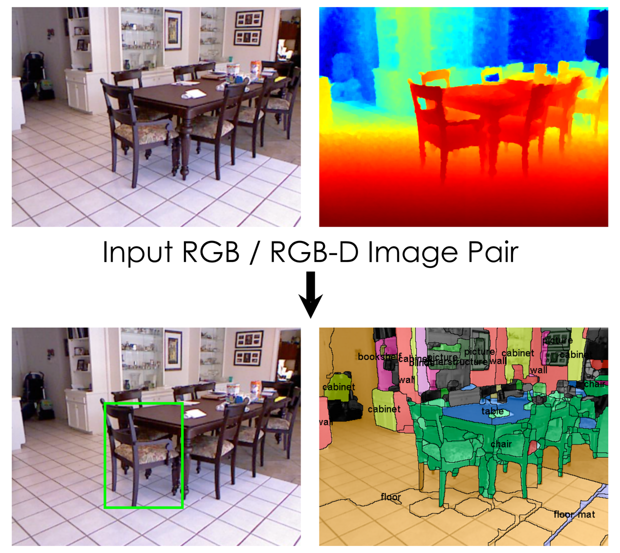
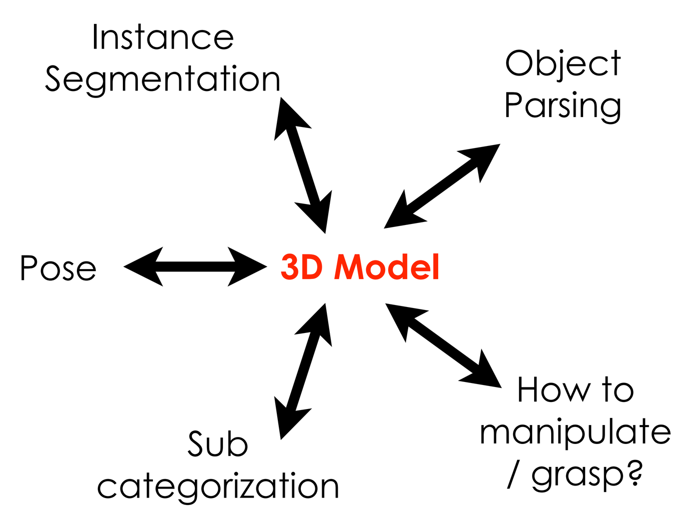
Aligning 3D Models to RGB-D Images of Cluttered Scenes

Saurabh Gupta[1] Pablo Arbeláez[2] Ross Girshick[3] Jitendra Malik[1]

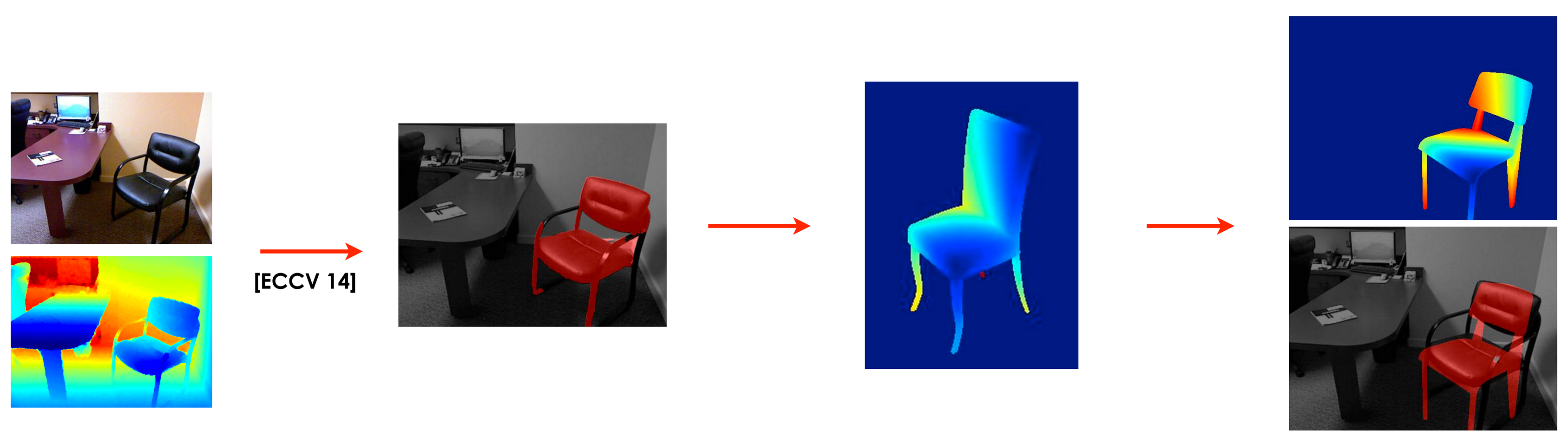
[1]UC Berkeley [2]Universidad de los Andes, Colombia [3]Microsoft Research

Replacing in-place with a 3D model





Overview



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

Learning from synthetic data and generalizing to real data

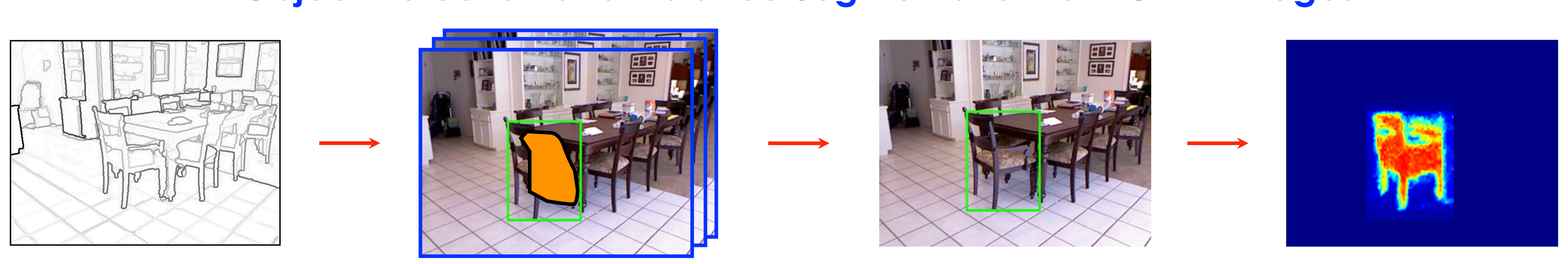
Starting with weak annotation (instance segmentation) able to produce a much richer output

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

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

Related Work

Object Detection and Instance Segmentation for RGB-D Images



Improved **contours** from RGB-D

Much better **region proposals**

Adapting CNNs trained on RGB images to Depth Images

A **geocentric embedding** for Depth images (HHA)

Extend this for task of **instance segmentation**

Acknowledgements: The authors would like to thank Shubham Tulsiani and Bharath Hariharan for their valuable comments. This work was supported by ONR SMARTS MURI N00014-09-1-1051, ONR MURI N00014-10- 10933, and a Berkeley Graduate Fellowship. We gratefully acknowledge NVIDIA corporation for the donation of Tesla and Titan GPUs used for this research.

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

[Girshick et al.] R. Girshick,J. Donahue, T. Darell, J. Malik **Rich feature hierarchies for accurate object detection and semantic segmentation**, CVPR 2014

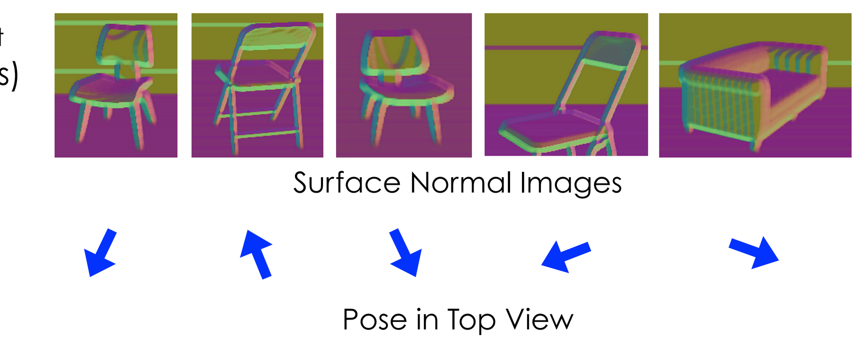
[Song et al.] S. Song and J. Xiao **Sliding shapes for 3D object detection in depth images**. In ECCV 14.

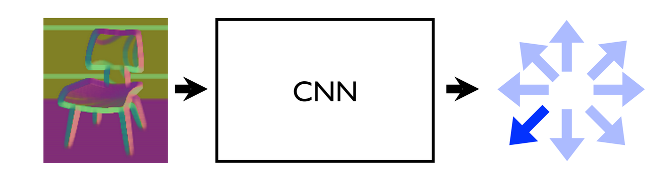
[Silberman et al.] N. Silberman, D. Hoiem, P. Kohli, R. Fergus **Indoor segmentation and support inference from RGBD images**, ECCV 2012

[Wu et al.] Z Wu, S Song, A Khosla, F Yu, L Zhang, X Tang, J Xiao **3D ShapeNets for 2.5D Object Recognition and Next-Best-View Prediction**. In CVPR 15

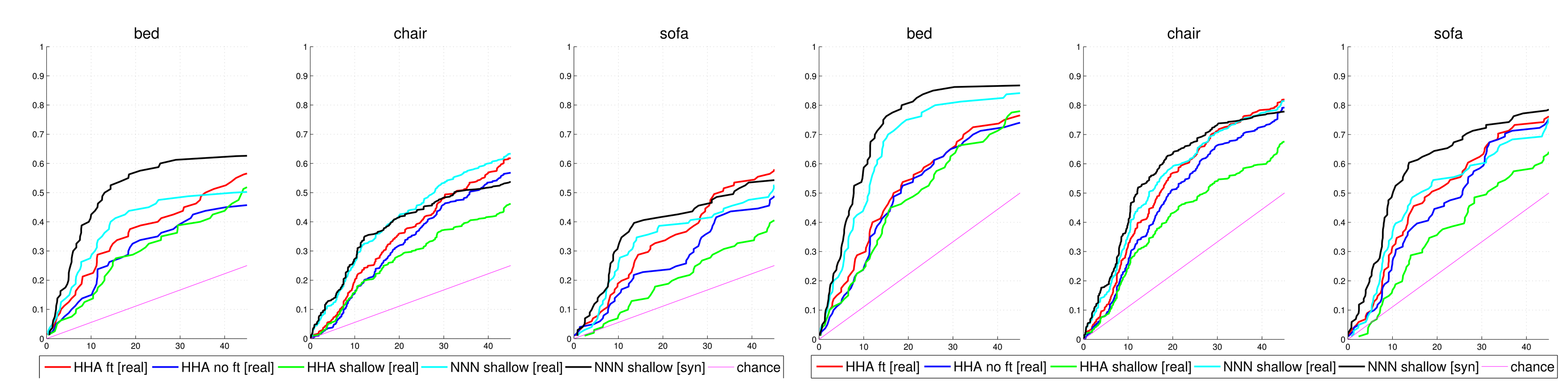
Coarse Pose Estimation

- 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)



1 Top Accuracy

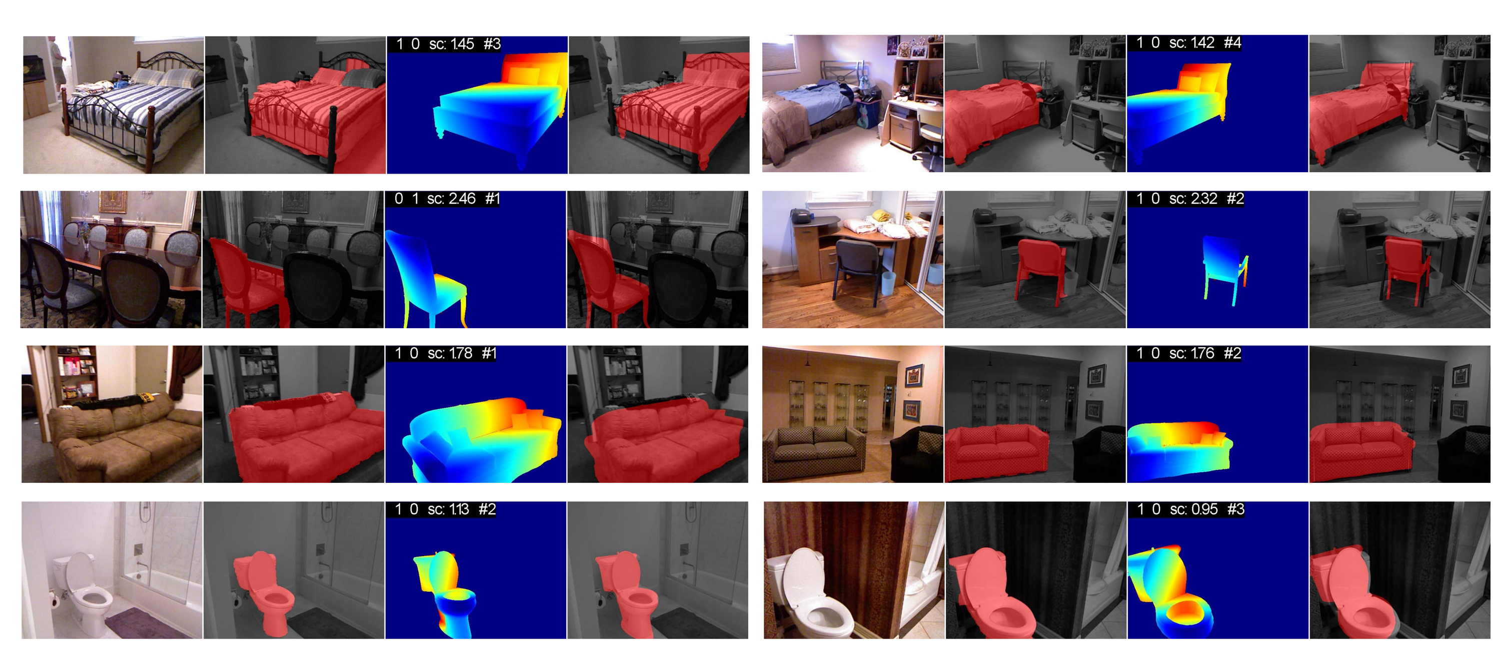
Top 2 Accuracy

Fine Pose 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**



Results

3D Object Detection

Putting a 3D Bounding box around the object in 3D [Song et al.]

	3D all						3D clean					
	mean	bed	chair	sofa	table	toilet	**mean**	bed	chair	sofa	table	toilet
Our (3D Box on instance segm. from [13])	48.4	74.7	18.6	50.3	28.6	69.7	66.1	90.9	45.9	68.2	25.5	100.0
Our (3D Box around estimated model)	58.5	73.4	44.2	57.2	33.4	84.5	71.1	82.9	72.5	75.3	24.6	100.0
Song and Xiao [34]	39.6	33.5	29.0	34.5	33.8	67.3	64.6	71.2	**78.7**	41.0	**42.8**	89.1
Our [no RGB[1]] (3D Box on instance segm. from [13])	46.5	71.0	18.2	49.6	30.4	63.4	62.3	**86.9**	43.6	57.4	26.6	**96.7**
Our [no RGB[1]] (3D Box around estimated model)	**57.6**	**72.7**	**47.5**	**54.6**	**40.6**	**72.7**	**70.7**	84.9	75.7	**62.8**	33.7	**96.7**

AP^m

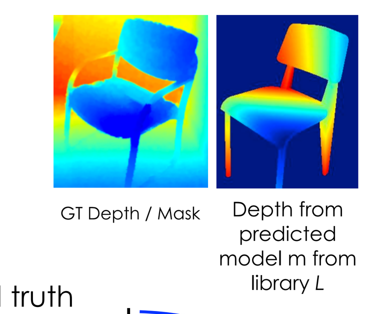
Algorithm outputs rendering of a model, m from a library L and an appropriate transformation, s, R, t

Render model, perform occlusion checking

Assign predicted **model** to ground truth **regions** based on **region** I/U overlap

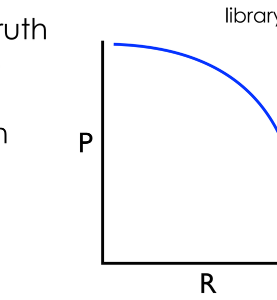
Pixels count in intersection only when within some distance of the ground truth depth value

AP^m = area under PR curve

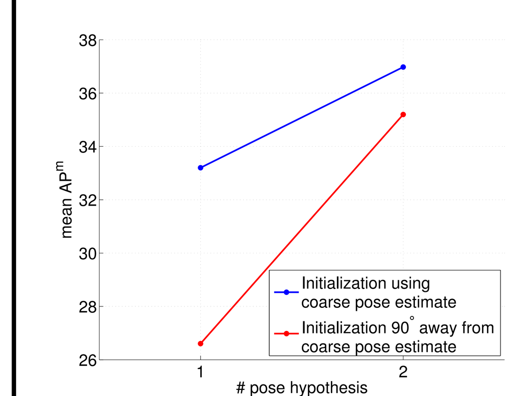


GT Depth / Mask

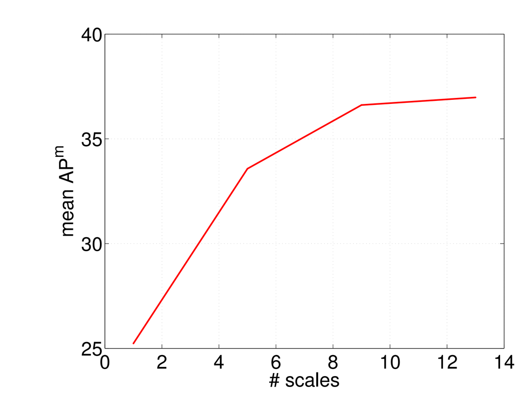
Depth from predicted model m from library L



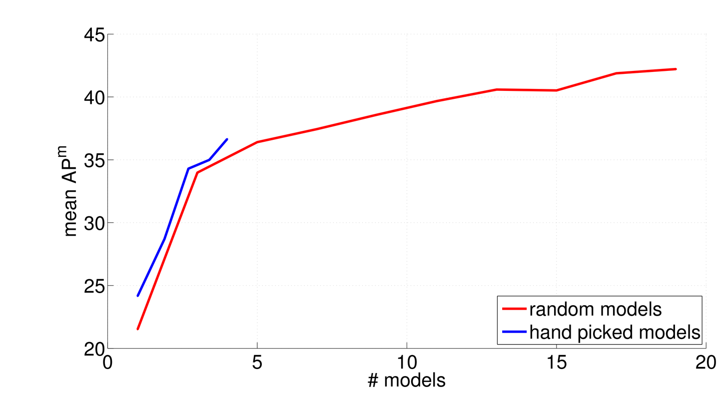
	val set								test set		
	ground truth segm		latent positive setting			detection setting			detection setting		
t_{agree}	0.5, 5 7	0.5, 5 ∞	0.5, 5 7	0.5, 5 ∞	AP^r upper bound	0.5, 5 7	0.5, 5 ∞	AP^r upper bound	0.5, 5 7	0.5, 5 ∞	AP^r upper bound
bathtub	57.4	76.8	55.3	83.3	94.7	6.7	19.4	25.7	7.9	50.4	42.0
bed	42.3	87.3	28.8	86.0	96.1	25.8	63.2	57.0	31.8	68.7	65.0
chair	45.3	74.1	29.0	56.9	70.1	11.8	25.2	30.4	14.7	35.6	42.9
desk	33.9	67.4	20.3	40.9	55.7	3.0	4.0	6.2	4.1	10.8	12.0
dresser	82.7	92.0	76.1	96.0	100.0	13.3	21.1	21.1	26.3	35.0	36.1
monitor	31.4	39.8	18.4	20.8	41.3	12.5	12.5	26.8	5.7	7.4	11.4
night-stand	62.5	77.6	51.3	65.2	87.9	18.9	21.6	25.5	28.1	33.7	34.8
sofa	45.1	85.0	28.5	72.0	92.4	10.5	30.4	37.7	21.8	48.5	47.4
table	18.8	52.2	15.8	34.3	46.8	5.5	11.9	13.3	5.6	12.3	15.0
toilet	66.0	100.0	46.0	86.0	100.0	35.9	72.4	73.2	41.8	68.4	68.4
mean	**48.5**	**75.2**	**37.0**	**64.1**	**78.5**	**14.4**	**28.2**	**31.7**	**18.8**	**37.1**	**37.5**



Importance of Initialization



scales



3D Models