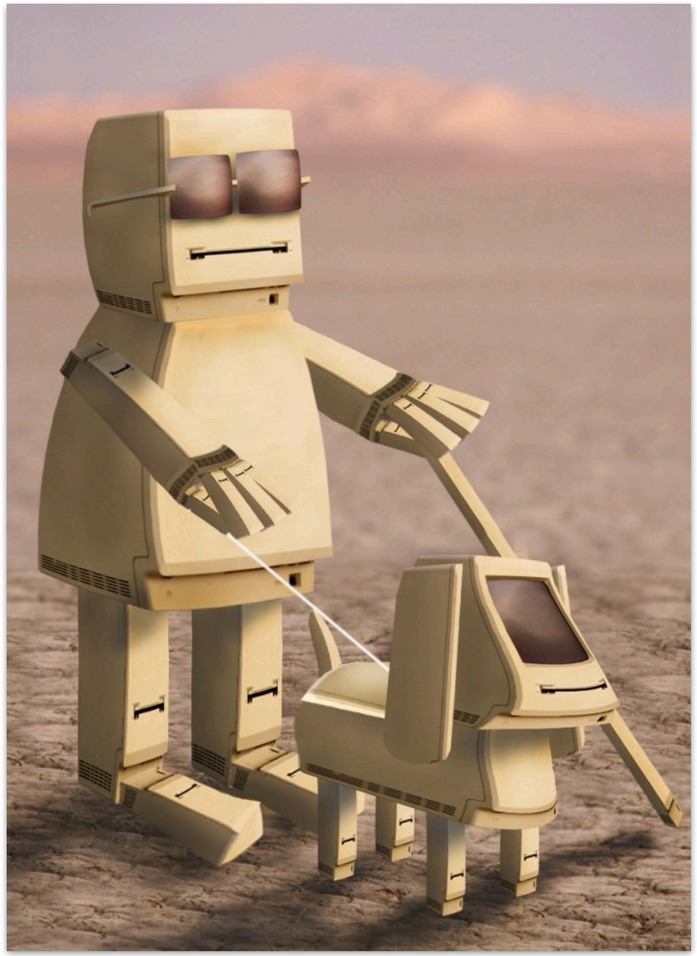


3D Deep Learning for Robot Perception

Jianxiong Xiao

PRINCETON UNIVERSITY

Today's Robotics

What journalists think how well they work

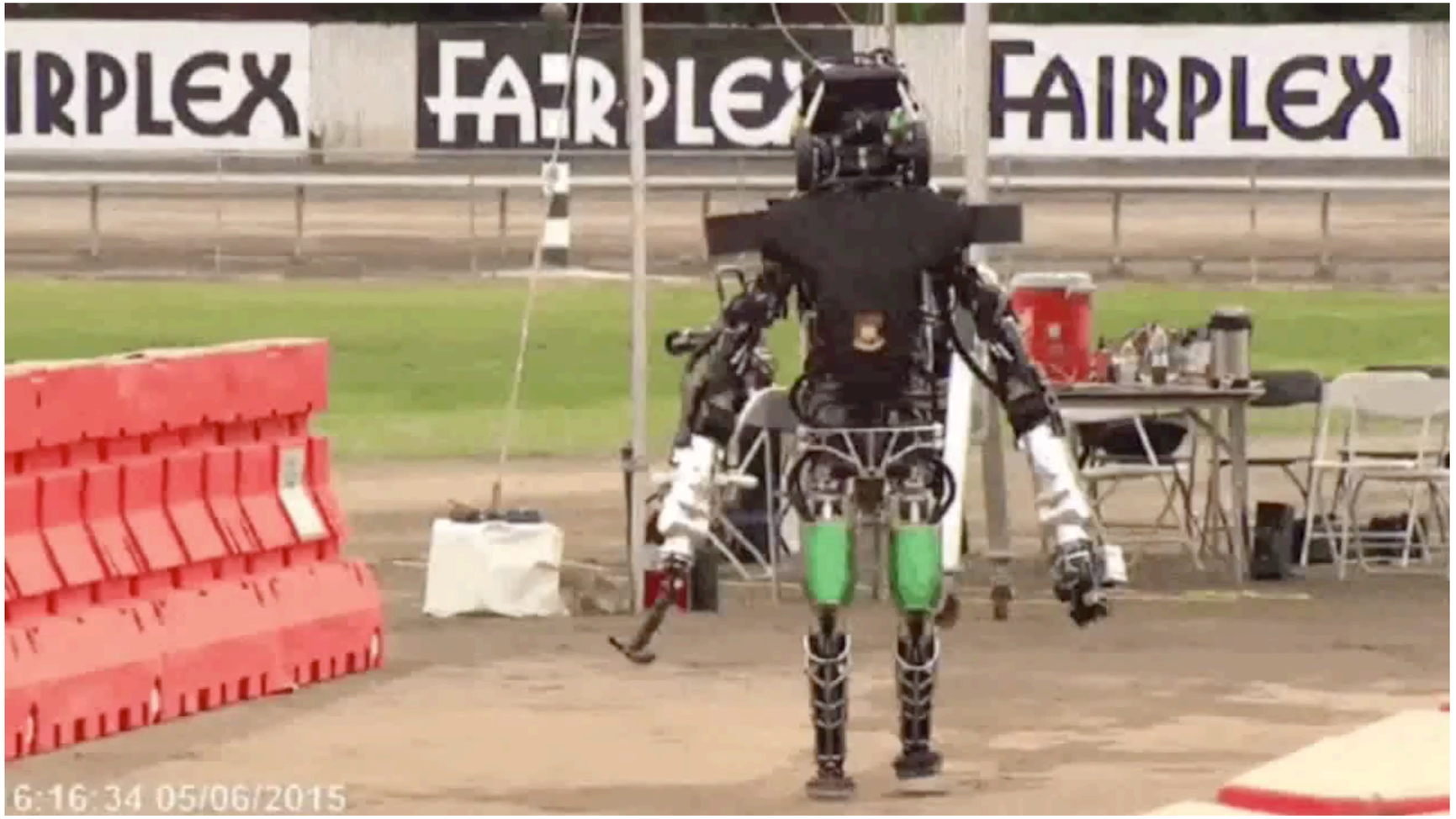
Today's Robotics

What we think how well they work

A 20 Million Dollar Question

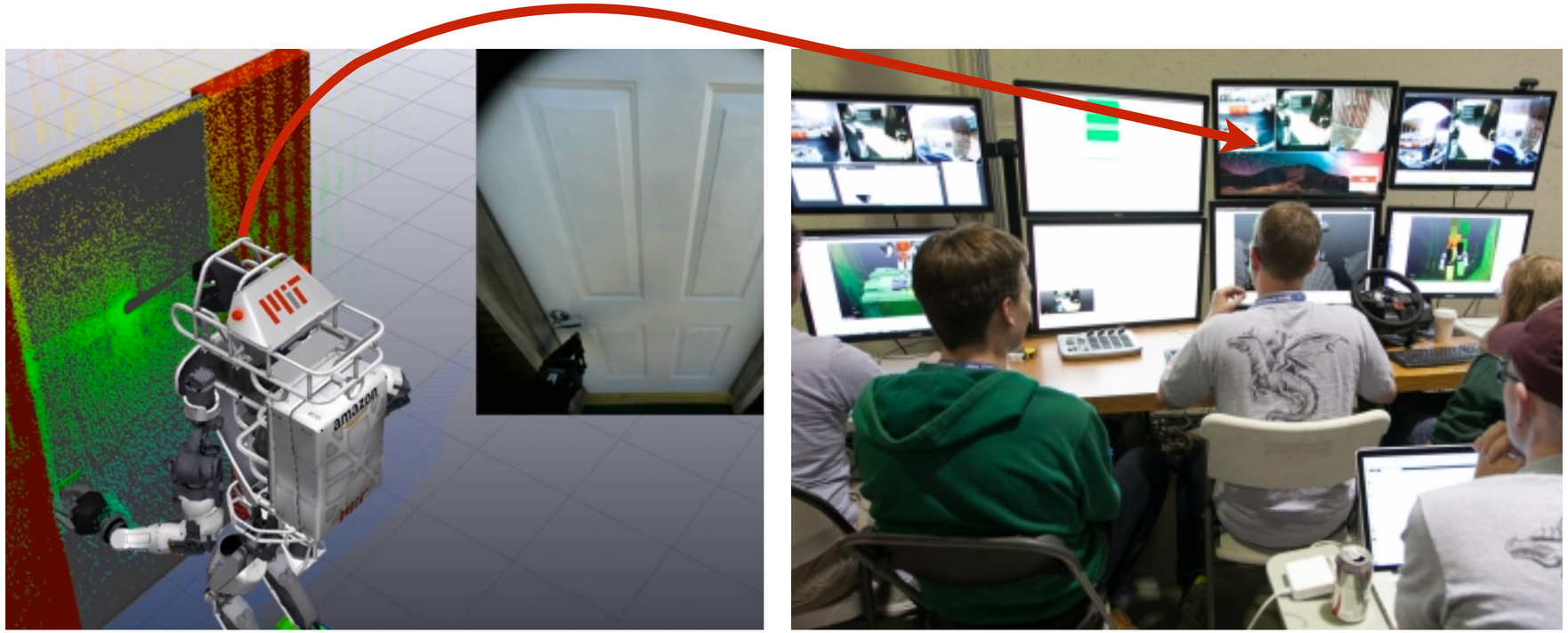
How well they really work (from DRC)

Today's Robotics



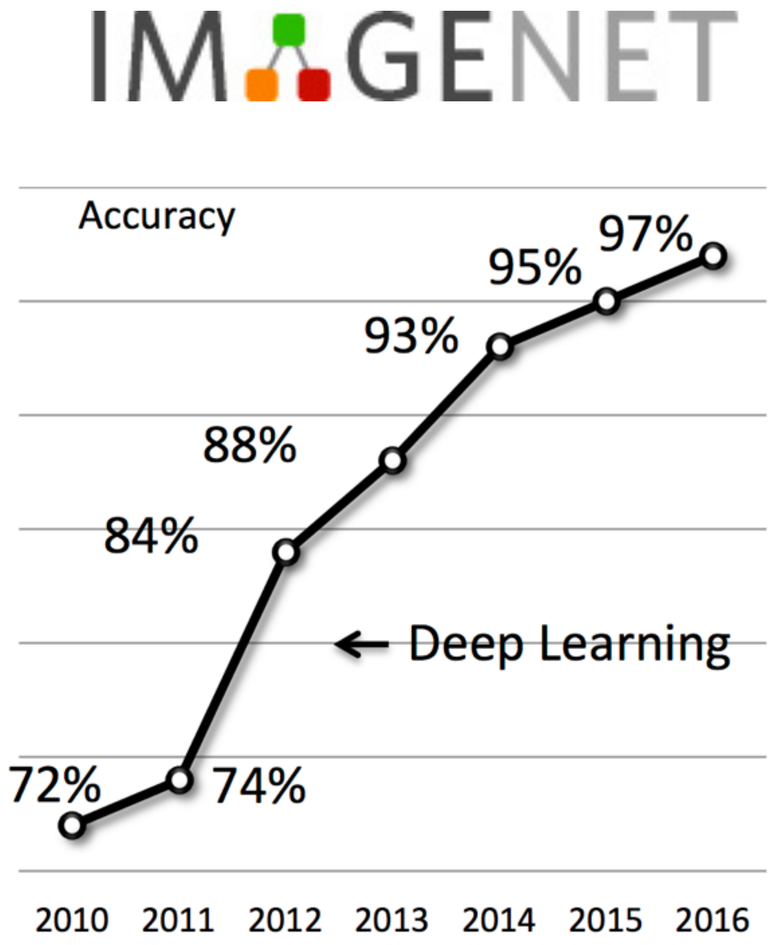
How well they really work (from DRC)

Today's Robotics



Today's robots are still mostly blind

Wait! Hold on…


IMAGENET
Accuracy
97%
95%
93%
88%
84%
Deep Learning
72%
74%
2010
2011
2012
2013
2014
2015
2016

Robotics has a Perception Problem!

"invited AI and vision specialists
to introduce their specialties to the mission,
but found that the most (actually all) famous algorithms
are not very effective in real situations."

Robotics has a Perception Problem!

Lessons from the Amazon Picking Challenge

Nikolaus Correll, *Senior Member*, Kostas E. Bekris, *Member*, Dmitry Berenson, *Member*, Oliver Brock, *Senior Member*, Albert Causo, *Member*, Kris Hauser, *Member*, Kei Okada, *Member*, Alberto Rodriguez, *Member*, Joseph M. Romano and Peter R. Wurman, *Member*

***Abstract*—This paper summarizes lessons learned from the first Amazon Picking Challenge in which 26 international teams designed robotic systems that competed to retrieve items from warehouse shelves. This task is currently performed by human workers, and there is hope that robots can someday help increase efficiency and throughput while lowering cost. We report on a 28-question survey posed to the teams to learn about each team's background, mechanism design, perception apparatus, planning and control approach. We identify trends in this data, correlate it with each team's success in the competition, and discuss observations and lessons learned.**

Note to Practitioners: *Abstract*—Perception, motion planning, grasping, and robotic system engineering has reached a level of maturity that makes it possible to explore automating simple warehouse tasks in semi-structured environments that involve high-mix, low-volume picking applications. This survey summarizes lessons learned from the first Amazon Picking Challenge, highlighting mechanism design, perception, and motion planning algorithms, as well as software engineering practices that were most successful in solving a simplified order fulfillment task. While the choice of mechanism mostly affects execution speed, the competition demonstrated the systems challenges of robotics and illustrated the importance of combining reactive control with deliberative planning.

I. Introduction

The first Amazon Picking Challenge (APC) was held during two days at the 2015 IEEE International Conference on Robotics and Automation (ICRA) in Seattle, Washington. The objective of the competition was to provide a challenge problem to the robotics research community that involved integrating the state of the art in object perception, motion planning, grasp planning, and task planning to manipulate real-world items in industrial settings in the spirit of a long tradition of competitions as a benchmark for Artificial Intelligence [1] with the long-term goal of warehouse automation [2], [3]. This paper presents the results of a survey of the 26 teams that participated in the challenge and synthesizes lessons learned by the participants.

Fig. 1. The RBO team's robot placing a pack of Oreo cookies that it retrieved from the warehouse shelf into a tote. Image courtesy of RBO team.

The diversity of the solutions employed was impressive at a hardware, software and algorithms level. They ranged from large, single robot arms to multiple small robots each assigned to one bin on the shelf, from simple suction cups to anthropomorphic robotic hands, and from fully reactive approaches to fully deliberative sense-plan-act approaches. In surveying the details of each team's approach and questioning them on what they learned from the experience, we hope to extract trends that help us (1) understand how to eventually solve the problem, and (2) discover what future robotics research directions are most promising for solving the general problems of perception, manipulation, and planning.

Extracting such trends, however, is not straightforward. Different teams got comparable results by following almost orthogonal approaches, sometimes stretching the limits of one technology as seen in Table III. Moreover, available data on successful grasps, i.e., removing a specific item from the bin and delivering it to a tote, is sparse, likely due to the numerous idiosyncratic ways that complex robotic systems can break down during a single evaluation trial outside of a lab environment. Still, it is possible to make some observations about the strengths and weaknesses of individual approaches, including both mechanisms and algorithms, and how they should be combined to improve the generality of solutions. We can also draw some conclusions about the process. For instance:

- some of the teams reported that they developed too many components from scratch and did not have time to make them robust,
- others reported that the off-the-shelf software components they used as "black-boxes" hid important functionality

N. Correll is with the Department of Computer Science, University of Colorado at Boulder, Boulder, CO 80309-430. Phone +1 330 717-1436, email: ncorrell@colorado.edu.
K. E. Bekris is with the Computer Science Department of Rutgers University, Piscataway, NJ, USA.
D. Berenson is with the Robotics Engineering Program at Worcester Polytechnic Institute (WPI), Worcester, MA, USA.
O. Brock is with the Robotics and Biology Laboratory at the Technische Universität Berlin, Germany.
A. Causo is with the Robotics Research Centre, School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore.
K. Hauser is with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708.
K. Okada is with JSK Robotics Laboratory, the University of Tokyo, Japan.
A. Rodriguez is with the Mechanical Engineering Department at MIT, Cambridge, USA.
J. Romano was a member of the advanced research team of Kiva Systems.
P. Wurman was CTO and Technical Co-Founder of Kiva Systems.

Amazon Picking Challenge

Team	Platform	Gripper	Sensor	Perception	Motion Planing
RBO	Single arm (Barrett) + mobile base (XR4000)	Suction	3D imaging on Arm, Laser on Base, Pressure sensor, Force-torque sensor	Multiple features (color, edge, height) for detection and filtering 3D bounding box for grasp selection	No
MIT	Single arm (ABB 1600ID)	Suction + gripper + spatula	Both 2D and 3D imaging on Head and Arm	3D RGB-D object matching	No
Grizzly	Dual arm (Baxter) + mobile base (Dataspeed)	Suction and gripper	2D imaging at End-effector, 3D imaging for head, and laser for base	3D bounding box segmentation and 2D feature based localization	Custom motion planning algorithm
NUS Smart Hand	Single arm (Kinova)	Two-finger gripper	3D imaging on Robot	Foreground subtraction and color histogram classification	Predefined path to reach and online cartesian planning inside the bin using MoveIt.
Z.U.N.	Dual arm (Custom)	Suction	(respondent skipped response)	(respondent skipped response)	MoveIt RRT Planning for reaching motion and use pre-defined motion inside bin
C^2M	Single arm (MELFA) on custom gantry	Custom gripper	3D imaging on End-effector and force sensor on arm	RGB-D to classify object and graspability	No
Rutgers U. Pracsys	Dual arm (Yaskawa Motoman)	Unigripper vacuum gripper & Robotiq 3-finger hand	3D imaging on Arm	3D object pose estimation	Pre computed PRM paths using PRACSYS software & grasps using GraspIt
Team K	Dual arm (Baxter)	Suction	3D imaging on Arm and Torso	Color and BoF for object verification	No
Team Nanyang	Single arm (UR5)	Suction and gripper	3D imaging on End-effector	Histogram to identify object and 2D features to determine pose	No
Team A.R.	Single arm (UR-10)	Suction	3D imaging on End-effector	Filtering 3D bounding box and matching to a database	No
Georgia Tech	Single arm	SCHUNK 3 finger hand	3D imaging on Head and Torso	Histogram data to to recognize and 3D perception to determine pose	Pre-defined grasp using custom software and OpenRave
Team Duke	Dual arm (Baxter)	Righthand 3 finger hand	3D imaging on End-effector	3D model to background subtraction and use color	Klamp't planner to reaching motion

Object Detection

2D Detection

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

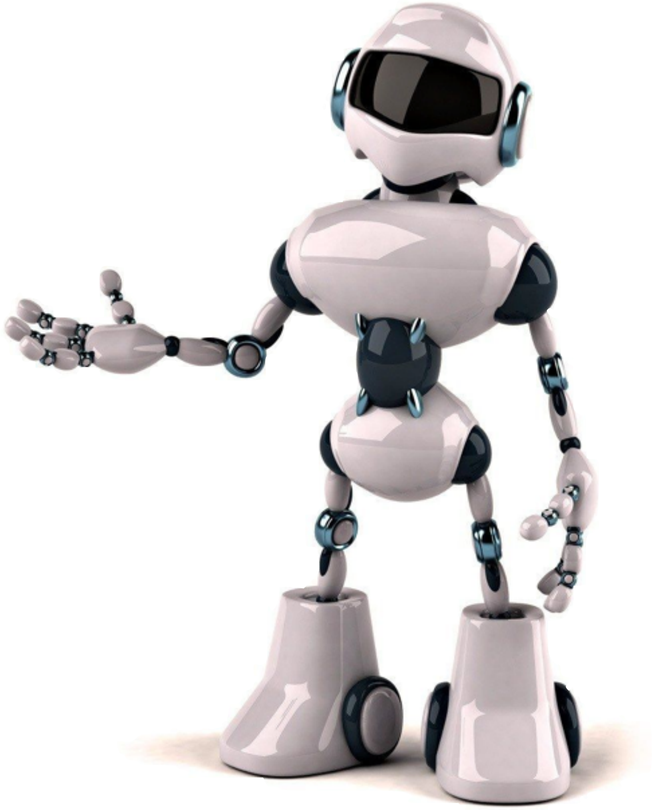
COCO

IMAGENET

Object Detection

2D Detection

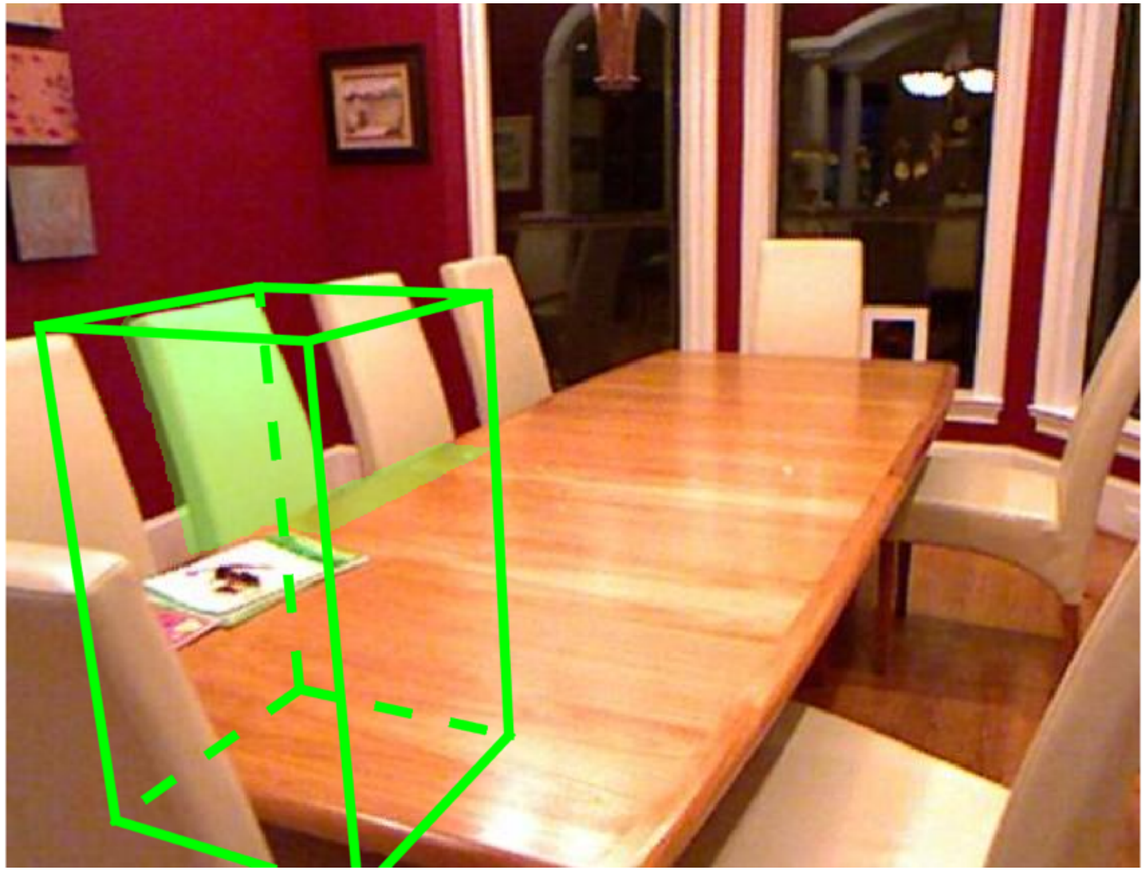
Where to sit?



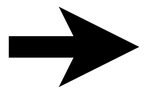
Object Detection

2D Detection

3D Detection

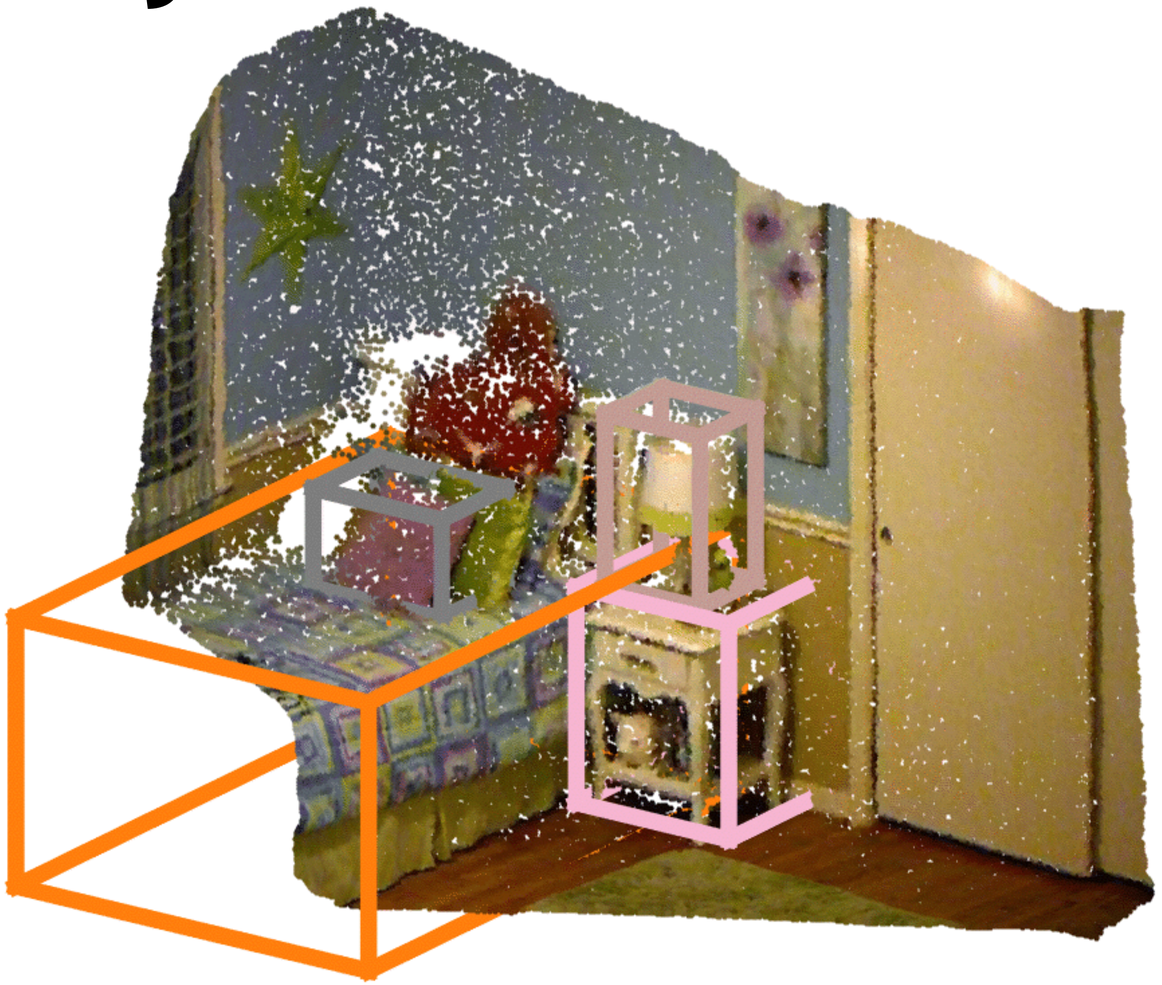


3D Amodal Object Detection



Input: Single RGB-D

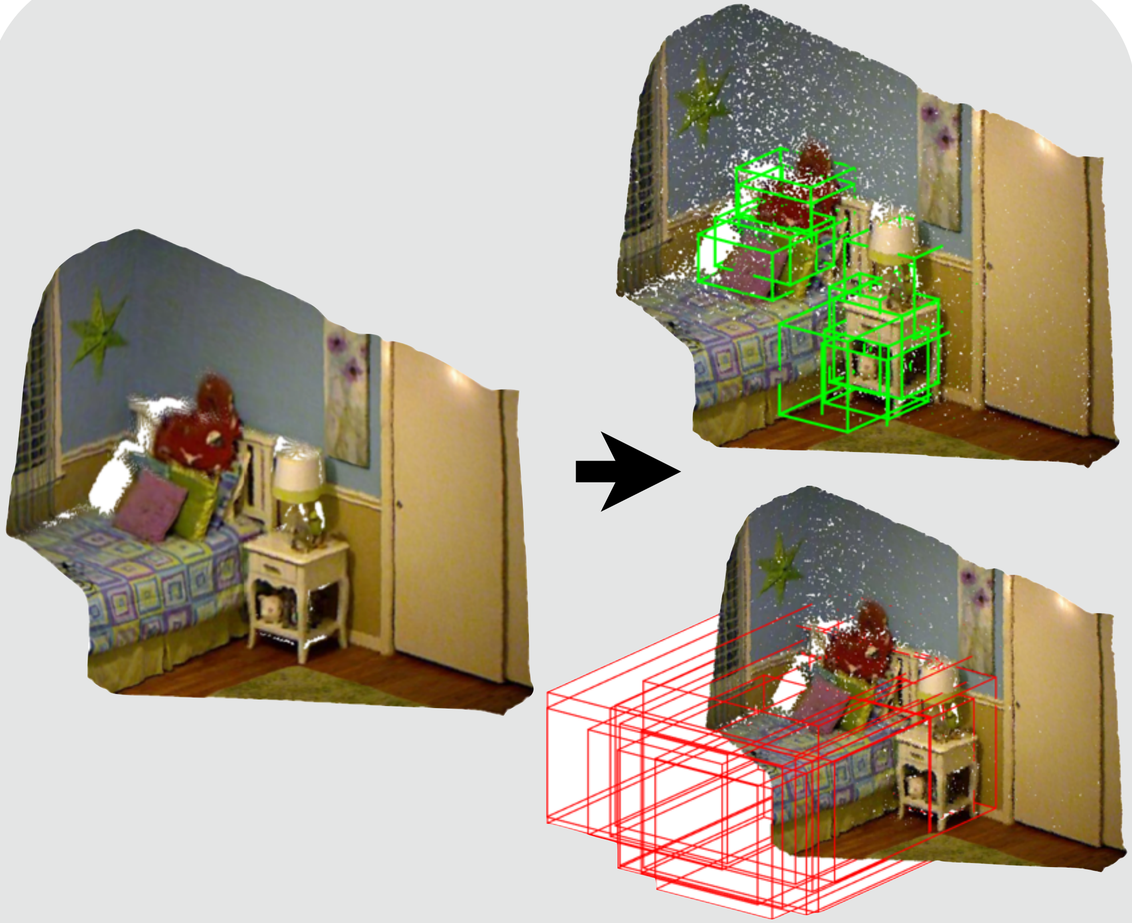
3D Amodal Object Detection



Input: Single RGB-D

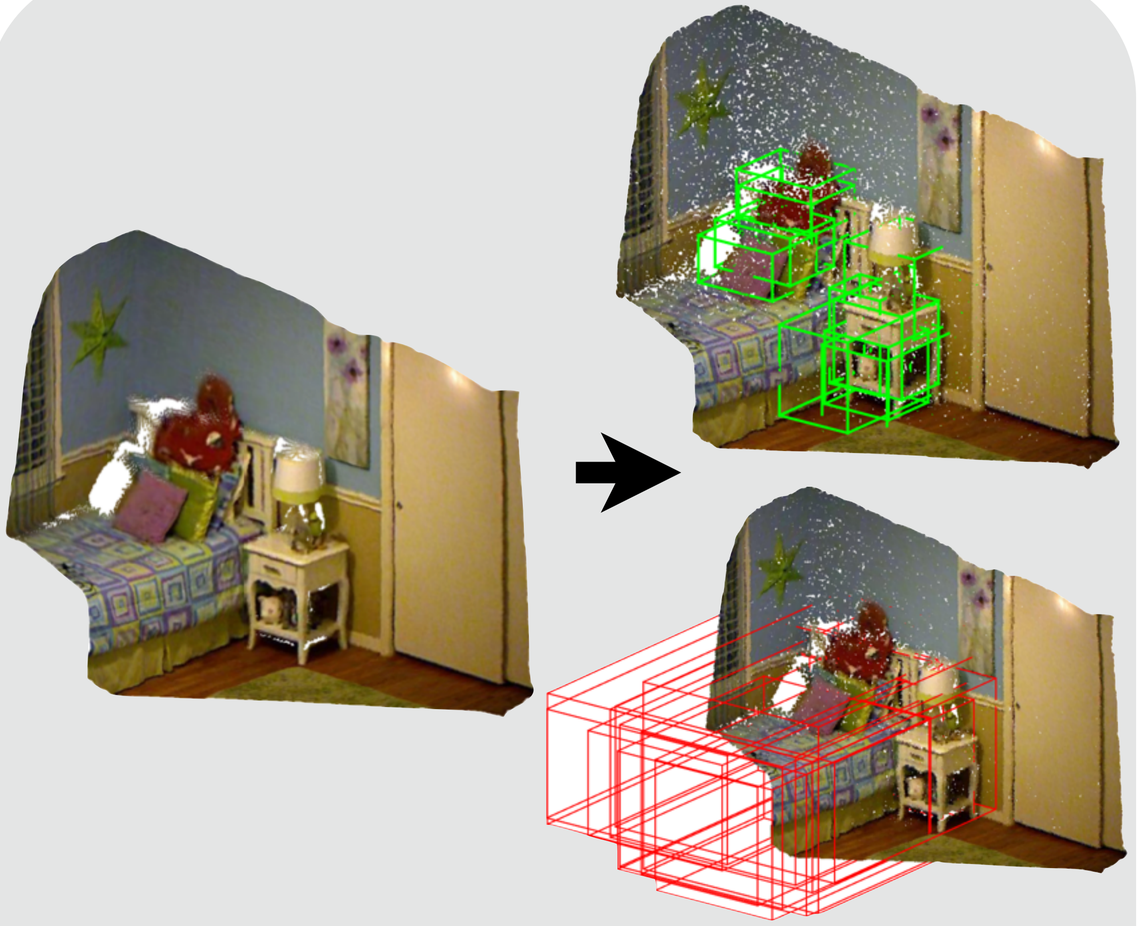
Output: 3D Amodal Boxes

Deep Sliding Shapes

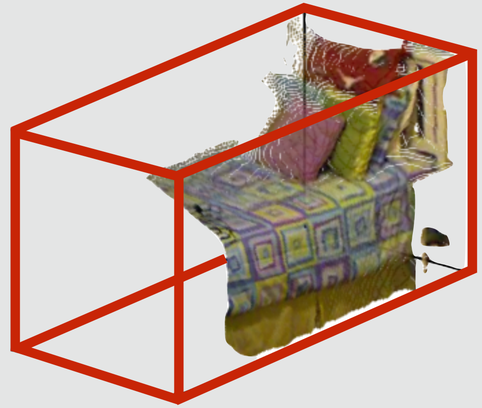


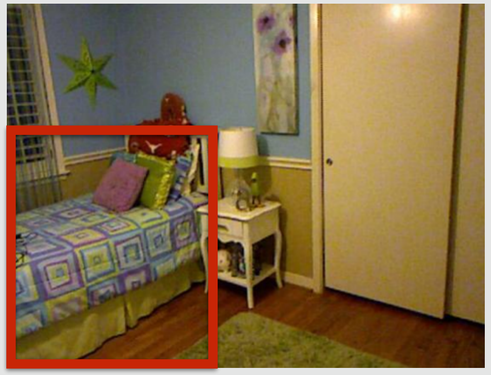
Region Proposal

Deep Sliding Shapes



Region Proposal



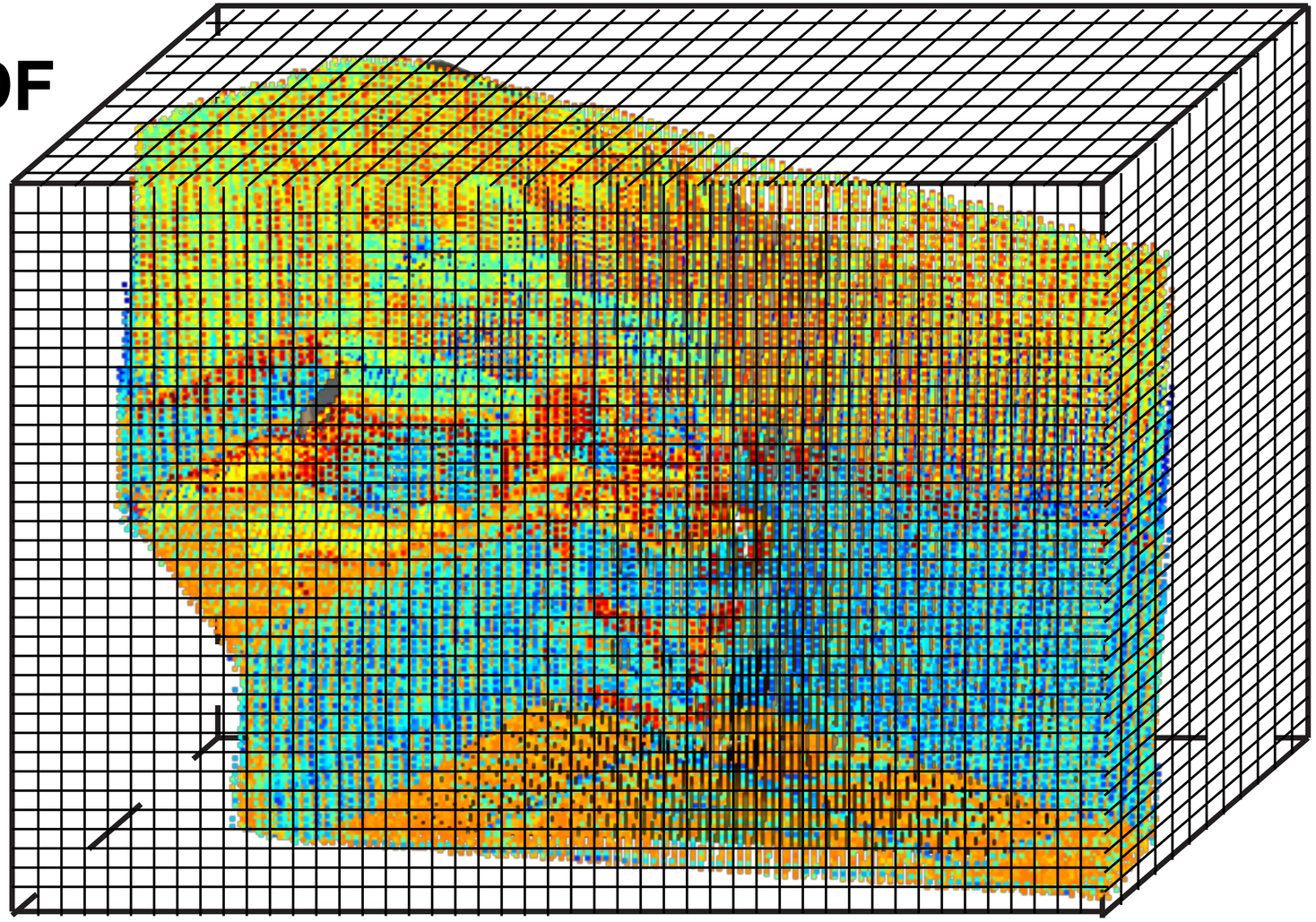


bed

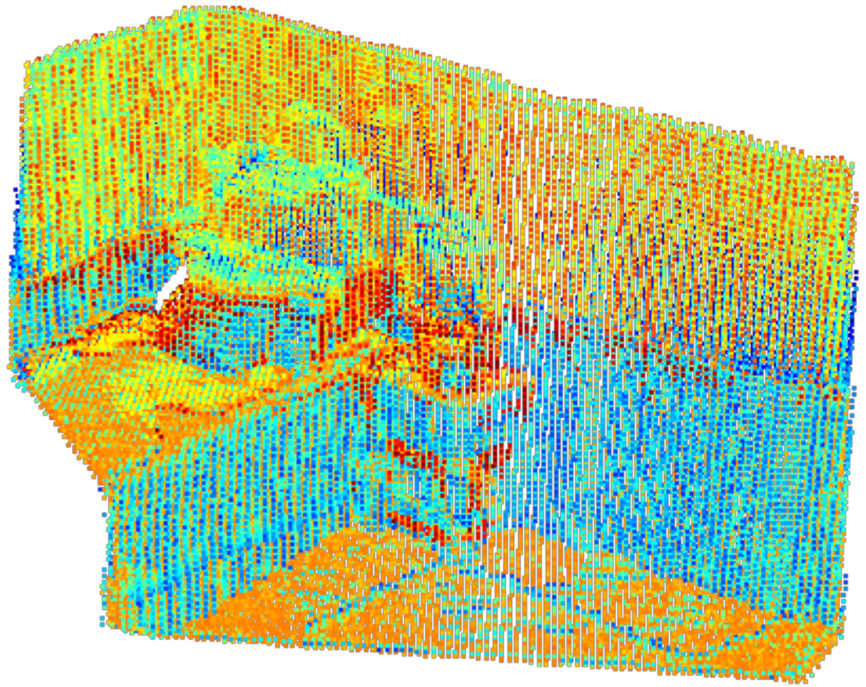
Recognition

Encoding 3D Representation

Compute TSDF

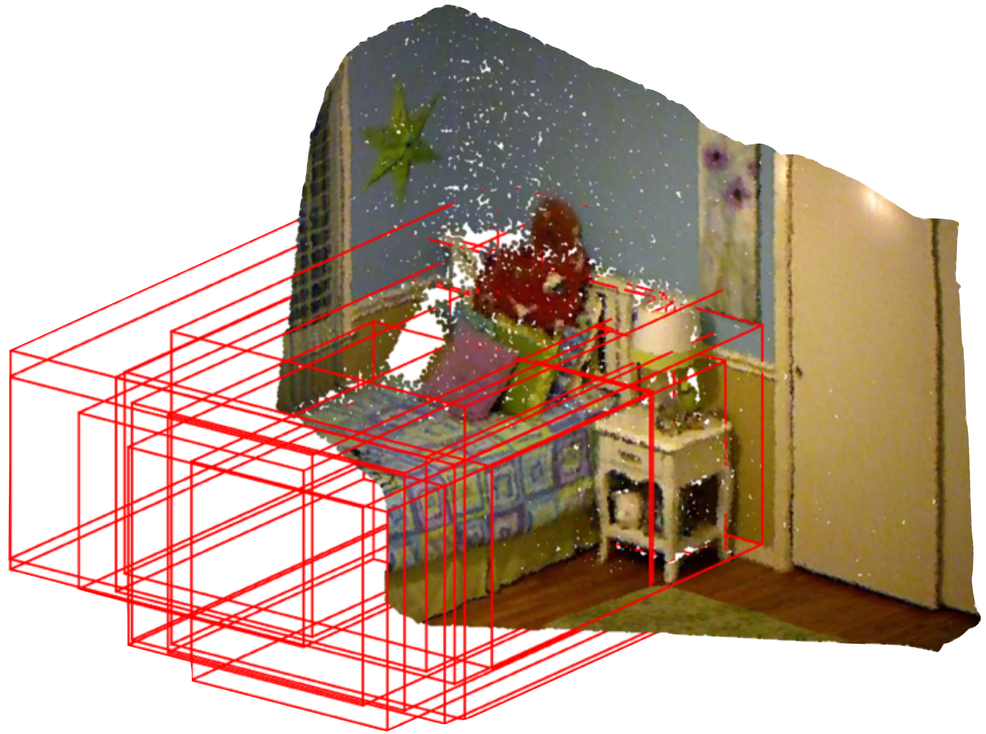


3D Region Proposal Network



TSDF

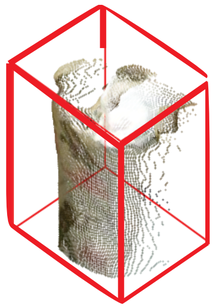
Region Proposal Network



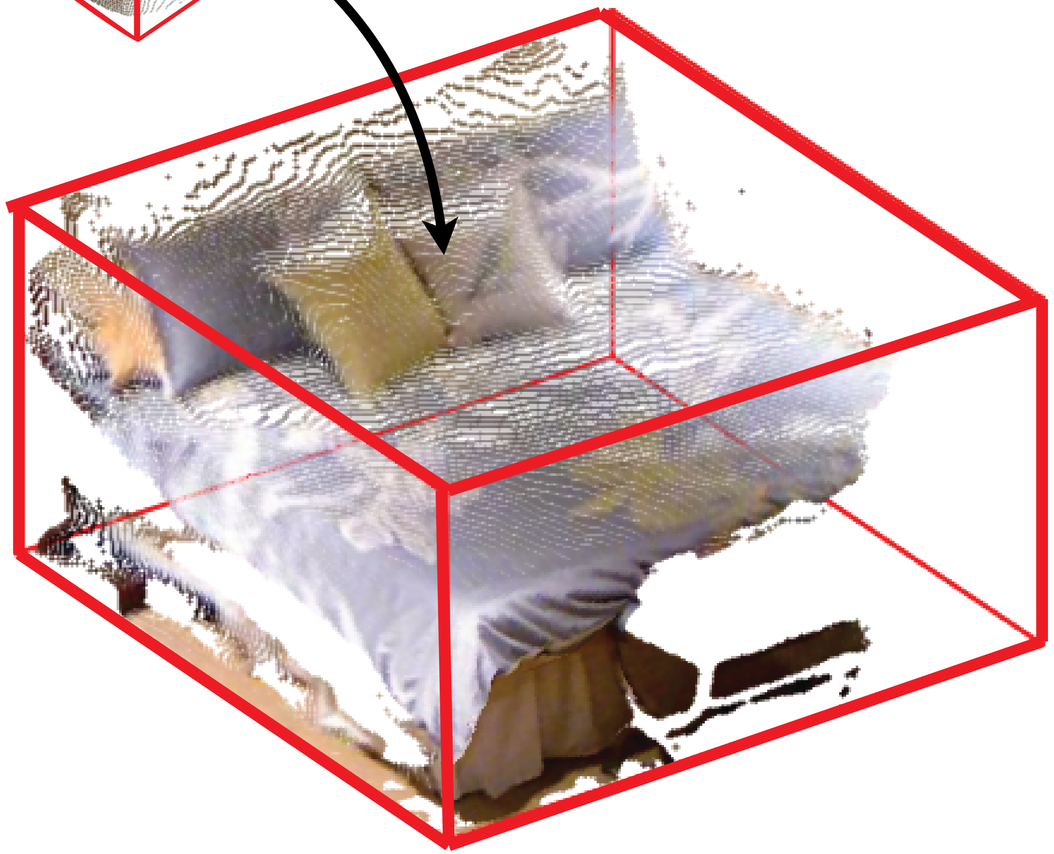
3D Region Proposals

3D Region Proposal Network

Pixel Area
×3

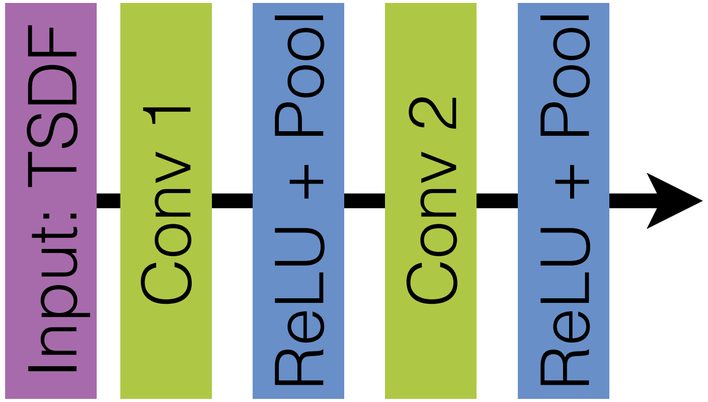


Physical Size
×50

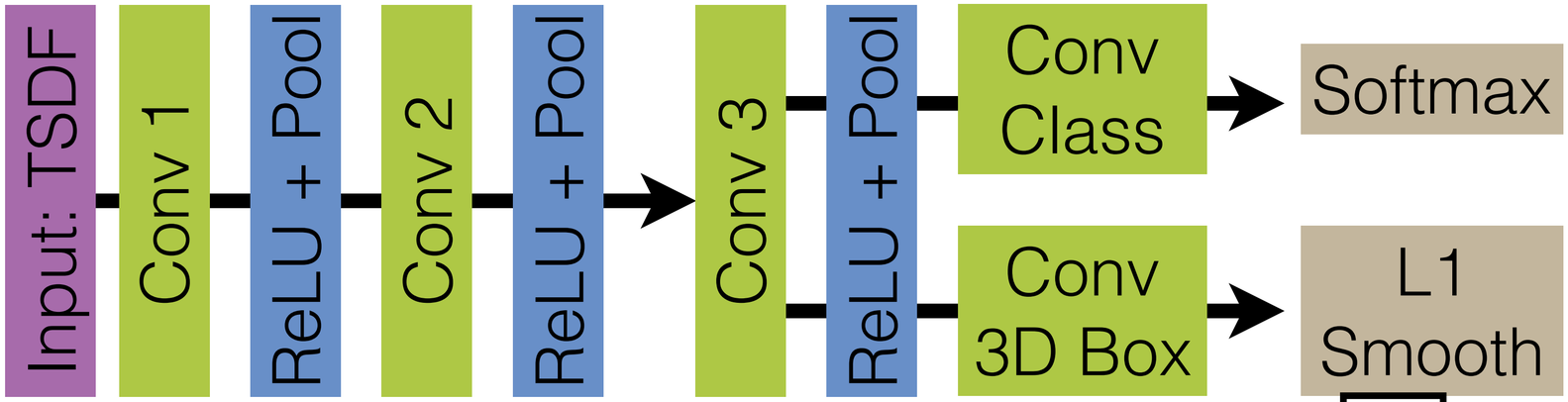


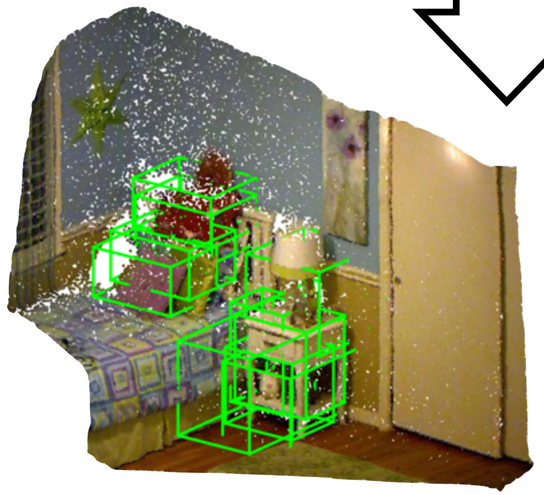
Multi-scale 3D Region Proposal Network

Multi-scale 3D Region Proposal Network



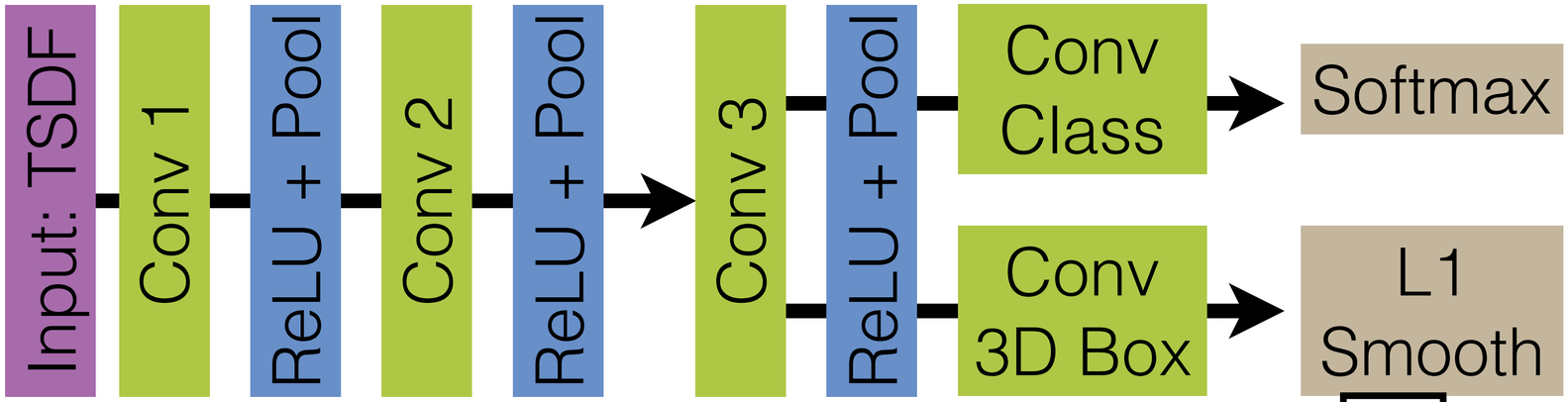
Multi-scale 3D Region Proposal Network

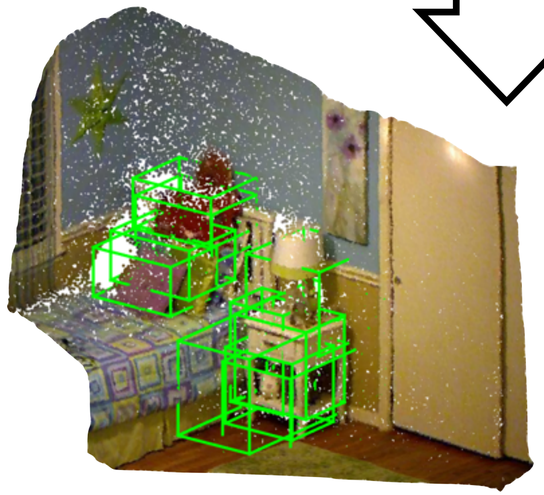




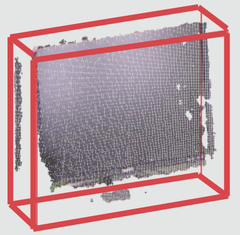
Receptive field: 0.4 m^3

Multi-scale 3D Region Proposal Network

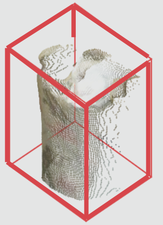




Receptive field: 0.4 m^3



0.6×0.2×0.4 m

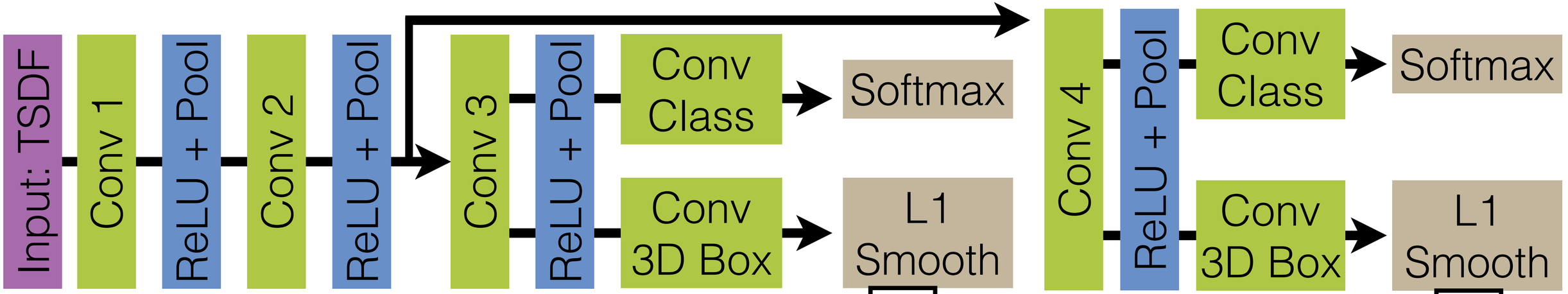


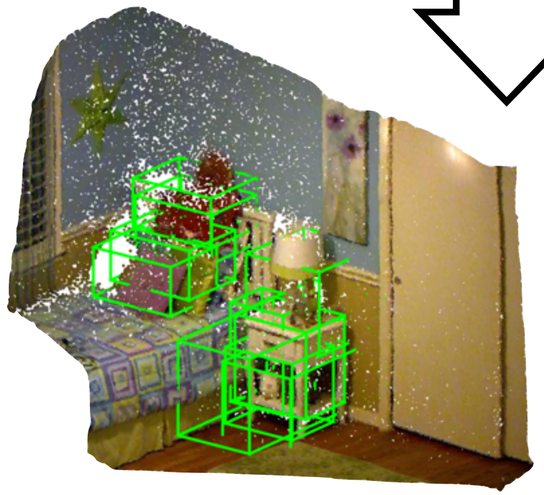
0.6×0.2×0.4 m

0.5×0.5×0.2 m

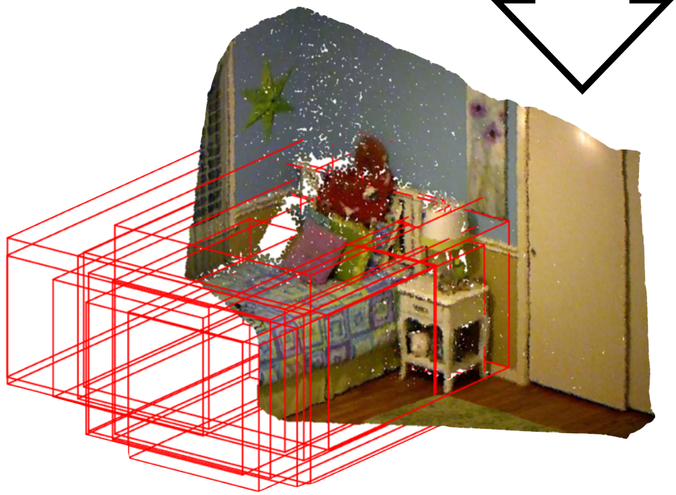
Level 1 Anchors

Multi-scale 3D Region Proposal Network





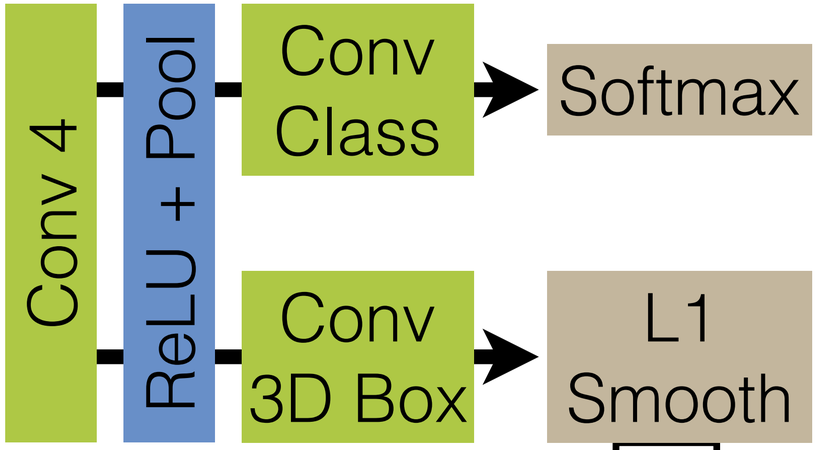
Receptive field: 0.4 m^3

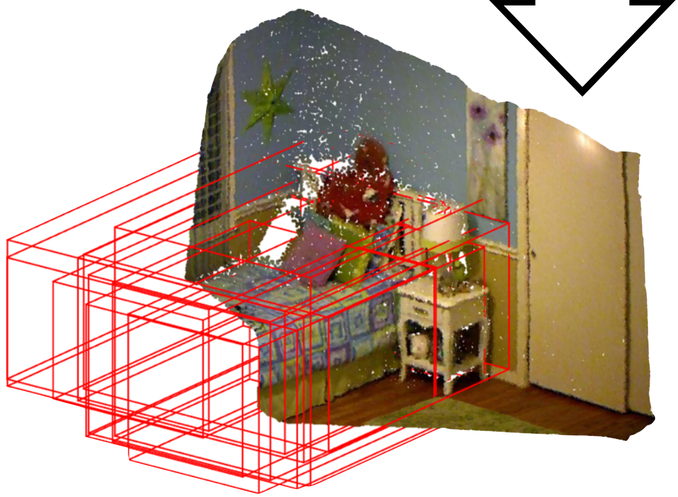


Receptive field: 1 m^3

Multi-scale 3D Region Proposal Network

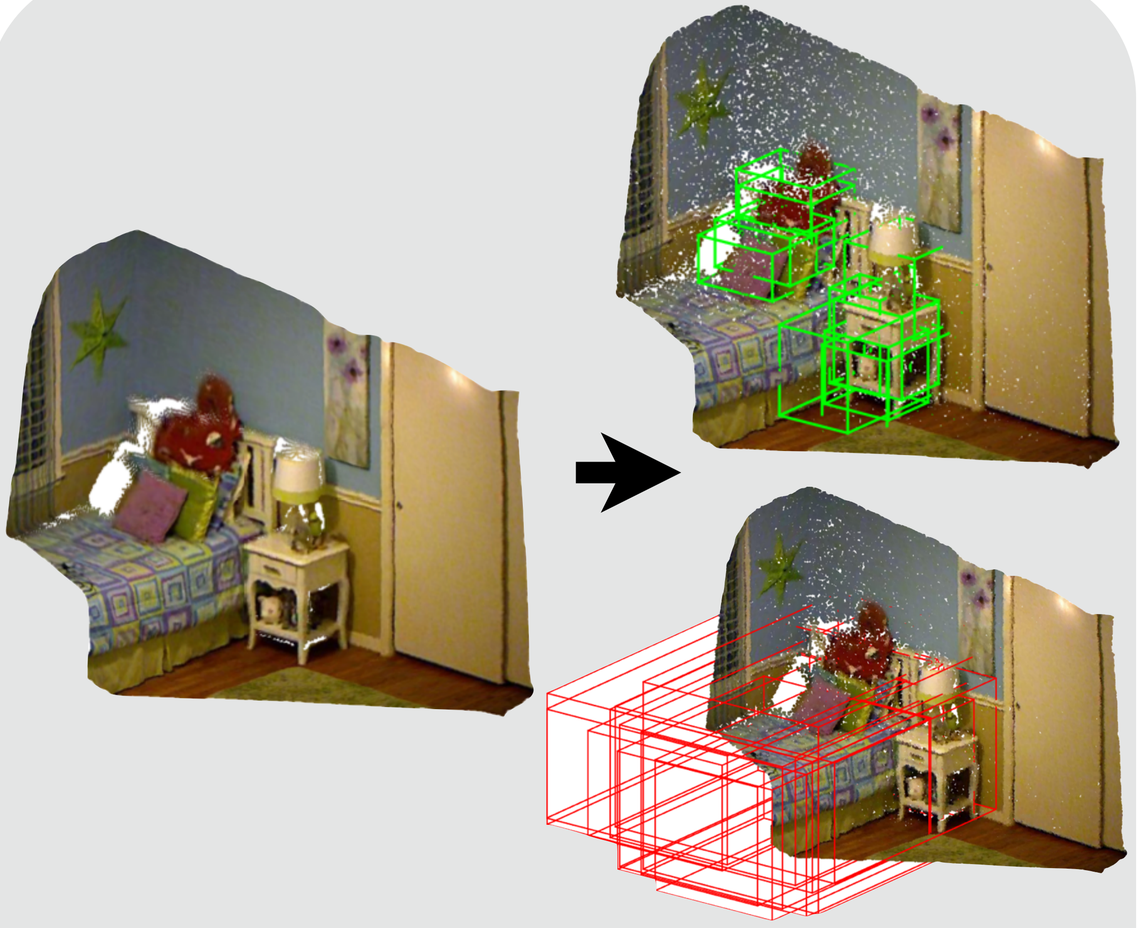
Level 2 Anchors



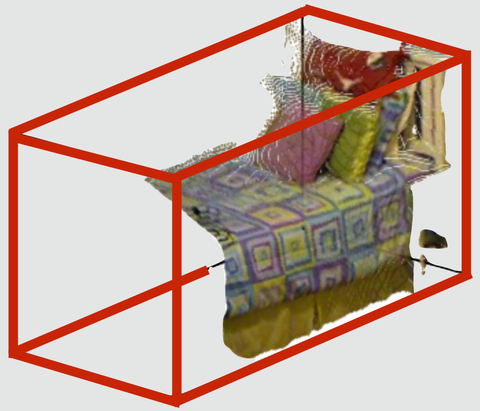


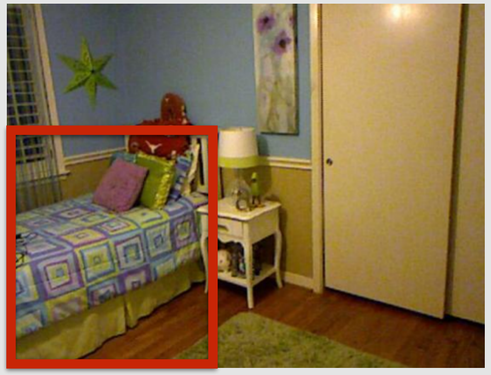
Receptive field: 1 m^3

Algorithm



Region Proposal

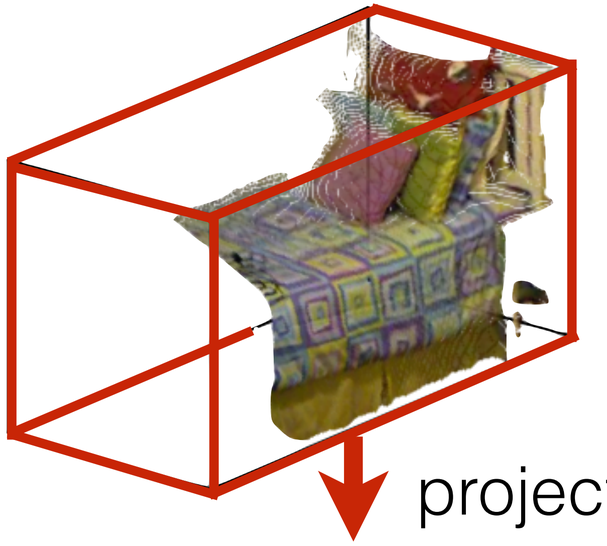




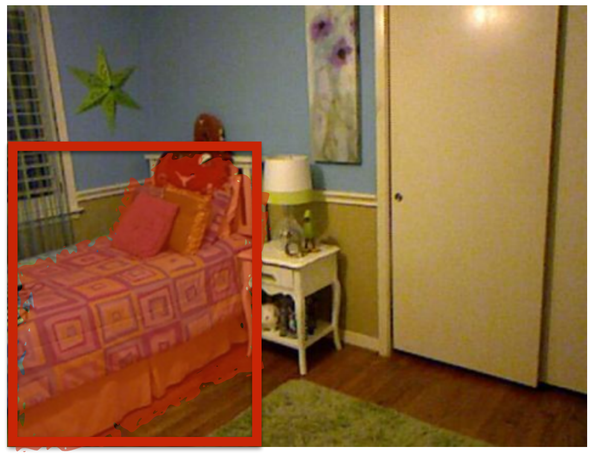
bed

Recognition

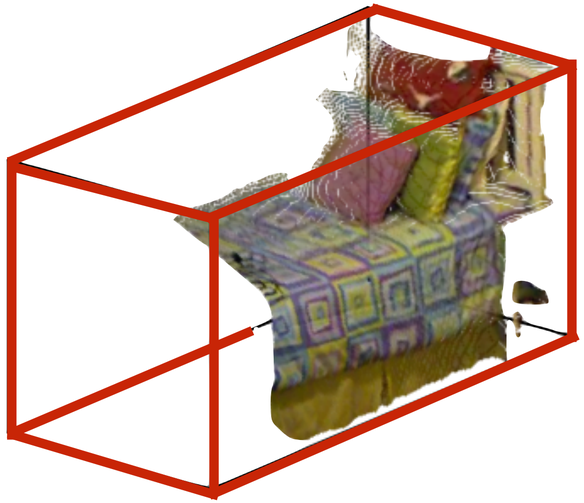
Joint Object Recognition Network

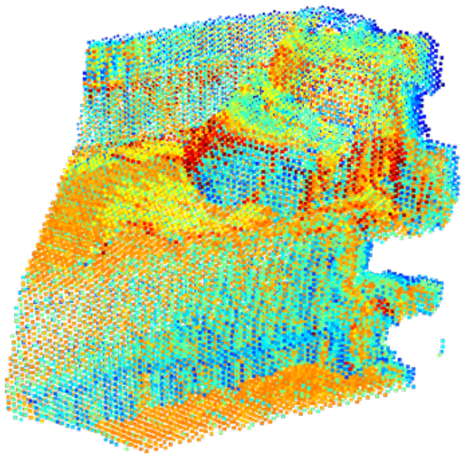


project to 2D



Joint Object Recognition Network





TSDF

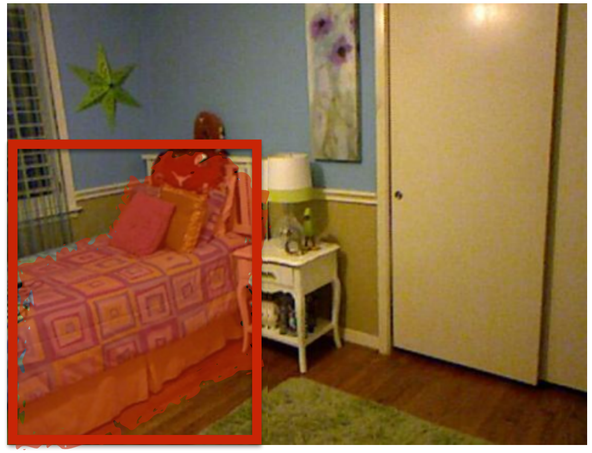
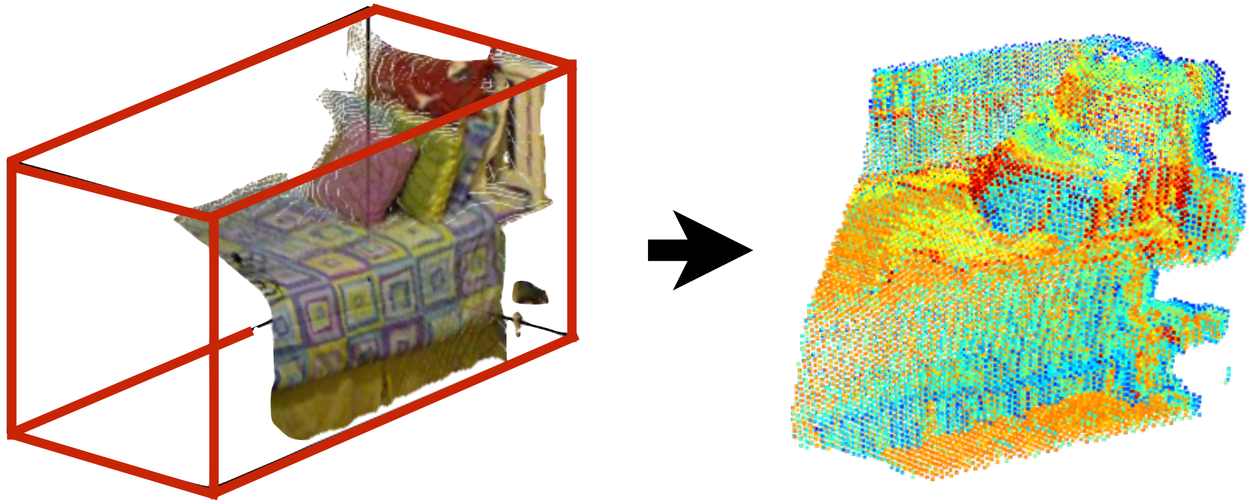
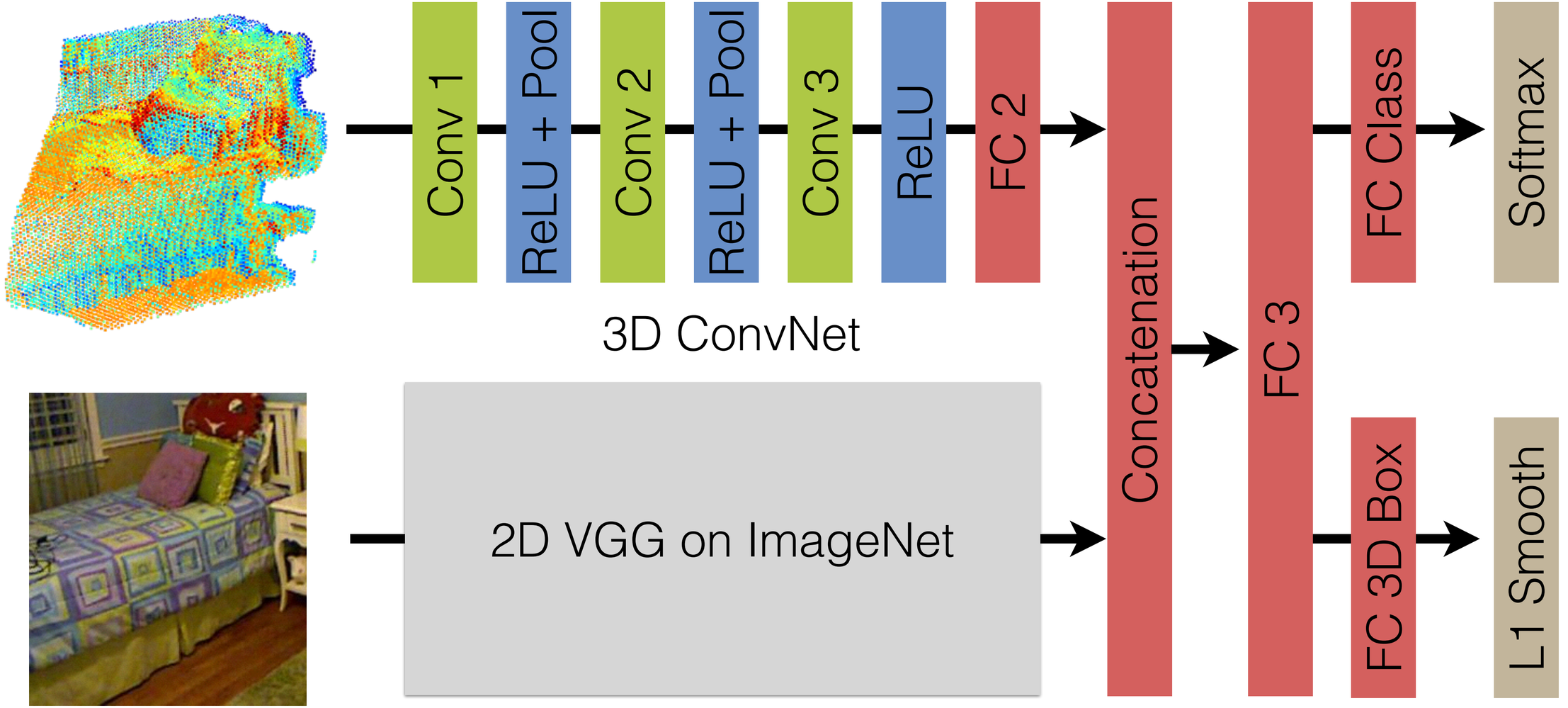


Image Patch

Joint Object Recognition Network



Joint Object Recognition Network



Marvin

A minimalist GPU-only N-dimensional ConvNet framework

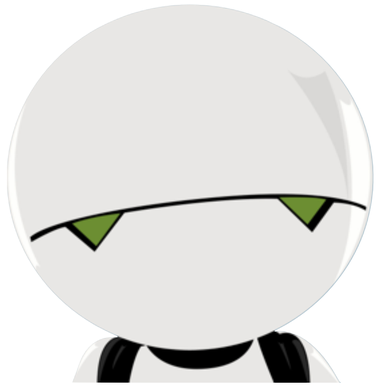
Marvin: A Minimalist GPU-only N-Dimensional ConvNets Framework — Edit

24 commits | **4** branches | **1** release | **5** contributors

Branch: **master** | **New pull request** | **New file** | **Find file** | **HTTPS** | https://github.com/Princetc | **Download ZIP**

danielsuo Update README.md — Latest commit `1a2e0f0` 11 hours ago

data	add demo support for places	14 hours ago
examples	Update README.md	11 hours ago
models	add demo	15 hours ago
python/marvin	Move the python package to a separate folder	10 days ago
tools	Move the python package to a separate folder	10 days ago
.gitignore	I'm alive!	22 days ago
CHANGELOG	Add CHANGELOG	12 days ago
LICENSE	I'm alive!	22 days ago
README.md	Fix BibTeX entry.	20 days ago
compile.sh	I'm alive!	22 days ago
marvin.cu	marvin.cu: convert tab to space	12 hours ago
marvin.hpp	move data type selection micro	12 hours ago



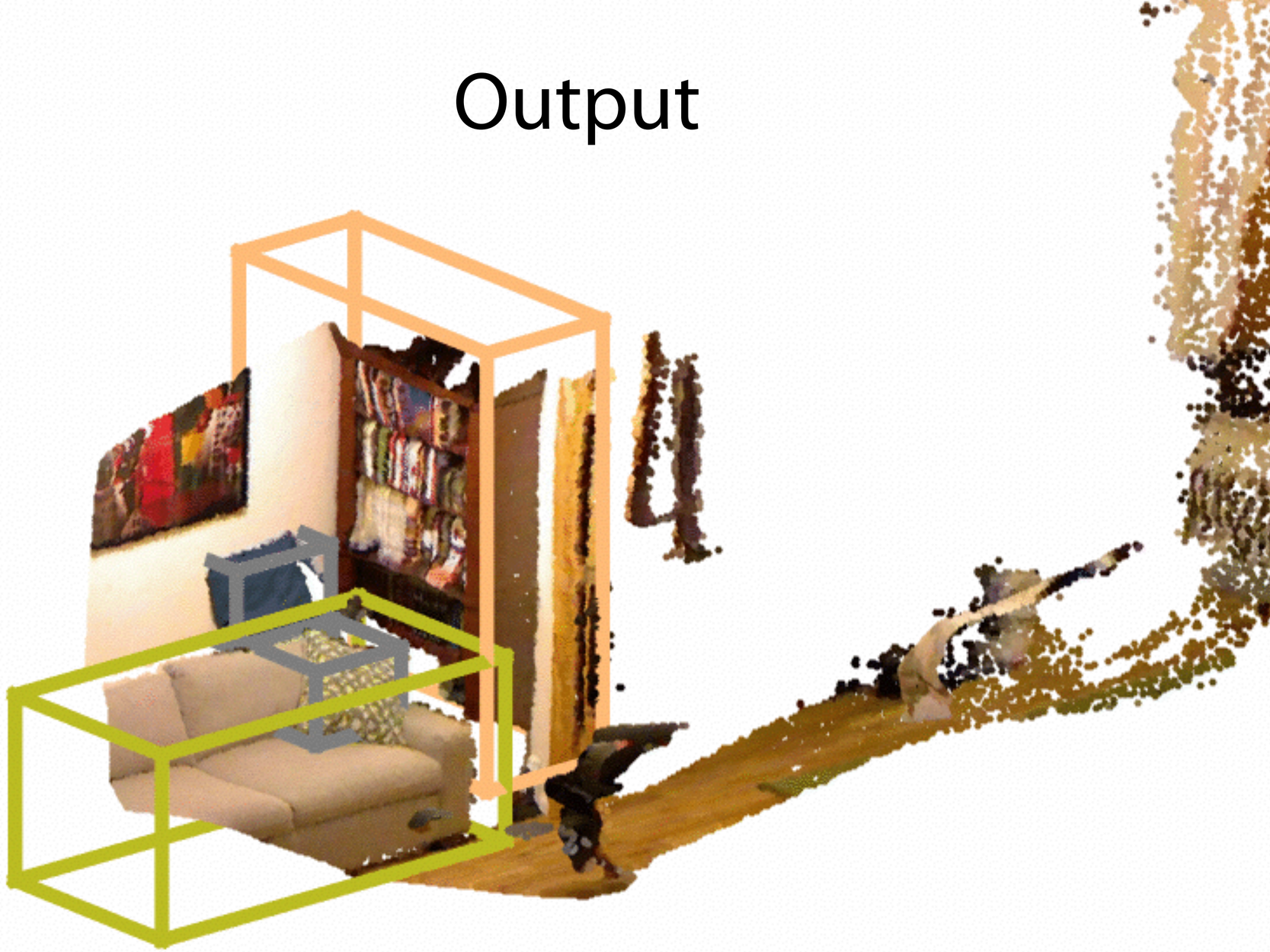
http://marvin.is

Best 3D Object Detector

	Algorithm	input						mAP
3D Non-Deep Learning	Sliding Shapes [20]	d	33.5	29	34.5	33.8	67.3	39.6
2D Deep Learning	[8] on instance seg	d	71	18.2	49.6	30.4	63.4	46.5
	[8] on instance seg	rgbd	74.7	18.6	50.3	28.6	69.7	48.4
	[8] on estimated model	d	72.7	47.5	54.6	40.6	72.7	57.6
	[8] on estimated model	rgbd	73.4	44.2	57.2	33.4	84.5	58.5
3D Deep Learning	ours [depth only]	d	83.0	58.8	68.6	49.5	79.2	67.8
	ours [depth + img]	rgbd	**84.7**	**61.1**	**70.5**	**55.4**	**89.9**	**72.3**

Input

Output



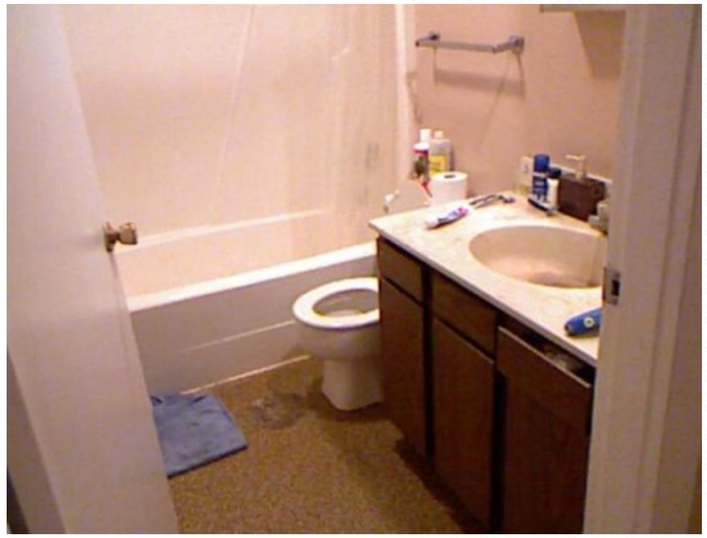
sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

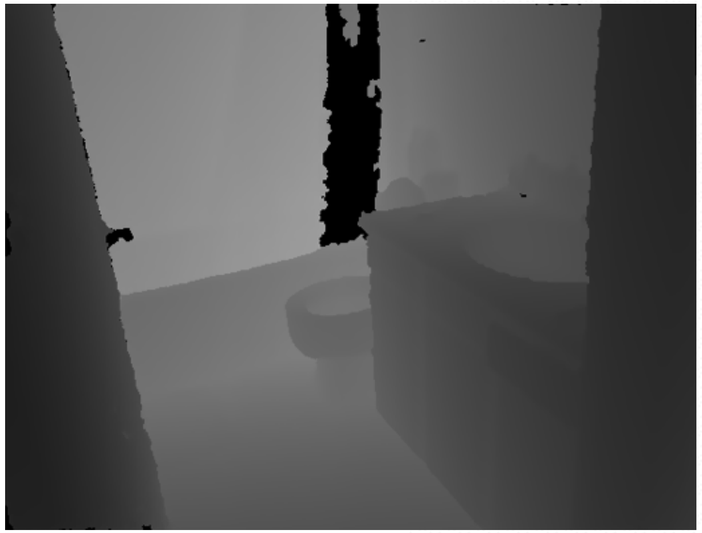
Input

Output

sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

Input

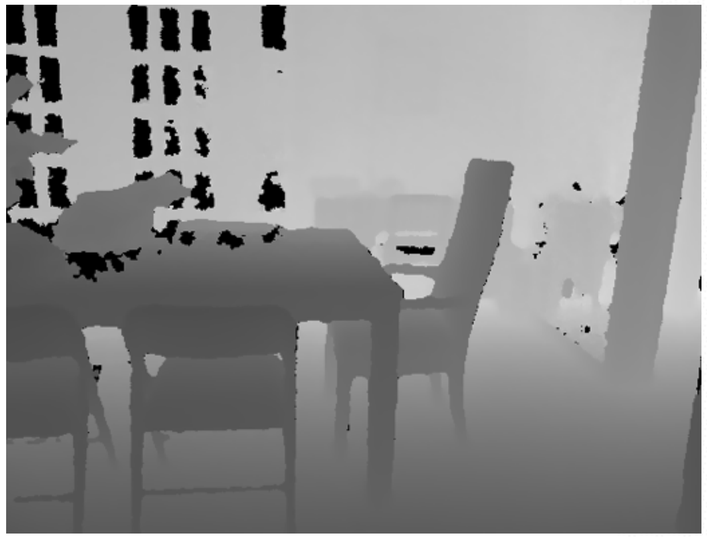




Output

sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

Input

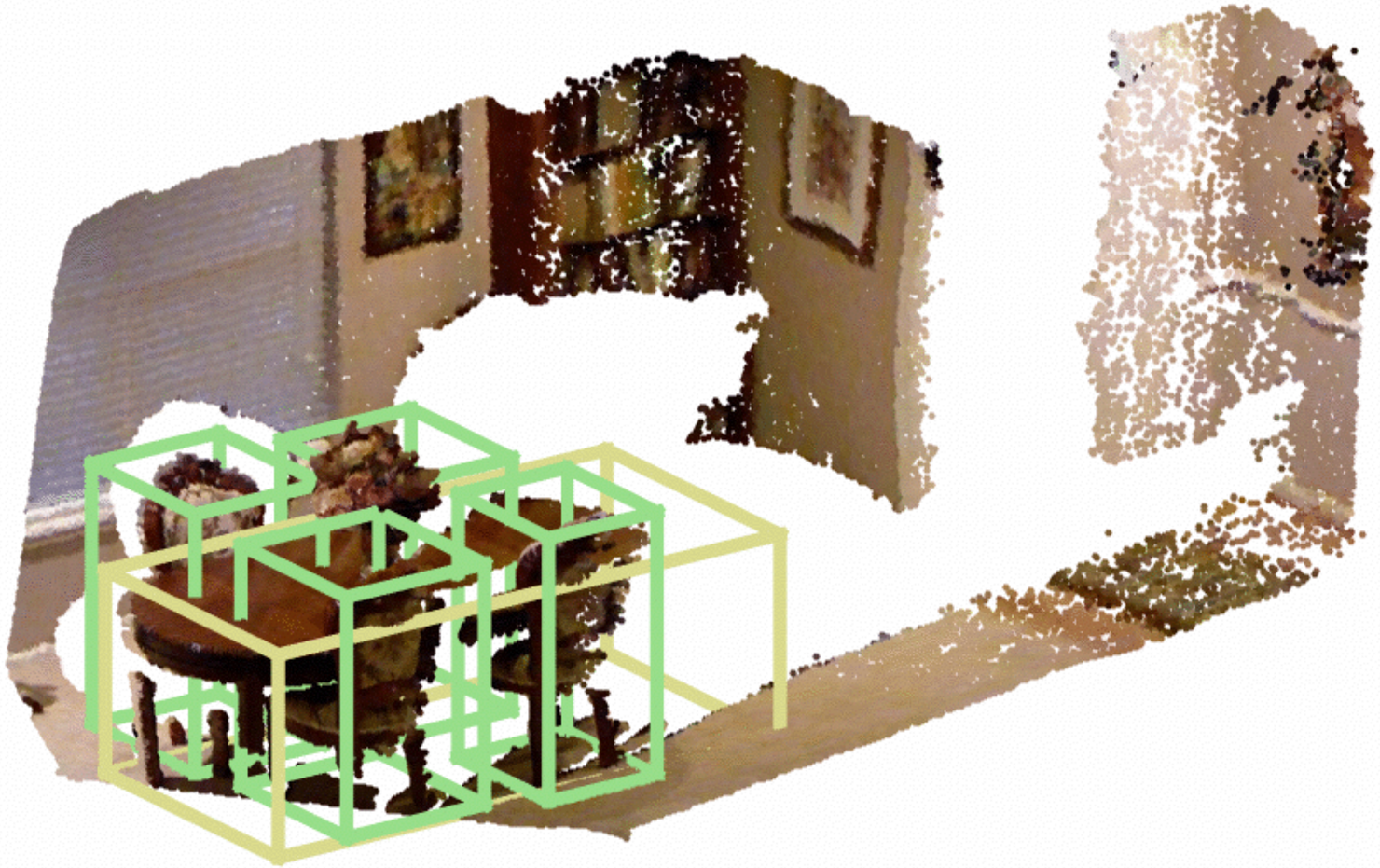


Output

sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

Input

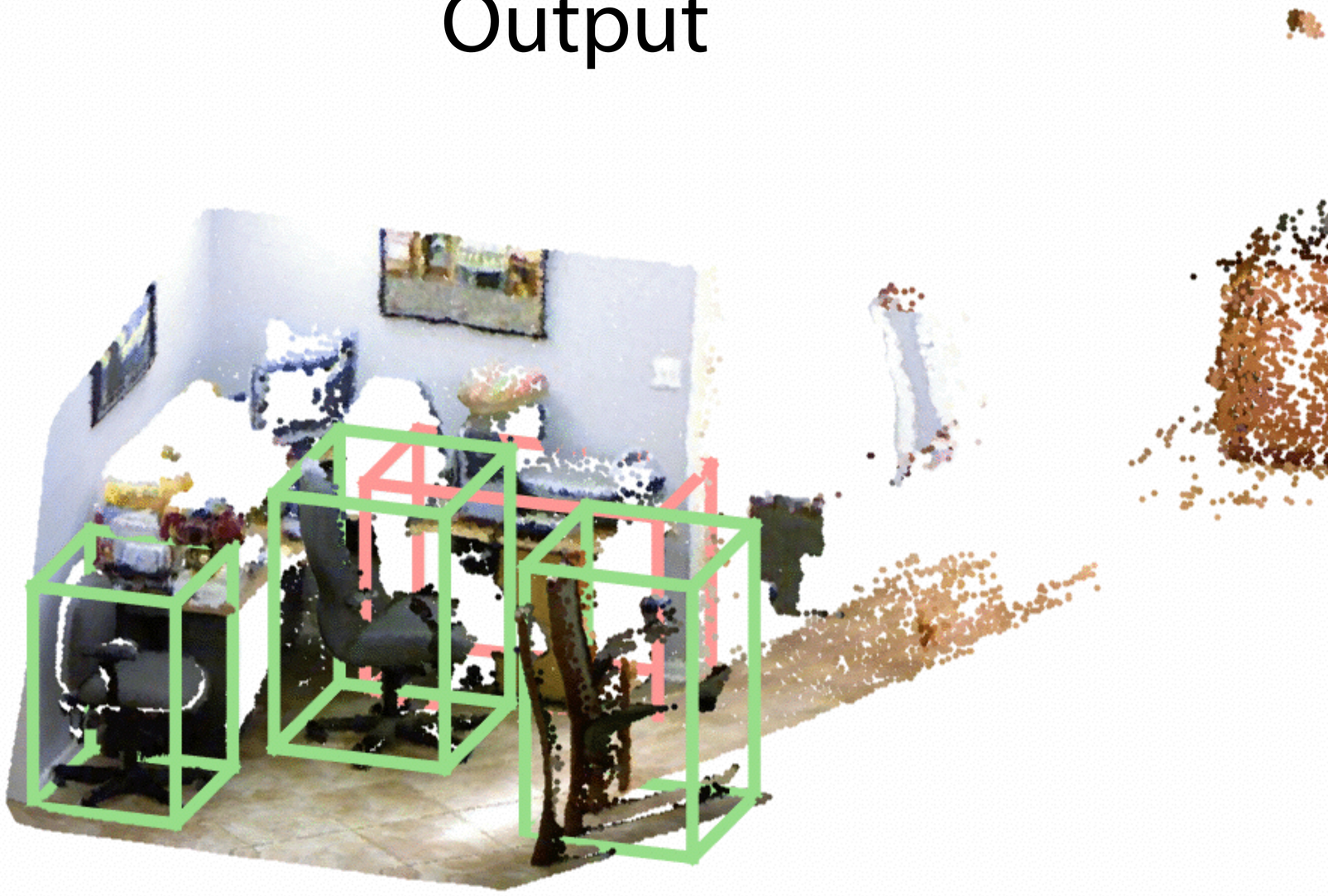
Output



sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

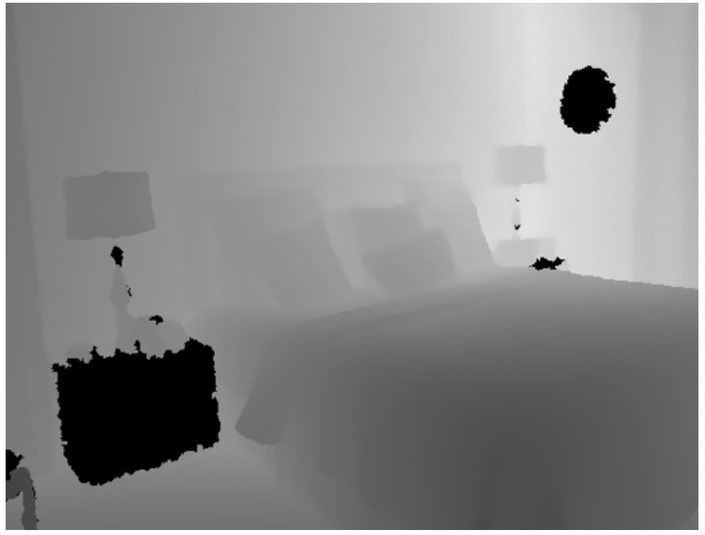
Input

Output

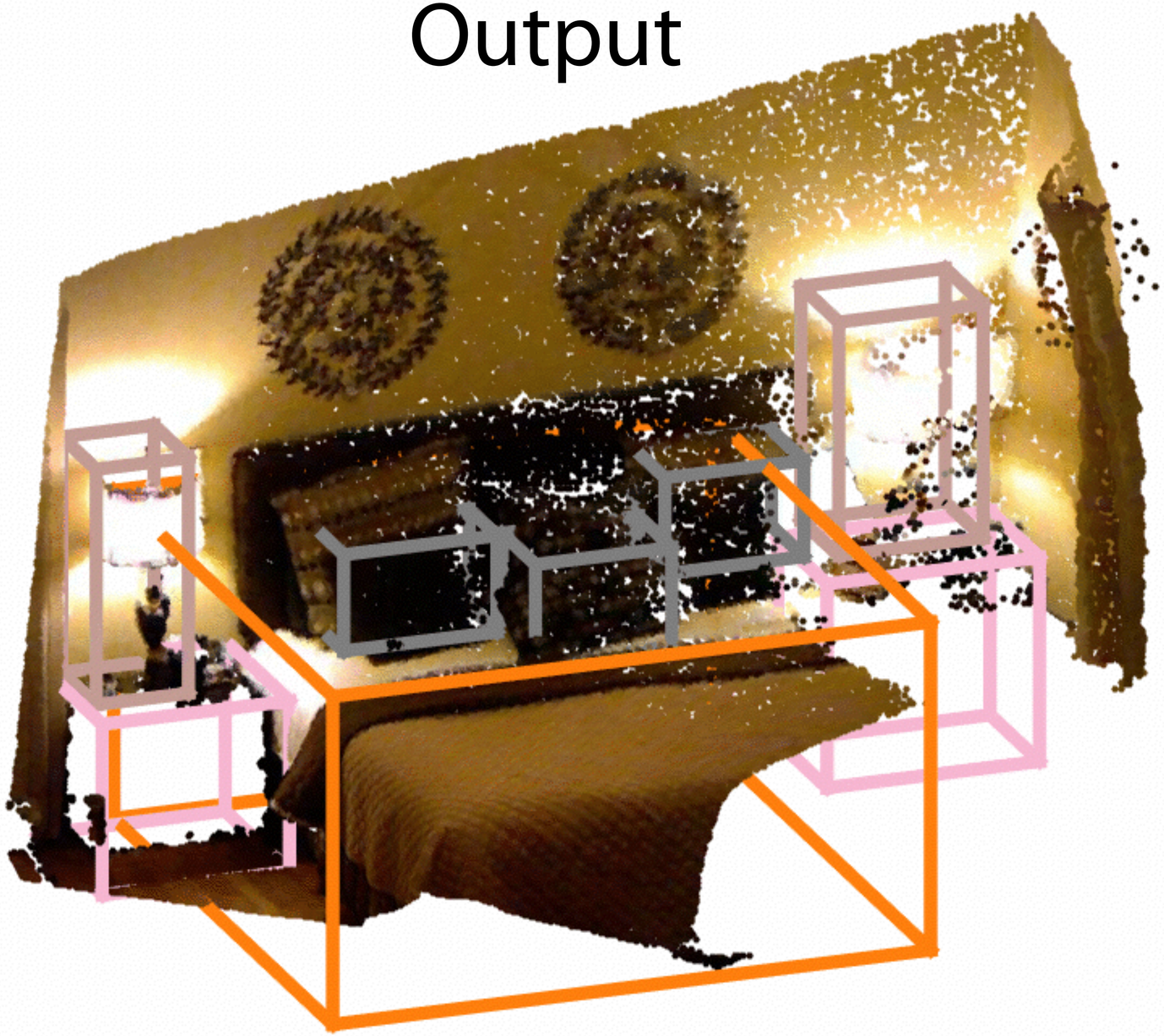


sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

Input

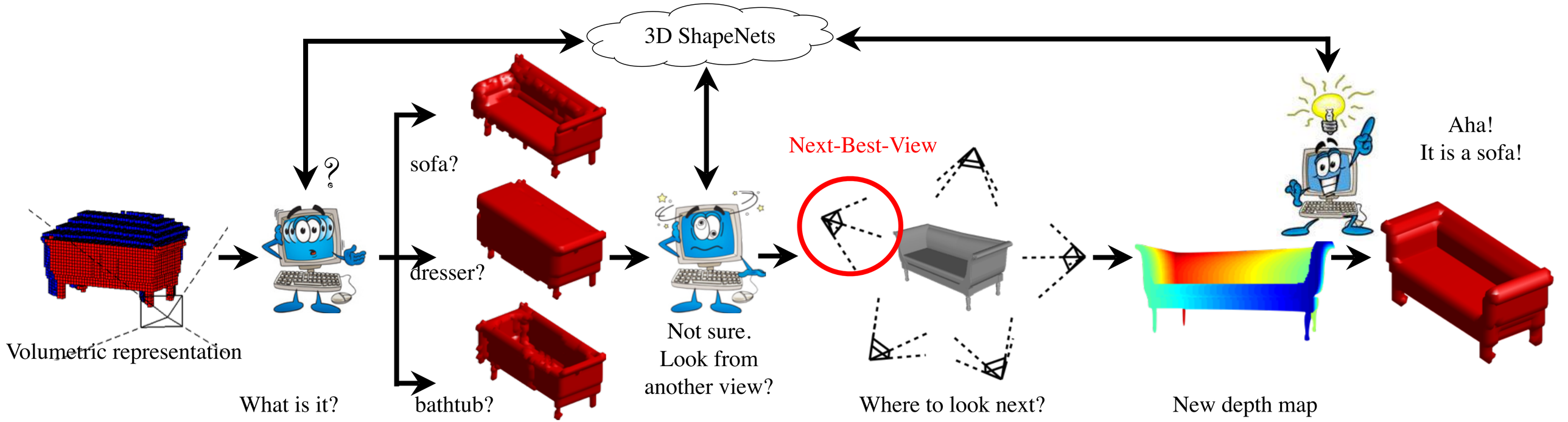


Output



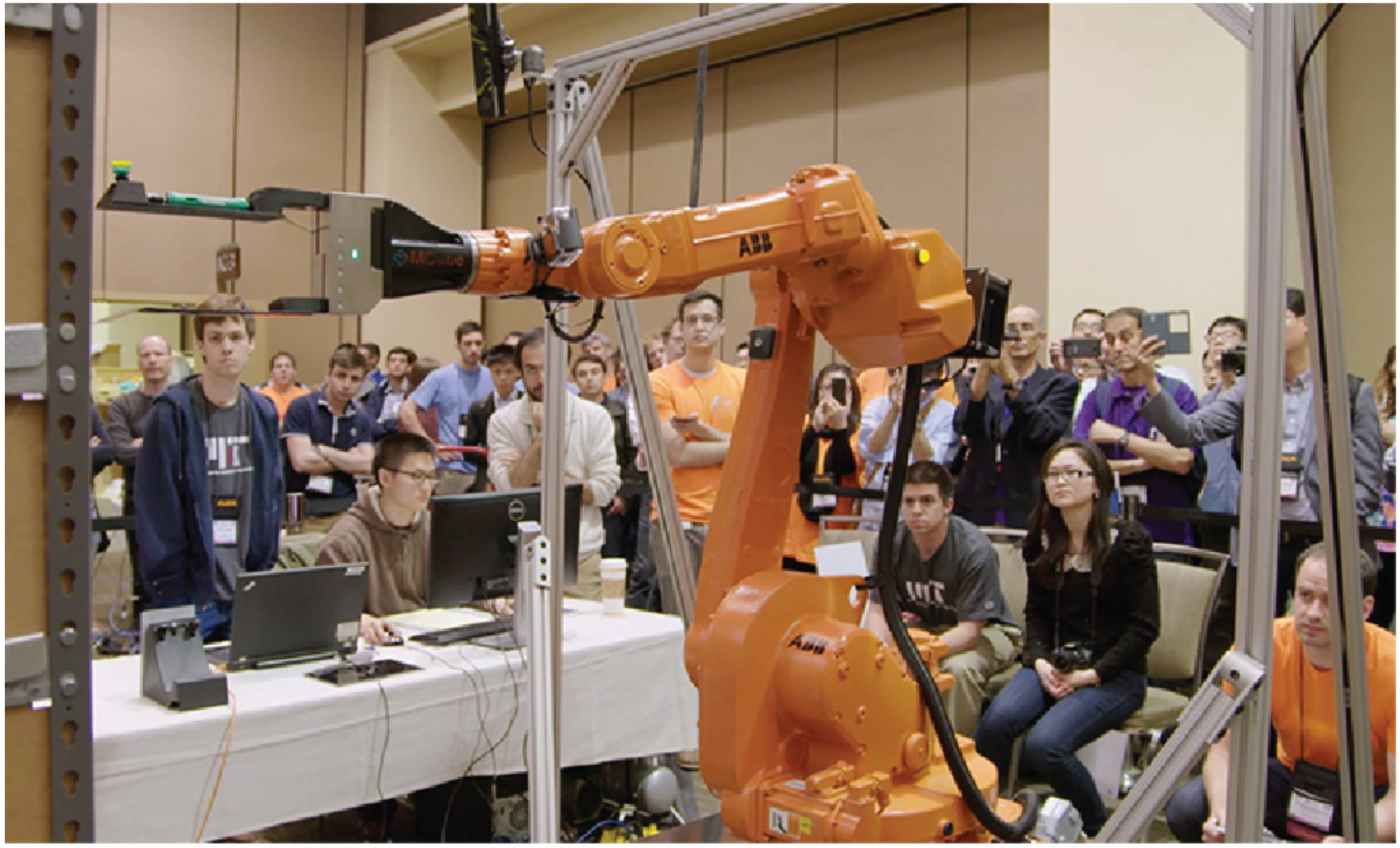
sofa bed bathtub garbage bin chair desk pillow bookshelf
table box monitor night stand door lamp sink toilet tv

Deep View Planning



Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang and J. Xiao
3D ShapeNets: A Deep Representation for Volumetric Shapes
Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015)

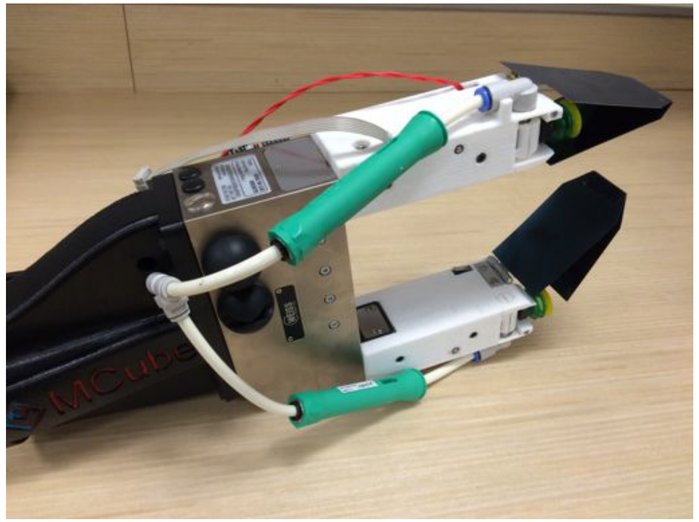
Amazon Picking Challenge



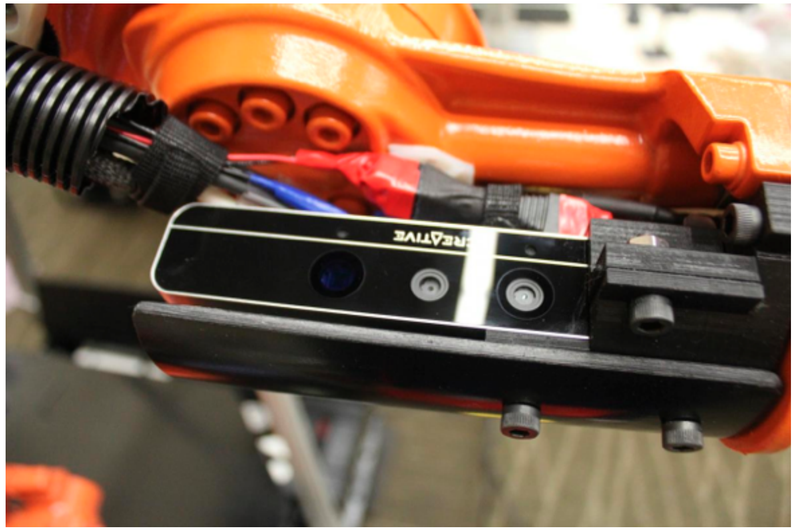
Team MIT+Princeton

Alberto Rodriguez

ABB 1600ID Robot Arm

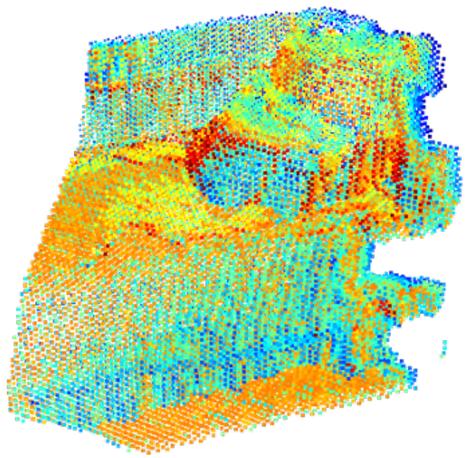


MIT Gripper



Princeton Camera Array

Multi-view +3D for 6D Pose Estimation



Conv 1

ReLU + Pool

Conv 2

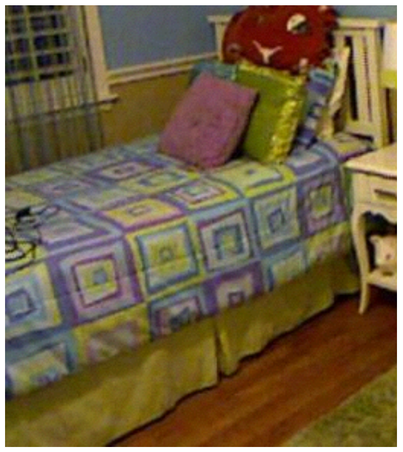
ReLU + Pool

Conv 3

ReLU

FC 2

3D ConvNet



2D VGG on ImageNet

Concatenation

FC 3

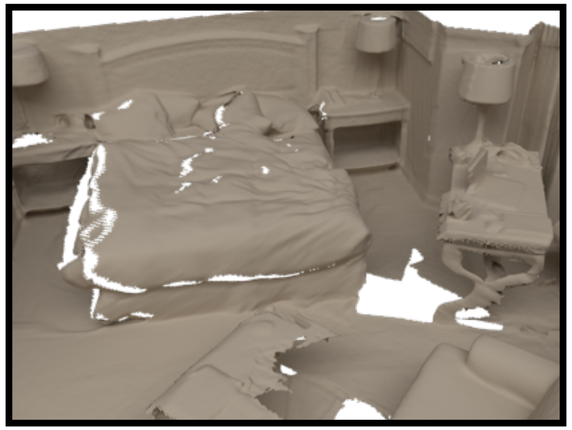
FC Class

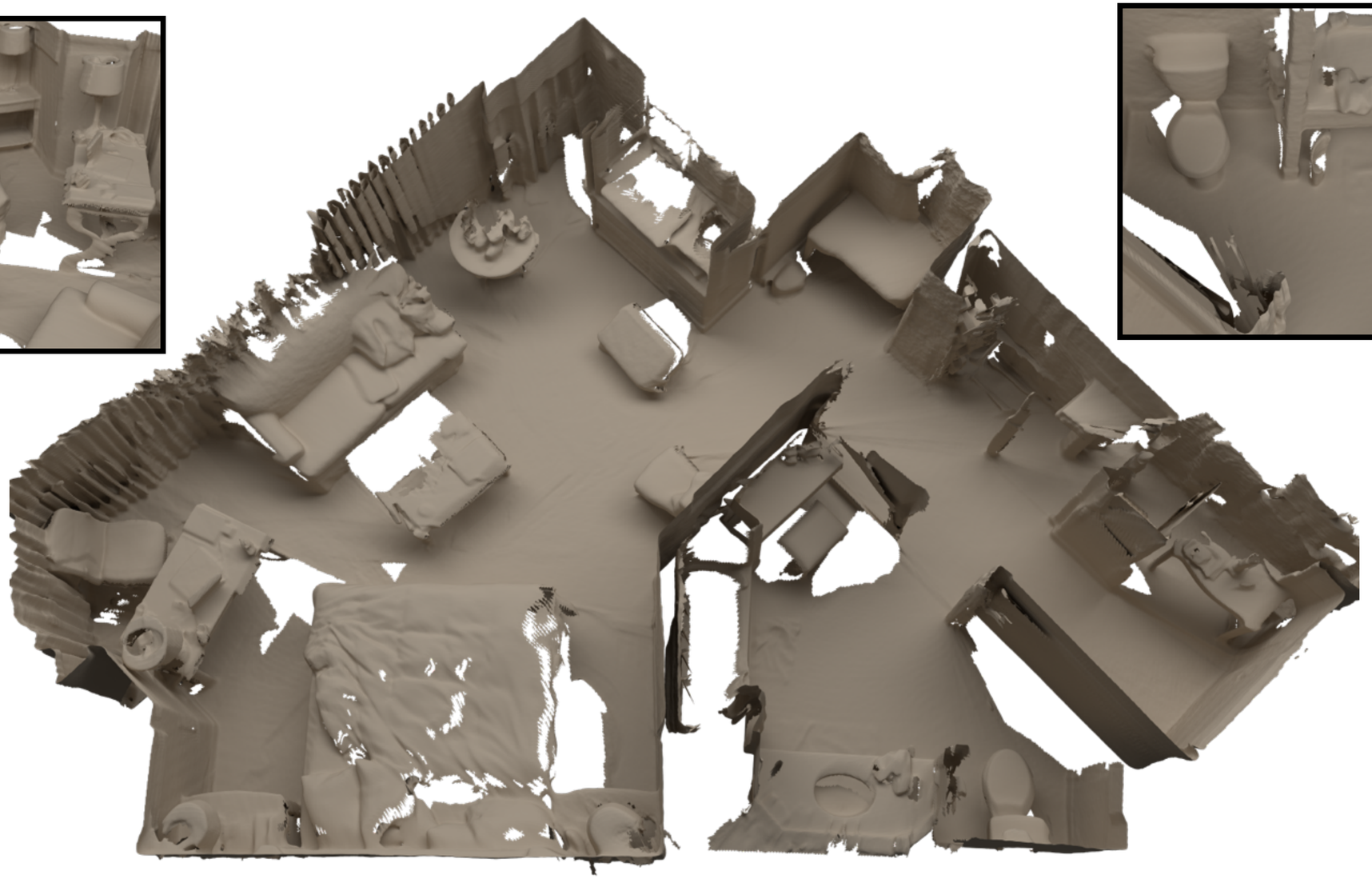
Softmax

FC 6D Pose

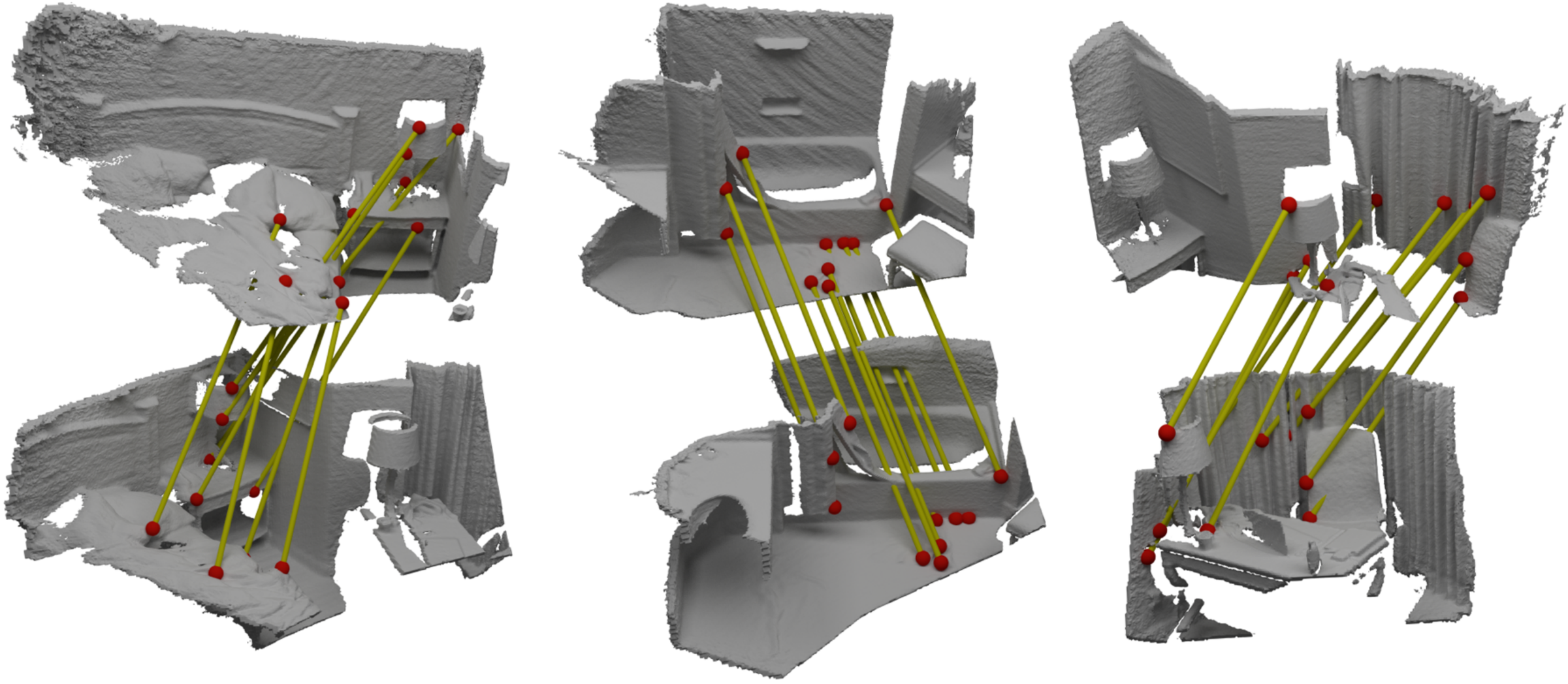
L1 Smooth

Deep Learning for SLAM

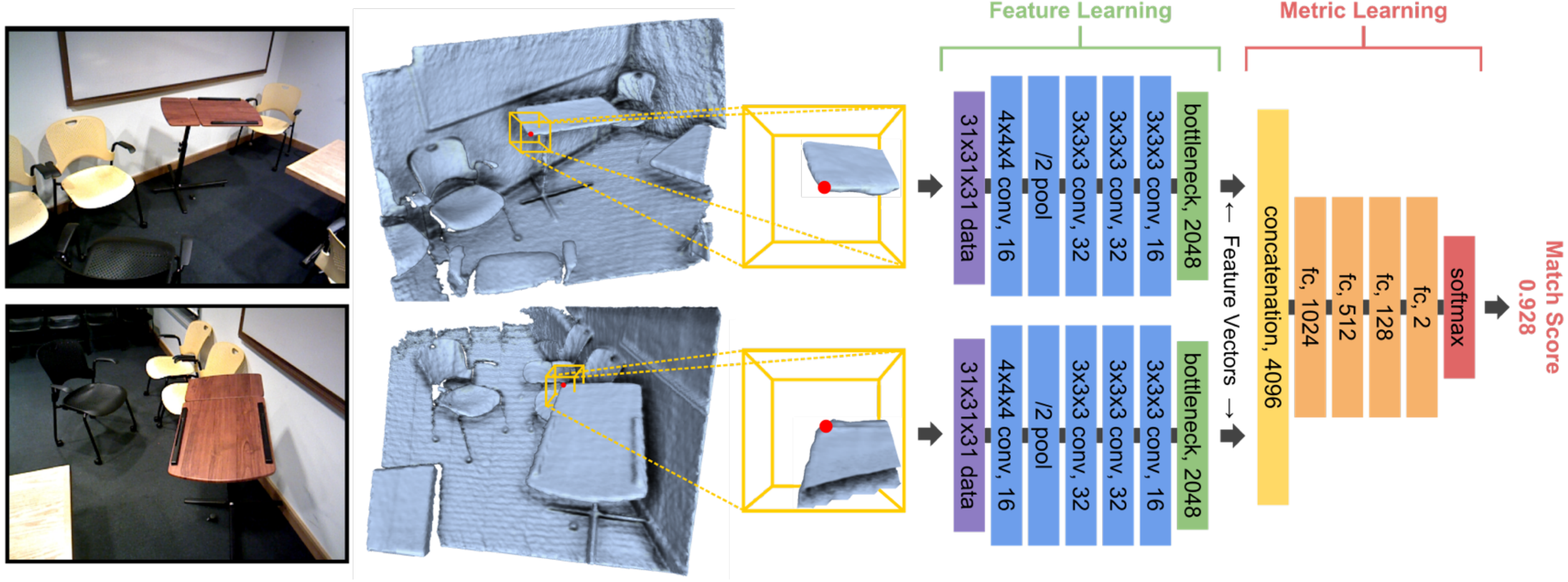




Deep Learning for SLAM

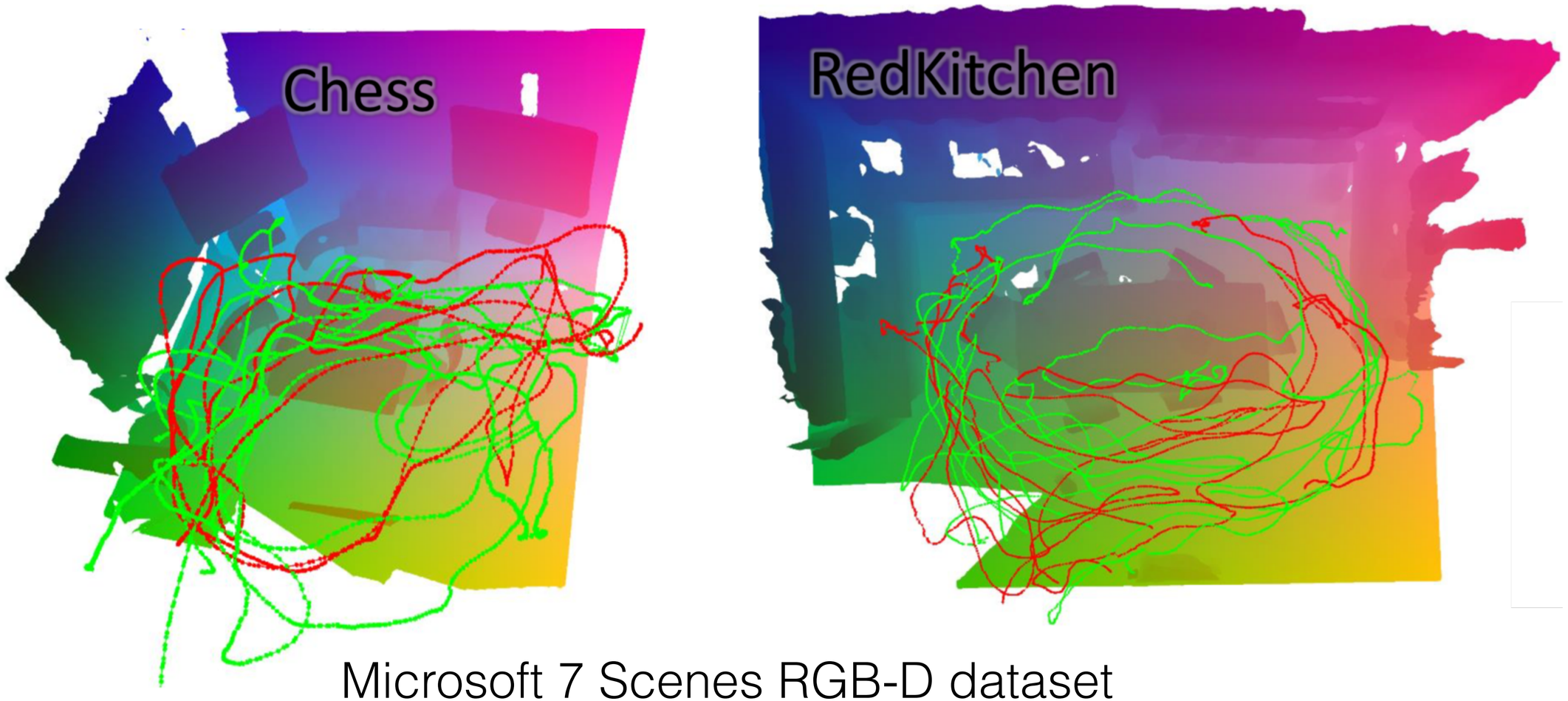


3D Deep Learning for Geometric Keypoints

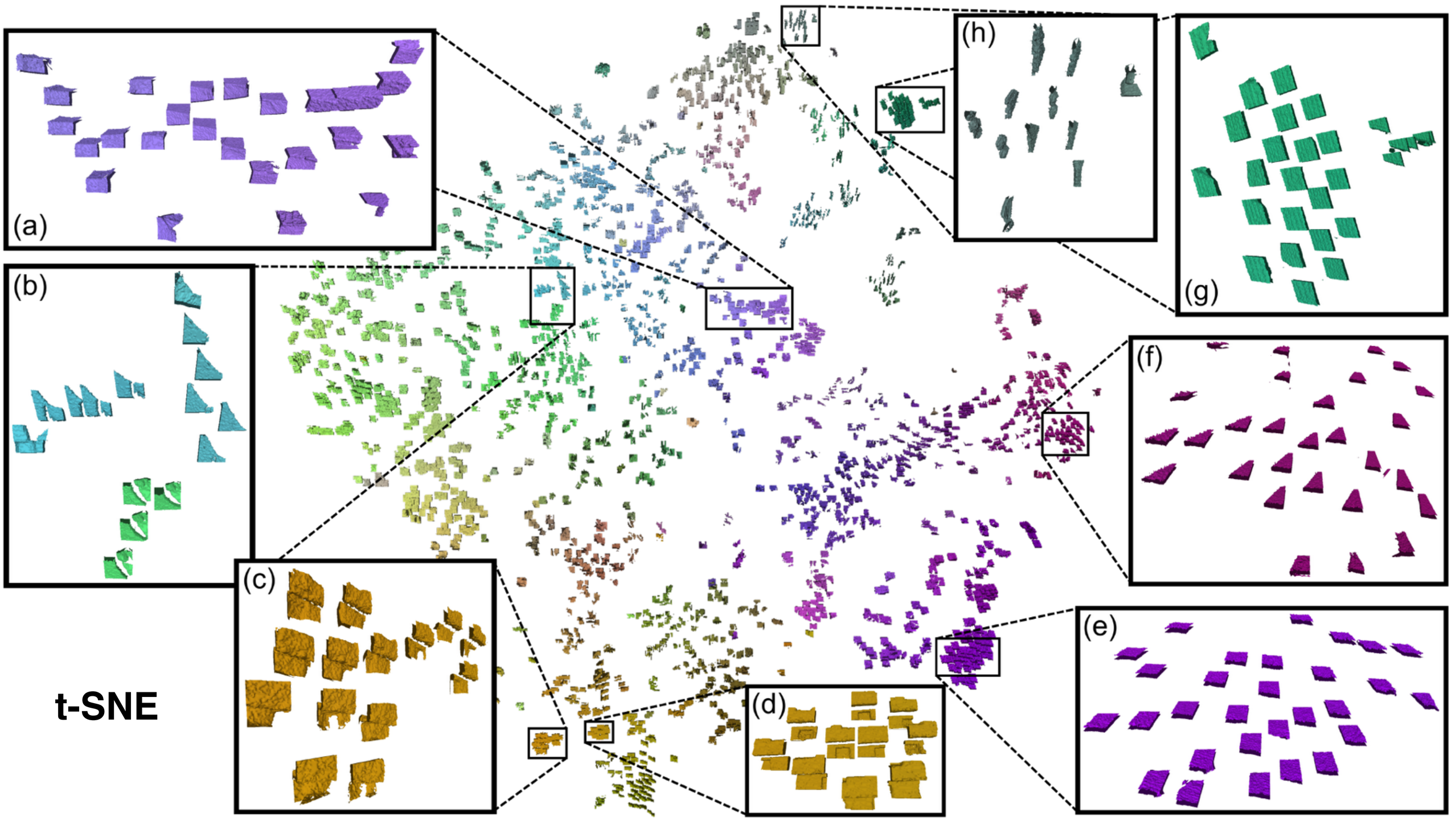


Under Review

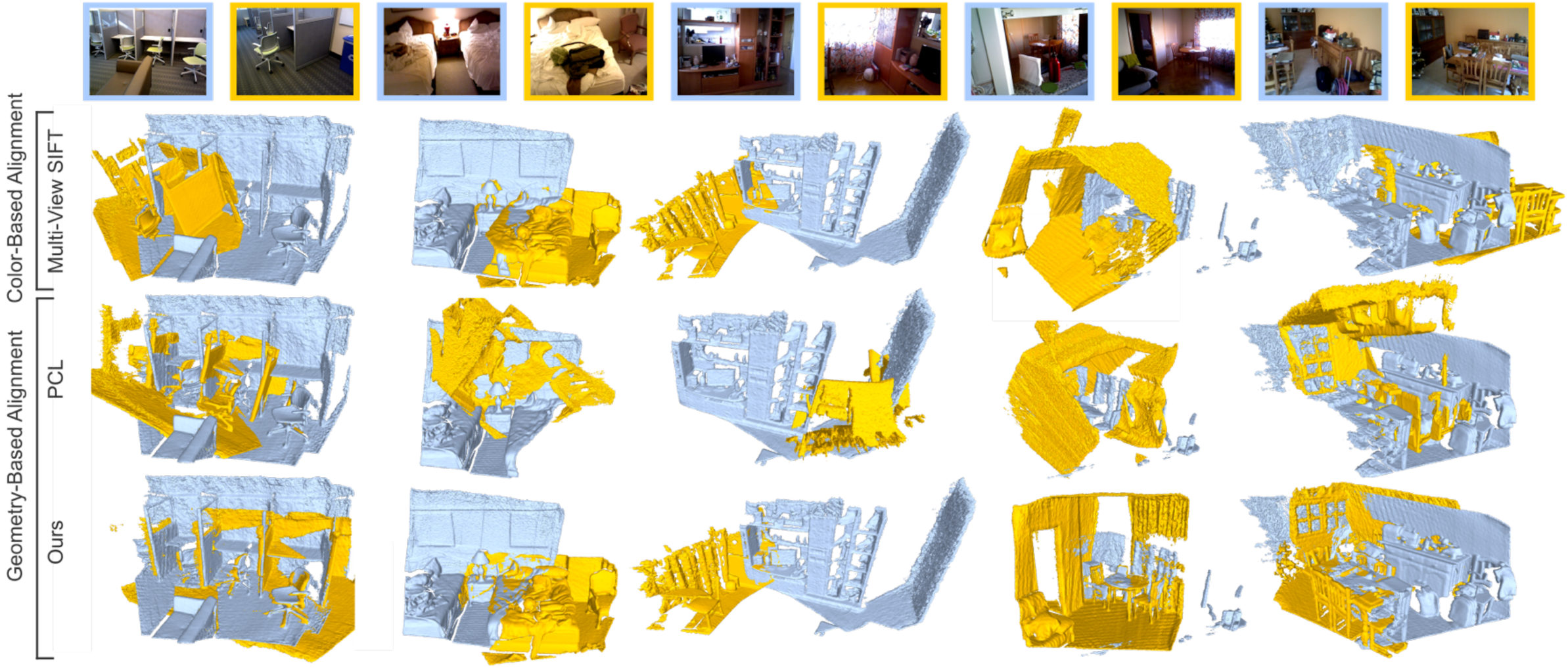
SLAM for Deep Learning



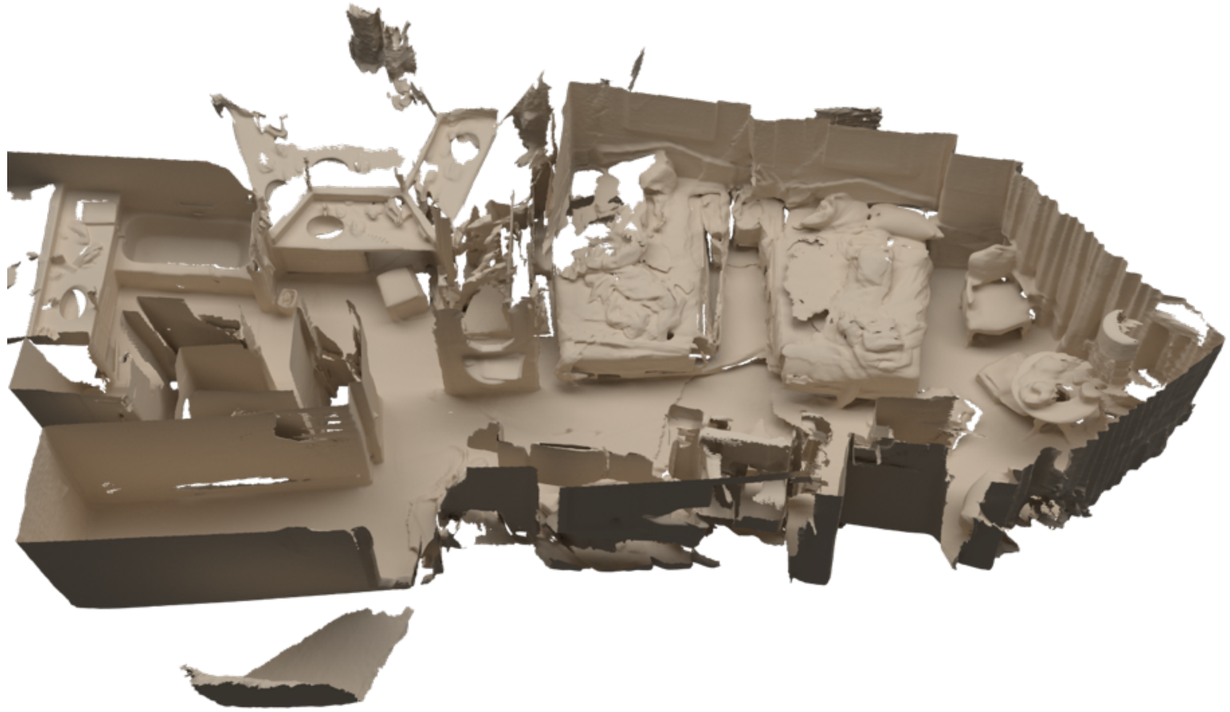
Microsoft 7 Scenes RGB-D dataset


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
t-SNE

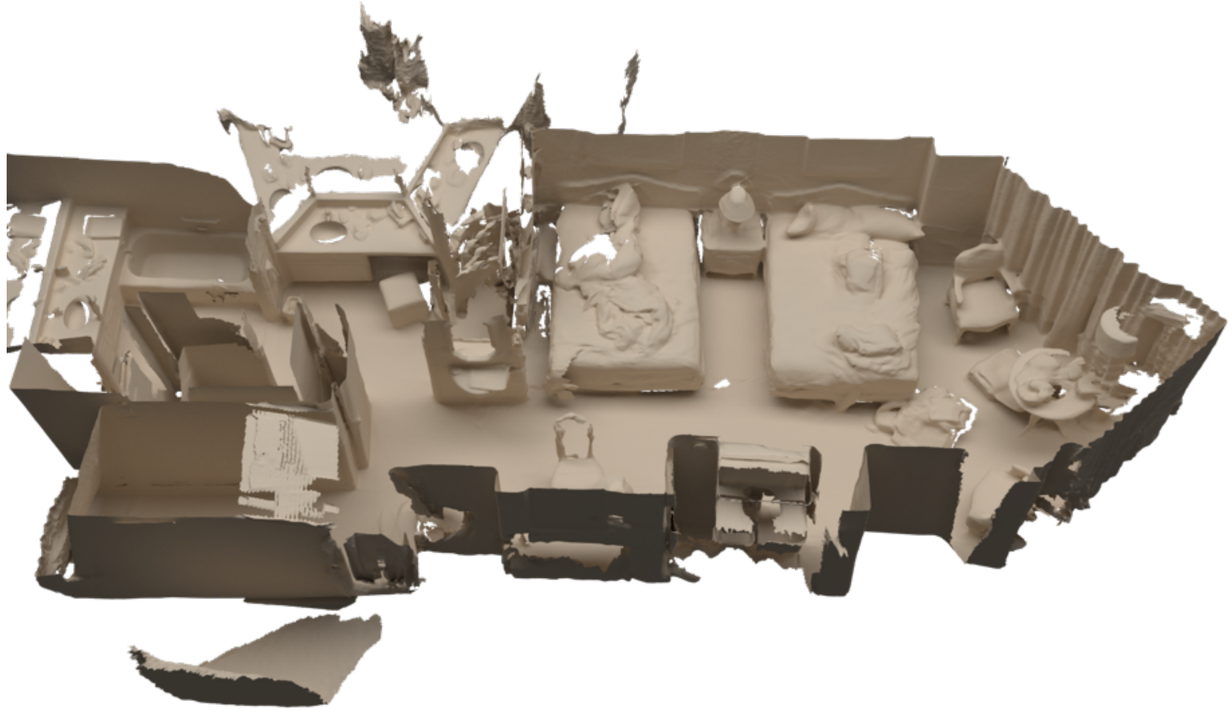
Compare with PCL



Robust Large-scale SLAM



SIFT-based SLAM

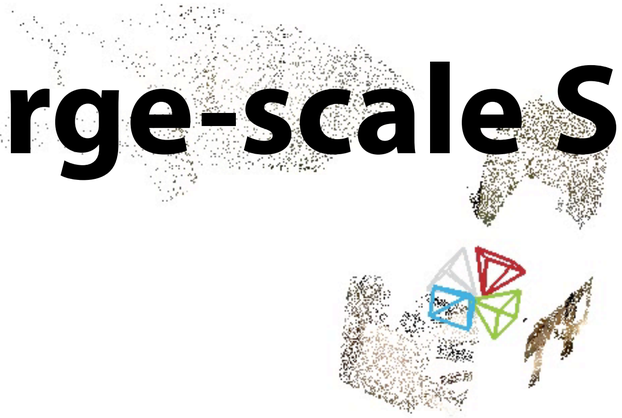


Our 3D Geometric Descriptors

3D Mapping Robots

Extremely Robust Large-scale SLAM

Extremely Robust Large-scale SLAM



3D Deep Learning for Robot Perception

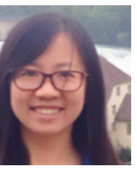
3D Amodal Object Detection

View Planning

Instance Recognition and Pose Recognition

Mapping and Reconstruction

Acknowledgements

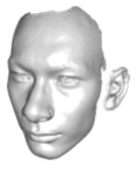

Shuran Song


Chenyi Chen


Daniel Suo

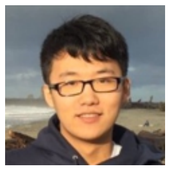

Ari Seff

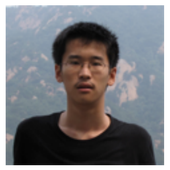

Andy Zeng


Fisher Yu


Yinda Zhang


Pingmei Xu


Mingru Bai


Zhirong Wu


Samuel Lichtenberg

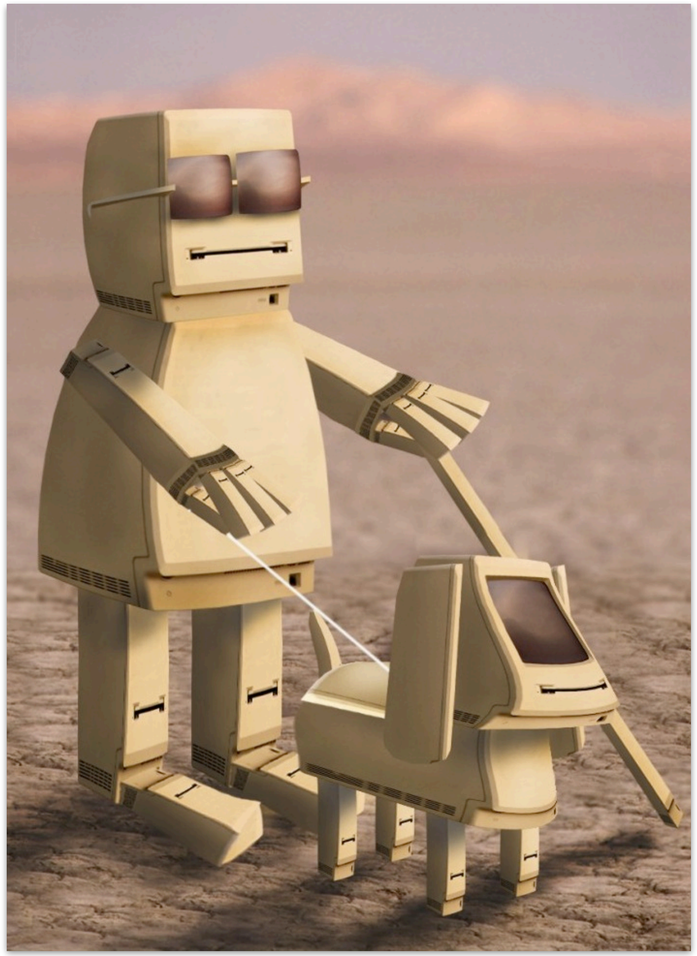

Aarav Chavda


Allan Jabri


Pallavi Koppol


Ed Walker Jr.

Alain Kornhauser, Thomas Funkhouser, Adam Finkelstein, Sanjeev Kulkarni, Aditya Khosla, Xiaoou Tang, Matthias Nießner, Matthew Fisher, Pushmeet Kohli, Shahram Izadi, …



3D Deep Learning for Robot Perception

Jianxiong Xiao

PRINCETON UNIVERSITY