3D Object Recognition and Scene Understanding

Yu Xiang

University of Washington

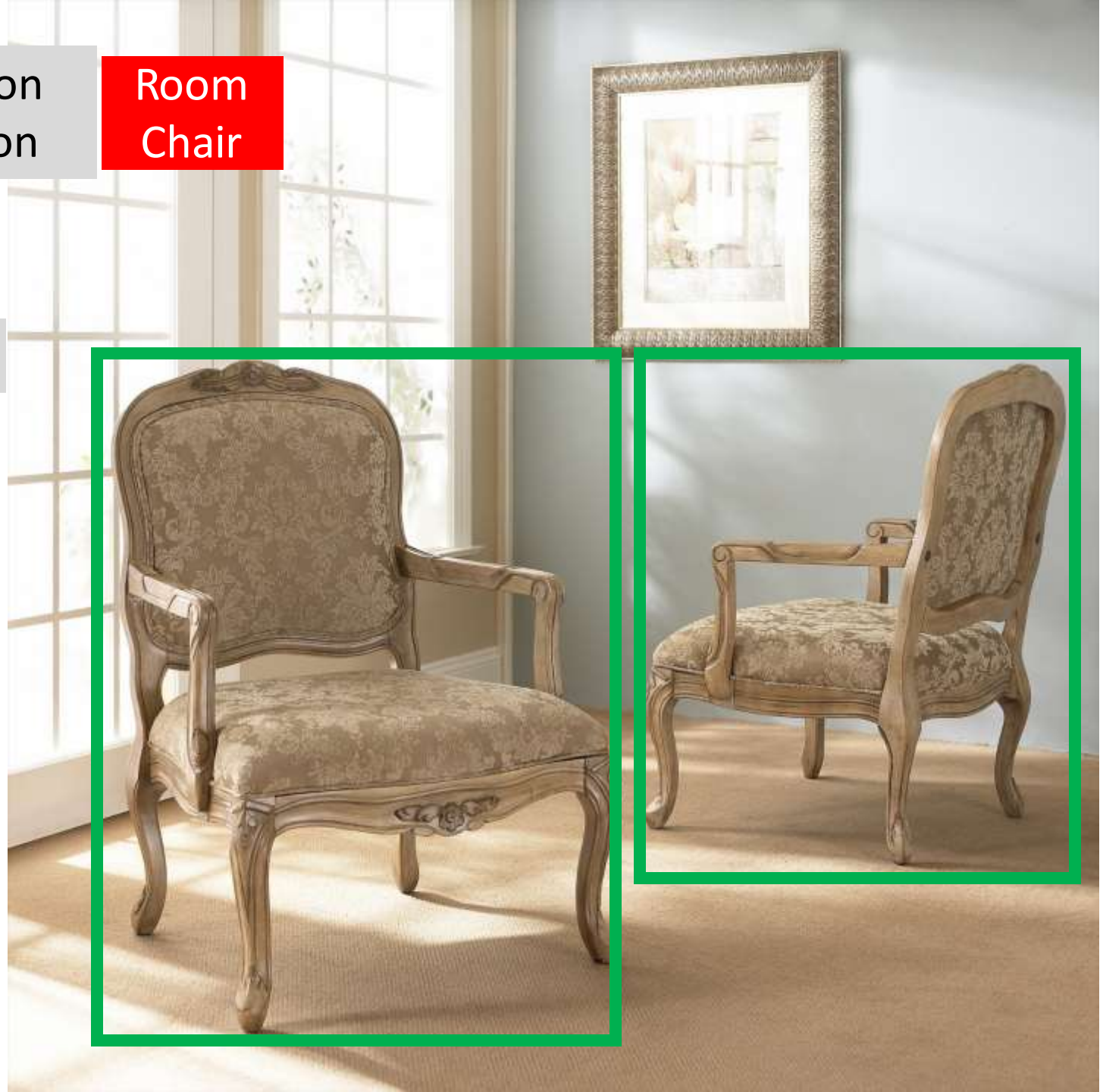
- Image classification tagging/annotation

Fergus et al. CVPR'03
Fei-Fei et al. CVPRW' 04
Chua et al. CIVR'09
Xiang et al. CVPR'10
Russakovsky et al. ECCV'12
Ordonez et al. ICCV'13
Deng et al. ECCV'14
…

- Image classification tagging/annotation

Room
Chair

- Object detection



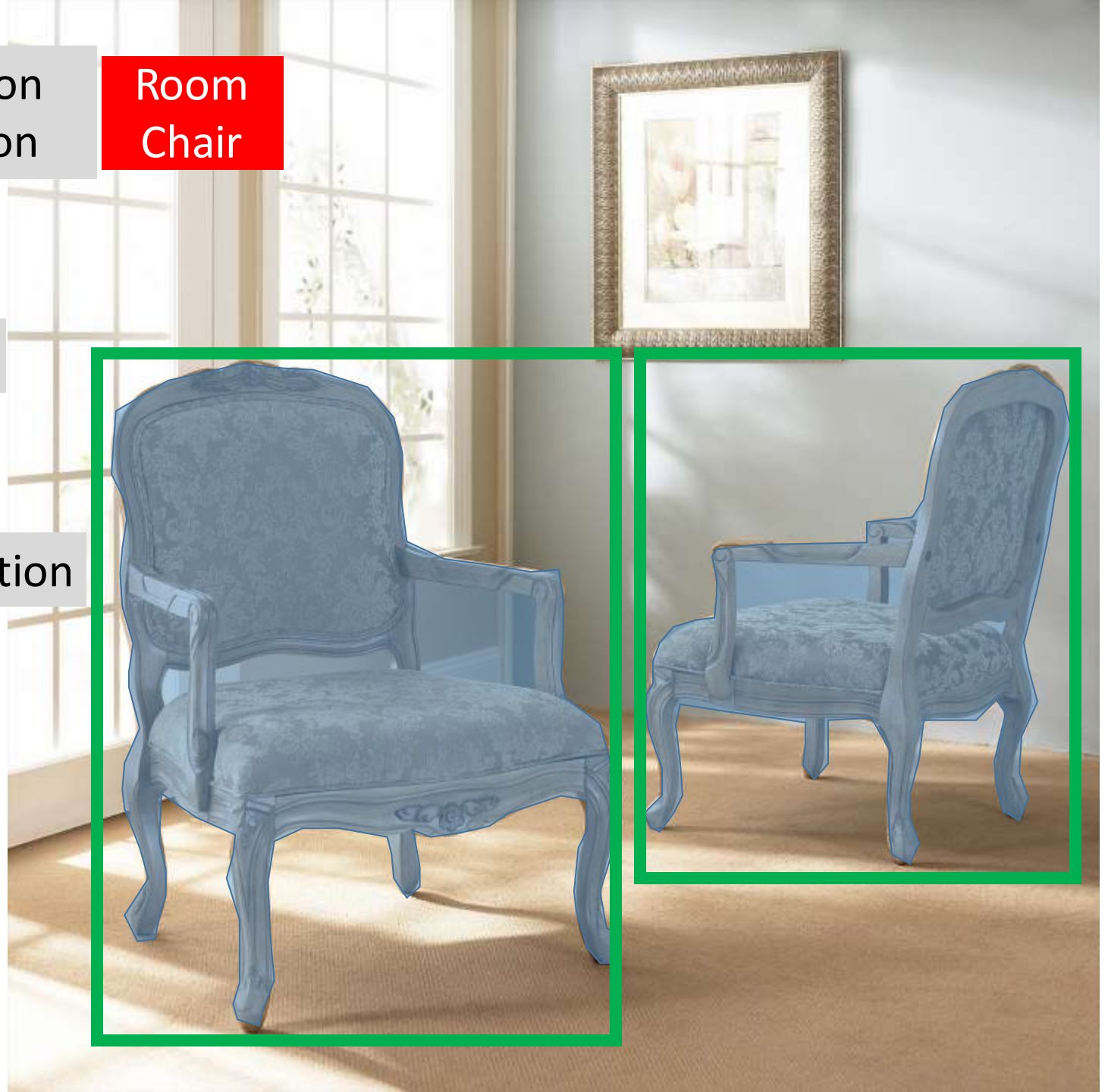
Viola & Jones. IJCV'04
Leibe et al. ECCVW'04
Dalal & Triggs. CVPR'05
Felzenszwalb et al. TPAMI' 10
Girshick et al. CVPR'14
Ren et al. NIPS'15
…

- Image classification tagging/annotation

Room
Chair

- Object detection

- Object segmentation



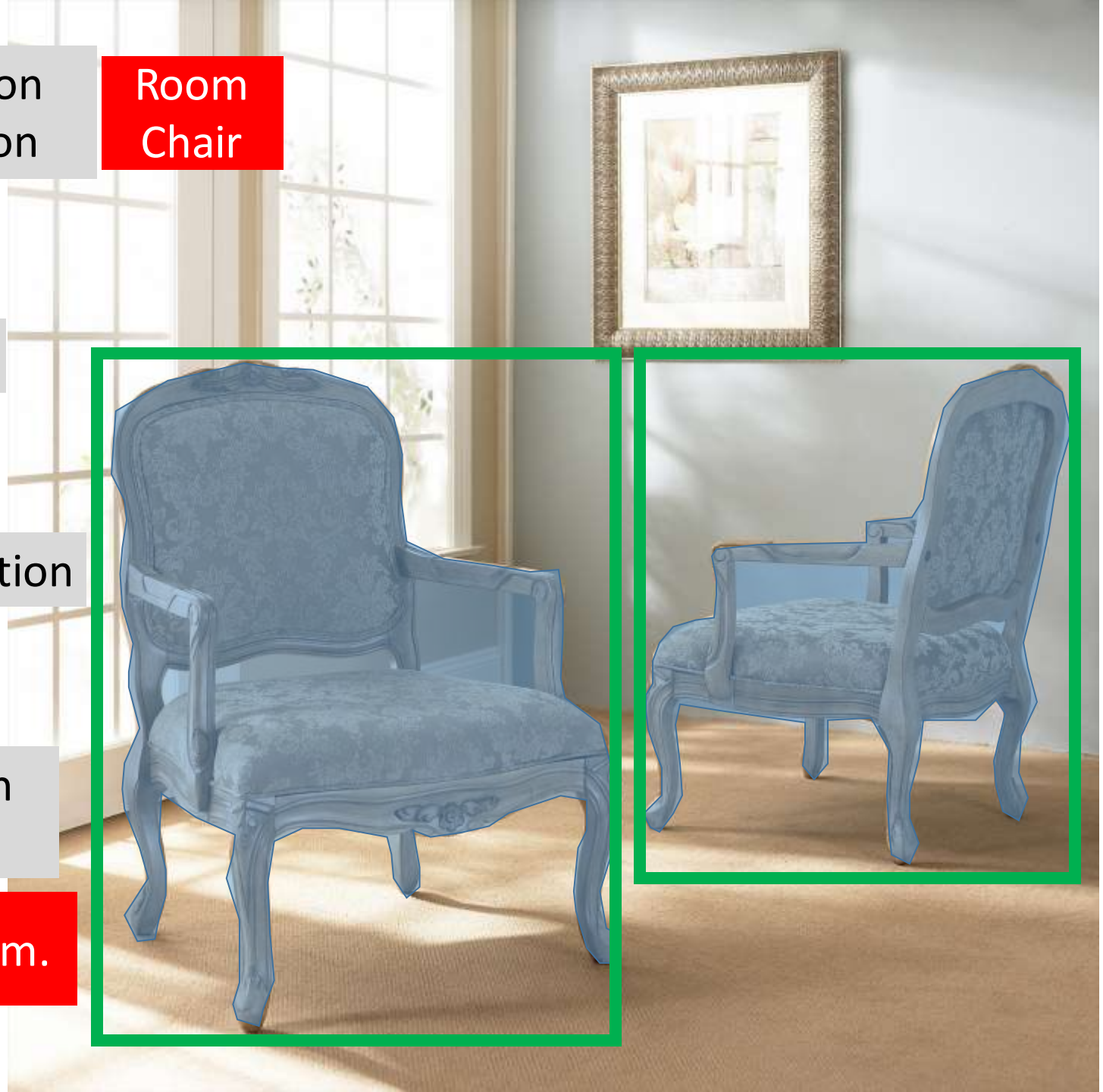
Shotton et al. IJCV'07
Pushmeet et al. IJCV'09
Ladicky et al. ECCV'10
Carreira et al. ECCV'12
Chen et al. ICLR'15
Long et al. CVPR'15
…

- Image classification tagging/annotation

Room
Chair

- Object detection
- Object segmentation
- Image description generation

Two chairs in a room.



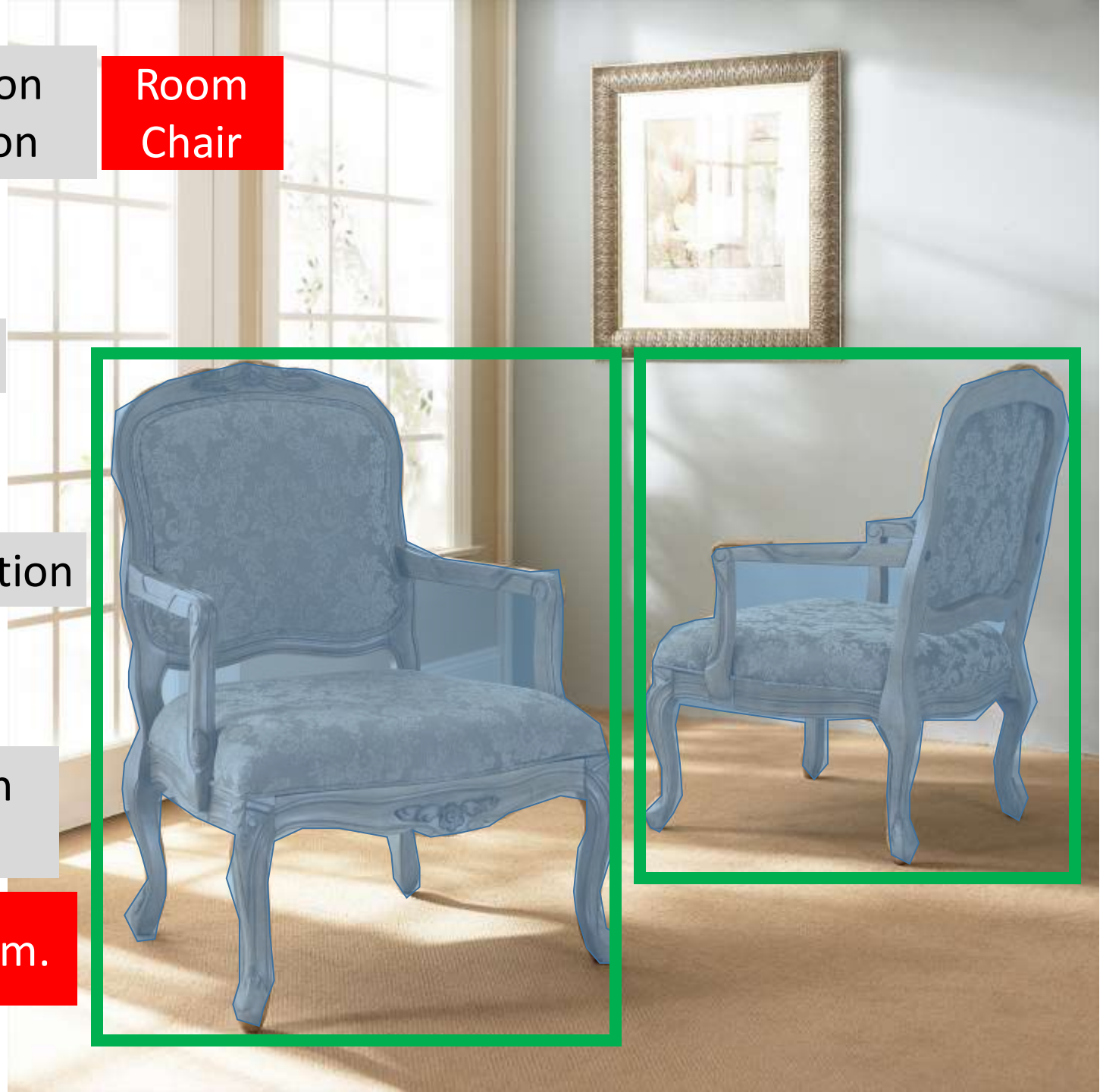
Kulkarni et al. CVPR'11
Karpathy & Fei-Fei. CVPR'15
Chen & Zitnik. CVPR'15
Gregor et al. ICML'15
Johnson et al. CVPR'16
…

- Image classification tagging/annotation

Room
Chair

- Object detection
- Object segmentation
- Image description generation

Two chairs in a room.



2D Recognition

- Image classification tagging/annotation

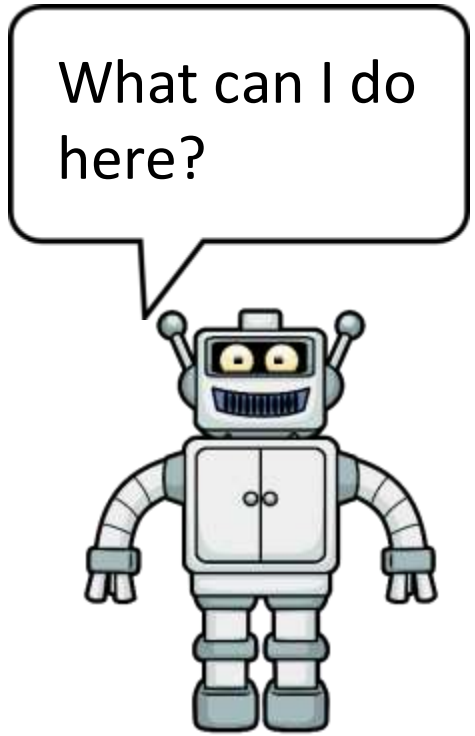
Room
Chair

- Object detection

- Object segmentation

- Image description generation

Two chairs in a room.



- Image classification tagging/annotation

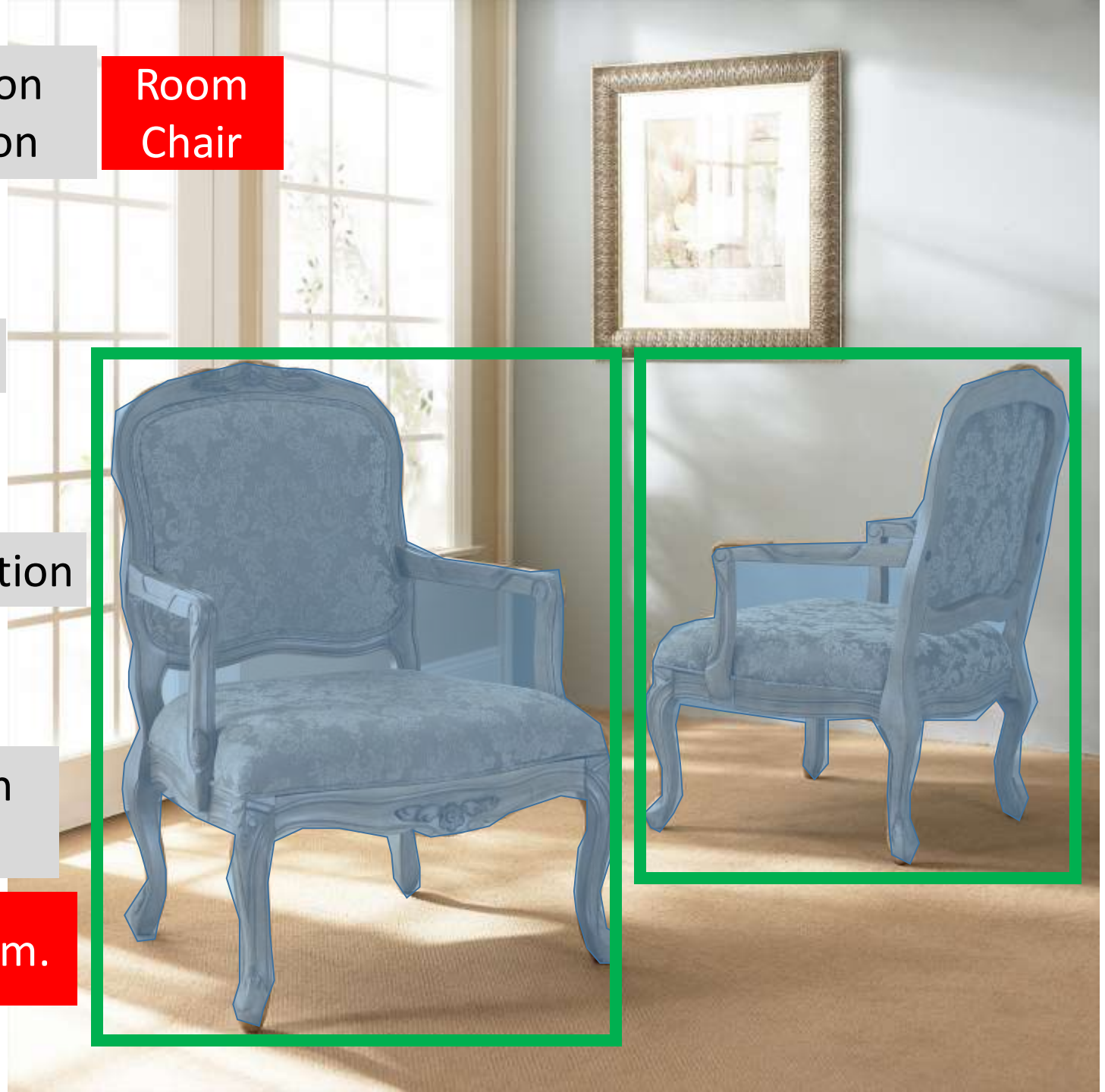
Room
Chair

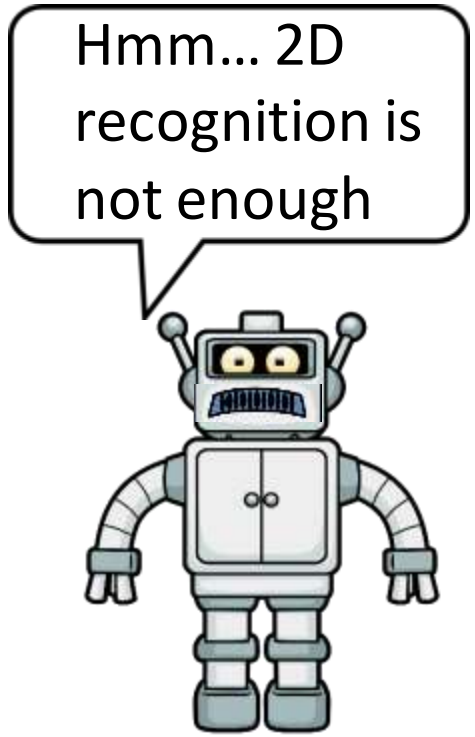
- Object detection

- Object segmentation

- Image description generation

Two chairs in a room.

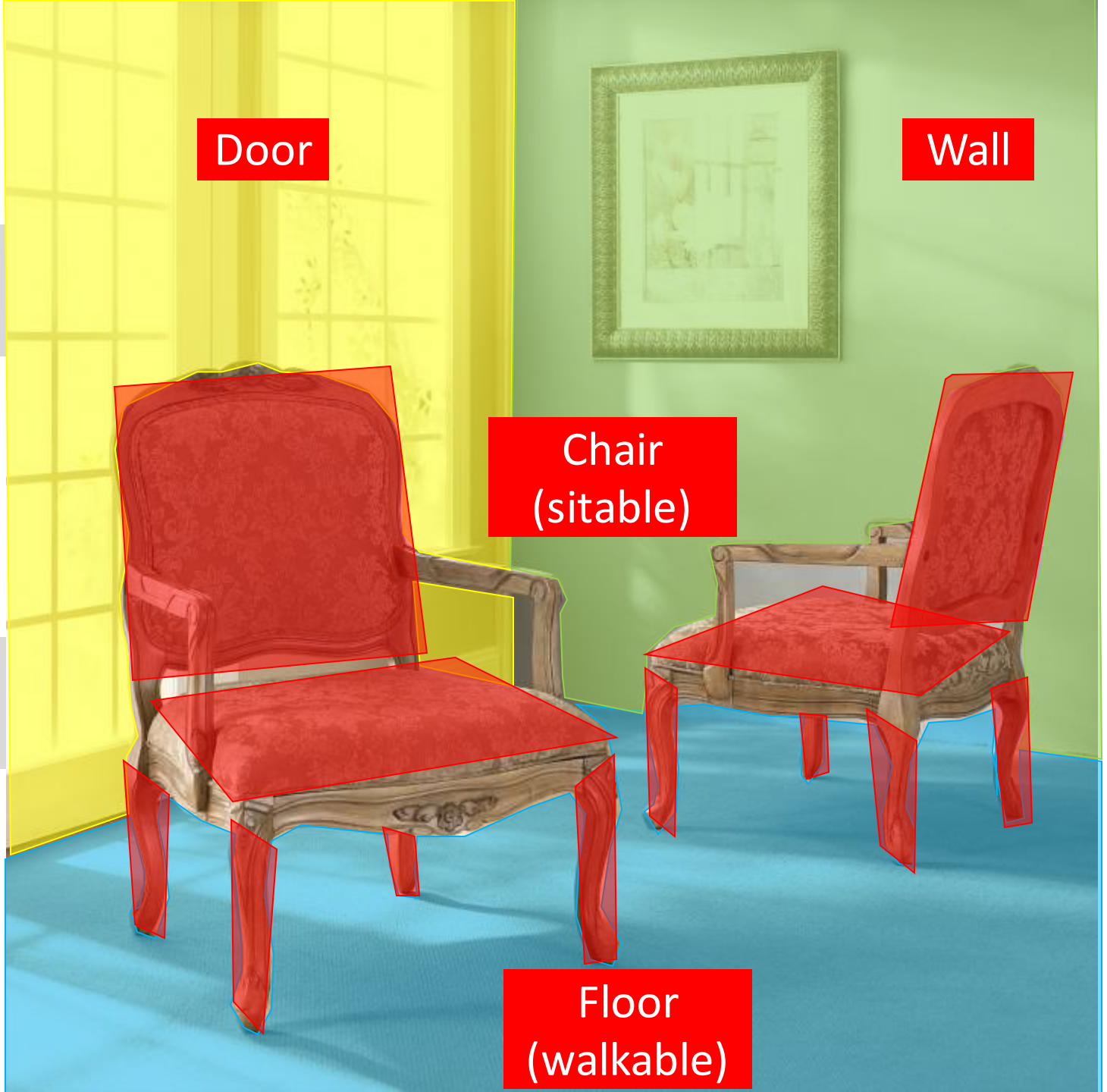




- 3D Scene Understanding

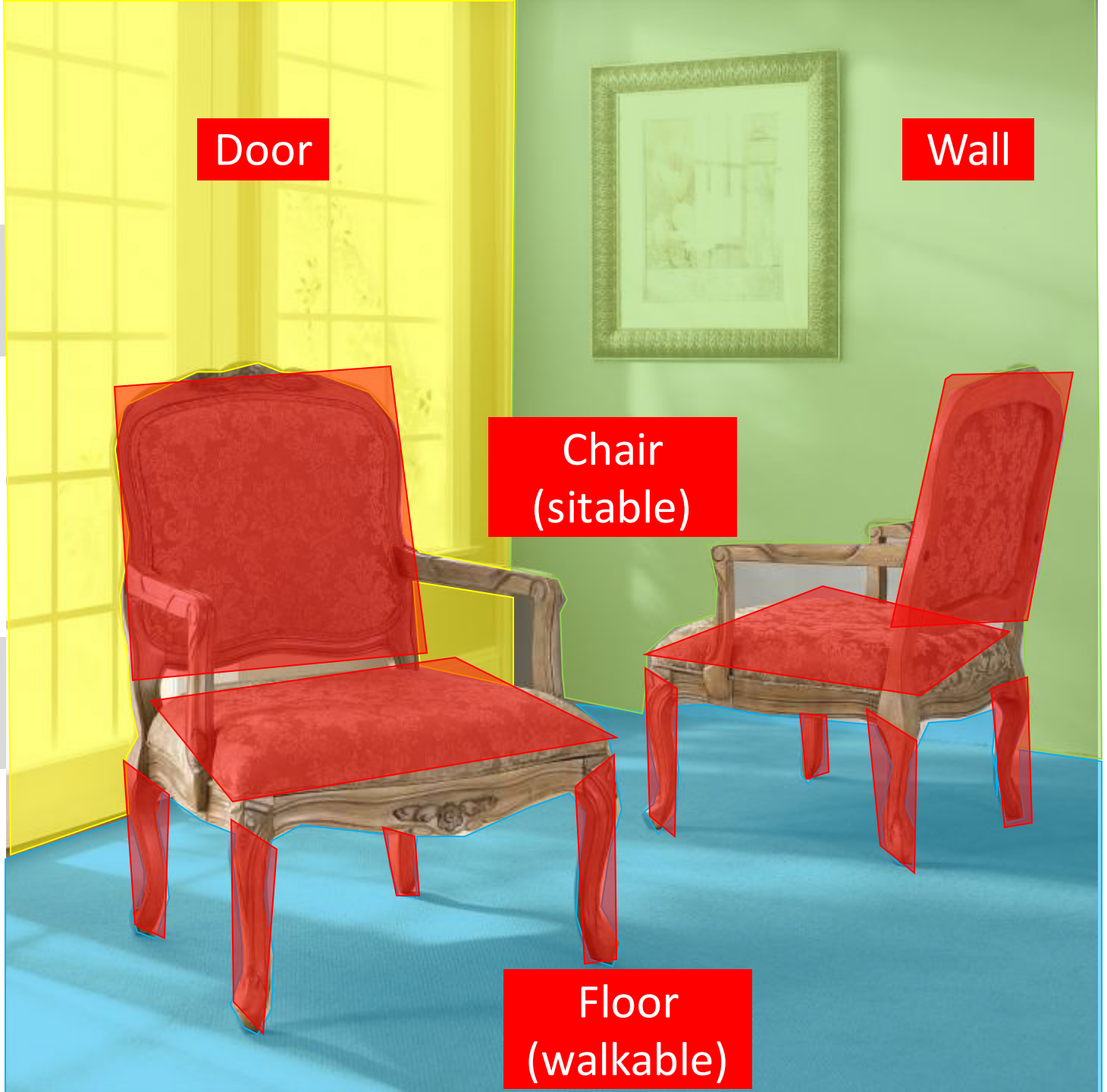
Hoiem et al., ICCV'05
Lee et al. CVPR'09
Hedau, el al., ICCV'09
Fouhey et al. ICCV'13
Schwing et al. ICCV'13
Lai, Bo & Fox. ICRA'14
Mallya & Lazebnik, ICCV'15
…

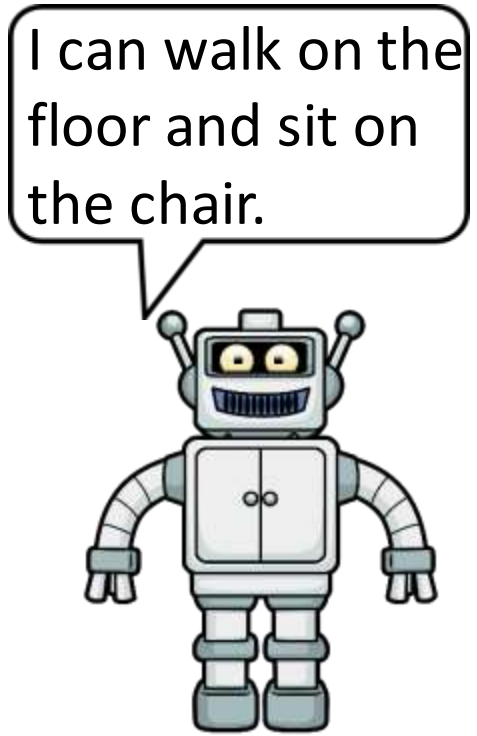
- 3D Scene Understanding
- 3D Object Recognition



Savarese & Fei-Fei, ICCV'07
Sun et al. CVPR'09
Stark et al. BMVC'10
Glasner et al. ICCV'11
Pepik et al. CVPR'12
Xiang & Savarese, CVPR'12
Kar et al., ICCV'15
Tulsiani & Malik, CVPR'15
…

- 3D Scene Understanding
- 3D Object Recognition





Applications that need 3D recognition

Autonomous Driving

Robotics

Any application that requires interaction with the 3D world!

Augmented Reality

Gaming

Outline

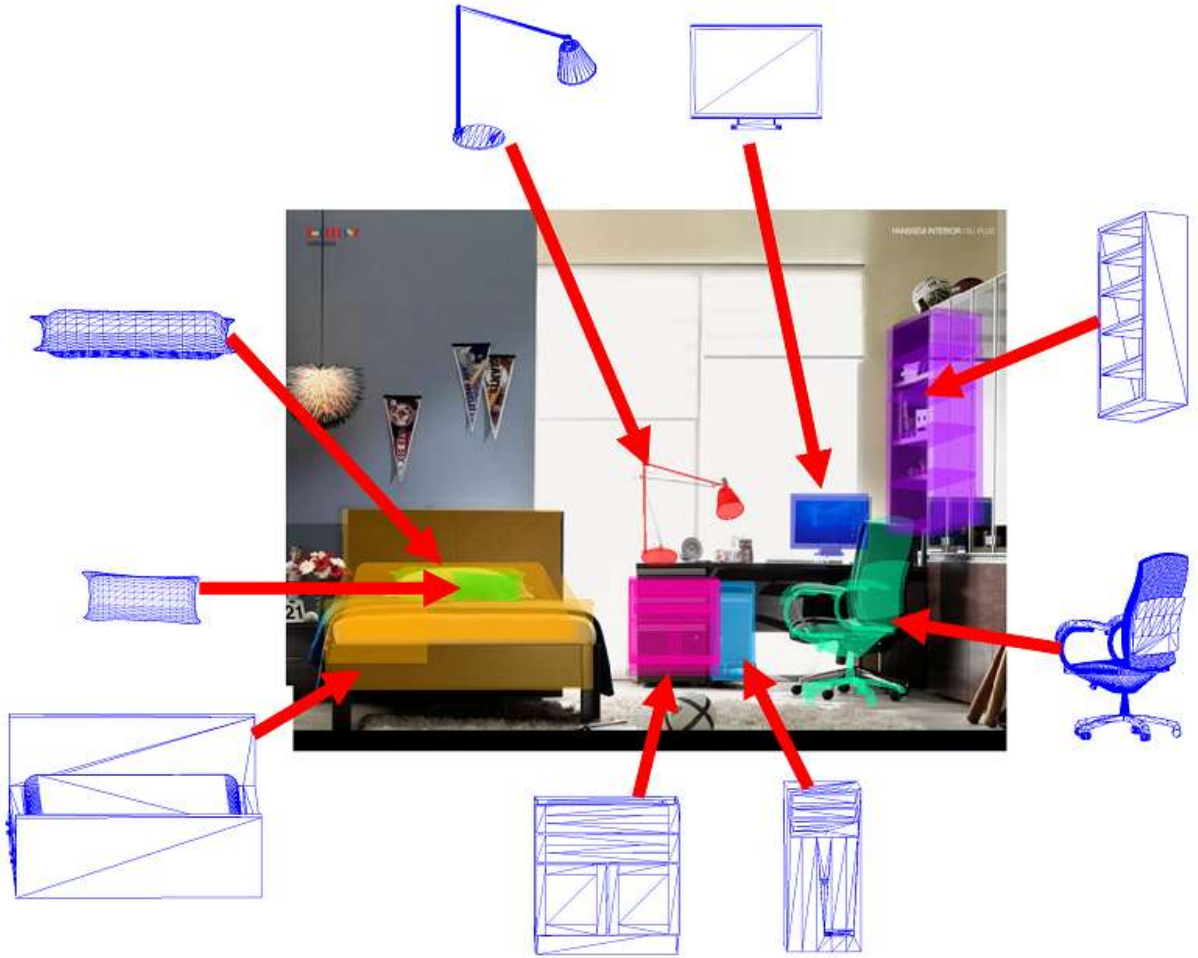
- ObjectNet3D: A Large Scale Database for 3D Object Recognition

- DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks

ObjectNet3D Database

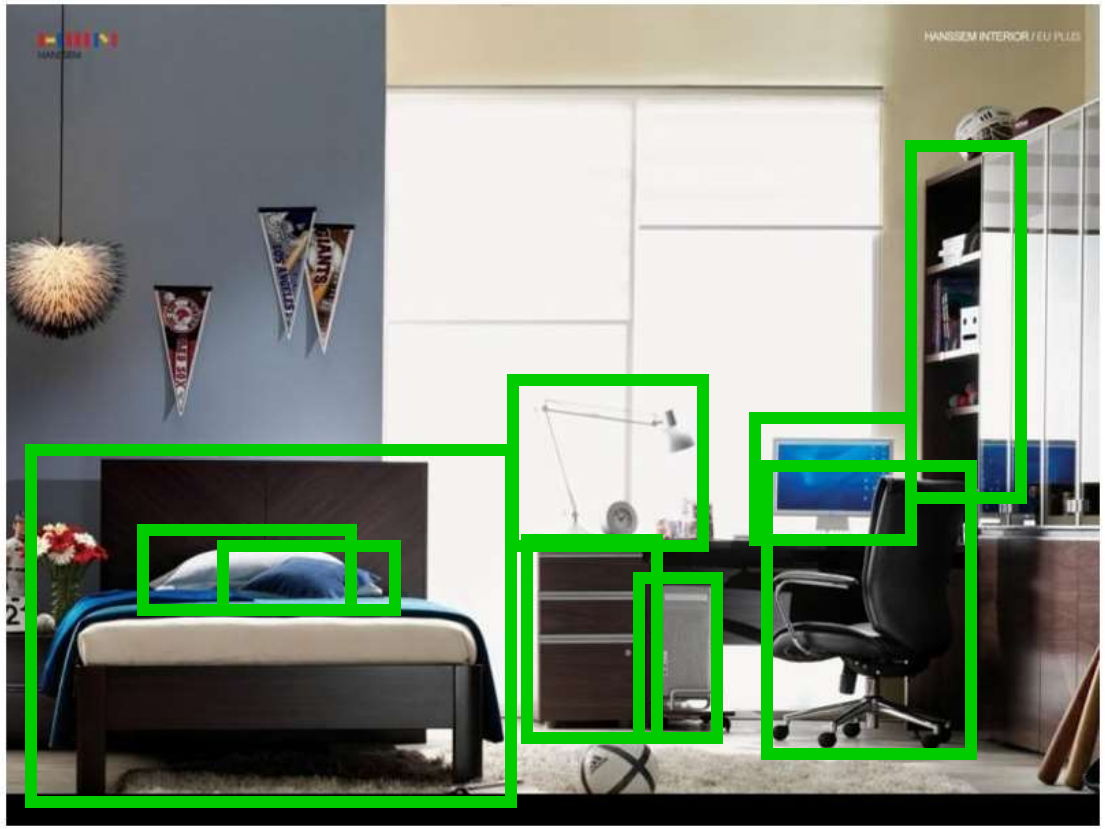
Xiang et al. ECCV'16

- A large scale database for 3D object recognition



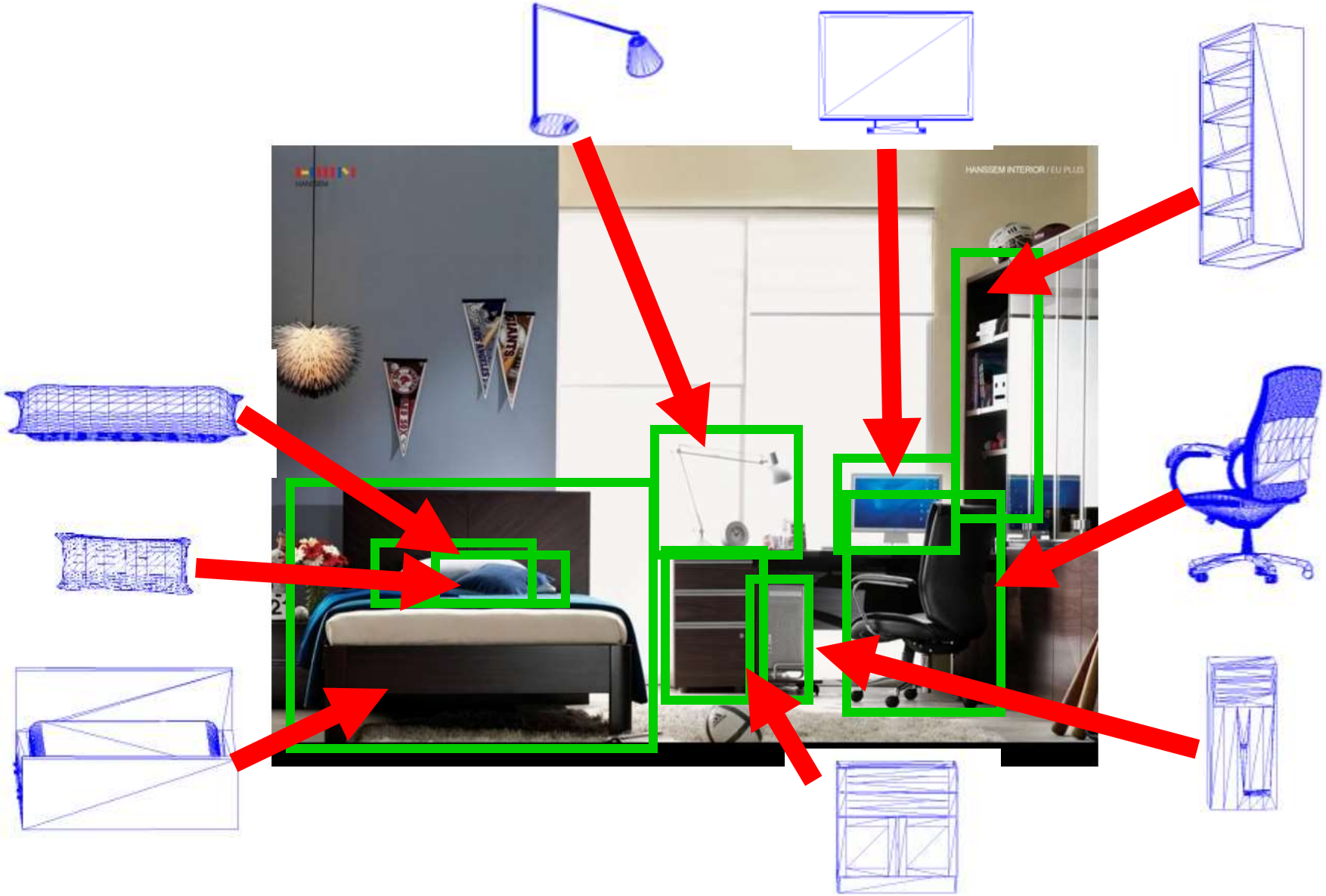
3D Annotation: 2D-3D Alignment

3D Annotation: 2D-3D Alignment

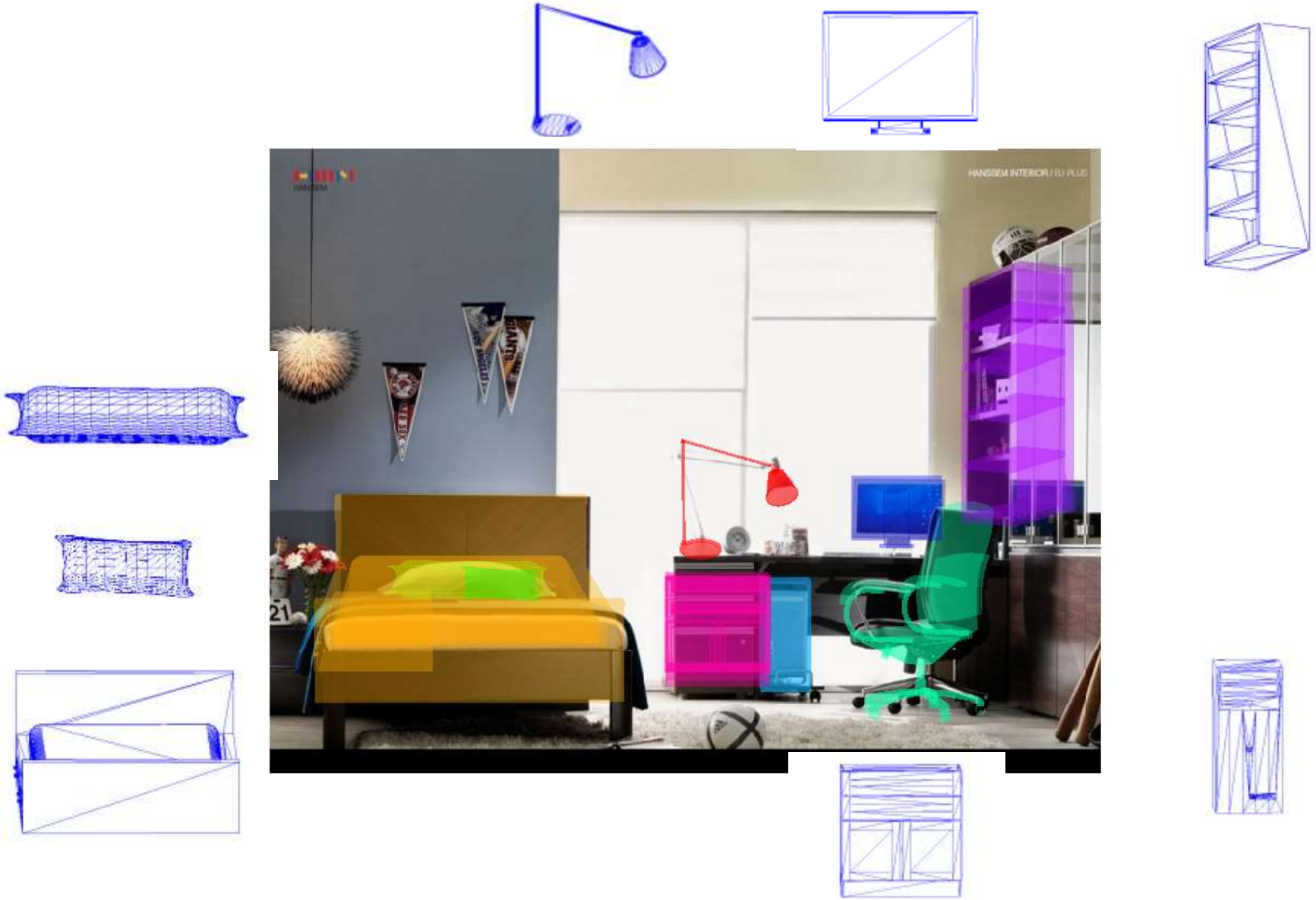


3D Annotation: 2D-3D Alignment

3D Annotation: 2D-3D Alignment



3D Annotation: 2D-3D Alignment



Comparison with Previous Datasets

	#category
3D Object [1]	10
EPFL Car [2]	1
RGB-D Object [3]	51
PASCAL VOC [4]	20
KITTI [5]	3
PASCAL3D+ [6]	12
ObjectNet3D (Ours)	**100**

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[3] K. Lai, L. Bo, X. Ren and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In ICRA, 2011.
[4] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[6] Y. Xiang, R. Mottaghi and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV, 2014.

Comparison with Previous Datasets

	#category	#instance
3D Object [1]	10	100
EPFL Car [2]	1	20
RGB-D Object [3]	51	300
PASCAL VOC [4]	20	27,450
KITTI [5]	3	80,256
PASCAL3D+ [6]	12	35,672
ObjectNet3D (Ours)	**100**	**201,888**

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[3] K. Lai, L. Bo, X. Ren and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In ICRA, 2011.
[4] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[6] Y. Xiang, R. Mottaghi and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV, 2014.

Comparison with Previous Datasets

	#category	#instance	Non-centered objects
3D Object [1]	10	100	✘
EPFL Car [2]	1	20	✘
RGB-D Object [3]	51	300	✘
PASCAL VOC [4]	20	27,450	✓
KITTI [5]	3	80,256	✓
PASCAL3D+ [6]	12	35,672	✓
ObjectNet3D (Ours)	**100**	**201,888**	✓

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[3] K. Lai, L. Bo, X. Ren and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In ICRA, 2011.
[4] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[6] Y. Xiang, R. Mottaghi and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV, 2014.

Comparison with Previous Datasets

	#category	#instance	Non-centered objects	Dense viewpoint
3D Object [1]	10	100	✗	✗
EPFL Car [2]	1	20	✗	✓
RGB-D Object [3]	51	300	✗	✓
PASCAL VOC [4]	20	27,450	✓	✗
KITTI [5]	3	80,256	✓	✓
PASCAL3D+ [6]	12	35,672	✓	✓
ObjectNet3D (Ours)	**100**	**201,888**	✓	✓

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[3] K. Lai, L. Bo, X. Ren and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In ICRA, 2011.
[4] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[6] Y. Xiang, R. Mottaghi and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV, 2014.

Comparison with Previous Datasets

	#category	#instance	Non-centered objects	Dense viewpoint	3D Shape
3D Object [1]	10	100	✗	✗	✗
EPFL Car [2]	1	20	✗	✓	✗
RGB-D Object [3]	51	300	✗	✓	✗
PASCAL VOC [4]	20	27,450	✓	✗	✗
KITTI [5]	3	80,256	✓	✓	✗
PASCAL3D+ [6]	12	35,672	✓	✓	✓79
ObjectNet3D (Ours)	**100**	**201,888**	✓	✓	✓**44,147**

[1] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In ICCV, 2007.
[2] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.
[3] K. Lai, L. Bo, X. Ren and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In ICRA, 2011.
[4] M. Everingham, L. Van Gool, C. K. I.Williams, J.Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. IJCV, 2010.
[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[6] Y. Xiang, R. Mottaghi and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In WACV, 2014.

Database Construction: 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	

Database Construction: 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	

Vehicles
Furniture
Container
Tools
Electronics
Personal items

Database Construction: Images

- 2D images from the ImageNet database [1]

[1] Russakovsky et al. ImageNet Large Scale Visual Recognition Challenge, IJCV 2015

Database Construction: 3D Shapes

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

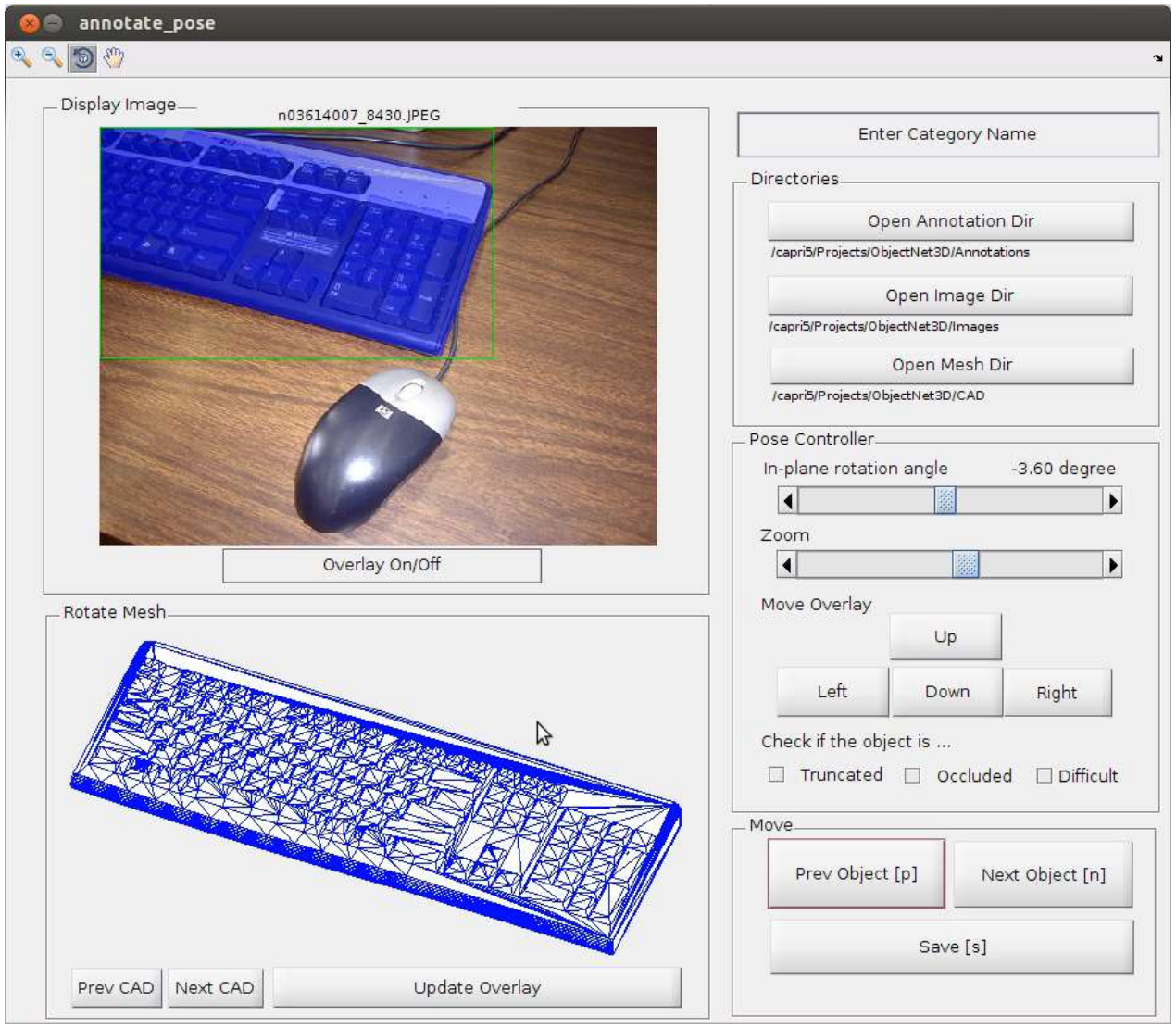
3D Shapes from Trimble 3D Warehouse

3D Shapes from ShapeNet

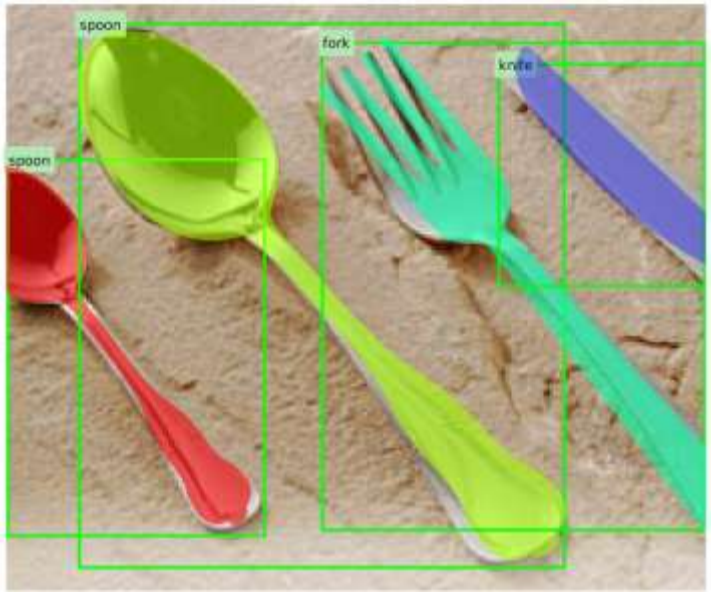
[1] https://3dwarehouse.sketchup.com

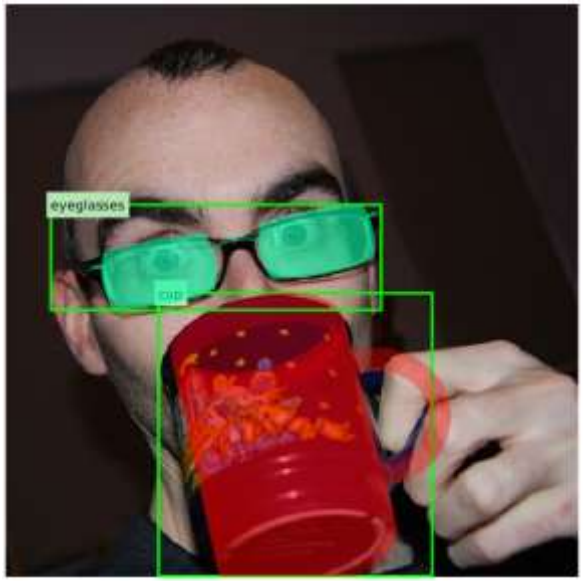
[2] Chang et al. ShapeNet: An Information-Rich 3D Model Repository, arXiv 2015

Database Construction: Annotation Demo

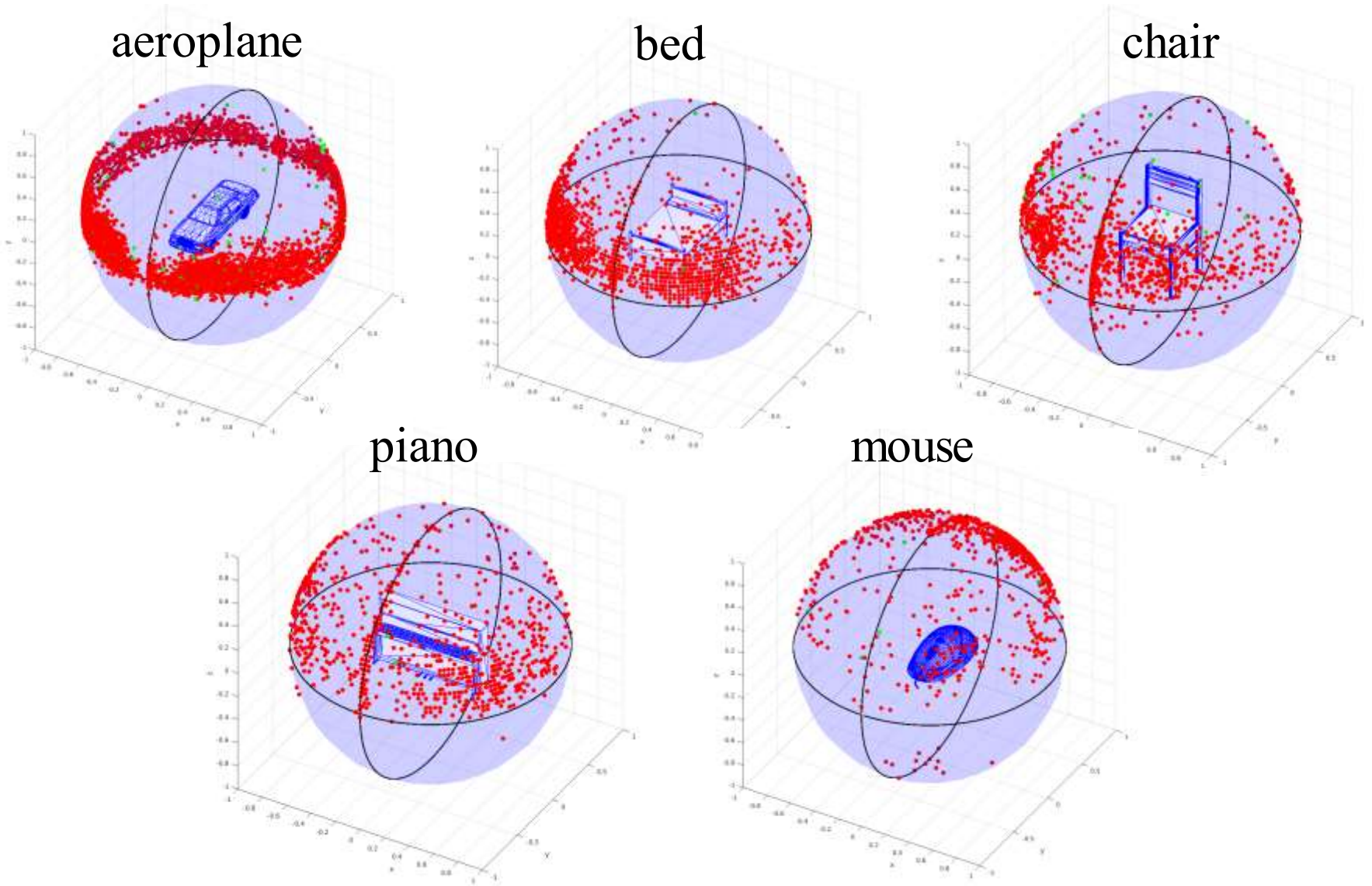


3D Pose Annotation Examples





Viewpoint Distributions



Database Construction: Image-based 3D Shape Retrieval

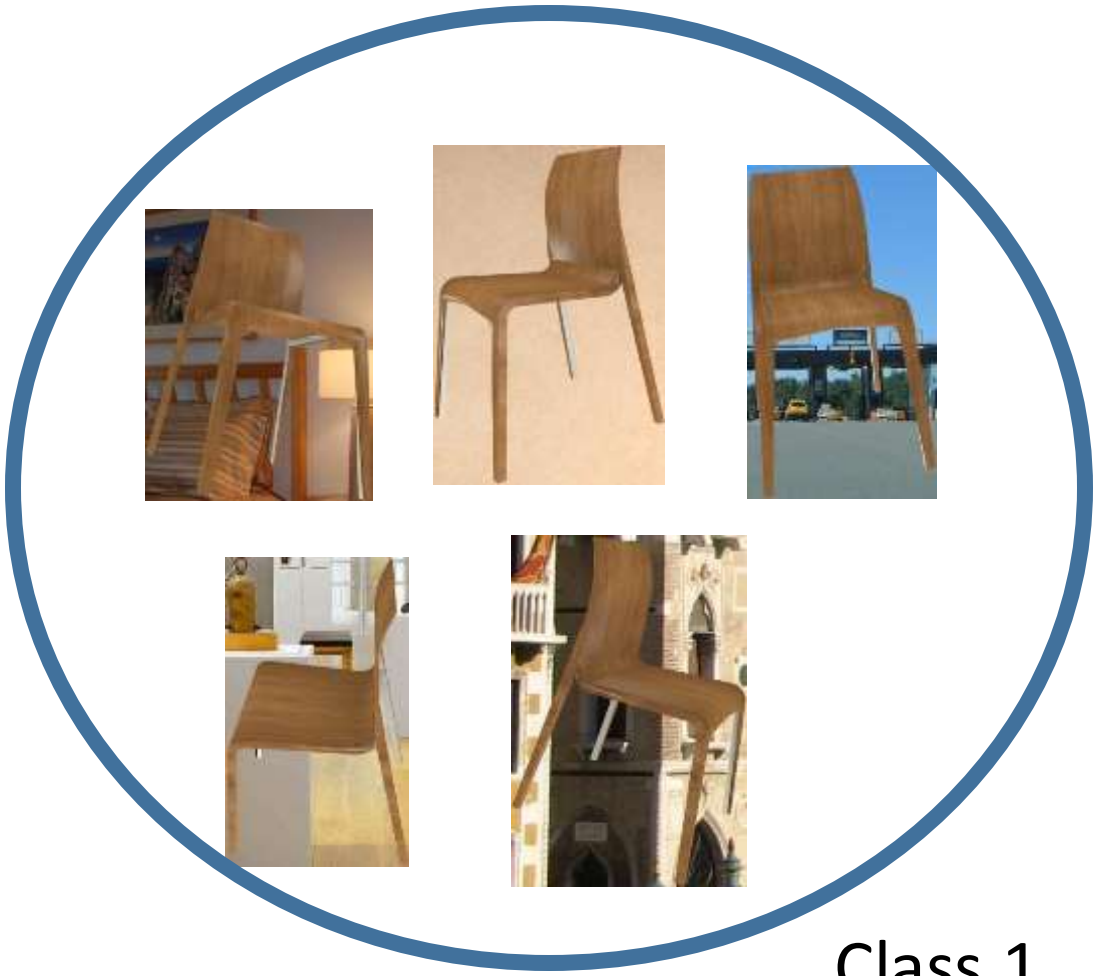
Database Construction: Image-based 3D Shape Retrieval

Database Construction: Image-based 3D Shape Retrieval

Database Construction: Image-based 3D Shape Retrieval

Database Construction: Image-based 3D Shape Retrieval

Metric Learning for Image-based 3D Shape Retrieval



Class 1

Class 2

Deep Metric Learning for Image-based 3D Shape Retrieval

CNN

CNN

Loss function

- Contrastive
- Triplet
- LiftedStruct [1]

[1] H.O. Song, Y. Xiang, S. Jegelka and S. Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. In CVPR, 2016.

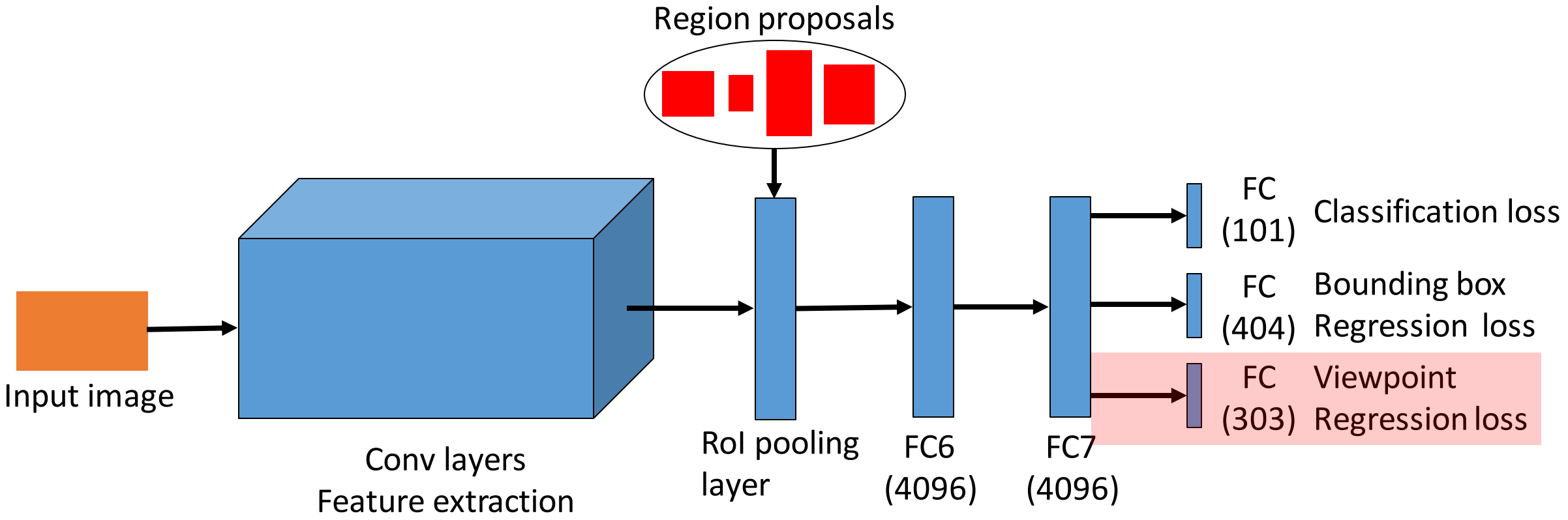
Database Construction: Image-based 3D Shape Retrieval

Test Object

Database Construction: Image-based 3D Shape Retrieval

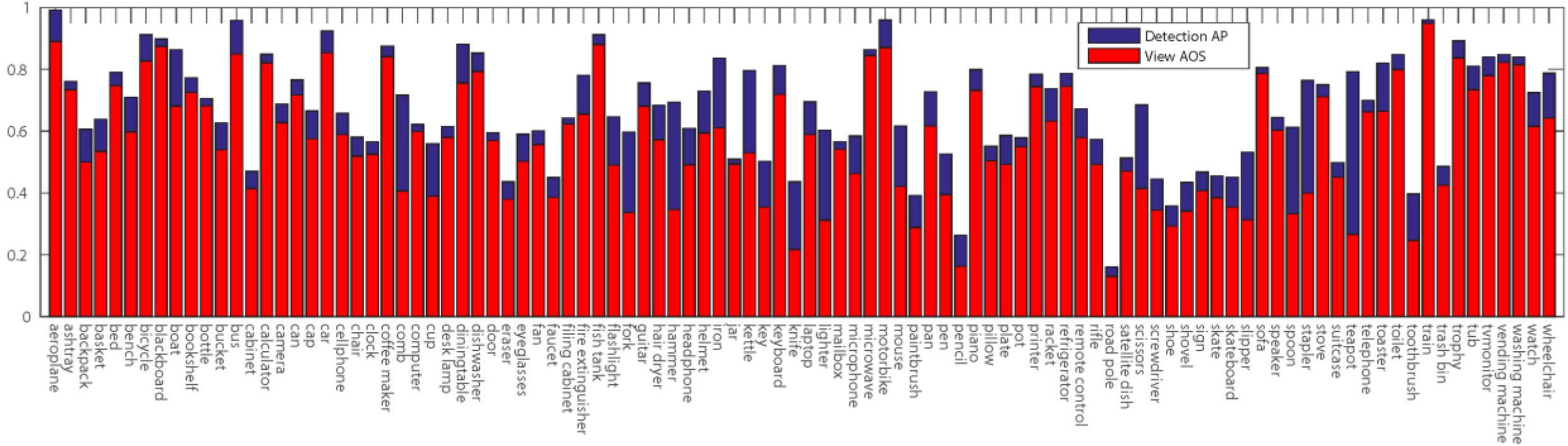
Database Construction: Image-based 3D Shape Retrieval

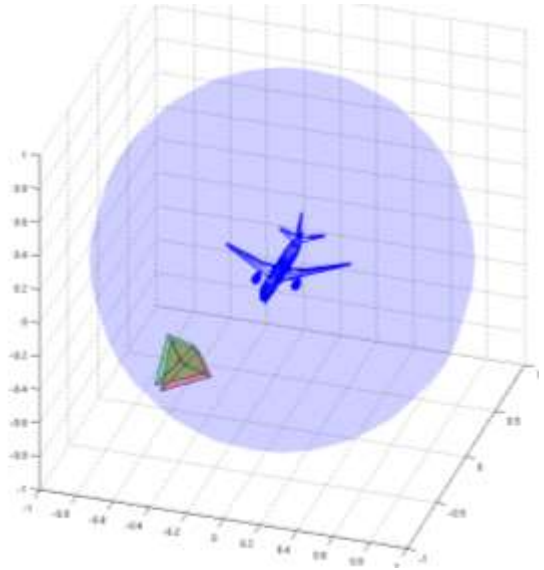
Joint Object Detection and Pose Estimation

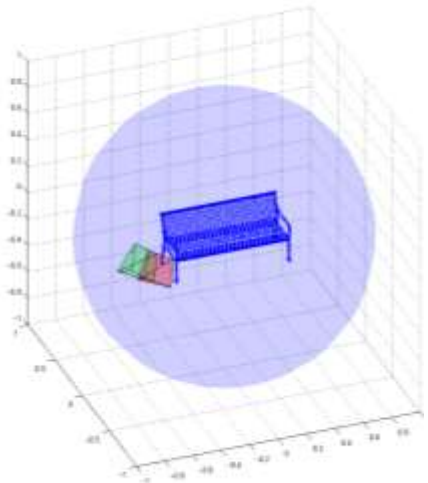


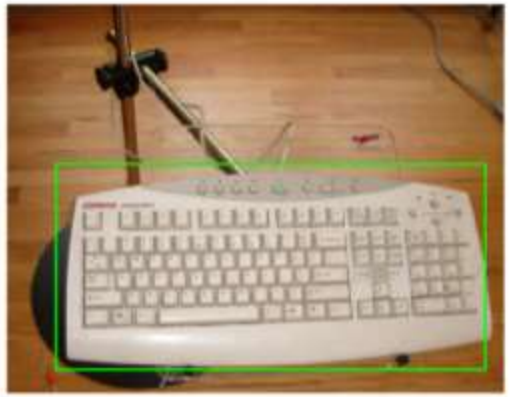
R. Girshick. Fast R-CNN. ICCV'15.

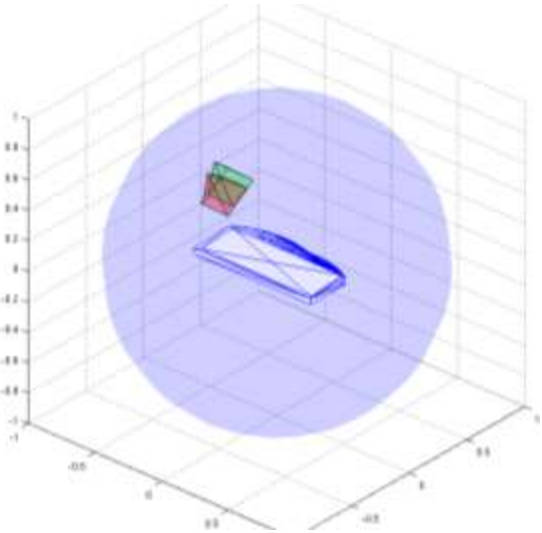
Joint Object Detection and Pose Estimation

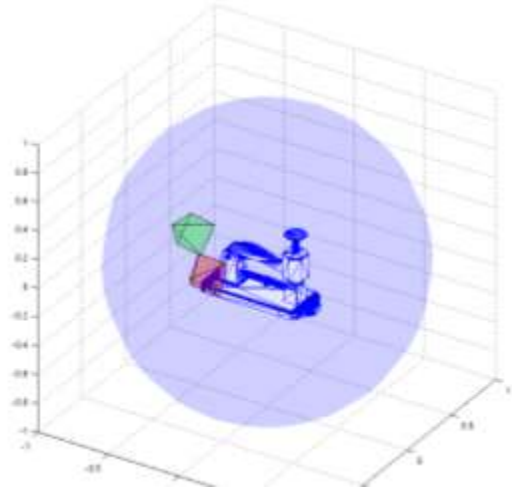




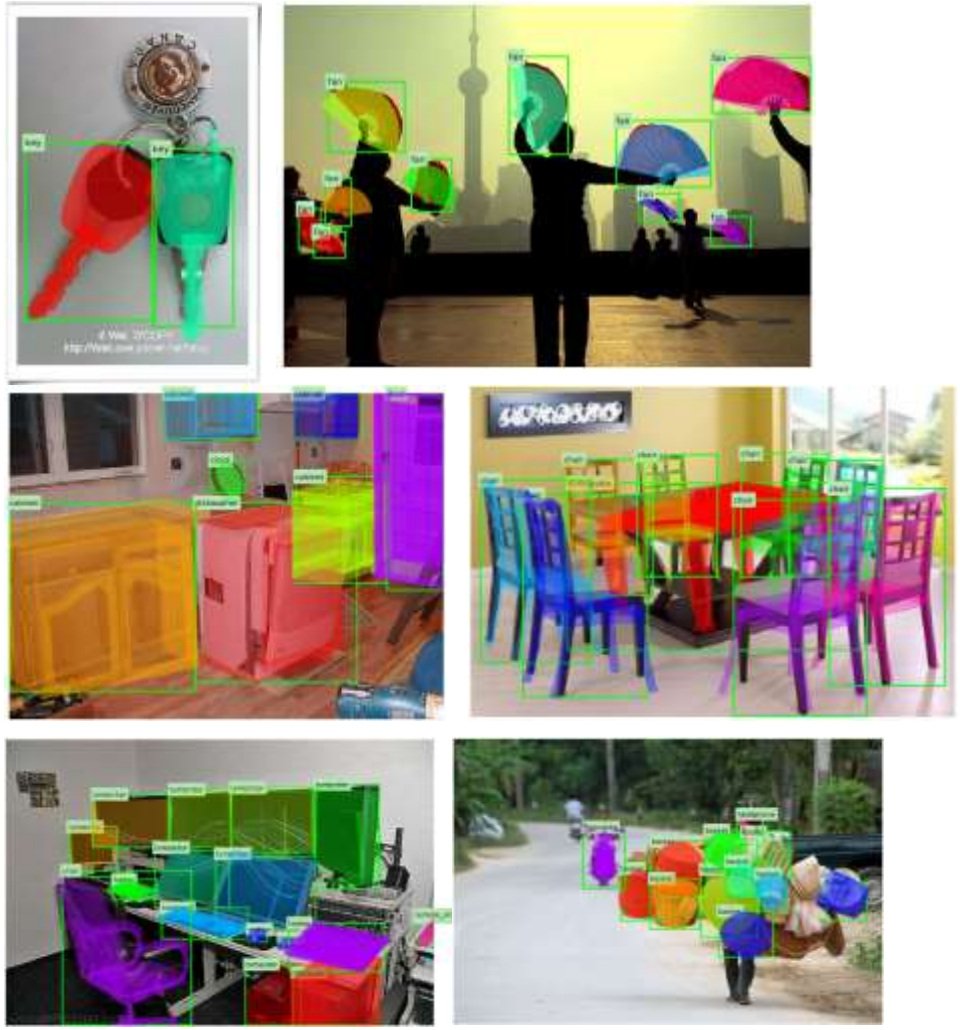








ObjectNet3D



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

Outline

- ObjectNet3D: A Large Scale Database for 3D Object Recognition

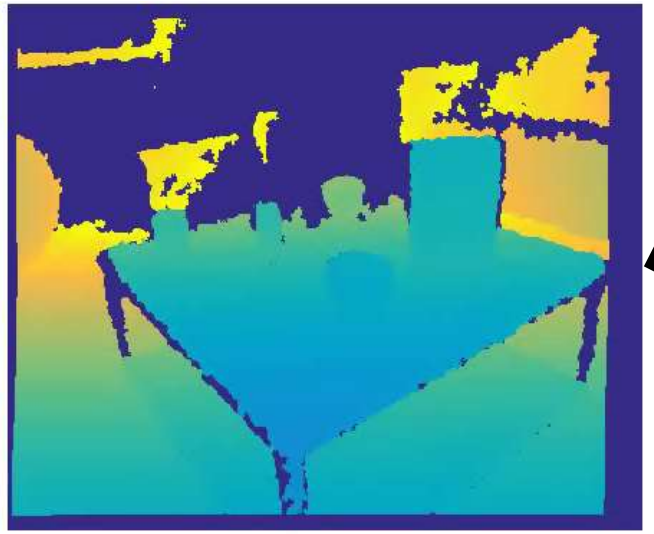
- DA-RNN: Semantic Mapping with Data Associated Recurrent Neural Networks

3D Scene Understanding

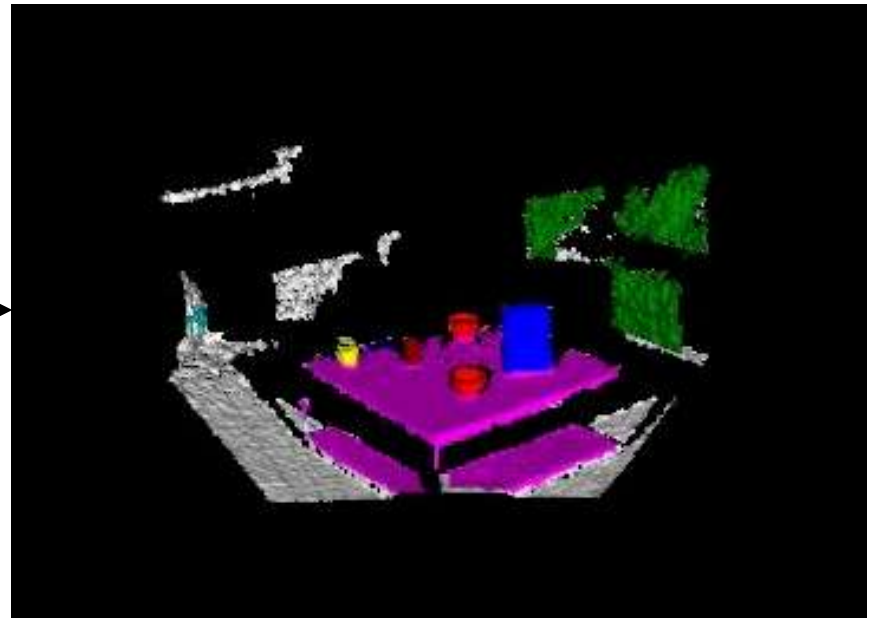
- Navigation
- Manipulation
- …

- Geometry
  - ✓ Free space
  - ✓ Surface
- Semantics
  - ✓ Objects
  - ✓ Affordances

Semantic Mapping with Data Associated Recurrent Neural Networks (DA-RNNs)

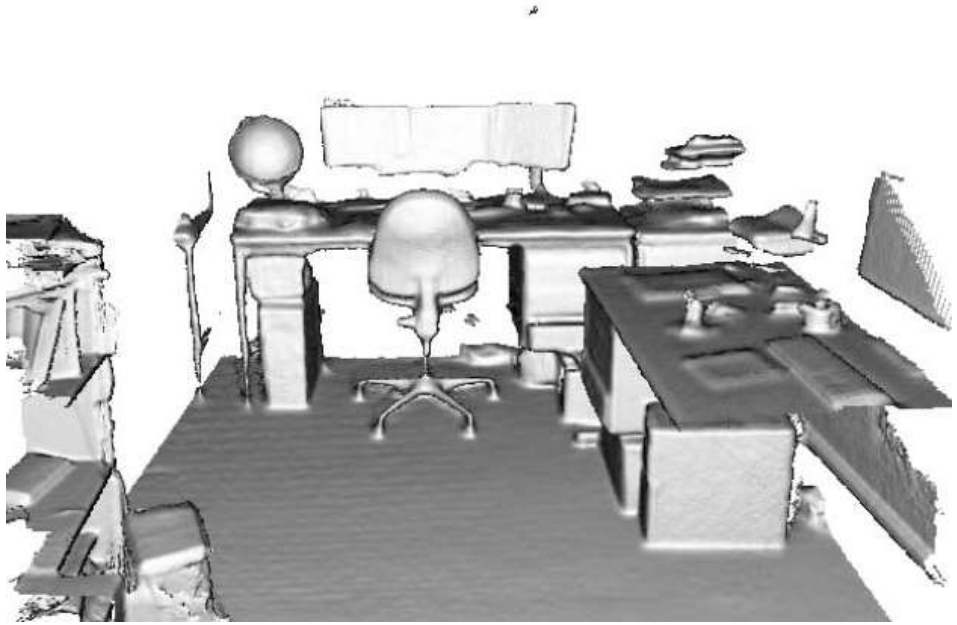


DA-RNN



Xiang & Fox. RSS'17

Related Work: 3D Scene Reconstruction



KinectFusion

- ✓ Geometry
- ✓ Data Association
- ✗ Semantics

- Newcombe et al., ISMAR'11
- Henry et al., IJRR'12, 3DV'13
- Whelan et al., RSS Workshop'12, RSS'15
- Keller et al., 3DV'13

Related Work: Semantic Labeling

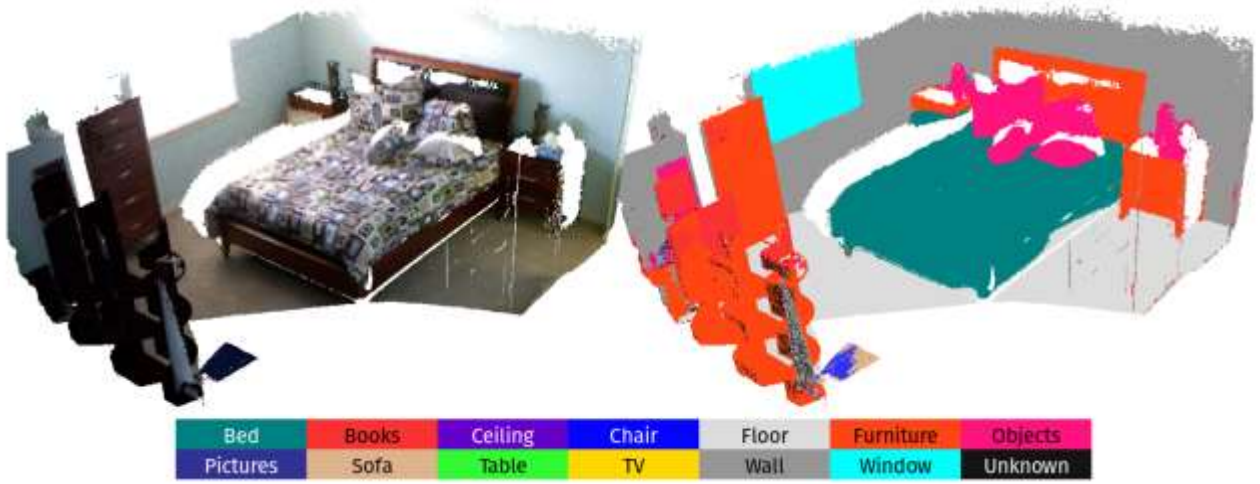
✘ Geometry

✘ Data Association

✓ Semantics

- Long et al., CVPR'12
- Zheng et al., ICCV'15
- Chen et al., ICLR'15
- Badrinarayanan et al., CVPR'15

Related Work: Semantic Mapping

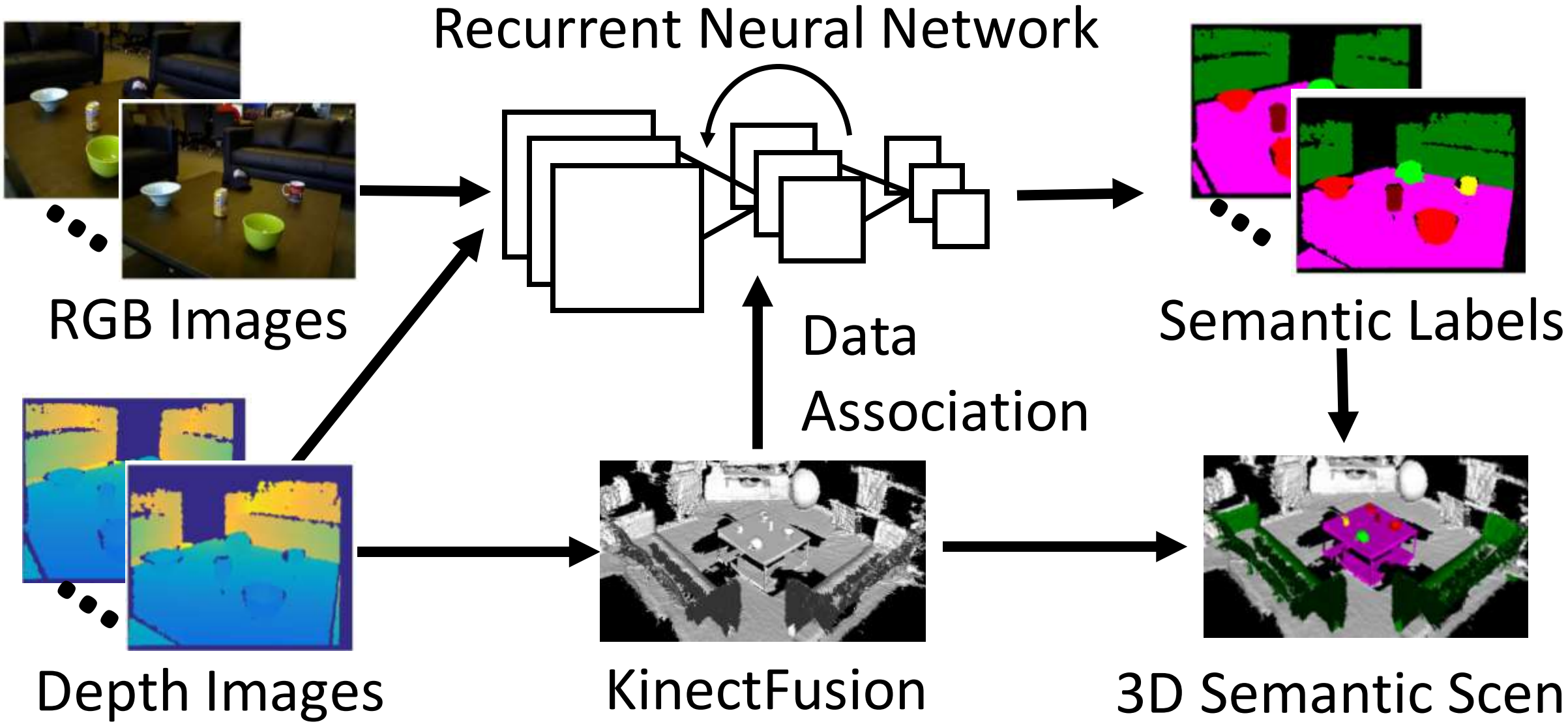


SemanticFusion

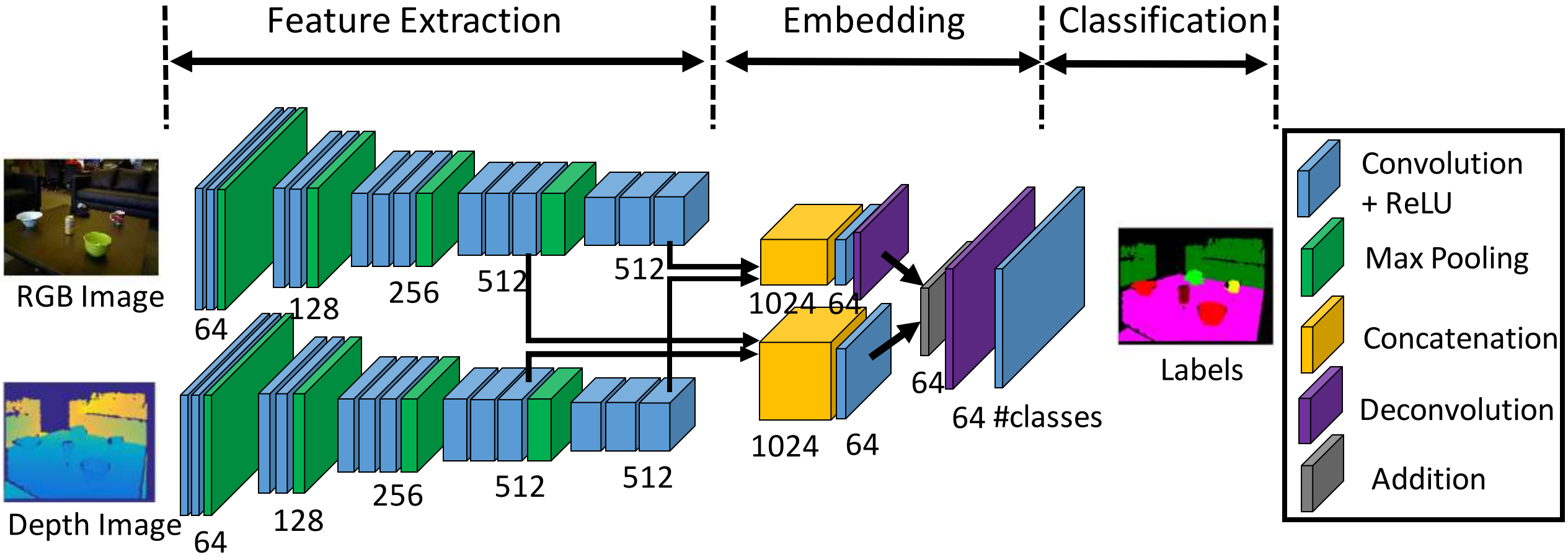
- ✓ Geometry
- ✓ Data Association
- ✓ Semantics

- Salas-Moreno et al., CVPR'13
- McCormac et al., ICRA'17

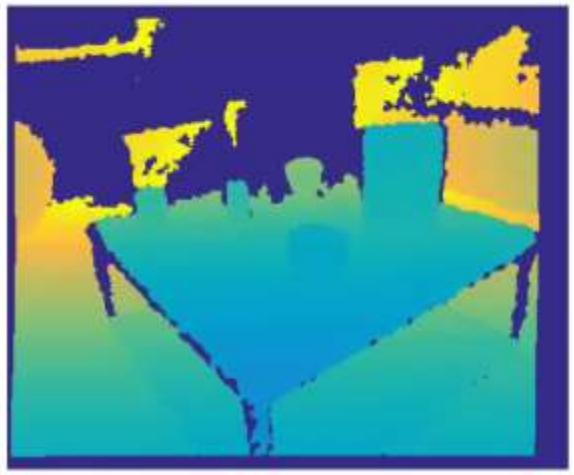
Our Contribution: DA-RNN

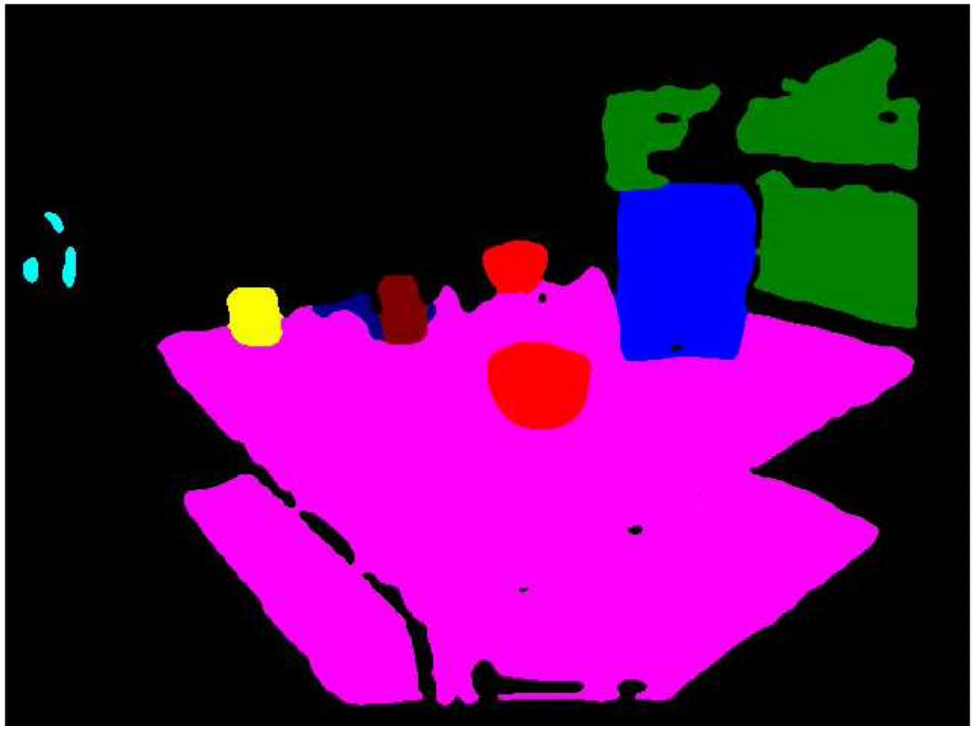


Single Frame Labeling with FCNs



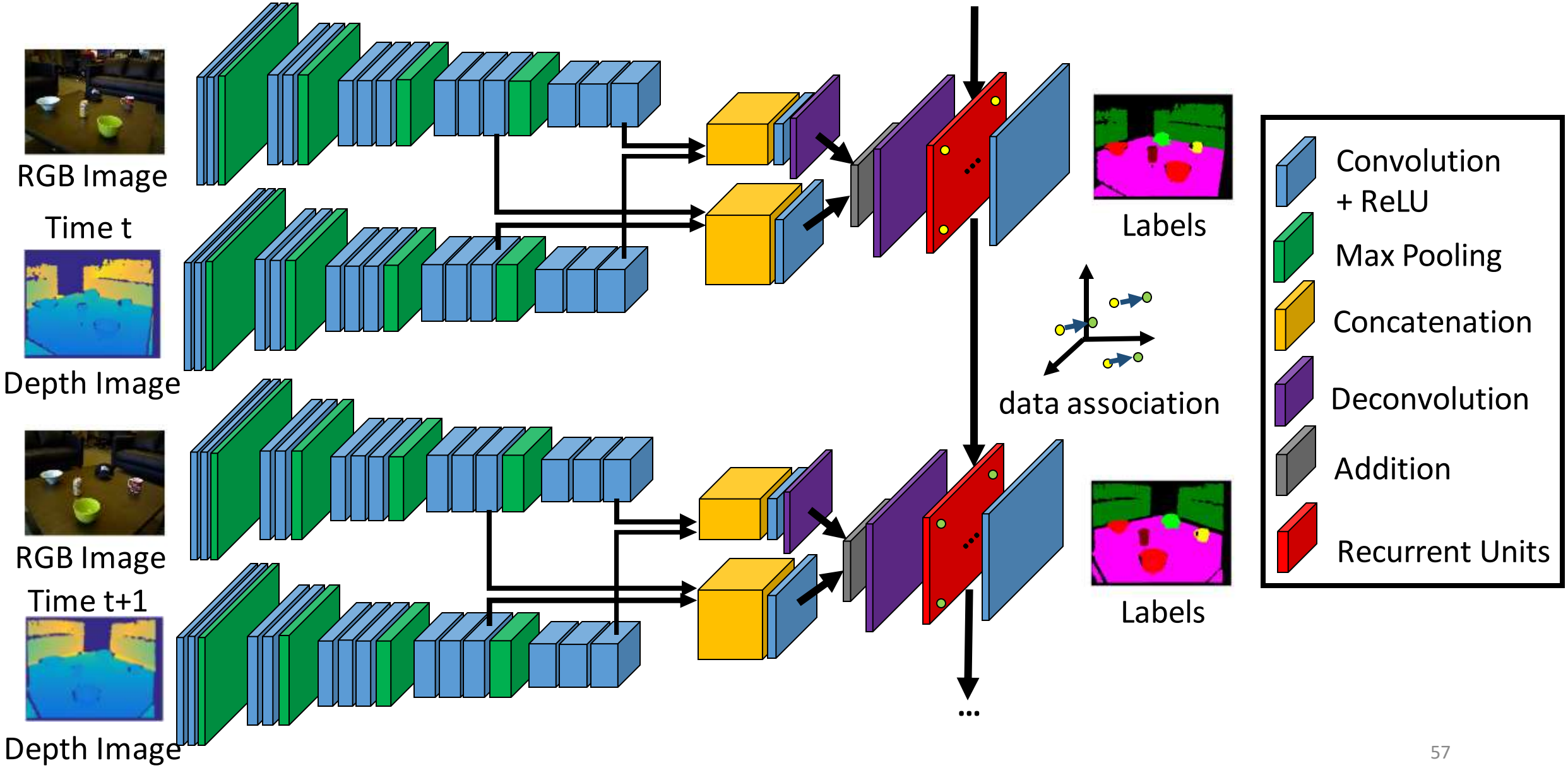
Results on RGB-D Scene Dataset [1]



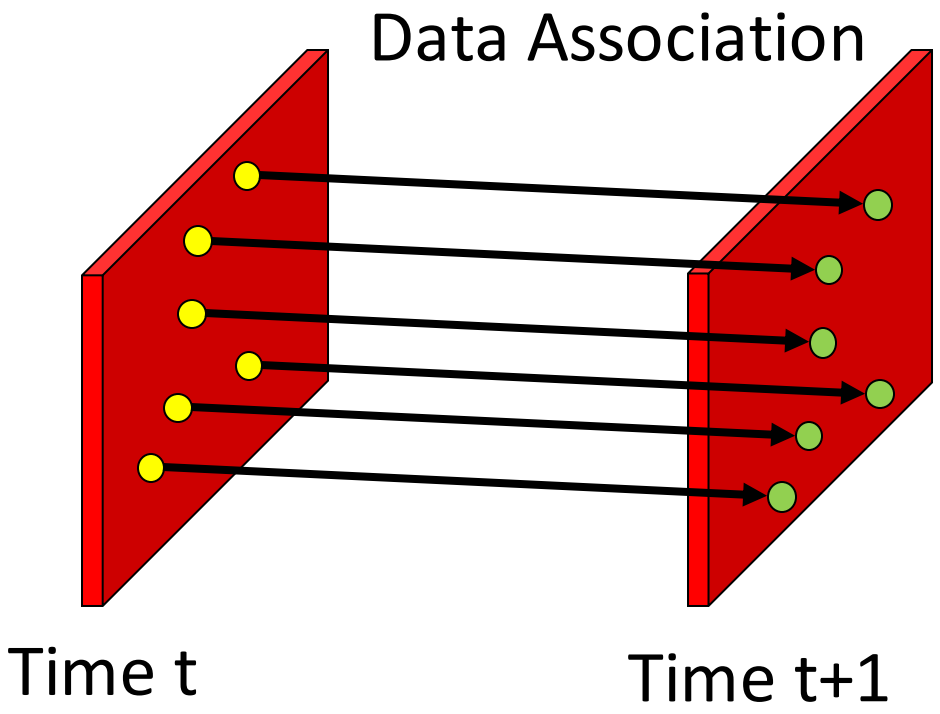


[1] K. Lai, L. Bo and D. Fox. Unsupervised feature learning for 3D scene labeling. In ICRA'14

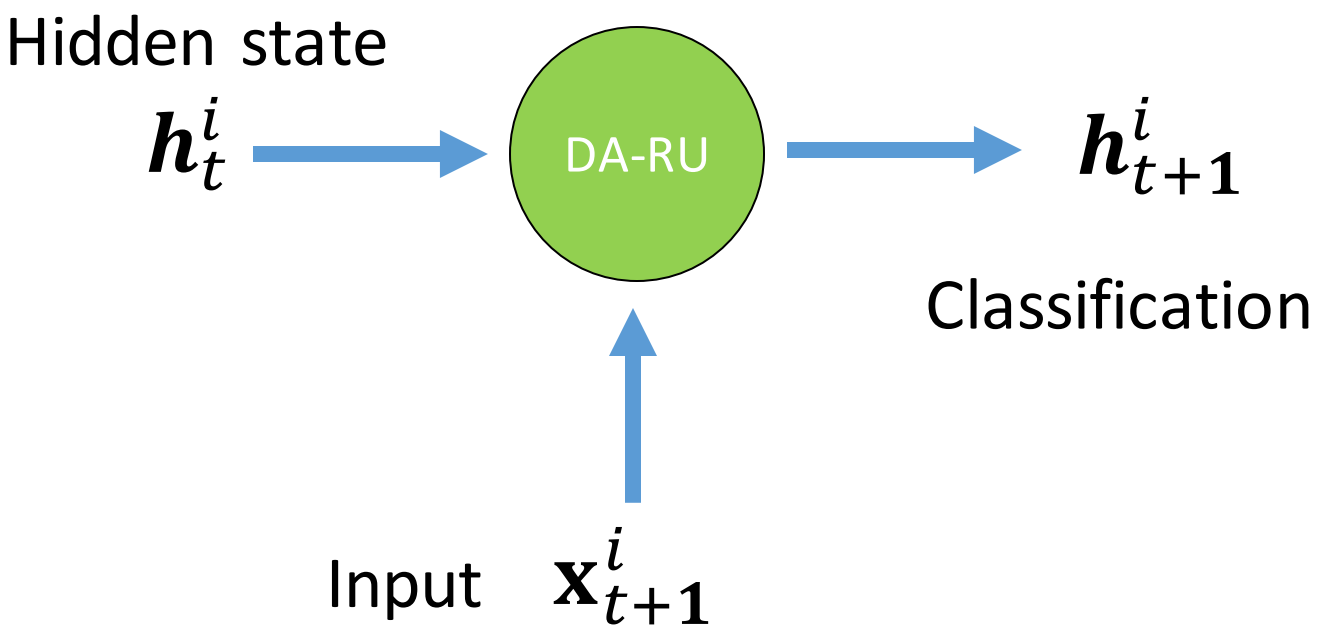
Video Semantic Labeling with DA-RNNs



Data Associated Recurrent Units (DA-RUs)

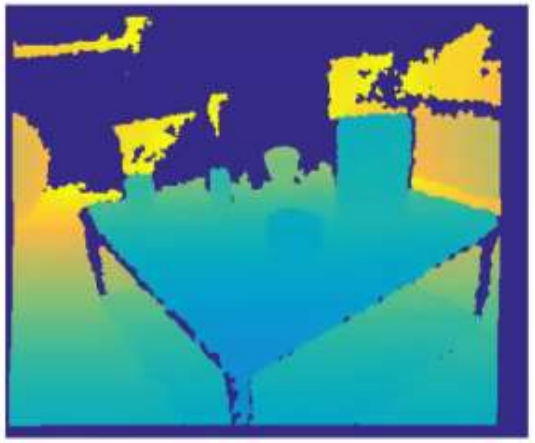


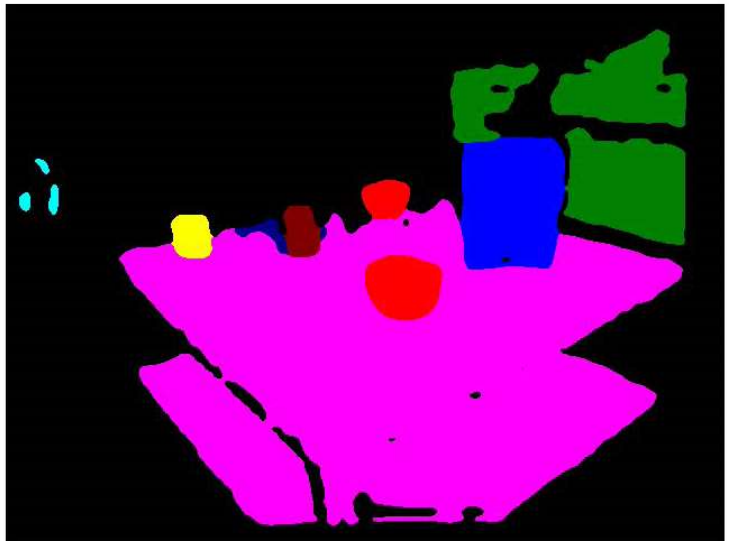
Recurrent layer

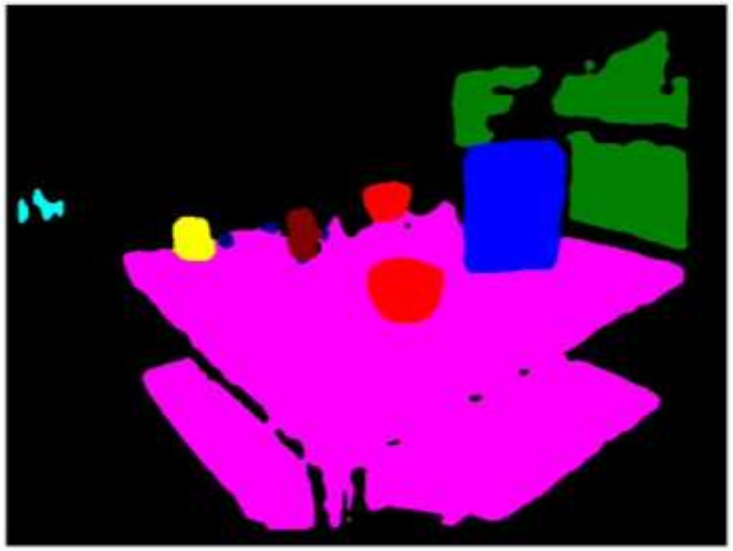


Weighted Moving Averaging
with learnable parameters

Results on RGB-D Scene Dataset [1]







FCN

DA-RNN

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

Experiments: Datasets

- RGB-D Scene Dataset [1]
  - 14 RGB-D videos of indoor scenes
  - 9 object classes

- ShapeNet Scene Dataset [2]
  - 100 RGB-D videos of virtual table-top scenes
  - 7 object classes

[1] K. Lai, L. Bo and D. Fox. Unsupervised feature learning for 3D scene labeling. In ICRA'14.
[2] Chang et al., ShapeNet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012, 2015.

Experiments: Comparison on Network Architectures

RGB-D Scenes

Methods	FCN [1]
Background	94.3
Bowl	78.6
Cap	61.2
Cereal Box	80.4
Coffee Mug	62.7
Coffee Table	93.6
Office Chair	67.3
Soda Can	73.5
Sofa	90.8
Table	84.2
MEAN	78.7

Metric: segmentation intersection over union (IoU)

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.

Experiments: Comparison on Network Architectures

RGB-D Scenes

Methods	FCN [1]	Our FCN
Background	94.3	**96.1**
Bowl	78.6	**87.0**
Cap	61.2	**79.0**
Cereal Box	80.4	**87.5**
Coffee Mug	62.7	**75.7**
Coffee Table	93.6	**95.2**
Office Chair	67.3	**71.6**
Soda Can	73.5	**82.9**
Sofa	90.8	**92.9**
Table	84.2	**89.8**
MEAN	78.7	**85.8**

Metric: segmentation intersection over union (IoU)

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.

Experiments: Comparison on Network Architectures

RGB-D Scenes

Methods	FCN [1]	Our FCN	Our GRU-RNN
Background	94.3	96.1	**96.8**
Bowl	78.6	**87.0**	86.4
Cap	61.2	79.0	**82.0**
Cereal Box	80.4	**87.5**	**87.5**
Coffee Mug	62.7	75.7	**76.1**
Coffee Table	93.6	95.2	**96.0**
Office Chair	67.3	71.6	**72.7**
Soda Can	73.5	**82.9**	81.9
Sofa	90.8	92.9	**93.5**
Table	84.2	89.8	**90.8**
MEAN	78.7	85.8	**86.4**

Metric: segmentation intersection over union (IoU)

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.

Experiments: Comparison on Network Architectures

RGB-D Scenes

Methods	FCN [1]	Our FCN	Our GRU-RNN	Our DA-RNN
Background	94.3	96.1	96.8	**97.6**
Bowl	78.6	87.0	86.4	**92.7**
Cap	61.2	79.0	82.0	**84.4**
Cereal Box	80.4	87.5	87.5	**88.3**
Coffee Mug	62.7	75.7	76.1	**86.3**
Coffee Table	93.6	95.2	96.0	**97.3**
Office Chair	67.3	71.6	72.7	**77.0**
Soda Can	73.5	82.9	81.9	**88.7**
Sofa	90.8	92.9	93.5	**95.6**
Table	84.2	89.8	90.8	**92.8**
MEAN	78.7	85.8	86.4	**90.1**

Metric: segmentation intersection over union (IoU)

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.

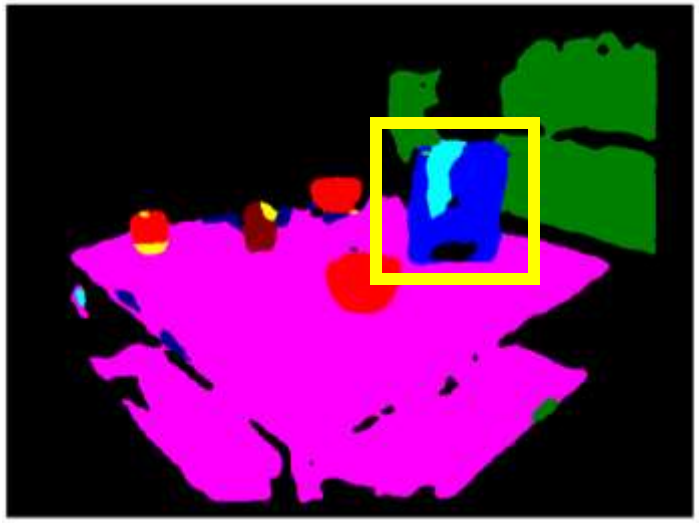
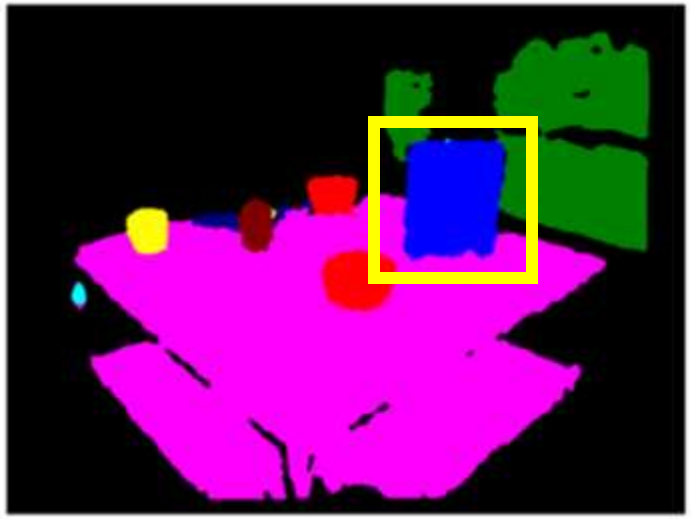
Experiments: Comparison on Network Architectures

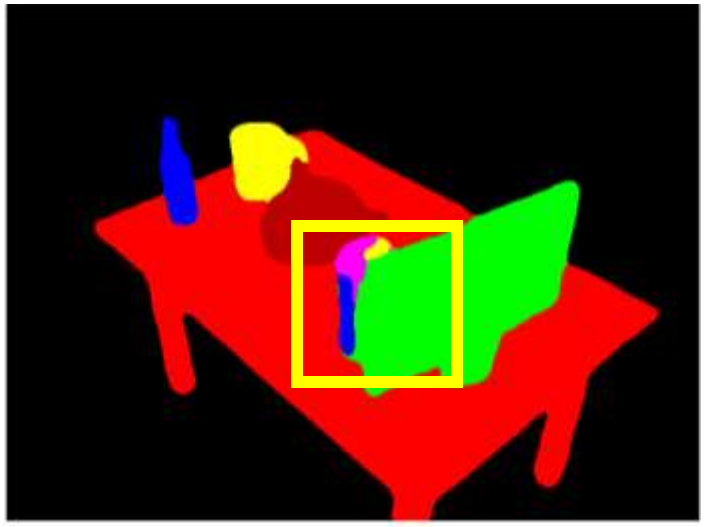
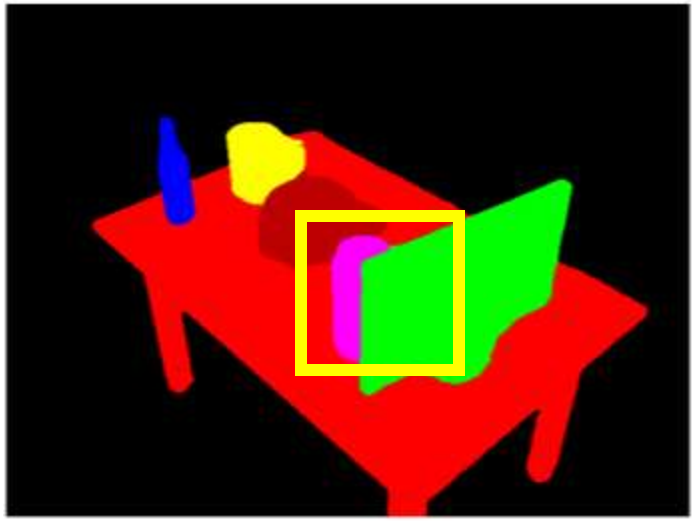
RGB-D Scenes

Methods	FCN [1]	Our FCN	Our GRU-RNN	Our DA-RNN	No Data Association
Background	94.3	96.1	96.8	**97.6**	69.1
Bowl	78.6	87.0	86.4	**92.7**	3.6
Cap	61.2	79.0	82.0	**84.4**	9.9
Cereal Box	80.4	87.5	87.5	**88.3**	14.0
Coffee Mug	62.7	75.7	76.1	**86.3**	4.5
Coffee Table	93.6	95.2	96.0	**97.3**	68.0
Office Chair	67.3	71.6	72.7	**77.0**	13.6
Soda Can	73.5	82.9	81.9	**88.7**	5.9
Sofa	90.8	92.9	93.5	**95.6**	35.6
Table	84.2	89.8	90.8	**92.8**	20.1
MEAN	78.7	85.8	86.4	**90.1**	24.4

Metric: segmentation intersection over union (IoU)

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR'15.

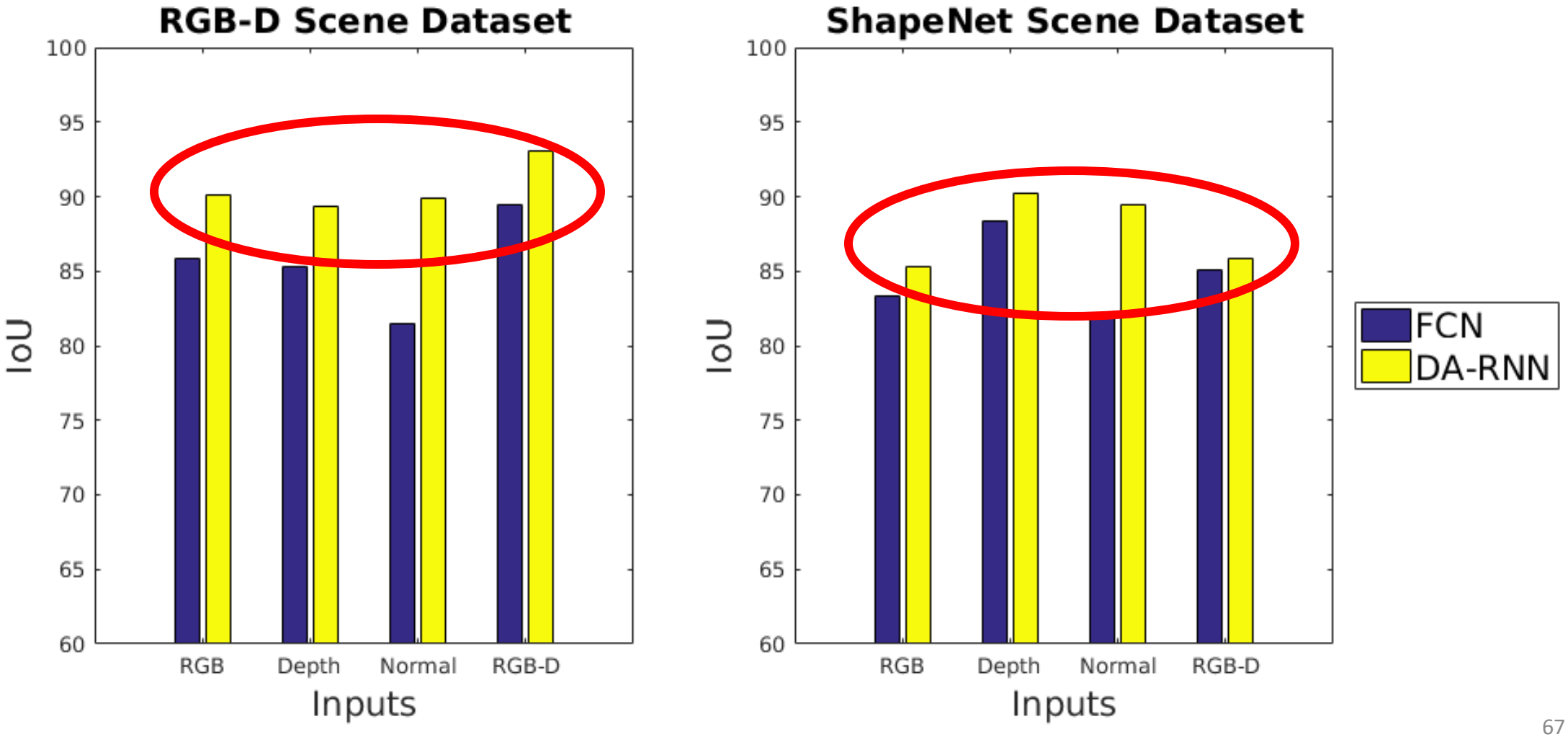



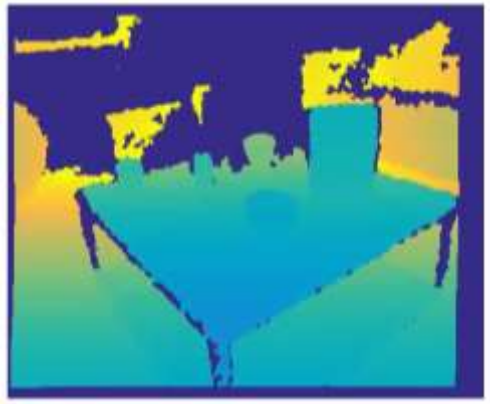
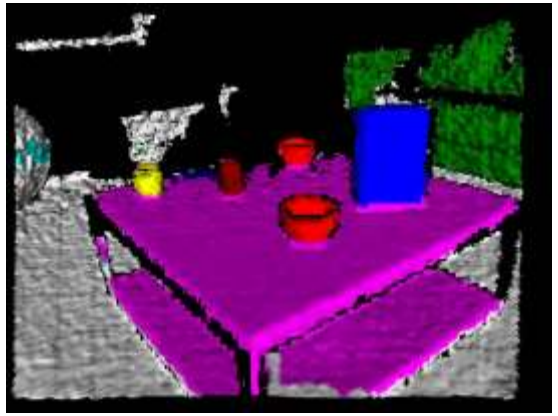
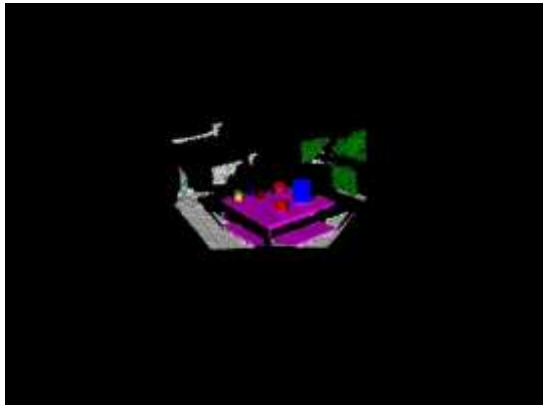
RGB Image

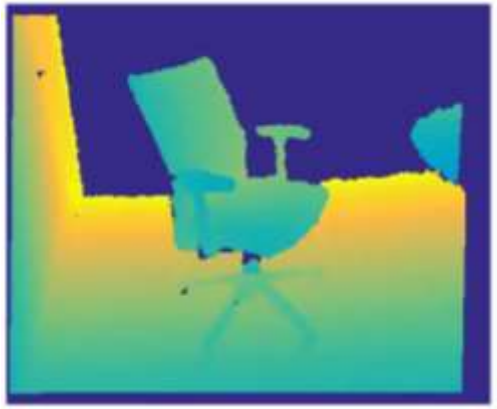
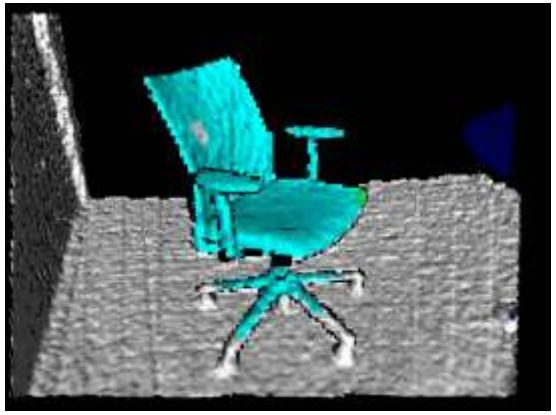
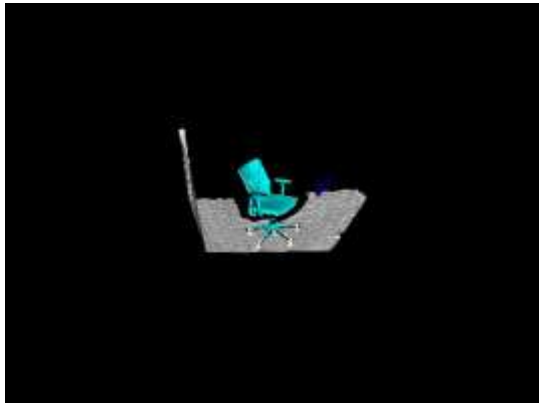
Our FCN

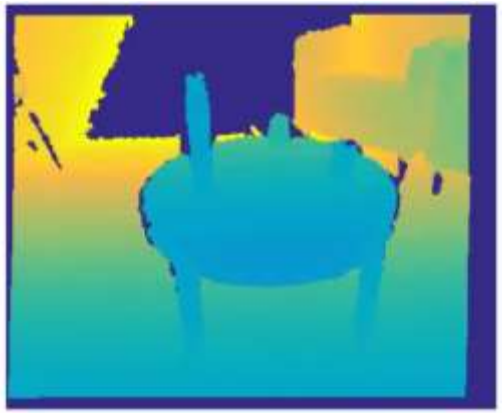
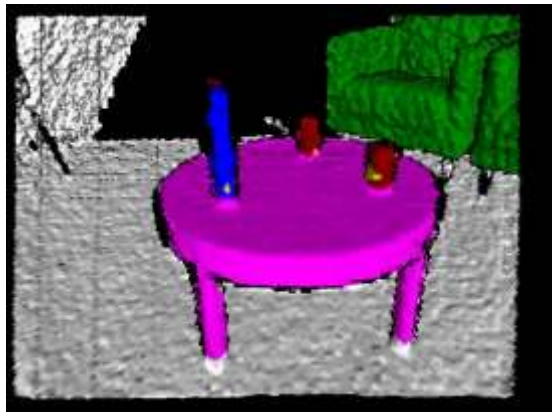
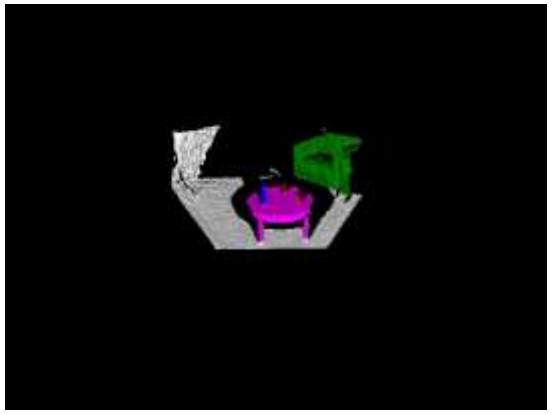
Our DA-RNN

Experiments: Analysis on Network Inputs



RGB Images

Depth Images

Semantic Mapping

Conclusion

- ObjectNet3D, a large scale dataset with 2D objects aligned with 3D shapes

- DA-RNN, A novel framework for joint 3D mapping and semantic labeling

Thank you!