Scene Understanding from RGB-D Images

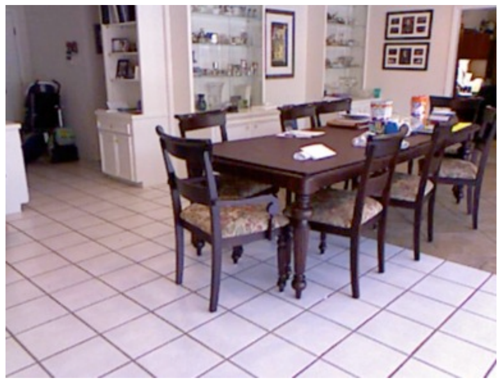
Object Detection, Semantic and Instance Segmentation
Pose Estimation

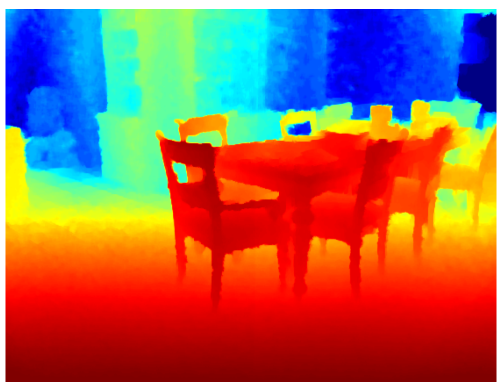
Saurabh Gupta, Ross Girshick, Pablo Arbeláez,
Jitendra Malik

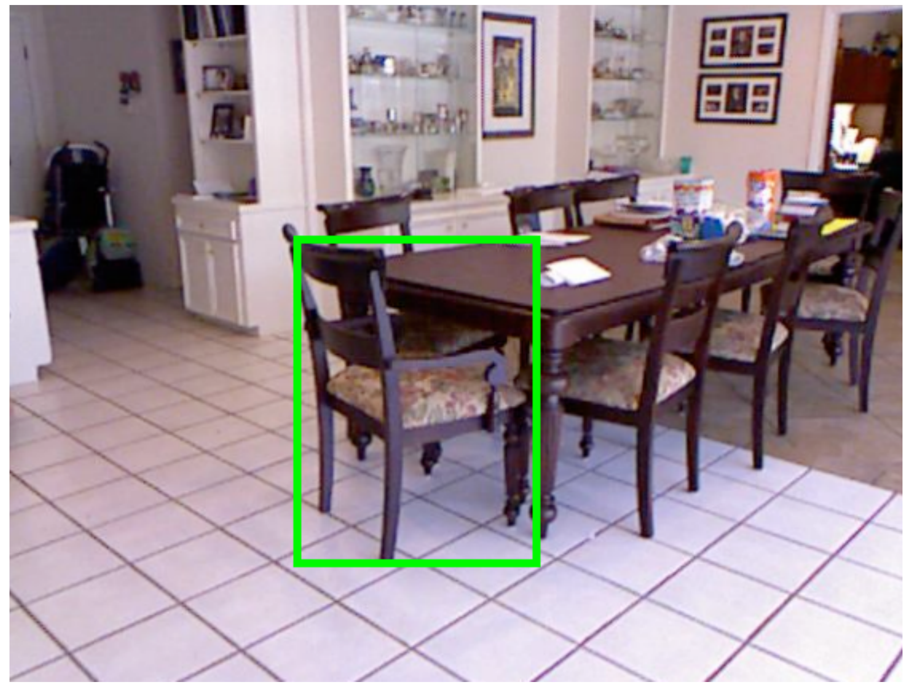
UC Berkeley

Scene Understanding

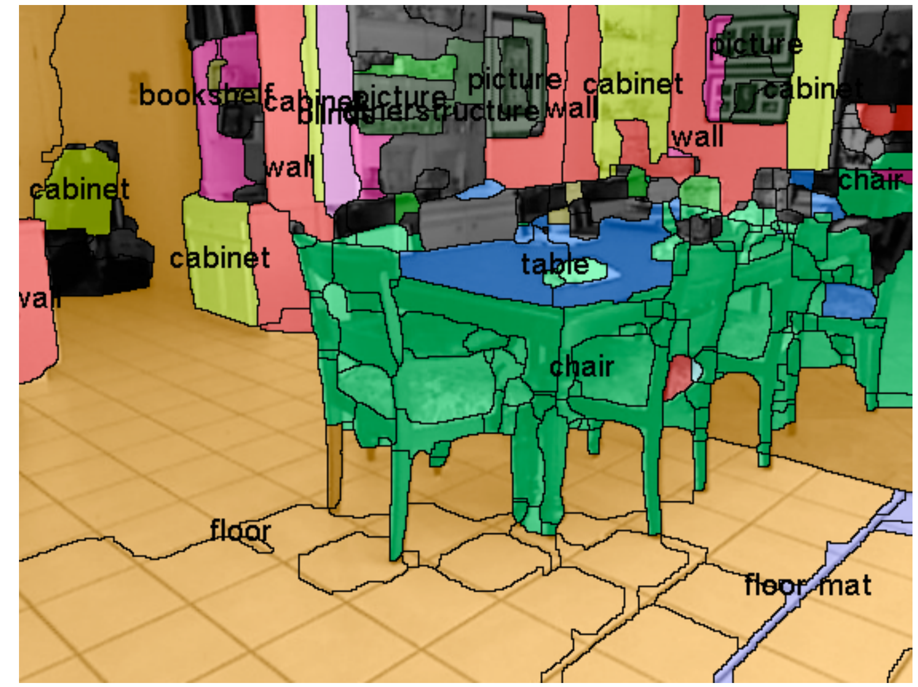
Motivation







Object Detection



Semantic Segm.

Good first steps

But we want to know much more

Instance Segmentation

Object Parsing

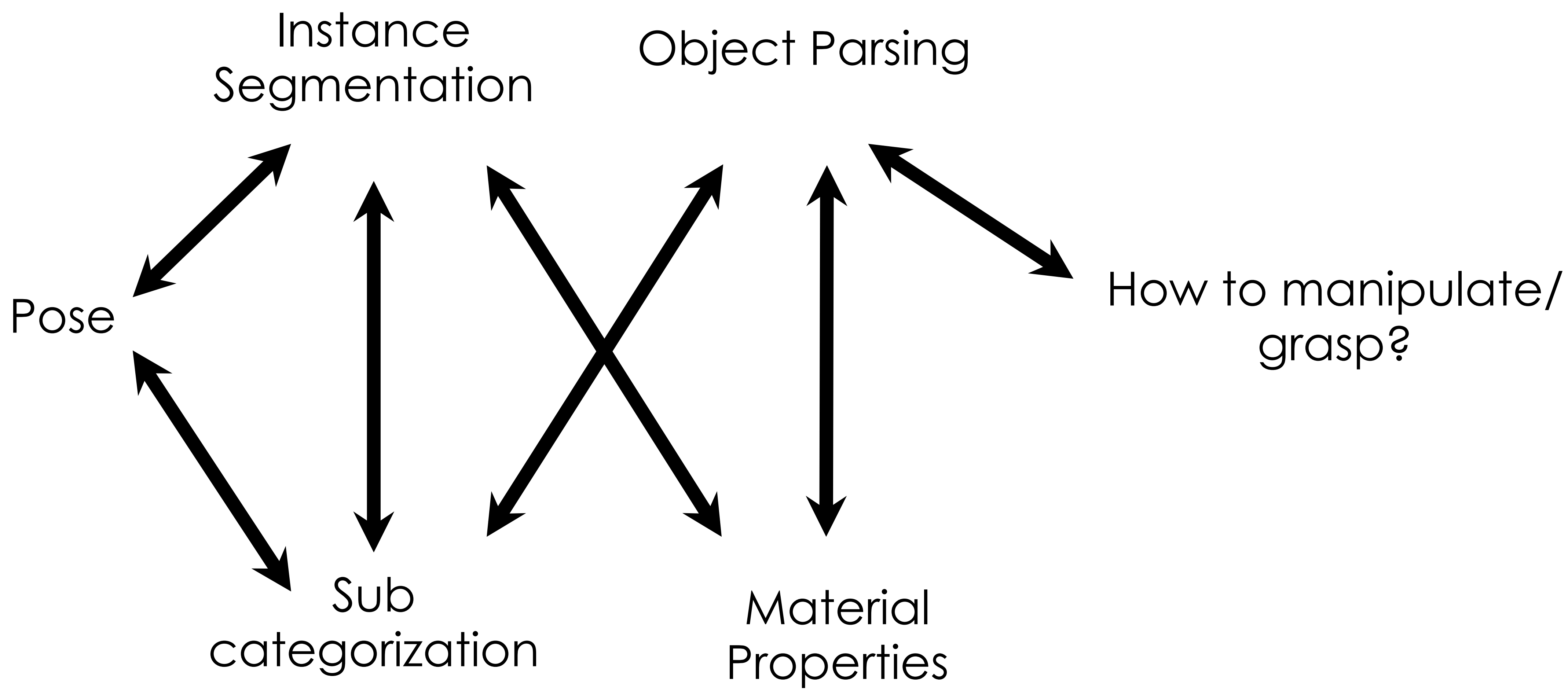
Sub categorization

How to manipulate/ grasp?

Material Properties

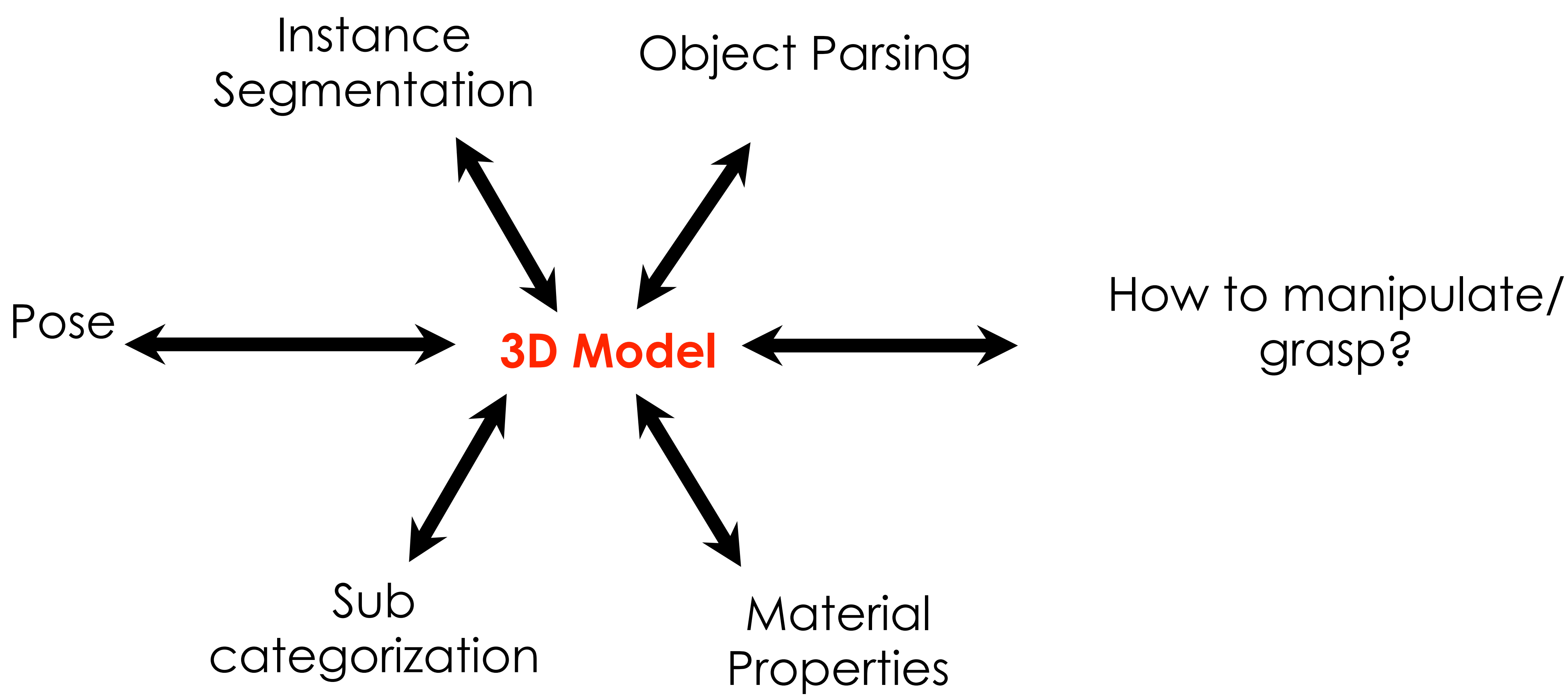
Pose

Detailed 3D Understanding



All these tasks are related, doing one will help the other

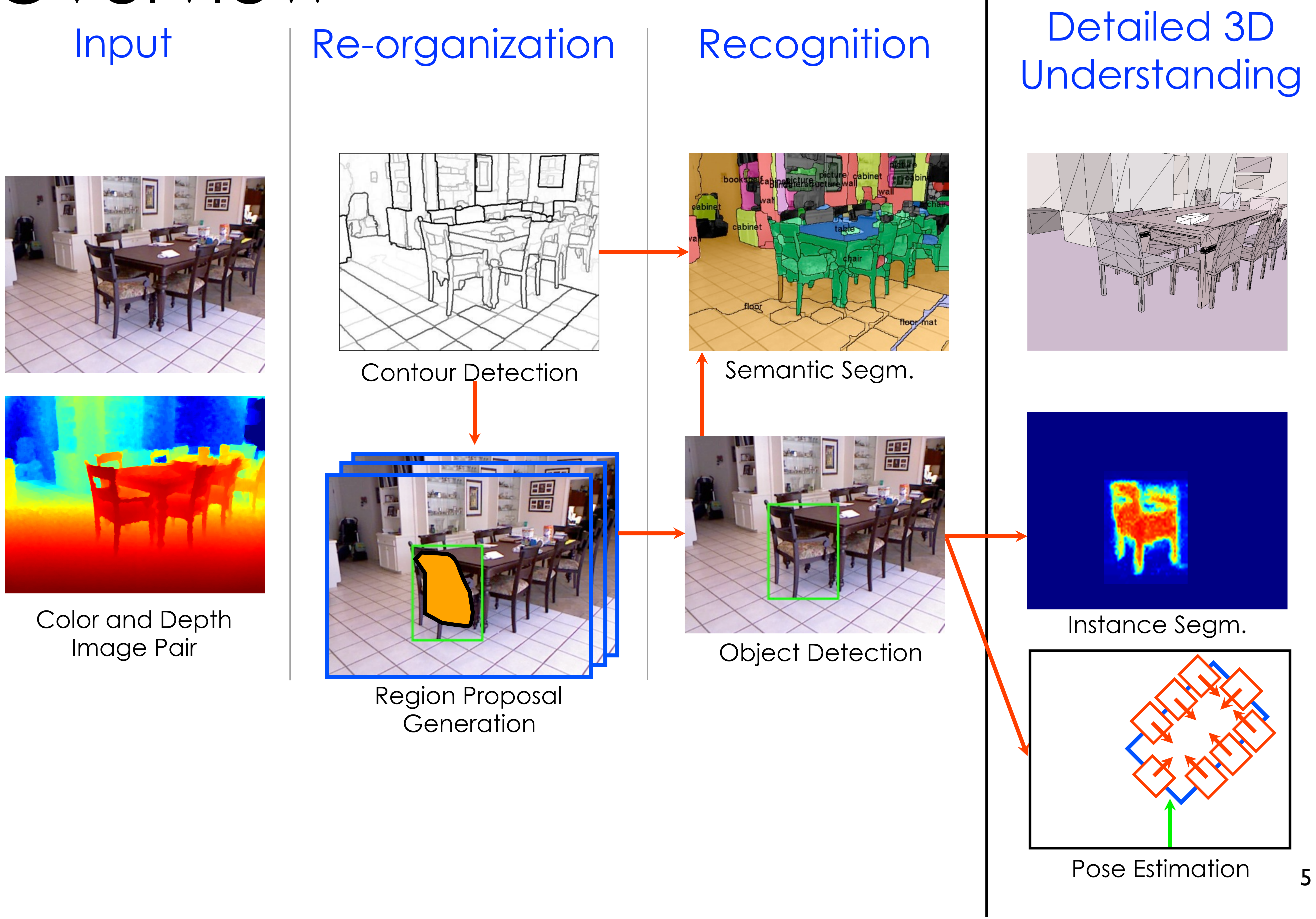
Detailed 3D Understanding



All these tasks are related, doing one will help the other

Estimating the 3D model explains all of these

Overview



Input
Re-organization
Recognition
Detailed 3D Understanding
Contour Detection
Semantic Segm.
Color and Depth Image Pair
Region Proposal Generation
Object Detection
Instance Segm.
Pose Estimation

Object Detection, Segmentation and Pose Estimation for RGB-D Images

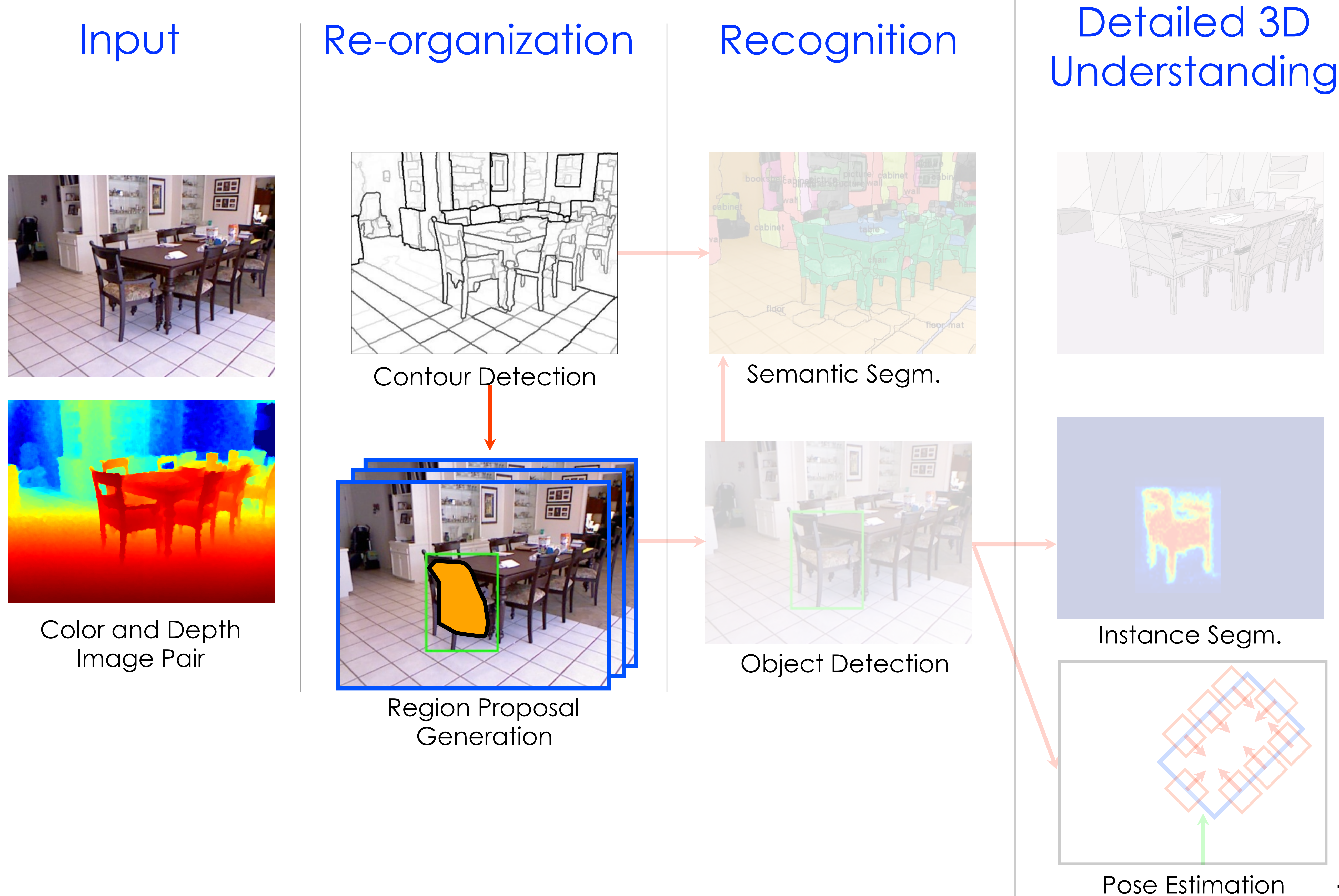
- S. Gupta, P. Arbeláez and J. Malik
  Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images,
  CVPR 2013 (oral)

- S. Gupta, R. B. Girshick, P. Arbeláez, and J. Malik
  Object Detection and Segmentation using Semantically Rich Image and Depth Features
  ECCV 2014

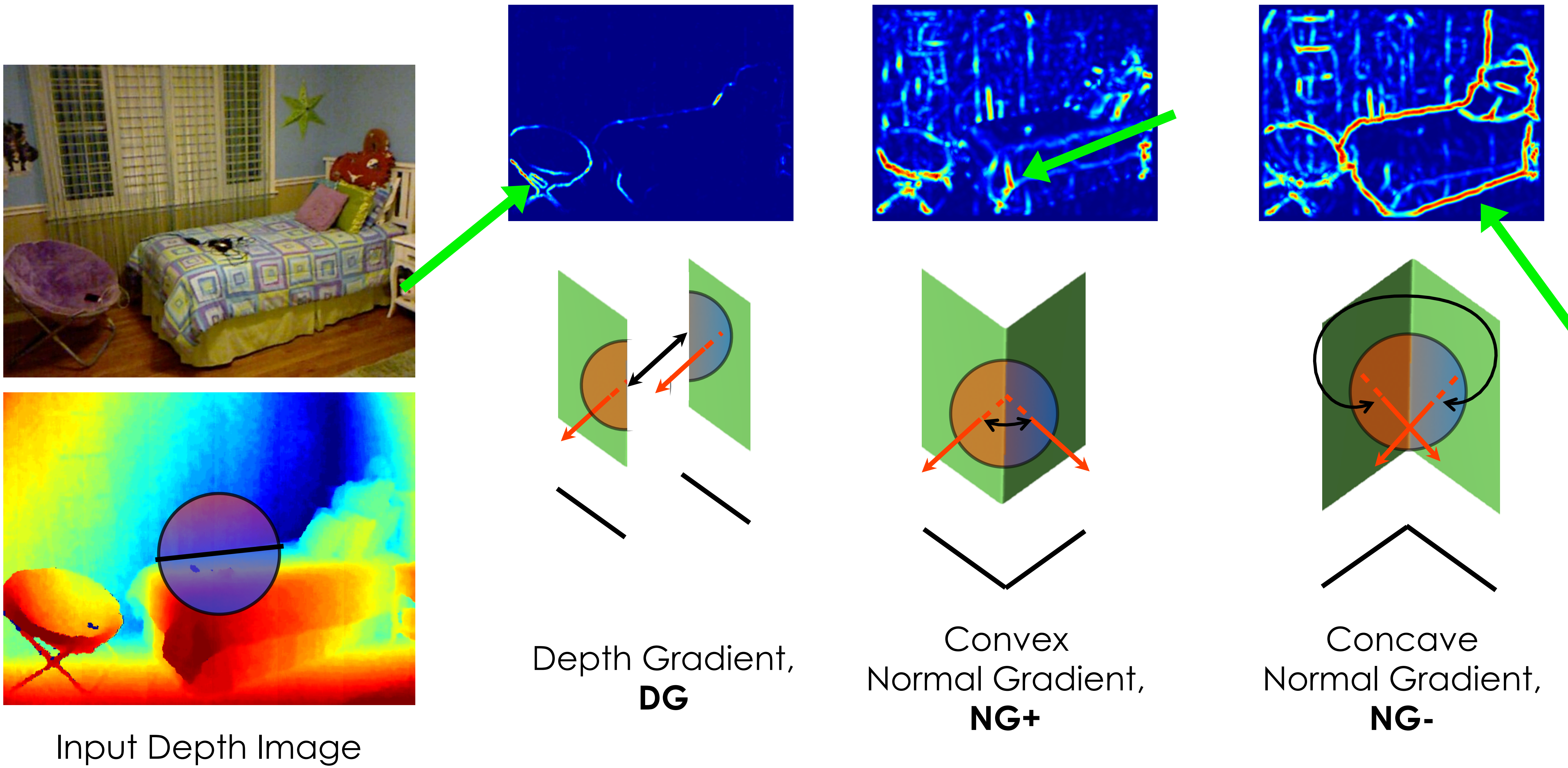
- S. Gupta, P. Arbeláez, R. B. Girshick, and J. Malik
  Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation
  IJCV 2014

- S. Gupta, P. Arbeláez, R. B. Girshick, and J. Malik
  Aligning 3D Models to RGB-D Images of Cluttered Scenes
  CVPR 2015, available on arXiv

Overview


Input
Re-organization
Recognition
Detailed 3D Understanding
Contour Detection
Semantic Segm.
Color and Depth
Image Pair
Region Proposal
Generation
Object Detection
Instance Segm.
Pose Estimation

Local Gradients on Depth Images



Multi-scale Local Gradients from Depth Images

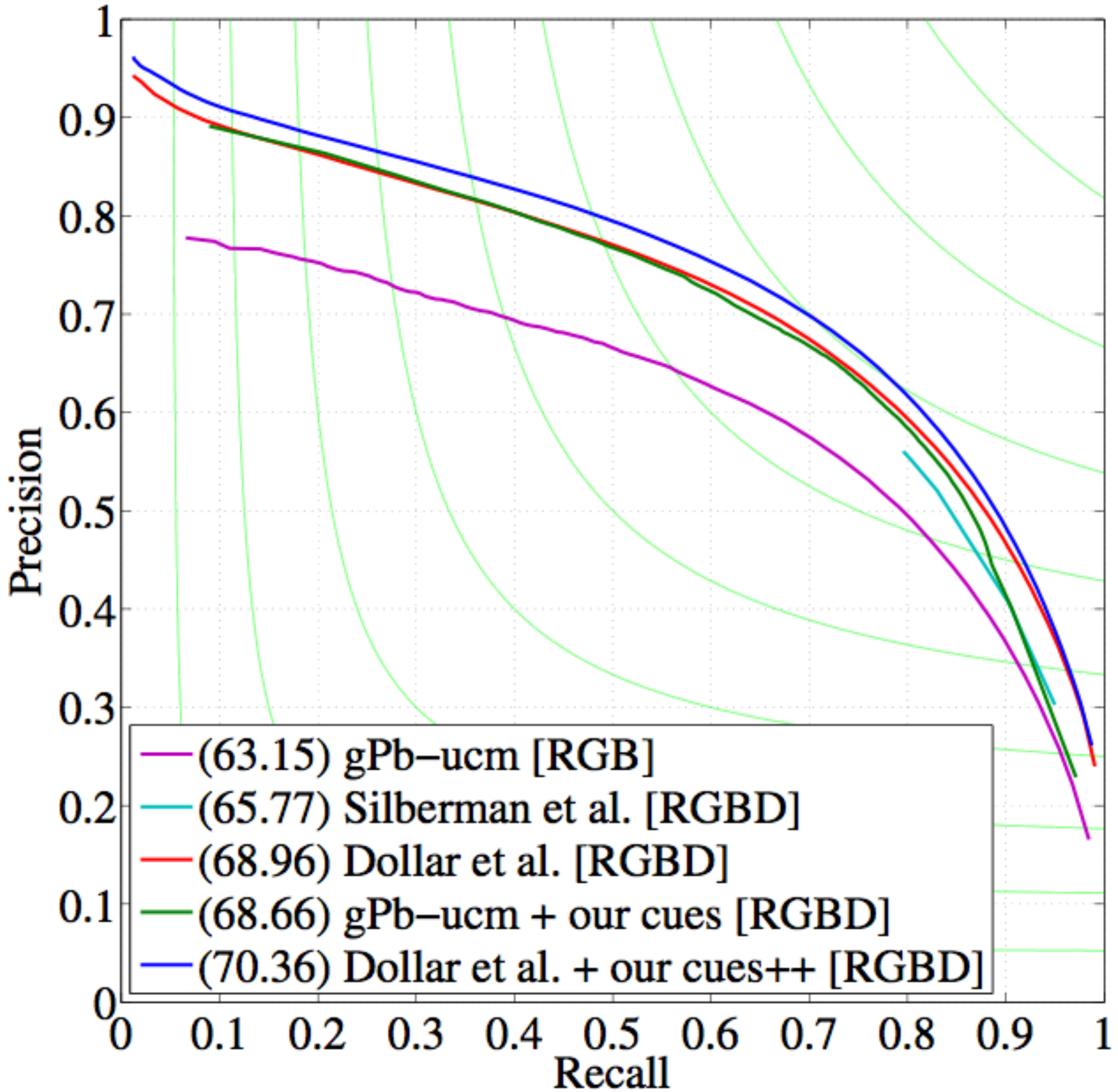
Important to differentiate between convex and concave normal gradients

Using Local Gradients for Contour Detection

Use with gPb-UCM

Use with Dollar et al.'s structured edges

Method		max F
gPb-UCM	RGB	63.15
Silberman et al.	RGB-D	65.77
Dollar et al.	RGB-D	68.96
Our **(gPb-UCM + our cues)**	RGB-D	68.66
Our **(Dollar et al. + our cues++)**	RGB-D	**70.36**



Arbeláez et al. Contour Detection and Hierarchical Image Segmentation, PAMI 2011

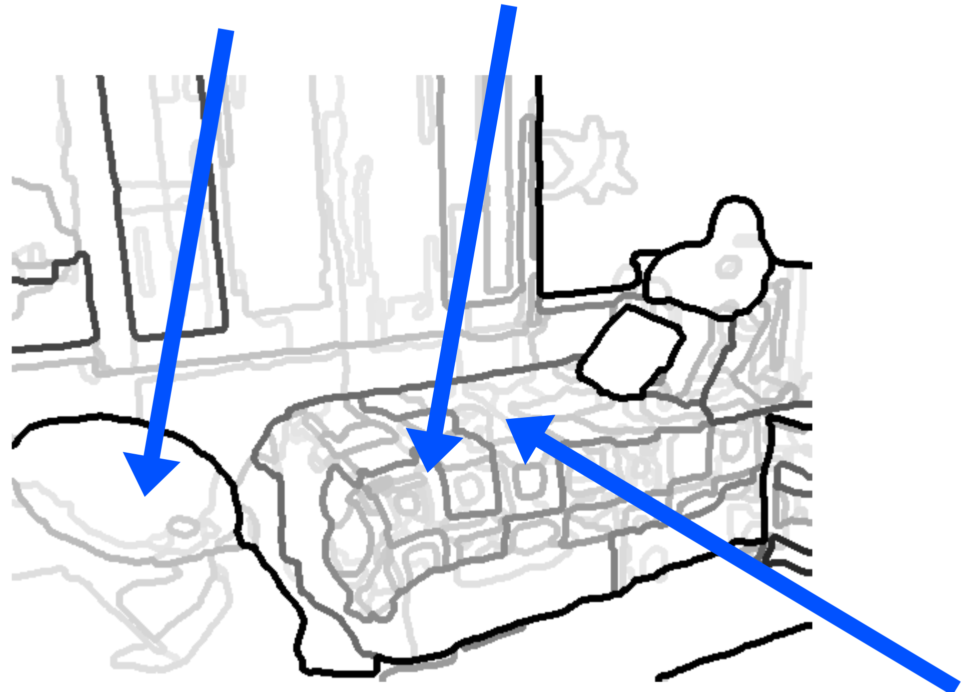
P. Dollar and L. Zitnick Structured Forests for fast edge detection, ICCV 2013

S. Gupta, P Arbeláez, J. Malik Perceptual Organization and Recognition in Indoor RGB-D Images, CVPR 2013

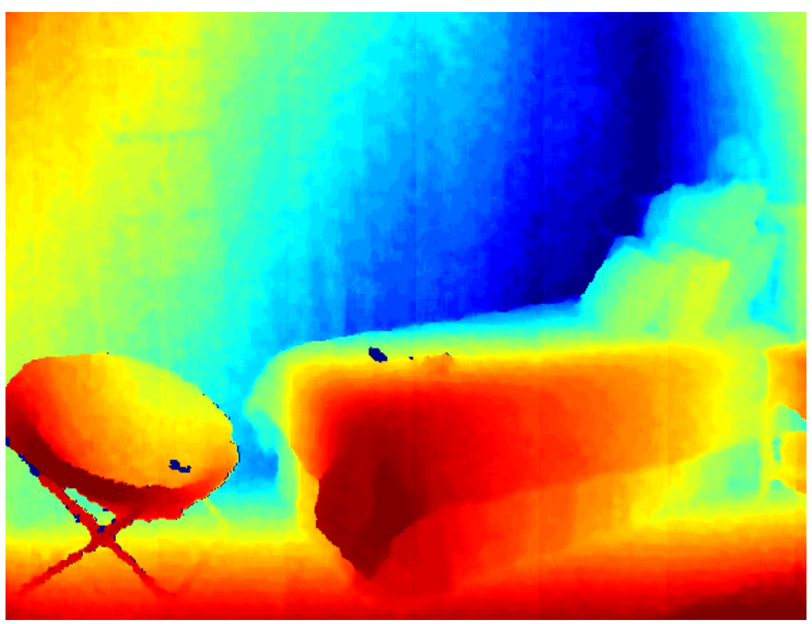
S. Gupta, R. Girshick, P Arbeláez, J. Malik , Object Detection and Segmentation using Semantically Rich Image and Depth Features, ECCV 2014

Results

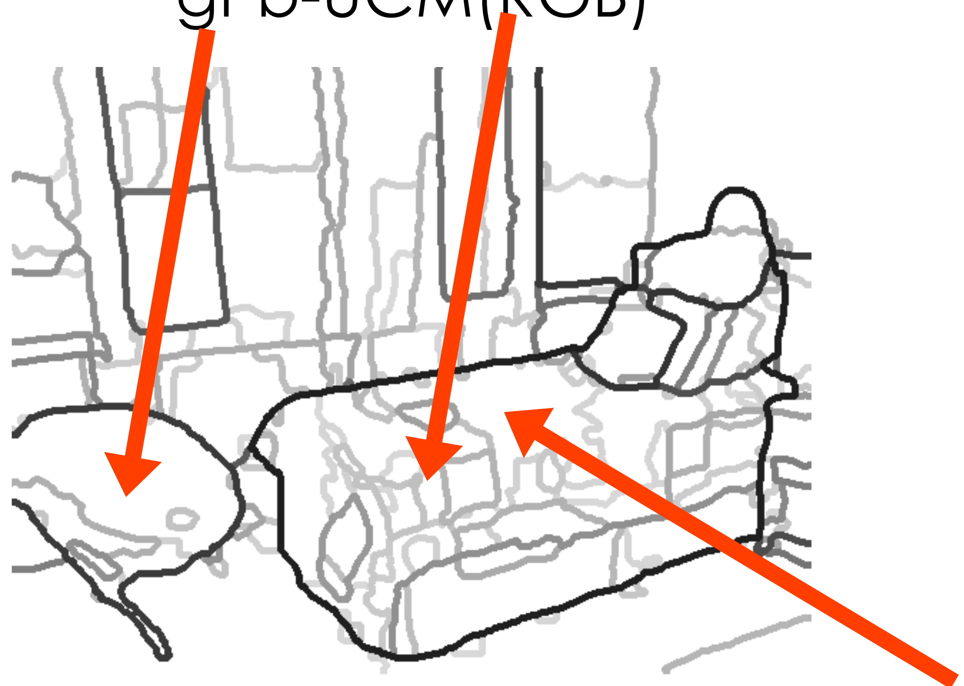
RGB



gPb-UCM(RGB)



D



This Work (RGB-D)

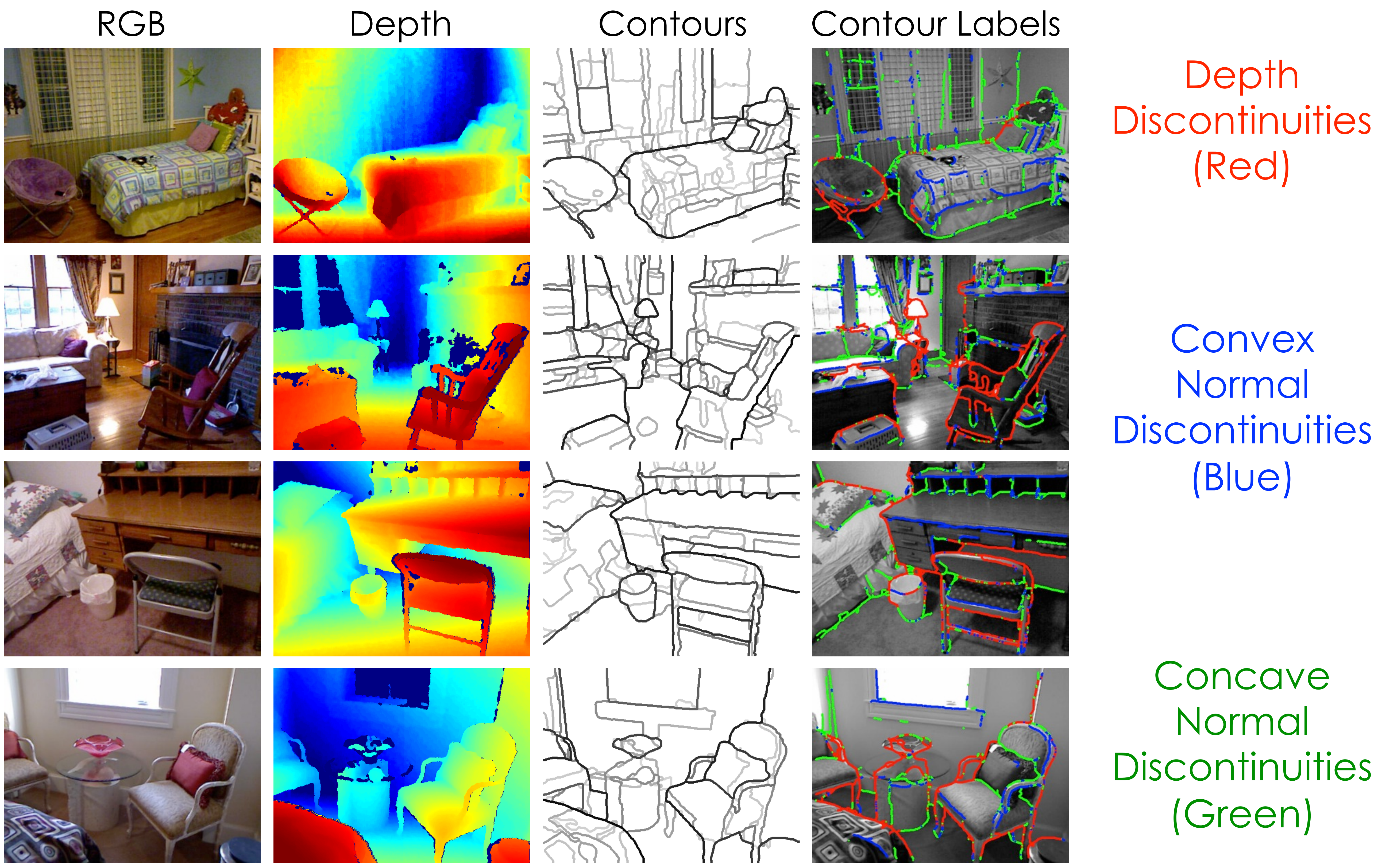
Less distracted by albedo

Higher Recall

Higher Precision

More Complete Objects

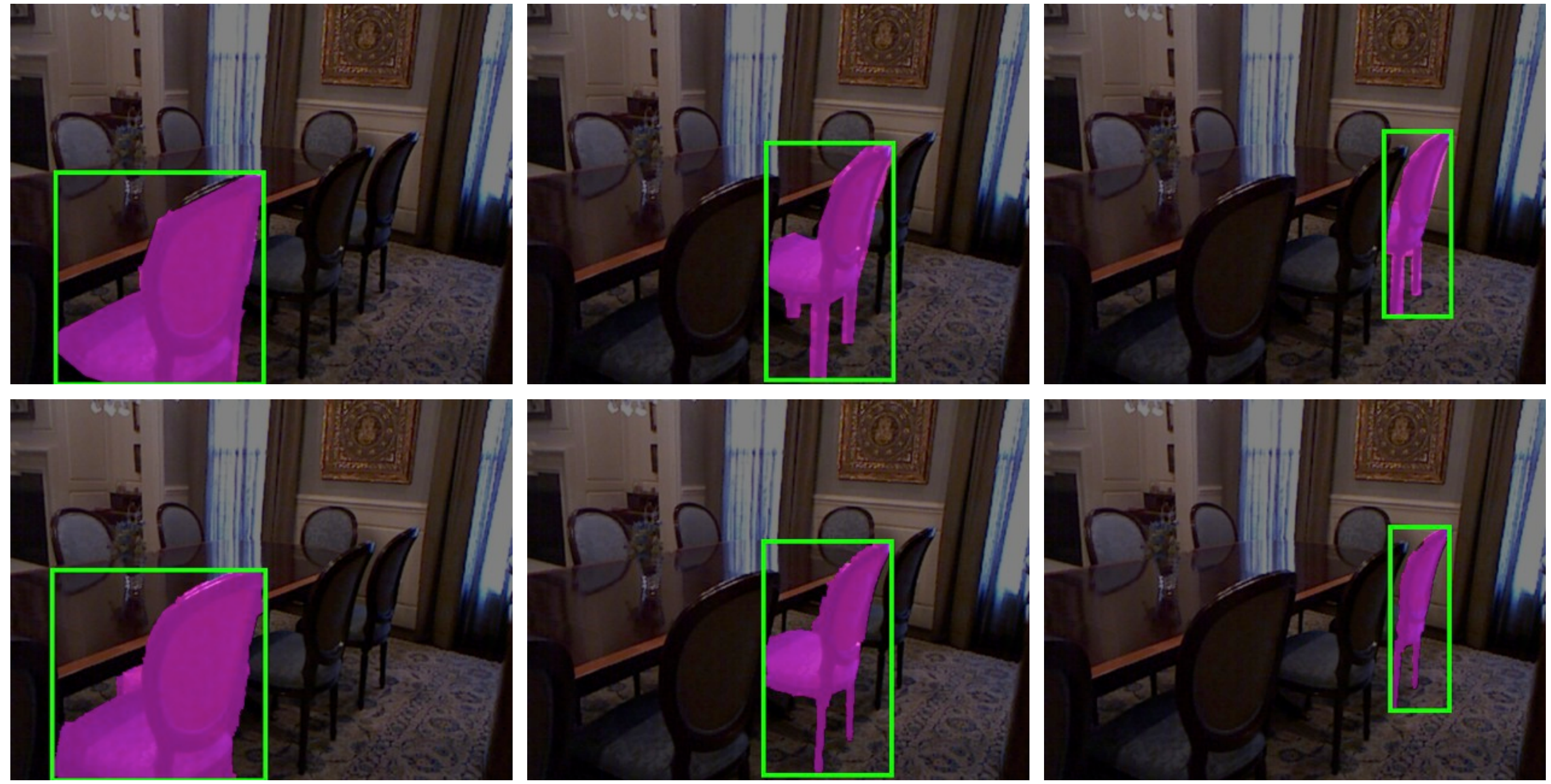
Results

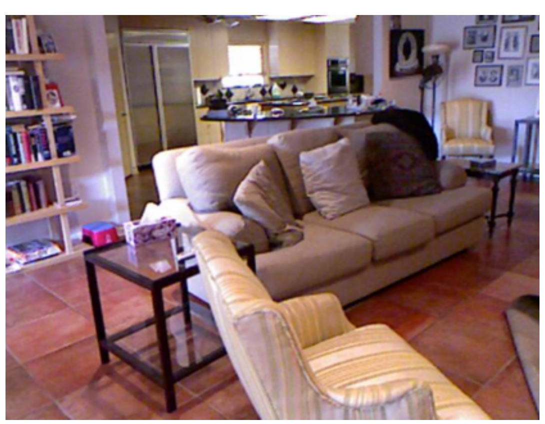

RGB
Depth
Contours
Contour Labels
Depth Discontinuities (Red)
Convex Normal Discontinuities (Blue)
Concave Normal Discontinuities (Green)

Examples

GT Mask

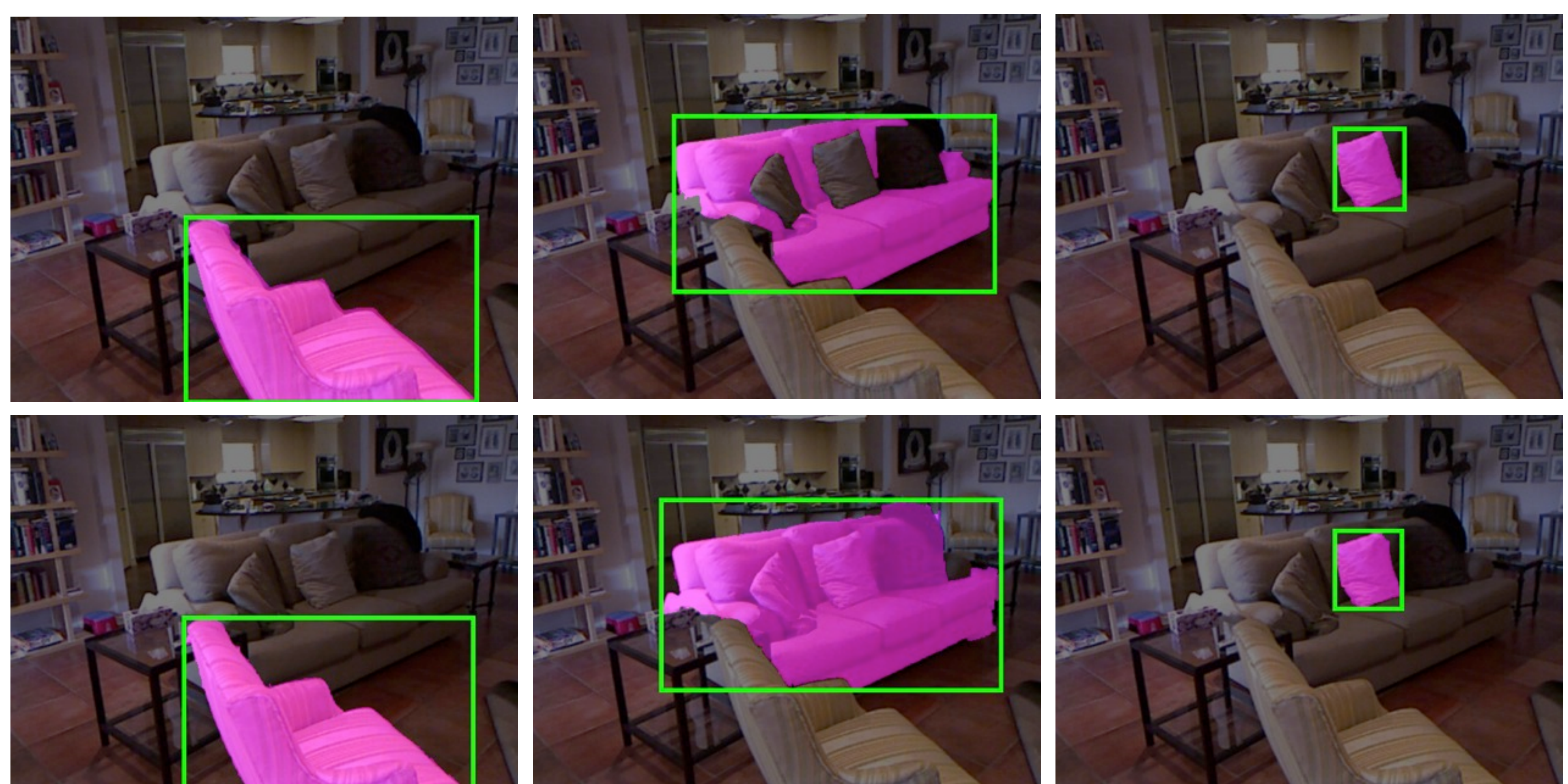
Best Proposal
@500





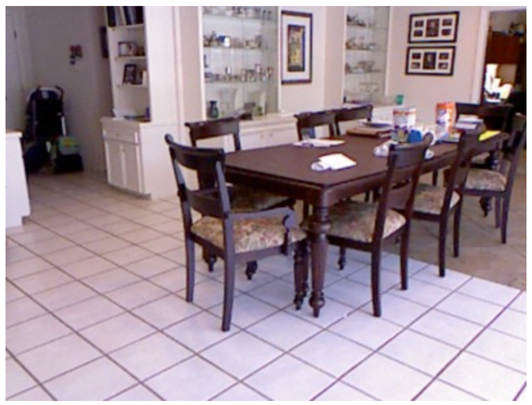
GT Mask

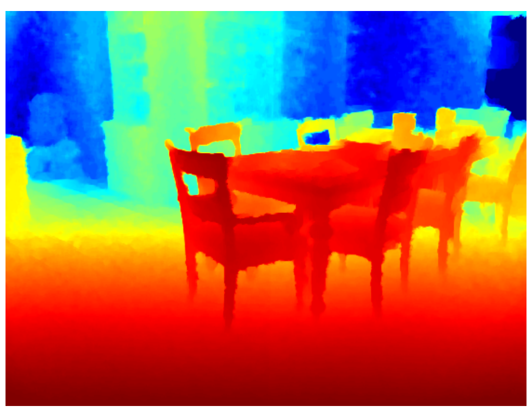
Best Proposal
@500



Overview

Input





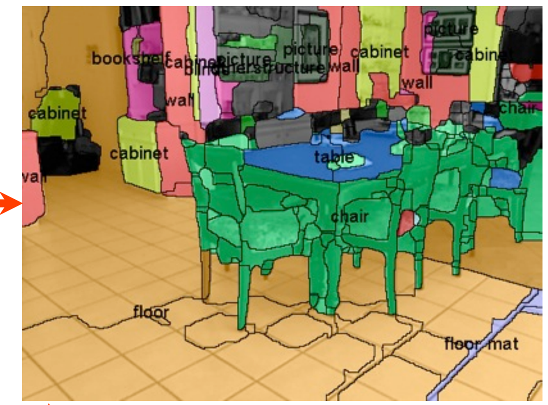
Color and Depth
Image Pair

Re-organization

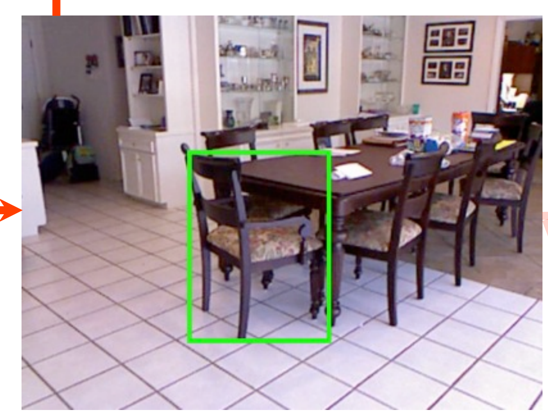
Contour Detection

Region Proposal
Generation

Recognition



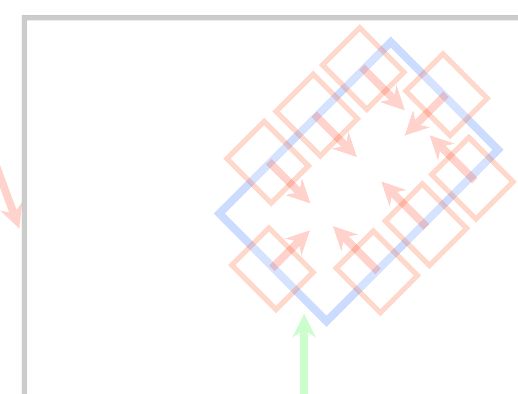
Semantic Segm.



Object Detection

Detailed 3D Understanding

Instance Segm.



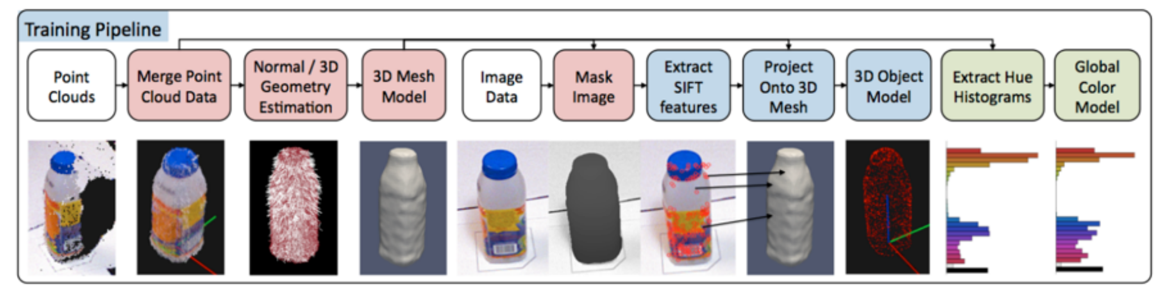
Pose Estimation

Object Detection

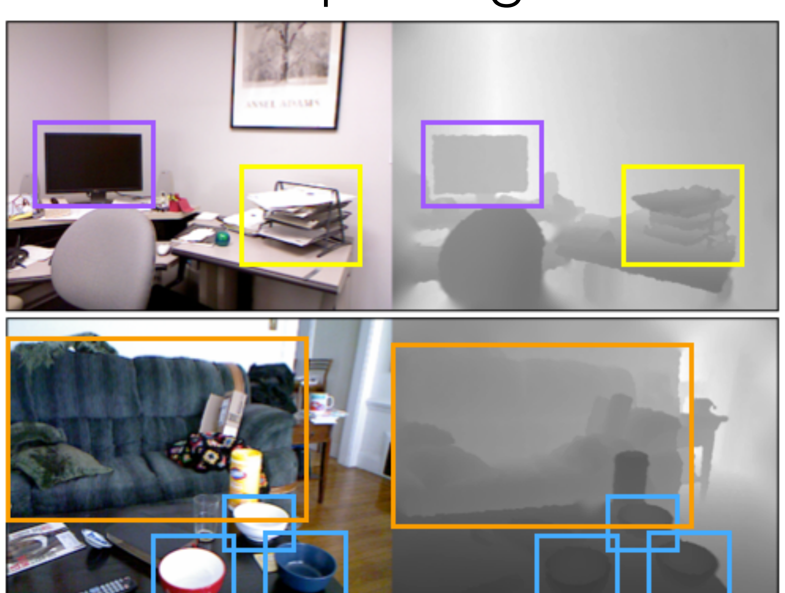
Related Work [RGB-D, Robotics]

Lai et al. ICRA 2011, A Large-Scale Hierarchical Multi-View RGB-D Object Dataset: RGB-D DPM, but instances and small table-top objects

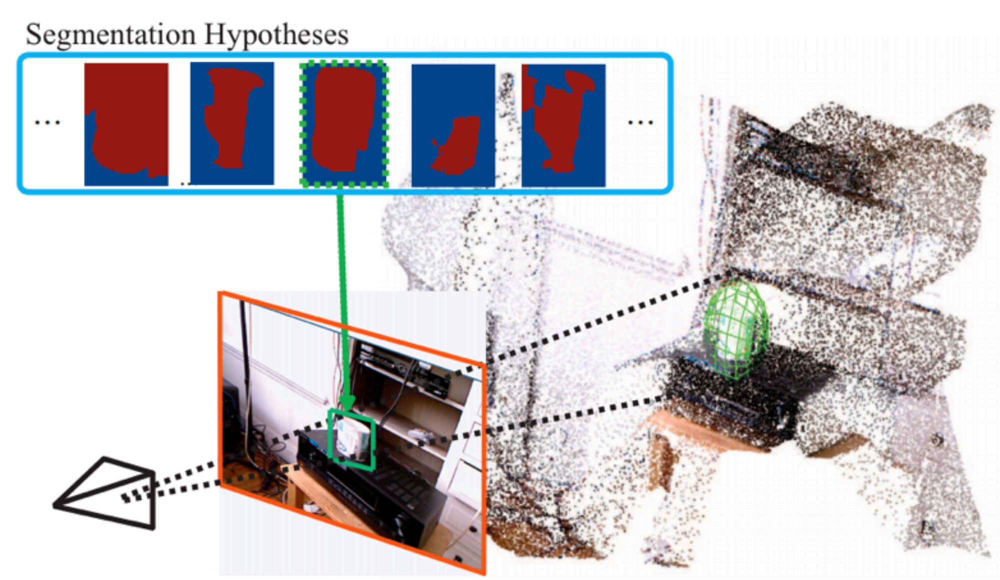
Tang et al. ICRA 2012, A Textured Object Recognition Pipeline for Color and Depth Image Data: Appearance matching, geometric verification



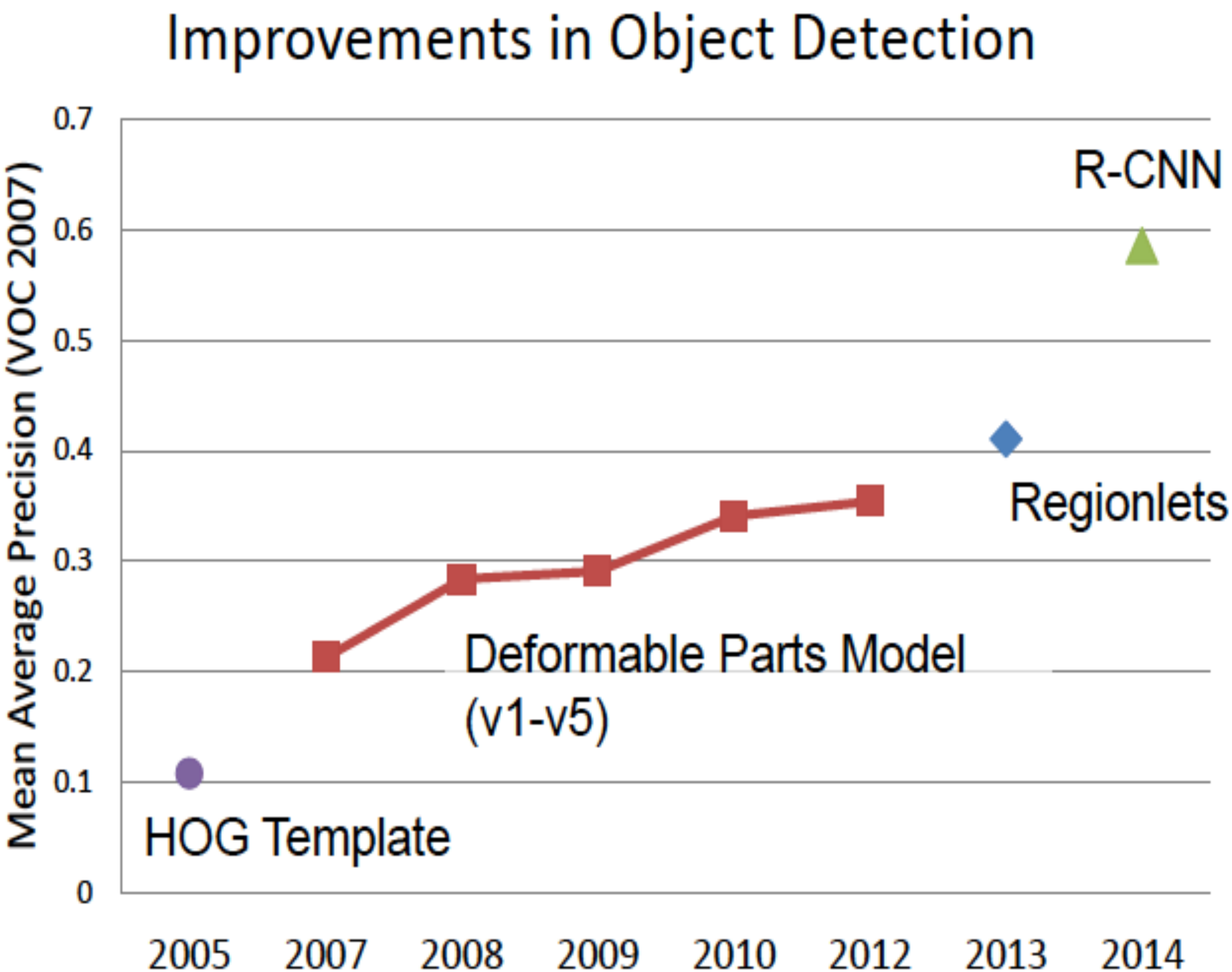
Janoch et al. ICCV-W 2011, A Category-Level 3-D Object Dataset: Putting the Kinect to Work, Absolute size based pruning and re-scoring with DPMs



Kim et al. CVPR 2013, Accurate Localization of 3D Objects from RGB-D Data using Segmentation Hypotheses, Extension to DPMs to model deformations in 3D



State of the Art in RGB Recognition

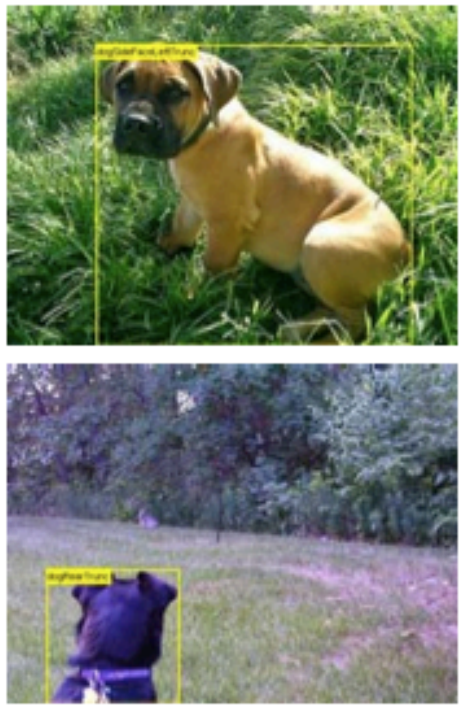


(Slide from D. Hoiem)

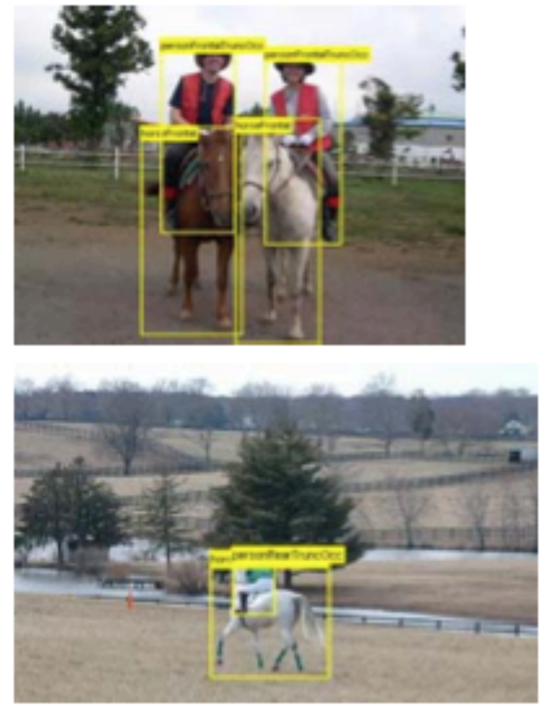
PASCAL Visual Object Challenge (Everingham et al)

Dining Table

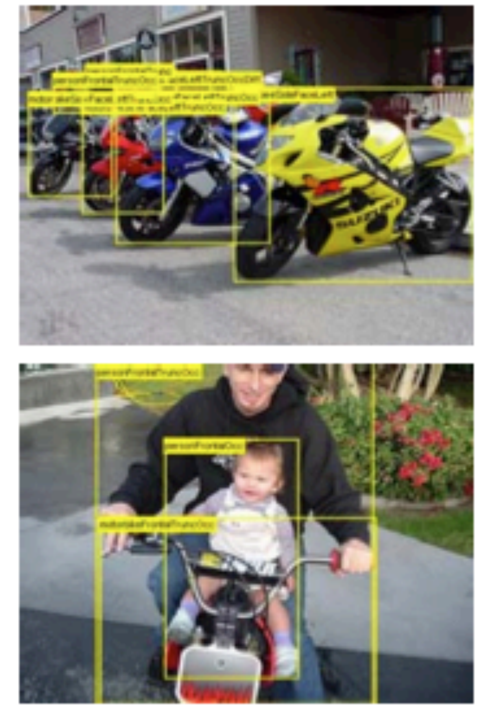
Dog



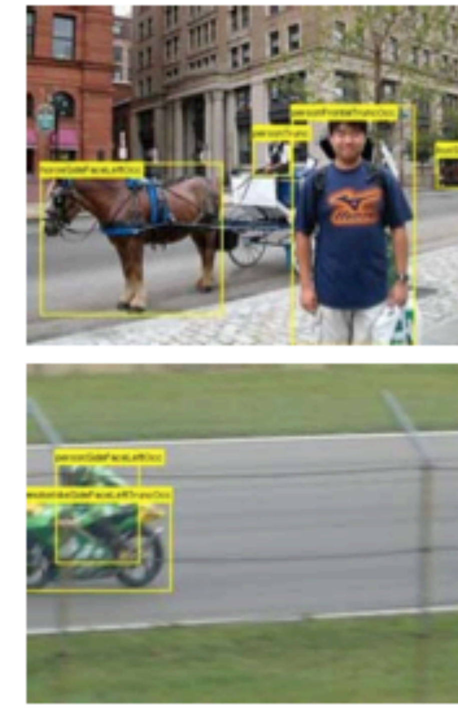
Horse



Motorbike

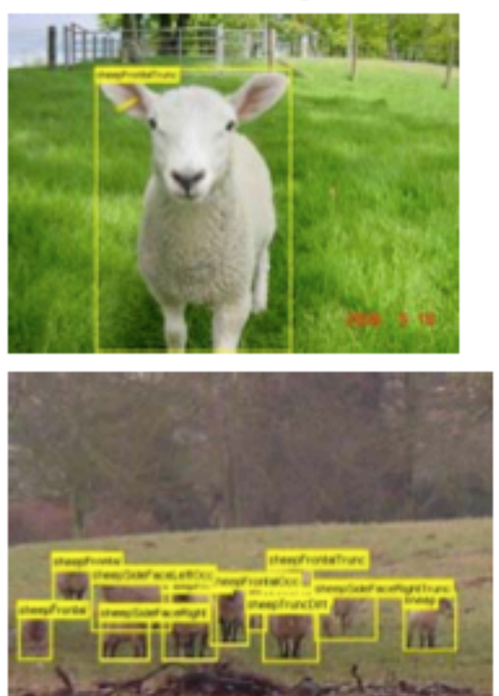


Person

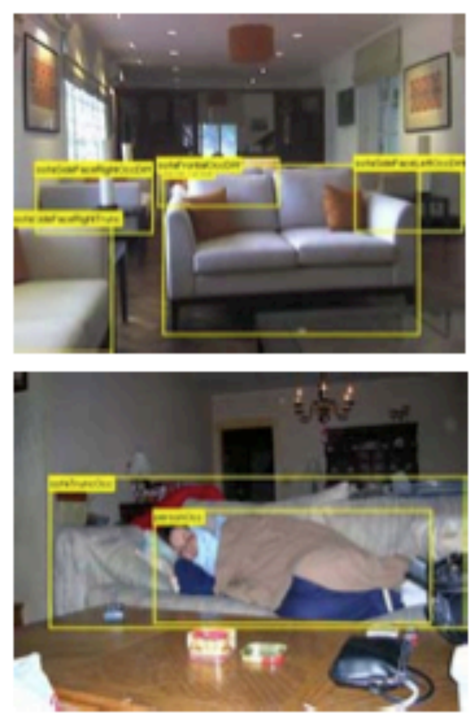


Potted Plant

Sheep



Sofa

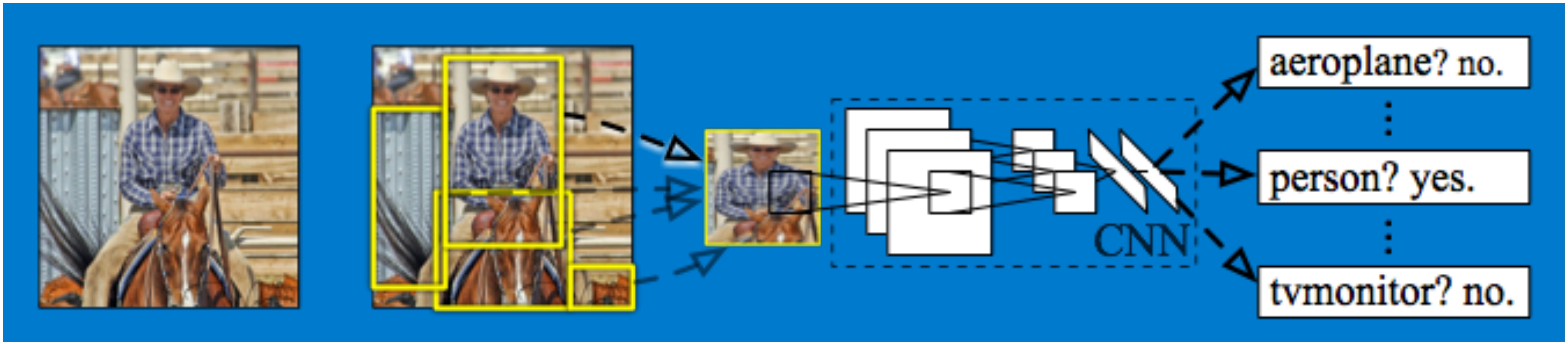


Train

TV/Monitor

R-CNN: Regions with CNN features

Girshick, Donahue, Darrell & Malik (CVPR 2014)



Input image

Extract region proposals (~2k / image)

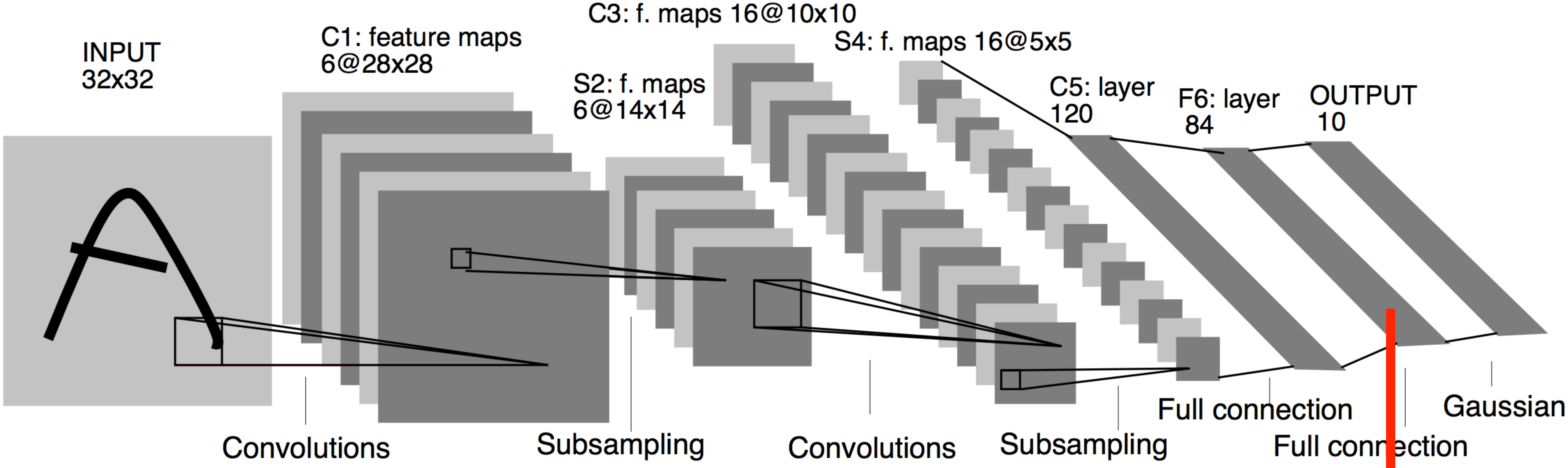
Compute CNN features

Classify regions (linear SVM)

CNN features are inspired by the architecture of the visual system

CNN Features ?

Convolutional Neural Network



Train on a large dataset

How to learn features for RGB-D Images ??

Generic representation useful for for a variety of tasks

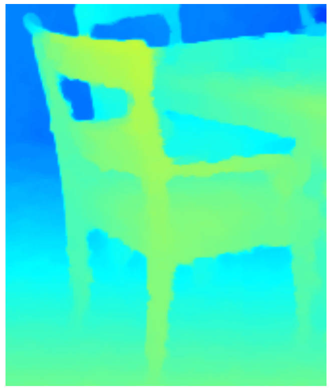
LeCun et al., Backpropagation applied to handwritten zip code recognition. Neural Computation (1989)
Krizhevsky et al., ImageNet classification with deep convolutional neural networks. In NIPS (2012)

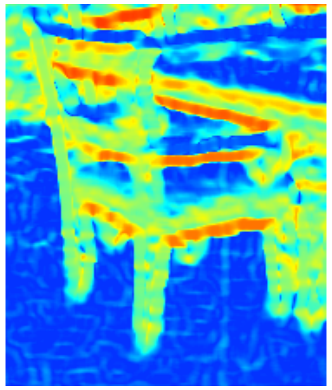
Object Detection in RGB-D images

Key Insights

Depth Images are **image-like enough** to use Convolutional Neural Network models

Geocentric embedding into *Horizontal Disparity, Height Above Ground,* and *Angle with Gravity* ***(HHA)*** works better than just raw disparity





Synthetic depth data can help

Object Detection

Test Set

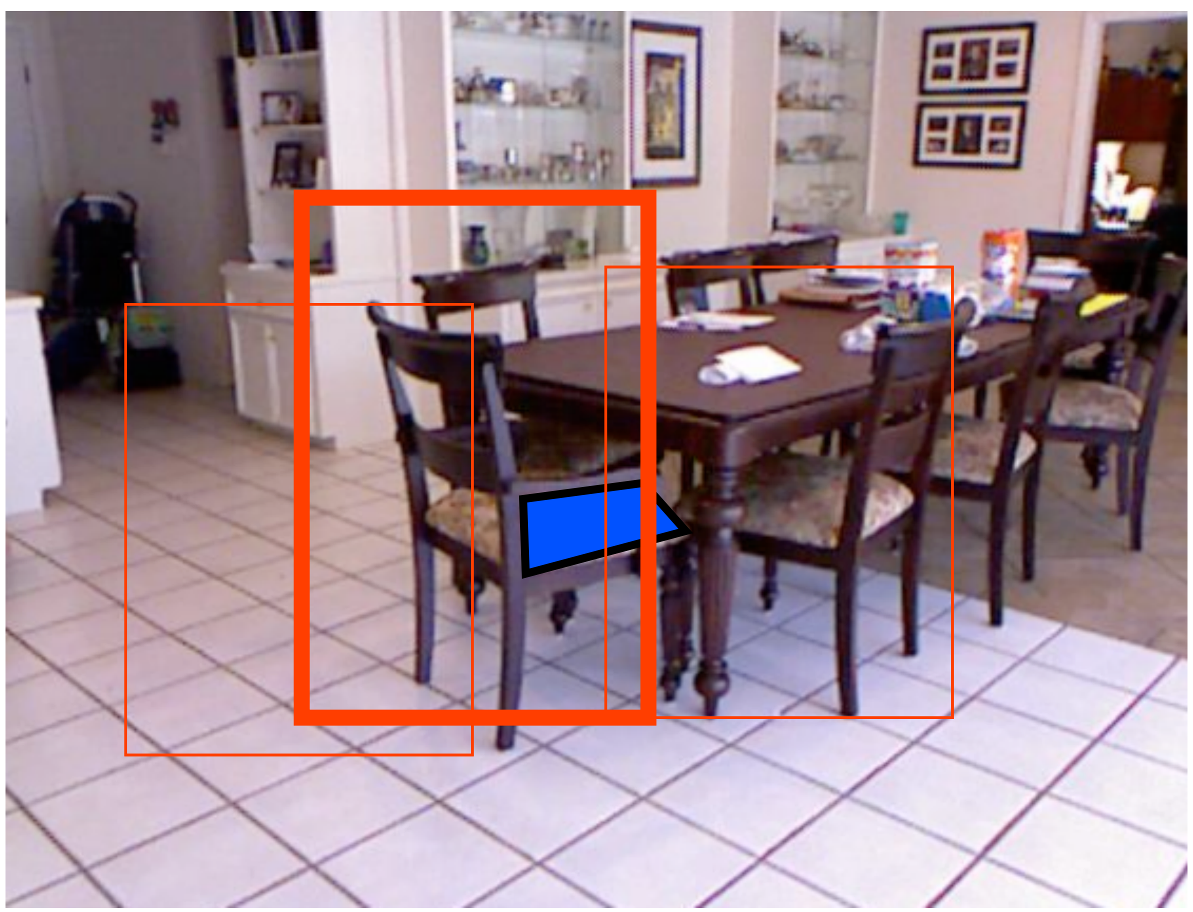
	mean	bath tub	bed	book shelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	television	toilet
RGB DPM	9	1	28	9	0	8	7	1	3	1	7	22	10	9	4	6	9	6	6	34
RGBD DPM	24	19	56	18	**1**	24	24	6	10	16	27	27	35	33	21	23	34	17	20	45
RGB RCNN	22	17	45	28	**1**	26	30	10	16	19	16	28	32	17	11	17	29	13	27	44
Our	**39**	**36**	**71**	**35**	**4**	**47**	**47**	**15**	**23**	**39**	**44**	**38**	**53**	**41**	**42**	**44**	**52**	**22**	**38**	**48**

Object Detection

For Semantic Segmentation

Use output from object detectors to compute **additional features** for superpixels

Feature Computation



1. Highest scoring detection
2. Use as features for the superpixel
   - detection score
   - overlap
   - difference in mean depth of superpixel and detection
   - non-linear combinations

Object Detection

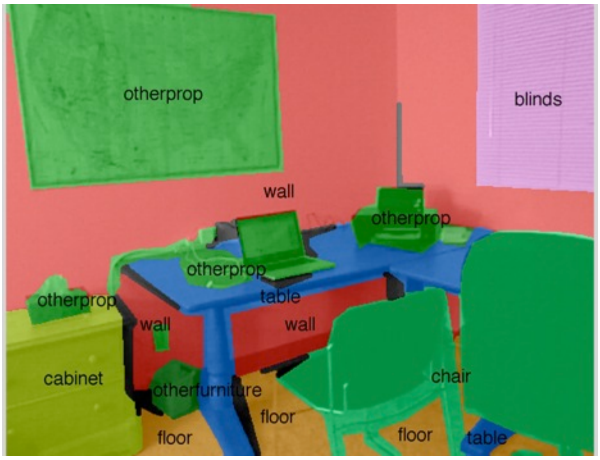
For Semantic Segmentation (Performance)

40 Class Task

Scene Surfaces - Floors, walls, ceiling, windows, doors, ...

Furniture - Beds, chairs, sofa, table, desks, ...

Objects - Pillow, books, bottles, ...

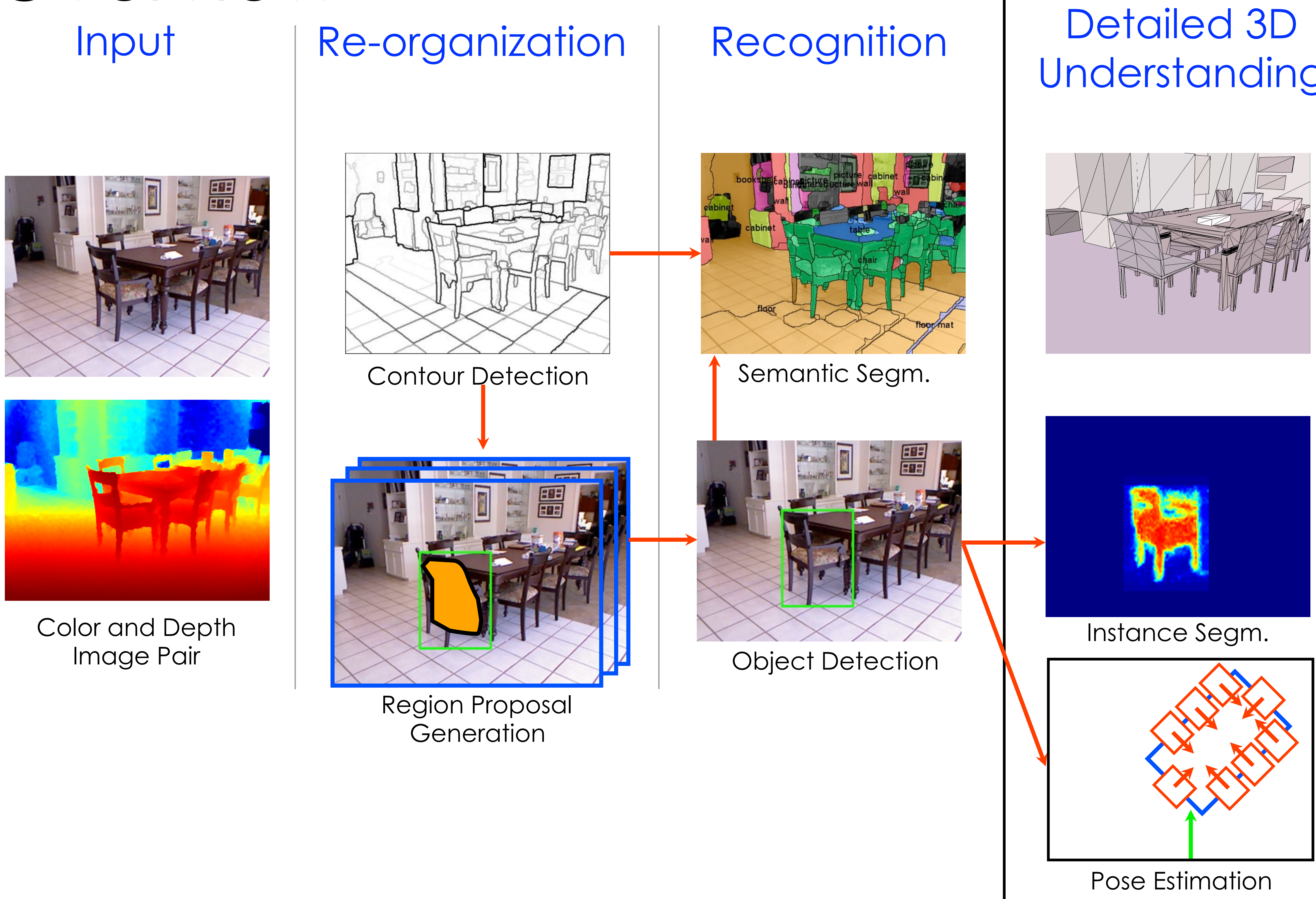


Ground Truth 40 Class

	Silberman et al. ECCV 12	Ren et al. CVPR 12	Gupta et al. CVPR 13	Gupta et al. (13) + RGB-D DPM	Gupta et al. (13) + Our Obj Det.
fwavacc	38.2	37.6	43.4	45.2	**47**
avacc	19	20.5	24.3	27.3	**28.6**
mean (maxIU)	-	21.4	27.9	29.6	**31.3**
pixacc	54.6	49.3	57.9	59	**60.3**
obj avg	18.4	21.1	26.4	31.1	**35.1**

Silberman et al., ECCV12, Indoor segmentation and support inference from RGBD images.
Ren et al., CVPR12, RGB-(D) scene labeling: Features and algorithms
Gupta et al., CVPR13, Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images.

Overview



Input
Re-organization
Recognition
Detailed 3D Understanding
Color and Depth Image Pair
Contour Detection
Region Proposal Generation
Semantic Segm.
Object Detection
Instance Segm.
Pose Estimation

Instance Segmentation

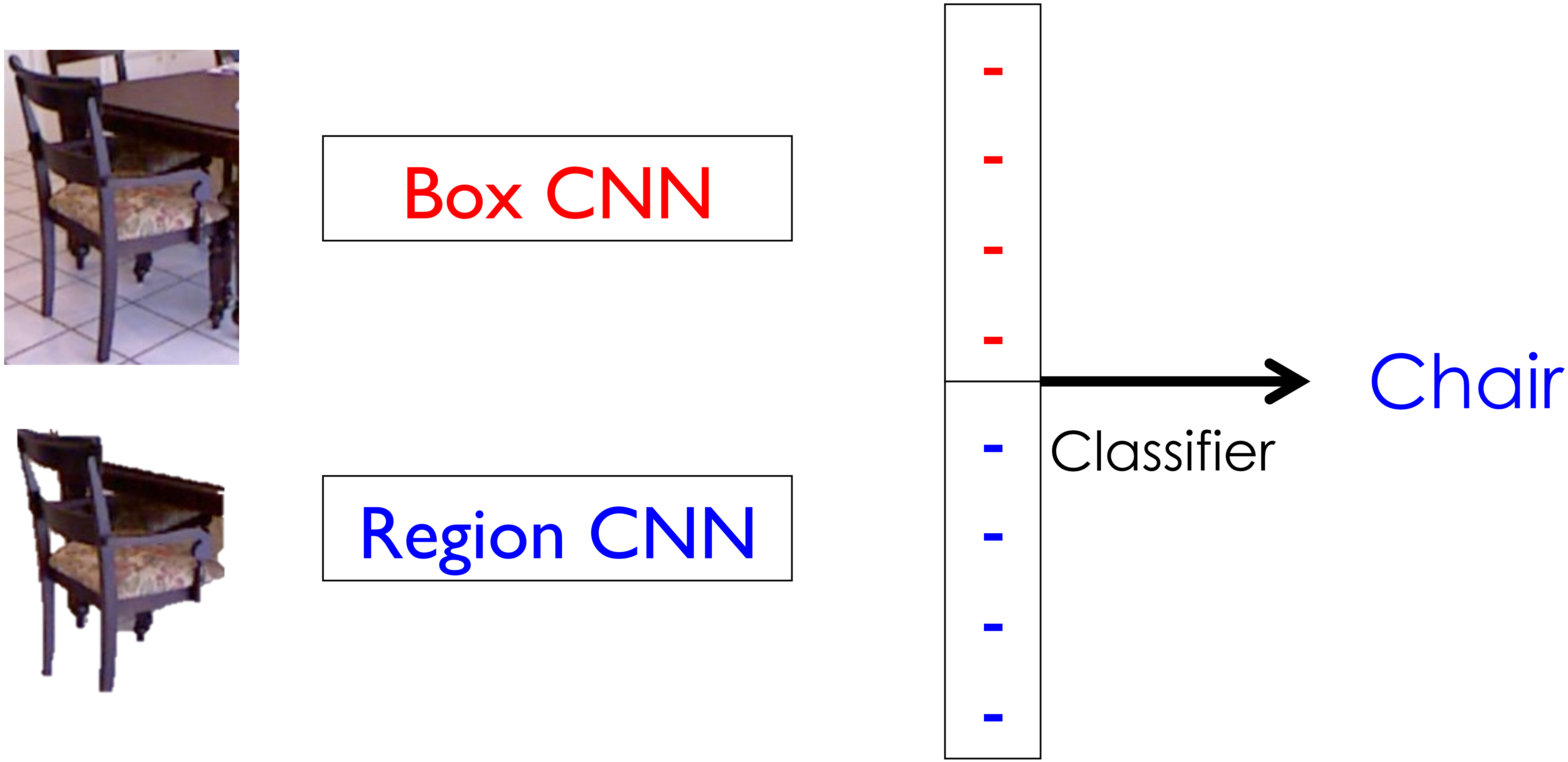
Task

Detect and segment objects

Hariharan et al., ECCV14, Simultaneous Localization and Detection

Instance Segmentation

Method



Instance Segmentation

Instance Segmentation

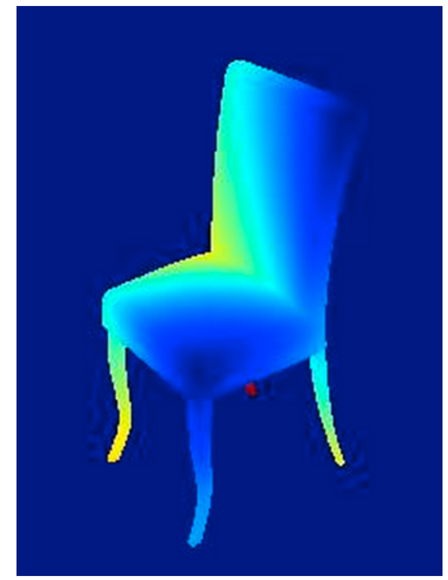
For Semantic Segmentation (Performance)

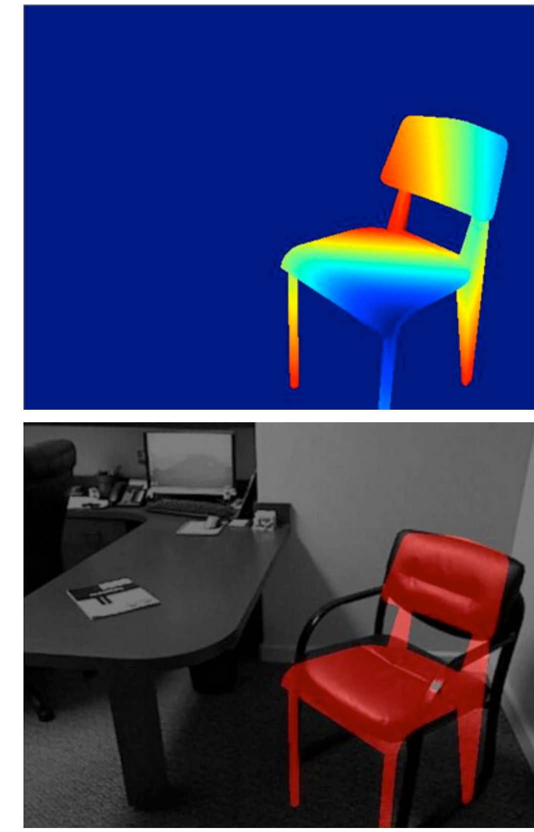
40 class task

	Silberman et al. ECCV 12	Ren et al. CVPR 12	Gupta et al. CVPR 13	Gupta et al. (13) + RGB-D DPM	Gupta et al. (13) + Our Obj Det.	+ Instance Segm.
fwavacc	38.2	37.6	43.4	45.2	47	**47.74**
avacc	19	20.5	24.3	27.3	28.6	**29.71**
mean (maxIU)	-	21.4	27.9	29.6	31.3	**32.90**
pixacc	54.6	49.3	57.9	59	60.3	**62.24**
obj avg	18.4	21.1	26.4	31.1	35.1	**37.50**

Pose Estimation

[ECCV 14]





Input

Instance Segmentation

Estimate Coarse Pose

Align to data

3D reasoning by initial 2D processing and then 'lifting' to 3D

Learning from synthetic data and generalizing to real data

Starting with weak annotation (instance segmentation) able to produce a much richer output

3 layer CNN on **normal images** trained on **synthetic** data

Search over **scale, placement** and **sub-type** to minimize re-projection error

[ECCV 14] Learning rich features from RGB-D images for object detection and segmentation. In ECCV 14.

Coarse Pose Estimation

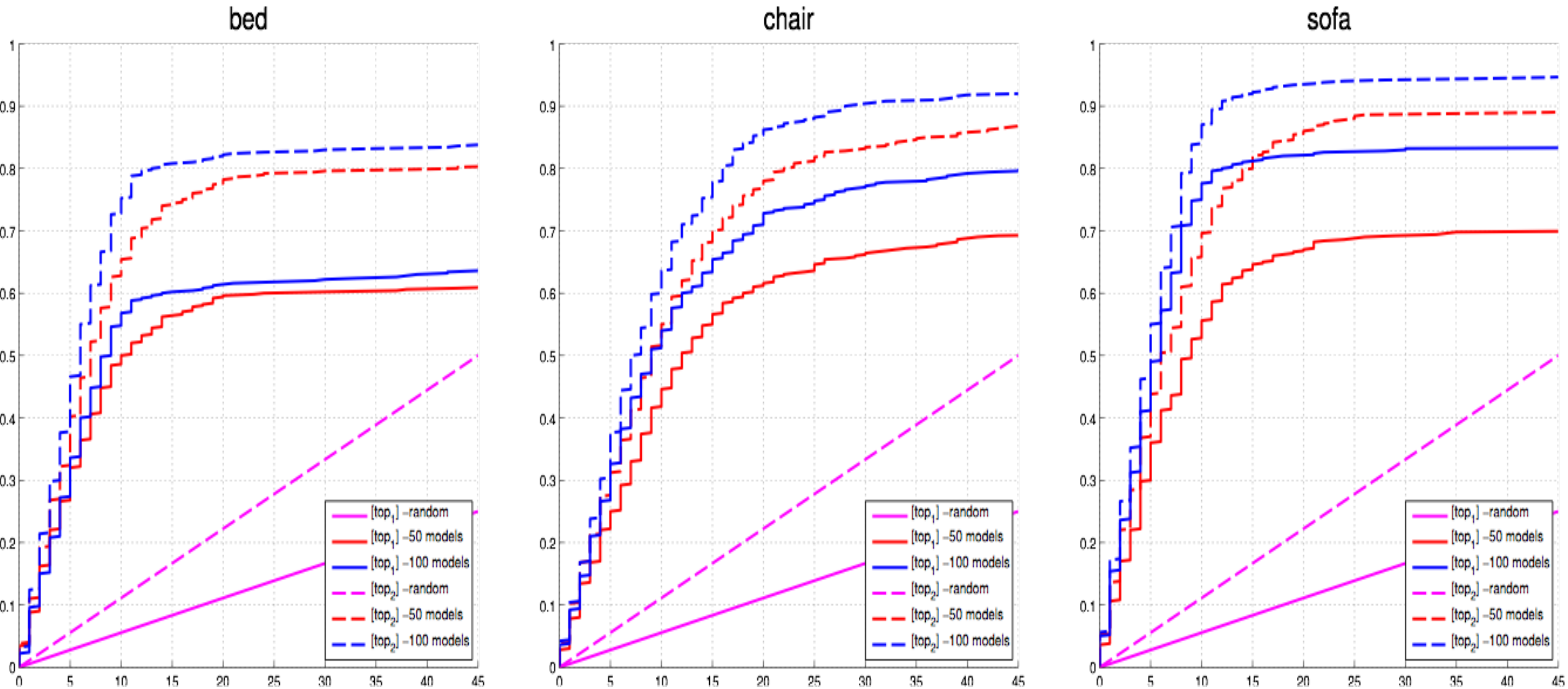
- Train on **synthetic data** (pose aligned CAD models [Wu et al.] rendered in scales and positions they occur in scenes)
- **Input representation**
  - HHA (depth, height above ground, angle with gravity) images don't have azimuth information
  - **Normal Images**
- Desirable to be **robust to occlusion**
- Depth images are 'simpler', so we use a **shallow network**

Use a shallow 3 layer fully convolutional network (average pooling to predict)

Wu et al. 3D ShapeNets for 2.5D Object Recognition and Next-Best-View Prediction. In arXiv 14

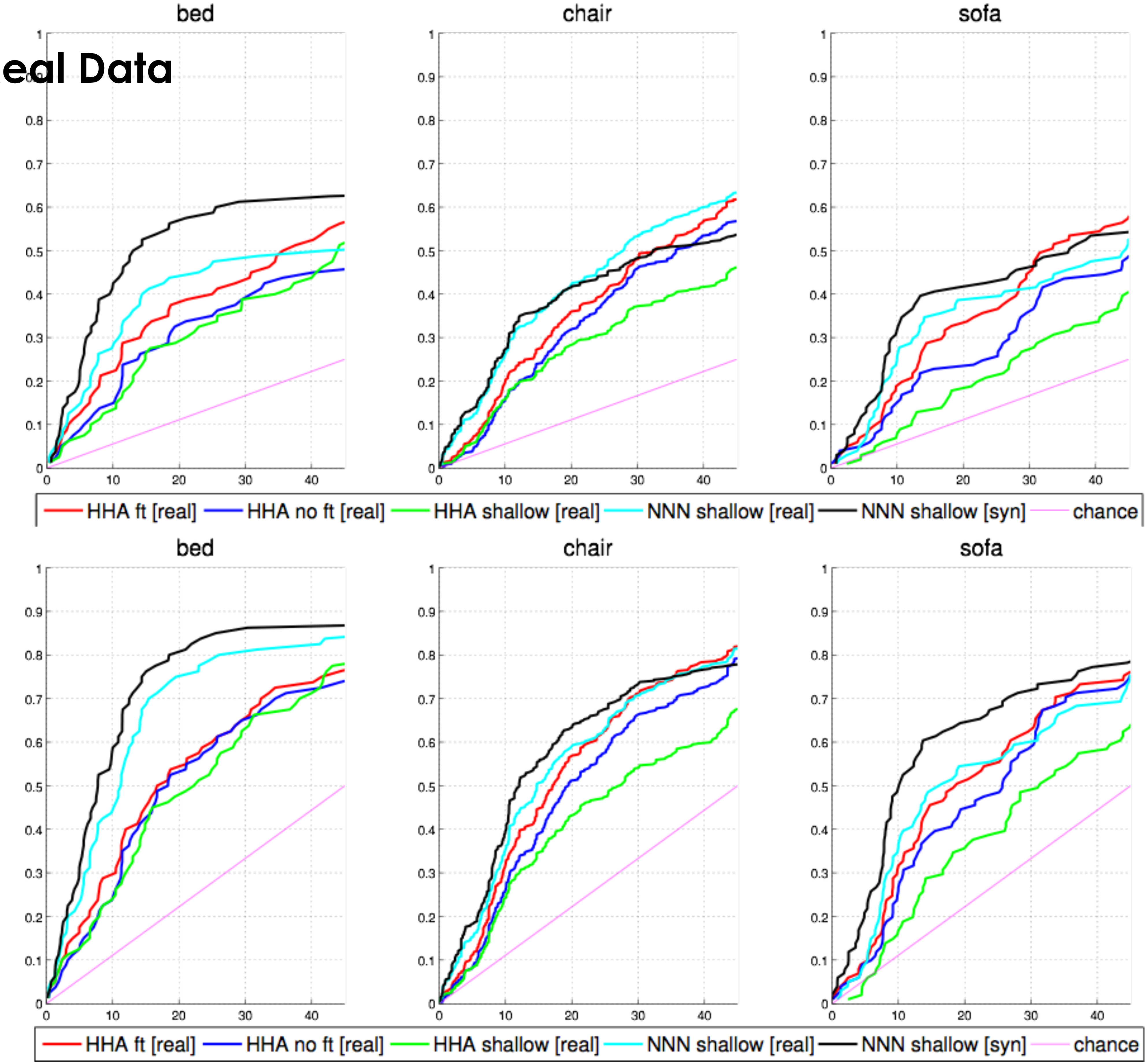
Coarse Pose Estimation

Test on Synthetic Data

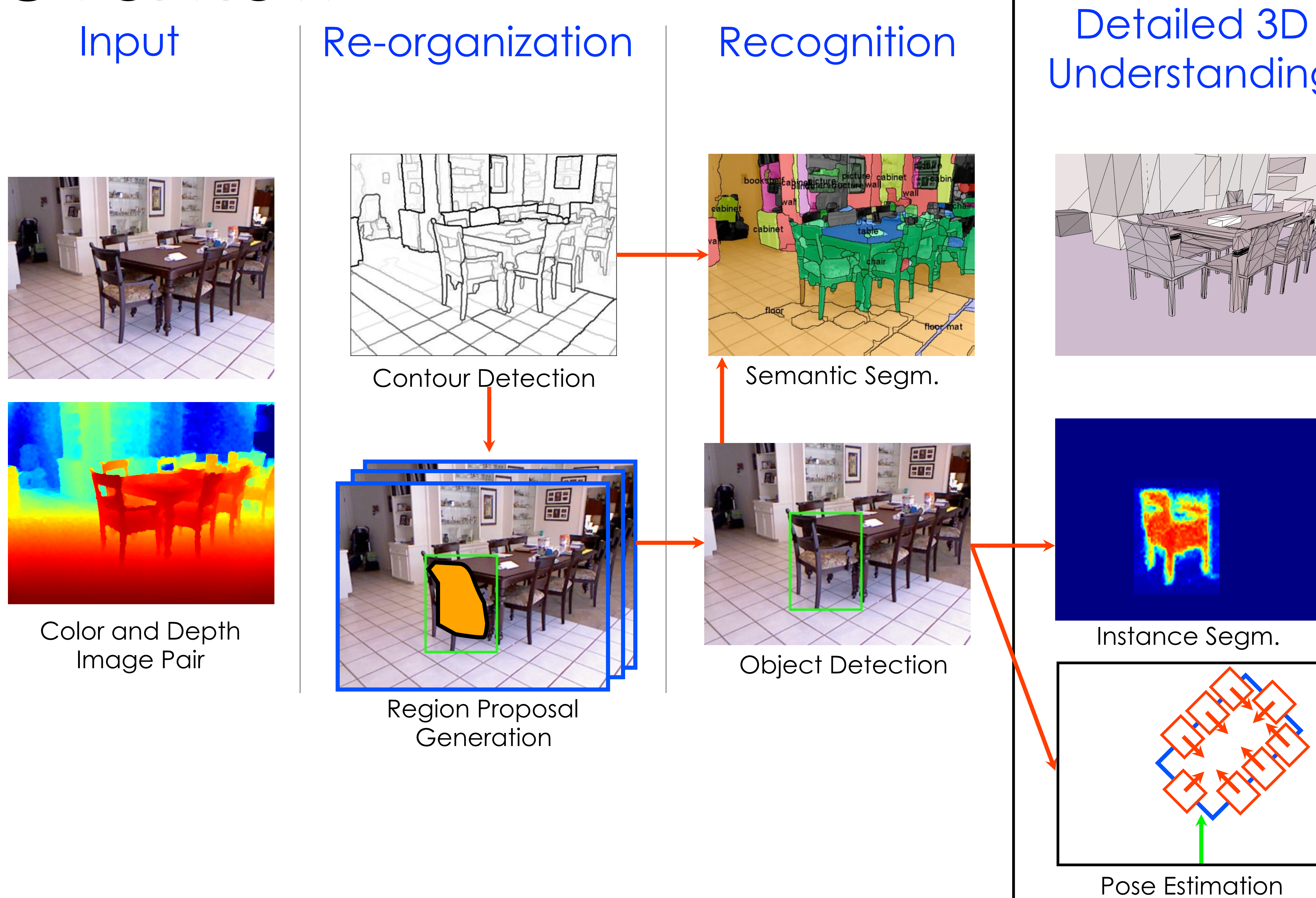


Coarse Pose Estimation

Test on Real Data



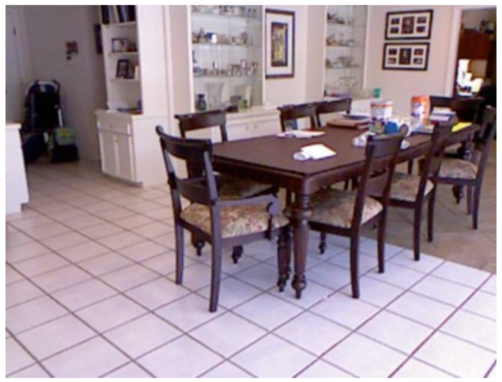
Overview

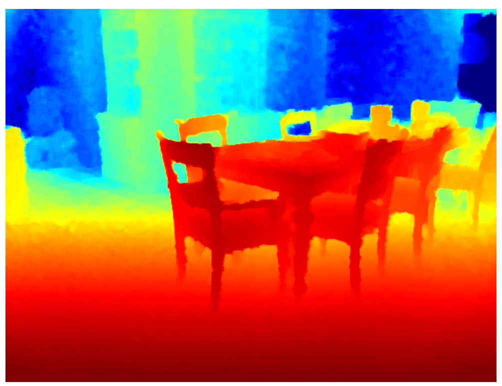


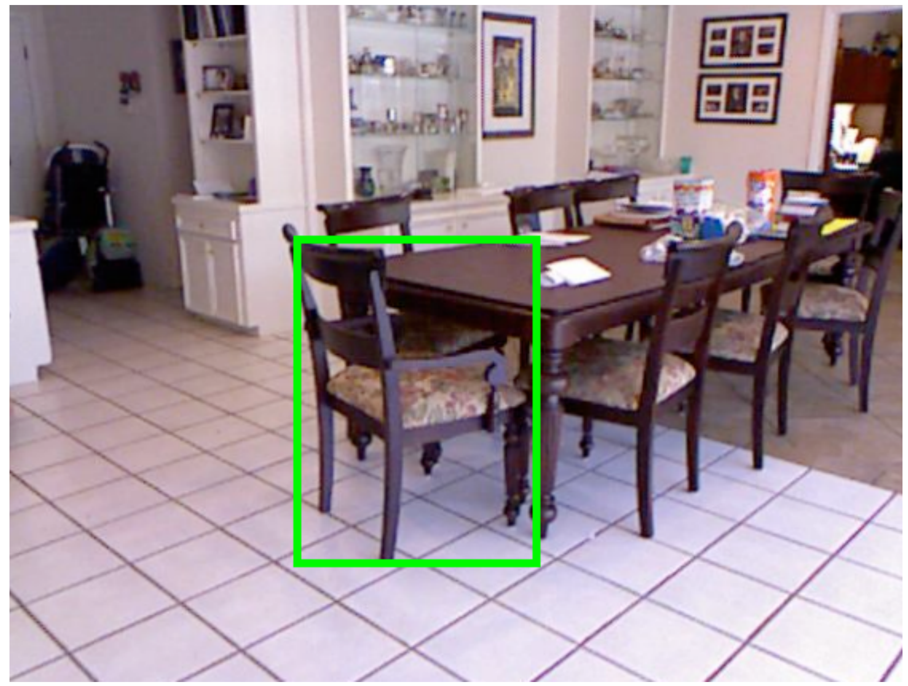
Input
Re-organization
Recognition
Detailed 3D Understanding
Contour Detection
Semantic Segm.
Color and Depth Image Pair
Region Proposal Generation
Object Detection
Instance Segm.
Pose Estimation

Detailed 3D Understanding

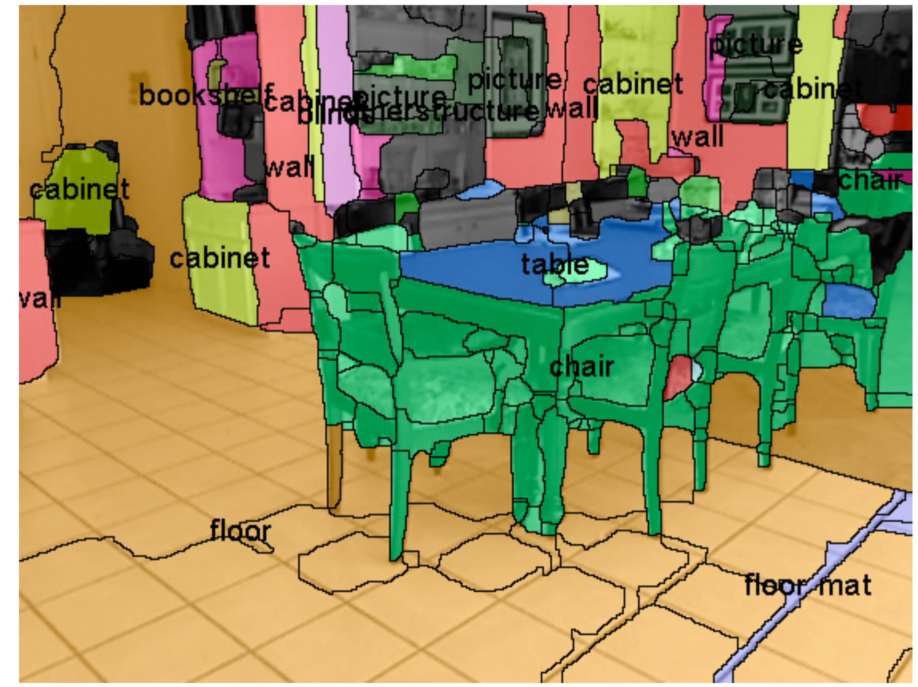
Motivation







Object Detection



Semantic Segm.

Good first steps

But not enough for a robot to manipulate objects

Instance Segmentation

Object Parsing

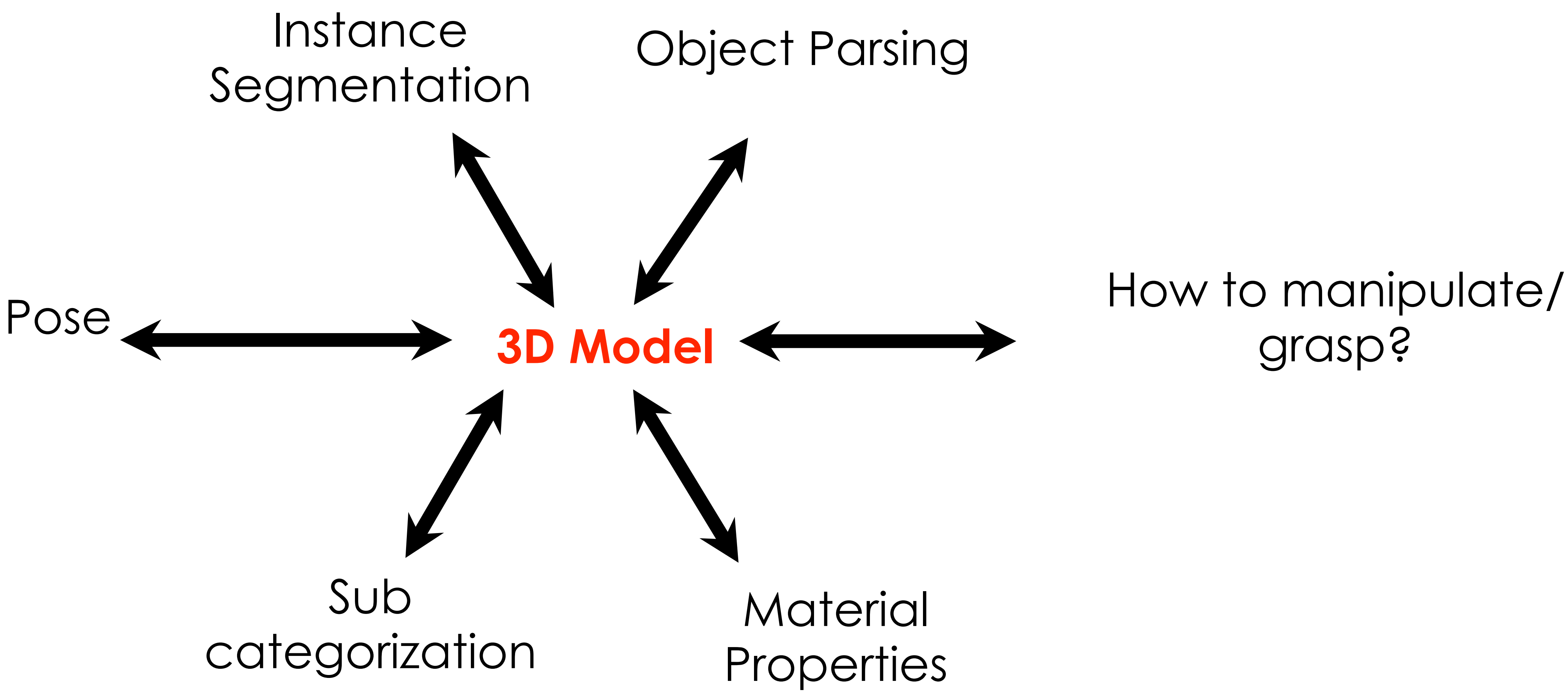
Sub categorization

How to manipulate/ grasp?

Material Properties

Pose

Detailed 3D Understanding



All these tasks are related, doing one will help the other

Estimating the 3D model explains all of these

Current Work / Preliminary Results

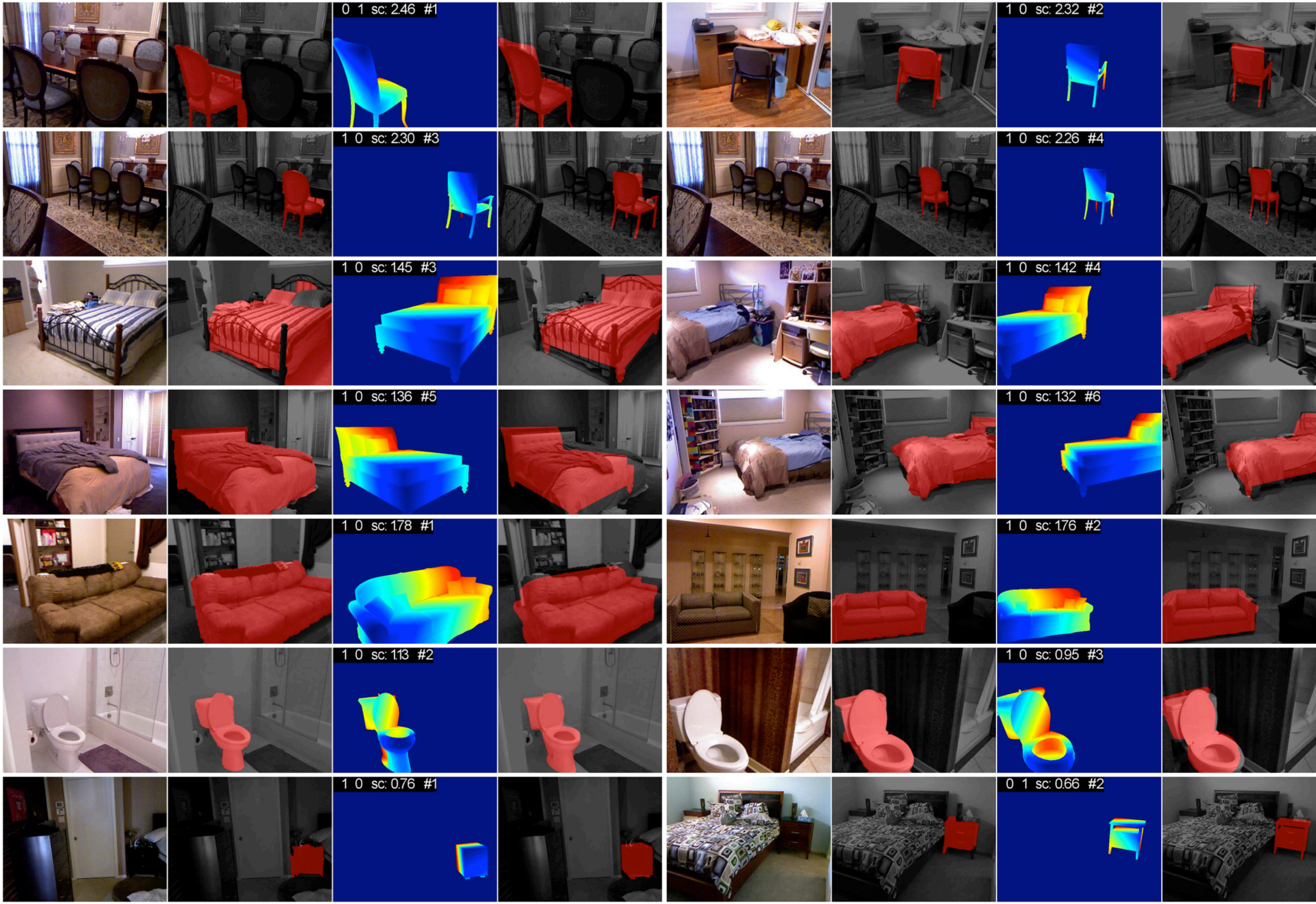
3D Model Estimation

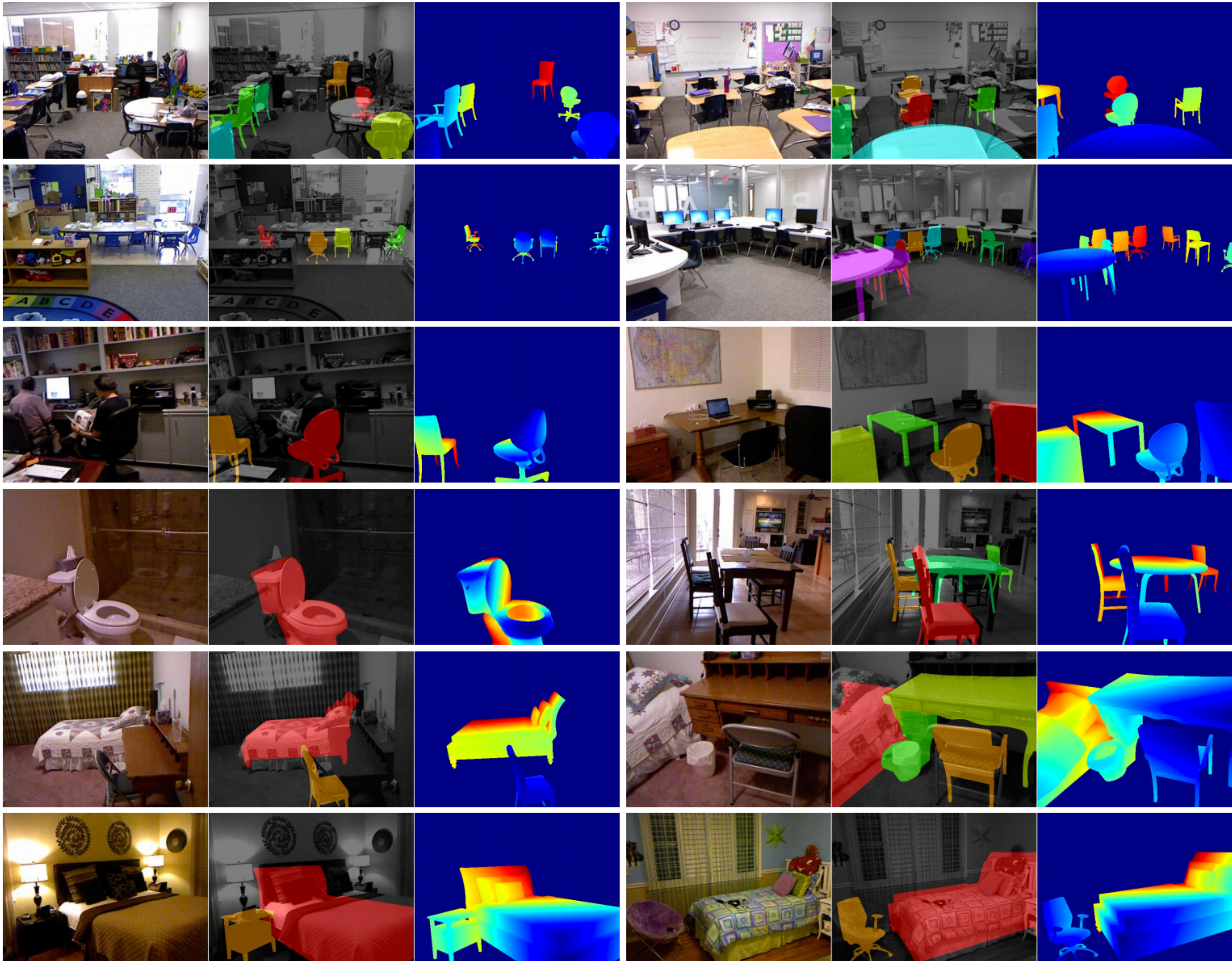
- Start with a model ***M***, at scale ***s***, an initial pose estimate ***R***
    - **Iterative Closest Point (ICP)** to optimize for ***R***, ***t*** (that aligns best to data)
        - Render model, use visible points, run ICP between these points, and points in the segmentation mask, re-estimate ***R***, ***t***, repeat
- Pick best model ***M****, scale ***s**** and pose ***R****, ***t**** based on fit to the data

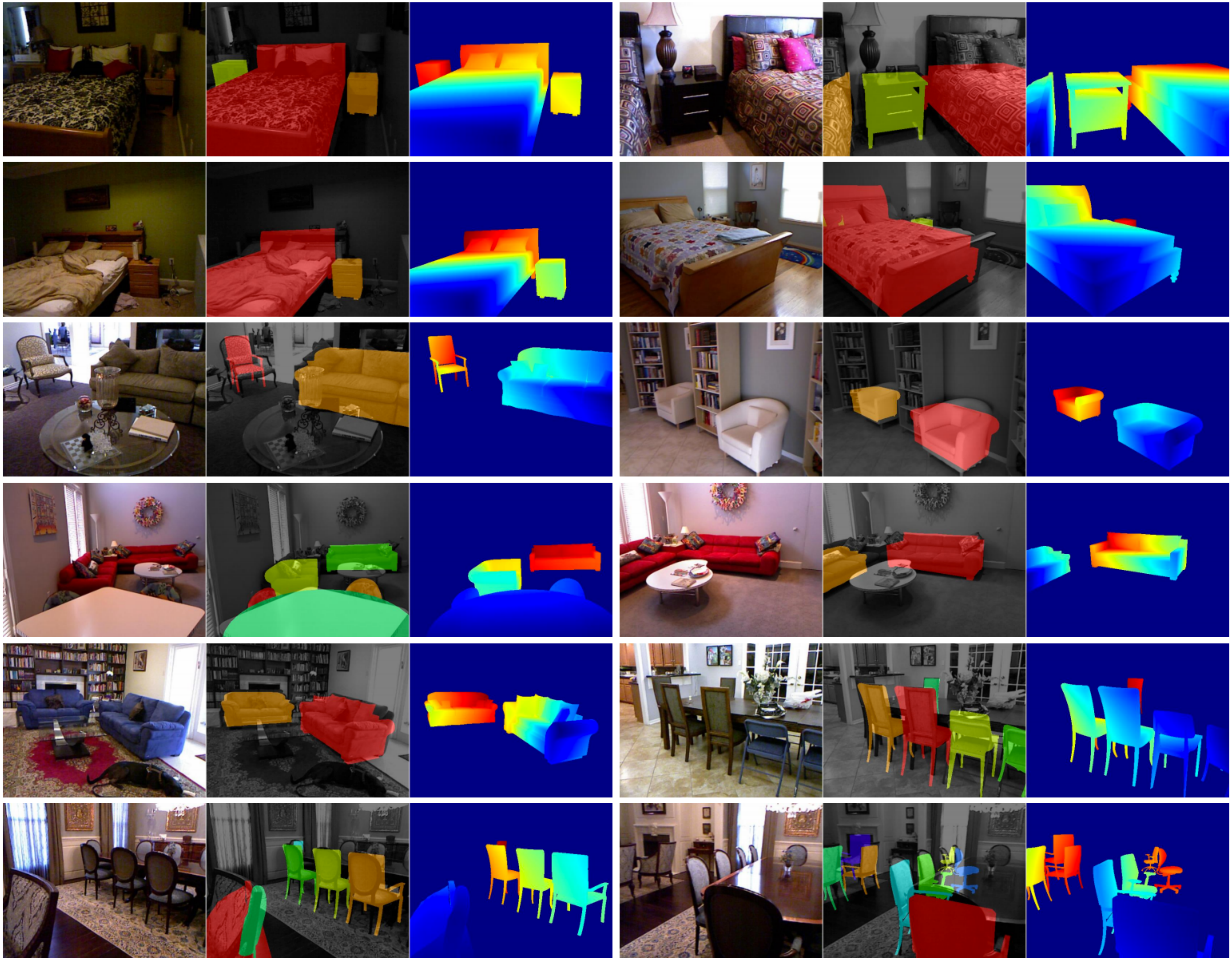
Works reasonably well even though

- **Inaccurate models**
- **Imperfect segmentation masks**

3D Model Estimation Results







3D Model Estimation

For 3D Detection

Put a 3D box around the 3D extent of the object

3D All (AP)	mean	bed	chair	sofa	table	toilet
Sliding Shapes	39.6	33.5	29.0	34.5	**33.8**	67.3
Our - 3D Box on Instance Segm.	48.4	**74.7**	18.6	50.3	28.6	69.7
Our - 3D Box on Model	**58.5**	73.4	**44.2**	**57.2**	33.4	**84.5**

3D Clean (AP)	mean	bed	chair	sofa	table	toilet
Sliding Shapes	64.6	71.2	**78.7**	41.0	**42.8**	89.1
Our - 3D Box on Instance Segm.	66.1	**90.9**	45.9	68.2	25.5	**100**
Our - 3D Box on Model	**71.1**	82.9	72.5	**75.3**	24.6	**100**

[Sliding Shapes] S. Song and J. Xiao Sliding shapes for 3D object detection in depth images. In ECCV 14.
[arXiv 15] S. Gupta et al. Inferring 3D Object Pose in RGB-D Images In arXiv 15.

3D Model Estimation

Results

AP^m

Prediction is an explicit placement of a model.

Pixels in intersection correct only when within some distance of the ground truth depth value

	detection setting		
t_{agree}	0.5, 5 7	0.5, 5 ∞	AP^r upper bound
bathtub	7.9	50.4	42.0
bed	31.8	68.7	65.0
chair	14.7	35.6	42.9
desk	4.1	10.8	12.0
dresser	26.3	35.0	36.1
monitor	5.7	7.4	11.4
night-stand	28.1	33.7	34.8
sofa	21.8	48.5	47.4
table	5.6	12.3	15.0
toilet	41.8	68.4	68.4
mean	**18.8**	**37.1**	**37.5**

Future Work

3D Object Context

Toilet in a bedroom

Chair overlapping with Table

Different chairs in a dinning set

Future Work

More Data

Current RGB-D datasets are really small

Dataset	# Training Images
NYUD2	0.8K
PASCAL	12K
MS COCO	120K
ImageNet	1000 K

Algorithms far from saturation

# Training Images	AP^b	AP^r
381	36.3	31.3
795	41.2	37.5

More Richly Annotated Data

New metrics and corresponding annotations for detailed tasks like

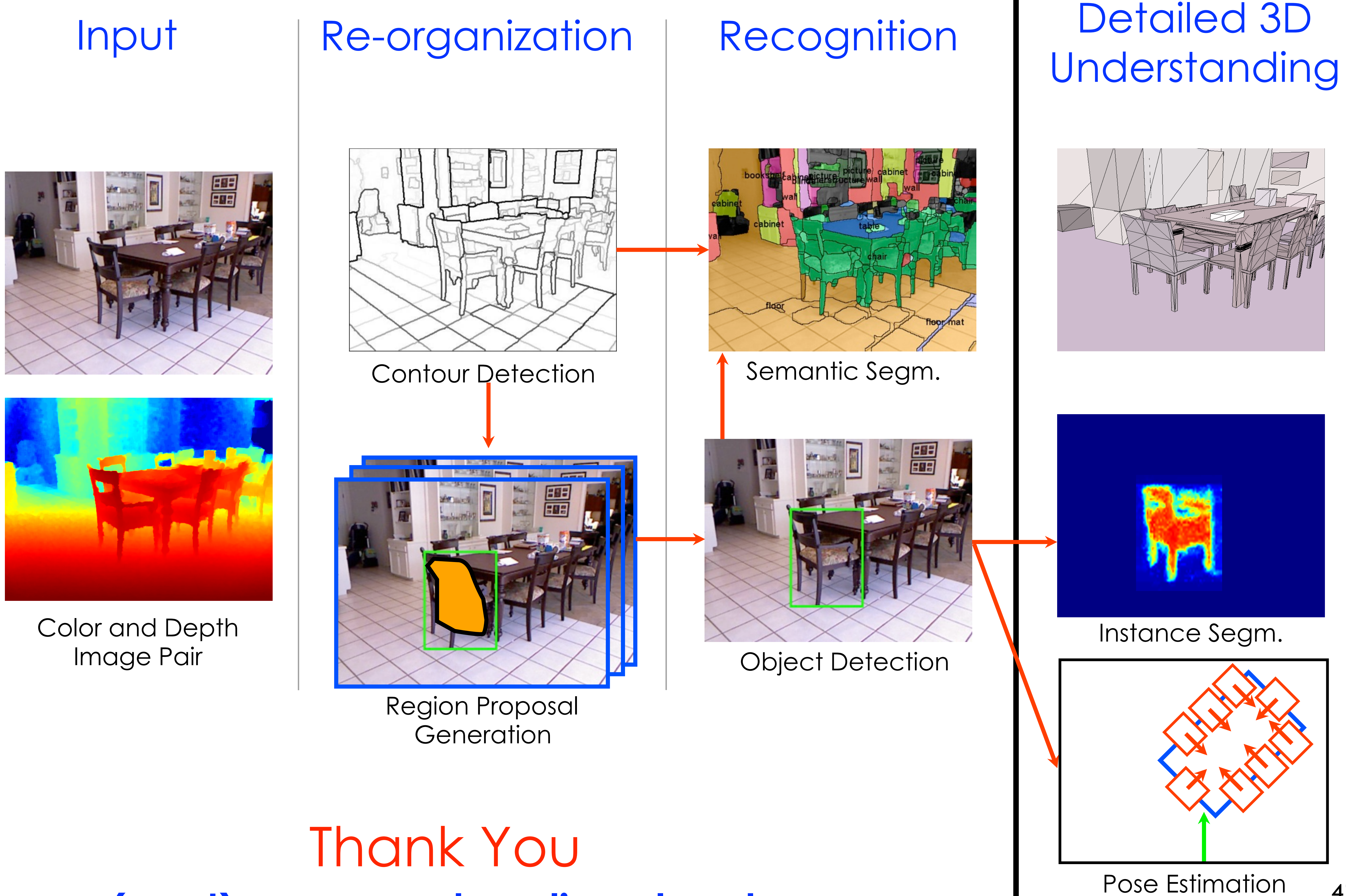
- pose estimation
- part labelling
- model placement

Realistic CAD models

Real high-fidelity models acquired using Kinect Fusion

Looking forward to new dataset from Princeton + Intel

Overview



Thank You

(most) source code online already