PRINCETON VISION GROUP

Stanford University

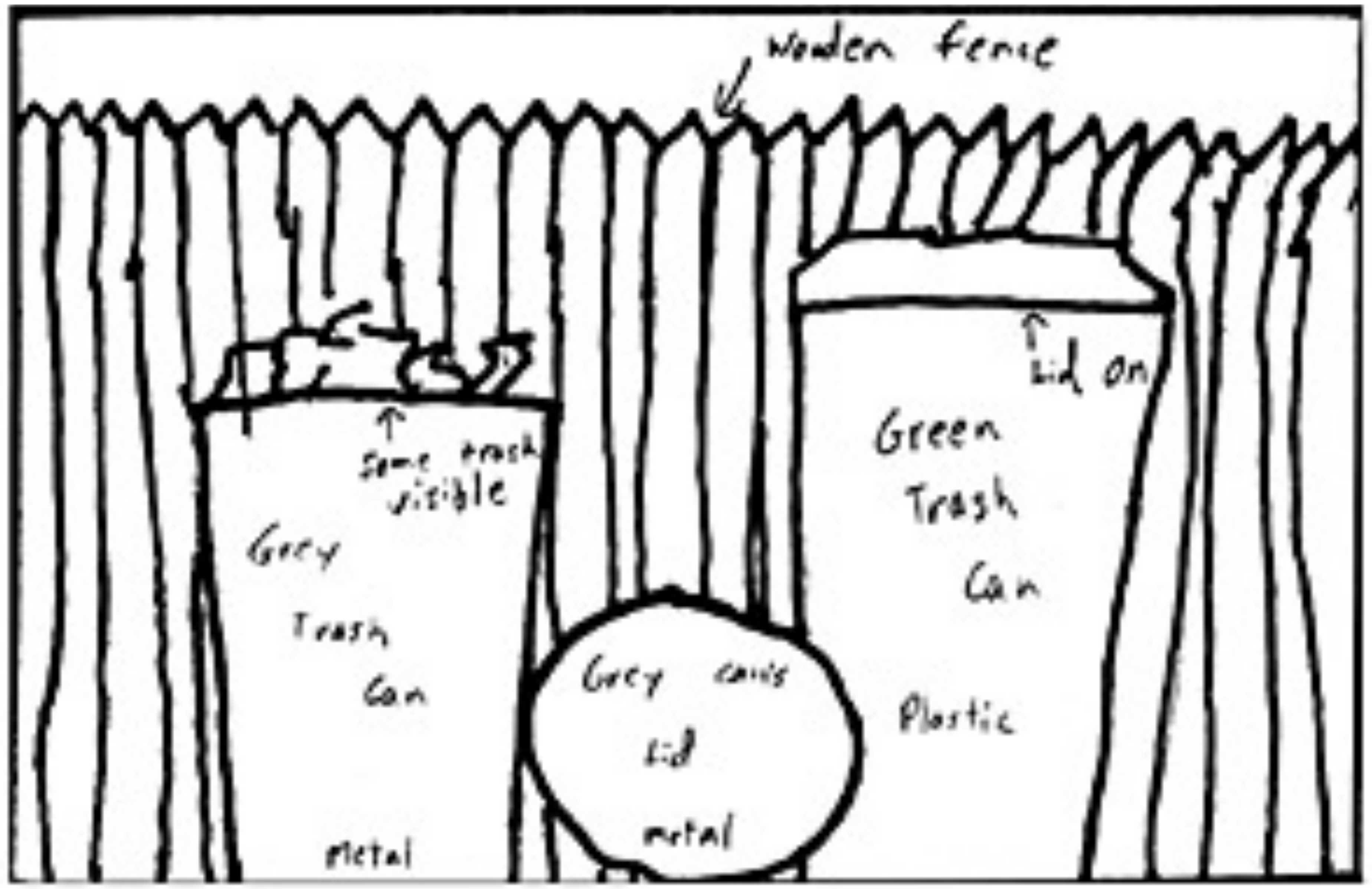
# Im2Pano3D:

## Extrapolating 360° Structure and Semantics Beyond the Field of View

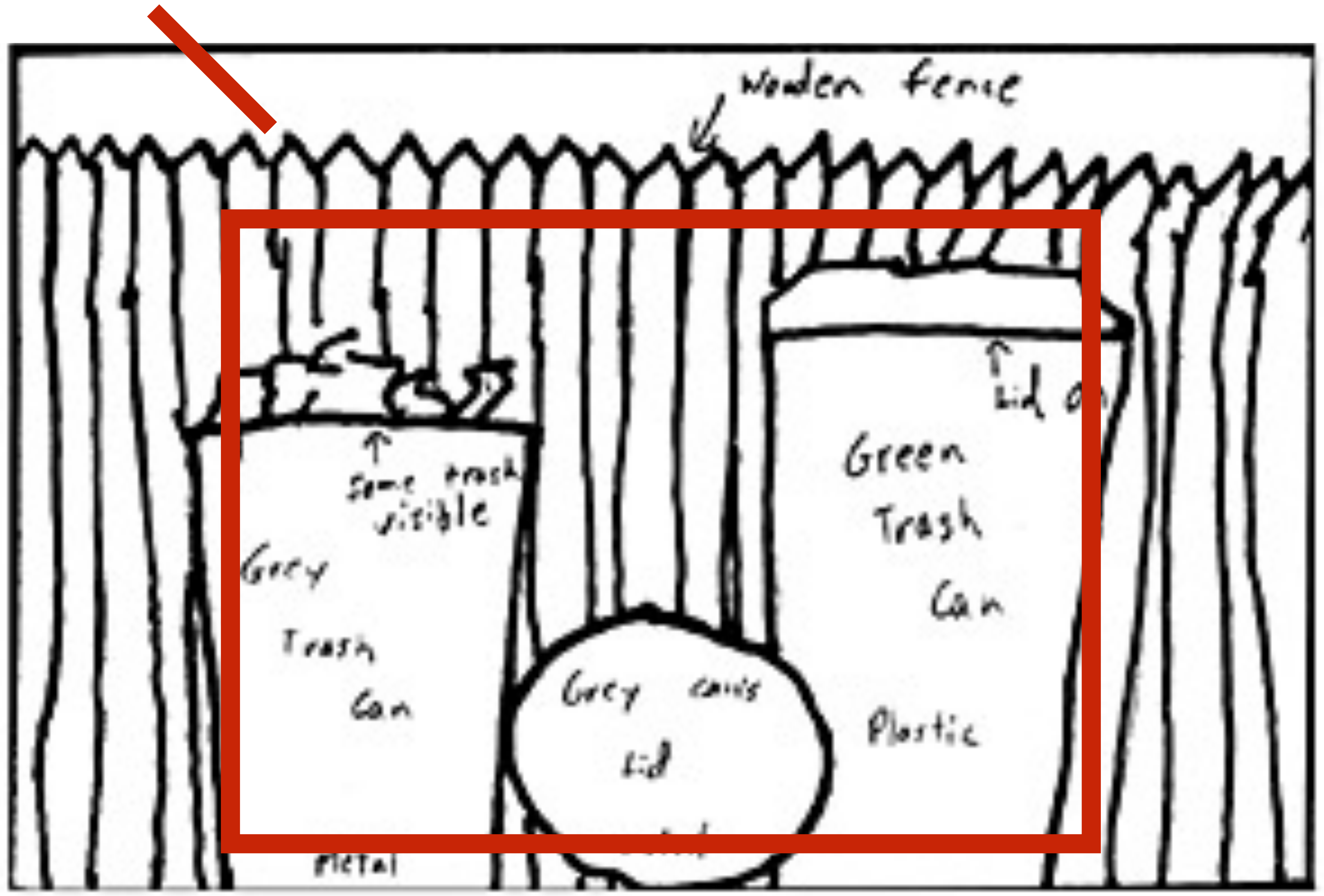
Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser

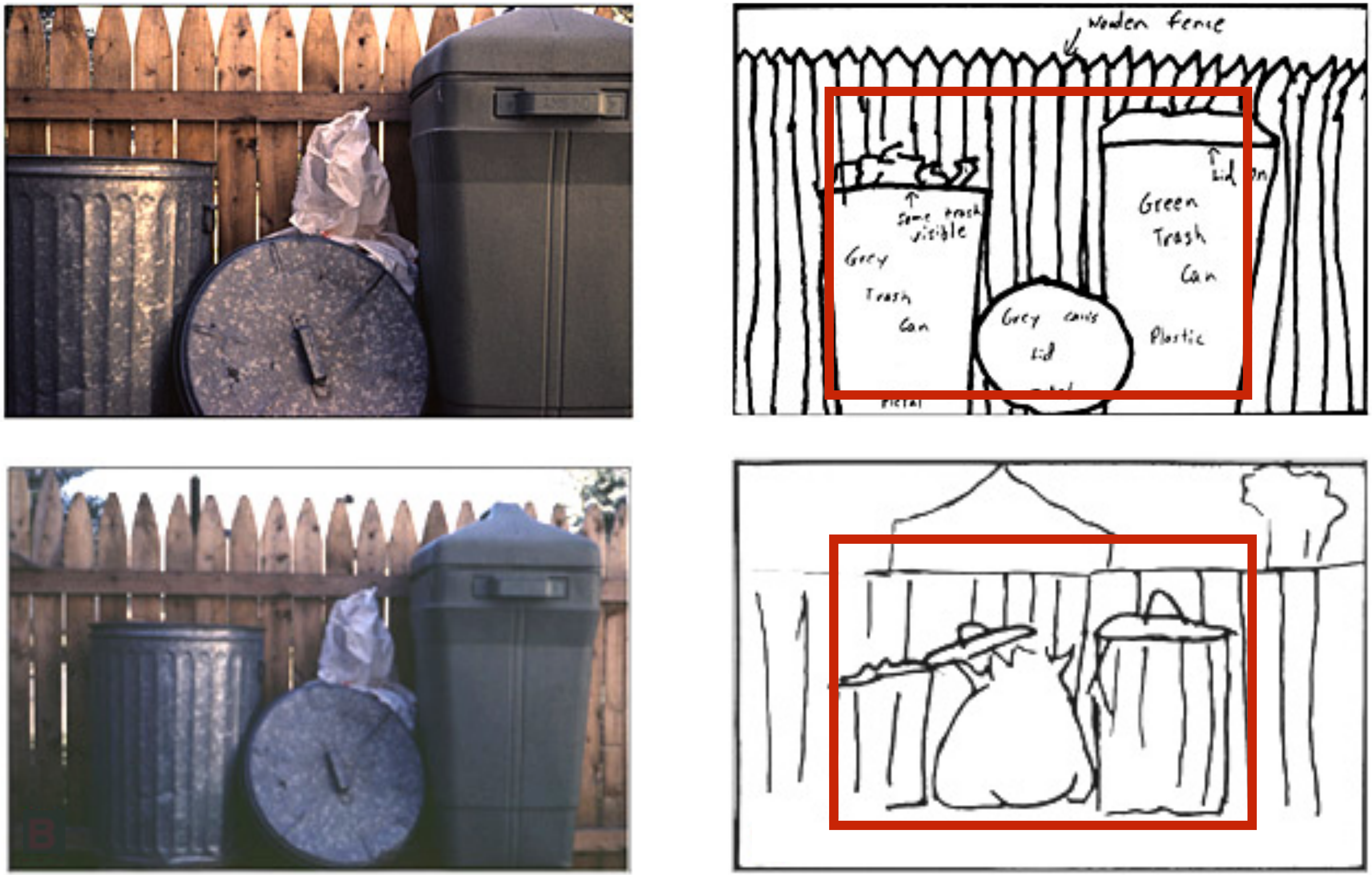
# Boundary Extension

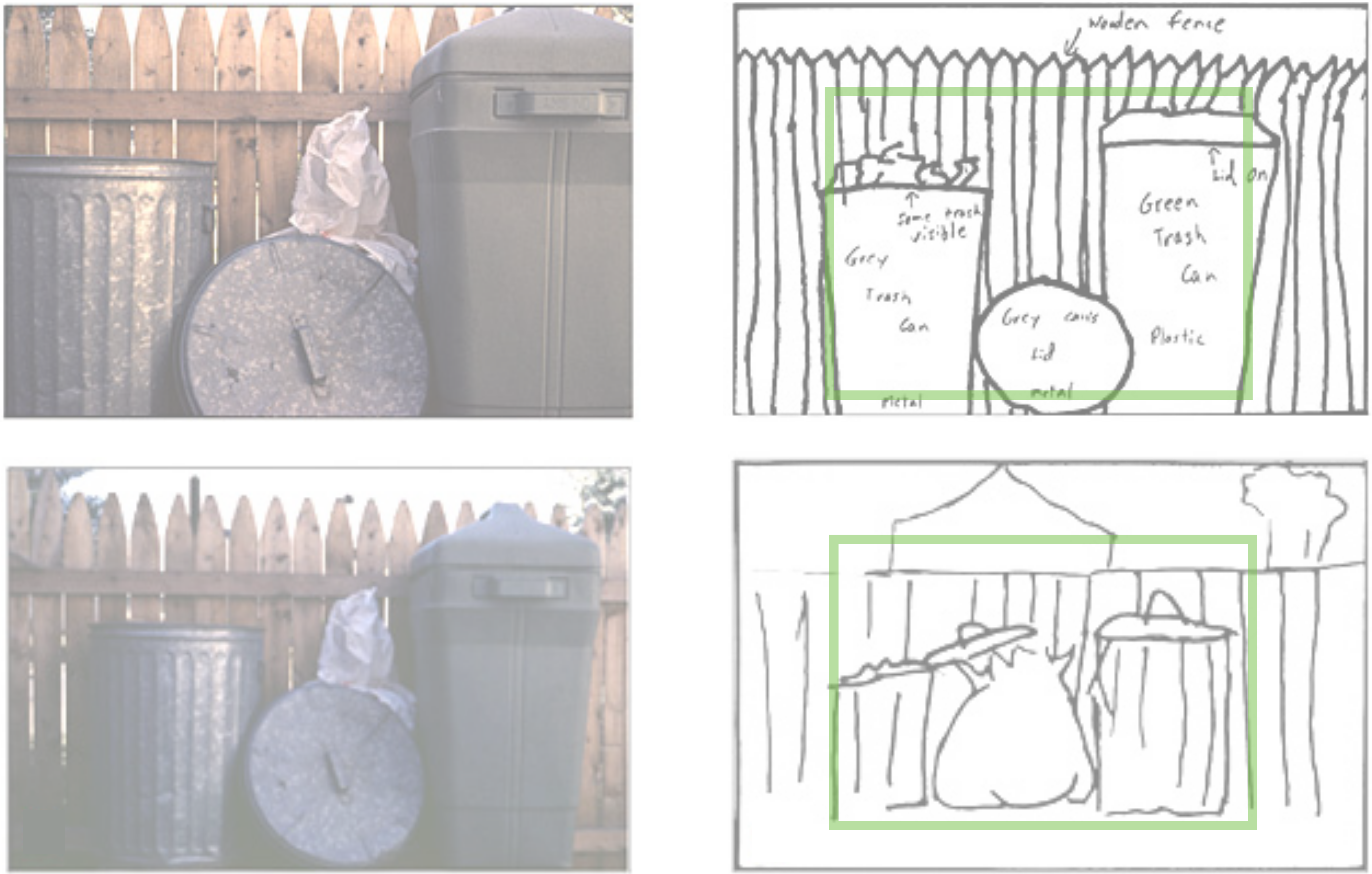
Intraub and Richardson, 1989


wooden fence
some trash visible
Grey Trash Can
metal
Grey cans lid
metal
lid on
Green Trash Can
Plastic

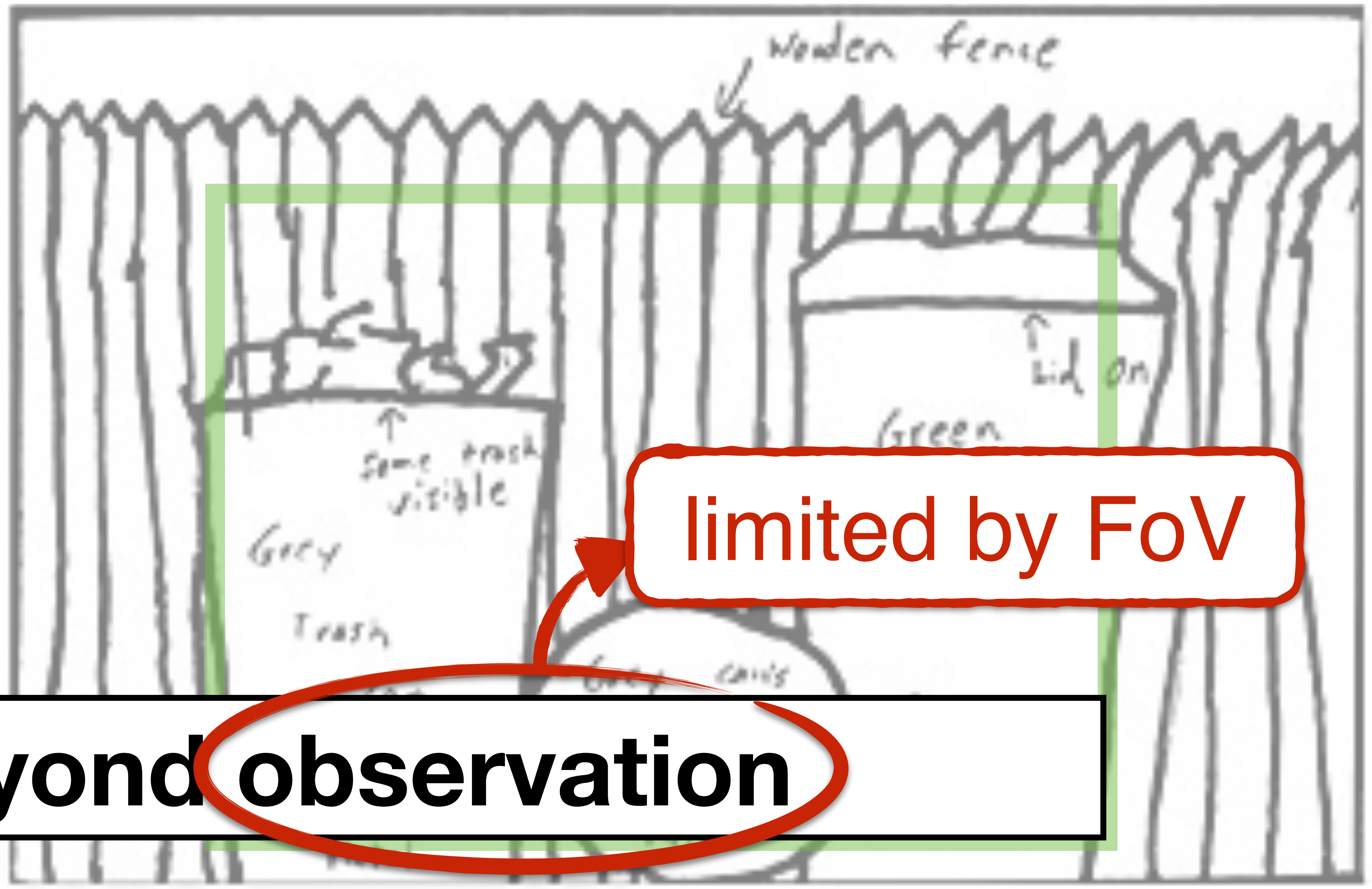
**20- 30 % More**


Woden fence
Some trash
visible
Grey
Trash
Can
Metal
Grey
cans
Lid
Green
Trash
Can
Plastic


Woden fence
lid on
Some trash visible
Grey Trash Can
Grey cans lid
Green Trash Can
Plastic

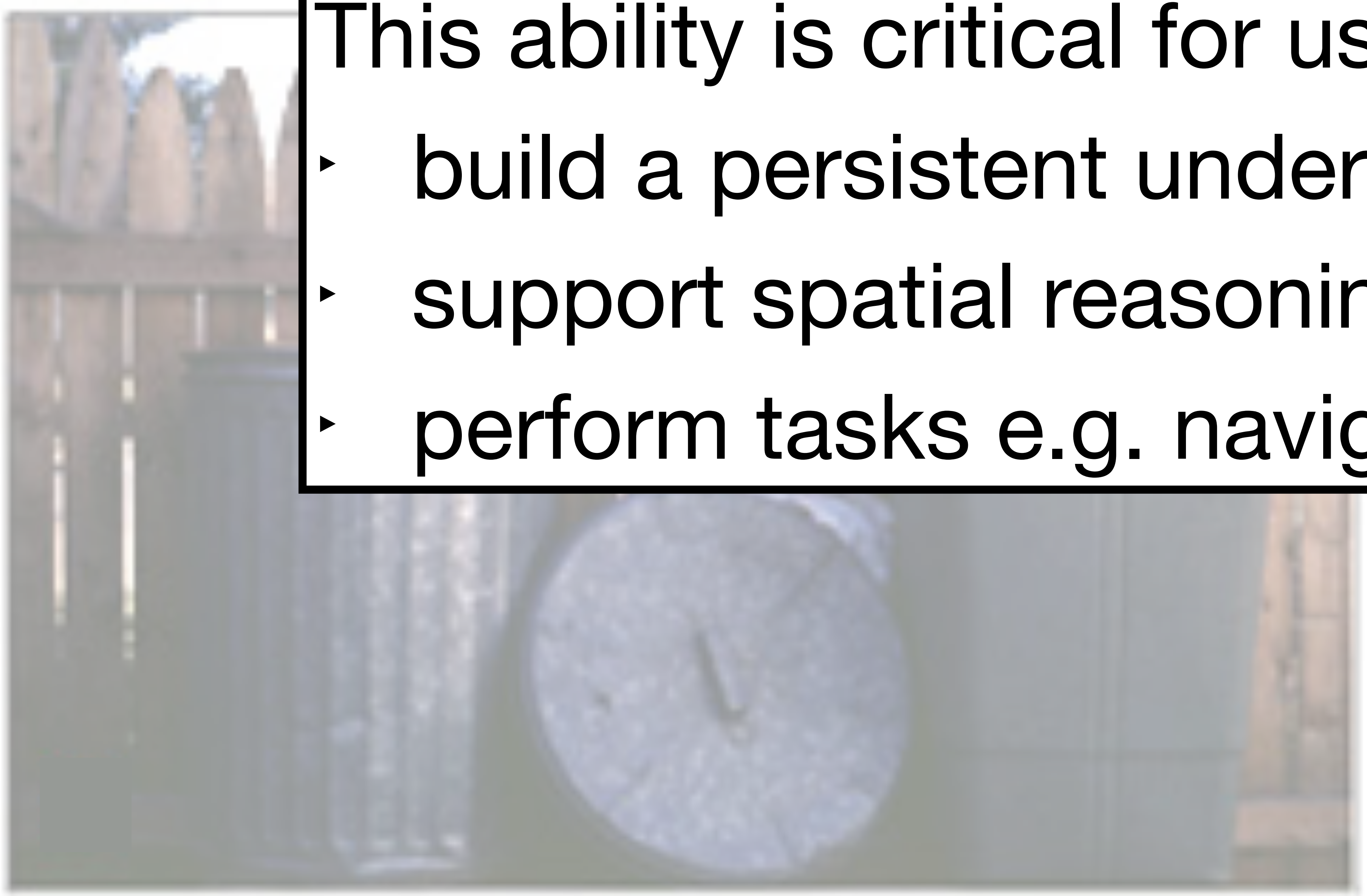


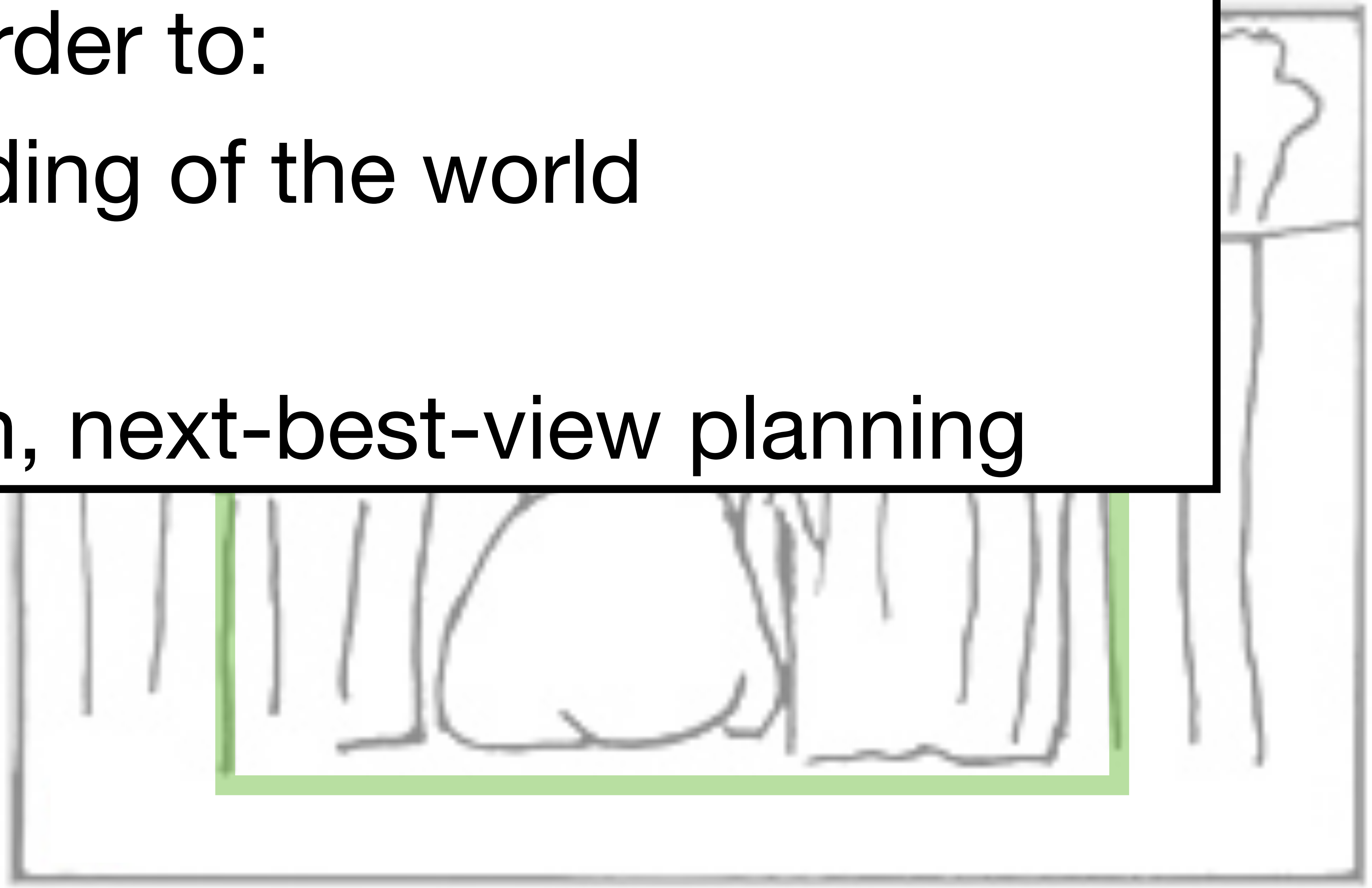

“False memory 1/20th of a second later: what the early onset of boundary extension reveals about perception.” Intraub and Dickinson




limited by FoV

**Infer the context beyond observation**

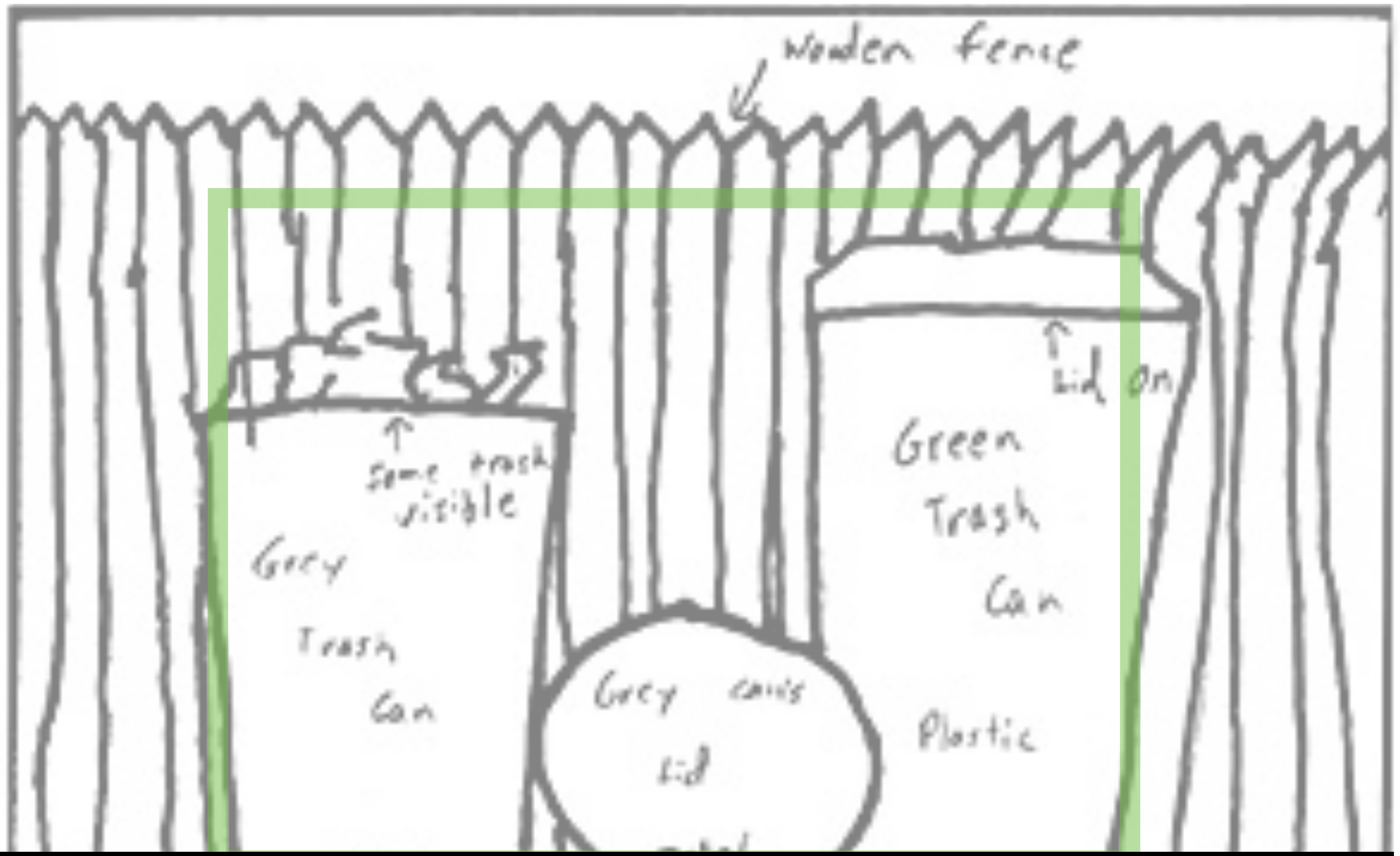




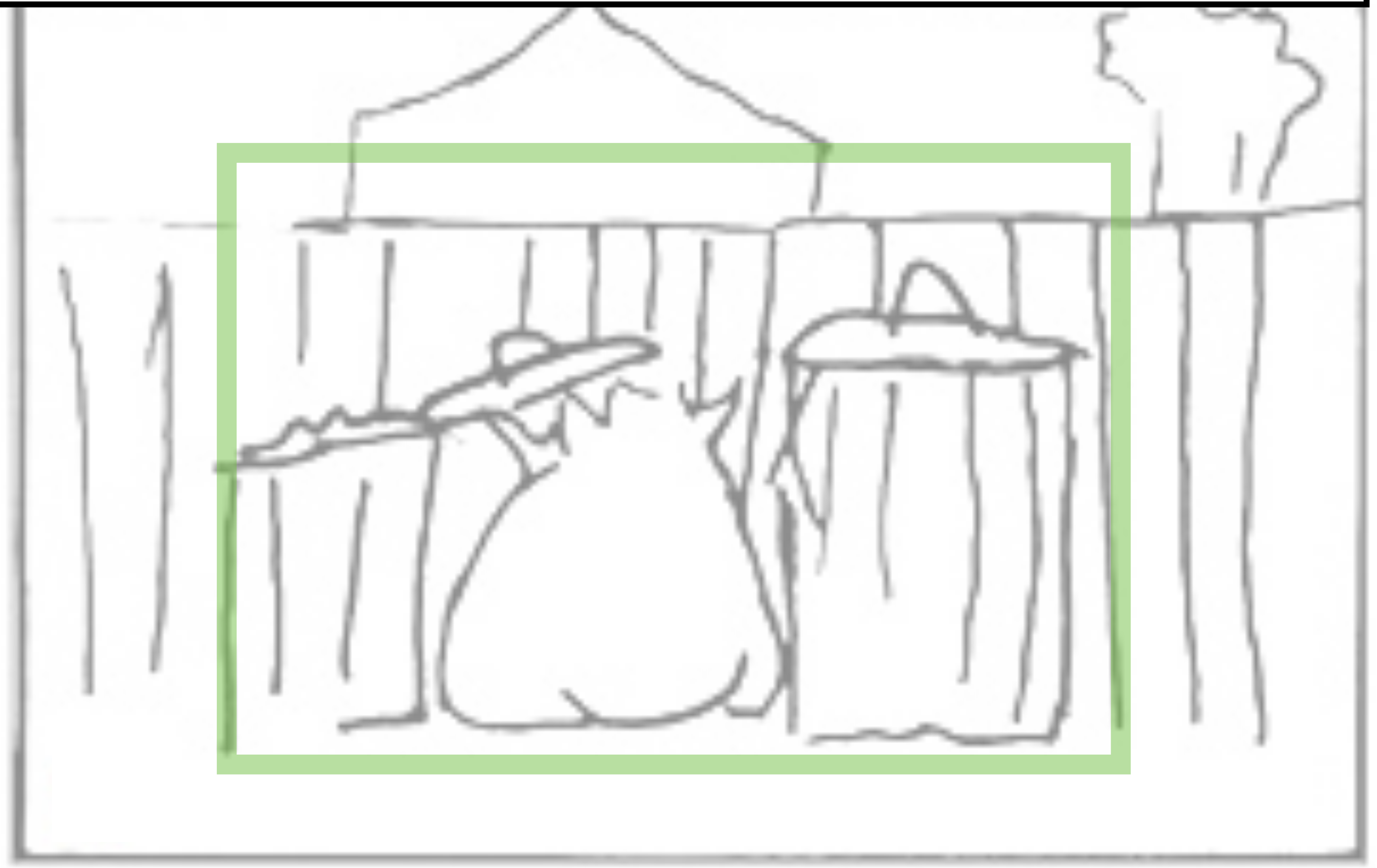
This ability is critical for us in order to:

- build a persistent understanding of the world
- support spatial reasoning
- perform tasks e.g. navigation, next-best-view planning

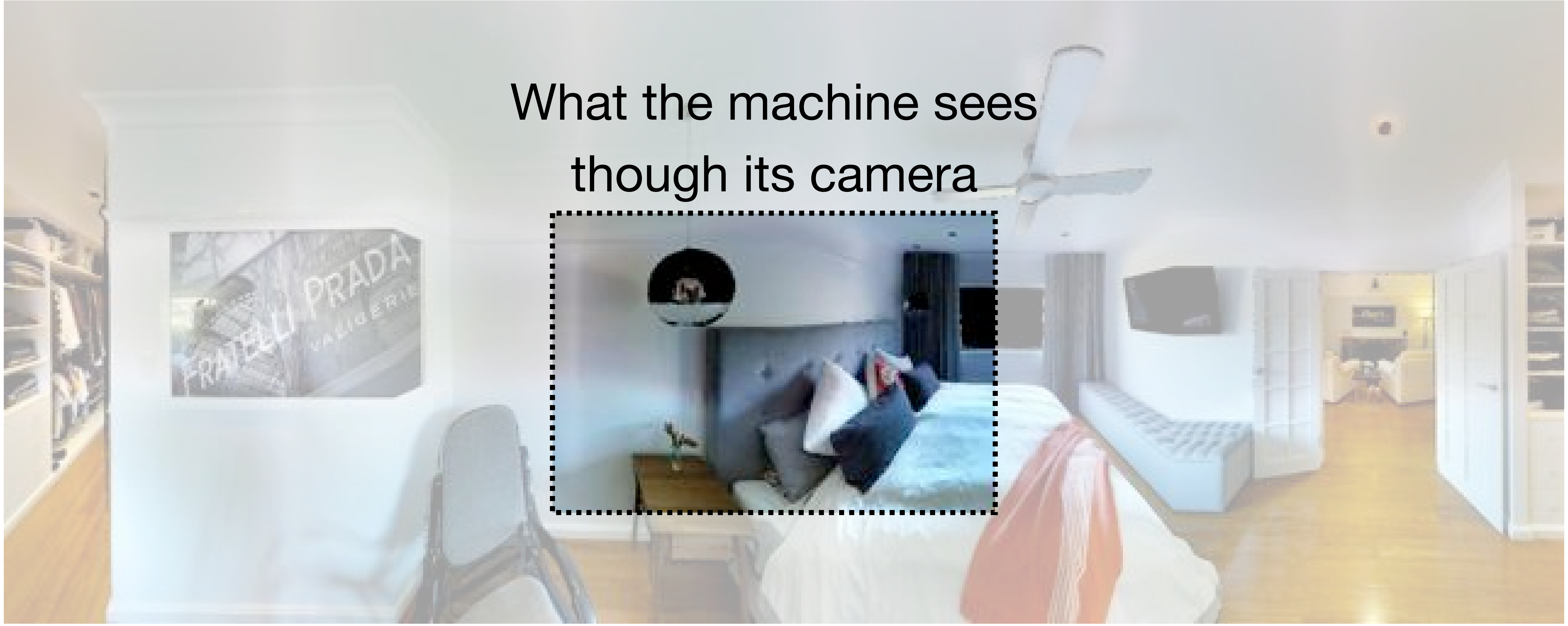
“False memory 1/20th of a second later: what the early onset of boundary extension reveals about perception.” Intraub and Dickinson




**Can we enable machines to do the same?**



# View Extrapolation




Complete surrounding environment

# View Extrapolation

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

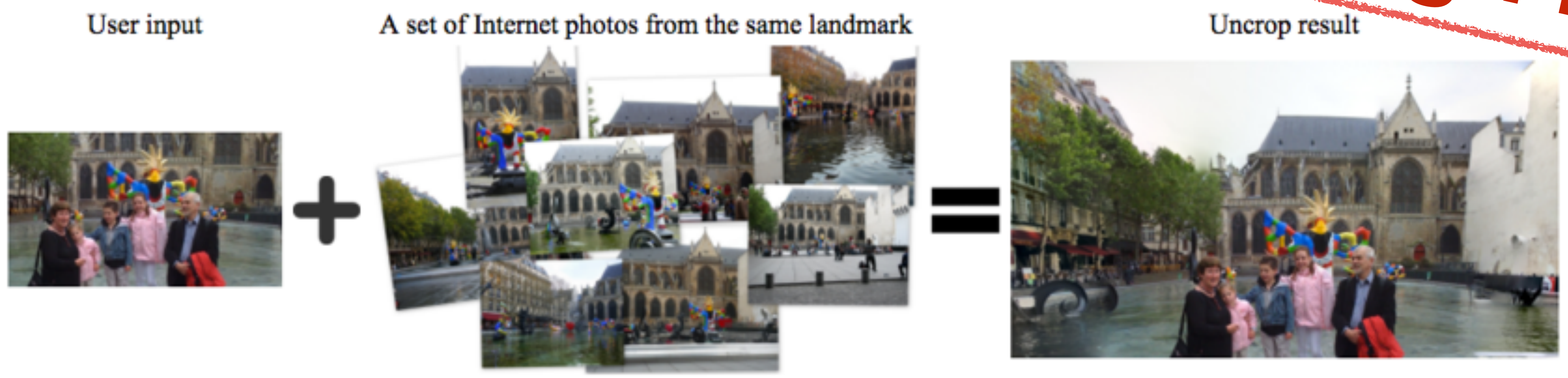
### Image Inpainting
[Pathak et. al]


(a) Input context

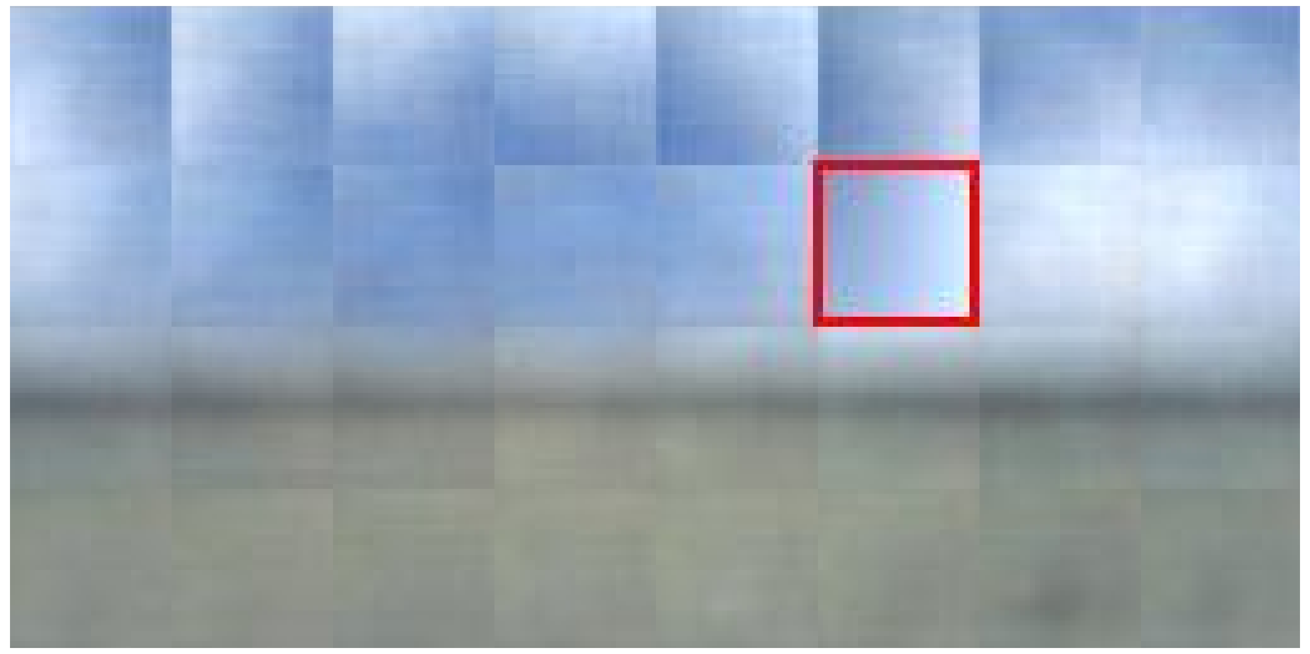

### User-guided view extrapolation [Zhang et al.]


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

### Stitching images from the Inte




### Learning to Look Around
[Jayaraman and Grauman]



# View Extrapolation

Input: RGB-D images

Output1: 3D Structures


Output2: Semantics

# View Extrapolation

Where can I move?

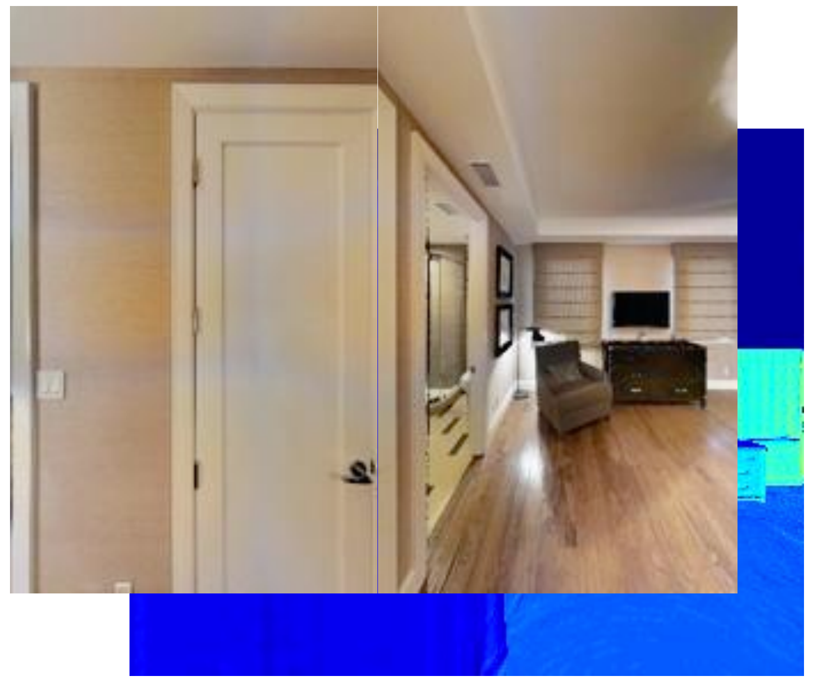
Output1: 3D Structures

Where should I turn to find a door?


Output2: Semantics

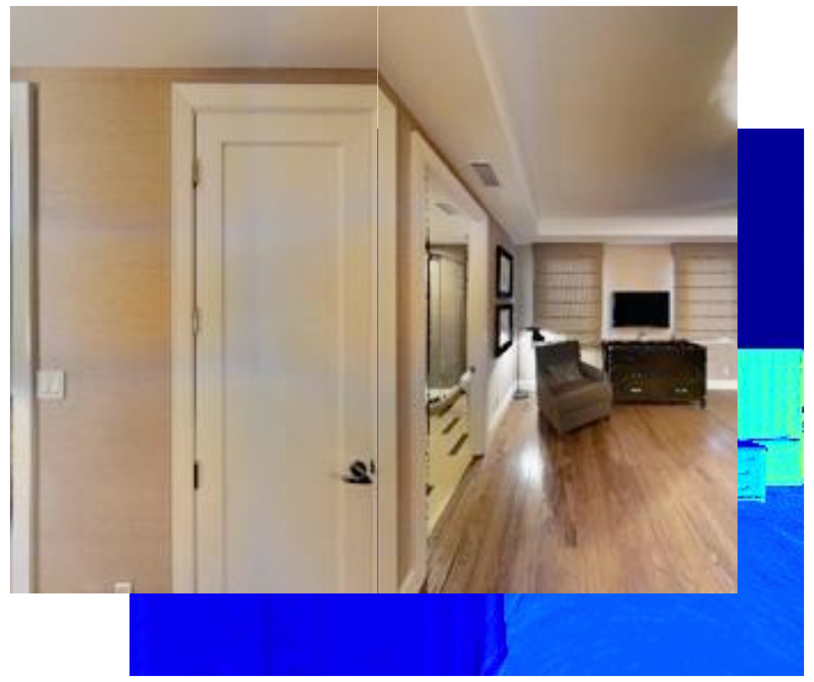
# Semantic-Structure View Extrapolation

**Input:** RGB-D images

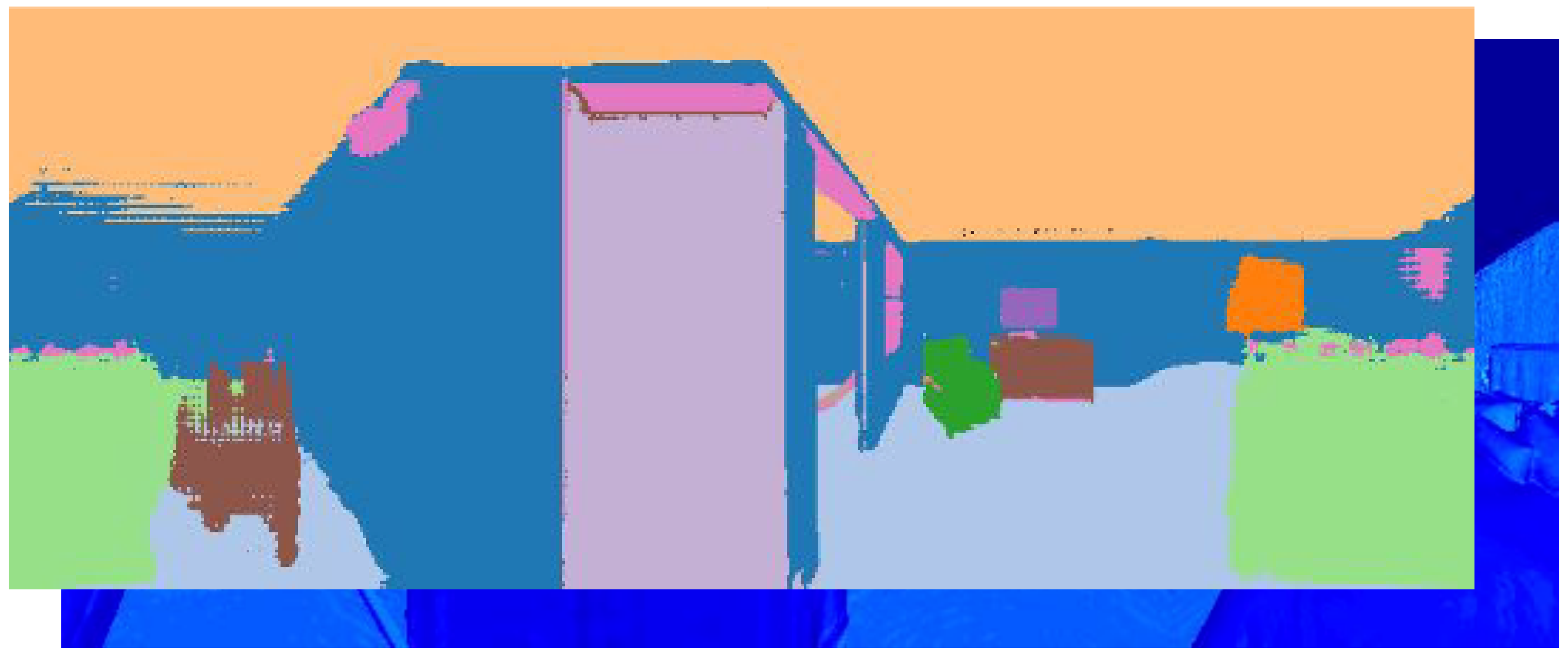


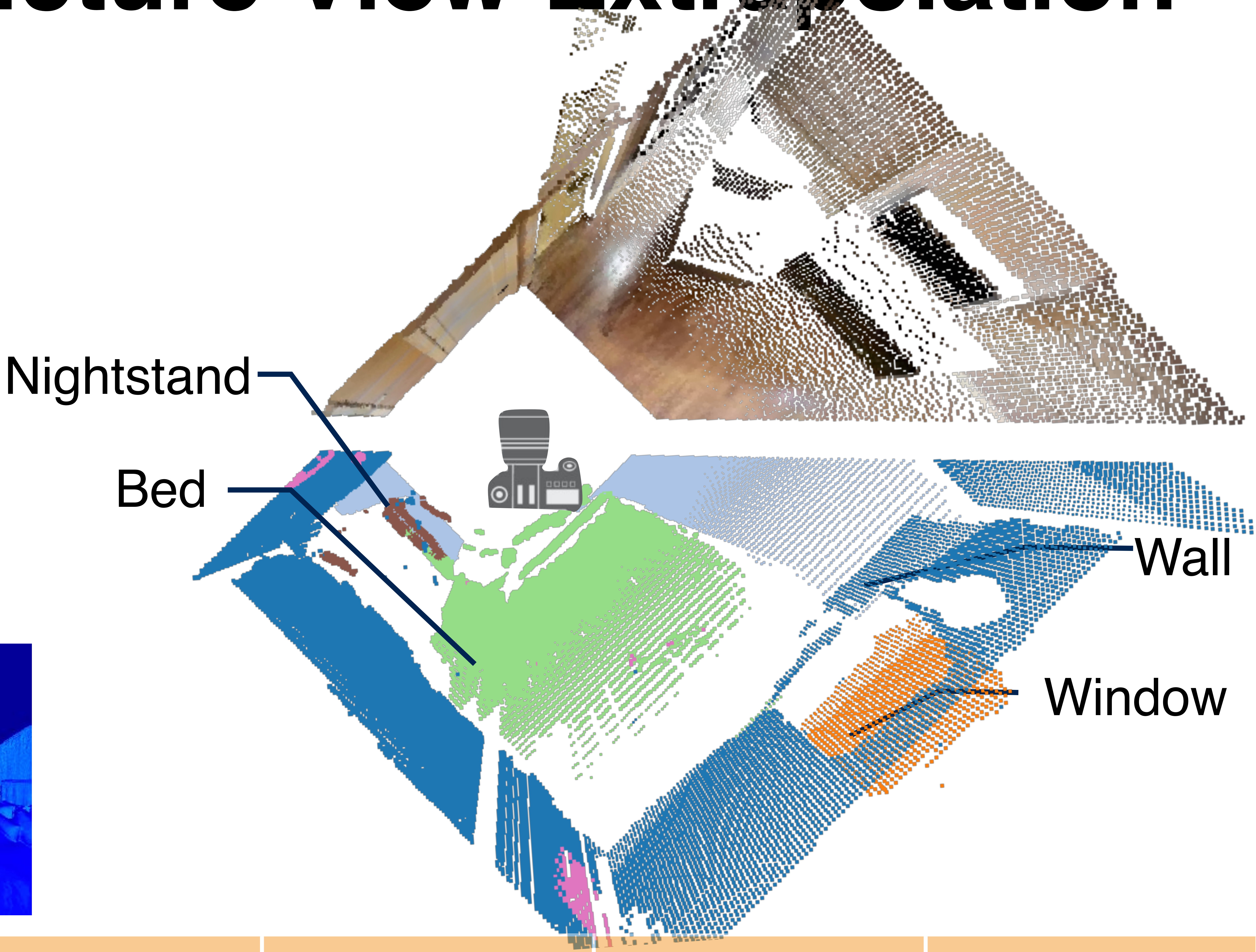
# Semantic-Structure View Extrapolation

**Input:** RGB-D images



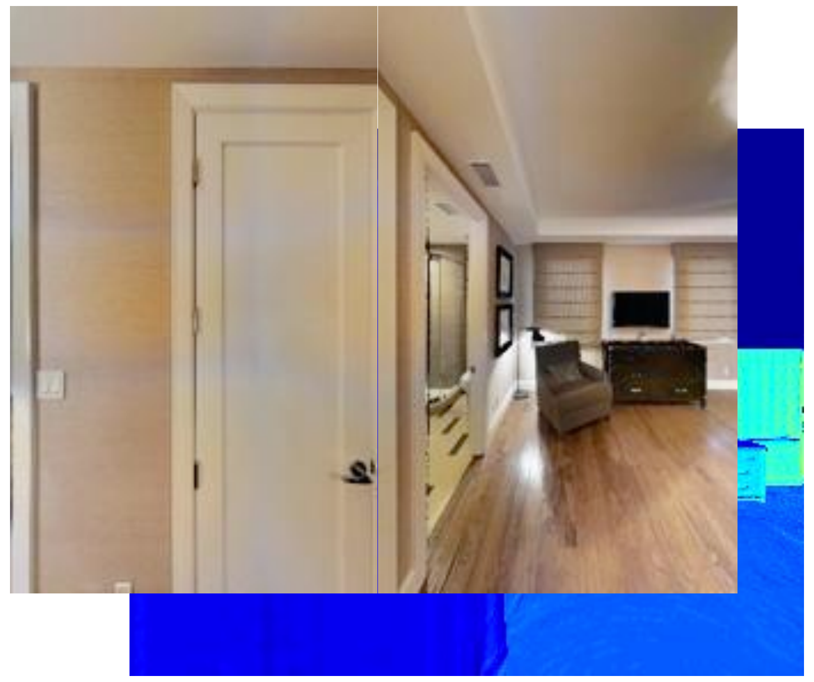
**Output:** 360° panorama
with 3D structure & semantics



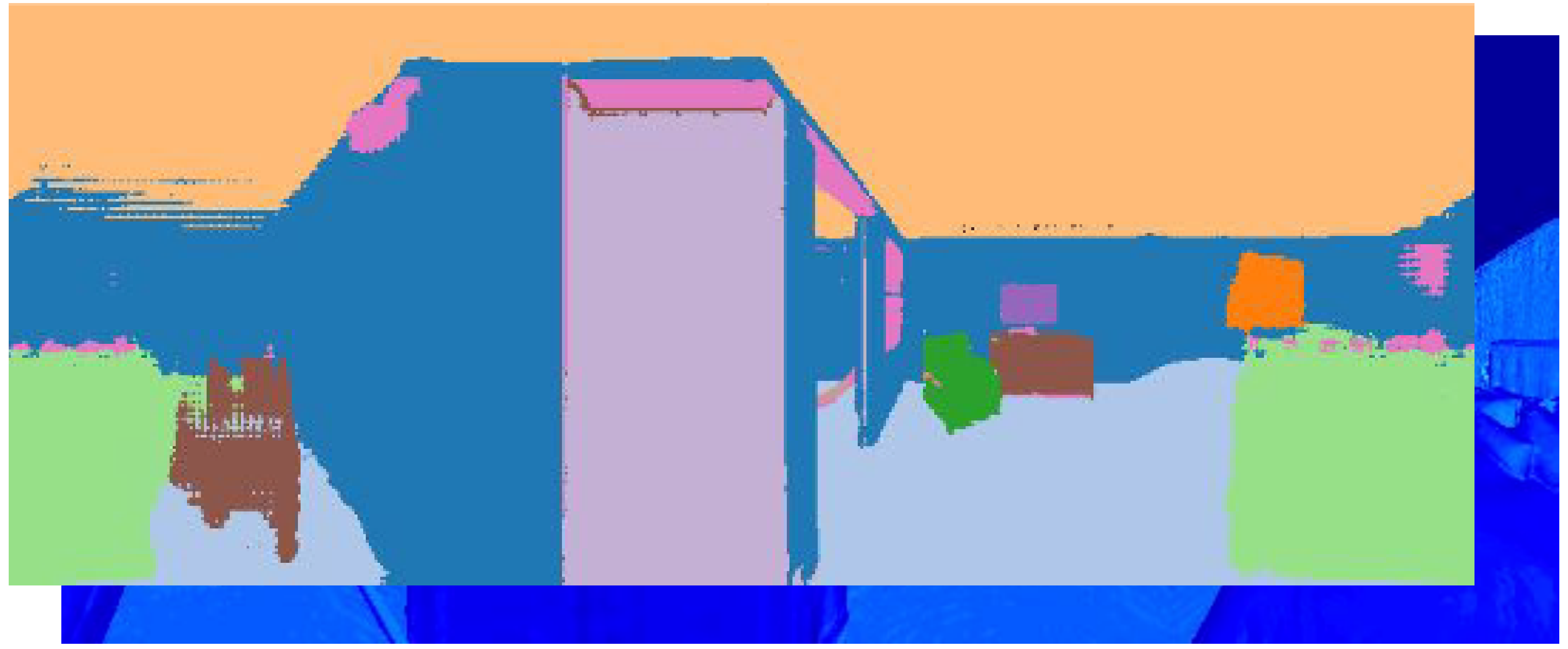


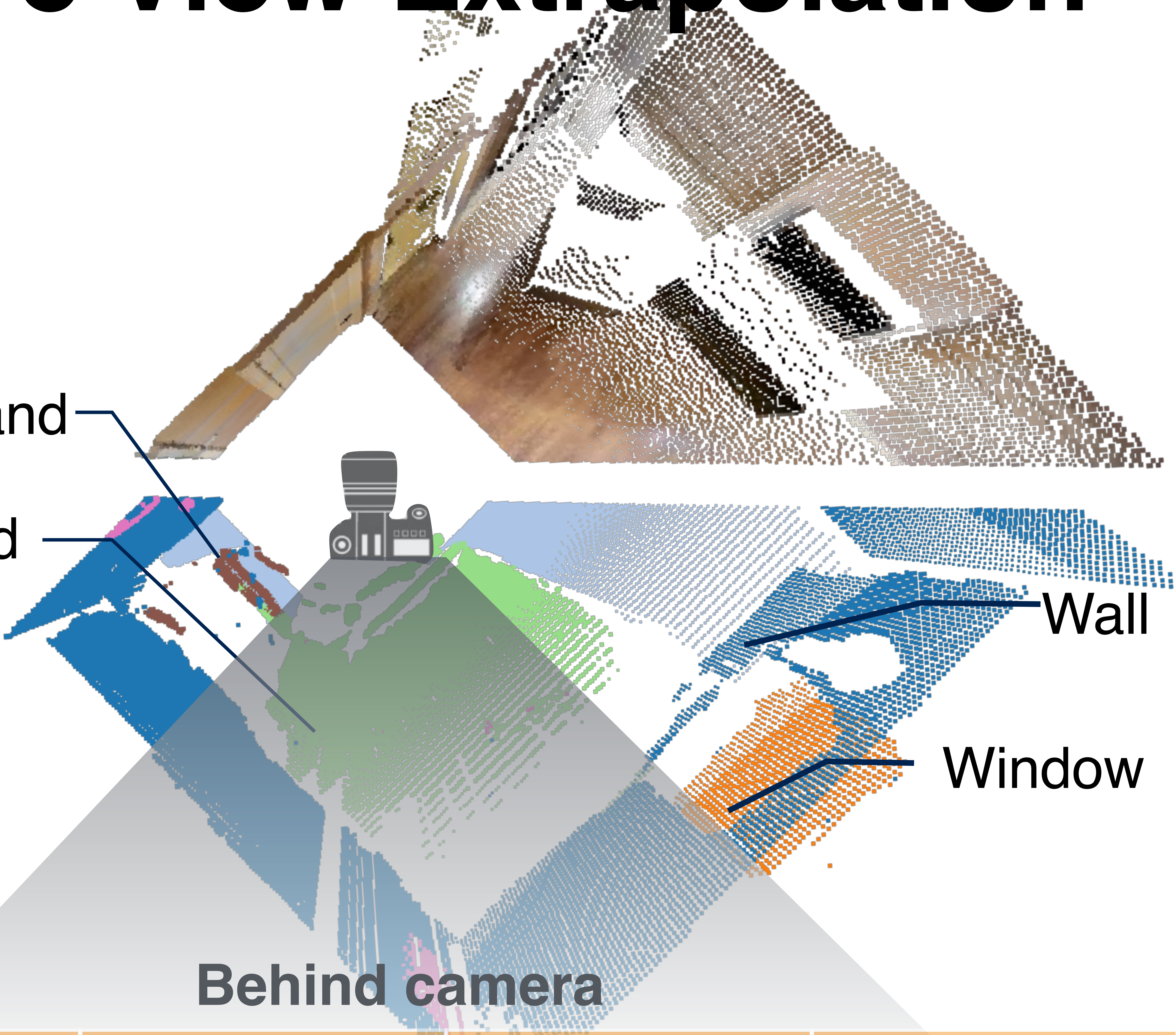
# Semantic-Structure View Extrapolation

**Input:** RGB-D images



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





# 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.

Data of indoor environments → Im2Pano3D Network → 3D Structure + Semantic

# Challenges

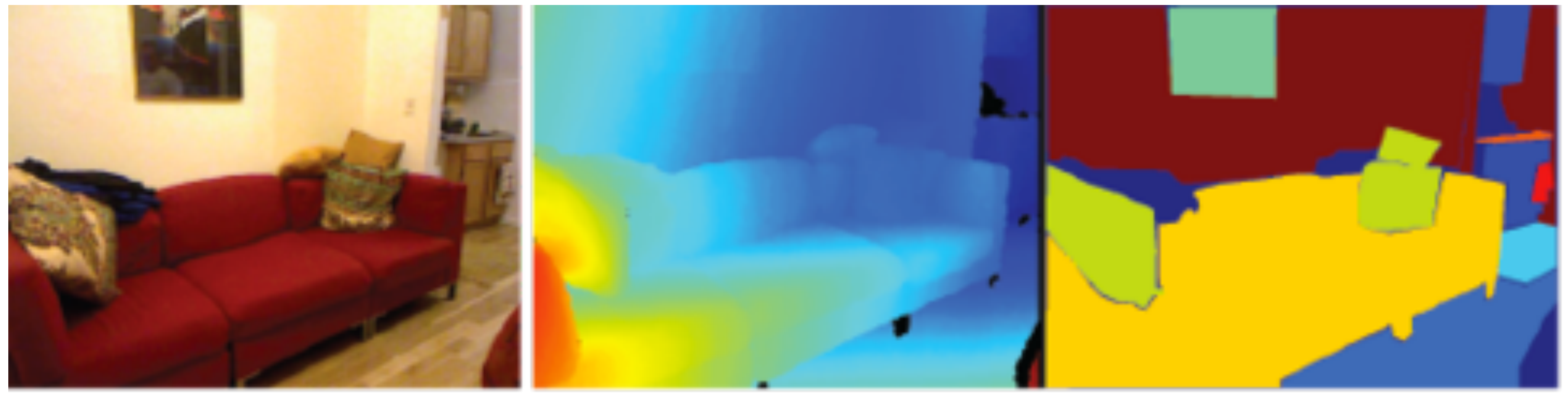
- How to obtain a large of amount training data?
- How to represent the 3D structure?
- How to provide meaningful supervision for training?

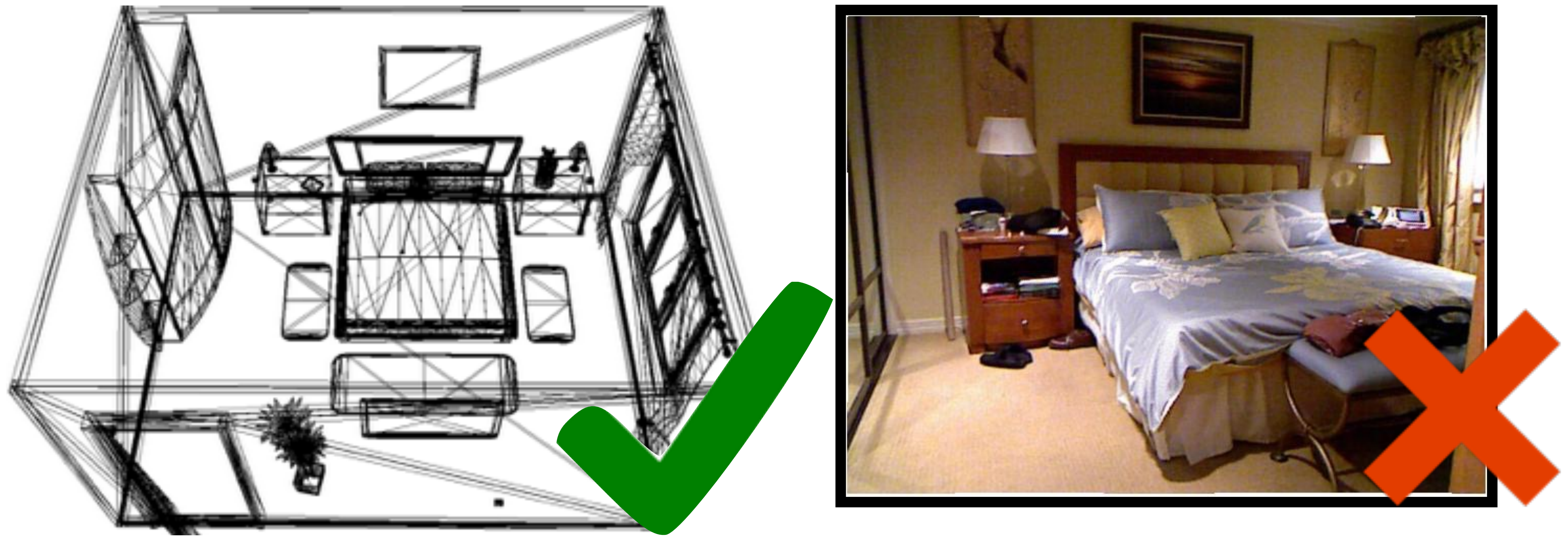
Data of indoor environments → Im2Pano3D Network → 3D Structure + Semantic

# Challenge 1: How to obtain the data?

**We need a dataset that has:**

- Semantic label.
- 3D structure information.
- Whole room context.
- Many examples.

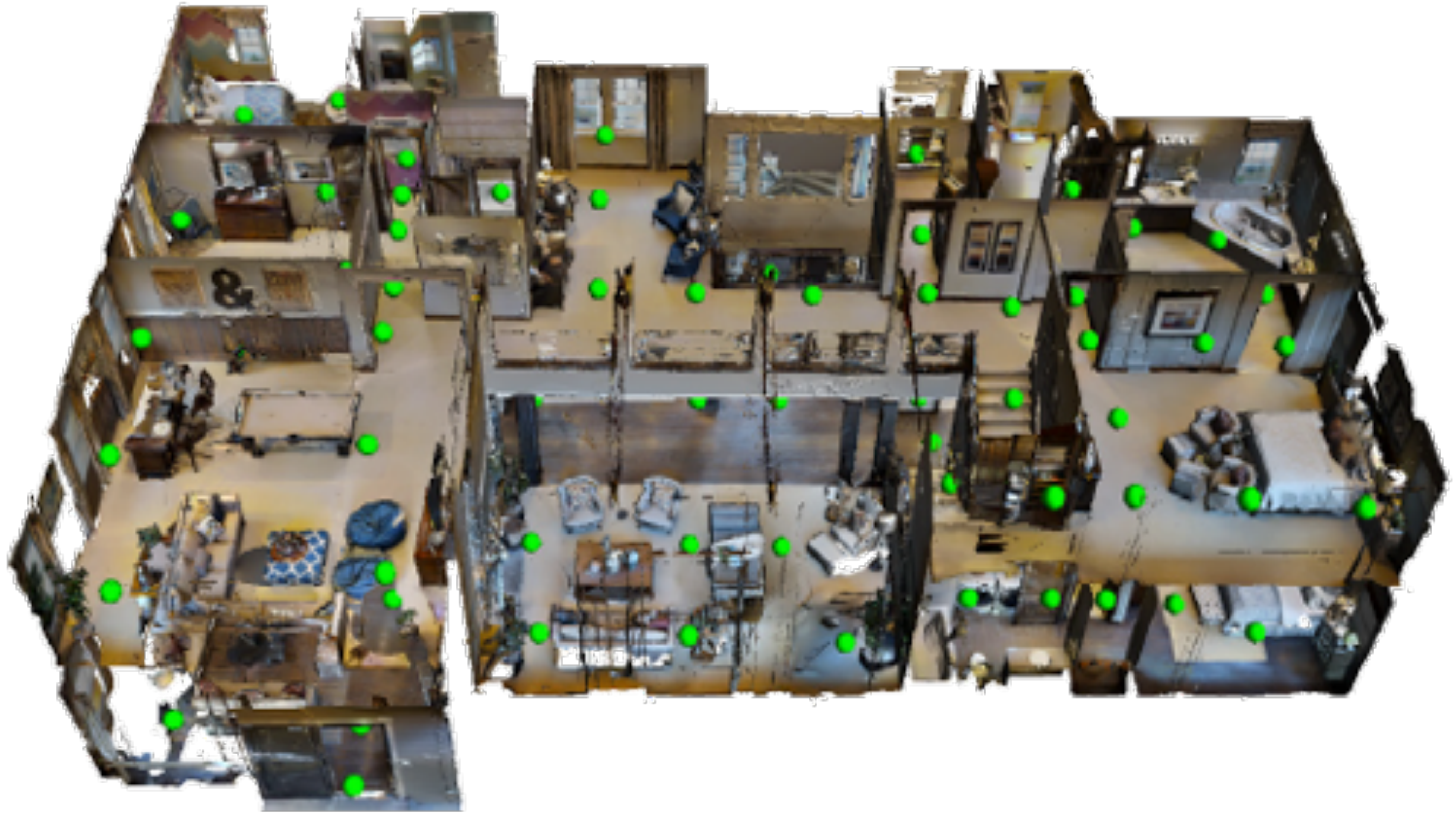




# Challenge 1: How to obtain the data?

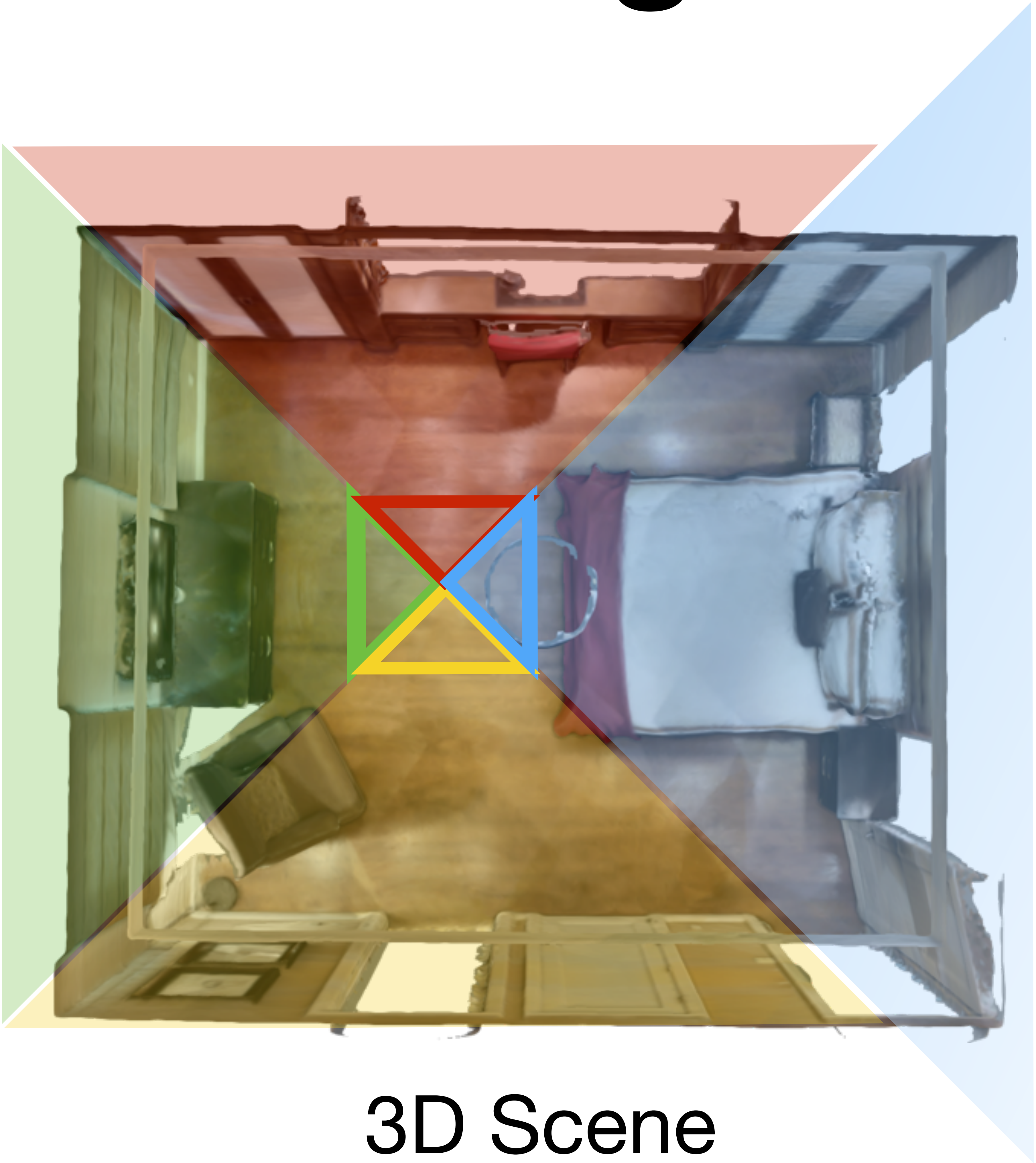
**3D Houses Datasets**

**Synthetic Houses (SUNCG):**
58,866 RGB-D panoramas
Pre-train

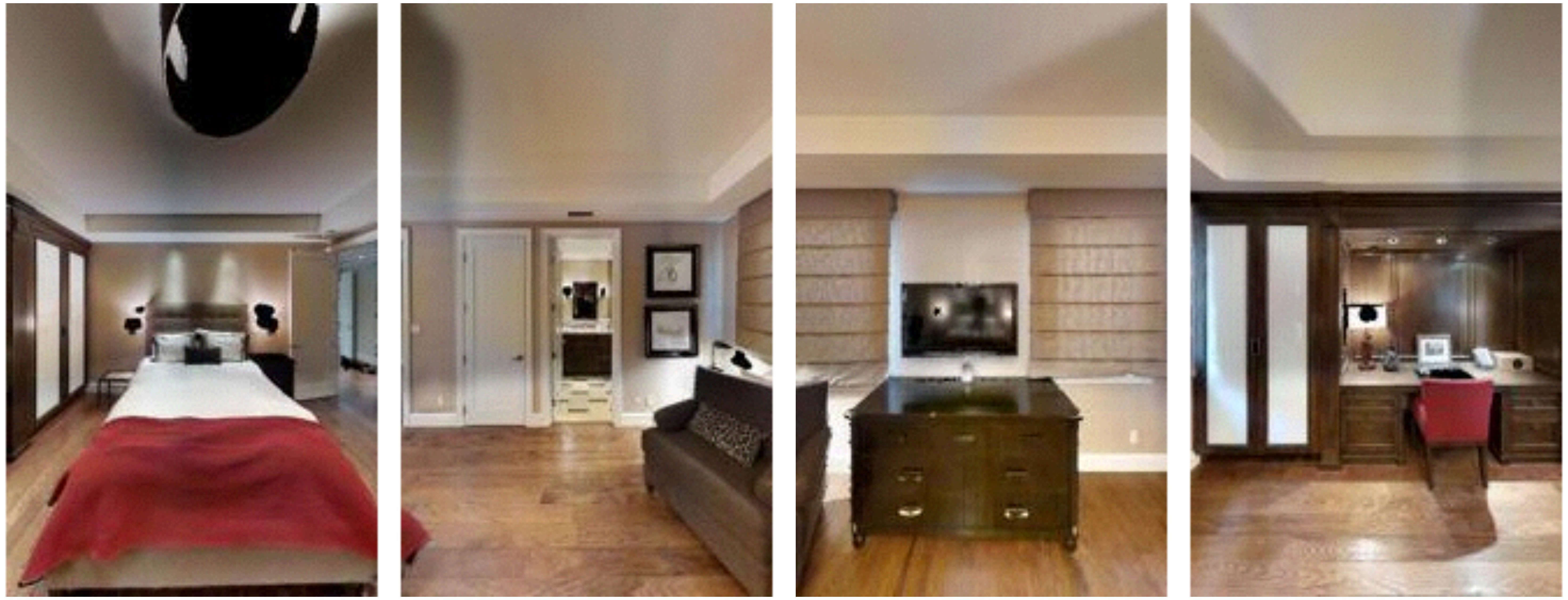


**Real-Word Houses (Matterport3D):**
5,315 RGB-D panoramas
Fine-tune and test

# Challenge 1: How to obtain the data?

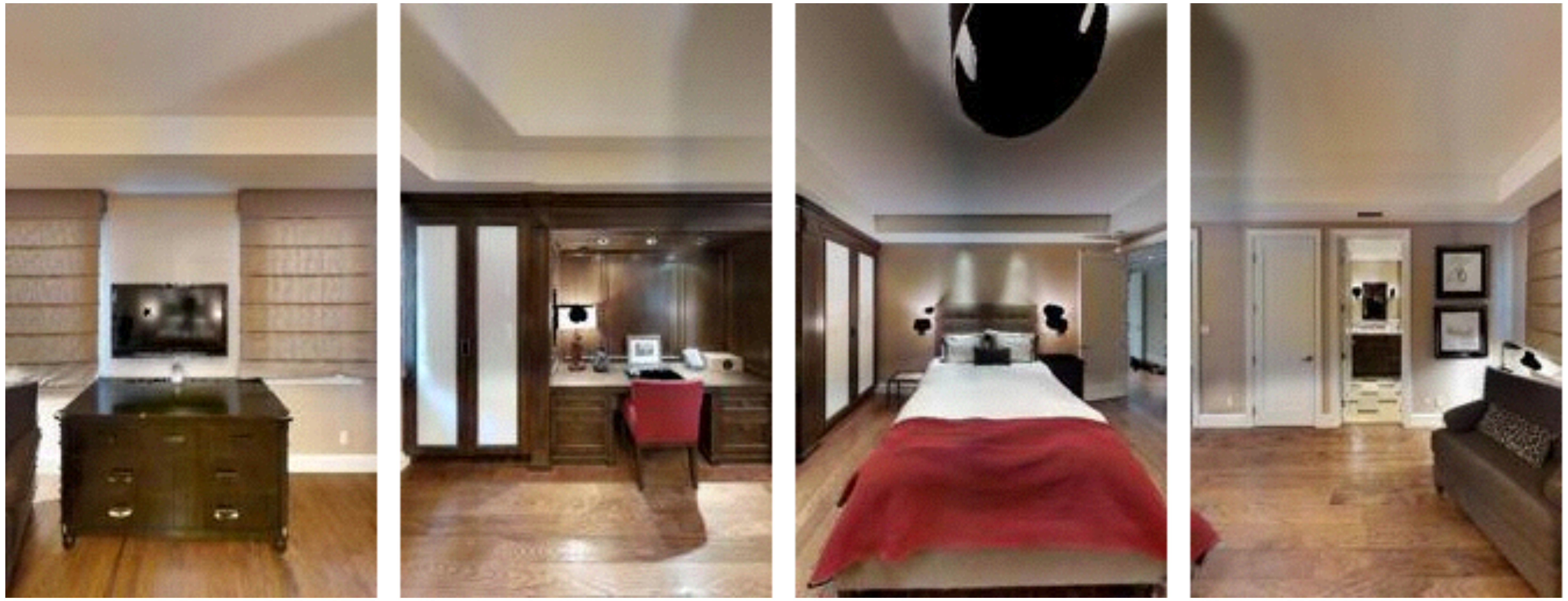


3D Scene



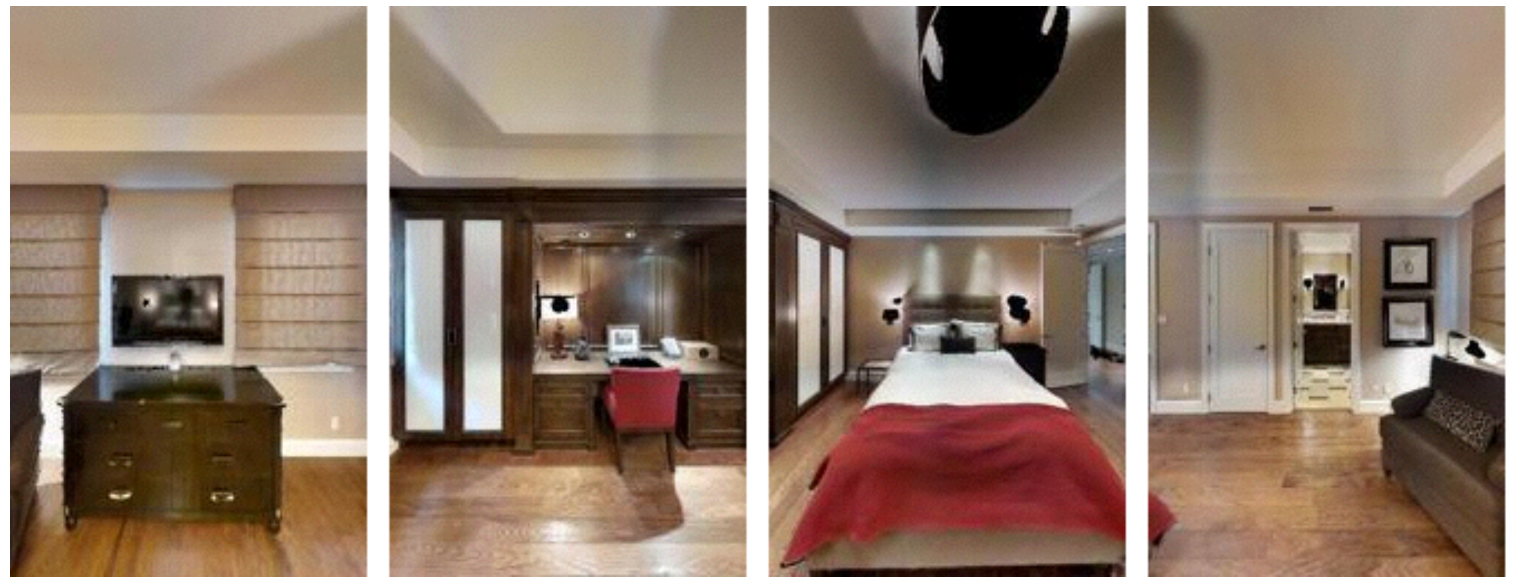
Whole Room Sky-box Panorama

# Challenge 1: How to obtain the data?

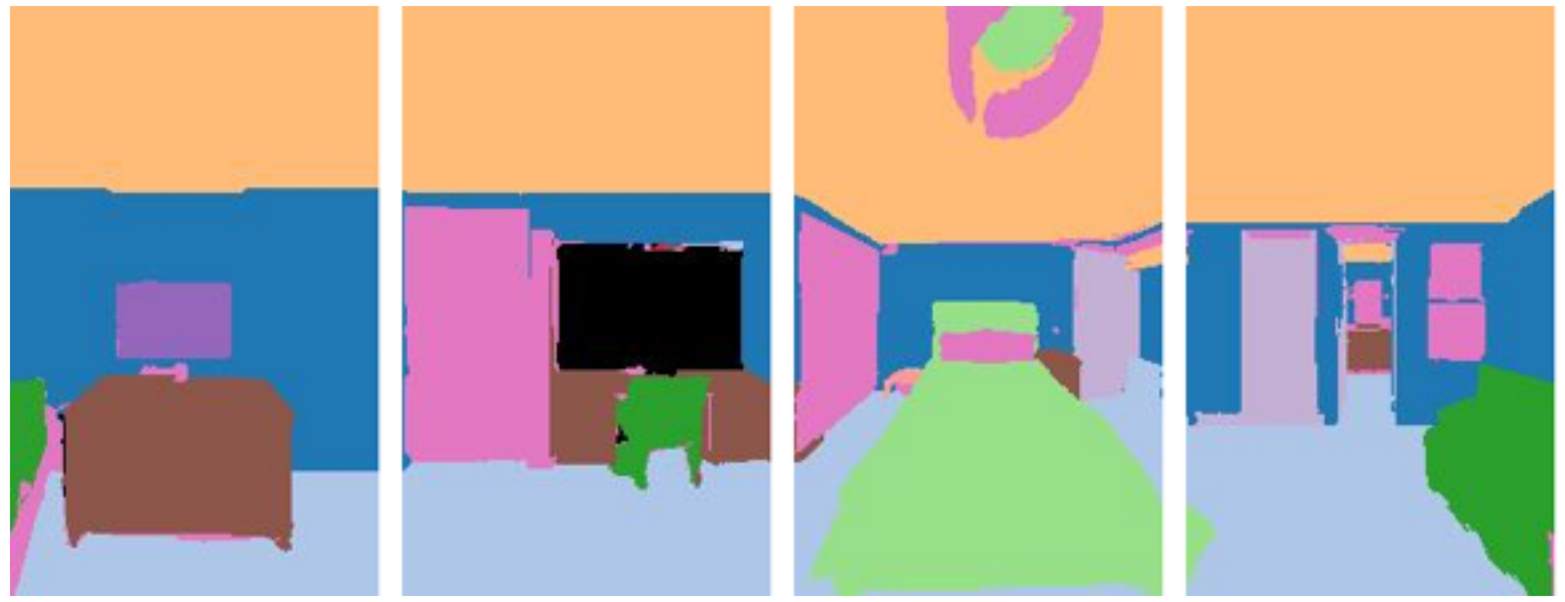


Whole Room Sky-box Panorama

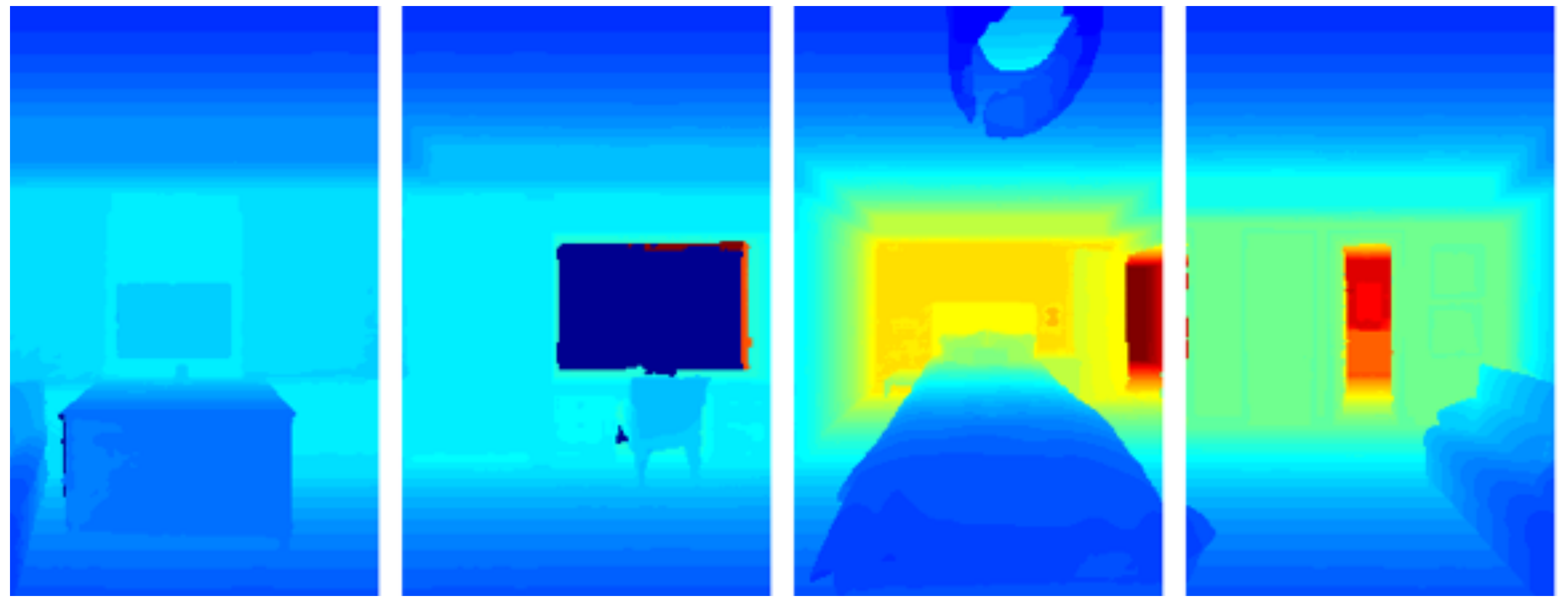
# Challenge 1: How to obtain the data?



Color Images

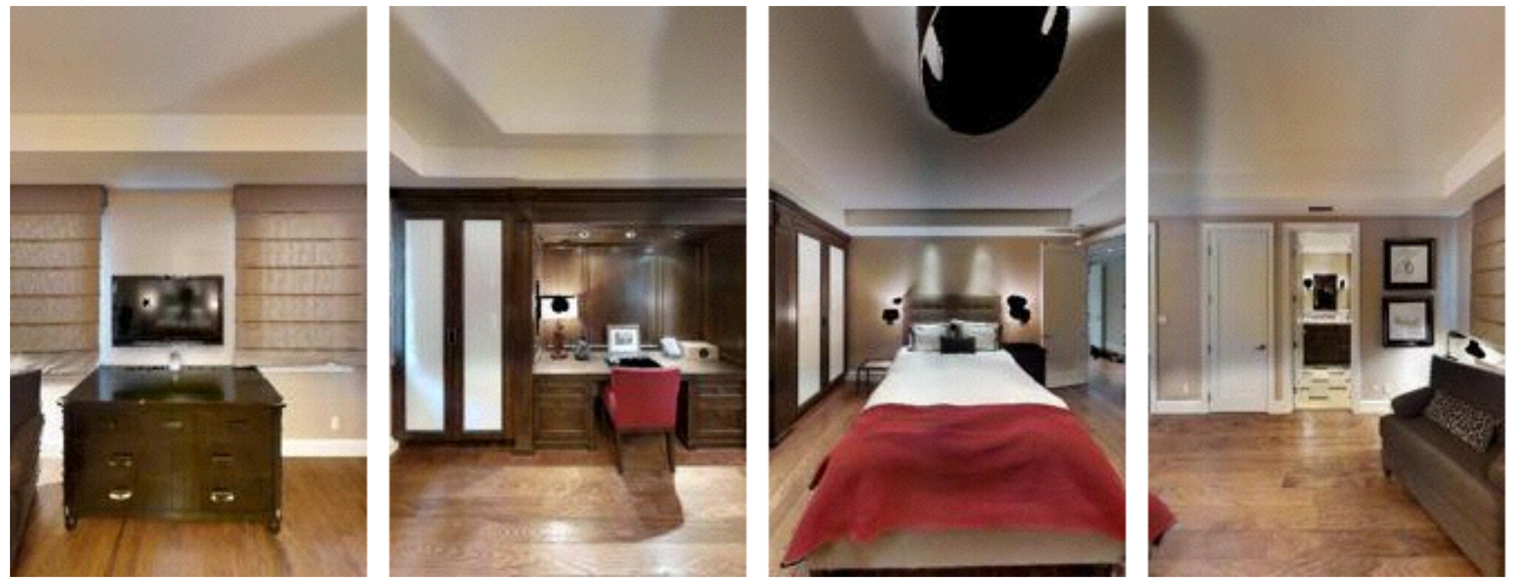


Semantics



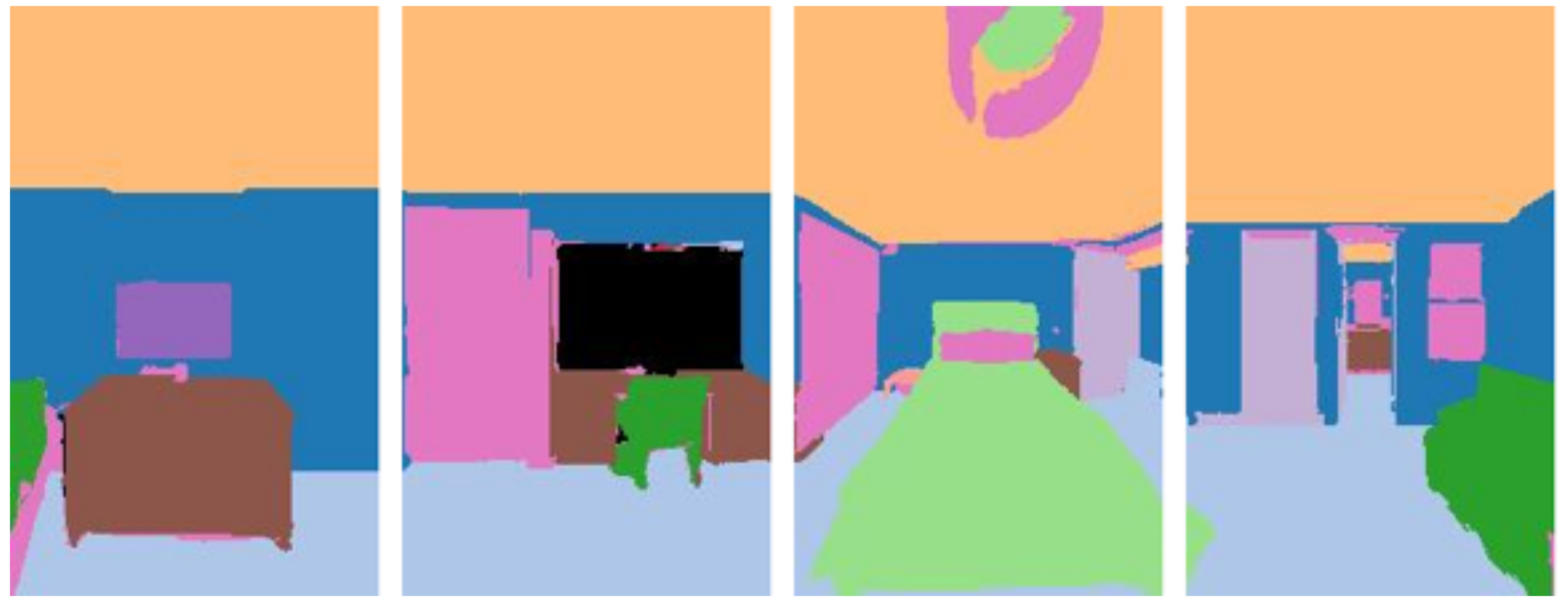
3D Structure

# Challenge 1: How to obtain the data?



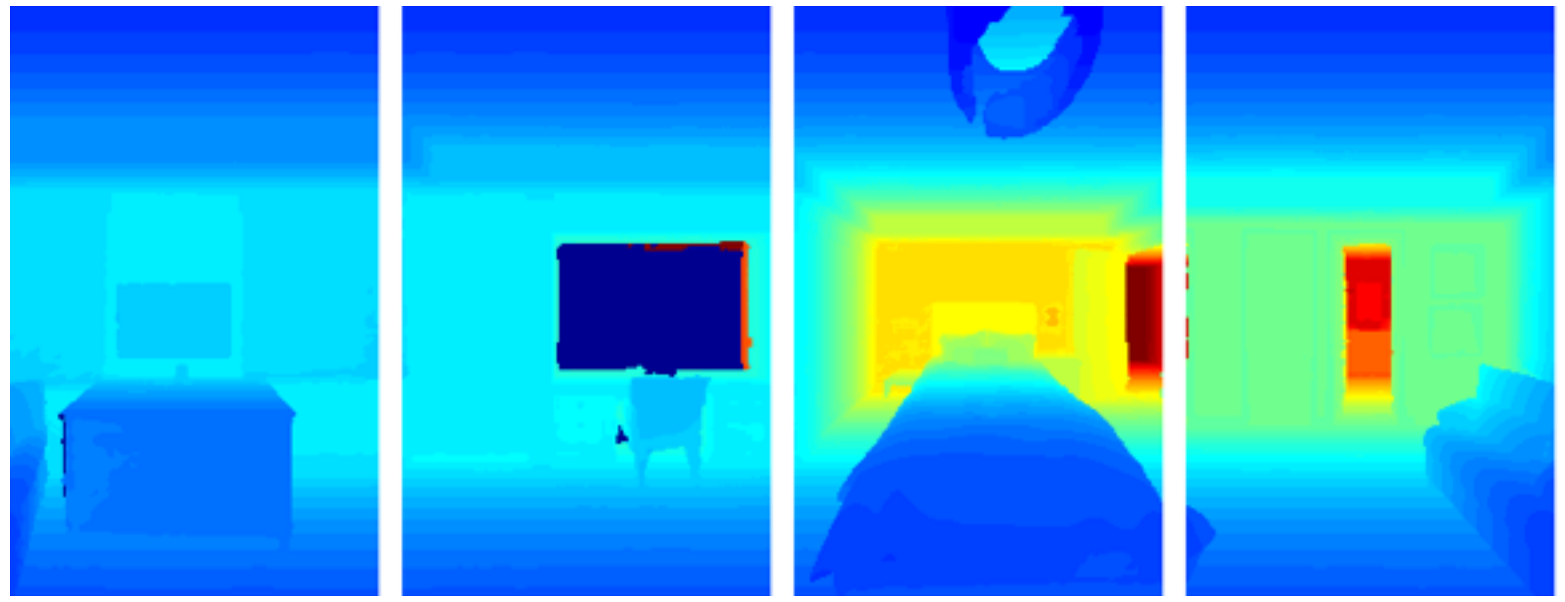
Color Images

**Color = R,G,B**



Semantics

**Semantics = ClassId**



3D Structure

**??**

**Standard Data Representation**

# Challenge 1: How to obtain the data?



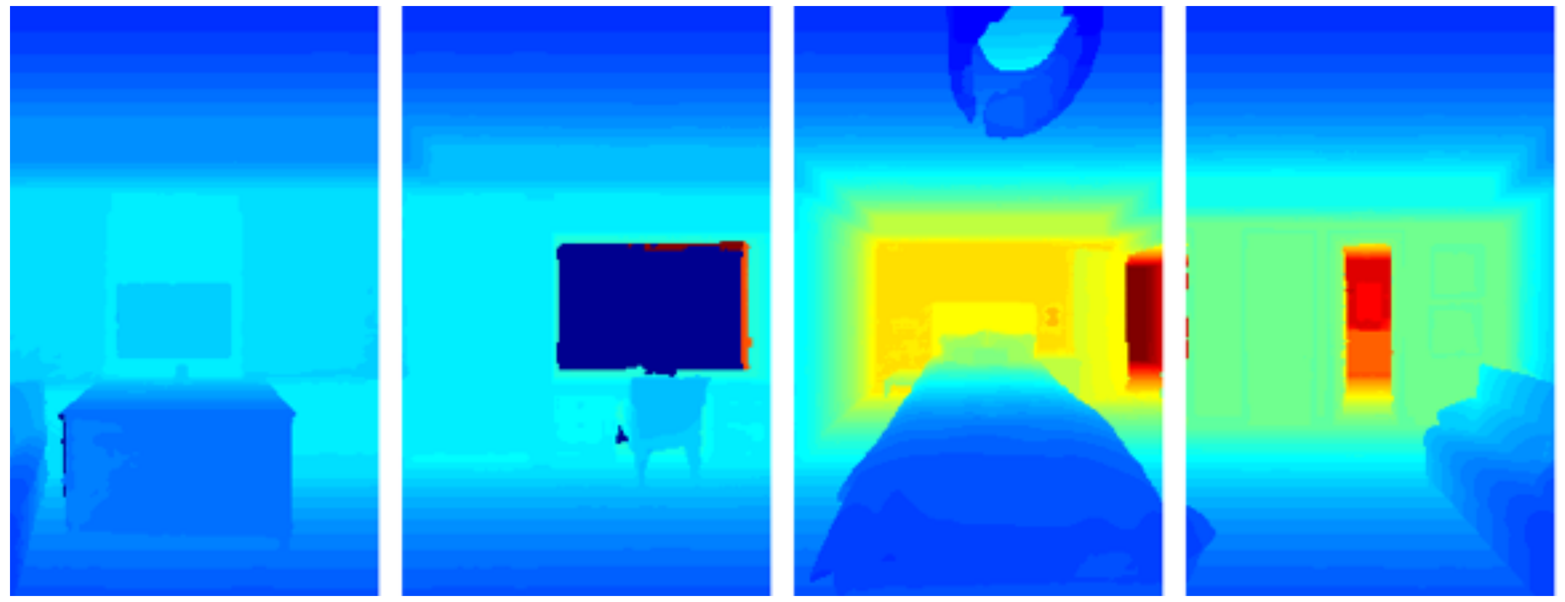
Color Images

**Color = R,G,B**



Semantics

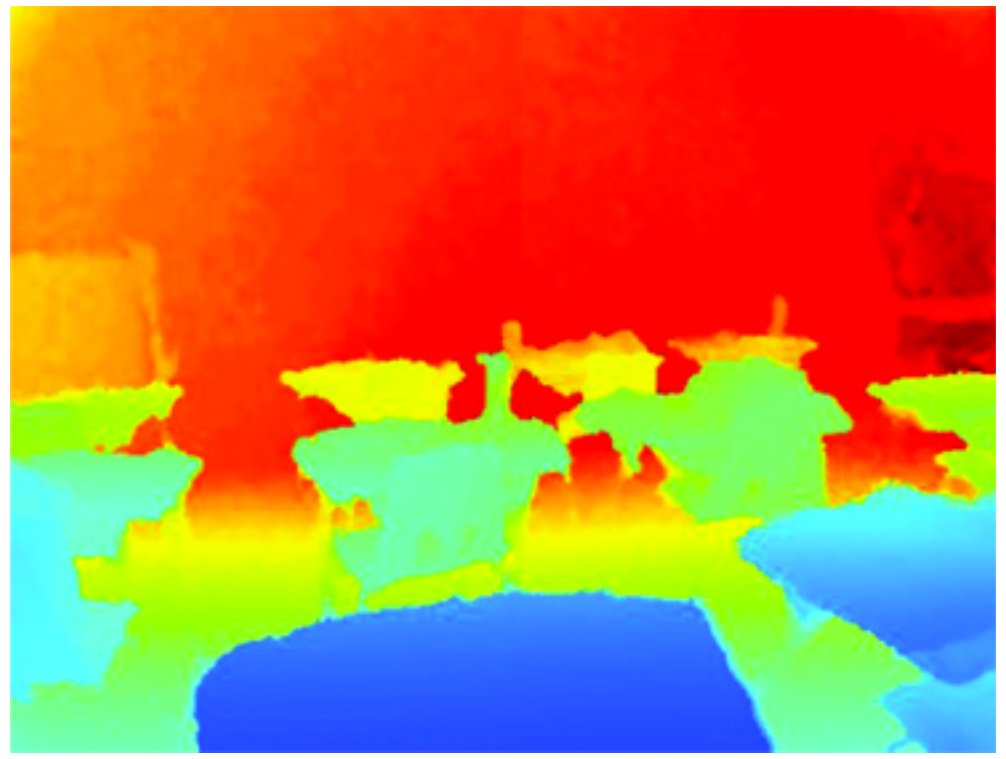
**Semantic = ClassId**



3D Structure

**??**

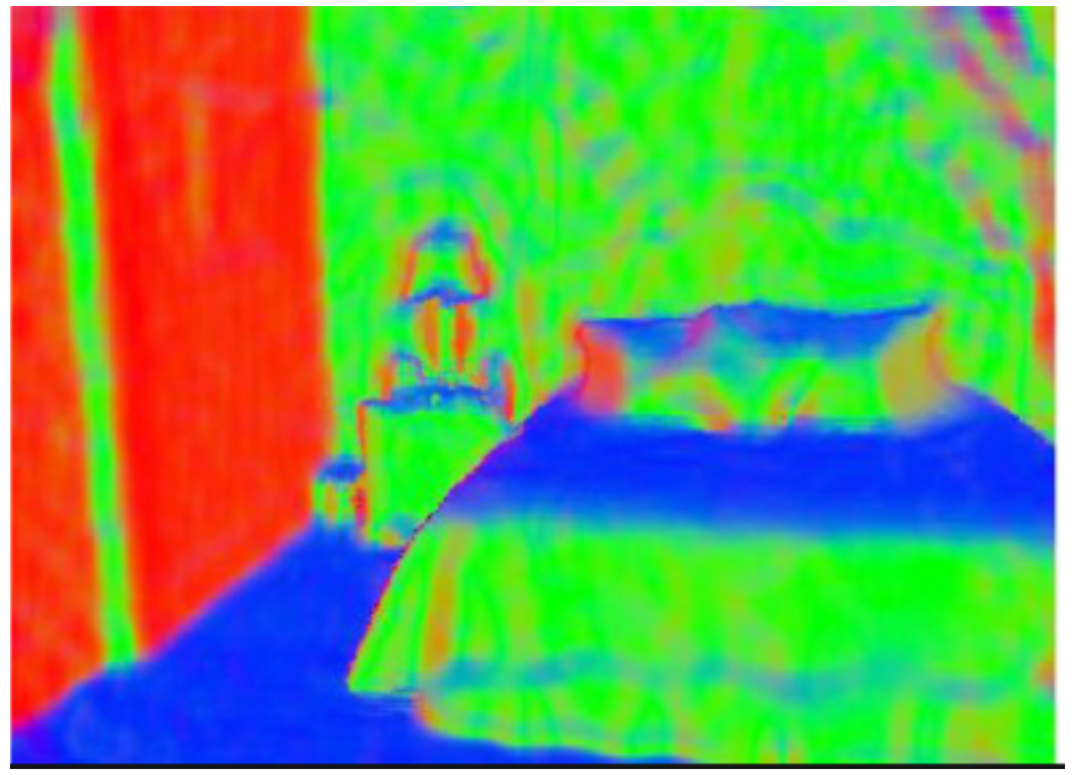
Depth?



Hard to predict.

- Viewpoint dependent.
- Large value variance even for nearby pixels on the same 3D plane

Normal?



Easier to predict.

- Solving back depth from normal is under constrained.

# Challenge 1: How to obtain the data?



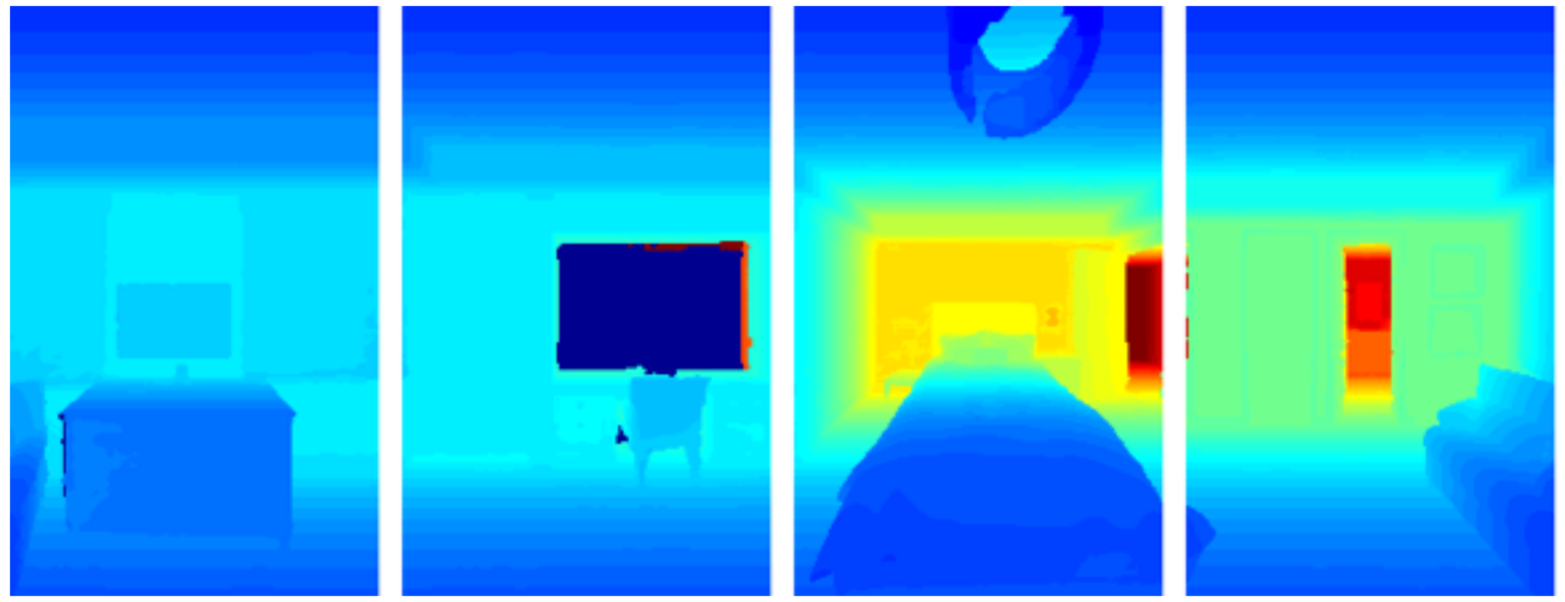
color images

Color = R,G,B



semantic maps

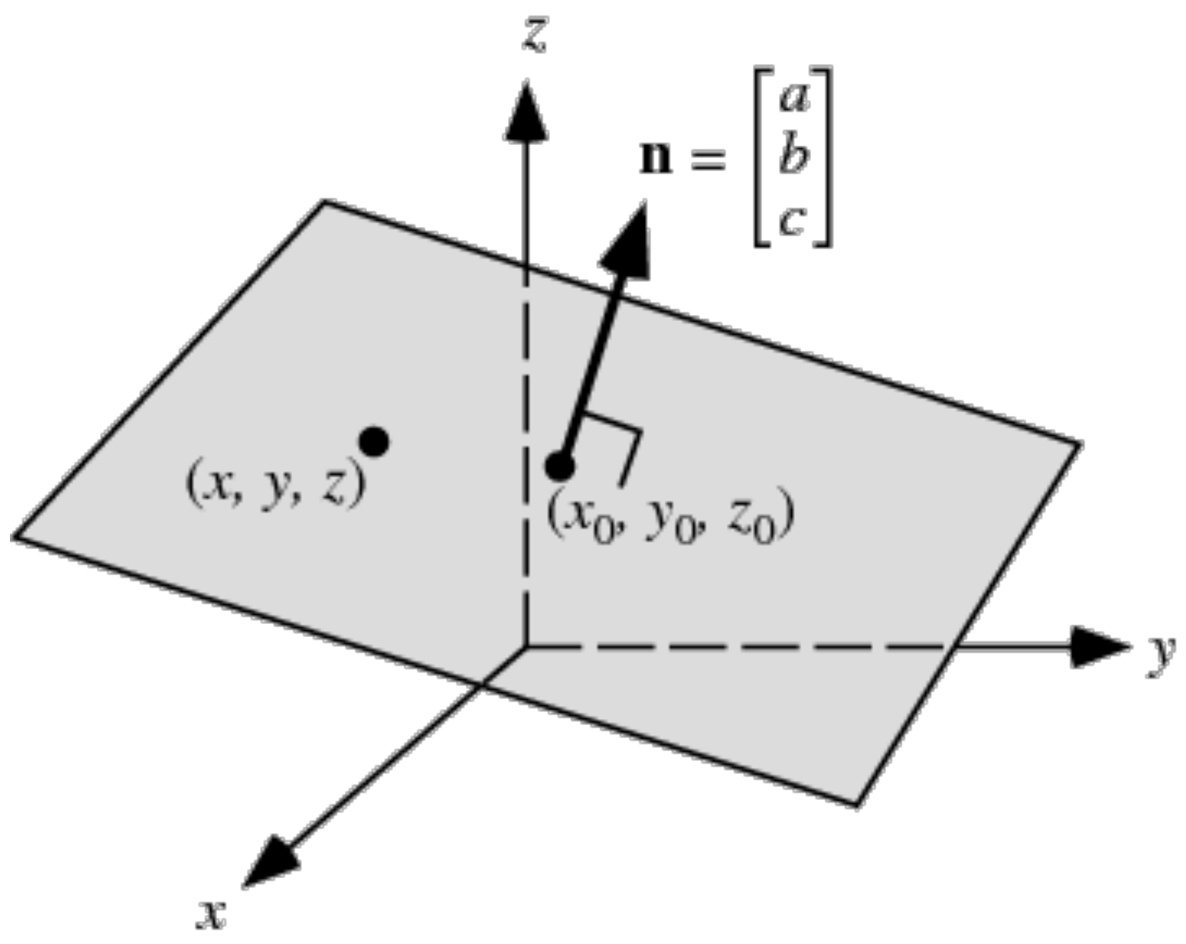
Semantic = ClassId



3D Structure

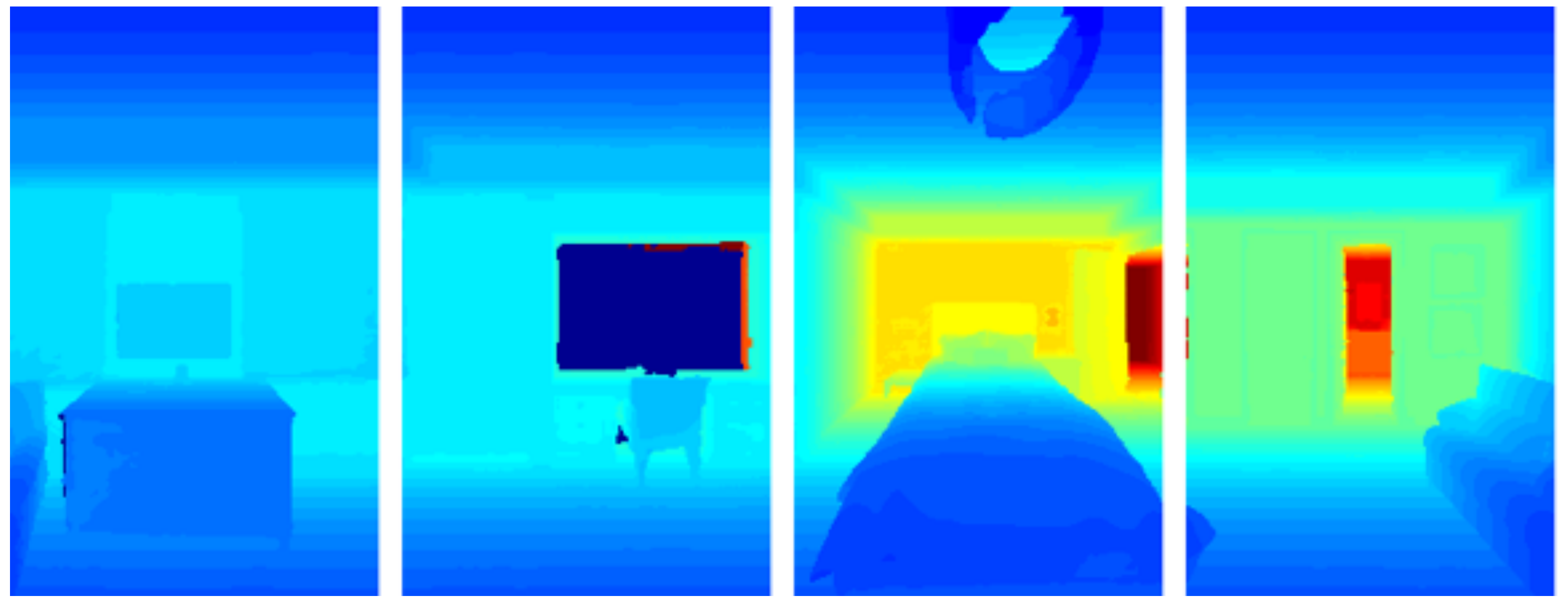
**Challenge 2: How to represent the 3D structure?**

# Challenge2: 3D Structure Representation

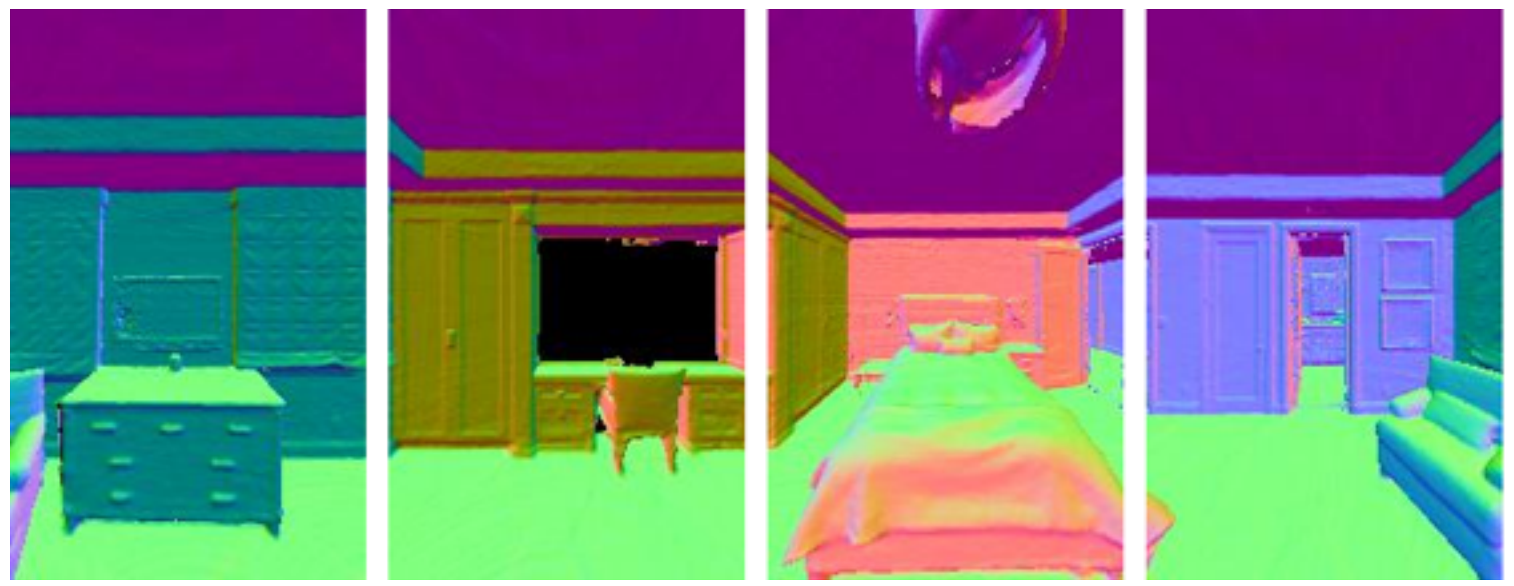



Plane Equation:

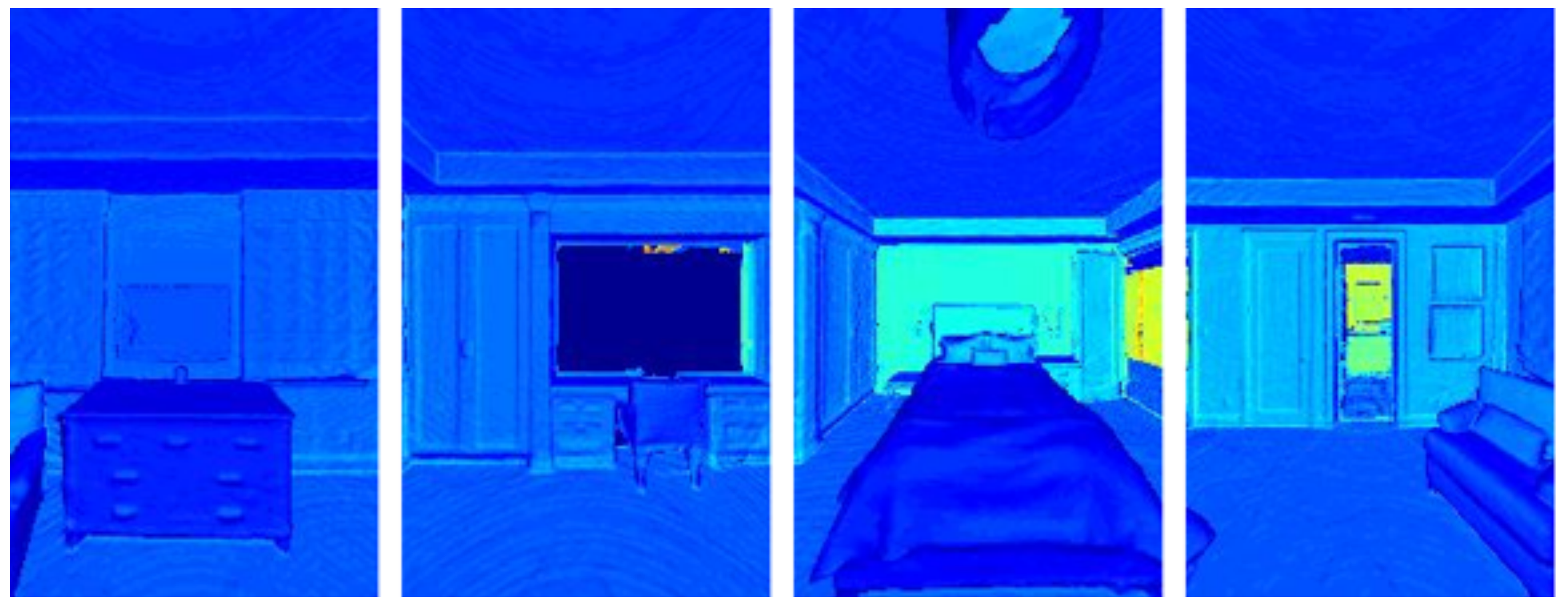
$ax + by + cz - p = 0$



3D Structure

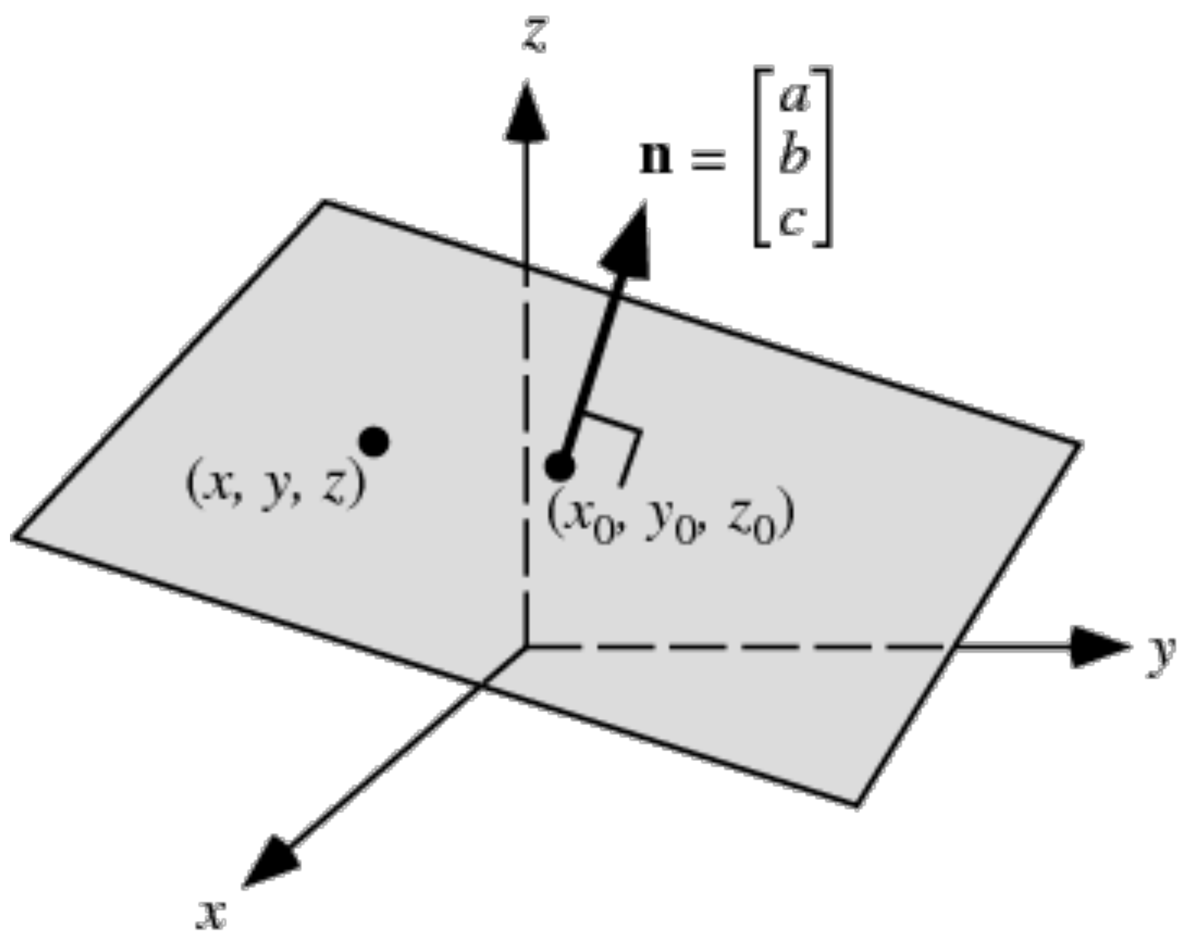


normal (a,b,c)

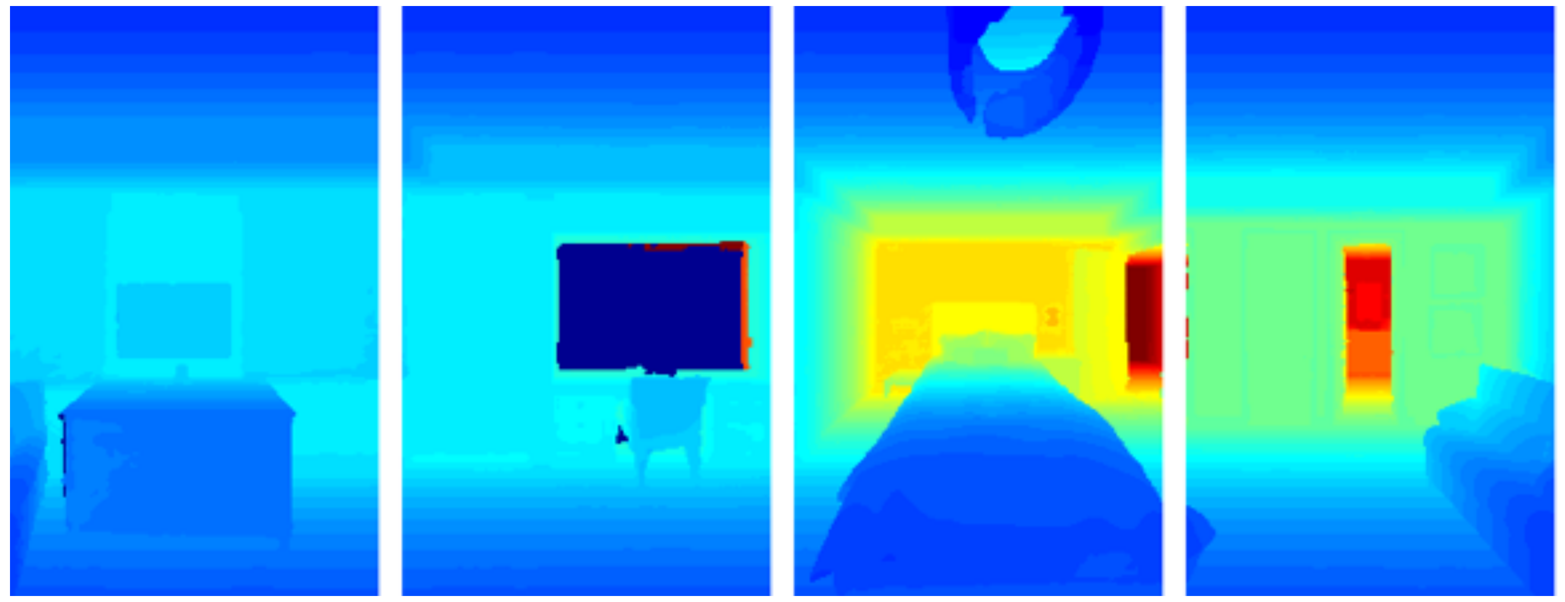


plane distance (p)

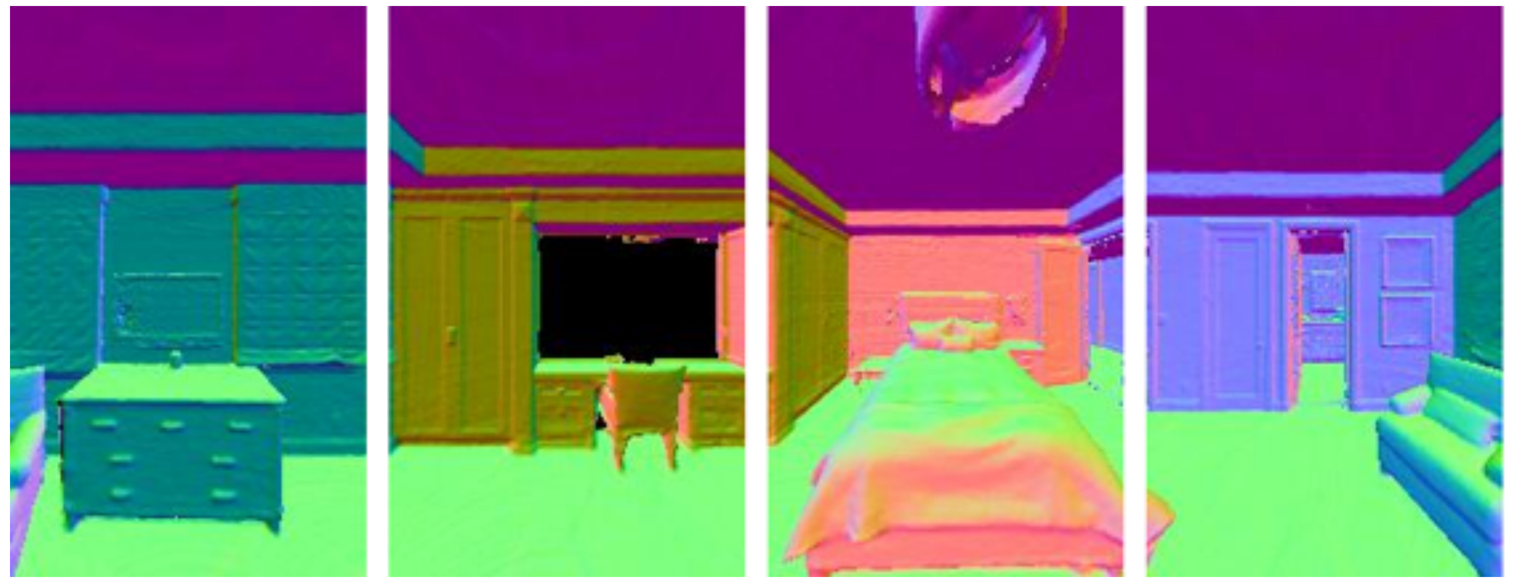
# Challenge2: 3D Structure Representation



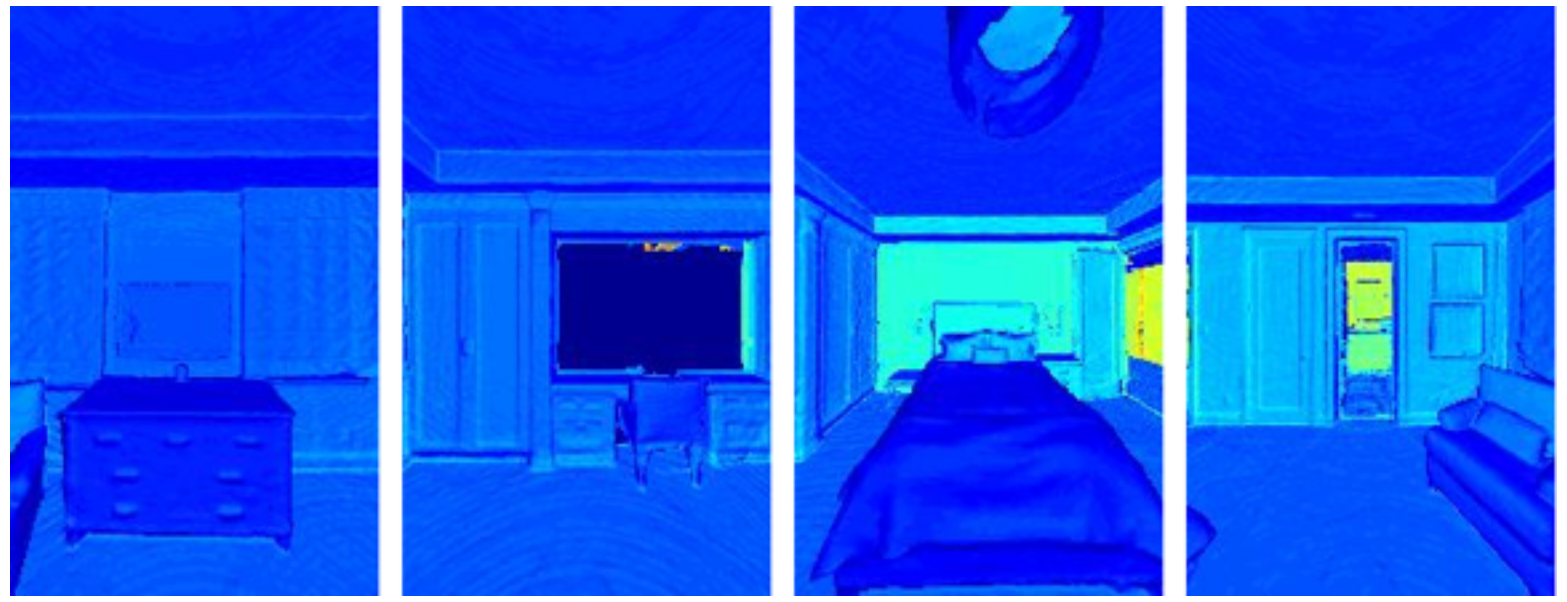

Plane Equation:
$$\underbrace{ax + by + cz}_{} - \underbrace{p}_{} = 0$$



3D Structure



normal (a,b,c)



plane distance (p)

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

✓Representation is piecewise constant

✓More robust

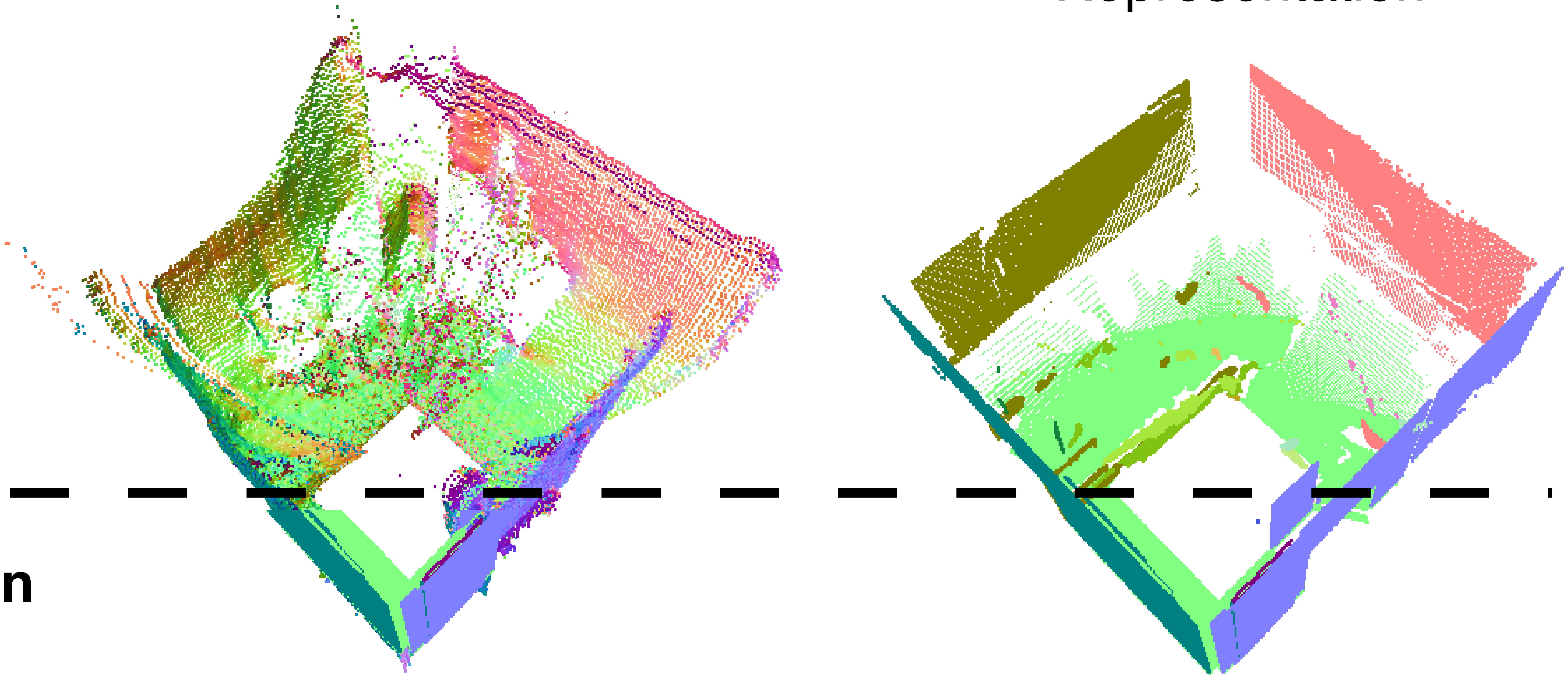
# Challenge2: 3D Structure Representation

Raw Depth Representation

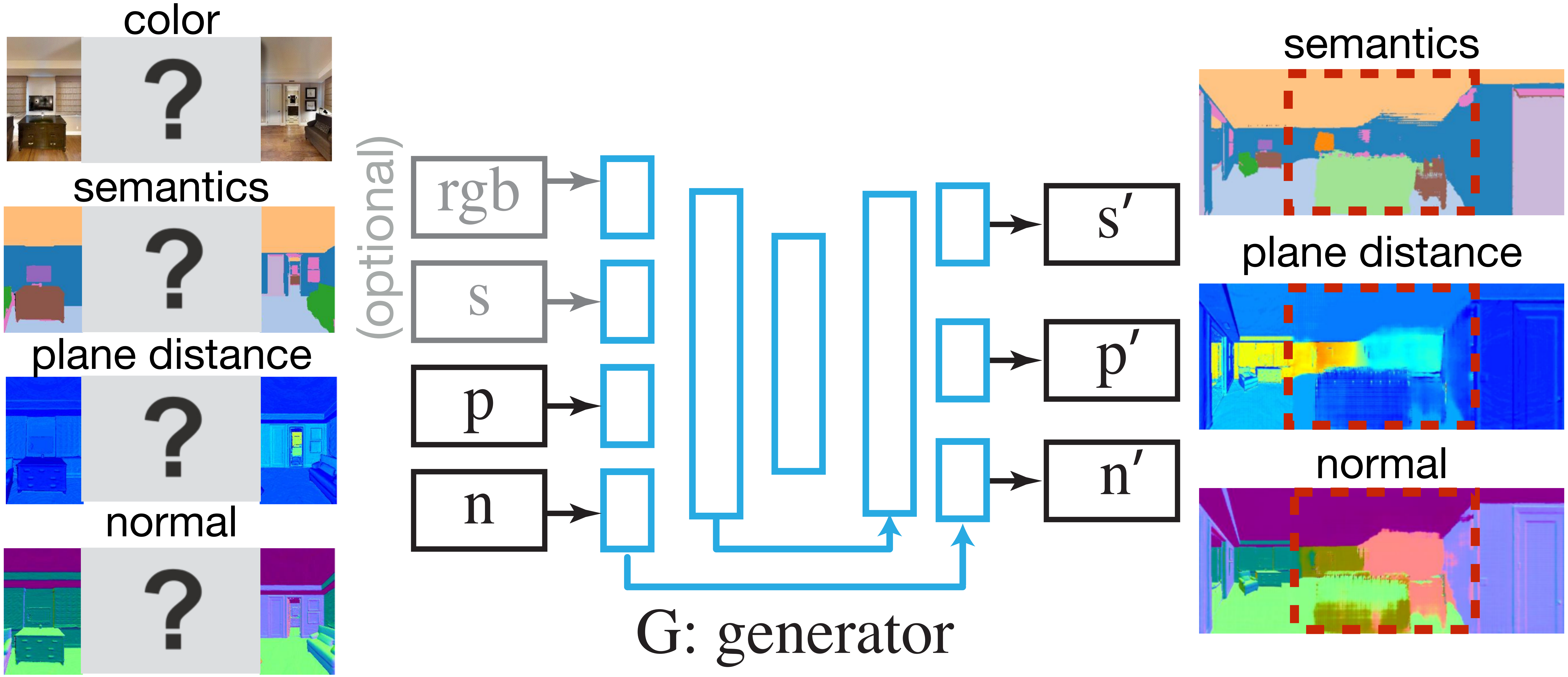
Plane Representation

Prediction

Observation

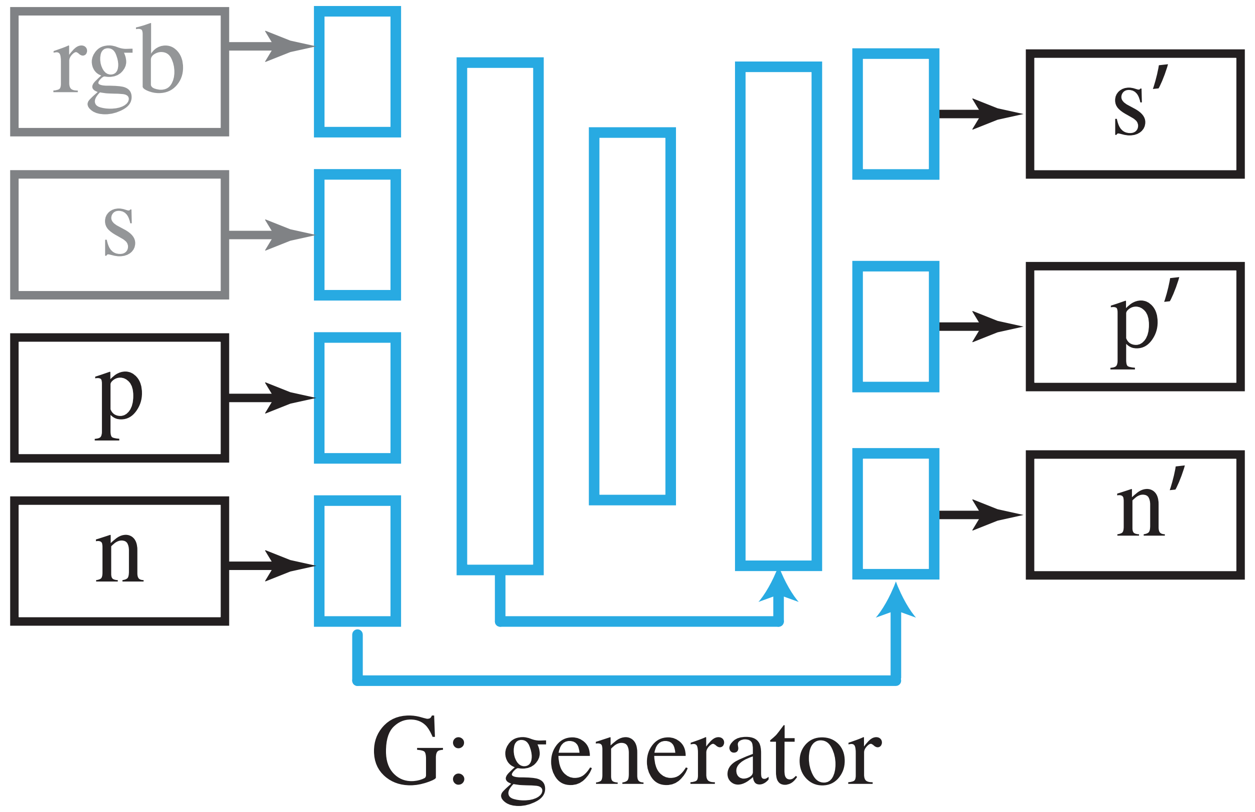


# Im2Pano3D Network



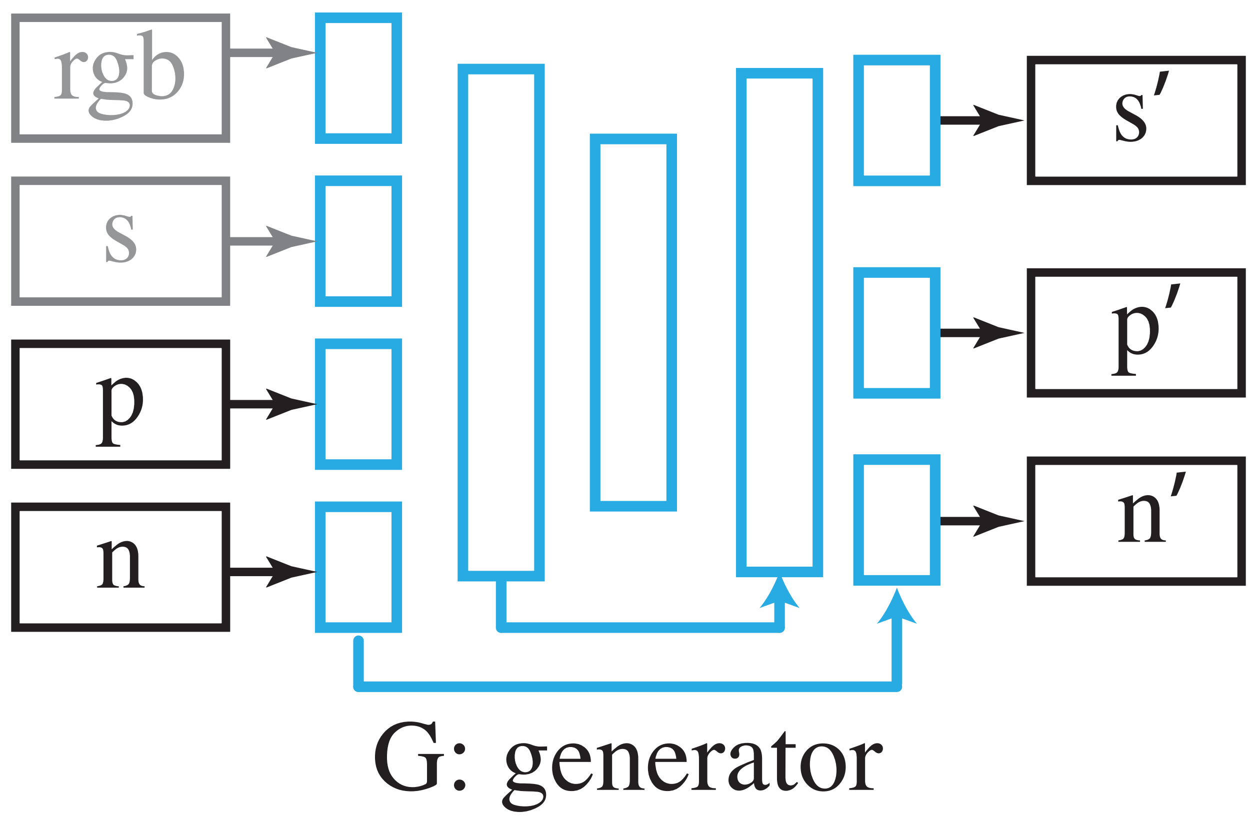

**Challenge 3: What training objectives should we use?**

# Challenge 3: Training Objectives



# Challenge 3: Training Objectives

Every Pixel is Correct




softmax

L1

cosine

# Challenge 3: Training Objectives


Adversarial loss
Goodfellow et al. 2014

# Challenge 3: Training Objectives

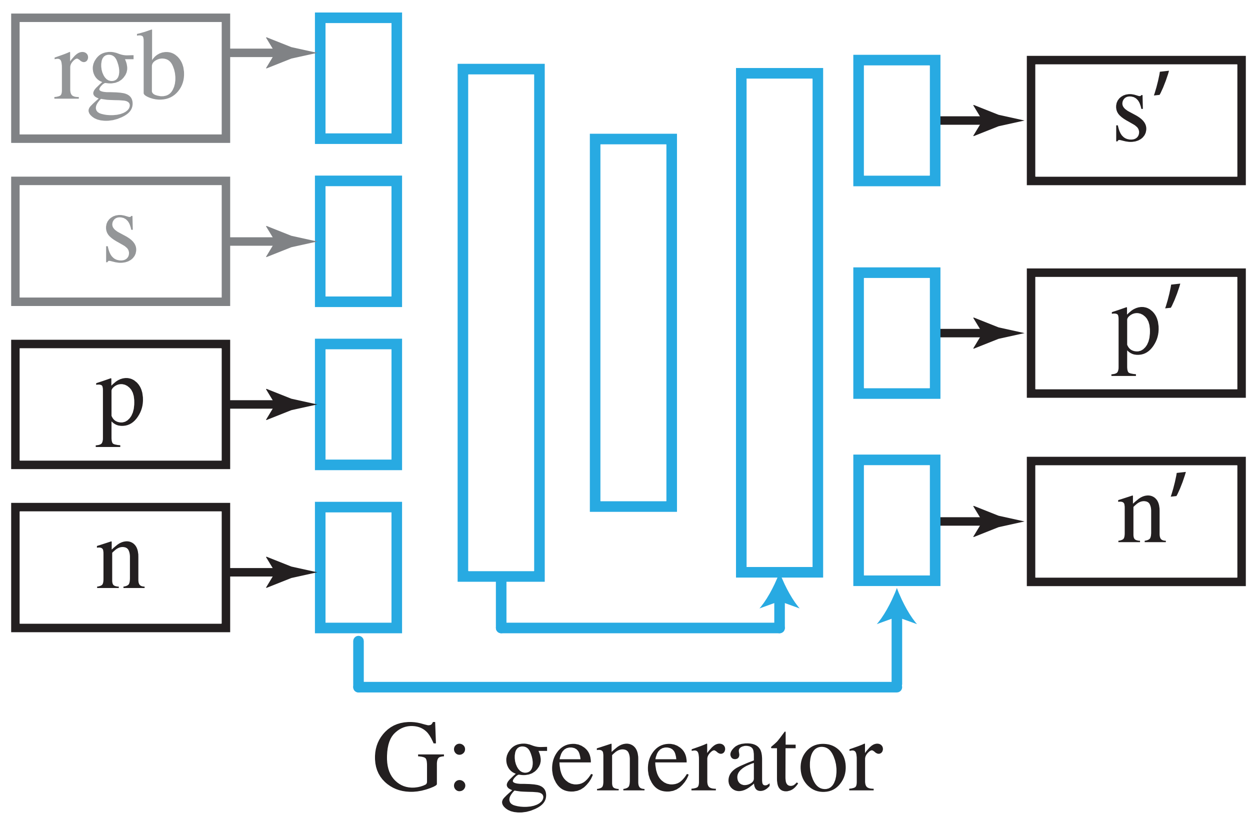

$$L_{dis} = \sum_{c} |y_c - h(x_c)|$$

# Challenge 3: Training Objectives

Every Pixel is Correct — $L_{recon}$

Similar Scene Attribute — $L_{attribute}$

Prediction is Plausible — $L_{adv}$



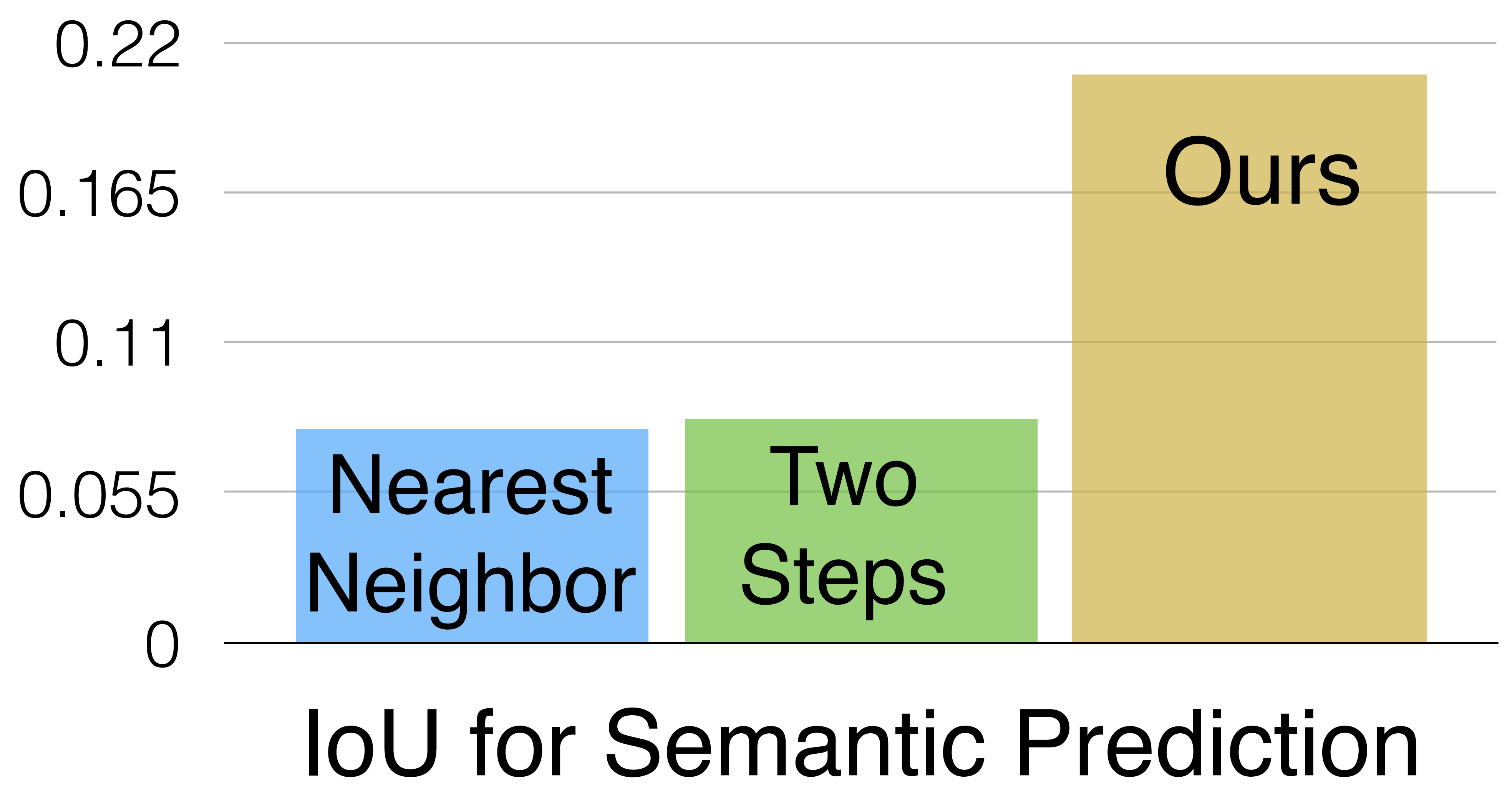

$$L = \lambda_1 L_{recon} + \lambda_2 L_{attribute} + \lambda_3 L_{adv}$$

# Evaluation

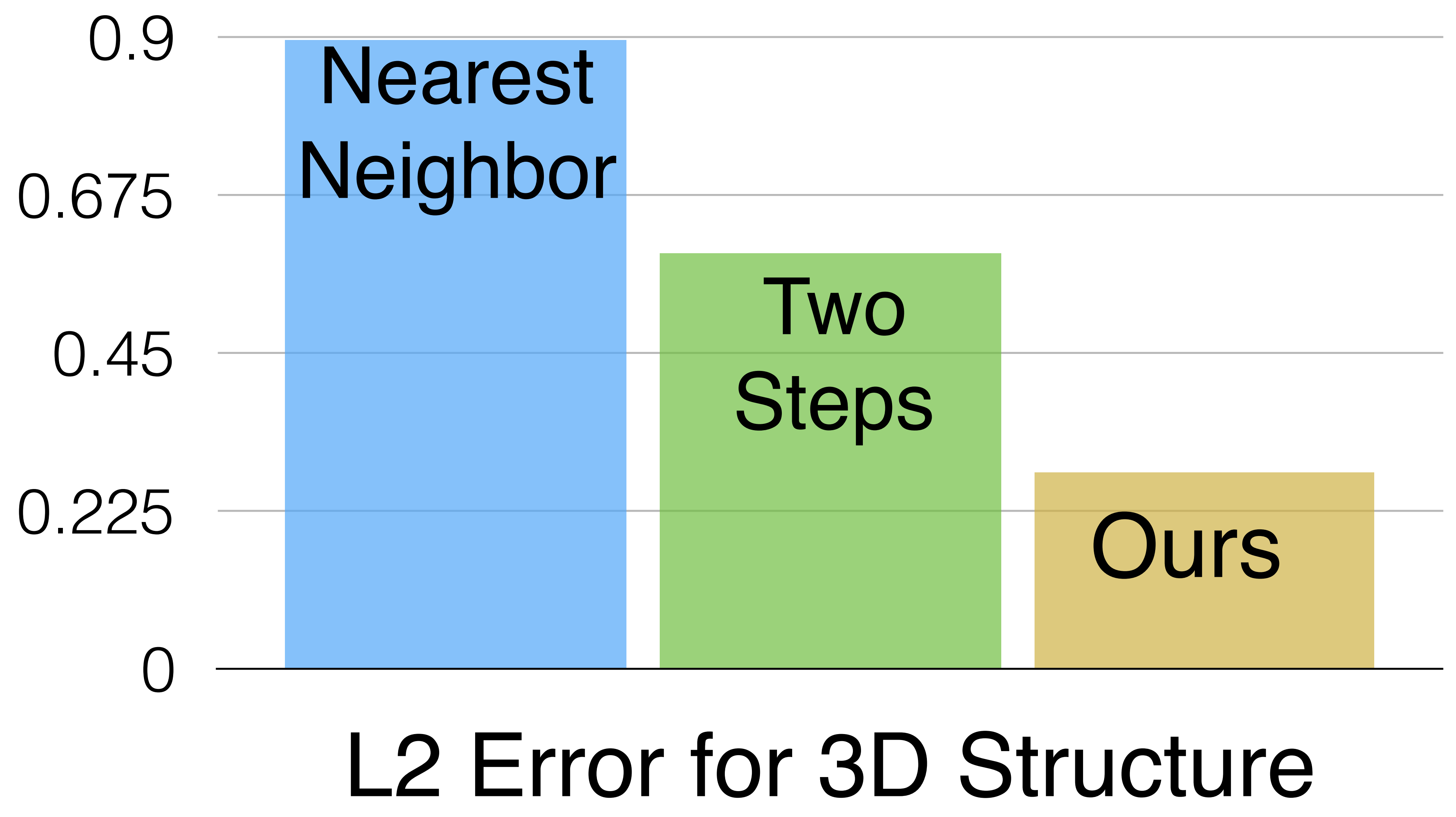
Every pixel is correct — Similar scene attribute — Prediction is plausible

| | Every pixel is correct | Similar scene attribute | Prediction is plausible |
|---|---|---|---|
| Semantic Prediction | Pixel-wise IoU | Probability over Groundtruth | Inception score (Scene classification) |
| 3D Structure | Pixel-wise L2 distance | Earth Mover Distance | |

# Evaluation

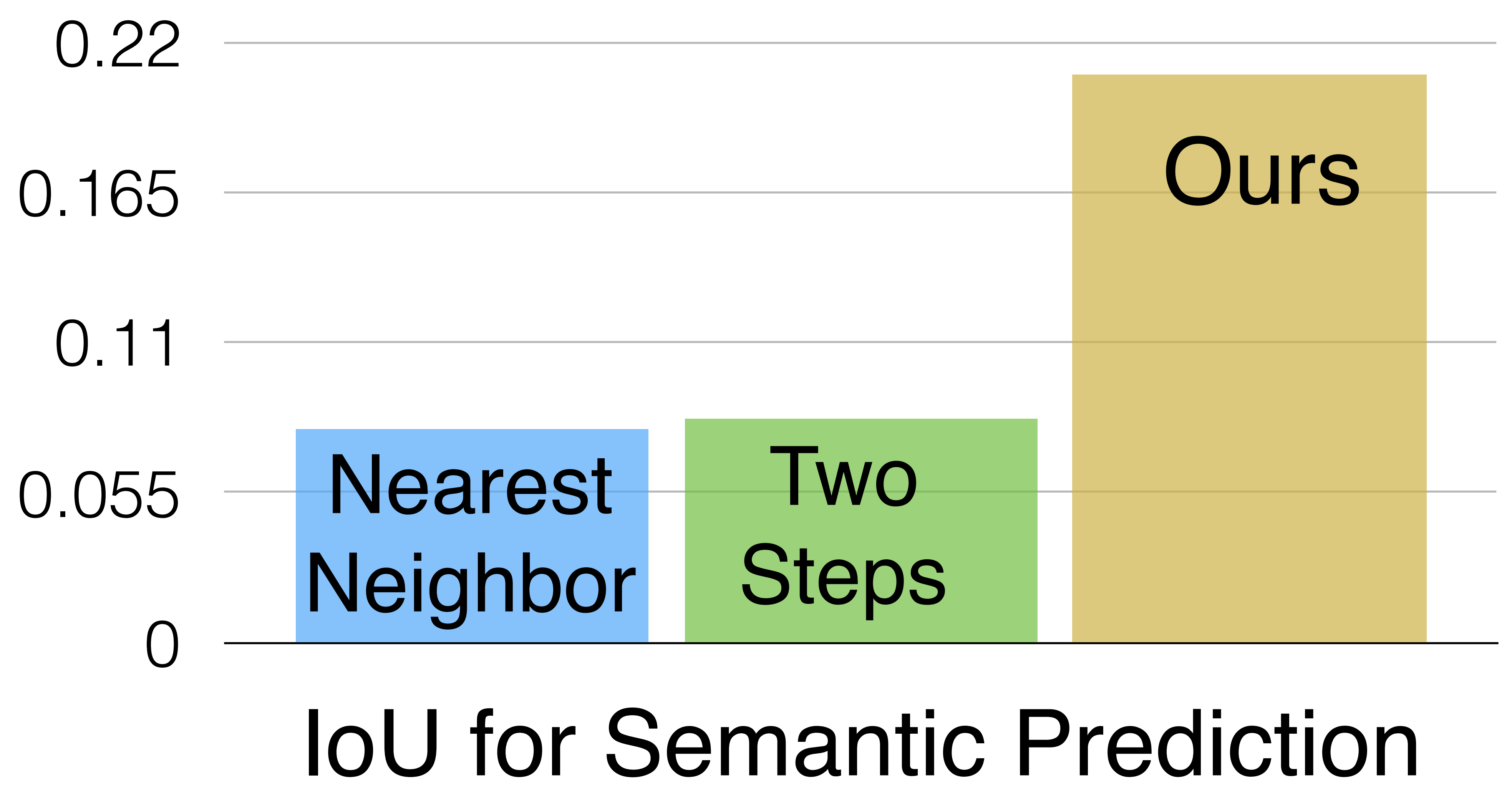


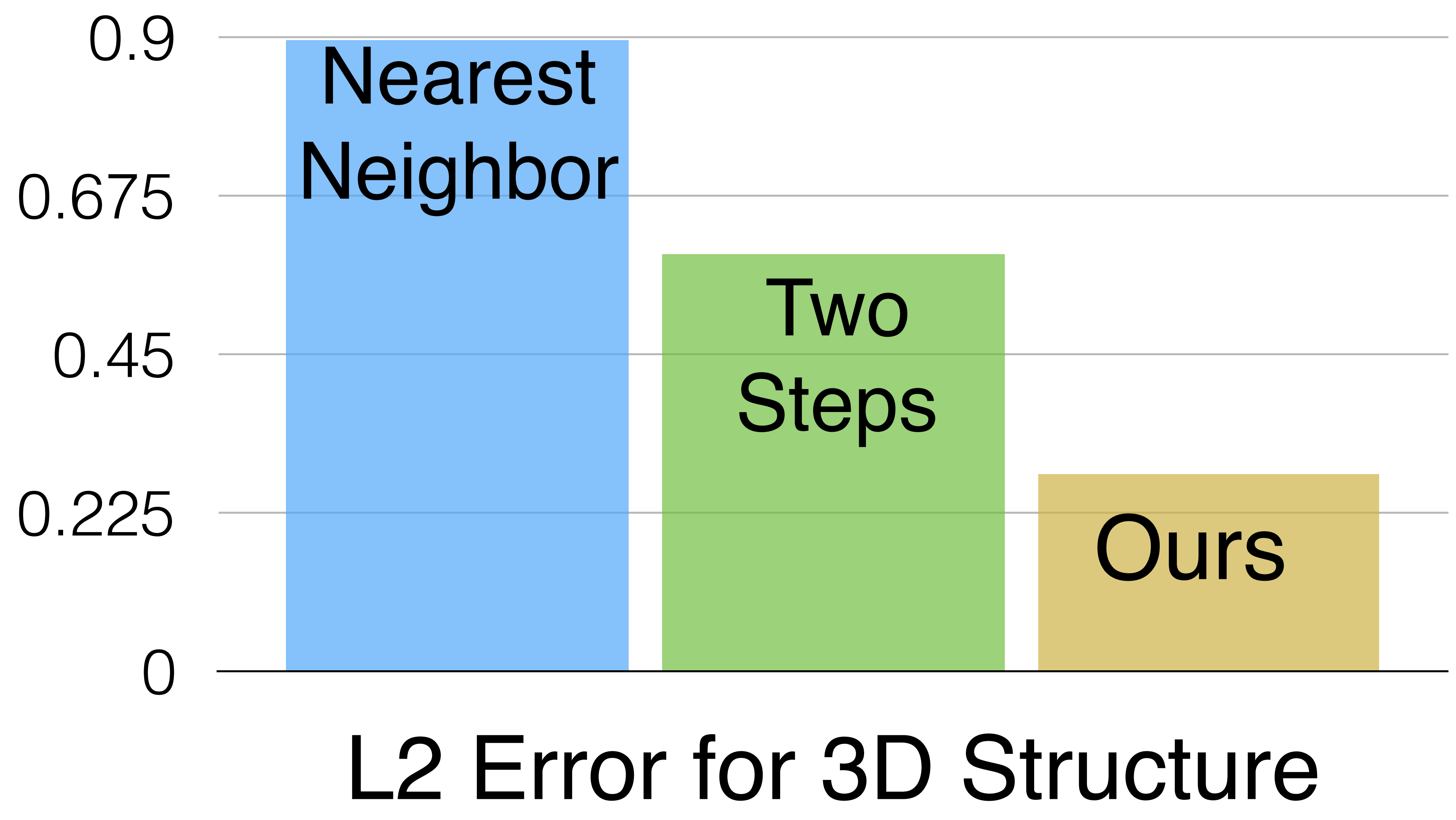

Higher is better

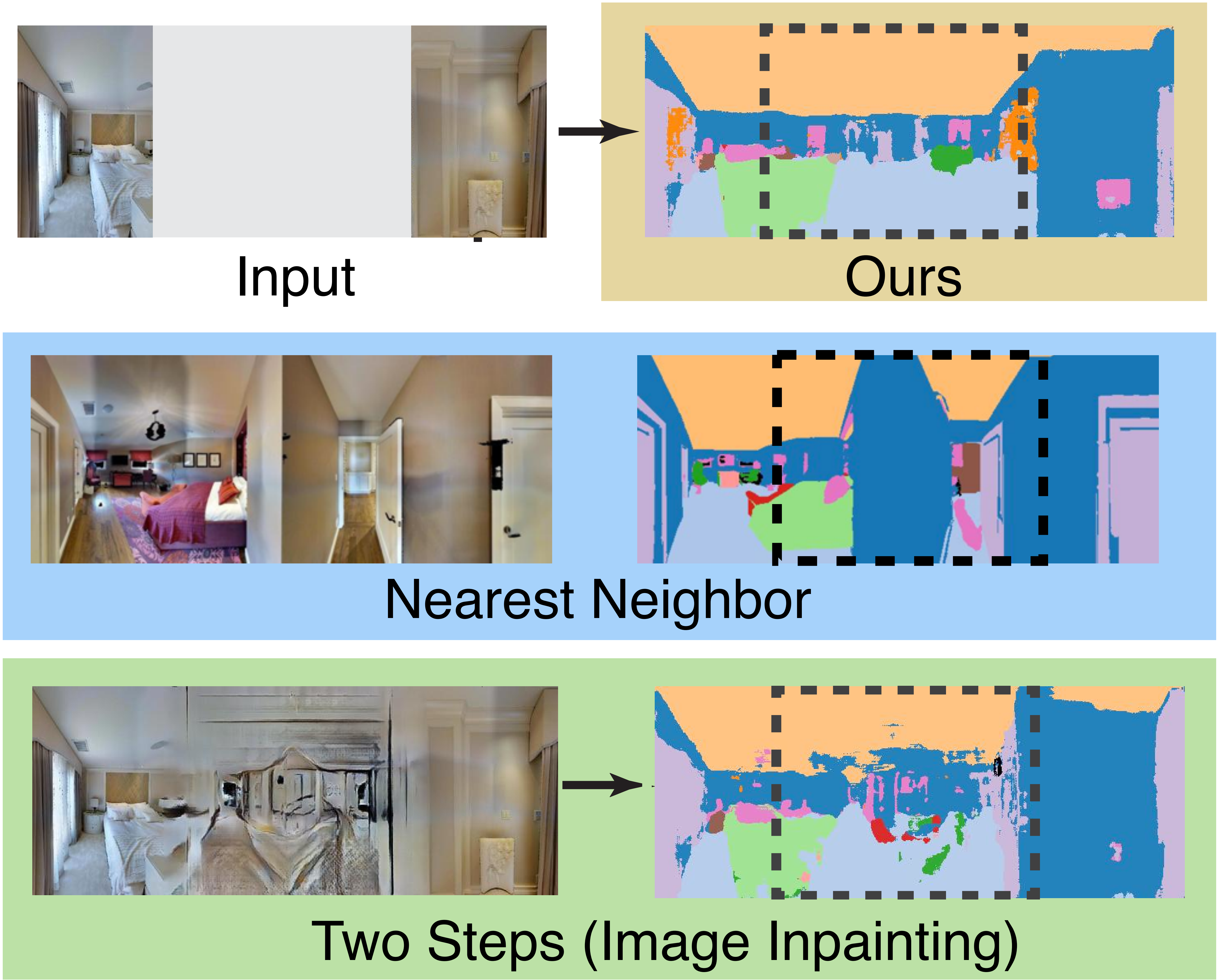



Lower is better

# Evaluation


0.22
0.165
0.11
0.055
0
Nearest Neighbor
Two Steps
Ours
IoU for Semantic Prediction


0.9
0.675
0.45
0.225
0
Nearest Neighbor
Two Steps
Ours
L2 Error for 3D Structure


Input
Ours
Nearest Neighbor
Two Steps (Image Inpainting)

# Example Results

# Observation

# Predicted Probability Distribution


Ceiling: Red indicates high probability

# Predicted Probability Distribution


Floor: Red indicates high probability

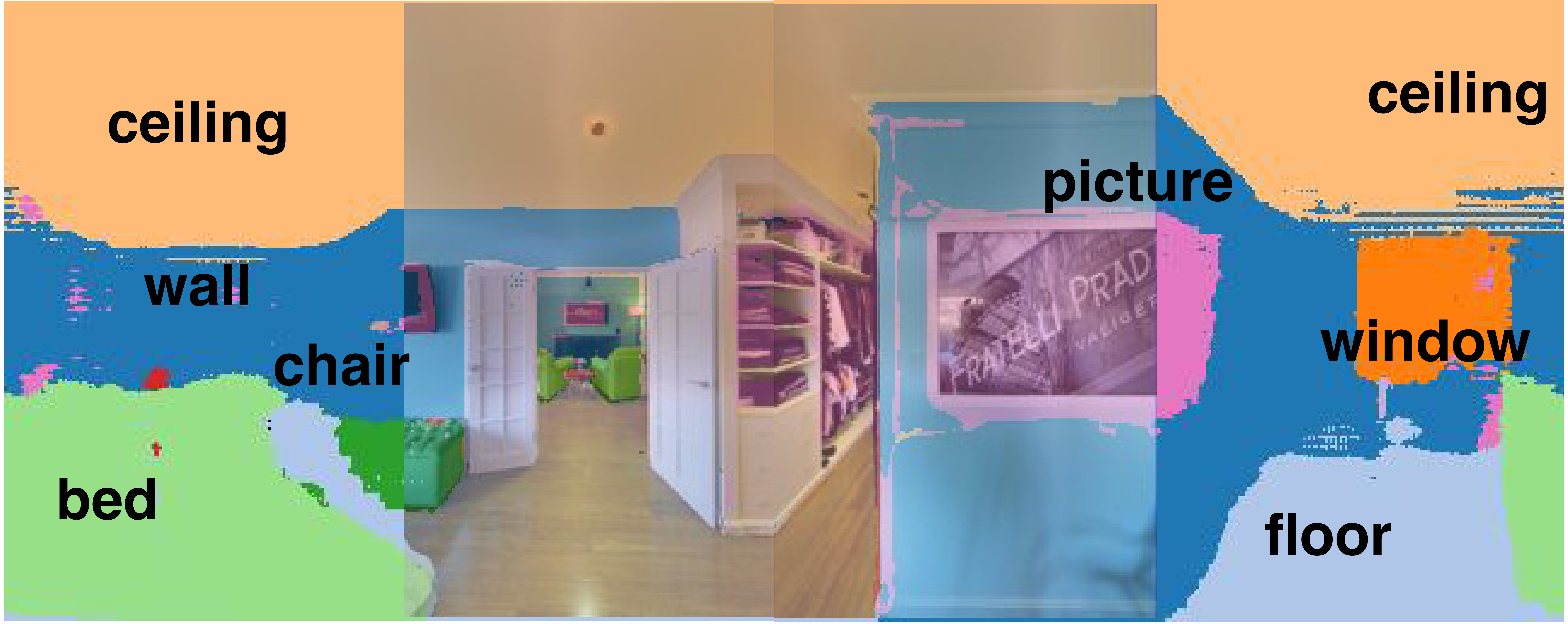
# Predicted Probability Distribution

Wall: Red indicates high probability

# Predicted Probability Distribution

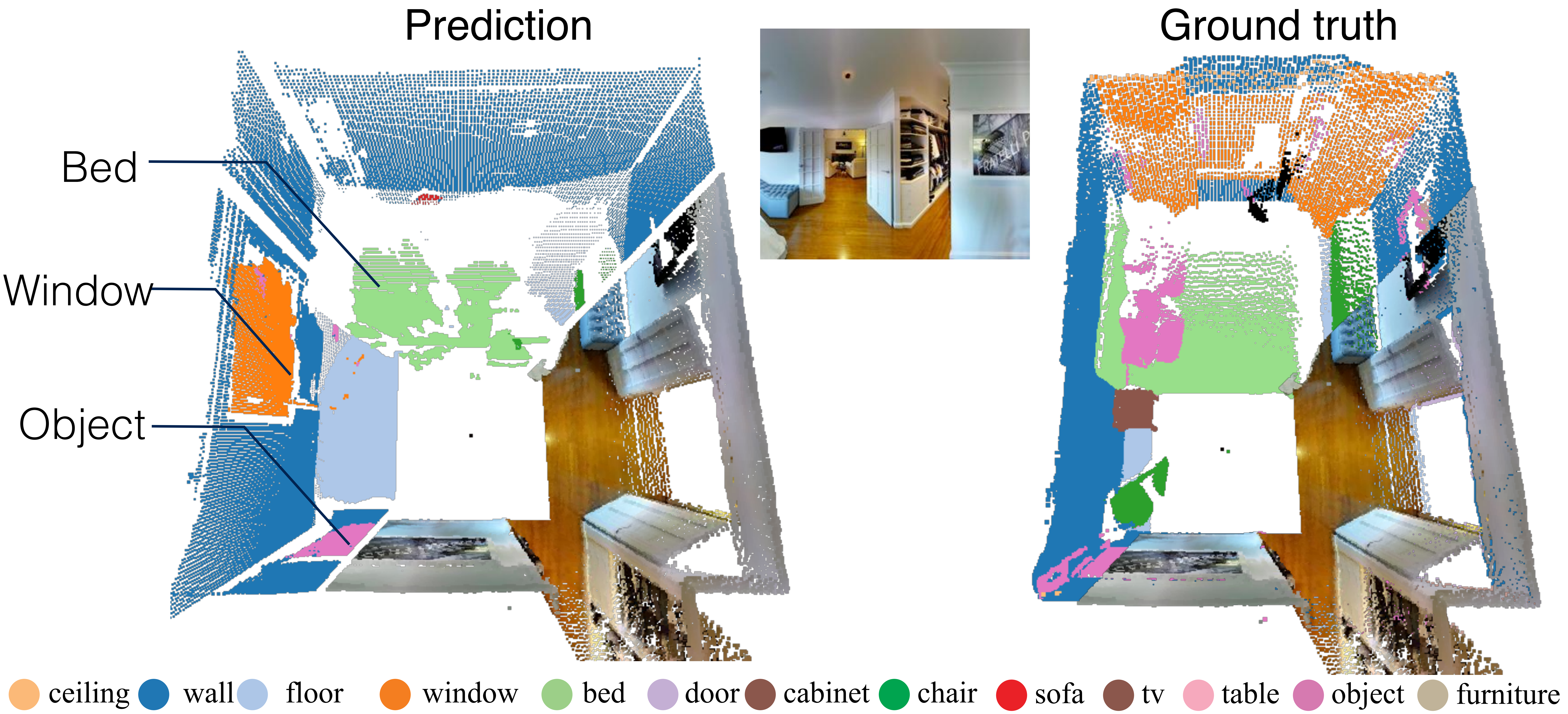

Bed: Red indicates high probability

# Semantic Prediction

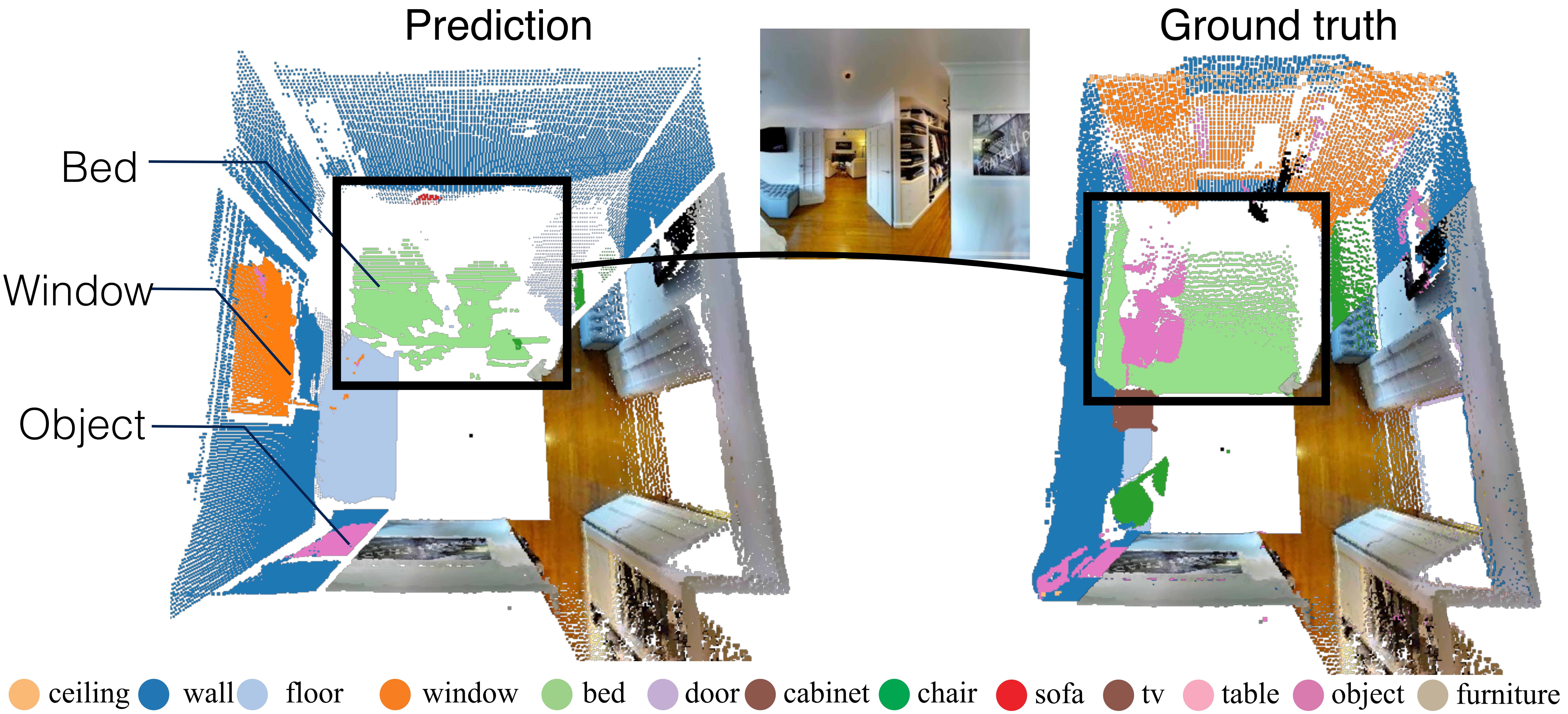



Semantic labels with highest probability per pixel

# Results

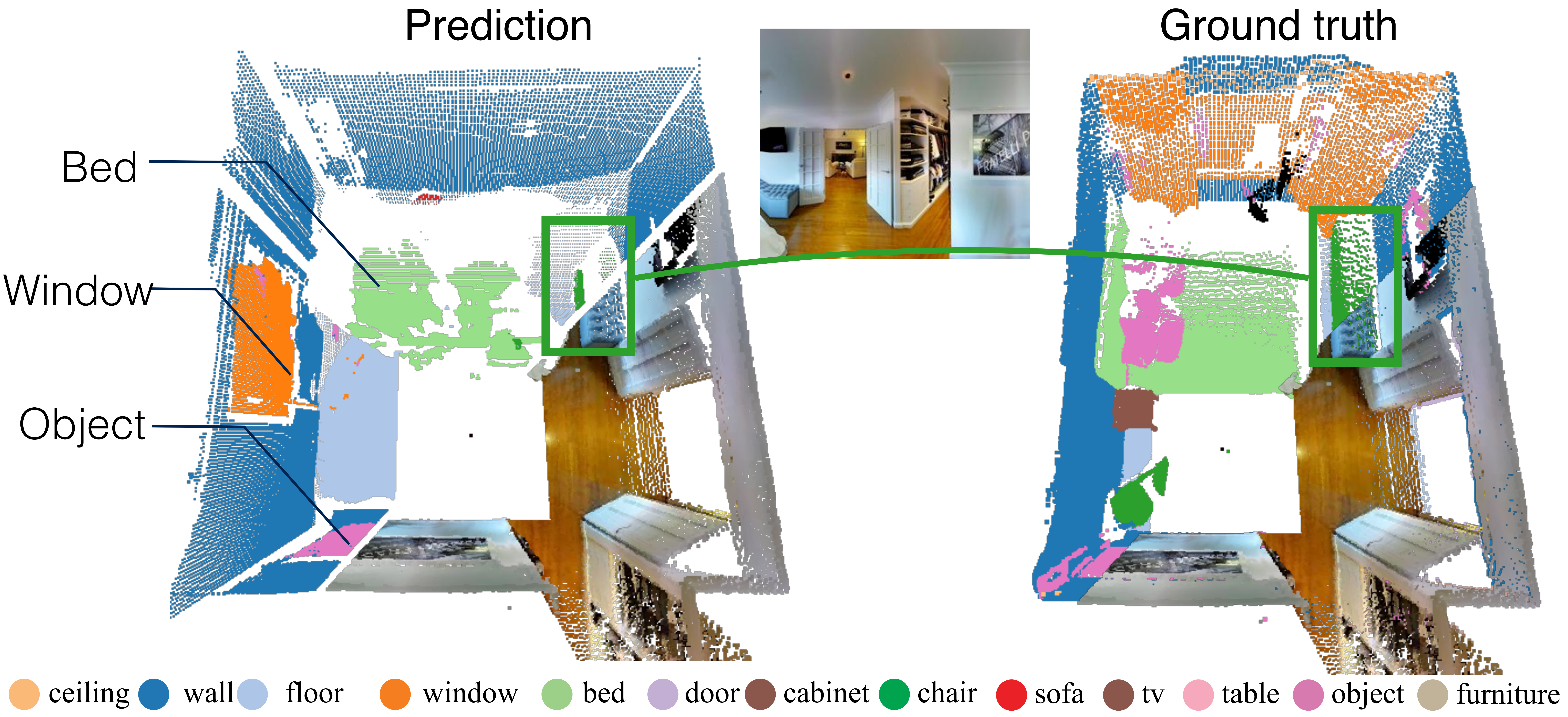

Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results



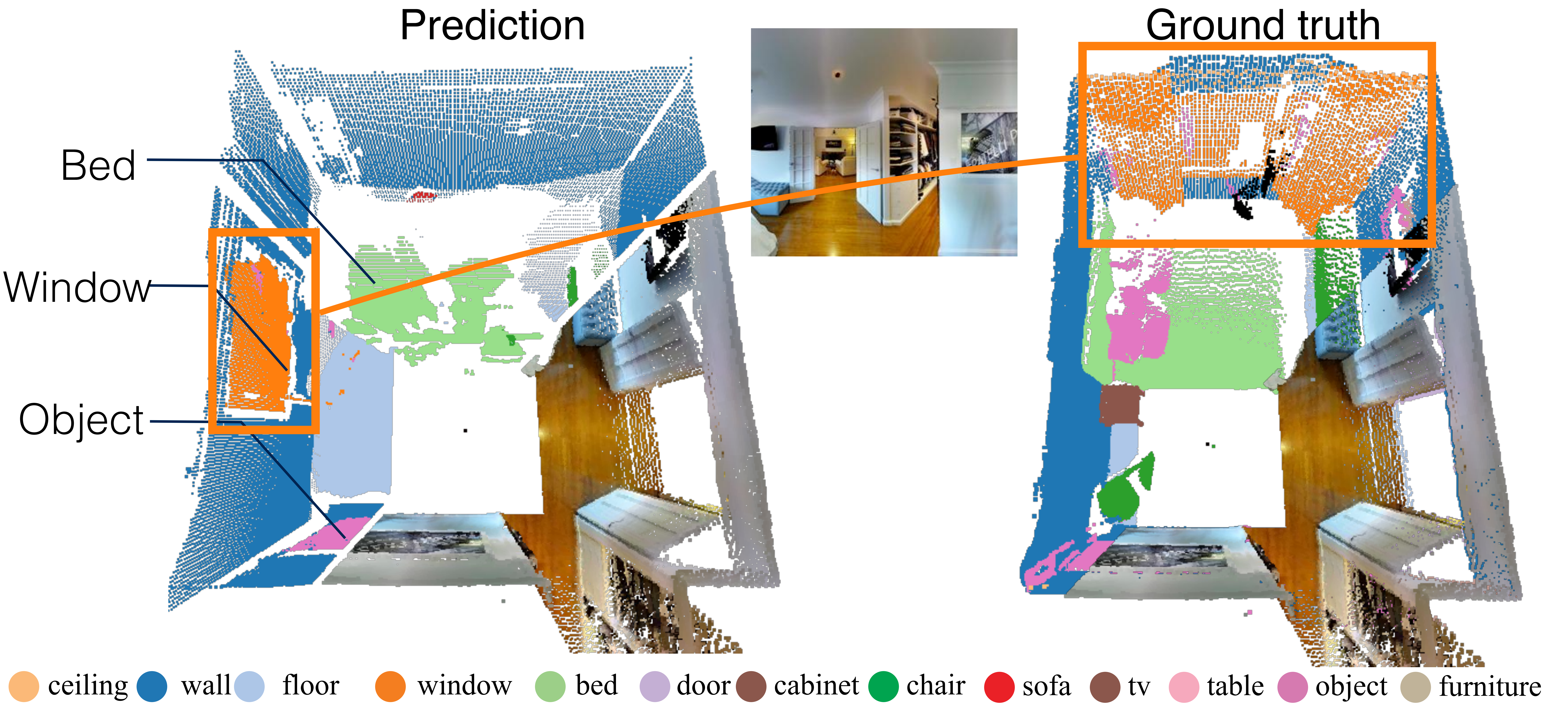
Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results

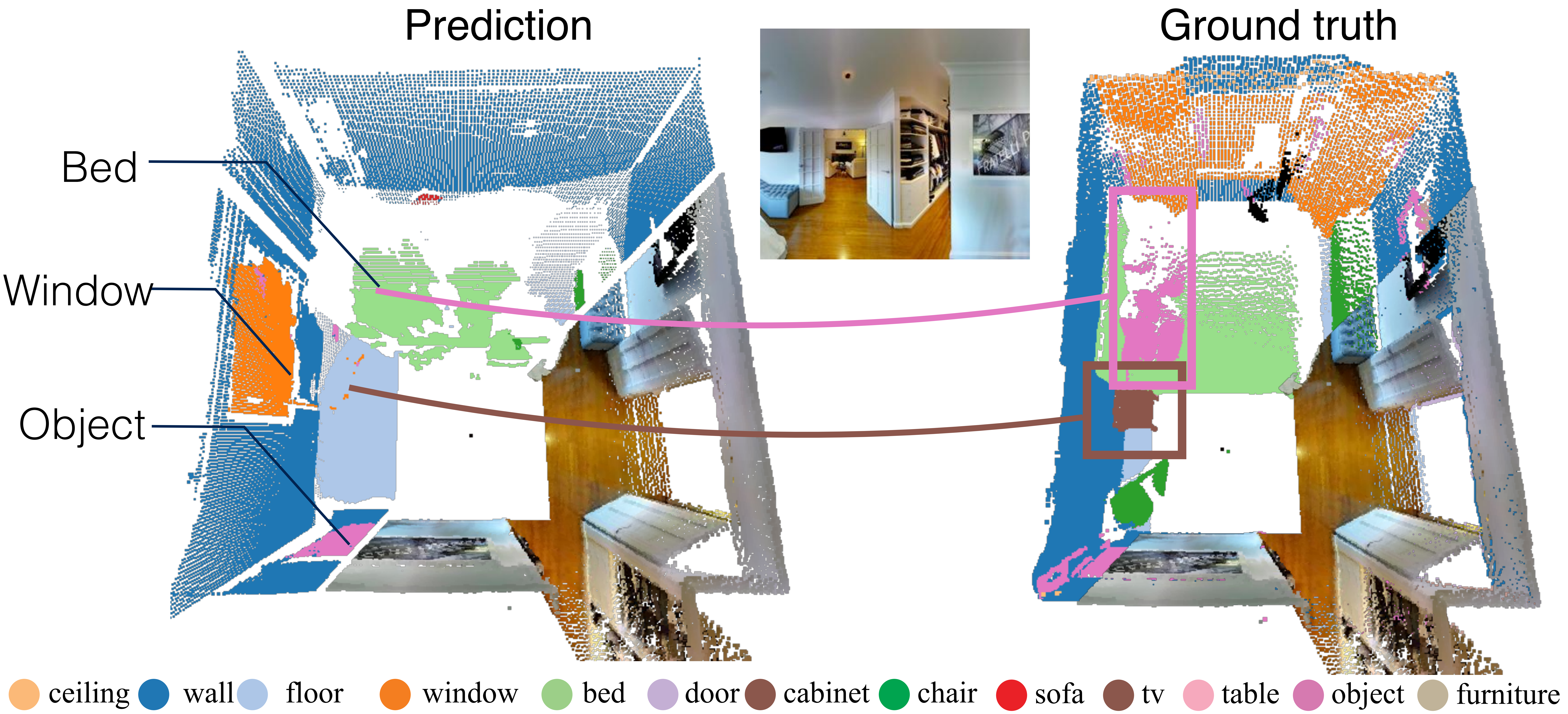


Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results

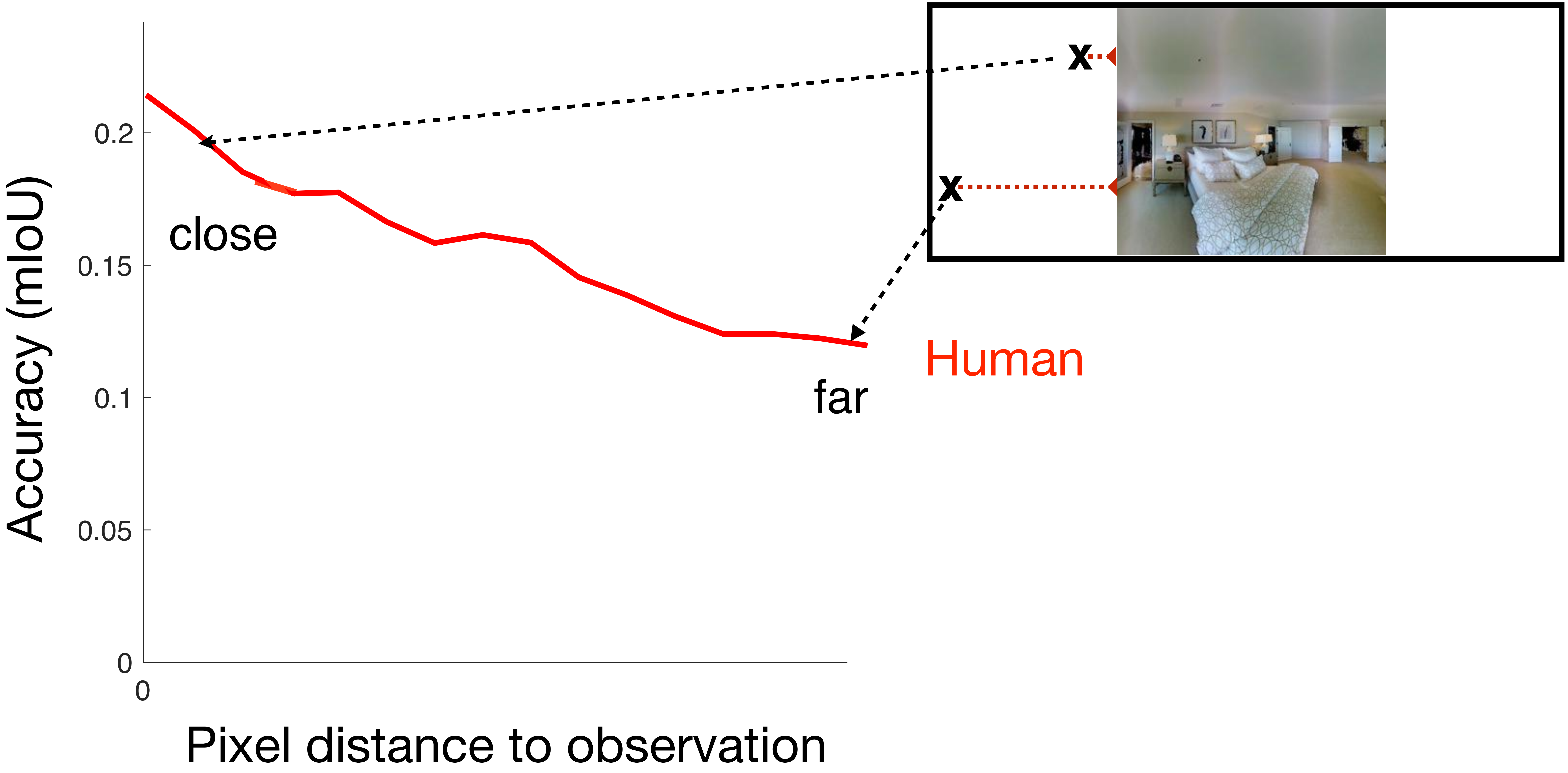

Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# Results

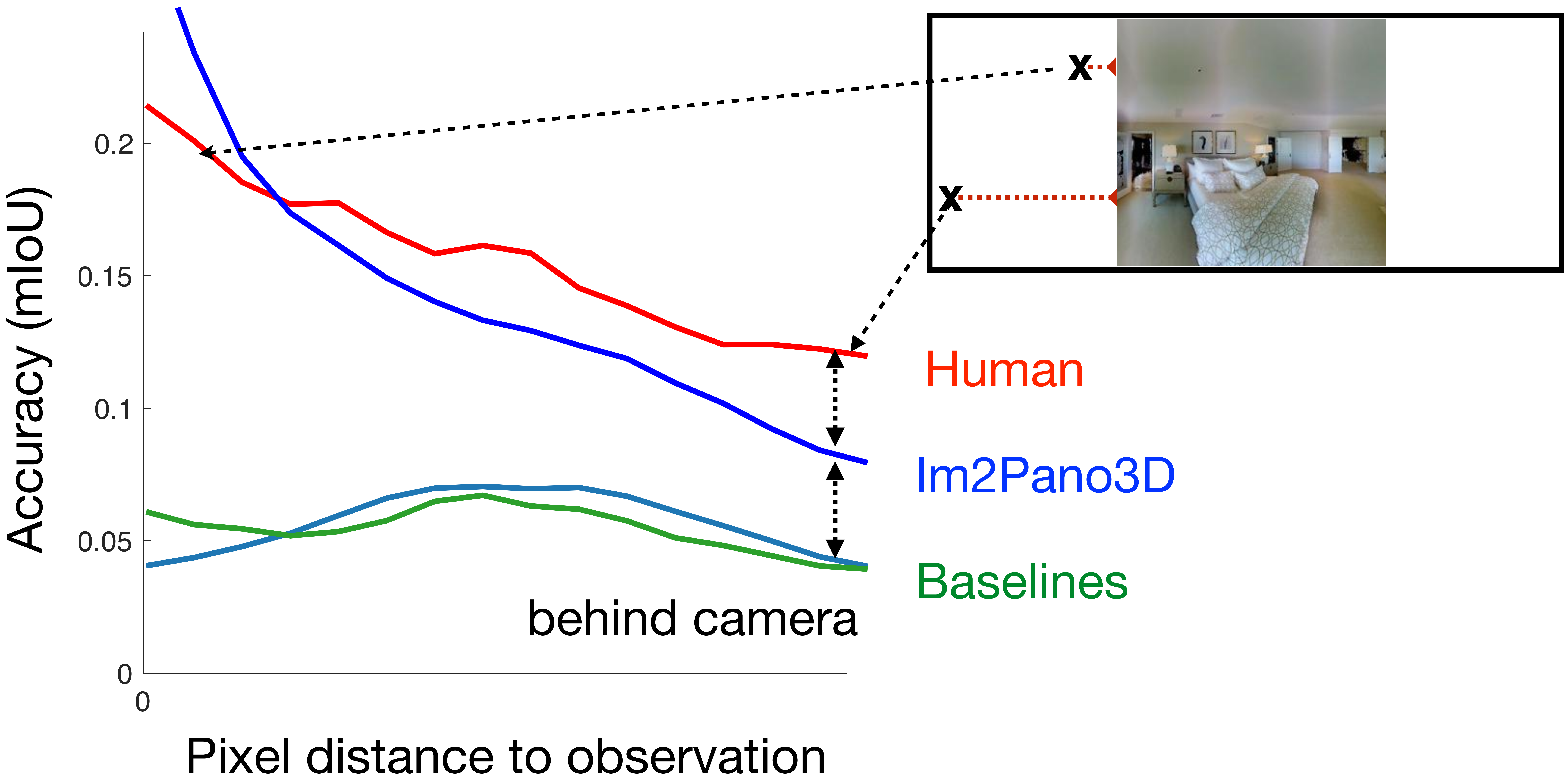

Prediction
Ground truth
Bed
Window
Object
ceiling
wall
floor
window
bed
door
cabinet
chair
sofa
tv
table
object
furniture

# How do we compare to people?

# How do we compare to people?

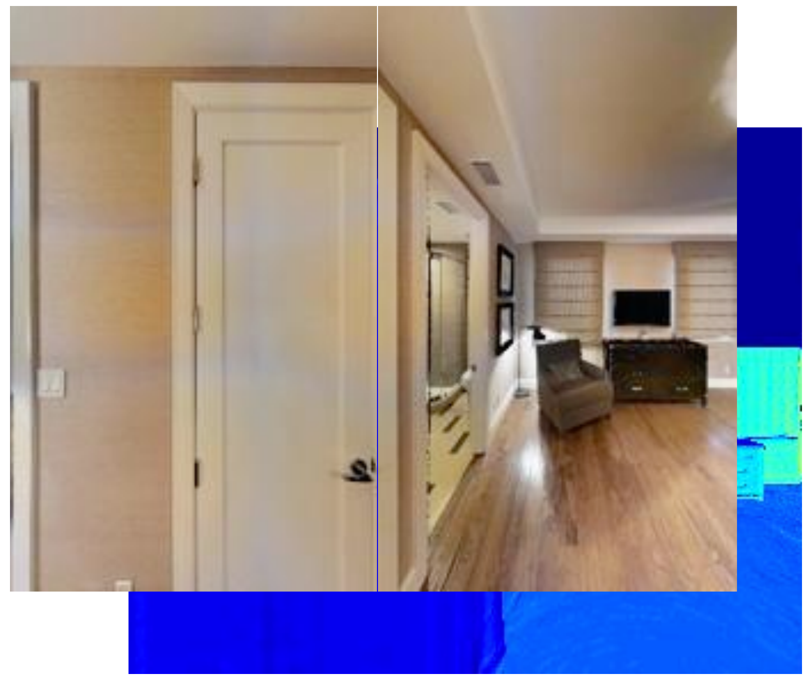


# How do we compare to people?

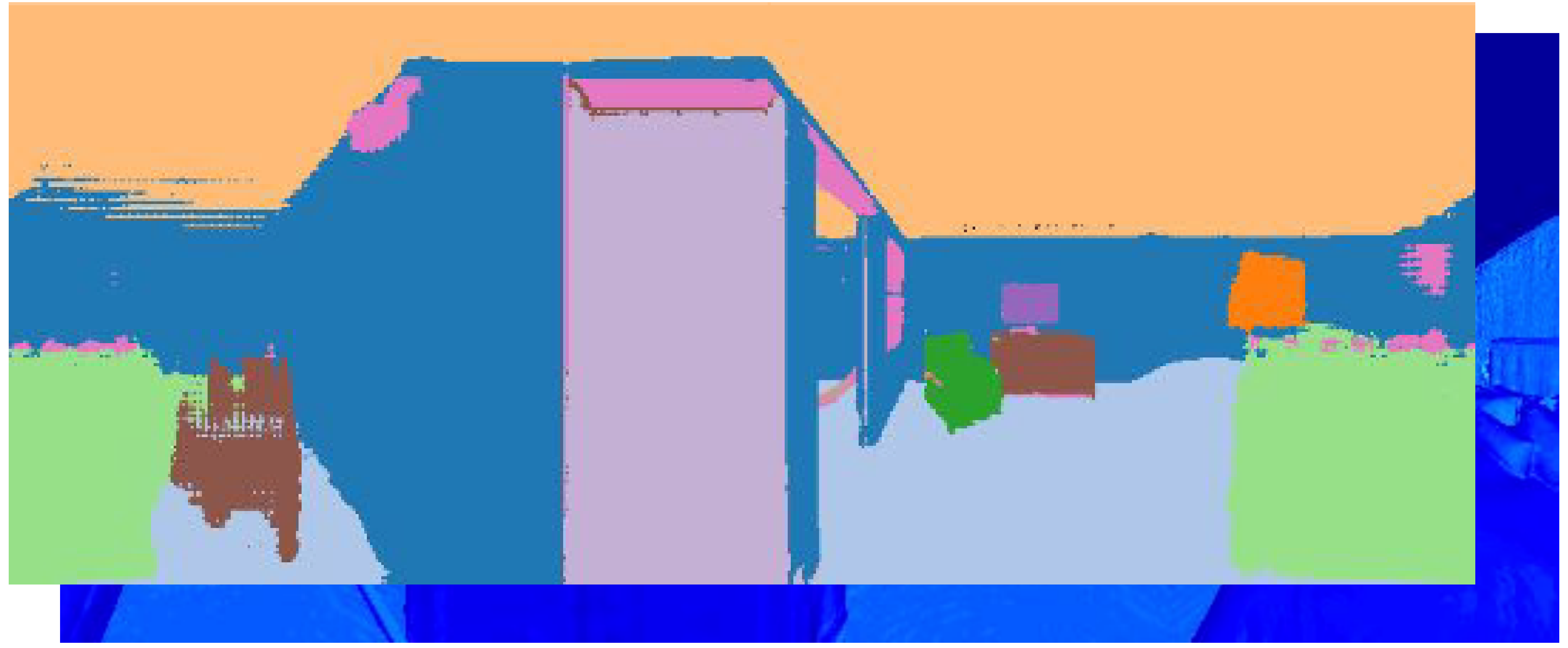


# Conclusion

**New Task:** Semantic-Structure View Extrapolation



**Im2Pano3D**



**Contextual priors** Two large-scale house level datasets

**Geometric priors** 3D plane equation representation

**Meaningful supervision** Multi-level loss functions

**Code & Data:** [im2pano3d.cs.princeton.edu](im2pano3d.cs.princeton.edu)

# Im2Pano3D:

## Extrapolating 360° Structure and Semantics Beyond the Field of View

Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, and Thomas Funkhouser