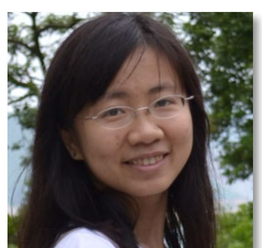
PRINCETON
VISION GROUP

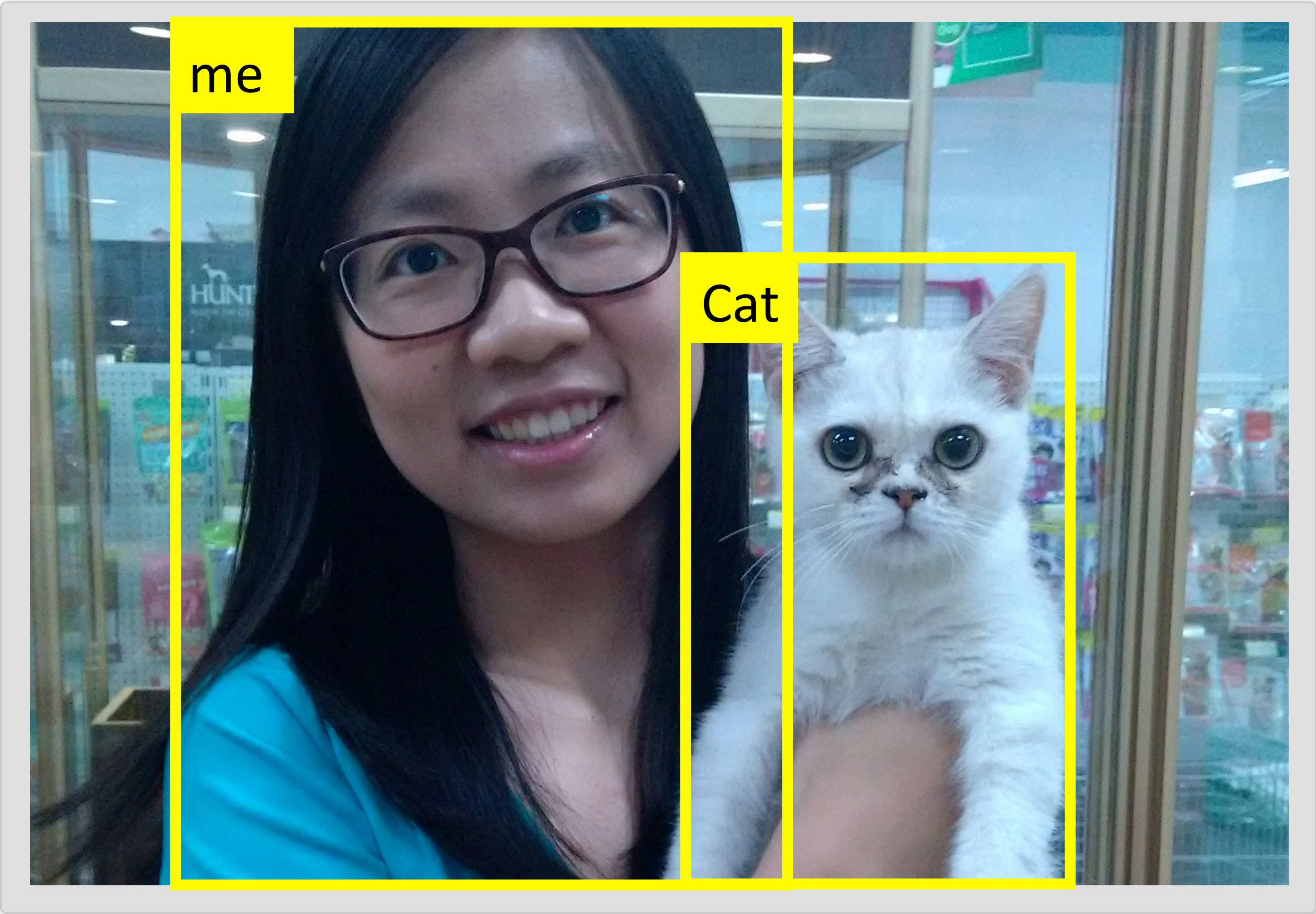
Sliding Shapes for 3D Object Detection in Depth Images




Shuran Song
Jianxiong Xiao

Object Detection

Object Detection



State-of-the-Art: Deep Learning

MIT Technology Review

10 BREAKTHROUGH TECHNOLOGIES 2013

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.

The New York Times

Scientists See Promise in Deep-Learning Programs

NYU Algorithm on NYU Dataset

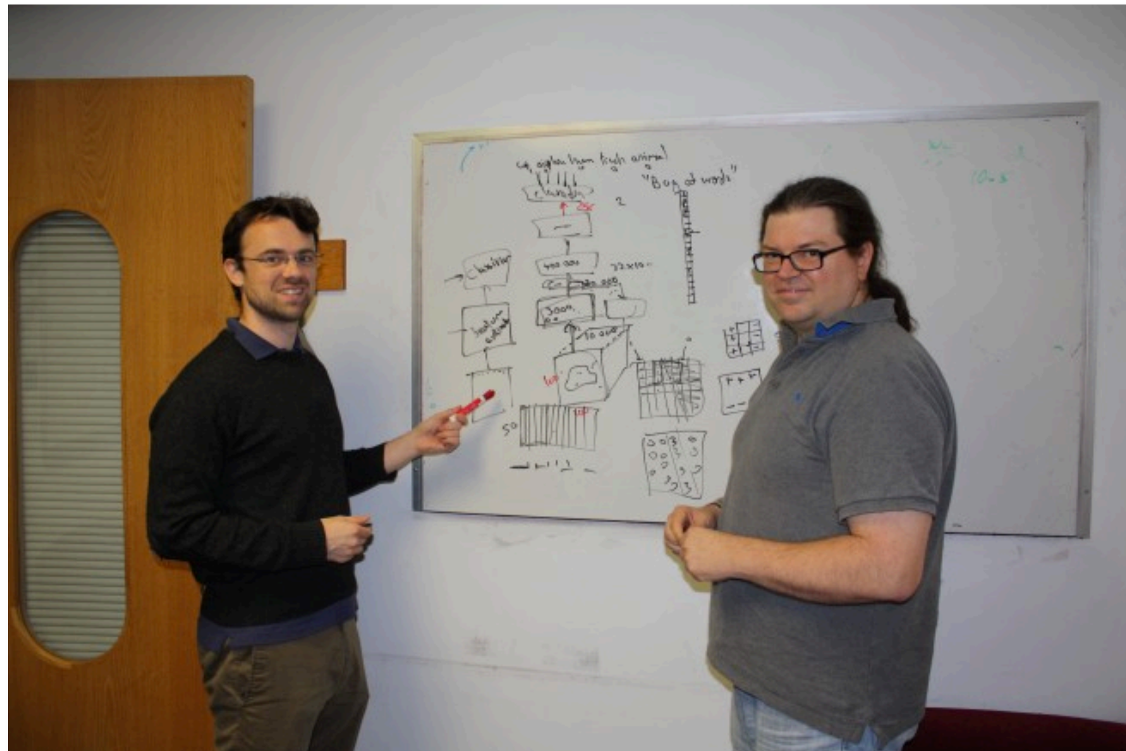
OverFeat:
Integrated Recognition, Localization and Detection using Convolutional Networks

Pierre Sermanet David Eigen
Xiang Zhang Michael Mathieu Rob Fergus Yann LeCun
Courant Institute of Mathematical Sciences, New York University
719 Broadway, 12th Floor, New York, NY 10003
sermanet,deigen,xiang,mathieu,fergus,yann@cs.nyu.edu

Indoor Segmentation and Support Inference from RGBD Images

Nathan Silberman[1], Derek Hoiem[2], Pushmeet Kohli[3], Rob Fergus[1]

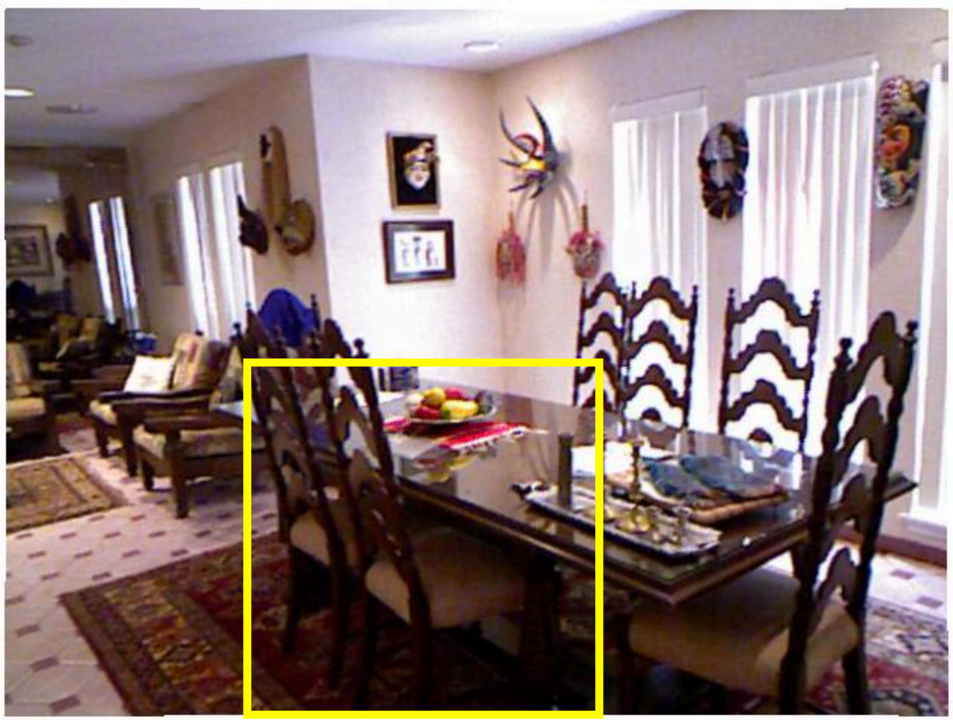
[1]Courant Institute, New York University
[2]Department of Computer Science, University of Illinois at Urbana-Champaign
[3]Microsoft Research, Cambridge



Rob Fergus
NYU

Yann LeCun
NYU

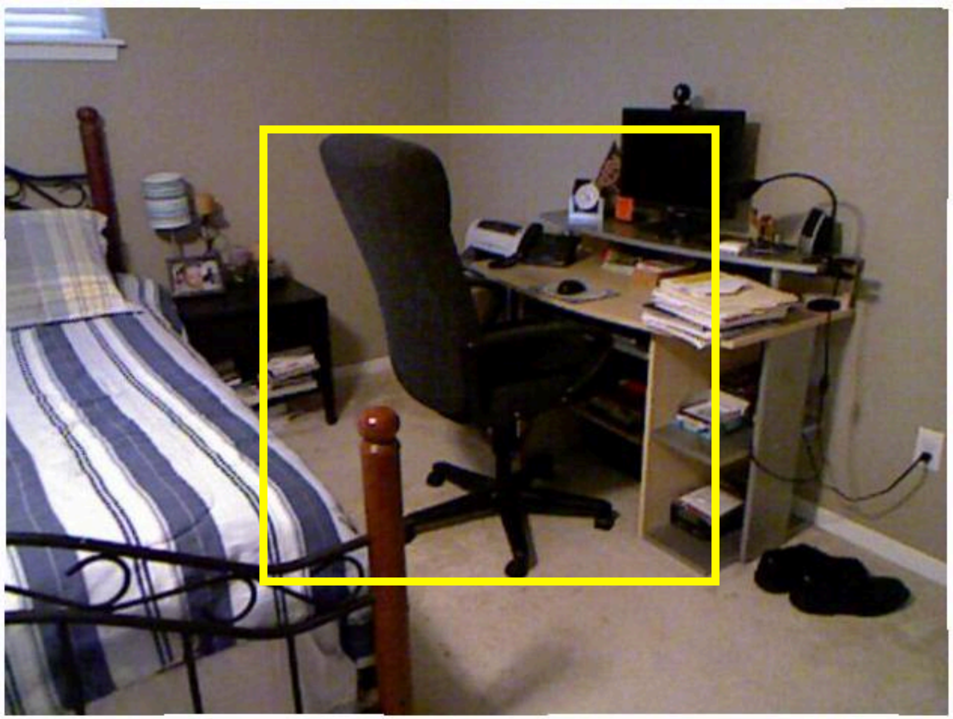
NYU Algorithm on NYU Dataset



NYU1418.jpg

5 most likely categories:
0.236223 shoe shop, shoe-shop, shoe store
0.027985 confectionery, confectionary
0.025233 cinema, movie theater
0.024637 butcher shop, meat market
0.024317 slot, one-armed bandit

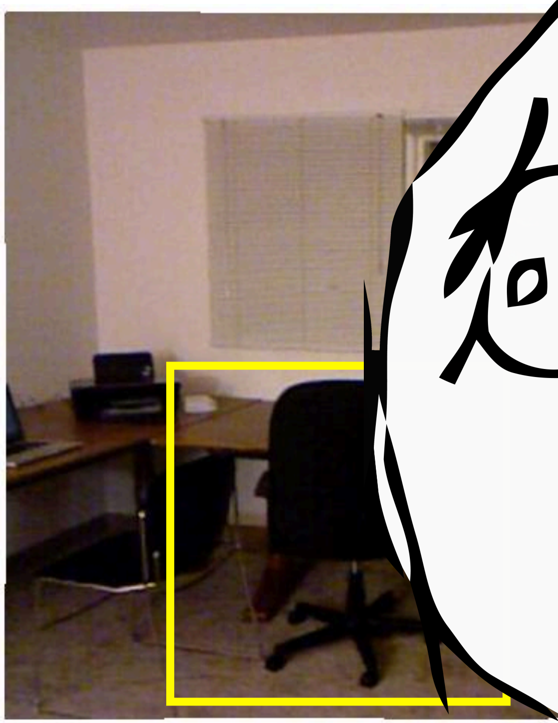
NYU Algorithm on NYU Dataset



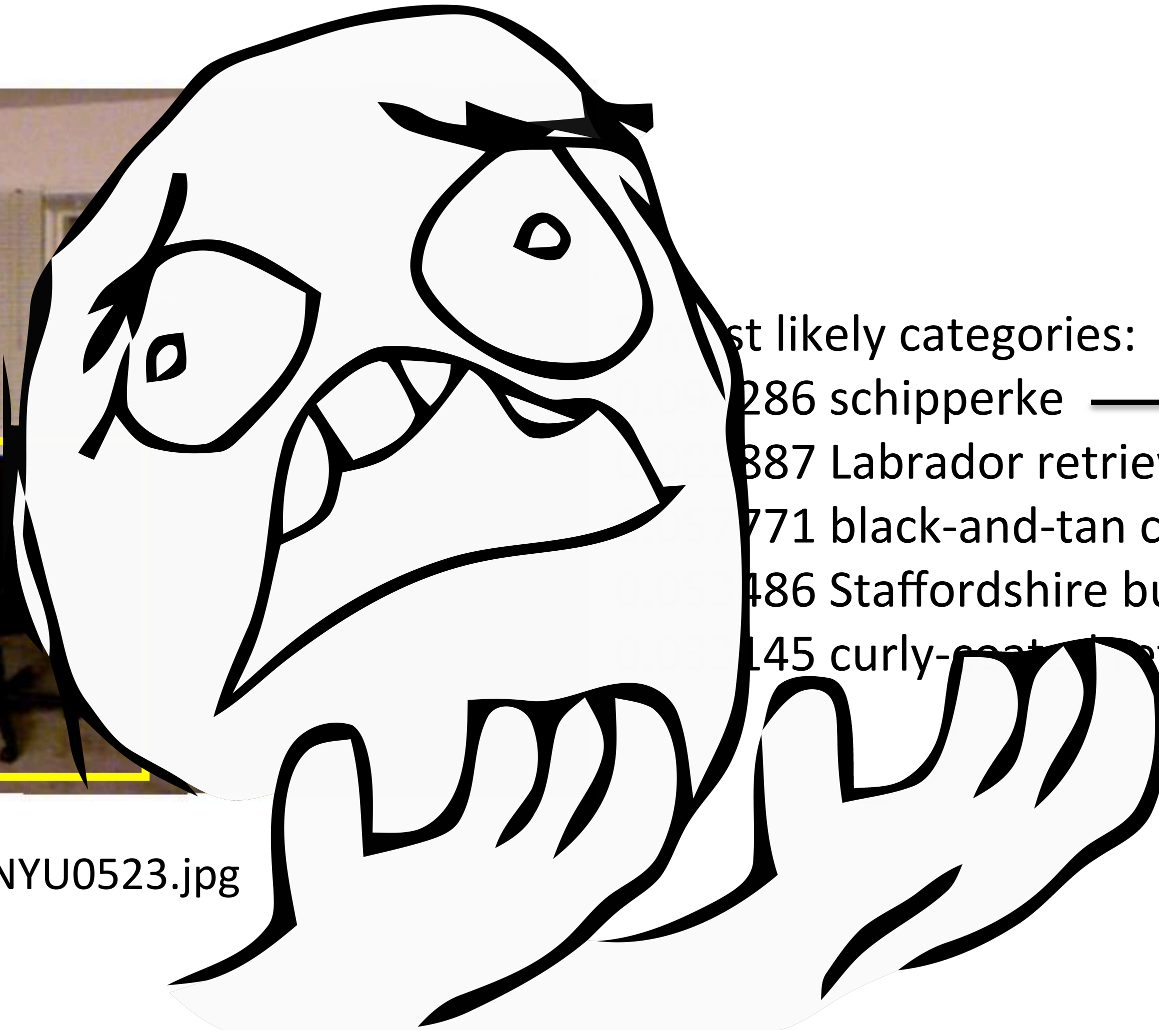
NYU0946.jpg

5 most likely categories:
0.181371 potter's wheel
0.175774 chiffonier, commode
0.126224 stove
0.060575 pedestal, plinth, footstall
0.044722 file, file cabinet, filing cabinet

NYU Algorithm on NYU Dataset



NYU0523.jpg



st likely categories:

286 schipperke

387 Labrador retriever

771 black-and-tan coonhound

486 Staffordshire bullterrier

145 curly- triever

Object Detection is Hard

Jia Deng's advisor's advisor's advisor's advisor

Nevatia & Binford, 1977.

IMAGENET

"These techniques are inadequate for three-dimensional scene analysis for many reasons:"

1. Variation

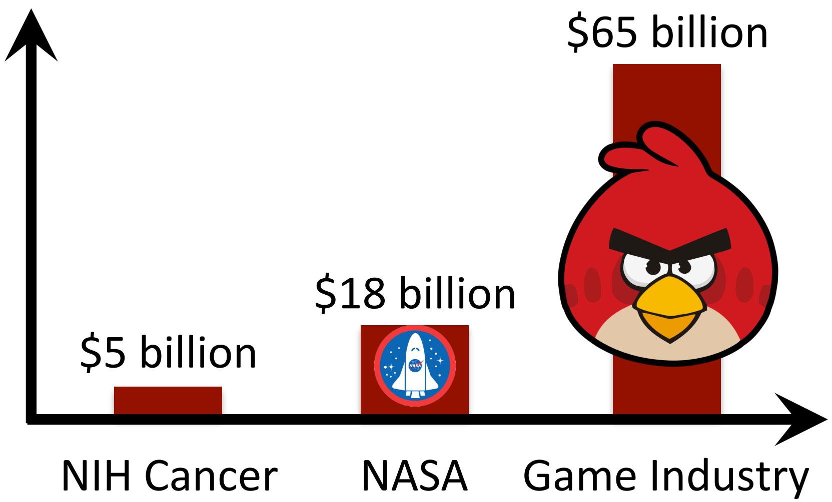
2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion

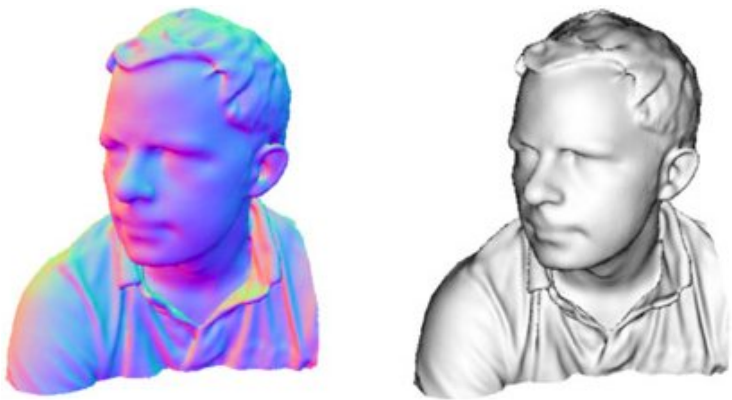
Maybe Depth Sensors can Help?



- Microsoft Kinect
- Intel RealSense
- Google Project Tango
- Apple Primesense
- Asus Xtion
- LEAP Motion
- Structure.io
- Stereo Cameras

Depth for Other Vision Tasks

Kinect Fusion



Newcombe et al. 2011

Intrinsic Image

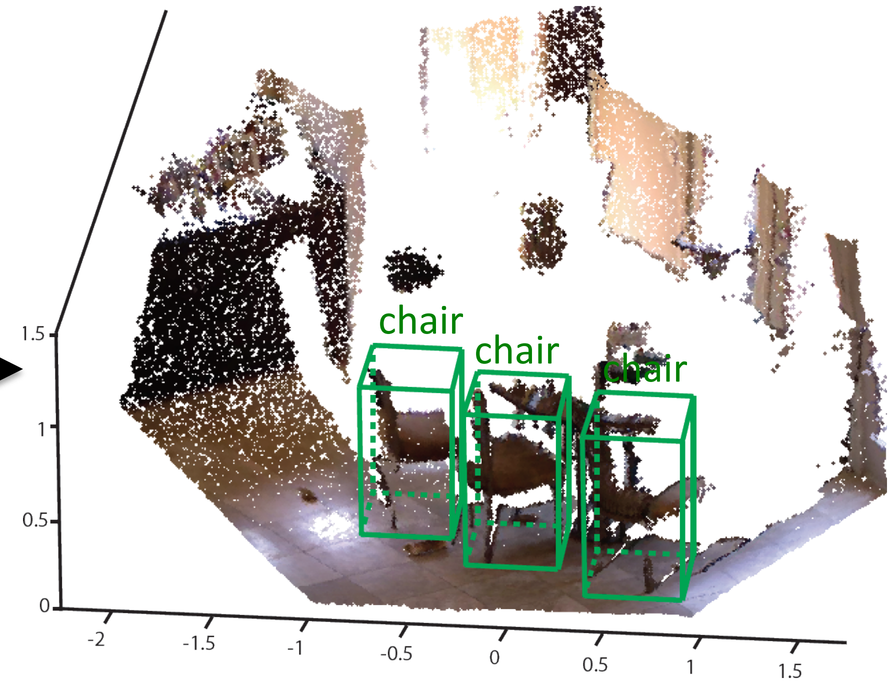
Barron & Malik, 2013

Human Pose Recognition

Shotton et al. 2011

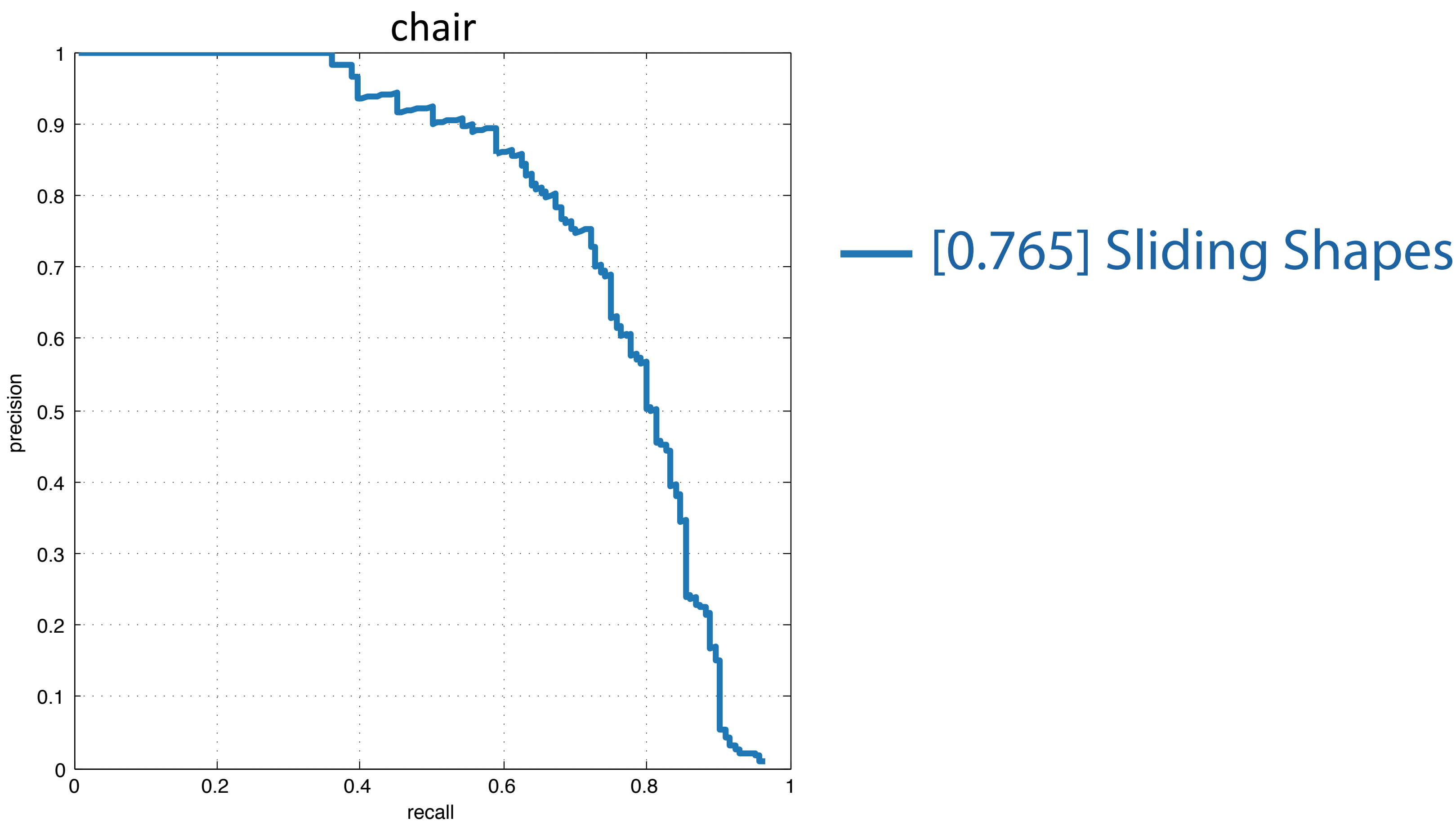
Sliding Shapes

Input: Kinect Depth Map

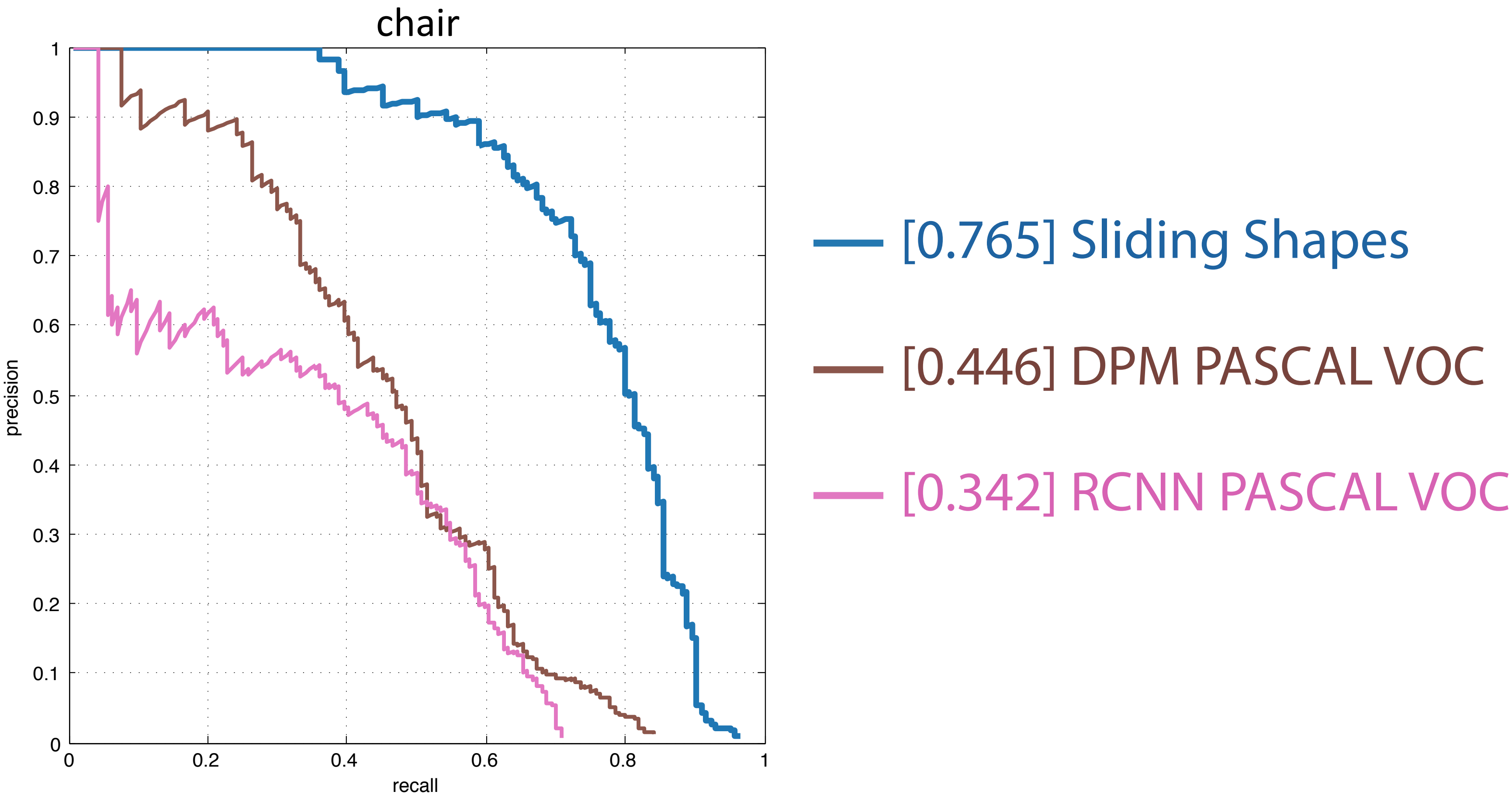


Output: 3D Bounding Box

Depth-based Object Detection



Depth-based Object Detection



×1.7 improvement on Average Precision compared to the best of DPM & R-CNN

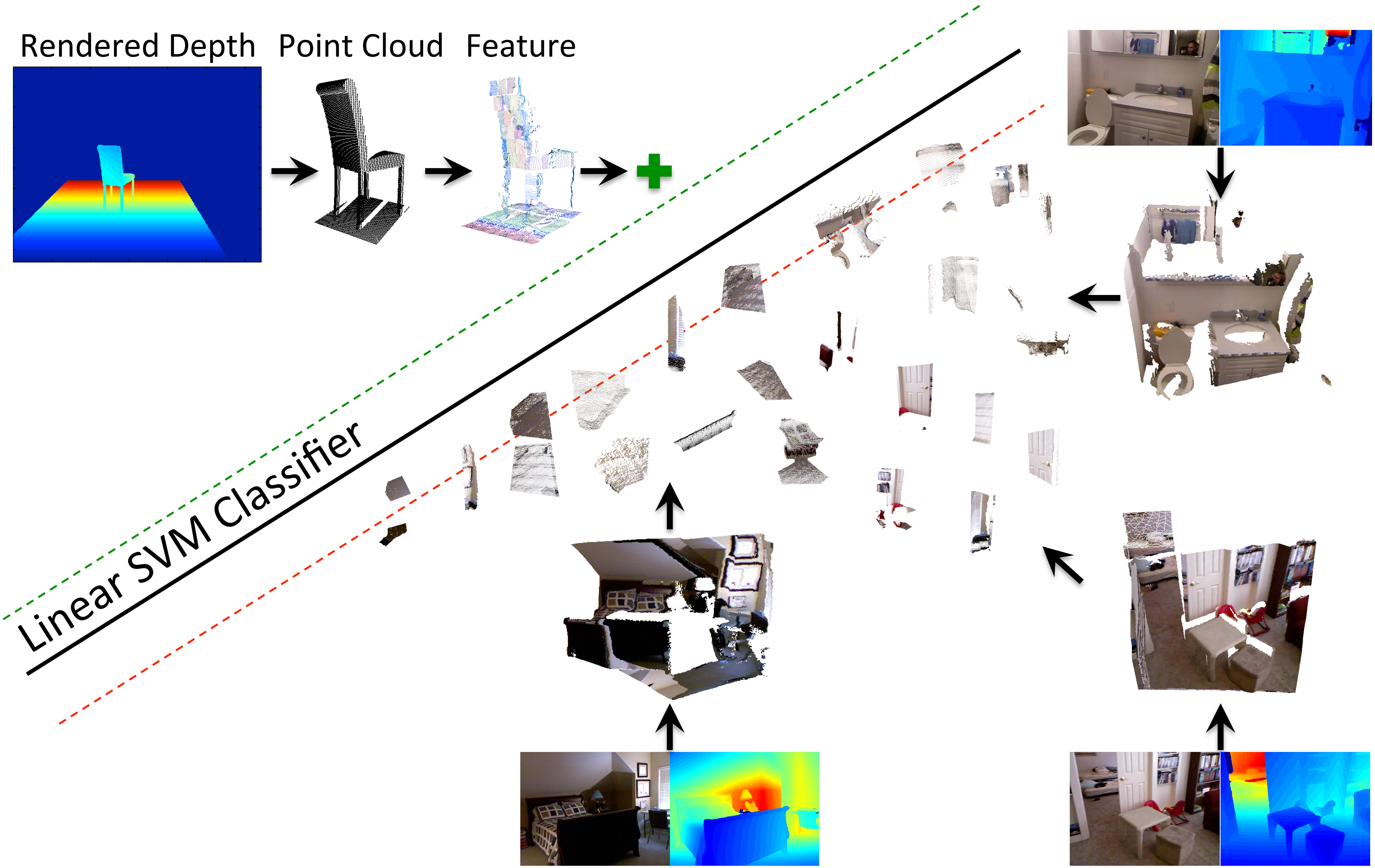
Algorithm

Training: CAD models

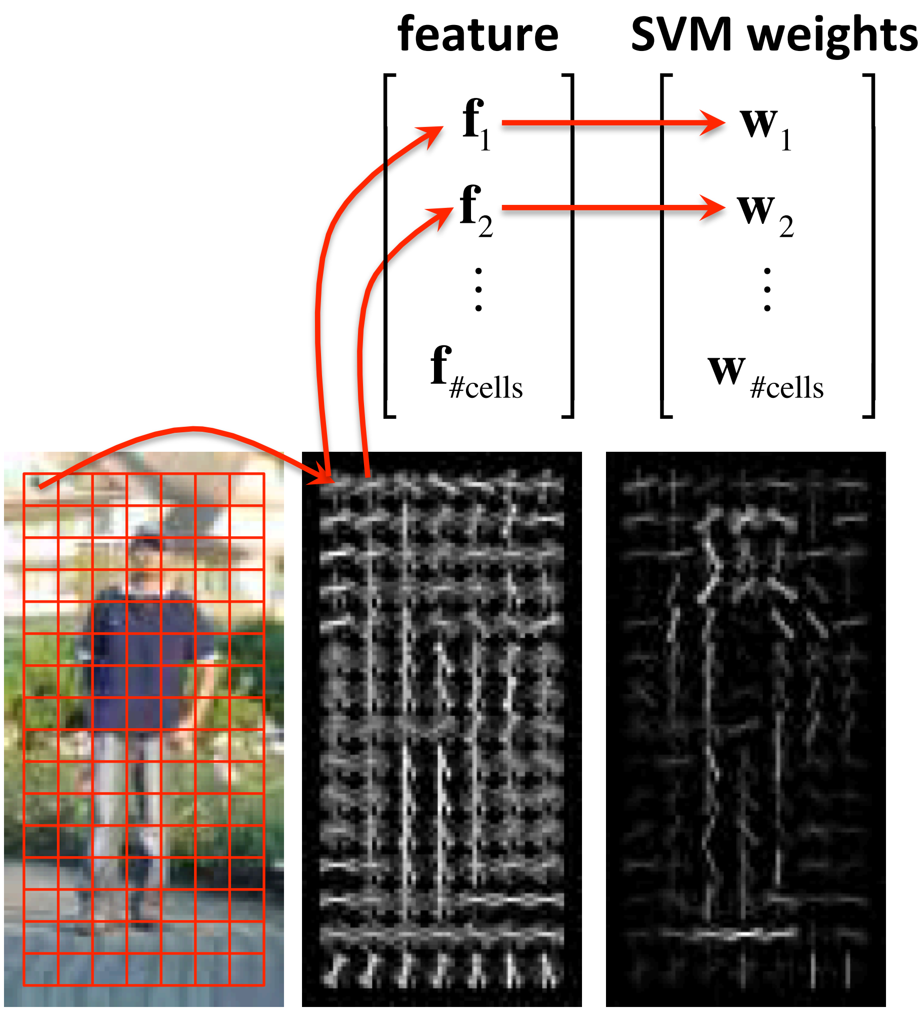
Training: CAD models

Training: Rendering Depth

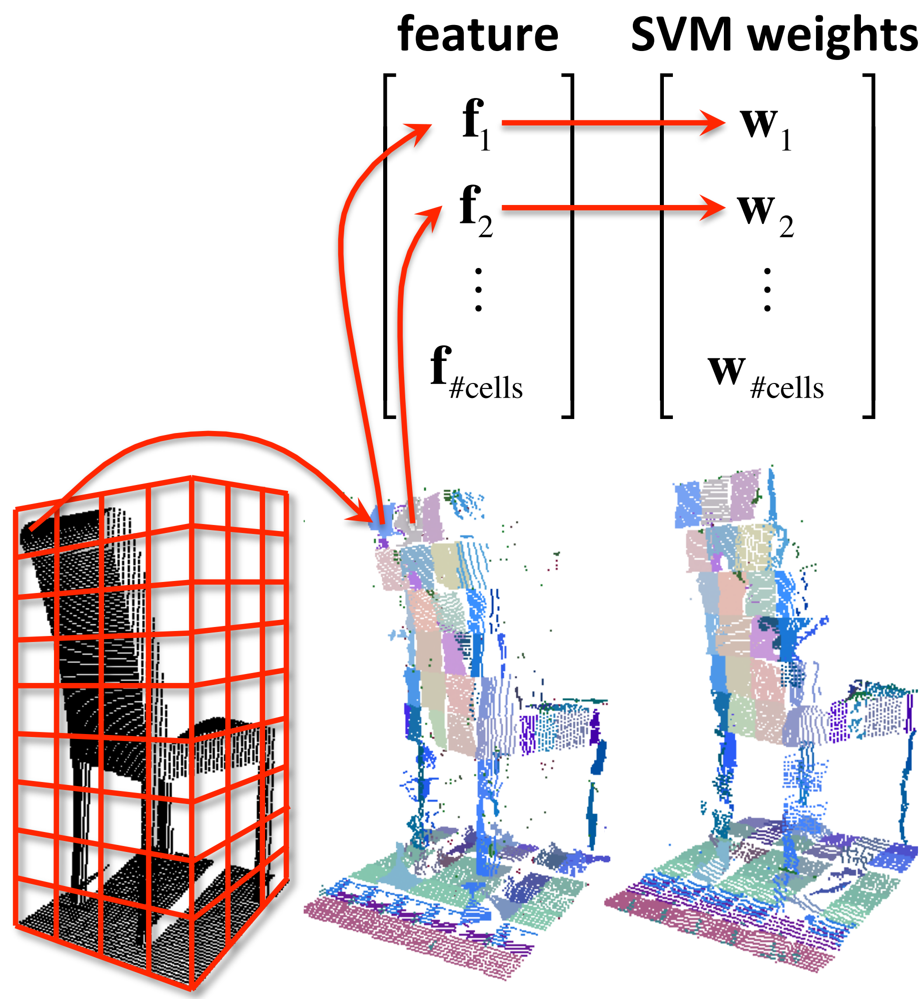
Training: 3D Exemplar SVM



Recipe of Detector Features

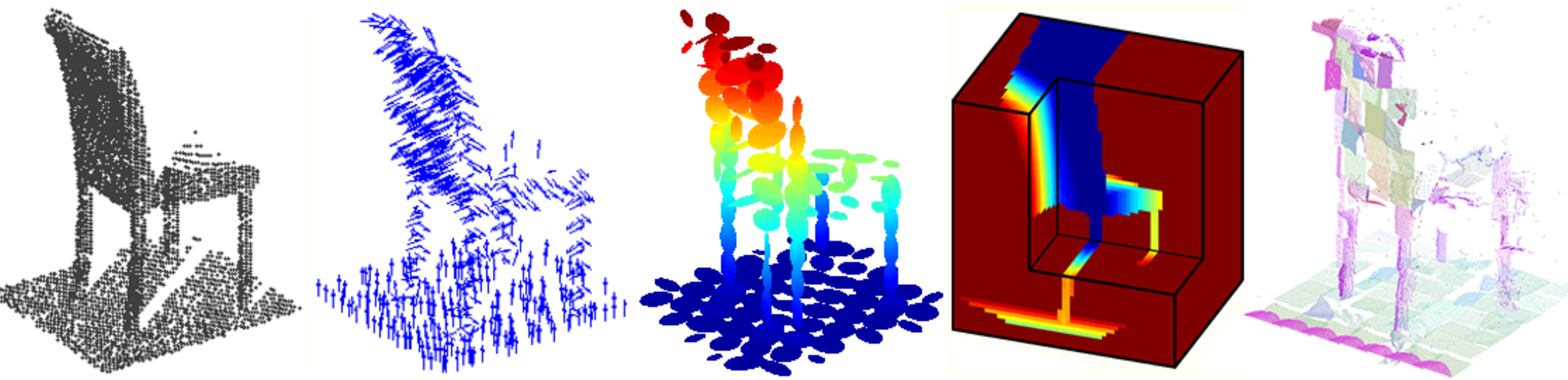


2D Object Detector



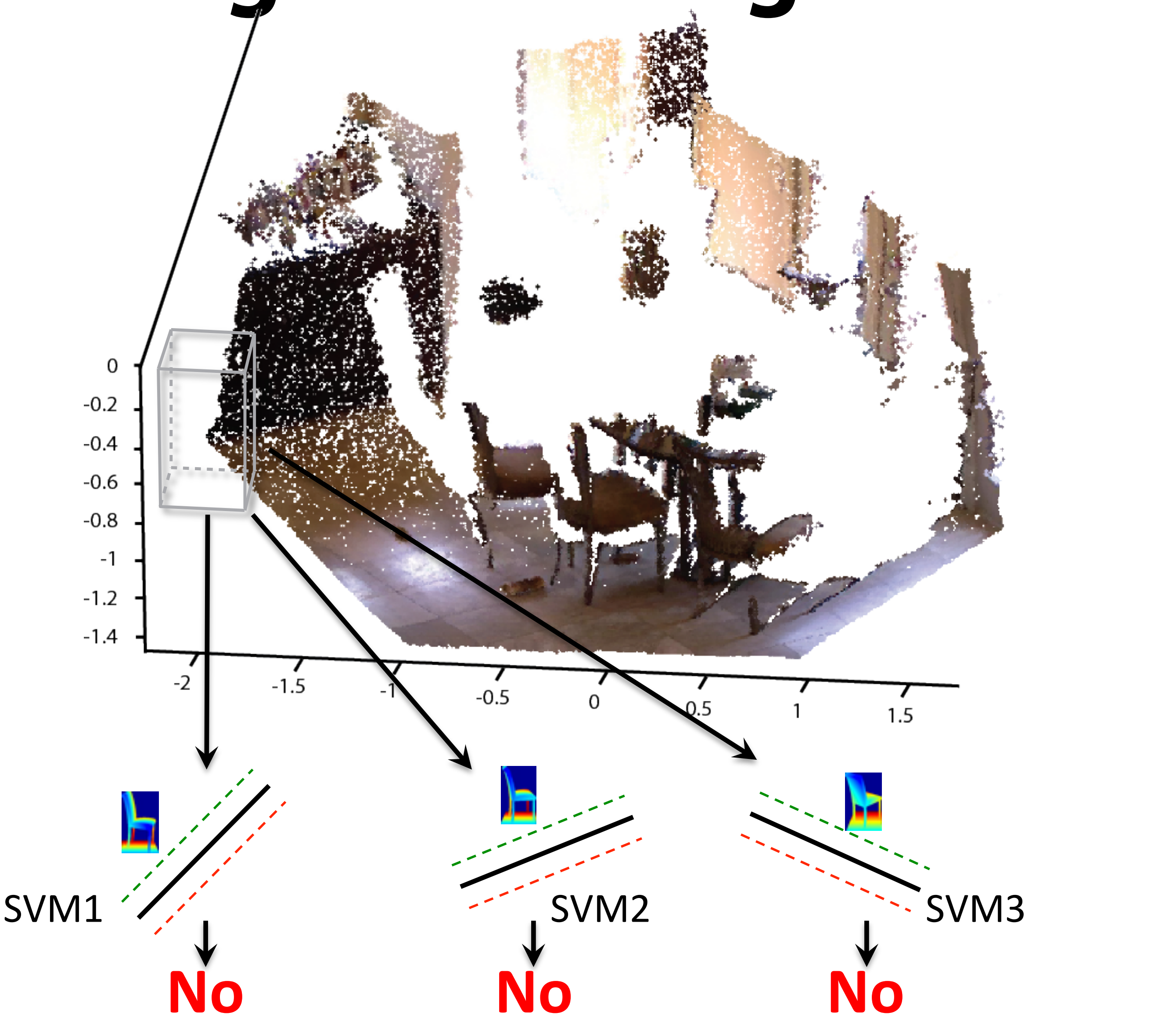
3D Object Detector

3D Features

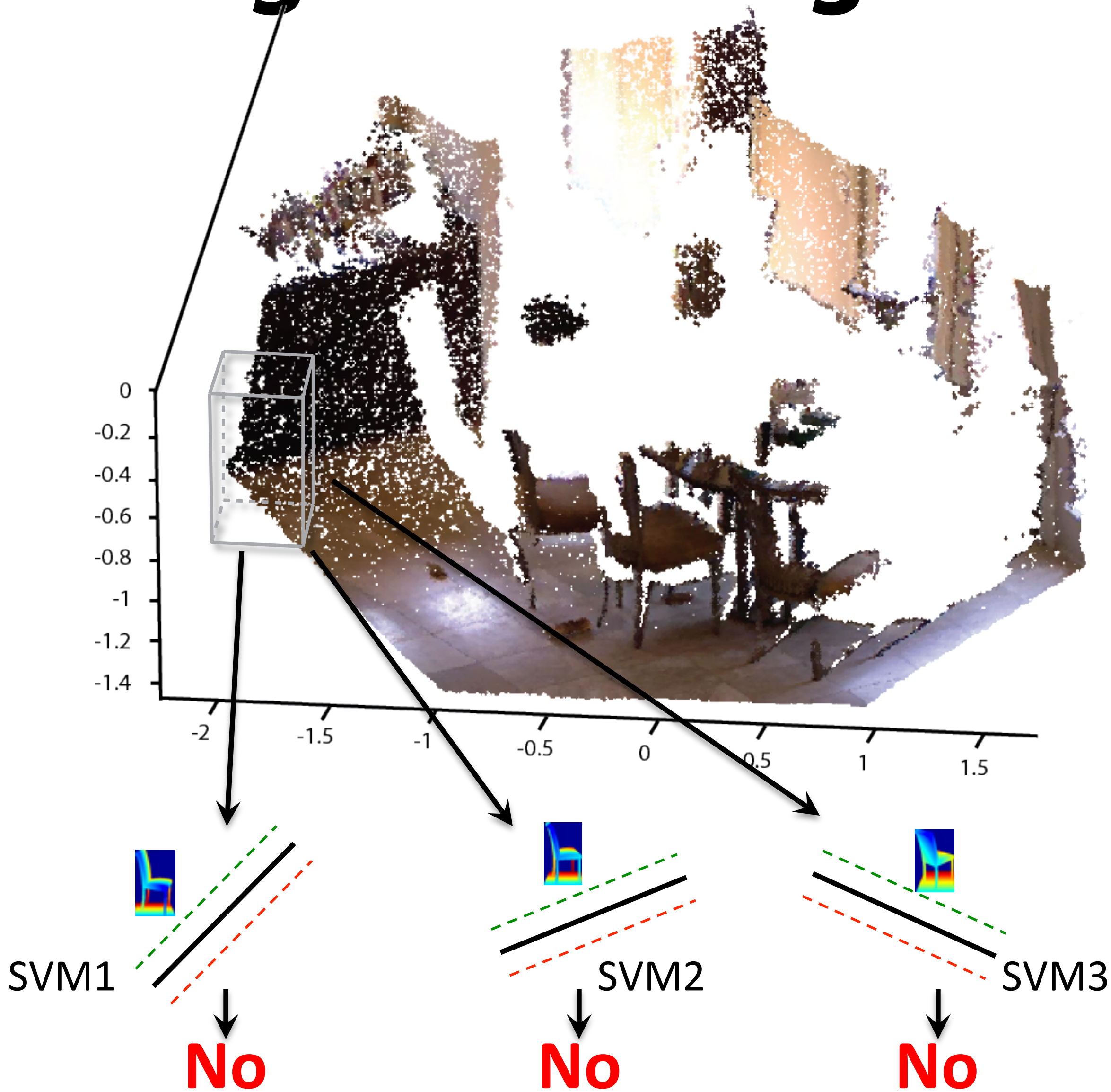


Points Normal Shape TSDF Combined

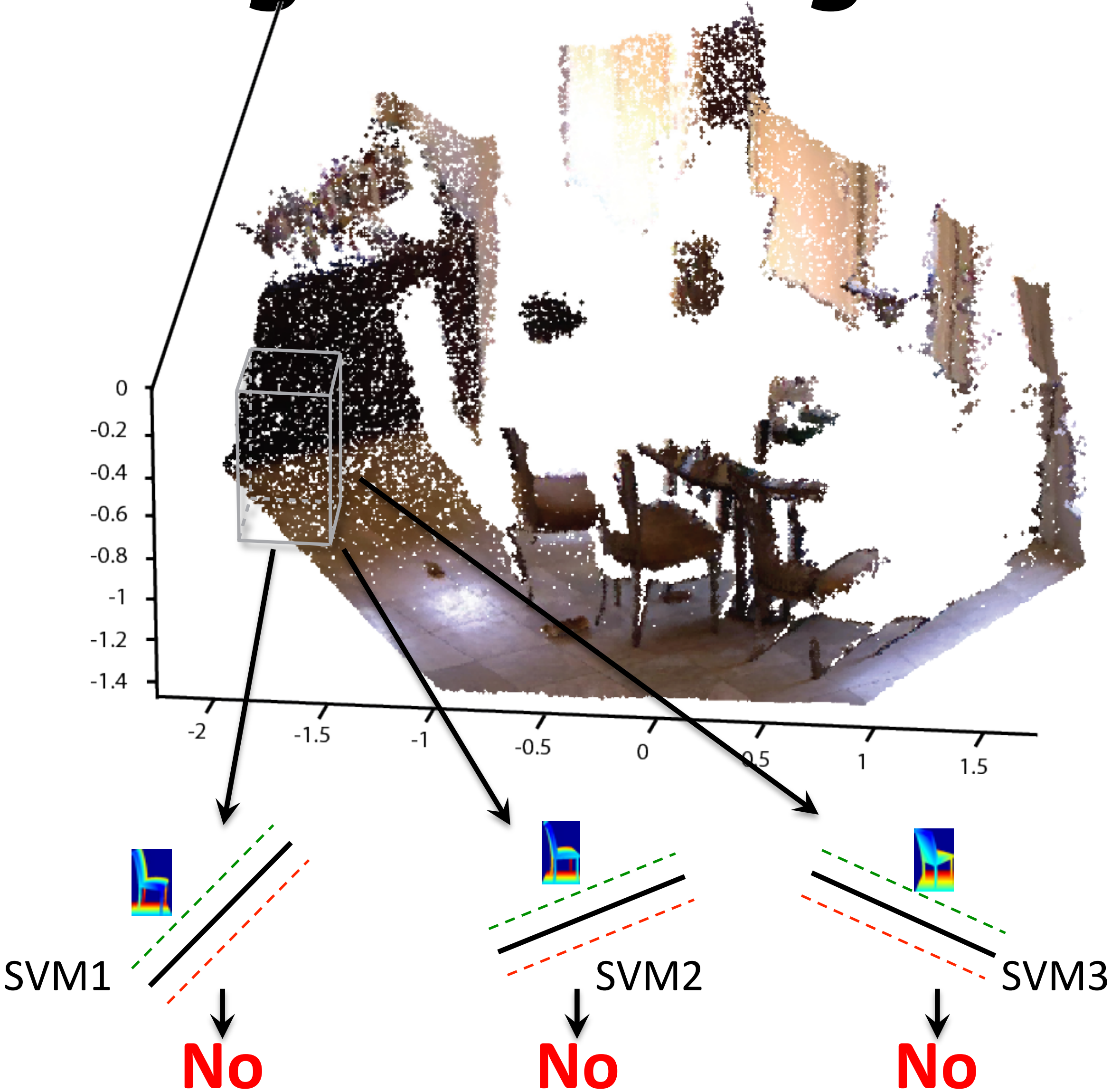
Testing: 3D Sliding Window



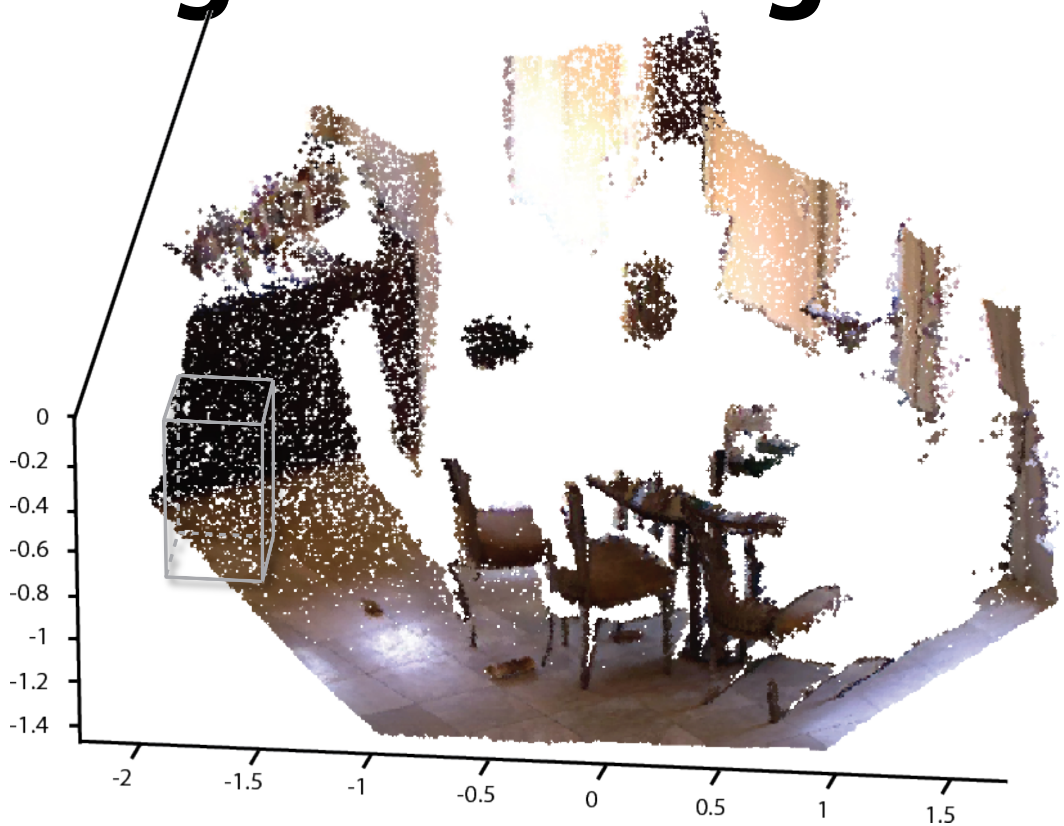
Testing: 3D Sliding Window

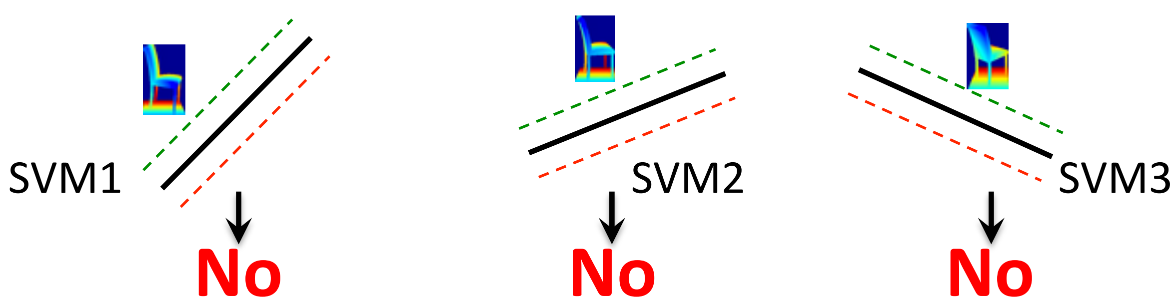


Testing: 3D Sliding Window

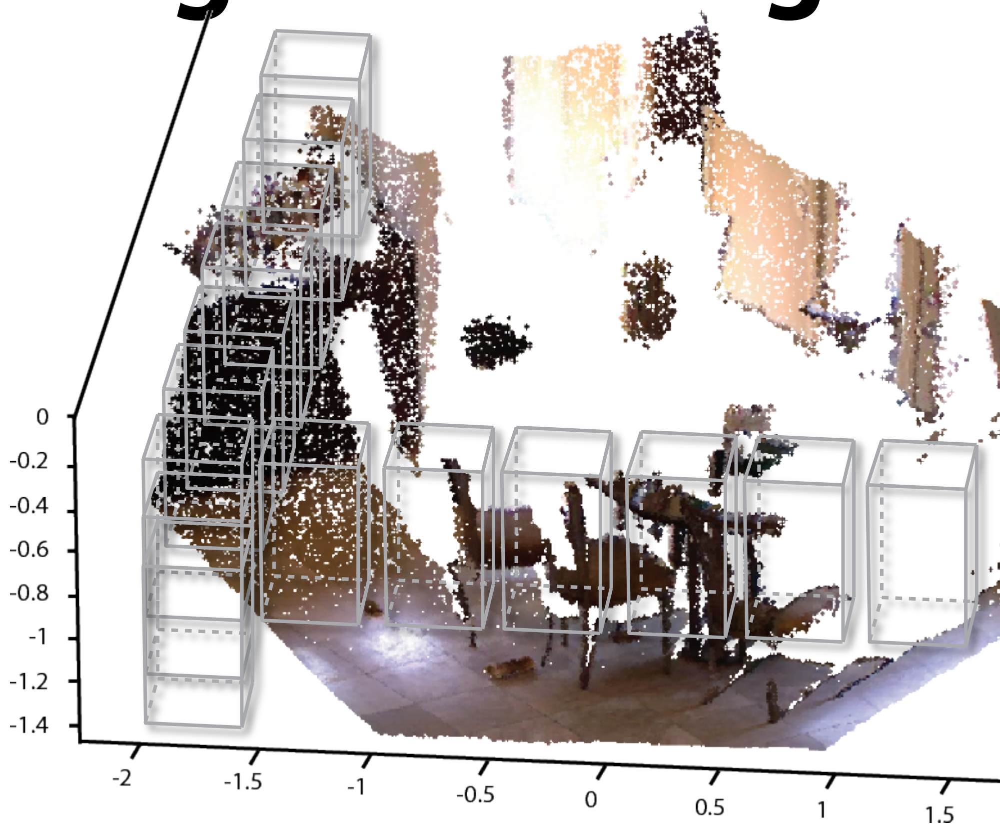


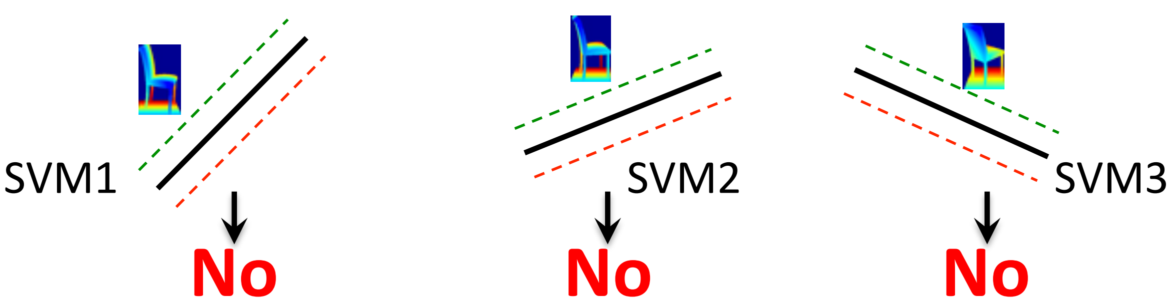
Testing: 3D Sliding Window



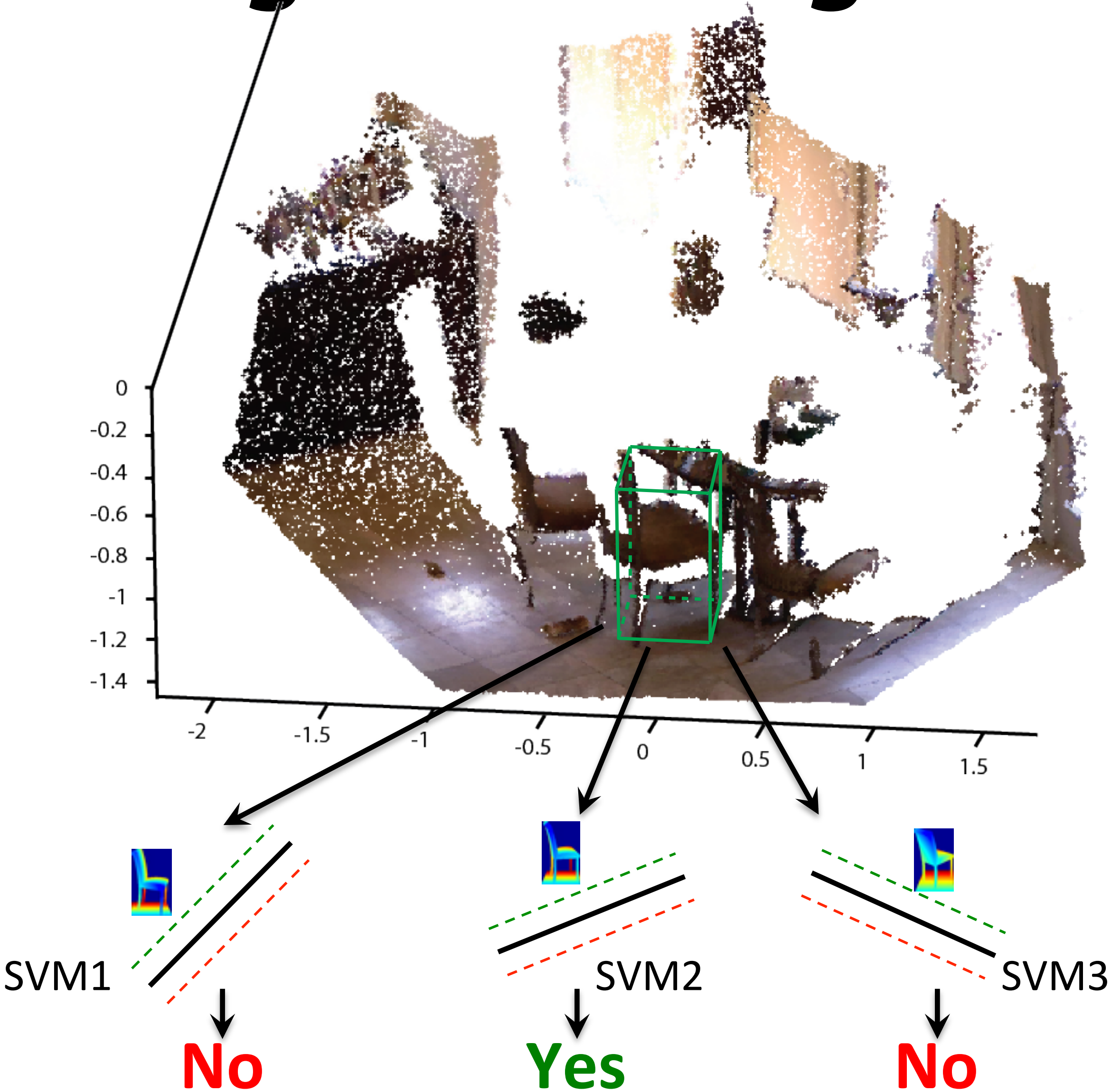


Testing: 3D Sliding Window

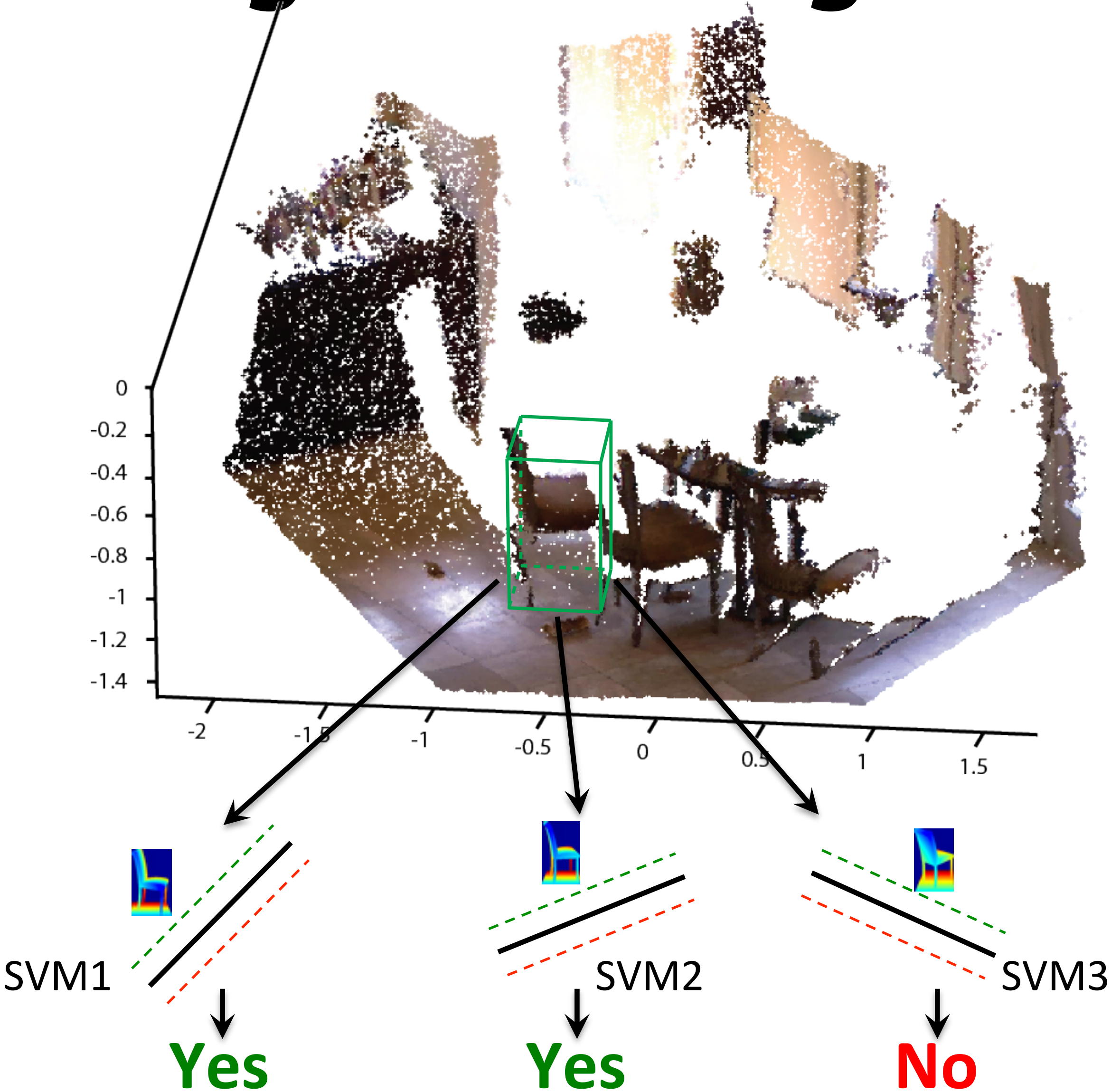




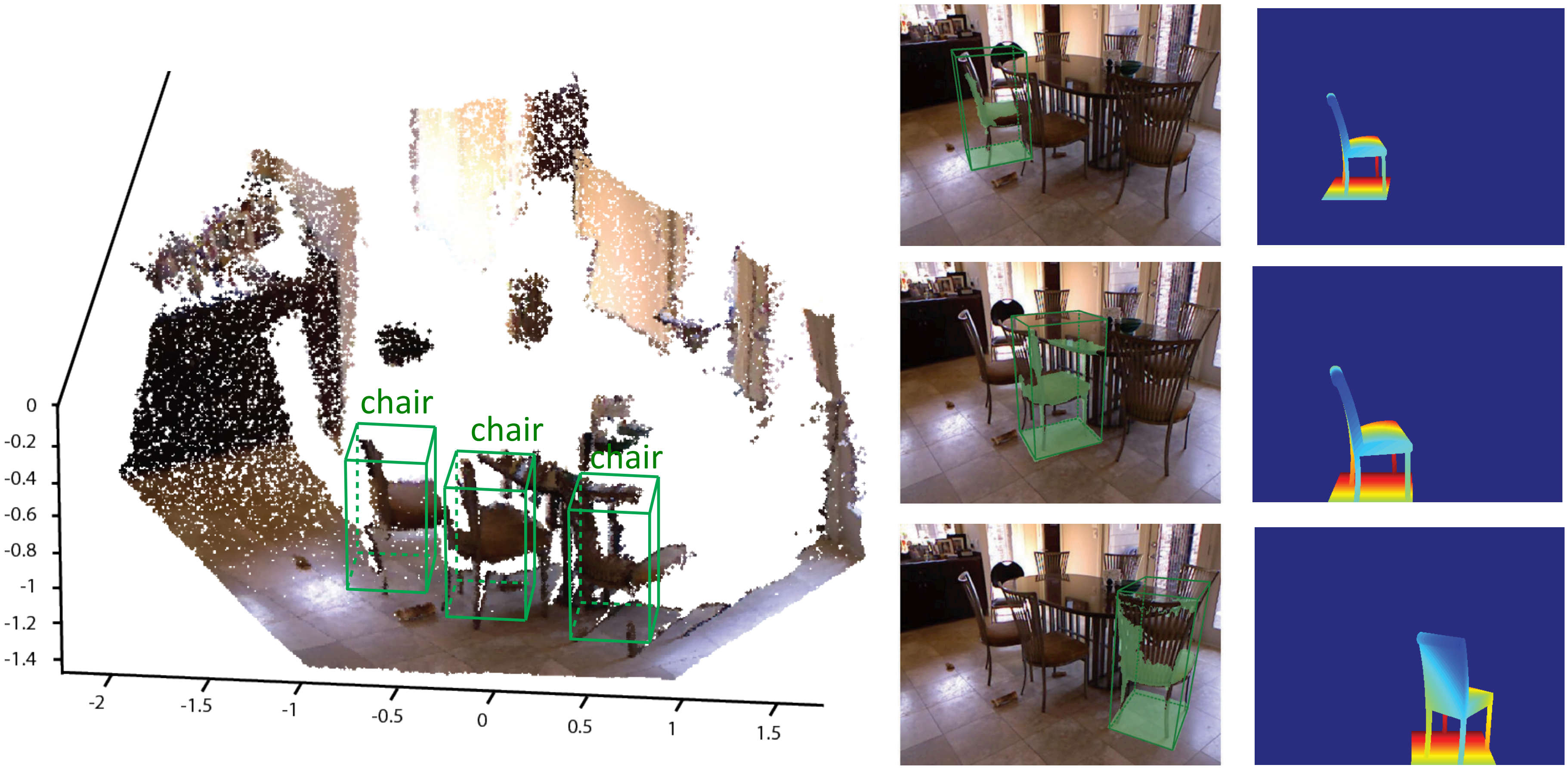
Testing: 3D Sliding Window



Testing: 3D Sliding Window

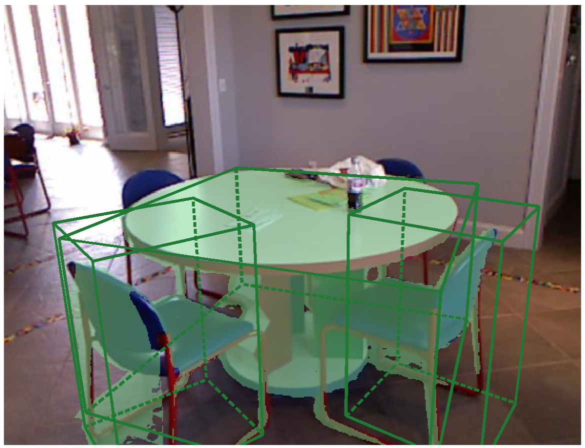


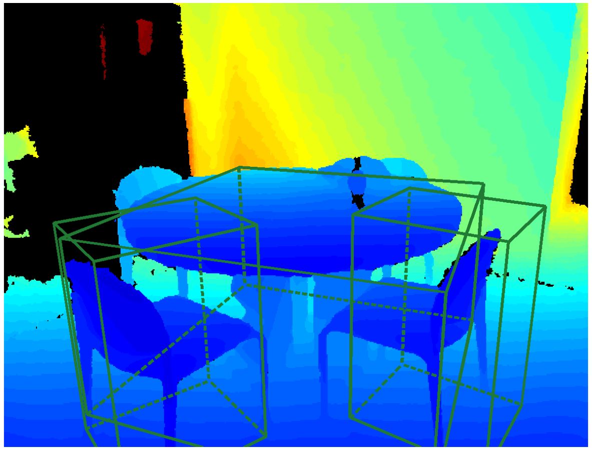
Testing: 3D Sliding Window



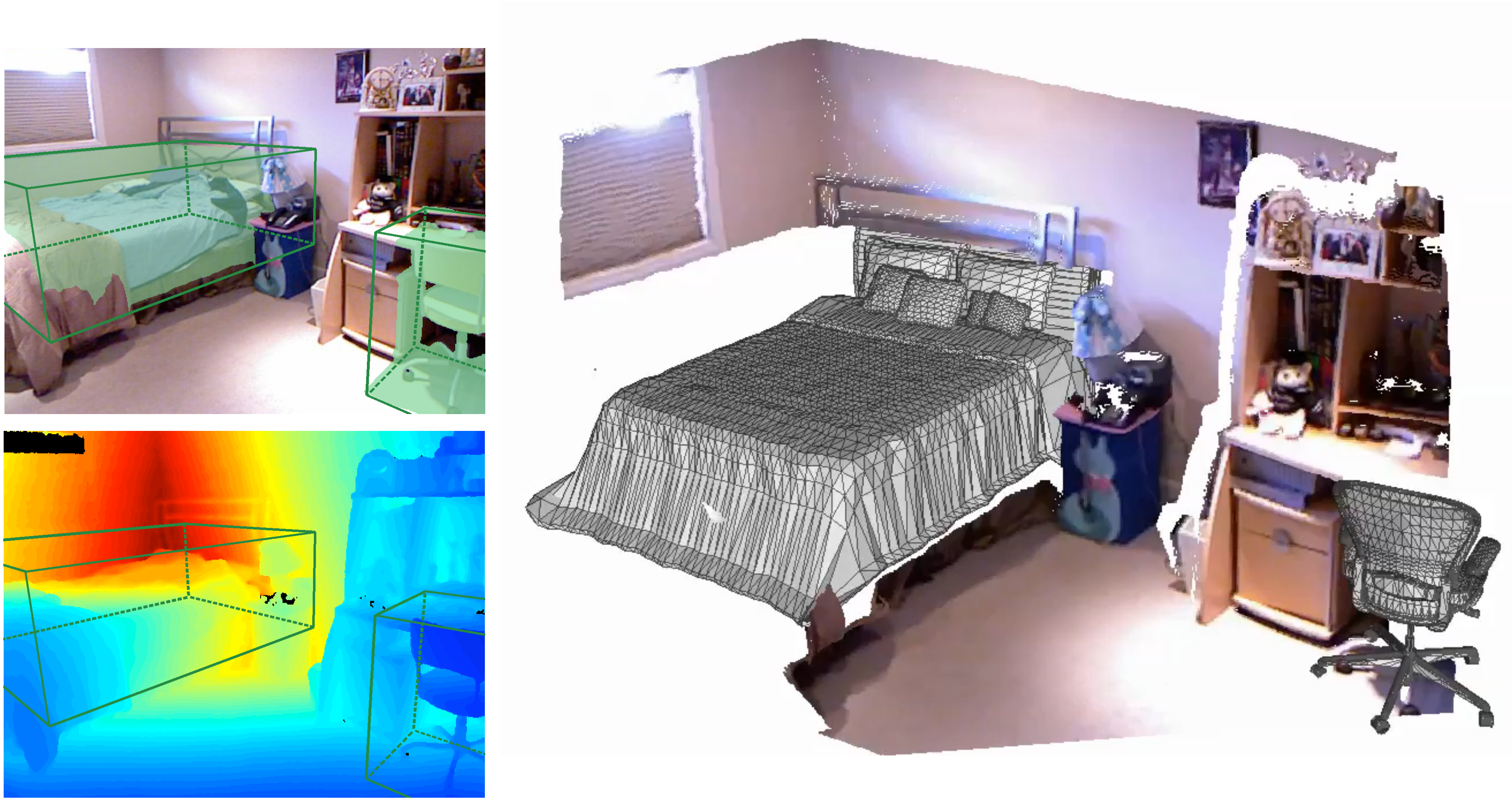
Exemplar Rendering

Results



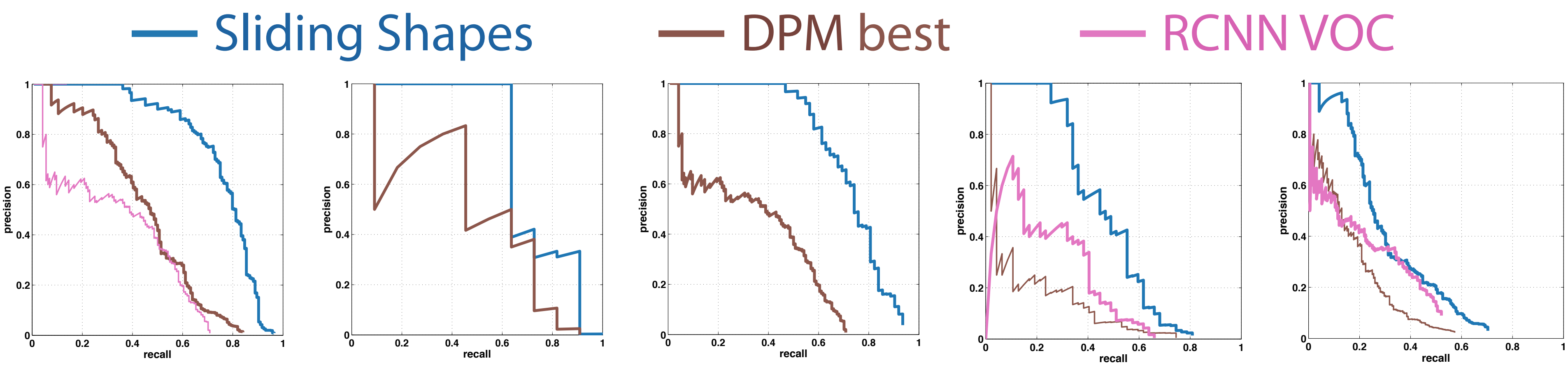


Results



Results

Evaluation

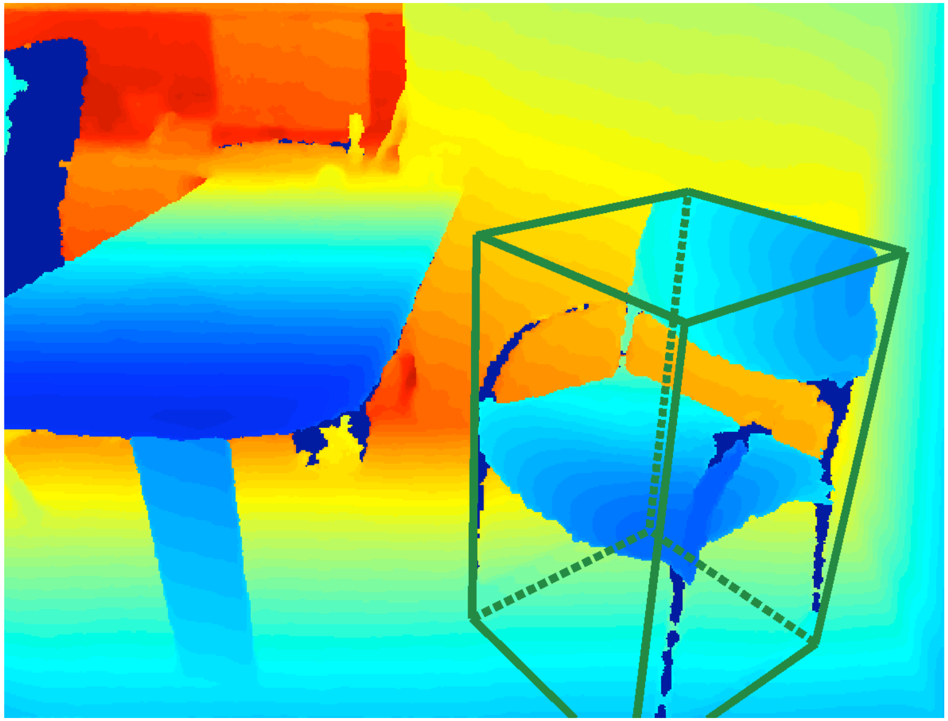


chair toilet bed sofa table

		chair				toilet				bed				sofa				table			
		3D+	3D	2D+	2D	3D+	3D	2D+	2D	3D+	3D	2D+	2D	3D+	3D	2D+	2D	3D+	3D	2D+	2D
	Sliding Shapes	**0.316**	**0.765**	**0.331**	**0.749**	**0.643**	**0.736**	**0.644**	**0.736**	**0.381**	**0.741**	**0.412**	**0.751**	**0.315**	**0.403**	**0.339**	**0.418**	**0.289**	**0.474**	**0.314**	**0.478**
RGB	DPM-VOC[2]	-	-	0.176	**0.446**	-	-	-	-	-	-	-	-	-	-	0.163	**0.213**	-	-	0.127	0.175
	DPM-SUN[2]	-	-	0.131	0.345	-	-	0.309	**0.532**	-	-	0.279	**0.503**	-	-	0.109	0.132	-	-	0.120	0.157
	DPM-RMRC[2]	-	-	0.115	0.269	-		**0.344**	0.419	-	-	**0.318**	0.427	-	-	0.099	0.137	-	-	0.045	0.048
	RCNN-VOC[6]	-	-	**0.182**	0.342	-	-	-	-	-	-	-	-	-	-	**0.200**	0.203	-	-	**0.213**	**0.237**

Comparison with DPM

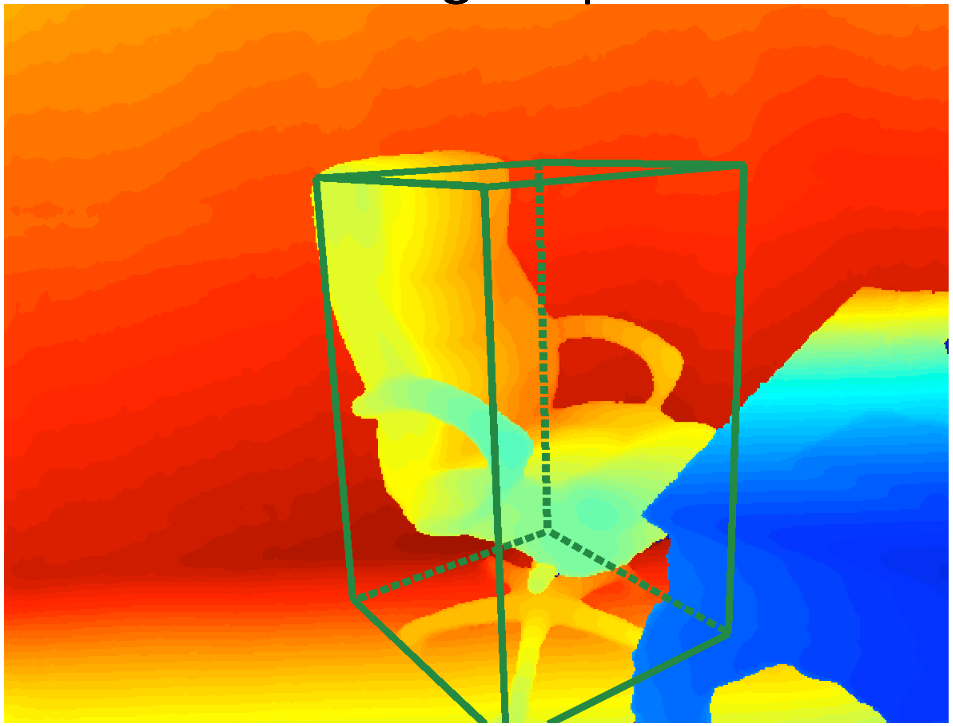
Sliding Shapes



DPM

Comparison with DPM

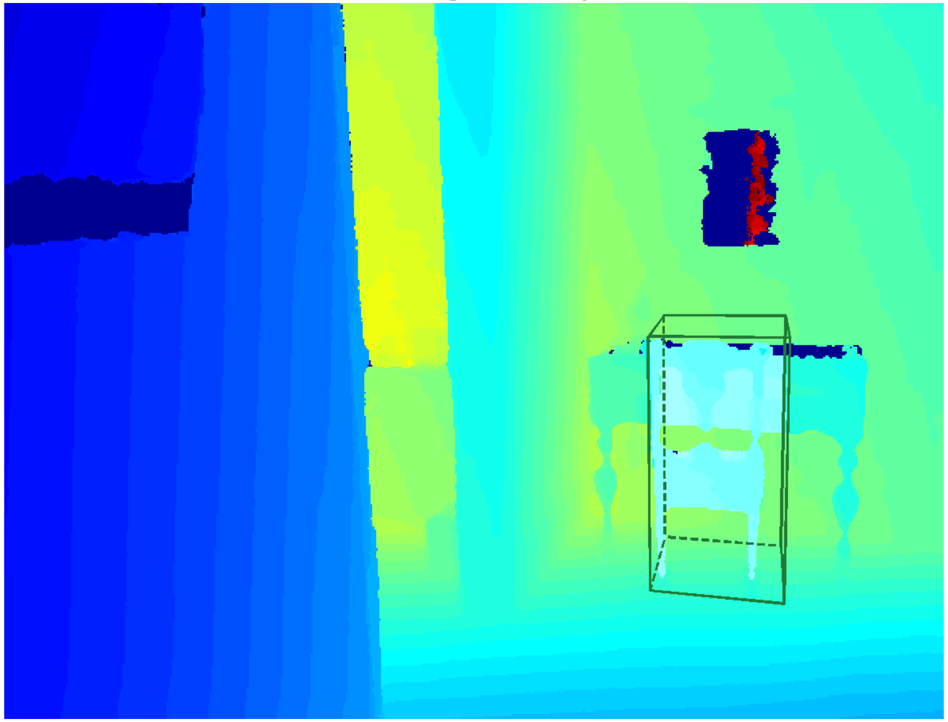
Sliding Shapes



DPM

Comparison with DPM

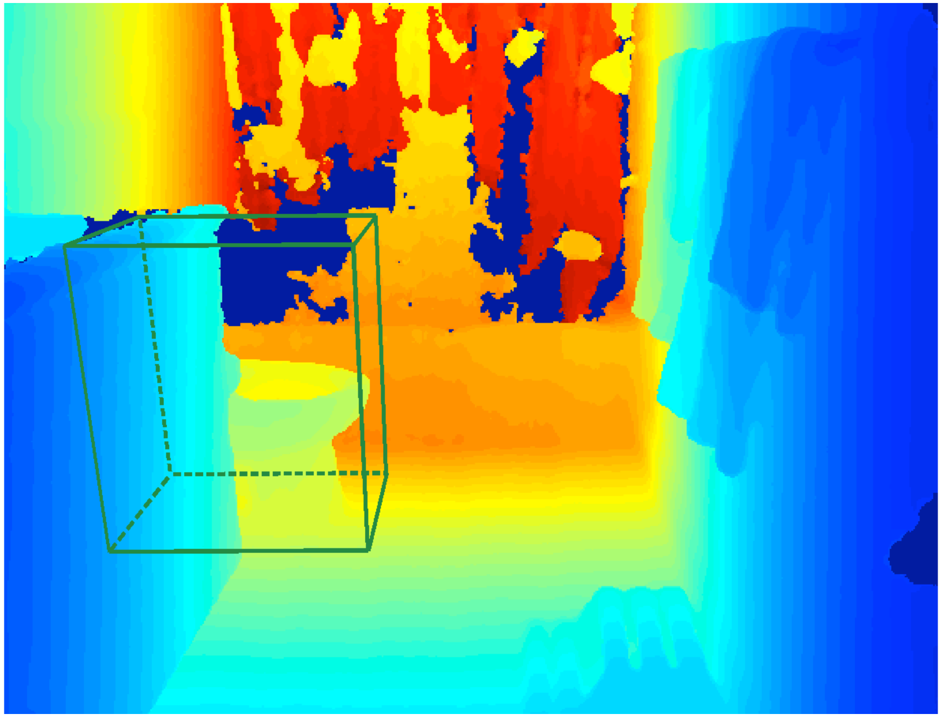
Sliding Shapes



DPM

Comparison with DPM

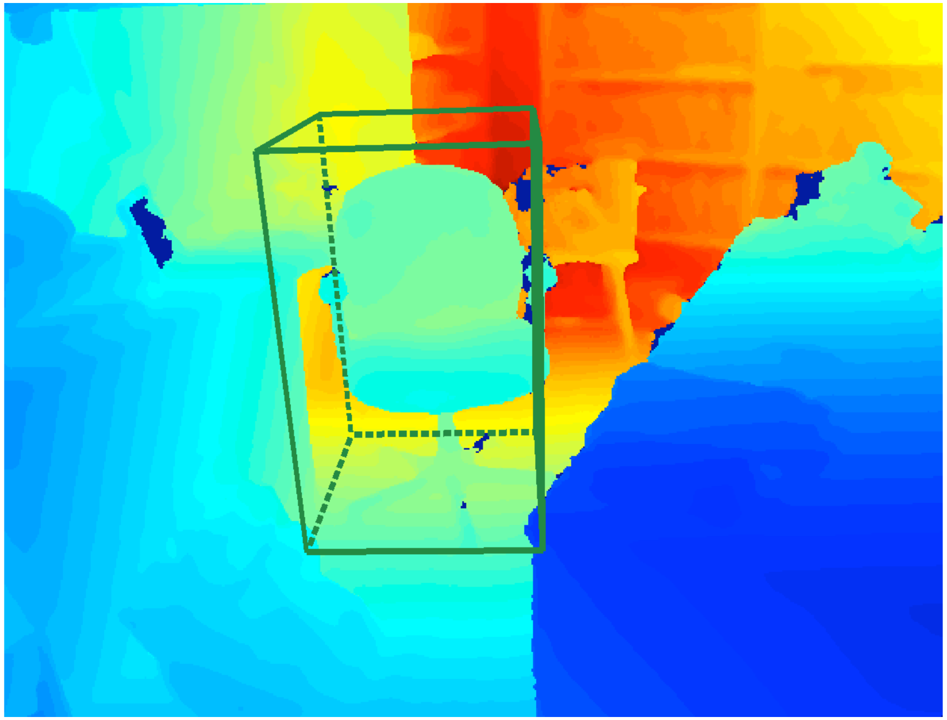
Sliding Shapes



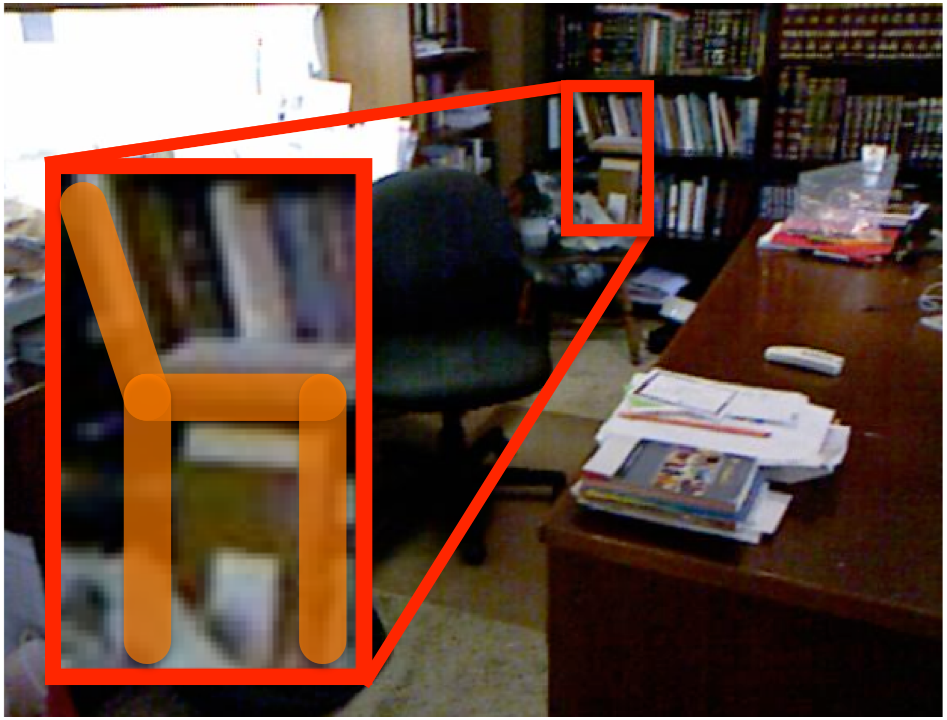
DPM

Comparison with DPM

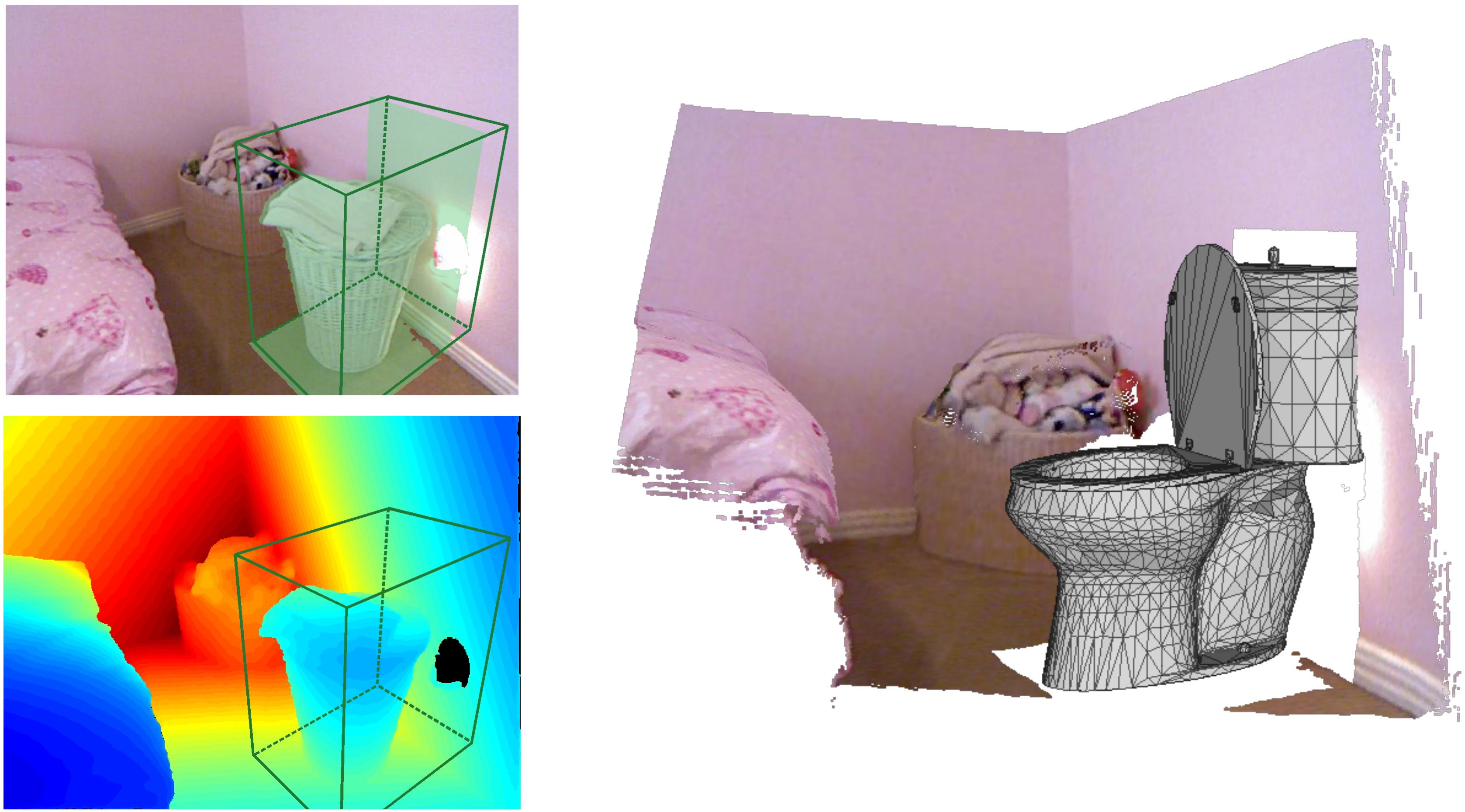
Sliding Shapes



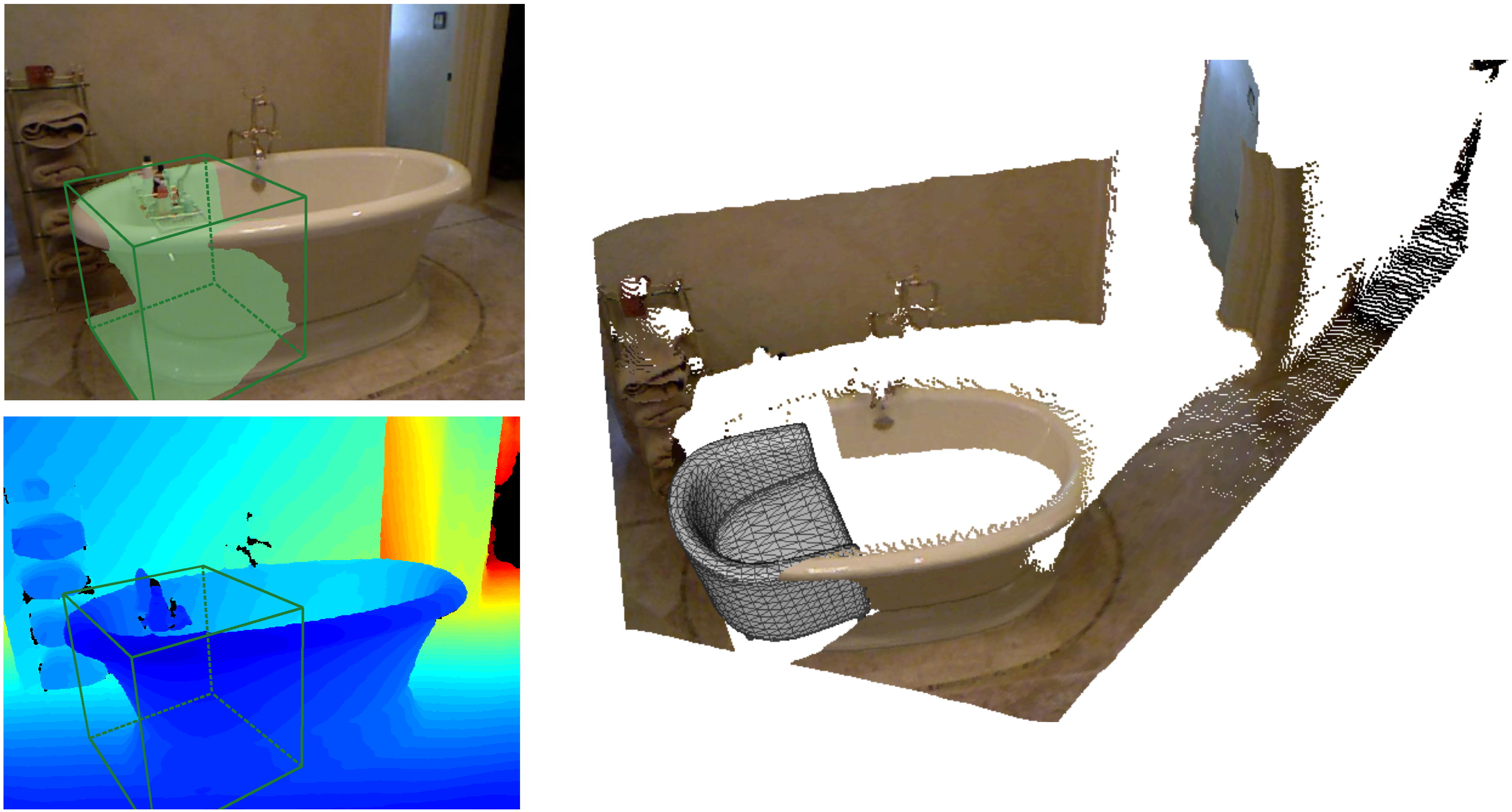
DPM



False Positives



False Positives



Analysis

Object Detection is Hard

Jia Deng's advisor's advisor's advisor's advisor

Nevatia & Binford, 1977.

IMAGENET

"These techniques are inadequate for three-dimensional scene analysis for many reasons:"


1. Variation


2. Viewpoint


3. Illumination


4. Clutter


5. Occlusion

1. Variation

2. Viewpoint

3. Illumination

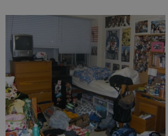
4. Clutter

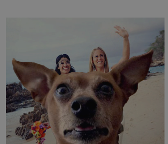
5. Occlusion


1. Variation


2. Viewpoint


3. Illumination


4. Clutter


5. Occlusion

Problem: Intra-class Variations

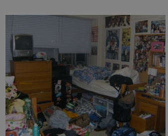
CAD Models Used for Training

1. Variation

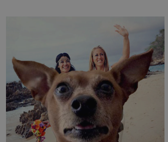
Solution: Data-driven Exemplar

2. Viewpoint

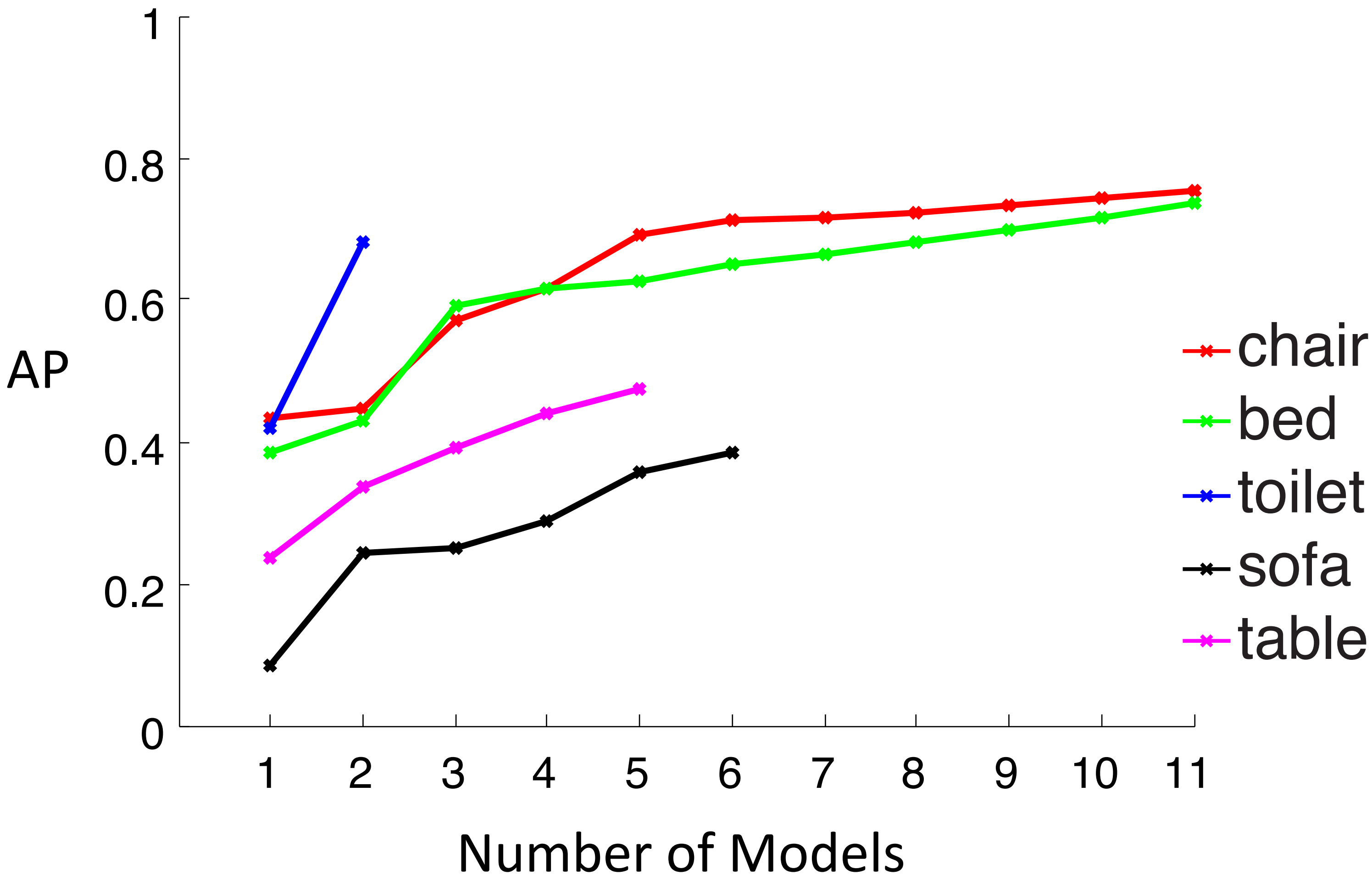
3. Illumination



4. Clutter



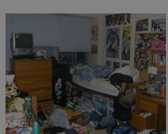
5. Occlusion

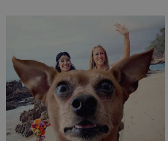



1. Variation


2. Viewpoint


3. Illumination


4. Clutter

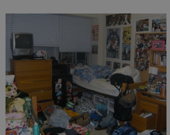

5. Occlusion

Problem: Viewpoints

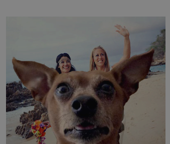
1. Variation

2. Viewpoint

3. Illumination

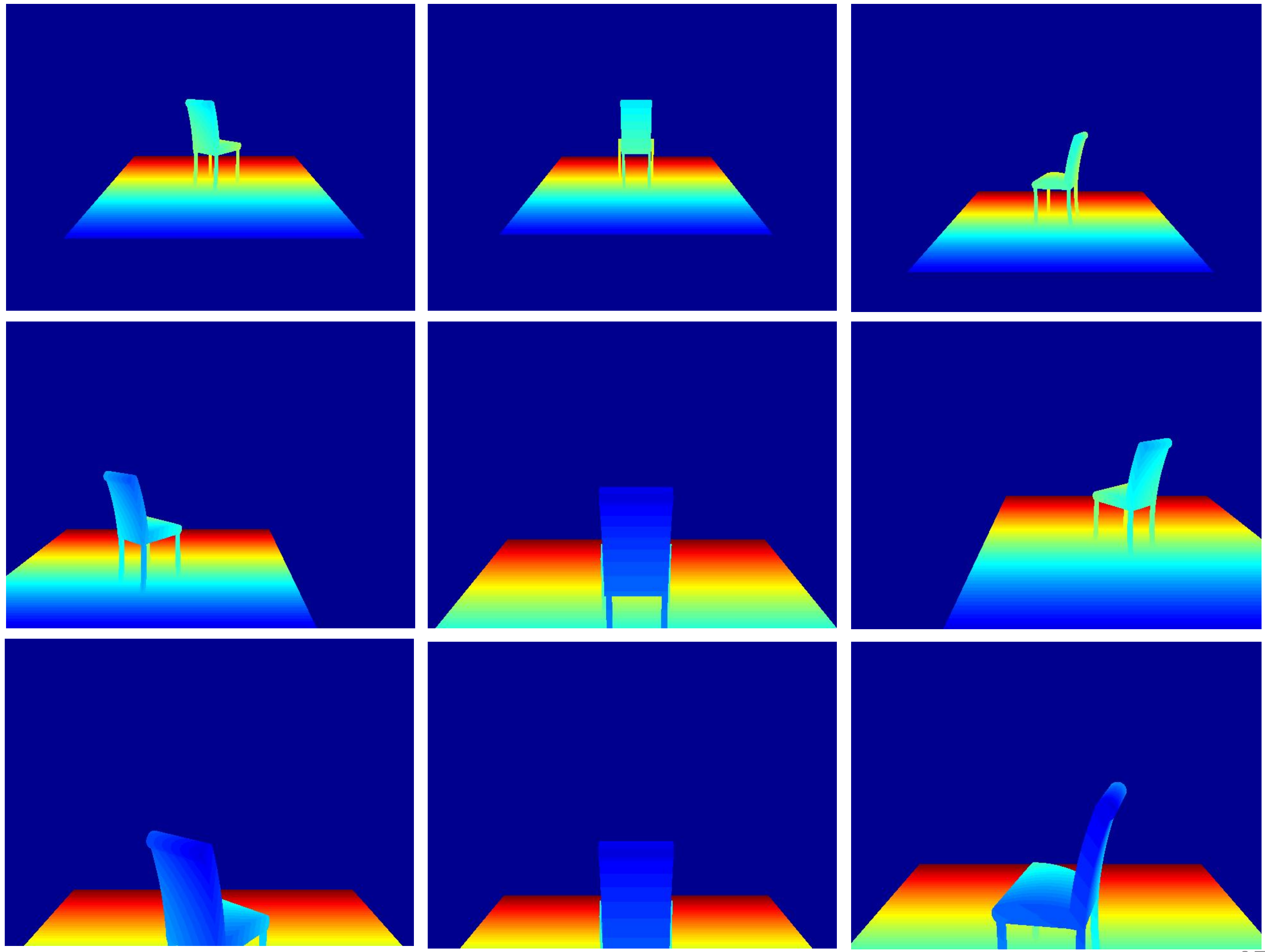


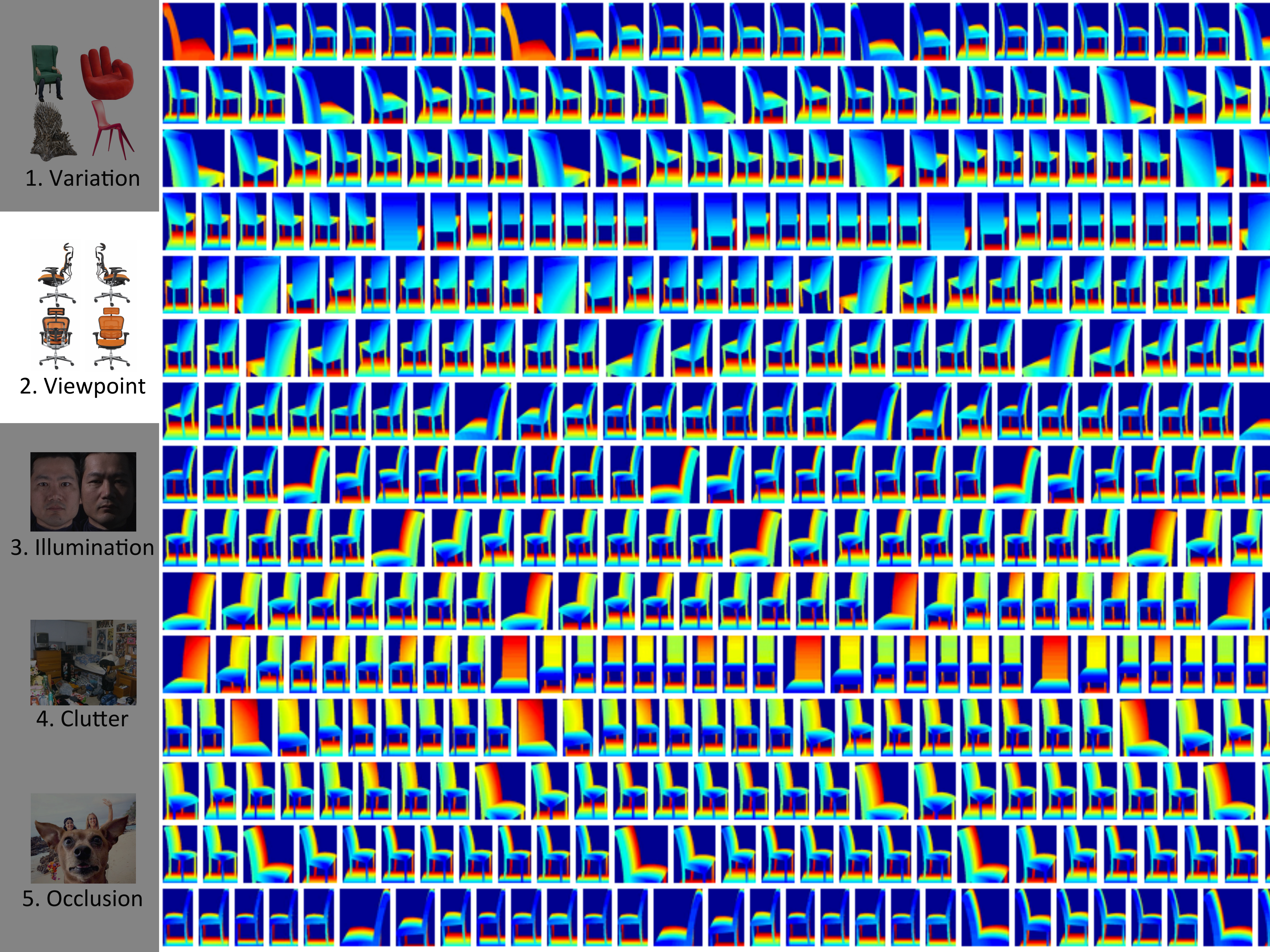
4. Clutter



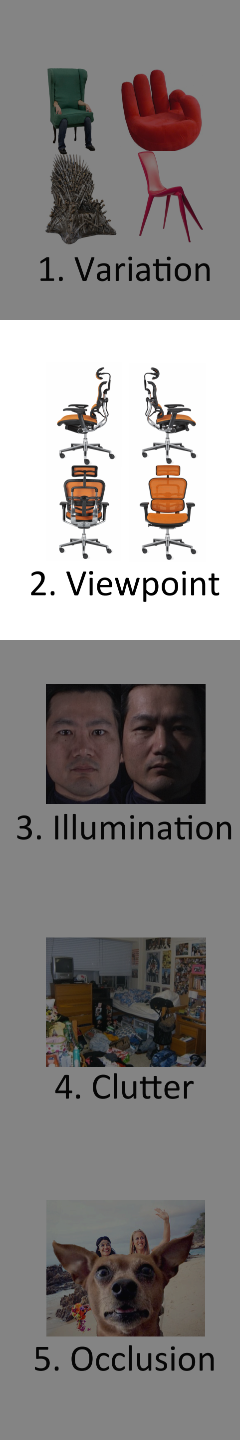
5. Occlusion

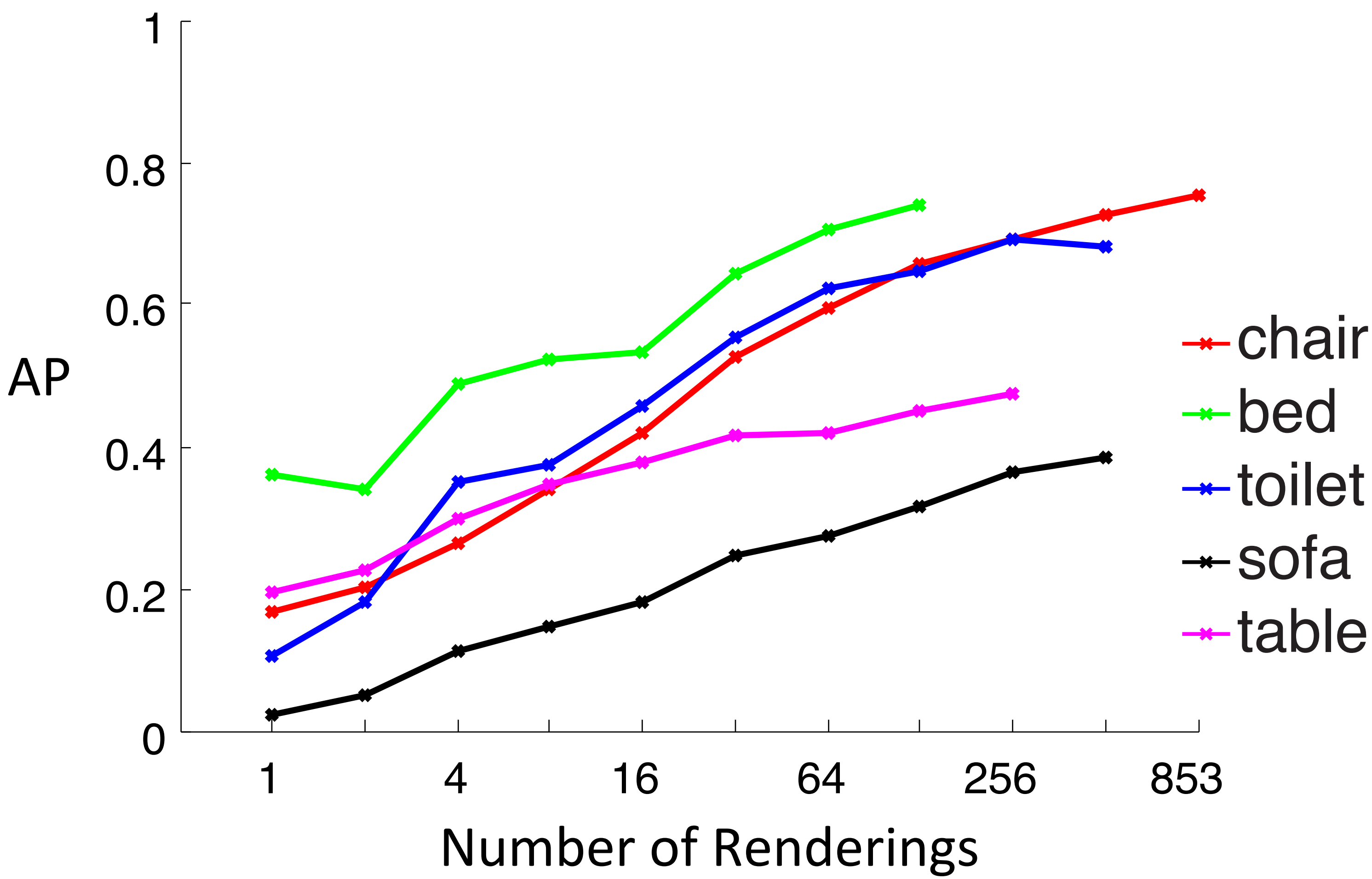
Solution: Numerate All Views




1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion

Solution: Numerate All Views





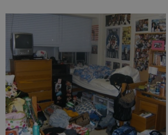
Problem: Illumination

- Color Rendering ≠ Real Photo

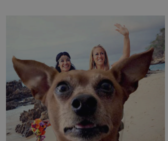
1. Variation

2. Viewpoint

3. Illumination



4. Clutter

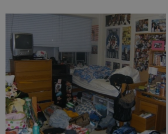


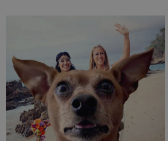
5. Occlusion


1. Variation


2. Viewpoint


3. Illumination


4. Clutter


5. Occlusion

Solution: 3D Depth

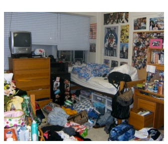
- Color Rendering $\neq$ Real Photo
- Depth Rendering $\approx$ Depth from Kinect

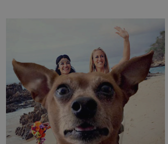
Kinect Body Pose Recognition [Shotton et al.]


1. Variation


2. Viewpoint


3. Illumination


4. Clutter


5. Occlusion

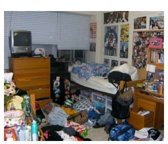
Problem: Clutter

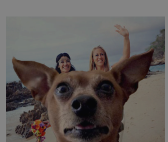

3D mesh


1. Variation


2. Viewpoint


3. Illumination

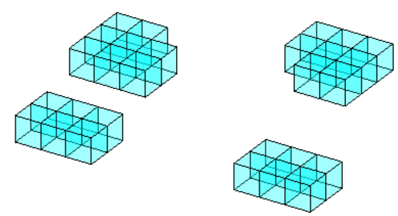

4. Clutter


5. Occlusion

Solution: Occupation Mask


3D mesh

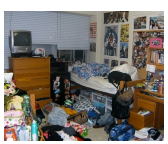
→

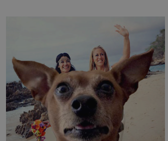

Occupation mask


1. Variation

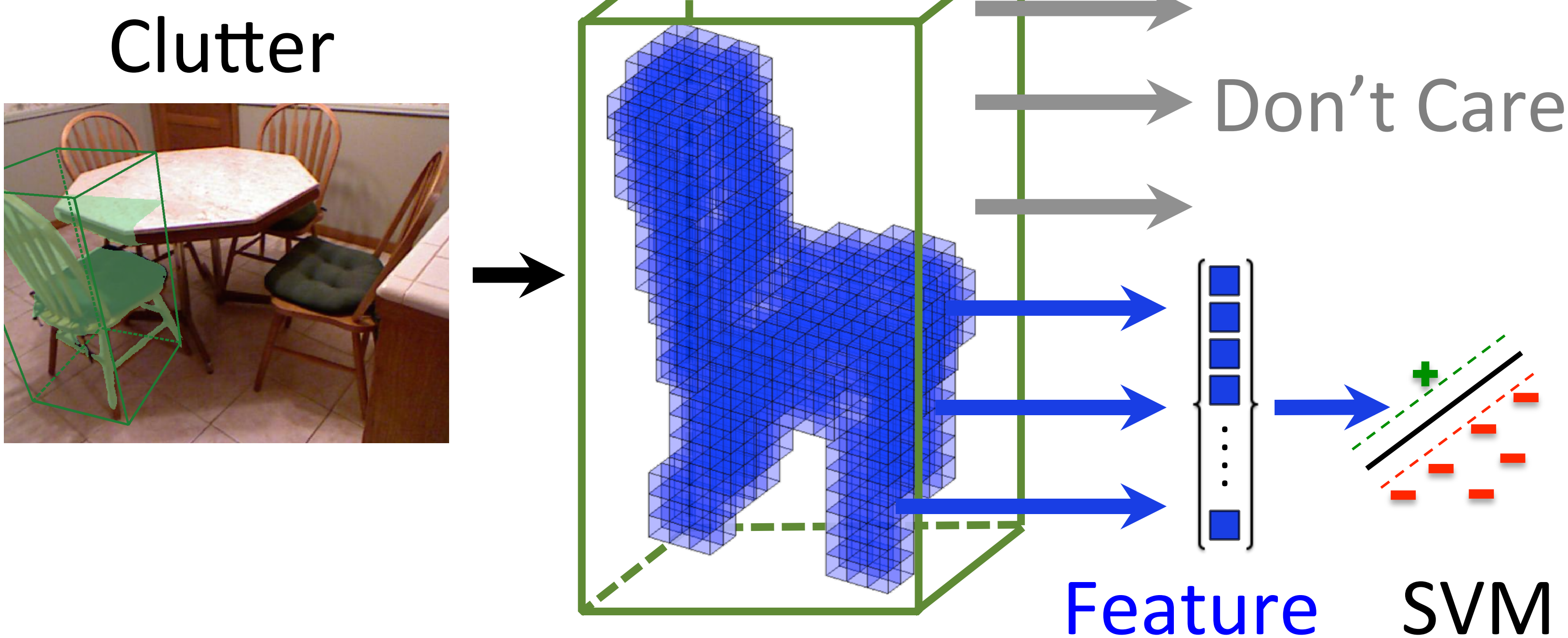

2. Viewpoint


3. Illumination


4. Clutter


5. Occlusion

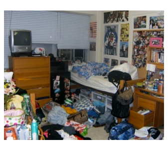
Solution: Occupation Mask



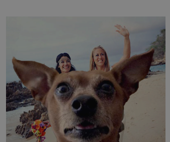
1. Variation

2. Viewpoint

3. Illumination

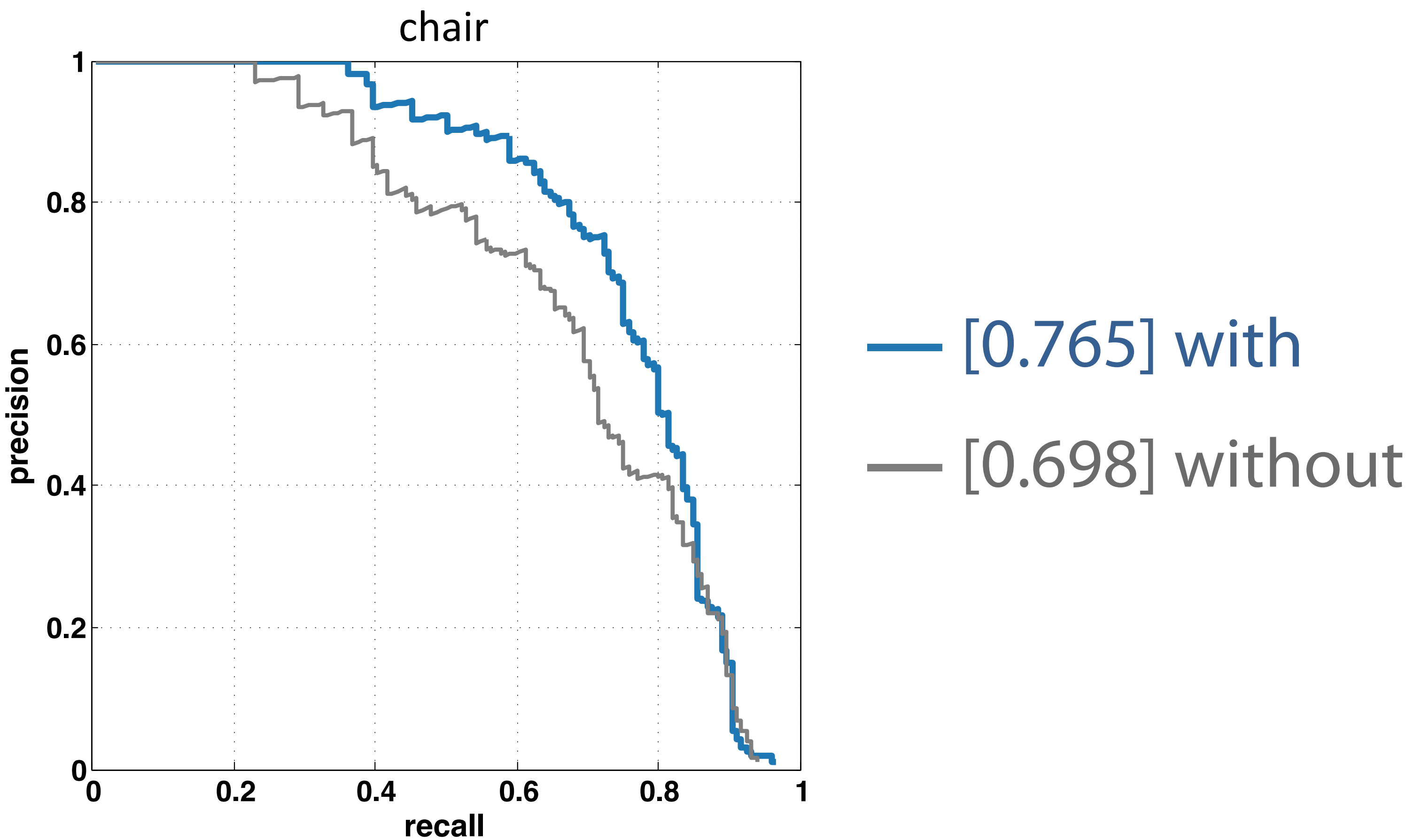


4. Clutter



5. Occlusion

Solution: Occupation Mask

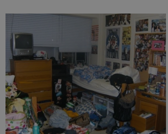


Problem: Occlusion

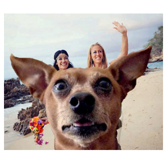
1. Variation

2. Viewpoint

3. Illumination



4. Clutter

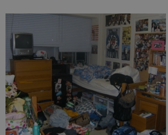


5. Occlusion

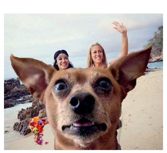
1. Variation

2. Viewpoint

3. Illumination

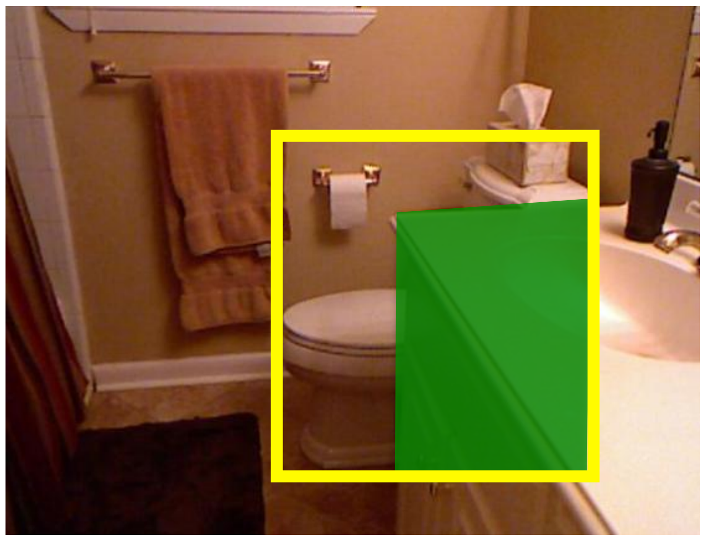


4. Clutter

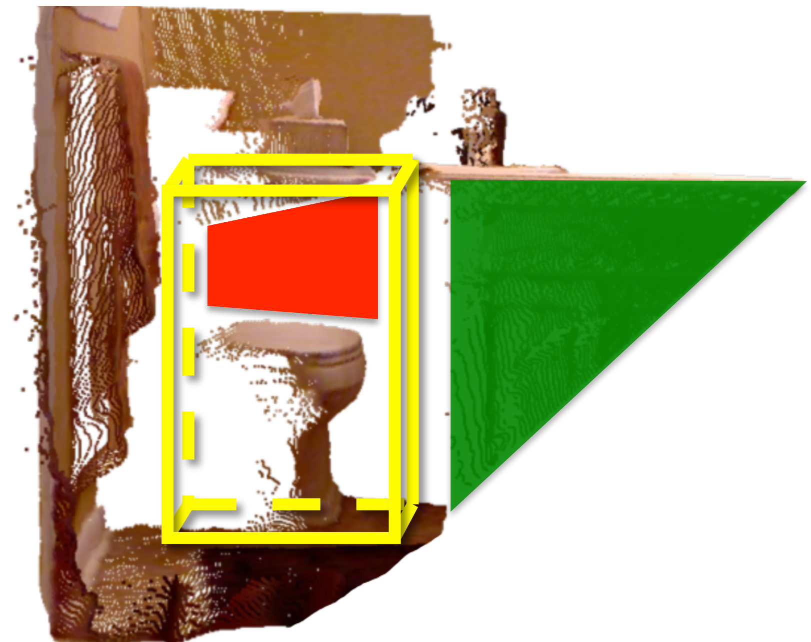


5. Occlusion

Solution: 3D Window



2D



3D

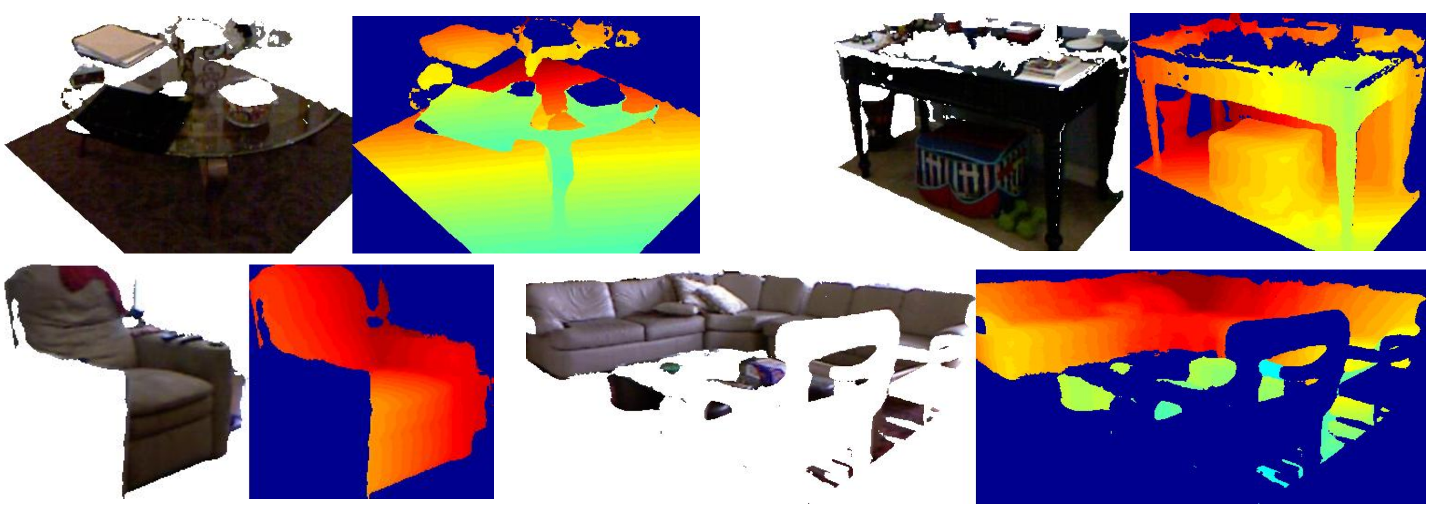
Using depth, we can know which part is **occluded**.

In 3D, we can separate the object from the **occluder**.

Data

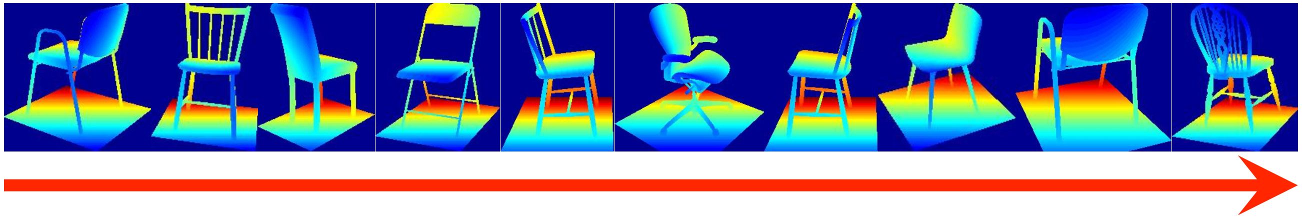
Quantity & Quality

- Kinect dataset is small
  - NYU (3D box): 1,074 images 4,258 objects
  - PASCAL VOC: 11,530 images 27,450 objects
  - SUN Database: 131,067 images 363,884 objects
- Kinect depth map is noisy



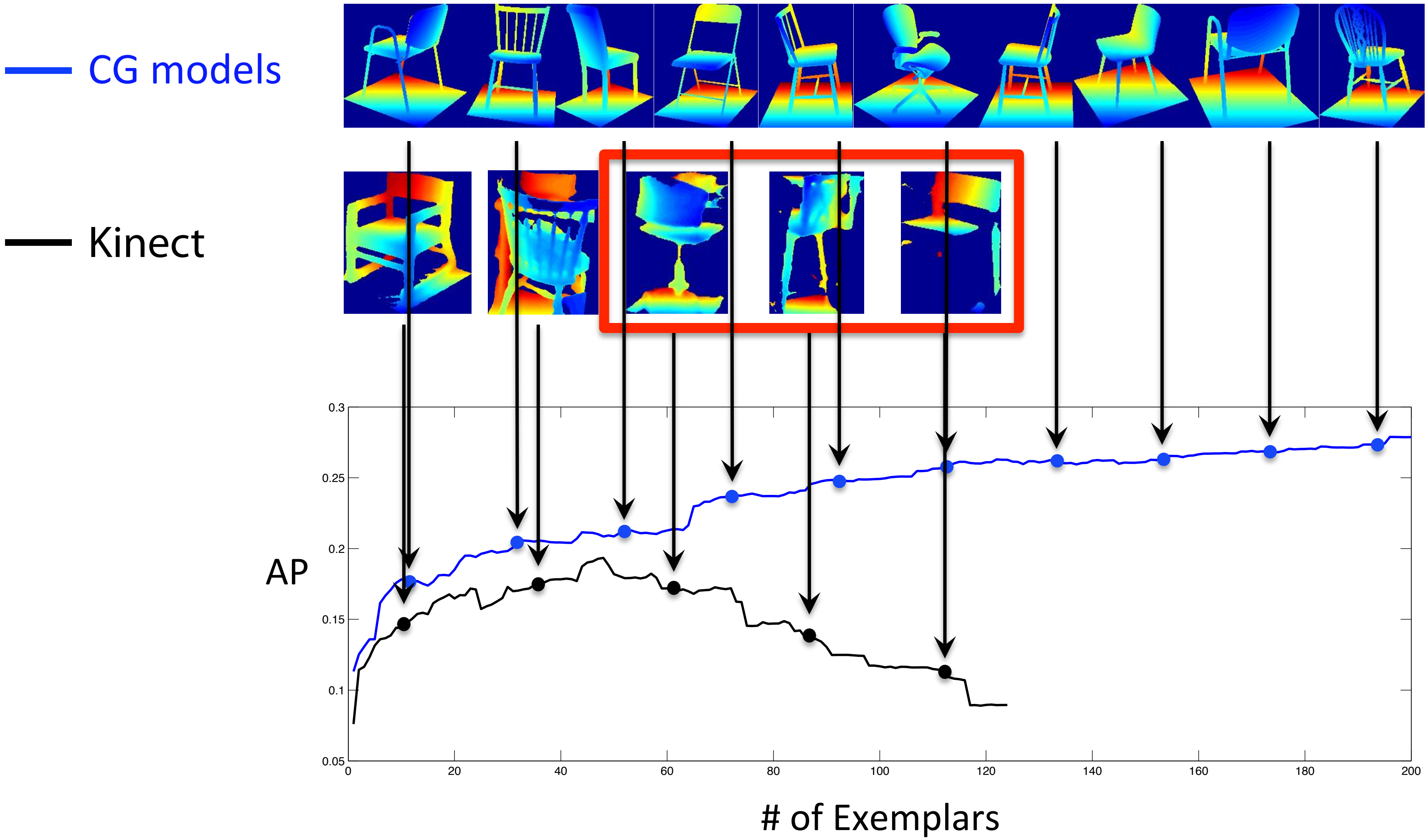
CG vs. Kinect

CG models



Individual AP : High Low

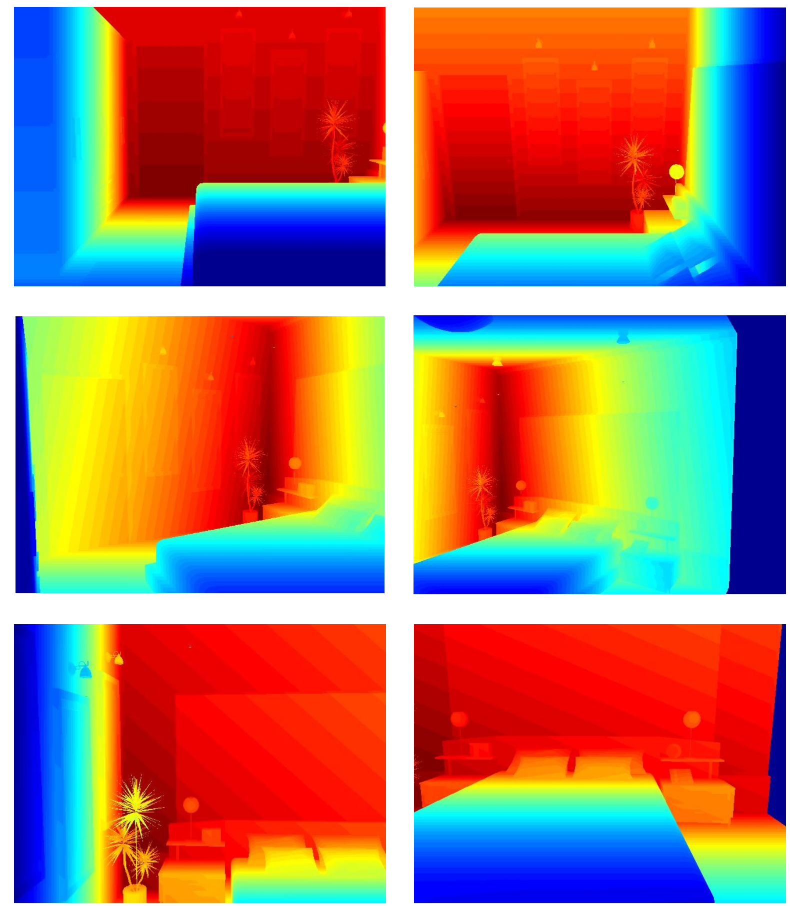
CG vs. Kinect



CG models
Kinect
AP
0.3
0.25
0.2
0.15
0.1
0.05
0
20
40
60
80
100
120
140
160
180
200
# of Exemplars

CG Rendering as Negatives

Whole-scene CG model
with no chair

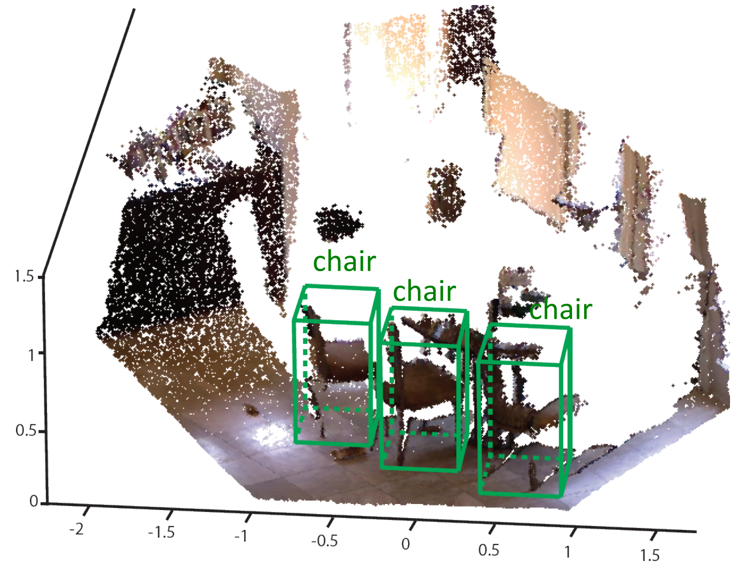


Conclusion

Usage of Kinect

Sliding Shapes

Input: Kinect Depth Map



Output: 3D Bounding Box

Code & Data Available
http://slidingshapes.cs.princeton.edu