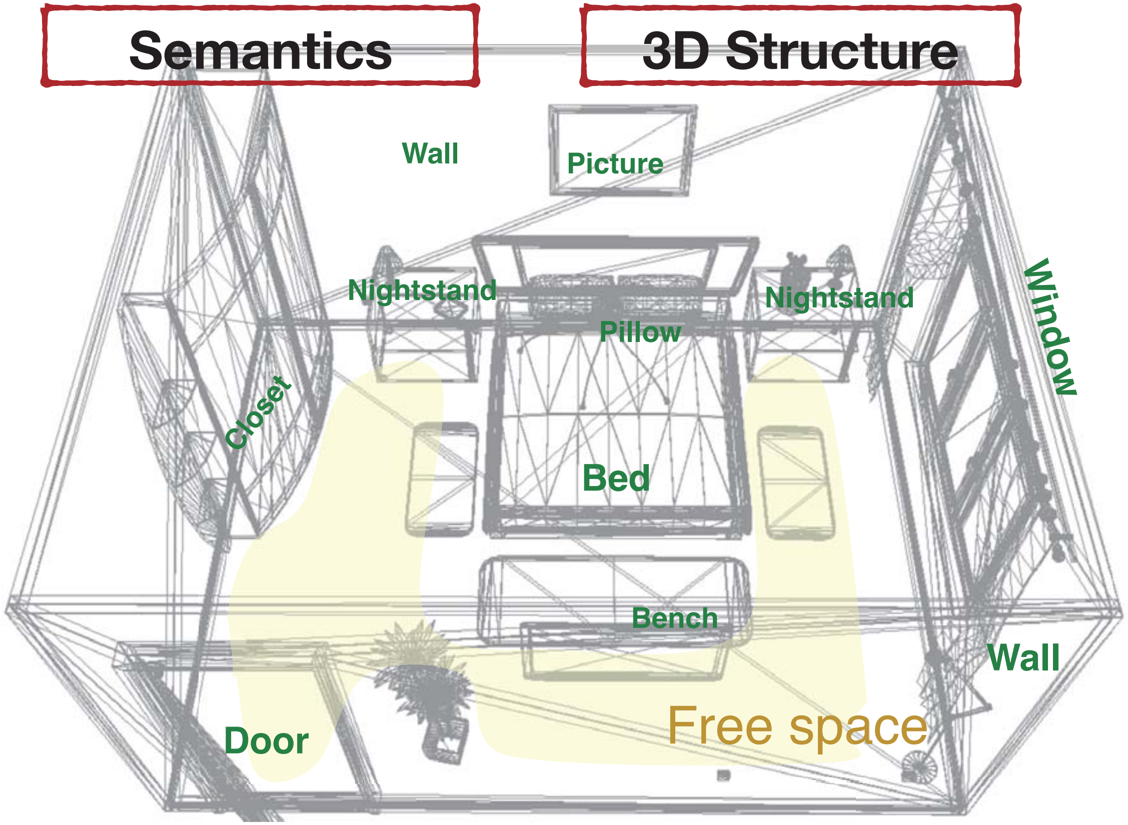
# Seeing the Unseen:
## Comprehensive 3D Scene Understanding

Shuran Song

Princeton —> Google —> Columbia

# Comprehensive 3D Scene Understanding

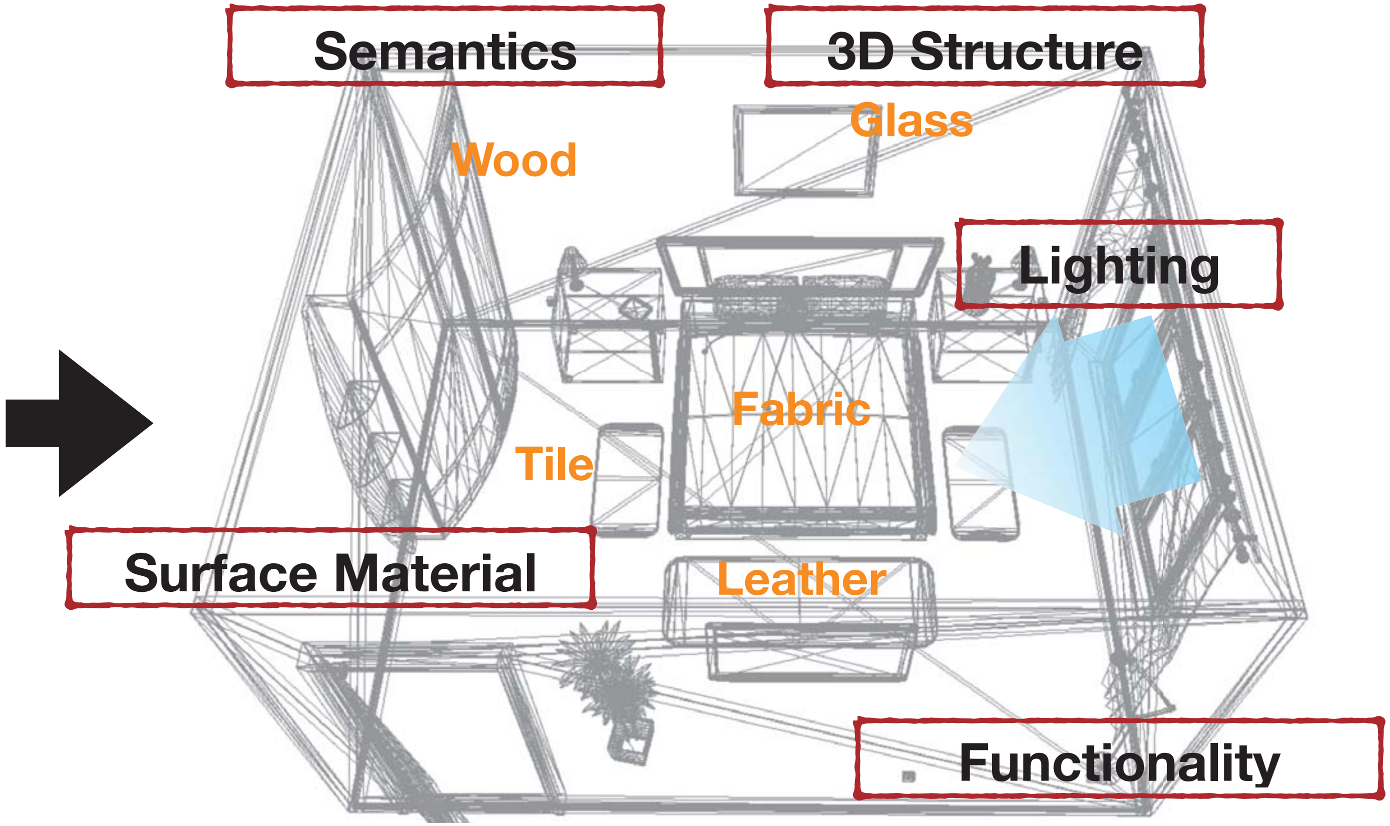
Partial Observation of the Environment




Complete Representation of the 3D Scene

# Comprehensive 3D Scene Understanding

Partial Observation of the Environment



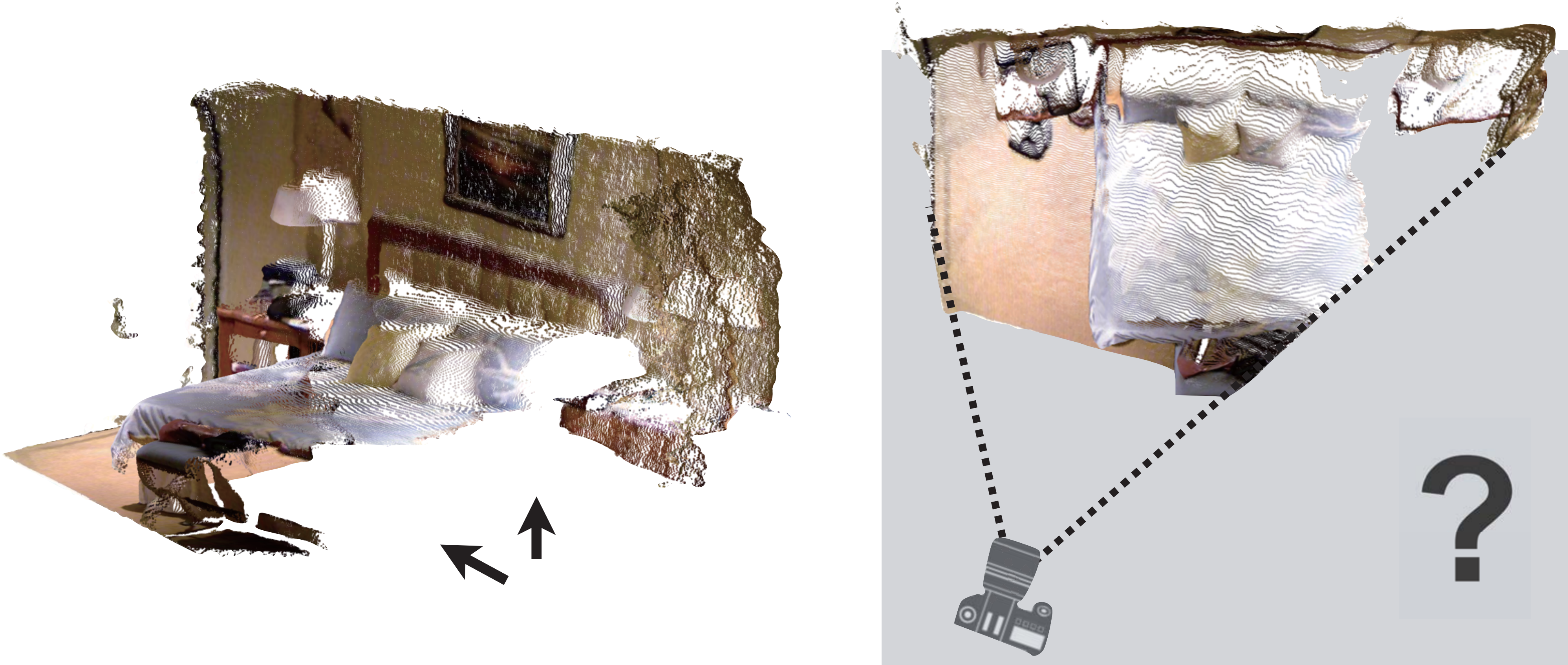

Complete Representation of the 3D Scene

# Challenge: Partial Observation

# Challenge: Partial Observation

**Sensors** **Partial Observation**

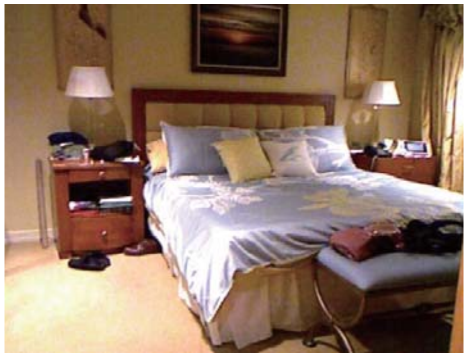
# Challenge: Partial Observation

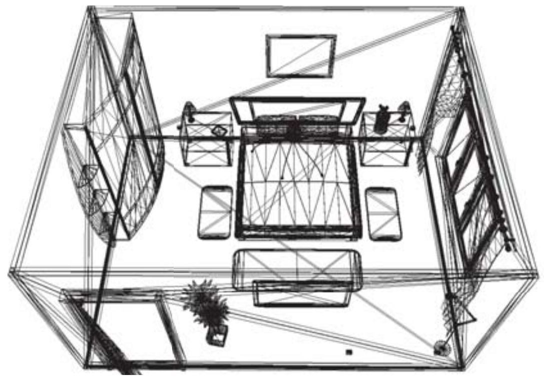



Occlusion

Limited Camera FOV

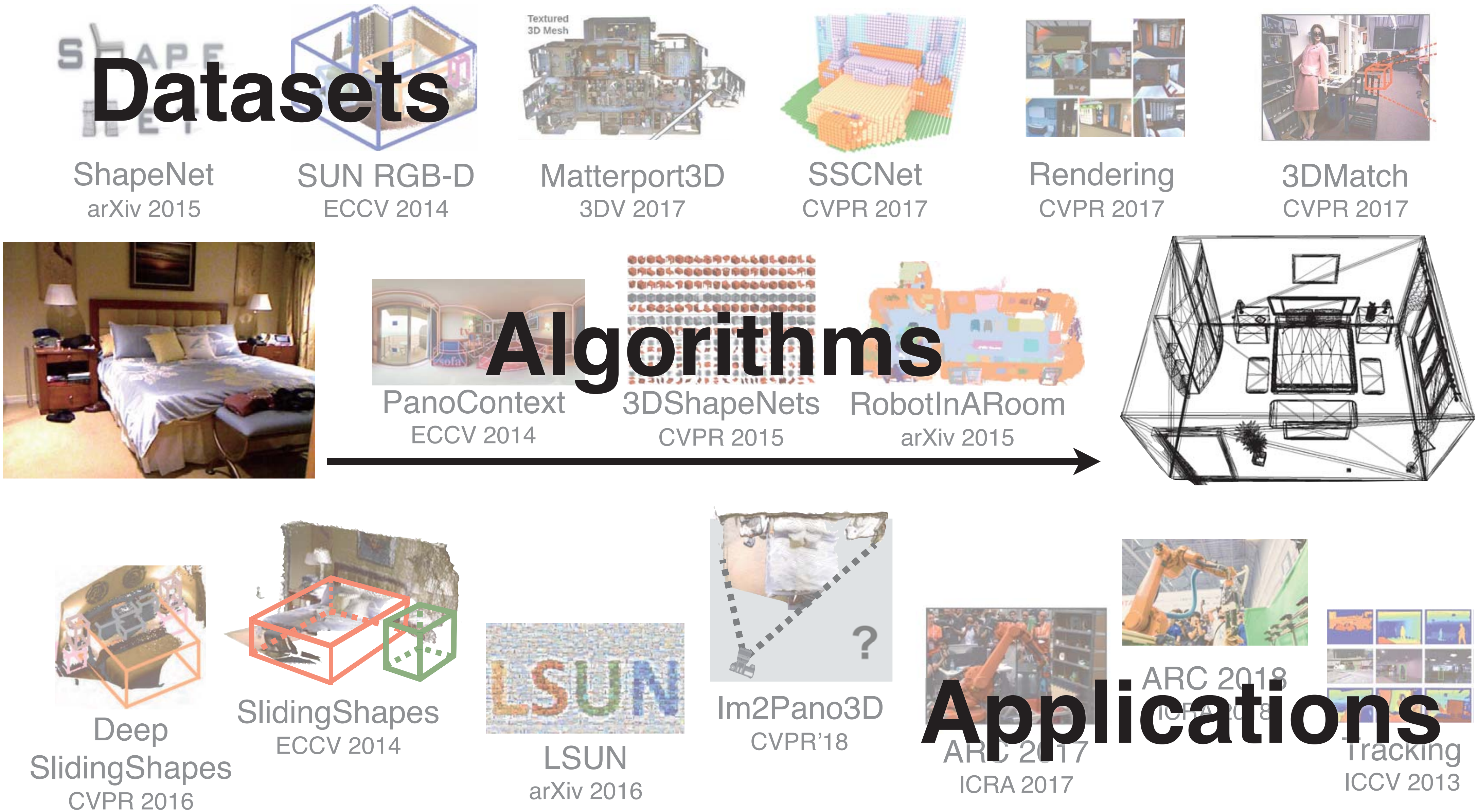
# Comprehensive 3D Scene Understanding



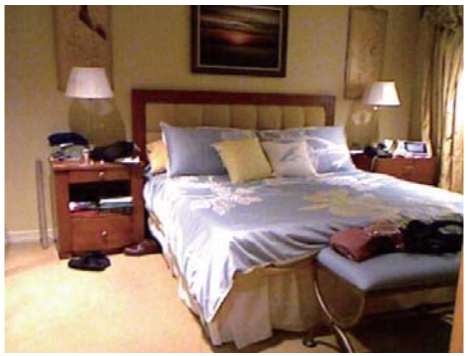


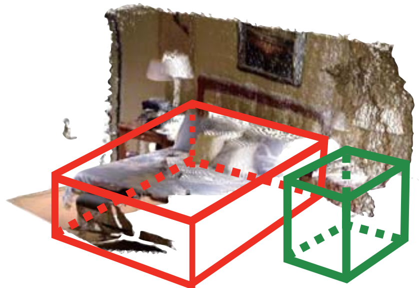
**Partial Observation**

**Complete 3D Scene**


Datasets
ShapeNet
arXiv 2015
SUN RGB-D
ECCV 2014
Textured
3D Mesh
Matterport3D
3DV 2017
SSCNet
CVPR 2017
Rendering
CVPR 2017
3DMatch
CVPR 2017
Algorithms
PanoContext
ECCV 2014
3DShapeNets
CVPR 2015
RobotInARoom
arXiv 2015
Deep
SlidingShapes
CVPR 2016
SlidingShapes
ECCV 2014
LSUN
arXiv 2016
Im2Pano3D
CVPR'18
ARC 2018
Applications
ICRA 2017
ICCV 2013

# Comprehensive 3D Scene Understanding





**Amodal 3D**

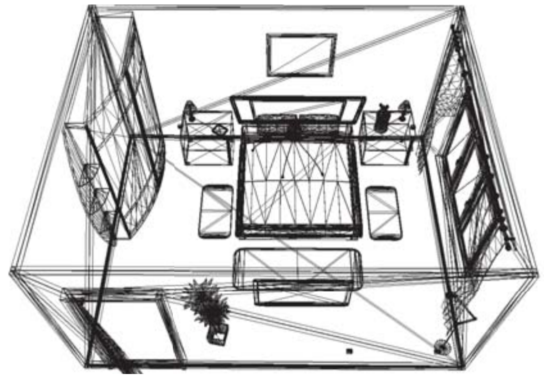
[Song and Xiao ECCV'14,CVPR'16]


**Beyond FoV**

[Song et al. CVPR'18]

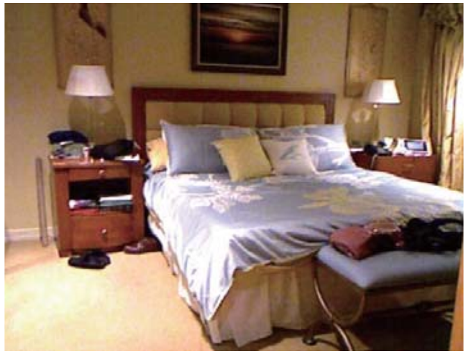
**Higher Fidelity**

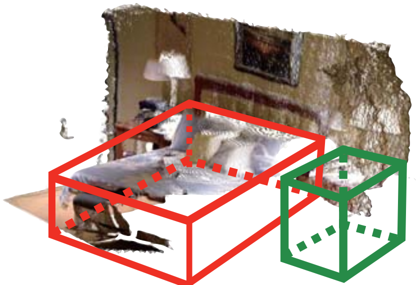
[Song et al. CVPR'17]



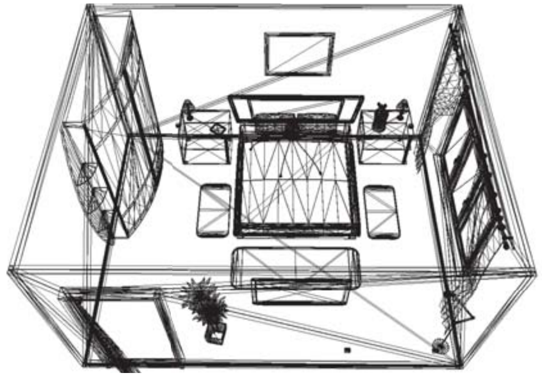
- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. Properties …

# Comprehensive 3D Scene Understanding





**Amodal 3D**
[Song and Xiao
ECCV'14,CVPR'16]



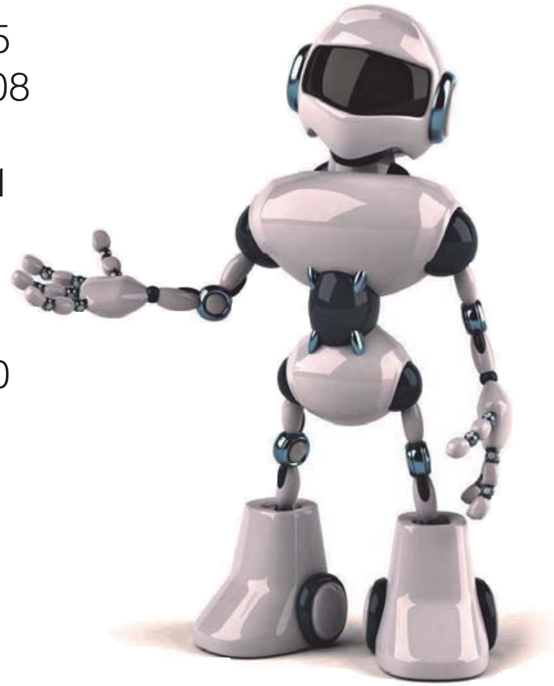
# Object Detection

2D Visible (Modal) Surface


Traditional Object Detection Output

Aubry et al. CVPR'14
Dalal and Triggs CVPR'05
Felzenszwalb et al. CVPR'08
Bo et al. CVPR'2011
Malisiewicz et al. ICCV'11
Girshick et al. CVPR'14
Ren et al. NIPS'15
Girshick, ICCV'15
Everingham et al. IJCV'10
He et al. ICCV'17
Liu et al. ECCV'16
Erhan et al. CVPR'14
He et al. ECCV'14
Szegedy NIPS'13
…

Where to sit?

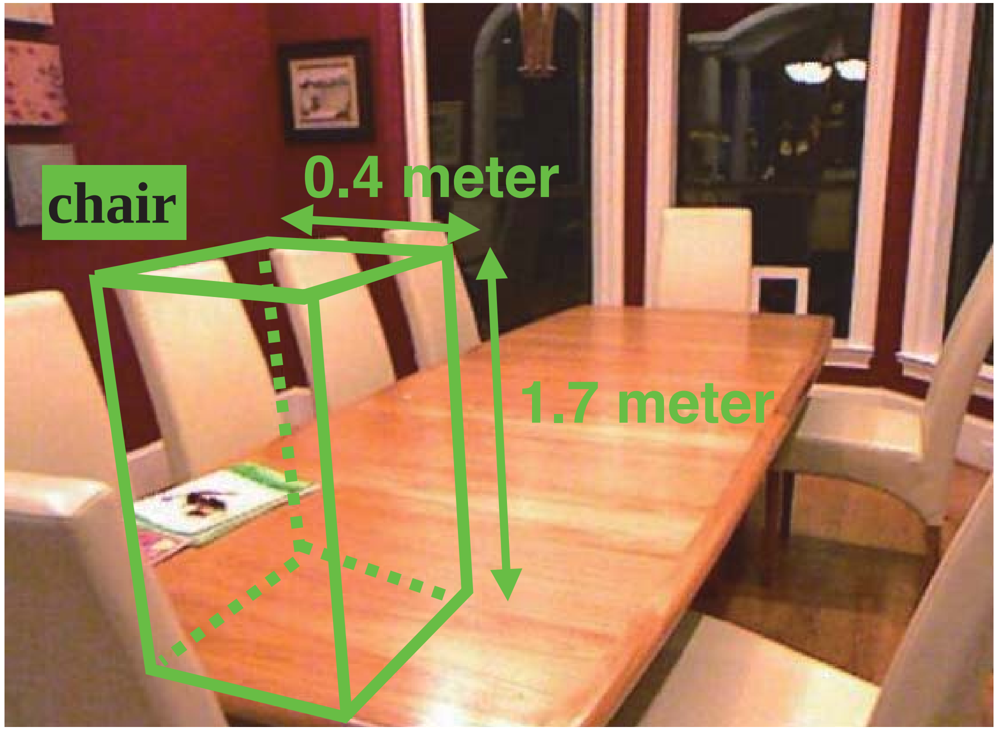


# Object Detection

2D Visible (Modal) Surface


Traditional Object Detection Output

3D Complete (Amodal) Shape

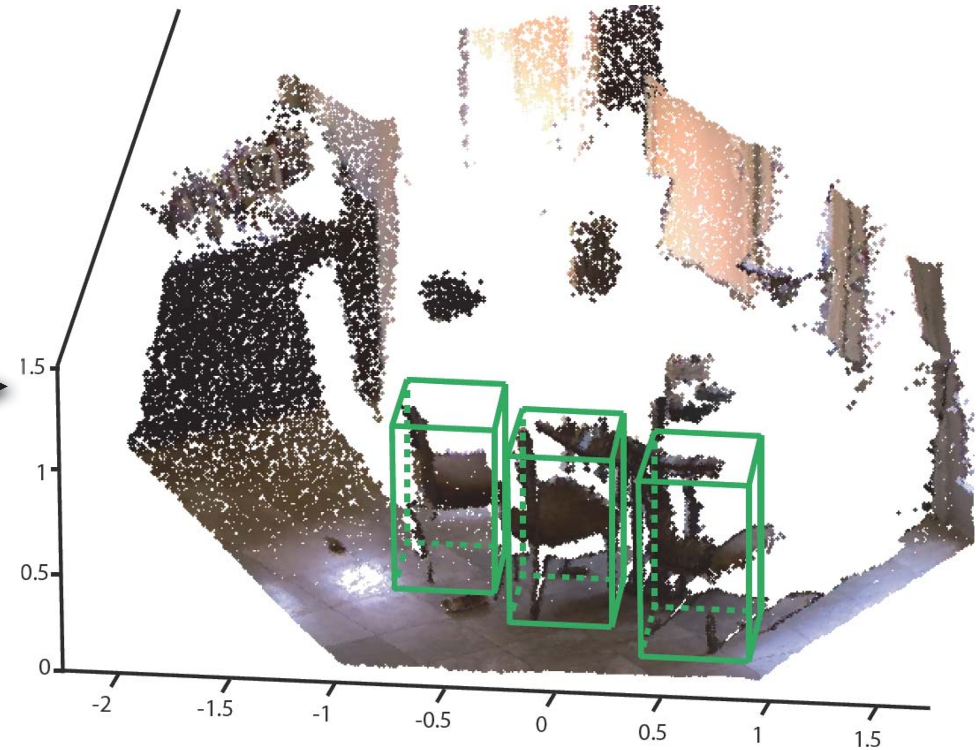



This work

S. Song and J. Xiao, Sliding Shapes for 3D Object Detection in Depth Images, ECCV 2014
S. Song and J. Xiao, Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images, CVPR 2016

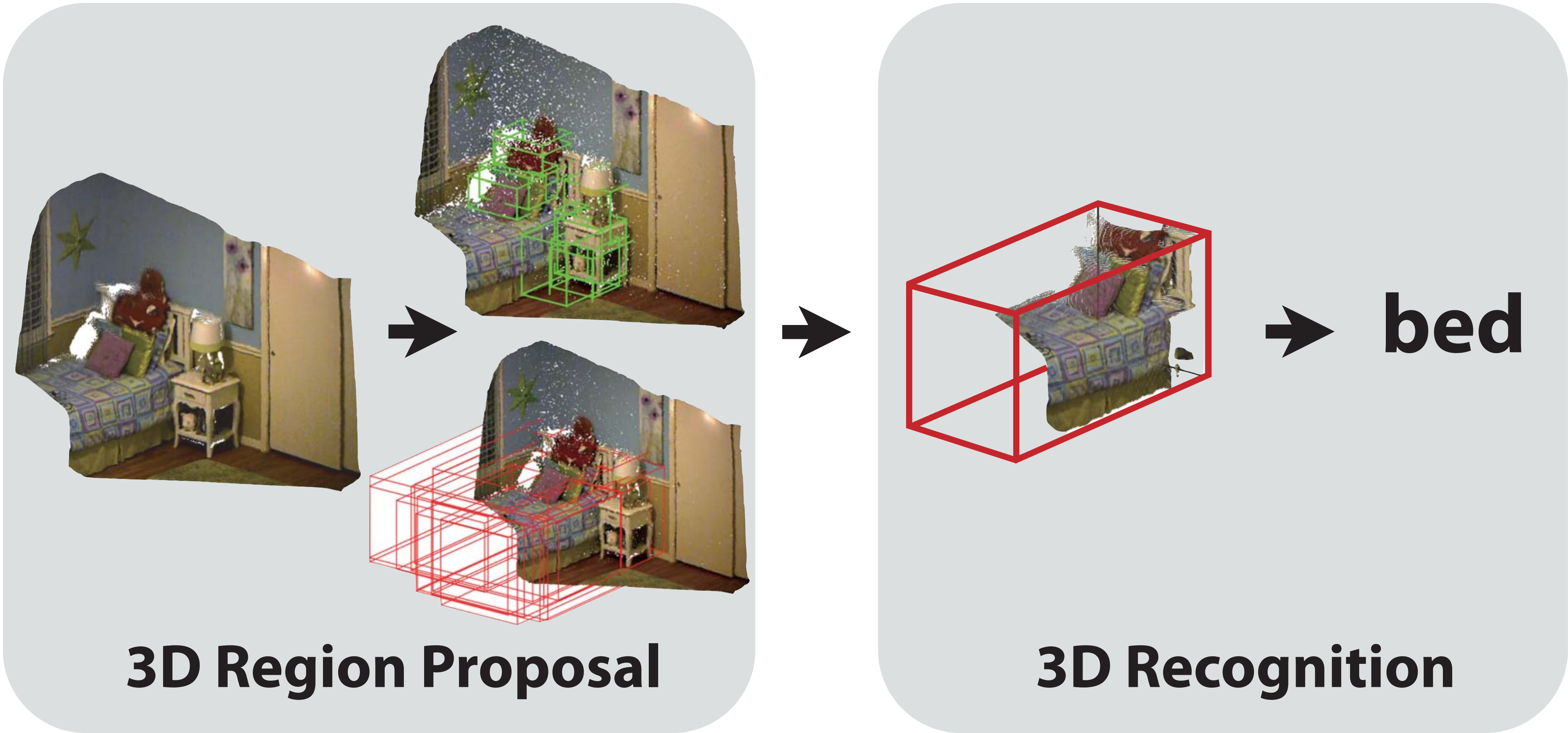
# Deep Sliding Shapes


Input: Kinect Depth Image




Output: 3D Bounding Box

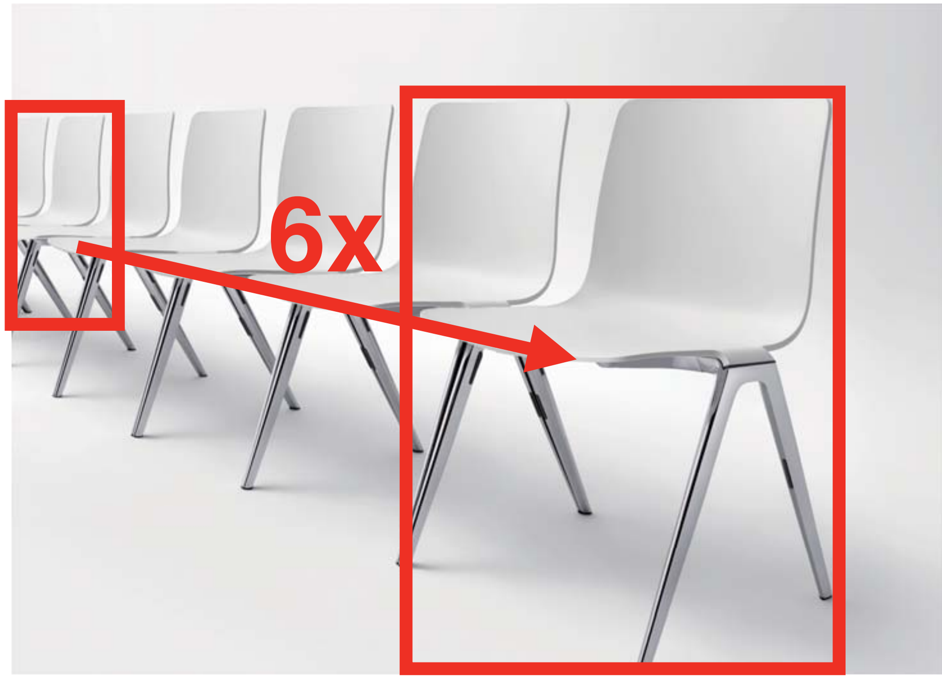
# 3D Deep learning



# Representation: 3D vs. 2D

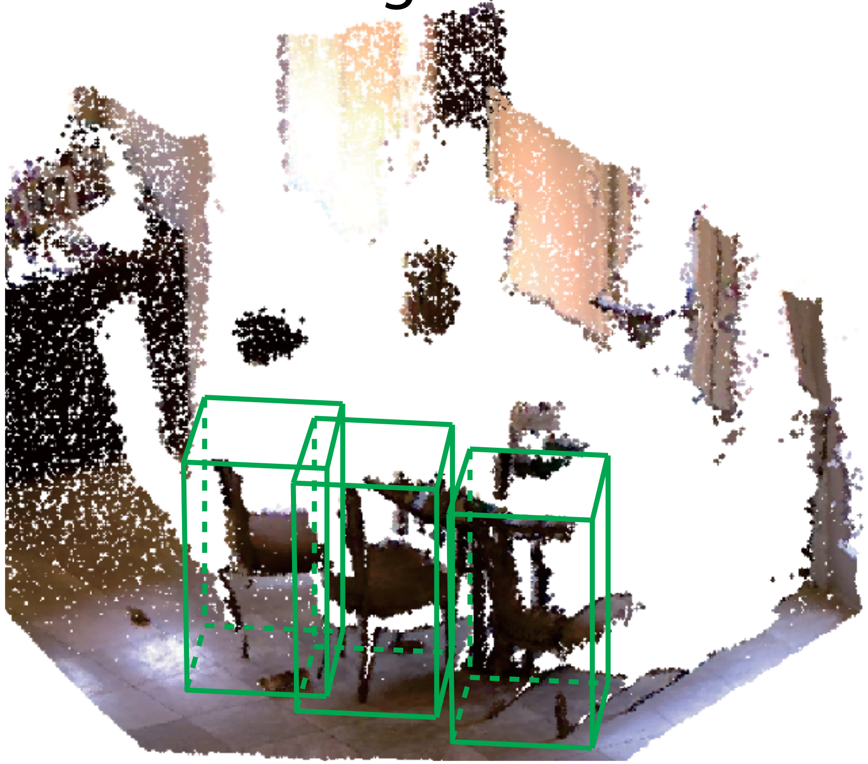
# Advantage: Exploiting Physical Size

2D Sliding Window




Multi-scale searching

3D Sliding Window

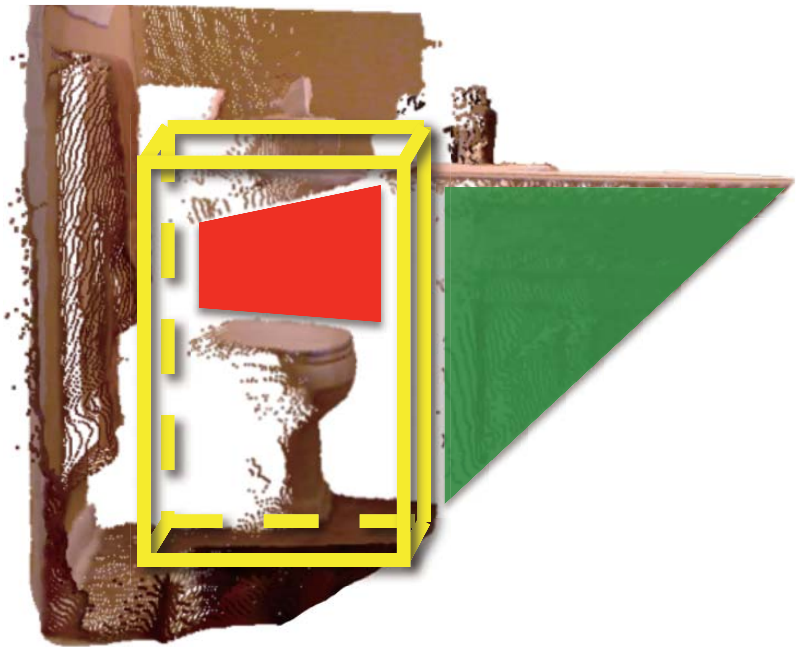


Physical size

# Advantage: Handling Occlusion

2D Sliding window

3D Sliding window



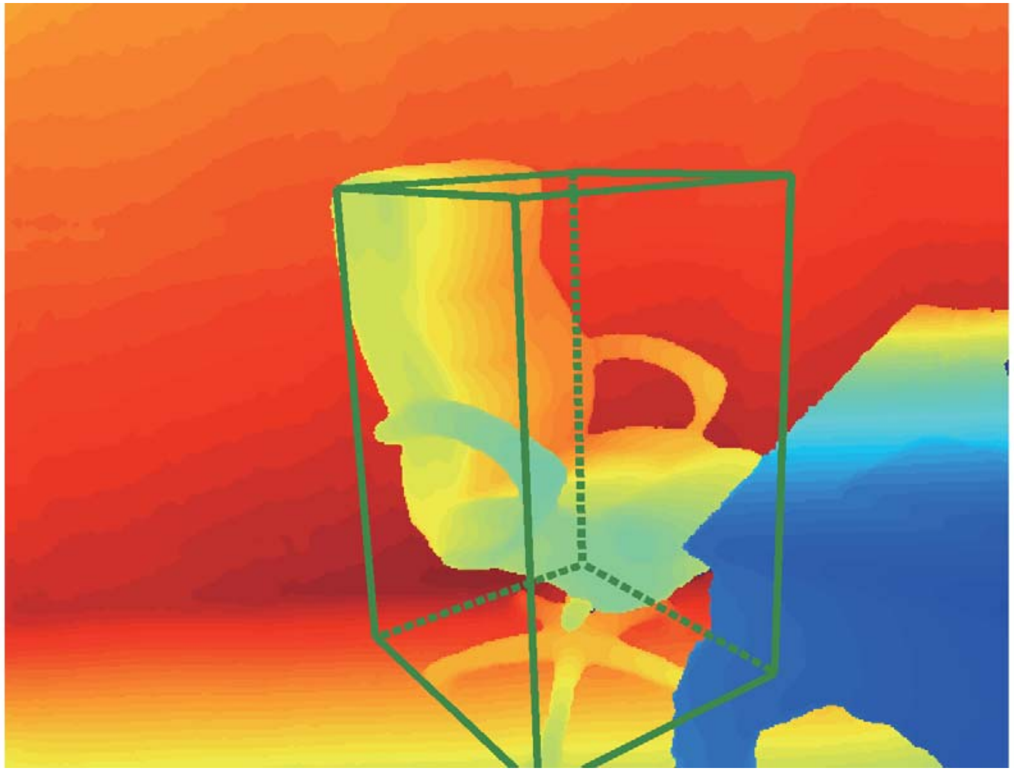
Using depth, we can know which part is **occluded**.

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

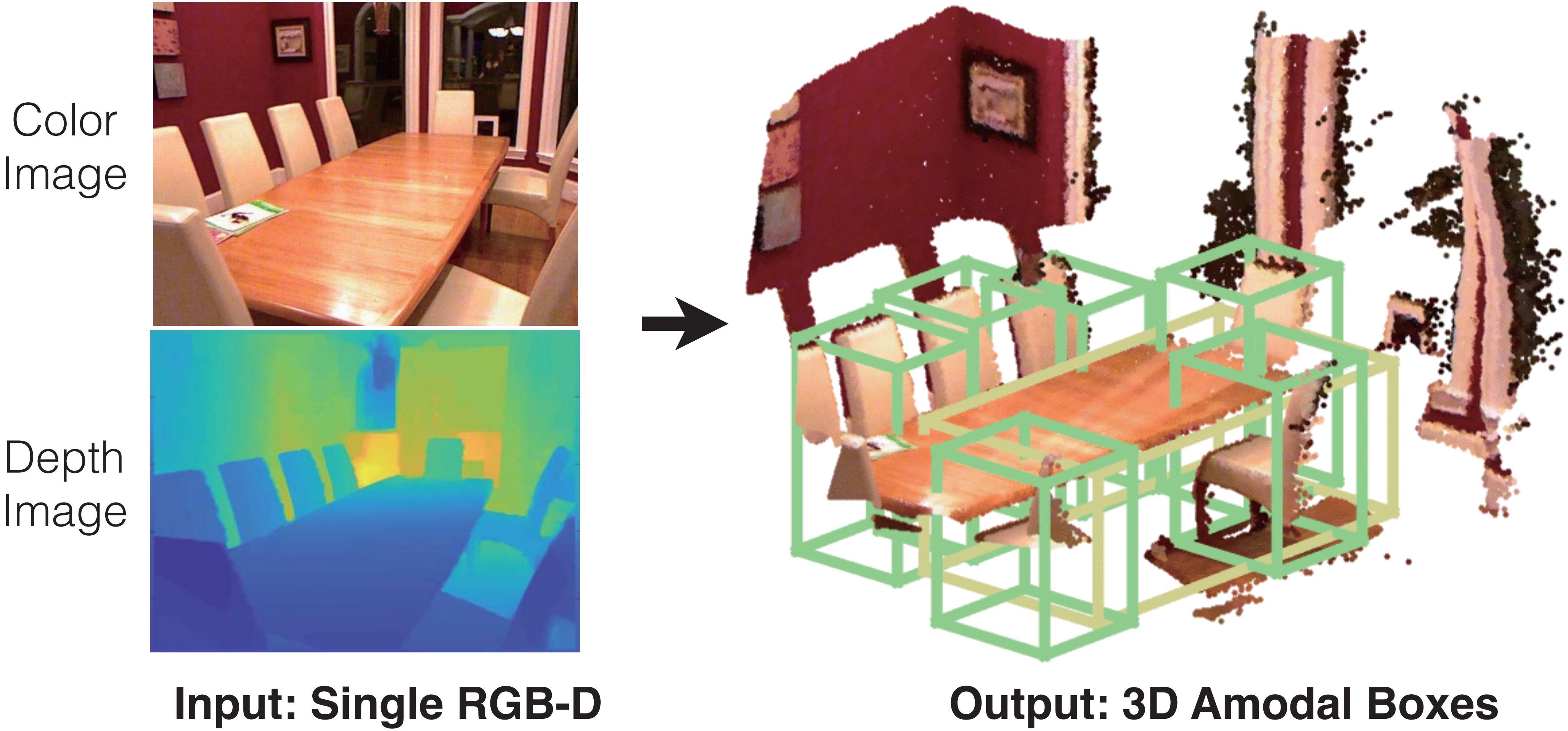
# Advantage: Insensitivity to Lighting

Color based detector: miss

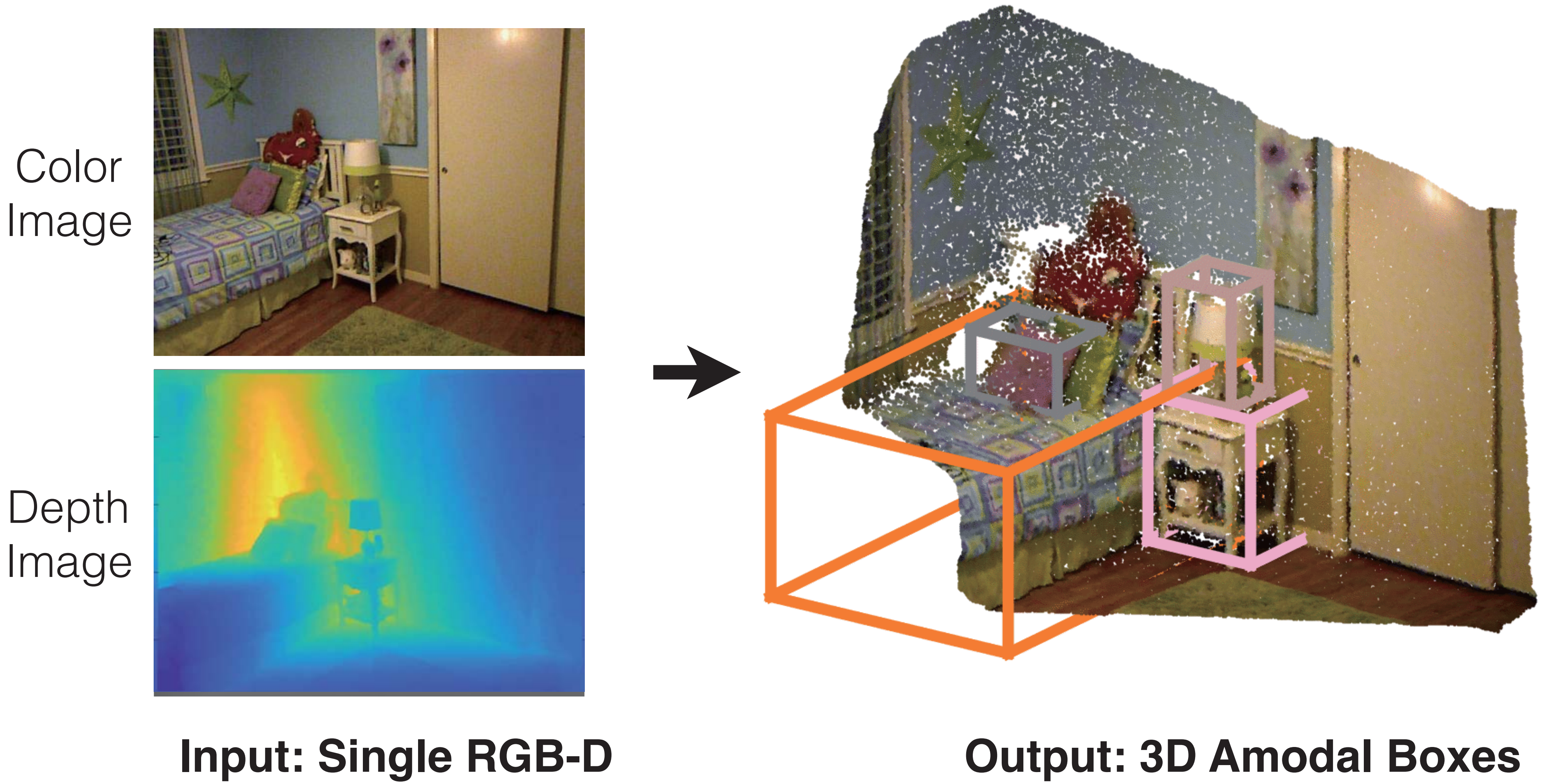
Sliding Shapes



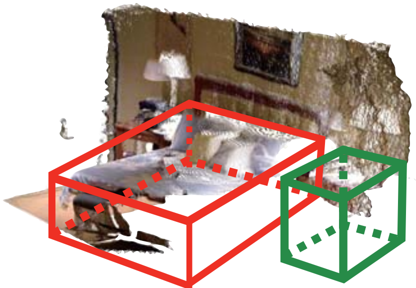
# Results: Deep Sliding Shapes



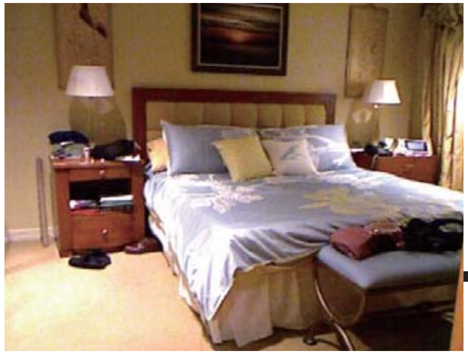
# Results: Deep Sliding Shapes

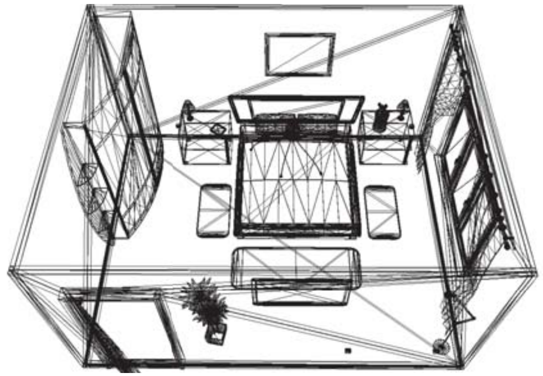


# Data-Driven 3D Scene Understanding



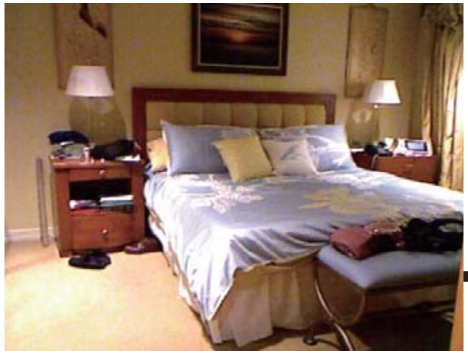
**Amodal 3D**
[Song and Xiao
ECCV'14,CVPR'16]

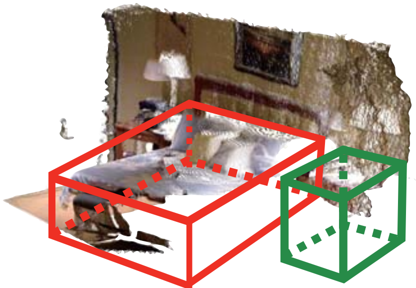




- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. properties …

# Data-Driven 3D Scene Understanding

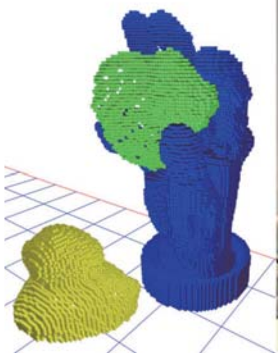


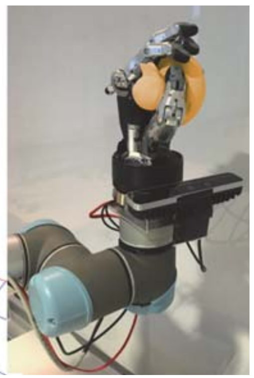


**Amodal 3D**

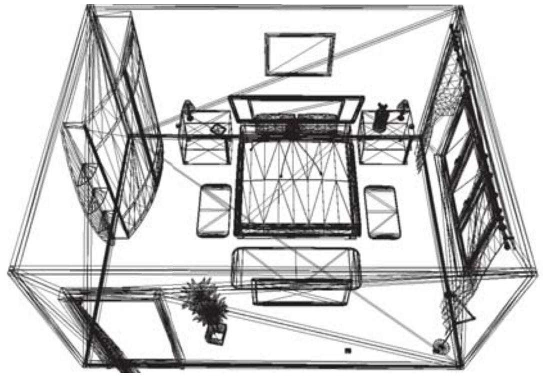
[Song and Xiao ECCV'14,CVPR'16]

• Only Boxes, No Detailed Geometry



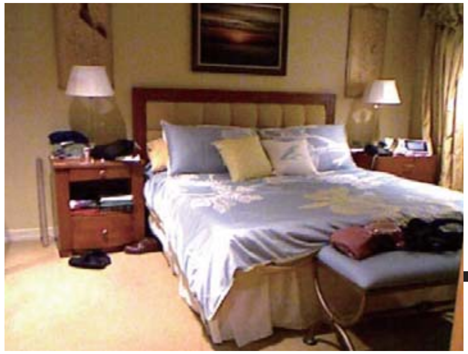


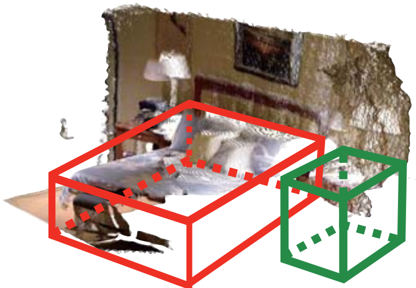
Not sufficient



- ✓ **Semantics Category**
- ✓ **3D Location, Size**
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. properties …

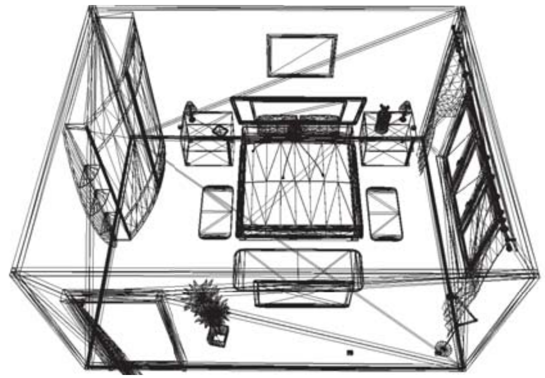
# Data-Driven 3D Scene Understanding





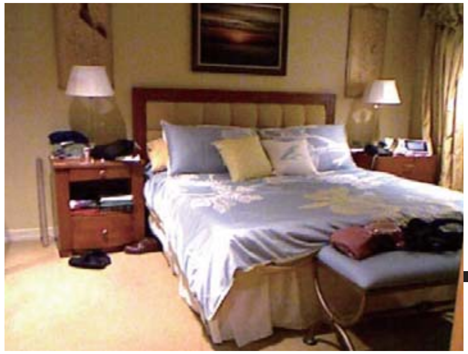
**Amodal 3D**
[Song and Xiao
ECCV'14,CVPR'16]

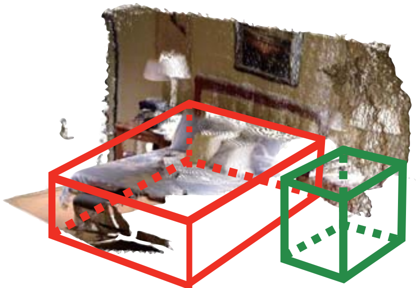
- Only Boxes, No Detailed Geometry
- Single Object, No Contextual Information



✓ **Semantics Category**
✓ **3D Location, Size**
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. properties …

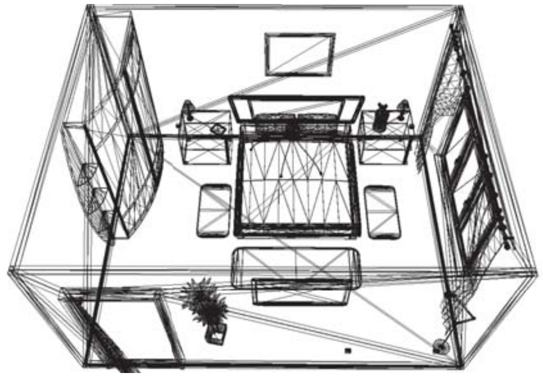
# Data-Driven 3D Scene Understanding





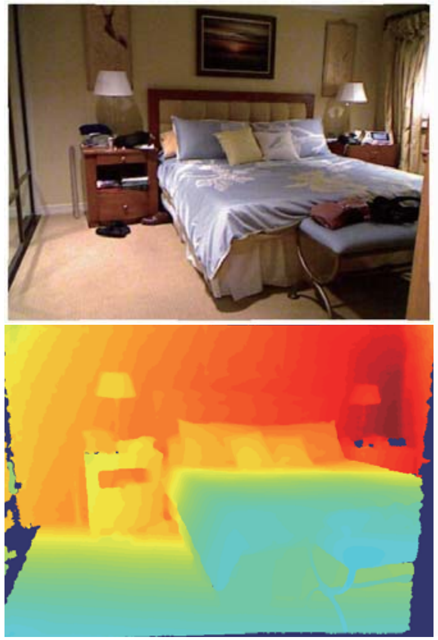
**Amodal 3D**
[Song and Xiao ECCV'14,CVPR'16]

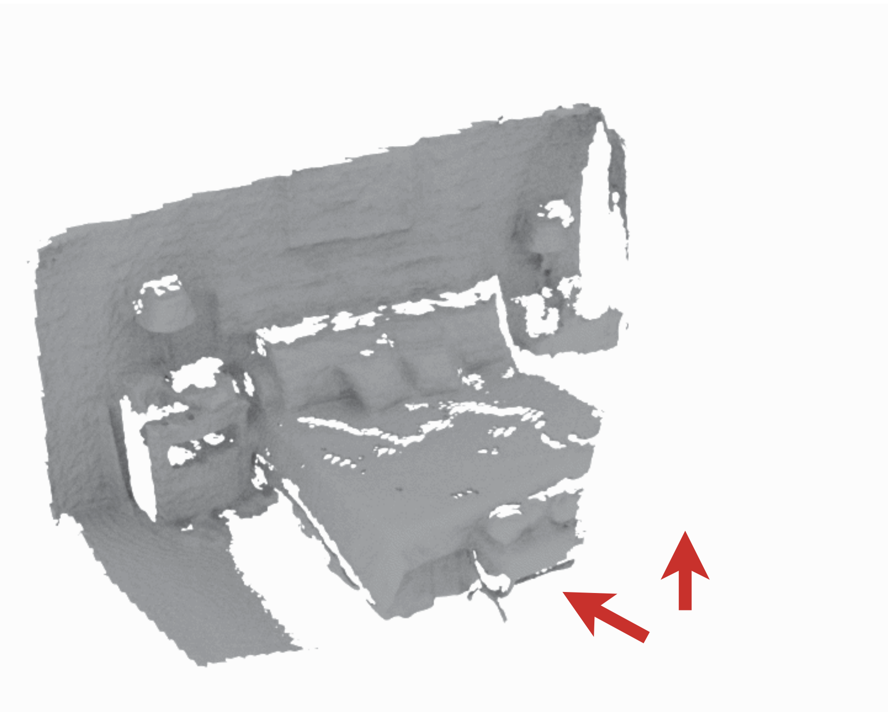
**Higher Fidelity**
[Song et al. CVPR'17]

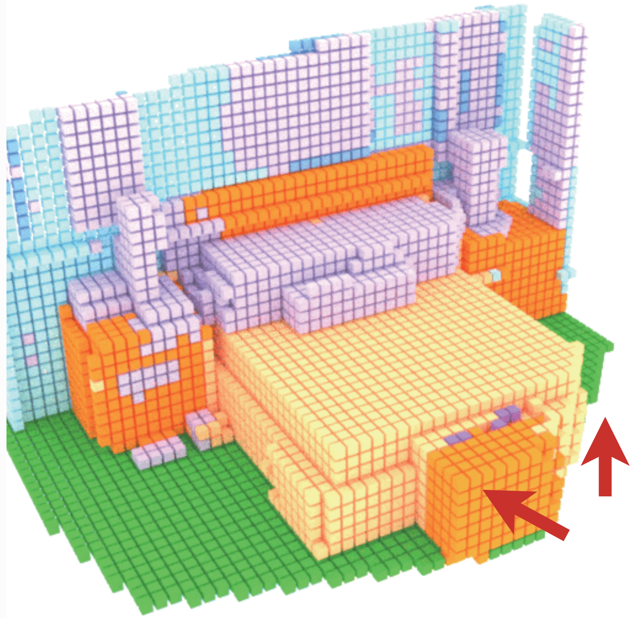


- ✓Semantics Category
- ✓3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. properties …

# Semantic Scene Completion





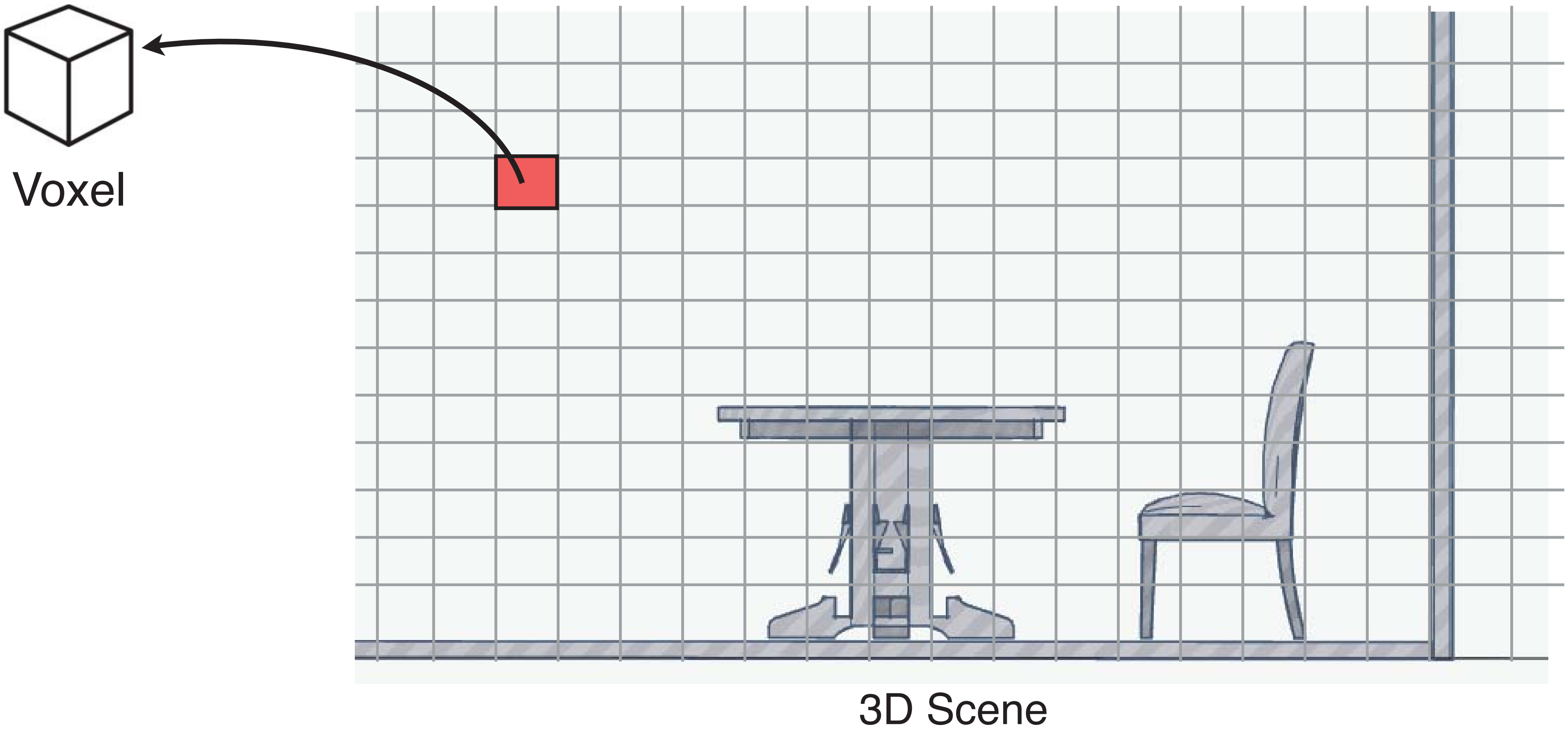


Input:
Single Depth Map

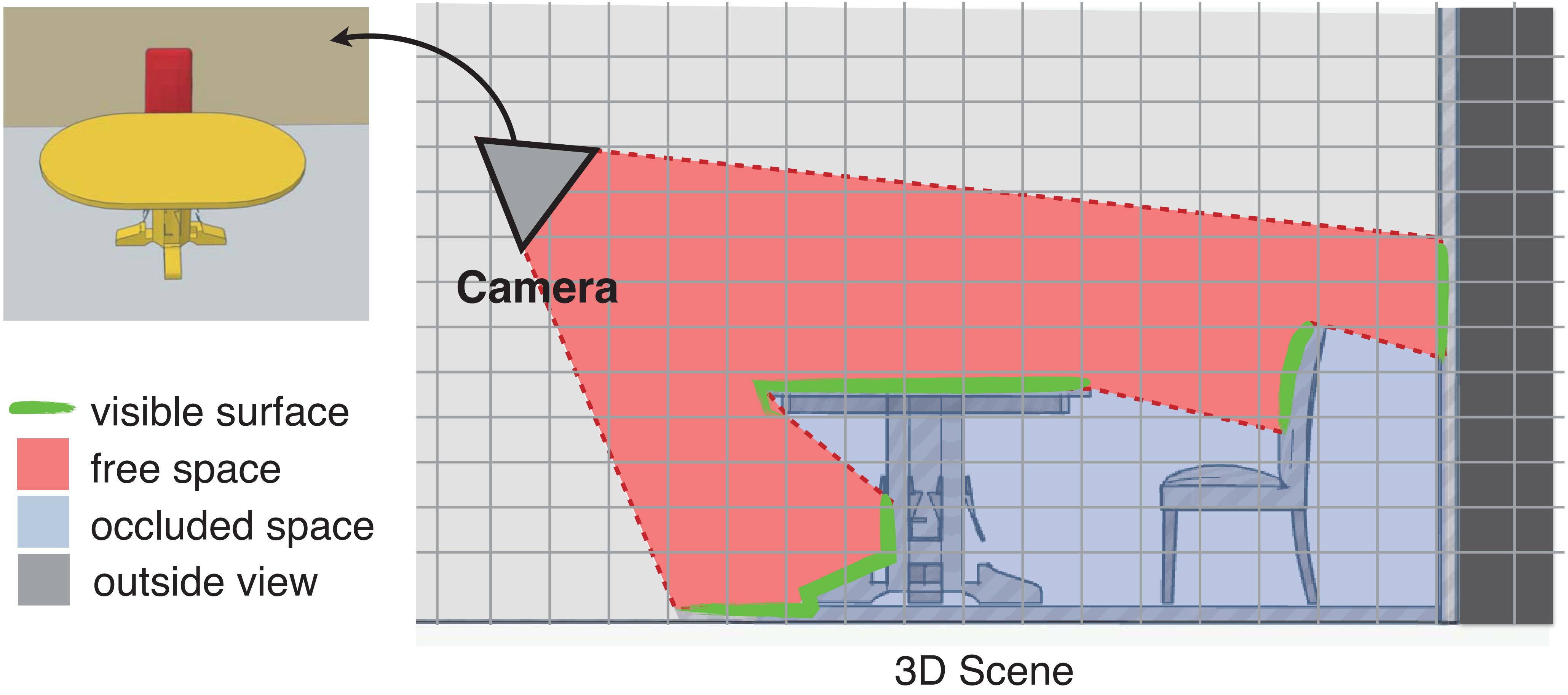
Output:
Volumetric Occupancy + Semantic

Song et al. Semantic Scene Completion from a Single Depth Image. CVPR'17

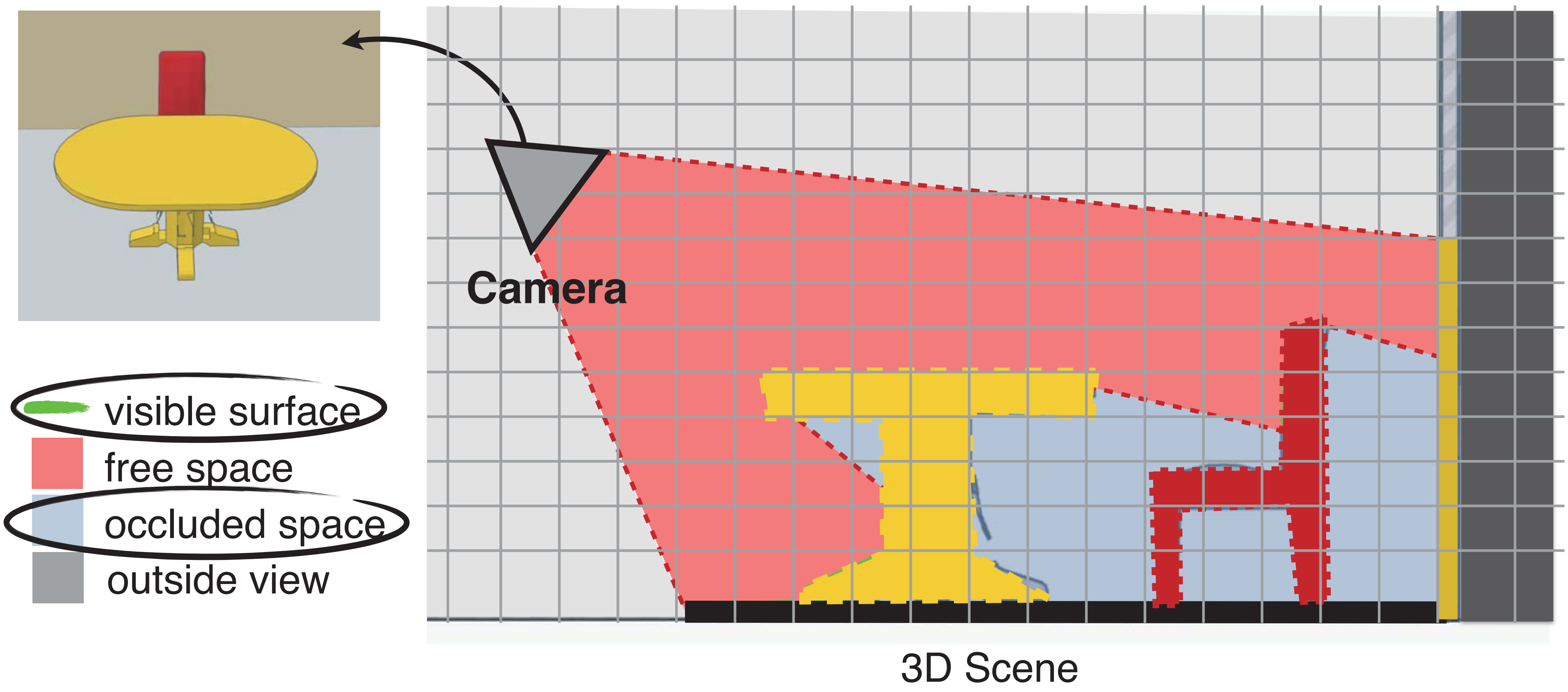
# Problem Definition



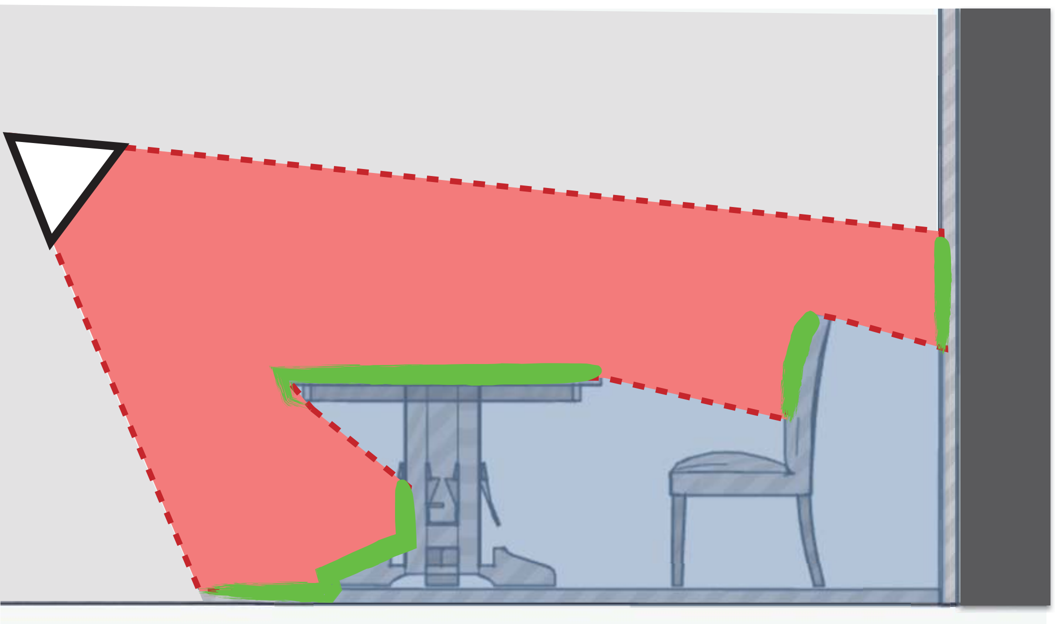
# Problem Definition



# Problem Definition



# Problem Definition



3D Scene

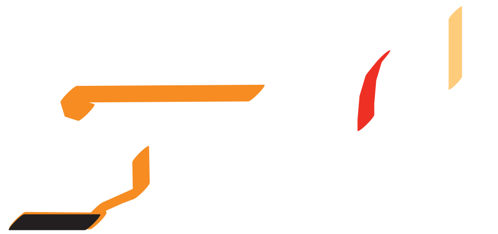
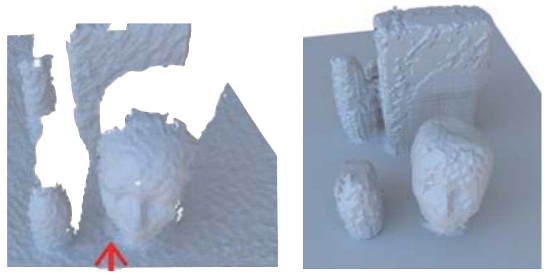


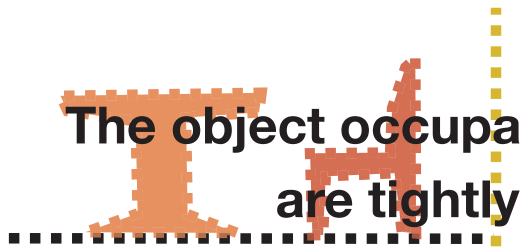
Image Segmentation

[Long *et al.* CVPR’15]

Shape Completion



[Firman *et al.* CVPR’17]

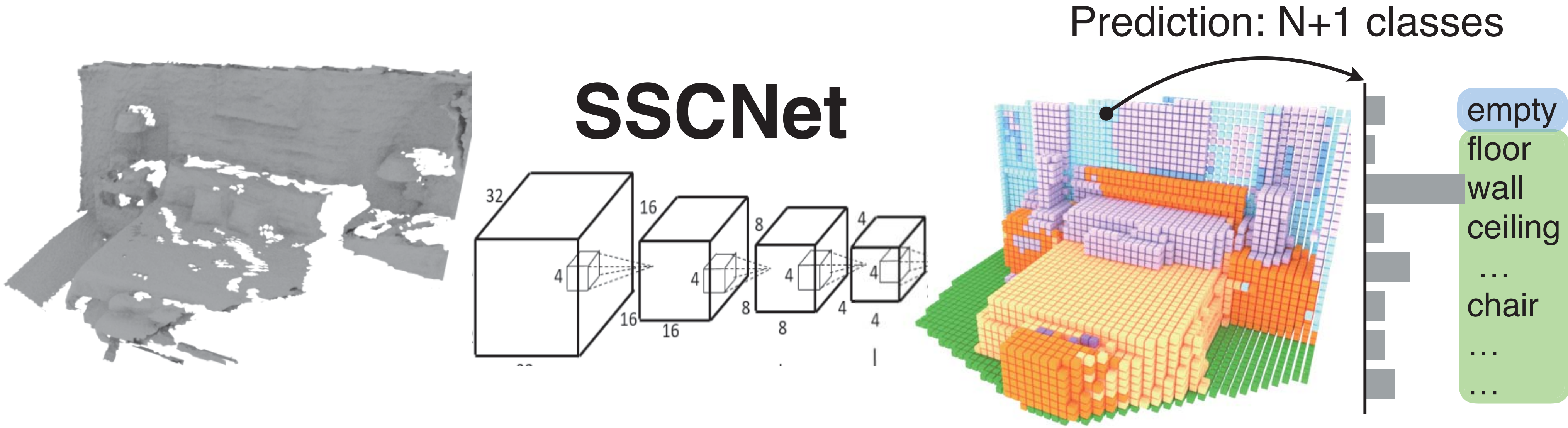


**The object occupancy and their identity are tightly intertwined !**

This paper

Semantic Scene Completion

# Semantic Scene Completion Network

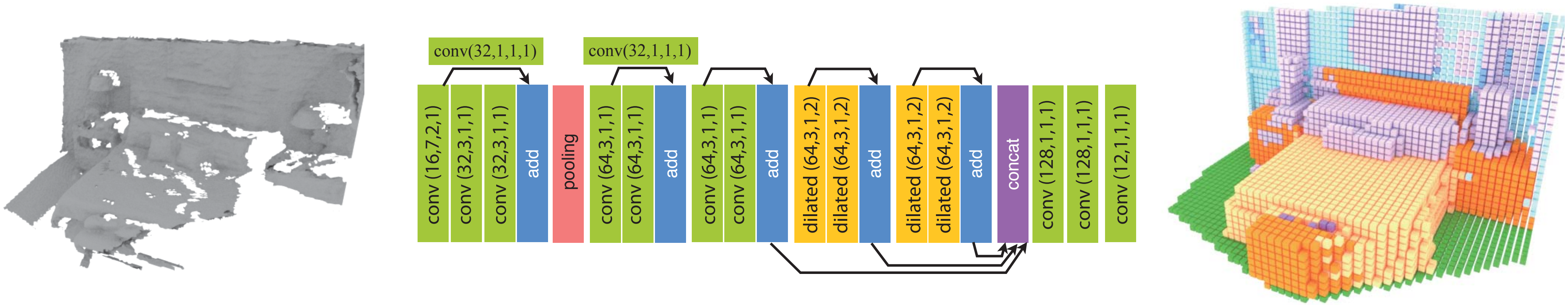



Input:
Single view depth map

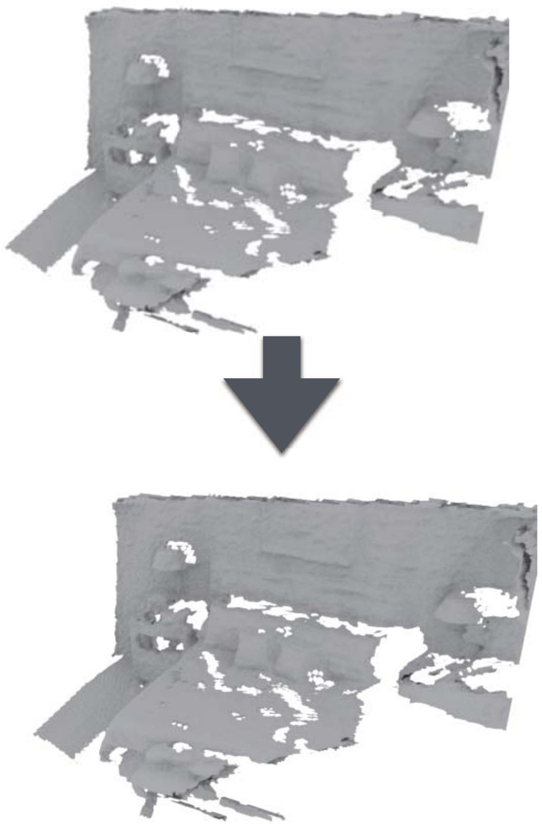
Output:
volumetric occupancy + semantic

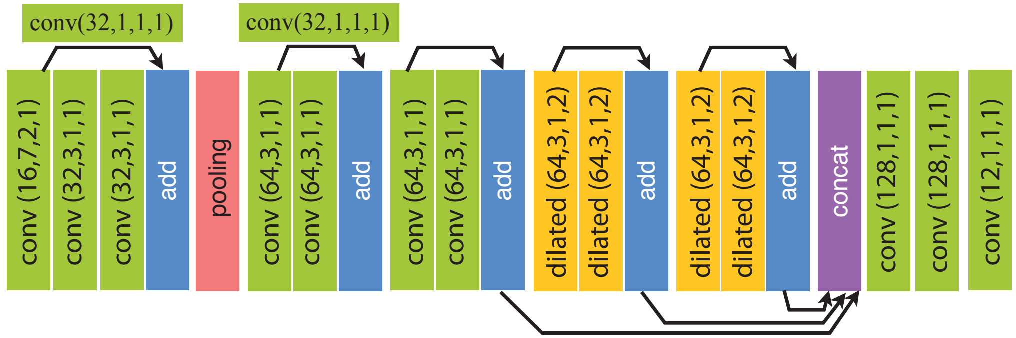
Simultaneously predict voxel occupancy and semantics classes by a single forward pass.

# Semantic Scene Completion Network

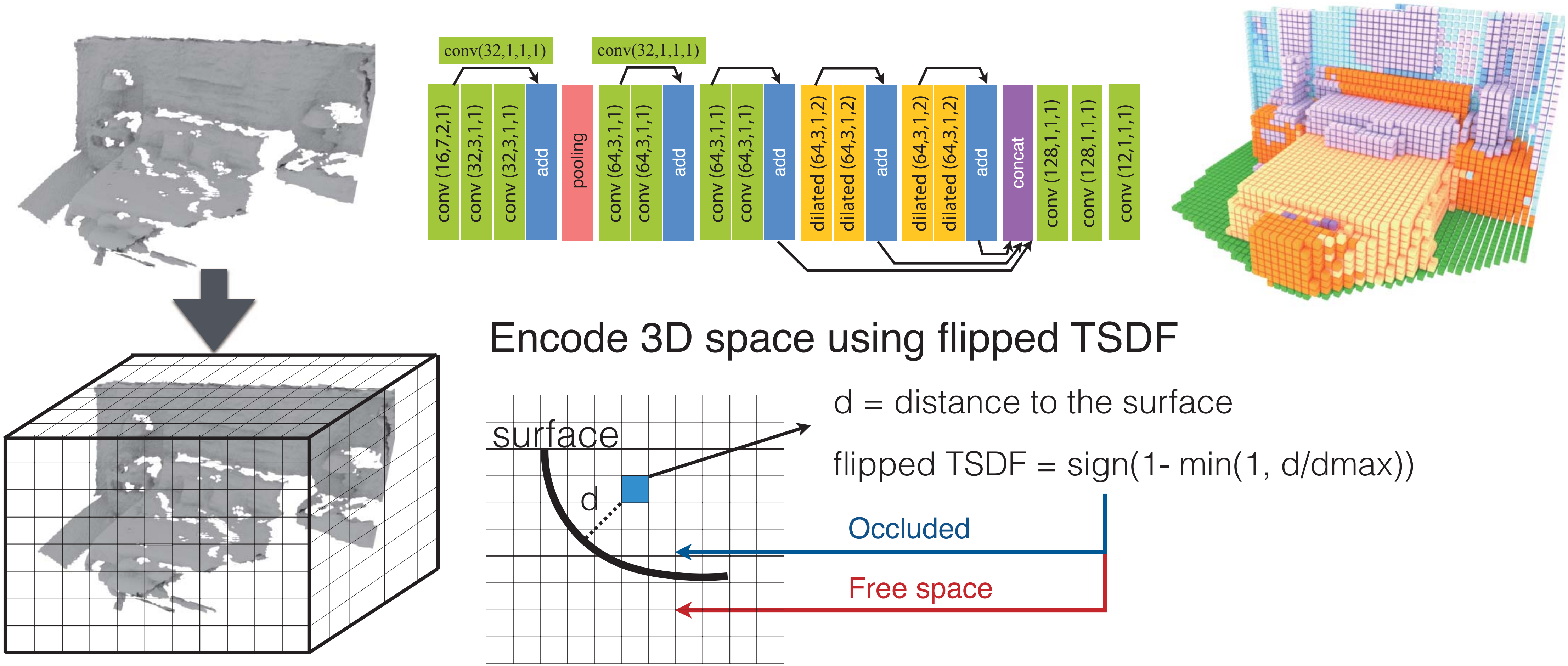


# Semantic Scene Completion Network

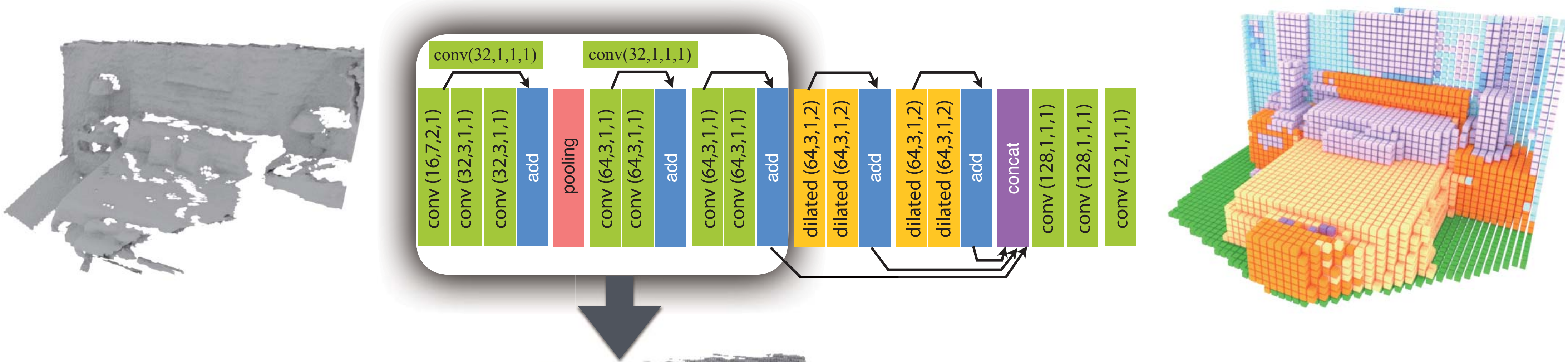


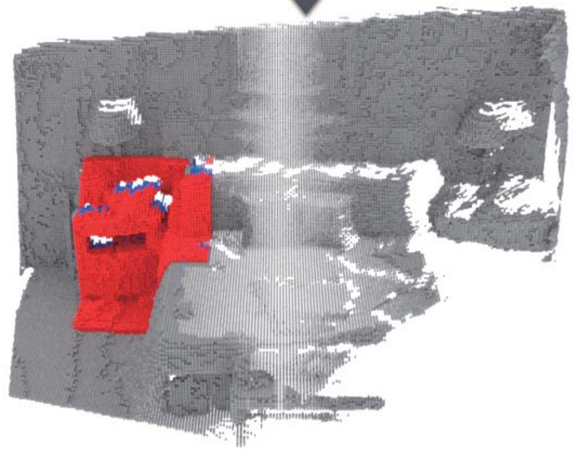



# Semantic Scene Completion Network



# Semantic Scene Completion Network

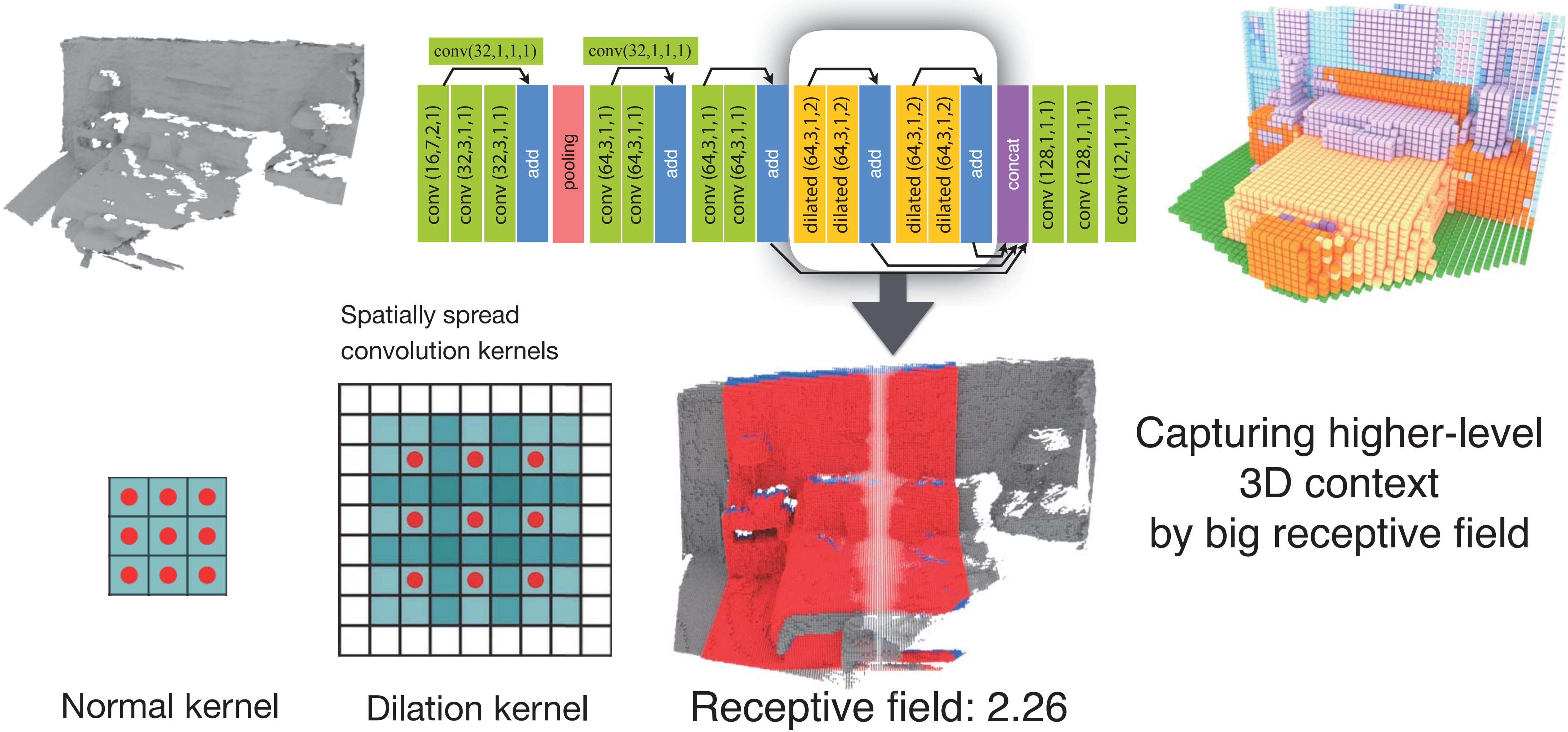


Local geometry

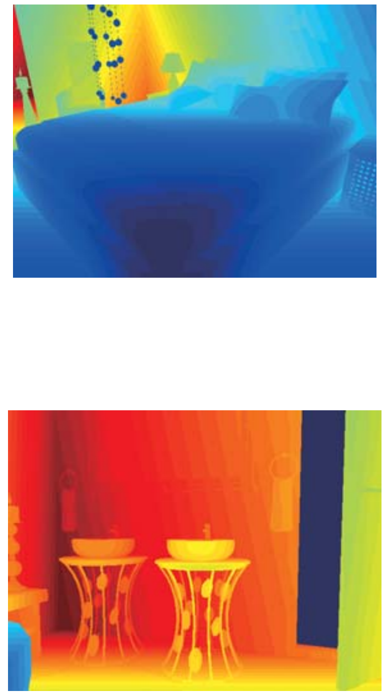
Receptive field: 0.98 m

# Semantic Scene Completion Network

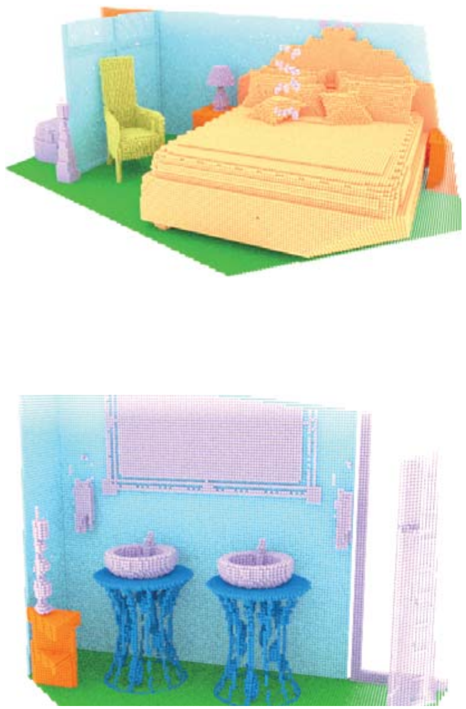


# Train on Synthetic 3D Scenes

Synthetic Scenes (SUNCG)

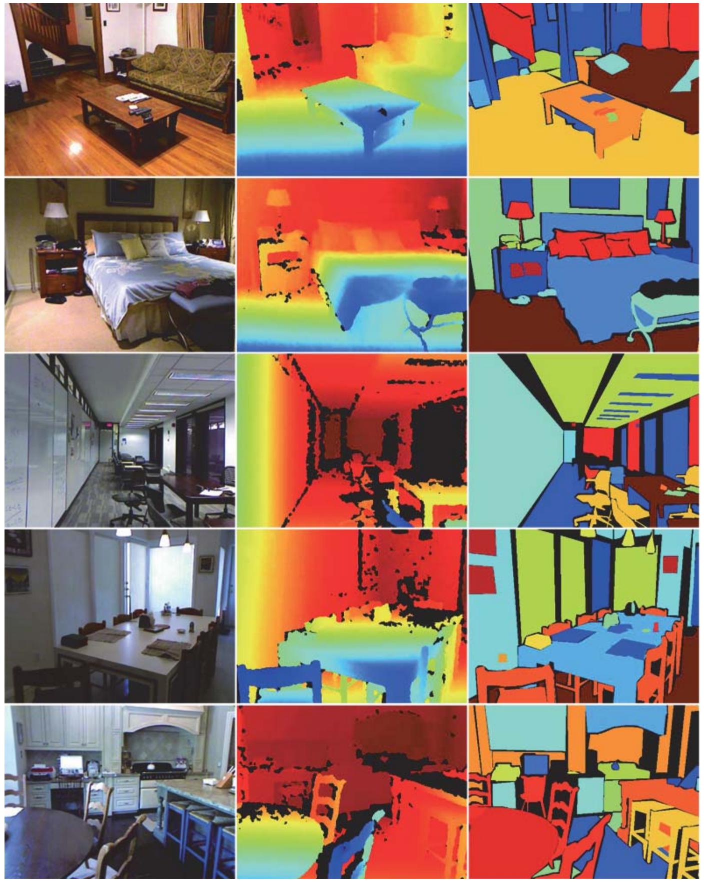


Depth



Ground Truth

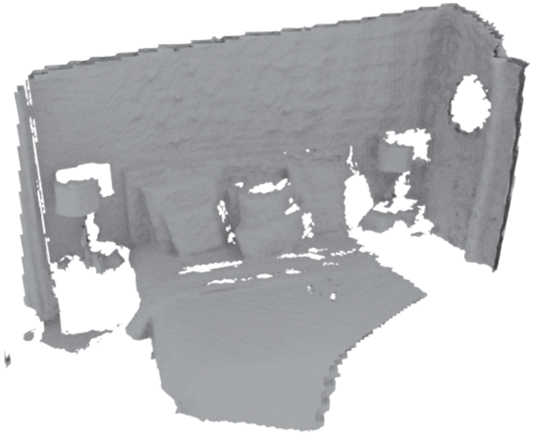
# Testing on Real-Word Data (NYU [1,2])



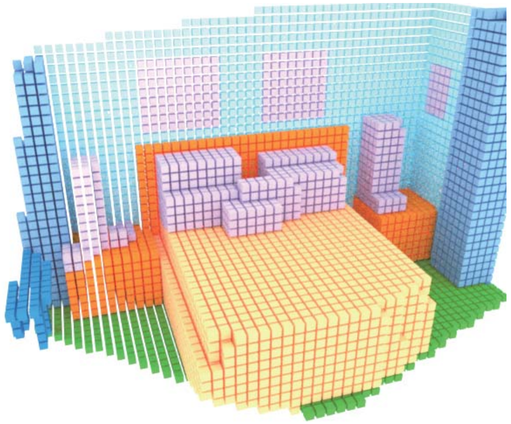
[1] **NYU depth v2:** Silberman et al. ECCV’12
[2] **Ground truth:** Guo and Hoiem IJCV’15

# Comparis on Real-world Dataset

# Comparison on Real-world Dataset

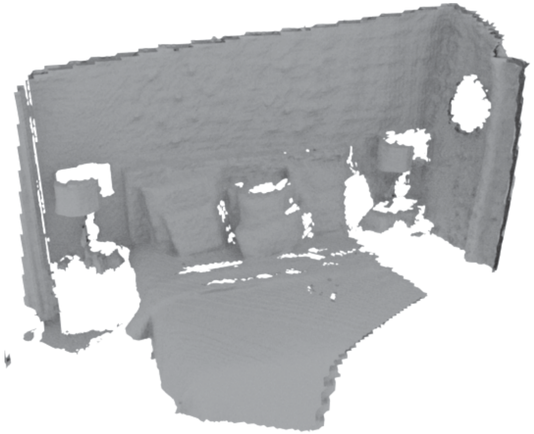


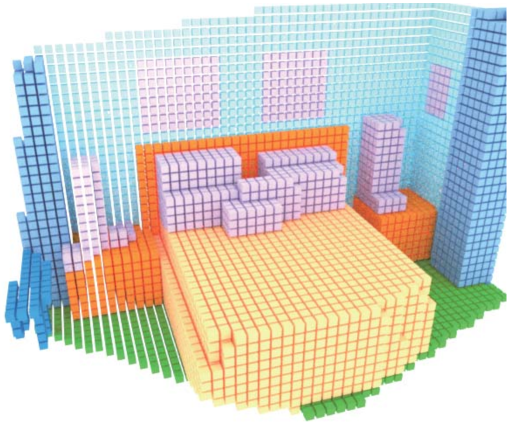
Observed Surface



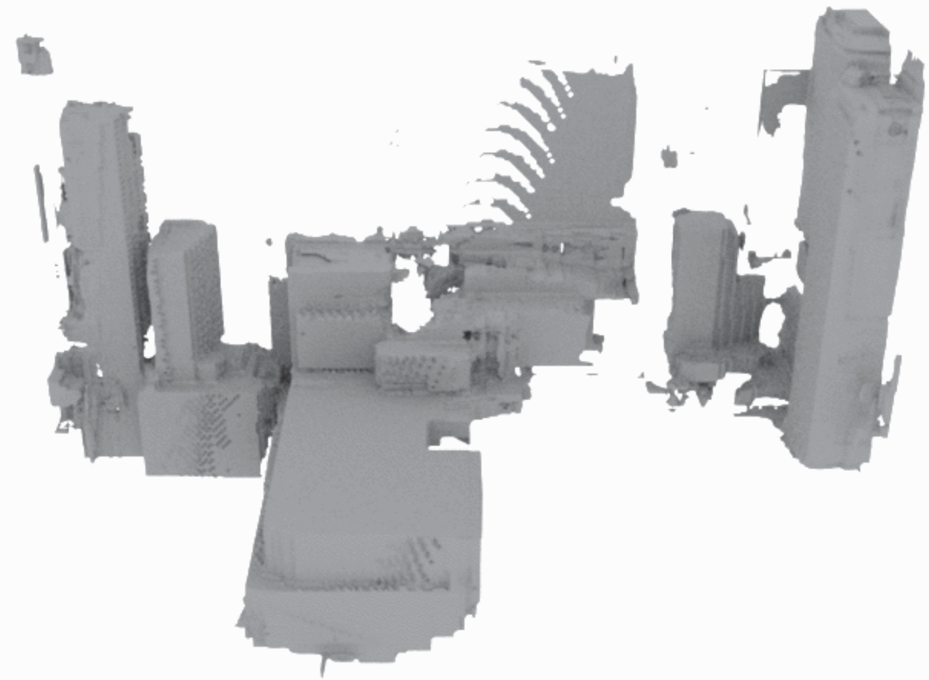
Ground Truth

# Comparison on Real-world Dataset

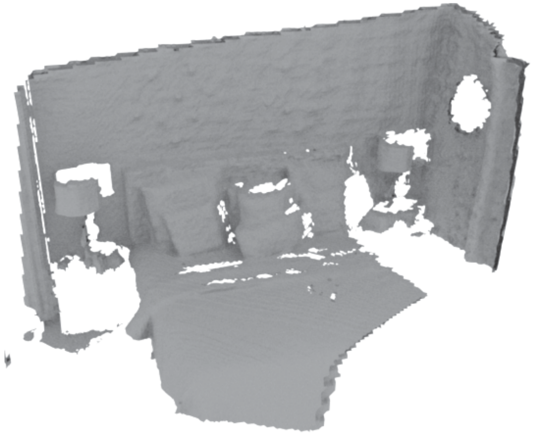

Observed Surface


Ground Truth

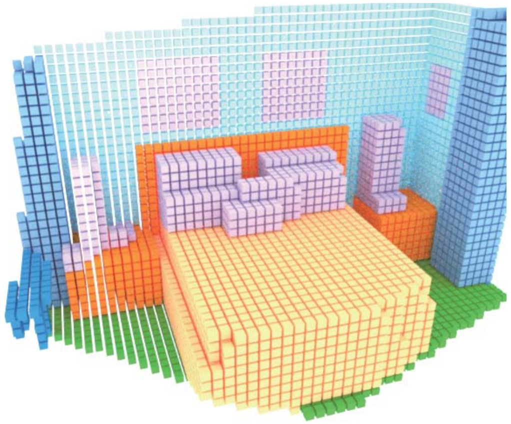
Shape Completion without Semantics
[Firman et al. CVPR'16]



# Comparison on Real-world Dataset

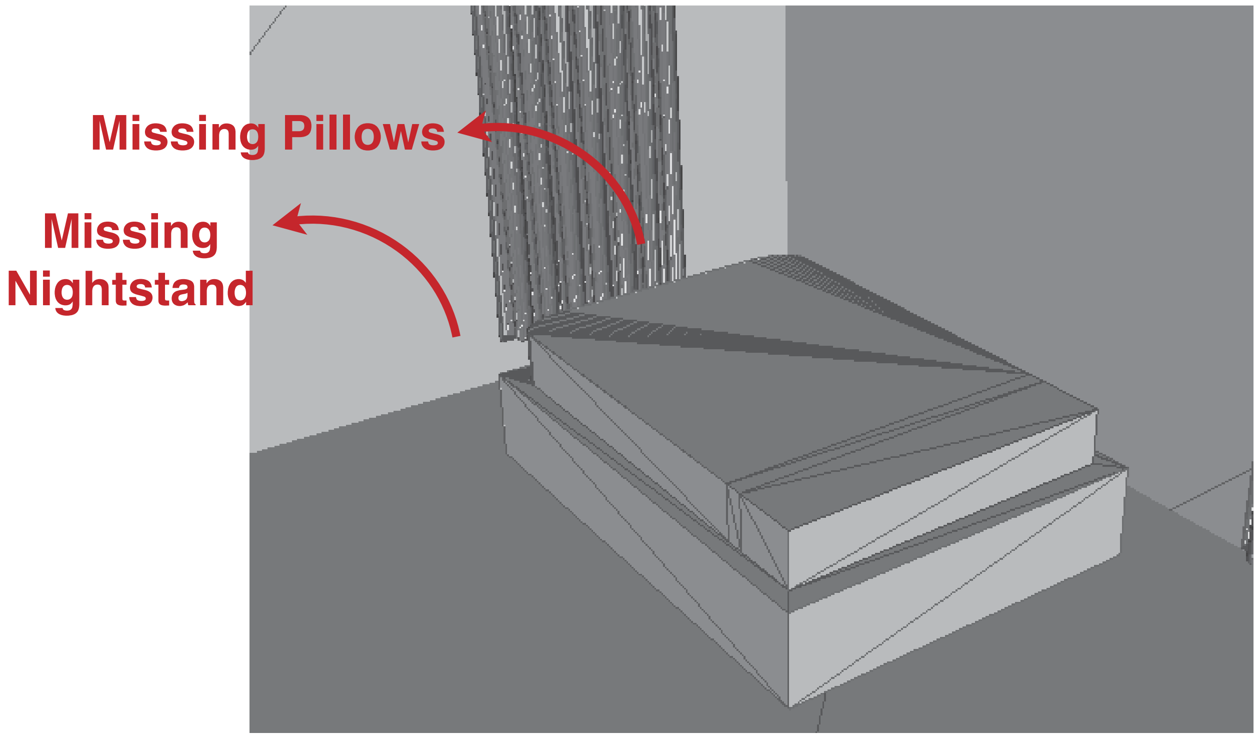


Observed Surface

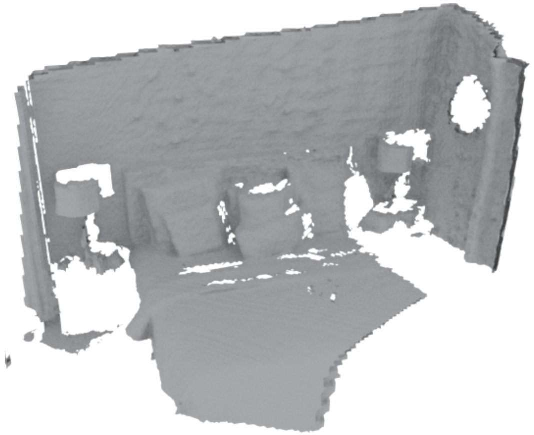


Ground Truth

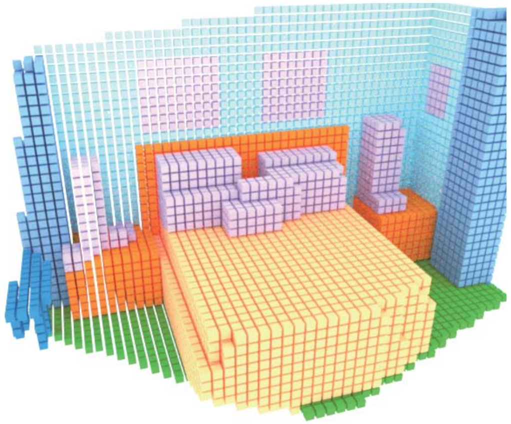
Model Retrieval+Fitting
[Geiger and Wang GCPR'15]



# Comparison on Real-world Dataset

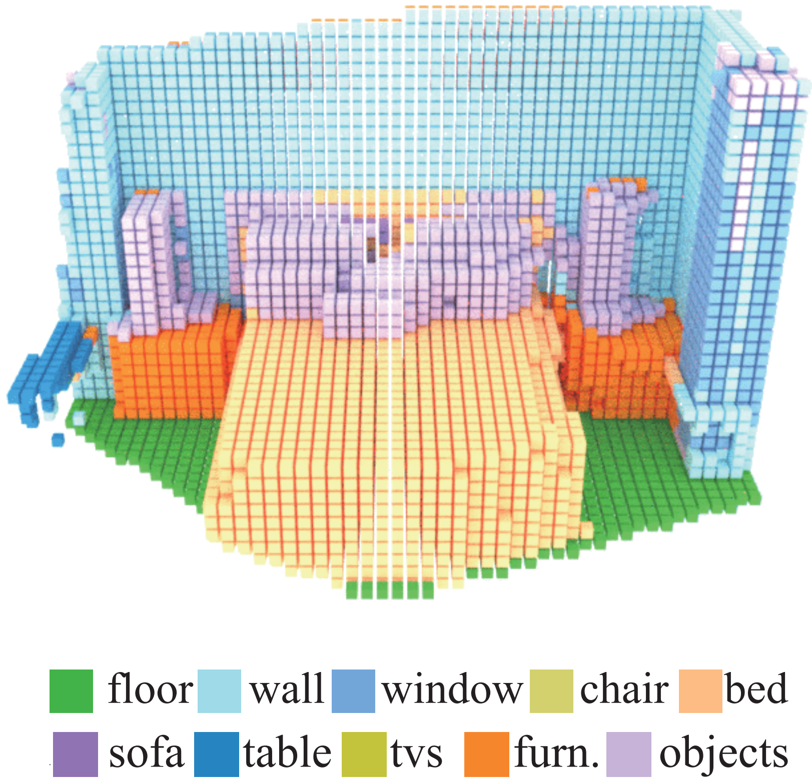


Observed Surface

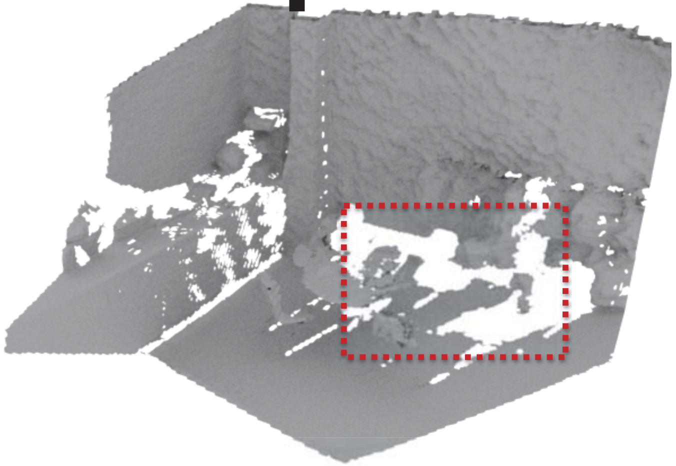


Ground Truth

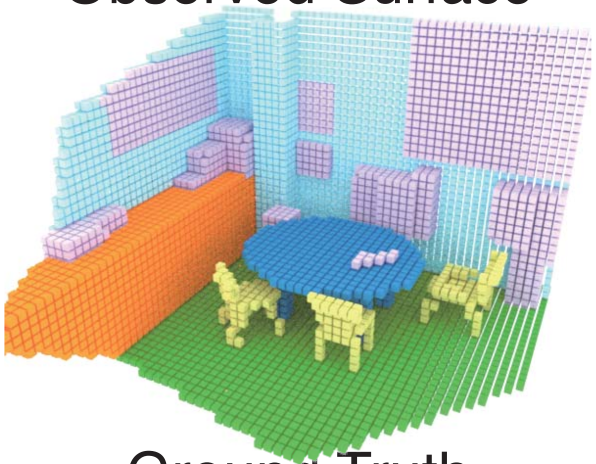
SSCNet
[Ours]



# Comparison on Real-world Dataset

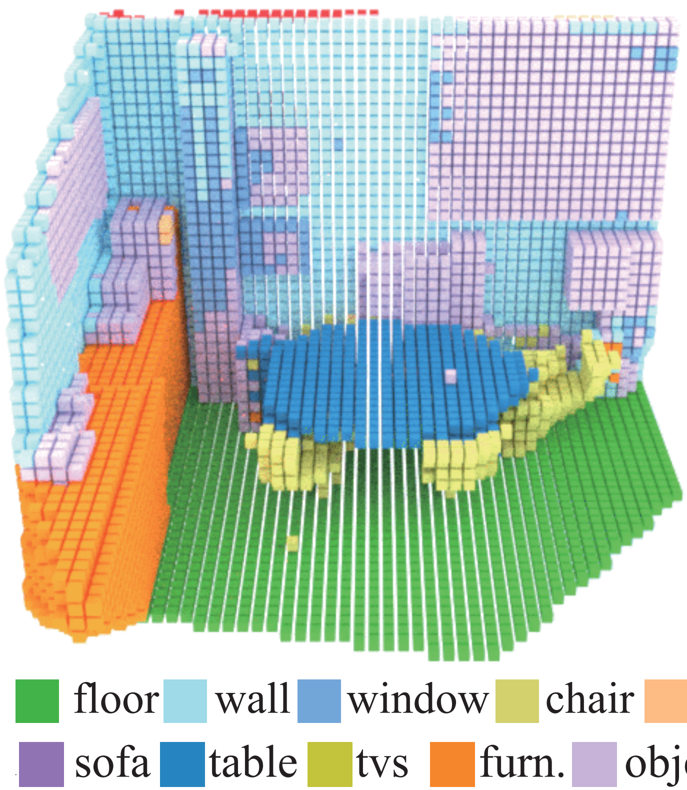


Observed Surface

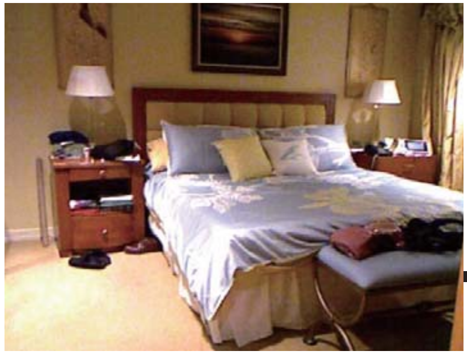


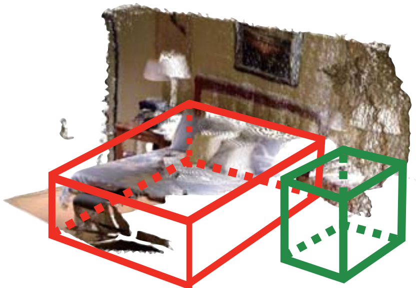
Ground Truth

SSCNet
[Ours]

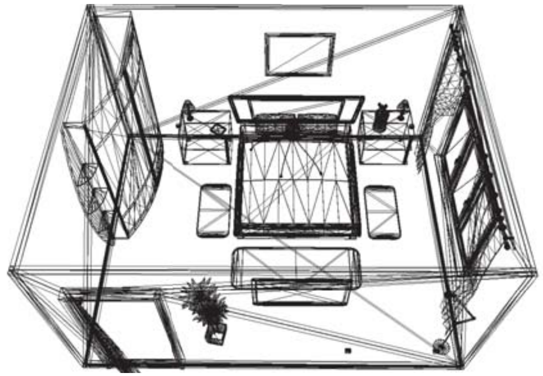


# Data-Driven 3D Scene Understanding





**Amodal 3D**
[Song and Xiao
ECCV'14,CVPR'16]



Prediction is limited by
Camera Field of View

**Higher Fidelity**
[Song et al. CVPR'17]

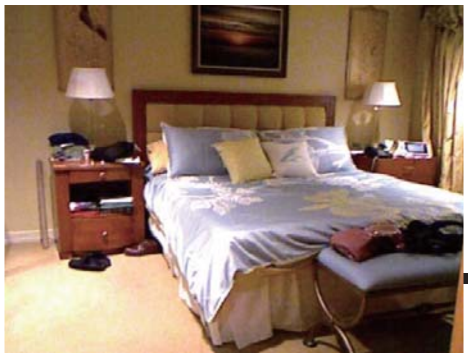
- ✓Semantics Category
- ✓3D Location, Size
- ✓Detailed Geometry
- ✓Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. properties …

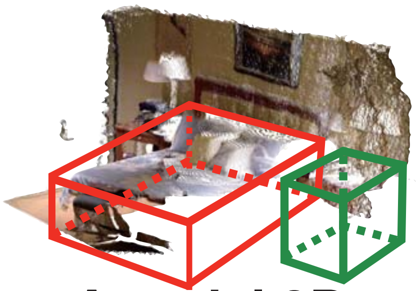
# Limited Camera FoV

Typical camera
FoV 60 degree

Small Portion of the Scene is
Observed due to Limited FoV

# Data-Driven 3D Scene Understanding



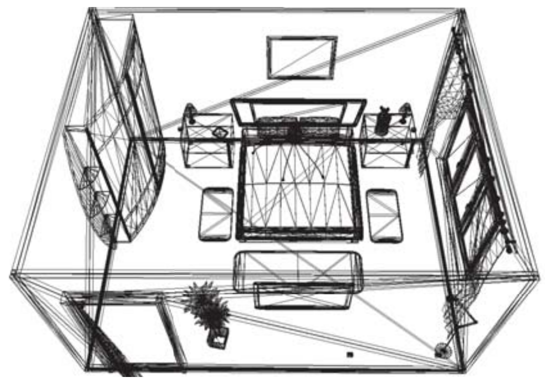


**Amodal 3D**
[Song and Xiao ECCV'14,CVPR'16]

**Higher Fidelity**
[Song et al. CVPR'17]



**Beyond FoV**
[song et al. CVPR'18]



✓Semantics Category
✓3D Location, Size
✓Detailed Geometry
✓Inter-Object Relationships
• **Not Limited by FoV**
• Action Affordances
• Phys. properties …

# View Extrapolation

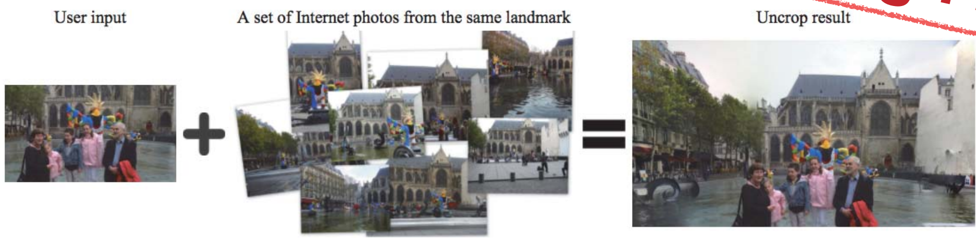
## Prior work: Predicting Scene Appearance (Only Colored Pixels)

Image Inpainting
[Pathak et. al]

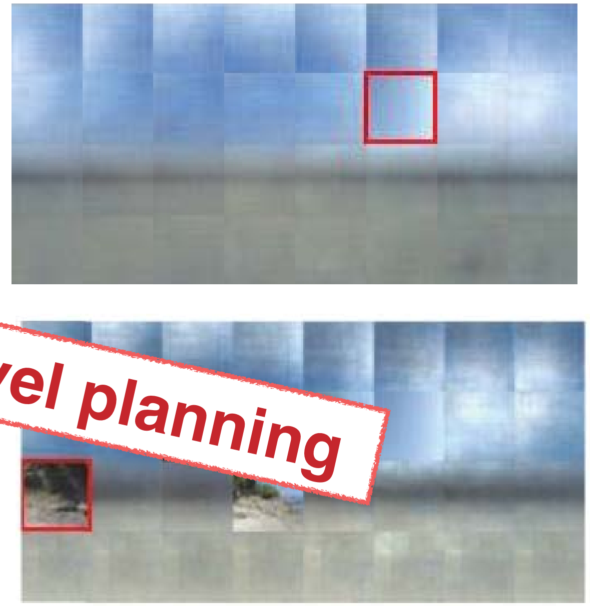
(a) Input context

User-guided view extrapolation [Zhang et al.]


Stitching images from the Inte...

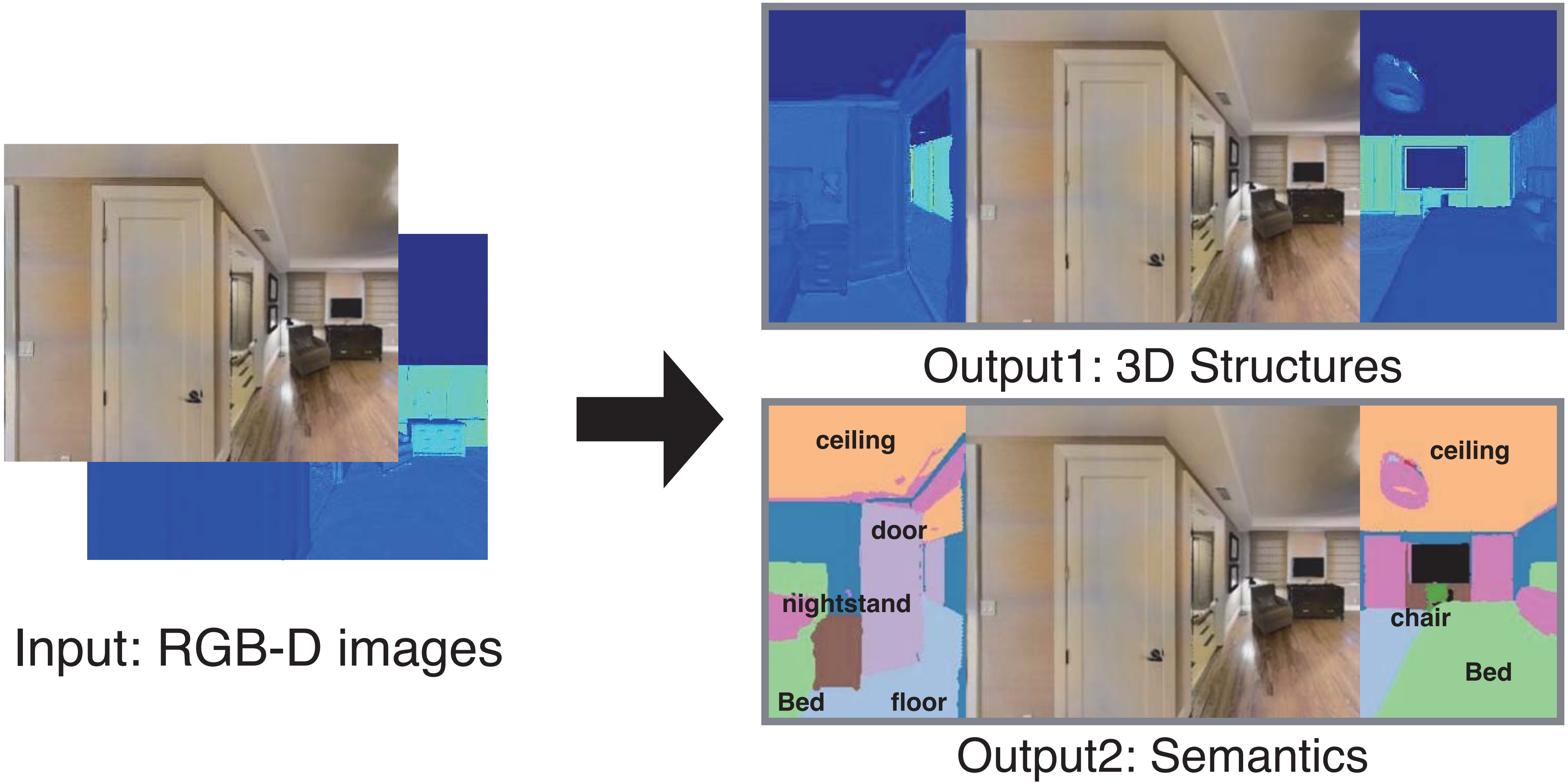



Learning to Look Around
[Jayaraman and Grauman]



**Hard to be used directly to support high level planning**

# View Extrapolation



# View Extrapolation

Where can I move?

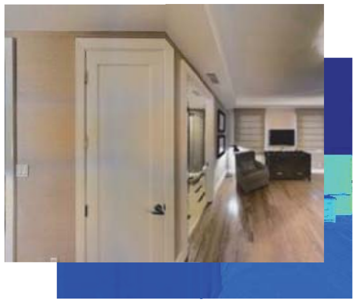
Output1: 3D Structures

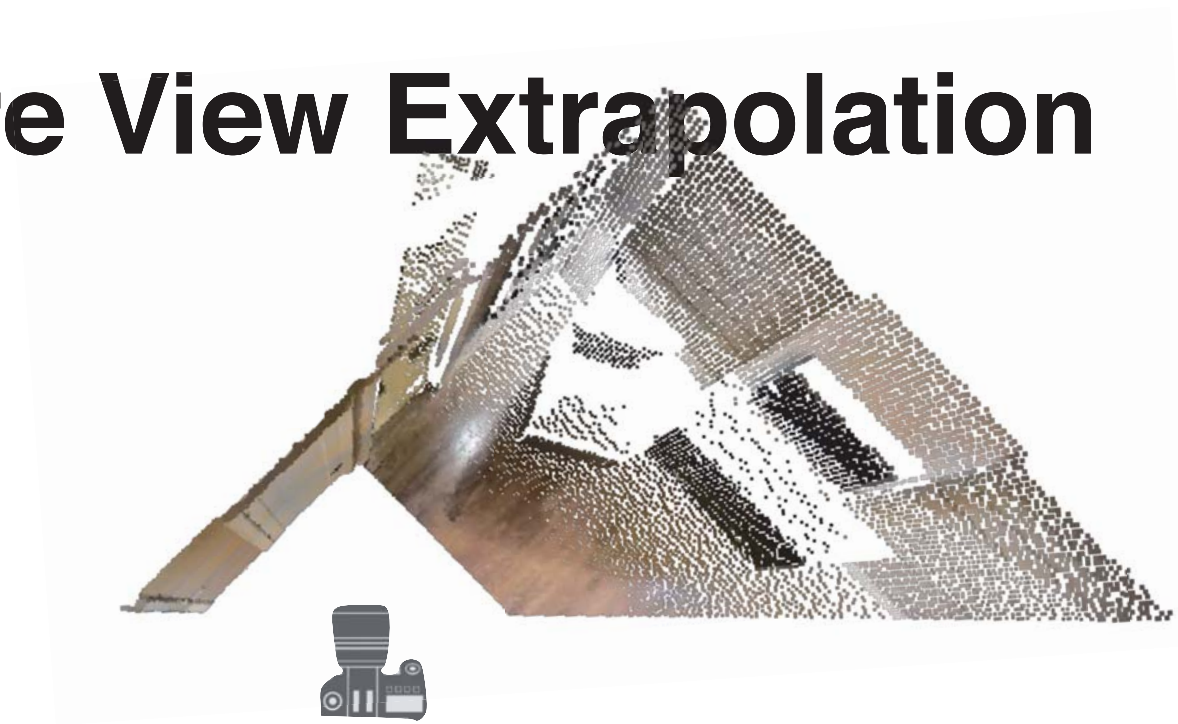
Where should I turn to find a door?


Output2: Semantics

# Semantic-Structure View Extrapolation

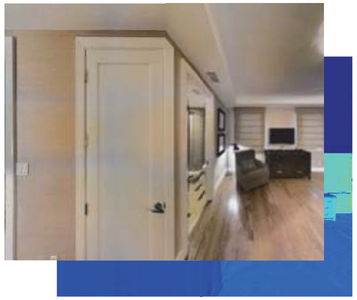
**Input:** RGB-D images



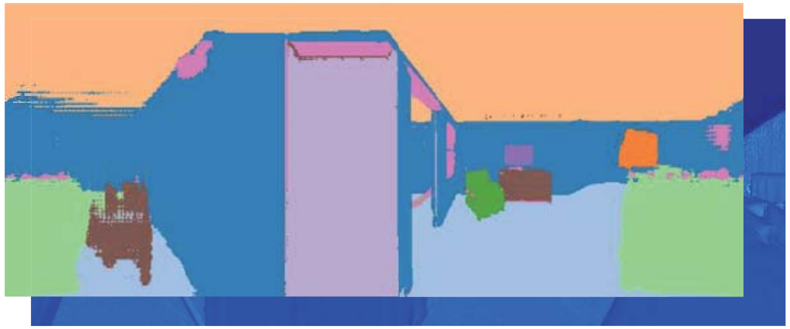


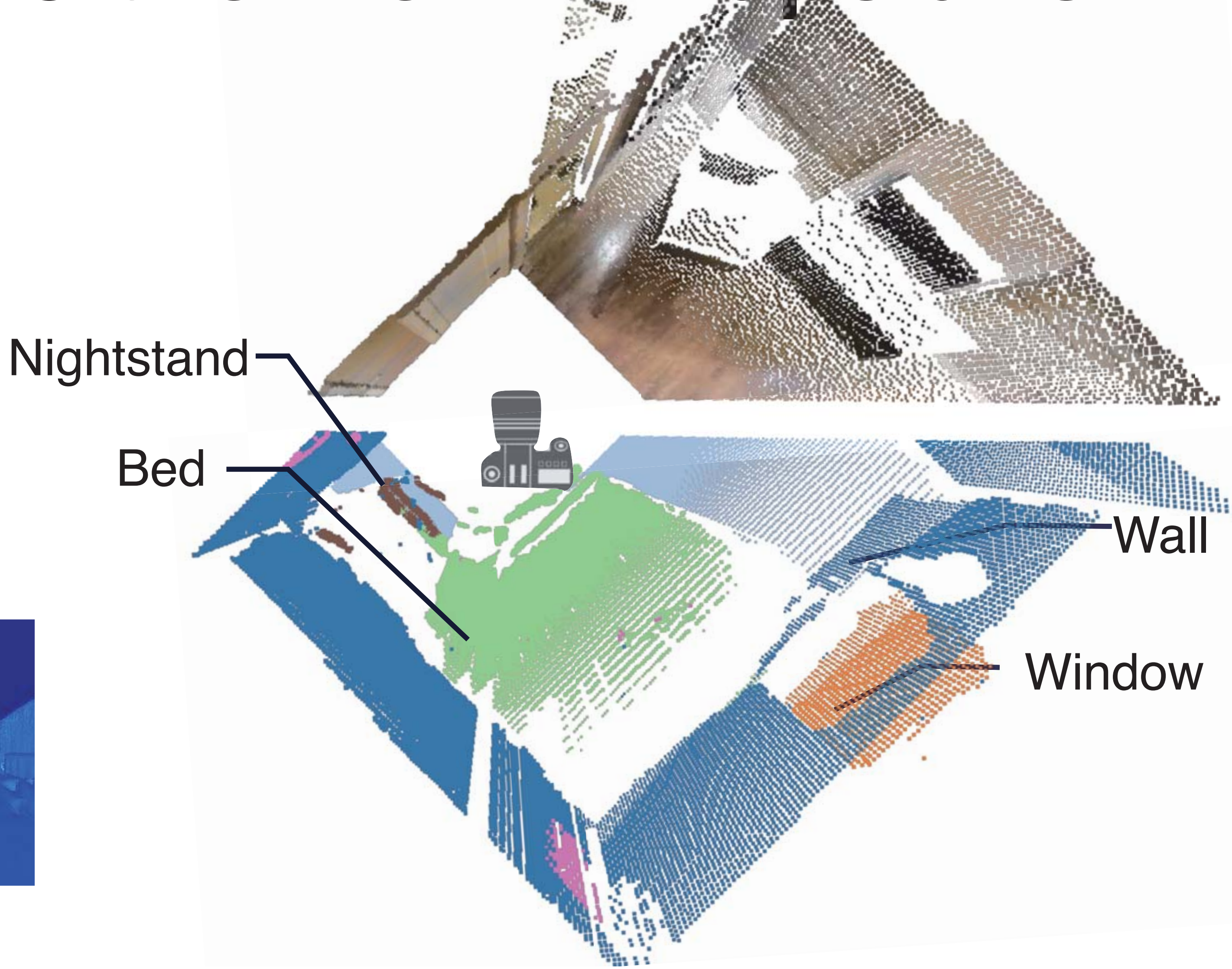
# Semantic-Structure View Extrapolation

**Input:** RGB-D images

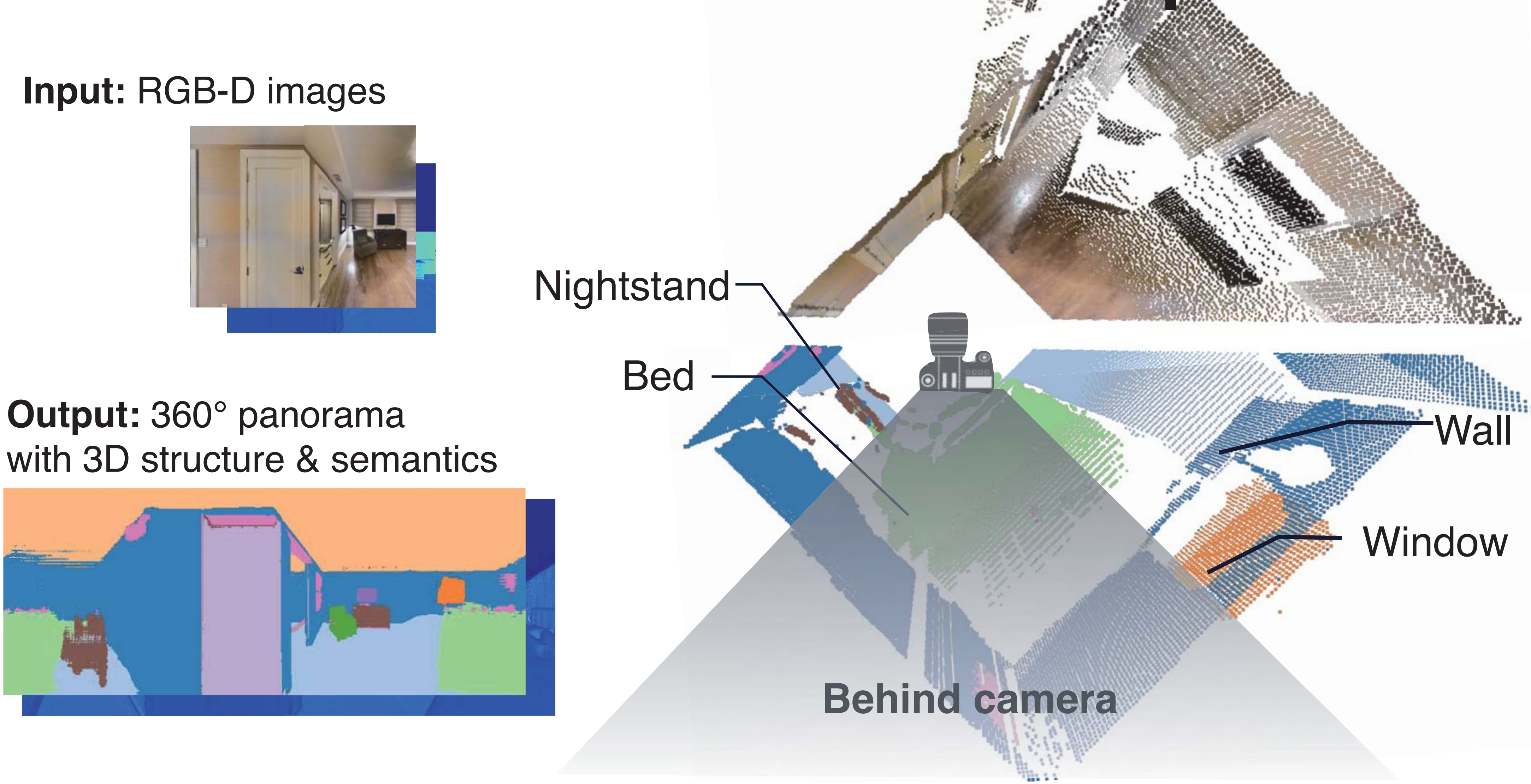


**Output:** 360° panorama
with 3D structure & semantics





# Semantic-Structure View Extrapolation



# Key idea

**Key idea:** Indoor environments are often **highly structured**.
By learning over the statistics of many typical scenes, the model should be able to leverage **strong contextual cues** inside the image to predict what is beyond the FoV.

# Training data

**3D House Datasets**

**Synthetic Houses (SUNCG):**

58,866 RGB-D panoramas

Pre-train

**Real-Word Houses (Matterport3D):**

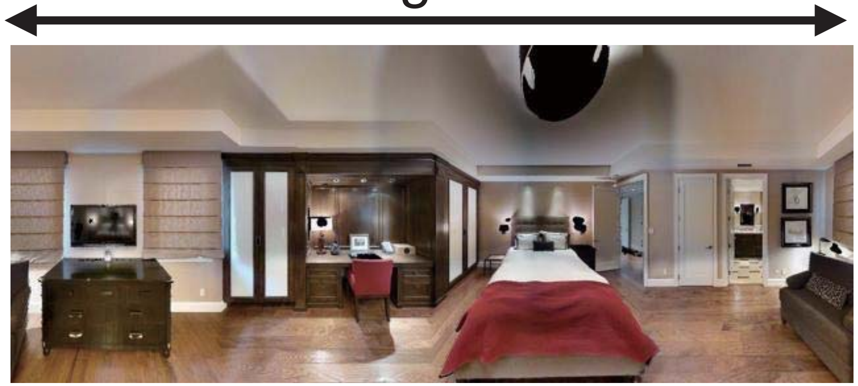
5,315 RGB-D panoramas

Fine-tune and test

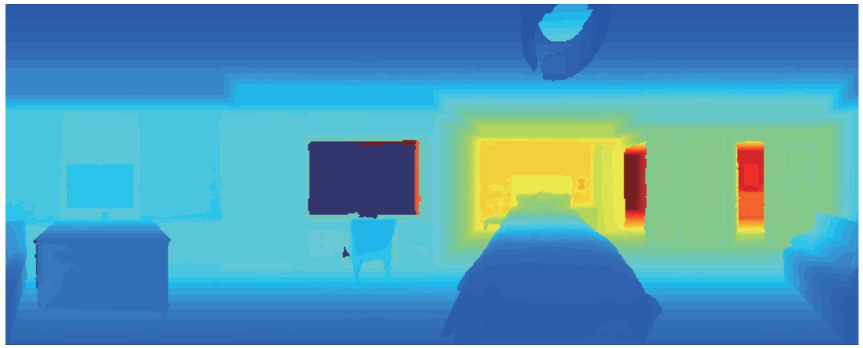
# Data Representation

3D Room

360 Degree FoV

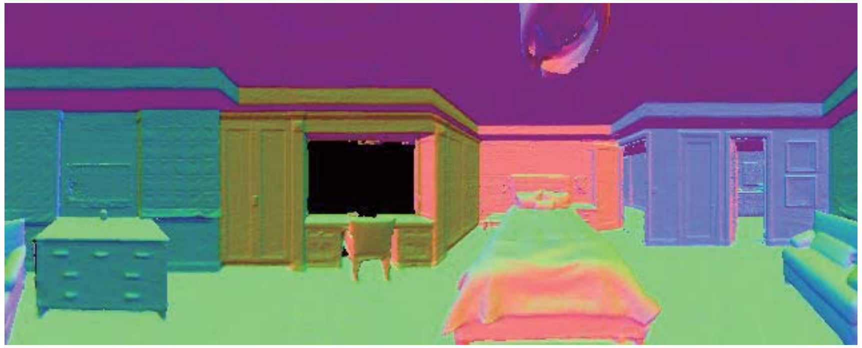


Color Panorama



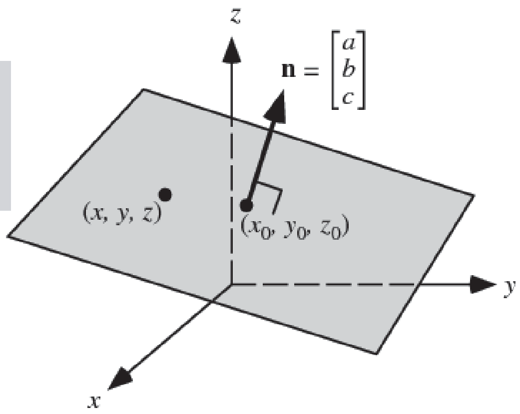
Depth Panorama

# Data Representation

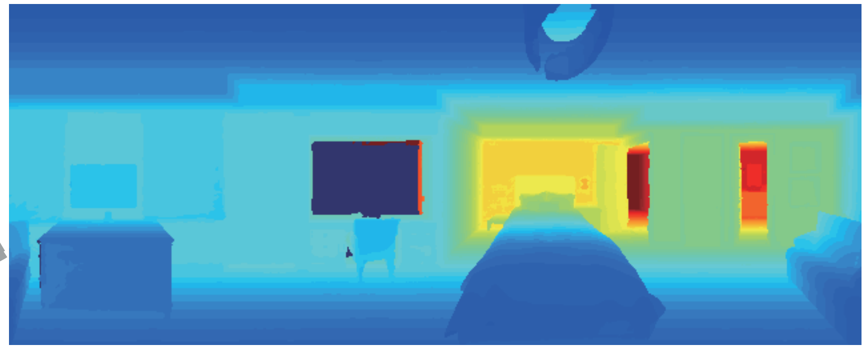


Surface Normal (a,b,c)

Plane Equation:
ax+by+cz-p=0

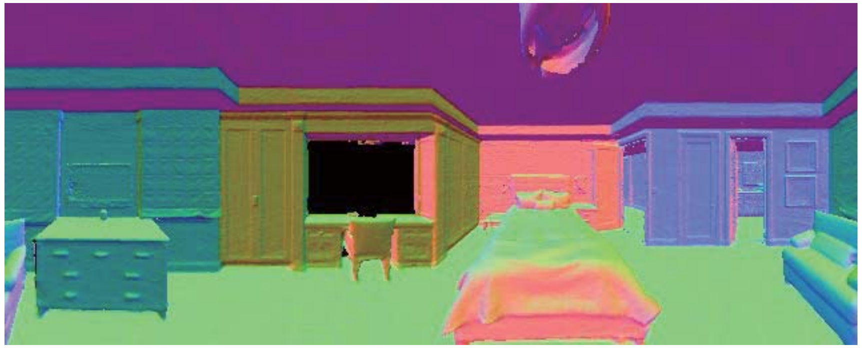



Plane Distance to Origin (p)



Depth Panorama

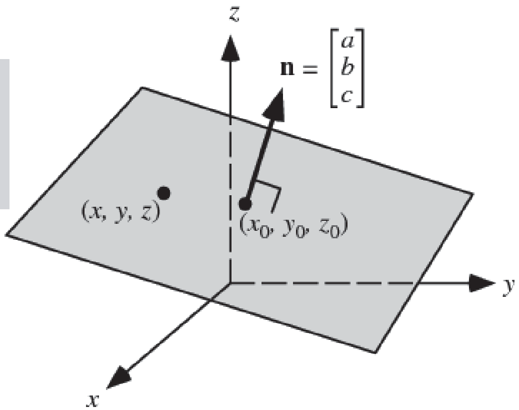
# Data Representation



Surface Normal (a,b,c)

Plane Distance to Origin (p)

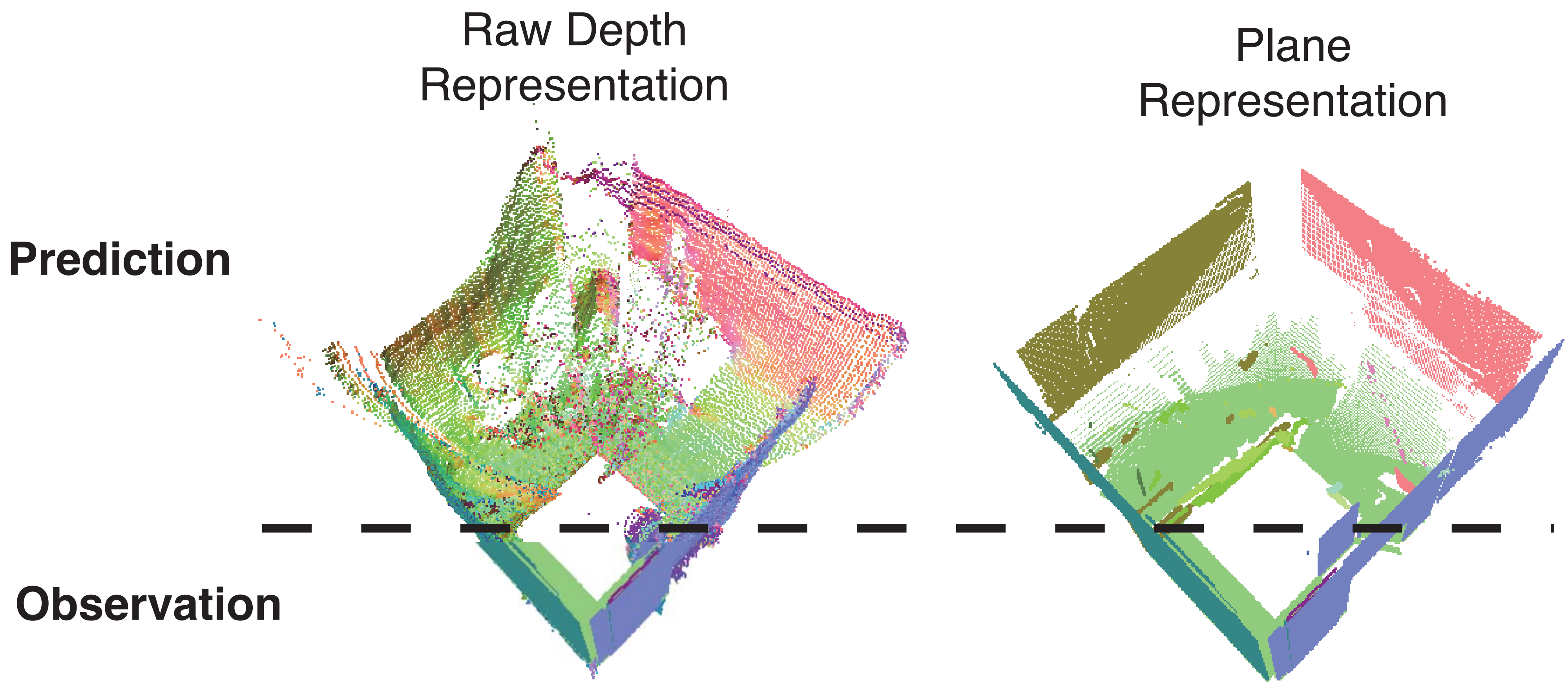
Plane Equation:
ax+by+cz-p=0



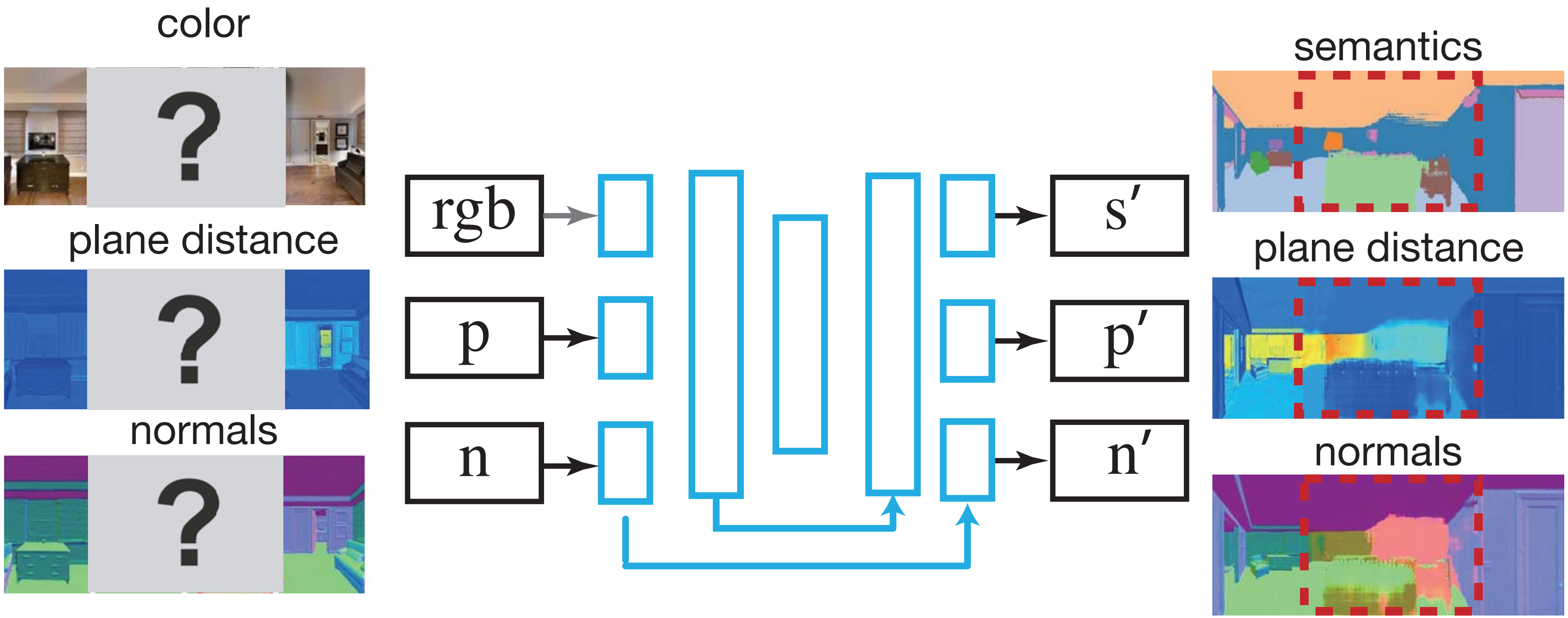

✓Pixels on the same planar surface share the same plane equation.

✓Representation is piecewise constant in a typical indoor environment.

# Data Representation

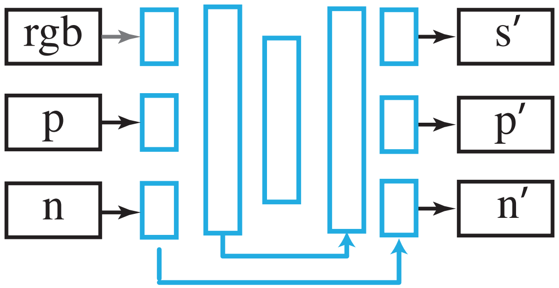


# Im2Pano3D Network



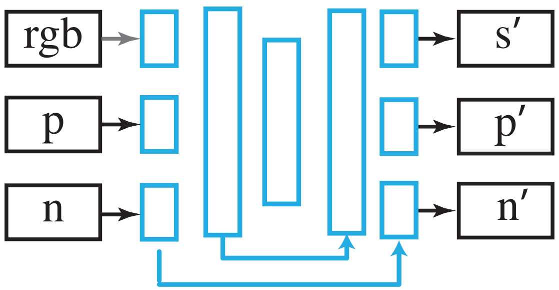

**What training objectives should we use?**

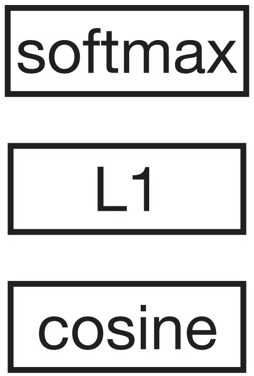
# Training Objectives



# Training Objectives

Every Pixel is Correct

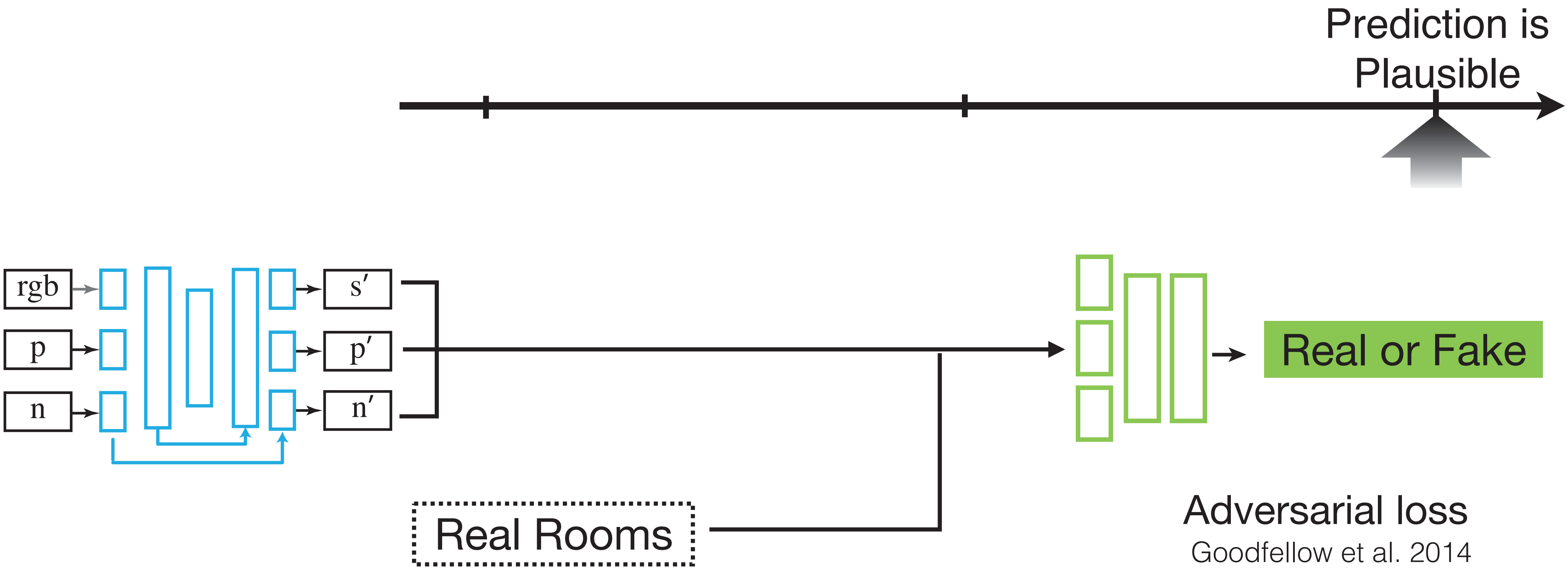


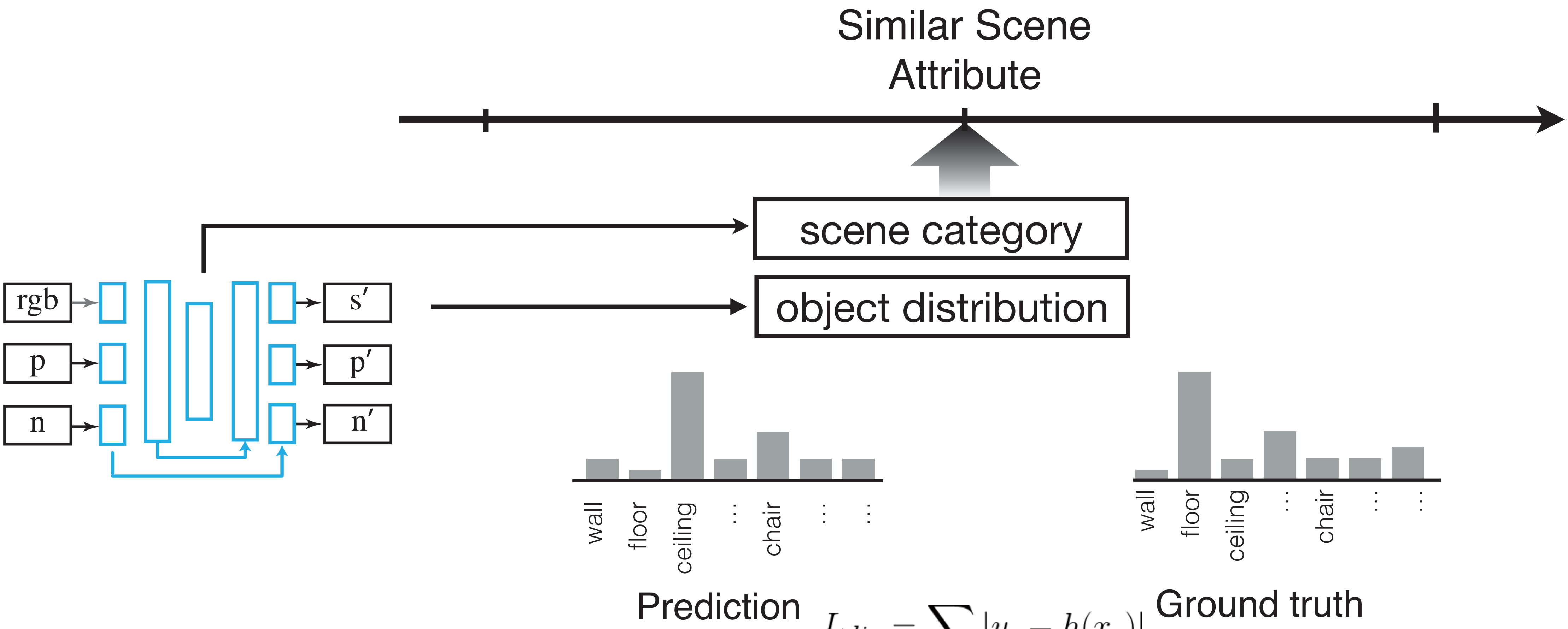

Prediction

Ground truth

# Training Objectives

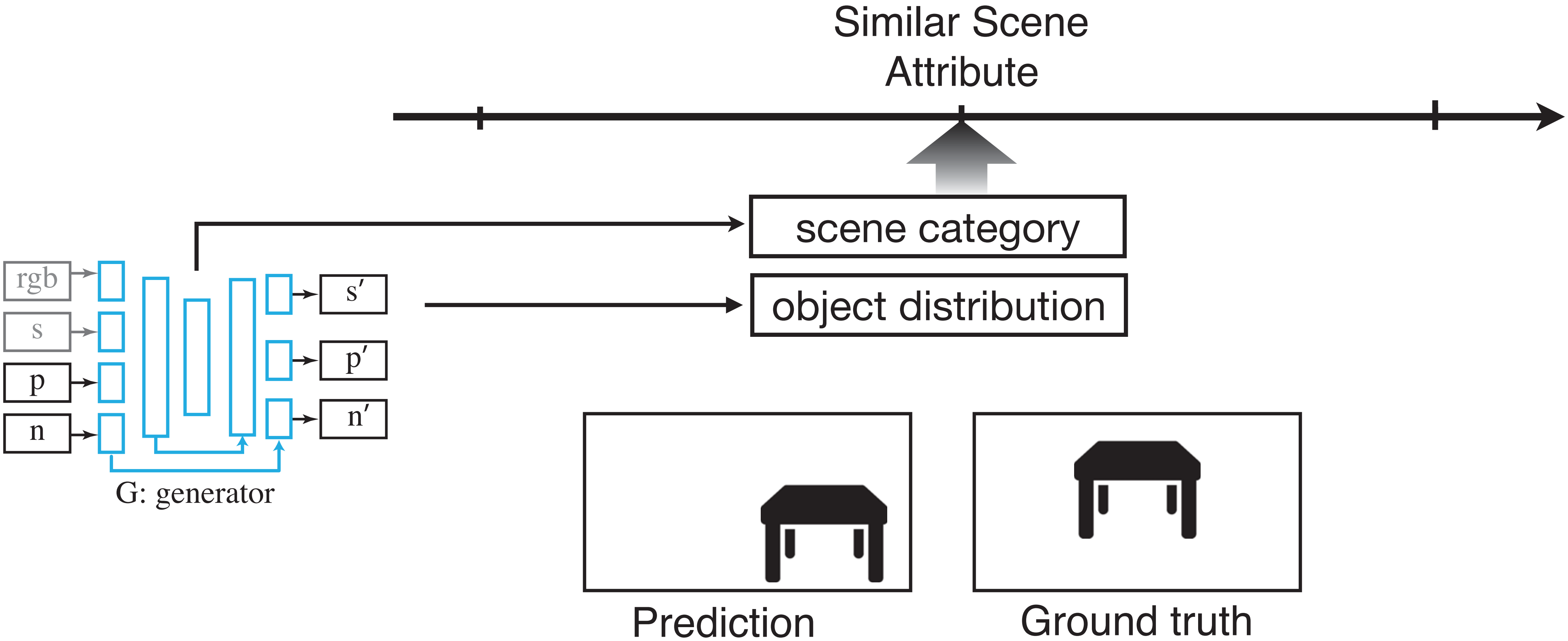


# Training Objectives




$$L_{dis} = \sum_{c} |y_c - h(x_c)|$$

# Training Objectives

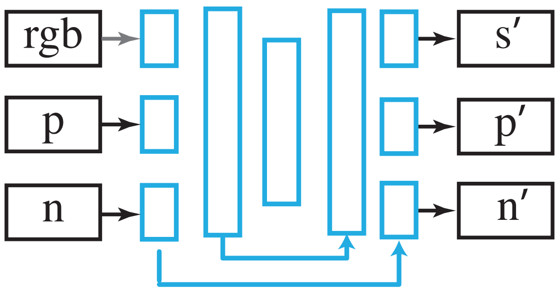


# Training Objectives

Every Pixel is Correct — $L_{recon}$

Similar Scene Attribute — $L_{attribute}$

Prediction is Plausible — $L_{adv}$

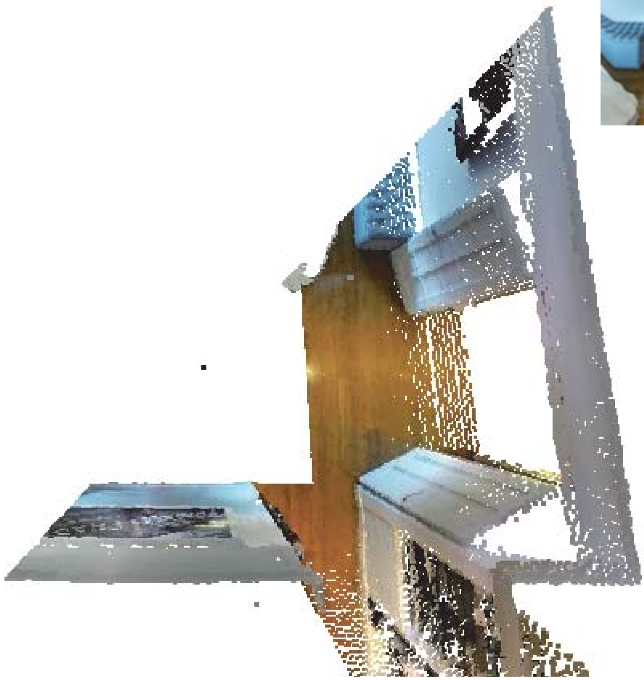



$$L = \lambda_1 L_{recon} + \lambda_2 L_{attribute} + \lambda_3 L_{adv}$$

# Results

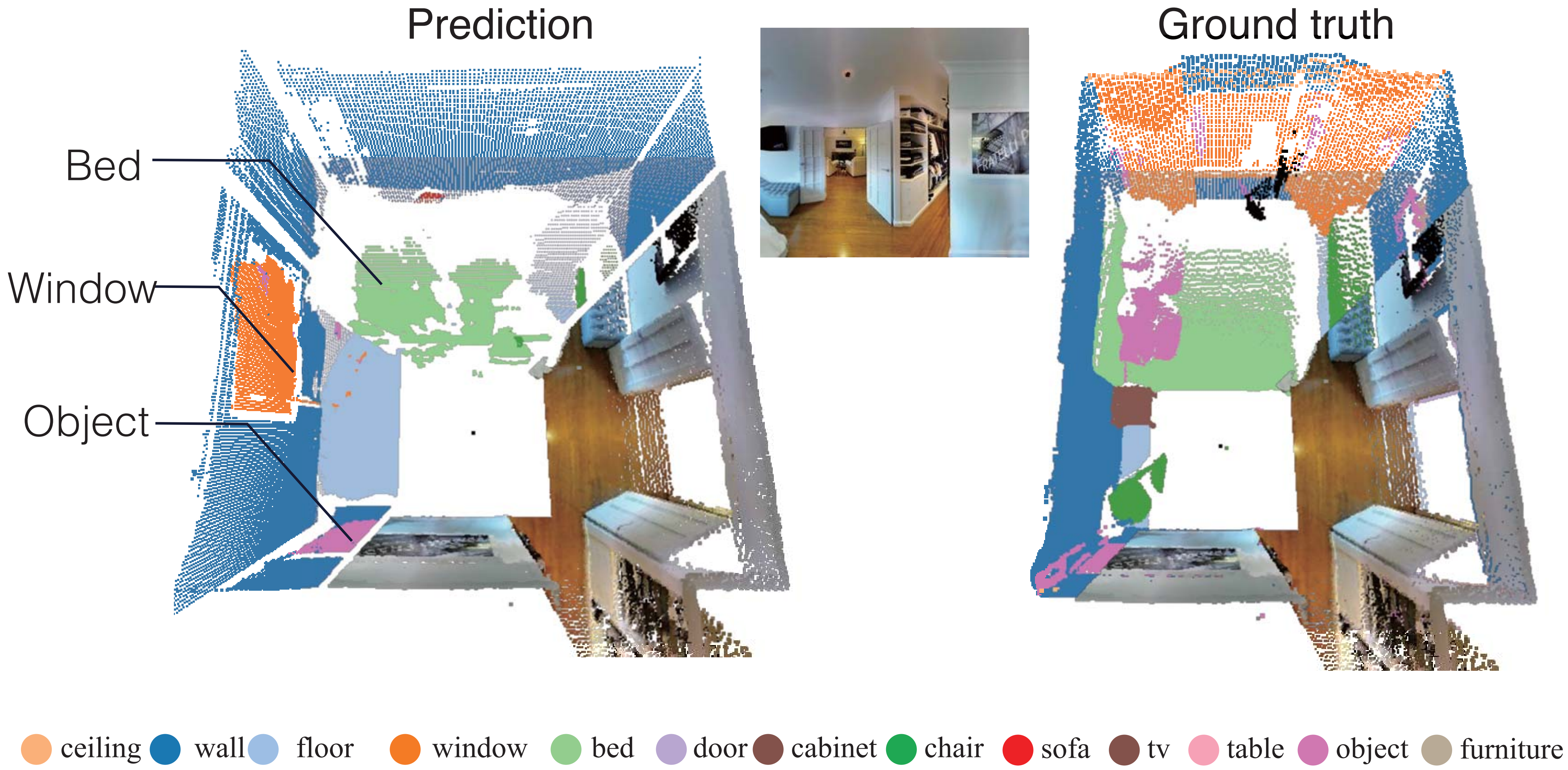
# Results

Input Observation

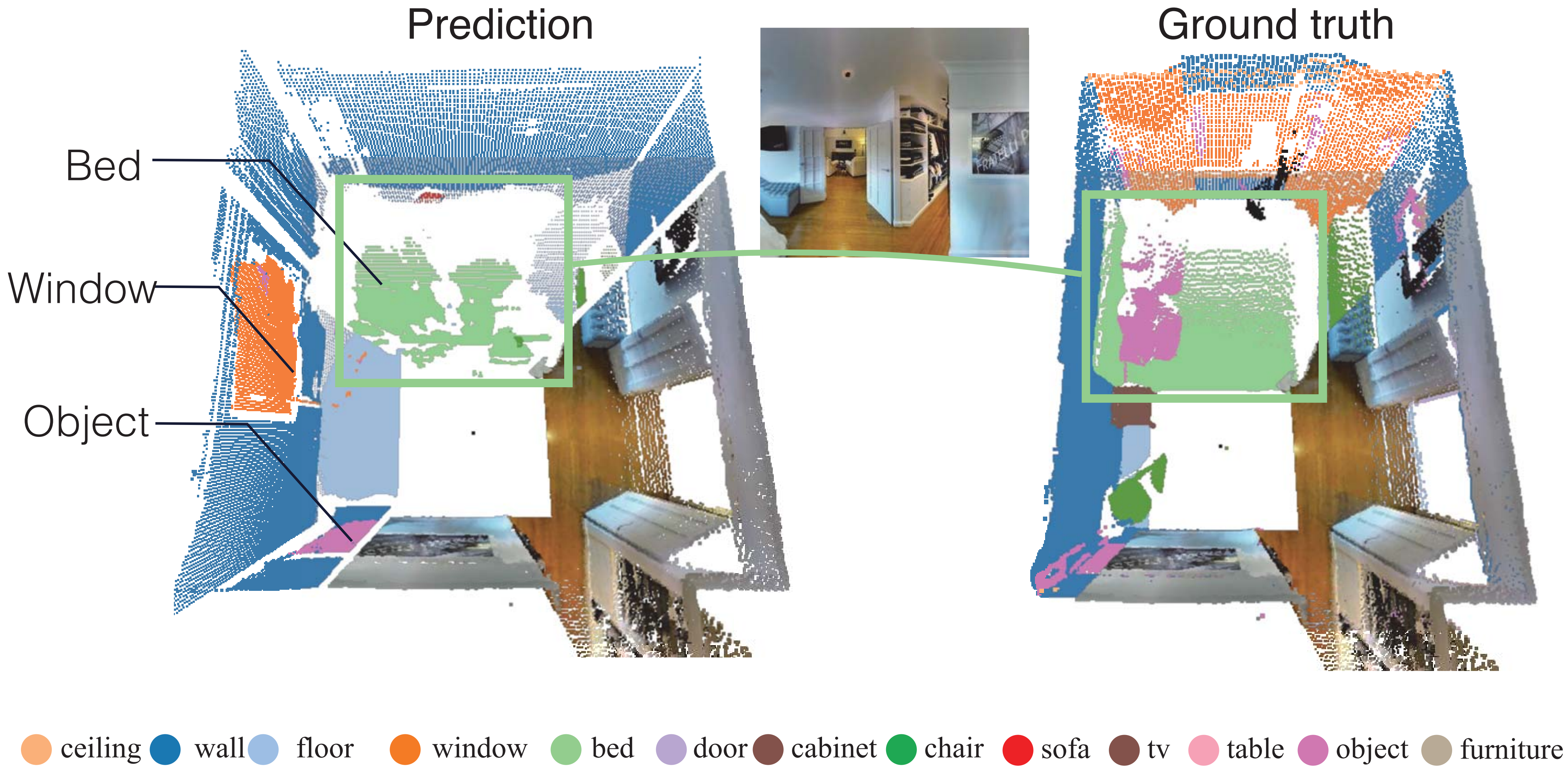


ceiling wall floor window bed door cabinet chair sofa tv table object furniture

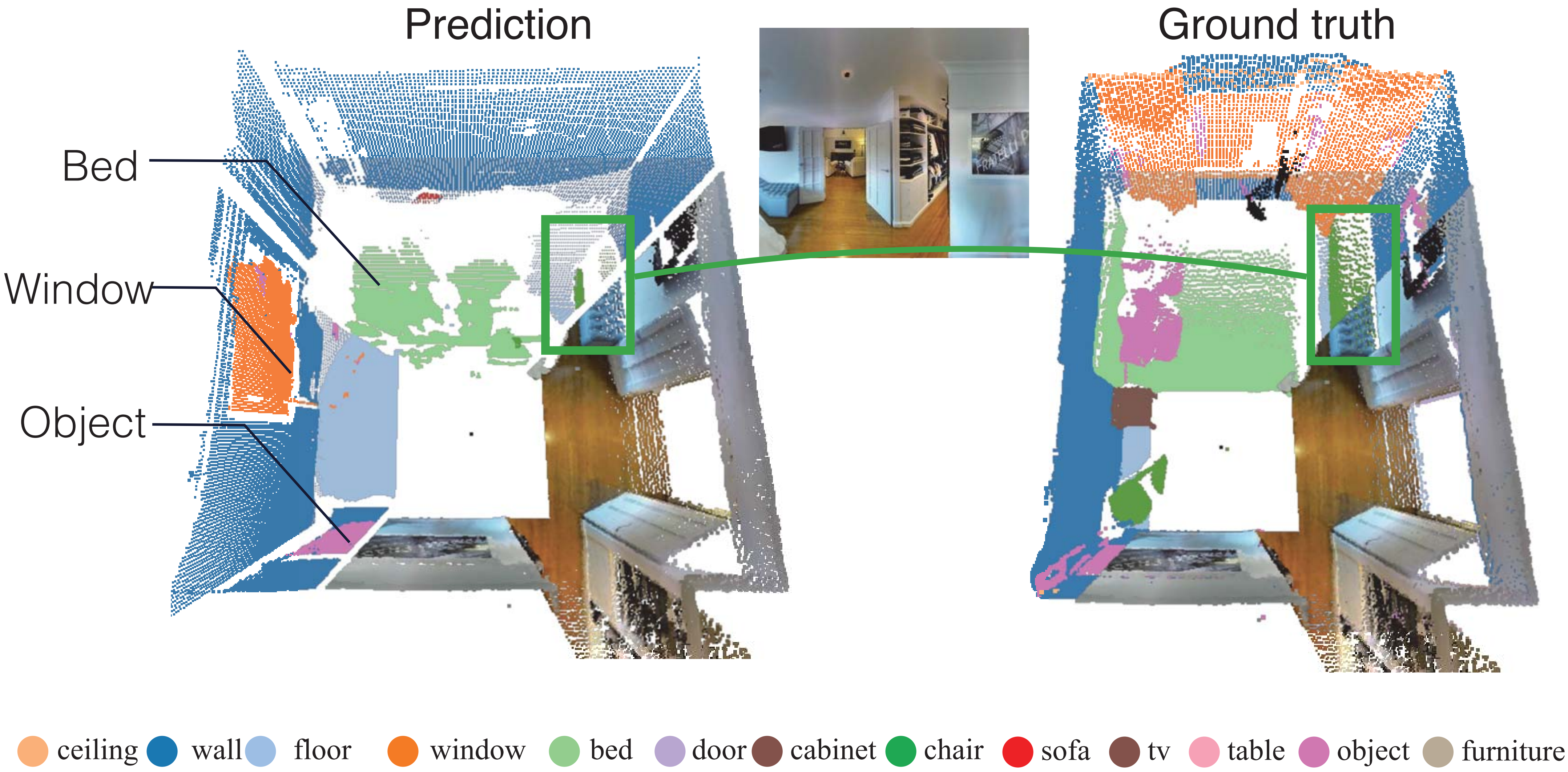
# Results


Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

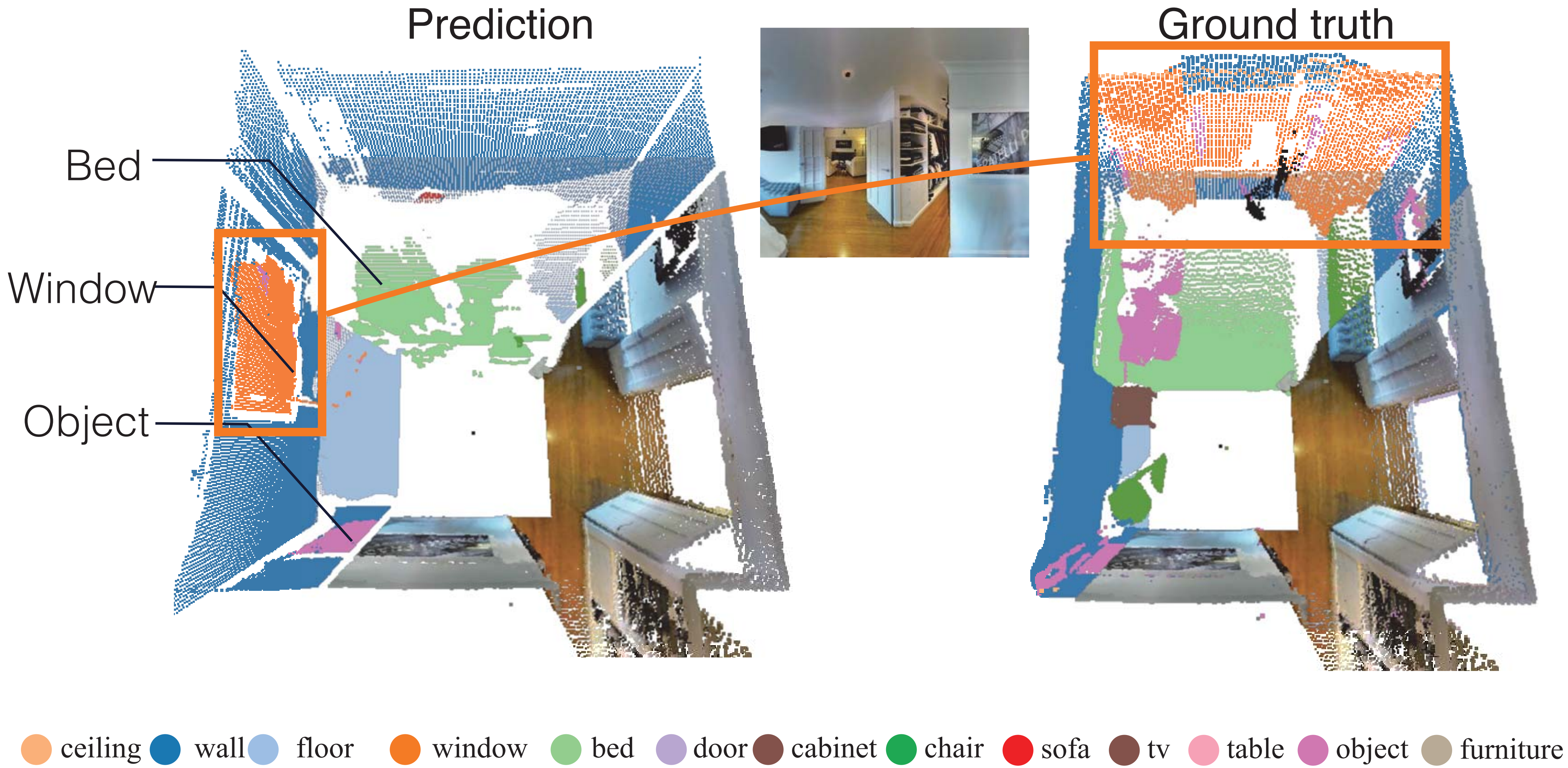
# Results


Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

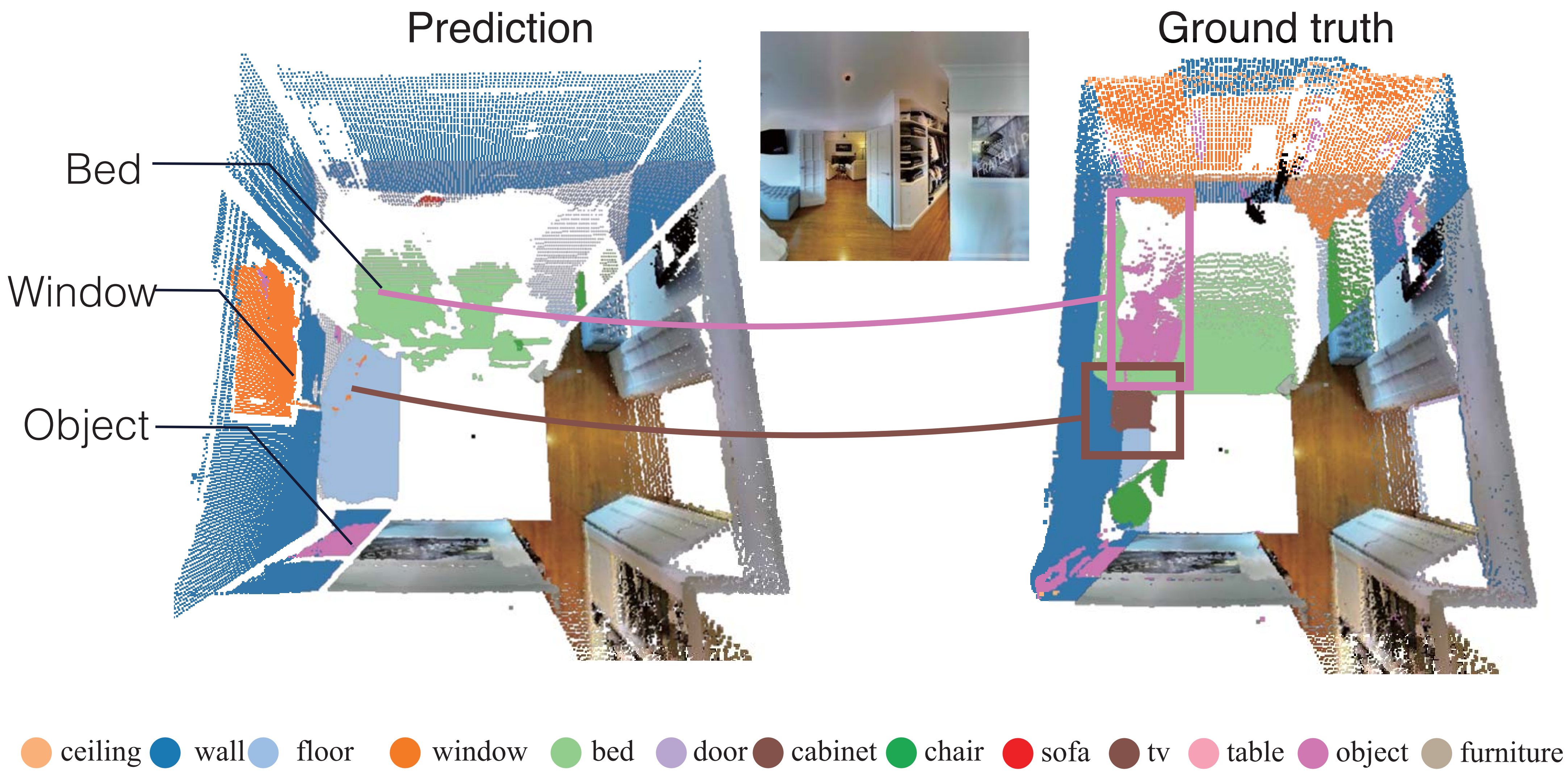
# Results


Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results

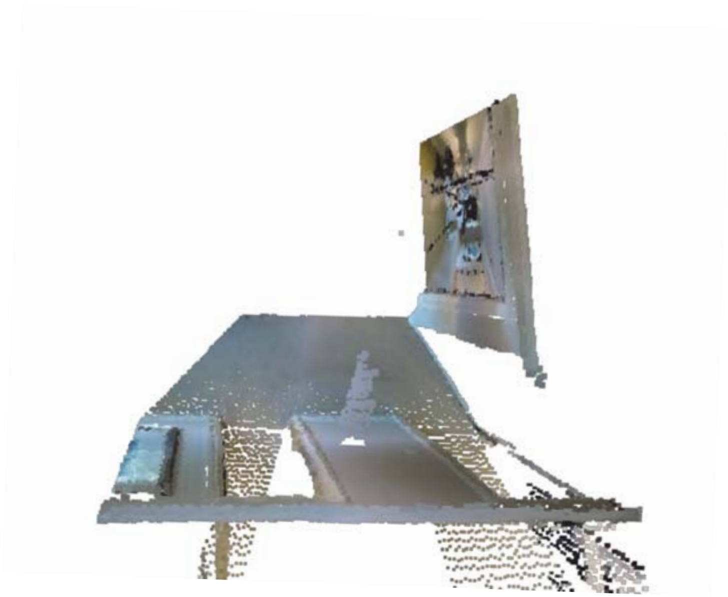

Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results


Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

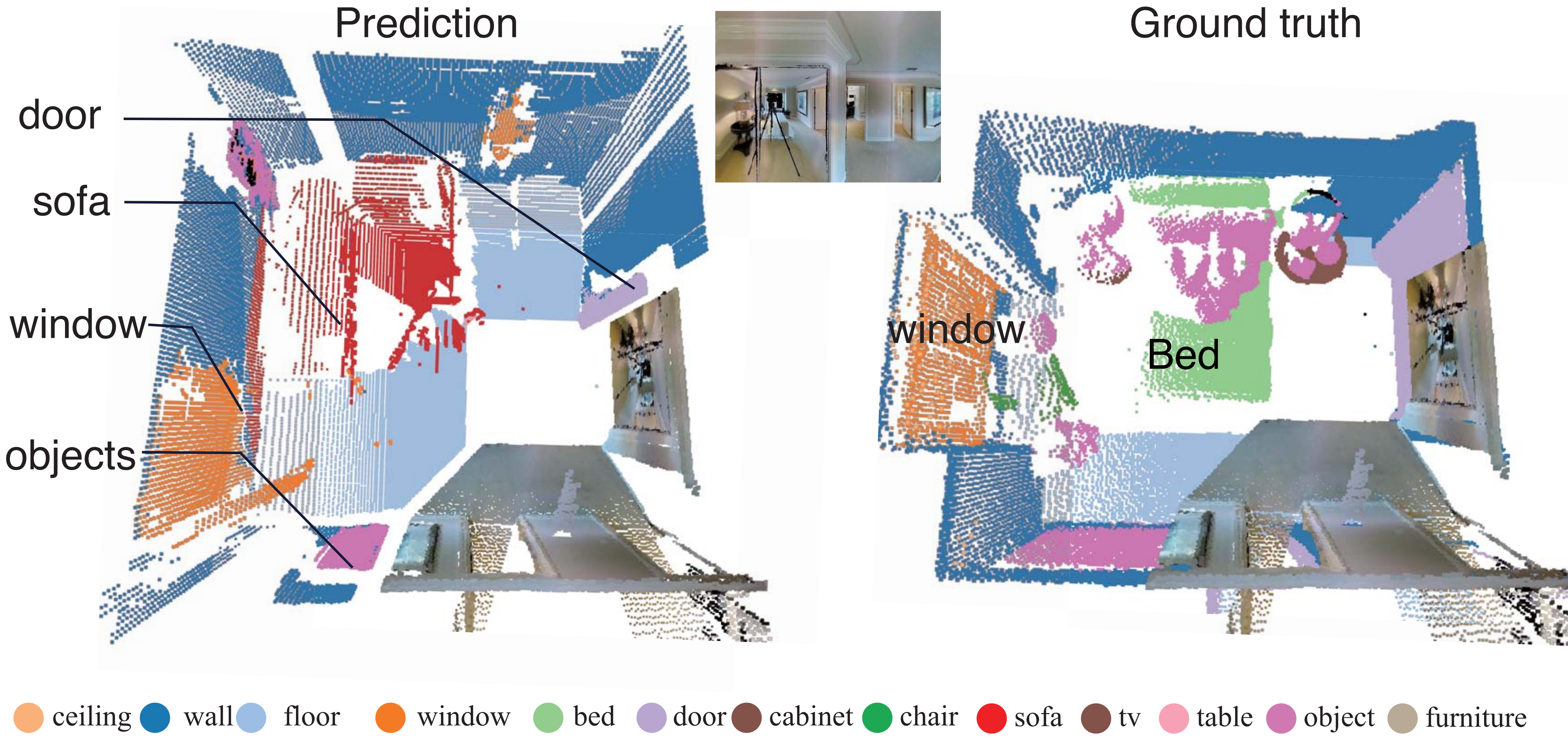
# Results

Input Observation

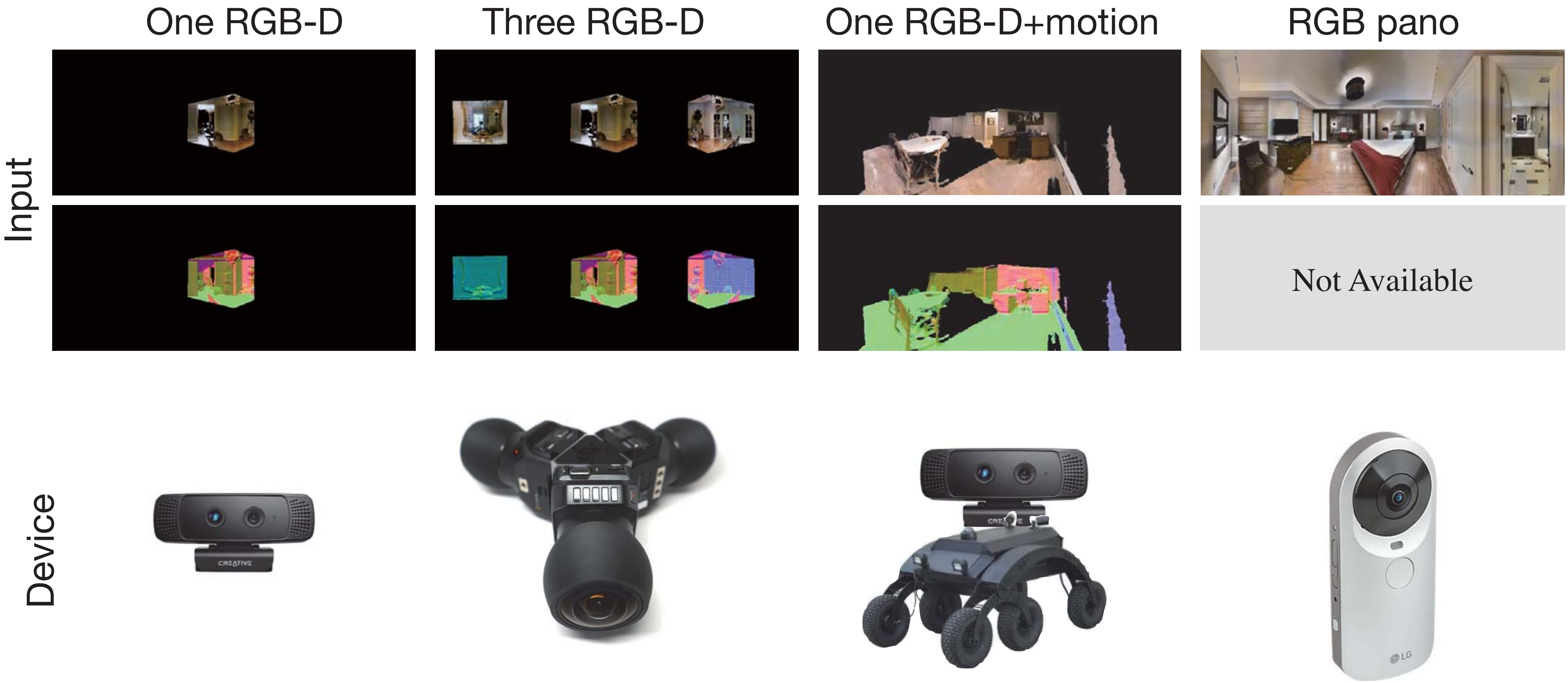


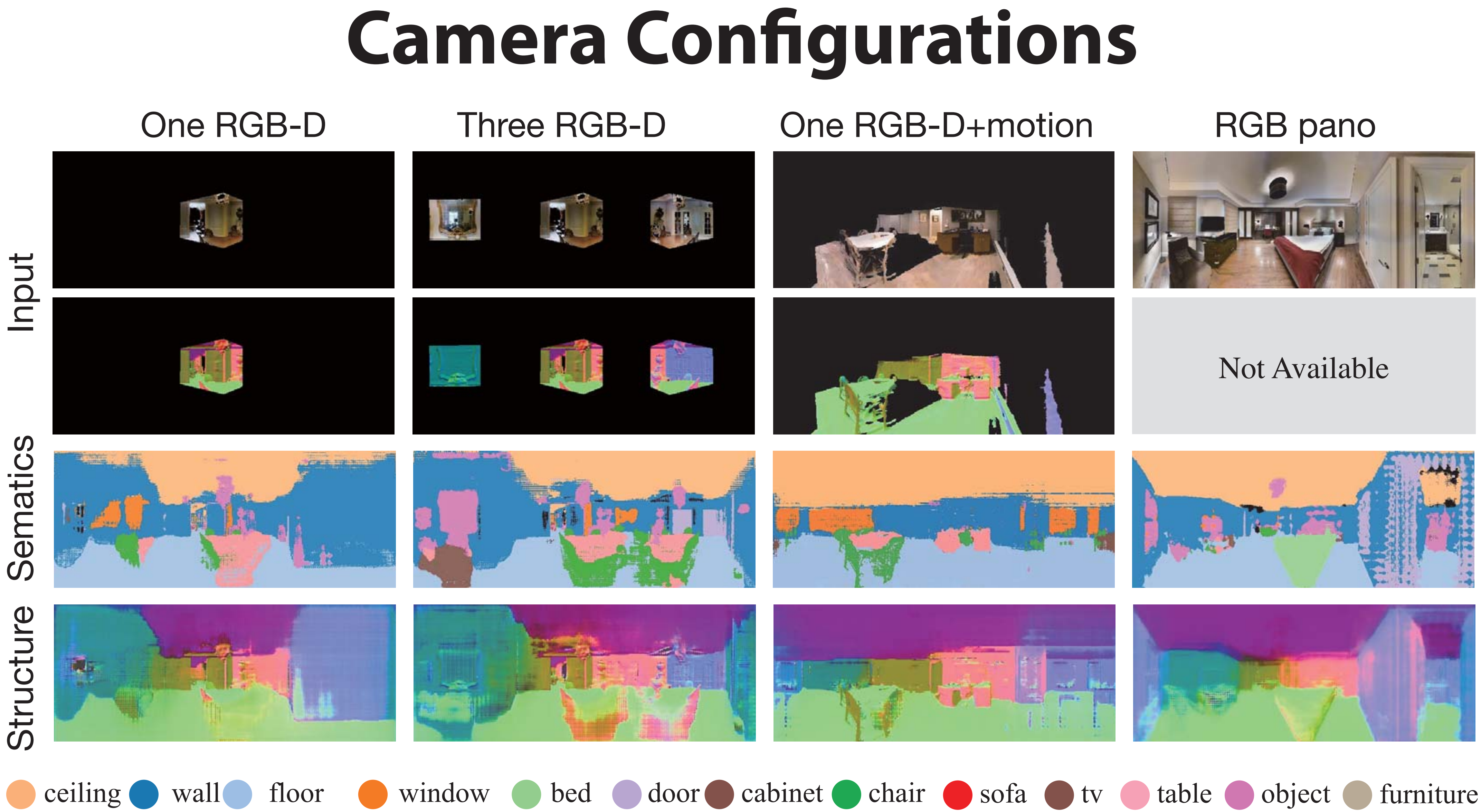
ceiling wall floor window bed door cabinet chair sofa tv table object furniture

# Results

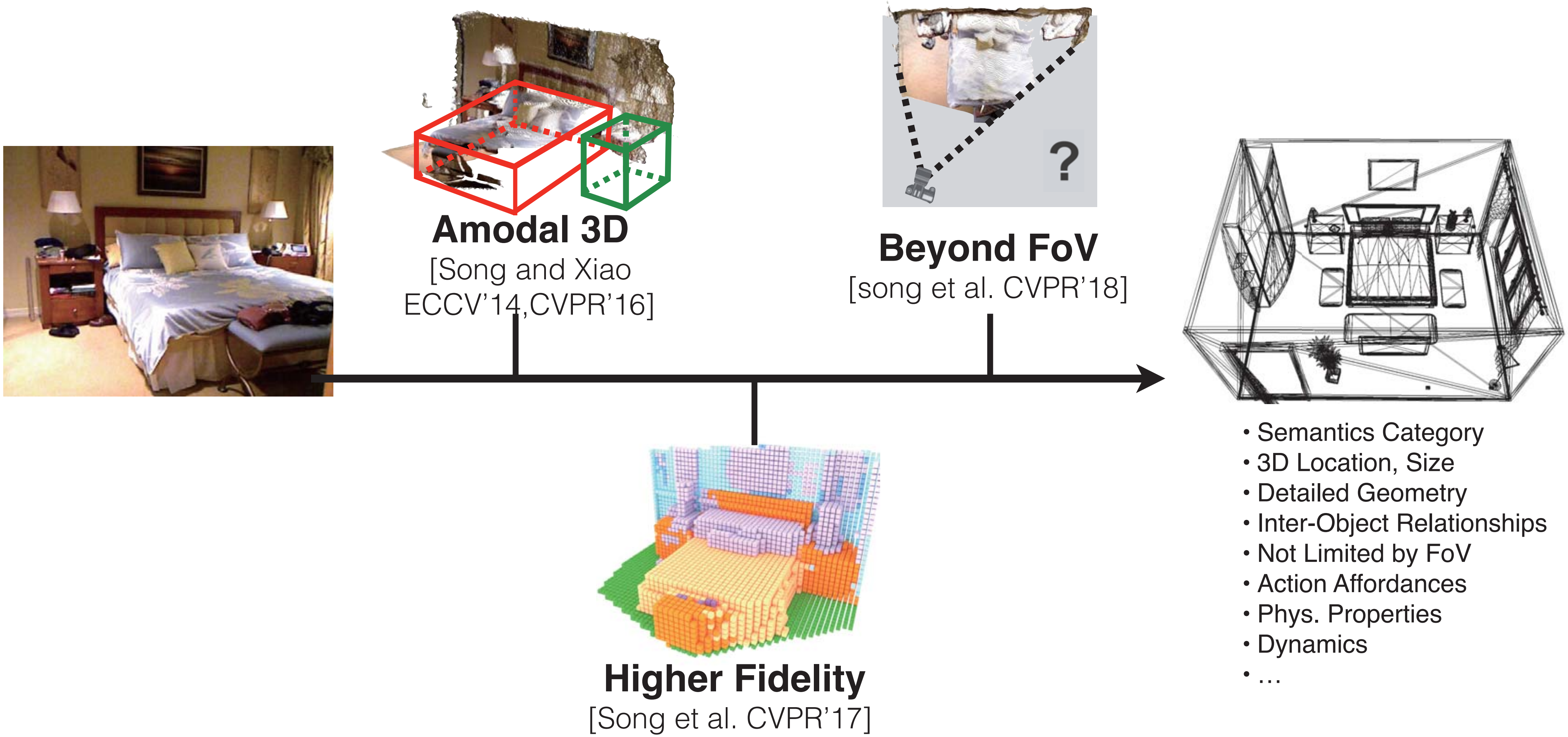

Prediction
Ground truth
door
sofa
window
objects
window
Bed
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Camera Configurations in real platforms

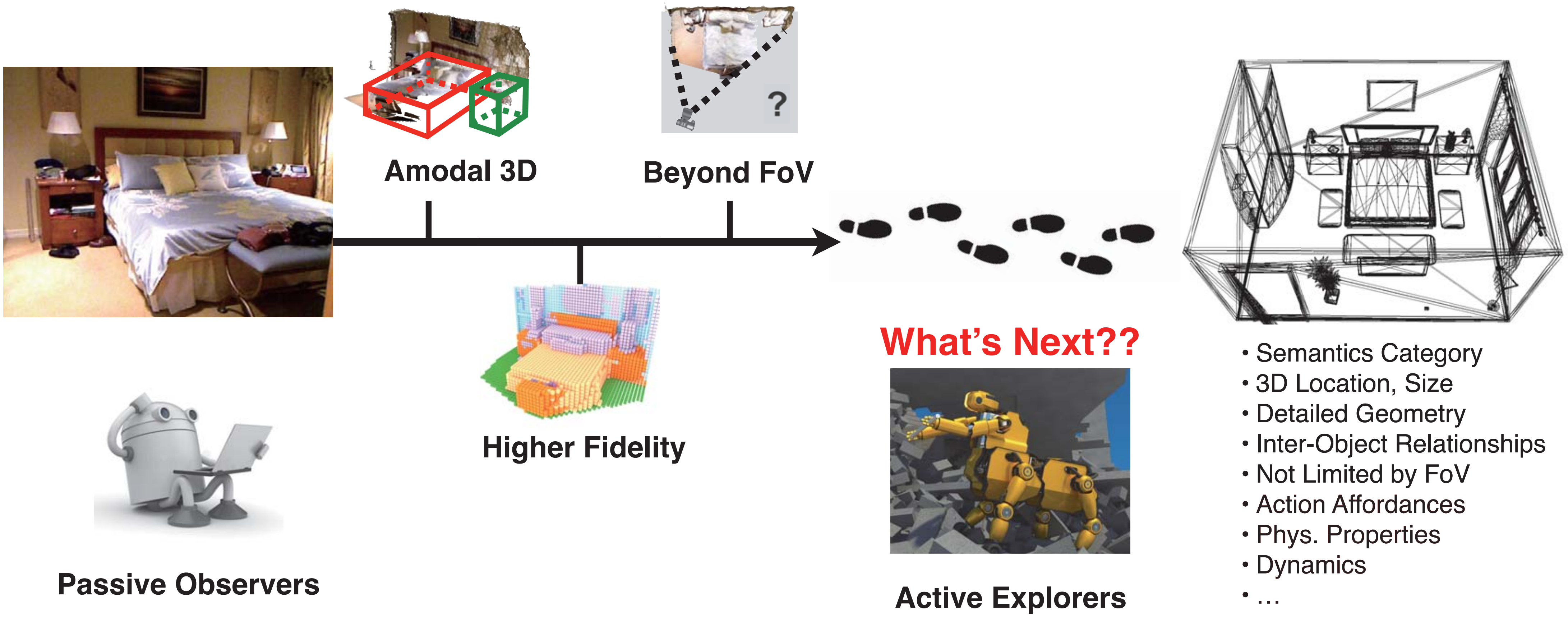



Camera Configurations
One RGB-D
Three RGB-D
One RGB-D+motion
RGB pano
Input
Not Available
Sematics
Structure
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Advances Towards 3D Scene Understanding

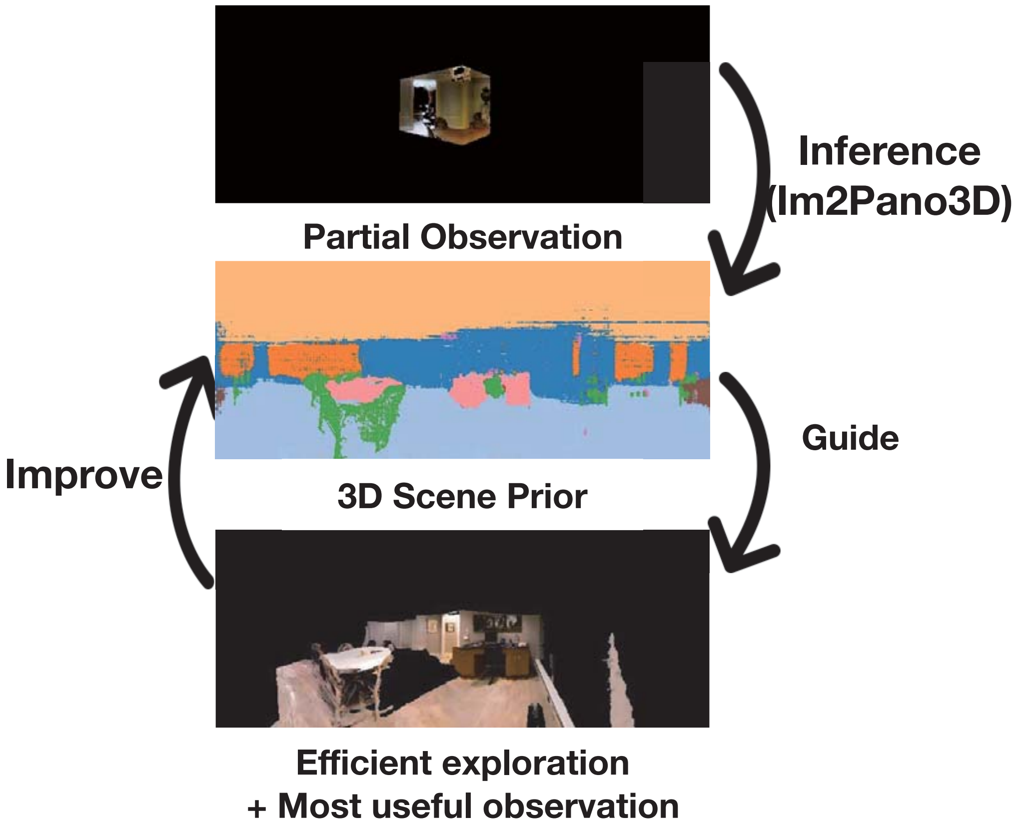


# Advances Towards 3D Scene Understanding



# Richer Representation through Interaction

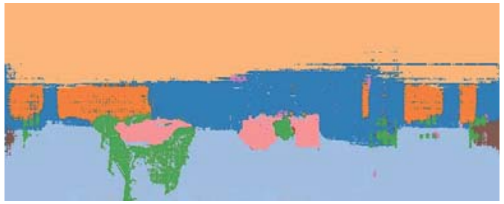
## Active Exploration




# Richer Representation through Interaction

## Active Exploration

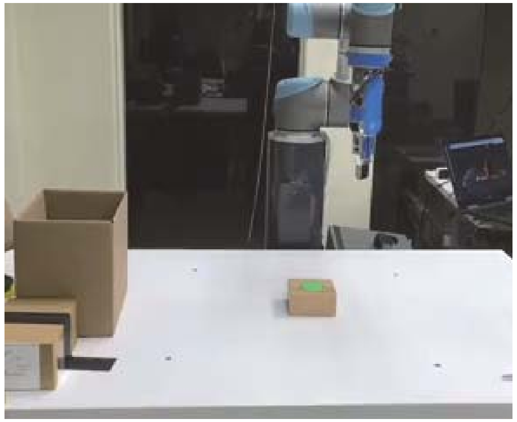
Partial Observation



3D Scene Prior

Efficient exploration
+ Most useful observation

## Active physical Interaction



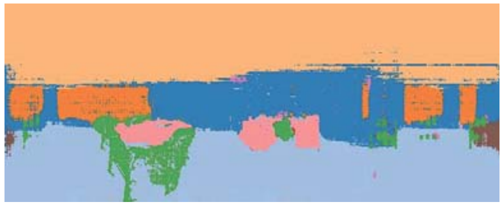
**Actions**: Poking,Grasping

**Physical properties:**
Surface material
Friction coefficient

# Richer Representation through Interaction

## Active Exploration

**Partial Observation**



**3D Scene Prior**

**Efficient Exploration**
**+ Most useful observation**

## Active physical Interaction

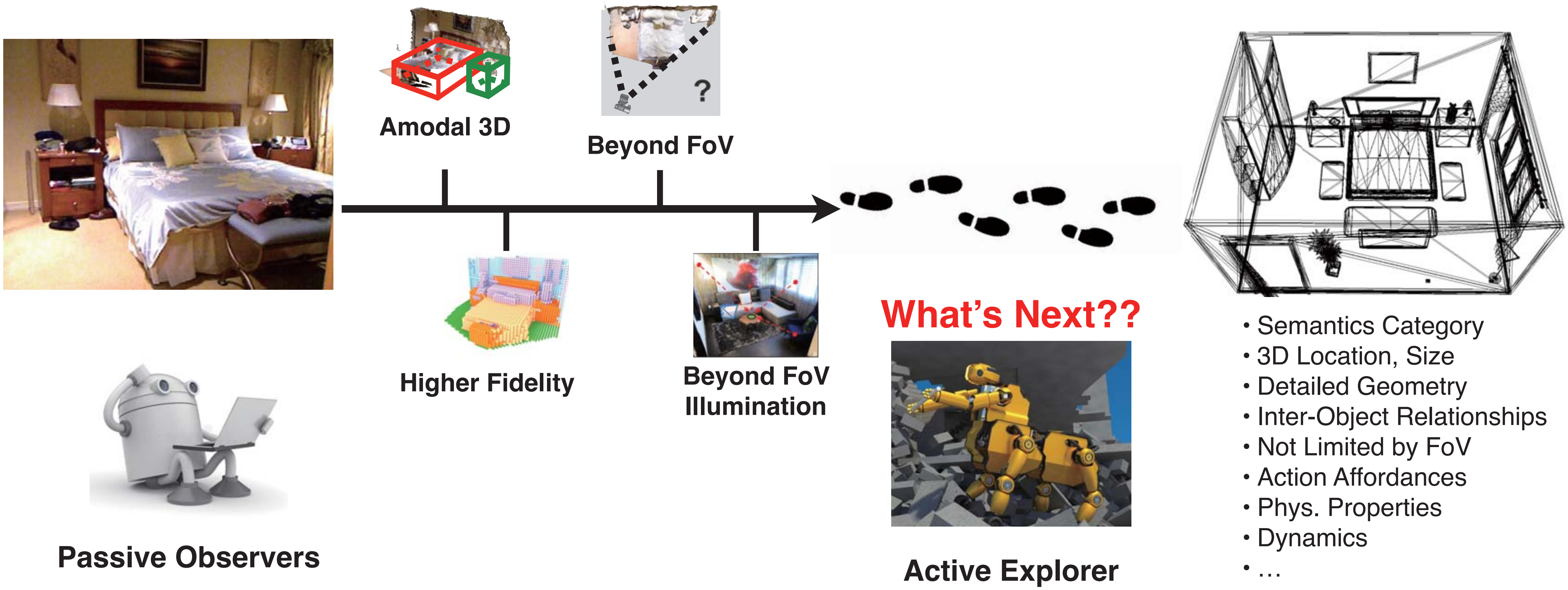
**Actions**: Pushing,Grasping

**Physical properties:**
Surface material
Friction coefficient

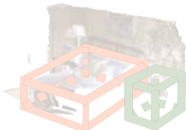
**Actions**: Tossing

**Physical properties:**
Mass distribution,
Aerodynamic

# Comprehensive 3D Scene Understanding



# Comprehensive 3D Scene Understanding



Amodal 3D

Beyond FoV

**We are looking for PhDs and Post-docs!**

Columbia University
IN THE CITY OF NEW YORK

Higher Fidelity

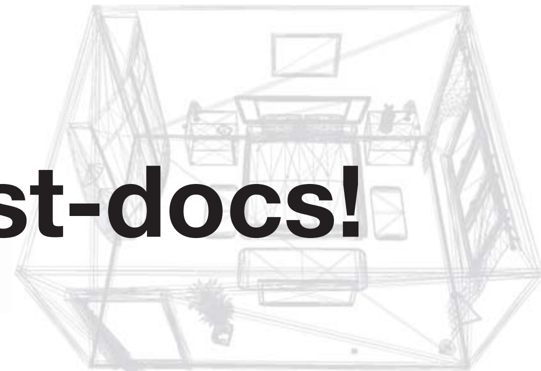
Illumination



Passive Observers



Active Explorer



- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. Properties
- Dynamics
- …

# Acknowledgements

## Collaborators

Ferran Alet
Maria Bauza
Angel Chang
Nikhil Chavan Dafle
Elliott Donlon
Nima Fazeli
Matthew Fisher
Thomas Funkhouser
Druck Green
Leonidas Guibas

Pat Hanrahan
Francois R. Hogan
Rachel Holladay
Qixing Huang
Hailin Jin
Joon-Young Lee
Zimo Li
Melody Liu
Weber Liu
Daolin Ma

Isabella Morona
Prem Qu Nair
Matthias Nießner
Alberto Rodriguez
Eudald Romo
Silvio Savarese
Manolis Savva
Ari Seff
Hao Su
Orion Taylor

Ian Taylor
Zhirong Wu
Jianxiong Xiao
Li Yi
Kuan-Ting Yu
Fisher Yu
Ersin Yumer
Andy Zeng
Linguang Zhang
Yinda Zhang

**Funding:**
NSF, Google, Intel, Facebook


# Thank You!