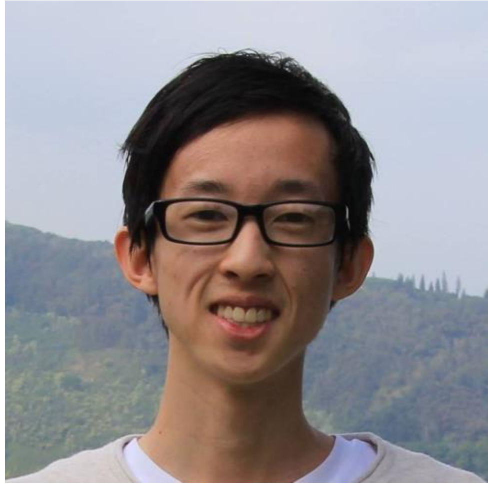
Scene Understanding with 3D Deep Networks

Thomas Funkhouser
Princeton University

Disclaimer: I am talking about the work of these people …

Shuran Song



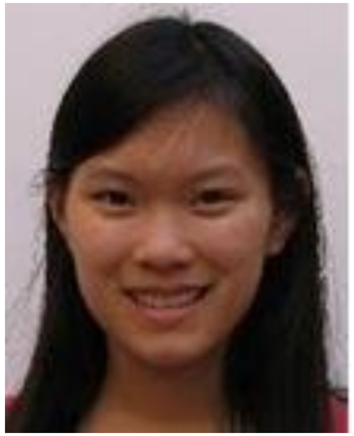
Andy Zeng

Fisher Yu

Yinda Zhang

Maciej Halber

Jianxiong Xiao



Angela Dai

Matthias Niessner

Matt Fisher

Goal

Understanding indoor scenes observed in RGB-D images

- Robotics
- Augmented reality
- Virtual tourism
- Surveillance
- Home remodeling
- Real estate
- Telepresence
- Forensics
- Games
- etc.

Goal

Understanding indoor scenes observed in RGB-D images

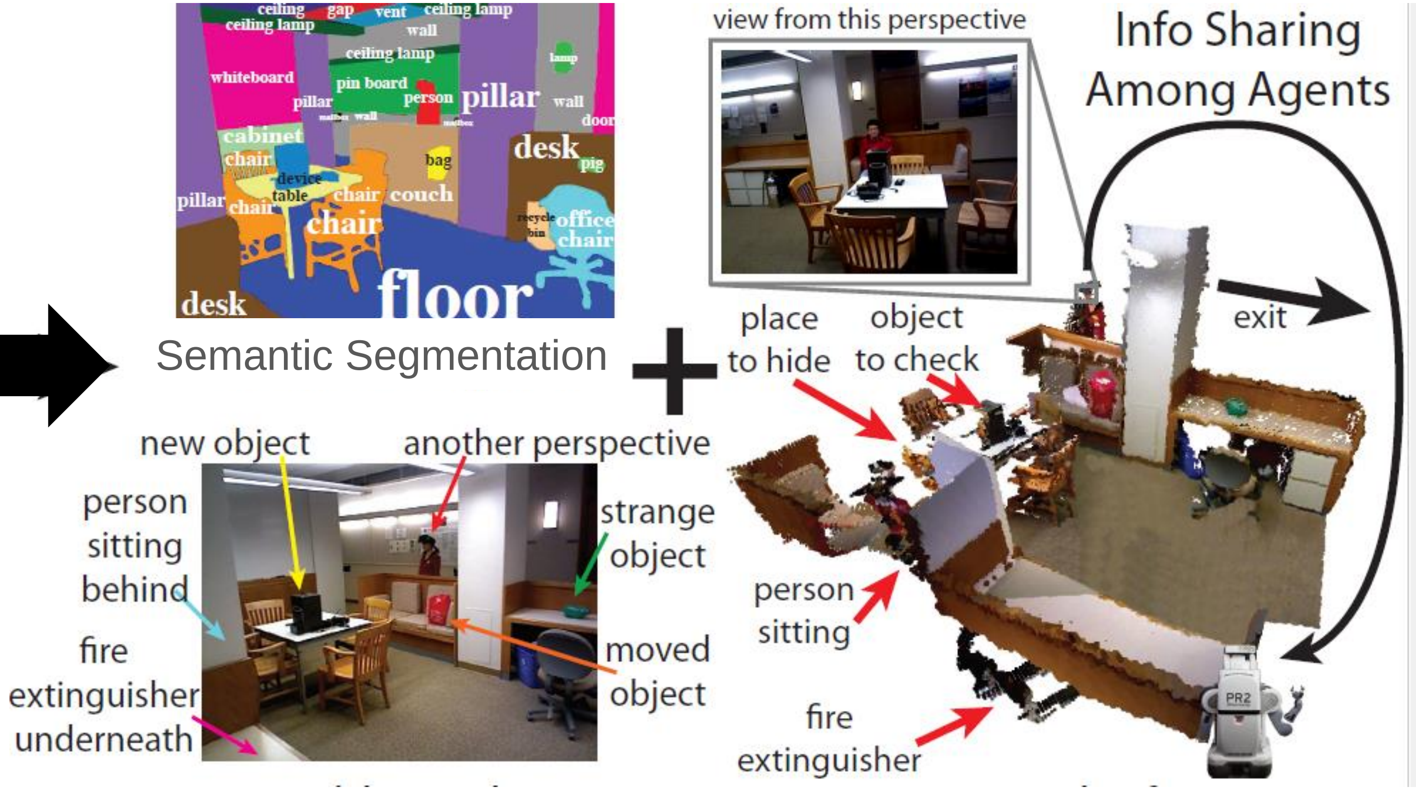
Input RGB-D Image(s)

Semantic Segmentation

Goal

Understanding indoor scenes observed in RGB-D images in 3D

Input RGB-D Image(s)

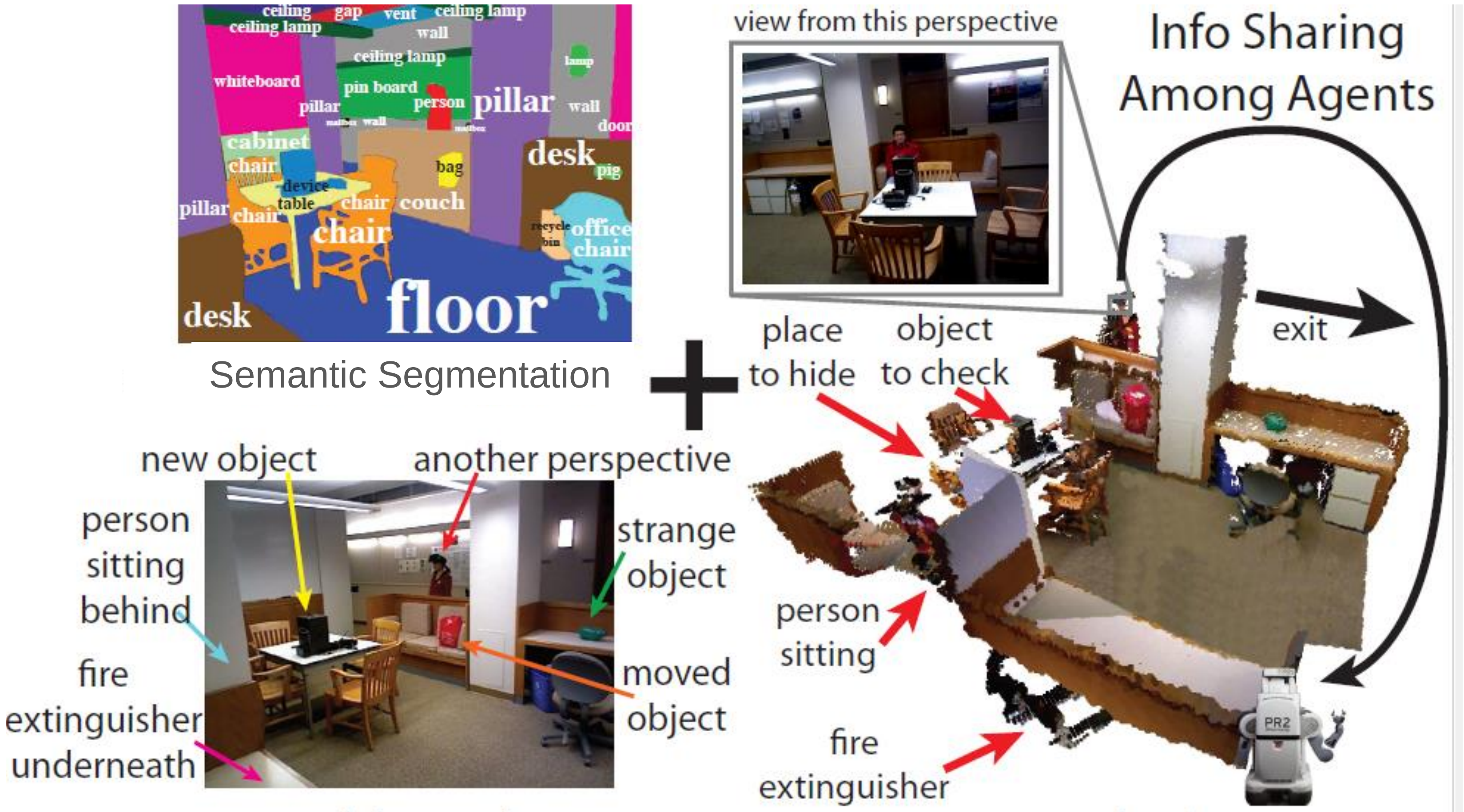


3D Scene Understanding

Goal

Understanding indoor scenes observed in RGB-D images in 3D

- Surface reconstruction
- Amodal object detection
- Object relationships
- Materials, lights, etc.
- Physical properties
- Novel views
- Info sharing
- Spatial inference
- Simulation
- etc.



Goal for This Talk

Learn ConvNets to recognize patterns in voxels

- Local shape descriptor
- Amodal object detection
- Semantic scene completion

Talk Outline

Small

Scale

Large

Local shape descriptor

Amodal object detection

Semantic scene completion

Talk Outline

Small

Scale

Local shape descriptor

Amodal object detection

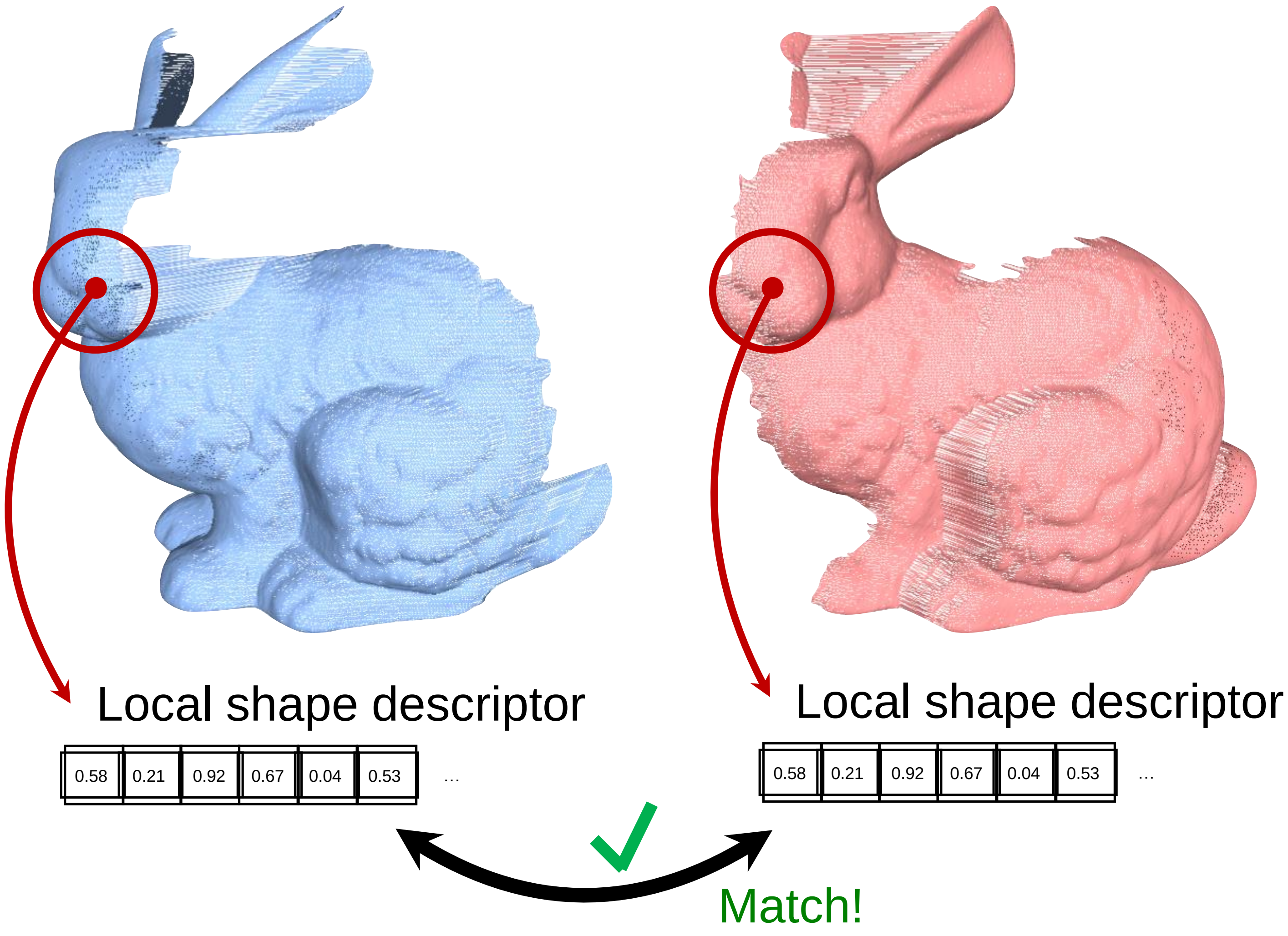
Semantic scene completion

Large

A. Zeng, S. Song, M. Niessner, M. Fisher, J. Xiao, T. Funkhouser,
"3DMatch: Learning Local Geometric Descriptors from 3D Reconstructions,"
submitted to CVPR 2017

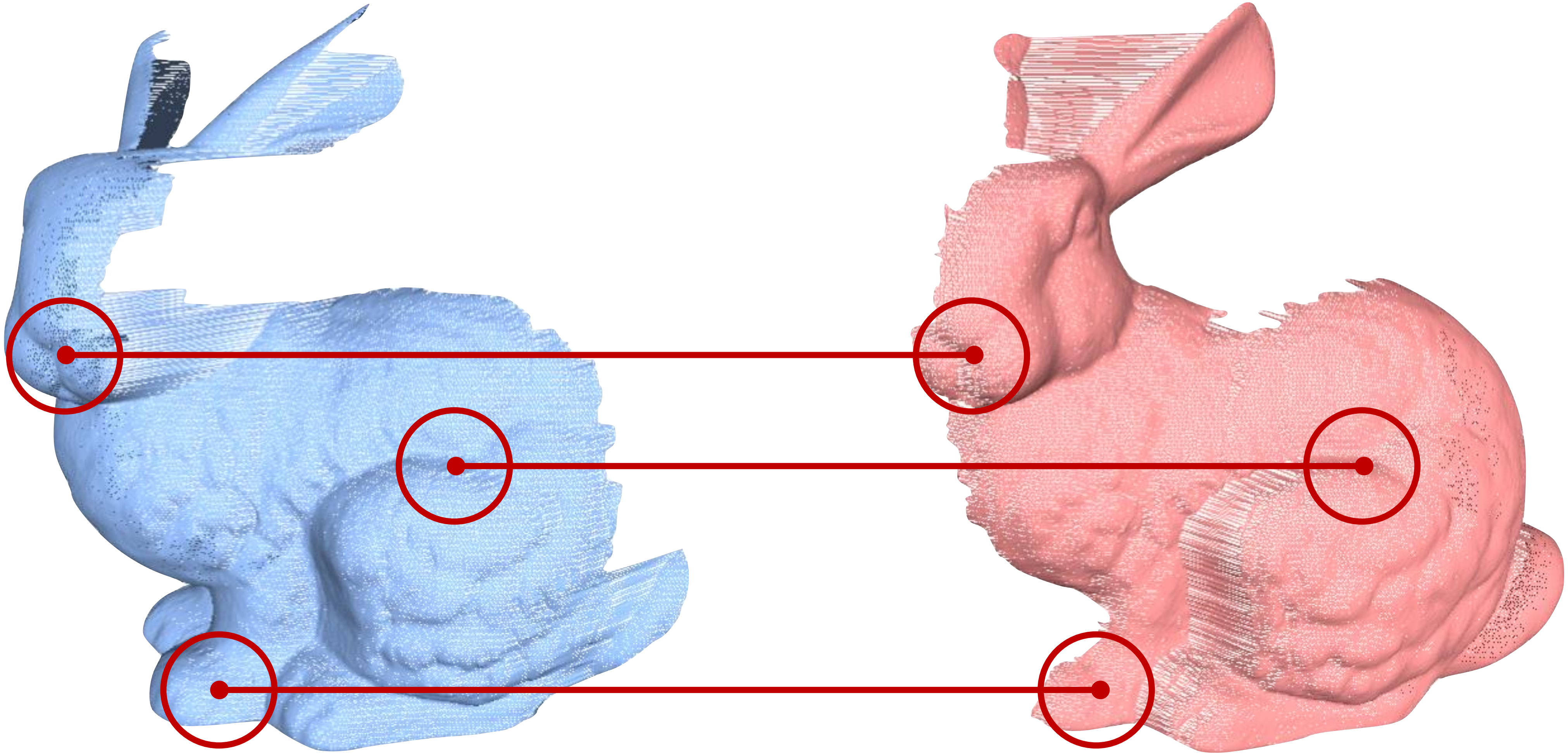
Local Shape Descriptor

Goal: train a discriminating 3D local shape descriptor from data



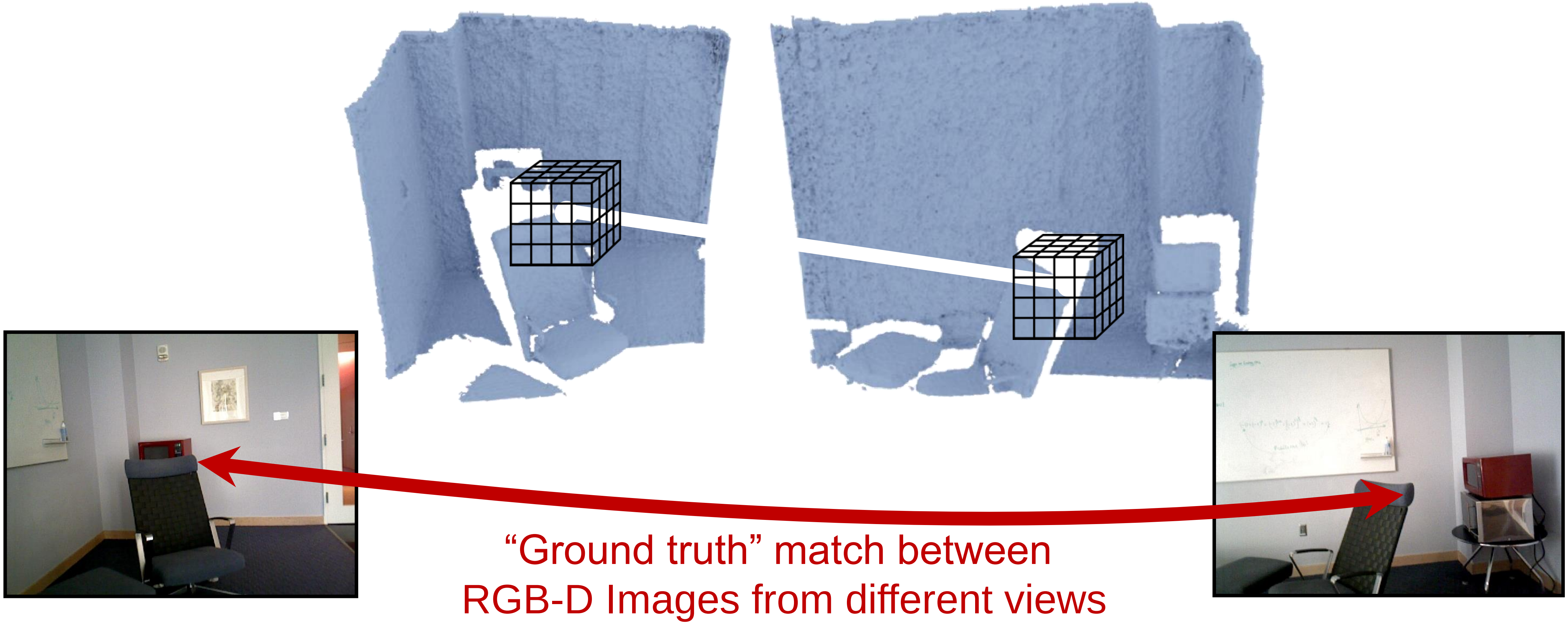
Local Shape Descriptor

Challenge: where to get training data?



Local Shape Descriptor: “3D Match”

Approach: train on wide-baseline correspondences in RGB-D reconstructions

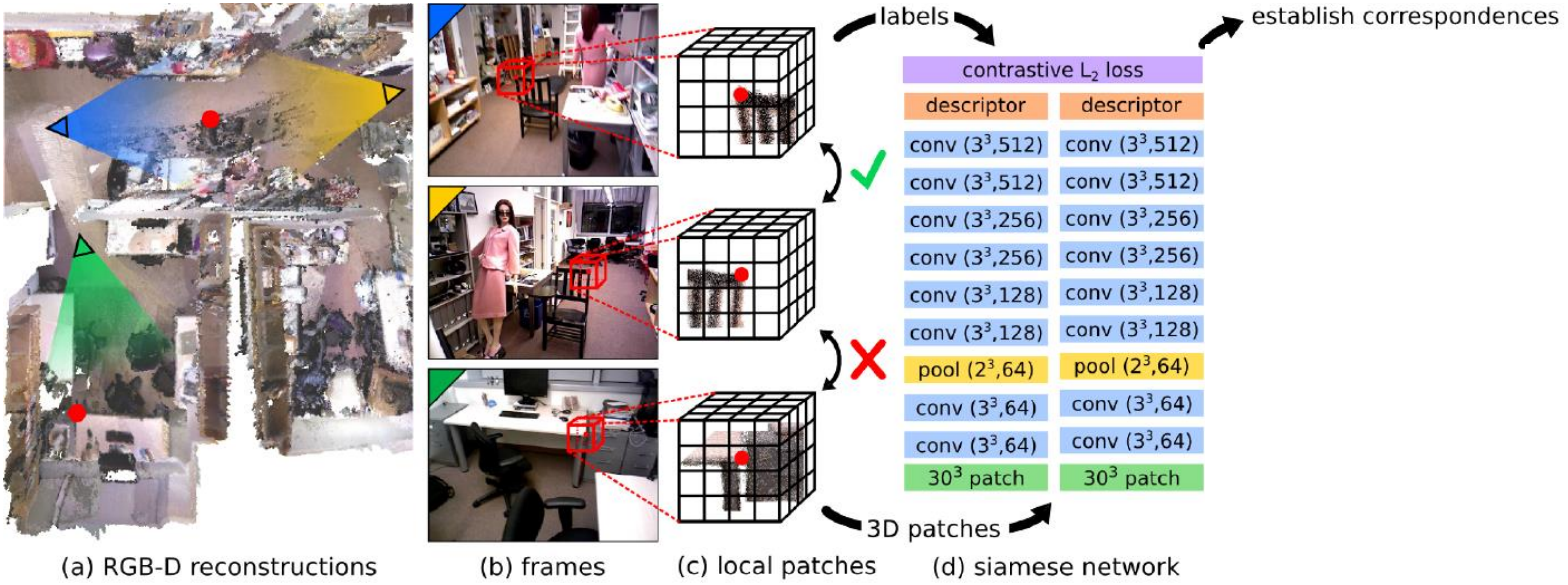


Local Shape Descriptor: “3D Match”

Approach: train on wide-baseline correspondences in RGB-D reconstructions

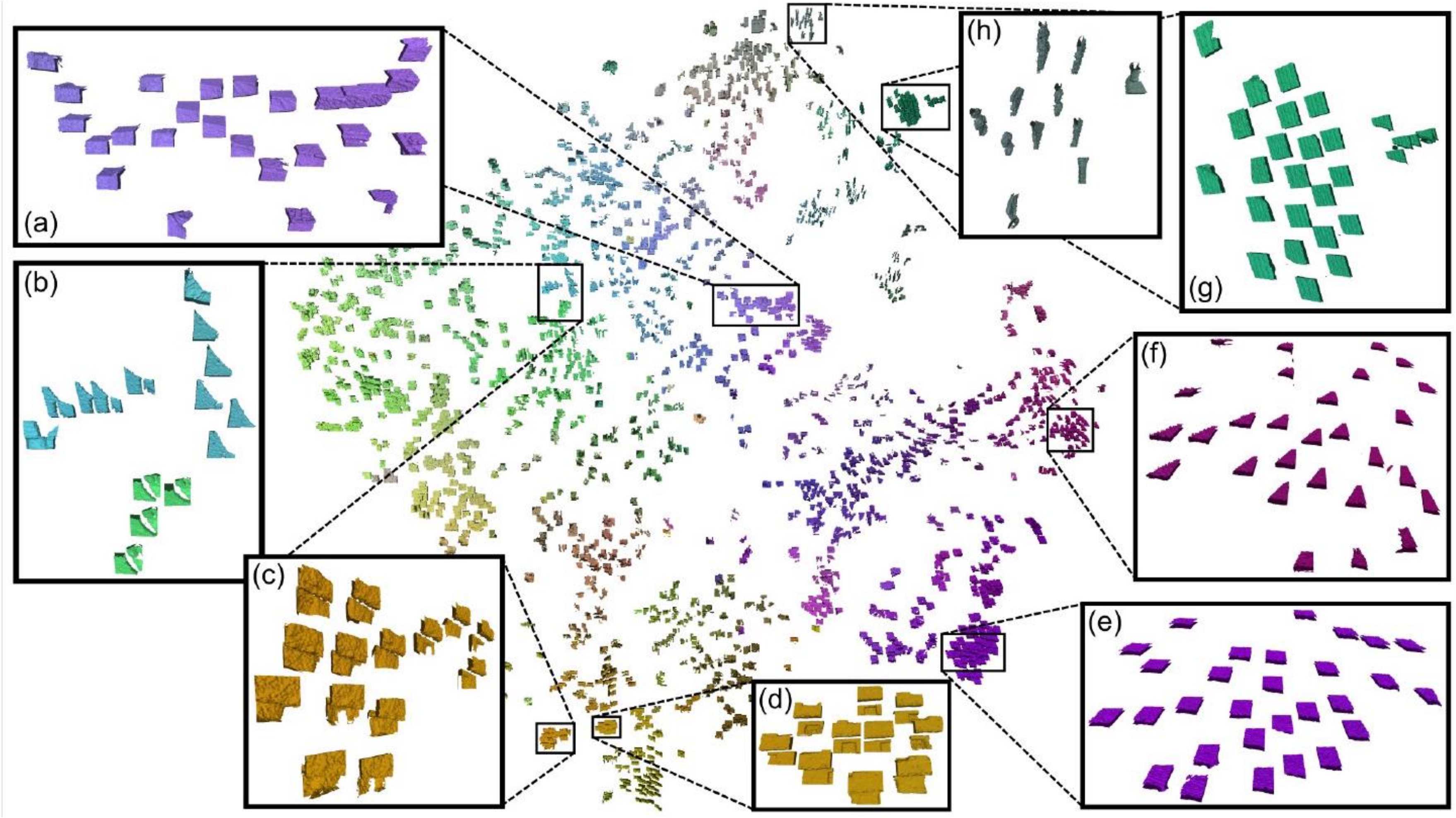
Local Shape Descriptor: “3D Match”

Method: sample true/false correspondences from RGB-D reconstructions, train Siamese network



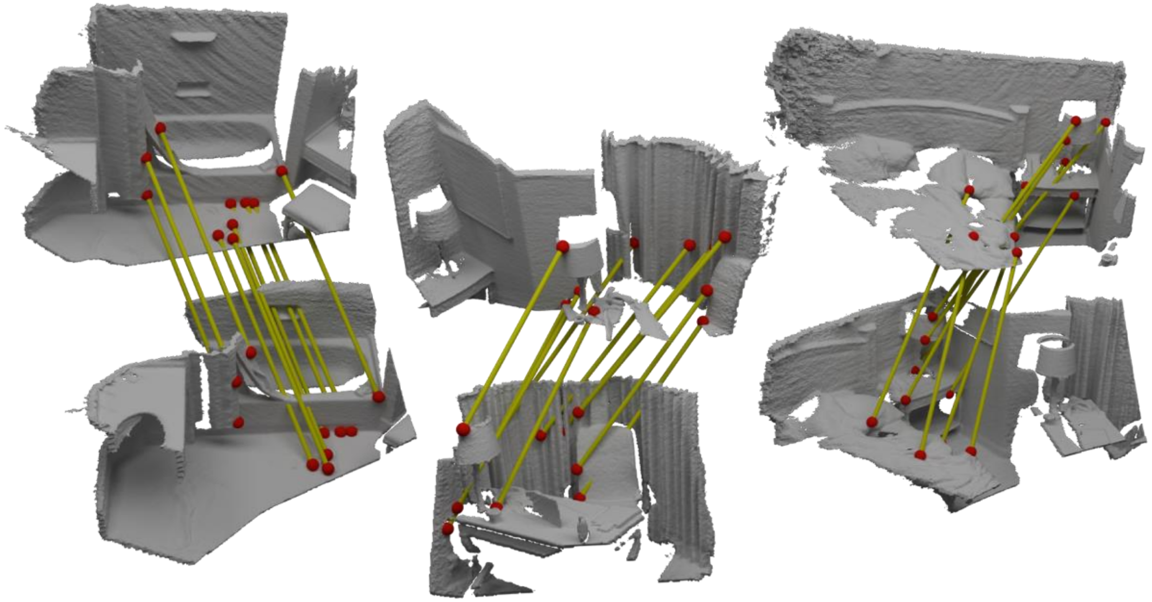
Local Shape Descriptor: “3D Match”

Result: learns to discriminate local shapes found in real-world data



Local Shape Descriptor: "3D Match" Results

Result 1: learned feature descriptor predicts RGB-D point correspondences more accurately than hand-tuned descriptors



Method	Error
Johnson *et al.* (Spin-Images) [18]	83.7
Rusu *et al.* (FPFH) [27]	61.3
2D ConvNet on Depth	38.5
Ours (3DMatch)	**28.5**

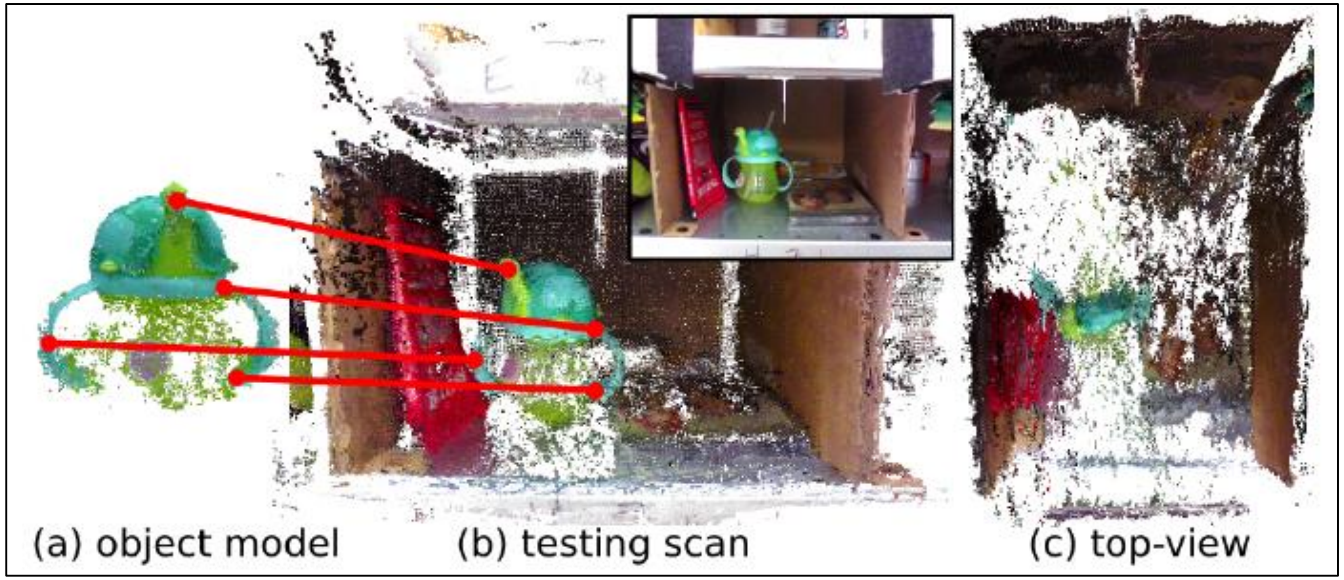
Match classification error at 95% recall

Method	Recall (%)	Precision (%)
Rusu *et al.* [27] + RANSAC	44.2	30.7
Johnson *et al.* [18] + RANSAC	51.8	31.6
Ours + RANSAC	**60.1**	**36.0**

Fragment Alignment Success Rate

Local Shape Descriptor: “3D Match” Results

Result 2: feature descriptor learned from RGB-D reconstructions provides matching for recognizing poses of small objects in Amazon Picking Challenge



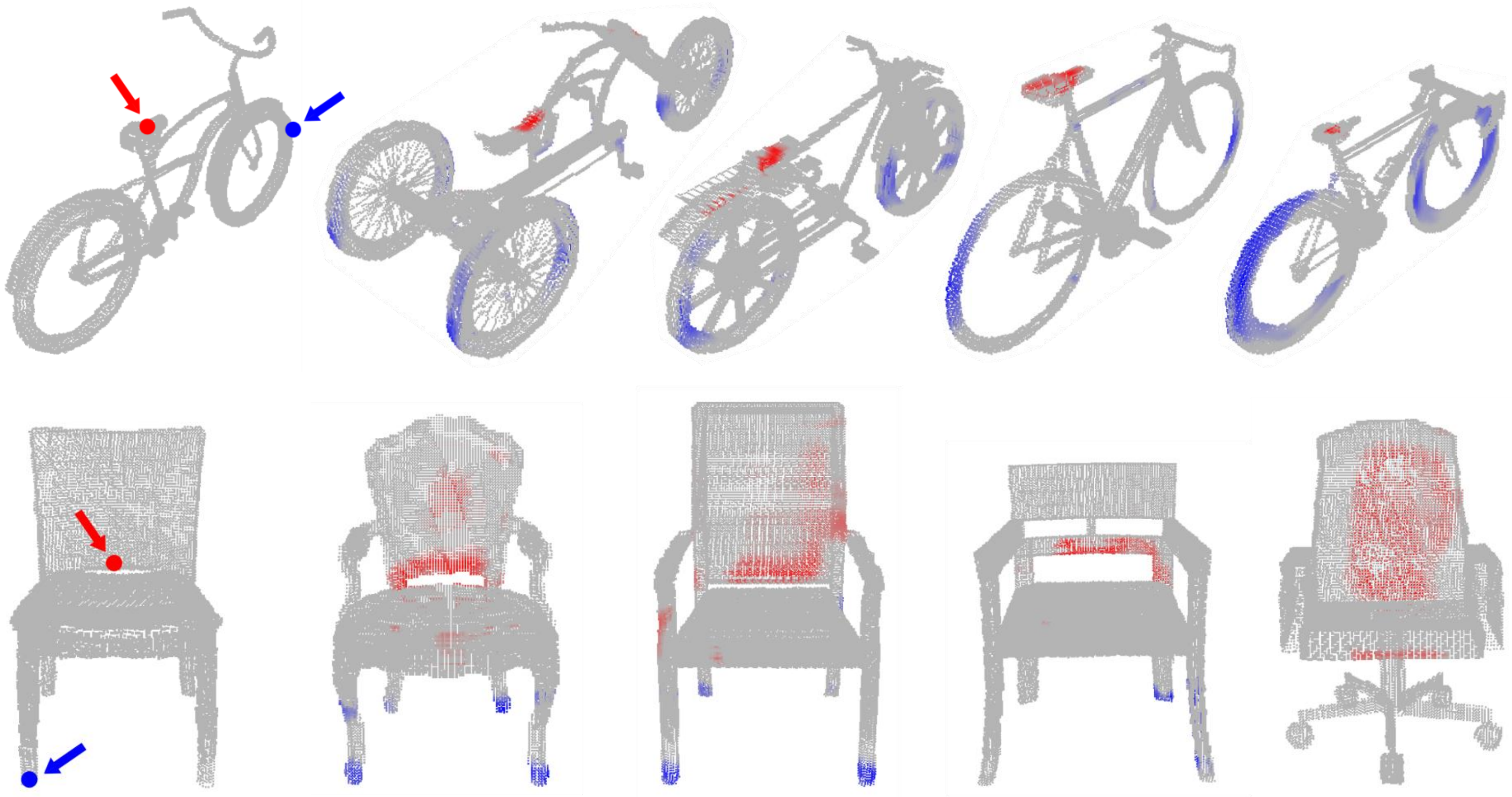
Predicting pose of 3D object model in RGB-D scan

Method	Rotation (%)	Translation (%)
Baseline [41]	49.0	67.6
Johnson *et al.* [18] + RANSAC	45.5	65.9
Rusu *et al.* [27] + RANSAC	43.5	65.6
Ours (no pretrain) + RANSAC	49.3	69.0
Ours + RANSAC	**61.0**	**71.7**

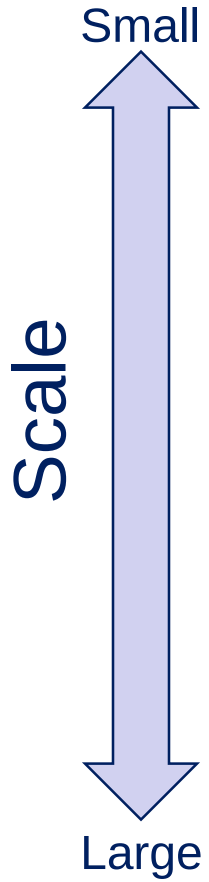
Object pose prediction accuracy

Local Shape Descriptor: “3D Match” Results

Result 3: feature descriptor learned from RGB-D reconstructions provides discriminative matching of semantic correspondences on 3D meshes



Talk Outline



Local Shape Descriptor

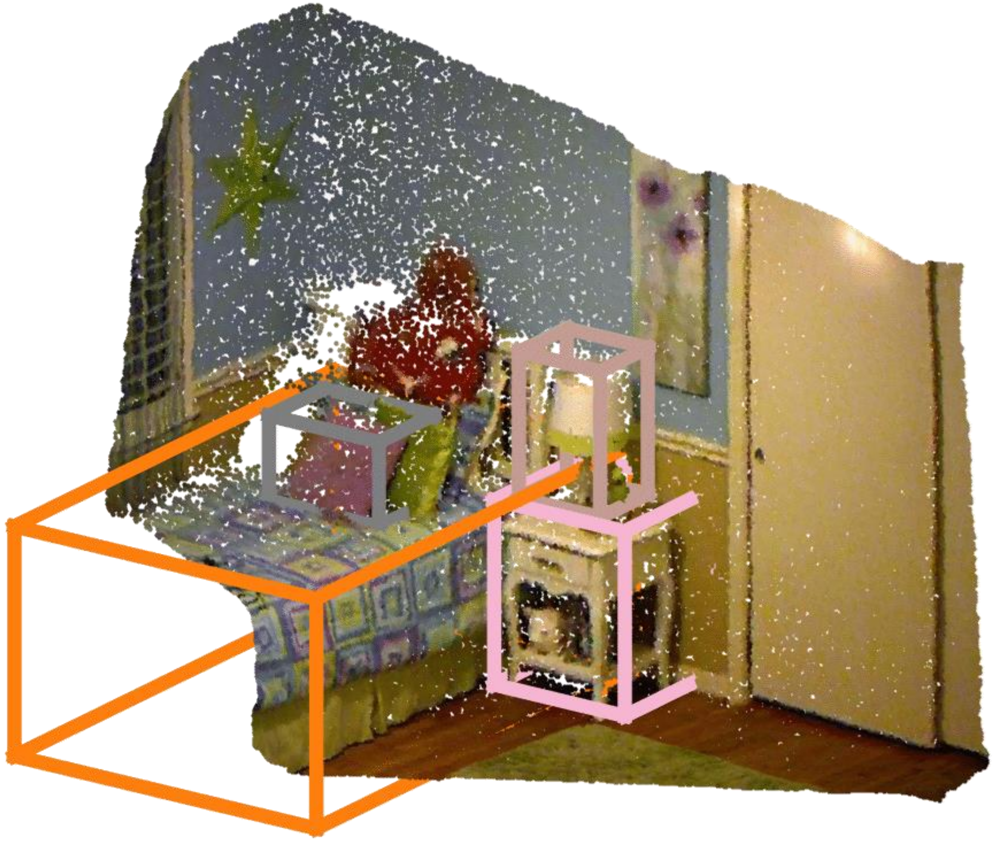
Amodal object detection

Semantic scene completion

S. Song and J. Xiao,
“Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images,”
CVPR 2016

Object Detection

Goal: given a RGB-D image, find objects (labeled 3D amodal bounding boxes)

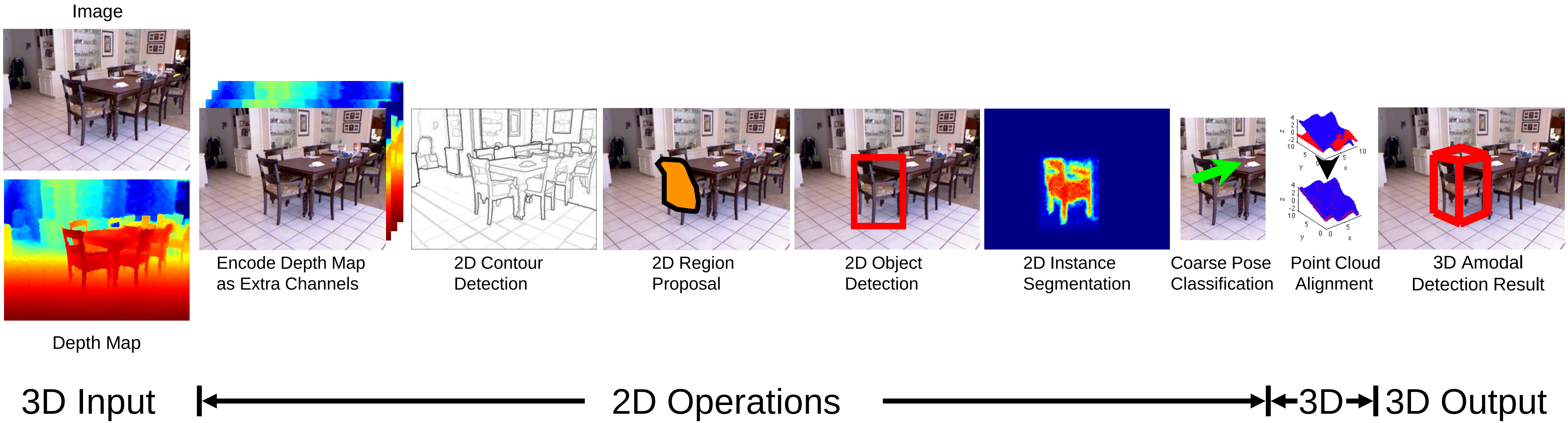


Input: Single RGB-D

Output: labeled 3D Amodal Boxes

Object Detection

Most previous work:

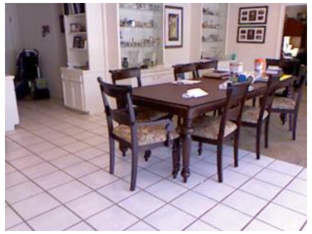


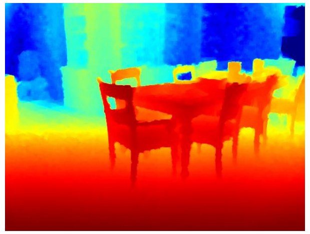
[CVPR13] Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images
[IJCV14] Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and semantic segmentation
[ECCV14] Object Detection and Segmentation using Semantically Rich Image and Depth Features
[CVPR15] Aligning 3D Models to RGB-D Images of Cluttered Scenes
[CVPR16] Cross Modal Distillation for Supervision Transfer

Object Detection: “Deep Sliding Shapes”

Approach:

Image





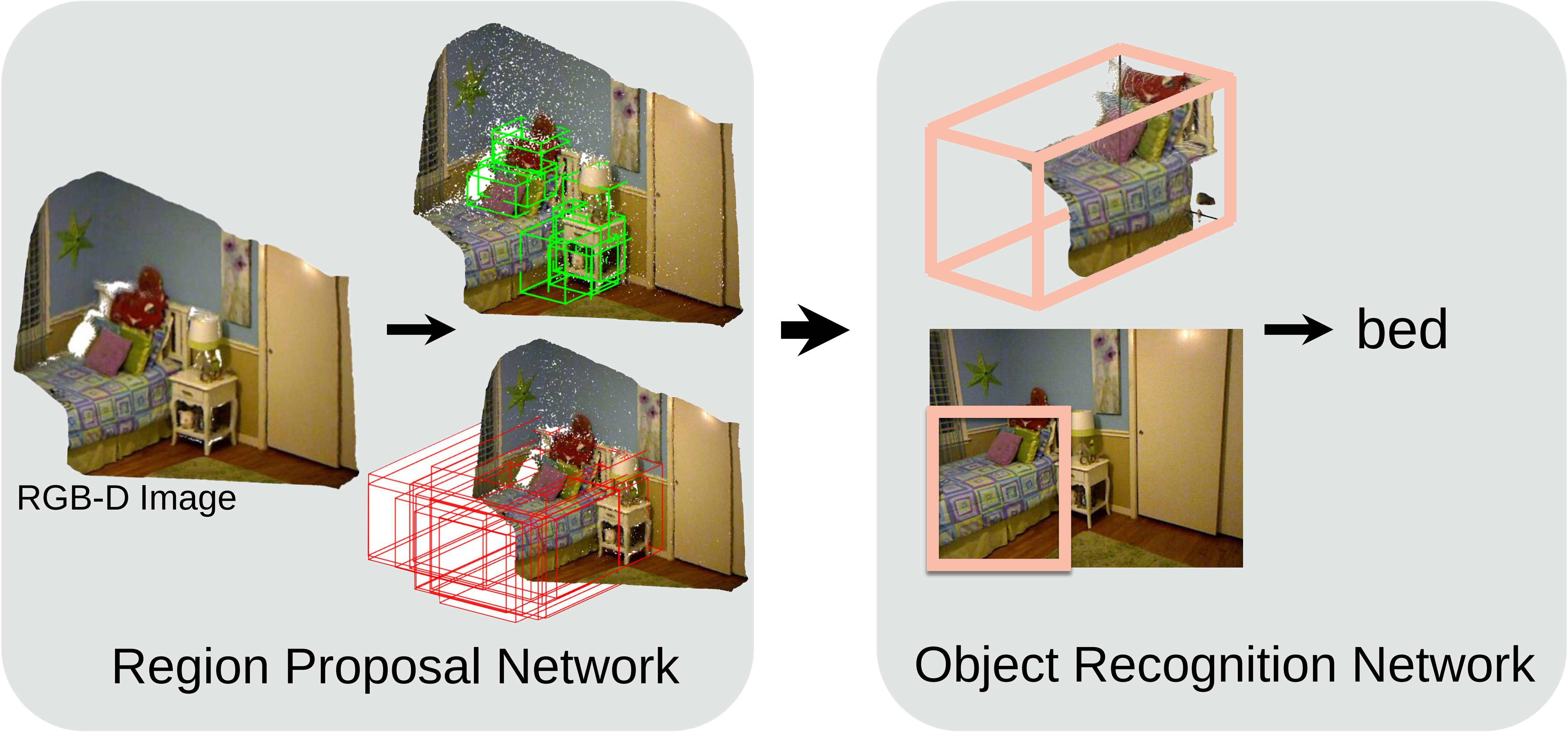
Depth Map

3D Deep Learning

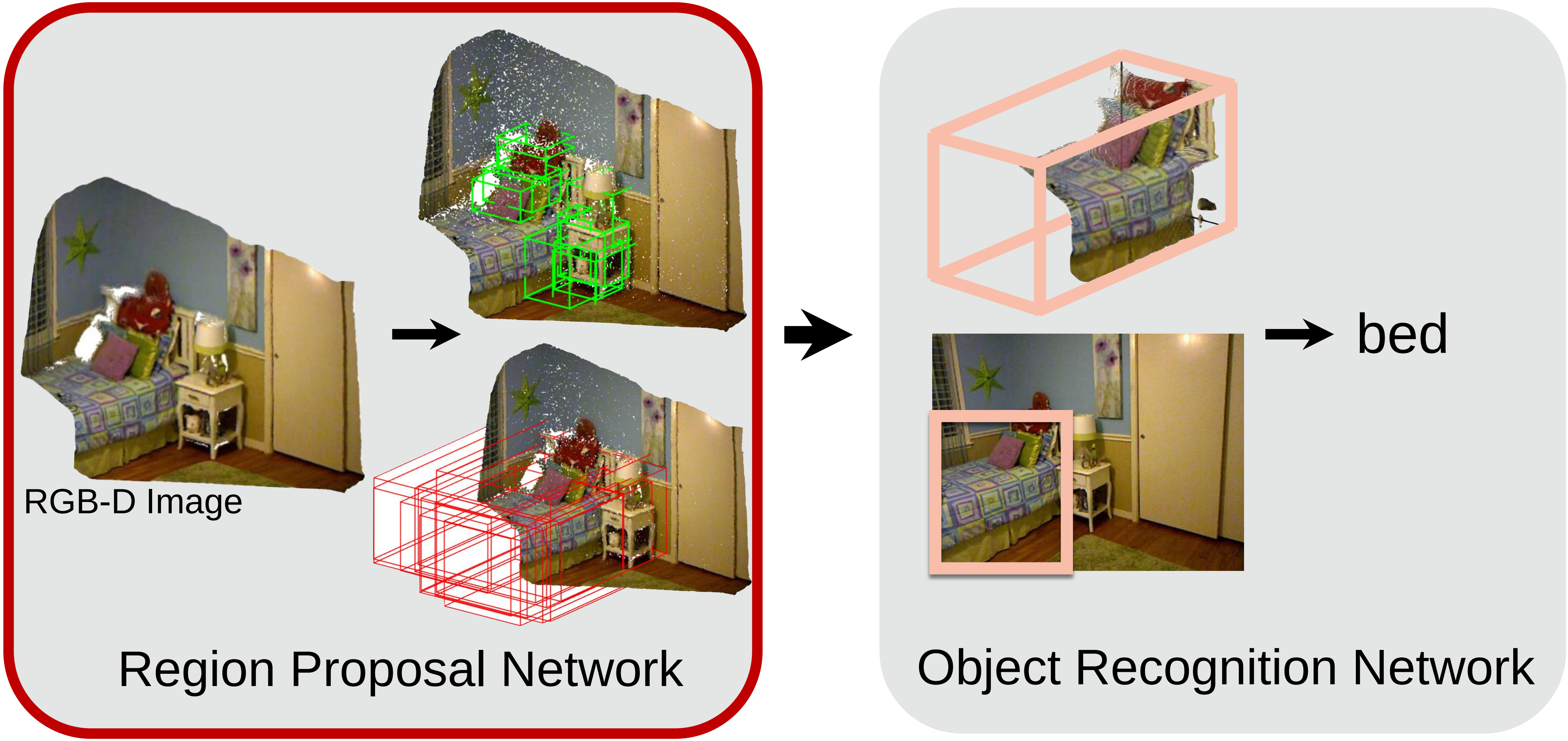
3D Amodal
Detection Result

3D Input ⟷ 3D Operations ⟷ 3D Output

Object Detection: “Deep Sliding Shapes”



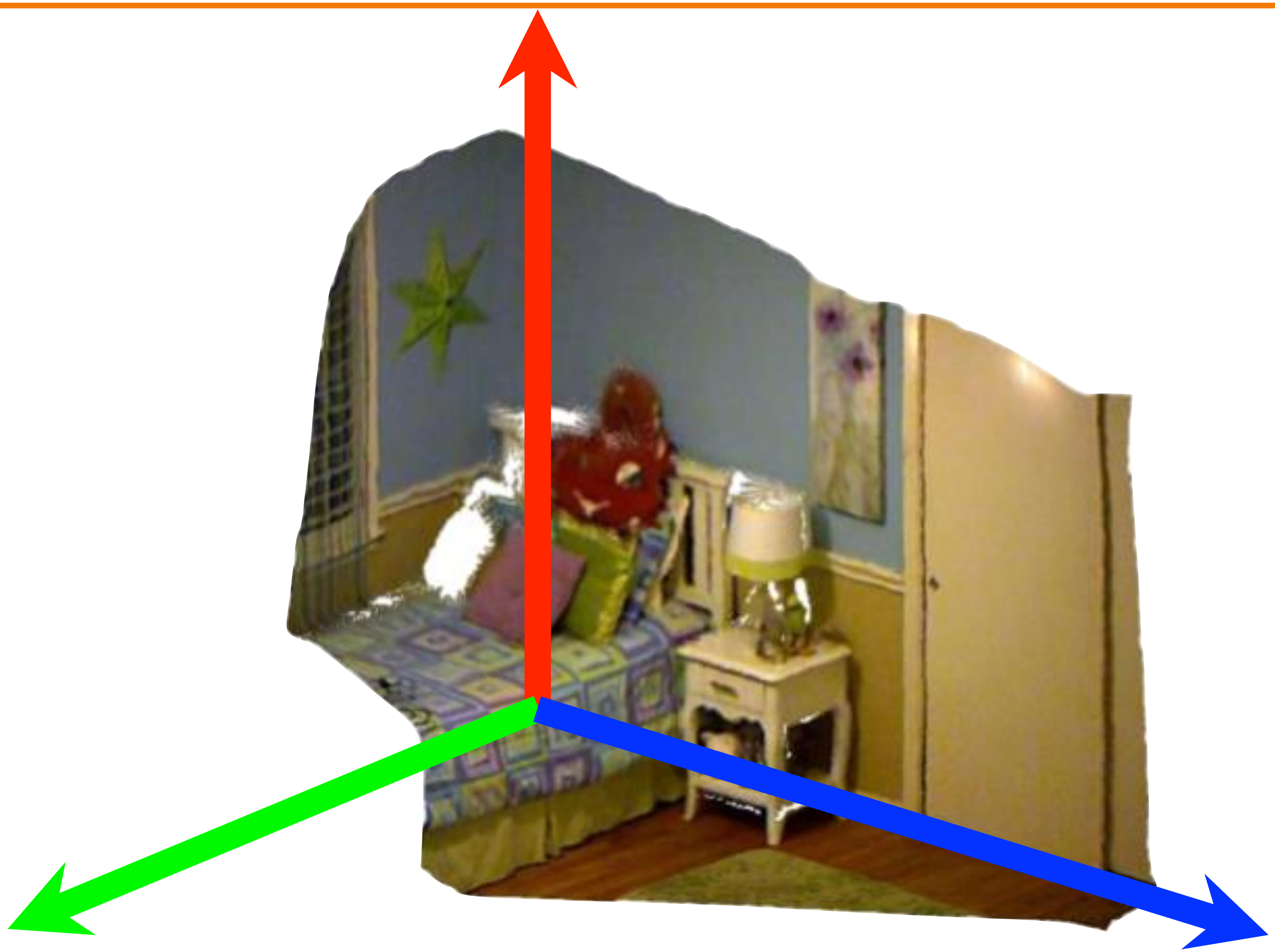
Object Detection: “Deep Sliding Shapes”



Object Detection: “Deep Sliding Shapes”

Data encoding:

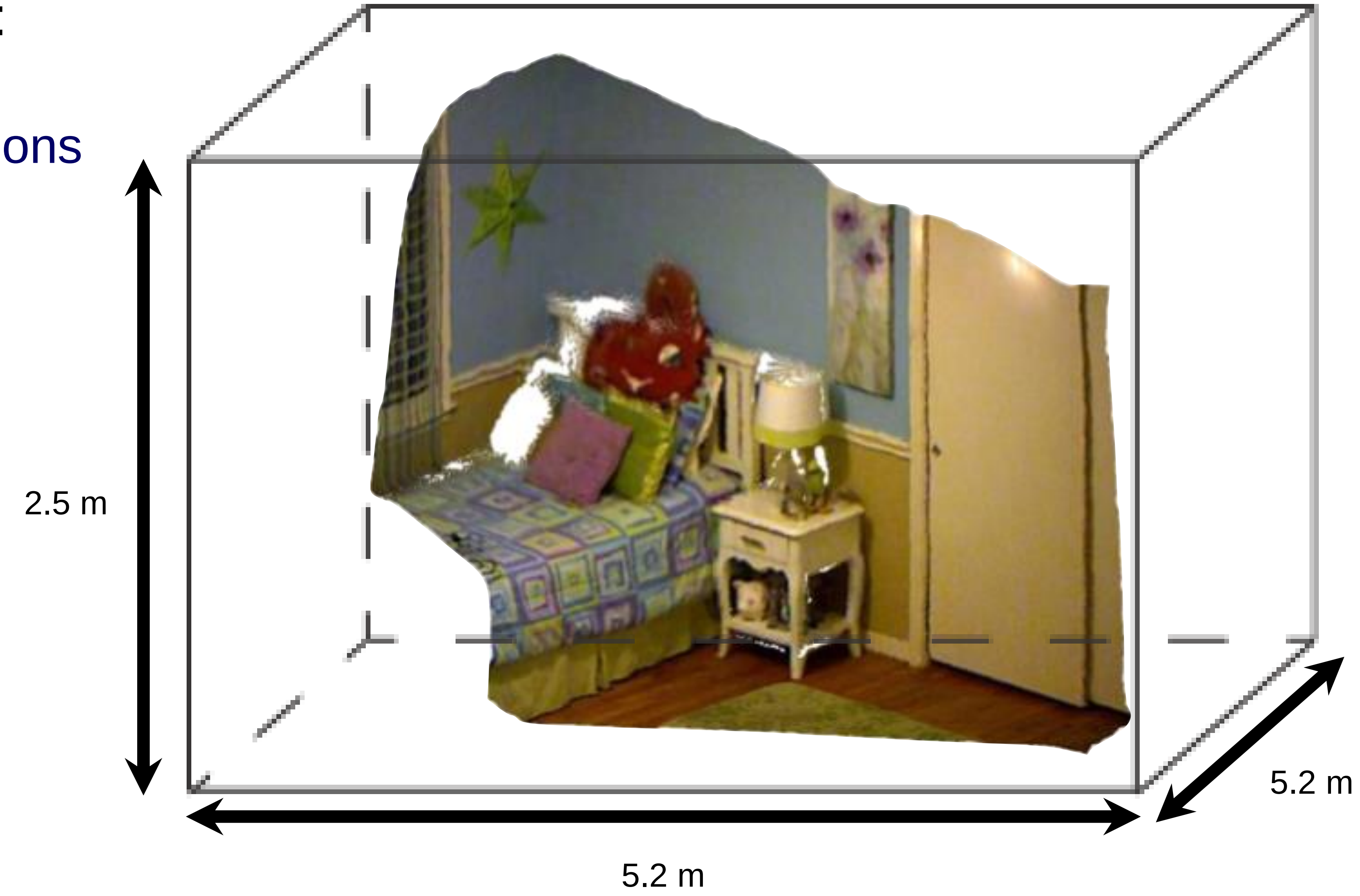
1) Estimate major directions of room
2) Compute TSDF



Object Detection: “Deep Sliding Shapes”

Data encoding:

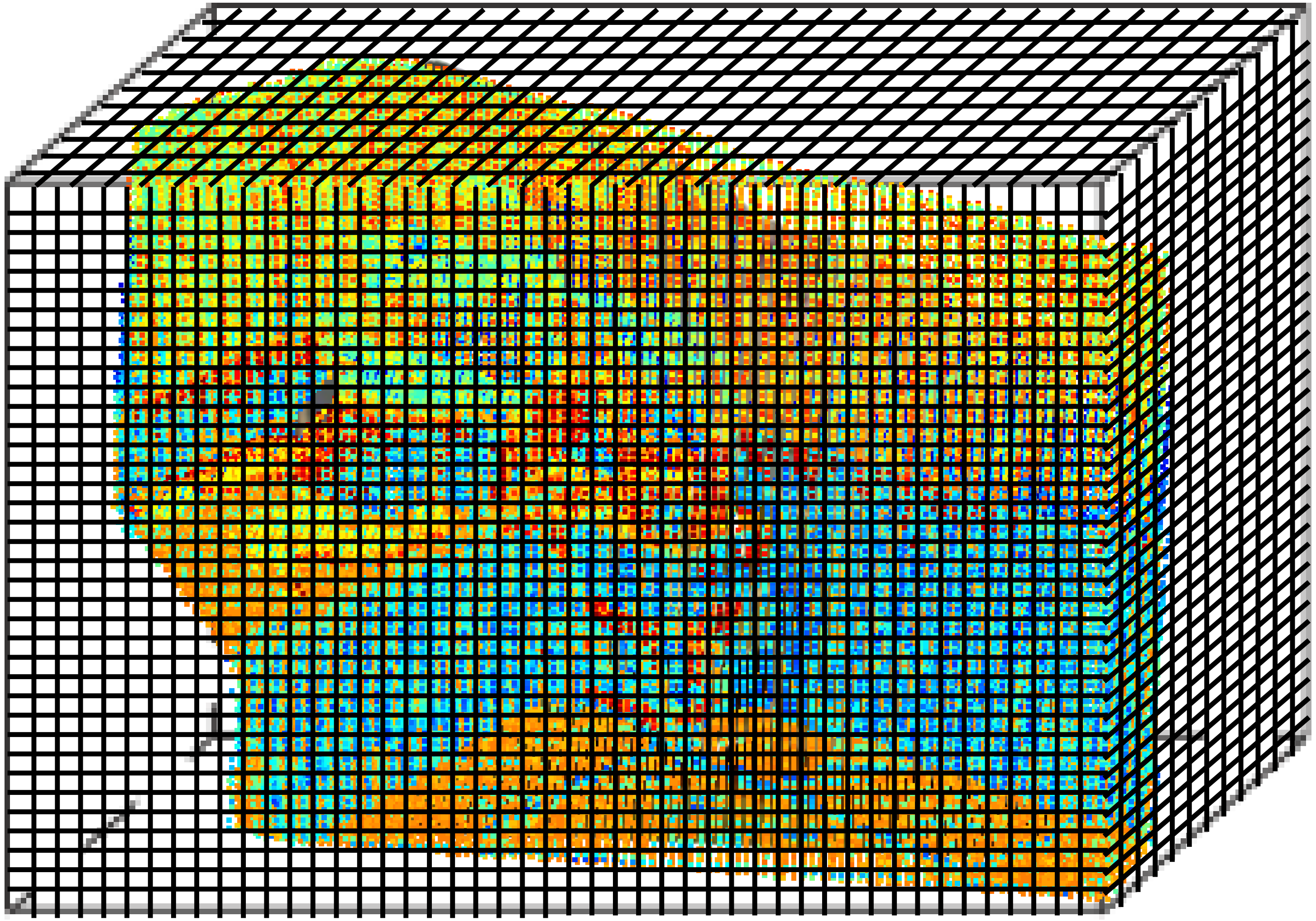
1) Estimate major directions of room
2) Compute TSDF



Object Detection: “Deep Sliding Shapes”

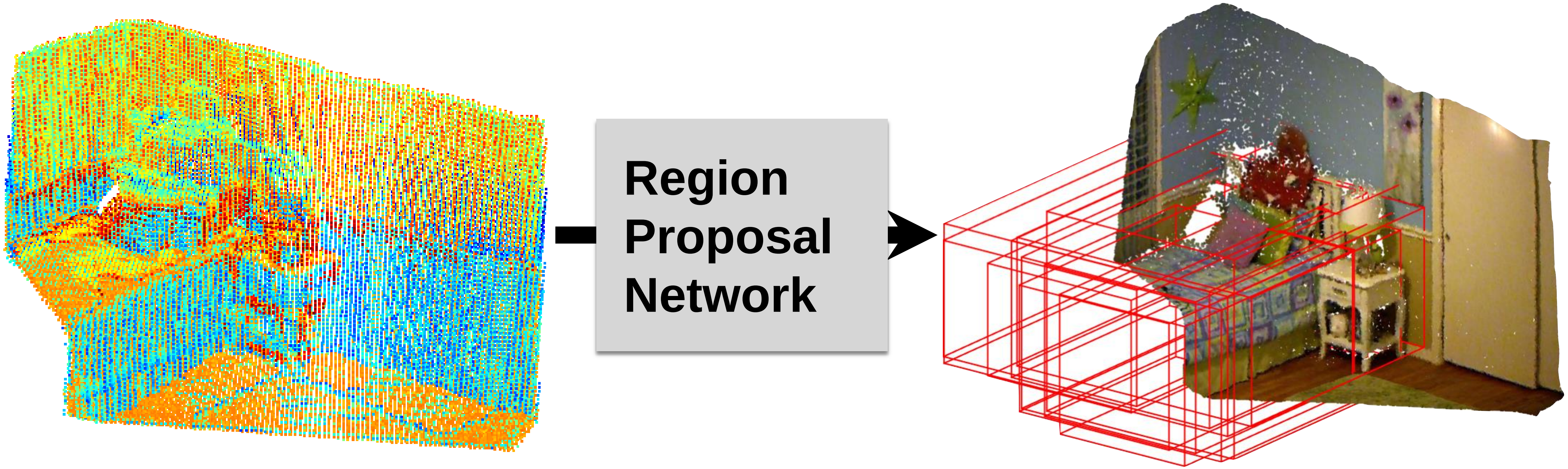
Data encoding:

1) Estimate major directions of room
2) Compute TSDF



Object Detection: “Deep Sliding Shapes”

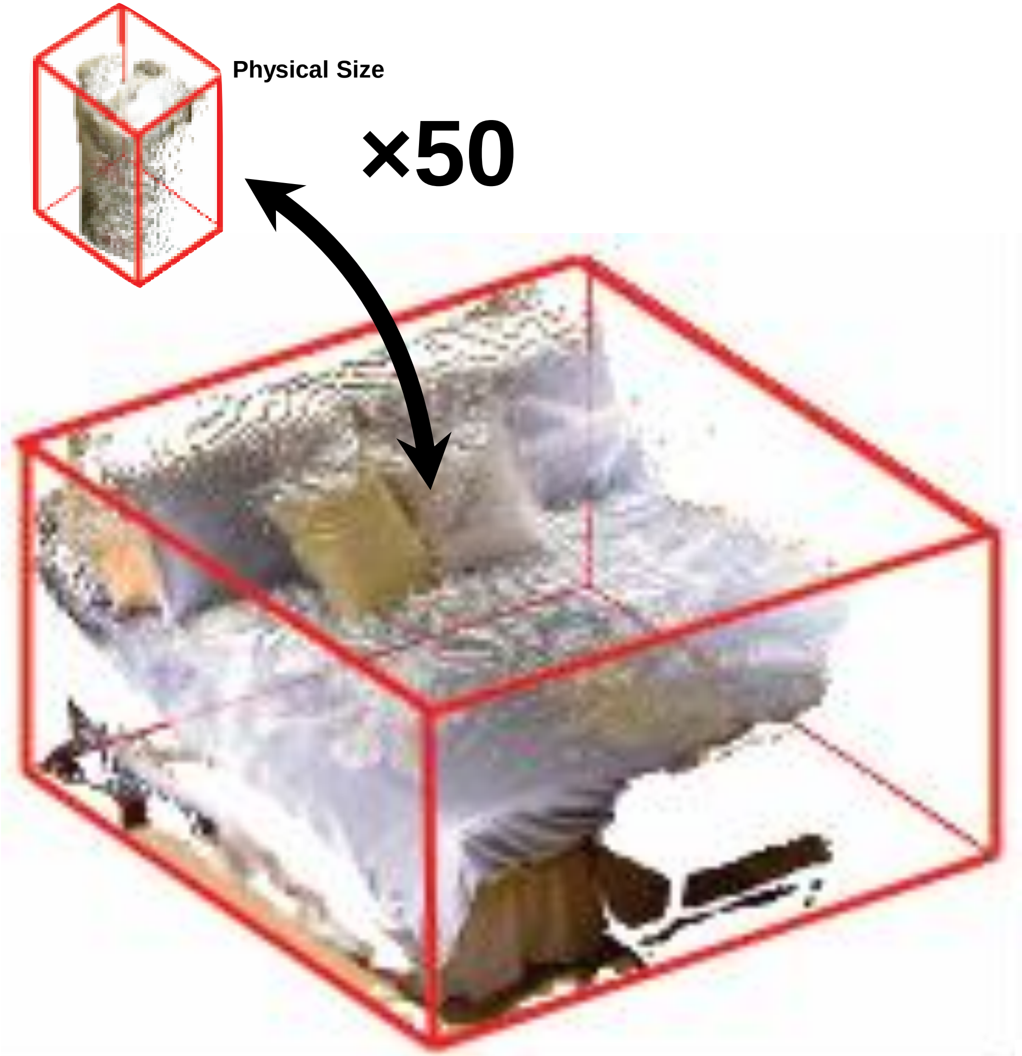
3D region proposal network:



TSDF **3D Region Proposals**

Object Detection: “Deep Sliding Shapes”

3D region proposal network:

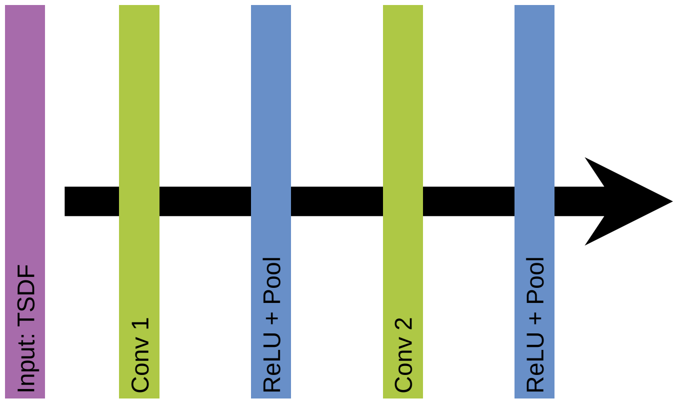


Object Detection: "Deep Sliding Shapes"

Multiscale 3D region proposal network:

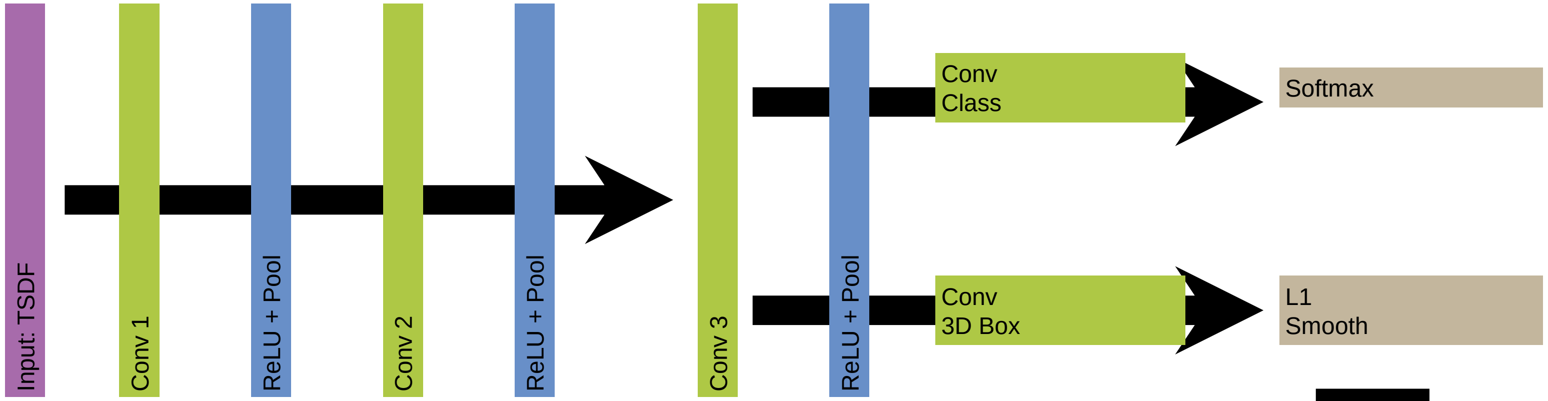
Object Detection: “Deep Sliding Shapes”

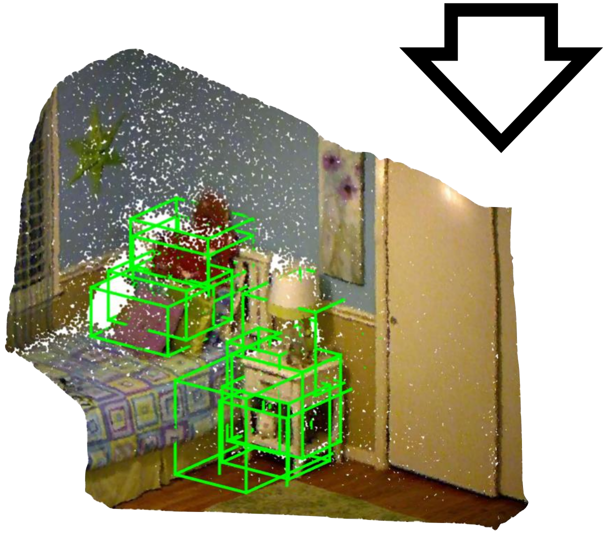
Multiscale 3D region proposal network:



Object Detection: “Deep Sliding Shapes”

Multiscale 3D region proposal network:

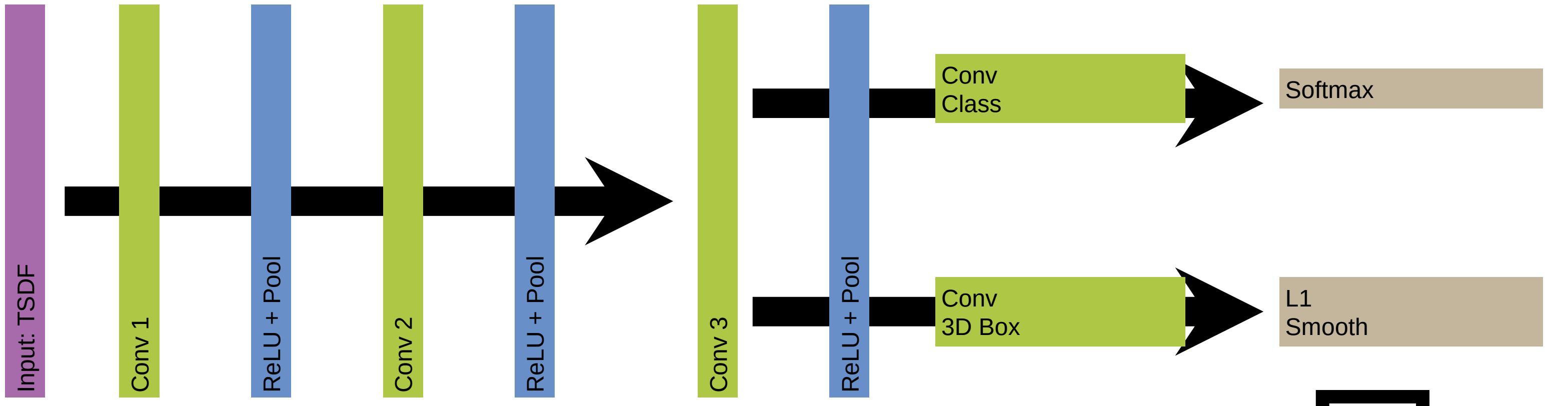


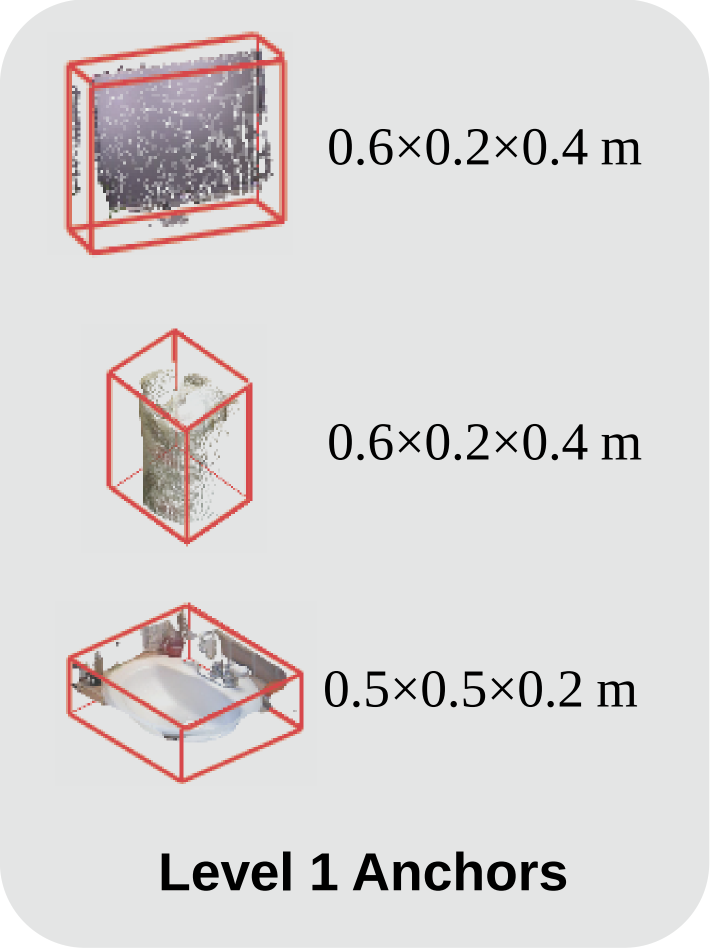


Receptive field: 0.4 m^3

Object Detection: “Deep Sliding Shapes”

Multiscale 3D region proposal network:

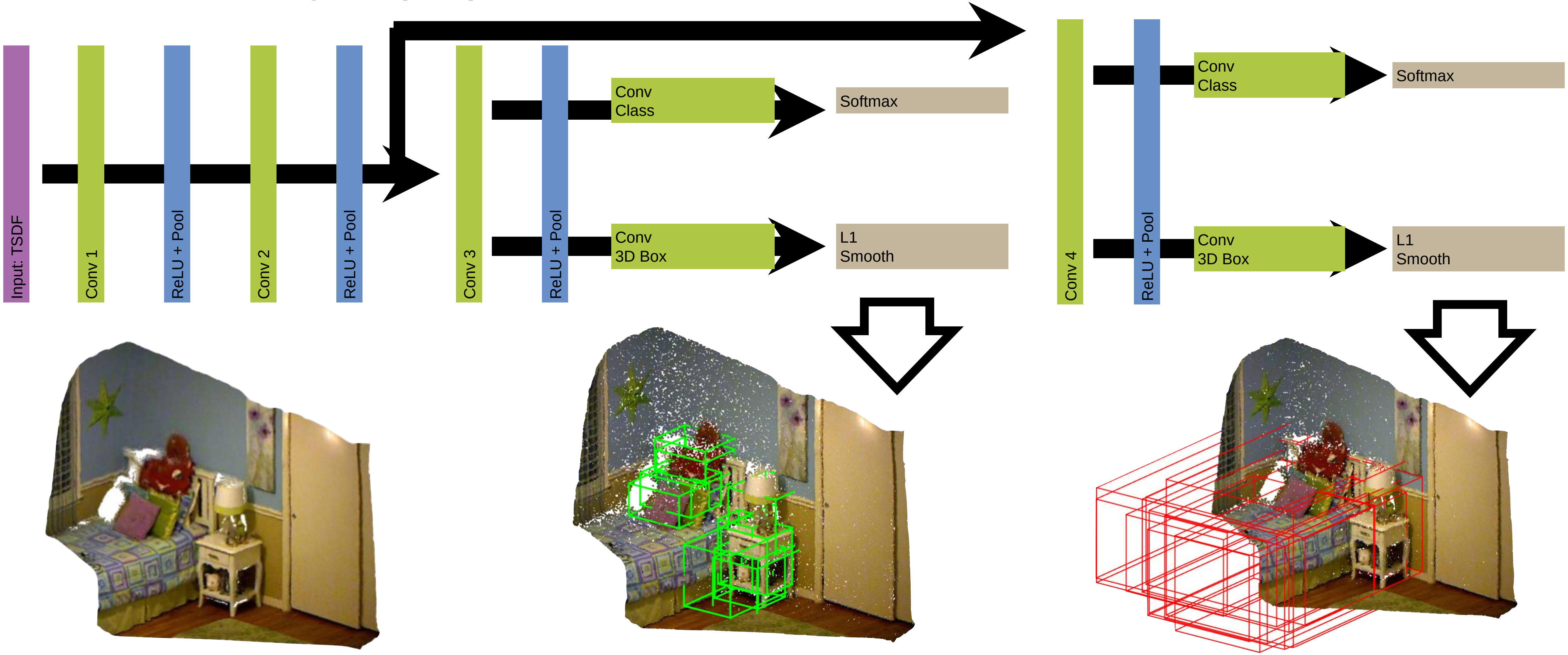




Receptive field: 0.4 m^3

Object Detection: “Deep Sliding Shapes”

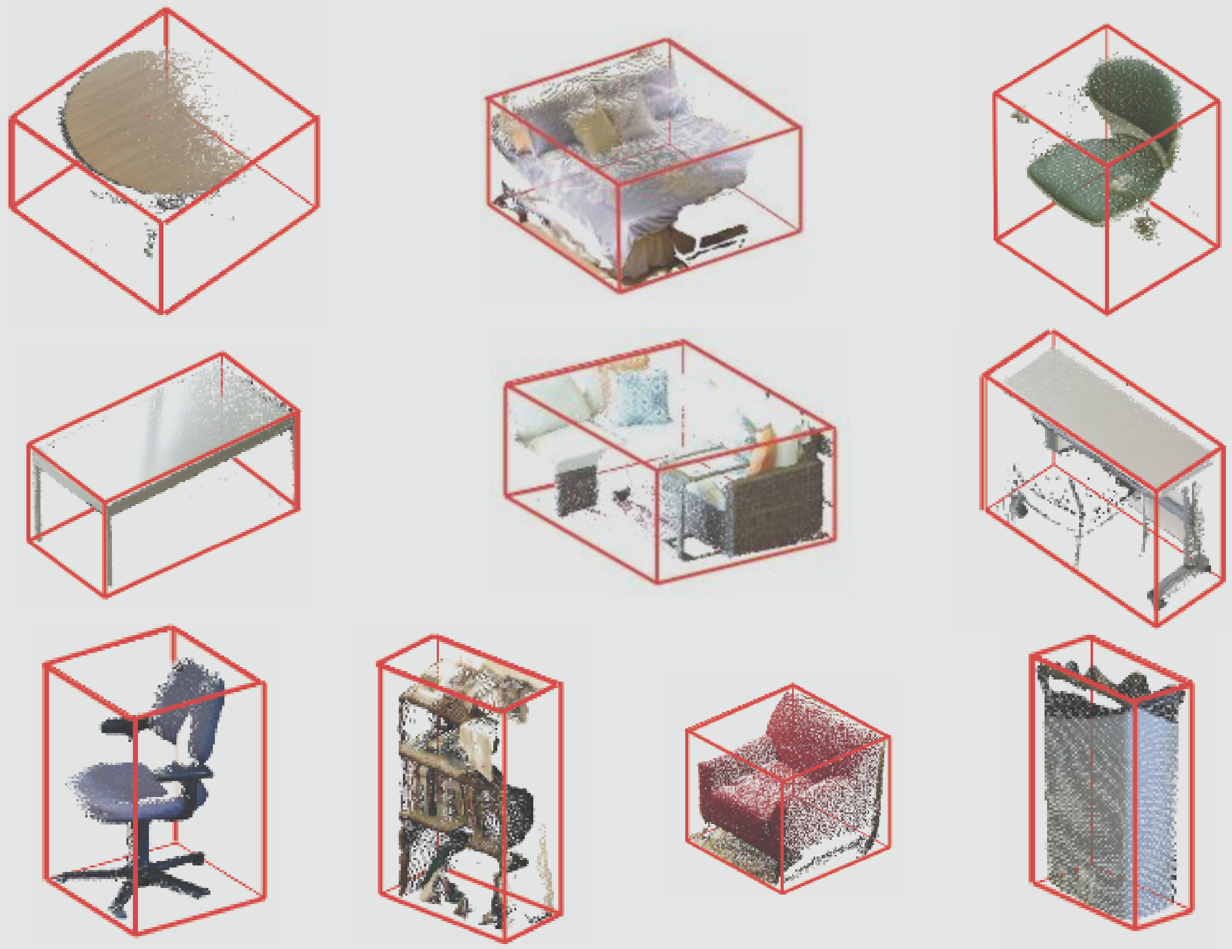
Multiscale 3D region proposal network:



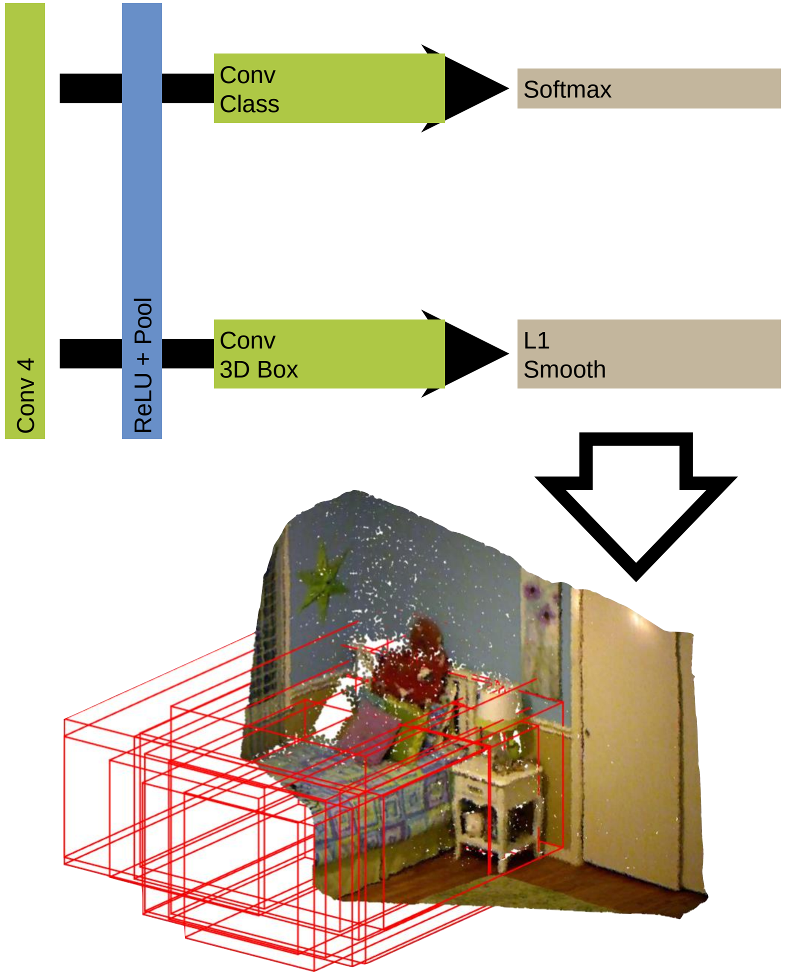
Receptive field: 0.4 m^3

Receptive field: 1 m^3

Object Detection: “Deep Sliding Shapes”

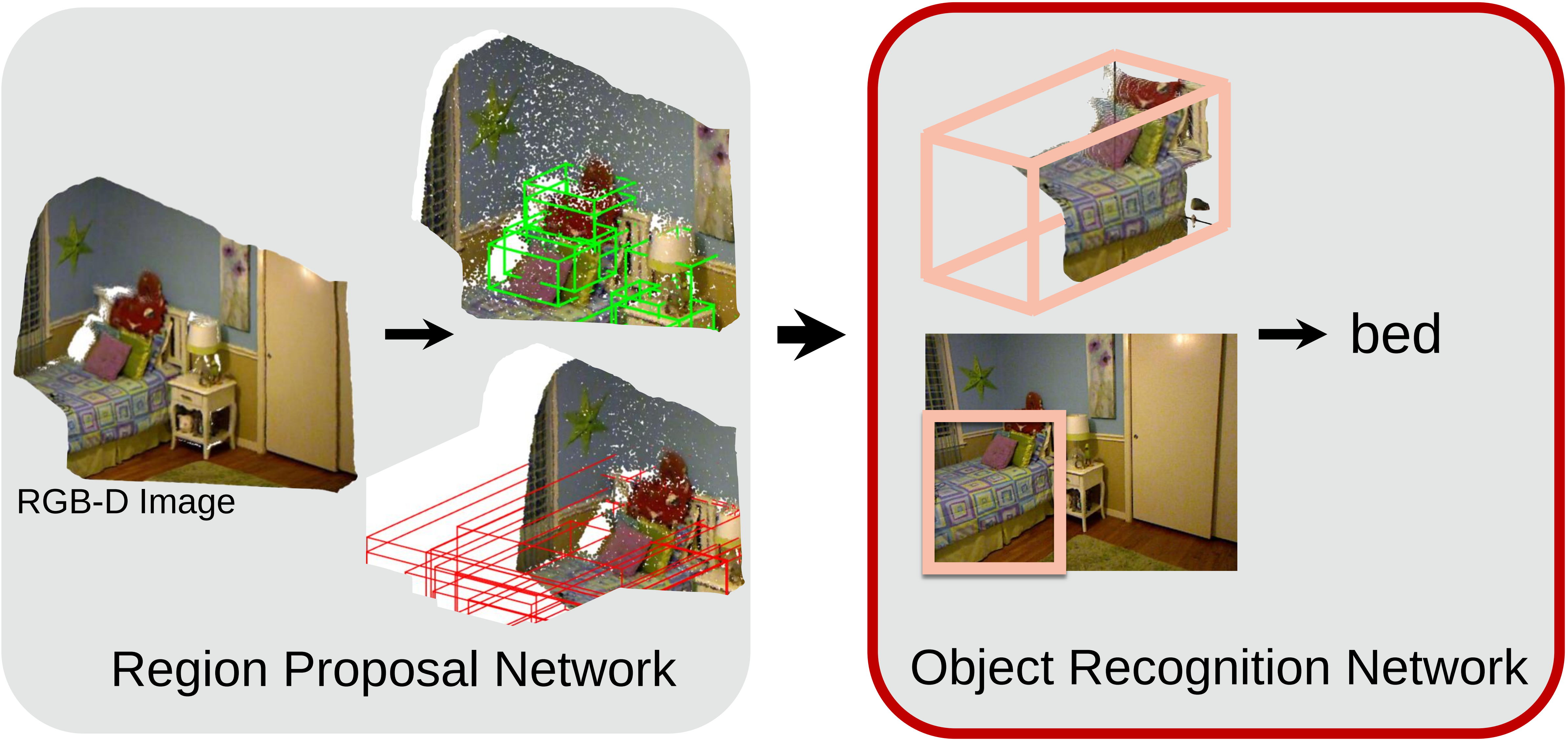


Level 2 Anchors



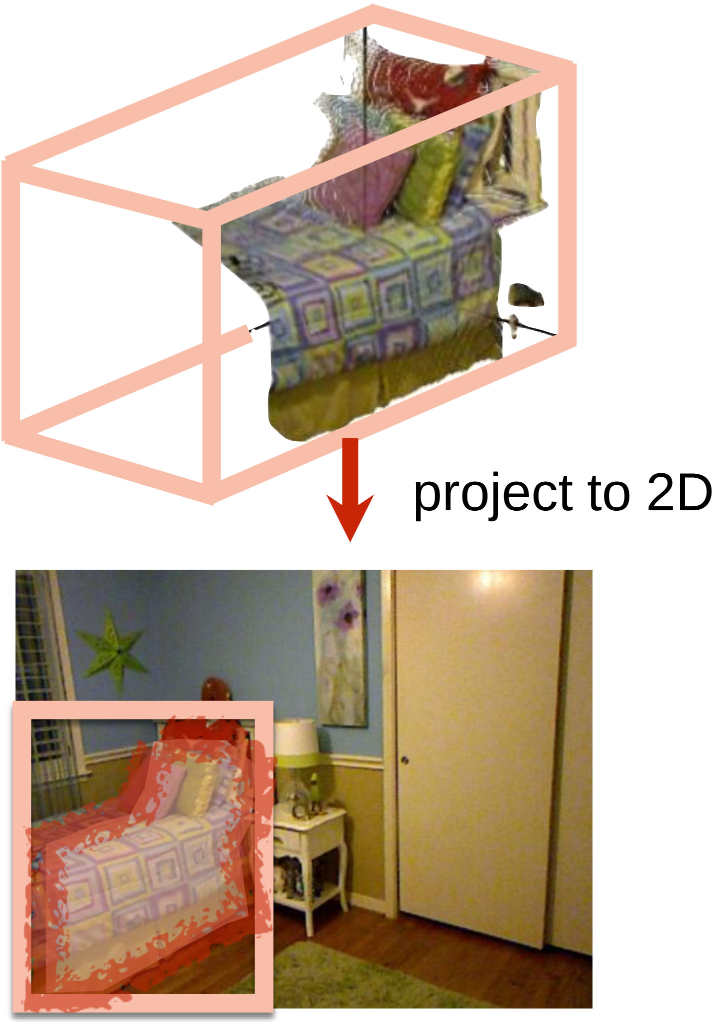
Receptive field: 1 m^3

Object Detection: “Deep Sliding Shapes”



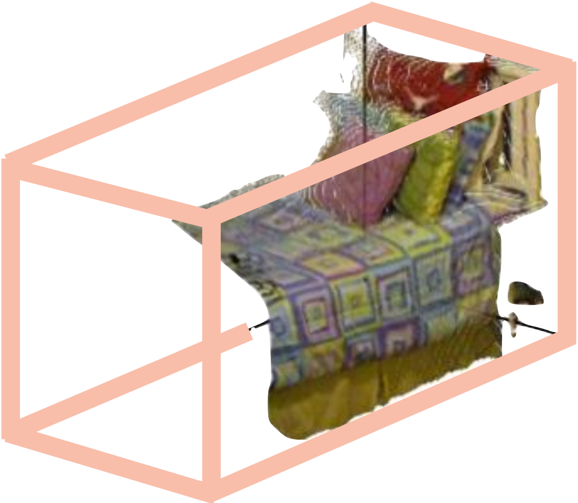
Object Detection: “Deep Sliding Shapes”

Joint object recognition network:

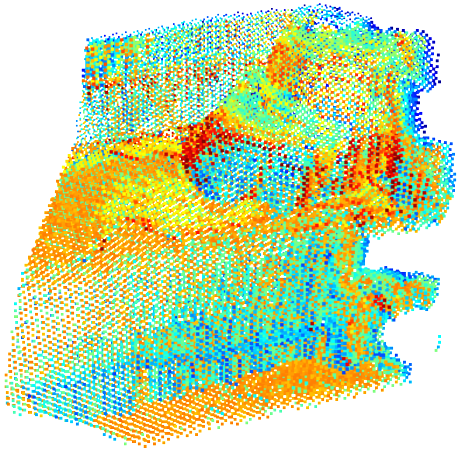


Object Detection: “Deep Sliding Shapes”

Joint object recognition network:



→



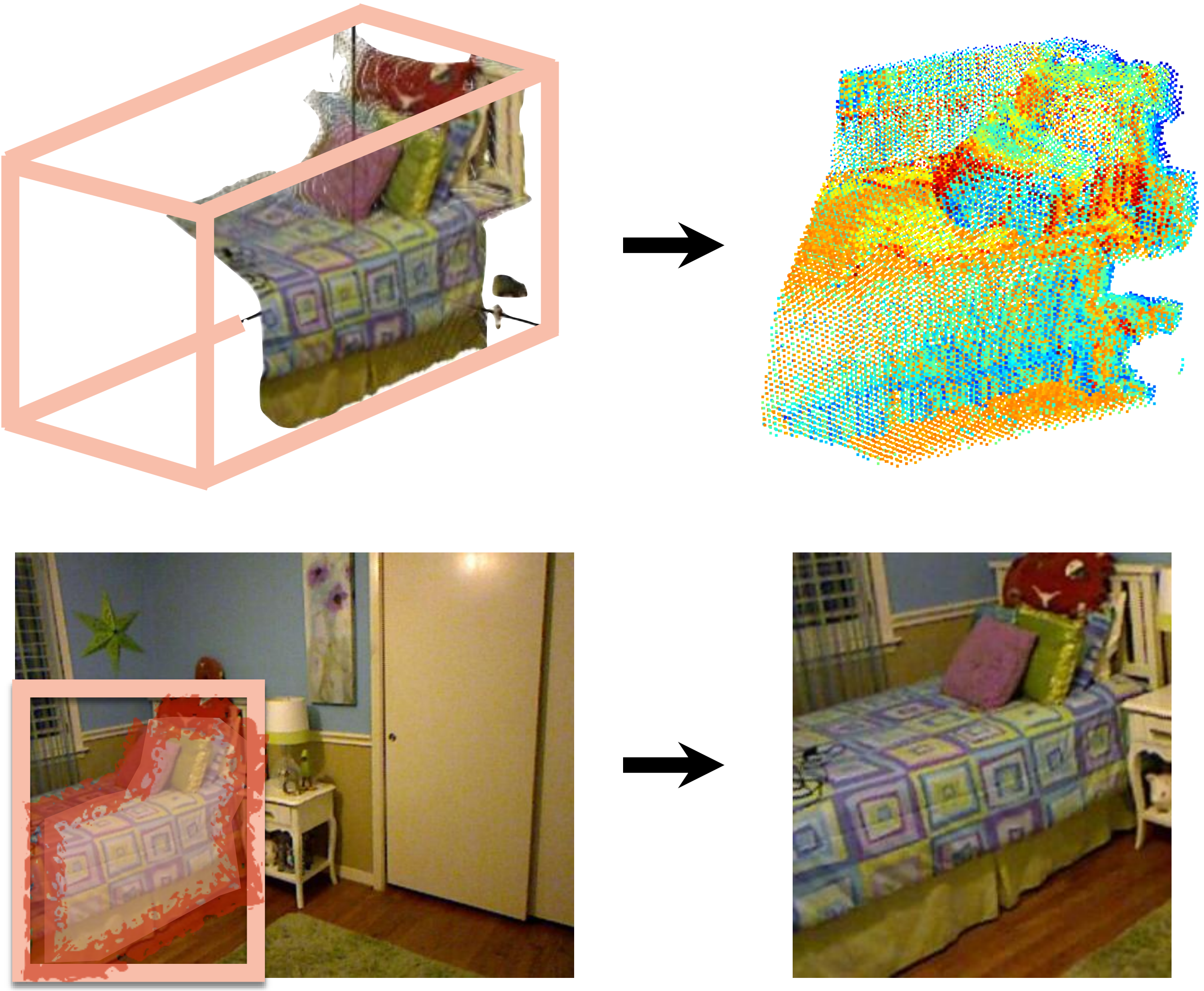
TSDF

→

Image Patch

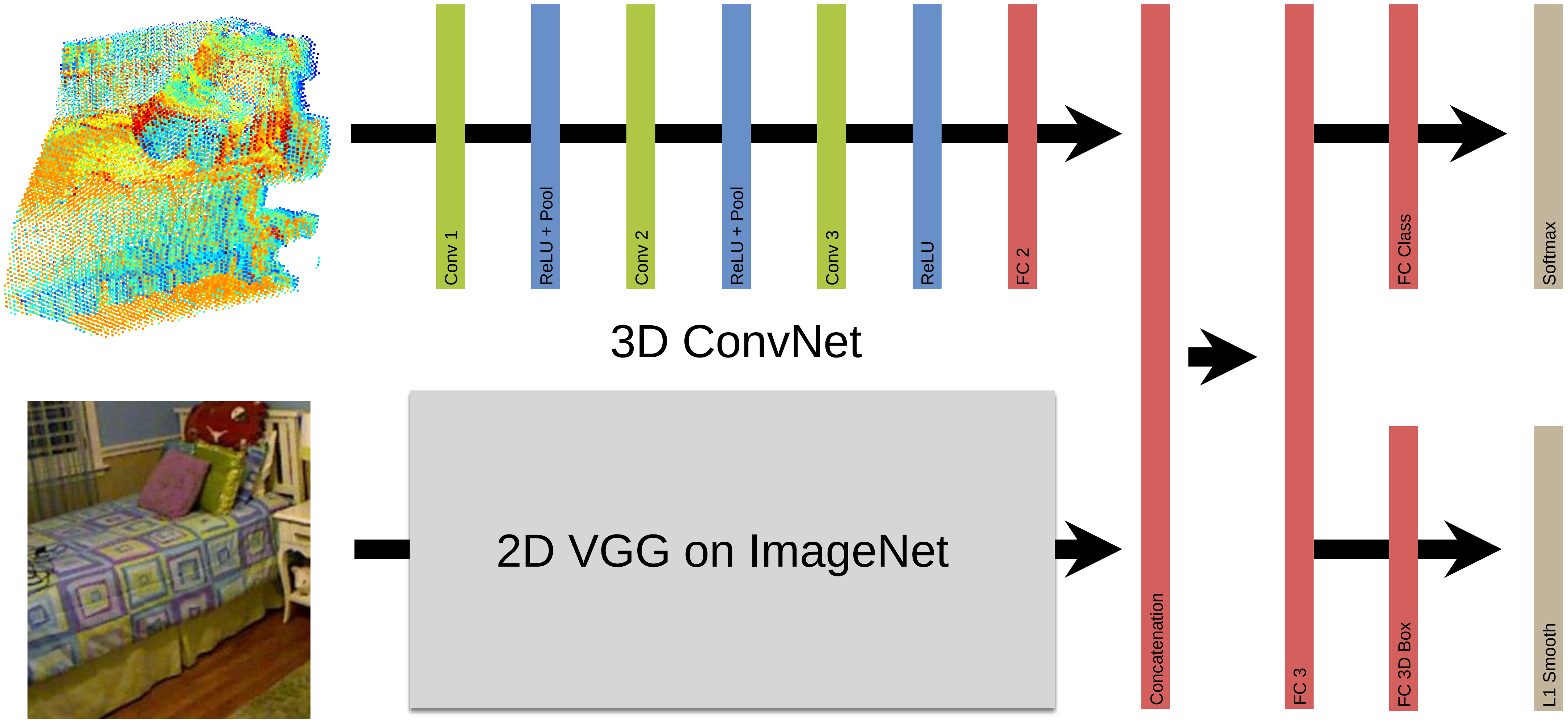
Object Detection: “Deep Sliding Shapes”

Joint object recognition network:



Object Detection: “Deep Sliding Shapes”

Joint object recognition network:



Object Detection: “Deep Sliding Shapes” Experiments

Train and test on amodal boxes provided in SUN RGB-D

S. Song, S. Lichtenberg, and J. Xiao, “SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite,” CVPR 2015

Object Detection: “Deep Sliding Shapes” Results

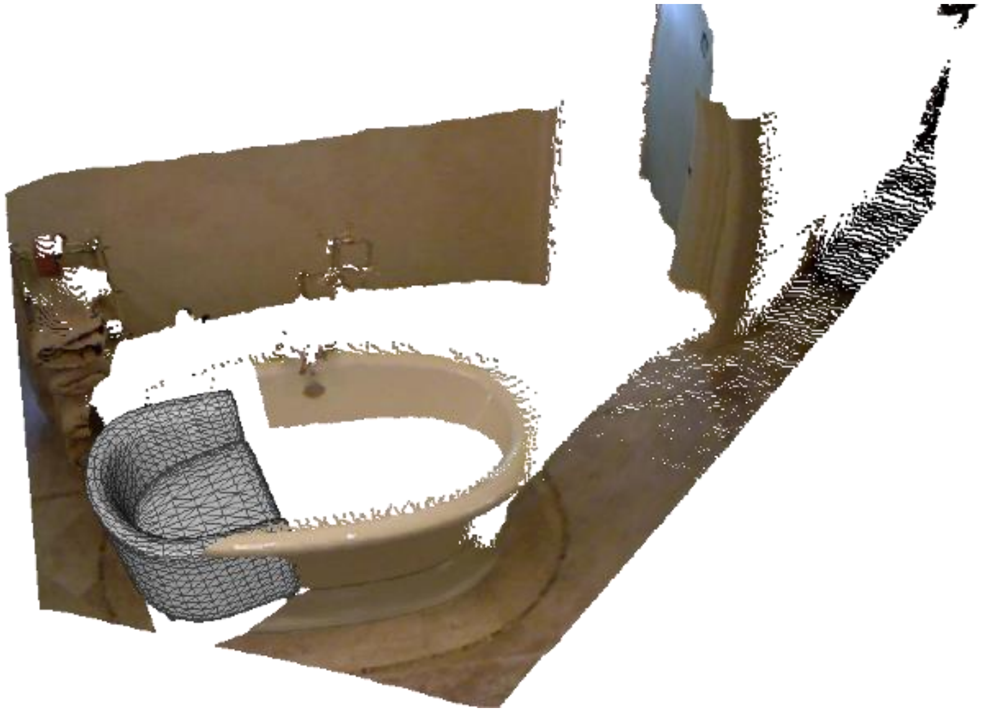
Quantitative comparisons:

	Algorithm	Input						mAP
3D Non-Deep Learning	Sliding Shapes	Depth	33.5	29	34.5	33.8	67.3	39.6
2D Deep Learning	Depth-RCNN (segment)	Depth	71	18.2	49.6	30.4	63.4	46.5
	Depth-RCNN (segment)	RGB-D	74.7	18.6	50.3	28.6	69.7	48.4
	Depth-RCNN (CAD fit)	Depth	72.7	47.5	54.6	40.6	72.7	57.6
	Depth-RCNN (CAD fit)	RGB-D	73.4	44.2	57.2	33.4	84.5	58.5
3D Deep Learning	Ours	Depth	83.0	58.8	68.6	49.5	79.2	67.8
	Ours	RGB-D	**84.7**	**61.1**	**70.5**	**55.4**	**89.9**	**72.3**

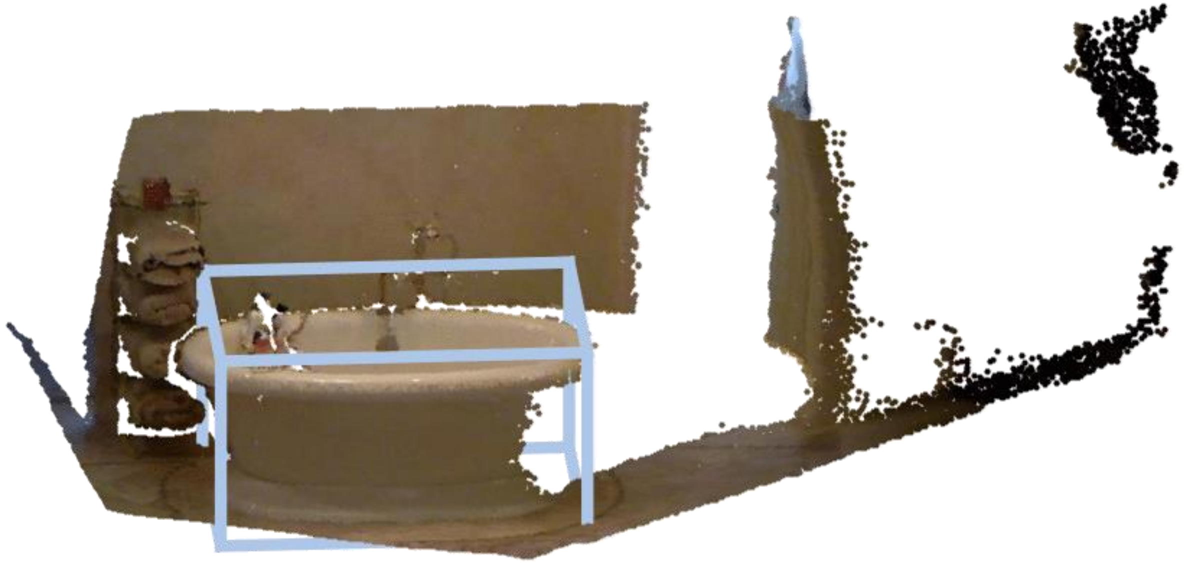
Object detection accuracy on NYU v2 dataset (mAP)

Object Detection: “Deep Sliding Shapes” Results

Qualitative comparisons:



Sliding Shapes: sofa

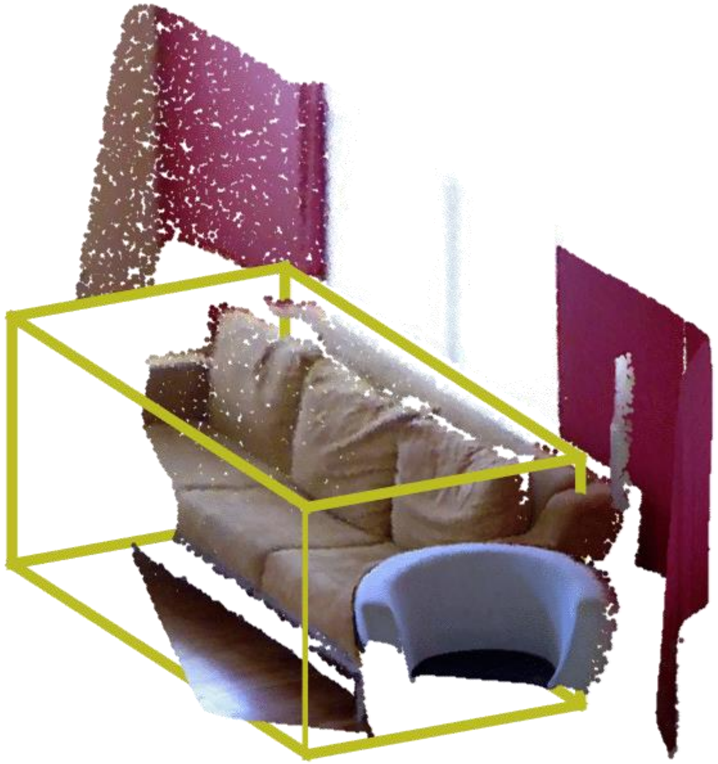


Ours: bathtub

Object Detection: “Deep Sliding Shapes” Results

Qualitative comparisons:

Sliding Shapes: chair



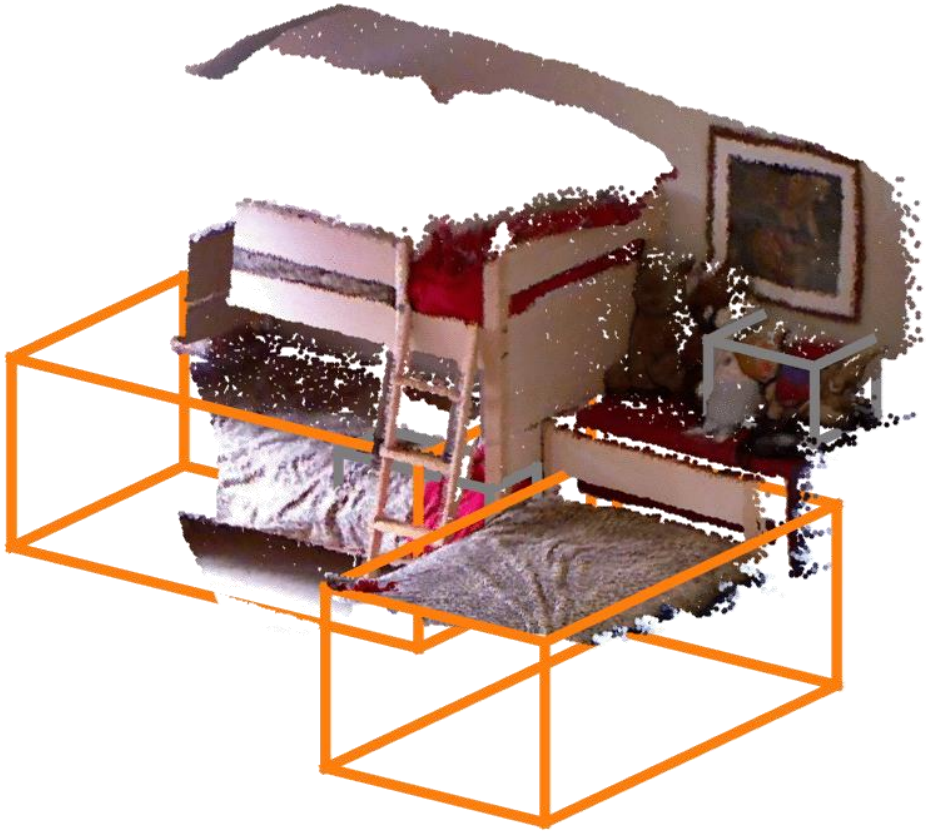
Ours: sofa

sofa bed bathtub garbage bin chair table night stand lamp pillow sink toilet bookshelf

Object Detection: “Deep Sliding Shapes” Results

Qualitative comparisons:

Sliding Shapes: table



Ours: bed

Object Detection: “Deep Sliding Shapes” Results

Qualitative comparisons:

Sliding Shapes: miss

Ours: table and chairs

sofa | bed | bathtub | garbage bin | chair | table | night stand | lamp | pillow | sink | toilet | bookshelf

Object Detection: “Deep Sliding Shapes” Results

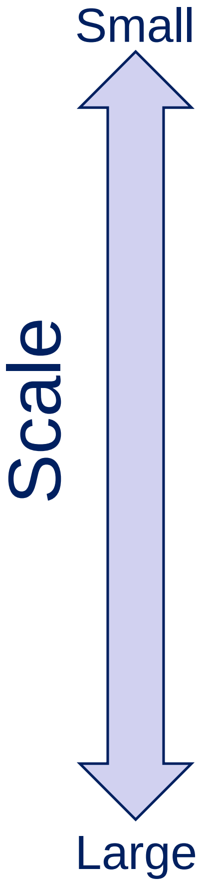
Qualitative comparisons:

Sliding Shapes: toilet

Ours: garbage bin+bed

sofa bed bathtub garbage bin chair table

Talk Outline



Local Shape Descriptor

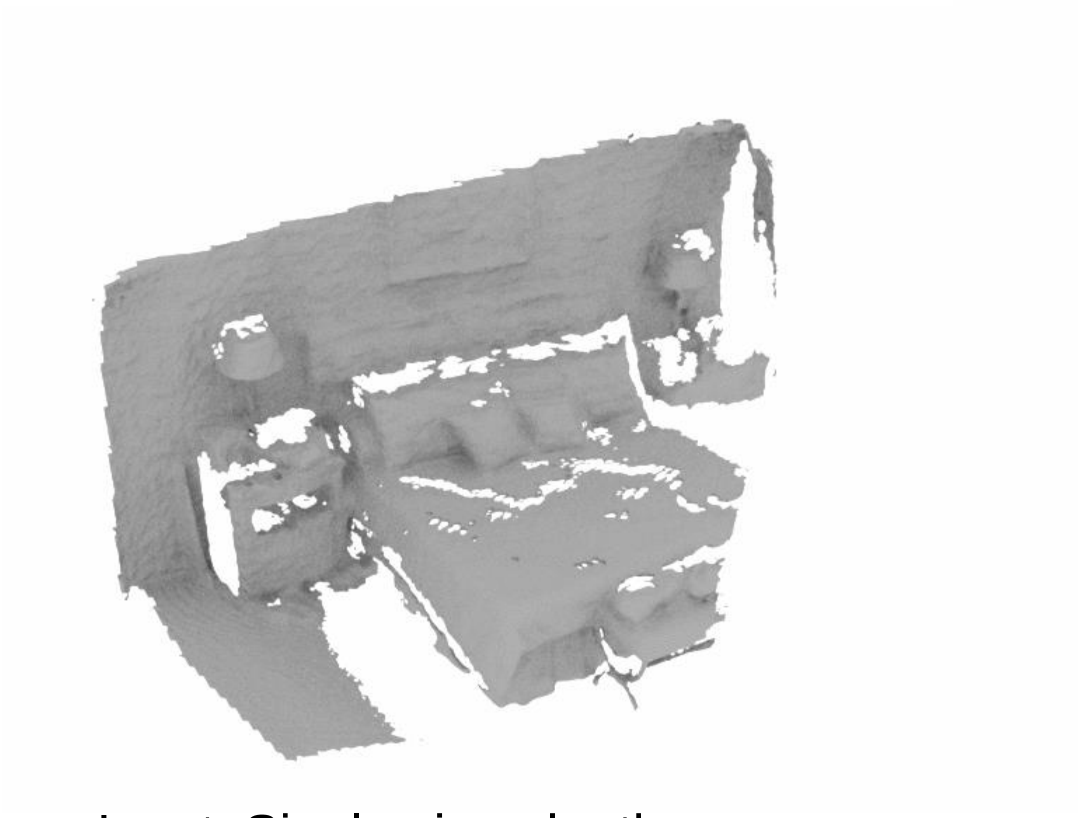
Amodal object detection

Semantic scene completion

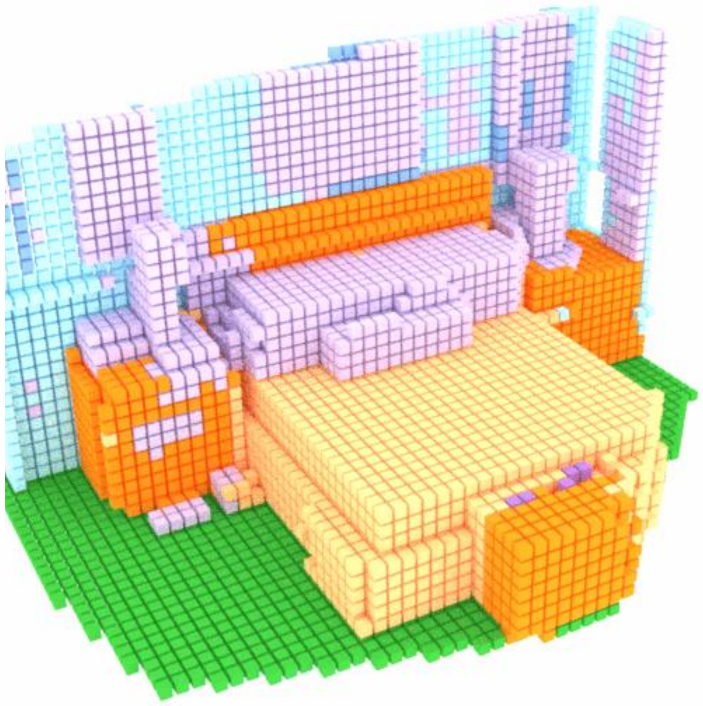
S. Song, F. Yu, A. Zeng, A. Chang, M. Savva, and T. Funkhouser, "Semantic Scene Completion from a Single Depth Image," submitted to CVPR 2017

Semantic Scene Completion

Goal: given an RGB-D image, label all voxels by semantic class



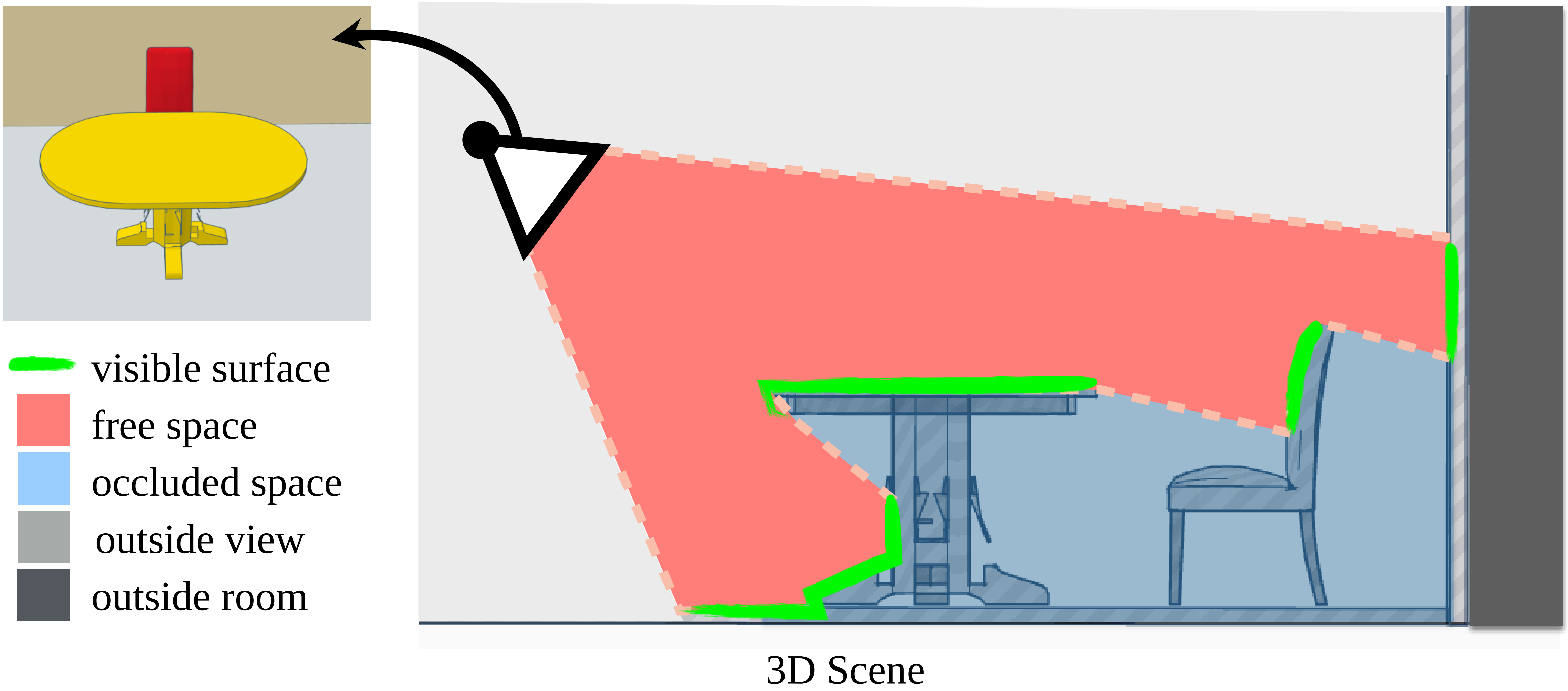
Input: Single view depth map



Output: Semantic scene completion

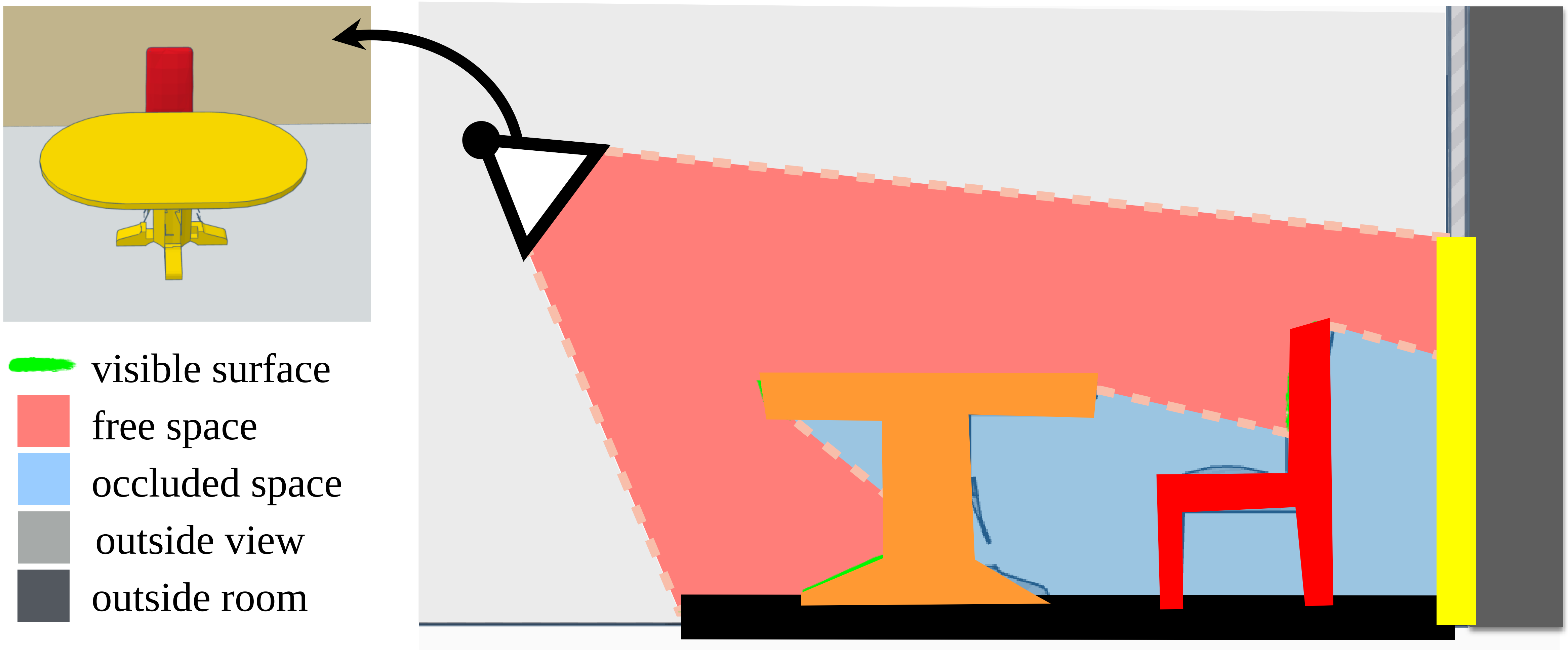
Semantic Scene Completion

Goal: given an RGB-D image, label all voxels by semantic class



Semantic Scene Completion

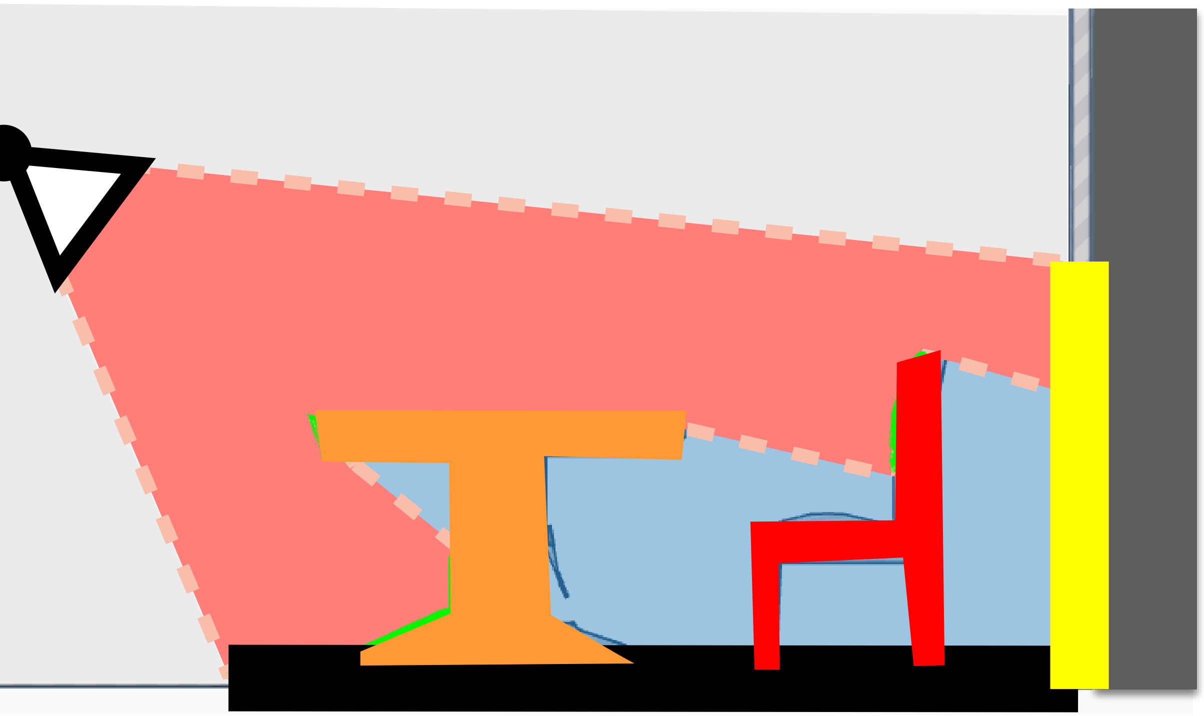
Goal: given an RGB-D image, label all voxels by semantic class



3D Scene

Semantic Scene Completion

Prior work: segmentation **OR** completion

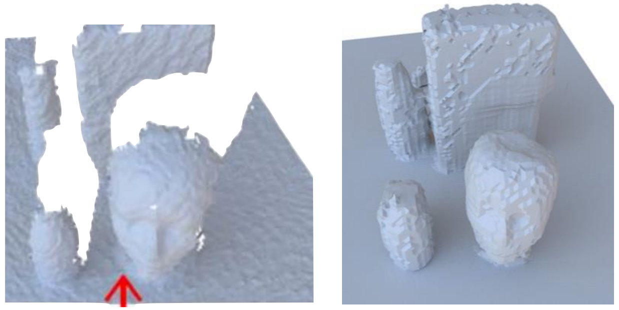


3D Scene

surface segmentation

Silberman *et al.*

scene completion



Firman *et al.*

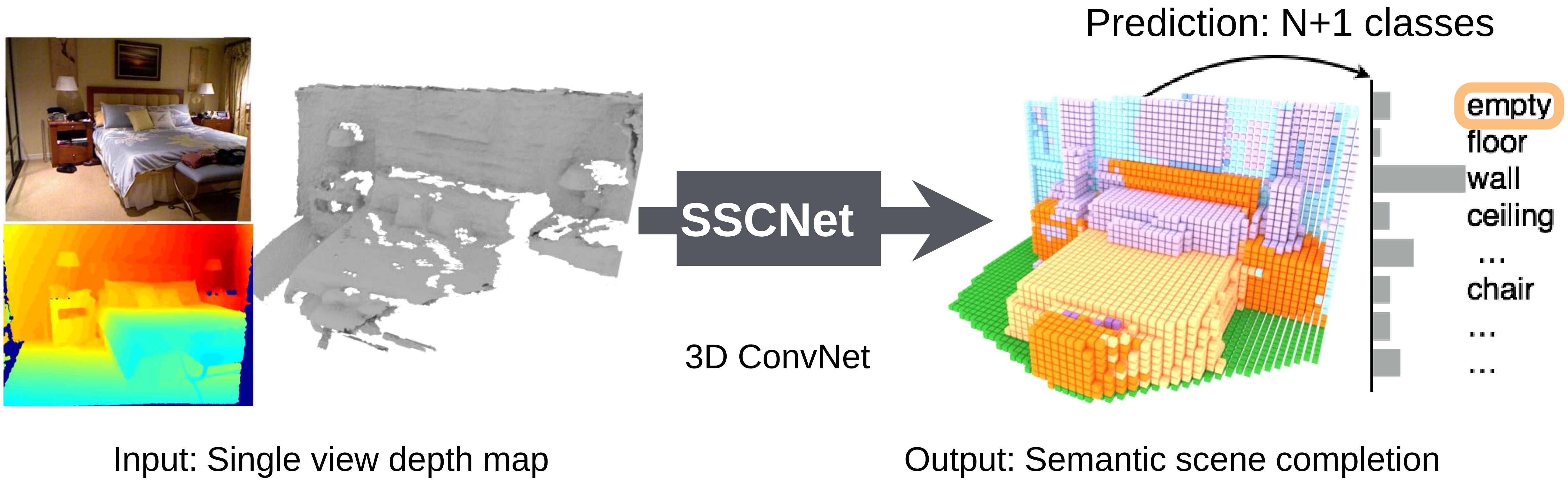
The occupancy and the object identity are tightly intertwined !

This paper

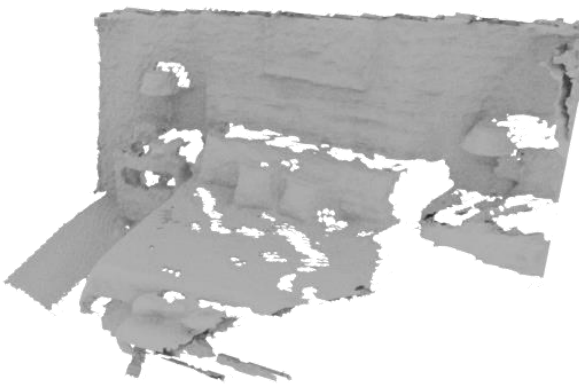
semantic scene completion

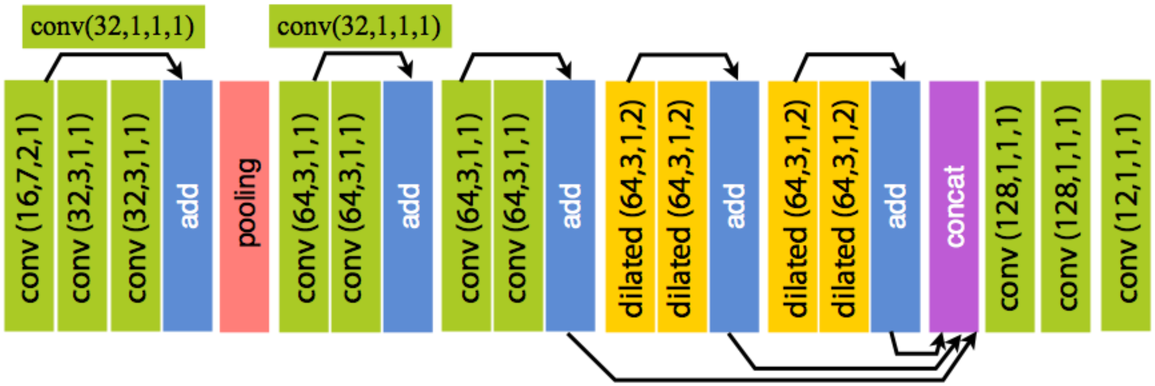
Semantic Scene Completion: “SSCNet”

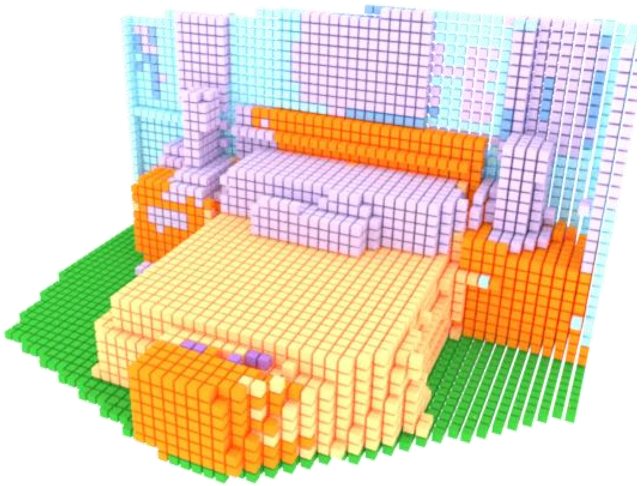
Approach: end-to-end deep network



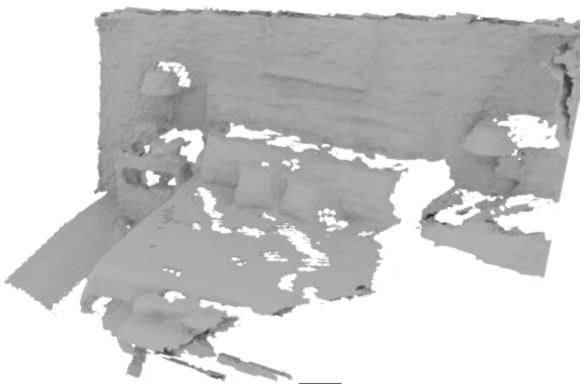
Semantic Scene Completion : “SSCNet”

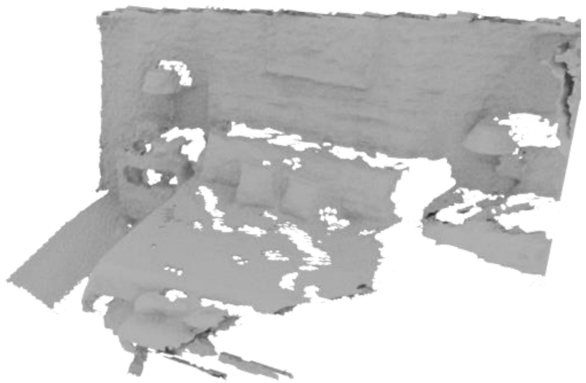


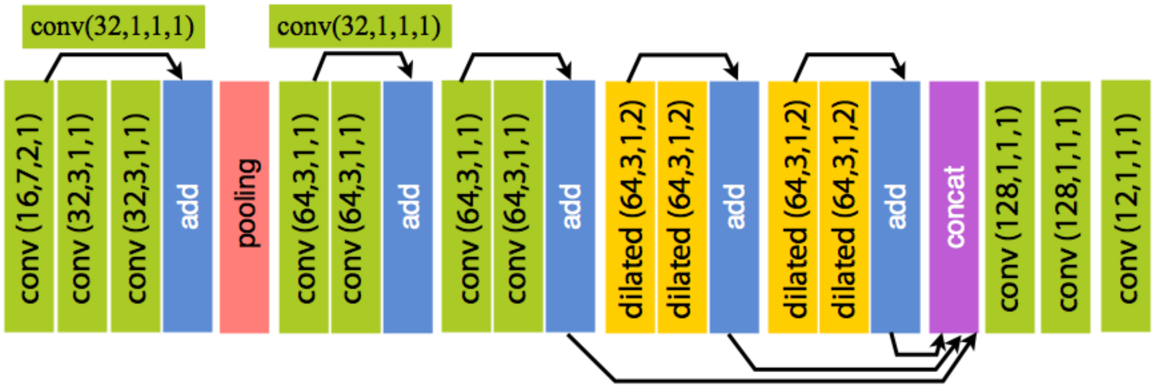


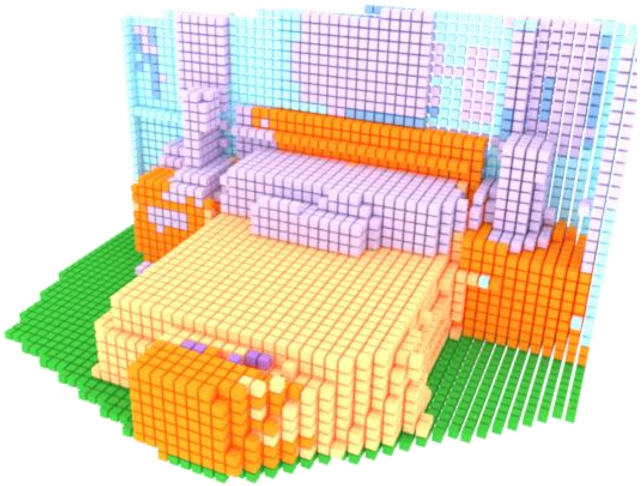


Semantic Scene Completion : “SSCNet”

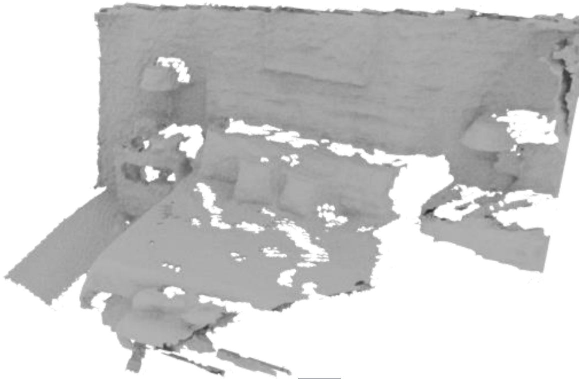


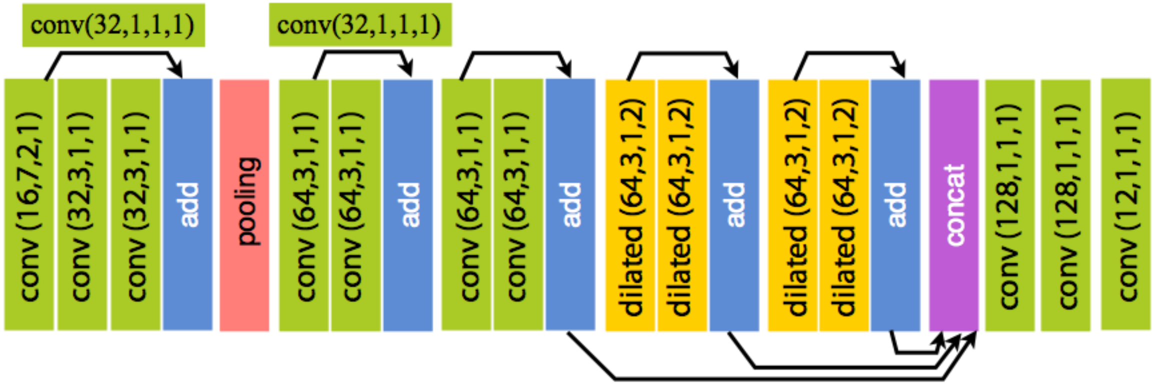


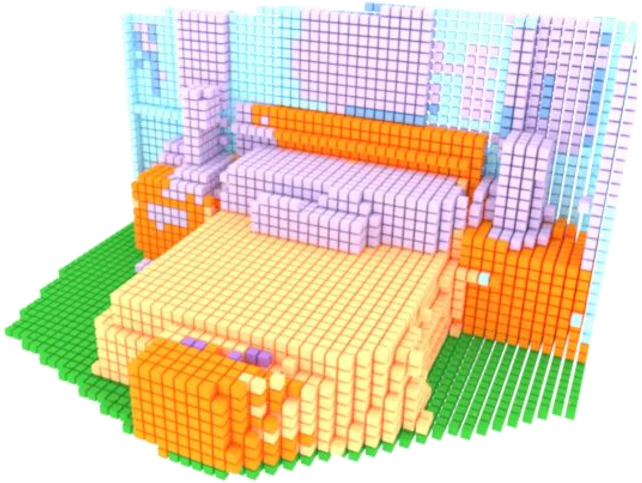


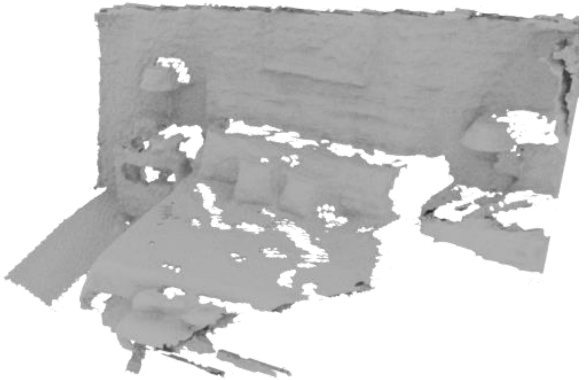


Semantic Scene Completion : “SSCNet”



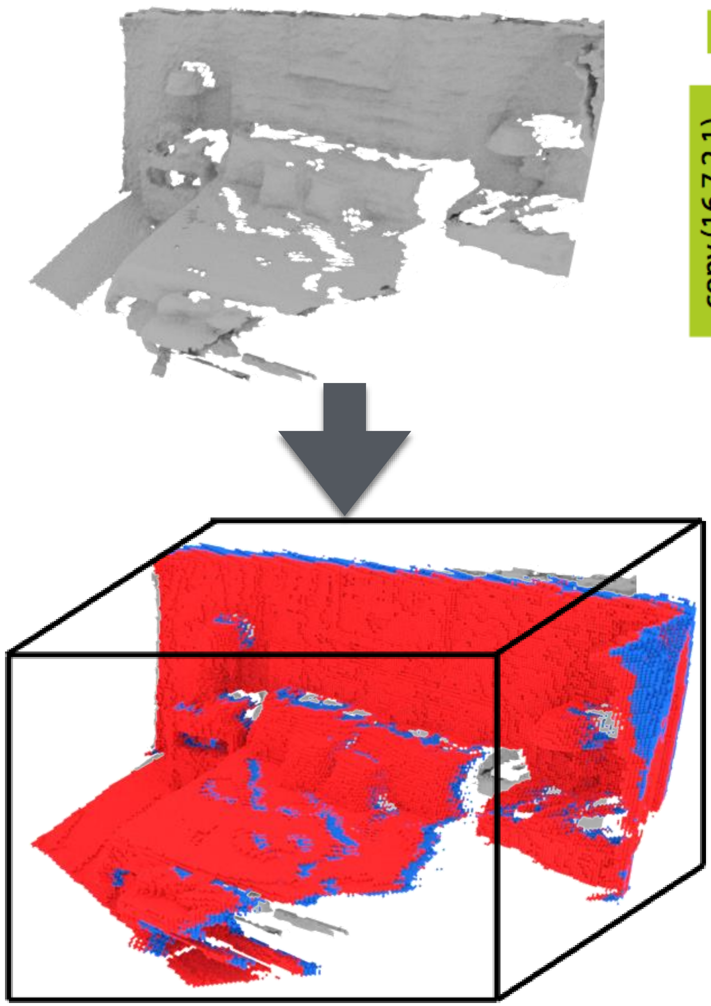


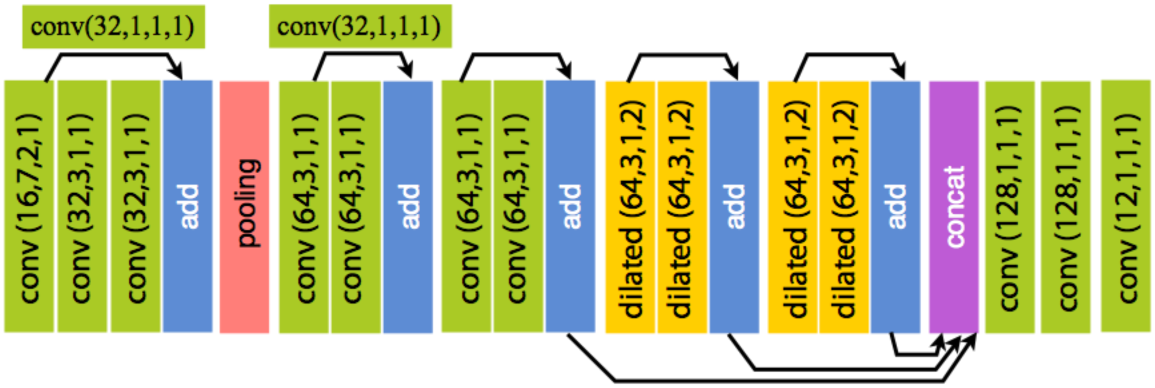


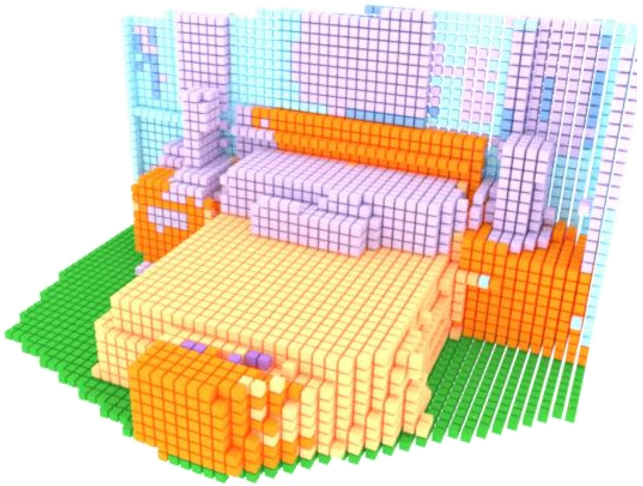


Encode 3D space using flipped TSDF

Semantic Scene Completion : “SSCNet”



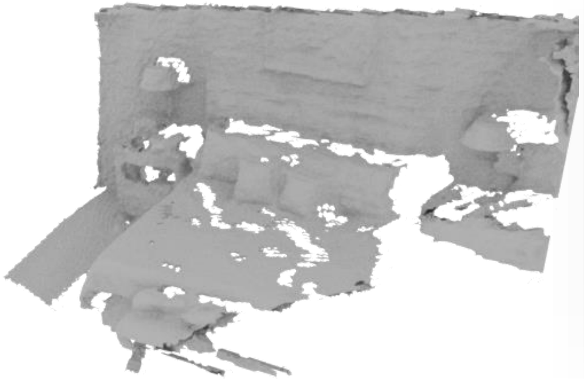


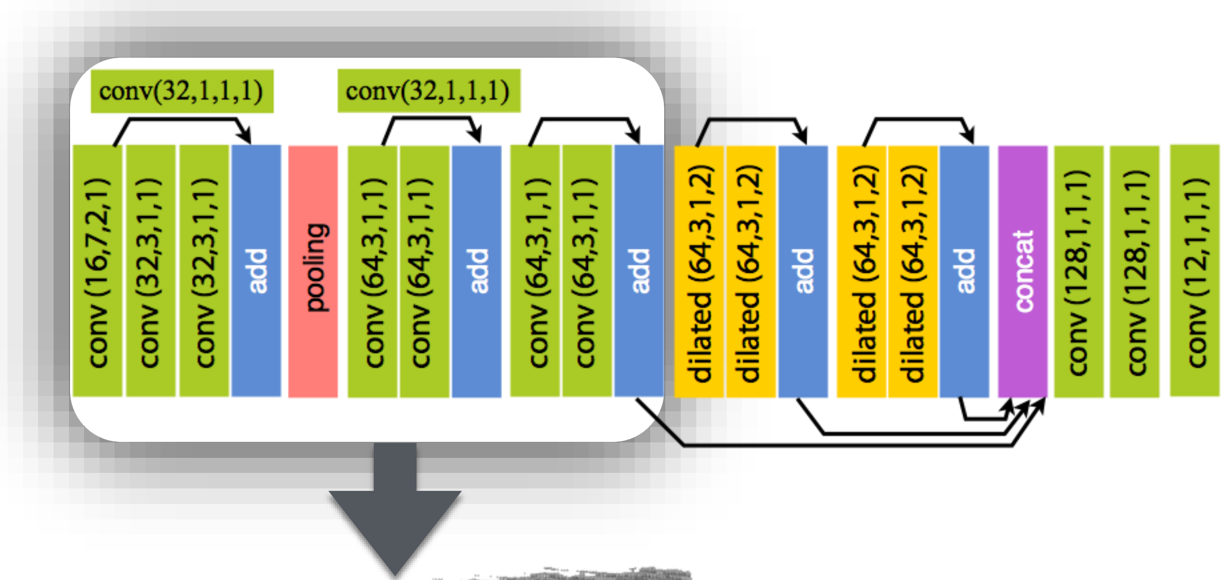


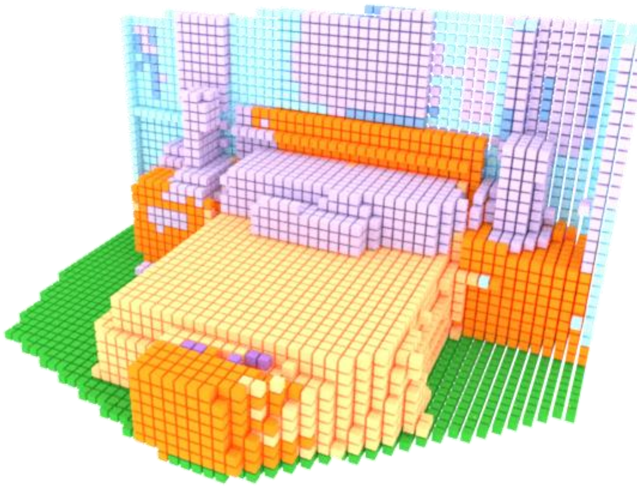
Encode 3D space using flipped TSDF

Voxel size: 0.02 m

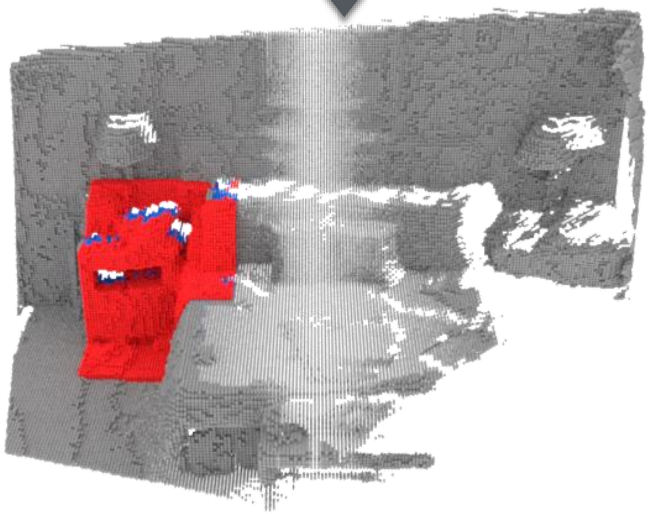
Semantic Scene Completion : “SSCNet”





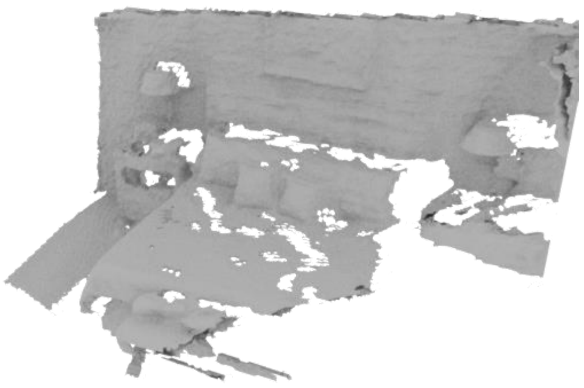


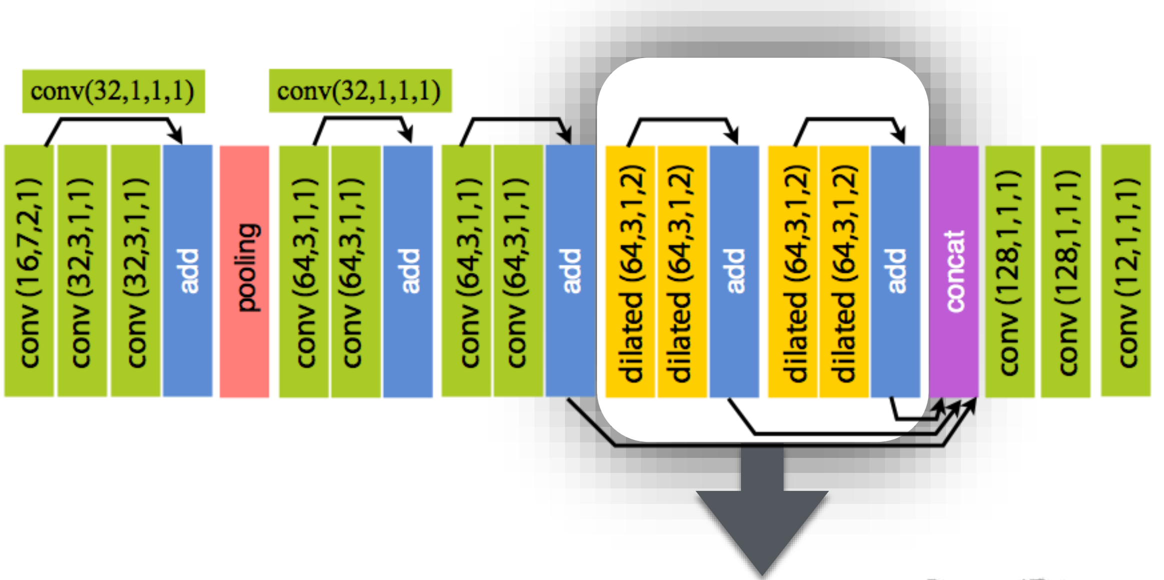
Local geometry

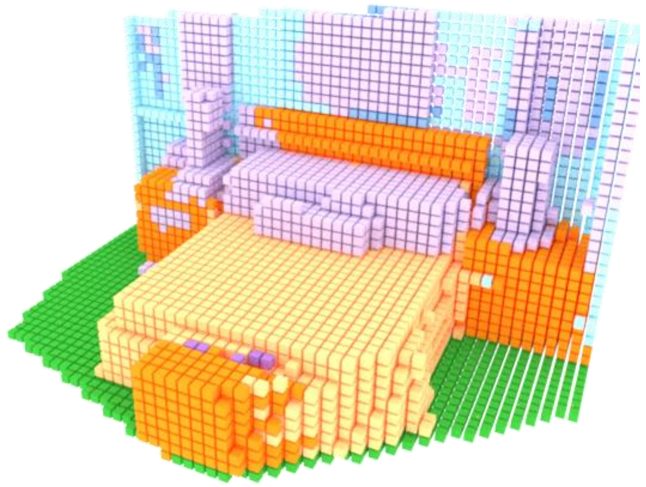


Receptive field: 0.98 m

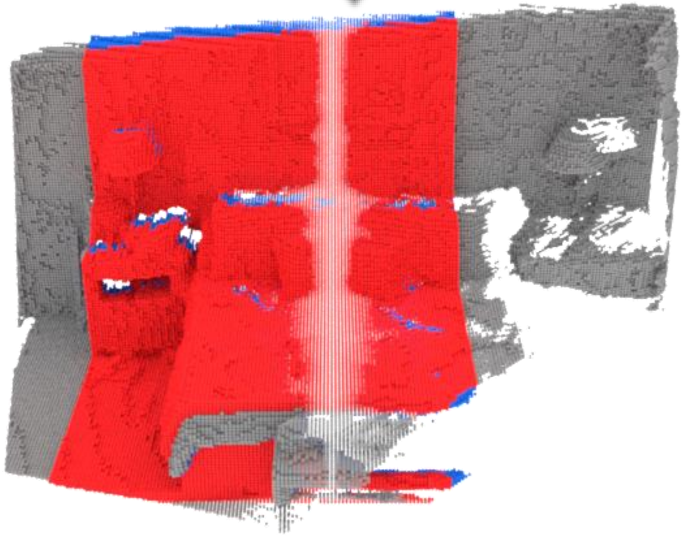
Semantic Scene Completion : “SSCNet”





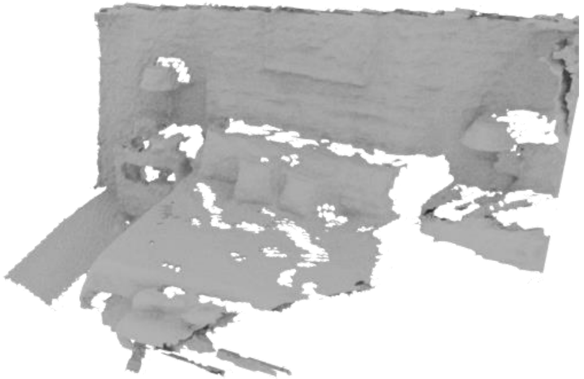


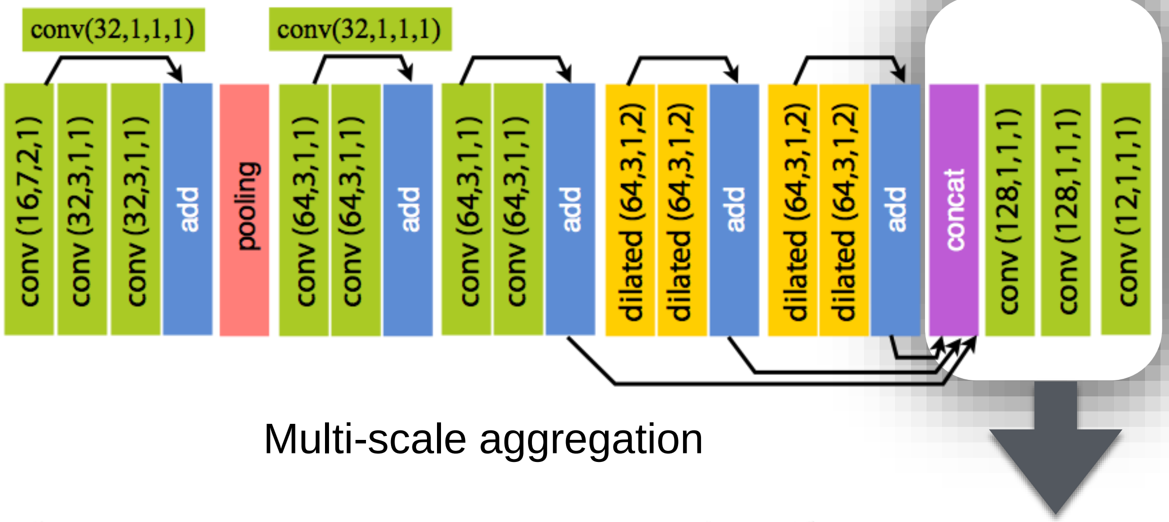
High-level 3D context
via big receptive field
provided by
dilated convolution

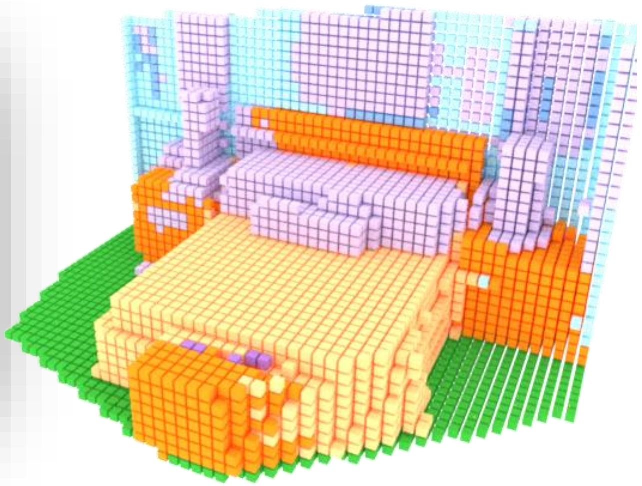


Receptive field: 2.26

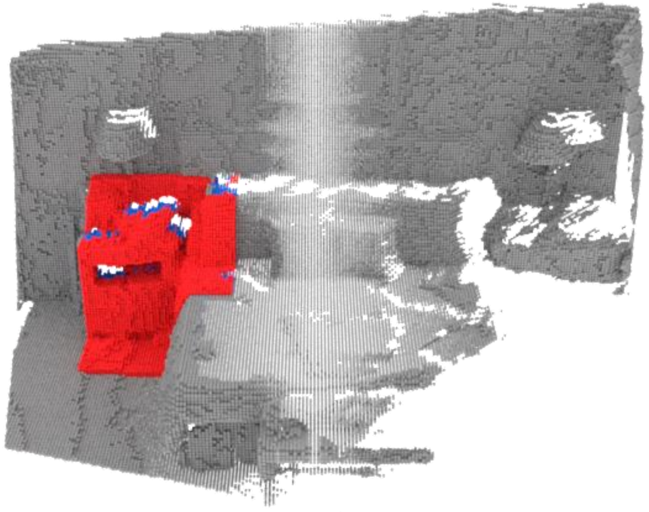
Semantic Scene Completion : “SSCNet”



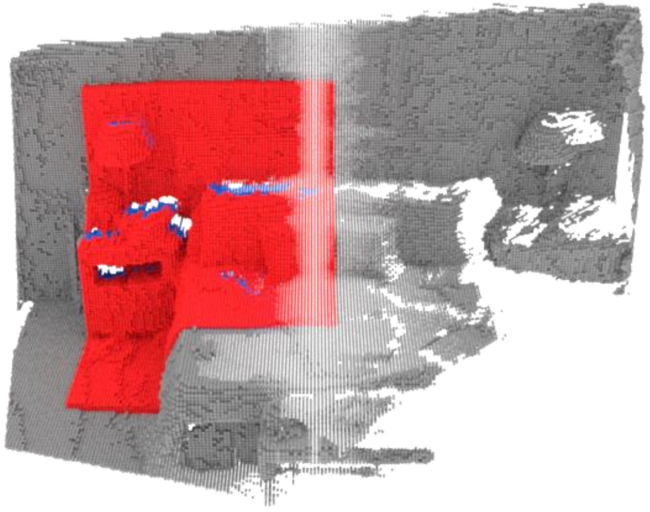




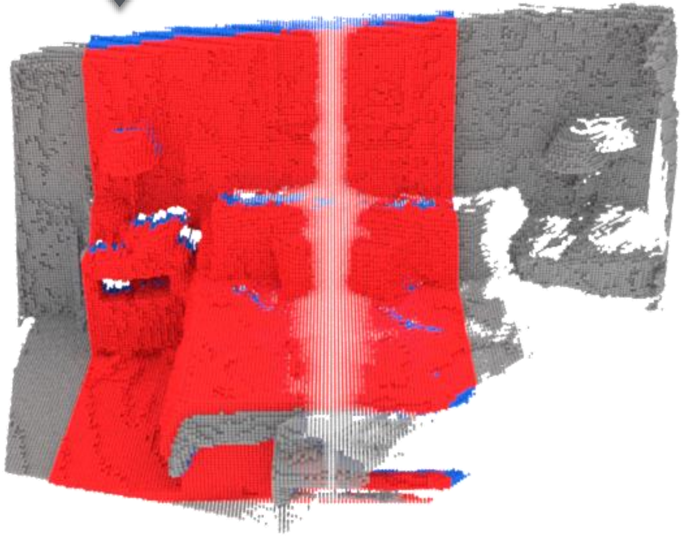
Multi-scale aggregation



Receptive field: 0.98 m



Receptive field:1.62 m



Receptive field: 2.26 m

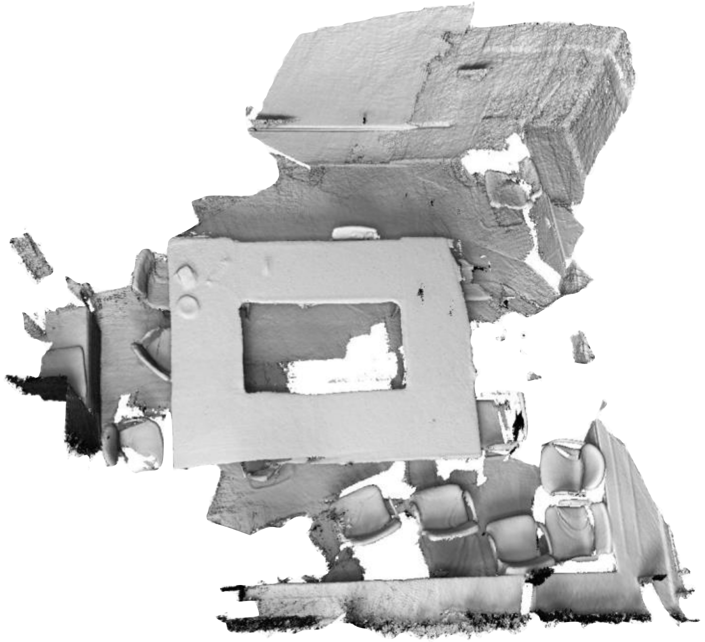
Semantic Scene Completion: “SSCNet” Experiments

Where to get training data?

Semantic Scene Completion: "SSCNet" Experiments

Where to get training data?

NYU: only visible surfaces



SUN3D: No semantic labels

No dense volumetric ground truth with semantic labels for a complete scene

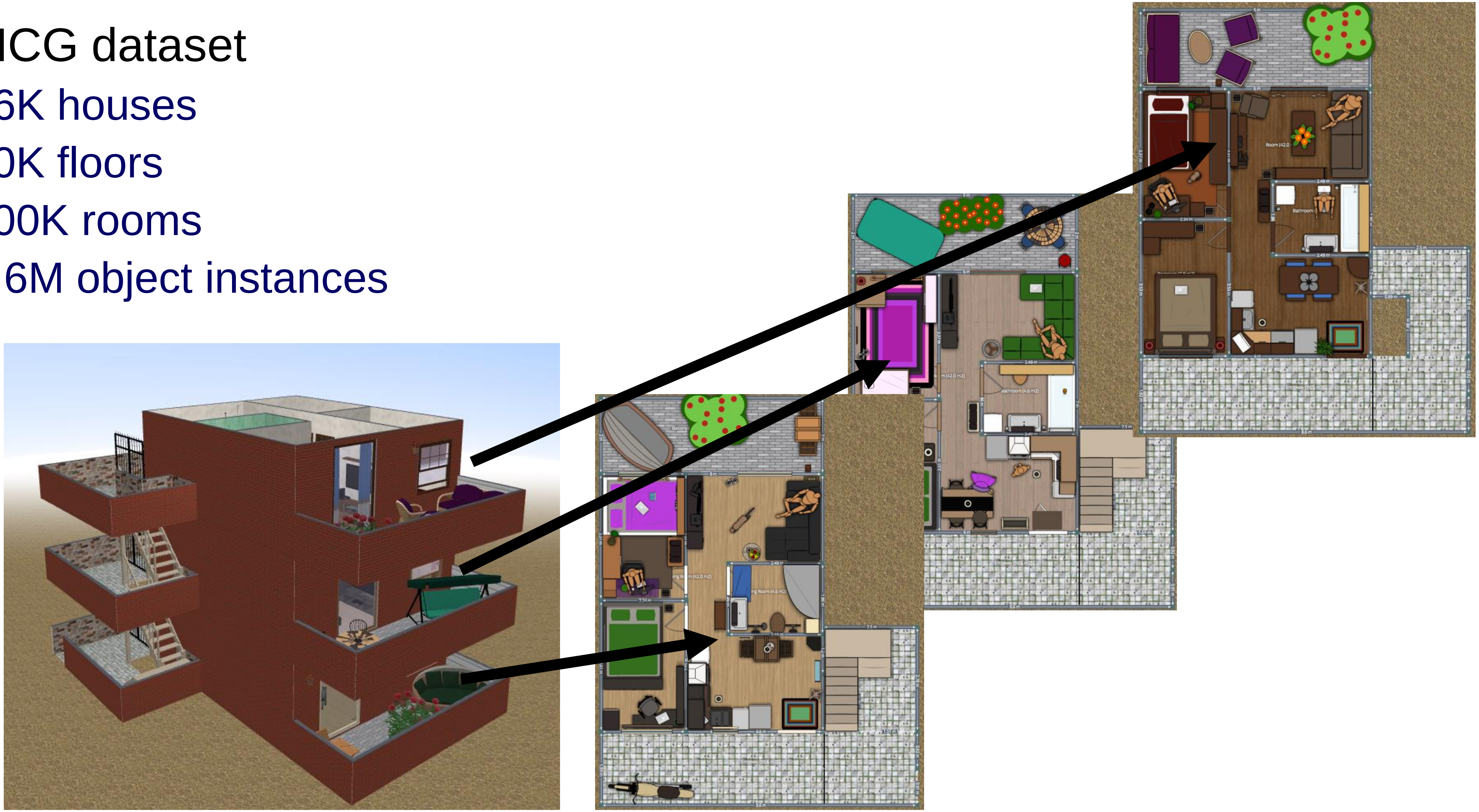
Semantic Scene Completion: “SSCNet” Experiments

SUNCG dataset

Semantic Scene Completion: “SSCNet” Experiments

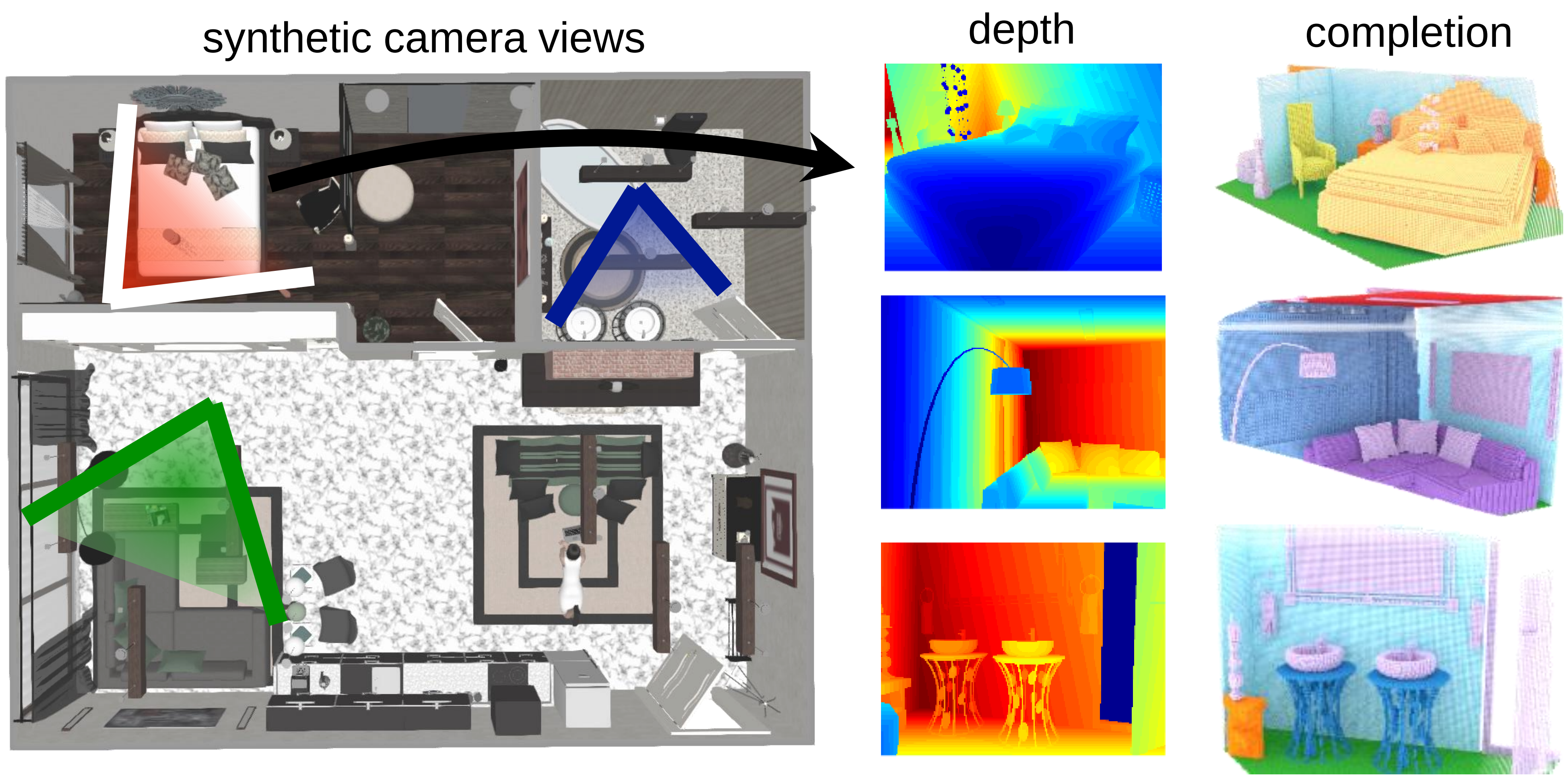
SUNCG dataset

- 46K houses
- 50K floors
- 400K rooms
- 5.6M object instances



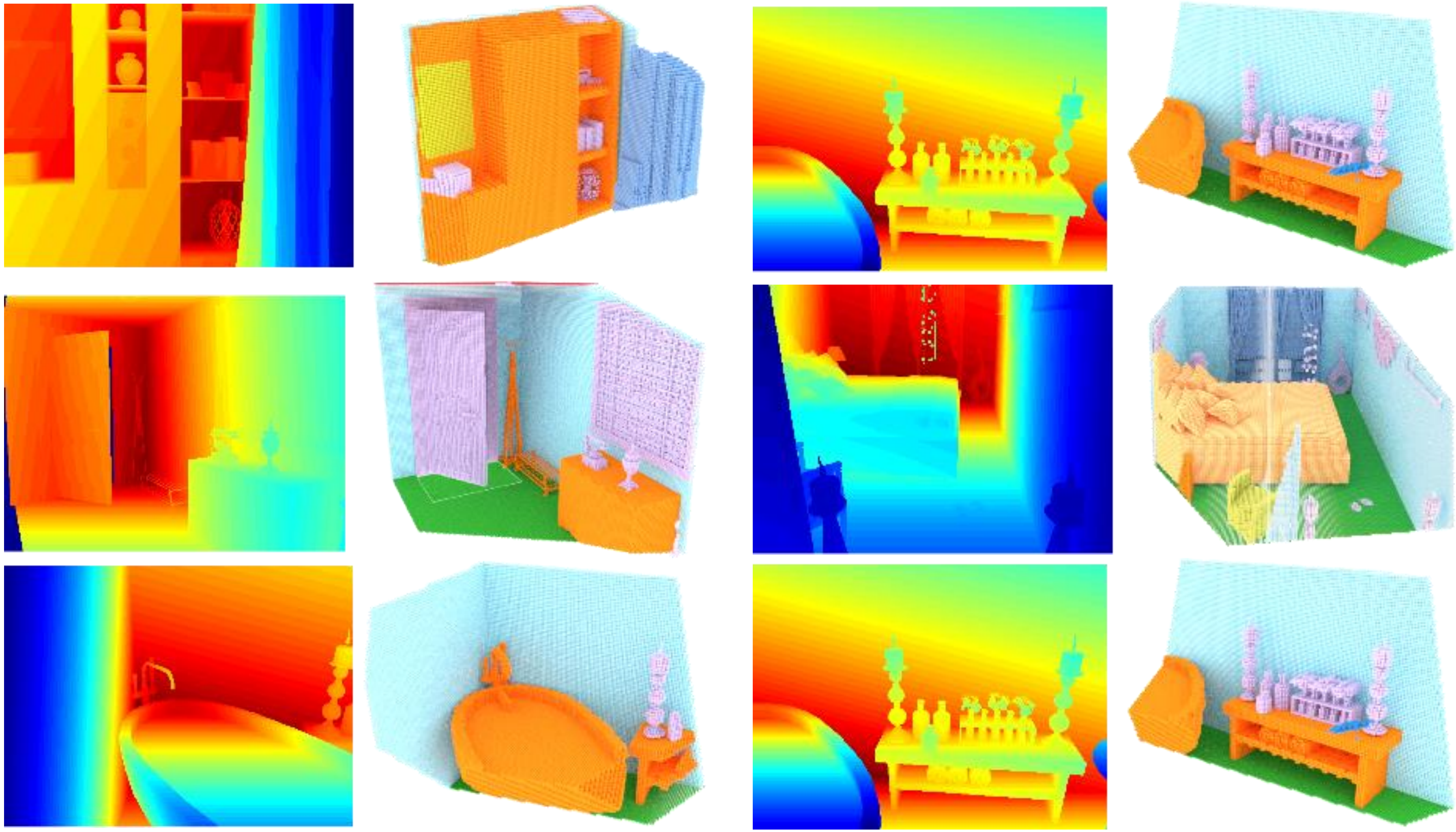
Semantic Scene Completion: “SSCNet” Experiments

SUNCG dataset



Semantic Scene Completion: “SSCNet” Experiments

SUNCG dataset



Semantic Scene Completion: “SSCNet” Experiments

Train on SUNCG

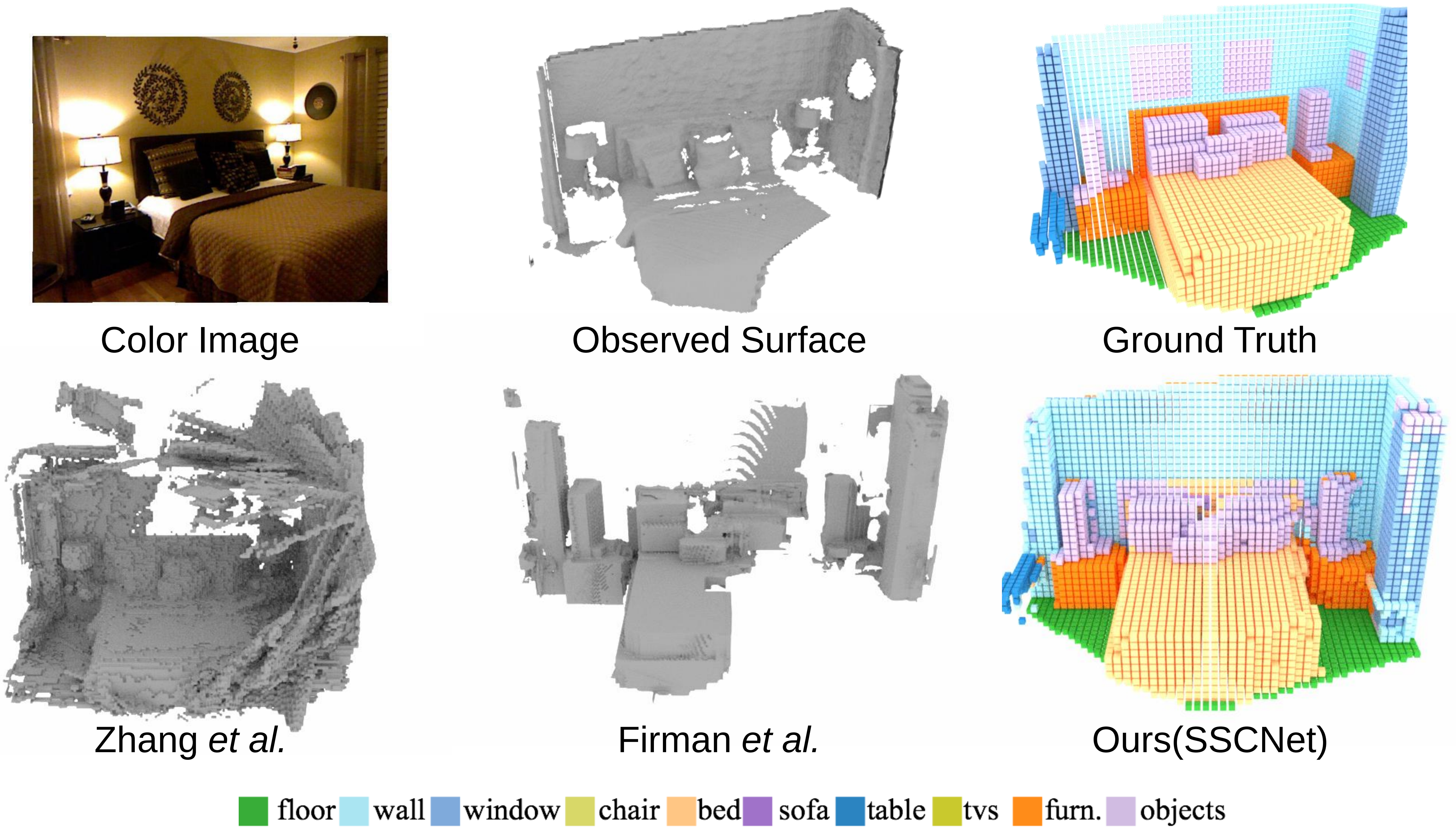
Test on NYU

Semantic Scene Completion: “SSCNet” Results

Result: better than previous volumetric completion algorithms

method	training	prec.	recall	IoU
Zheng *et al*. [36]	NYU	60.1	46.7	34.6
Firman *et al*. [3]	NYU	66.5	69.7	50.8
SSCNet completion	NYU	66.3	96.9	64.8
SSCNet joint	NYU	75.0	92.3	70.3
SSCNet joint	SUNCG+NYU	**75.0**	**96.0**	**73.0**

Comparison to previous algorithms for volumetric completion

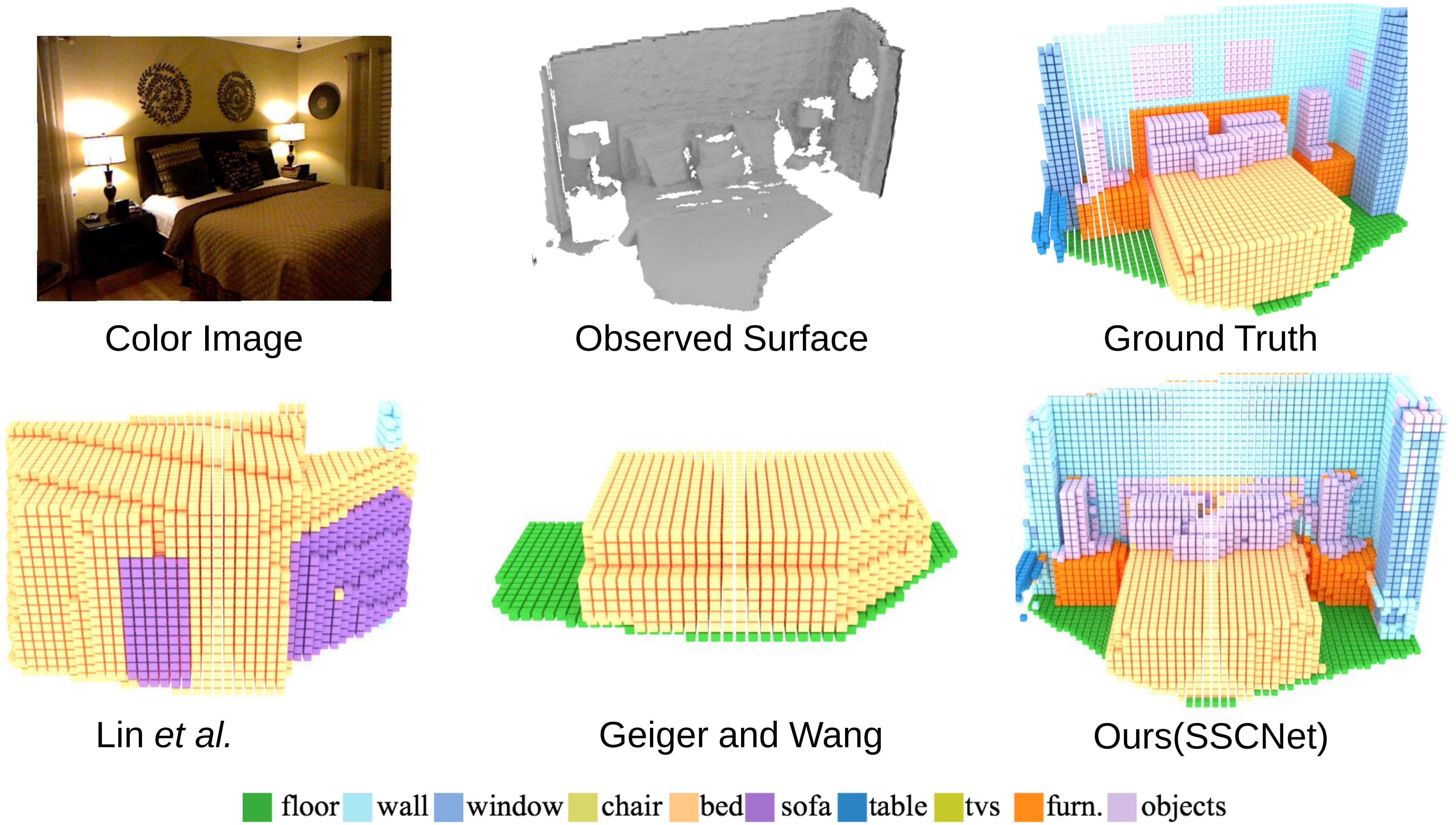

Color Image
Observed Surface
Ground Truth
Zhang et al.
Firman et al.
Ours(SSCNet)
floor
wall
window
chair
bed
sofa
table
tvs
furn.
objects

Semantic Scene Completion: “SSCNet” Results

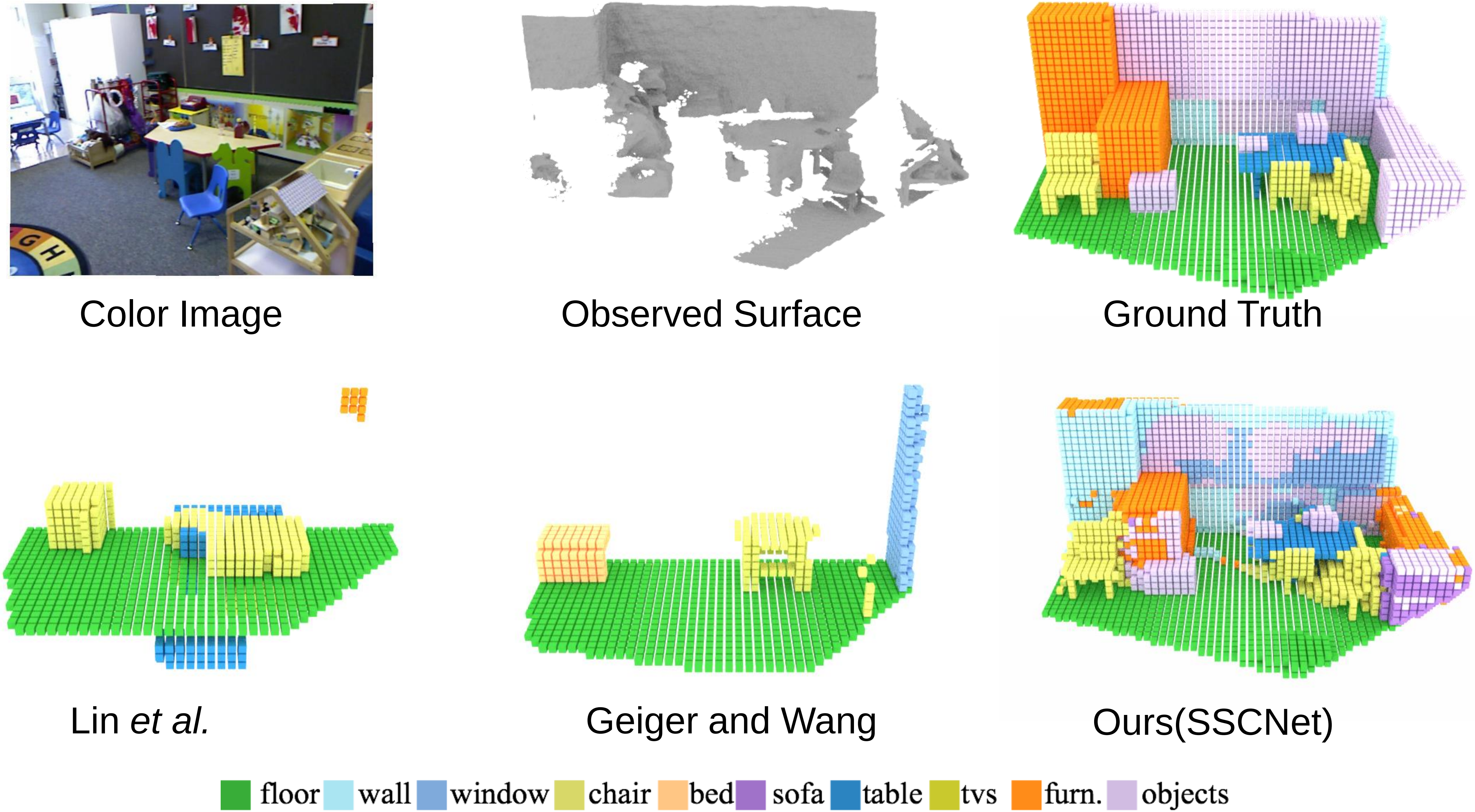
Result: better than previous 3D model fitting algorithms

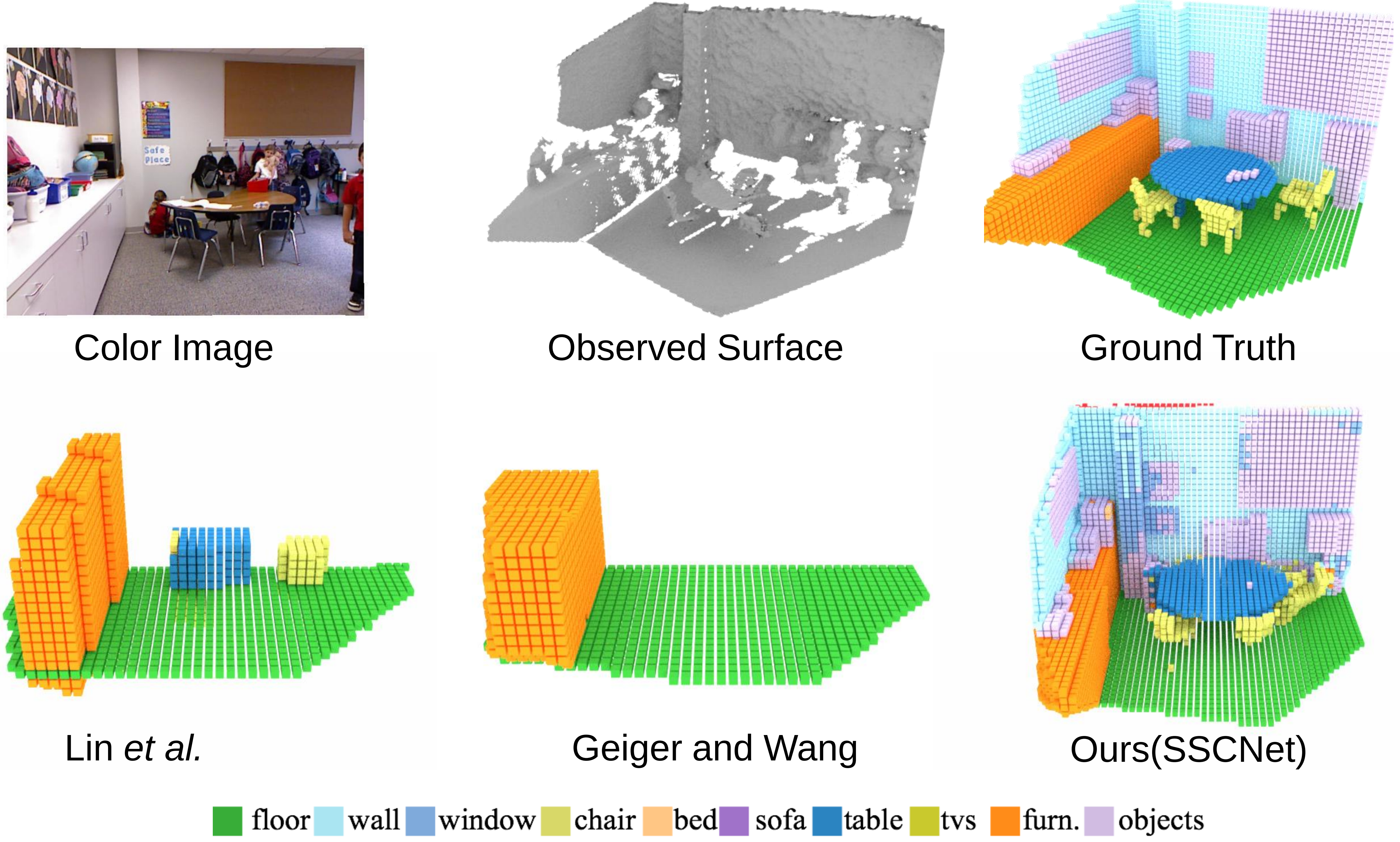
	scene completion			semantic scene completion											
method (train)	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
Lin *et al.* (NYU) [17]	58.5	49.9	36.4	0	11.7	13.3	**14.1**	9.4	29	24	6.0	7.0	16.2	1.1	12.0
Geiger and Wang (NYU) [4]	65.7	58	44.4	10.2	62.5	19.1	5.8	8.5	40.6	27.7	7.0	6.0	22.6	5.9	19.6
SSCNet (NYU)	57.0	**94.5**	55.1	15.1	**94.7**	24.4	0	12.6	32.1	35	13	7.8	27.1	10.1	24.7
SSCNet (SUNCG)	55.6	91.9	53.2	5.8	81.8	19.6	5.4	12.9	34.4	26	13.6	6.1	9.4	7.4	20.2
SSCNet (SUNCG+NYU)	**59.3**	92.9	**56.6**	**15.1**	94.6	**24.7**	10.8	**17.3**	**53.2**	**45.9**	**15.9**	**13.9**	**31.1**	**12.6**	**30.5**

Comparison to previous algorithms for 3D model fitting



Color Image
Observed Surface
Ground Truth
Lin *et al.*
Geiger and Wang
Ours(SSCNet)
floor wall window chair bed sofa table tvs furn. objects


Color Image
Observed Surface
Ground Truth
Lin et al.
Geiger and Wang
Ours(SSCNet)
floor
wall
window
chair
bed
sofa
table
tvs
furn.
objects


Color Image
Observed Surface
Ground Truth
Lin et al.
Geiger and Wang
Ours(SSCNet)
floor
wall
window
chair
bed
sofa
table
tvs
furn.
objects

Summary

Three projects where ConvNets are trained to recognize patterns in voxels with different …

- Tasks
- Scales
- Training data
- Loss functions
- Network architectures
- Training protocols

Future Challenges

Acquiring larger data sets

Leveraging geometric structure

Leveraging semantic structure

Better integration RGB and D

Better surface parameterizations

Finer-grained categories

Higher resolution

etc.

Future Challenges

Acquiring larger data sets

Leveraging geometric structure

Leveraging semantic structure

Better integration RGB and D

Better surface parameterizations

Finer-grained categories

Higher resolution

etc.

Future Challenges

- **Acquiring larger data sets**
- Leveraging geometric structure
- Leveraging semantic structure
- Better integration RGB and D
- Better surface parameterizations
- Finer-grained categories
- Higher resolution
- etc.

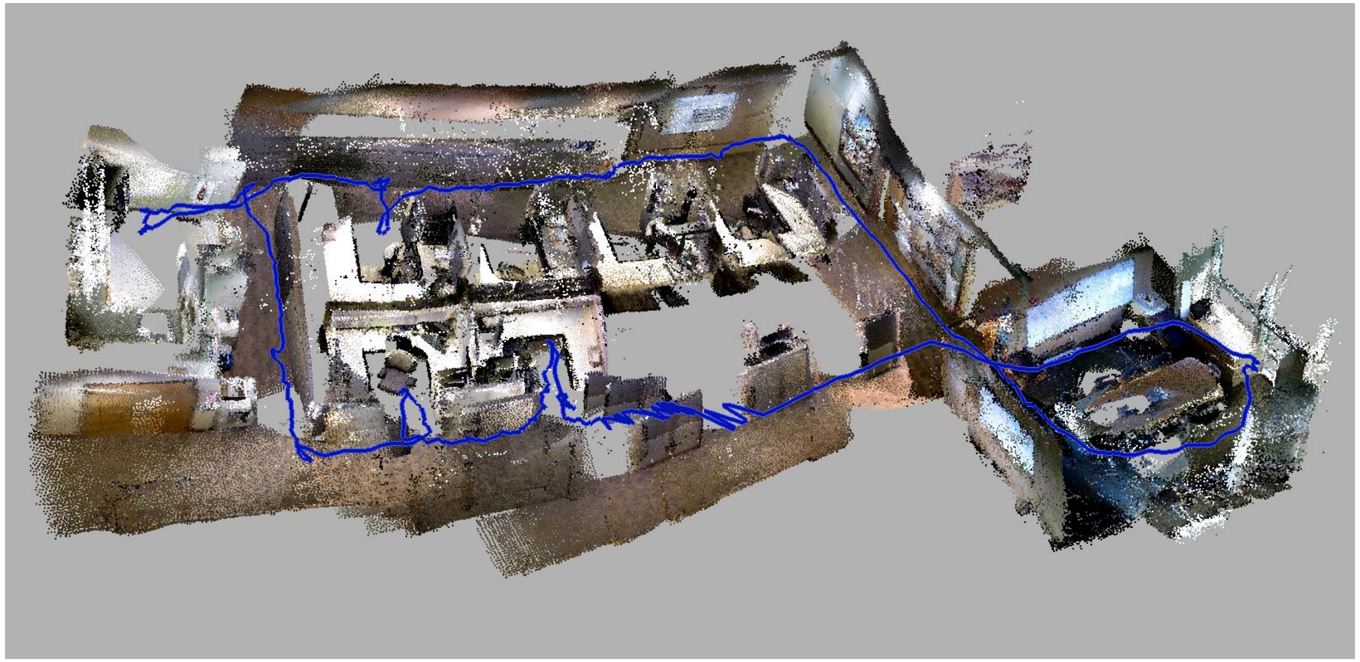
1,500 surface reconstructions

36,213 labeled objects

A. Dai, A. Chang, M. Savva,
M. Halber, T. Funkhouser, and M. Niessner,
"ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes,"
submitted to CVPR 2017.

Future Challenges

- Acquiring larger data sets
- ► Leveraging geometric structure
- Leveraging semantic structure
- Better integration RGB and D
- Better surface parameterizations
- Finer-grained categories
- Higher resolution
- etc.



M. Halber, T. Funkhouser,
"Fine-to-Coarse Registration of RGB-D Scans,"
submitted to CVPR 2017

Future Challenges

Acquiring larger data sets

► Leveraging geometric structure

Leveraging semantic structure

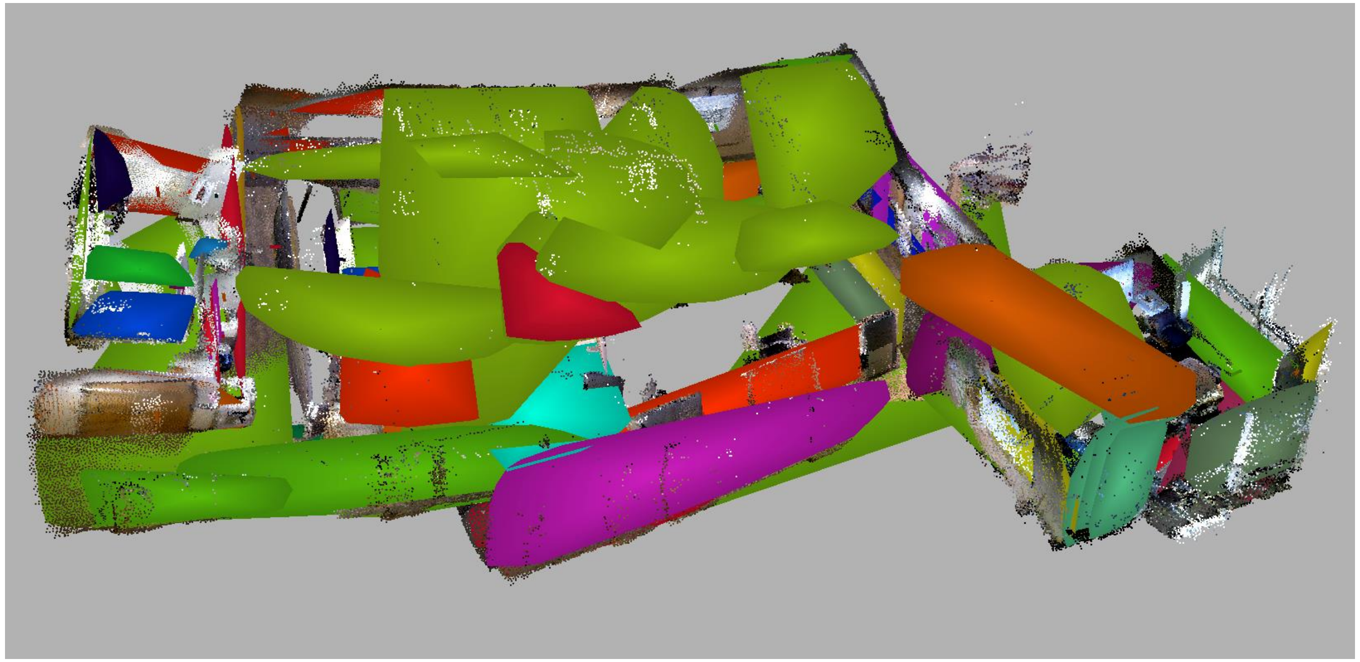
Better integration RGB and D

Better surface parameterizations

Finer-grained categories

Higher resolution

etc.



M. Halber, T. Funkhouser,
"Fine-to-Coarse Registration of RGB-D Scans,"
submitted to CVPR 2017

Future Challenges

Acquiring larger data sets

► Leveraging geometric structure

Leveraging semantic structure

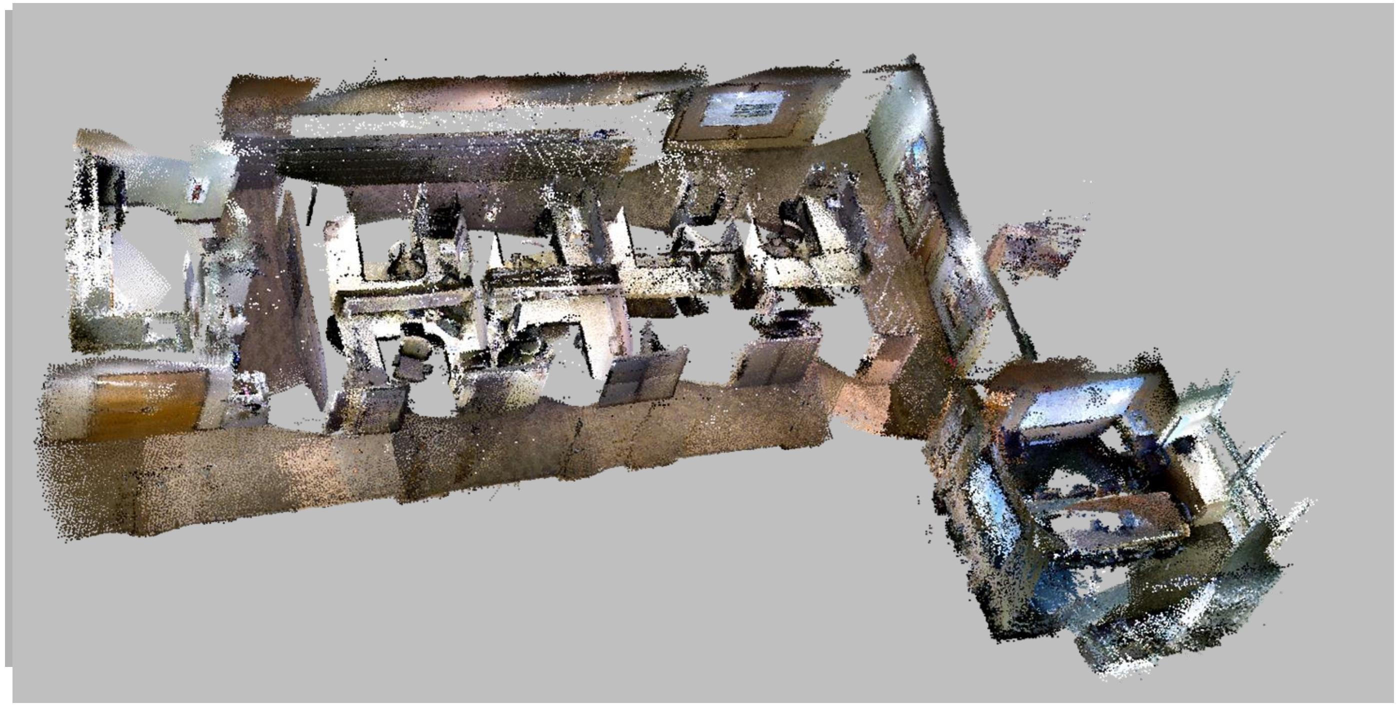
Better integration RGB and D

Better surface parameterizations

Finer-grained categories

Higher resolution

etc.

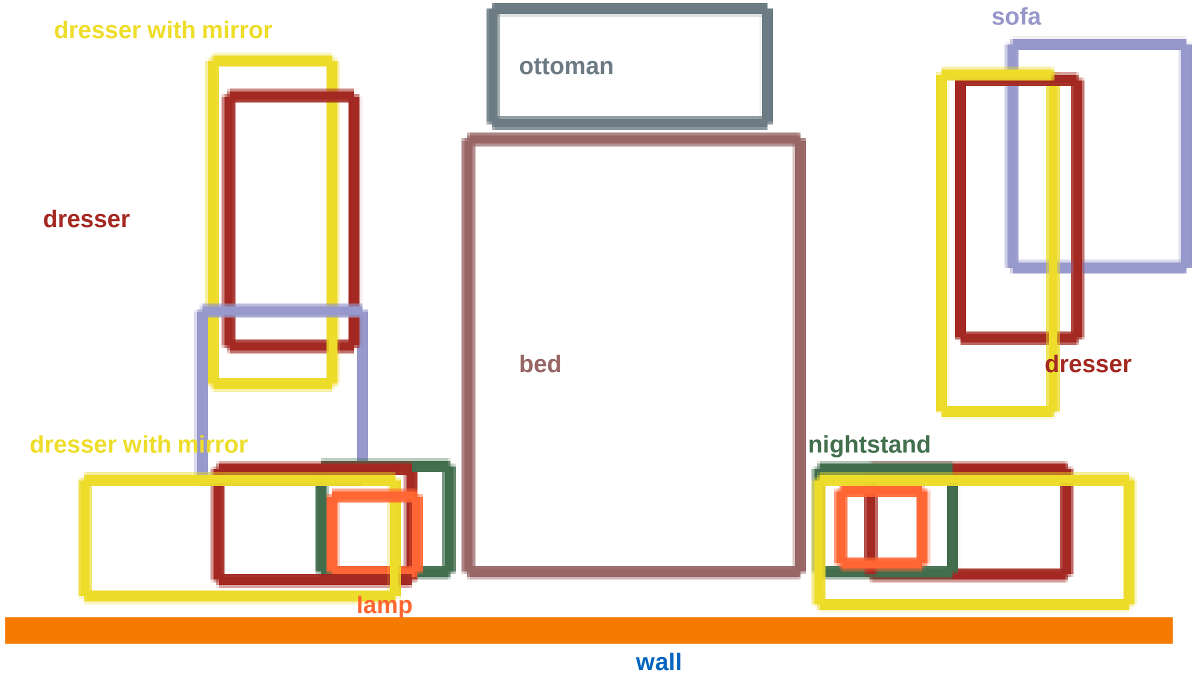


M. Halber, T. Funkhouser,
"Fine-to-Coarse Registration of RGB-D Scans,"
submitted to CVPR 2017

Future Challenges

- Acquiring larger data sets
- Leveraging geometric structure
- ► Leveraging semantic structure
- Better integration RGB and D
- Better surface parameterizations
- Finer-grained categories
- Higher resolution
- etc.

Sleeping Area



Y. Zhang, M. Bai, J. Xiao, P. Kohli, and S. Izadi, "DeepContext: Context-Encoding Neural Pathways for 3D Holistic Scene Understanding," submitted to CVPR 2017

Acknowledgments

Princeton:

- Angel Chang, Maciej Halber, Manolis Savva, Elena Sizikova, Shuran Song, Jianxiong Xiao, Fisher Yu, Yinda Zhang, Andy Zeng

Collaborators:

- Angela Dai, Matt Fisher, Matthias Niessner, Ersin Yumer

Data:

- SUN3D, 7-Scenes, Analysis-by-Synthesis, NYU, Trimble, Planner5D

Funding:

- Intel, NSF, Adobe

Thank You!