Perceiving the 3D World from Images and Videos

Yu Xiang
Postdoctoral Researcher
University of Washington & NVIDIA Research

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING


HONDA

Acting in the 3D World

Sensing

Camera

Sound sensor

Depth sensor

Touch sensor

Intelligent system

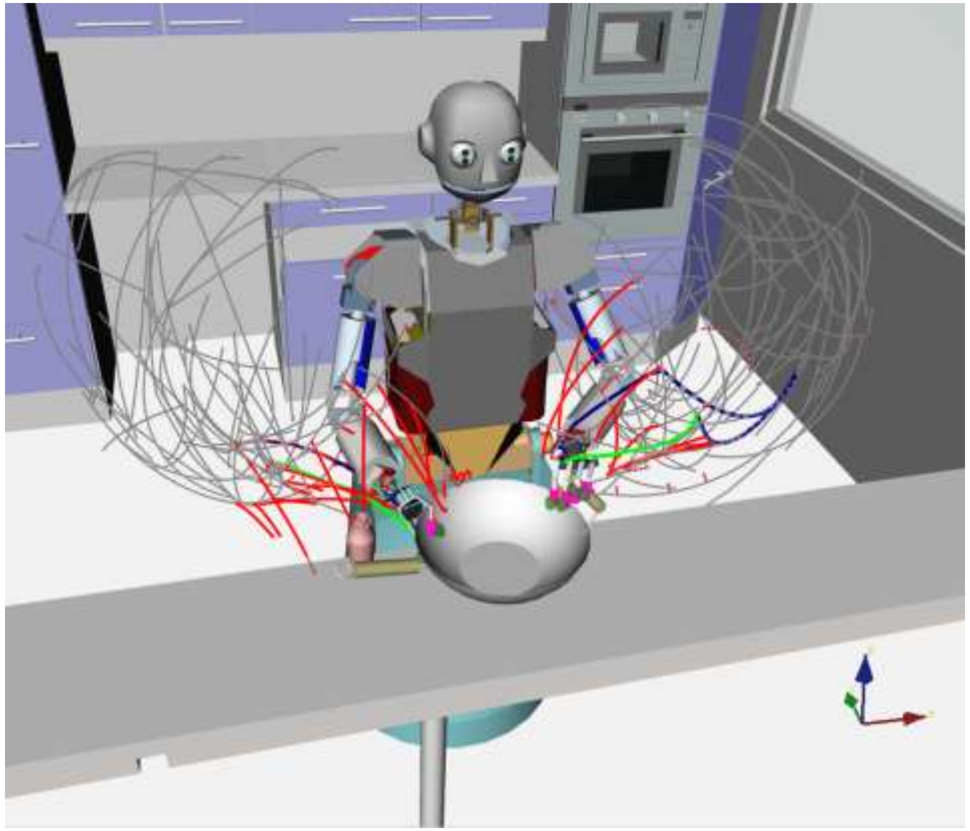
Acting in the 3D World

Perception

- Geometry
  - ❑ Free space
  - ❑ Surfaces
  - ❑ 3D shapes
- Semantics
  - ❑ Humans
  - ❑ Objects
  - ❑ Affordances

Acting in the 3D World

Planning and Control



Acting in the 3D World

Sensing

Perception

Planning & Control

Intelligent System

Intelligent visual models

3D World

3D Scene Understanding

3D Reconstruction

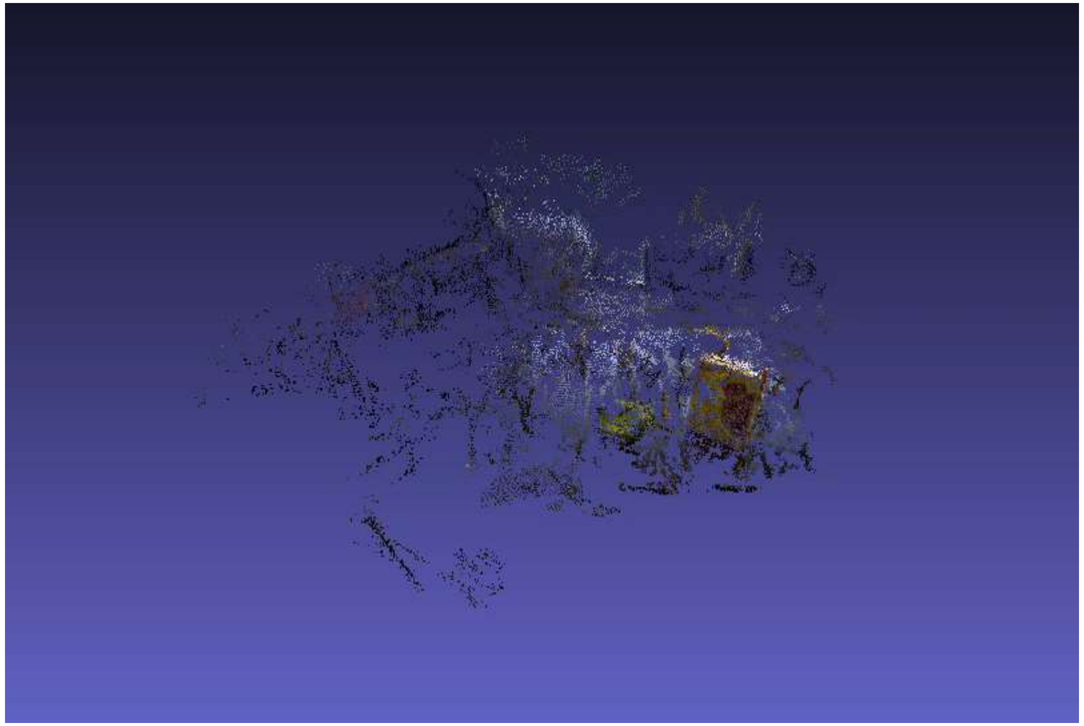
- Structure from Motion

VisualSFM

- Longuet-Higgins, Nature, 1981
- Tomasi & Kanade, IJCV, 1992
- Sturm & Triggs, ECCV, 1996
- Soatto, Automatica, 1997
- Snavely et al., SIGGRAPH, 2006
- Pollefeys et al., IJCV, 2008
- Agarwal et al., ICCV, 2009
- Furukawa et al., CVPR, 2010
- Sinha et al., RMLE, 2010
- Wu et al., CVPR, 2011
- Wilson & Snavely, ICCV, 2013

3D Reconstruction

- Dense structure from motion (multi-view stereo)

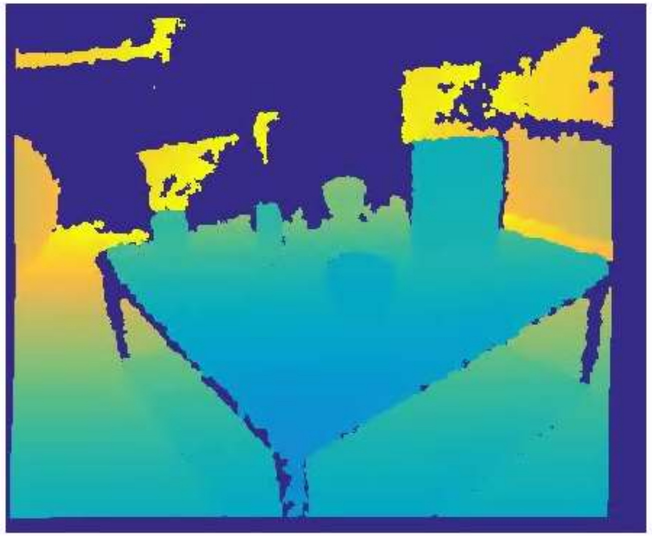


- Curless & Levoy, SIGGRAPH, 1996
- Jin et al, CVPR, 2003
- Hornung et al., ECCV, 2006
- Goesele et al., ICCV, 2007
- Furukawa & Ponce, PAMI, 2008
- Campbell et al., ECCV, 2008
- Kolev & Cremers, ECCV, 2008
- Hiep et al., CVPR, 2009
- Furukawa et al., CVPR, 2010
- Jancosek & Pajdla, CVPR, 2011
- Matzen & Snavely, ECCV, 2014

RGB-Depth Sensor

Kinect in 2010

3D Reconstruction using Depth



KinectFusion

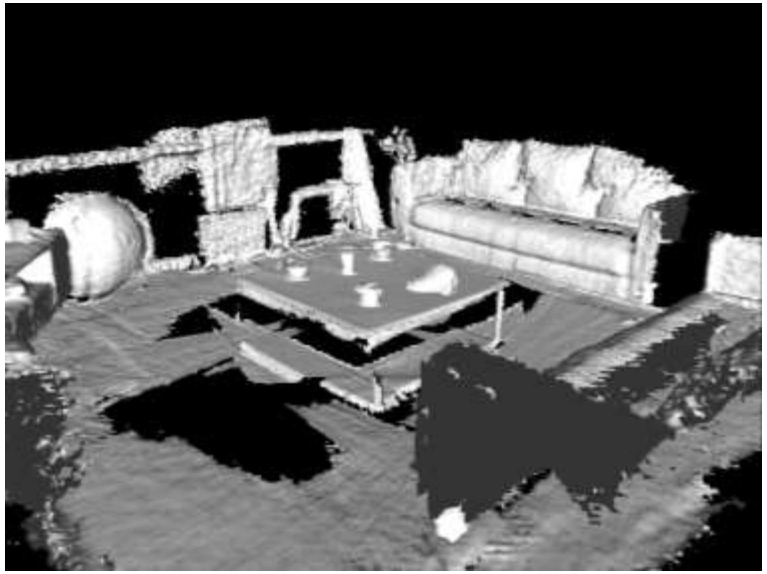
- Newcombe et al., ISMAR, 2011
- Izadi et al., UIST, 2011
- Henry et al., IJRR, 2012
- Whelan et al., RSS Workshop, 2012
- Henry et al., 3DV, 2013
- Keller et al., 3DV, 2013
- Salas-Moreno et al., CVPR, 2013
- Steinbrucker et al., ICCV, 2013
- Zollhöfer et al., TOG, 2014
- Whelan et al., RSS, 2015

Robot navigation

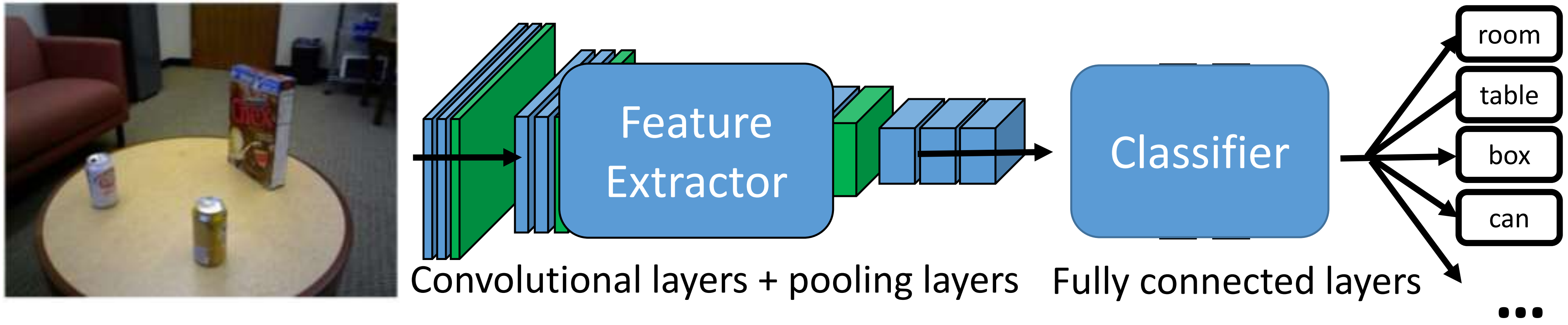
Semantics

Bring me the mug on the table?

???



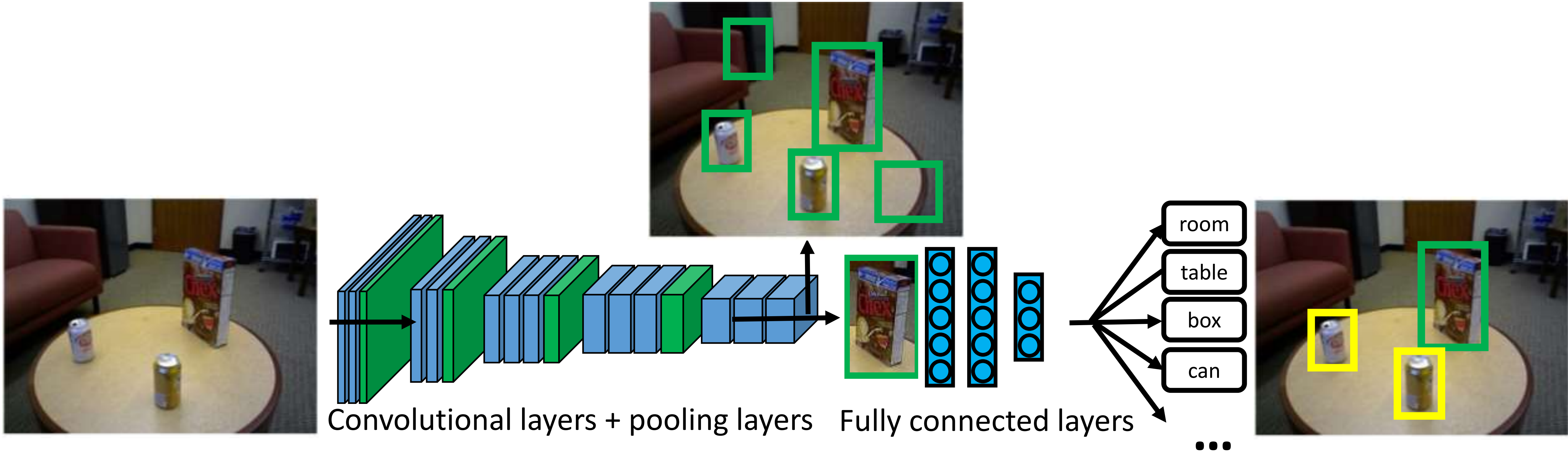
Recognition: Image Classification



Convolutional Neural Networks (CNNs)

- Krizhevsky et al., NIPS, 2012
- Ciregan et al., CVPR, 2012
- Karpathy et al., CVPR, 2014
- Simonyan & Zisserman, arXiv, 2014
- Lin et al., ICLR, 2014
- Zeiler & Fergus, ECCV, 2014
- He et al., ECCV, 2014
- Srivastava et al., JMLR, 2014
- Mahendran & Vedaldi, CVPR, 2015
- Jaderberg et al, NIPS, 2015
- Su et al., CVPR, 2015
- LeCun et al., Nature, 2015
- Szegedy et al., CVPR, 2015
- He et al., CVPR, 2016
- Rastegari et al., ECCV, 2016
- Huang et al., CVPR, 2017

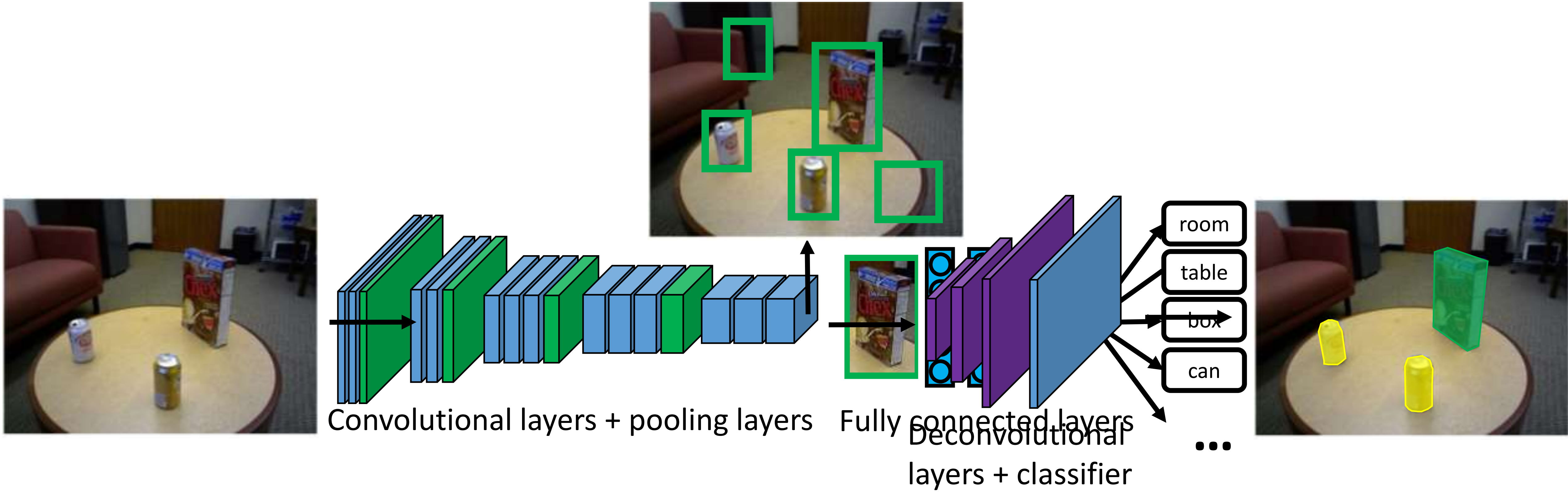
Recognition: Object Detection



CNN-based Object Detection

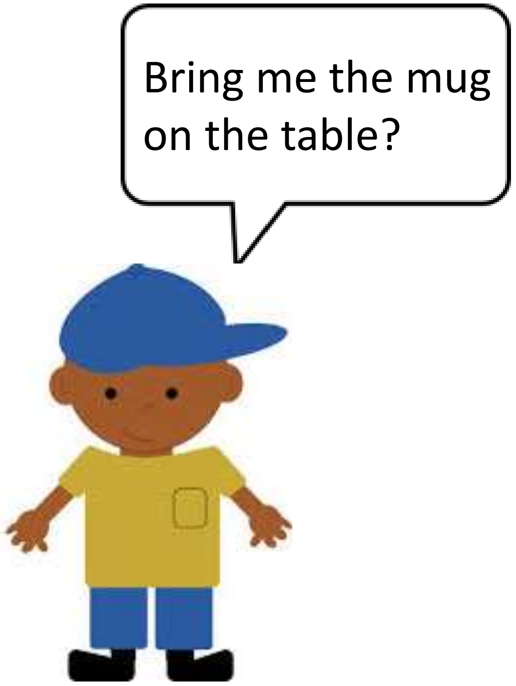
- Sermanet et al., arXiv, 2013
- Girshick et al., CVPR, 2014
- Gupta et al., ECCV, 2014
- Zhang et al., ECCV, 2014
- He et al., ECCV, 2014
- Erhan et al, CVPR, 2014
- Ren et al., NIPS, 2015
- Girshick, ICCV, 2015
- Bell et al., CVPR, 2016
- Liu et al., ECCV, 2016
- Yang et al, CVPR, 2016
- Cai et al., ECCV, 2016
- Redmon et al., CVPR, 2016
- Redmon et al., ECCV, 2016
- Dai et al., NIPS, 2016
- Xiang et al., WACV, 2017

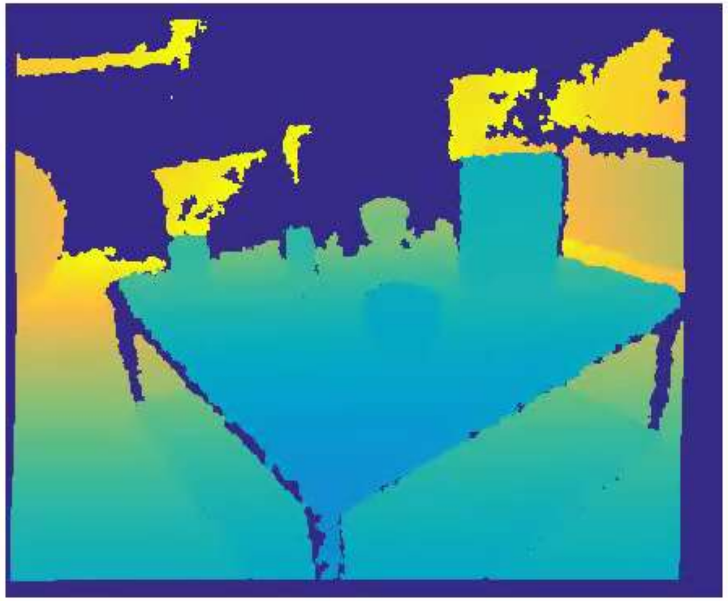
Recognition: Semantic Labeling

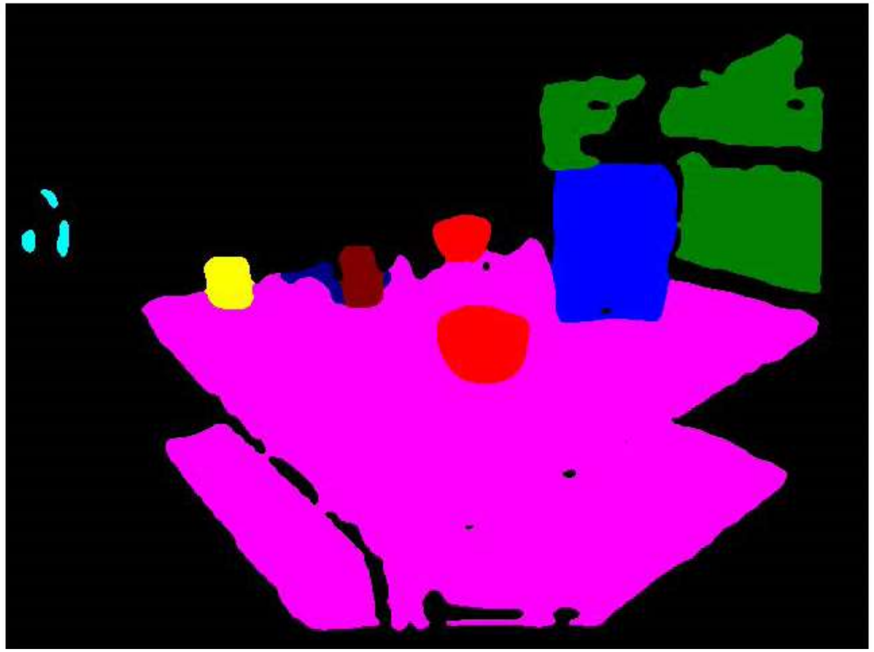


Fully Convolutional Networks (FCNs)

- Pinheiro & Collobert, JMLR, 2014
- Girshick et al., CVPR, 2014
- Hariharan et al., ECCV, 2014
- Zheng et al., CVPR, 2015
- Ronneberger et al., MICCAI, 2015
- Chen et al., ICLR, 2015
- Long et al., CVPR, 2015
- Noh et al., CVPR, 2015
- Papandreou et al., CVPR, 2015
- Liu et al., ICCV, 2015
- Hariharan et al., CVPR, 2015
- Dai et al., CVPR, 2015
- Mostajabi et al., CVPR, 2015
- Dai et al., CVPR, 2016
- Milletari et al., 3DV, 2016
- Badrinarayanan et al., PAMI, 2017

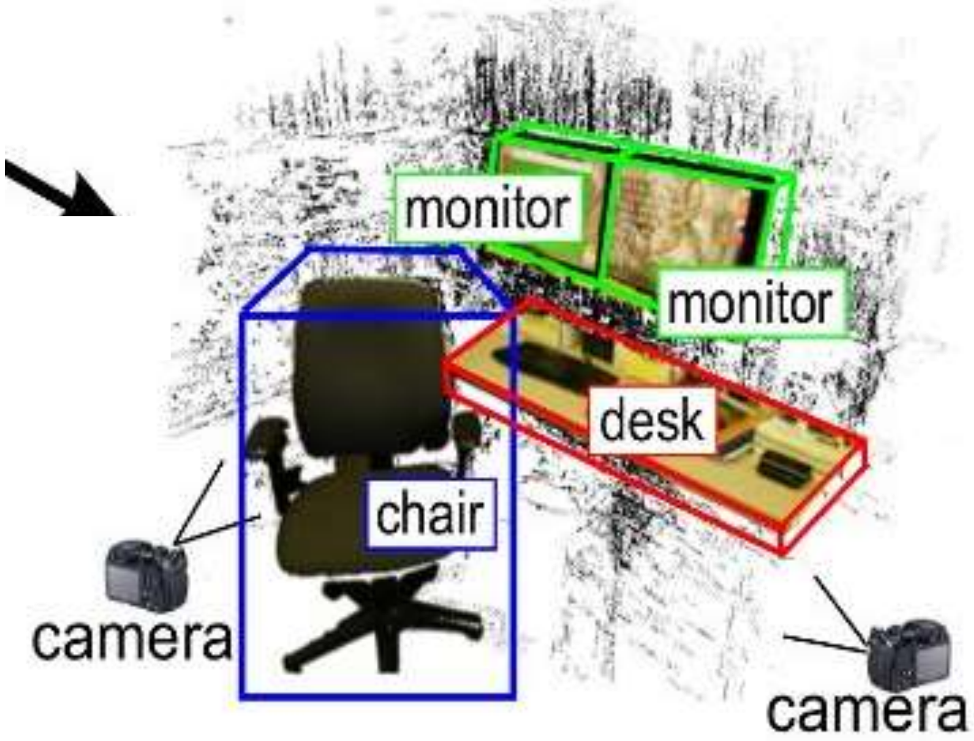




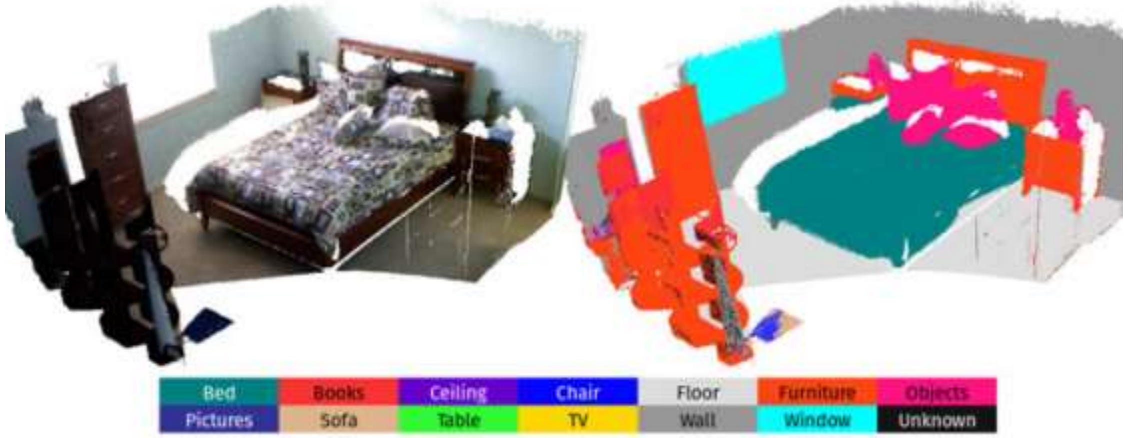


[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.

Semantic 3D Reconstruction



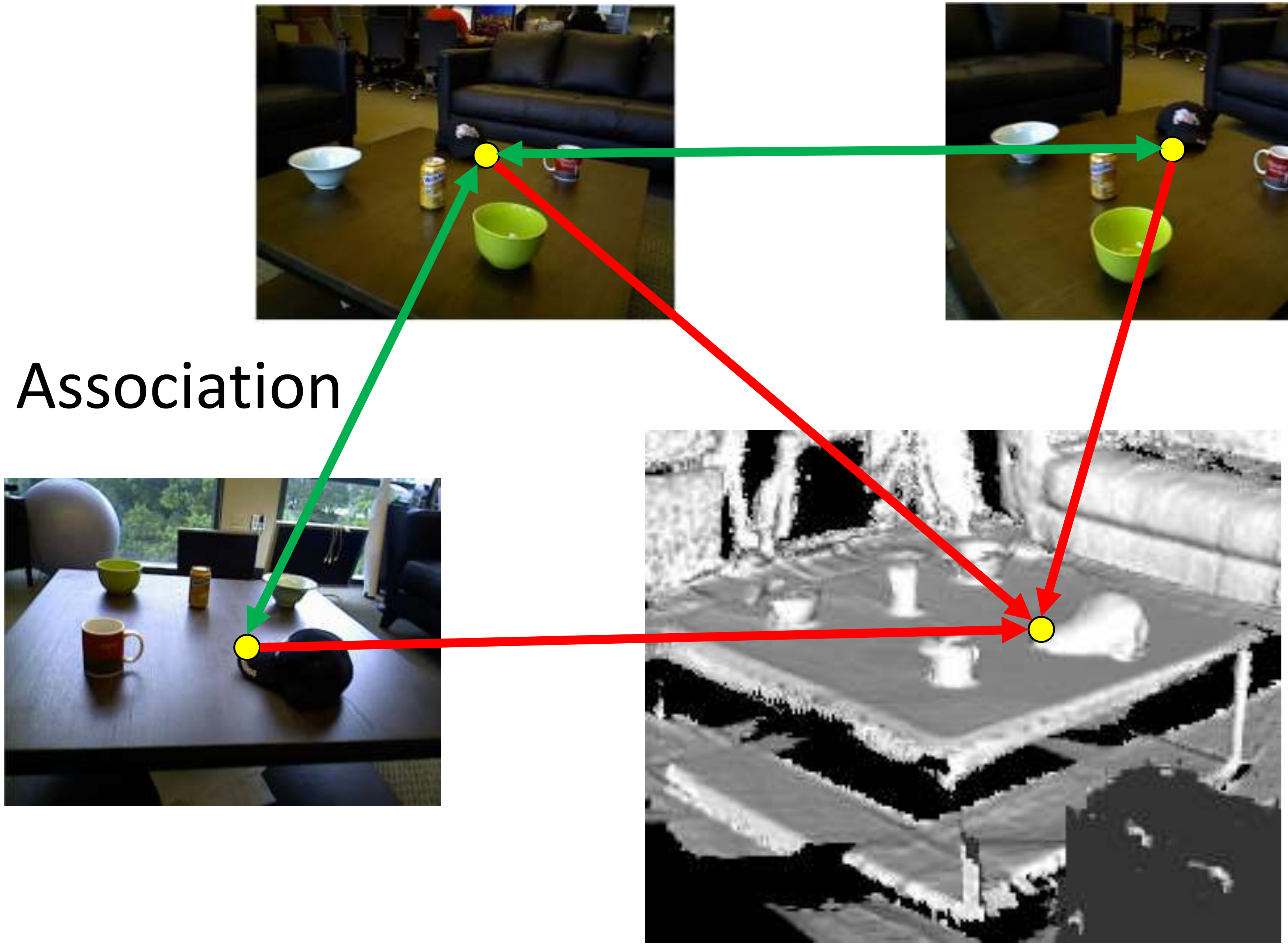
Semantic Structure from Motion
Bao & Savarese, CVPR, 2011



SemanticFusion
McCormac et al., ICRA, 2017

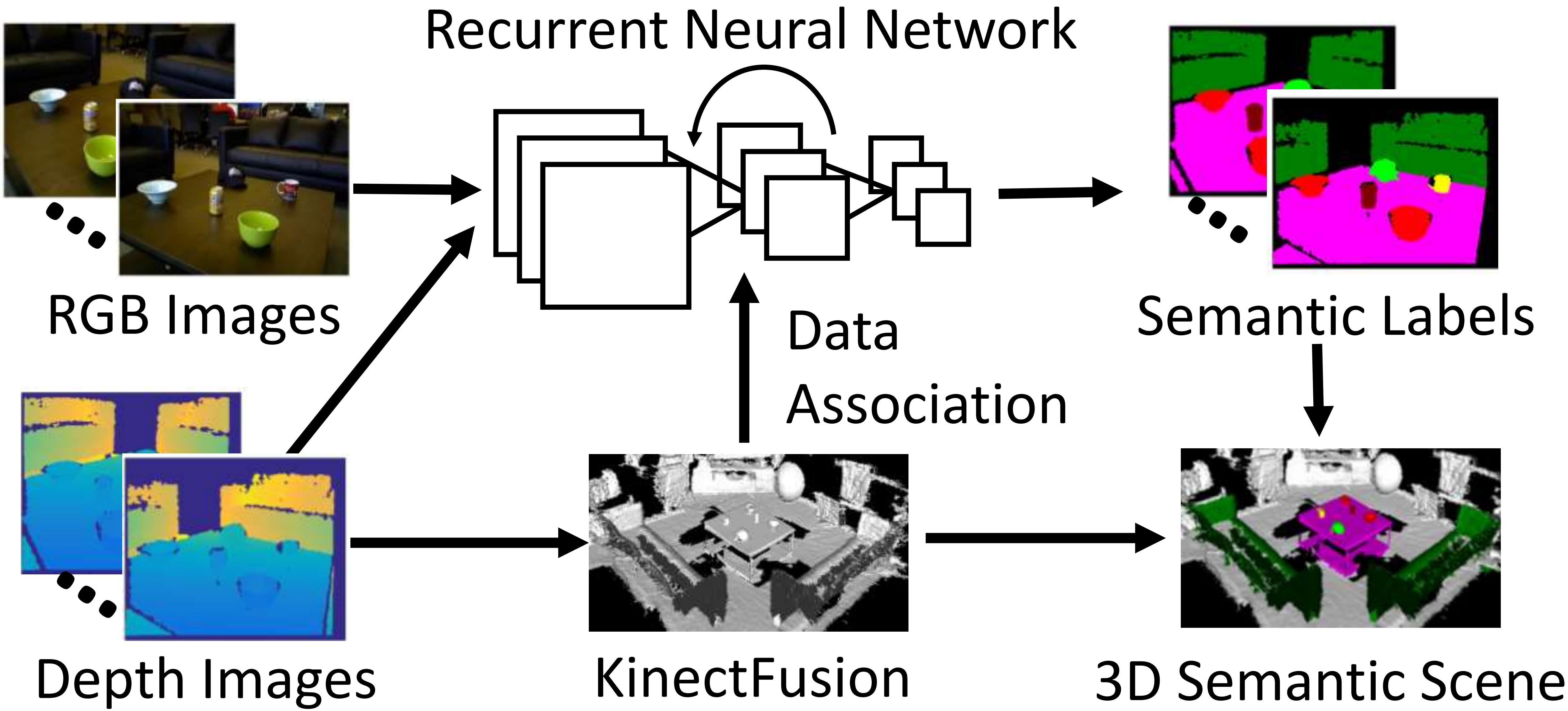
Can 3D Reconstruction Help Learning Semantics?

Data Association



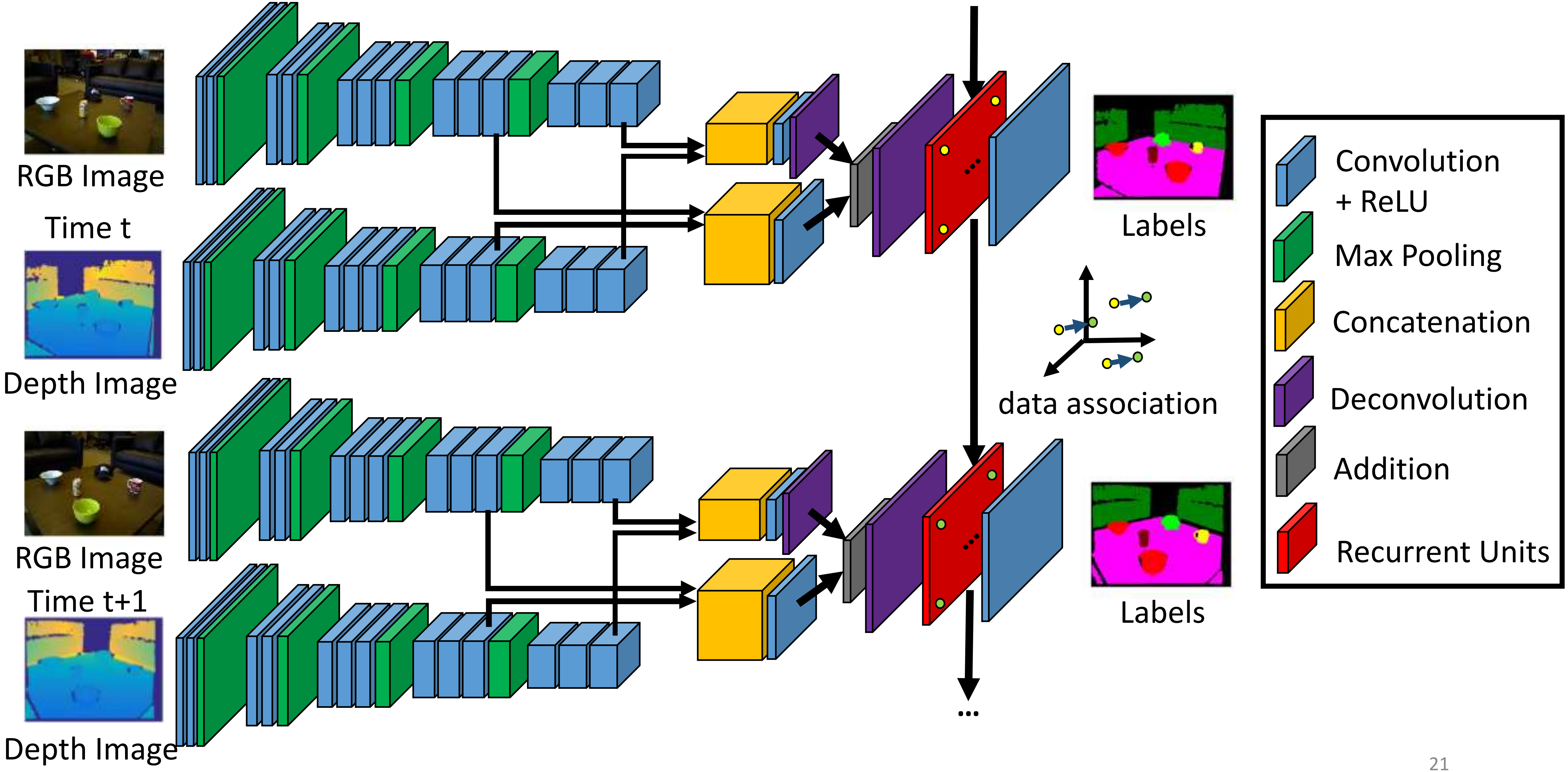
KinectFusion map

Our Contribution: DA-RNNs



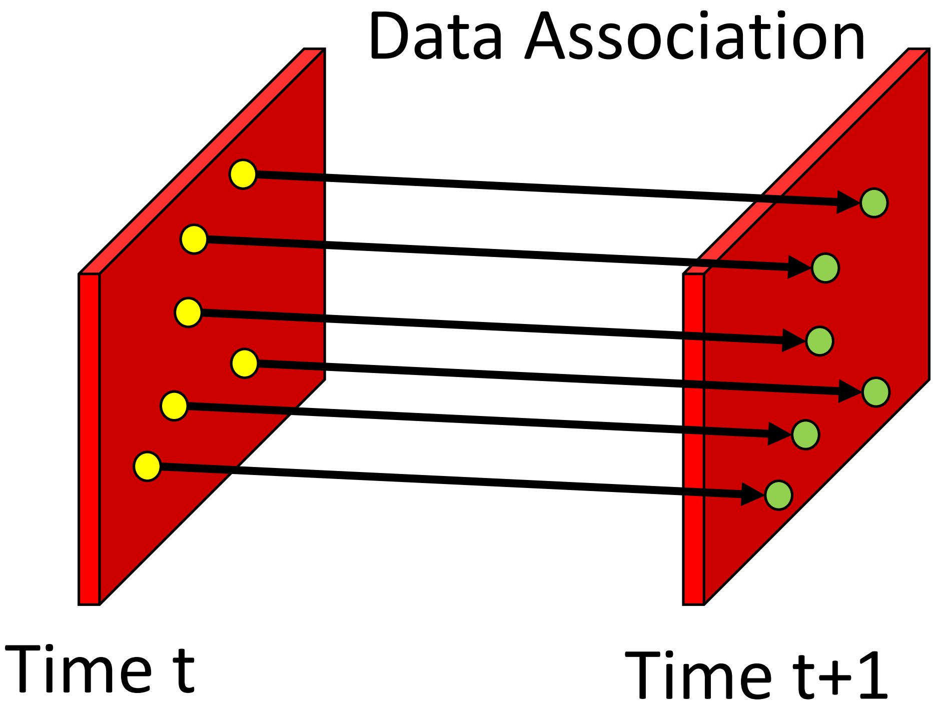
Y. Xiang and D. Fox. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. In RSS, 2017.

Video Semantic Labeling with DA-RNNs

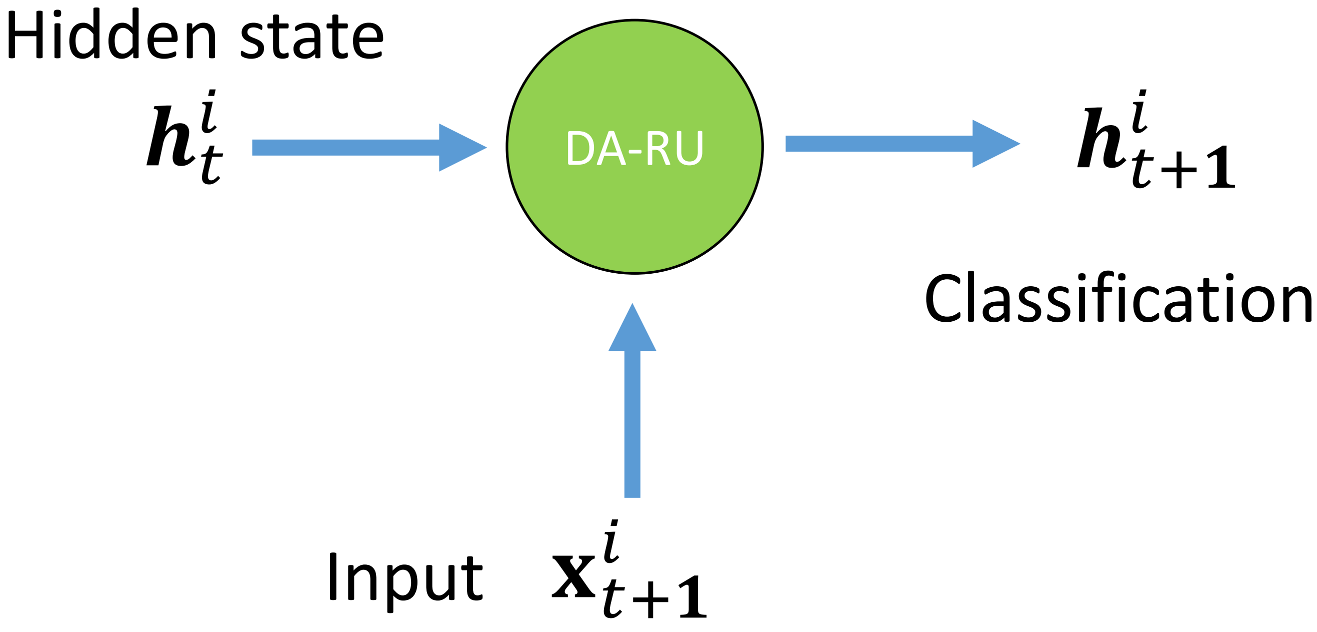


Y. Xiang and D. Fox. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. In RSS, 2017.

Data Associated Recurrent Units (DA-RUs)



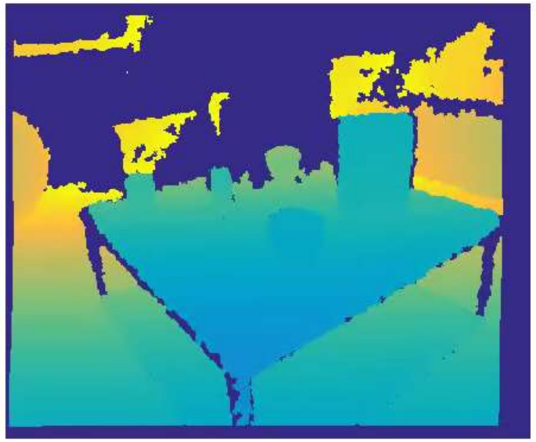
Recurrent layer

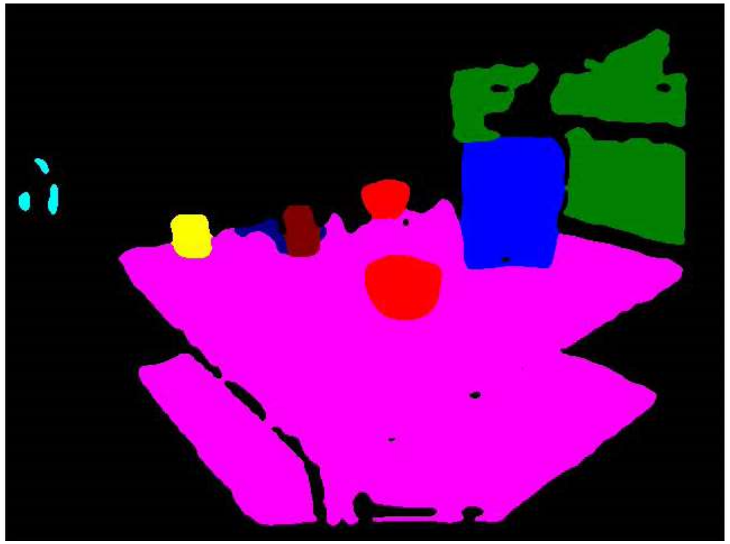


Weighted Moving Averaging
with learnable parameters

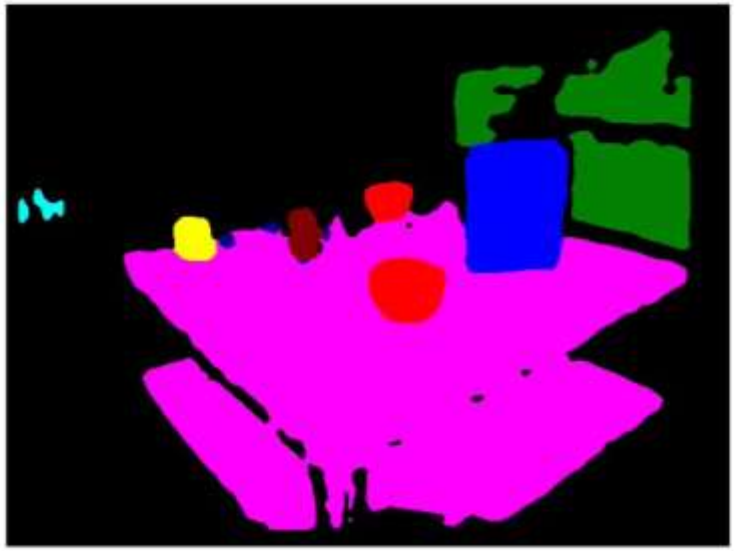
Y. Xiang and D. Fox. DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks. In RSS, 2017.

Results on RGB-D Scene Dataset [1]



FCN

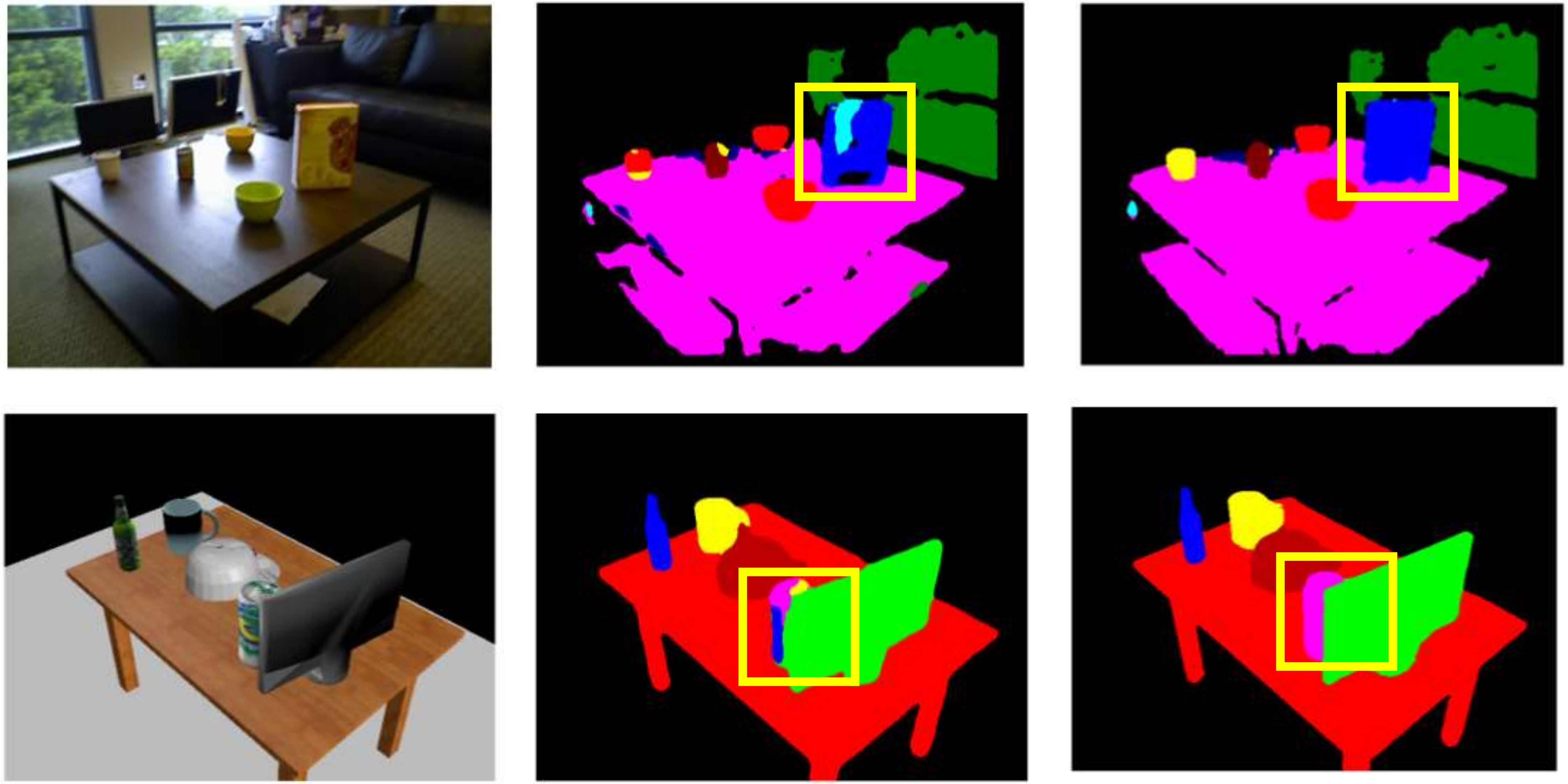


Our DA-RNN

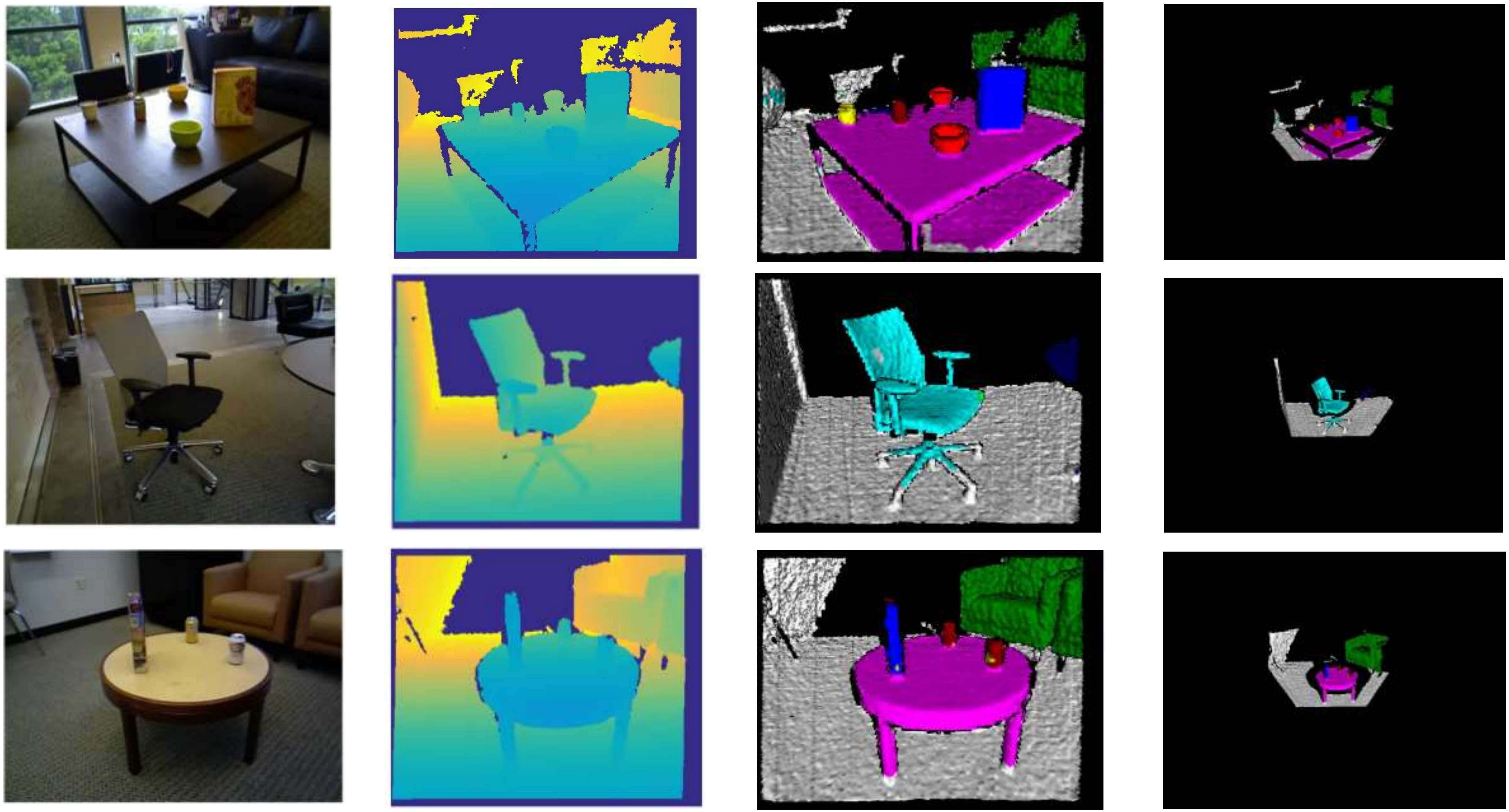
Methods	FCN	GRU-RNN	DA-RNN
Background	96.1	96.8	**97.6**
Bowl	87.0	86.4	**92.7**
Cap	79.0	82.0	**84.4**
Cereal Box	87.5	87.5	**88.3**
Coffee Mug	75.7	76.1	**86.3**
Coffee Table	95.2	96.0	**97.3**
Office Chair	71.6	72.7	**77.0**
Soda Can	82.9	81.9	**88.7**
Sofa	92.9	93.5	**95.6**
Table	89.8	90.8	**92.8**
MEAN	85.8	86.4	**90.1**

Metric: segmentation intersection over union (IoU)

[1] K. Lai, L. Bo and D. Fox. Unsupervised feature learning for 3D scene labeling. In ICRA’14.

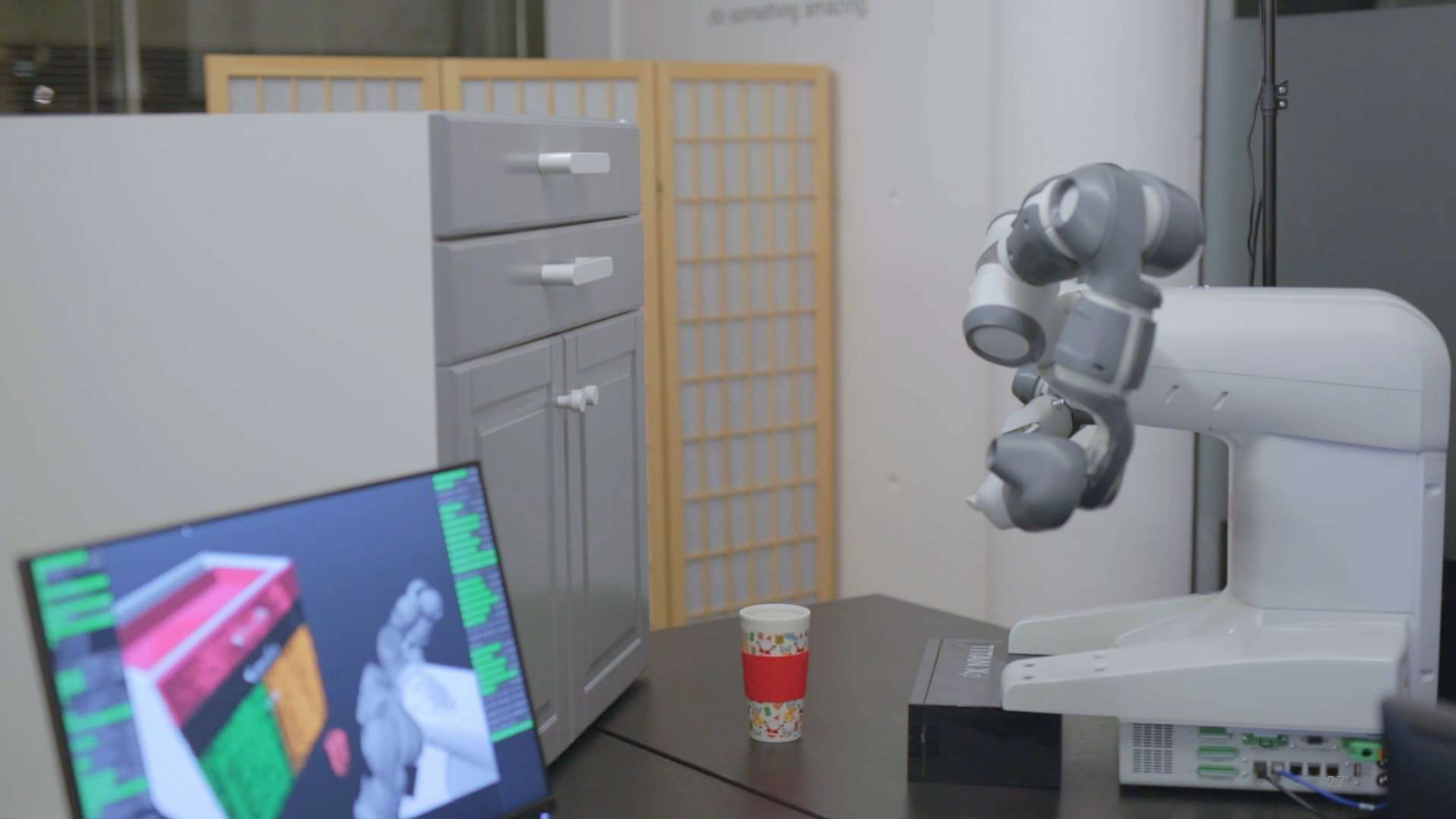


RGB Image FCN Our DA-RNN



RGB Images Depth Images Semantic Mapping

3D Object Recognition


do something amazing
TITAN Xp

3D Object Recognition

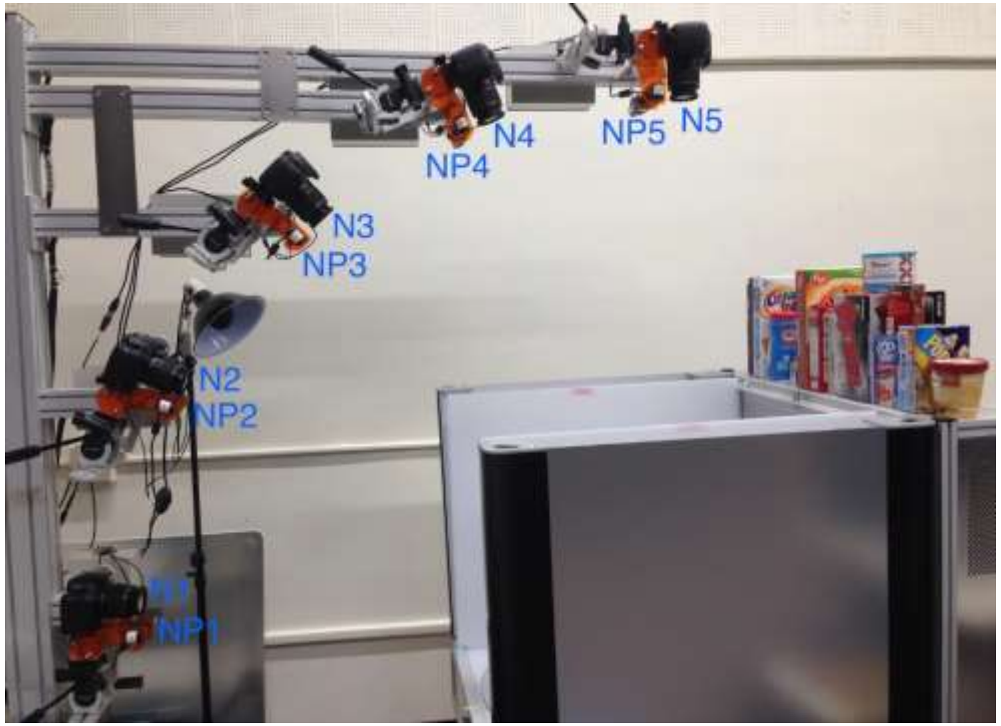

Camera

- 3D location
- 3D orientation


Input image


3D world

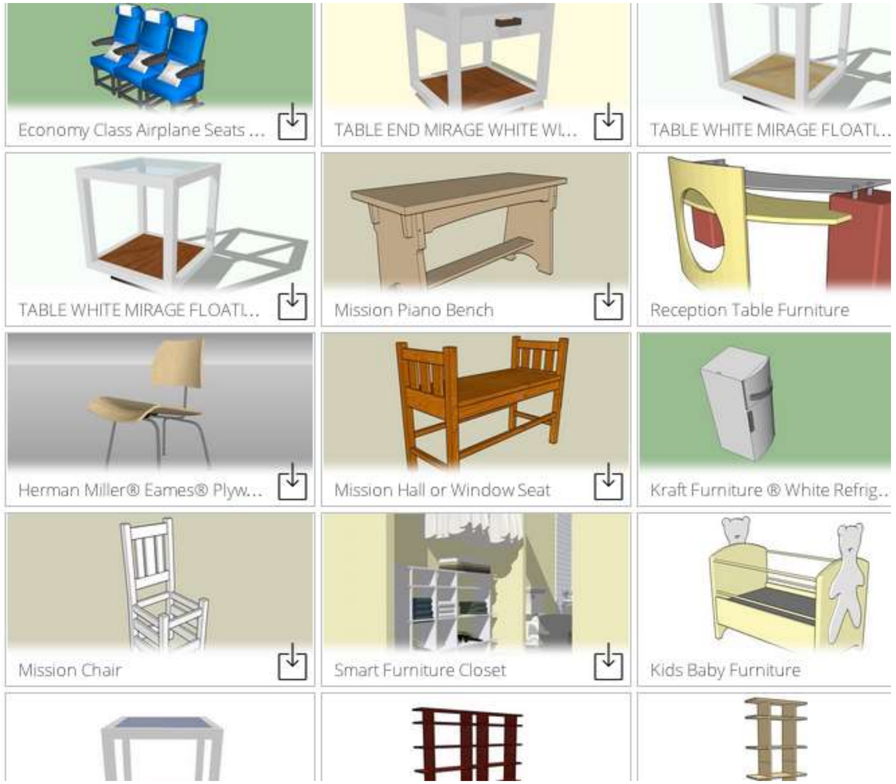
Building the 3D models: 3D Object Reconstruction



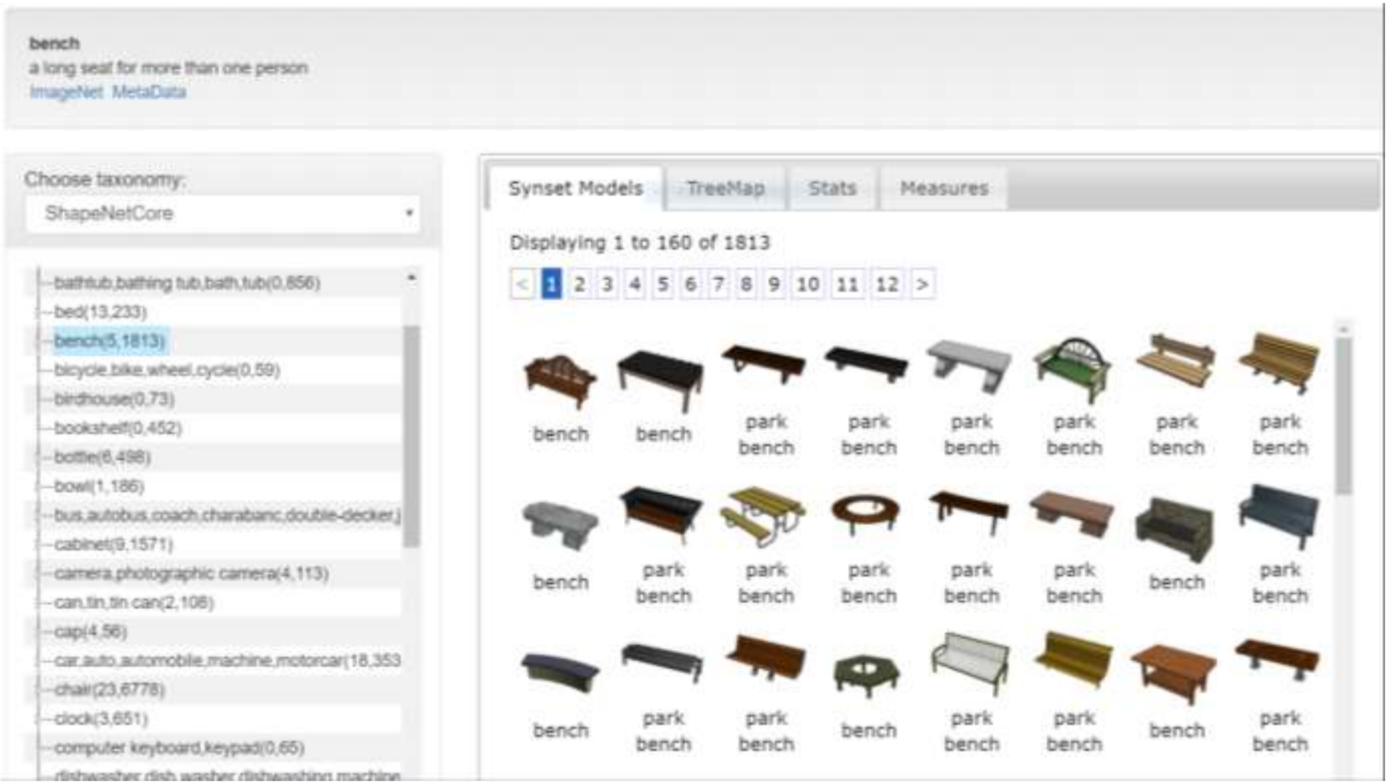
- Terzopoulos et al., IJCV, 1998
- Banta et al., SMC:Systems, 2000
- Esteban & Schmitt, 3DPVT, 2002
- Guan et al., 3DPVT, 2008
- Singh et al., ICRA, 2014
- Calli et al., RA Magazine, 2015

Berkeley Instance Recognition Dataset
Singh et al., ICRA, 2014

Building the 3D models: 3D CAD Models

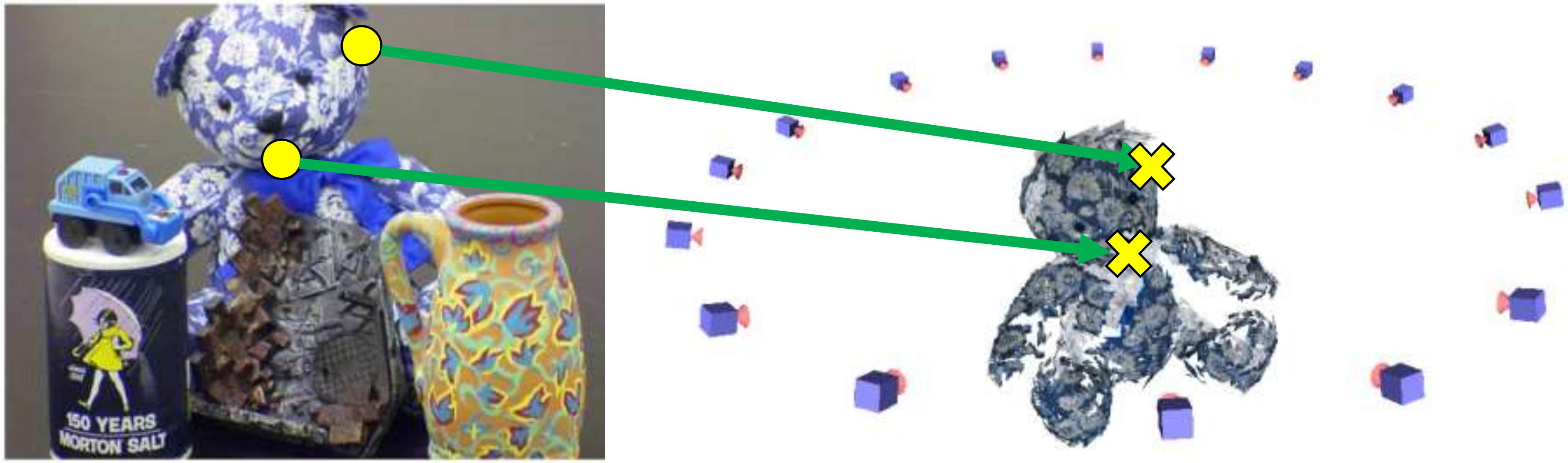


Trimble 3D Warehouse
https://3dwarehouse.sketchup.com



ShapeNet
https://www.shapenet.org/

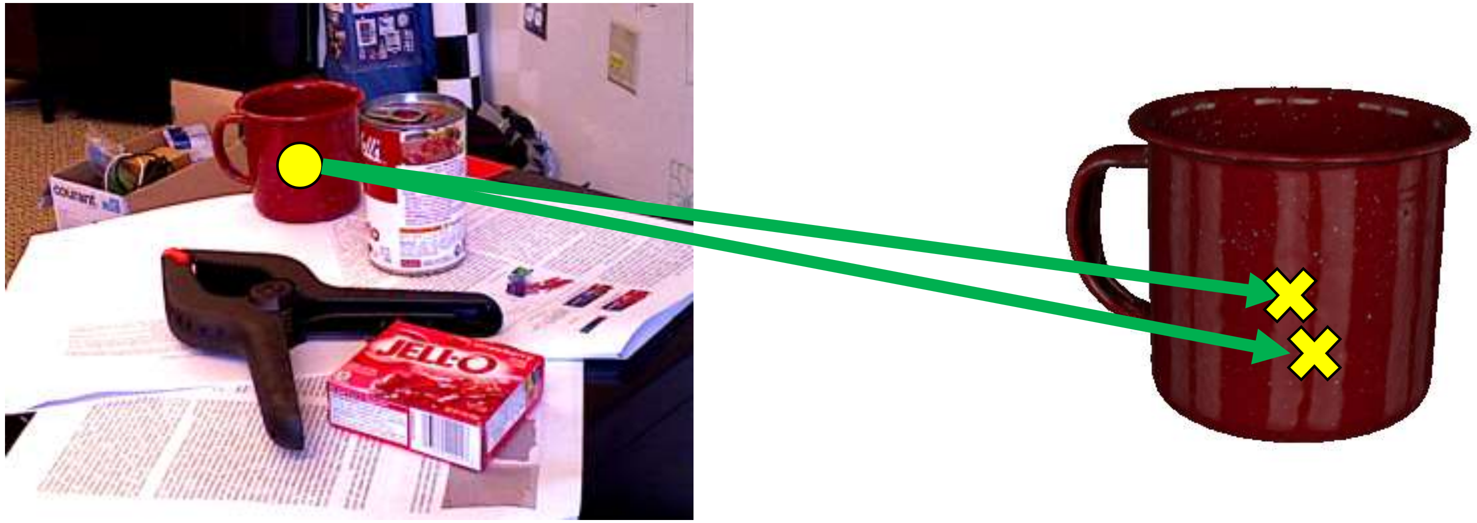
3D Object Recognition: Feature Matching



2D image

3D model

- Lowe, ICCV, 1999
- Rothganger et al., IJCV, 2006
- Savarese & Fei-Fei, ICCV, 2007
- Collet et al., IJRR, 2011
- Brachmann et al., ECCV, 2014
- Krull et al., ICCV, 2015
- Kehl et al., ECCV, 2016
- Michel et al., CVPR, 2017
- Pavlakos et al., ICRA, 2017
- Rad & Lepetit, ICCV, 2017

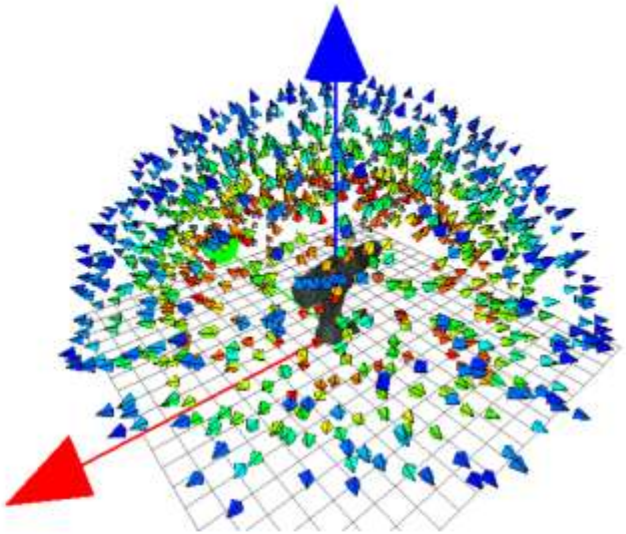


✘ Texture-less objects

✘ Symmetric objects

✓ Occlusions

3D Object Recognition: Template Matching



- Thomas et al., CVPR, 2006
- Ozuysal et al., CVPR, 2009
- Gu & Ren, ECCV, 2010
- Hinterstoisser et al., ACCV, 2012
- Xiang & Savarese, CVPR, 2012
- Pepik et al., CVPR, 2012
- Su et al., ICCV, 2015
- Cao et al., ICRA, 2016
- Tekin et al, CVPR, 2018

✓ Texture-less objects
✓ Symmetric objects
✘ Occlusions

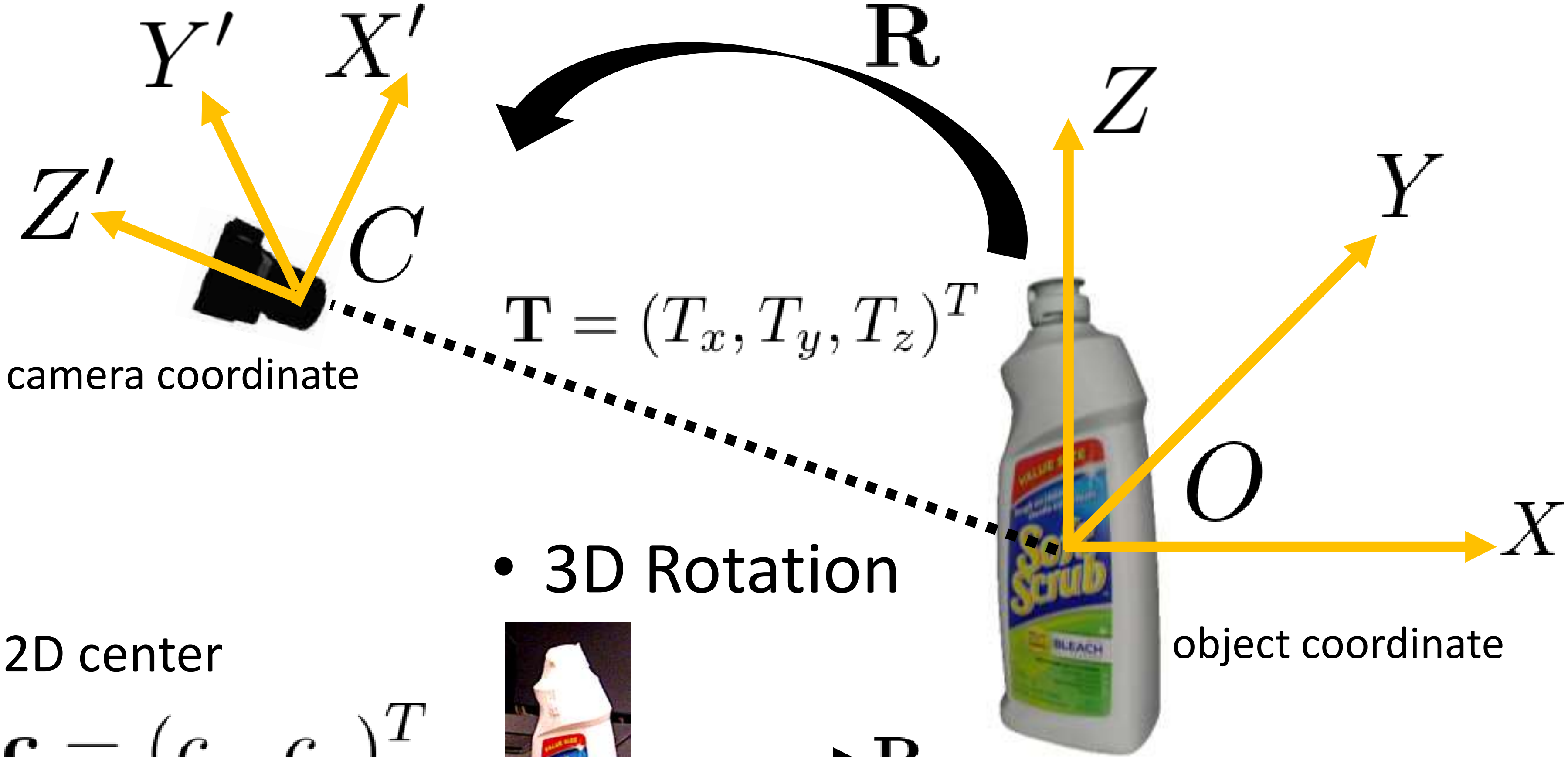
Our Contribution: A Generic Convolutional Neural Network for 6D Object Pose Estimation

PoseCNN

- ✓ Texture-less objects
- ✓ Symmetric objects
- ✓ Occlusions

Y. Xiang, T. Schmidt, V. Narayanan and D. Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In arXiv:1711.00199, 2018 (Under Review)

PoseCNN: Decouple 3D Translation and 3D Rotation



- 3D Translation

2D center

$\mathbf{c} = (c_x, c_y)^T$

Distance T_z

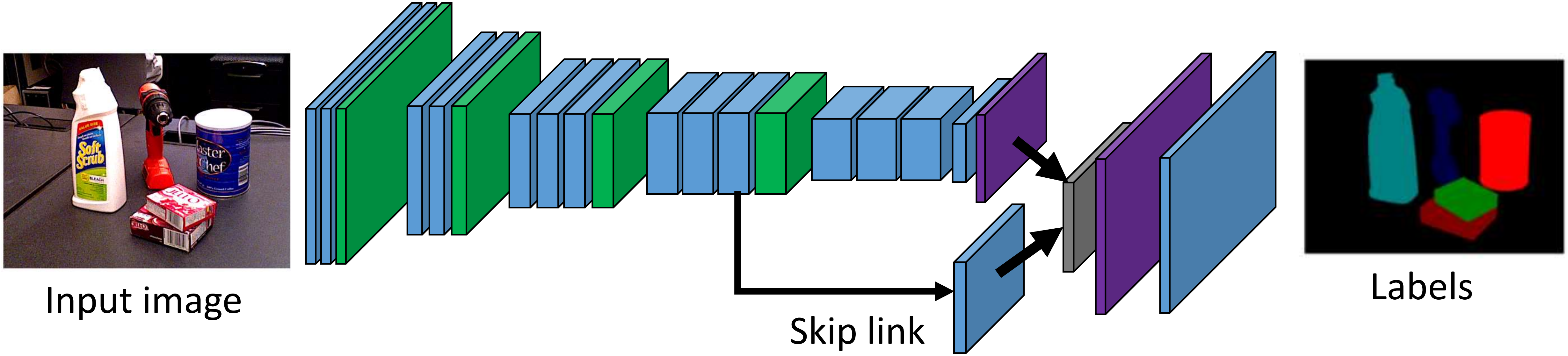
2D Center Localization

- 3D Rotation

$\longrightarrow \mathbf{R}$

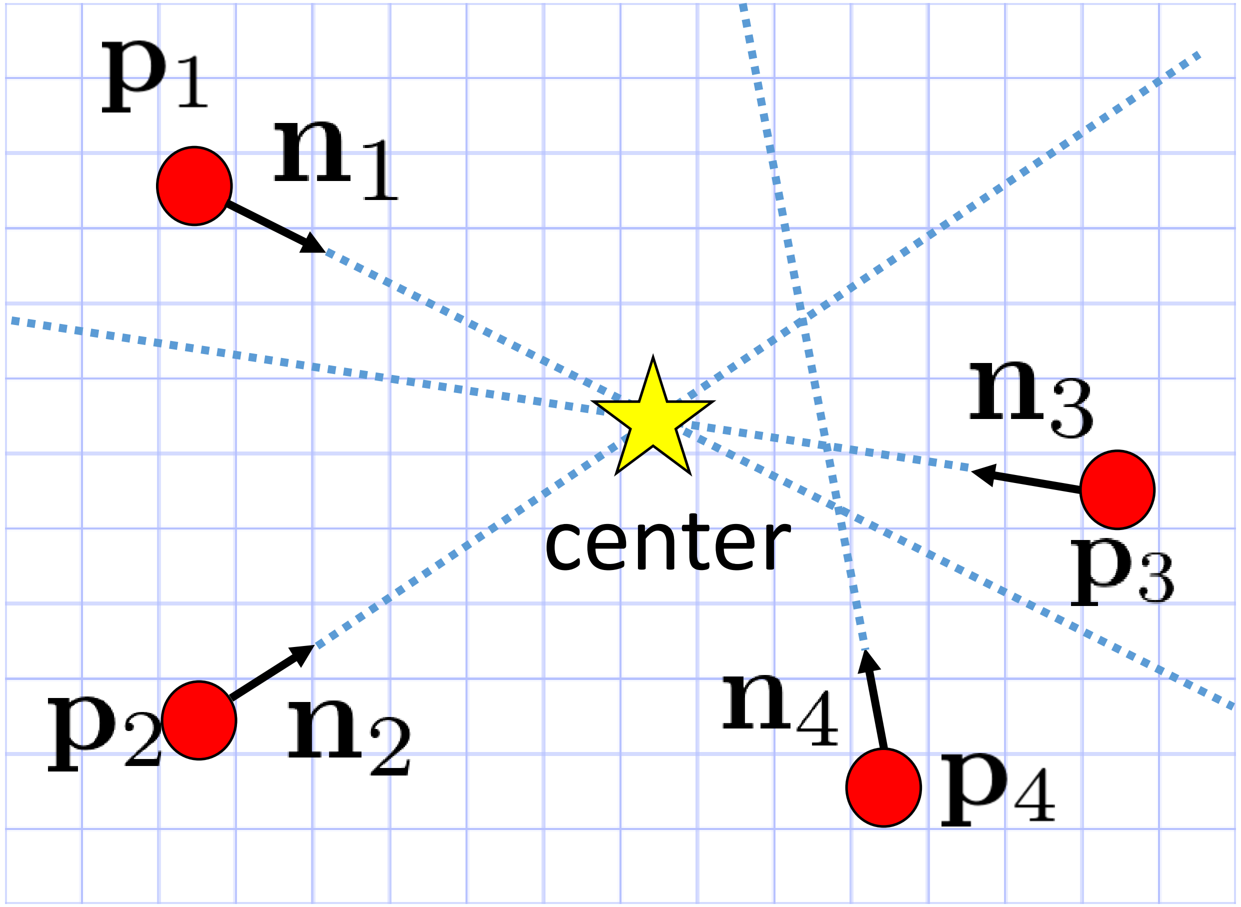
3D Rotation Regression

PoseCNN: Semantic Labeling

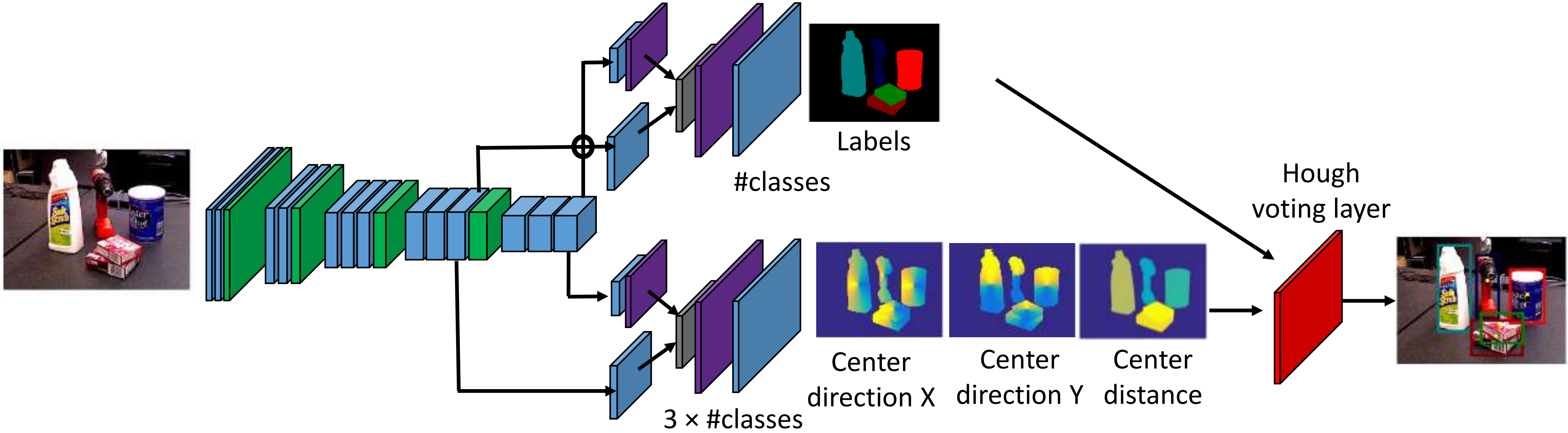


Fully convolutional network

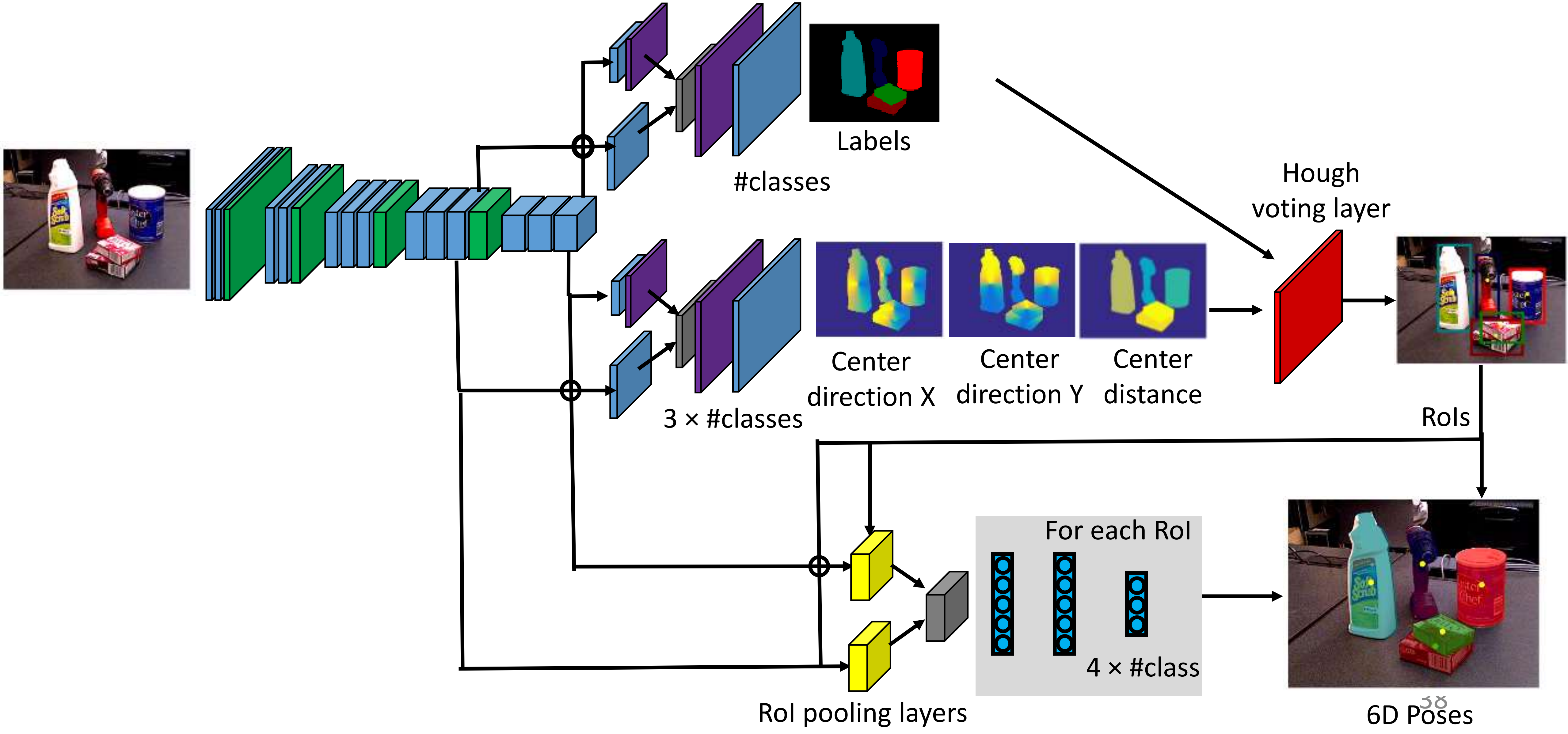
PoseCNN: 2D Center Voting for Handling Occlusions



PoseCNN: 3D Translation Estimation

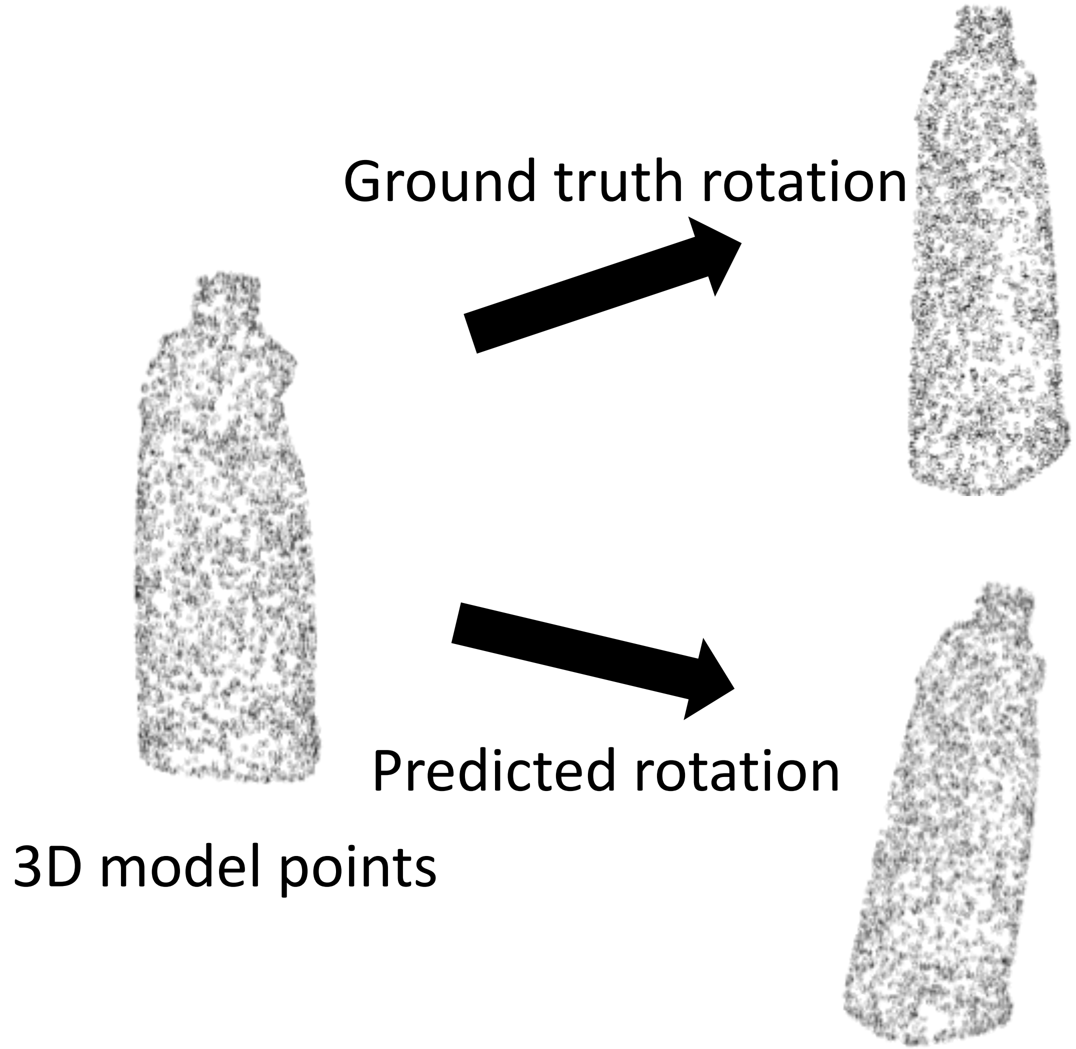


PoseCNN: 3D Rotation Regression



PoseCNN: Handle Symmetric Objects

PoseCNN: 3D Rotation Regression Loss Functions



Pose Loss (non-symmetric)

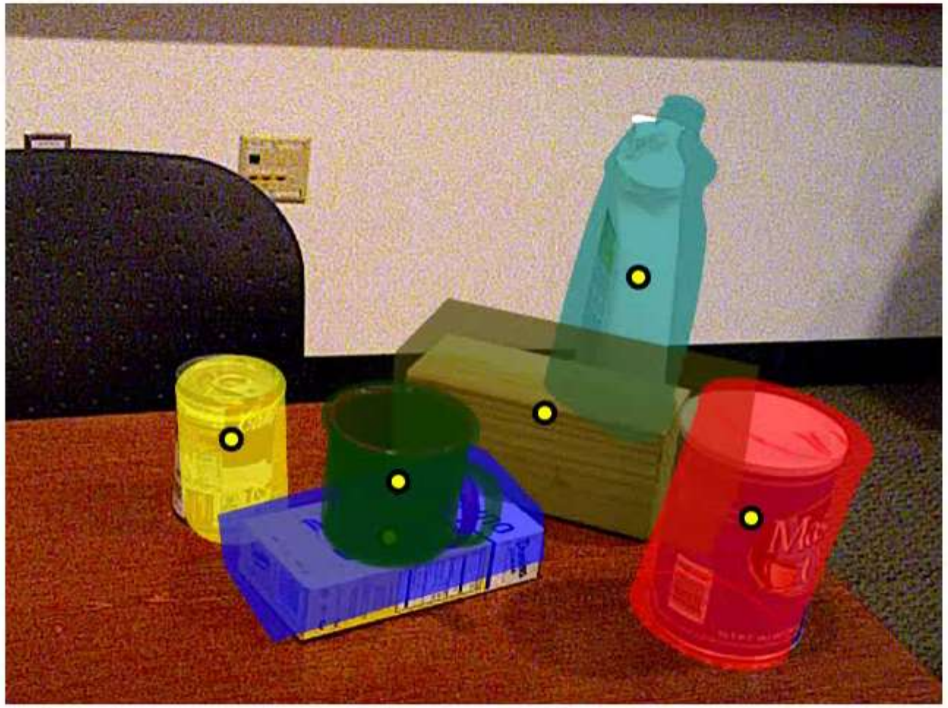
$$\mathrm{PLoss}(\tilde{\mathbf{q}}, \mathbf{q}) = \frac{1}{2m} \sum_{\mathbf{x} \in \mathcal{M}} \| R(\tilde{\mathbf{q}})\mathbf{x} - R(\mathbf{q})\mathbf{x} \|^2$$

Shape-Match Loss for symmetric objects (symmetric)

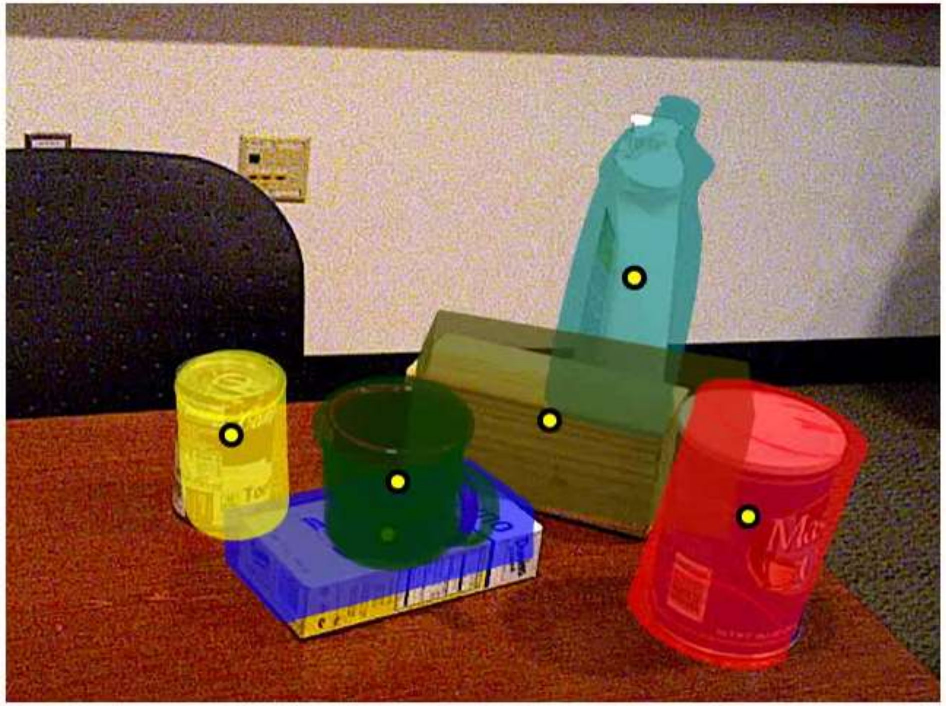
$$\mathrm{SLoss}(\tilde{\mathbf{q}}, \mathbf{q}) = \frac{1}{2m} \sum_{\mathbf{x}_1 \in \mathcal{M}} \min_{\mathbf{x}_2 \in \mathcal{M}} \| R(\tilde{\mathbf{q}})\mathbf{x}_1 - R(\mathbf{q})\mathbf{x}_2 \|^2$$

PoseCNN: Analysis on the Rotation Regression Loss

Symmetric loss for wood block



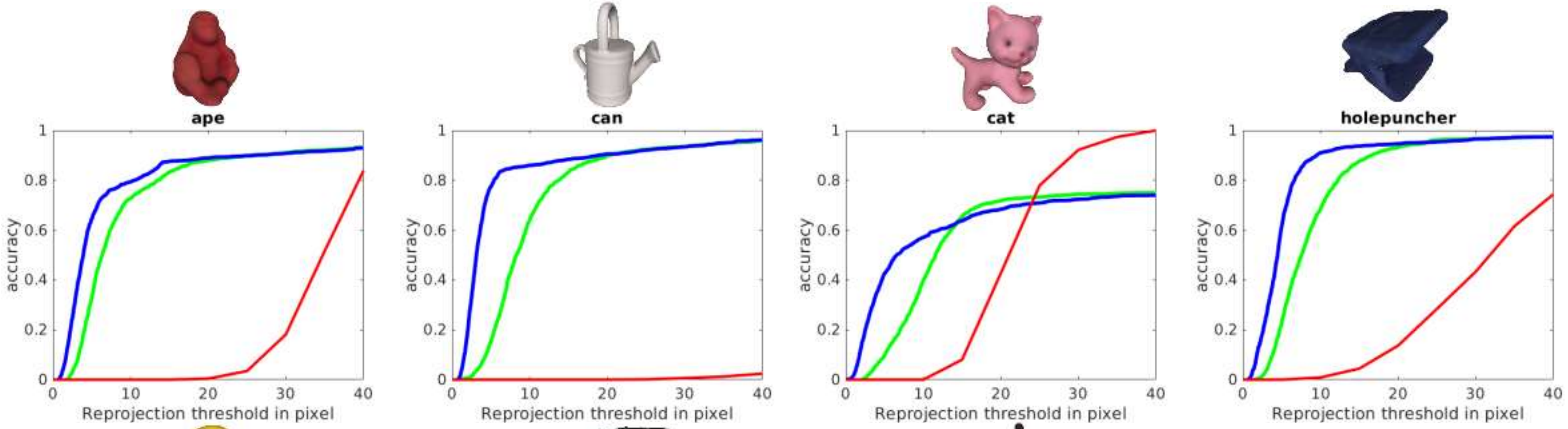
Non-symmetric loss for wood block

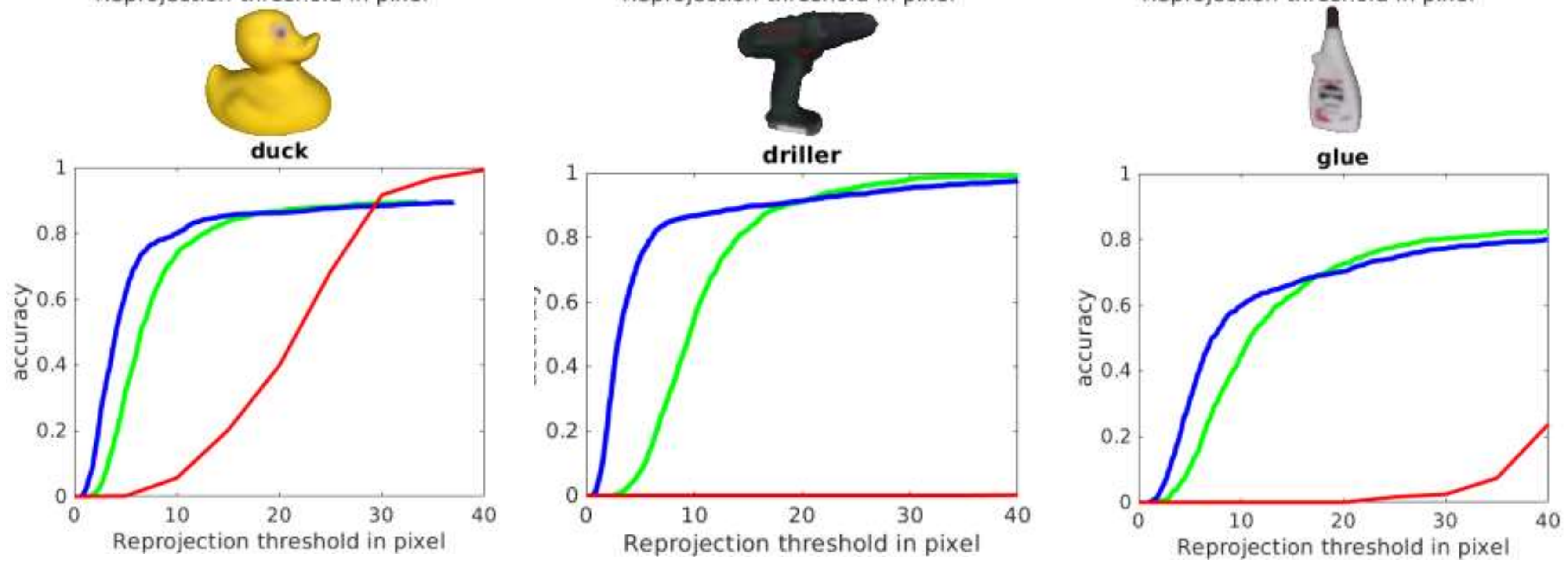


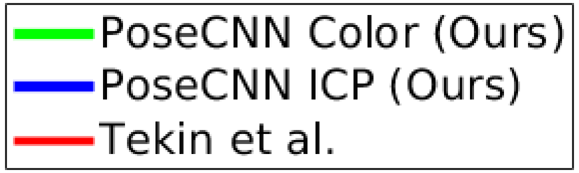
The LINEMOD Dataset [1]

[1] Hinterstoisser et al., Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In ACCV'12.

Results on the Occlusion LINEMOD Dataset







[1] B. Tekin, S. N Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In CVPR'18.

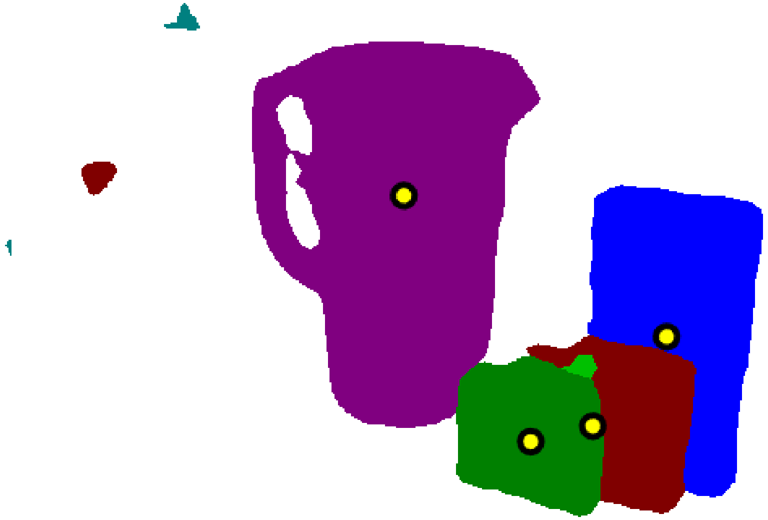
The YCB-Video Dataset

21 YCB Objects

92 Videos, 133,827 frames

Input Image

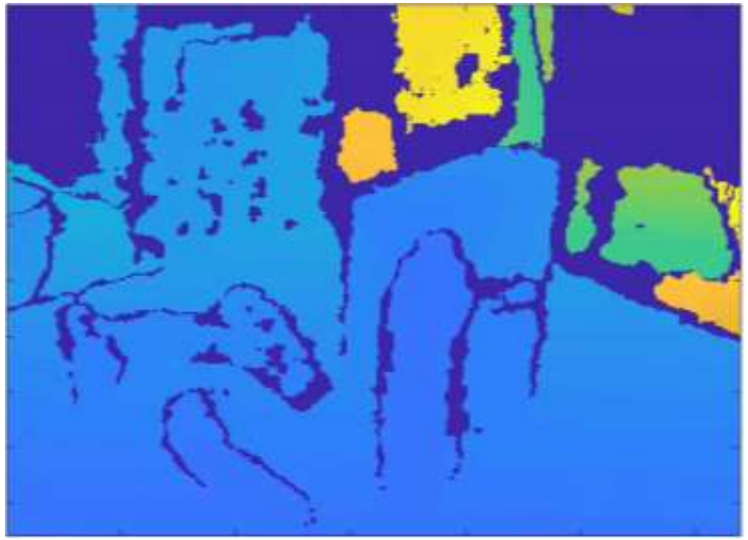
Labeling & Centers



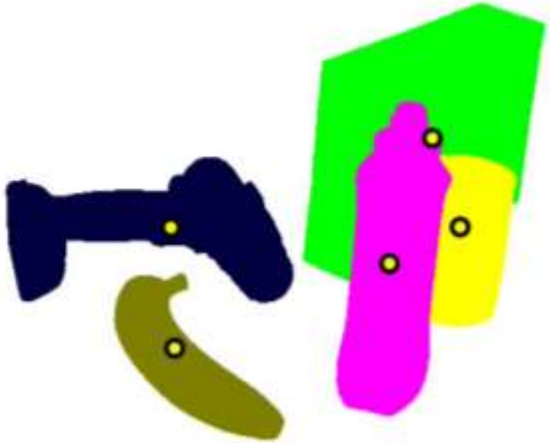
PoseCNN Color

PoseCNN ICP

RGB



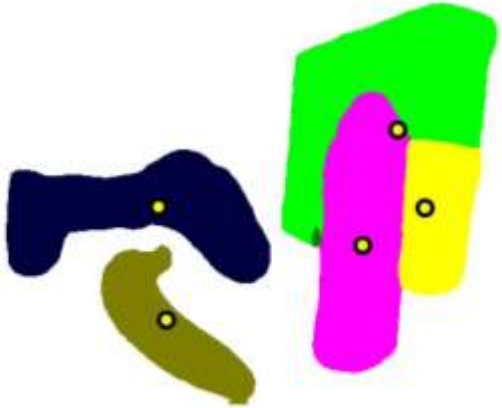
Depth



Groundtruth Labels

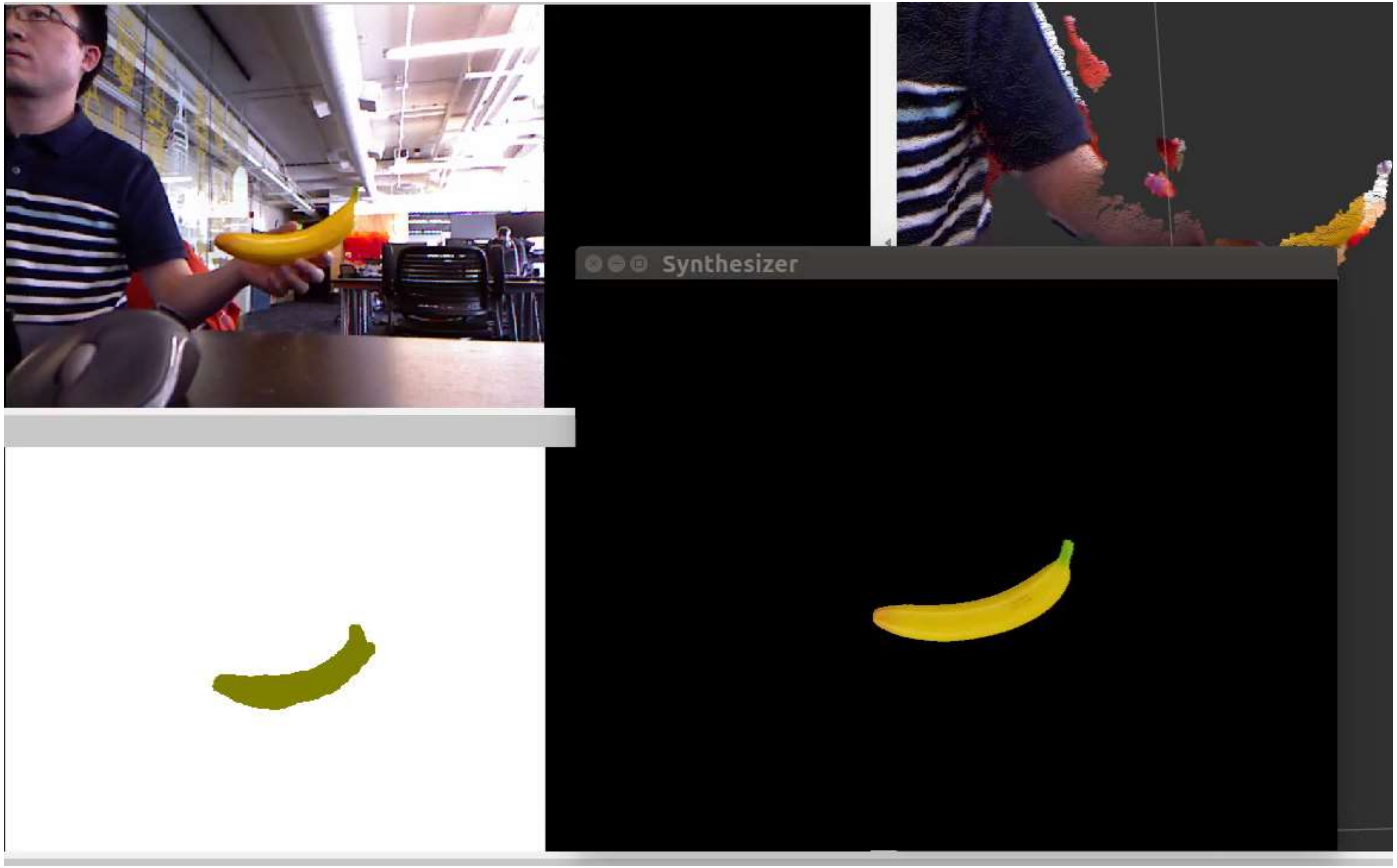
PoseCNN (RGB only)

PoseCNN + ICP



Predicted Labels

PoseCNN: Banana Pose Tracking Demo

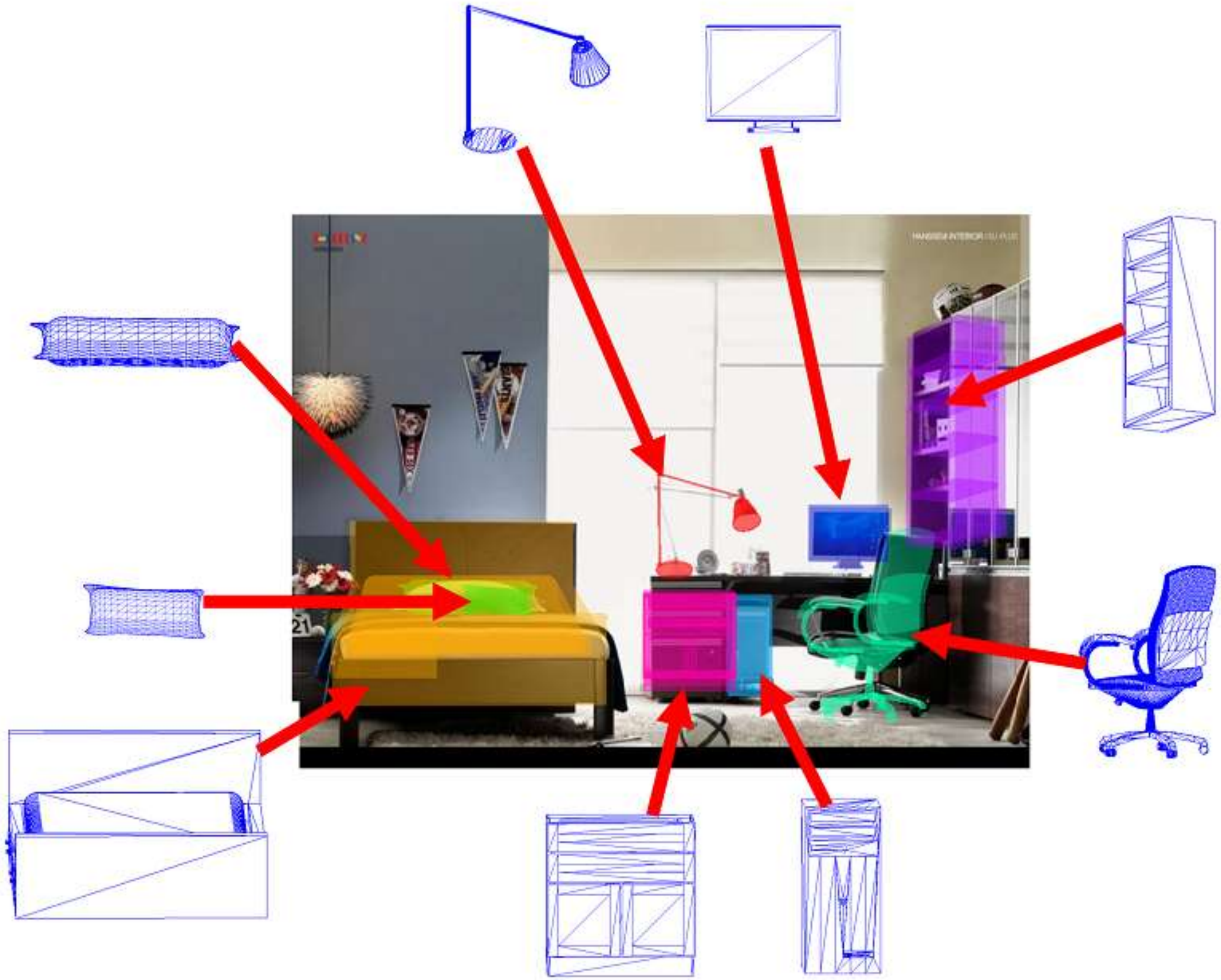


A Large Scale Database for 3D Object Recognition

Xiang et al., ObjectNet3D: A Large Scale Database for 3D Object Recognition. In ECCV, 2016 (Spotlight Oral).

3D Object Recognition for Object Categories

ObjectNet3D: A Large Scale Database for 3D Object Recognition



Xiang et al., ObjectNet3D: A Large Scale Database for 3D Object Recognition. In ECCV, 2016 (Spotlight Oral).

ObjectNet3D: Object Categories

- 100 rigid object categories

Aeroplane	Cap	Filing cabinet	Lighter	Remote control	Suitcase
Ashtray	Car	Fire extinguisher	Mailbox	Rifle	Teapot
Backpack	Cellphone	Fish tank	Microphone	Road pole	Telephone
Basket	Chair	Flashlight	Microwave	Satellite dish	Toaster
Bed	Clock	Fork	Motorbike	Scissors	Toilet
Bench	Coffee maker	Guitar	Mouse	Screwdriver	Toothbrush
Bicycle	Comb	Hair dryer	Paintbrush	Shoe	Train
Backboard	Computer	Hammer	Pan	Shovel	Trash bin
Boat	Cup	Headphone	Pen	Sign	Trophy
Bookshelf	Desk lamp	Helmet	Pencil	Skate	Tub
Bottle	Dining table	Iron	Piano	Skateboard	Tvmonitor
Bucket	Dishwasher	Jar	Pillow	Slipper	Vending machine
Bus	Door	Kettle	Plate	Sofa	Washing machine
Cabinet	Eraser	Key	Pot	Speaker	Watch
Calculator	Eyeglasses	Keyboard	Printer	Spoon	Wheelchair
Camera	Fan	Knife	Racket	Stapler	
Can	Faucet	Laptop	Refrigerator	Stove	

ObjectNet3D: Object Categories

- 100 rigid object categories

Aeroplane
Ashtray
Backpack
Basket
Bed
Bench
Bicycle
Backboard
Boat
Bookshelf
Bottle
Bucket
Bus
Cabinet
Calculator
Camera
Can

Cap
Car
Cellphone
Chair
Clock
Coffee maker
Comb
Computer
Cup
Desk lamp
Dining table
Dishwasher
Door
Eraser
Eyeglasses
Fan
Faucet

Filing cabinet
Fire extinguisher
Fish tank
Flashlight
Fork
Guitar
Hair dryer
Hammer
Headphone
Helmet
Iron
Jar
Kettle
Key
Keyboard
Knife
Laptop

Lighter
Mailbox
Microphone
Microwave
Motorbike
Mouse
Paintbrush
Pan
Pen
Pencil
Piano
Pillow
Plate
Pot
Printer
Racket
Refrigerator

Remote control
Rifle
Road pole
Satellite dish
Scissors
Screwdriver
Shoe
Shovel
Sign
Skate
Skateboard
Slipper
Sofa
Speaker
Spoon
Stapler
Stove

Suitcase
Teapot
Telephone
Toaster
Toilet
Toothbrush
Train
Trash bin
Trophy
Tub
Tvmonitor
Vending machine
Washing machine
Watch
Wheelchair

Vehicles
Furniture
Container
Tools
Electronics
Personal items

ObjectNet3D: Images

- 2D images from the ImageNet database [1]

[1] Deng et al., ImageNet: a Large Scale Hierarchical Image Database, CVPR, 2009

ObjectNet3D: 3D Shapes

- Trimble 3D Warehouse [1]
- ShapeNet database [2]

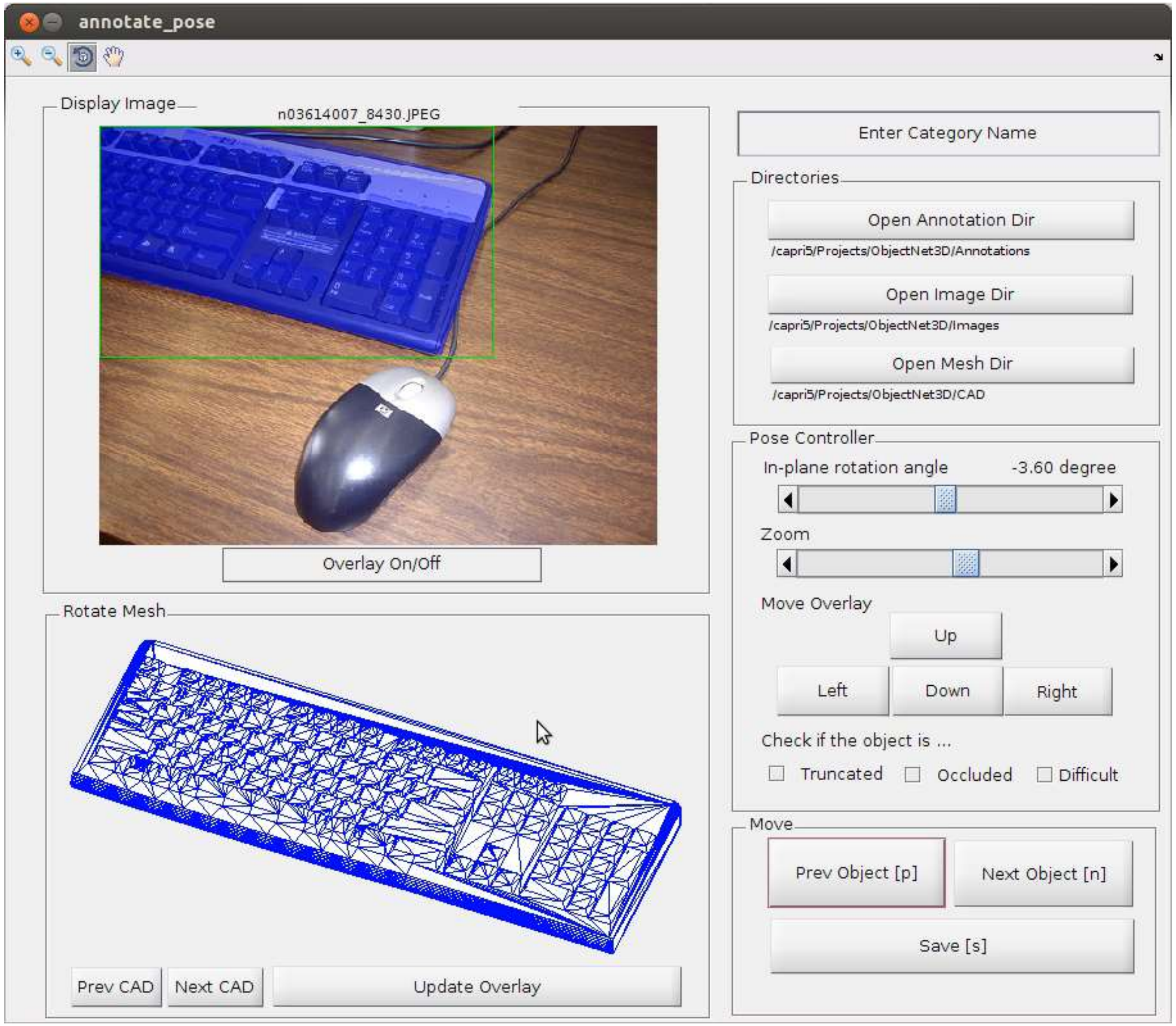
3D Shapes from Trimble 3D Warehouse

3D Shapes from ShapeNet

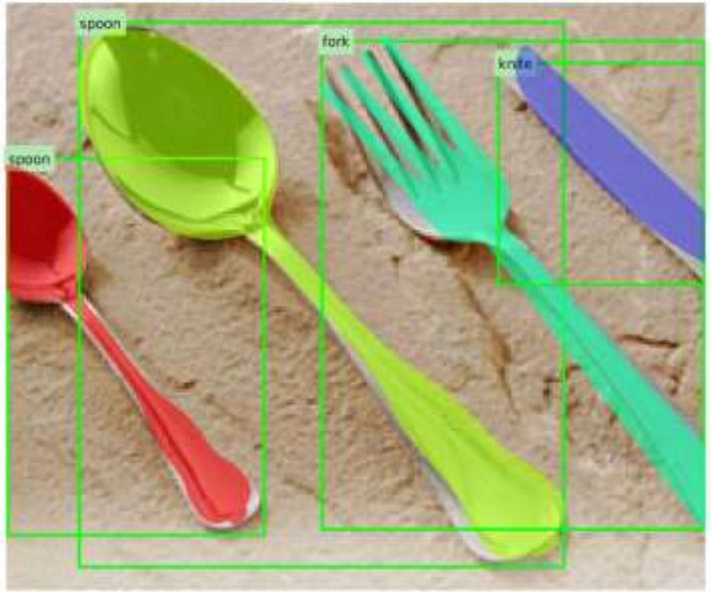
[1] https://3dwarehouse.sketchup.com

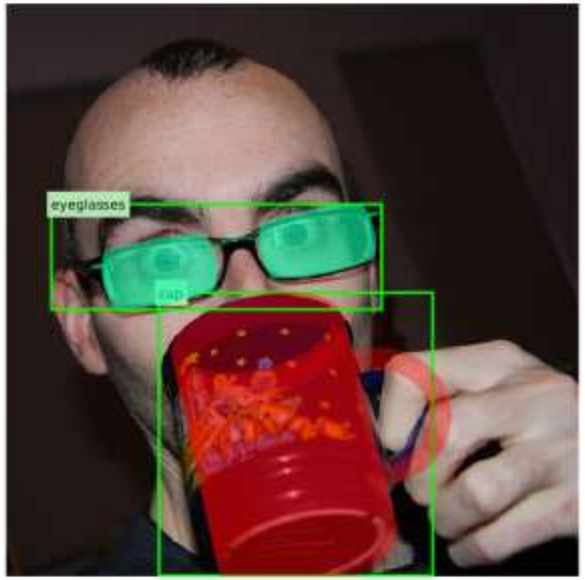
[2] Chang et al. ShapeNet: An Information-Rich 3D Model Repository, arXiv 2015

ObjectNet3D: Annotation Demo

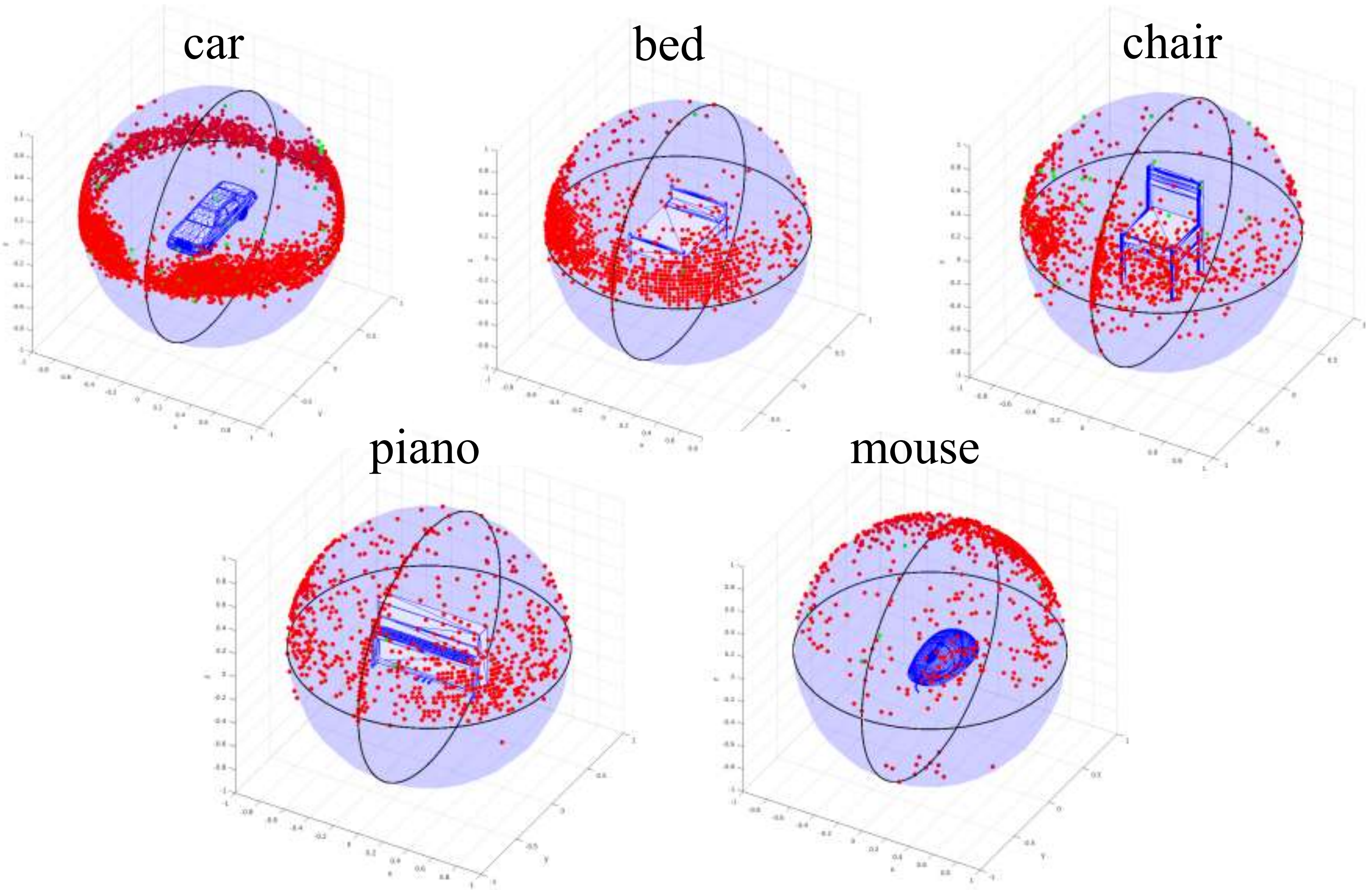


ObjectNet3D: 3D Pose Annotation Examples

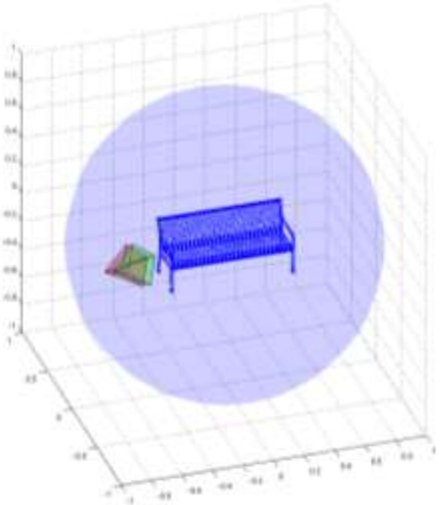


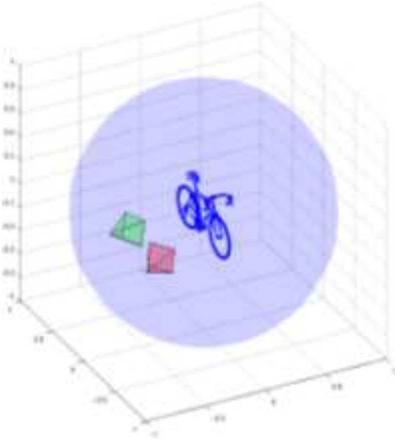


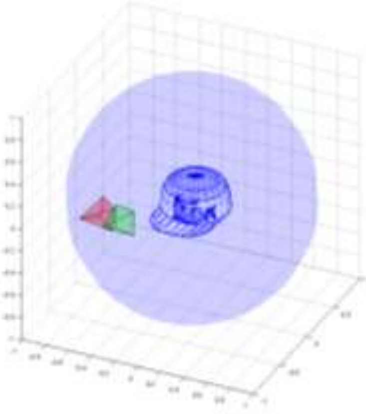
ObjectNet3D: Viewpoint Distributions

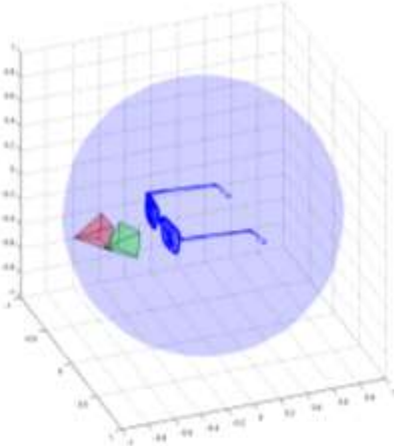


ObjectNet3D: Pose Estimation

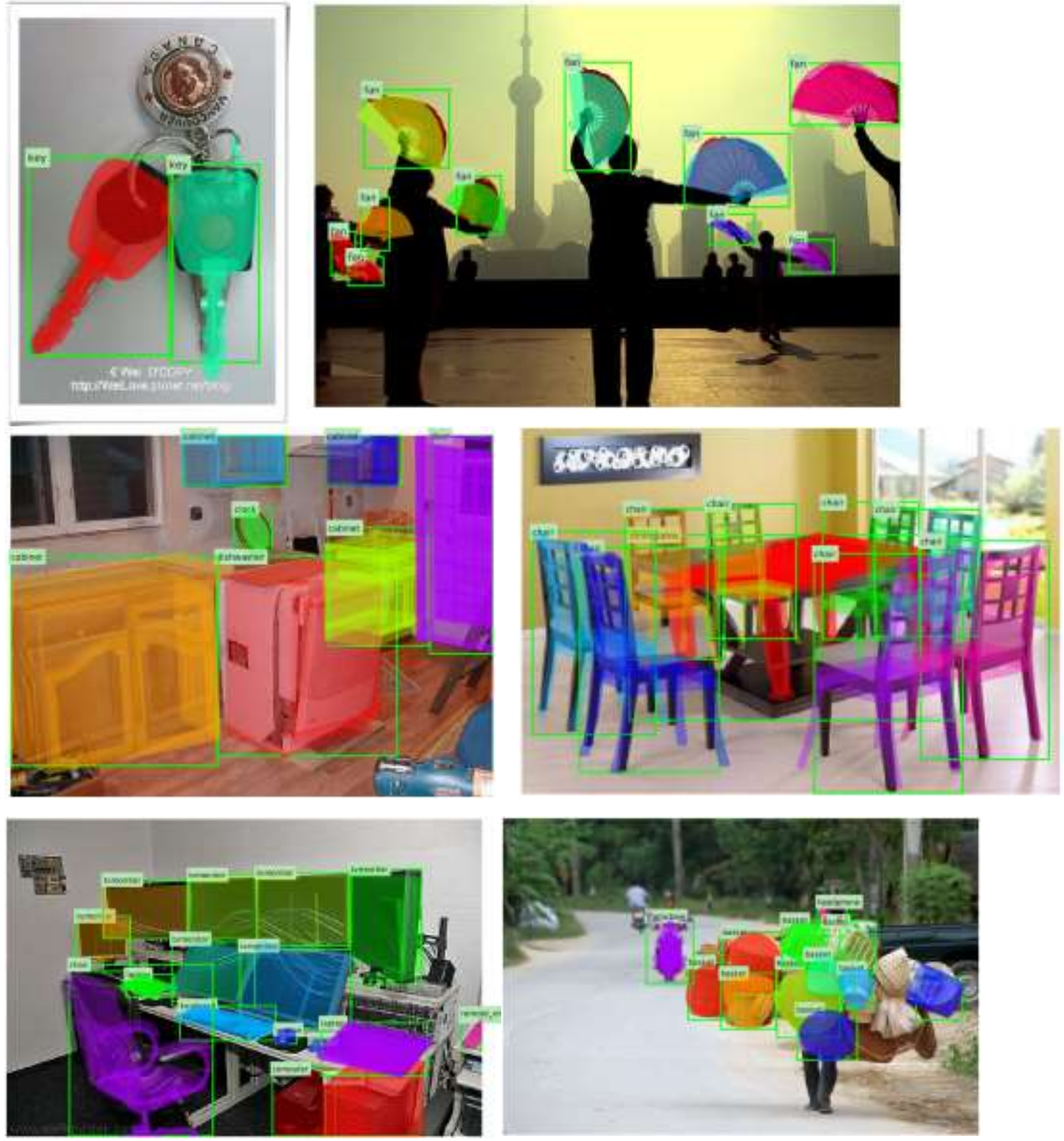








ObjectNet3D



- 100 object categories
- 90,127 images
- 201,888 objects
- 44,147 3D shapes
- 2D-3D alignments
- Baseline experiments on different recognition tasks

Conclusions

- DA-RNN: A recurrent neural network integrated with KinectFusion for 3D scene understanding

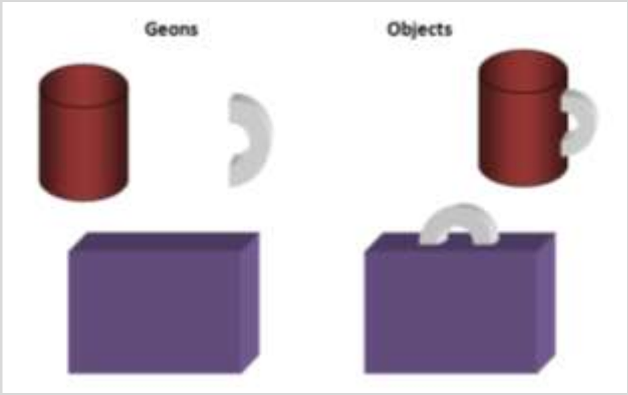
- PoseCNN: A generic convolutional neural network for 3D object recognition

- ObjectNet3D: A large scale database for 3D object recognition

- Deep neural networks with geometric representations

Future Work: Perception for Robotics

- Geometry

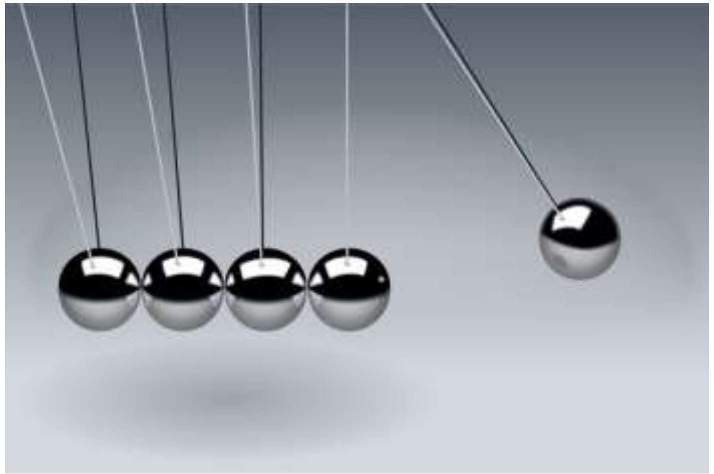


Biederman, Psychological Review, 1987

- Affordances

- Semantics & Language

- Physics & Common Sense

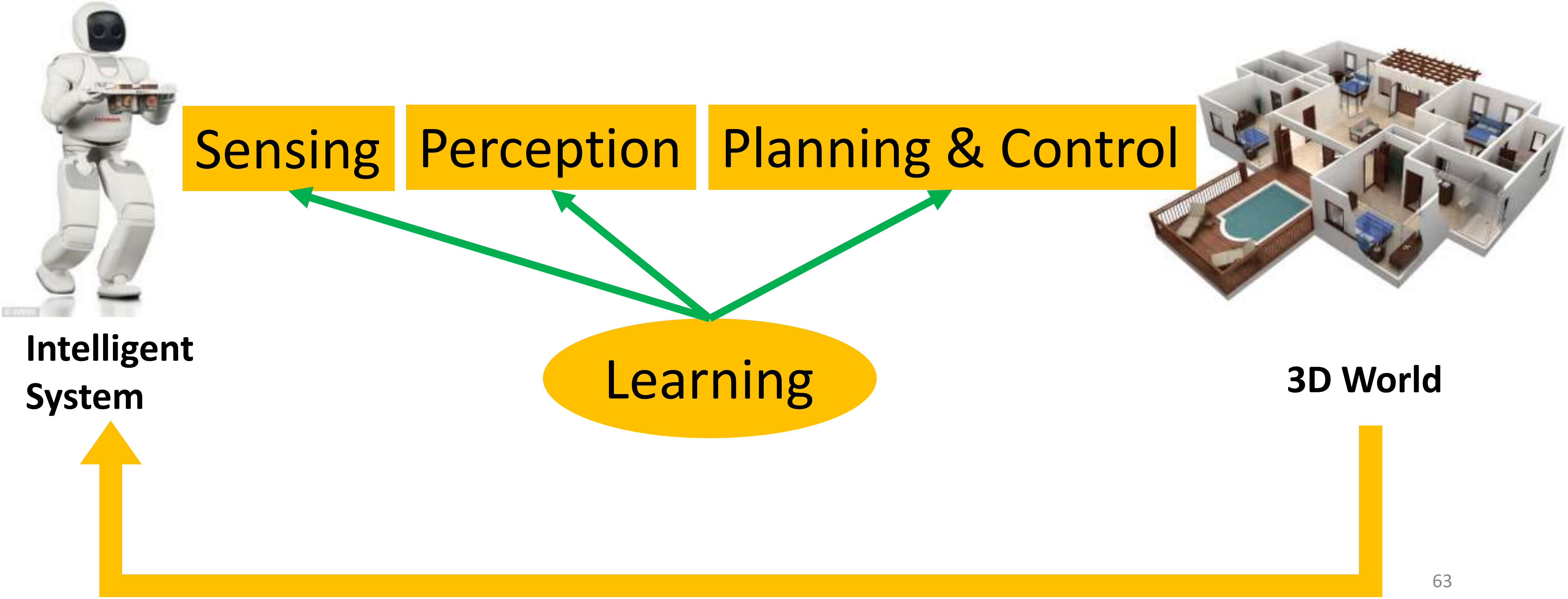


Future Work: Perception for Robotics

- Human behavior

Future Work: Perception for Robotics

- Integrating perception, planning and robot control



Acknowledgements










Thank you!