simulation, training could be performed with small number of physically realistic simulated scenes (~2k in our results).
- The smaller number helps avoid overfitting with respect to texture and scene illumination which is the most significant cause of dataset bias.
- We further train by simulating different illumination conditions.

Evaluation(Physics Simulation Object Detection)

Results are evaluated on the Princeton's APC "Shelf & Tote" dataset.

Training dataset	success (IoU > 50%)
synthetic data with uniform pose distribution	31%
synthetic data with known pose distribution	69%
physics-aware simulation (Our)	64%
physics-aware simulation, varying illumination (Our)	**70%**

- When using uniform distribution, a large dataset comprising different scales and views are required many of which do not represent natural scenes or views.
- In case the distribution of object poses is known in the test distribution, we can simulate these poses and the results indicate the upper limit of success with respect to varying object poses.
- For physics-aware simulation, training could be performed with small number of physically realistic simulated scenes (~2k in our results).
- The smaller number helps avoid overfitting with respect to texture and scene illumination which is the most significant cause of dataset bias.
- We further train by simulating different illumination conditions.

Evaluation(Physics Simulation Object Detection)

Results are evaluated on the Princeton's APC "Shelf & Tote" dataset.

Training dataset	success (IoU > 50%)
synthetic data with uniform pose distribution	31%
synthetic data with known pose distribution	69%
physics-aware simulation (Our)	64%
physics-aware simulation, varying illumination (Our)	**70%**

- When using uniform distribution, a large dataset comprising different scales and views are required many of which do not represent natural scenes or views.
- In case the distribution of object poses is known in the test distribution, we can simulate these poses and the results indicate the upper limit of success with respect to varying object poses.
- For physics-aware simulation, training could be performed with small number of physically realistic simulated scenes (~2k in our results).
- The smaller number helps avoid overfitting with respect to texture and scene illumination which is the most significant cause of dataset bias.
- We further train by simulating different illumination conditions.

Self-Learning(Motivation)

- complex properties like material reflectivity, type of light source, light intensity and position might not have been captured in the simulation based training.

- To overcome this, the system autonomously labels real-world scenes and adds these examples to the training dataset.

Self-Learning(Motivation)

- complex properties like material reflectivity, type of light source, light intensity and position might not have been captured in the simulation based training.

- To overcome this, the system autonomously labels real-world scenes and adds these examples to the training dataset.

Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.

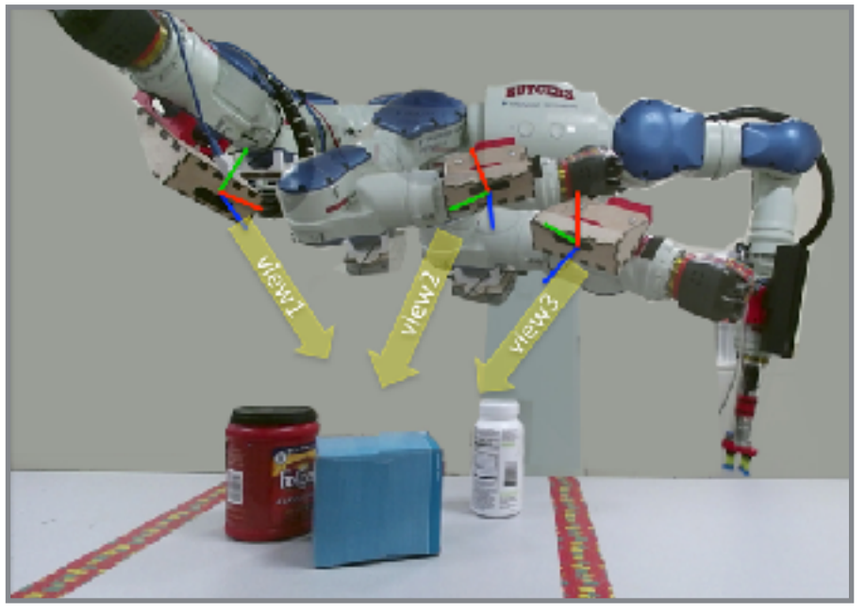
- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.

- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.

- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).

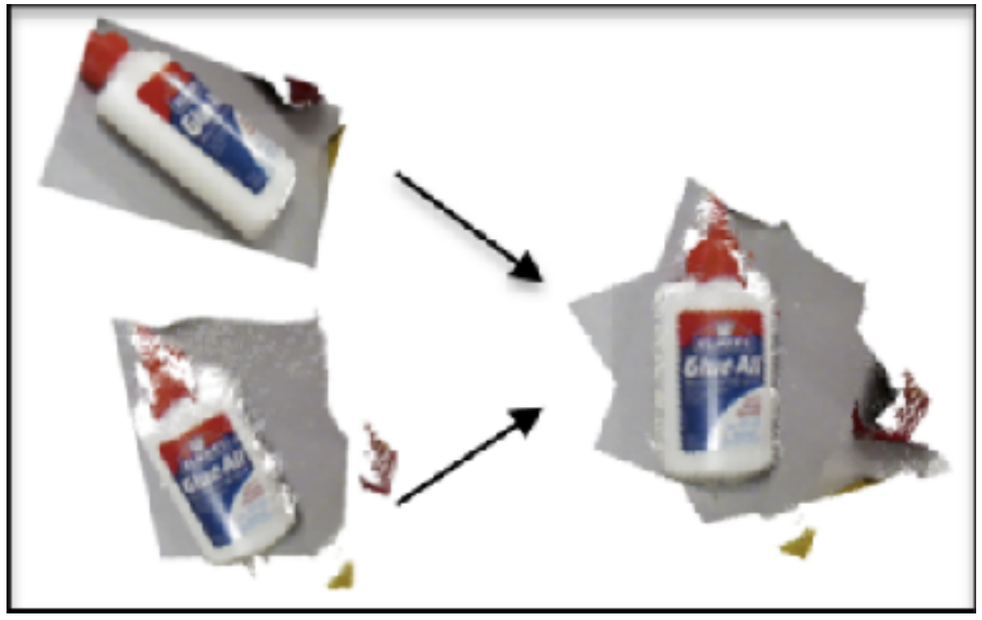
- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.

- We also use the manipulator to reconfigure the scenes by moving the objects around.



Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.
- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.
- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.
- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).
- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.
- We also use the manipulator to reconfigure the scenes by moving the objects around.



Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.

- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.

- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.

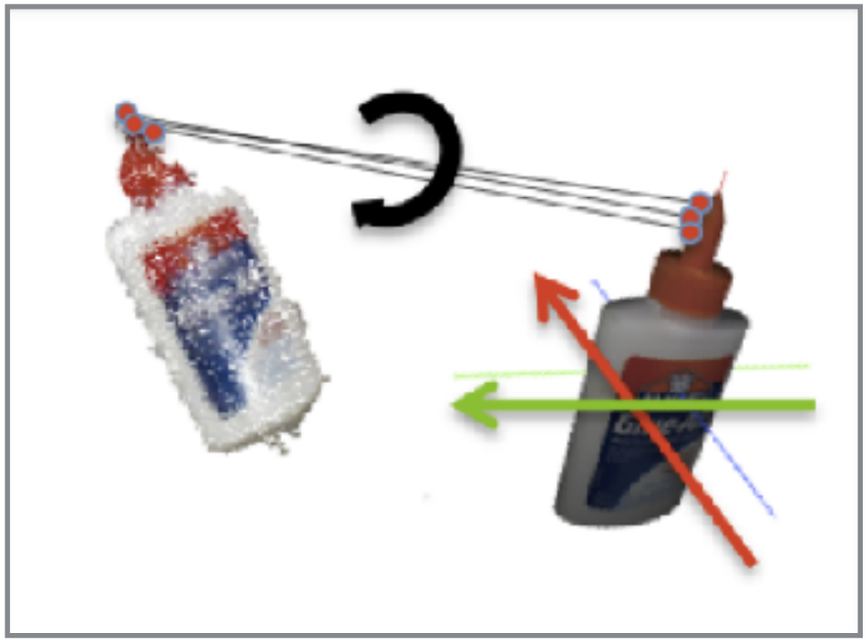
- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).

- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.

- We also use the manipulator to reconfigure the scenes by moving the objects around.

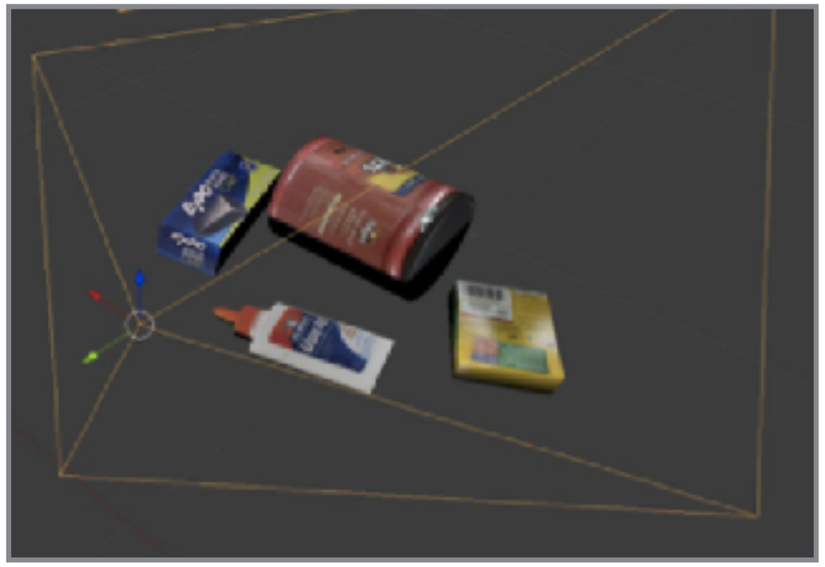
Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.
- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.
- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.
- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).
- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.
- We also use the manipulator to reconfigure the scenes by moving the objects around.



Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.
- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.
- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.
- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).
- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.
- We also use the manipulator to reconfigure the scenes by moving the objects around.



Self-Learning

- an RGBD sensor mounted on manipulator arm captures a scene (clutter of objects) from multiple views and the detector is used to get bounding-box for each object present in the scene.

- For each object the bounding-box detections with high confidence are projected to 3D using the depth data and merged to obtain point cloud for each object.

- Further 3D segmentation is performed which is also assisted by the knowledge of the location of resting surface.

- 6D pose is computed using a 3D model registration process (Super4PCS, followed by ICP in our case).

- The result of the pose estimation is projected back to the multiple-views in simulation and 2D bounding-box labels are obtained.

- We also use the manipulator to reconfigure the scenes by moving the objects around.

Evaluation (Self-Learning Object Detection)

Training dataset	success (IoU > 50%)
physics simulation, varying illumination (Our)	70%
Team MIT-Princeton, APC 2016 (Shelf & Tote dataset)	75%
adding data with multi-view self labeling (Our)	**82%**

- The training dataset used by Team MIT-Princeton comprises several real images of single objects labeled by background subtraction.
- Our approach provides labeled examples of cluttered scenes in simulation and in reality.

Evaluation (Self-Learning Object Detection)

Training dataset	success (IoU > 50%)
physics simulation, varying illumination (Our)	70%
Team MIT-Princeton, APC 2016 (Shelf & Tote dataset)	75%
adding data with multi-view self labeling (Our)	**82%**

- The training dataset used by Team MIT-Princeton comprises several real images of single objects labeled by background subtraction.
- Our approach provides labeled examples of cluttered scenes in simulation and in reality.

Evaluation (Pose Estimation)

Object recognition/model matching	success
FCN/PCA, ICP (Benchmark)	54.6%
ground-truth bounding-box/PCA, ICP (upper limit)	84.8%
RCNN/Super4PCS (Our-training)	75%
RCNN/PCA, ICP (Our-training)	**79.4%**

- The benchmark uses semantic segmentation(FCN) followed by PCA and ICP.
- Bounding-box detection has a higher recall as compared to semantic segmentation which combined with 3D segmentation can give better 3D object segments.
- In case we provide the system with ground truth bounding boxes, the upper limit for successful pose estimation is computed as above.
- Our training achieves a success which is very close to the upper limit using bounding-box detections.

Evaluation (Pose Estimation)

Object recognition/model matching	success
FCN/PCA, ICP (Benchmark)	54.6%
ground-truth bounding-box/PCA, ICP (upper limit)	84.8%
RCNN/Super4PCS (Our-training)	75%
RCNN/PCA, ICP (Our-training)	**79.4%**

- The benchmark uses semantic segmentation(FCN) followed by PCA and ICP.
- Bounding-box detection has a higher recall as compared to semantic segmentation which combined with 3D segmentation can give better 3D object segments.
- In case we provide the system with ground truth bounding boxes, the upper limit for successful pose estimation is computed as above.
- Our training achieves a success which is very close to the upper limit using bounding-box detections.

Evaluation (Pose Estimation)

Object recognition/model matching	success
FCN/PCA, ICP (Benchmark)	54.6%
ground-truth bounding-box/PCA, ICP (upper limit)	84.8%
RCNN/Super4PCS (Our-training)	75%
RCNN/PCA, ICP (Our-training)	**79.4%**

- The benchmark uses semantic segmentation(FCN) followed by PCA and ICP.
- Bounding-box detection has a higher recall as compared to semantic segmentation which combined with 3D segmentation can give better 3D object segments.
- In case we provide the system with ground truth bounding boxes, the upper limit for successful pose estimation is computed as above.
- Our training achieves a success which is very close to the upper limit using bounding-box detections.

Evaluation (Pose Estimation)

Object recognition/model matching	success
FCN/PCA, ICP (Benchmark)	54.6%
ground-truth bounding-box/PCA, ICP (upper limit)	84.8%
RCNN/Super4PCS (Our-training)	75%
RCNN/PCA, ICP (Our-training)	**79.4%**

- The benchmark uses semantic segmentation(FCN) followed by PCA and ICP.
- Bounding-box detection has a higher recall as compared to semantic segmentation which combined with 3D segmentation can give better 3D object segments.
- In case we provide the system with ground truth bounding boxes, the upper limit for successful pose estimation is computed as above.
- Our training achieves a success which is very close to the upper limit using bounding-box detections.

Future Work

- In the future, we would like to study how could learning the unknown parameters of the simulation such as illumination and model properties help improve the training.
- And how to efficiently use a physics-aware simulation for global reasoning in the pose estimation problem.

Future Work

- In the future, we would like to study how could learning the unknown parameters of the simulation such as illumination and model properties help improve the training.
- And how to efficiently use a physics-aware simulation for global reasoning in the pose estimation problem.

Webpage : https://www.cs.rutgers.edu/~cm1074/PHYSIM.html

The task-specific, simulated and self-labeled real data generated by our autonomous system can be used to train a CNN and get **state-of-the-art** performance in object detection and 6D pose estimation **without any human effort**.