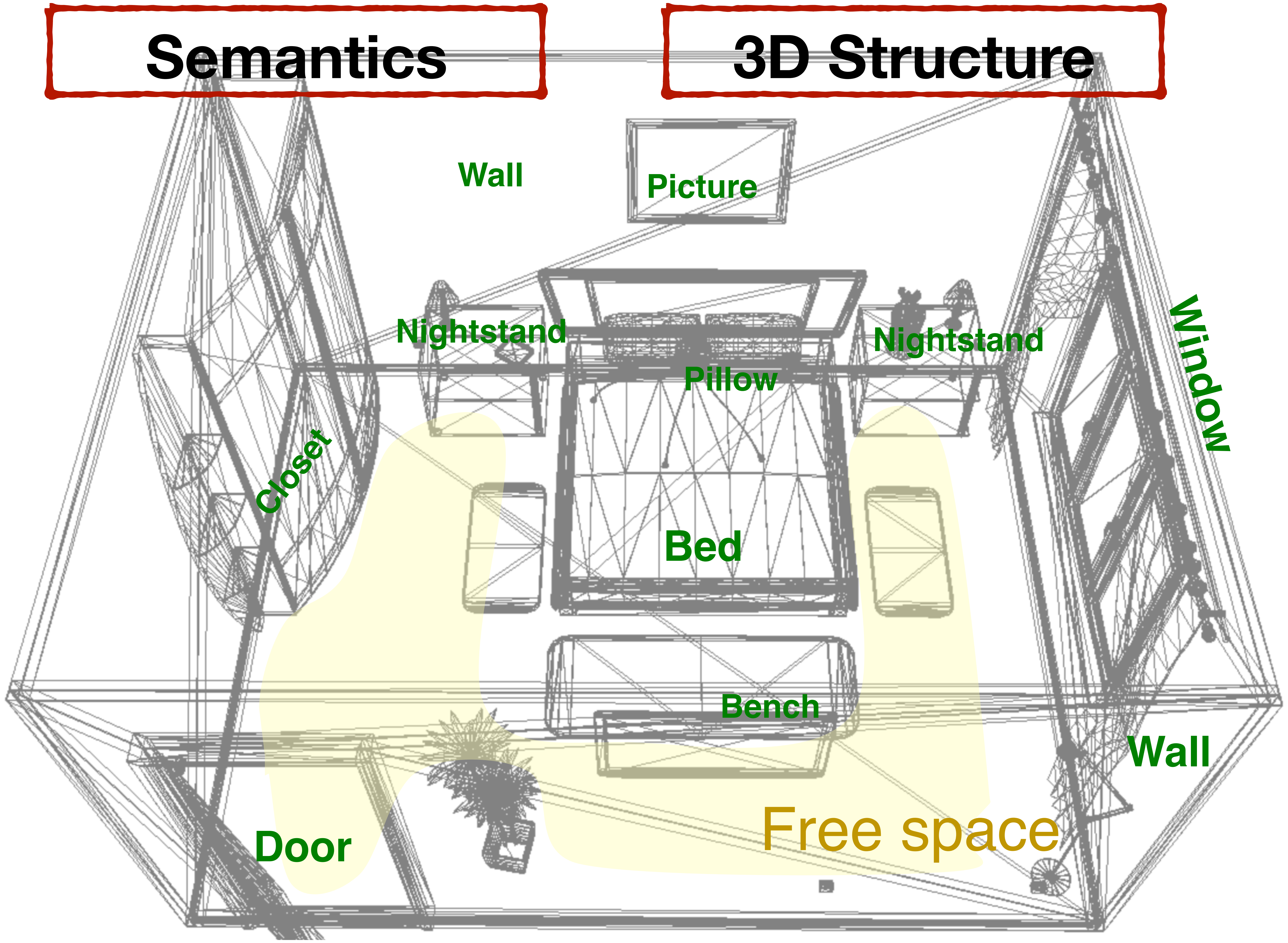
Comprehensive 3D Scene Understanding Beyond the Field of View

Shuran Song

Princeton —> Google —> Columbia

Comprehensive 3D Scene Understanding

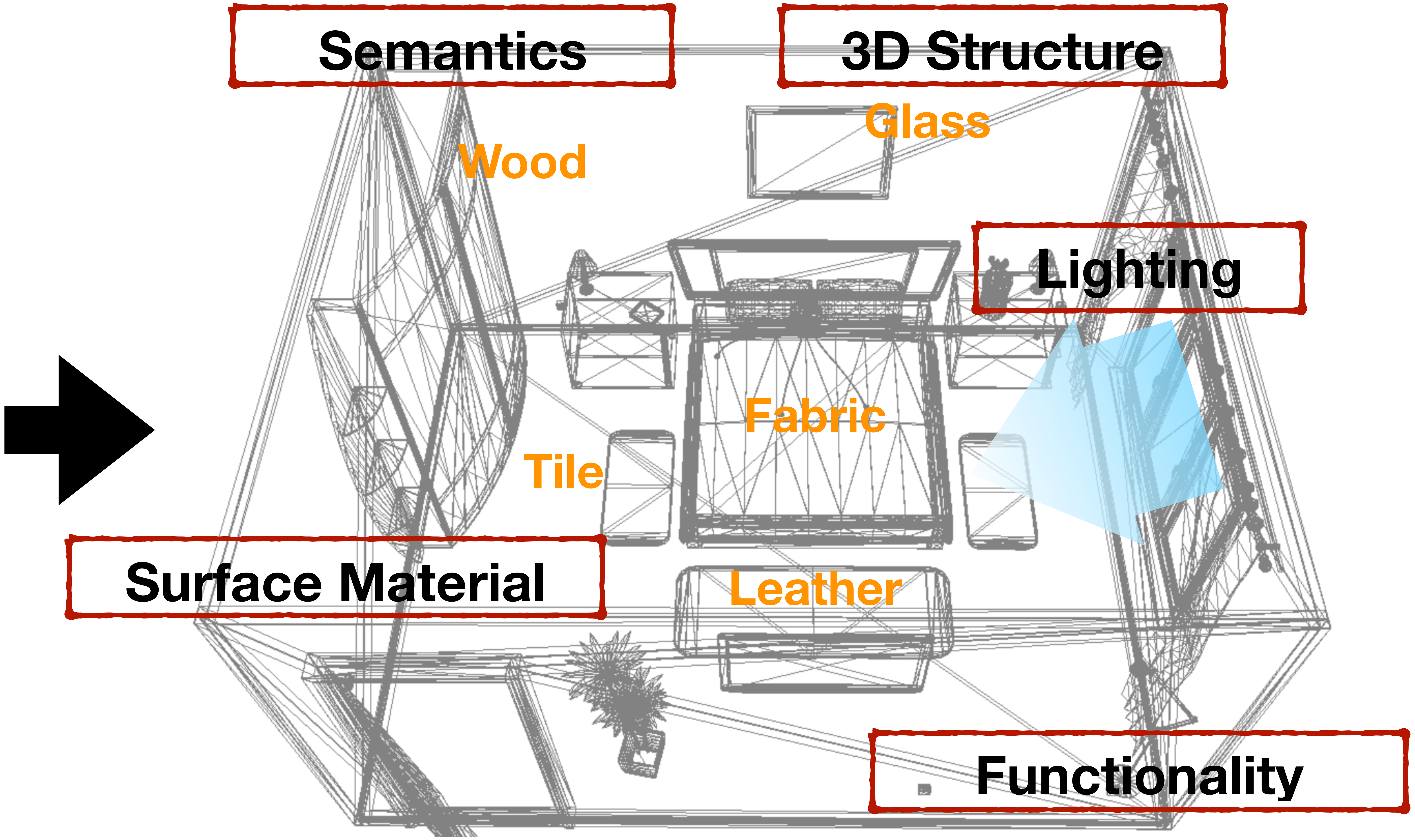
Partial Observation of the Environment



Complete Representation of the 3D Scene

Comprehensive 3D Scene Understanding

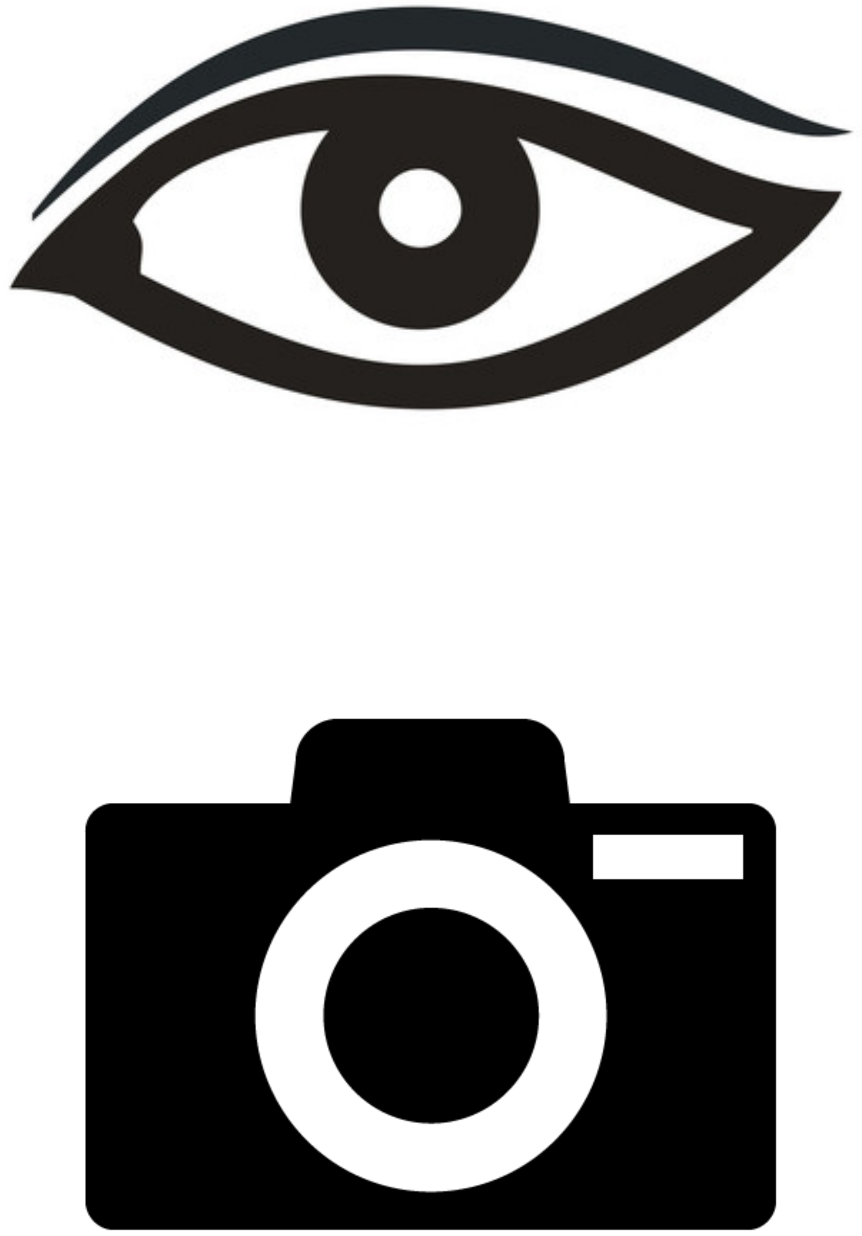
Partial Observation of the Environment



Complete Representation of the 3D Scene

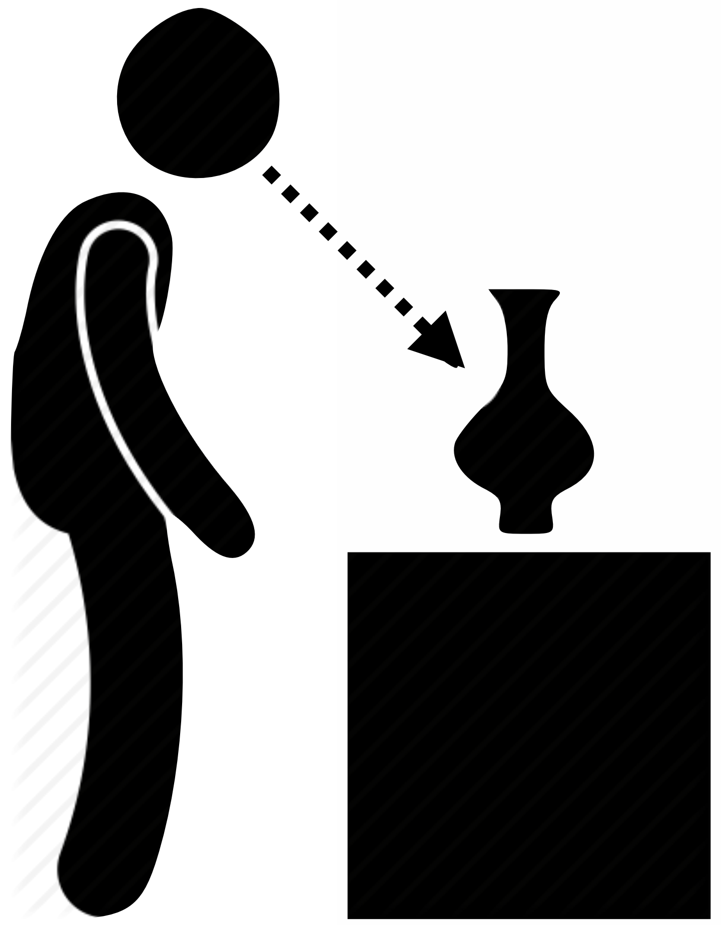
Challenges: Partial Observation

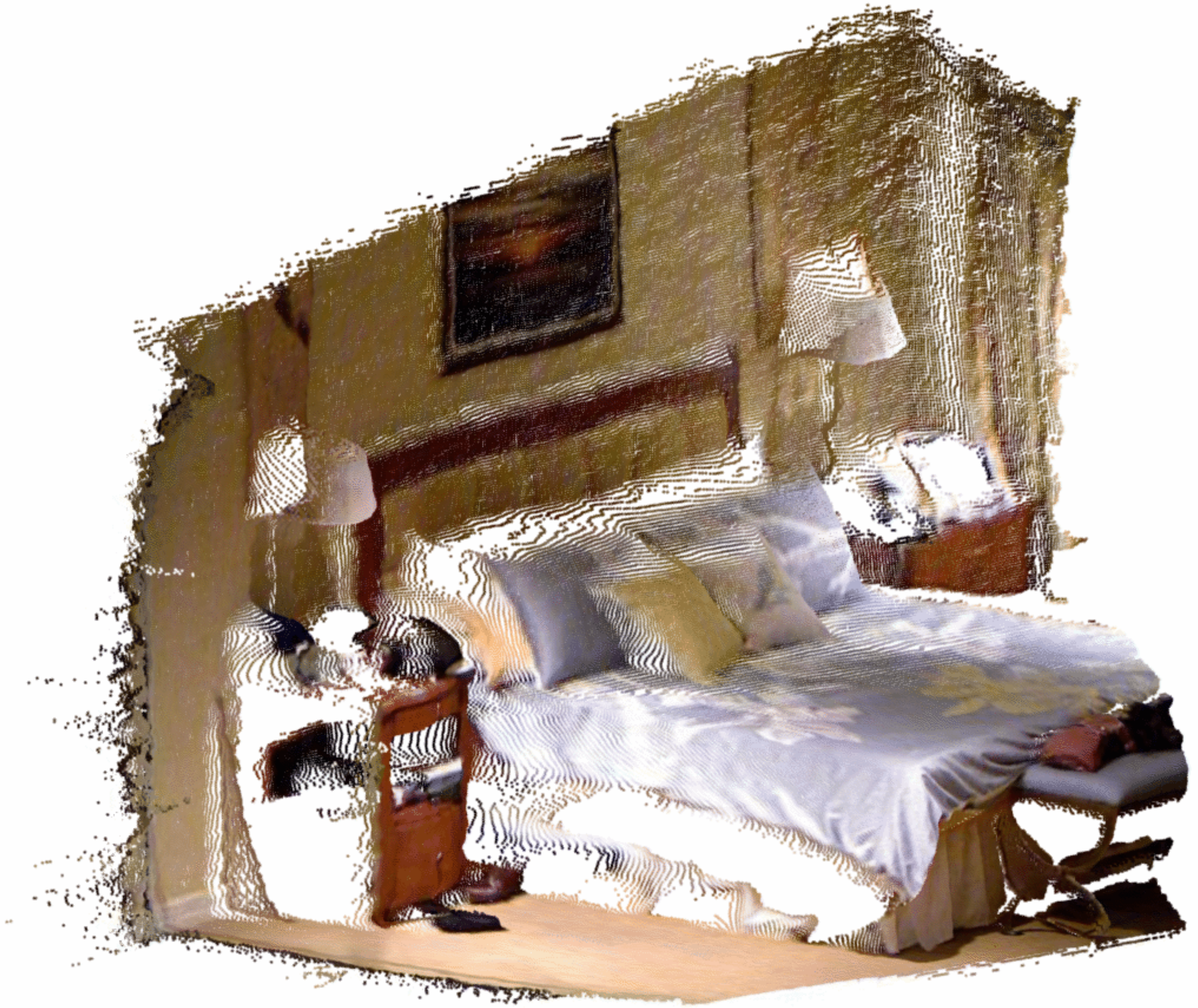
Challenges: Partial Observation



Sensors

Challenges: Partial Observation

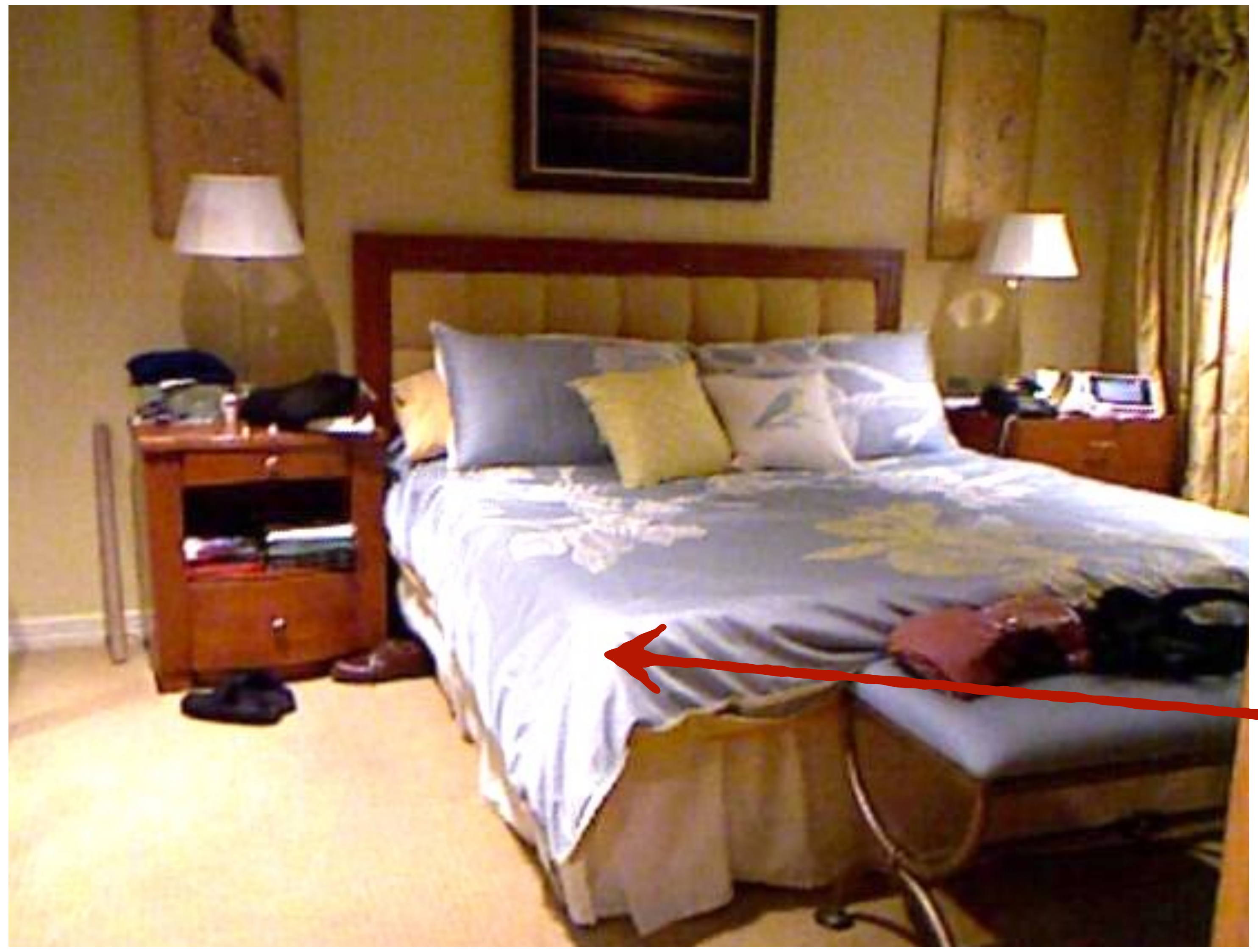




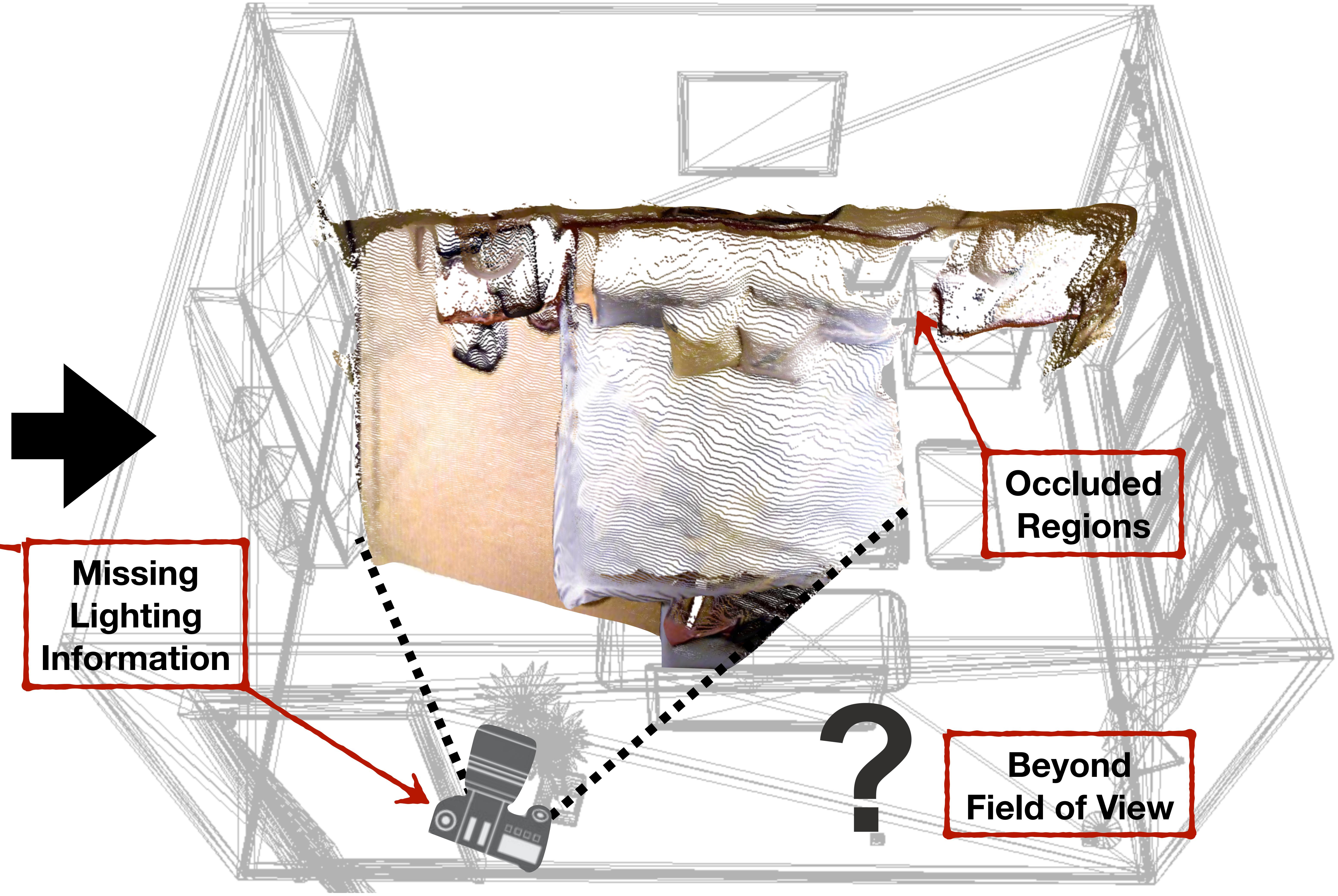
Sensors

Partial Observation

Challenges: Partial Observation



Partial Observation of the Environment



Missing Lighting Information

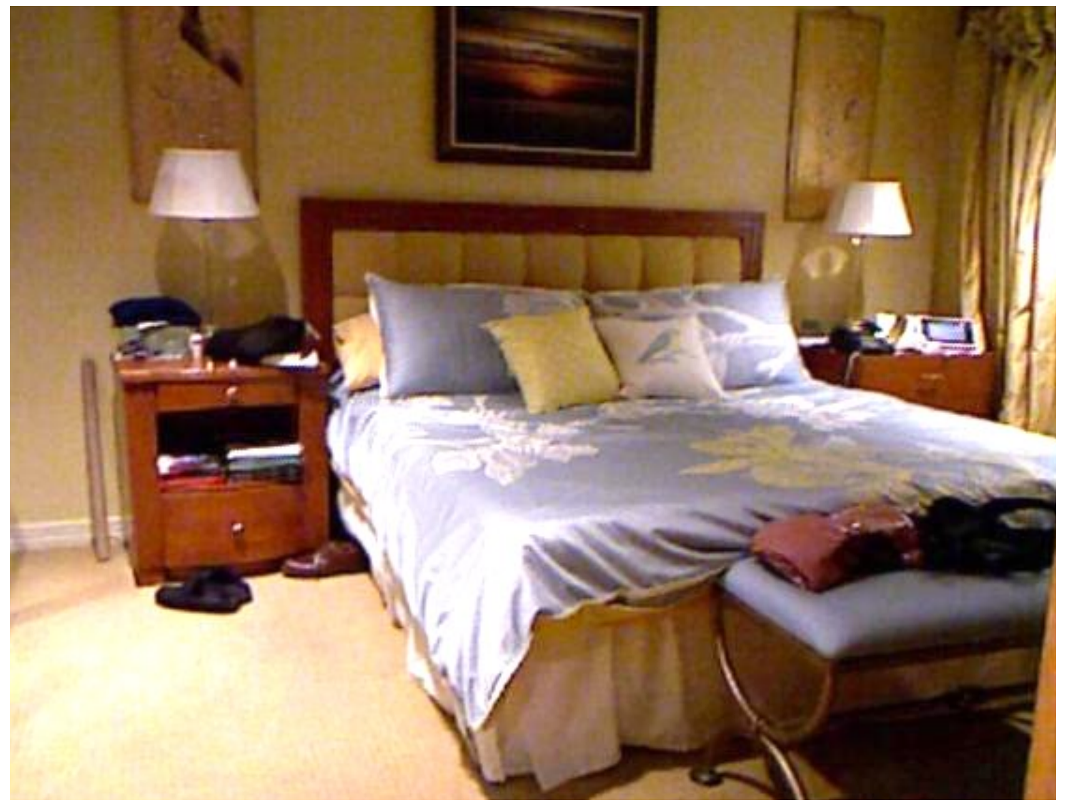
Occluded Regions

Beyond Field of View

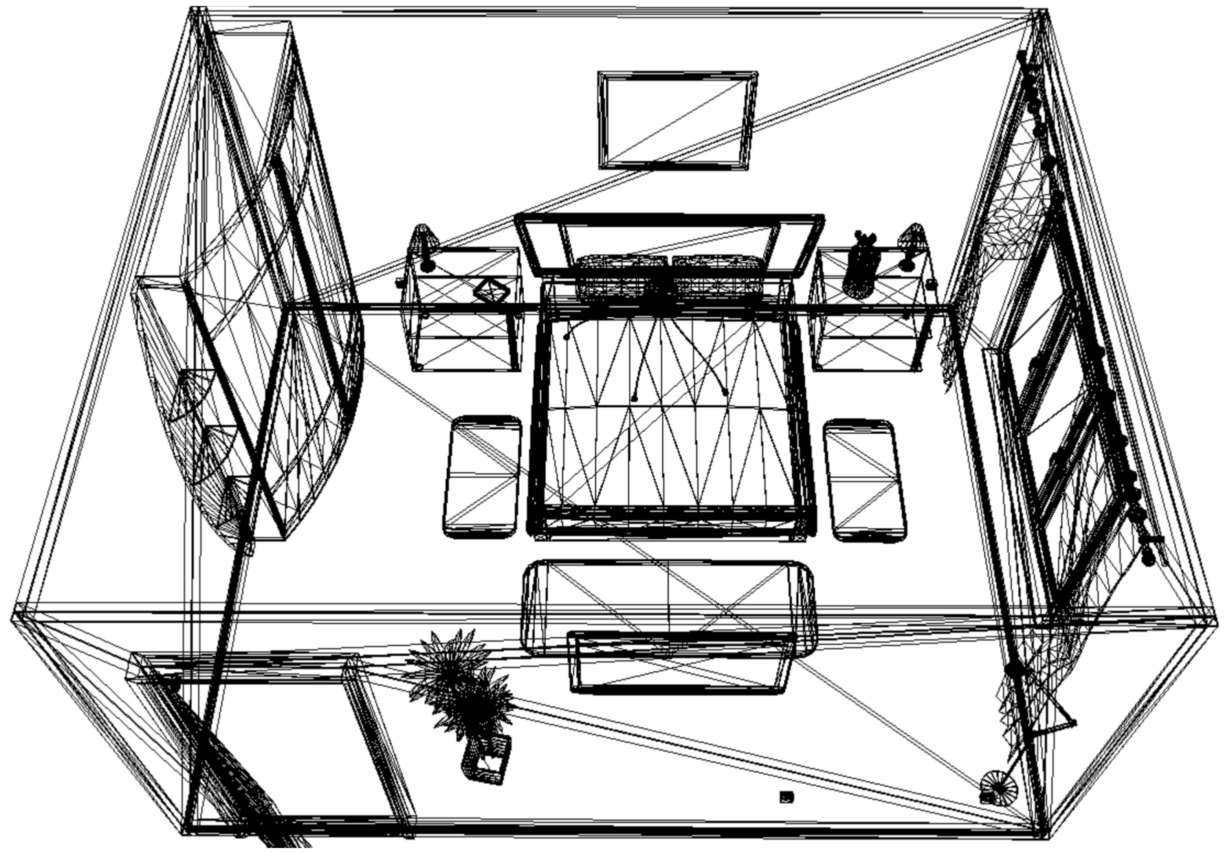
Top-down View

Advances Towards 3D Scene Understanding

Partial Observation



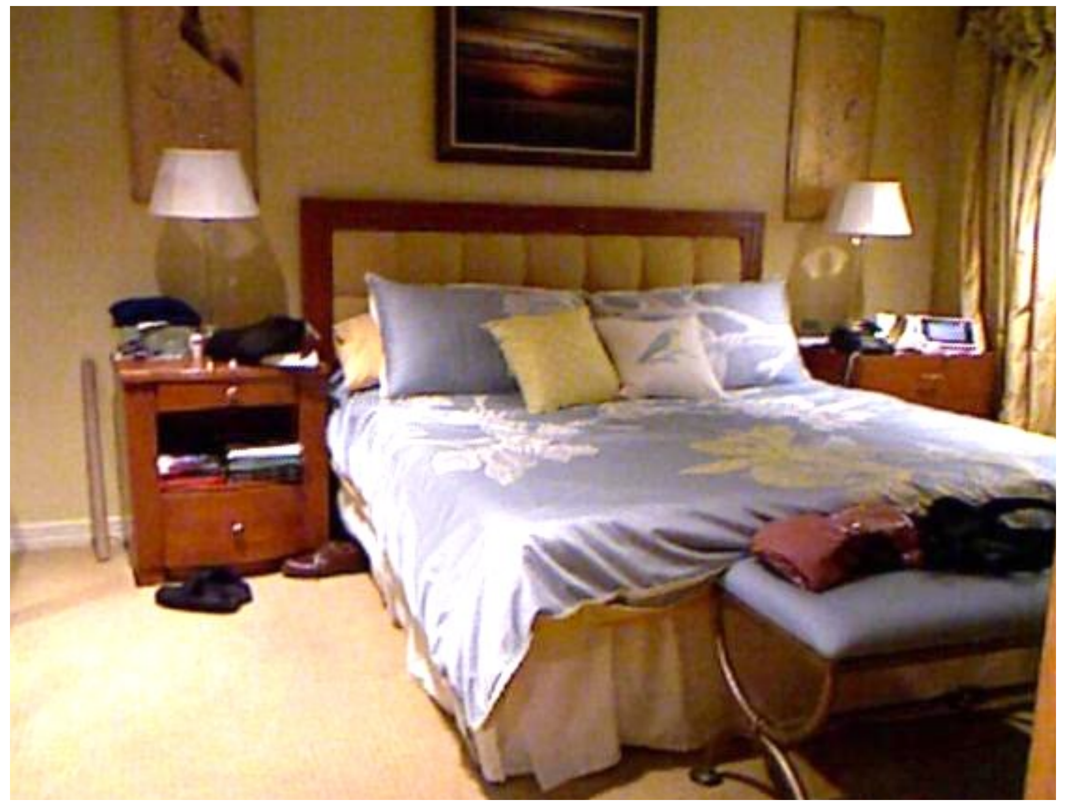
Complete 3D Scene

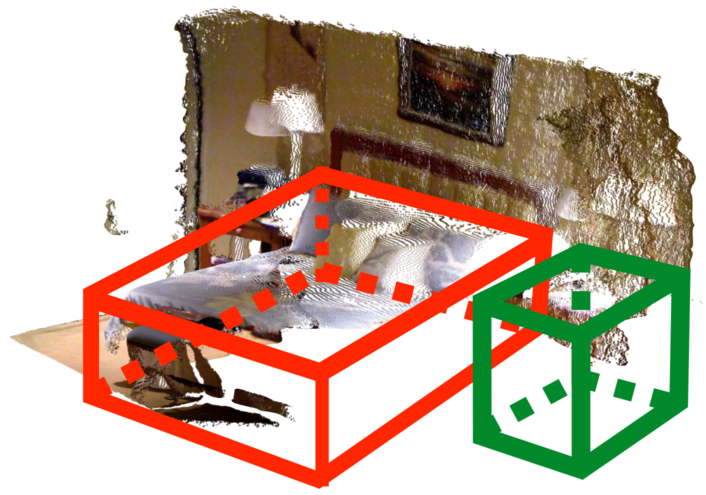


- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

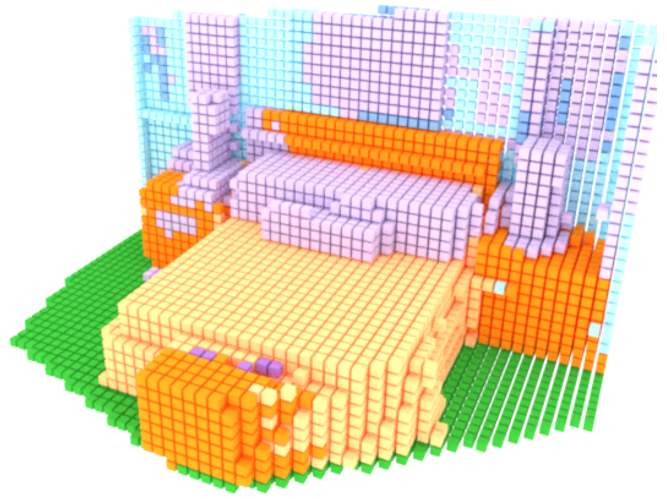
Advances Towards 3D Scene Understanding

Partial Observation

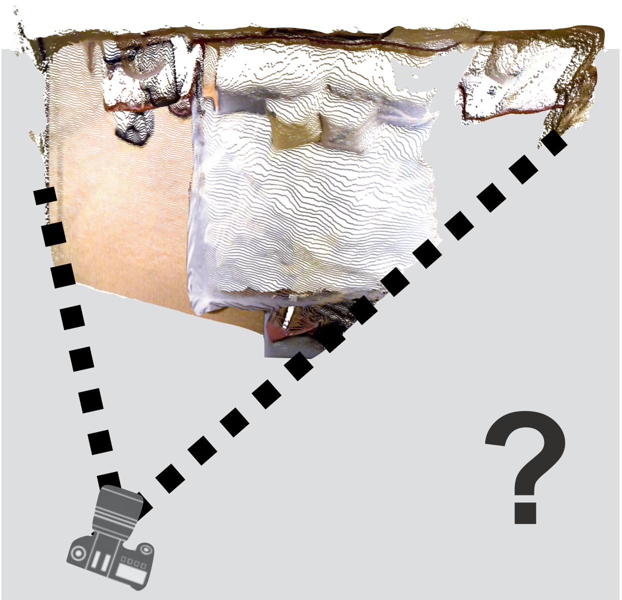




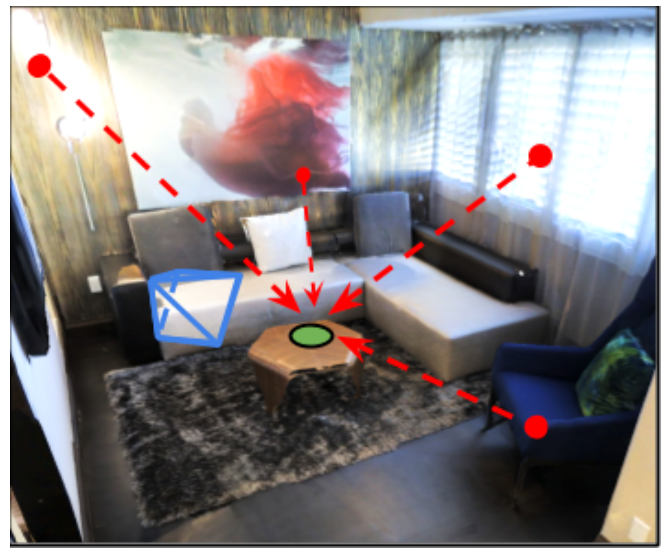
Amodal 3D Bounding Boxes
[Song and Xiao ECCV’14,CVPR’16]



Higher Fidelity 3D Voxels
[Song et al. CVPR’17]

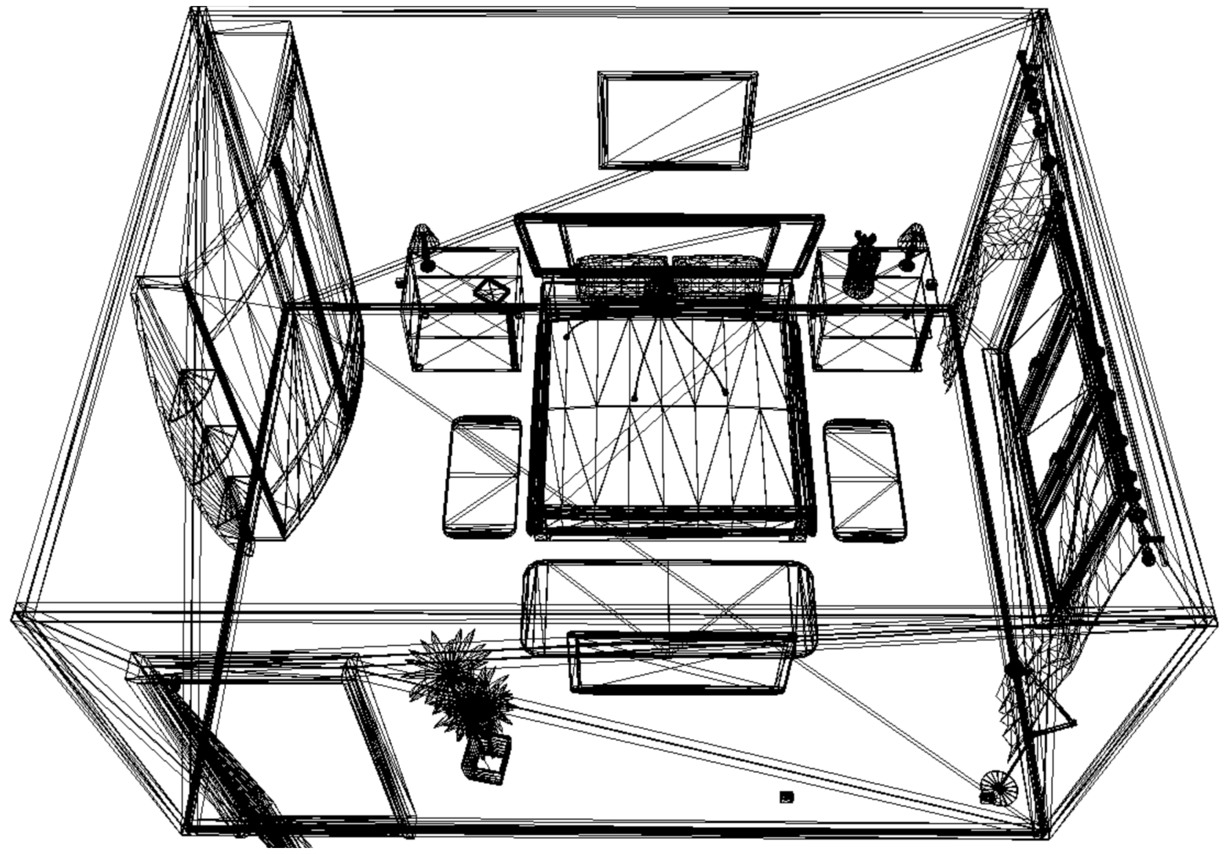


Beyond FoV Semantics&Structure
[Song et al. CVPR’18]



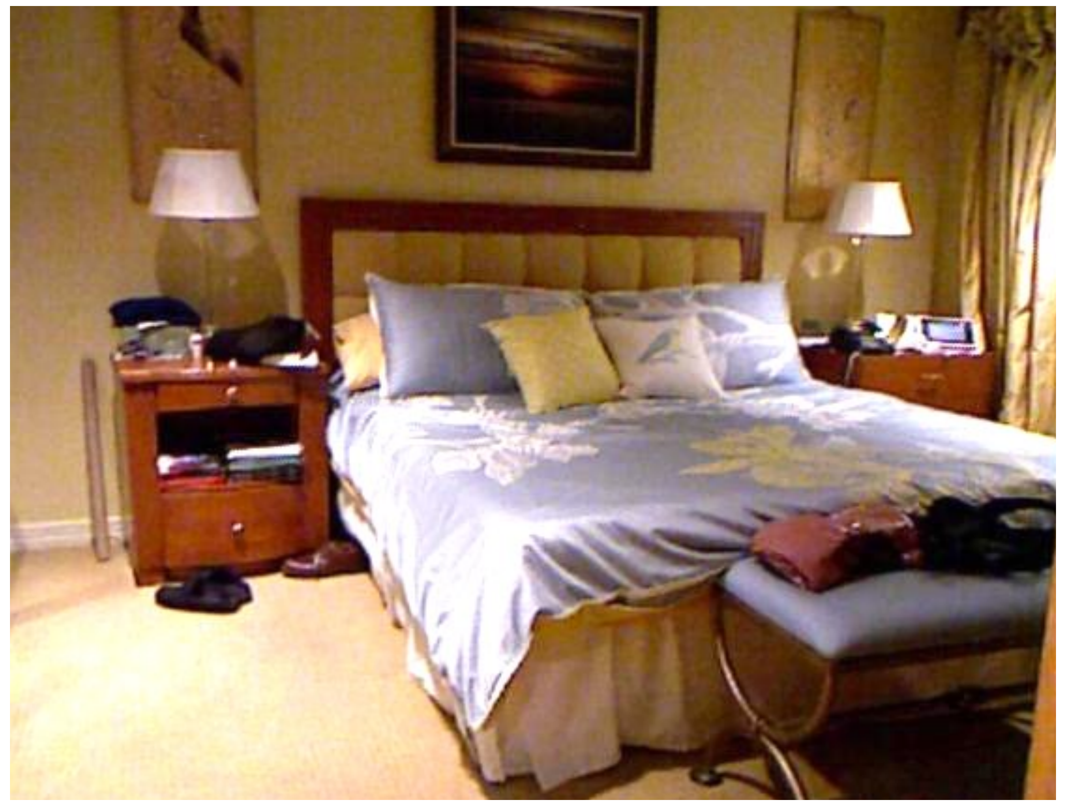
Beyond FoV Illumination
[Song and Funkhouser]

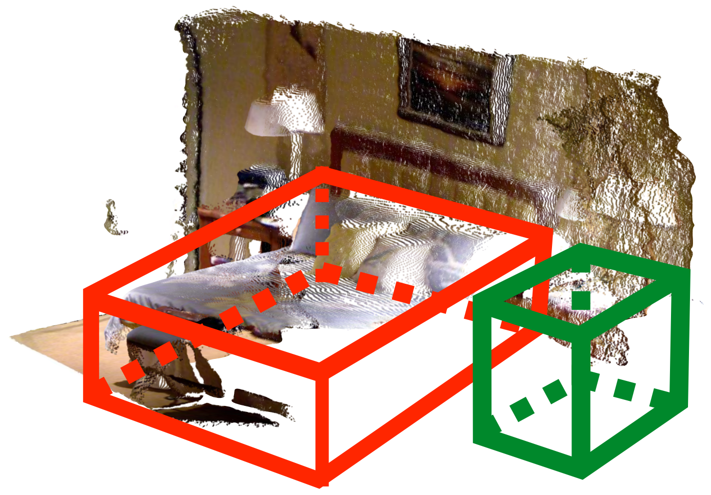
Complete 3D Scene



- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

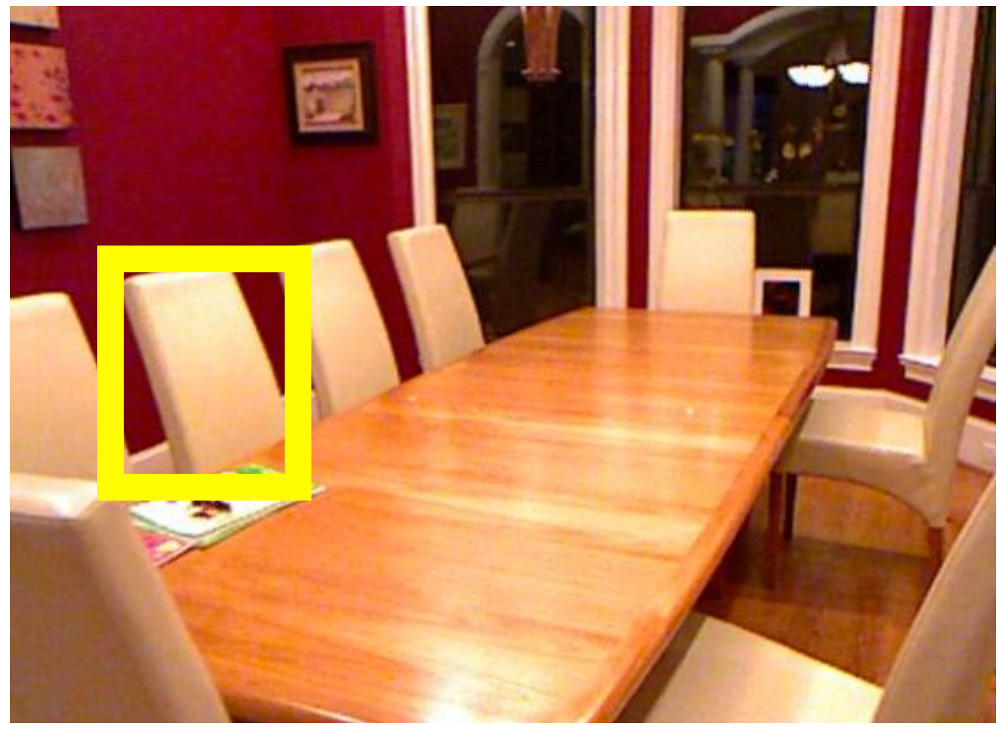
Advances Towards 3D Scene Understanding

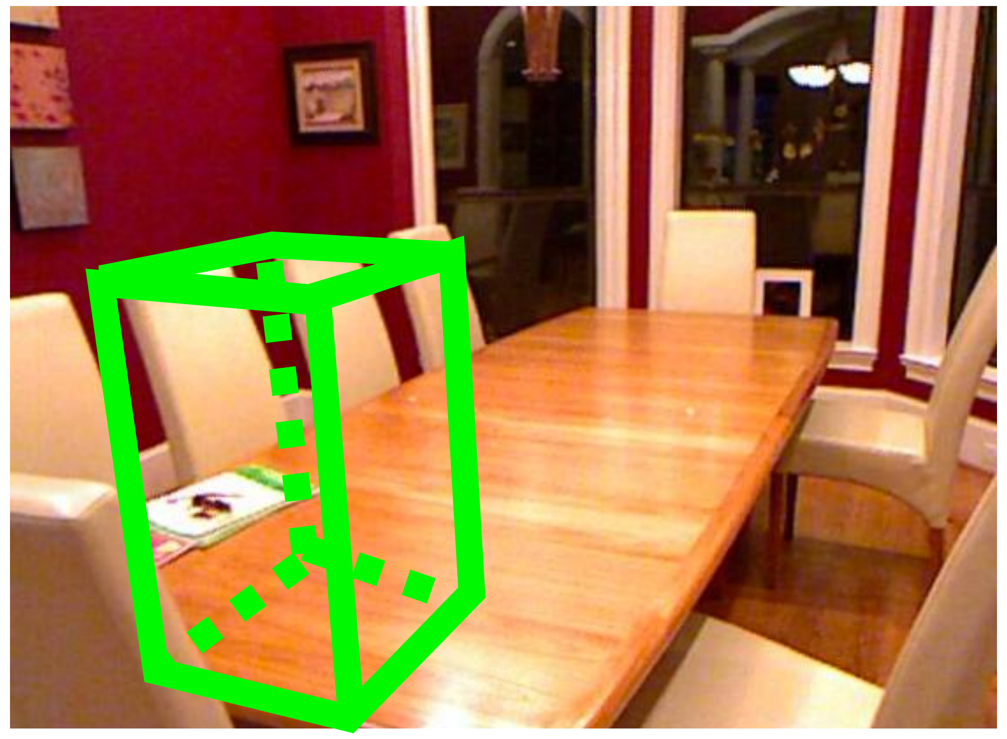




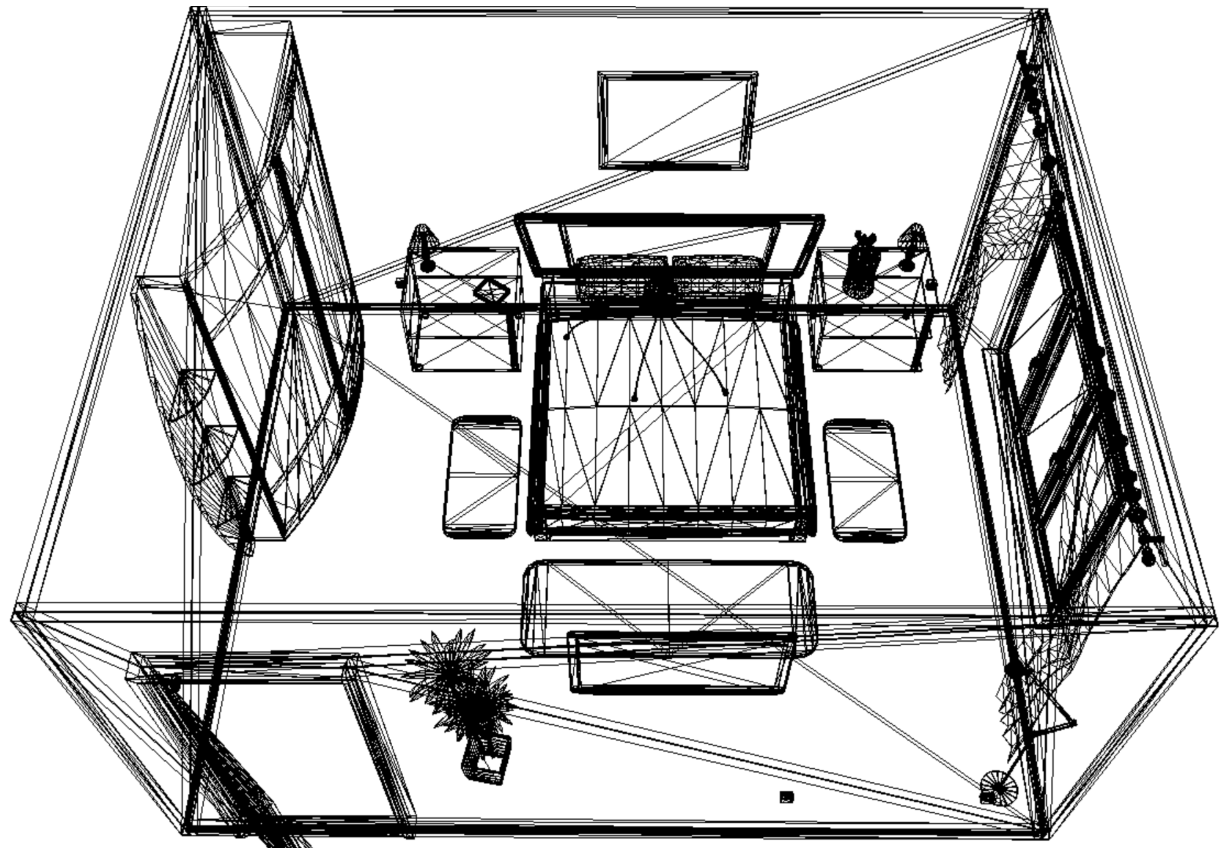
Amodal 3D Bounding Boxes
[Song and Xiao ECCV'14,CVPR'16]

3D Bounding Box track



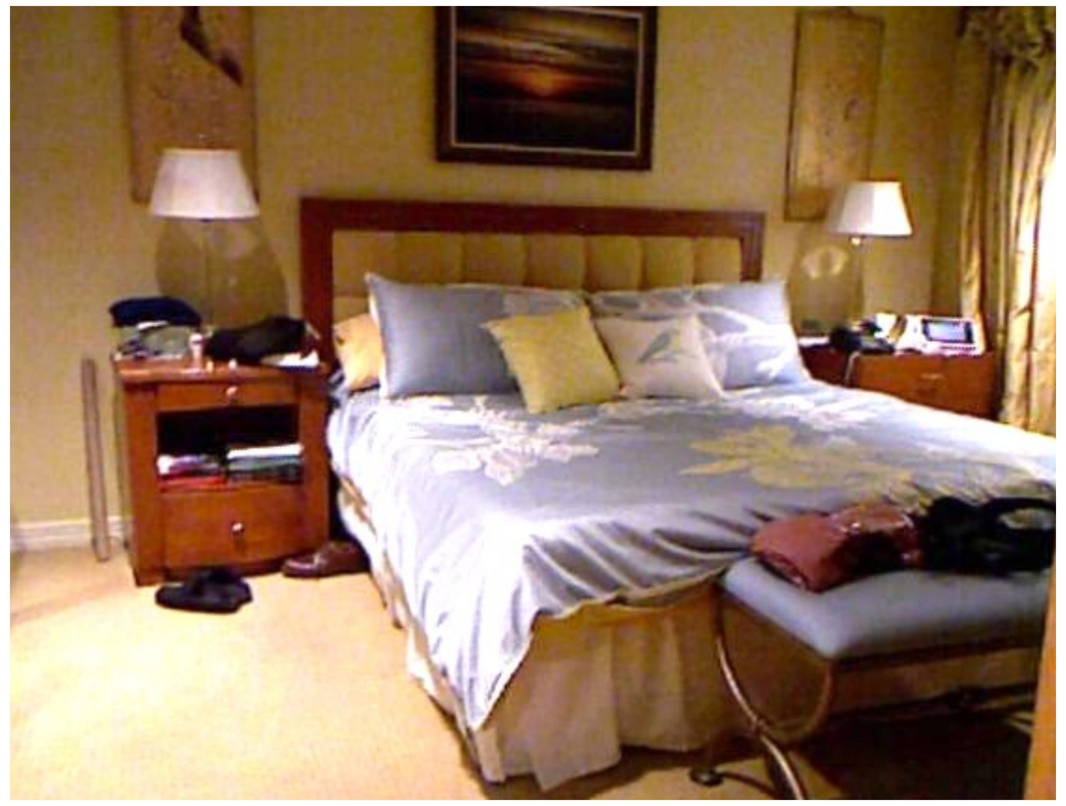


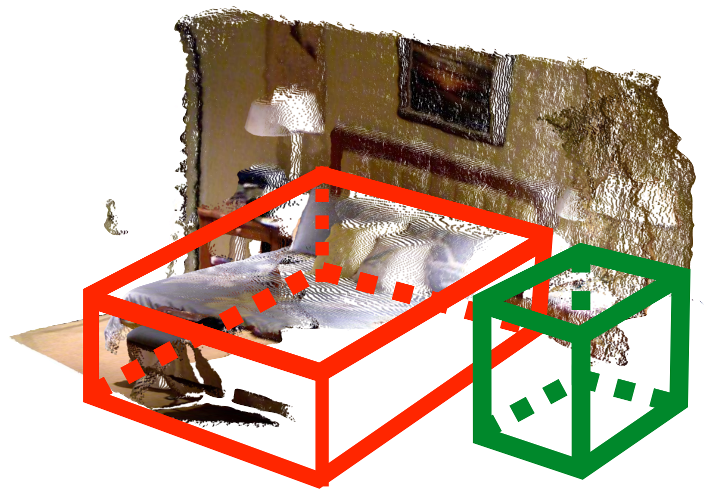
Sliding Shapes [ECCV'14]
Deep Sliding Shapes [CVPR'16]



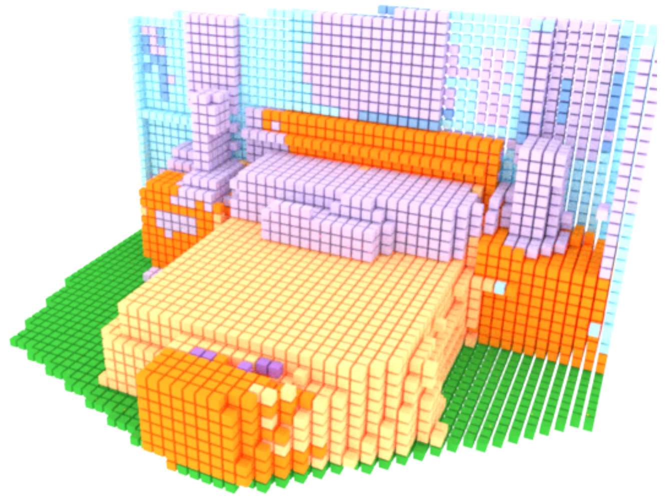
- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

Advances Towards 3D Scene Understanding

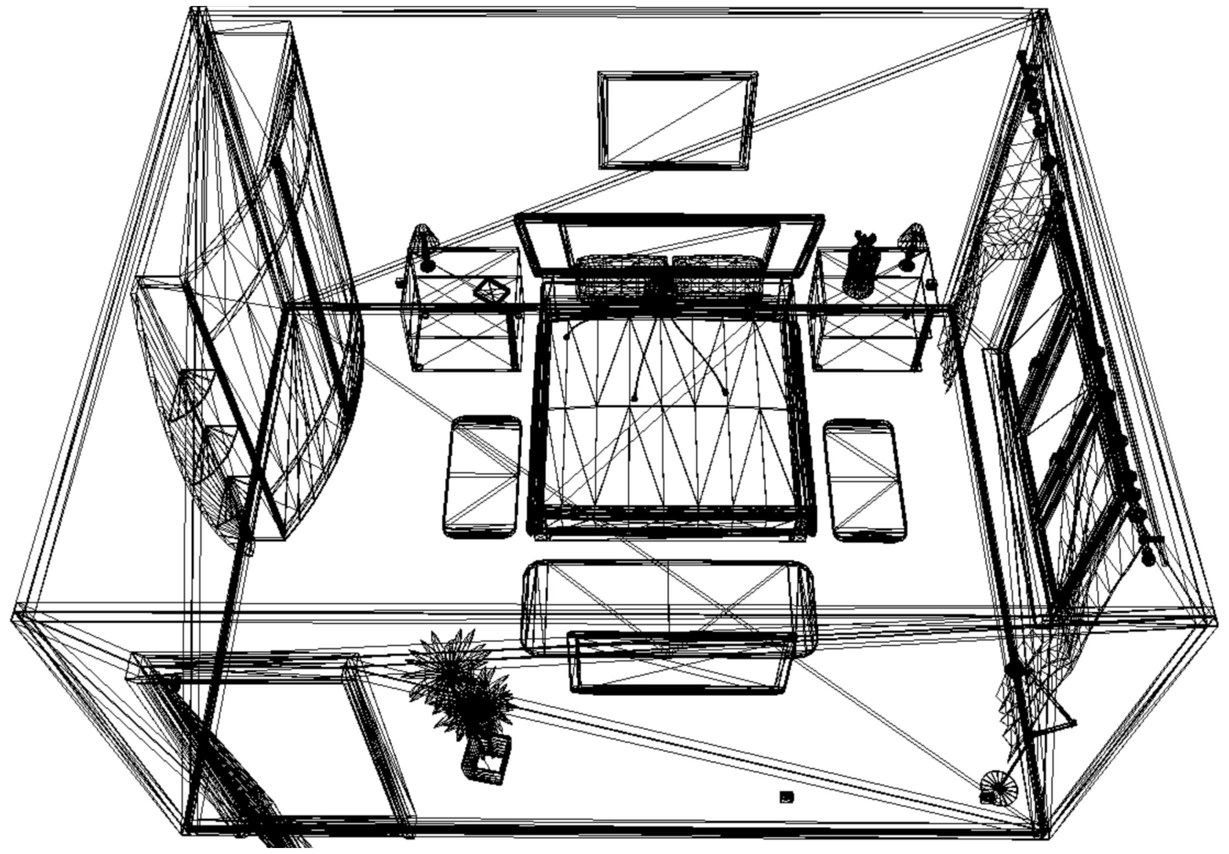




Amodal 3D Bounding Boxes
[Song and Xiao ECCV'14,CVPR'16]

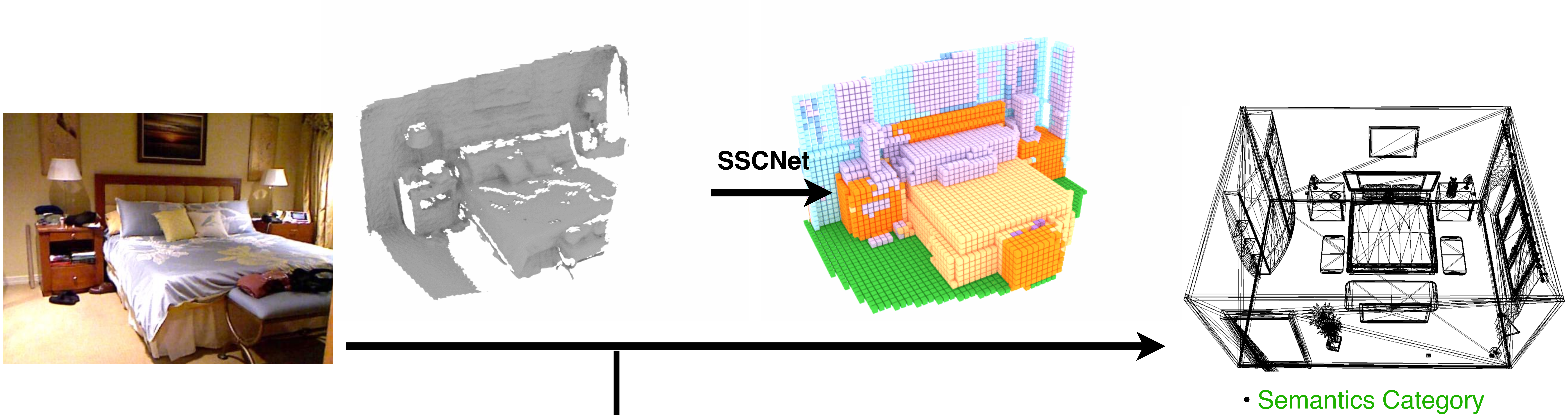


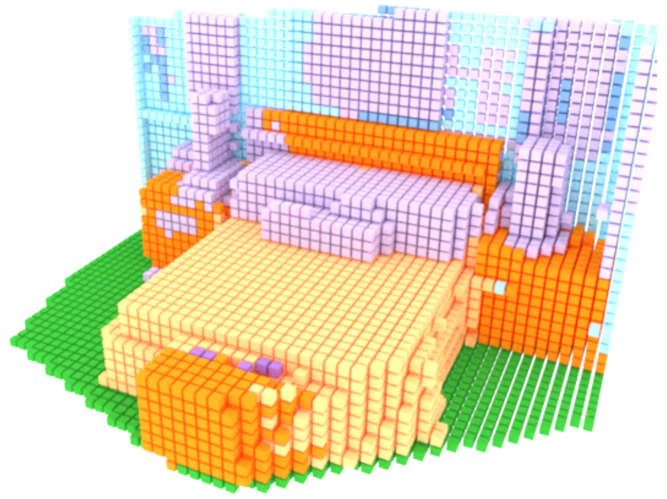
Higher Fidelity 3D Voxels
[Song et al. CVPR'17]



- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

Advances Towards 3D Scene Understanding



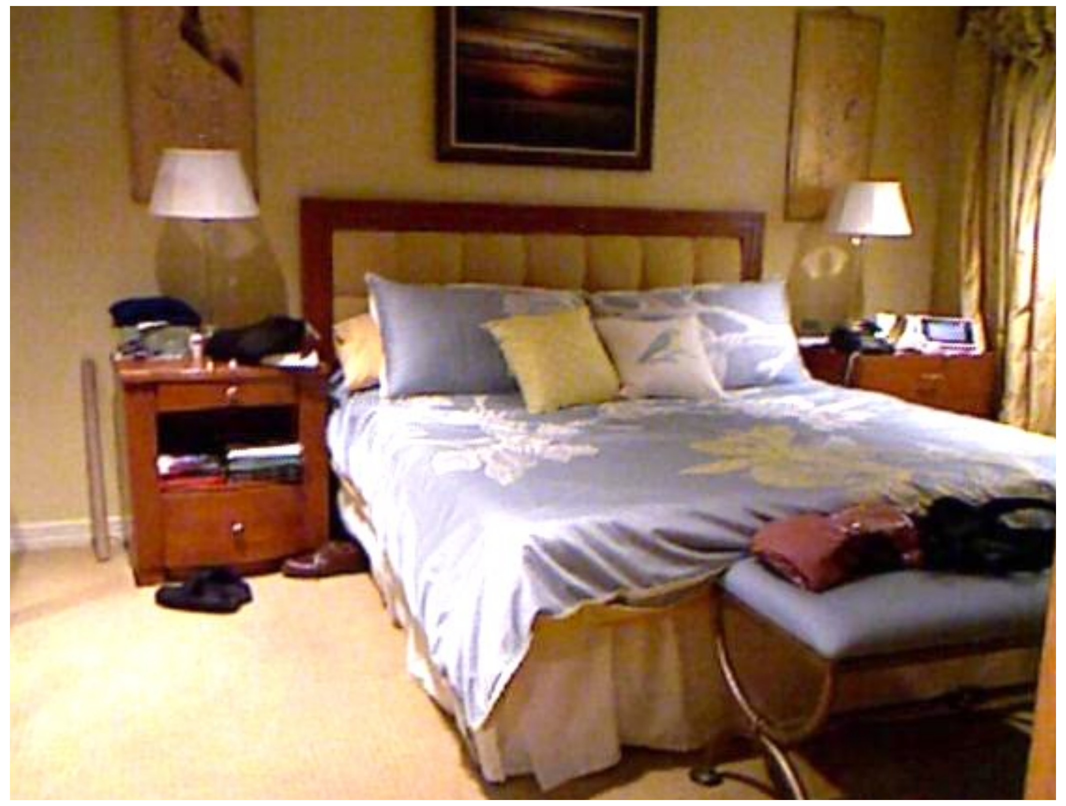


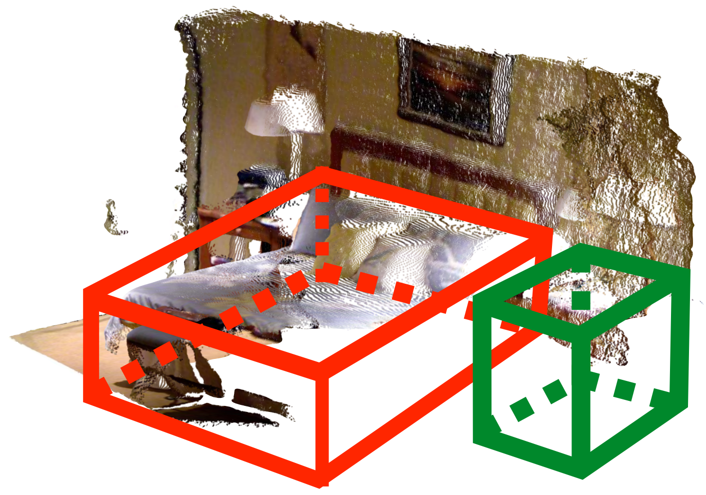
Higher Fidelity
3D Voxels
[Song et al. CVPR'17]

3D Voxel
Grid Track

- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

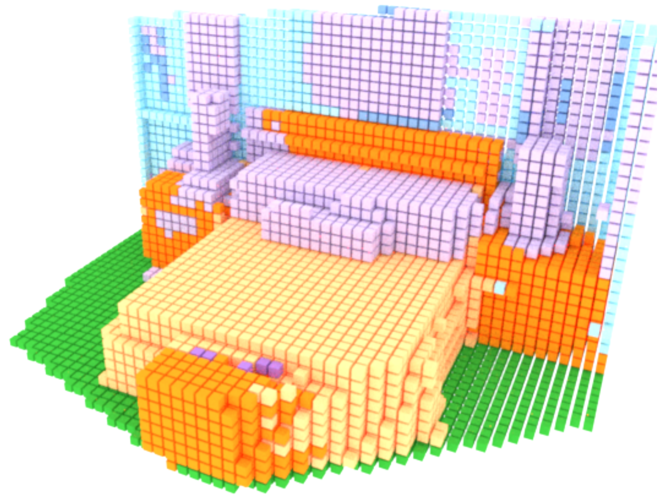
Advances Towards 3D Scene Understanding





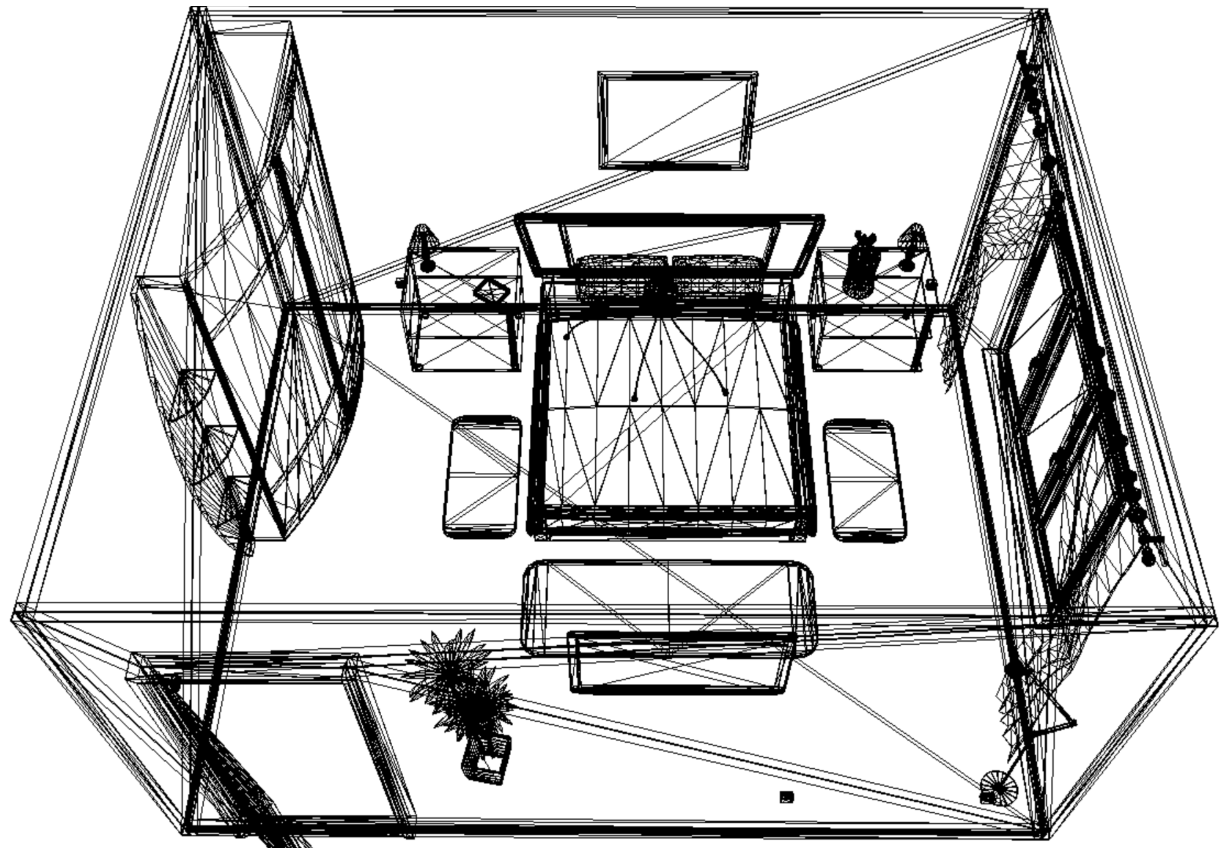
Amodal 3D Bounding Boxes
[Song and Xiao ECCV’14,CVPR’16]

3D Bounding Box track



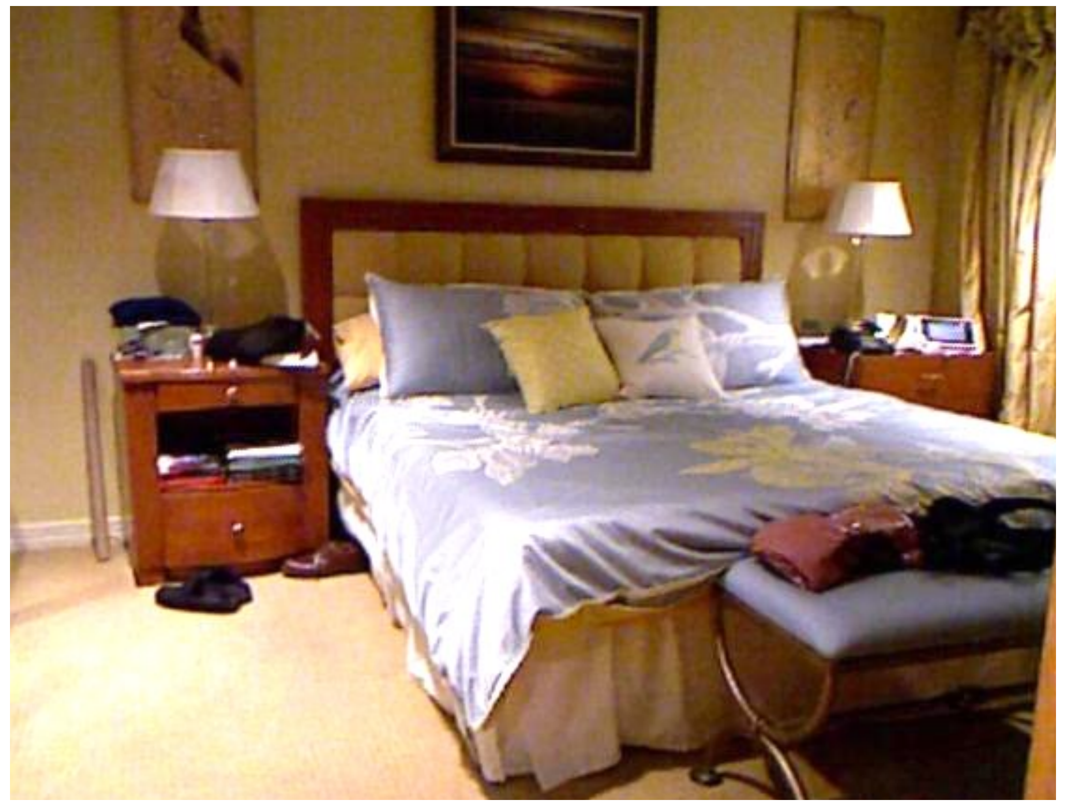
Higher Fidelity 3D Voxels
[Song et al. CVPR’17]

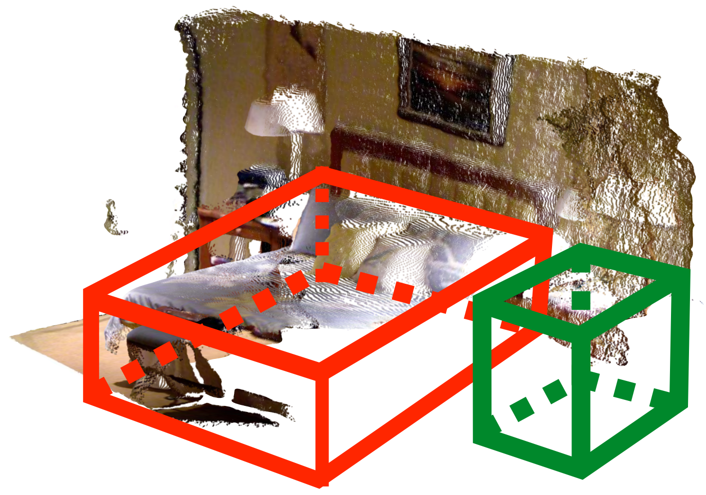
3D Voxel Grid Track



- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

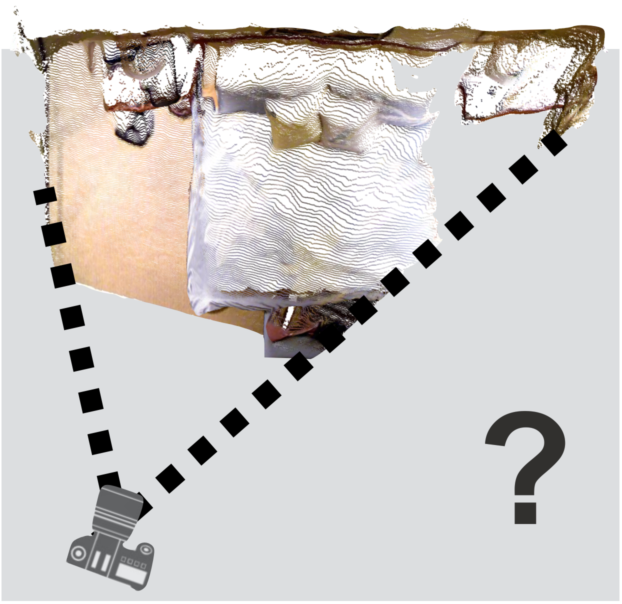
Advances Towards 3D Scene Understanding





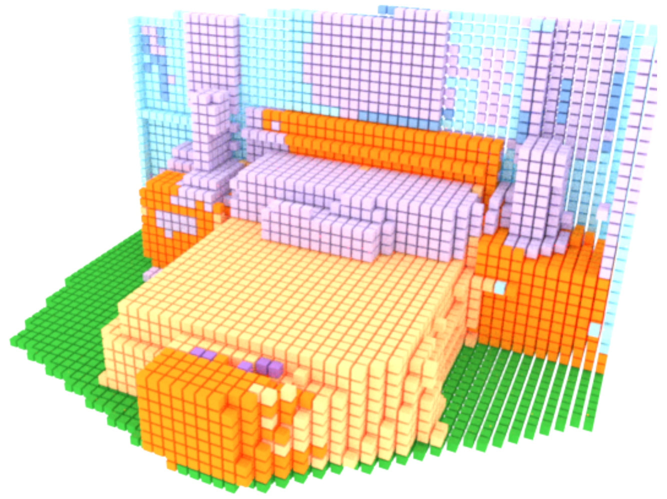
Amodal 3D Bounding Boxes

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



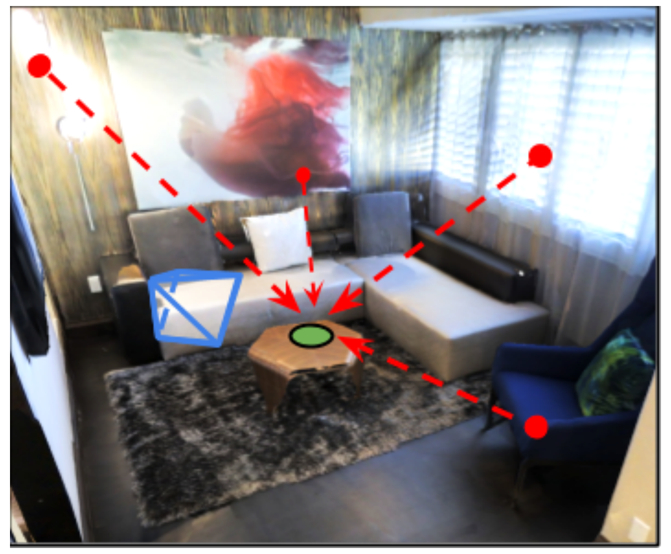
Beyond FoV Semantics&Structure

[Song et al. CVPR'18]



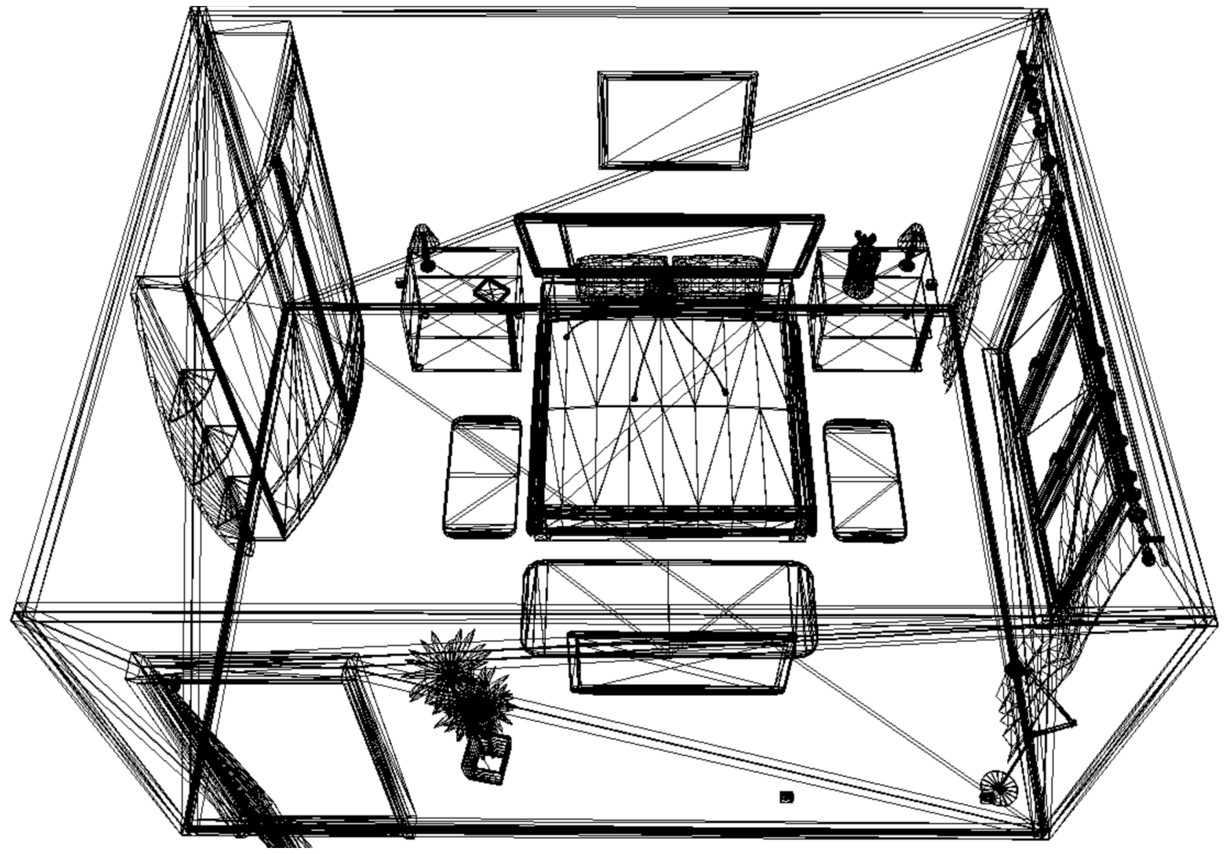
Higher Fidelity 3D Voxels

[Song et al. CVPR'17]



Beyond FoV Illumination

[Song and Funkhouser]



- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

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

Real-World RGB-D Panorama is Still Hard to Obtain

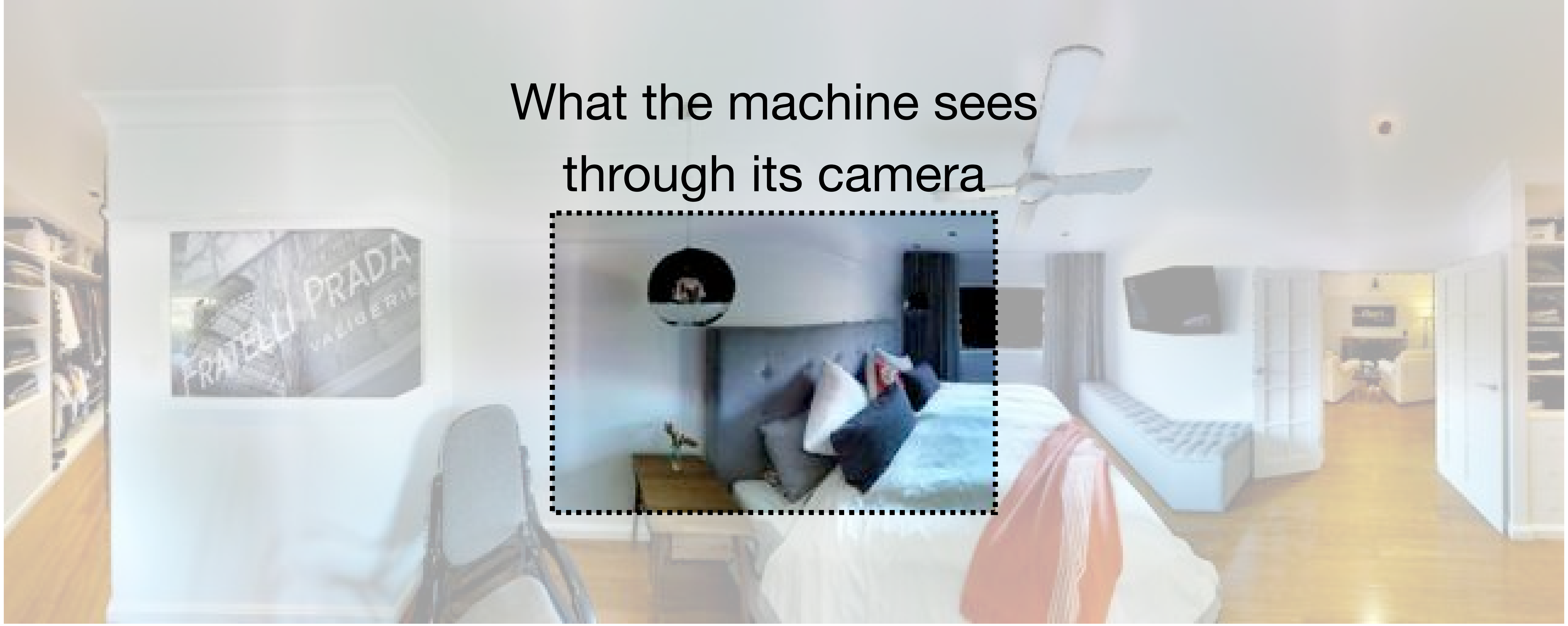
Matterport Camera
Expensive Device +
Time consuming process

Real-World Systems
has constrains on cost, power,
other physics constrains.

Cheap Devices
RGB-D image with limited FOV

Training → **Testing**

View Extrapolation



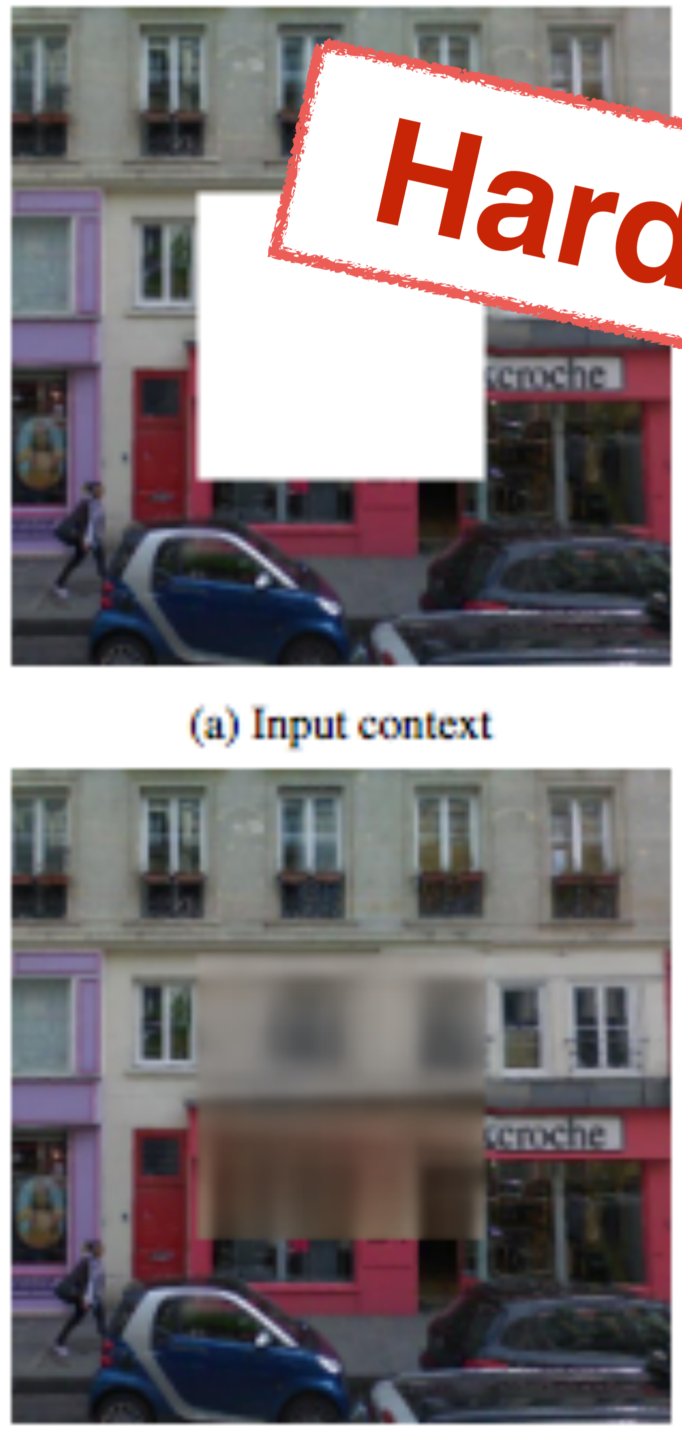
What the machine sees through its camera

Complete surrounding environment

View Extrapolation

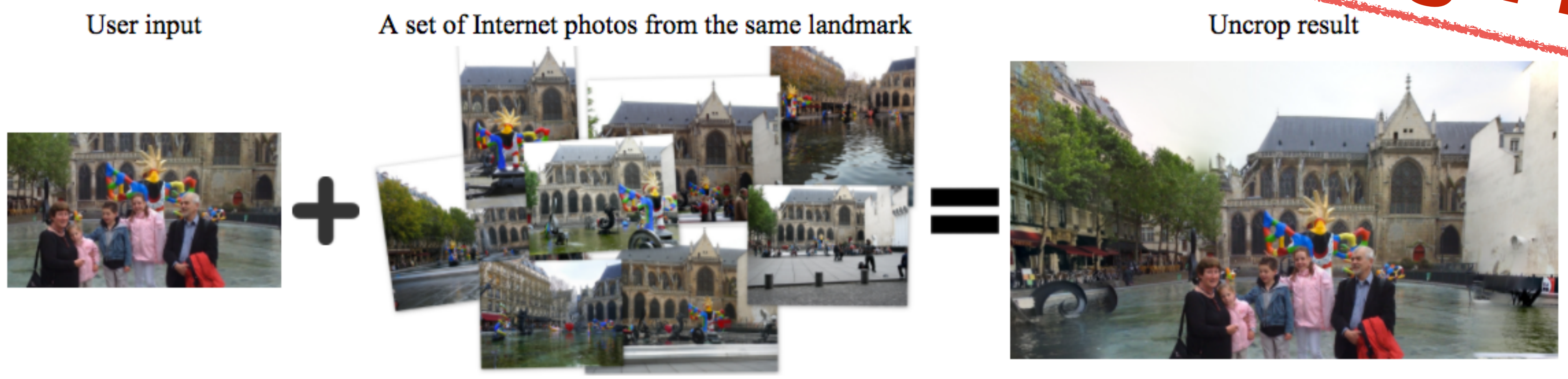
Prior work: Predicting Scene Appearance (Only Colored Pixels)

Image Inpainting
[Pathak et. al]

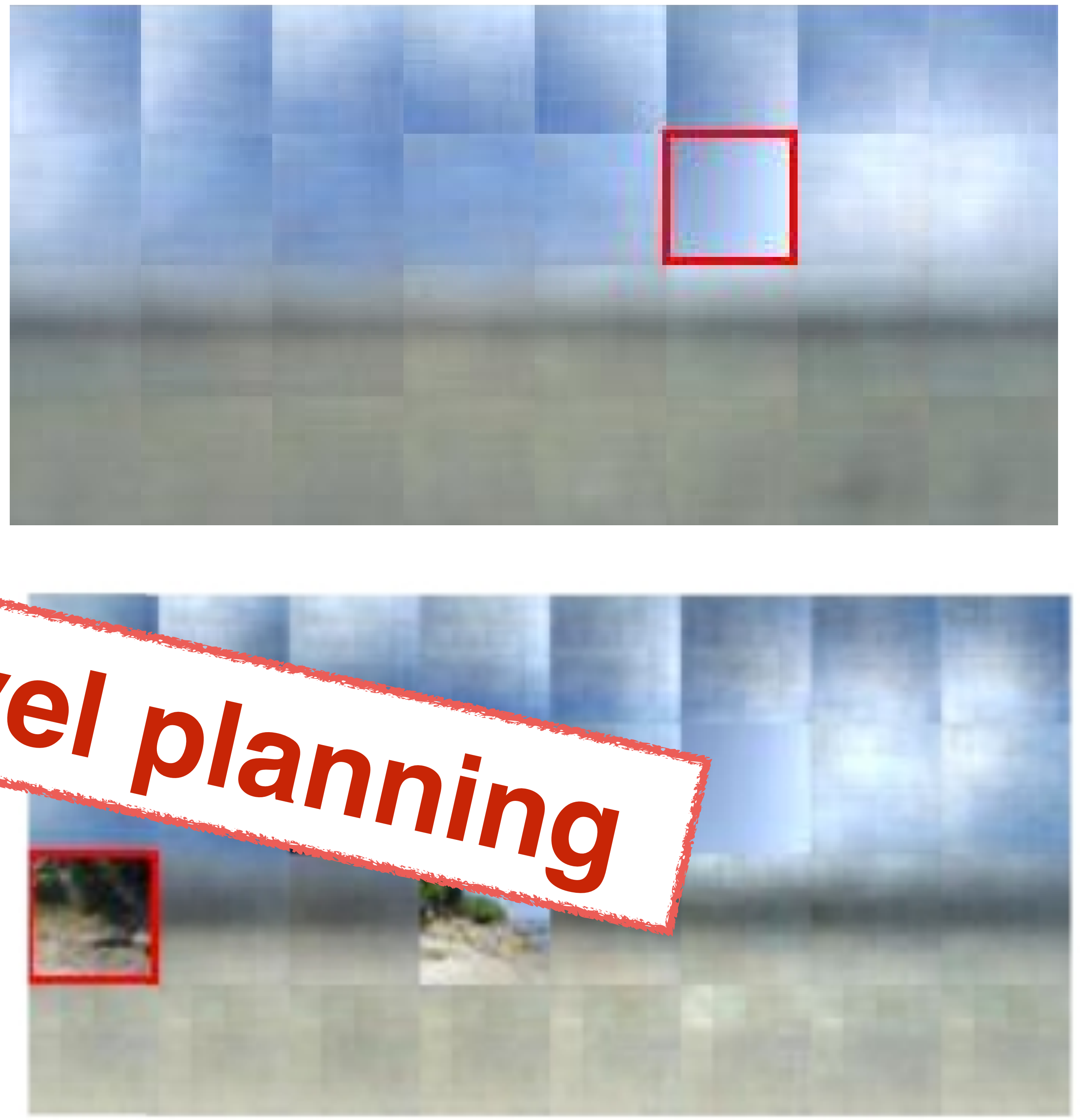

(a) Input context

User-guided view extrapolation [Zhang et al.]

Stitching images from the Inte

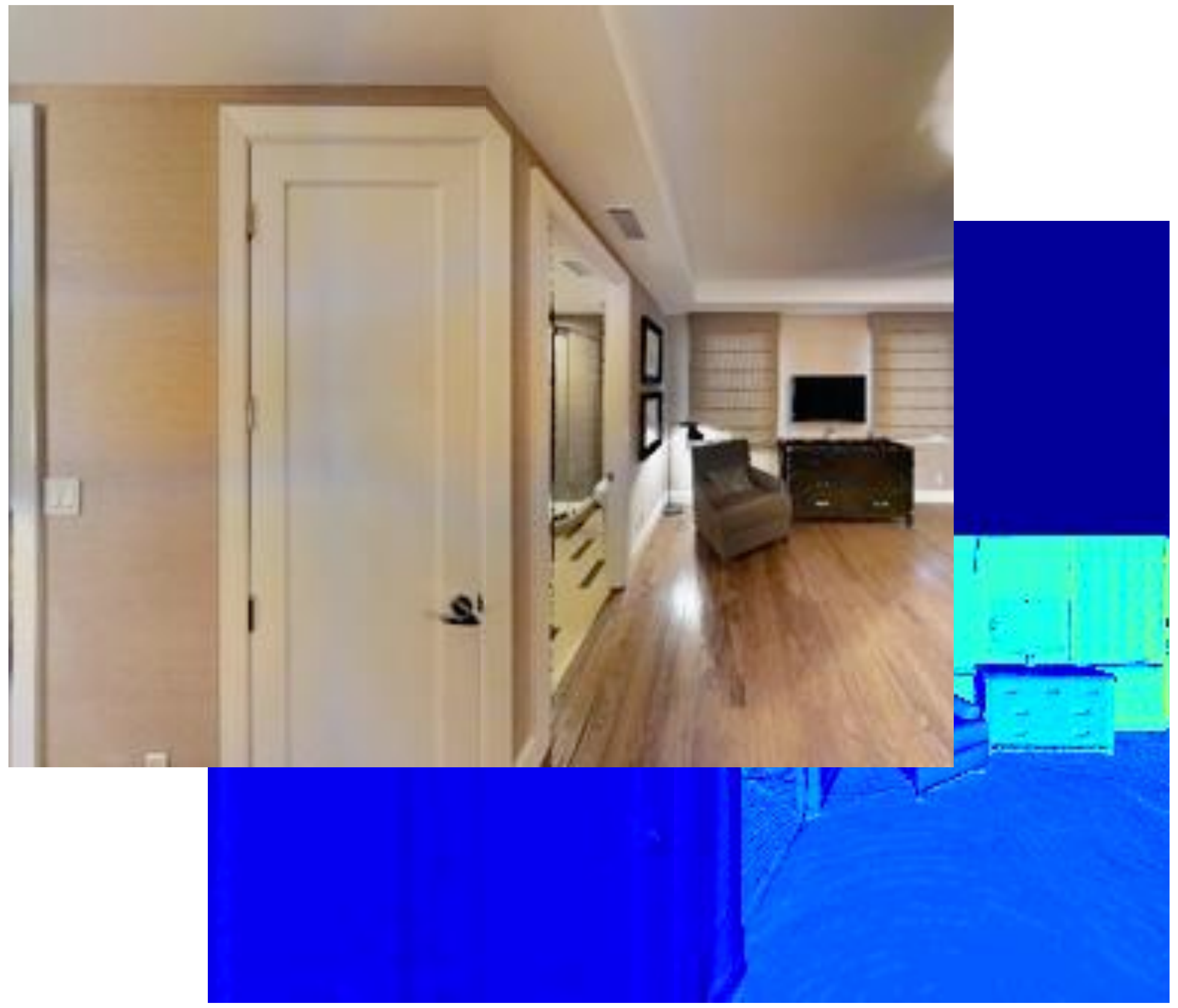


Learning to Look Around
[Jayaraman and Grauman]

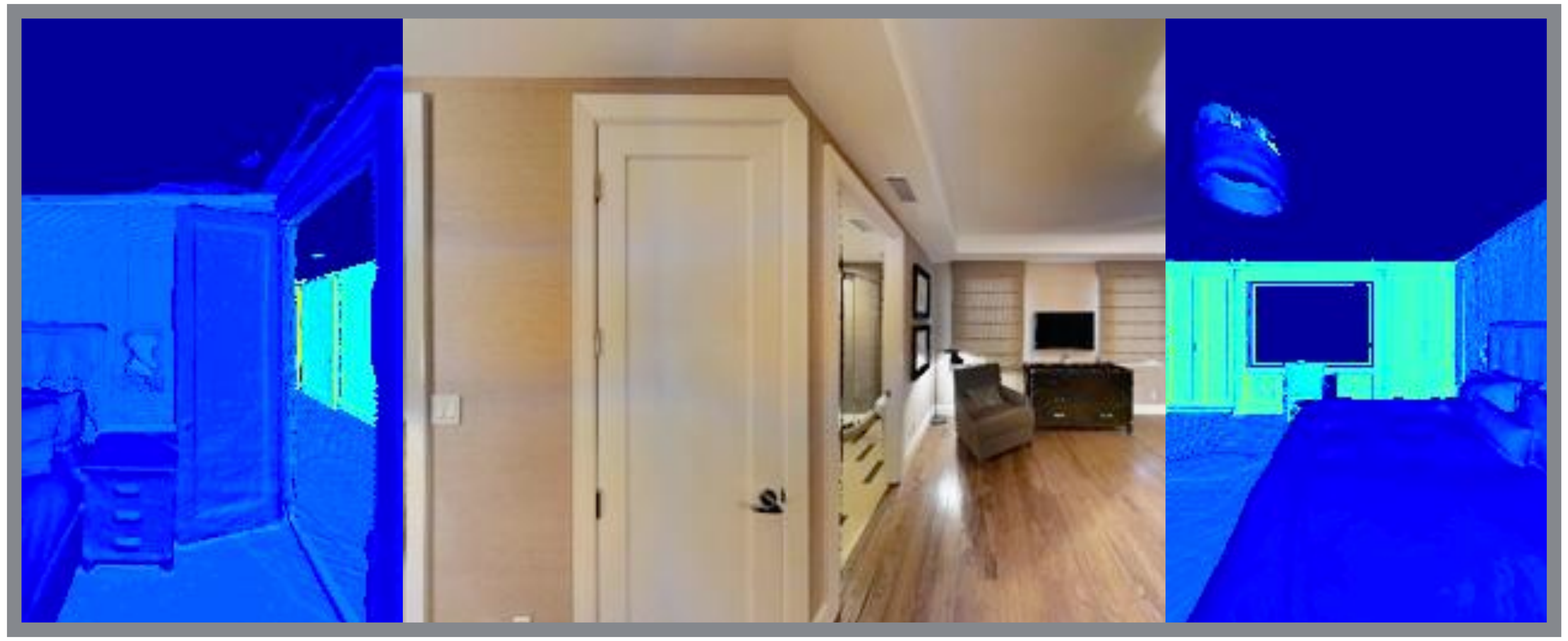


Hard to be used directly to support high level planning

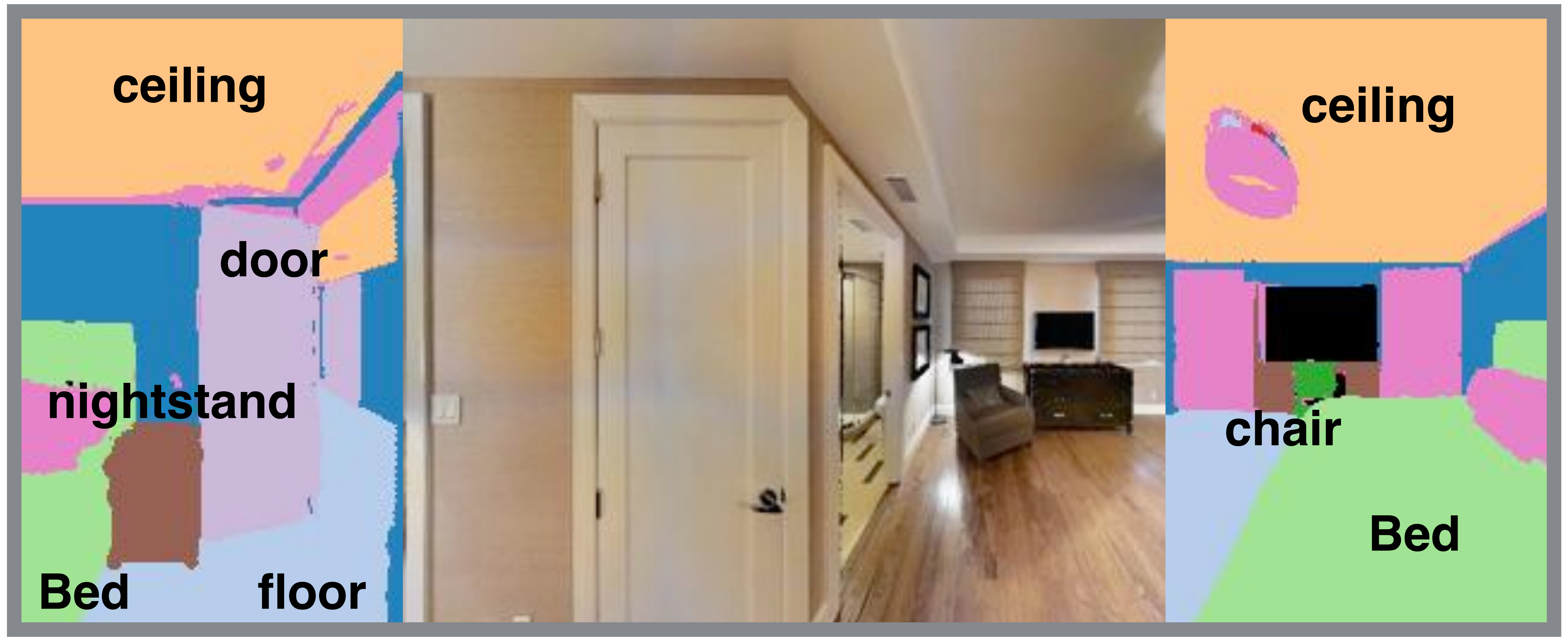
View Extrapolation



Input: RGB-D images



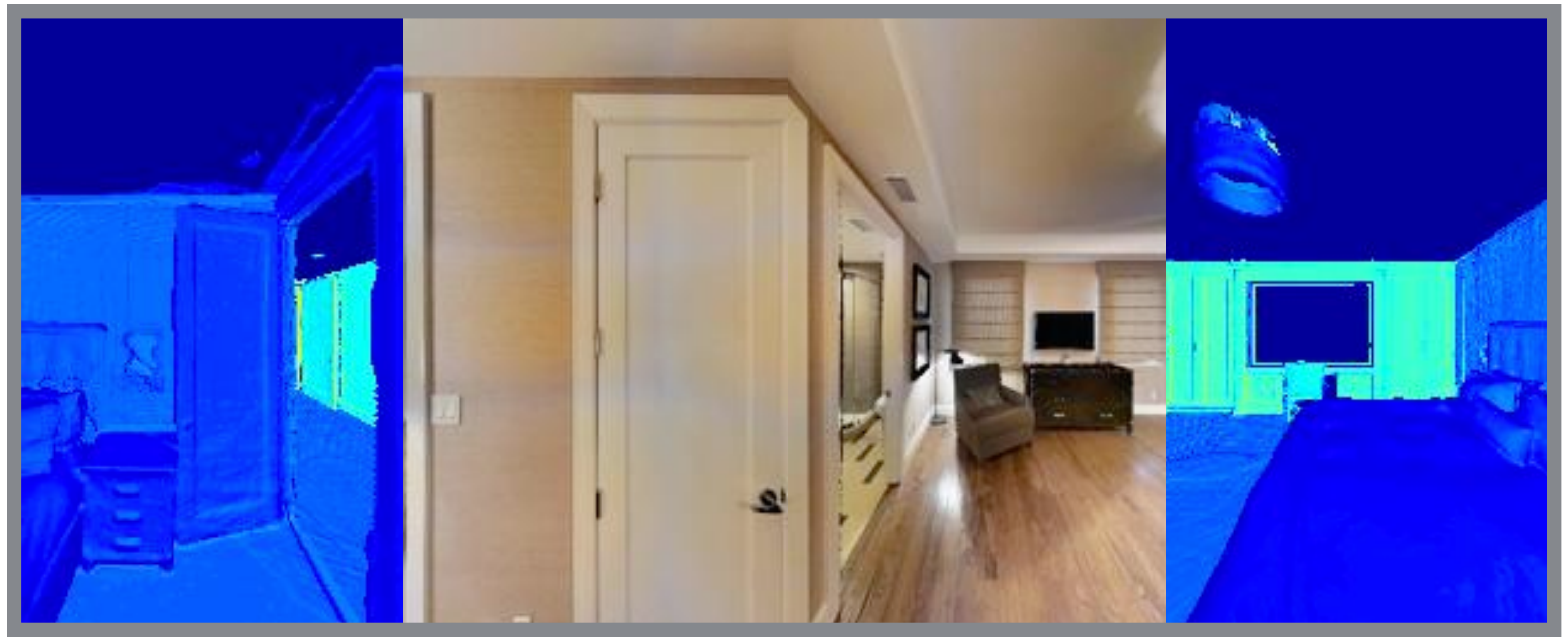
Output1: 3D Structures



Output2: Semantics

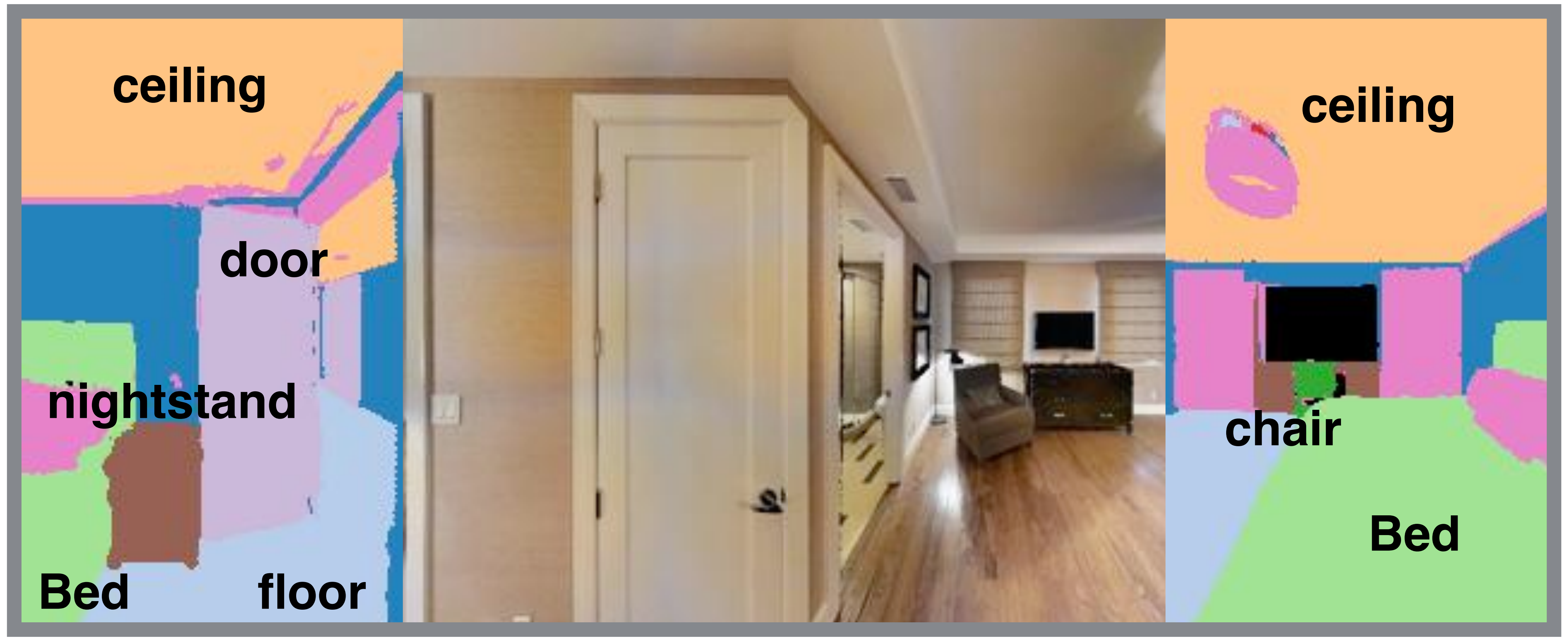
View Extrapolation

Where can I move?



Output1: 3D Structures

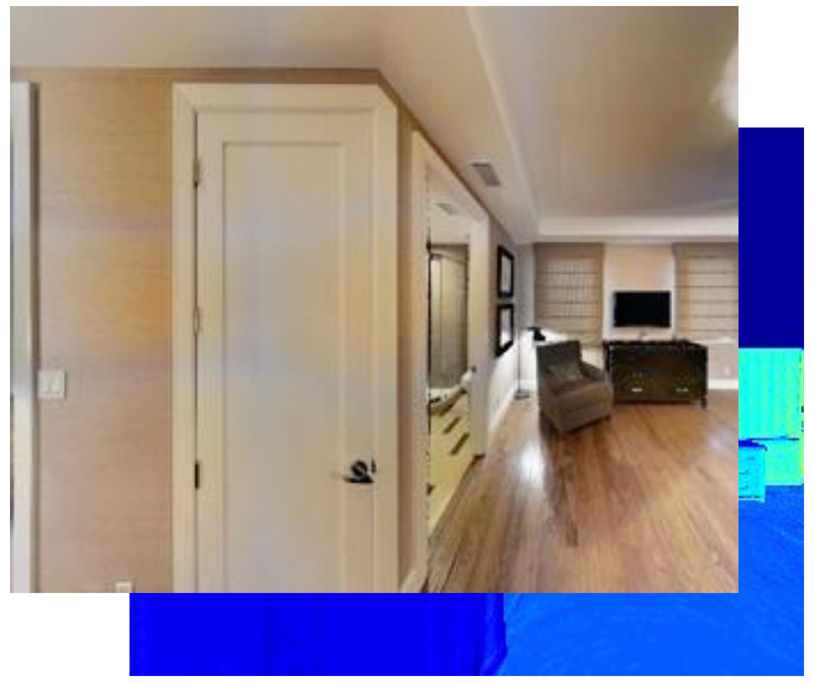
Where should I turn to find a door?

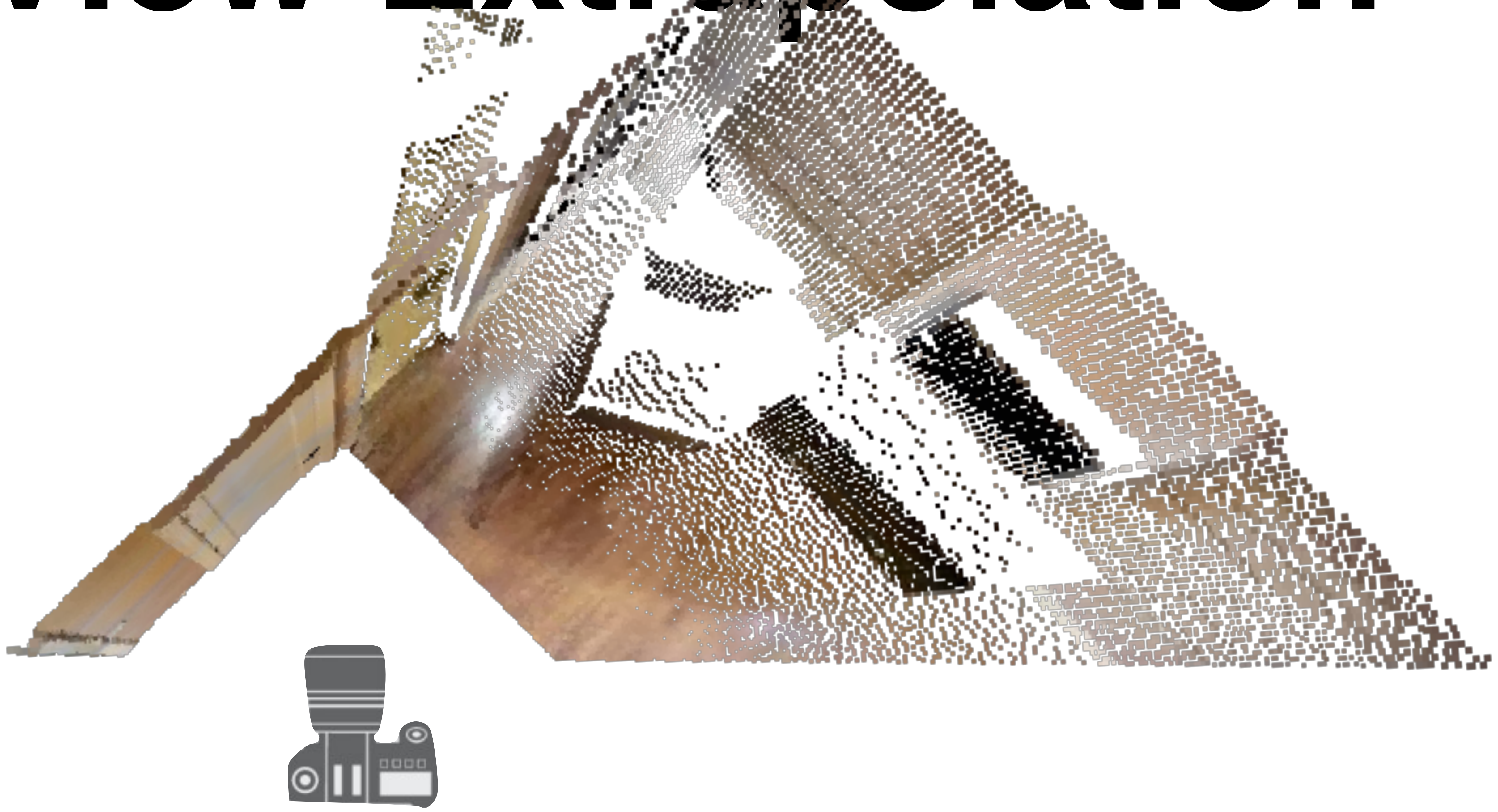


Output2: Semantics

Semantic-Structure View Extrapolation

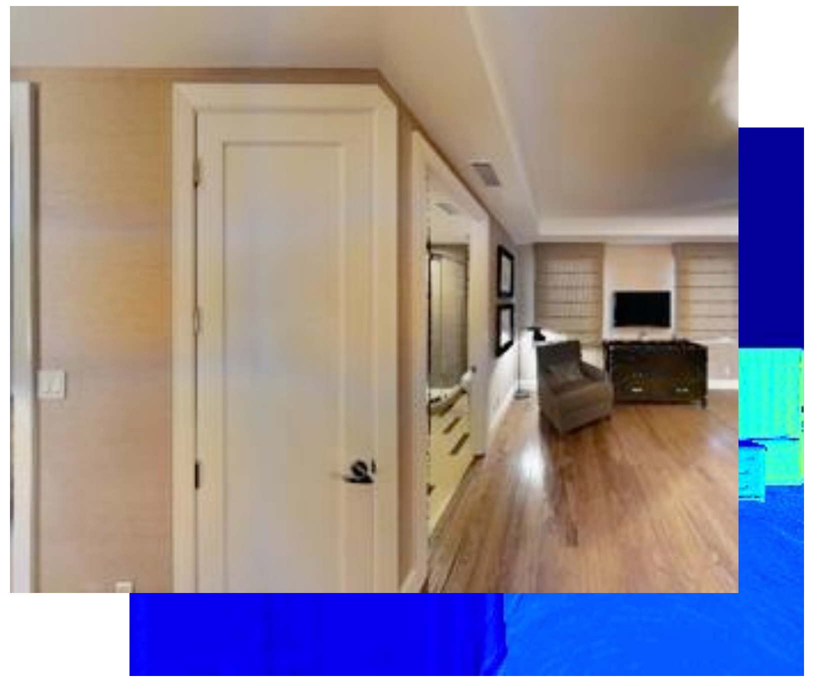
Input: RGB-D images



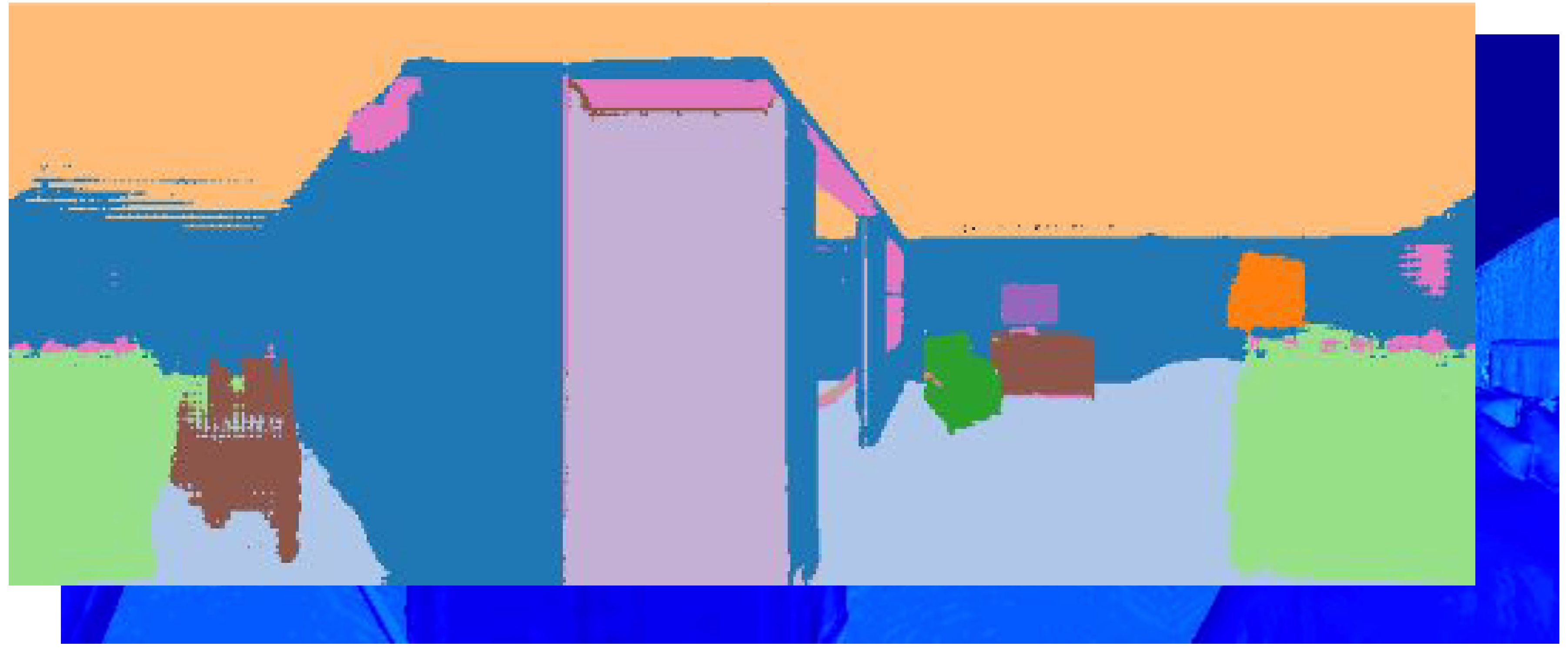


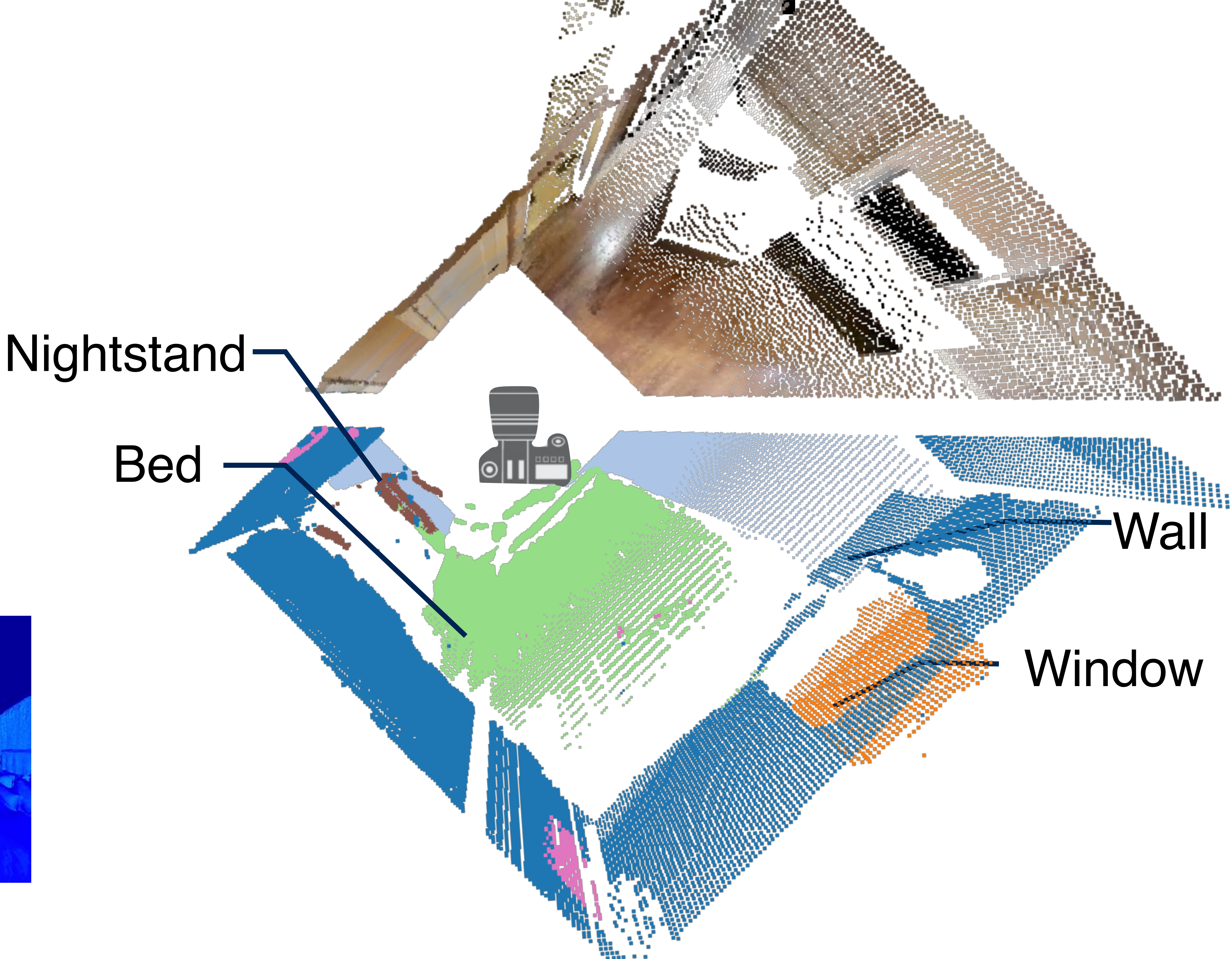
Semantic-Structure View Extrapolation

Input: RGB-D images



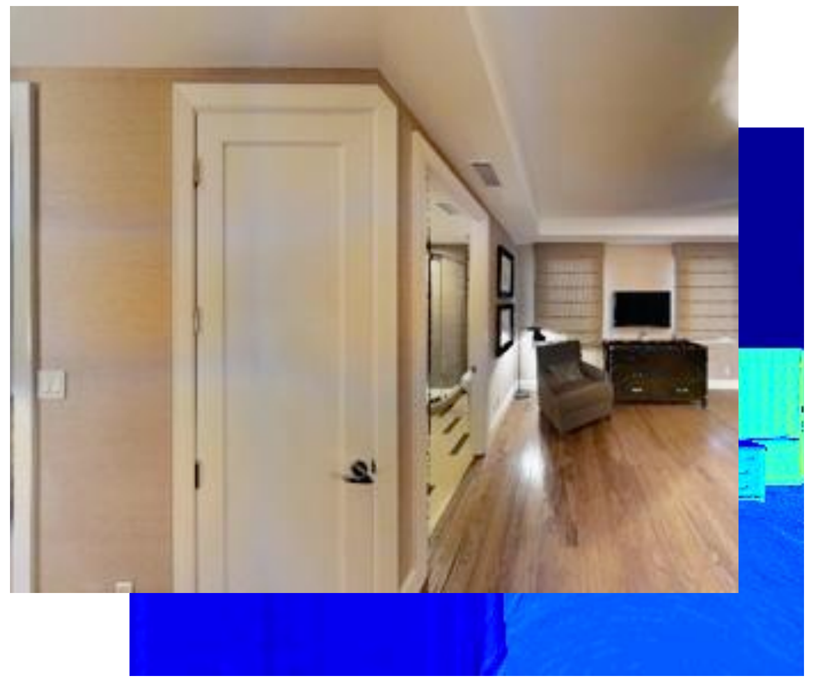
Output: 360° panorama
with 3D structure & semantics



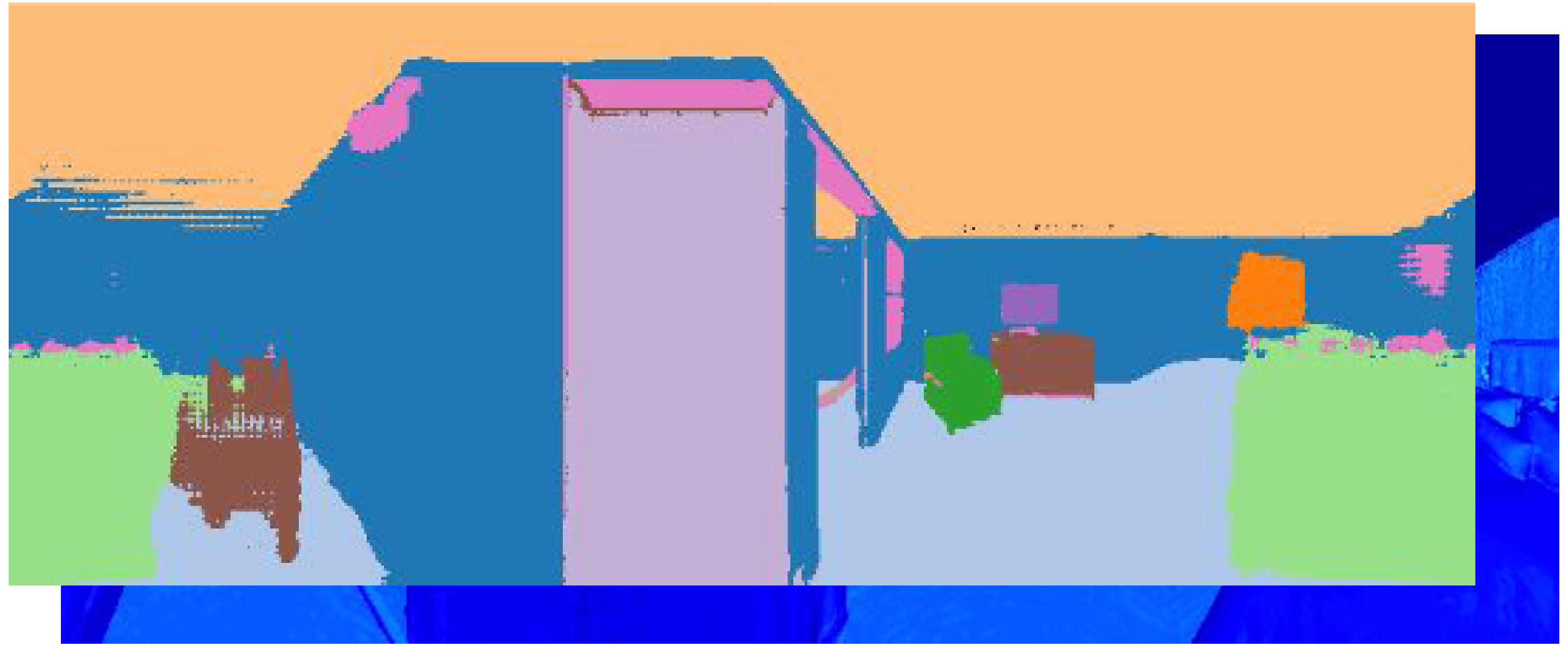


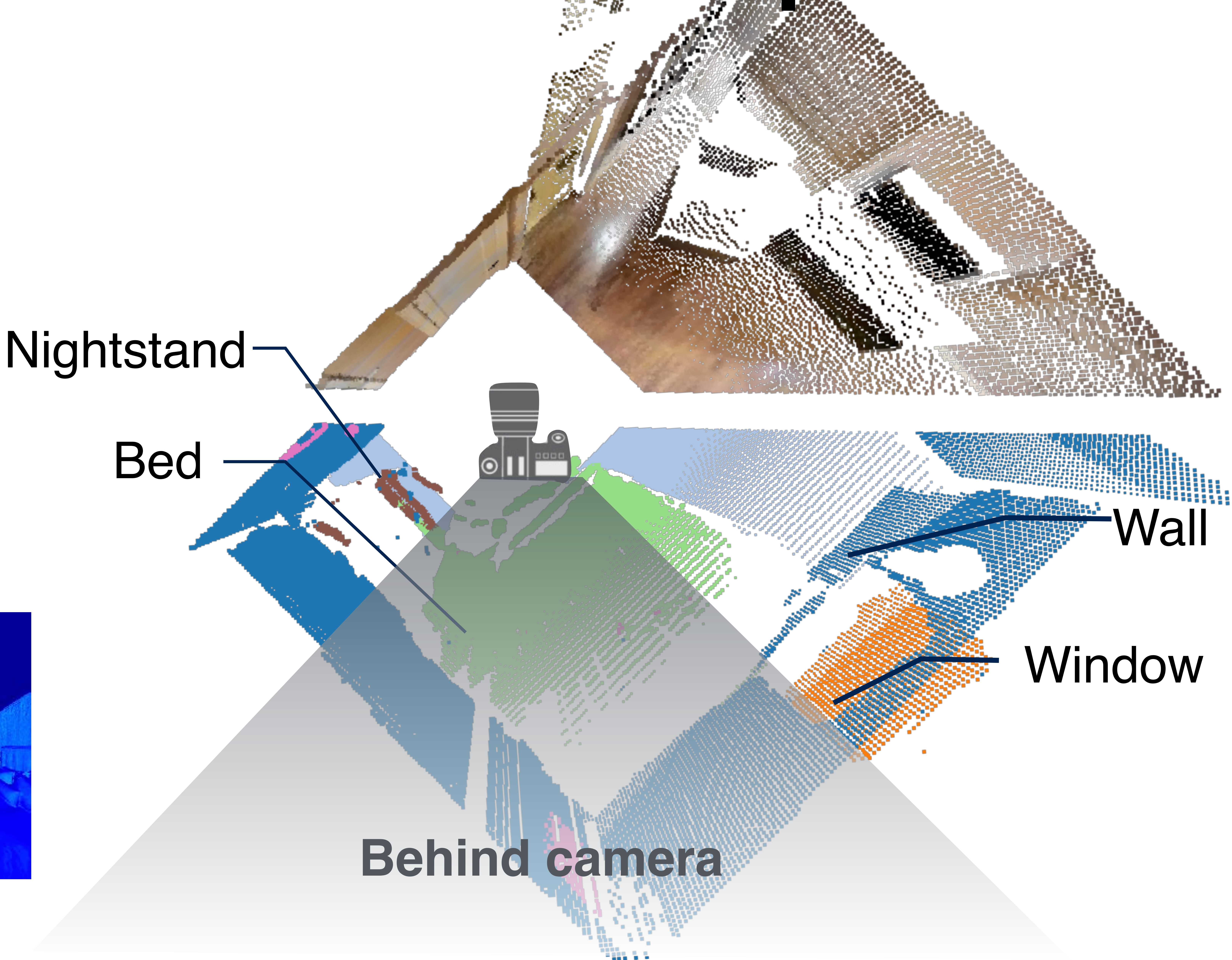
Semantic-Structure View Extrapolation

Input: RGB-D images



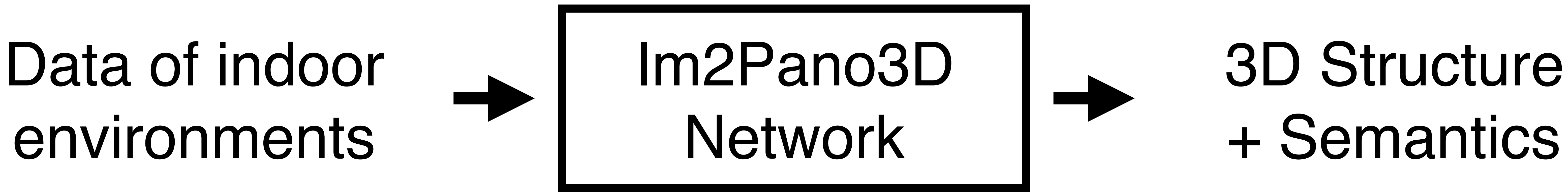
Output: 360° panorama with 3D structure & semantics





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

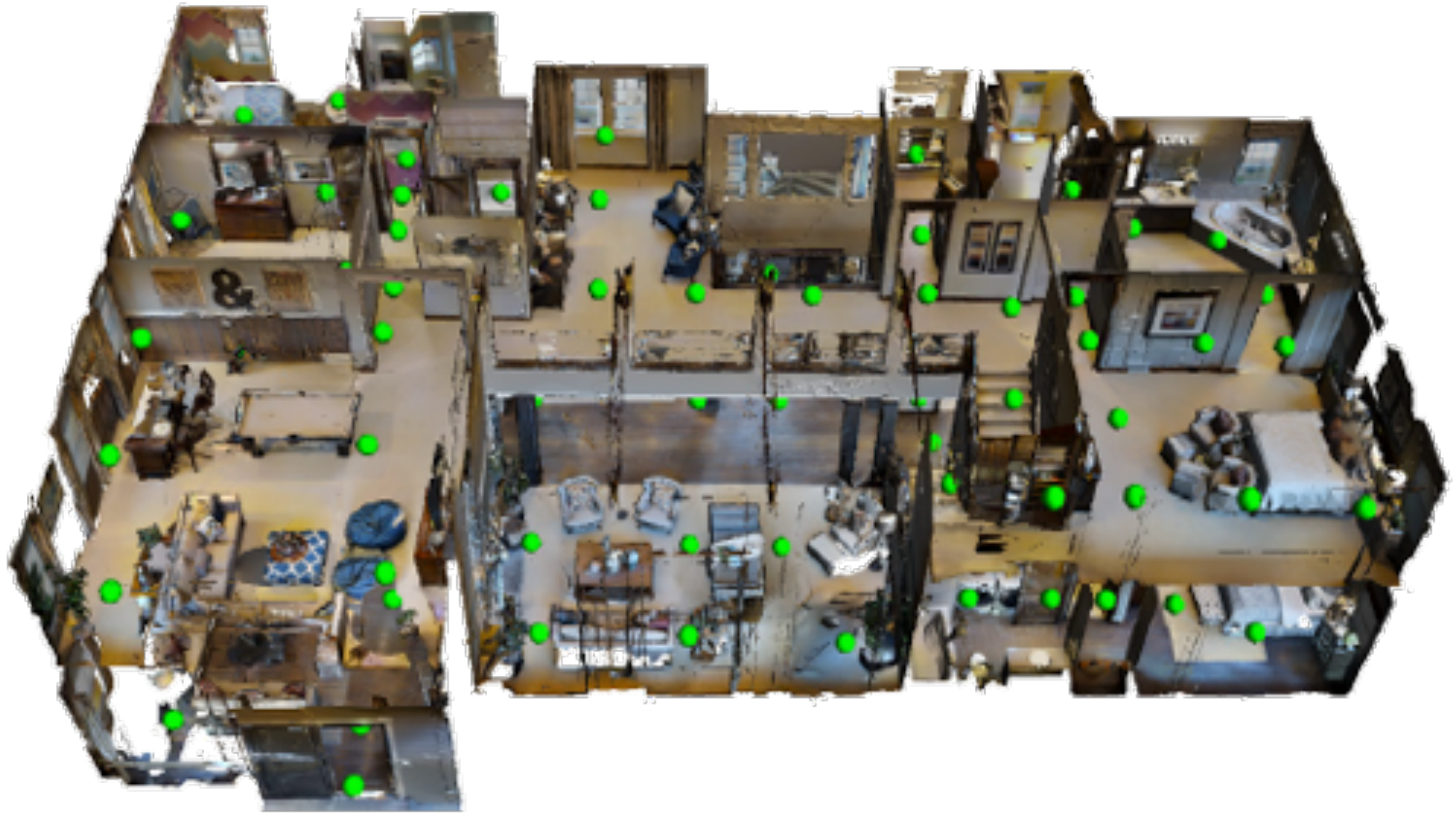
Im2Pano3D Network

3D Structure + Semantics

Training data

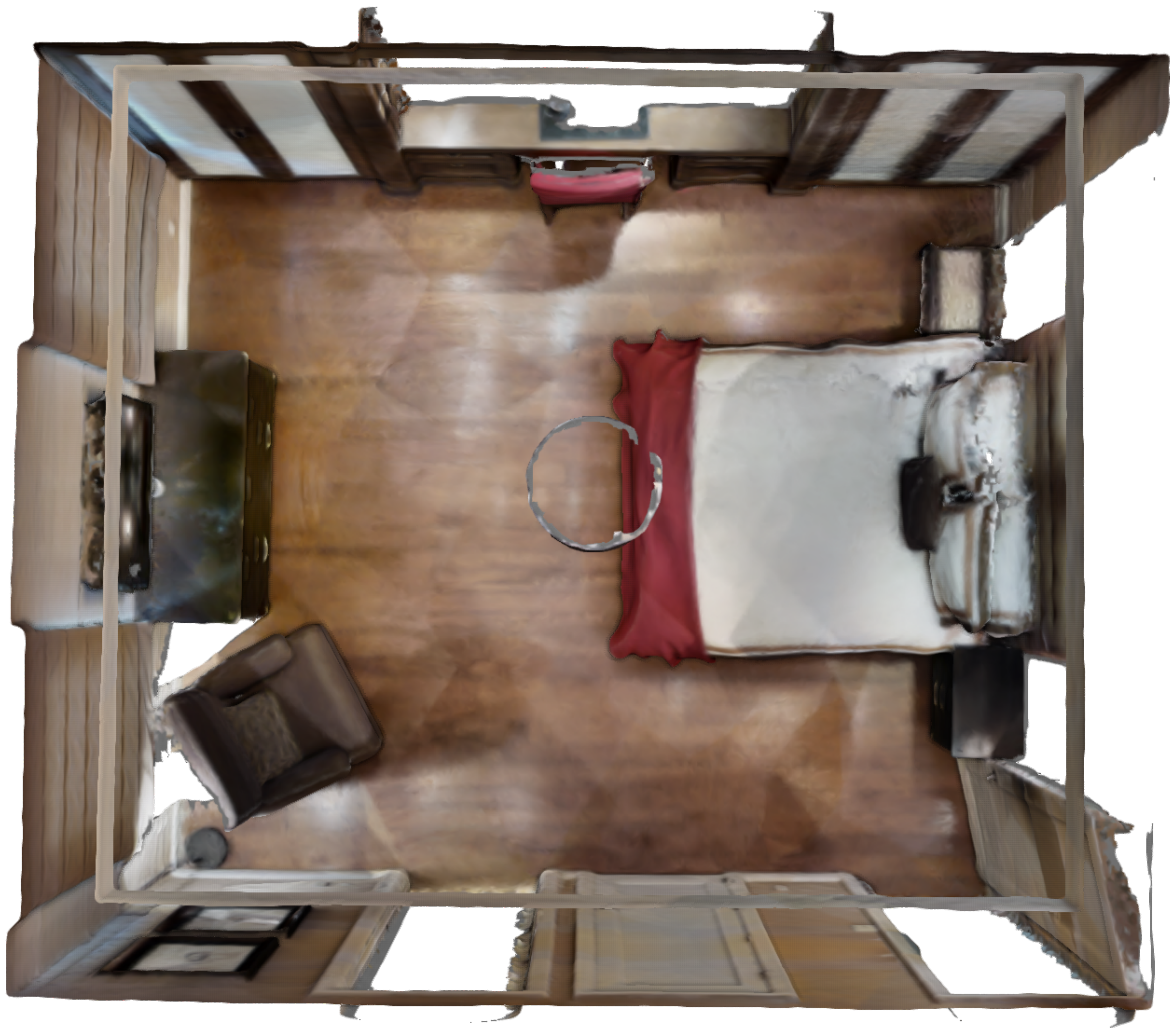
3D House Datasets

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



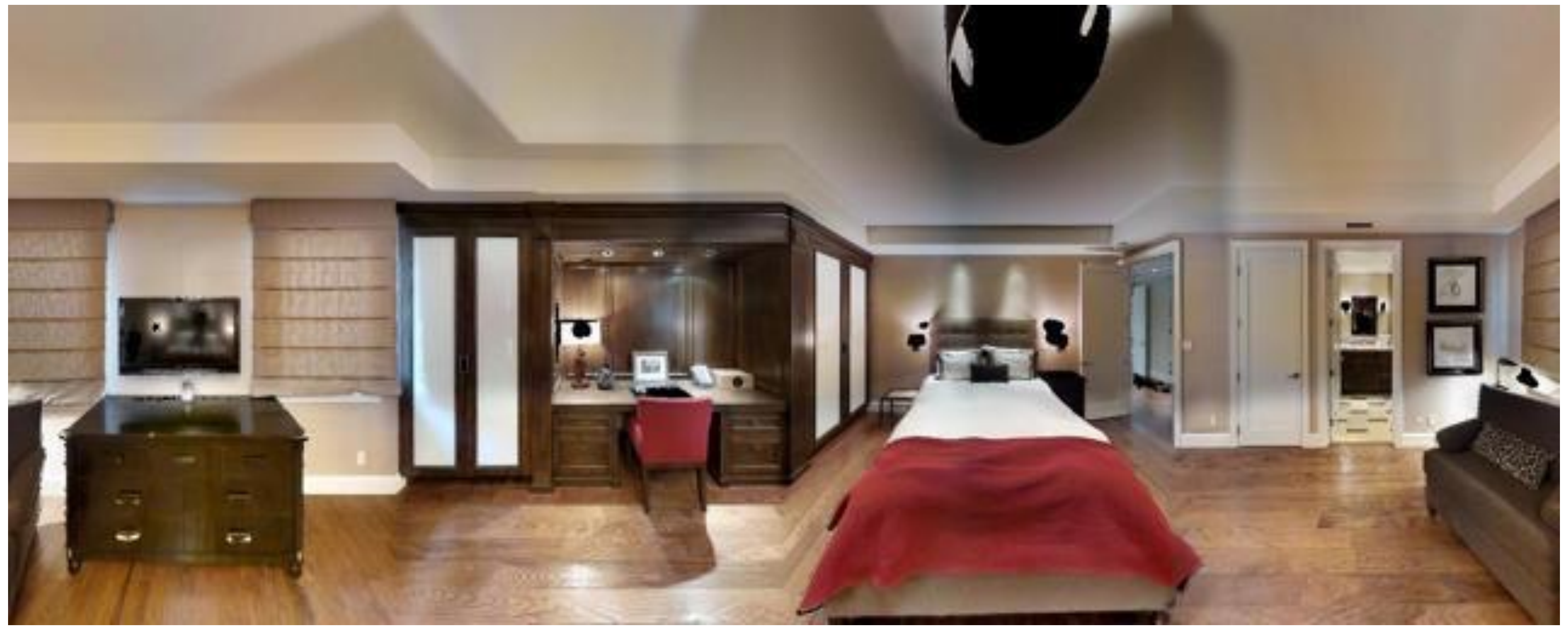
Real-Word Houses (Matterport3D):
5,315 RGB-D panoramas
Fine-tune and test

Data Representation

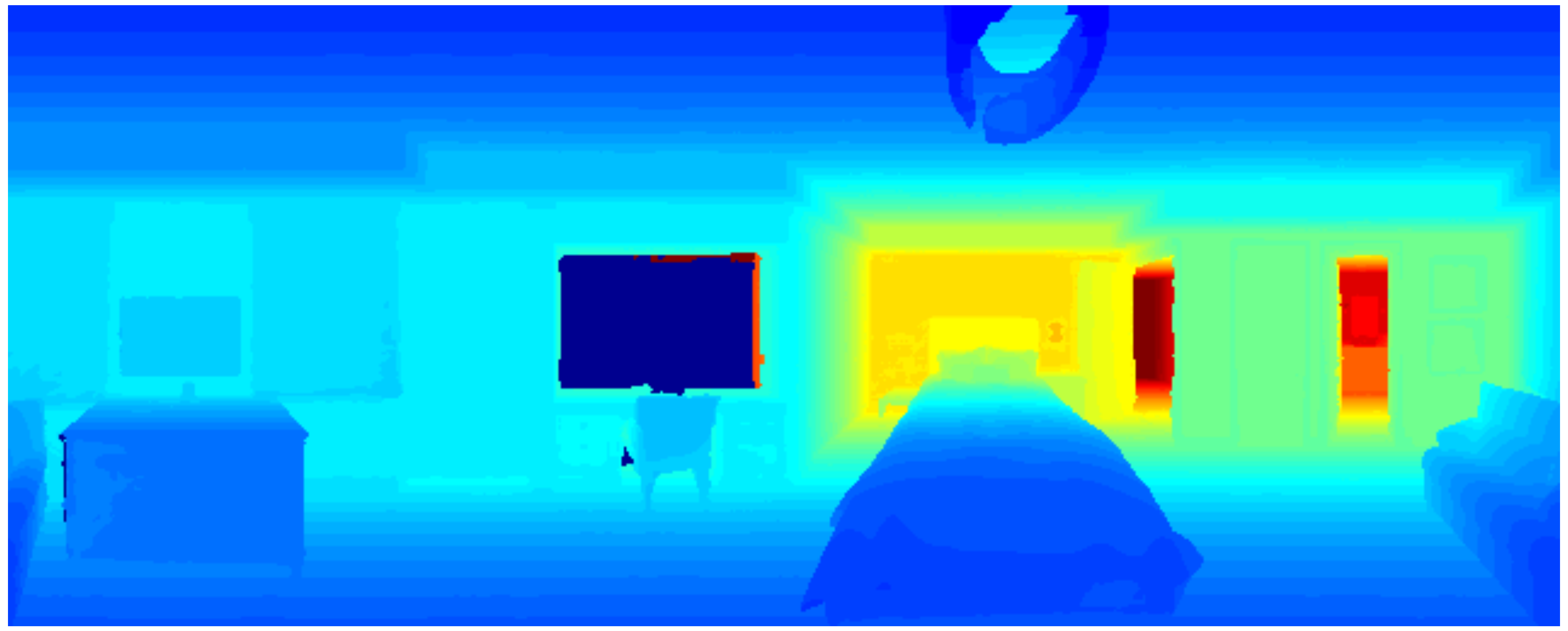


3D Room

360 Degree FoV

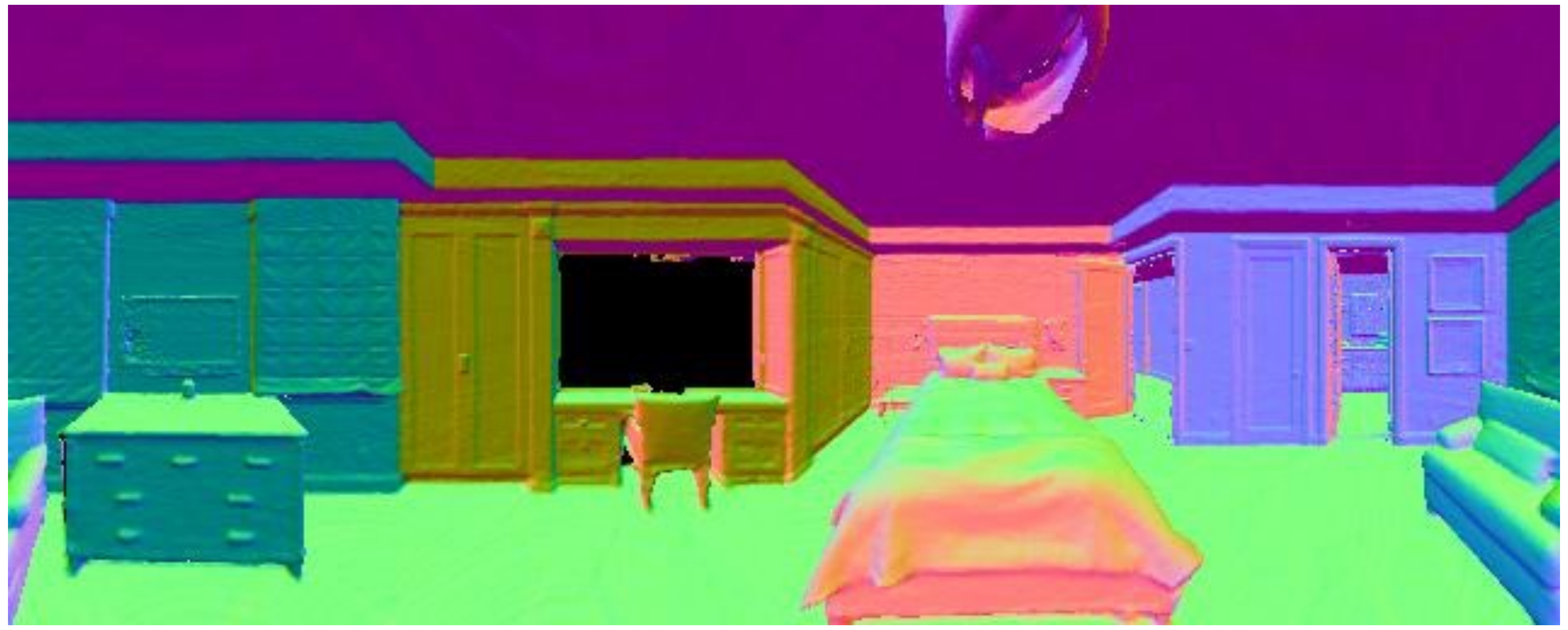


Color Panorama



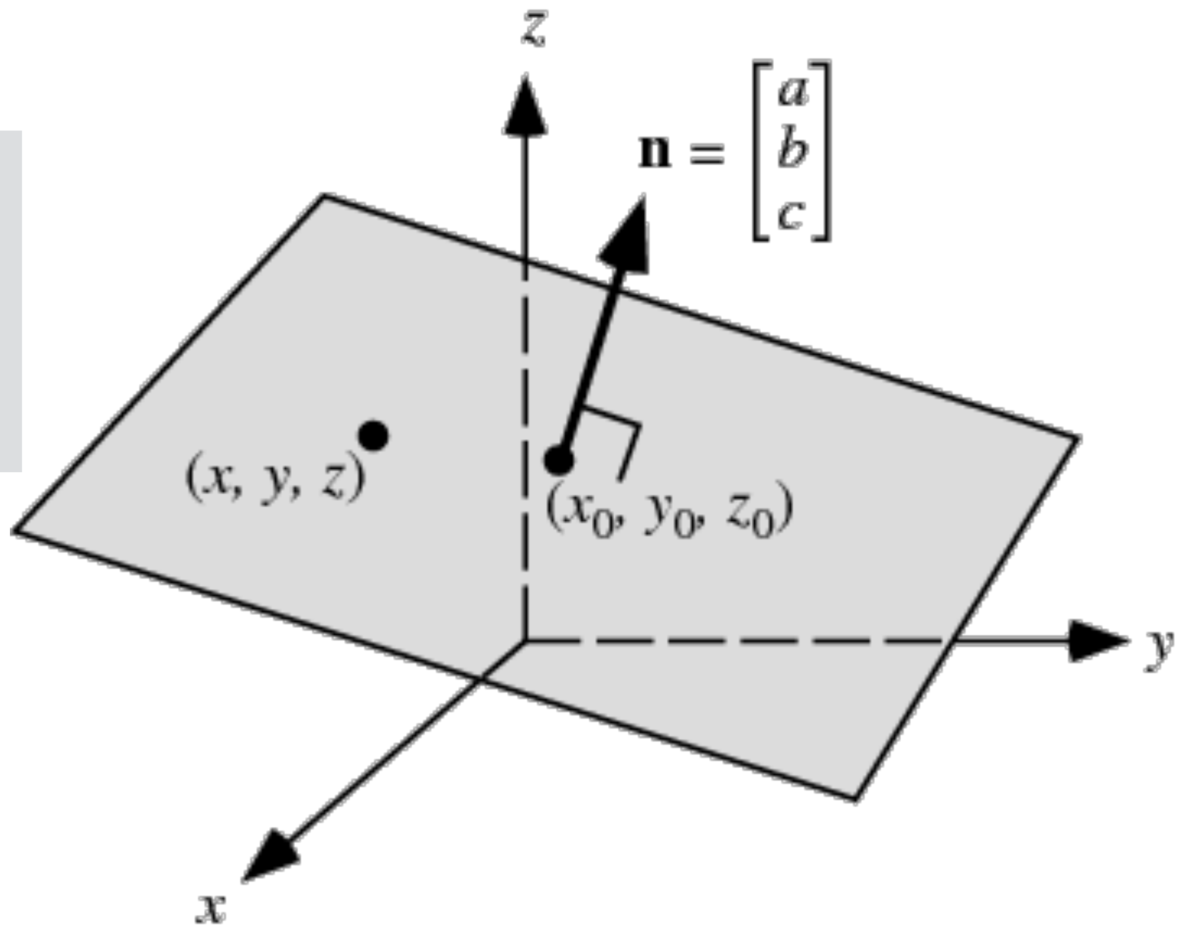
Depth Panorama

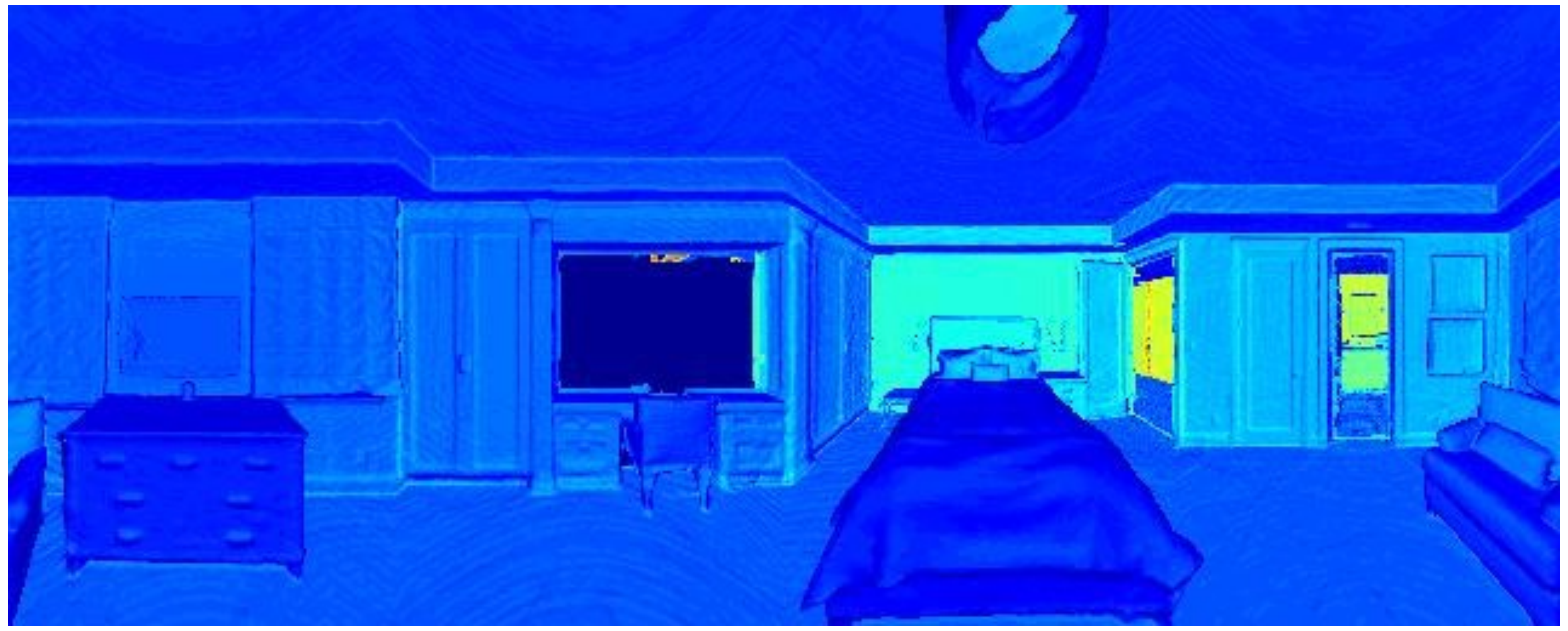
Data Representation



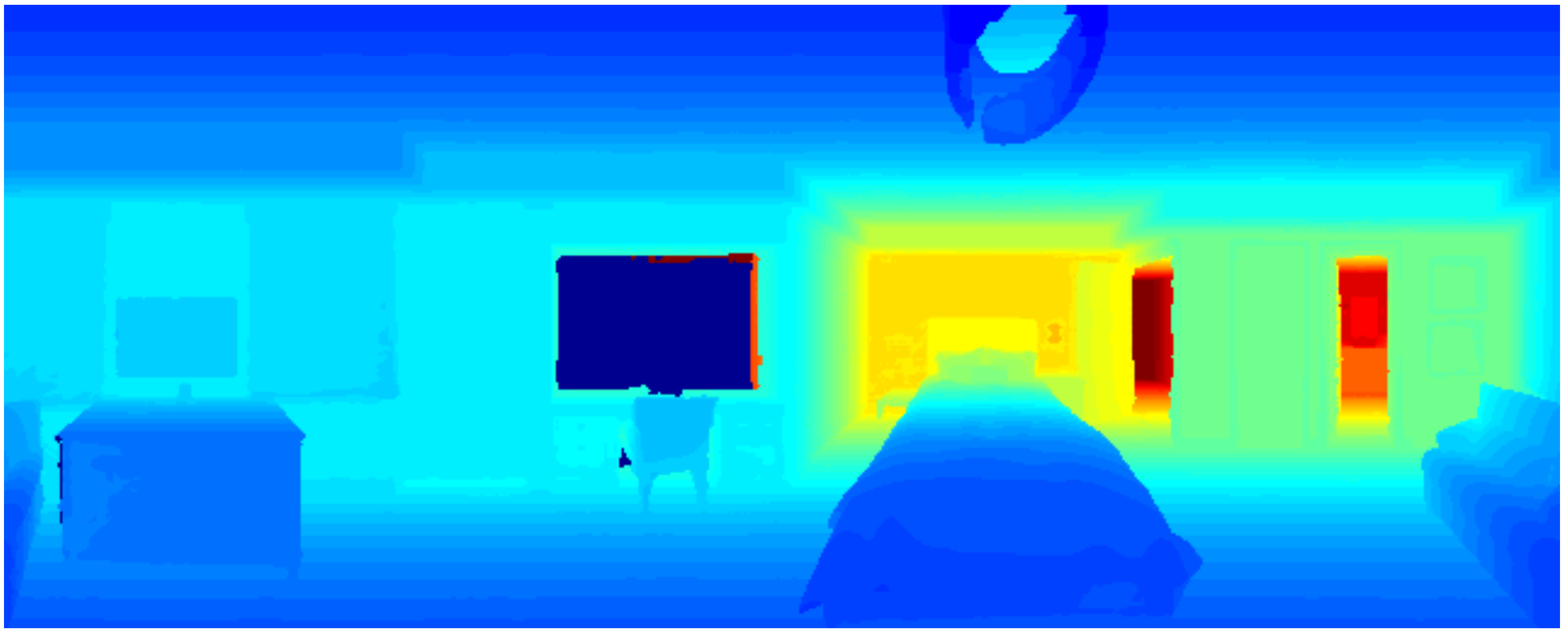
Surface Normal (a,b,c)

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



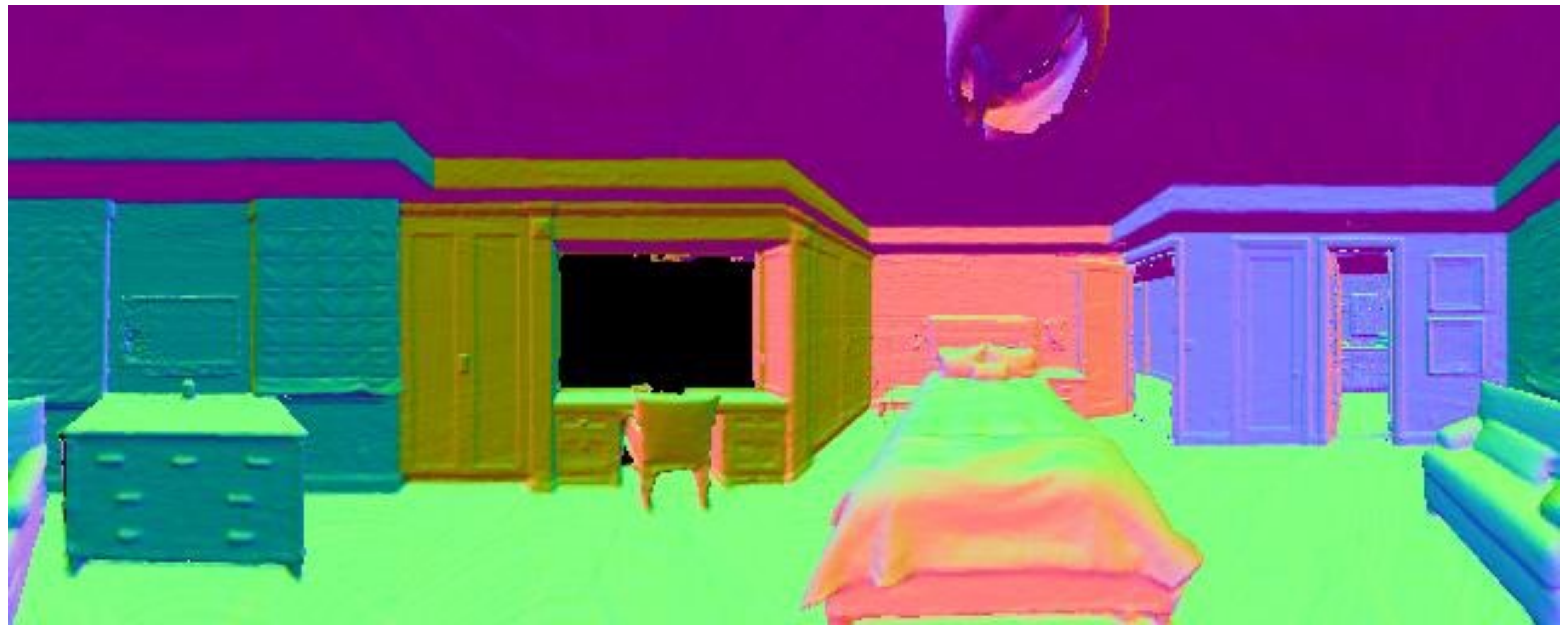


Plane Distance to Origin (p)

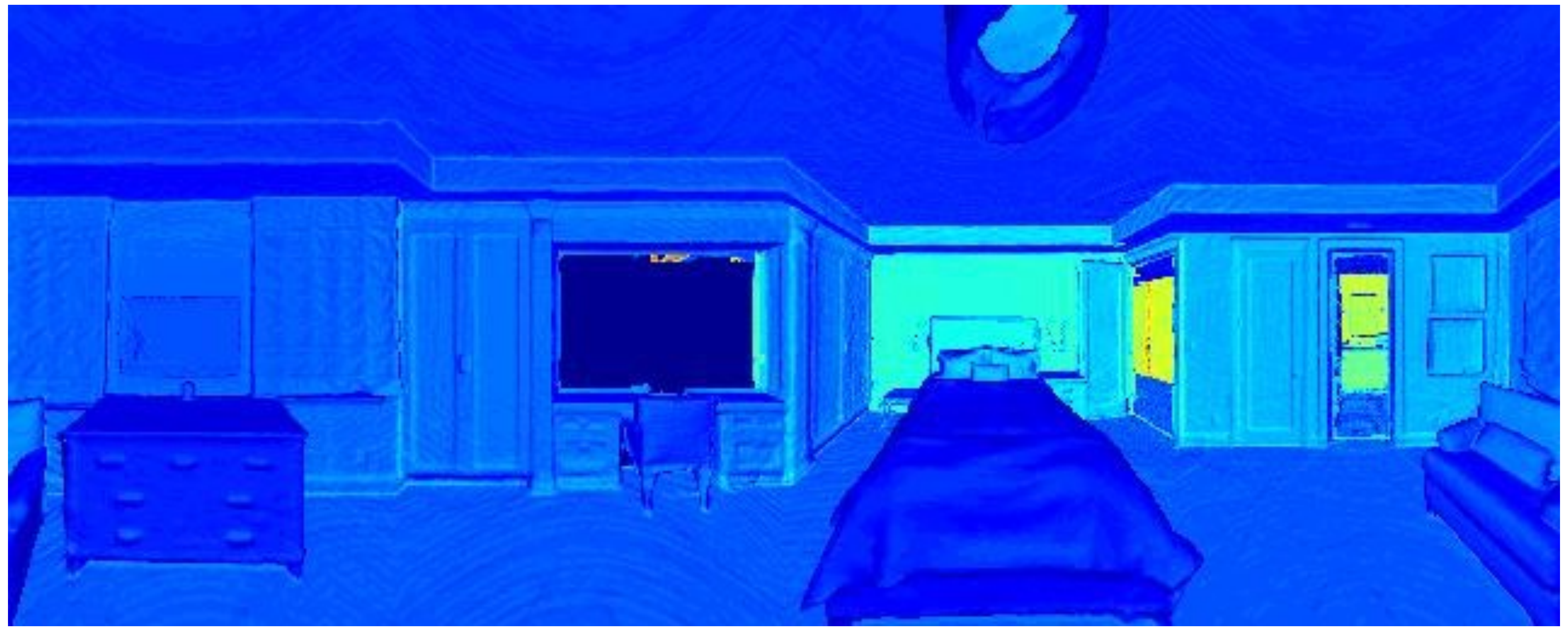


Depth Panorama

Data Representation

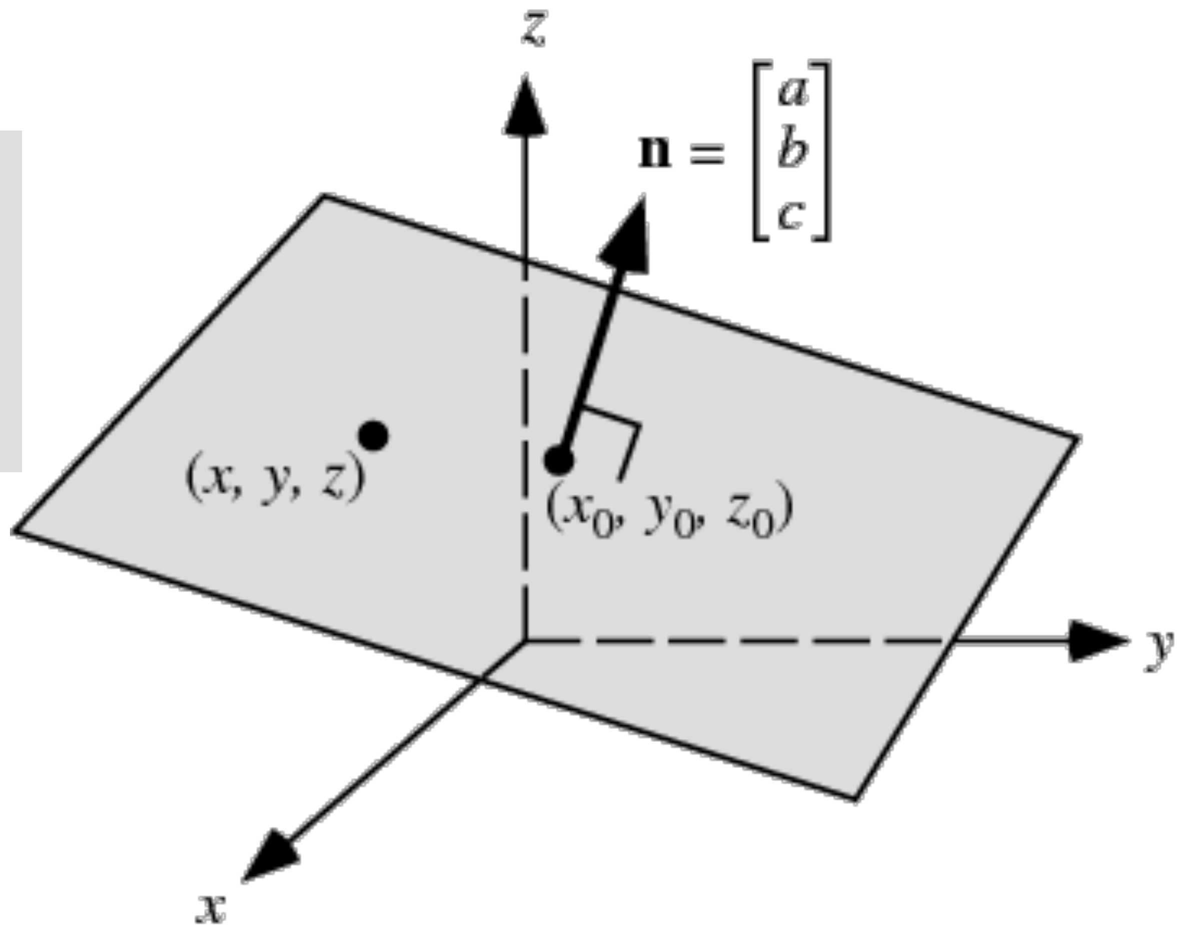


Surface Normal (a,b,c)



Plane Distance to Origin (p)

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



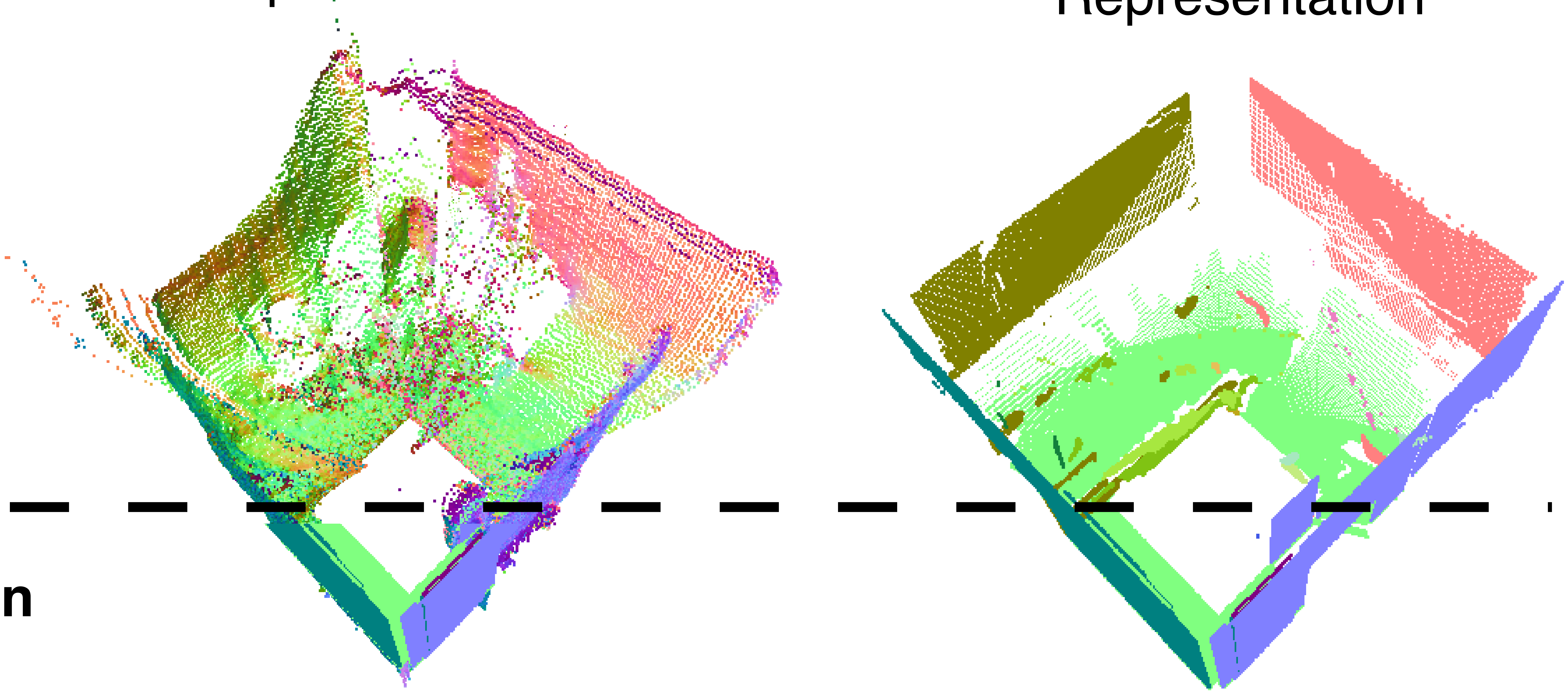
✓Pixels on the same planar surface share the same plane equation.

✓Representation is piecewise constant in a typical indoor environment.

Data Representation

Raw Depth Representation

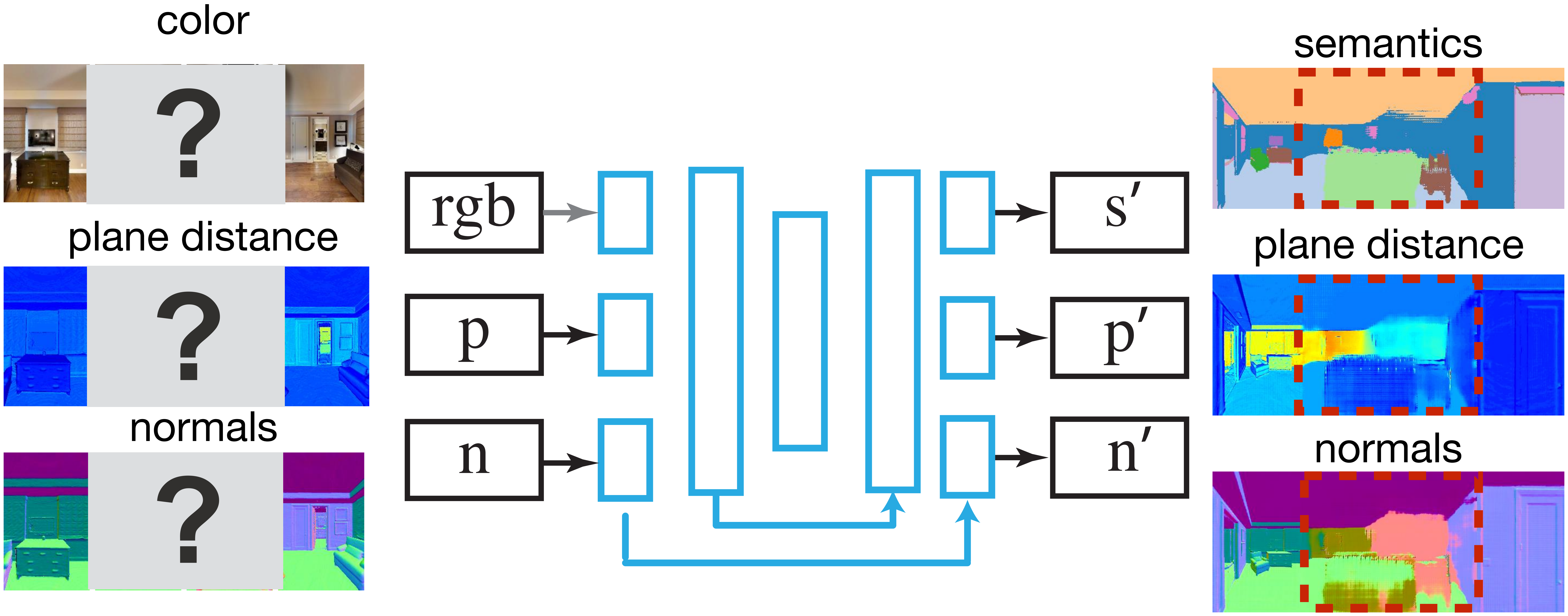
Plane Representation



Prediction

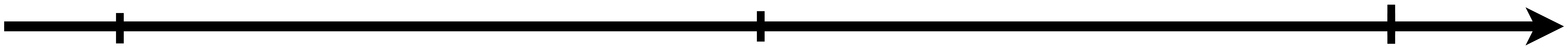
Observation

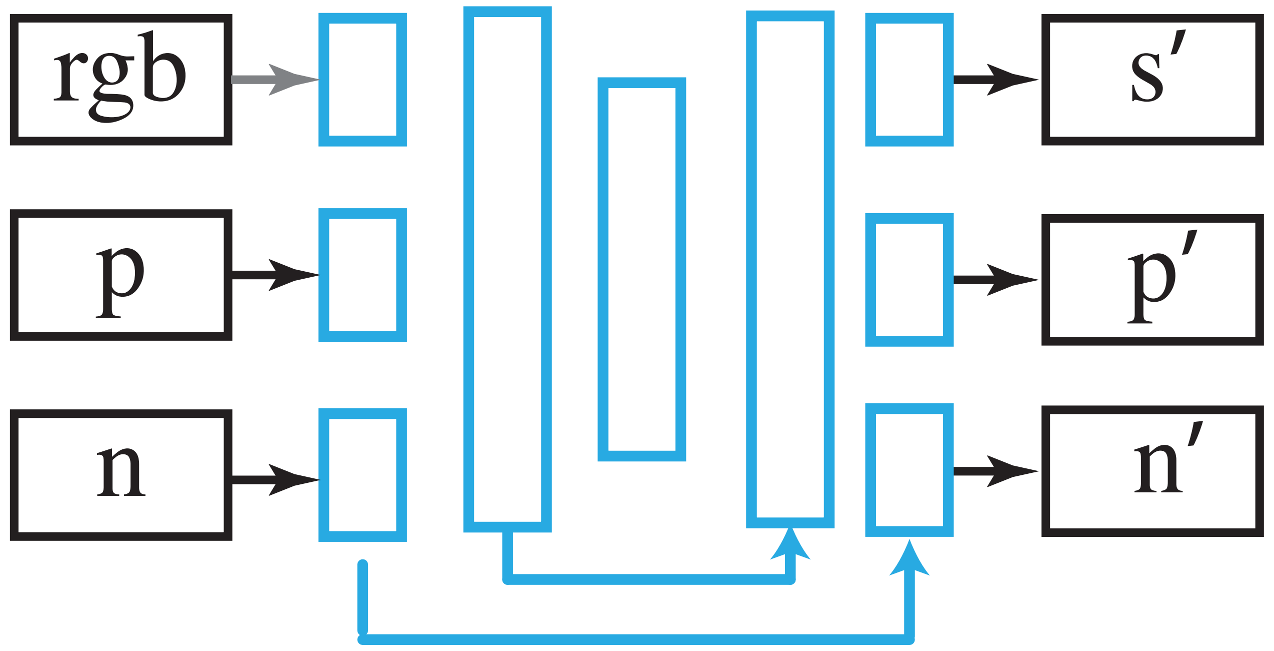
Im2Pano3D Network



What training objectives should we use?

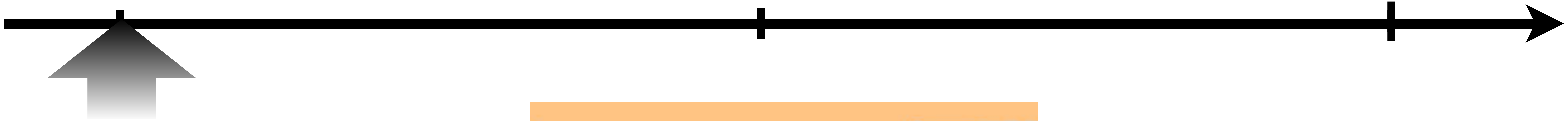
Training Objectives

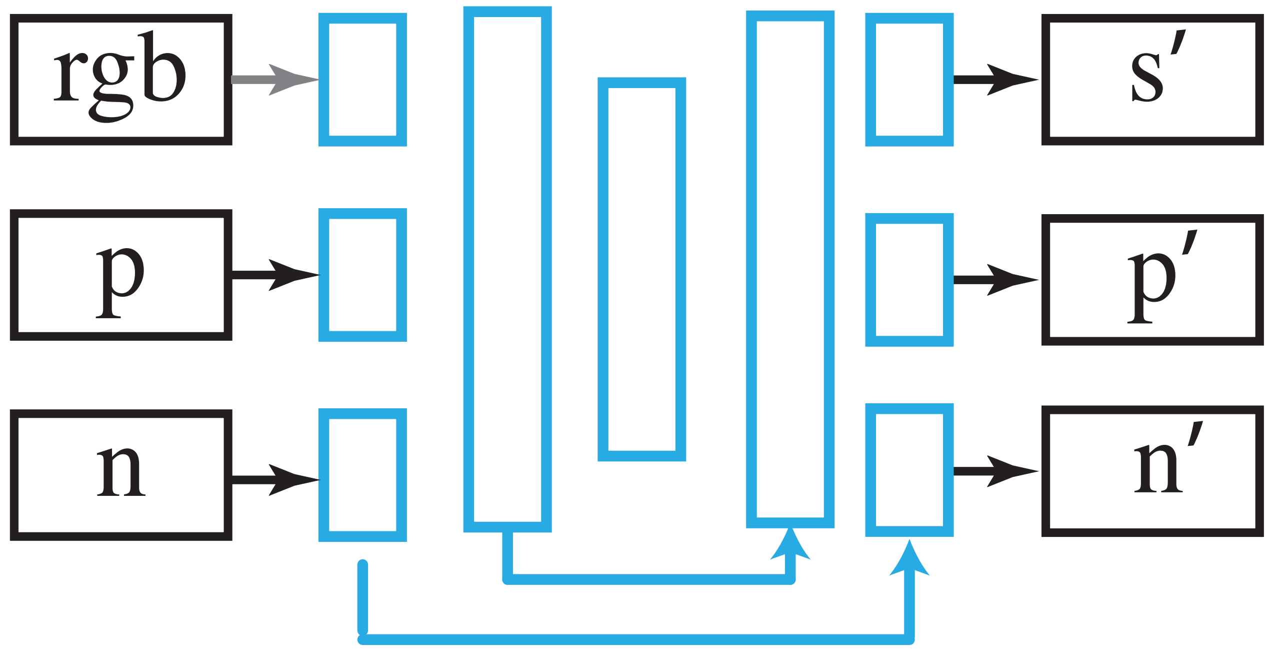




Training Objectives

Every Pixel is Correct

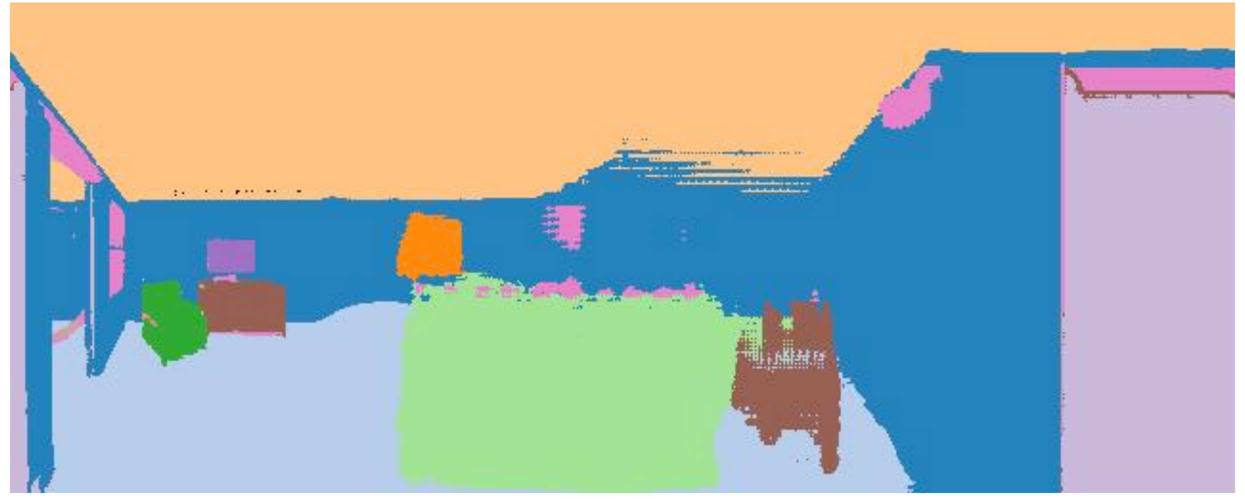




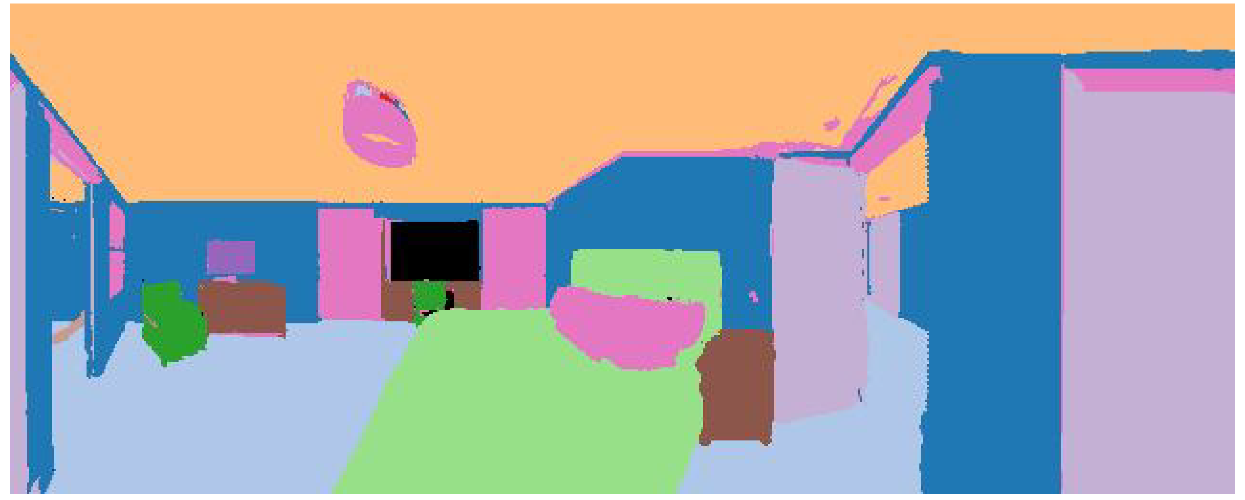
softmax

L1

cosine



Prediction



Ground truth

Training Objectives

Prediction is Plausible
rgb
p
n
s′
p′
n′
Real or Fake
Real Rooms

Adversarial loss

Goodfellow et al. 2014

Training Objectives

Similar Scene
Attribute
scene category
object distribution
rgb
p
n
s′
p′
n′
wall
floor
ceiling
⋮
chair
⋮
⋮
wall
floor
ceiling
⋮
chair
⋮
⋮

Prediction

Ground truth

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

Training Objectives

Similar Scene Attribute

scene category

object distribution

Prediction

Ground truth

rgb
s
p
n
s′
p′
n′
G: generator

Training Objectives

Every Pixel is Correct

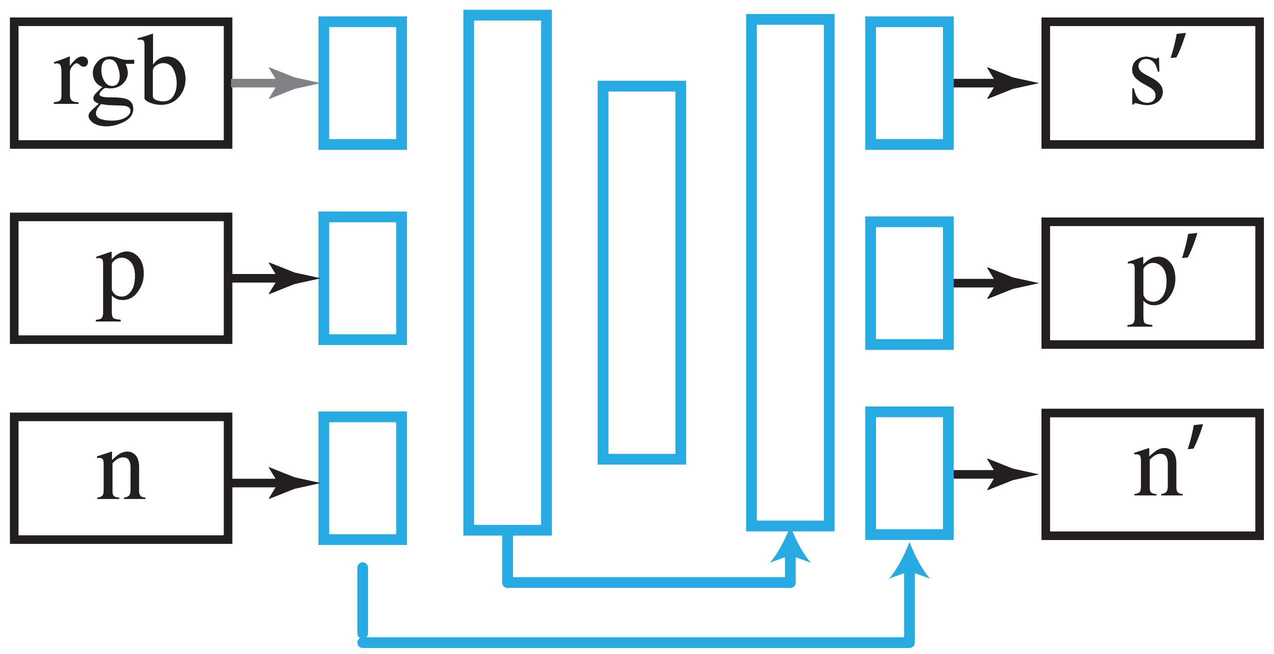
L_{recon}

Similar Scene Attribute

$L_{attribute}$

Prediction is Plausible

L_{adv}

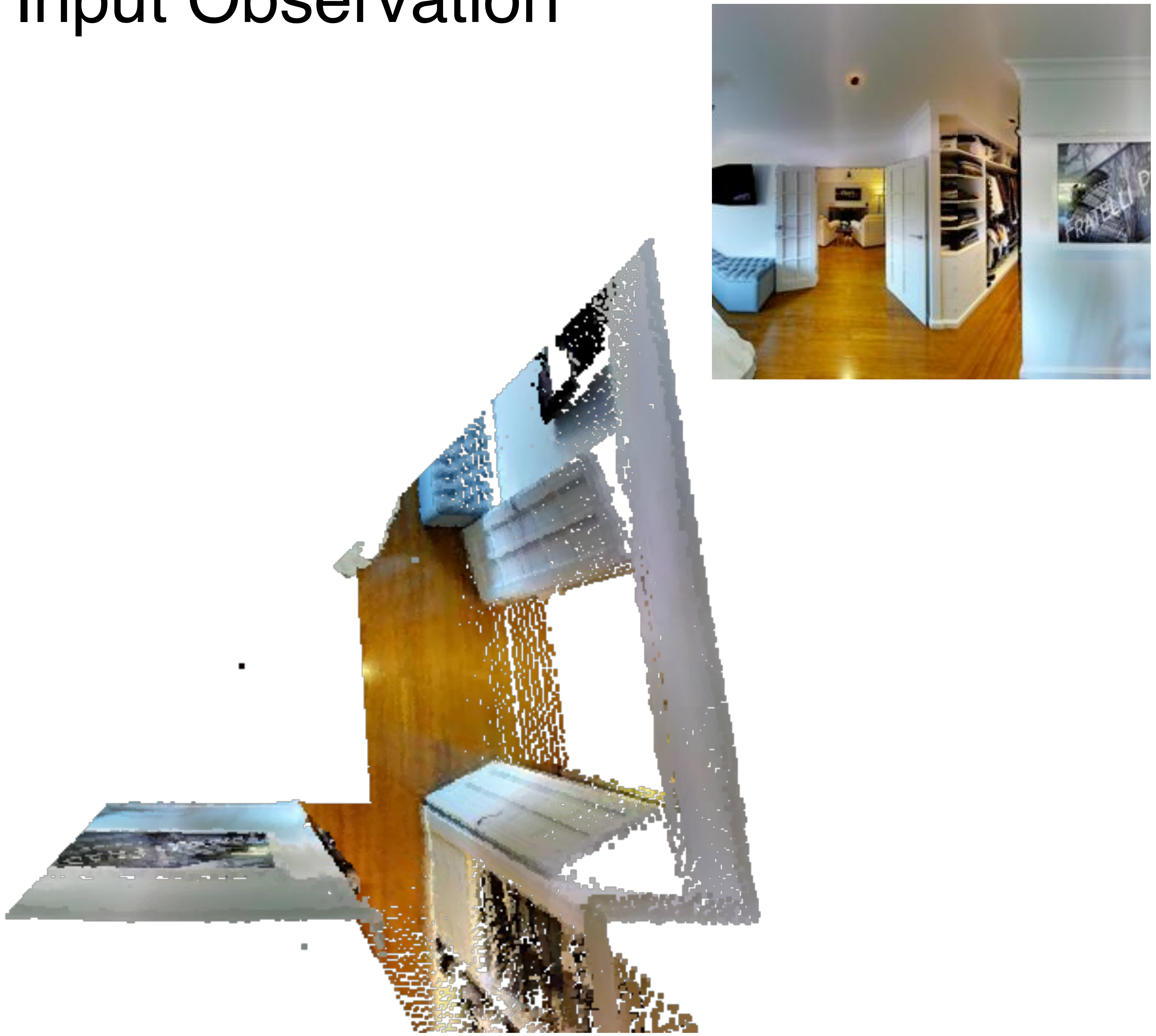


$$L = \lambda_1 L_{recon} + \lambda_2 L_{attribute} + \lambda_3 L_{adv}$$

Results

Results

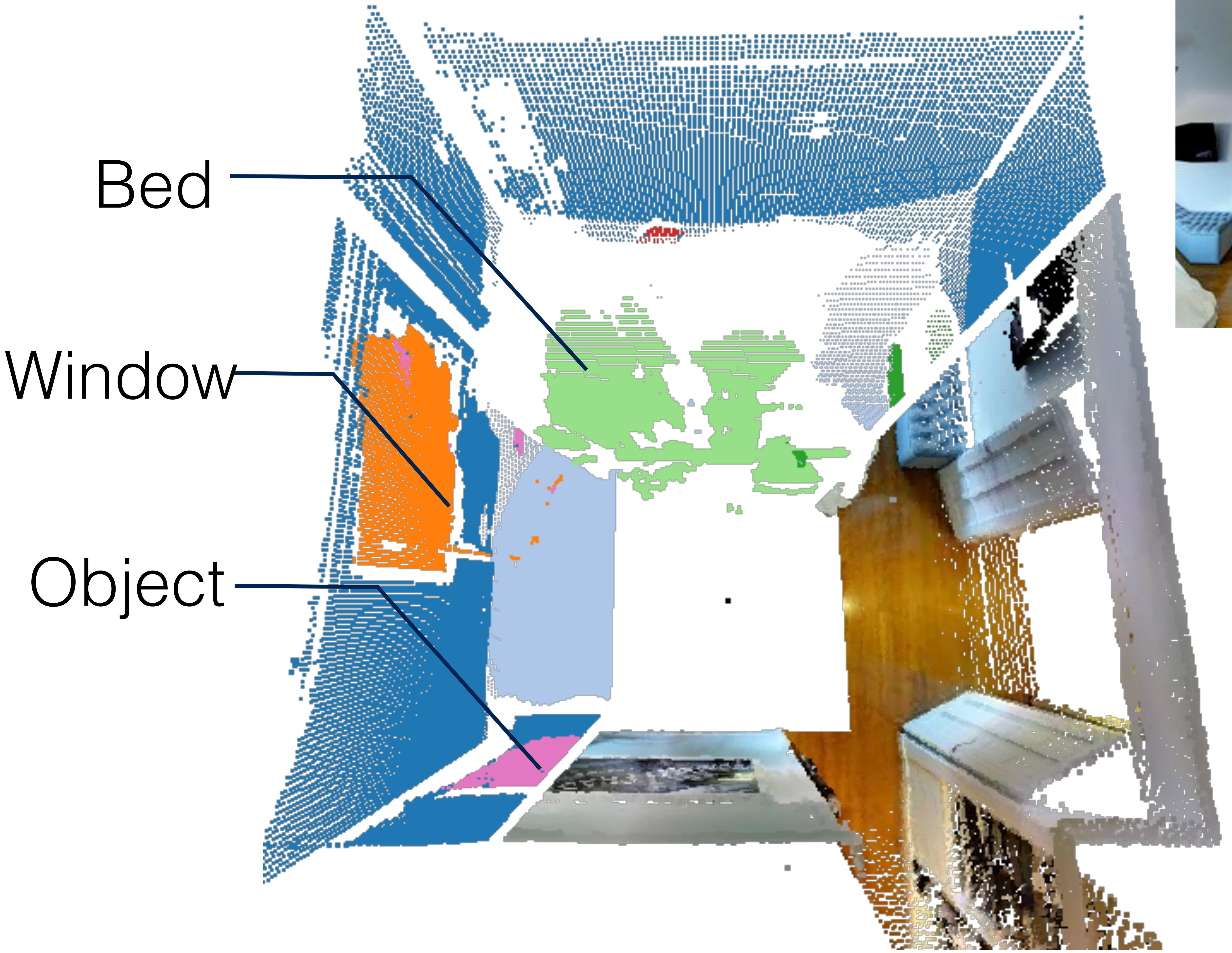
Input Observation



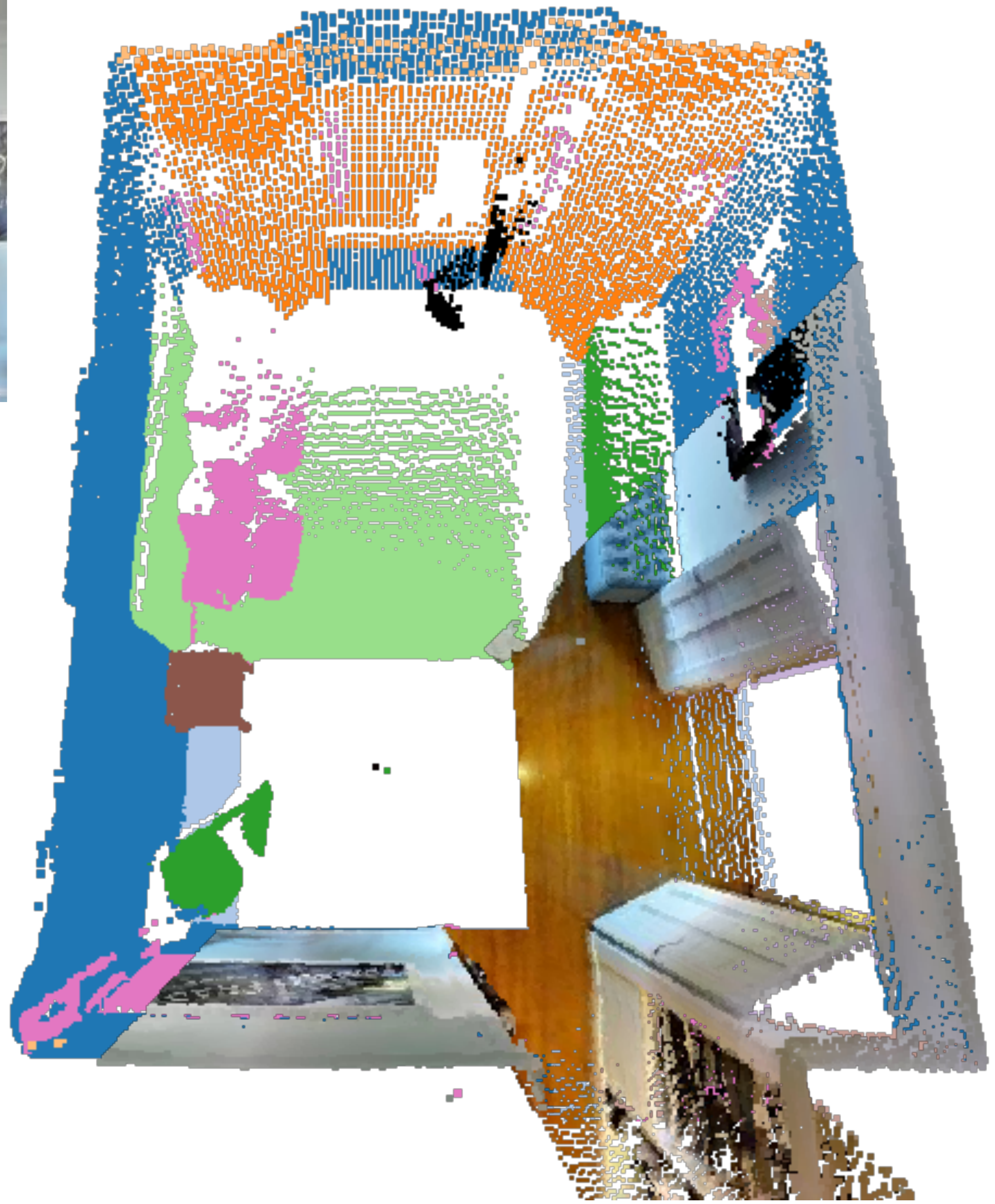
● ceiling ● wall ● floor ● window ● bed ● door ● cabinet ● chair ● sofa ● tv ● table ● object ● furniture

Results

Prediction

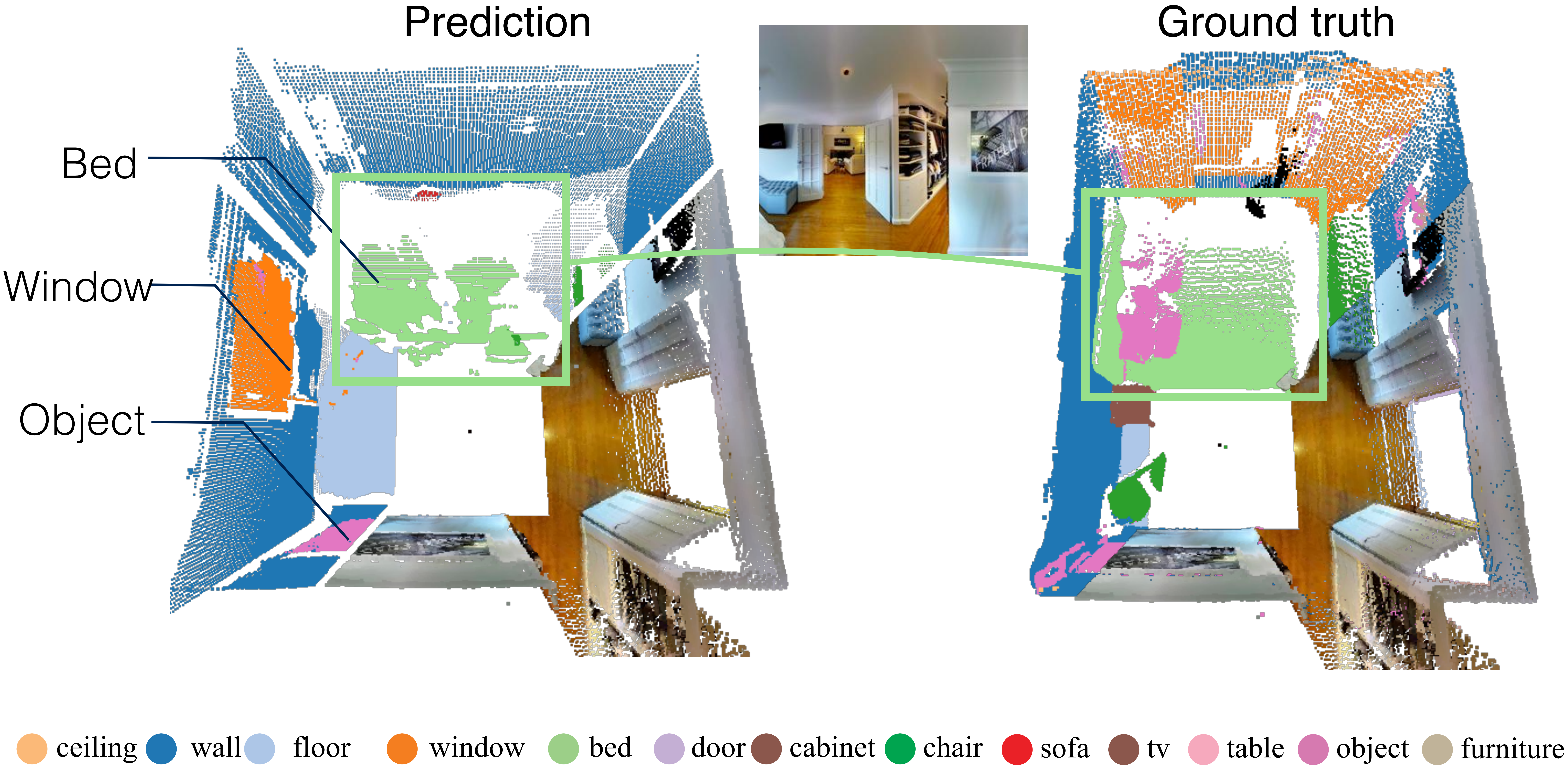


Ground truth

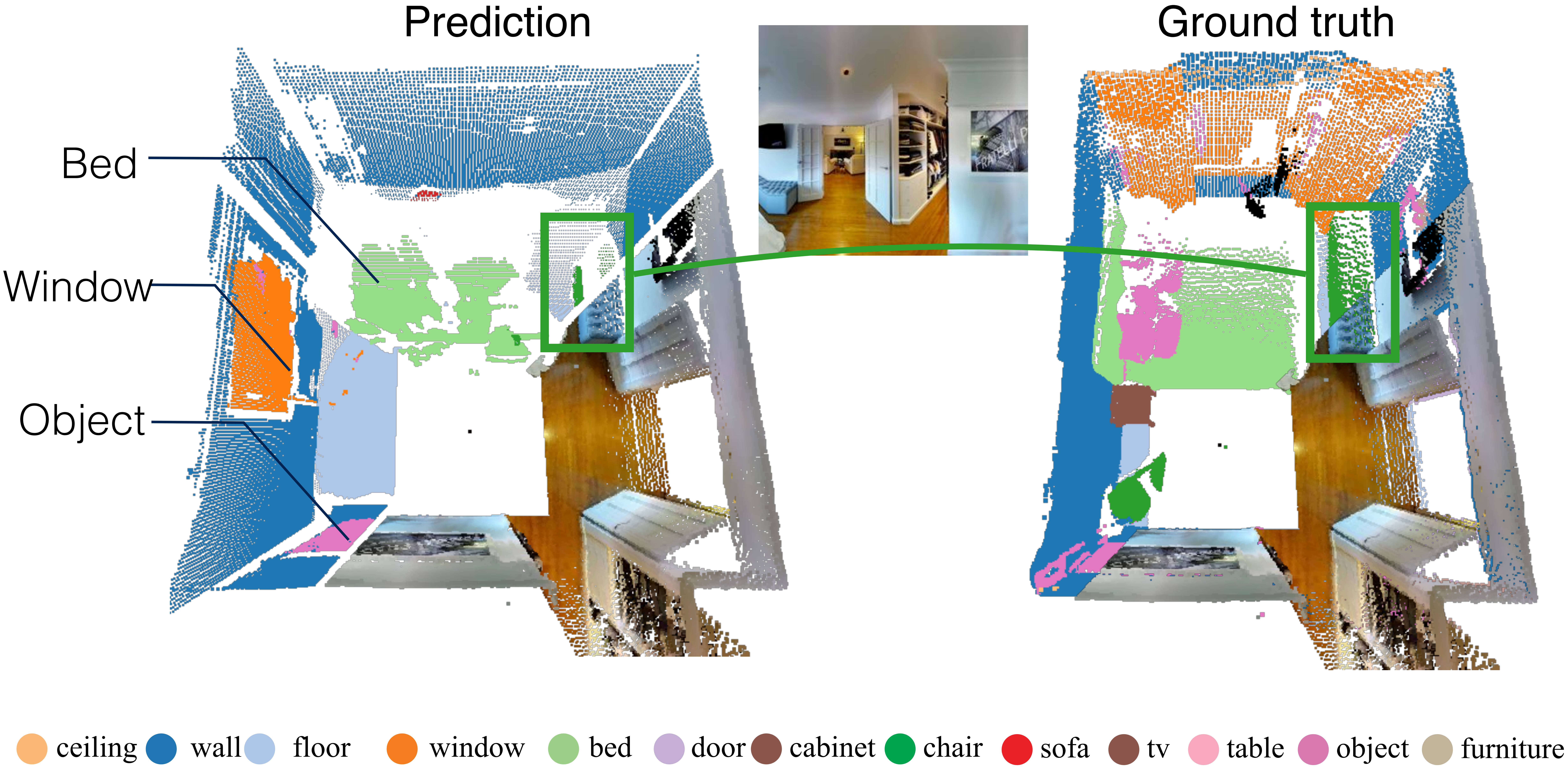


● ceiling ● wall ● floor ● window ● bed ● door ● cabinet ● chair ● sofa ● tv ● table ● object ● furniture

Results


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

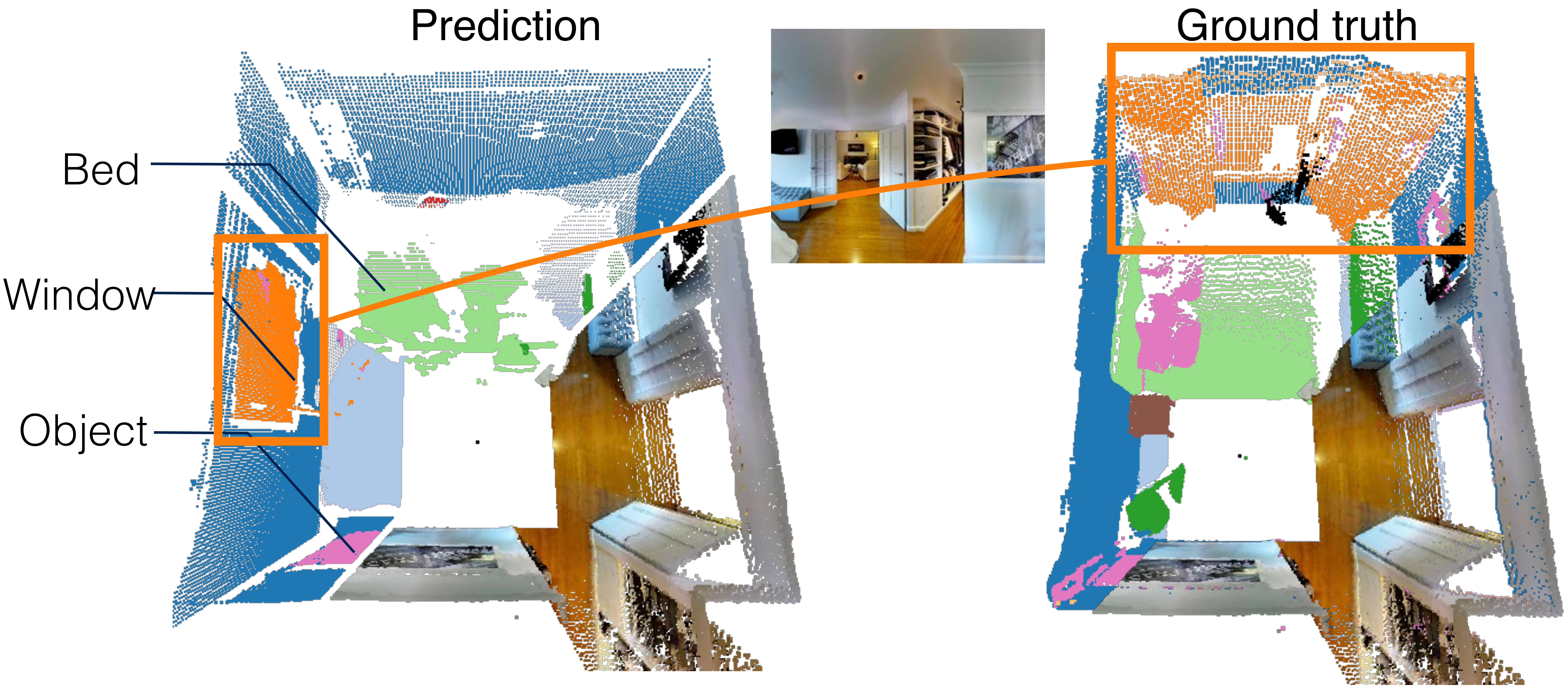
Results



Prediction

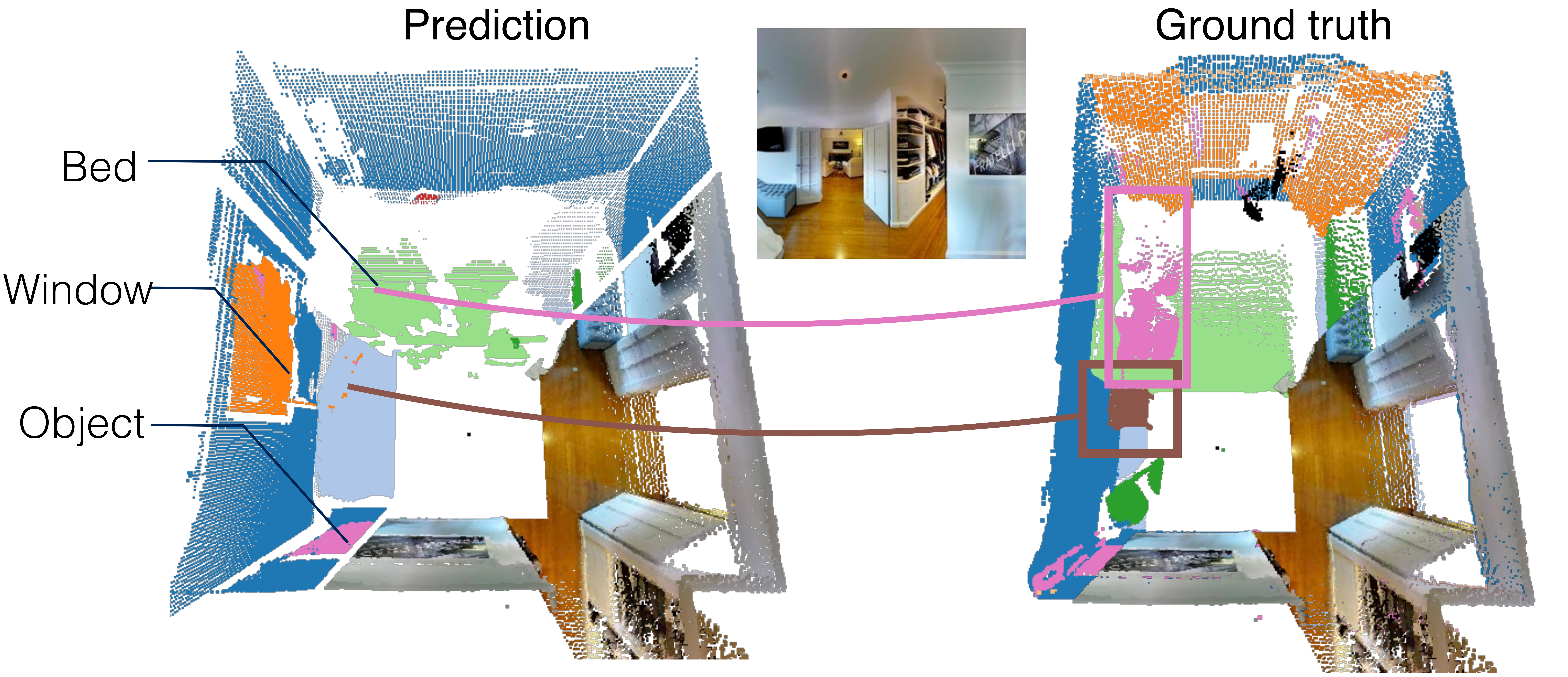
Ground truth

Results



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

Results



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

Results

Input Observation

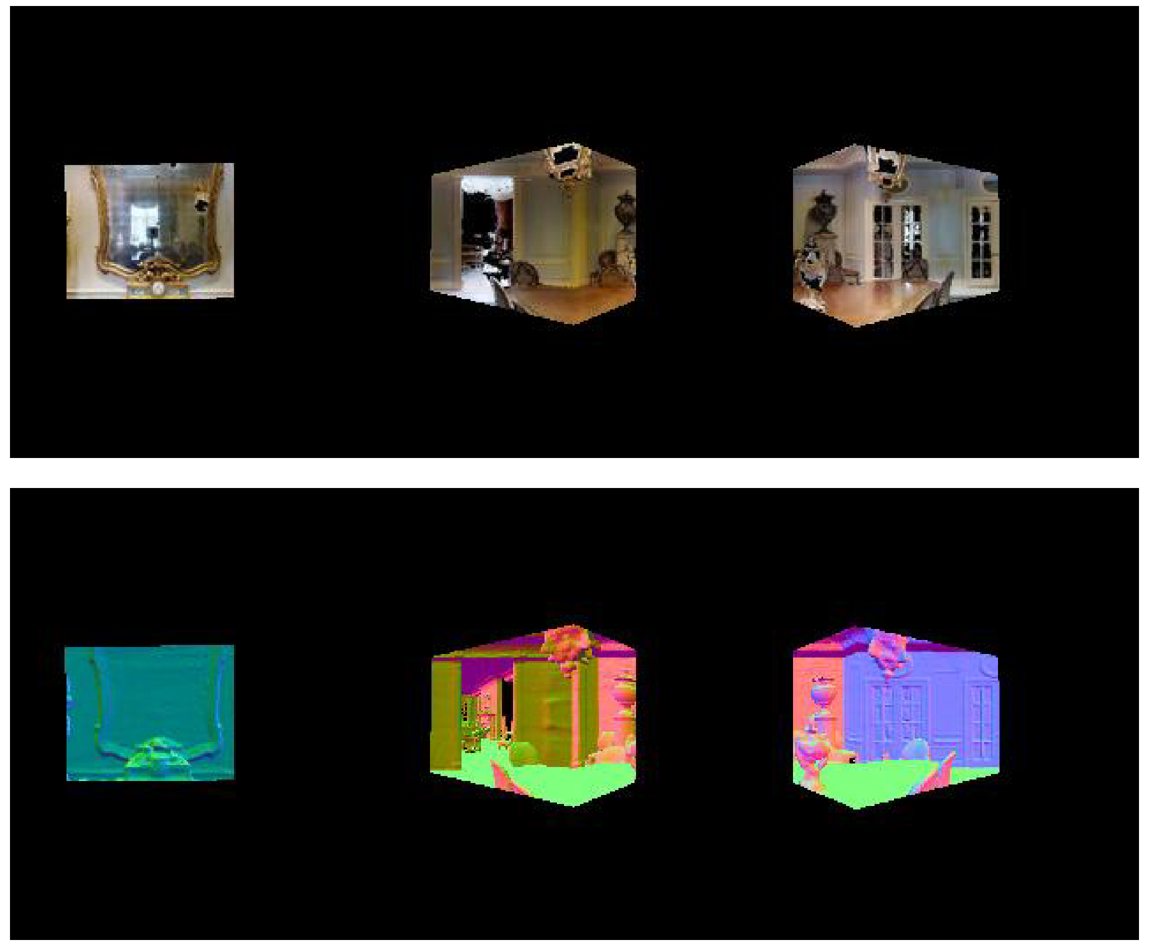
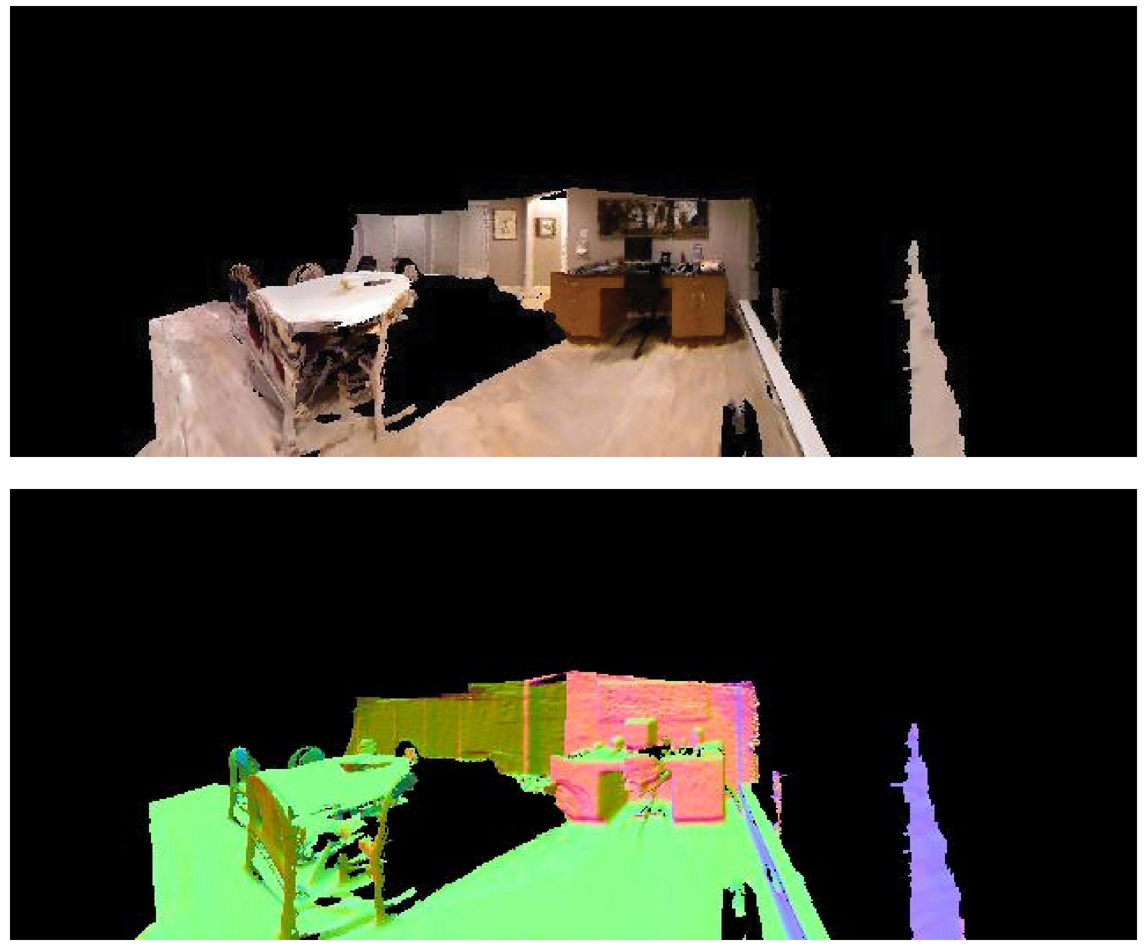
● ceiling ● wall ● floor ● window ● bed ● door ● cabinet ● chair ● sofa ● tv ● table ● object ● furniture

Results

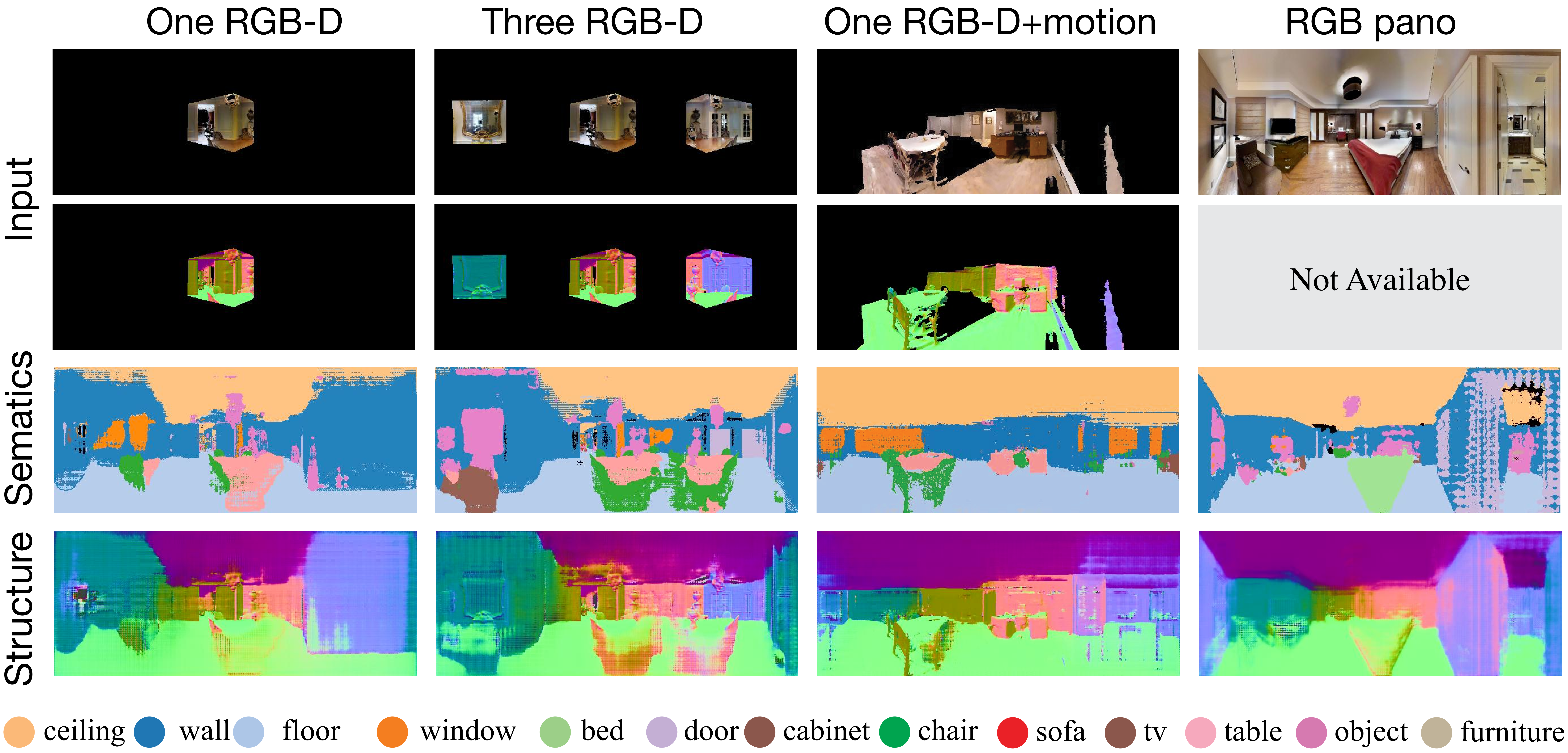
Prediction
door
sofa
window
objects
Ground truth
window
Bed

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

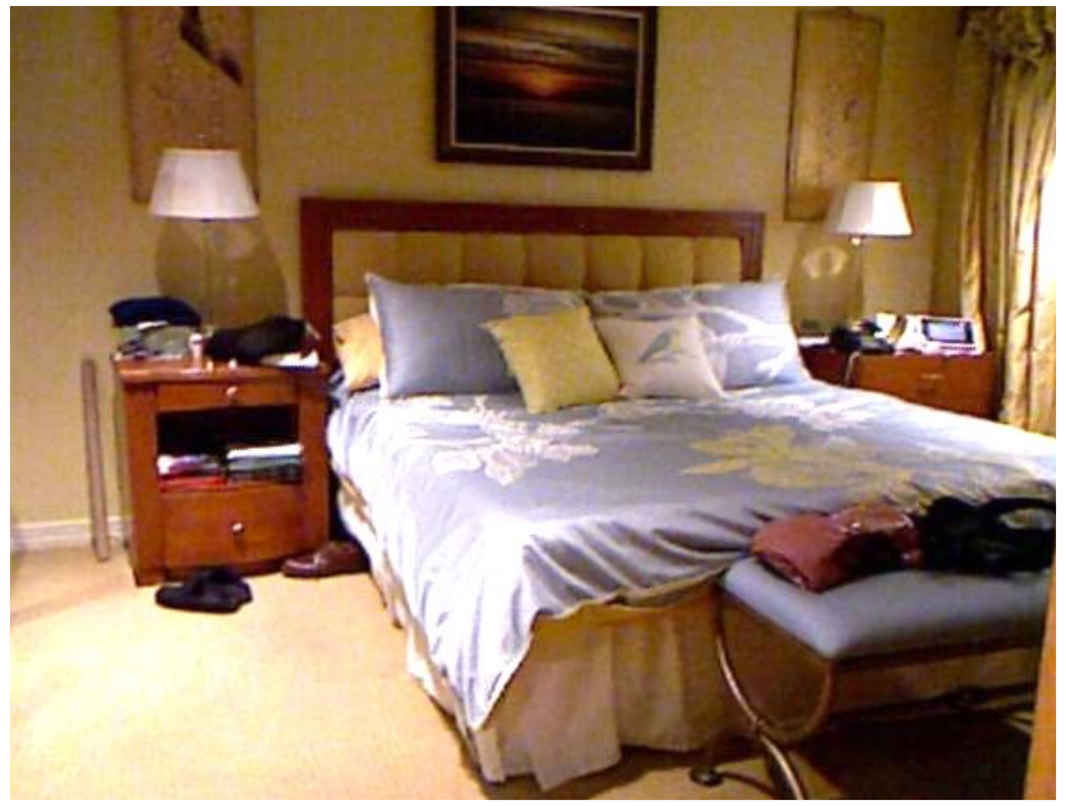
Camera Configurations in real platforms

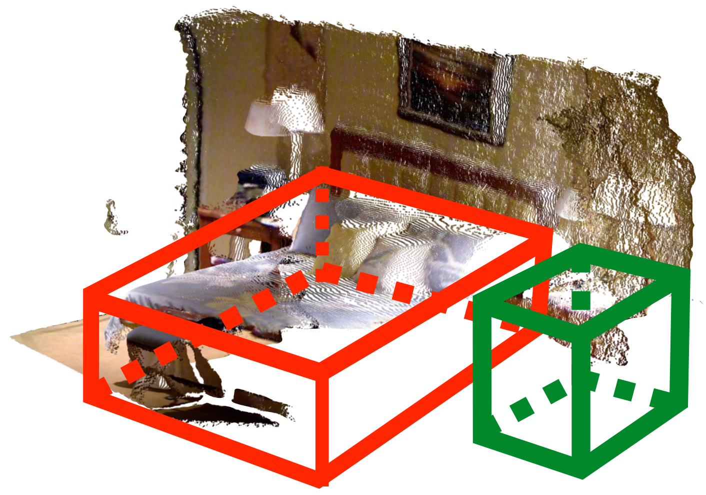
	One RGB-D	Three RGB-D	One RGB-D+motion	RGB pano
Input				 Not Available
Device				

Camera Configurations

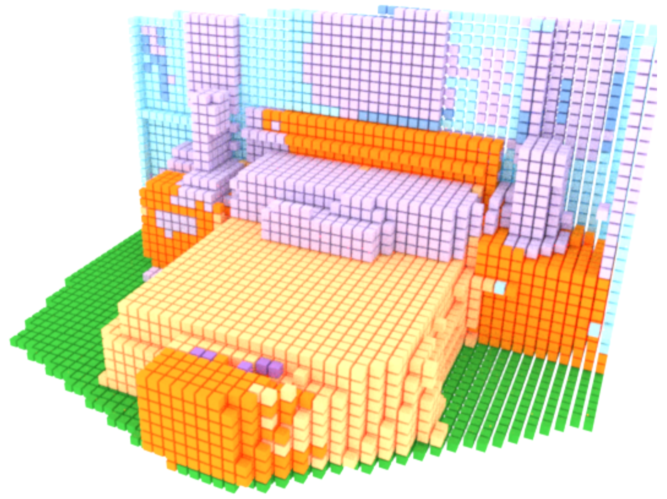

One RGB-D
Three RGB-D
One RGB-D+motion
RGB pano
Input
Not Available
Sematics
Structure
● ceiling ● wall ● floor ● window ● bed ● door ● cabinet ● chair ● sofa ● tv ● table ● object ● furniture

Advances Towards 3D Scene Understanding

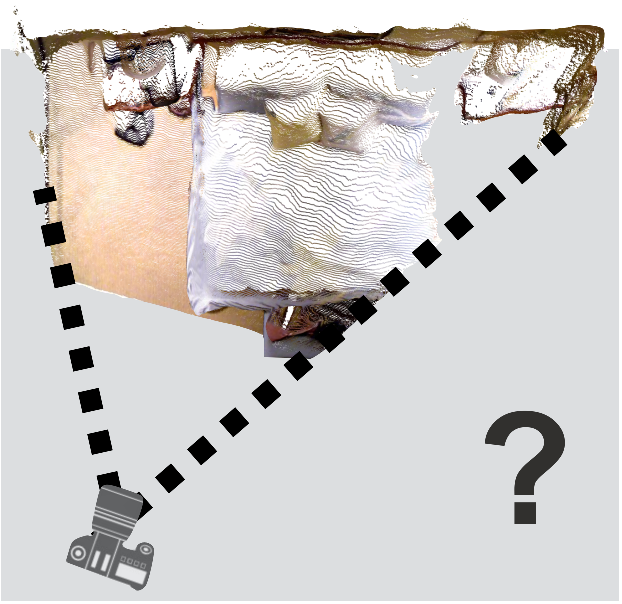




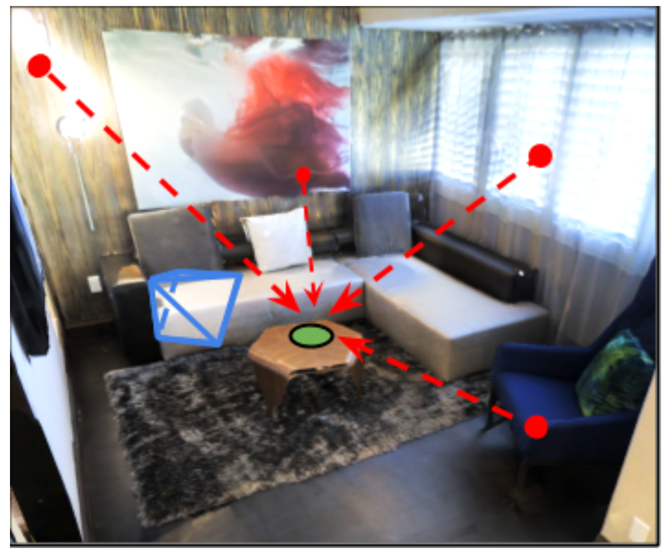
Amodal 3D Bounding Boxes
[Song and Xiao ECCV’14,CVPR’16]



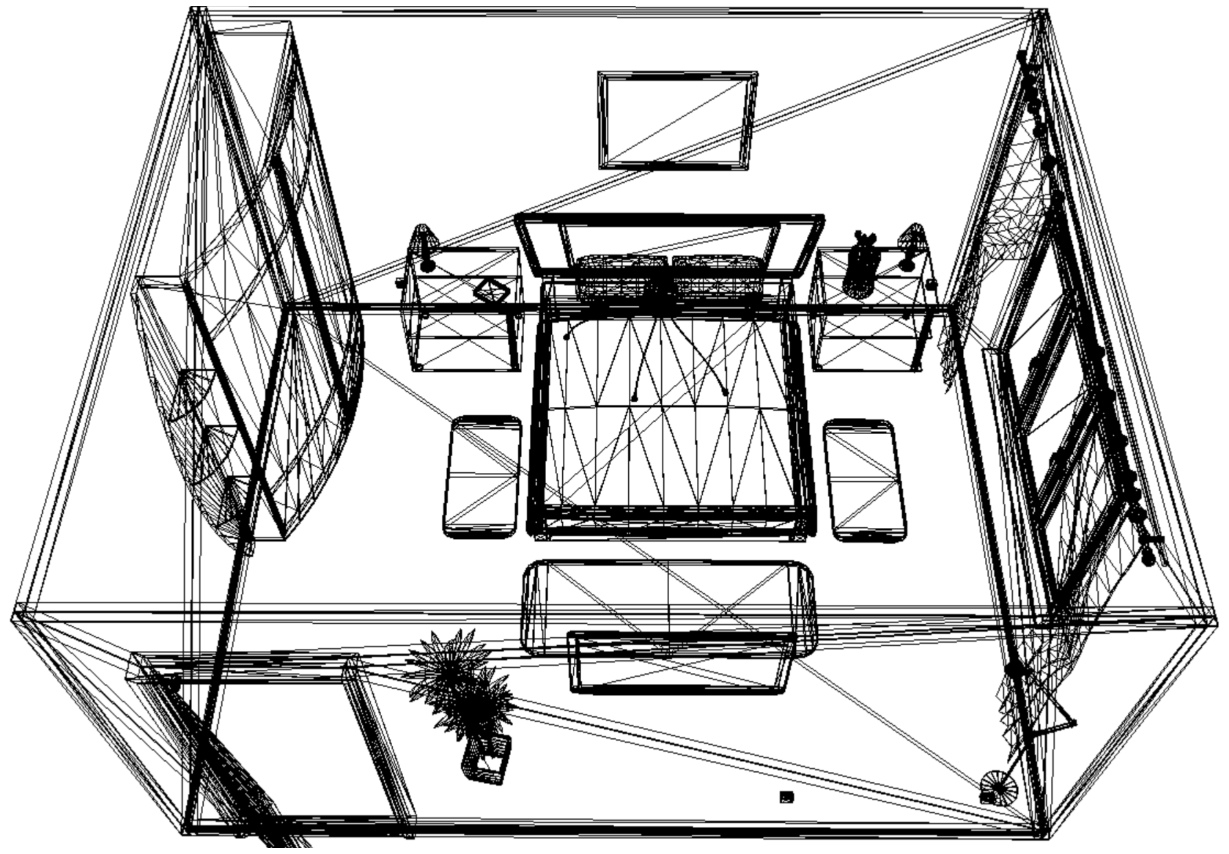
Higher Fidelity 3D Voxels
[Song et al. CVPR’17]



Beyond FoV Semantics&Structure
[Song et al. CVPR’18]



Beyond FoV Illumination
[Song and Funkhouser]



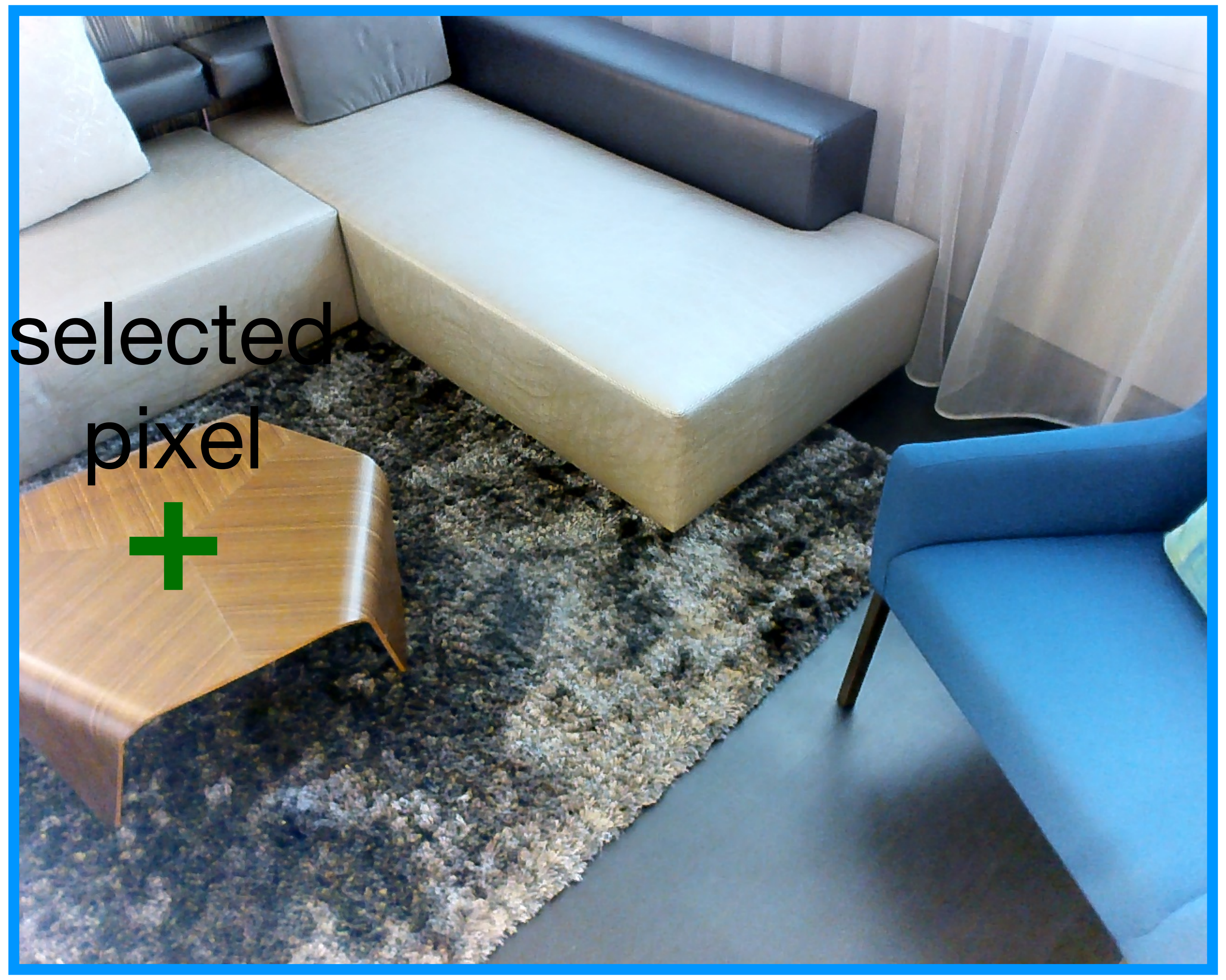
- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Lighting information
- Surface materials
- Phys. Properties
- …

Neural Illumination

Lighting Prediction for Indoor Environments

Shuran Song and Thomas Funkhouser

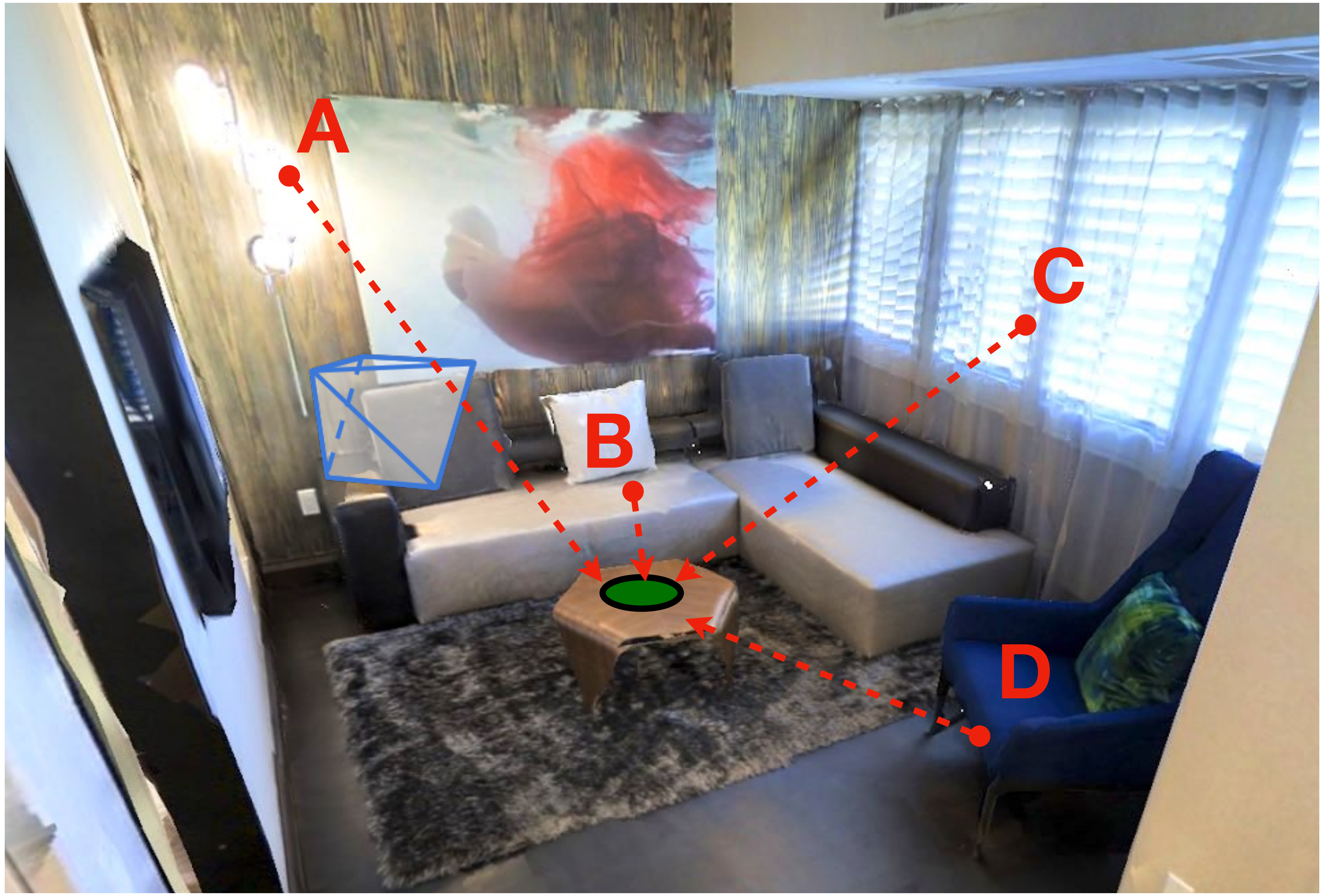
Neural Illumination



Input: Image + Selected pixel

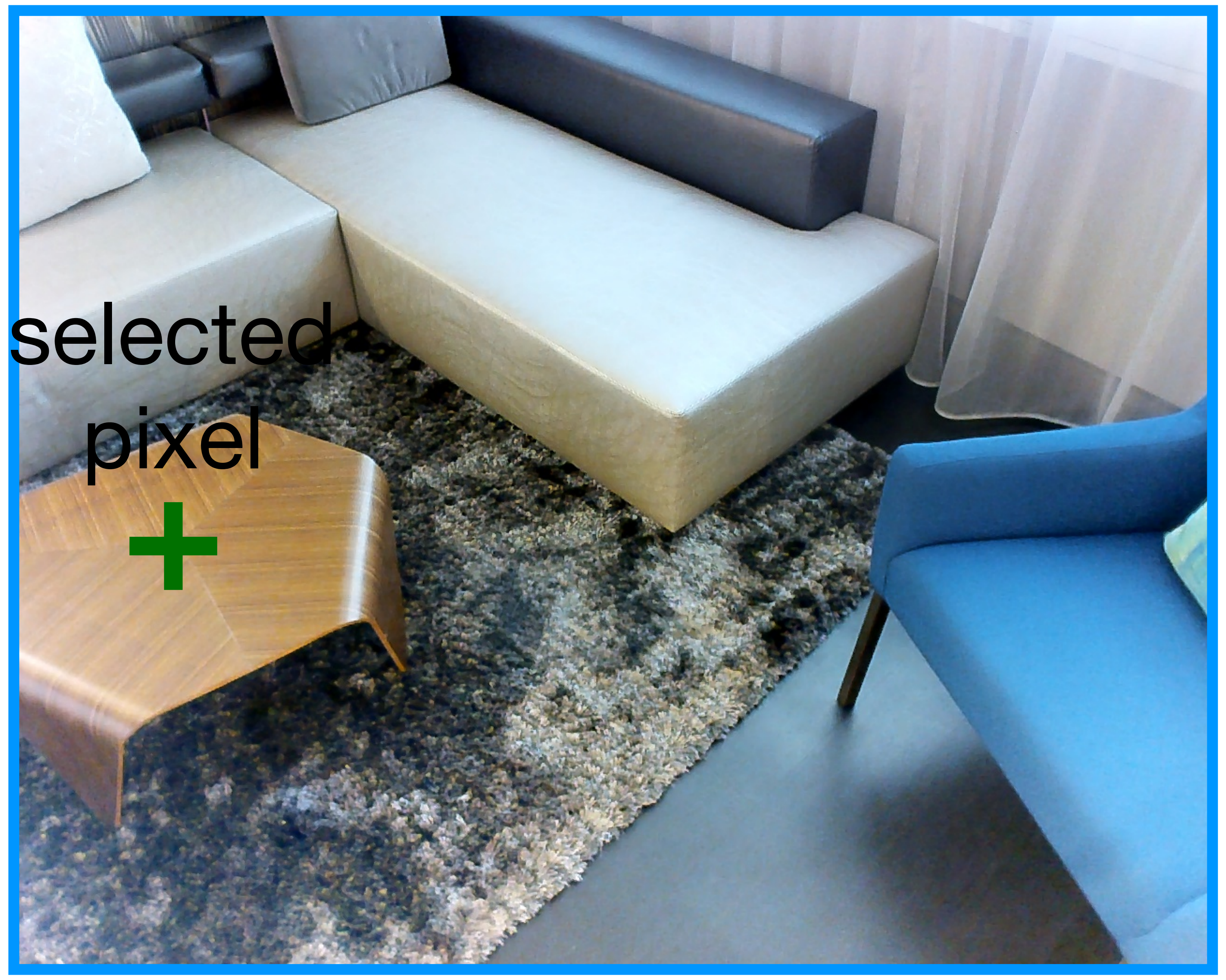
Neural Illumination

Input: Image + Selected pixel

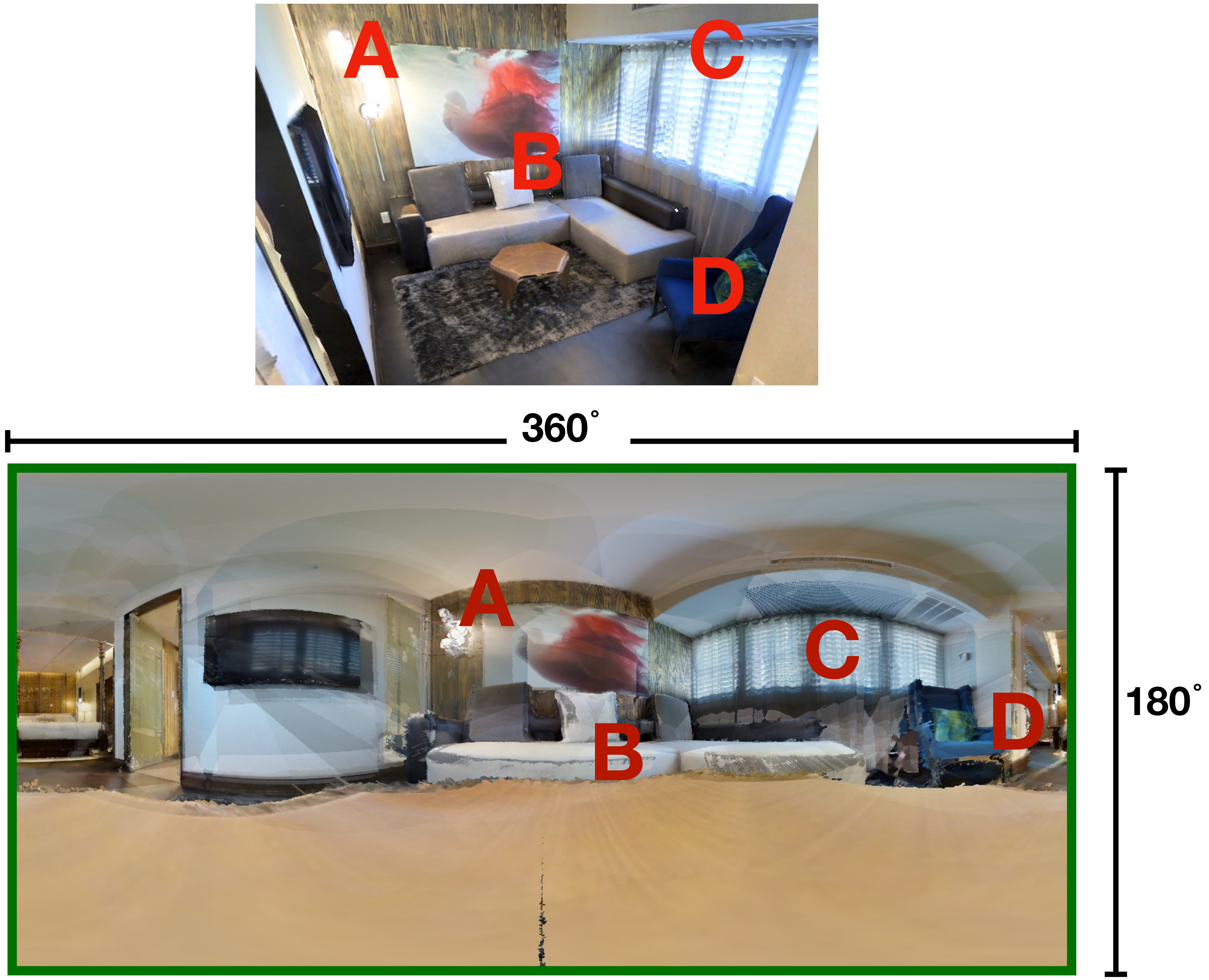


Goal: output estimated the incoming light from all directions to the selected locale

Neural Illumination



Input: Image + Selected pixel



Goal: estimating the incoming light from all directions to the selected locale

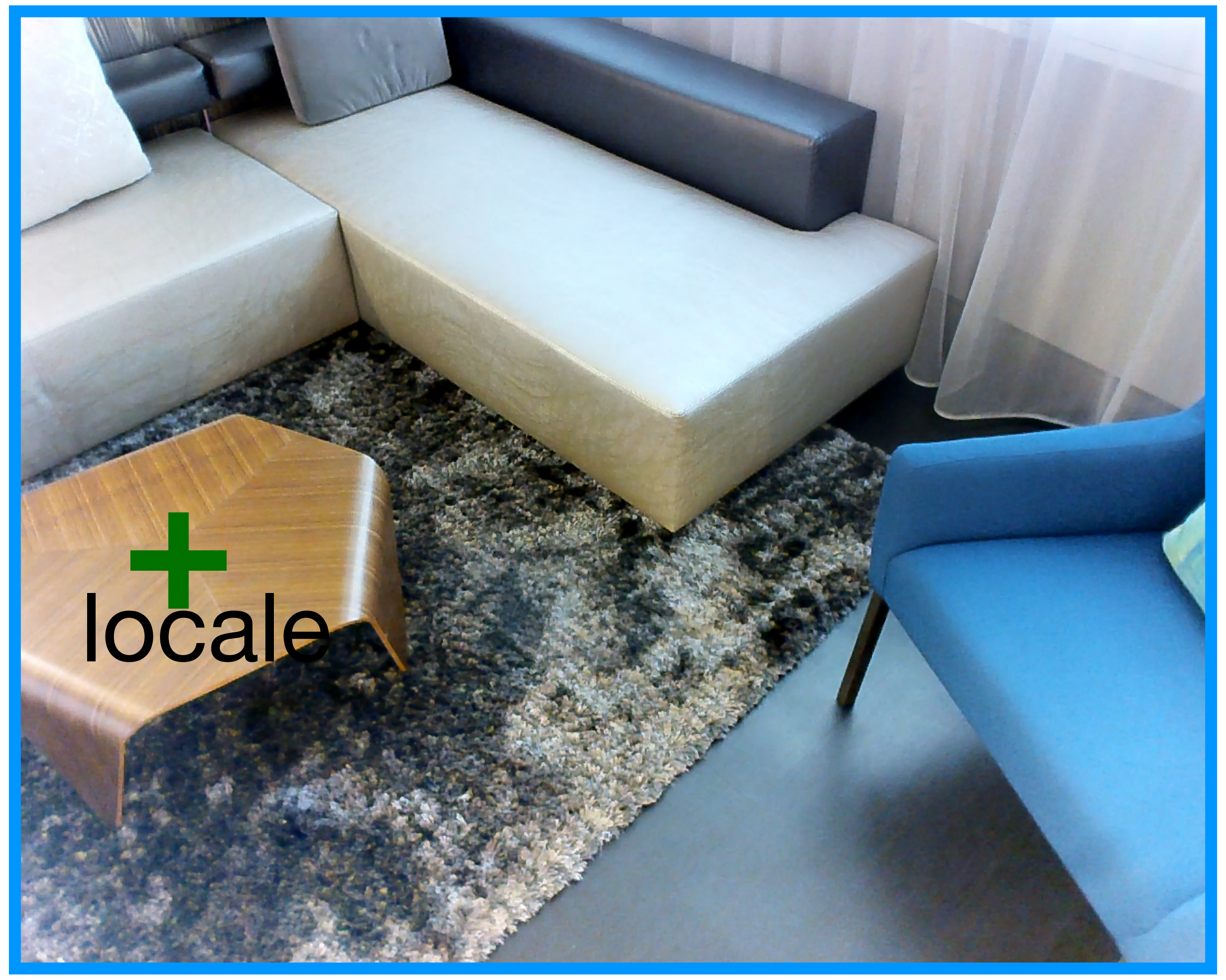
Neural Illumination

Virtual Object Relighting

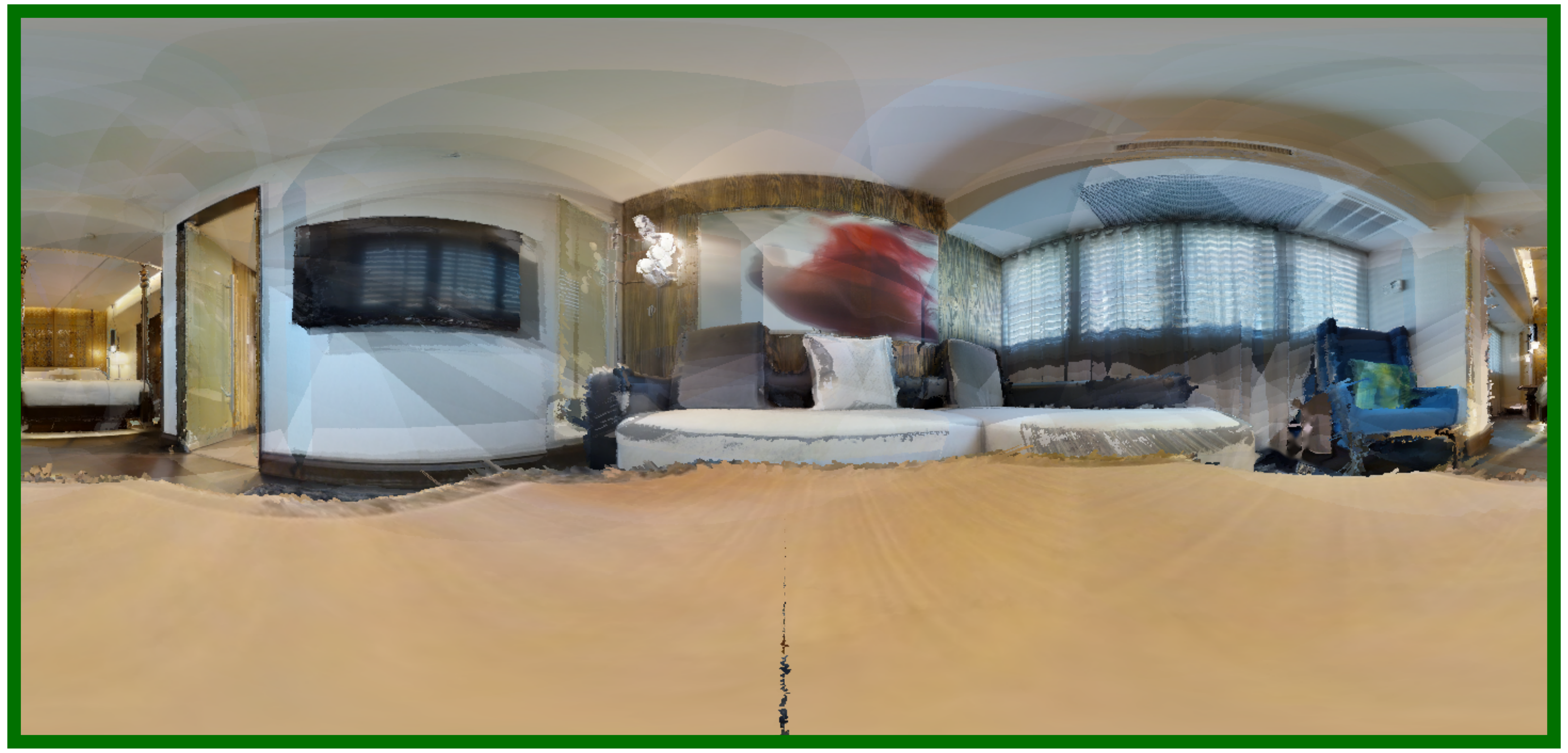
Input: Image + Selected p

Goal: estimating the incoming light from all directions to the selected locale

Illumination Estimation



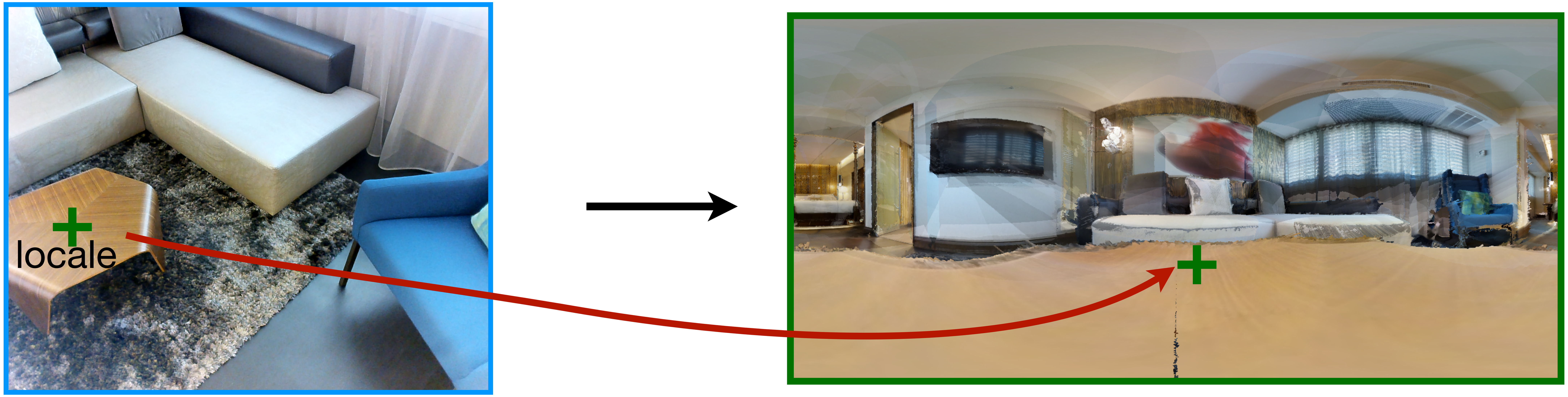
Input: Image+Selected pixel



Output: Illumination map

Requires a **comprehensive** understanding of the environment,
in order to predict a ***complete*** illumination map from a ***partial*** RGB observation.

Illumination Estimation

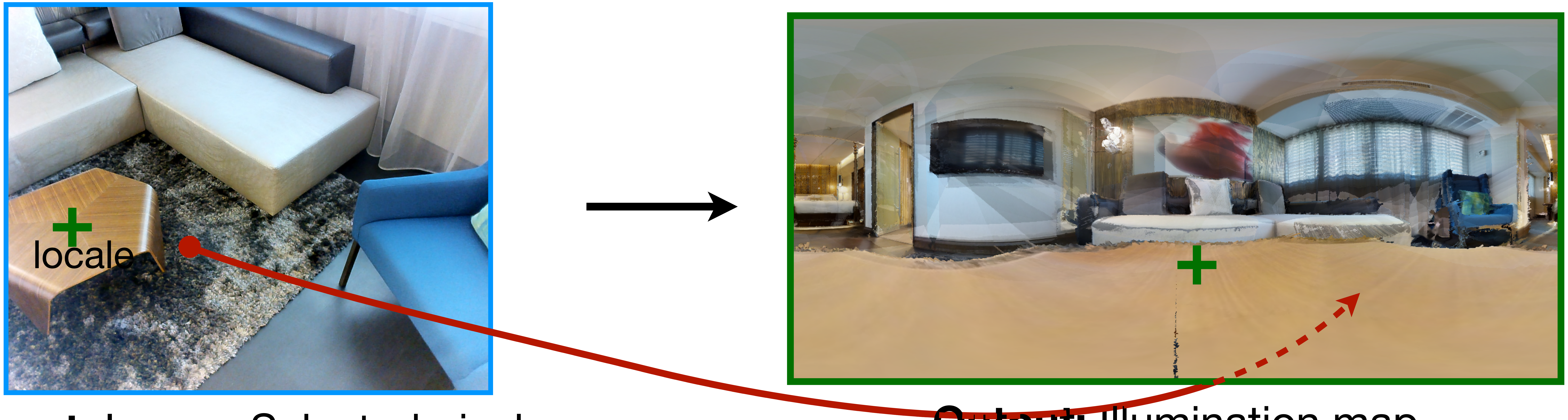


Input: Image+Selected pixel

Output: Illumination map

- The 3D location of the selected pixel

Illumination Estimation

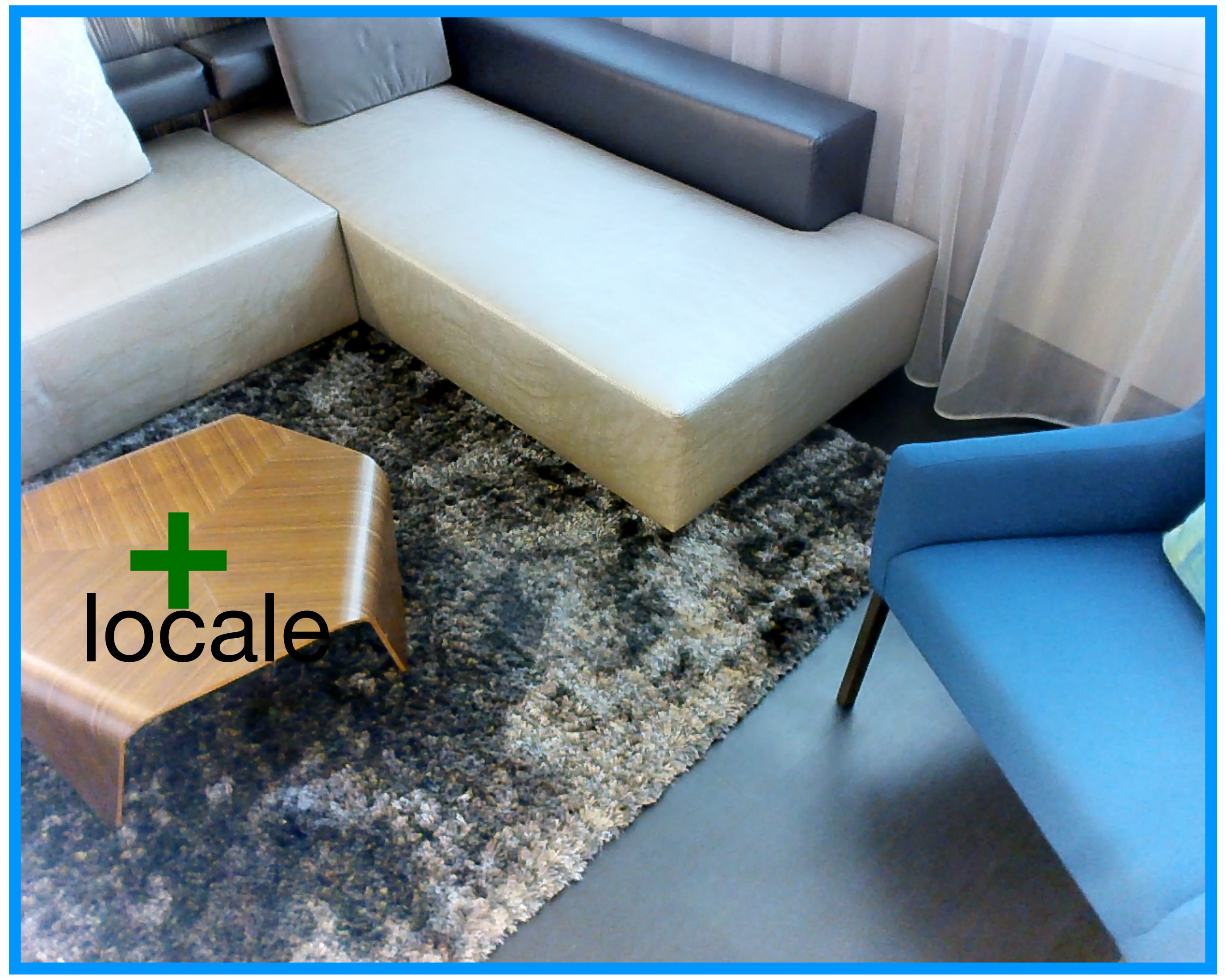


Input: Image+Selected pixel

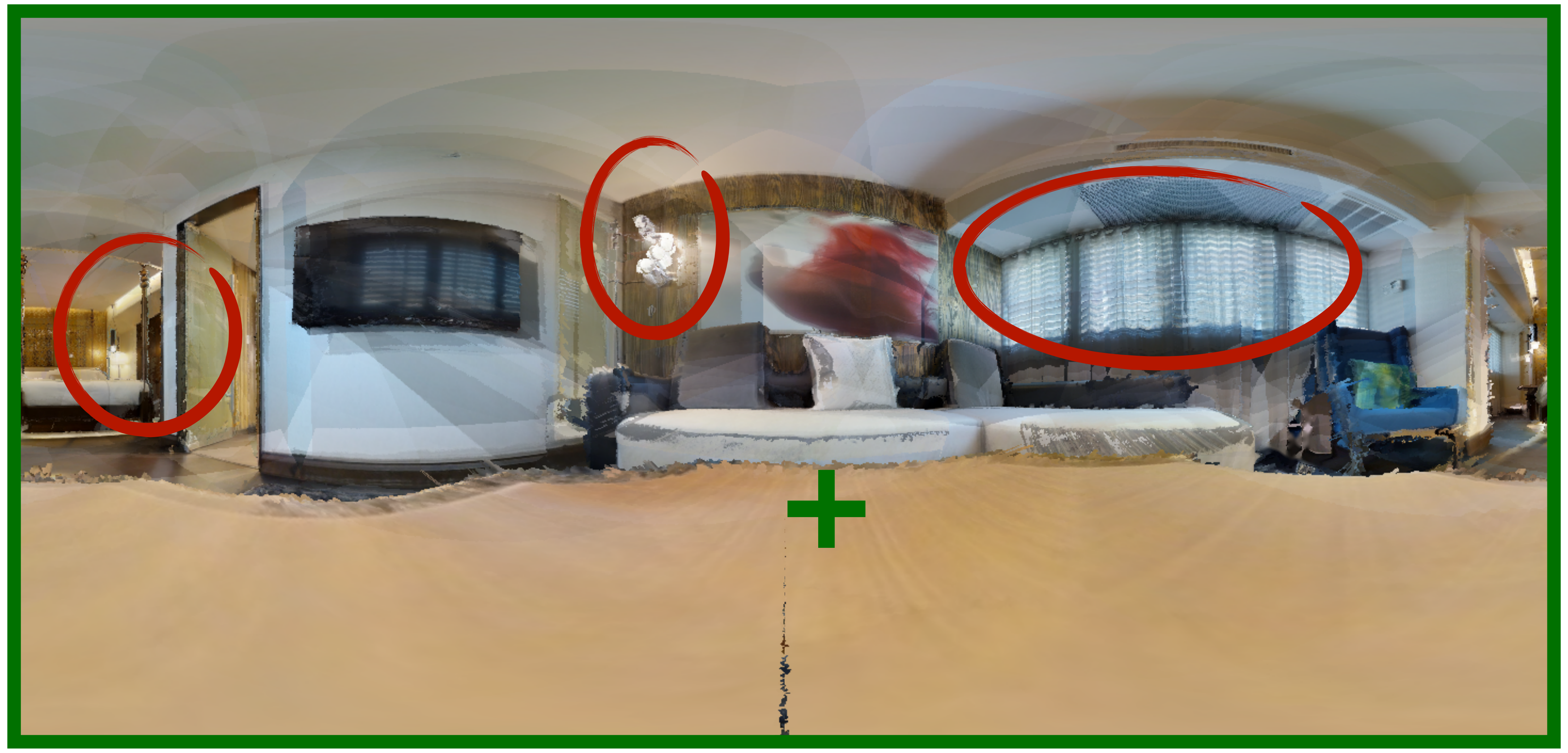
Output: Illumination map

- The 3D location of the selected pixel
- The occluded light source caused by scene geometry

Illumination Estimation



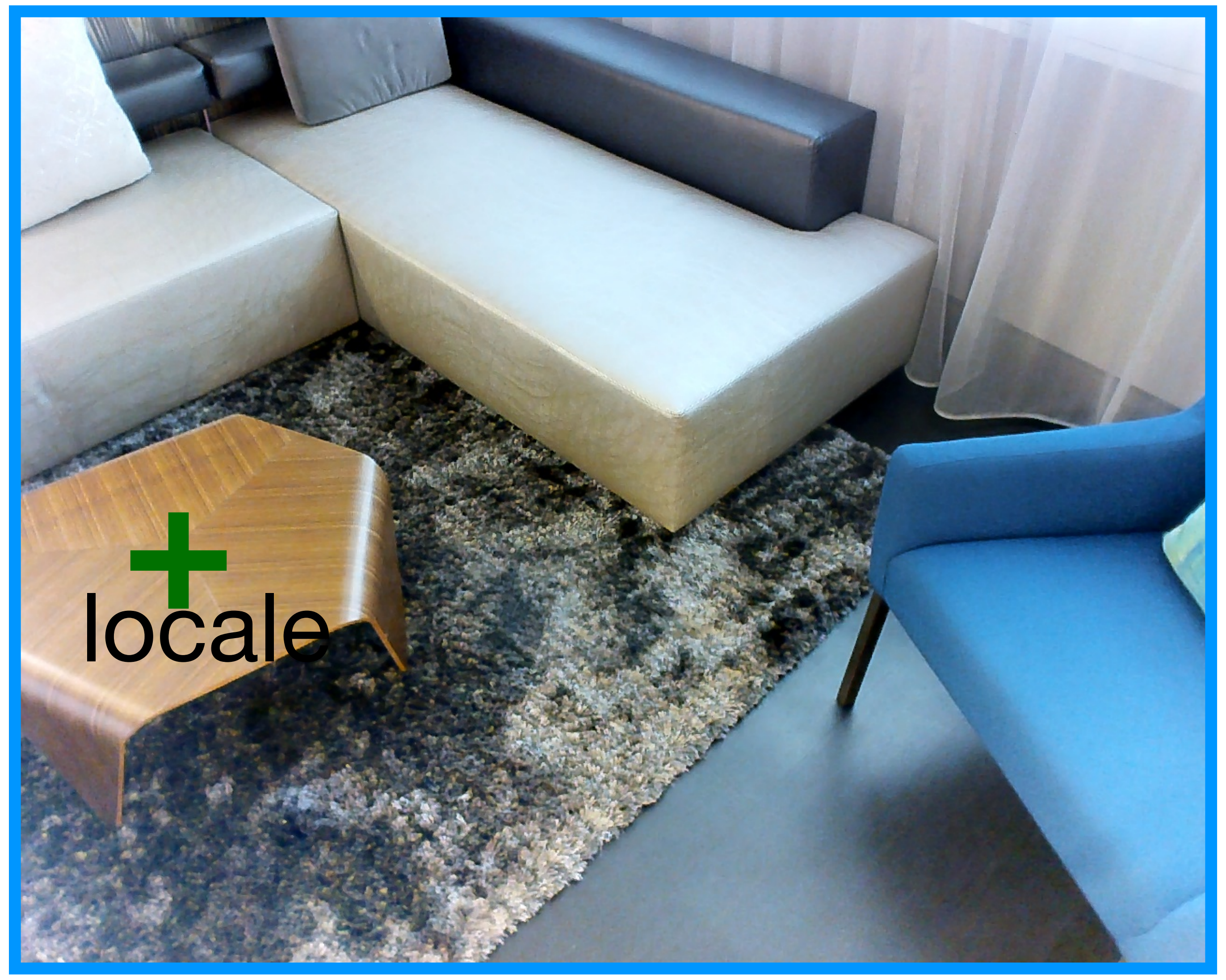
Input: Image+Selected pixel



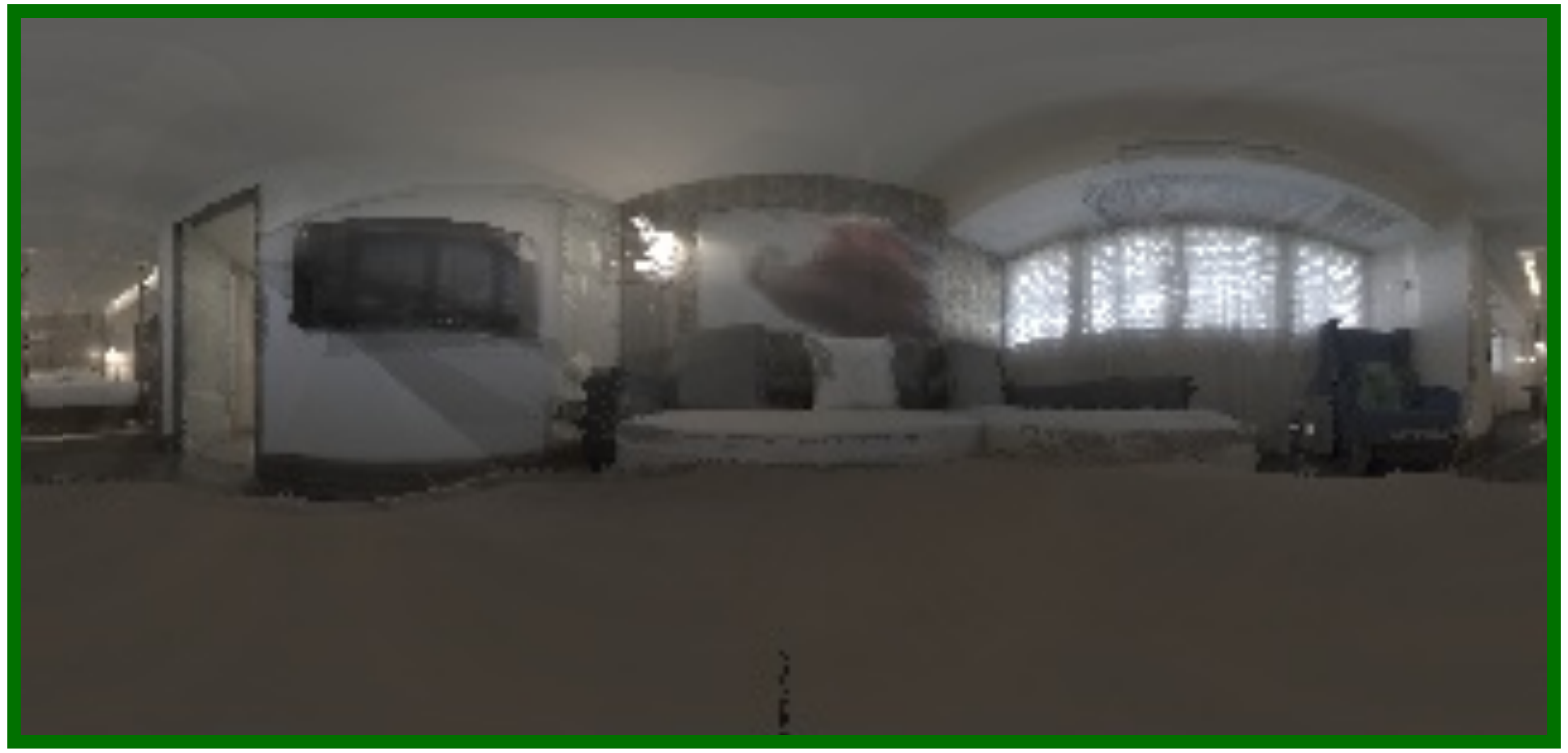
Output: Illumination map

- The 3D location of the selected pixel
- The occluded light source caused by scene geometry
- The distribution of unobserved light sources

Illumination Estimation



Input: Image+Selected pixel

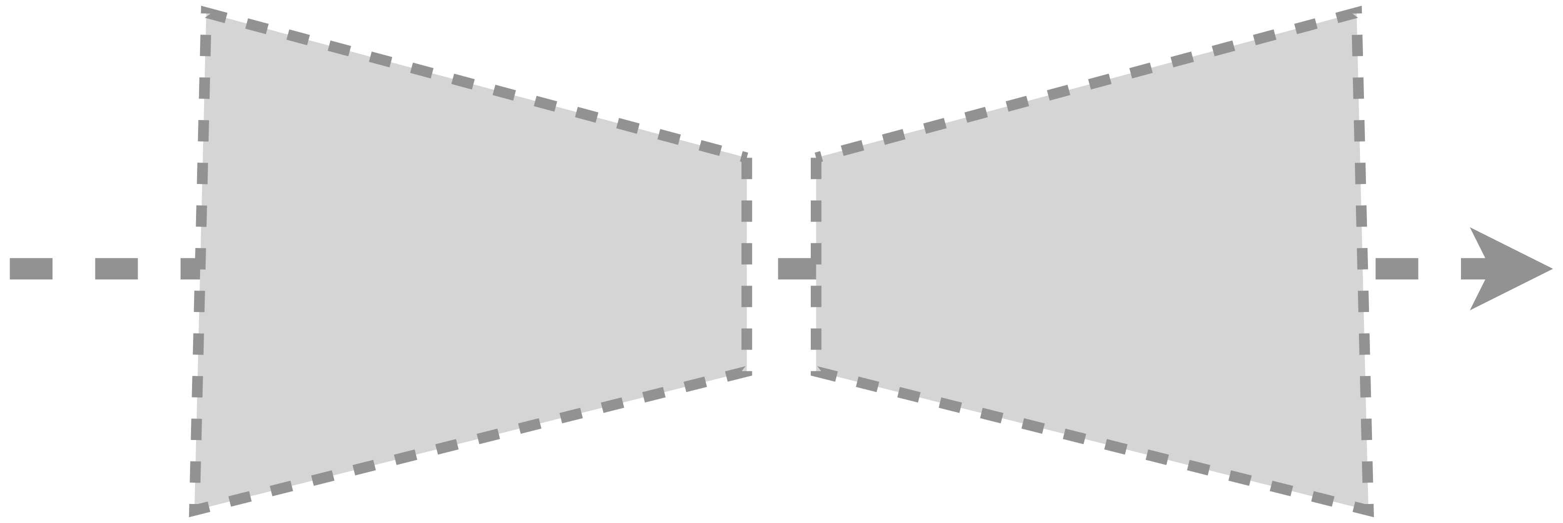


Output: Illumination map

- The 3D location of the selected pixel
- The occlusions caused by scene geometry
- The distribution of unobserved light sources
- The missing high dynamic range information

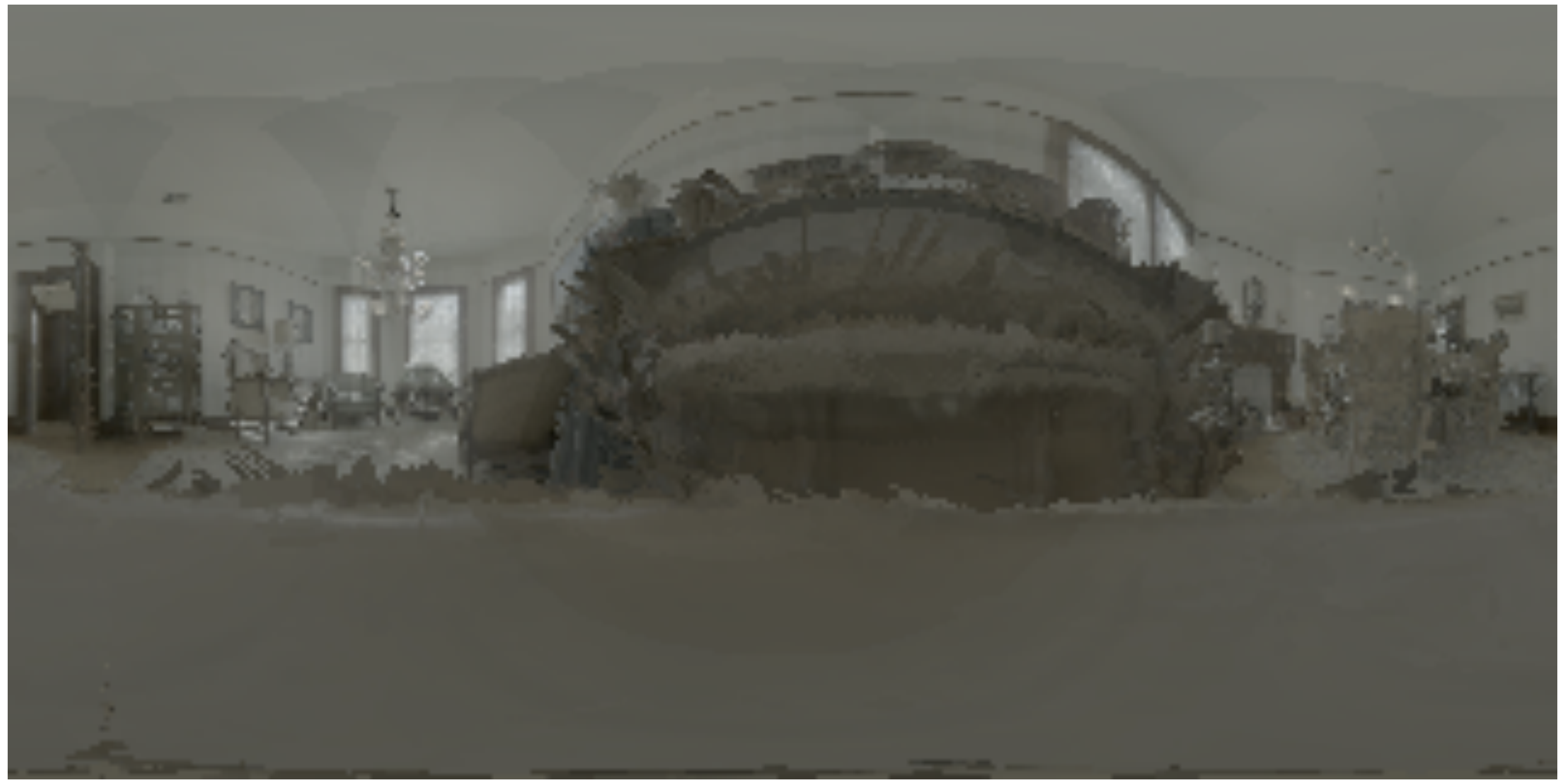
Prior work

Input: Image+selected pixel

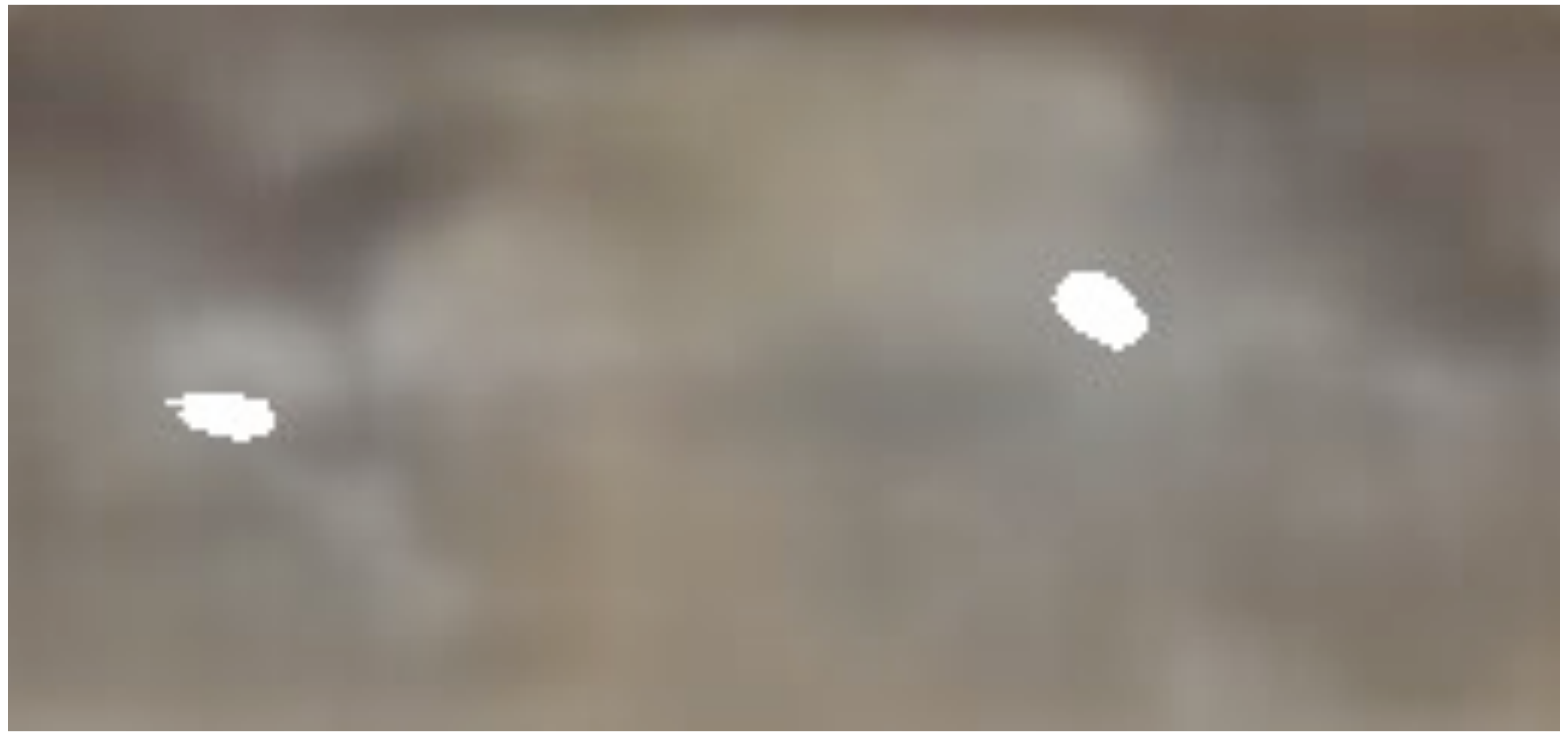


A Single Black-Box Network

(in log scale)

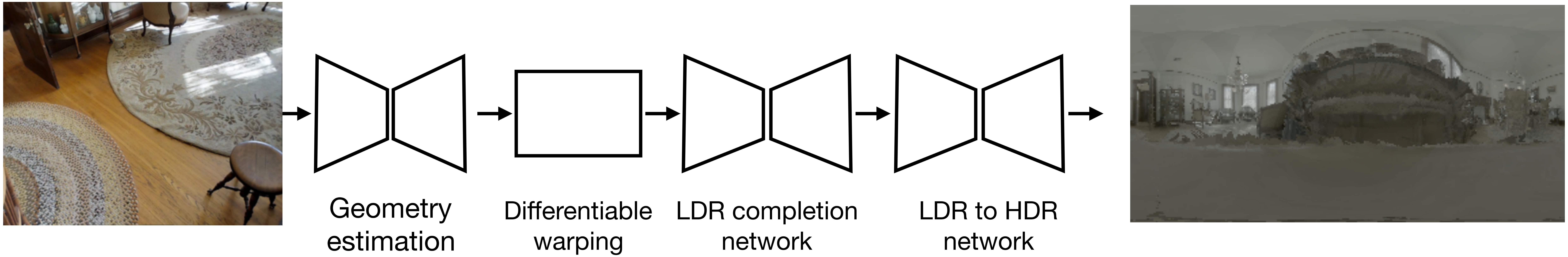


Output: HDR Illumination map



Gardner *et al.*

Neural Illumination



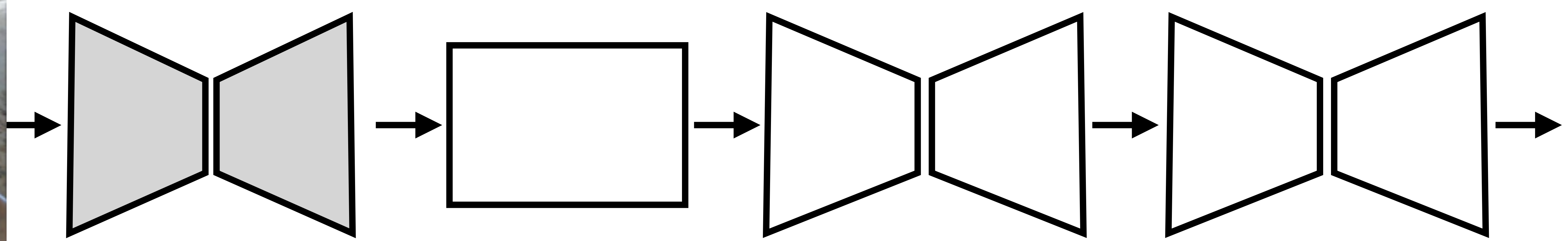
Input:
Image+selected pixel

Output:
HDR Illumination map

Each sub-module is able to focus on a relatively easier task
and can be trained with direct supervision.

Neural Illumination

Input:
Image+selected pixel

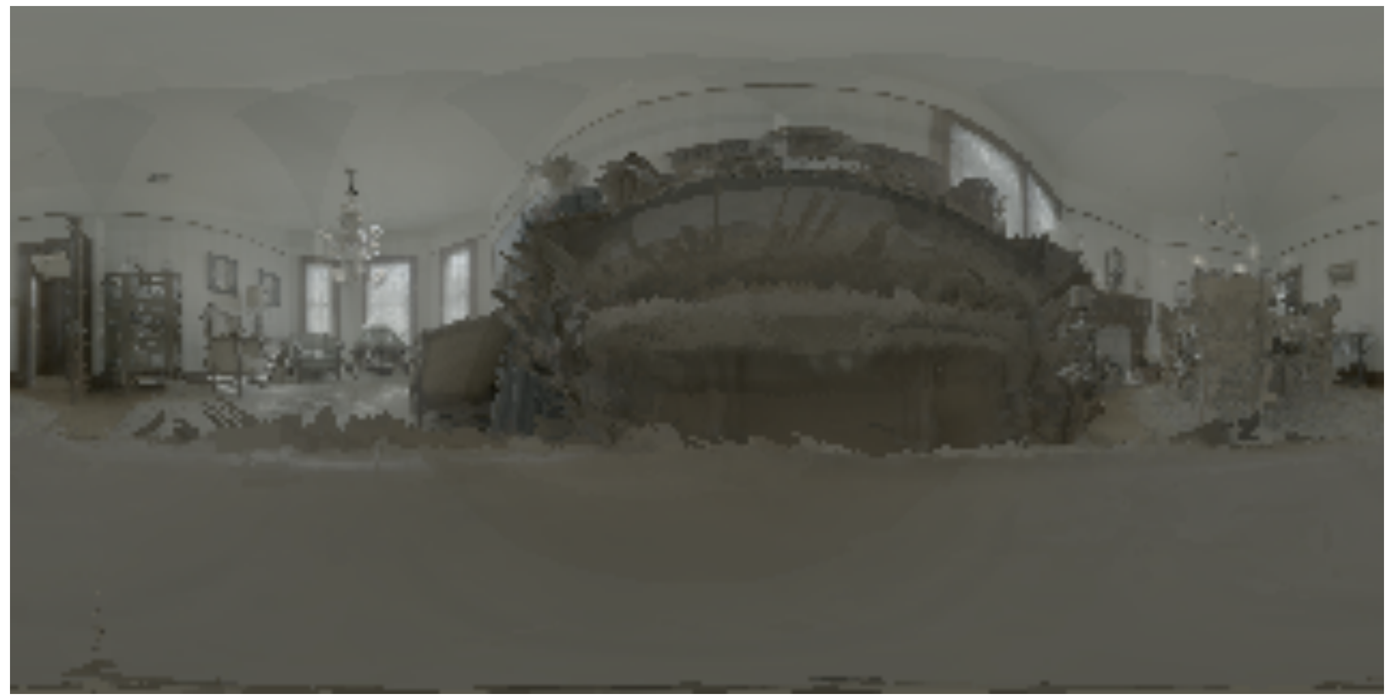


Geometry estimation

Differentiable warping

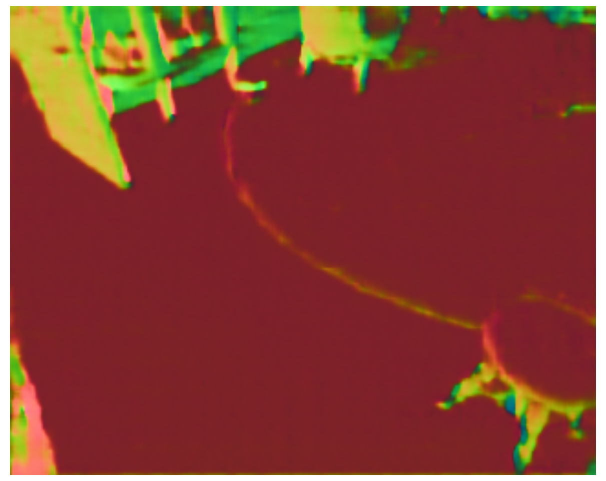
LDR completion network

LDR to HDR network



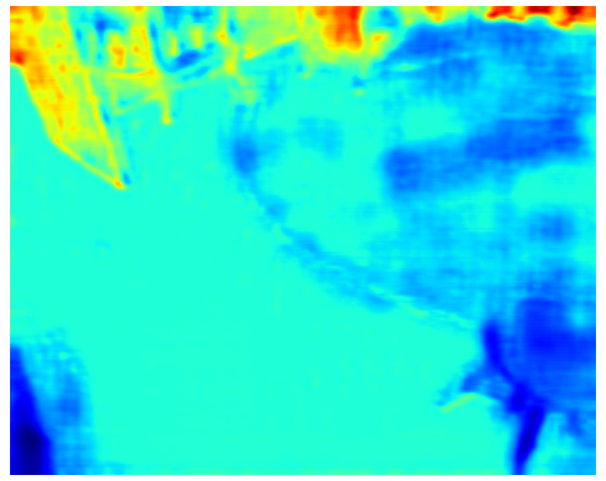
Output:
HDR Illumination map

Surface normal



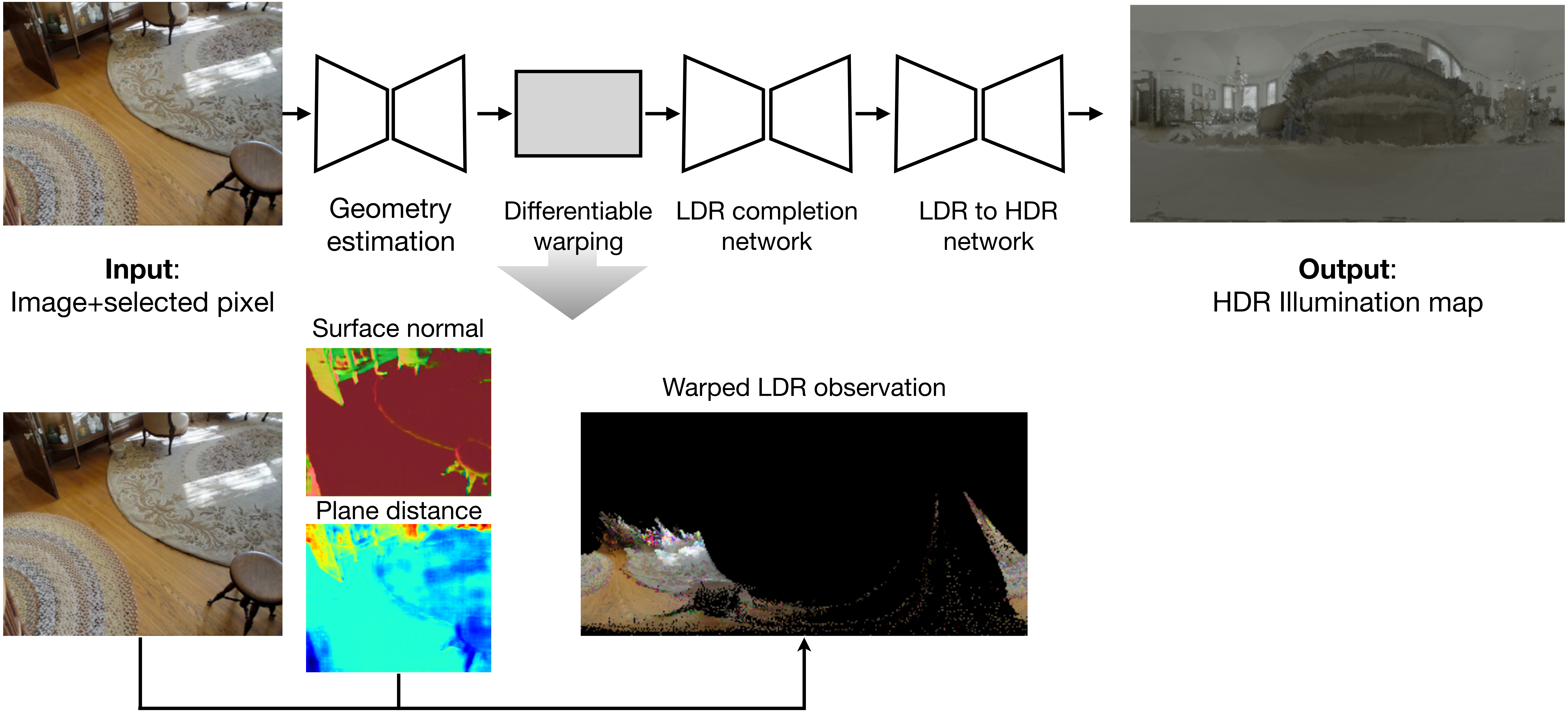
→ Cosine loss

Plane distance



→ L1 loss

Neural Illumination



Input:
Image+selected pixel

Geometry estimation

Differentiable warping

LDR completion network

LDR to HDR network

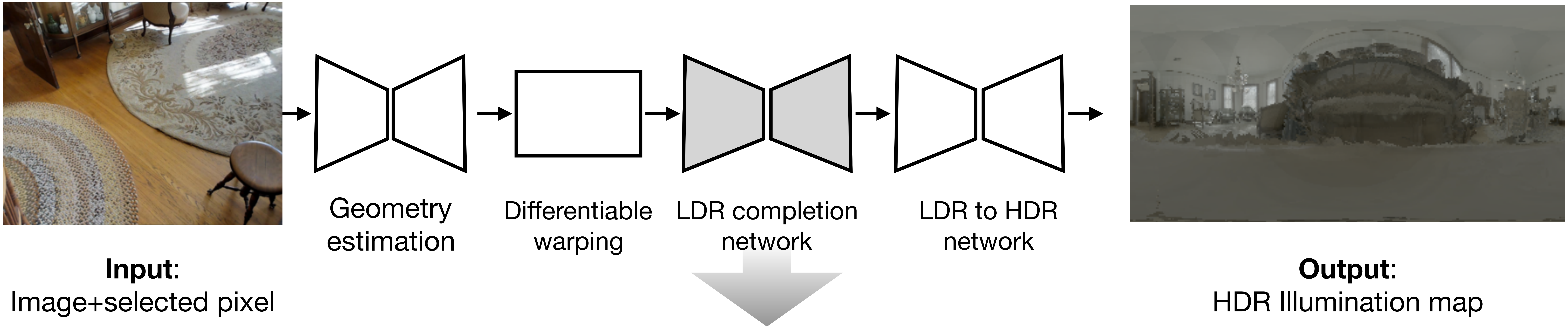
Output:
HDR Illumination map

Surface normal

Plane distance

Warped LDR observation

Neural Illumination



Geometry estimation

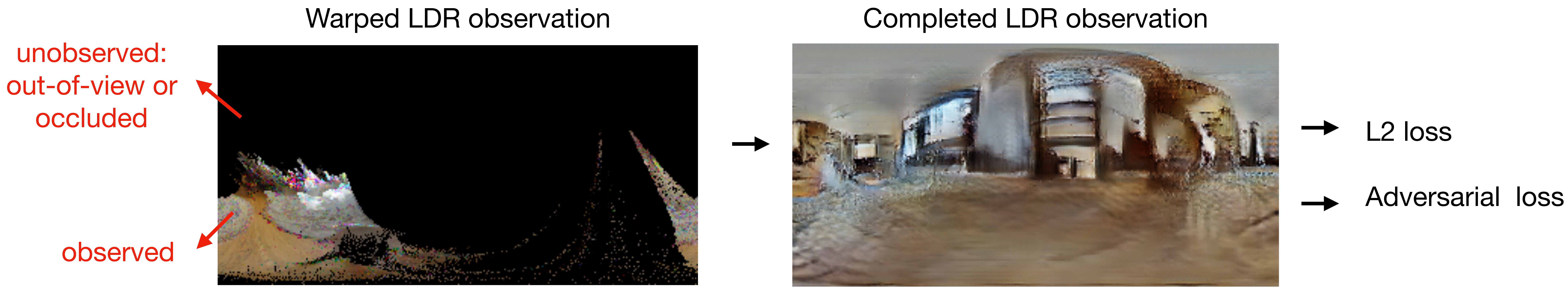
Differentiable warping

LDR completion network

LDR to HDR network

Input:
Image+selected pixel

Output:
HDR Illumination map



Warped LDR observation

Completed LDR observation

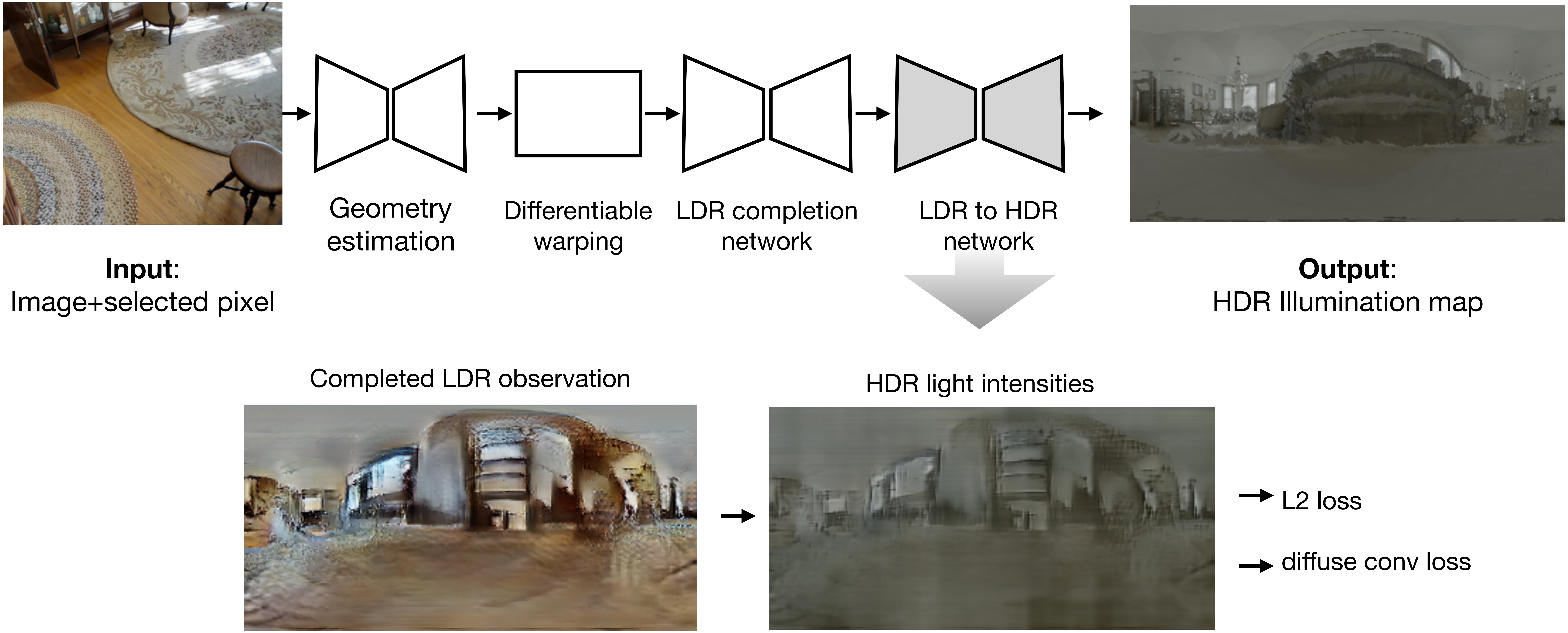
unobserved:
out-of-view or
occluded

observed

L2 loss

Adversarial loss

Neural Illumination



Geometry estimation

Differentiable warping

LDR completion network

LDR to HDR network

Input:
Image+selected pixel

Output:
HDR Illumination map

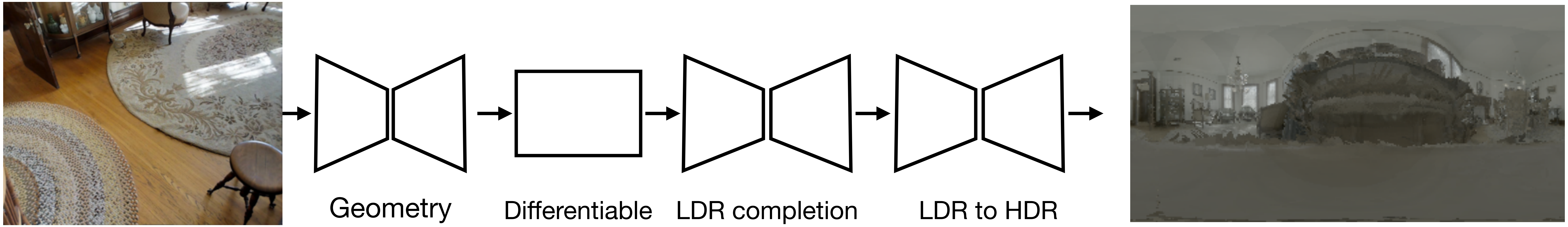
Completed LDR observation

HDR light intensities

L2 loss

diffuse conv loss

Neural Illumination



Geometry estimation

Differentiable warping

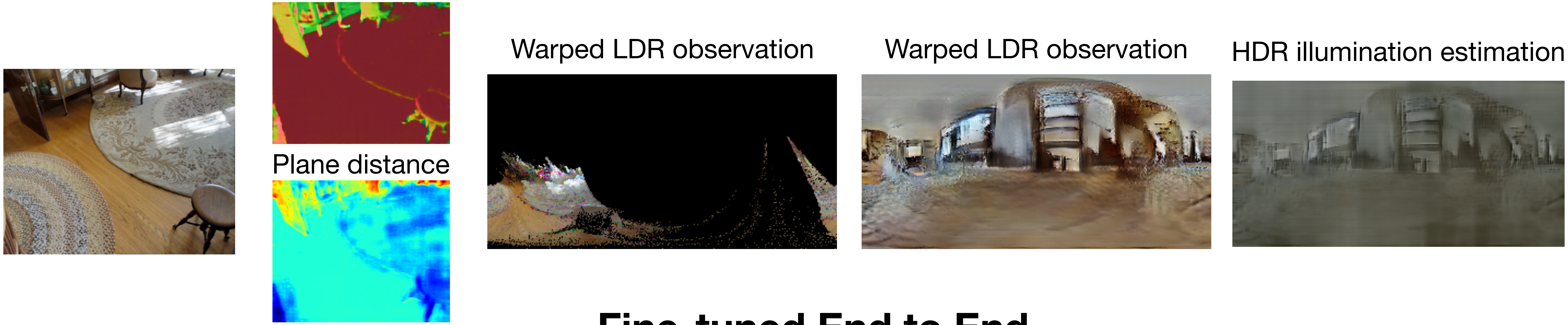
LDR completion network

LDR to HDR network

Input:
Image+selected pixel

Output:
HDR Illumination map

Surface normal



Plane distance

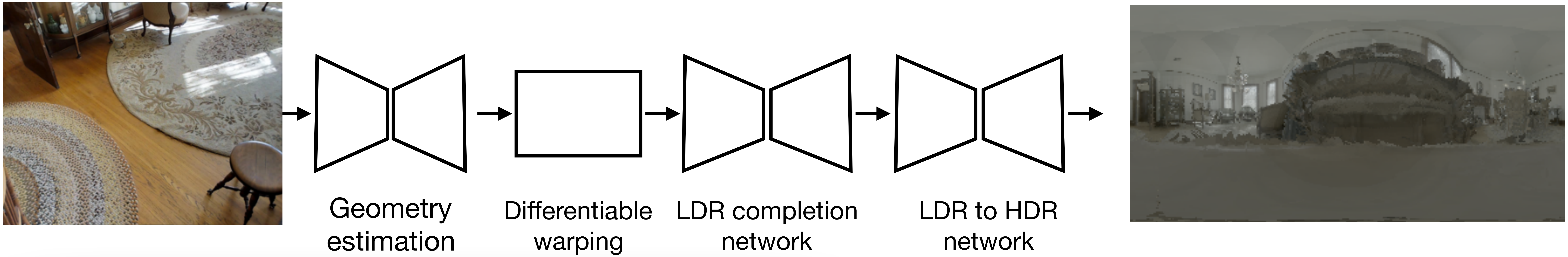
Warped LDR observation

Warped LDR observation

HDR illumination estimation

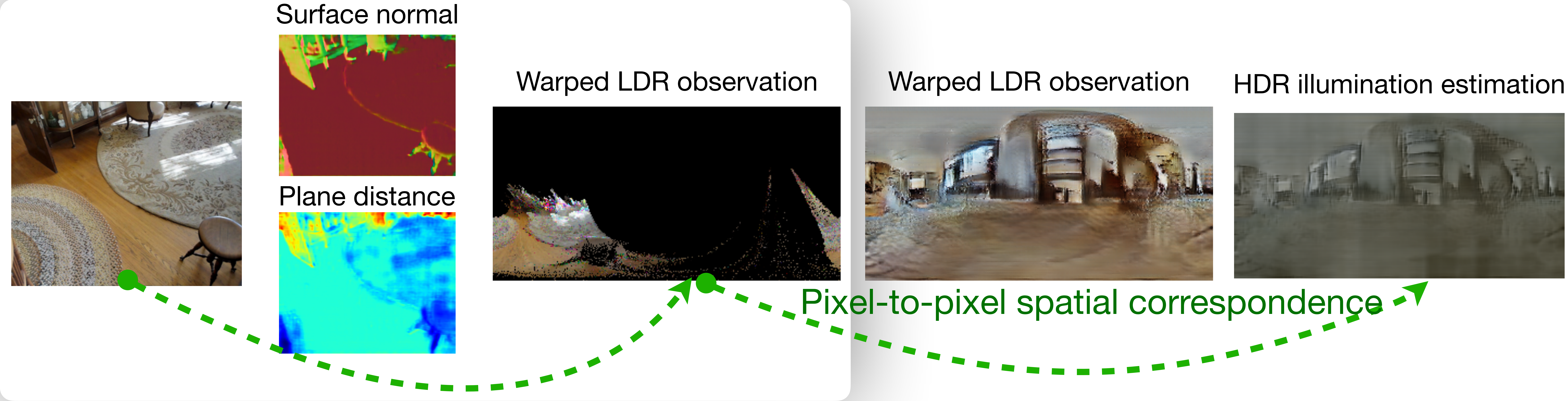
Fine-tuned End to End

Neural Illumination

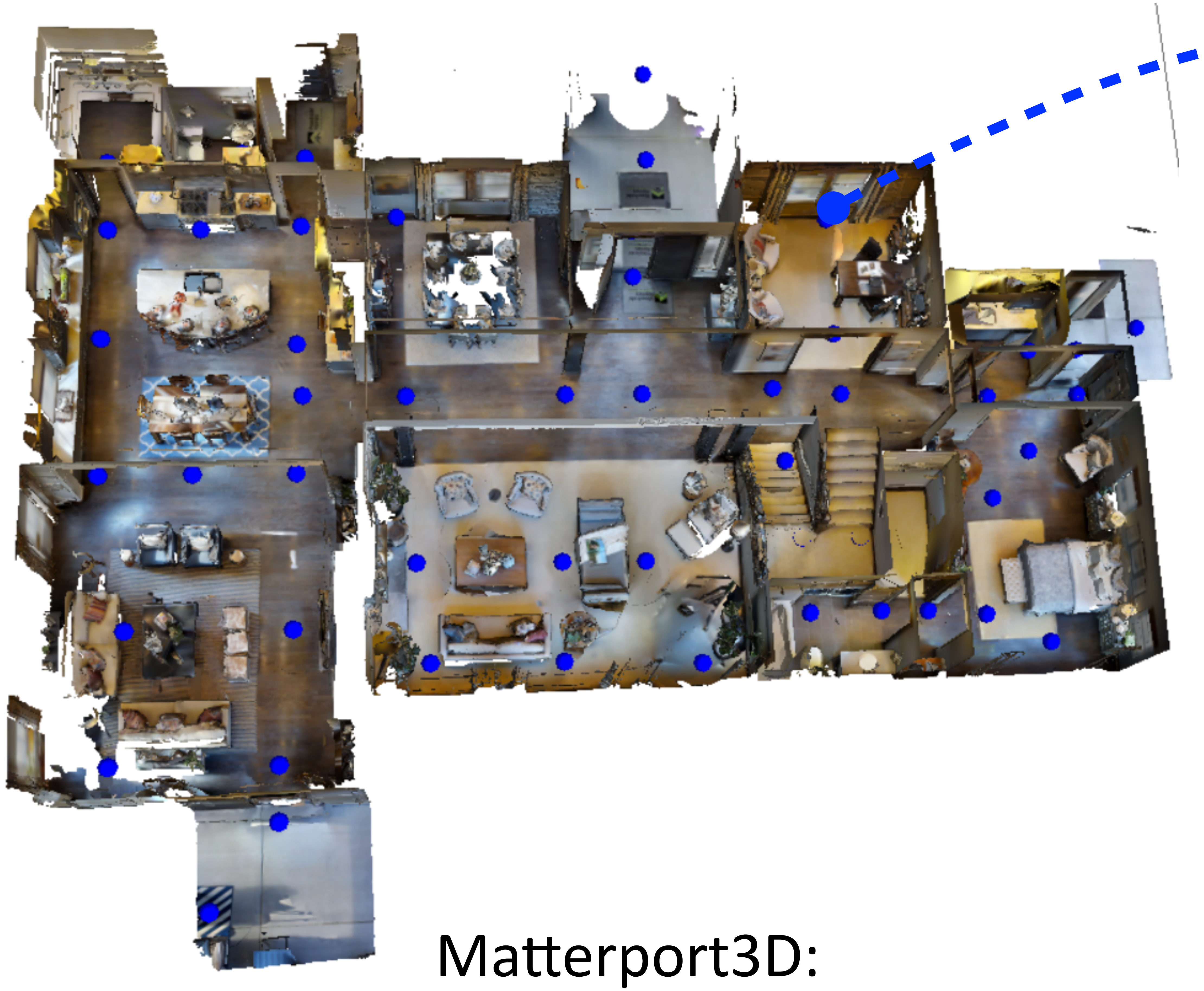


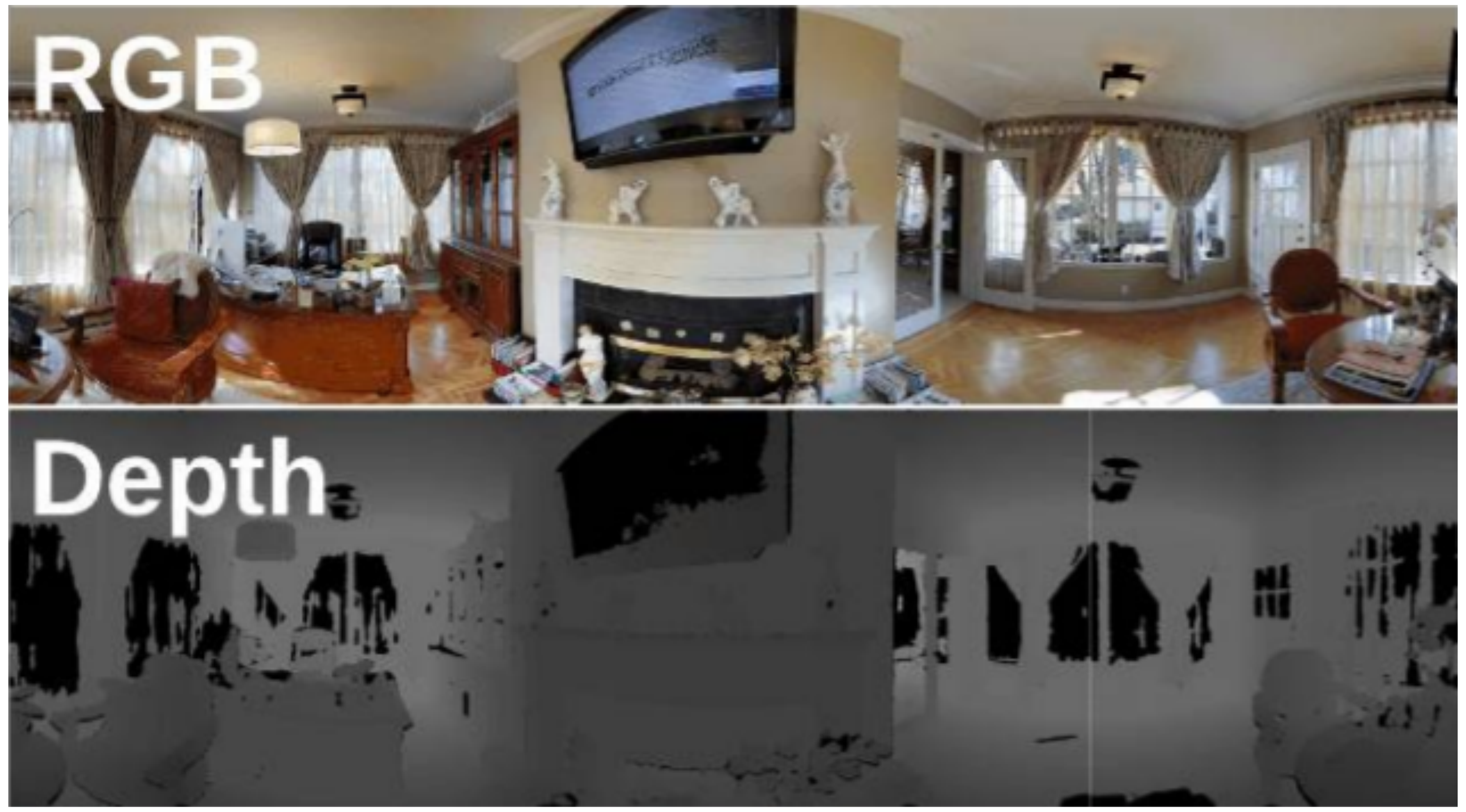
Input:
Image+selected pixel

Output:
HDR Illumination map



Training Data Generation

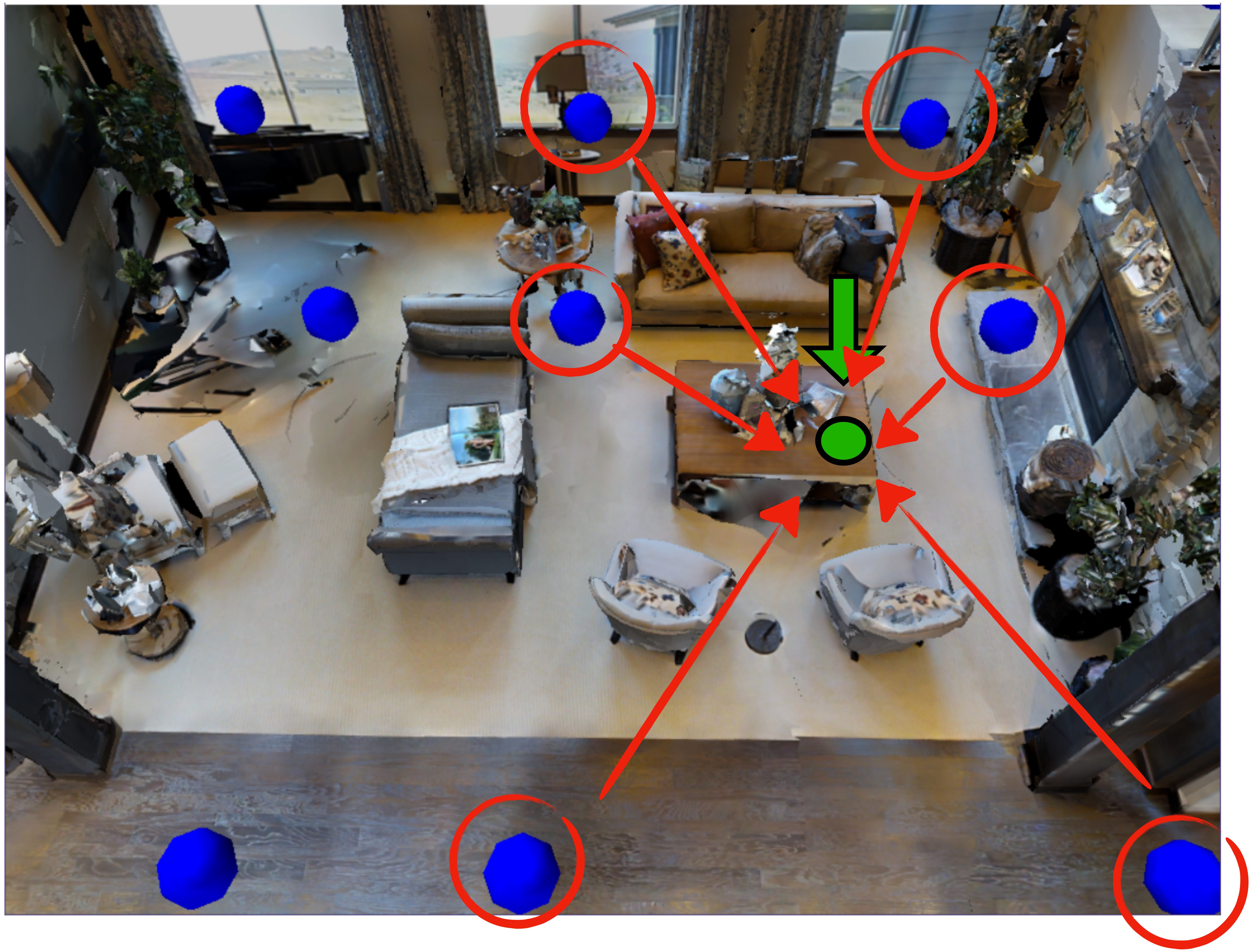


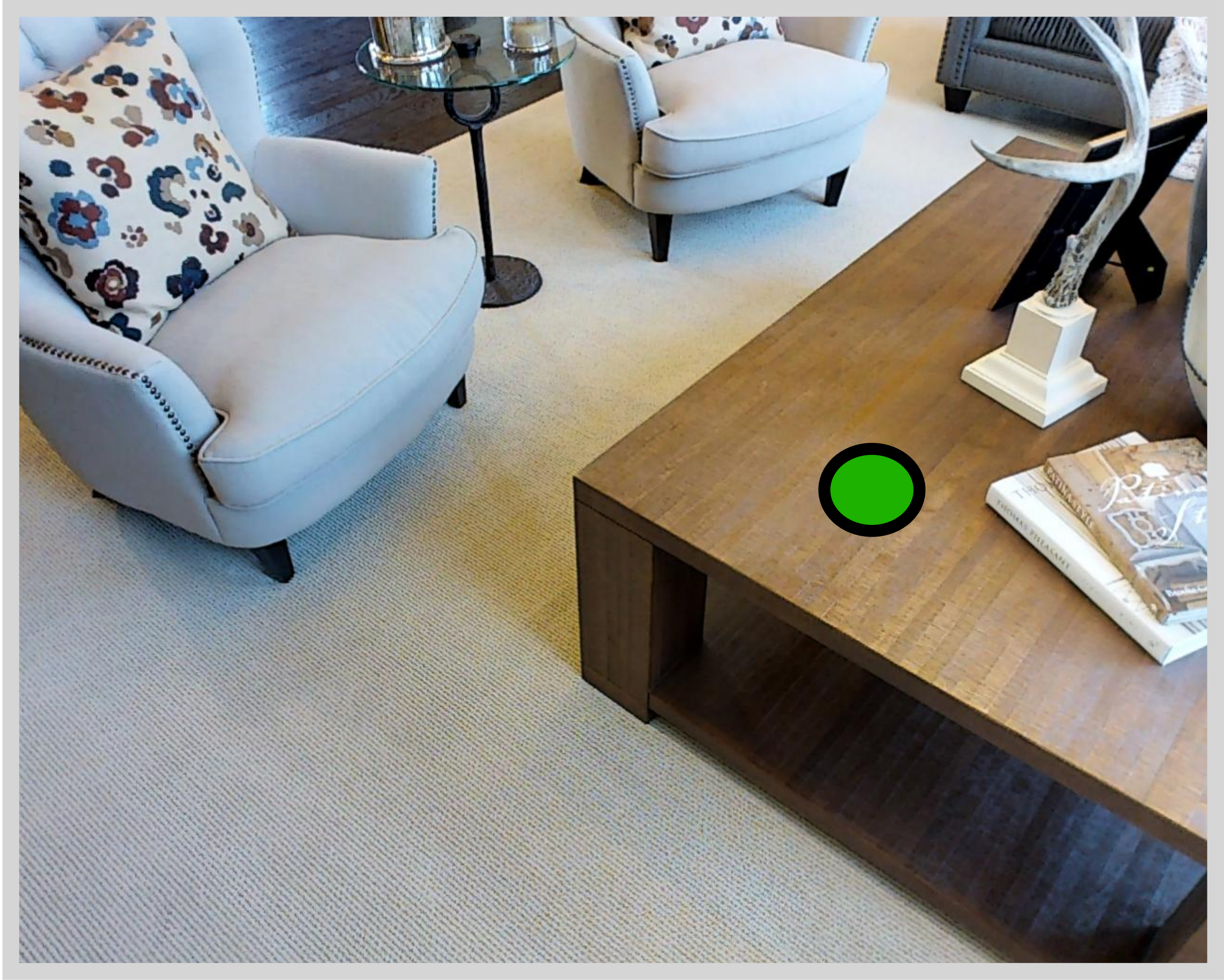


Matterport3D:

HDR RGB-D panoramas throughout 90 houses

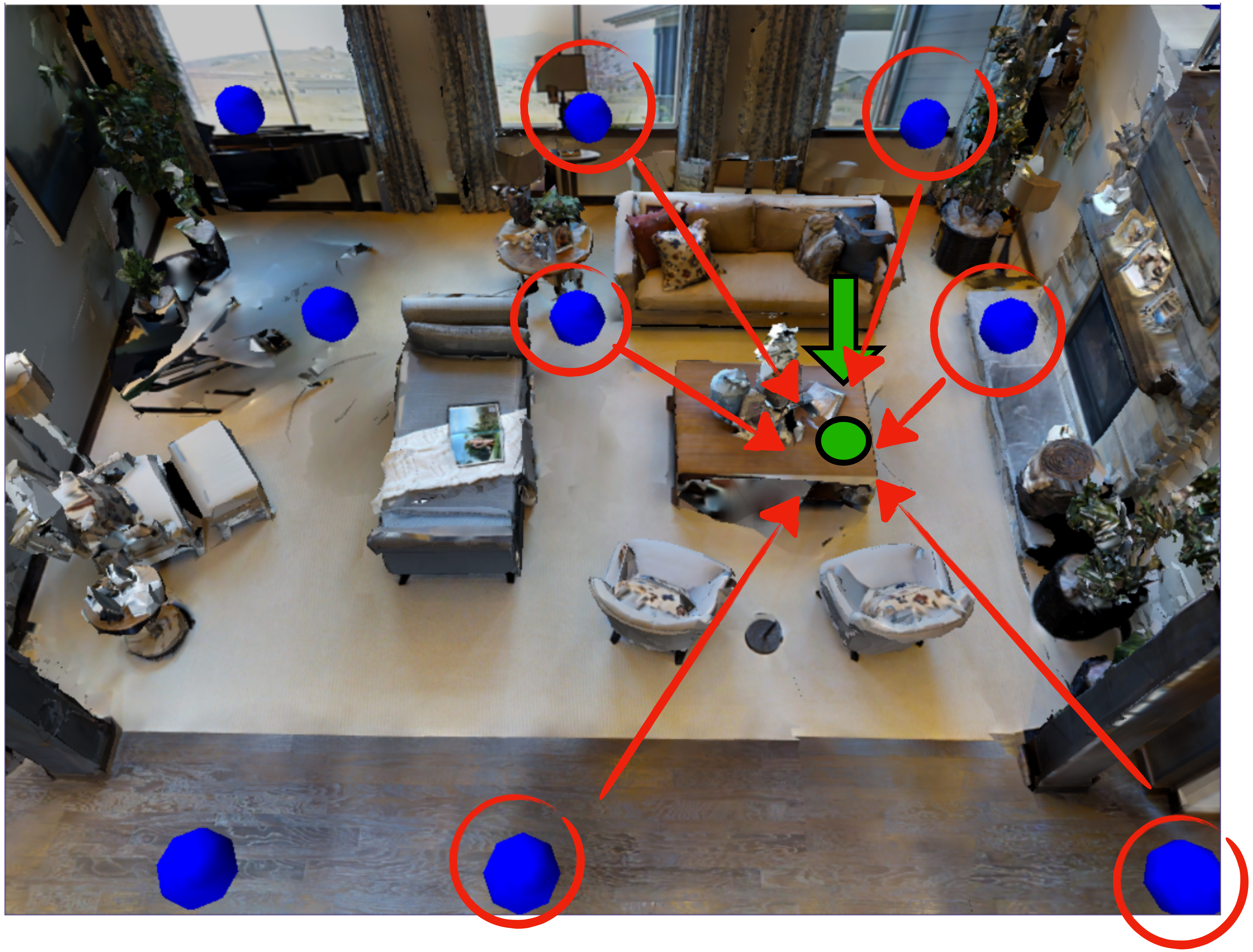
Training Data Generation

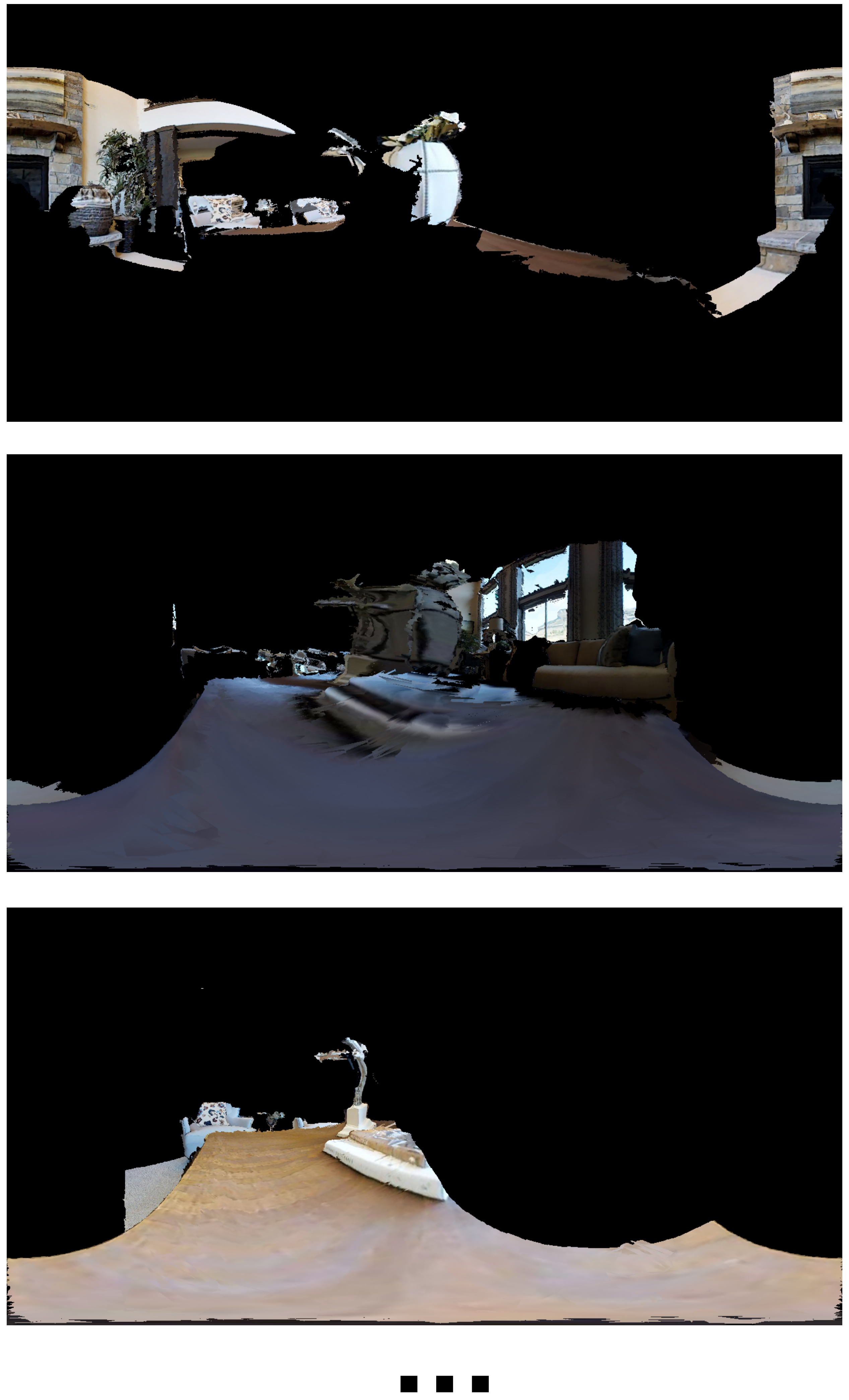




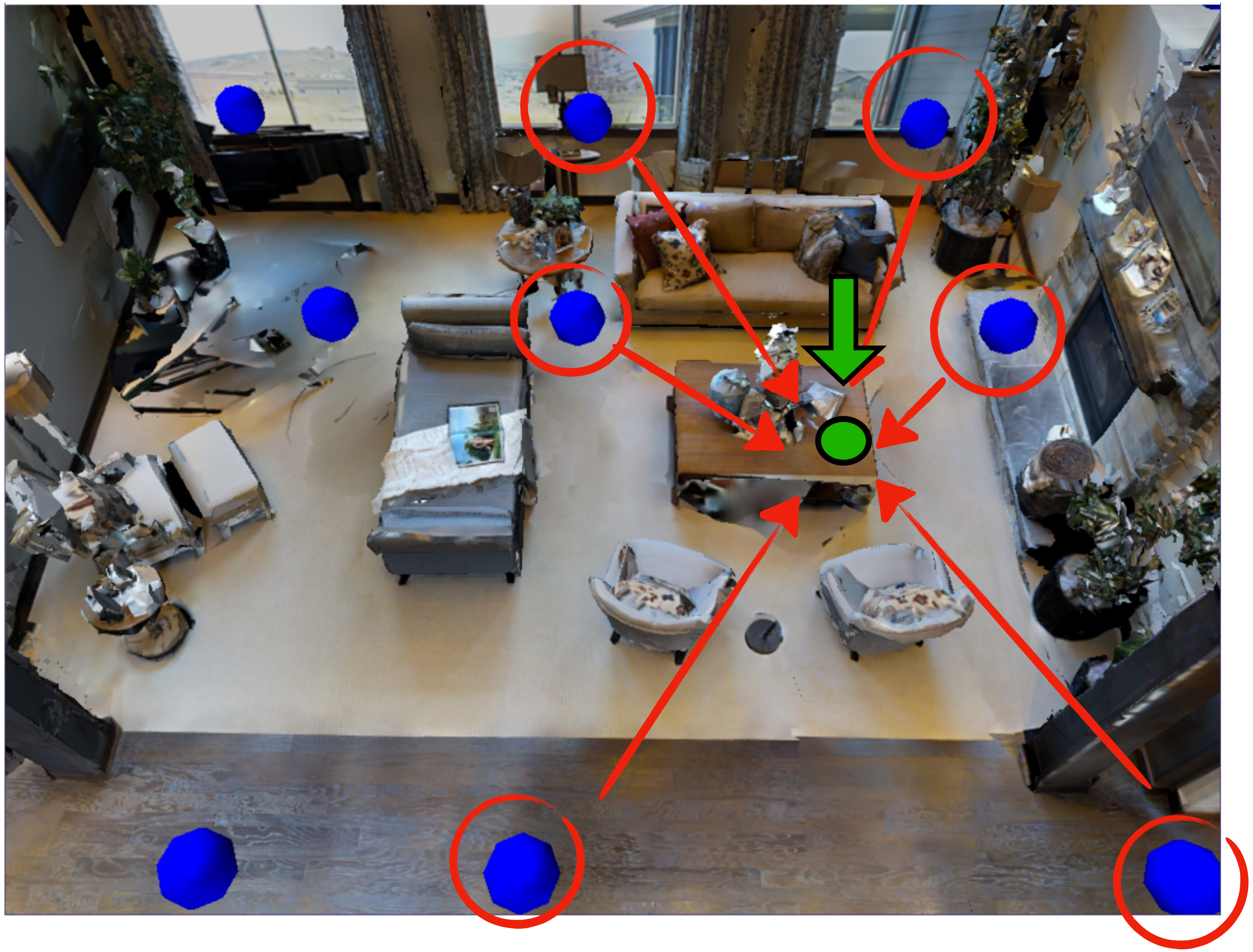
Input:
Image+selected pixel

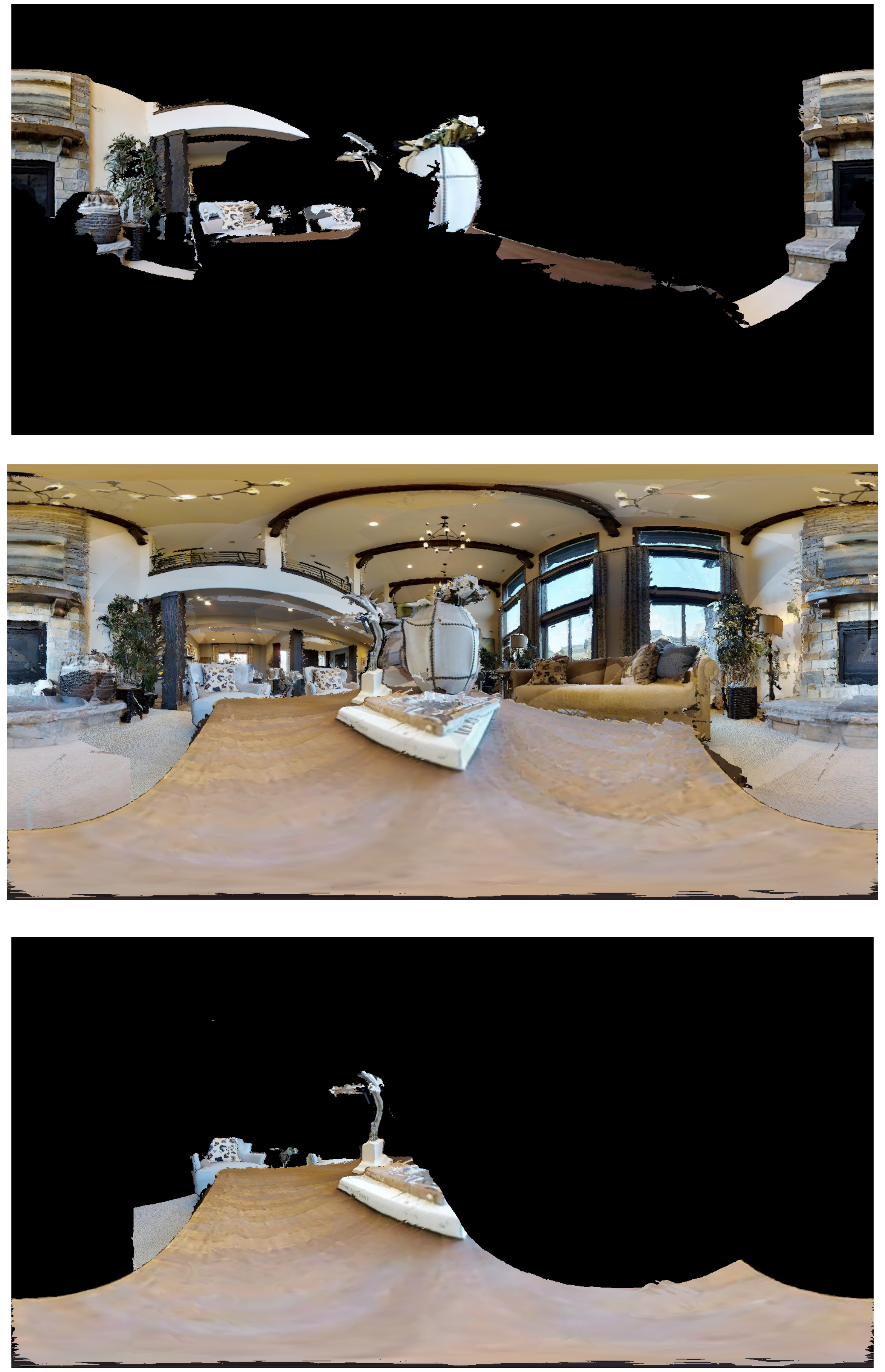
Training Data Generation





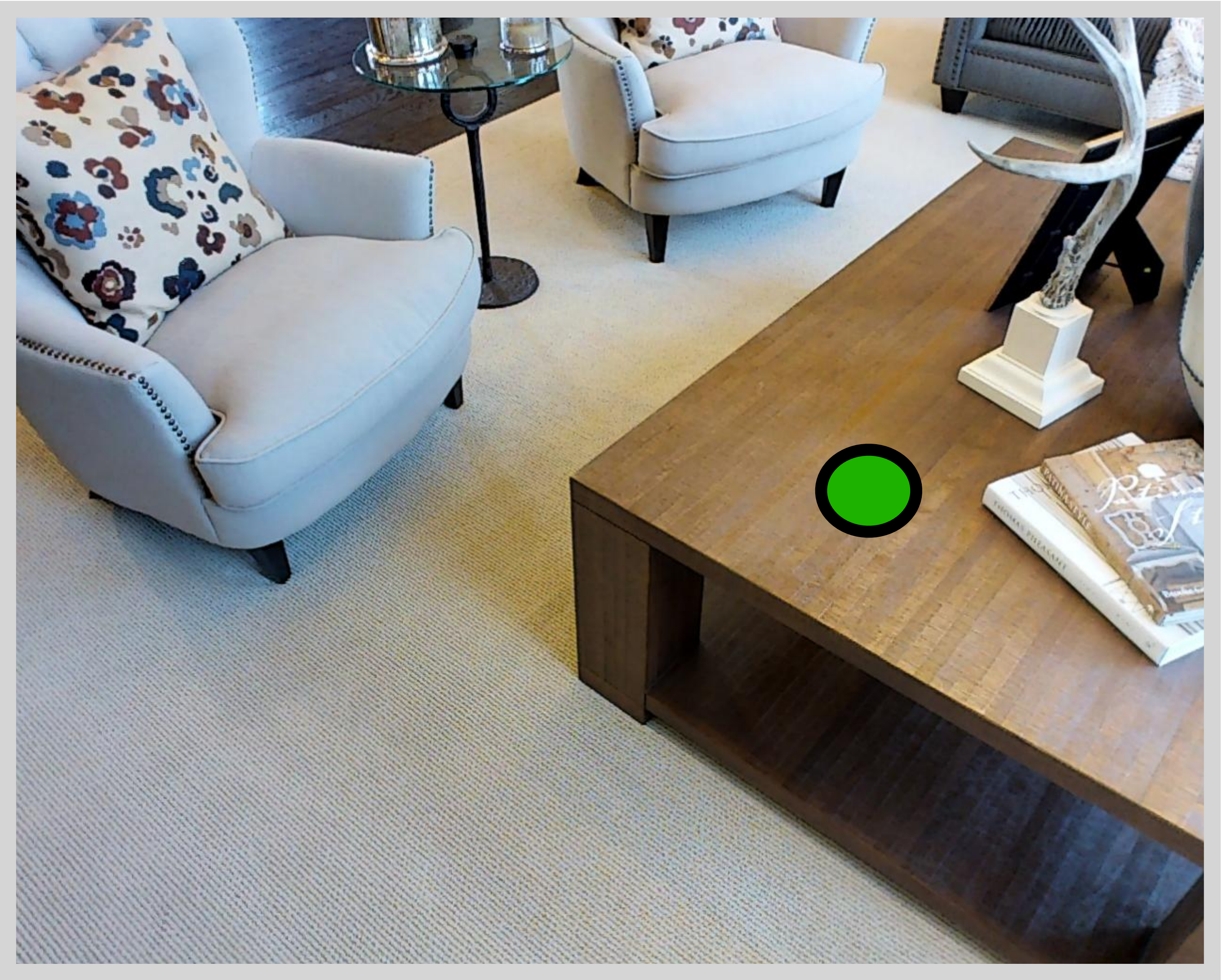
Training Data Generation





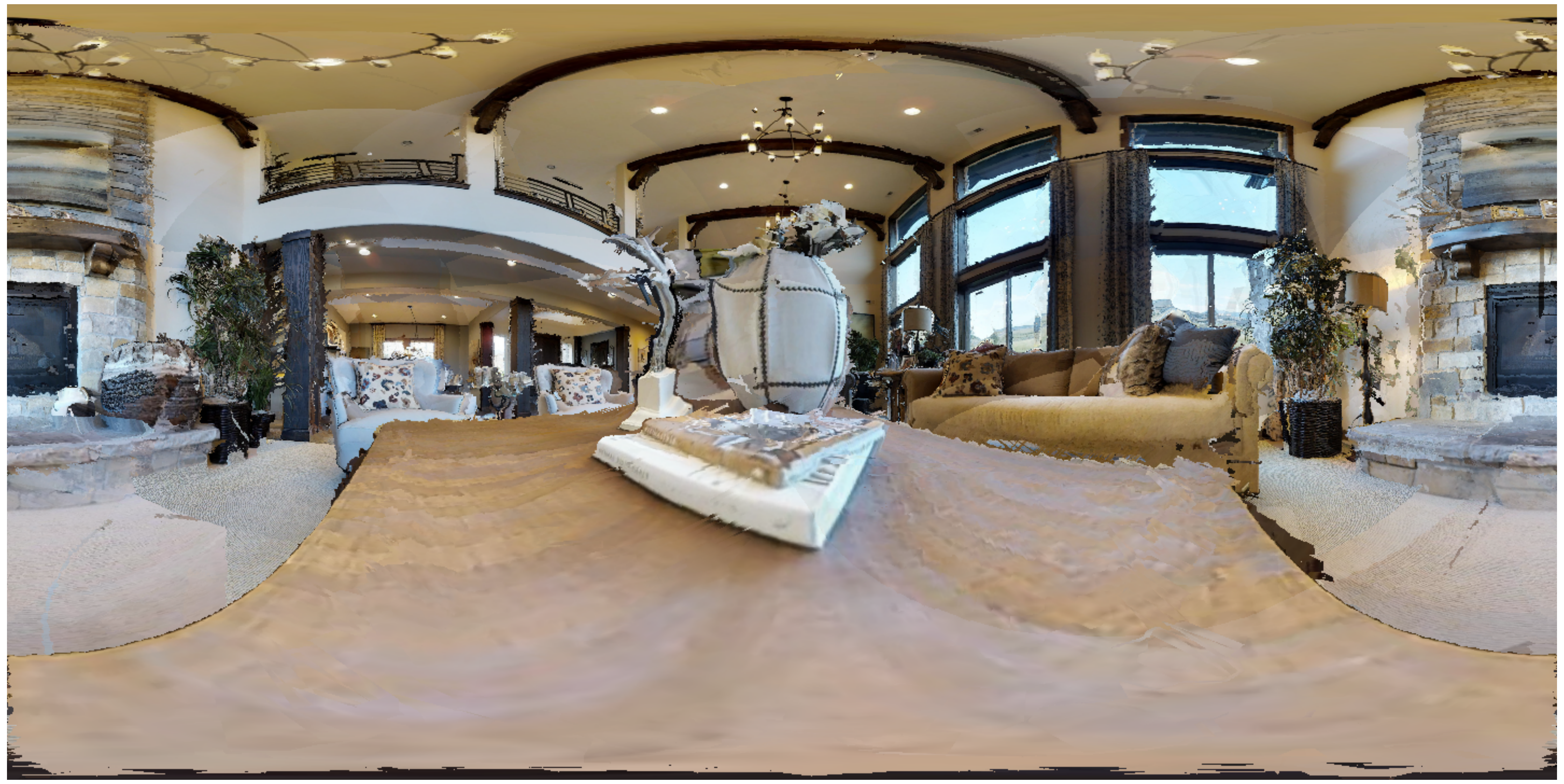
Training Data Generation

{



LDR images+target pixel

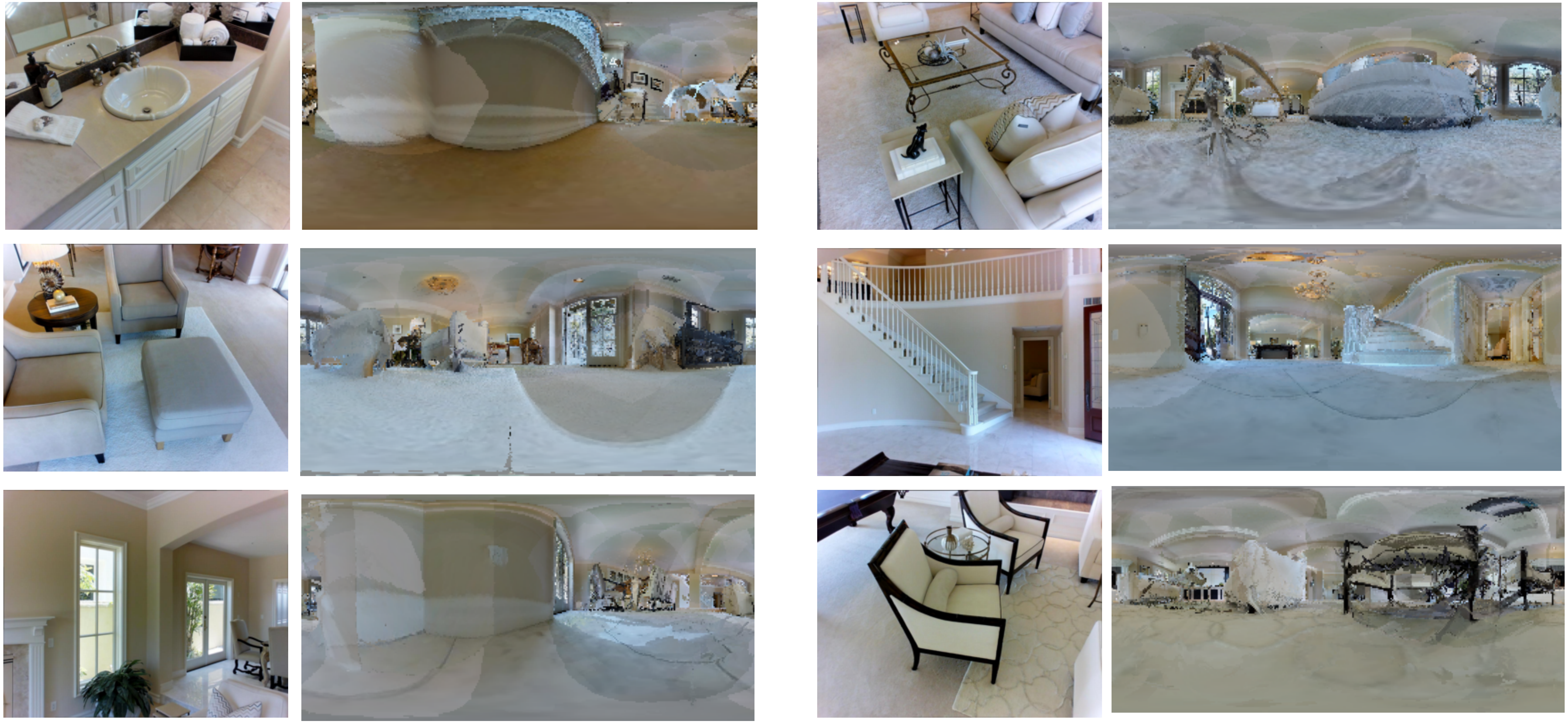
,



Output HDR illumination map

}

Training Data Generation



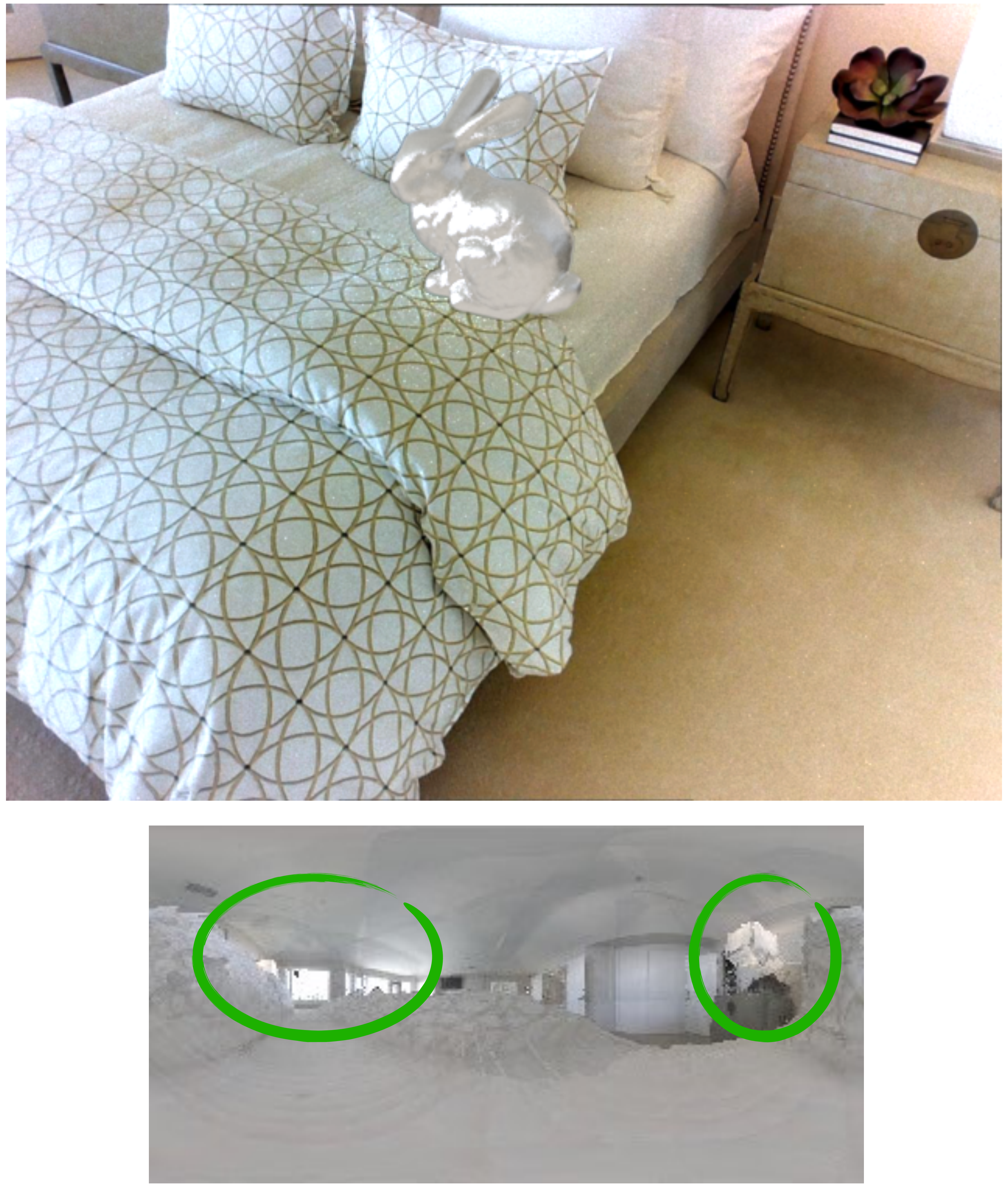
Overall, we generate >90K locales and >360K illumination pairs

Virtual object re-lighting

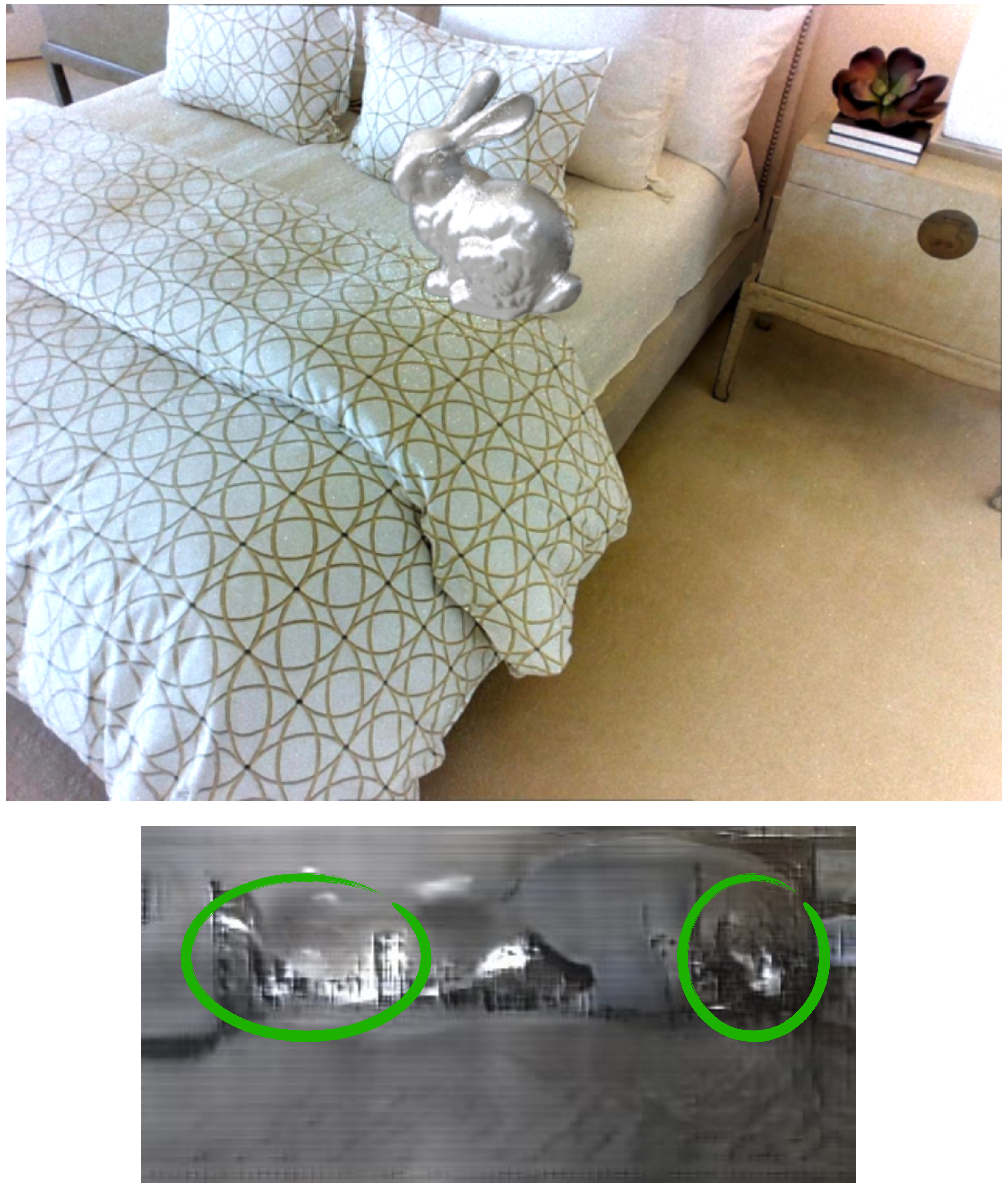
Input:

LDR images+selected pixel

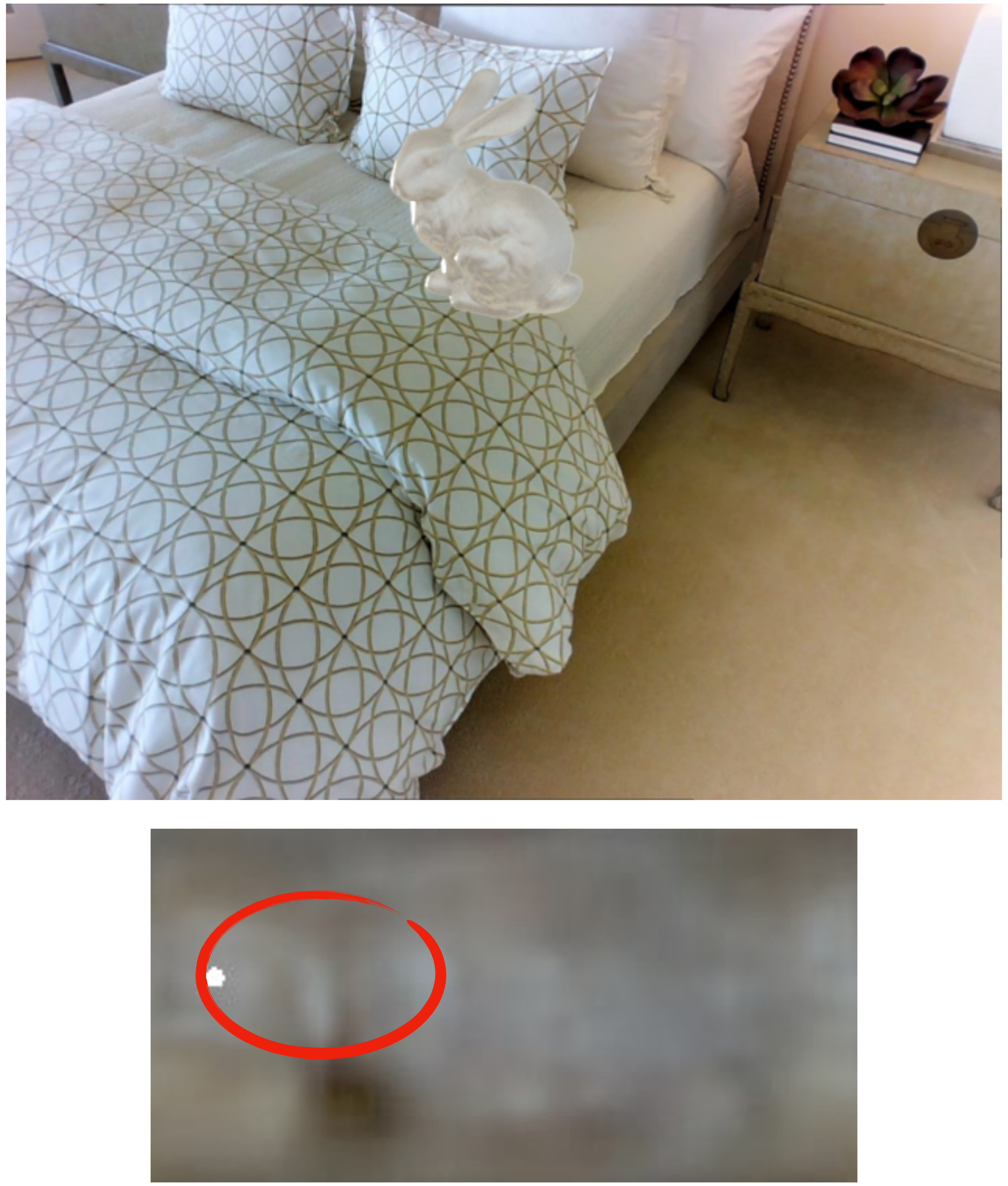
Virtual object re-lighting



Groundtruth

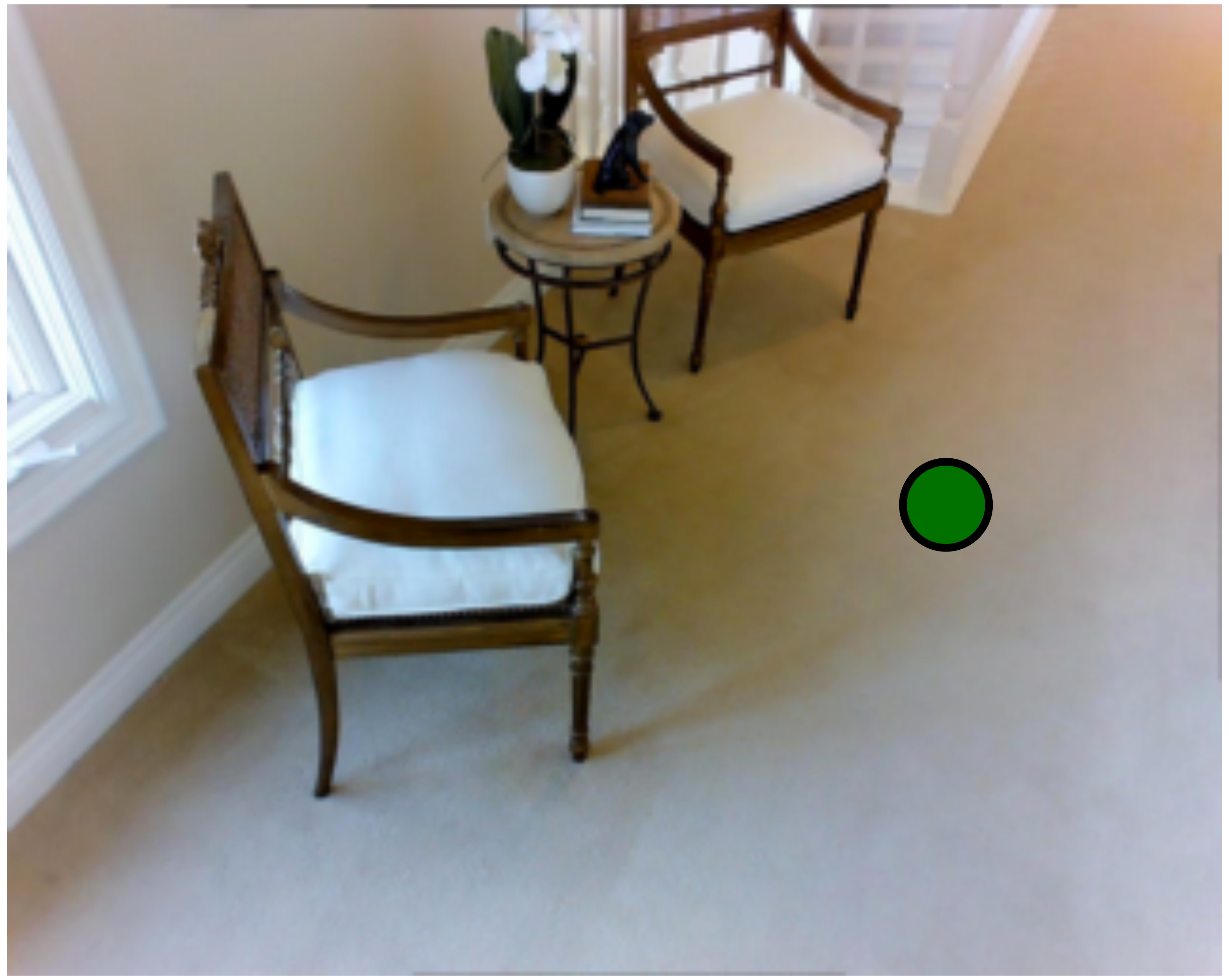


Ours



Gardner *et al.*

Virtual object re-lighting

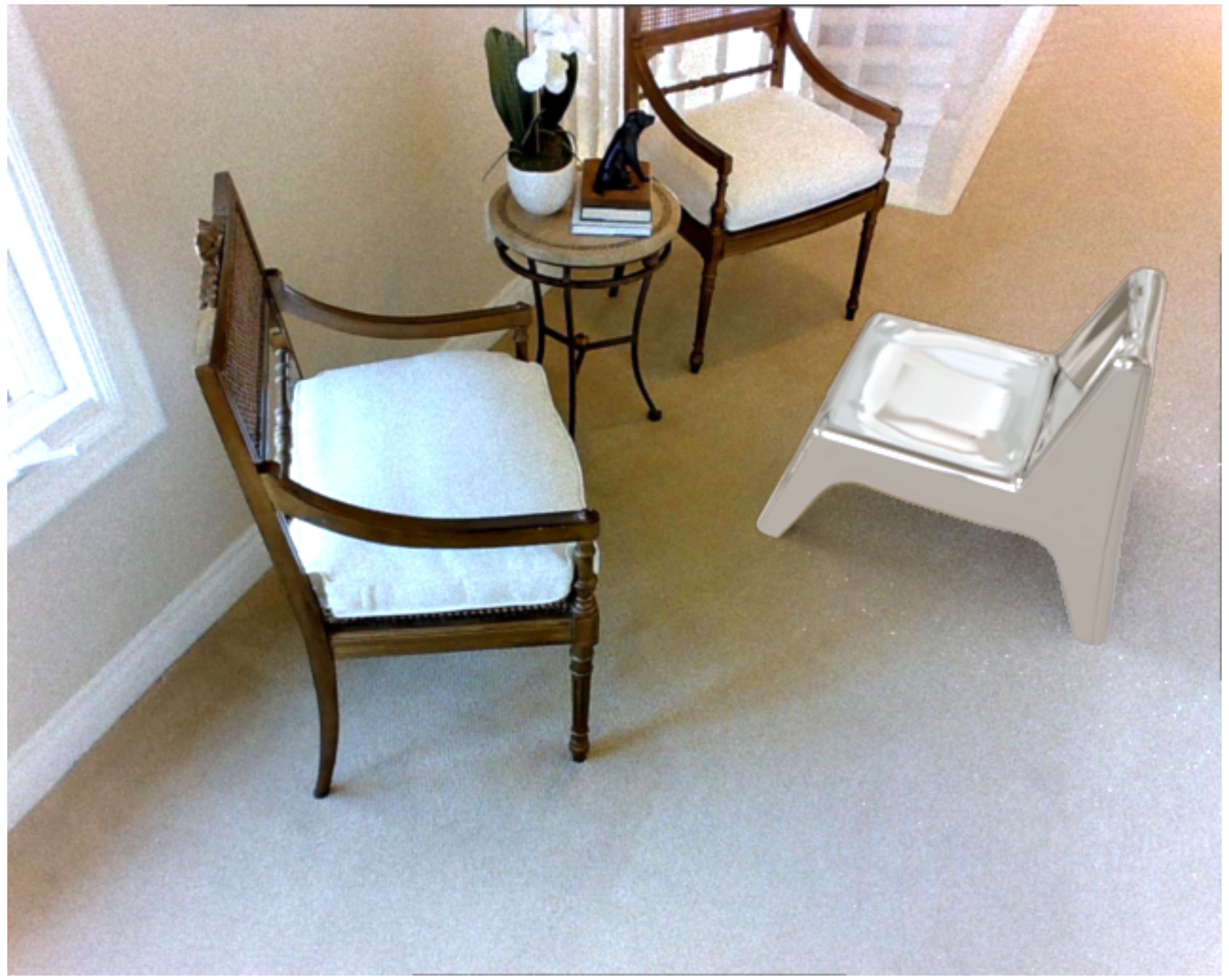


Input:

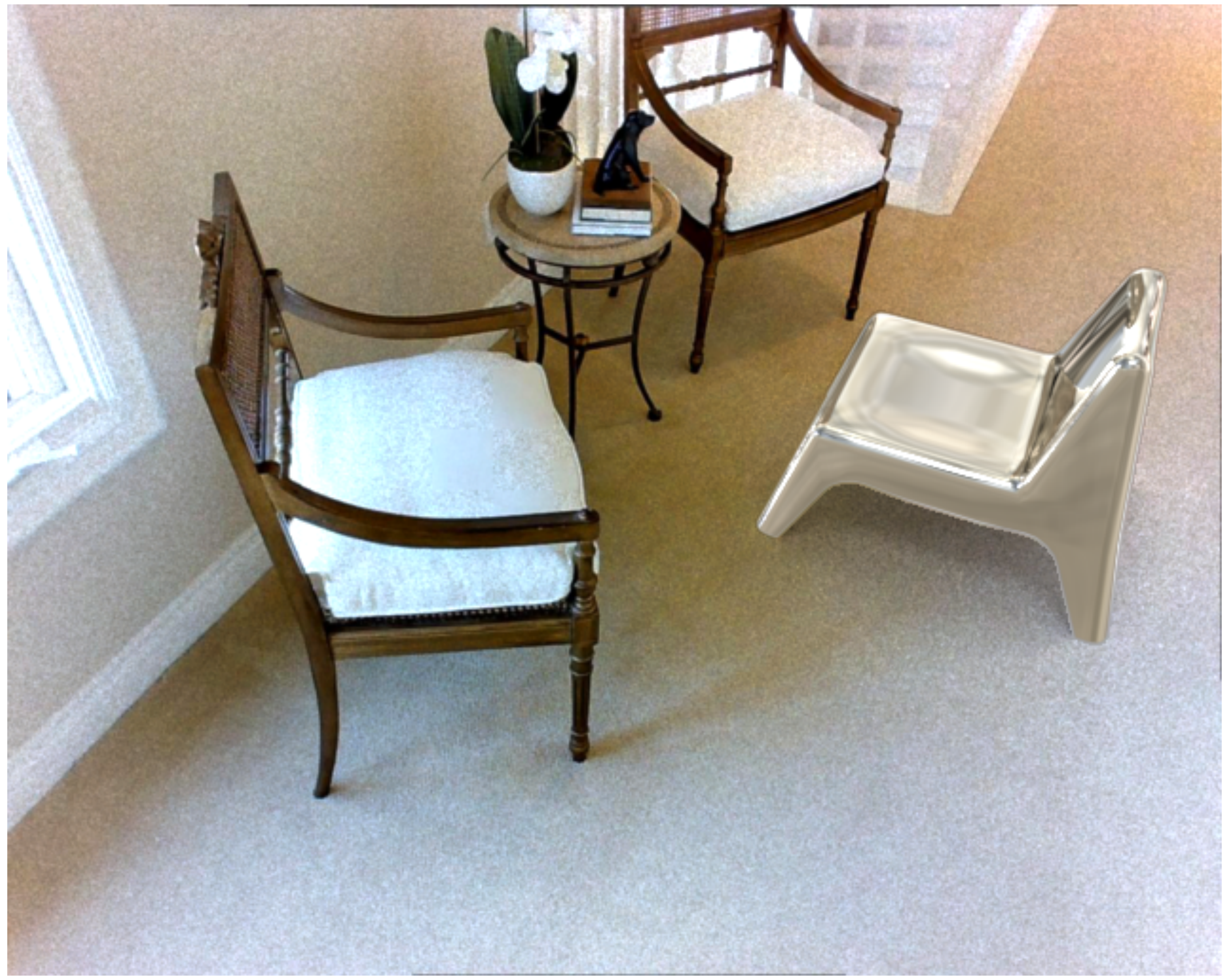
LDR images+selected pixel

Virtual object re-lighting

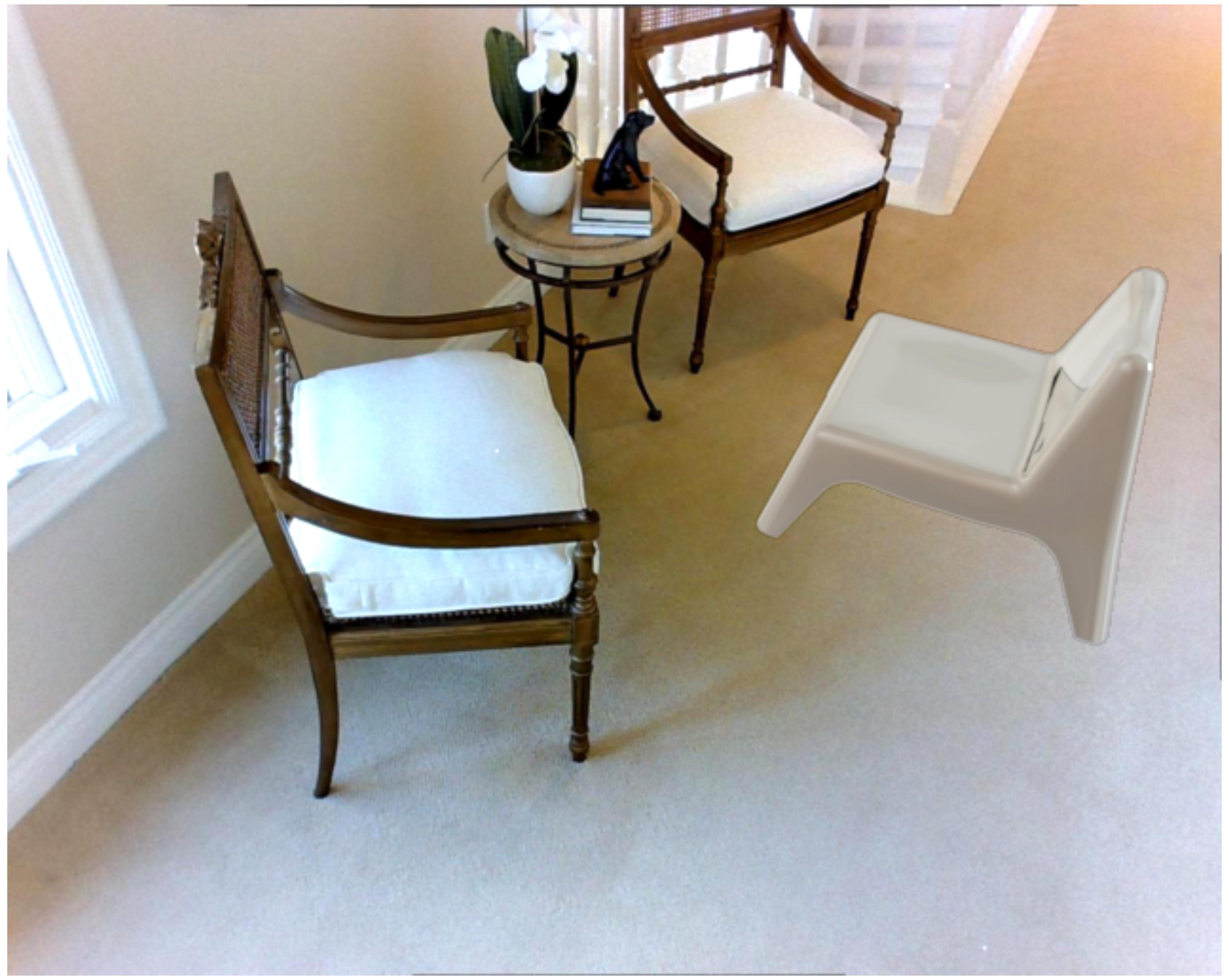
Ground truth

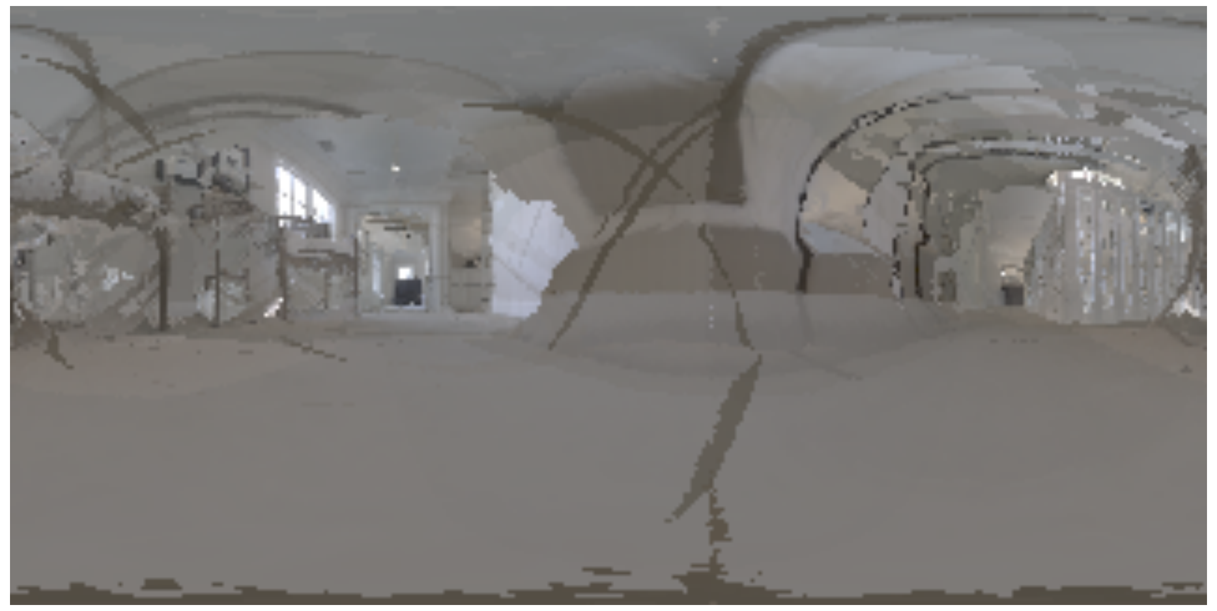


Ours

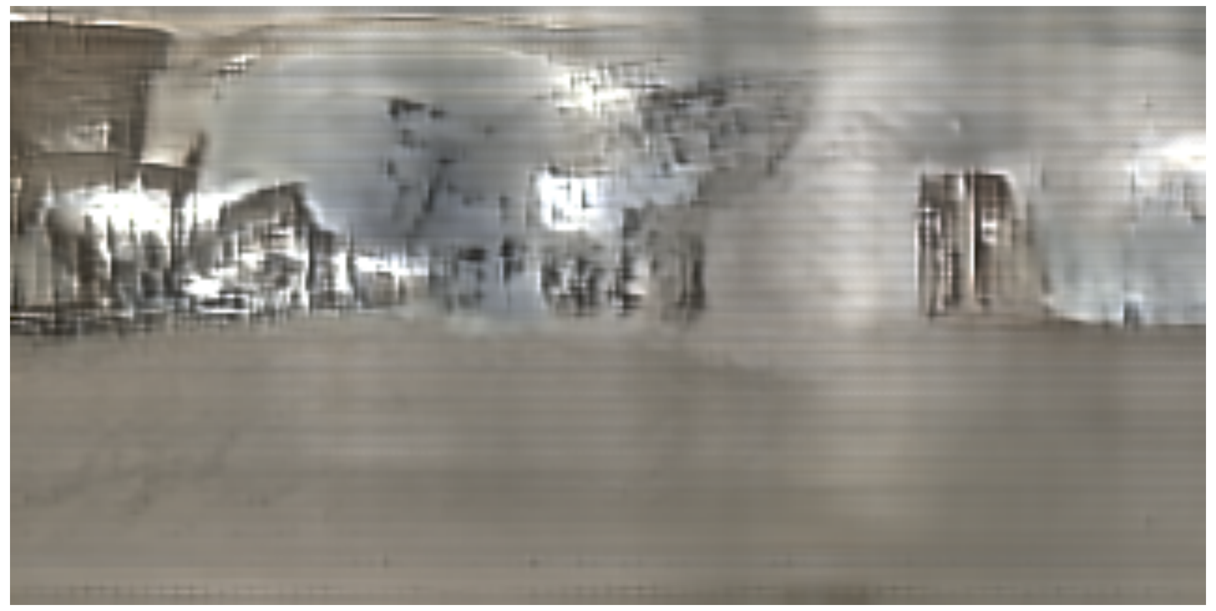


Gardner et al.

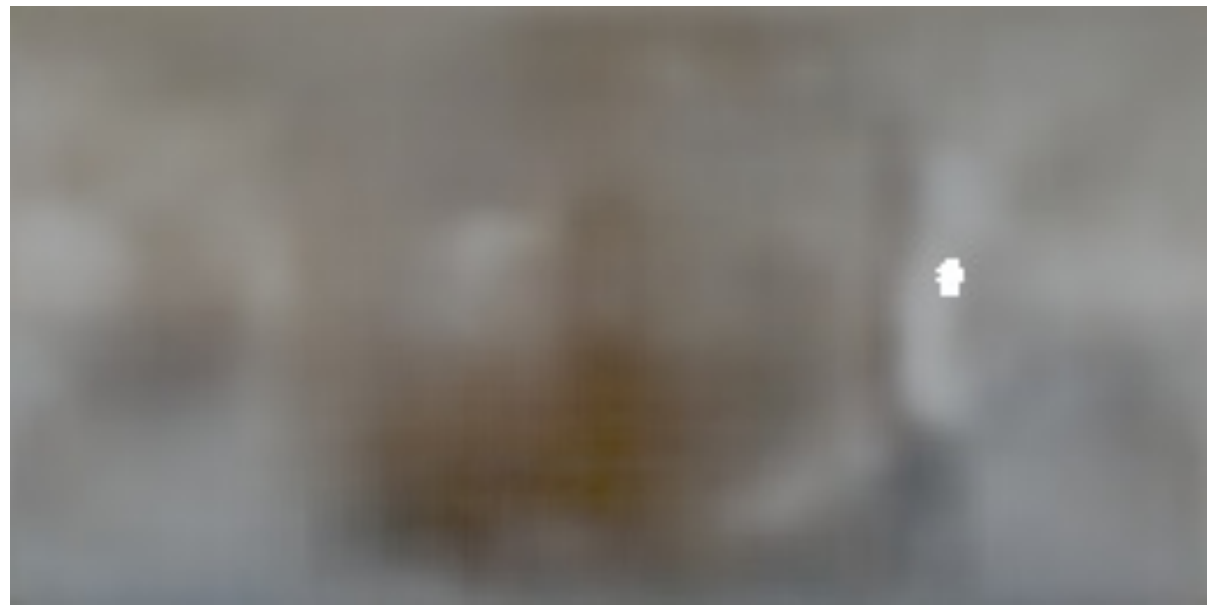




Groundtruth

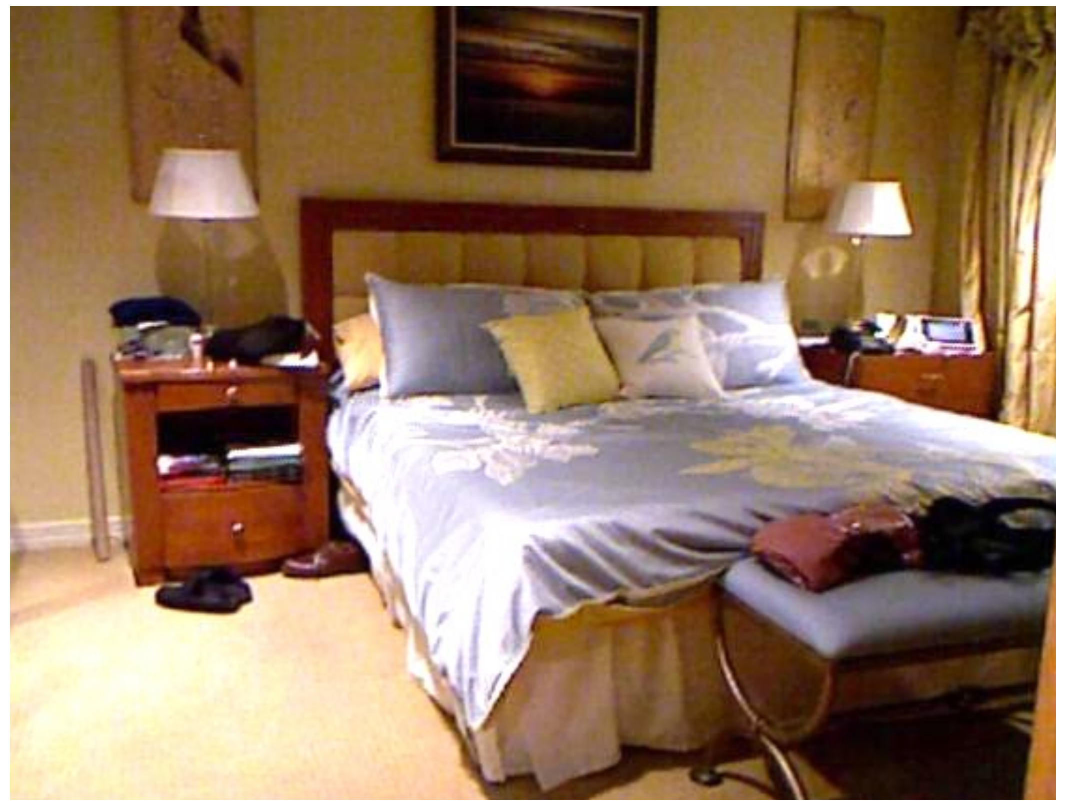


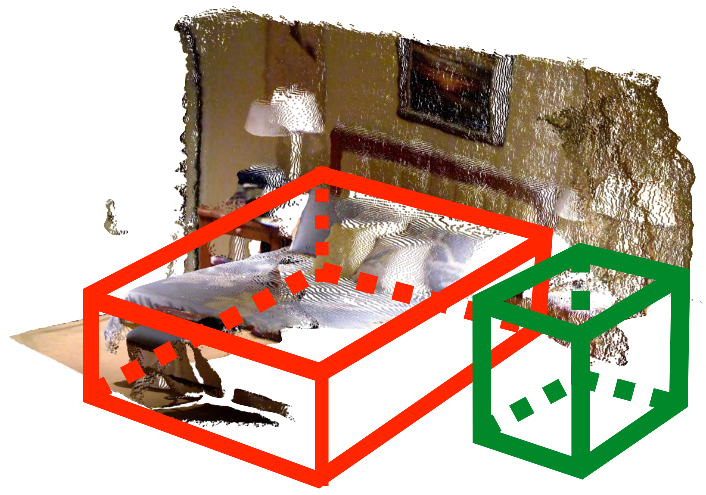
Ours



Gardner *et al.*

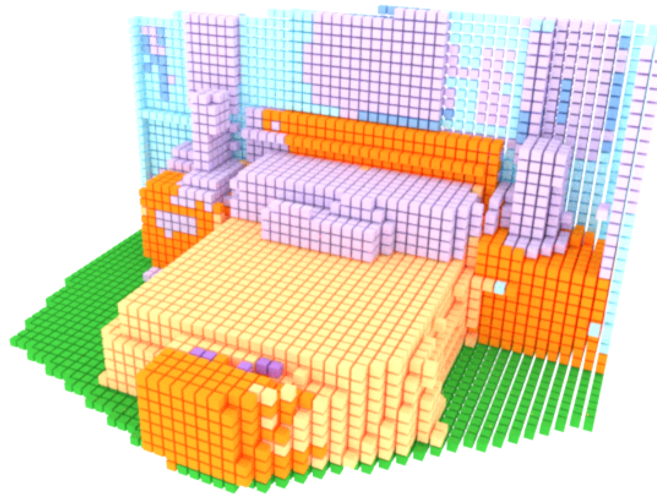
Advances Towards 3D Scene Understanding





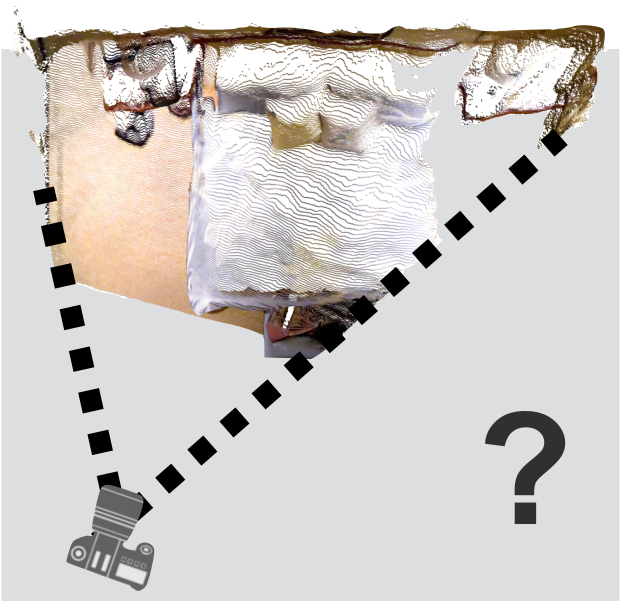
Amodal 3D Bounding Boxes

[Song and Xiao ECCV’14,CVPR’16]



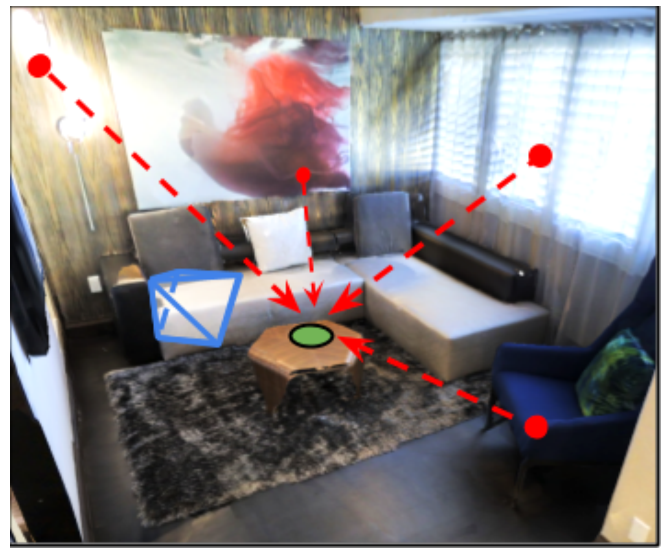
Higher Fidelity 3D Voxels

[Song et al. CVPR’17]



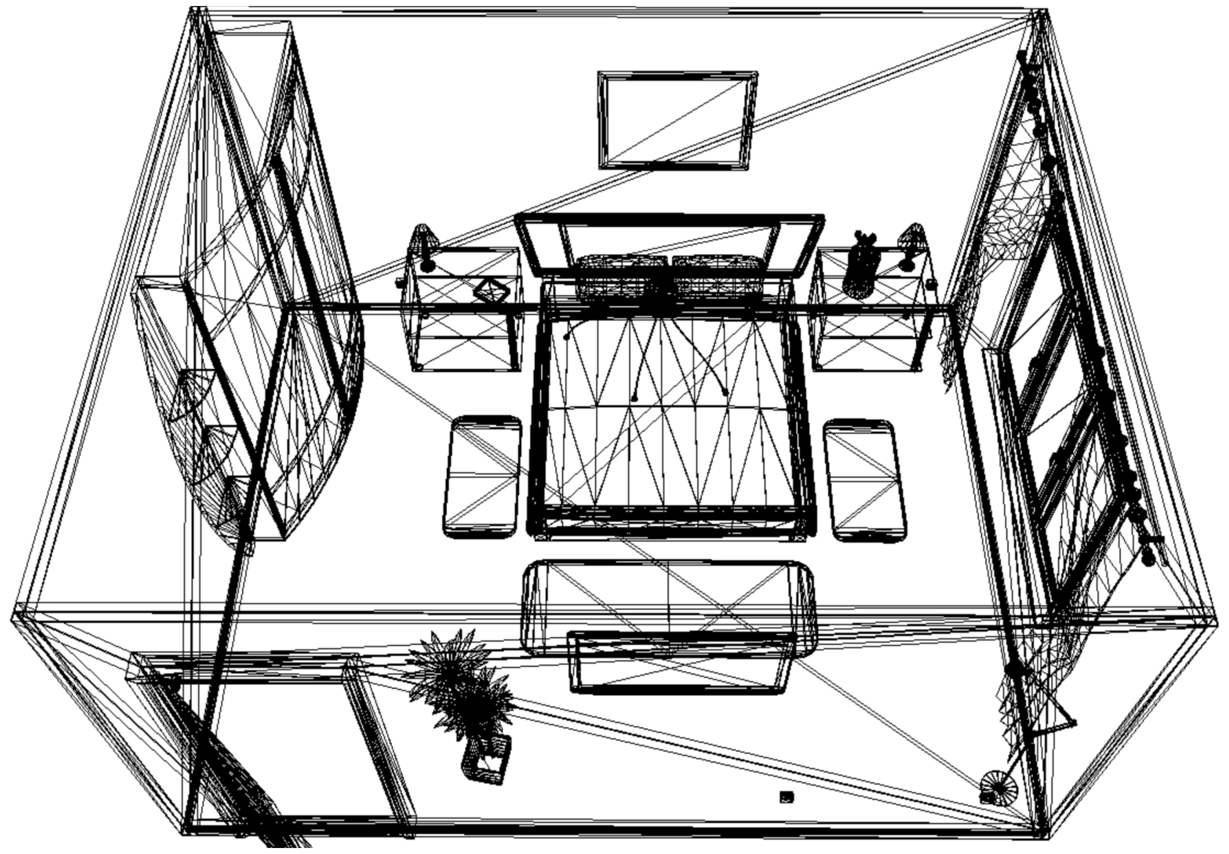
Beyond FoV Semantics&Structure

[Song et al. CVPR’18]



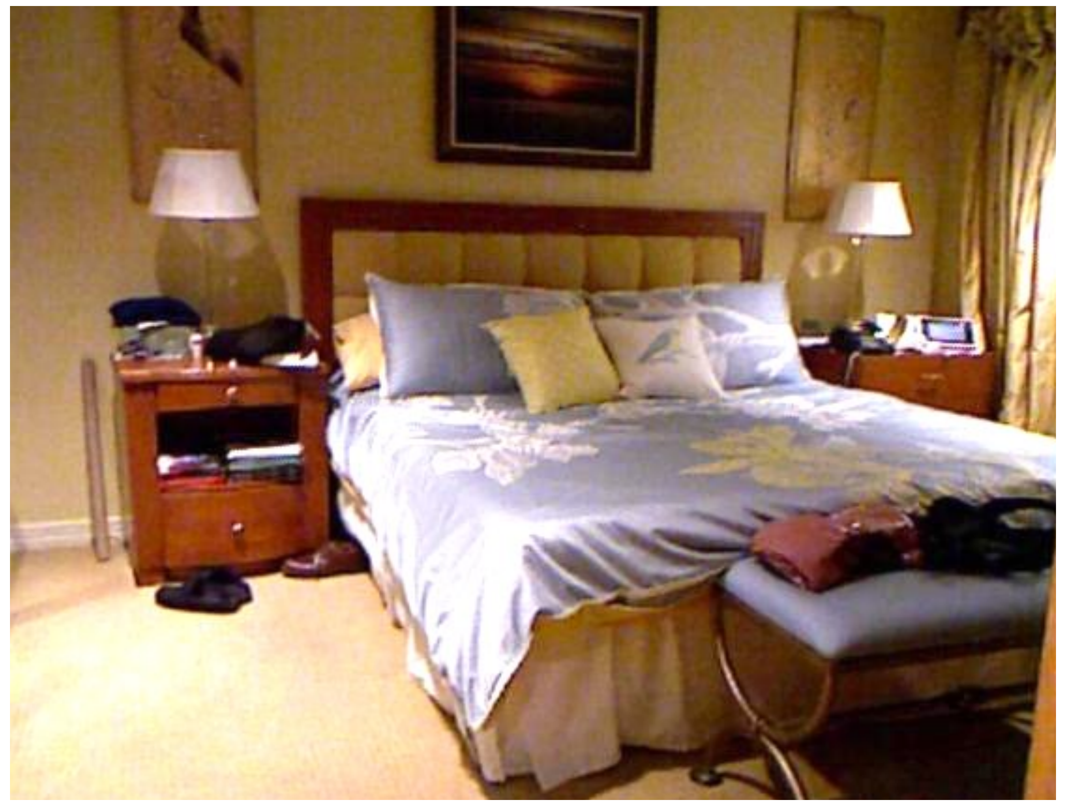
Beyond FoV Illumination

[Song and Funkhouser]

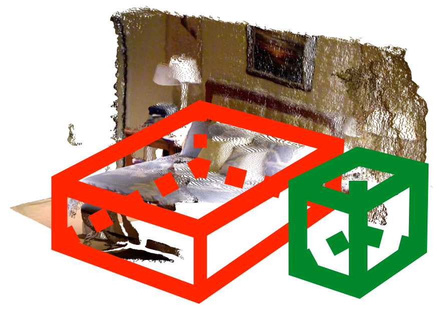


- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. Properties
- Dynamics
- …

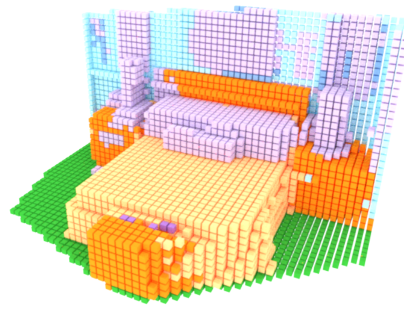
Advances Towards 3D Scene Understanding



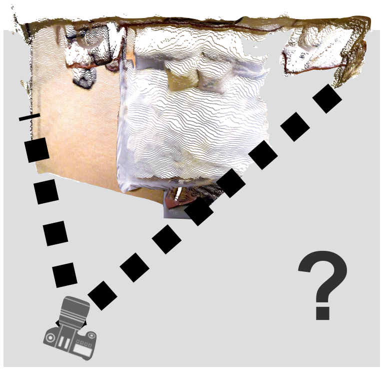
Passive Observers



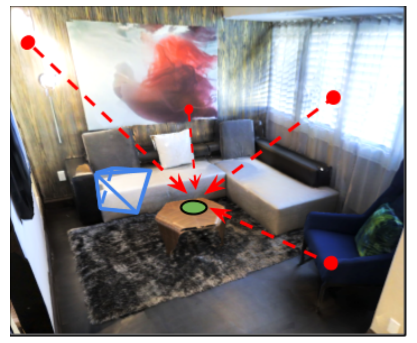
Amodal 3D



Higher Fidelity

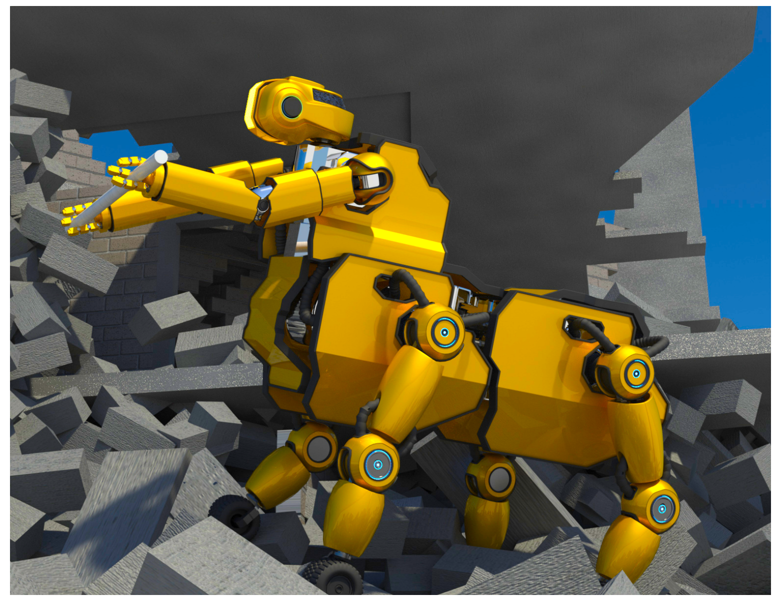


Beyond FoV

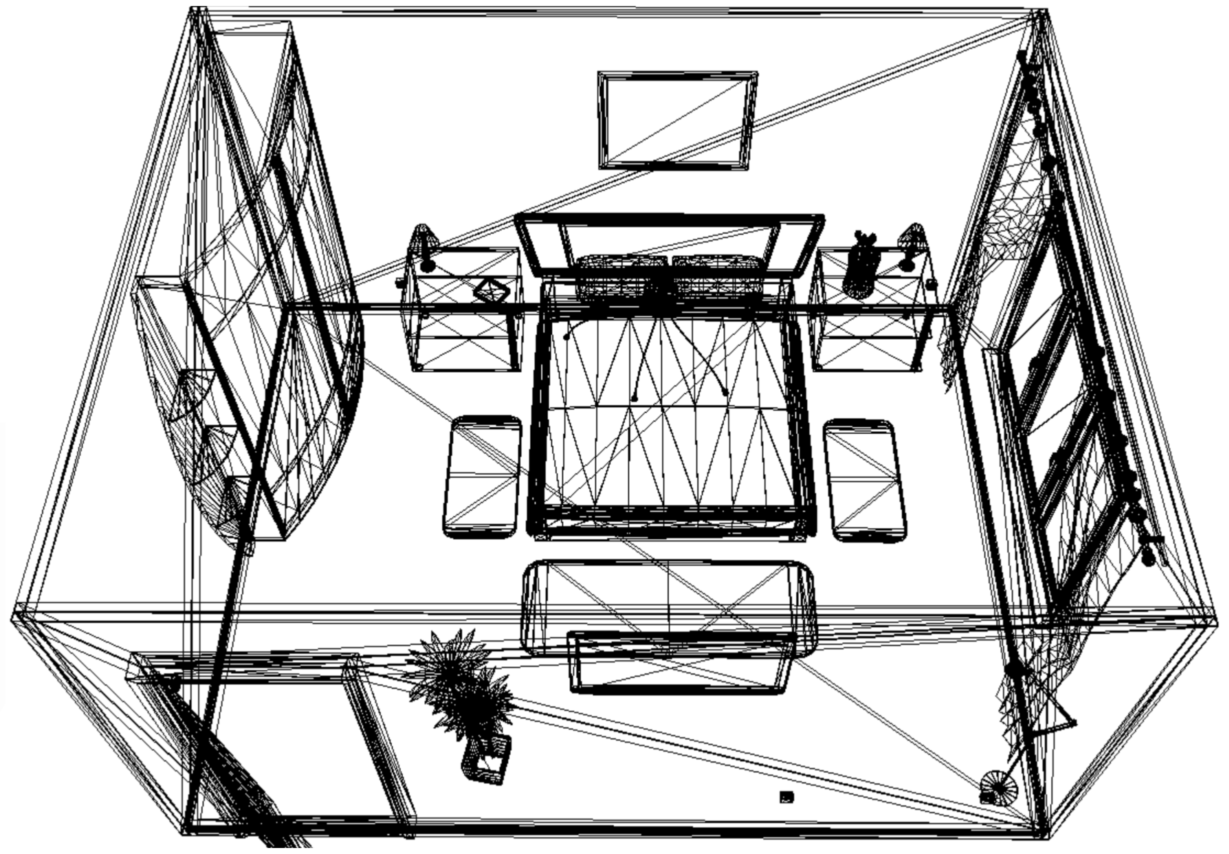


Beyond FoV Illumination

What's Next??



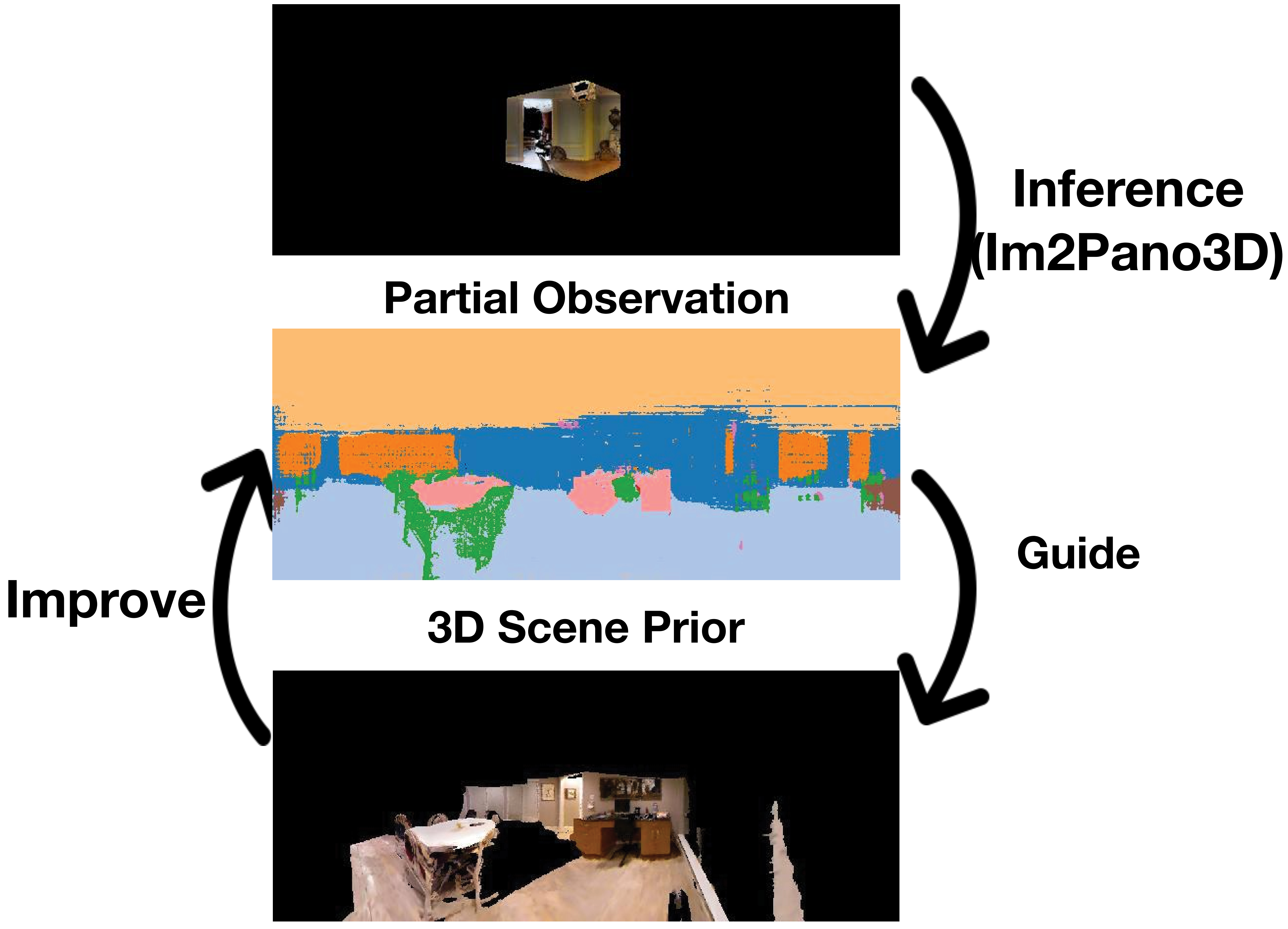
Active Explorers



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

Richer Representation through Interaction

Active Exploration



Partial Observation

Inference (Im2Pano3D)

3D Scene Prior

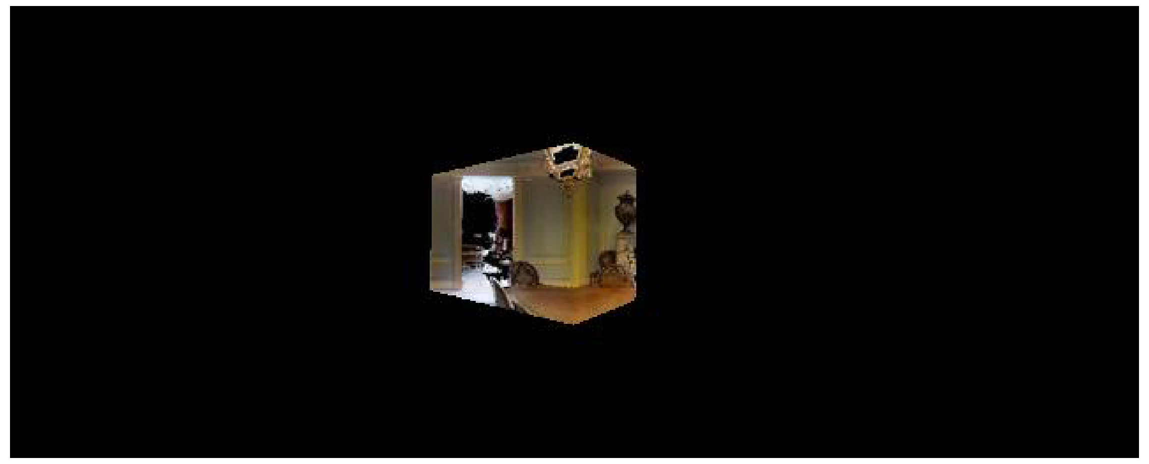
Guide

Improve

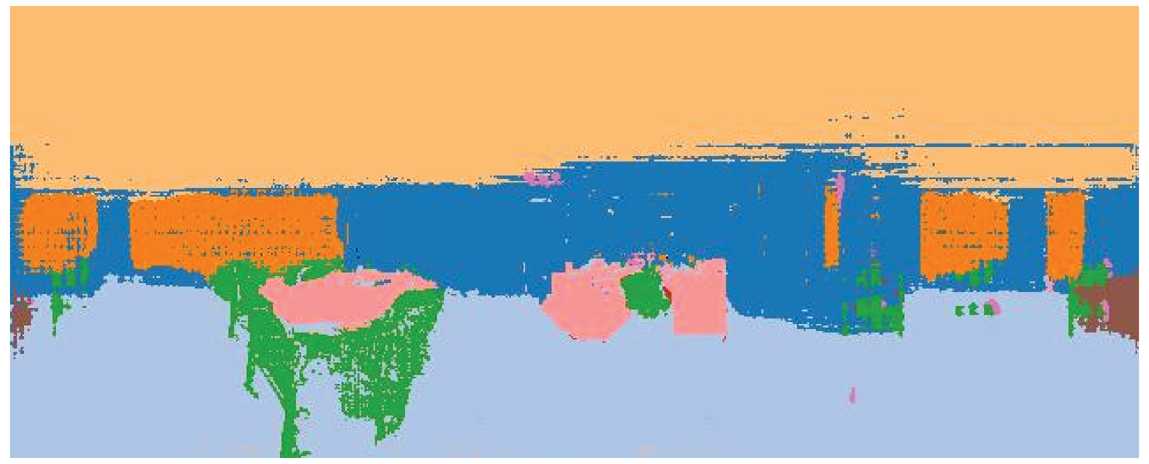
Efficient exploration
+ Most useful observation

Richer Representation through Interaction

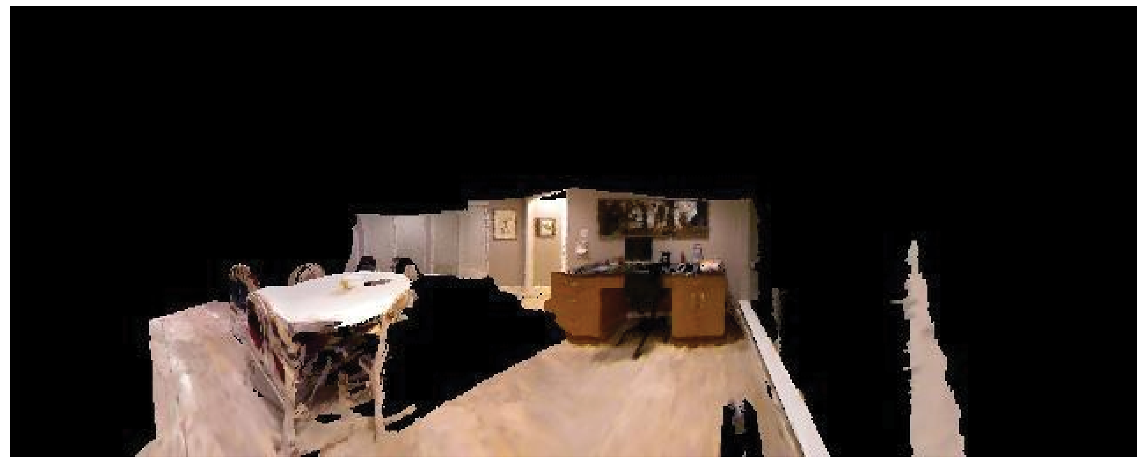
Active Exploration



Partial Observation

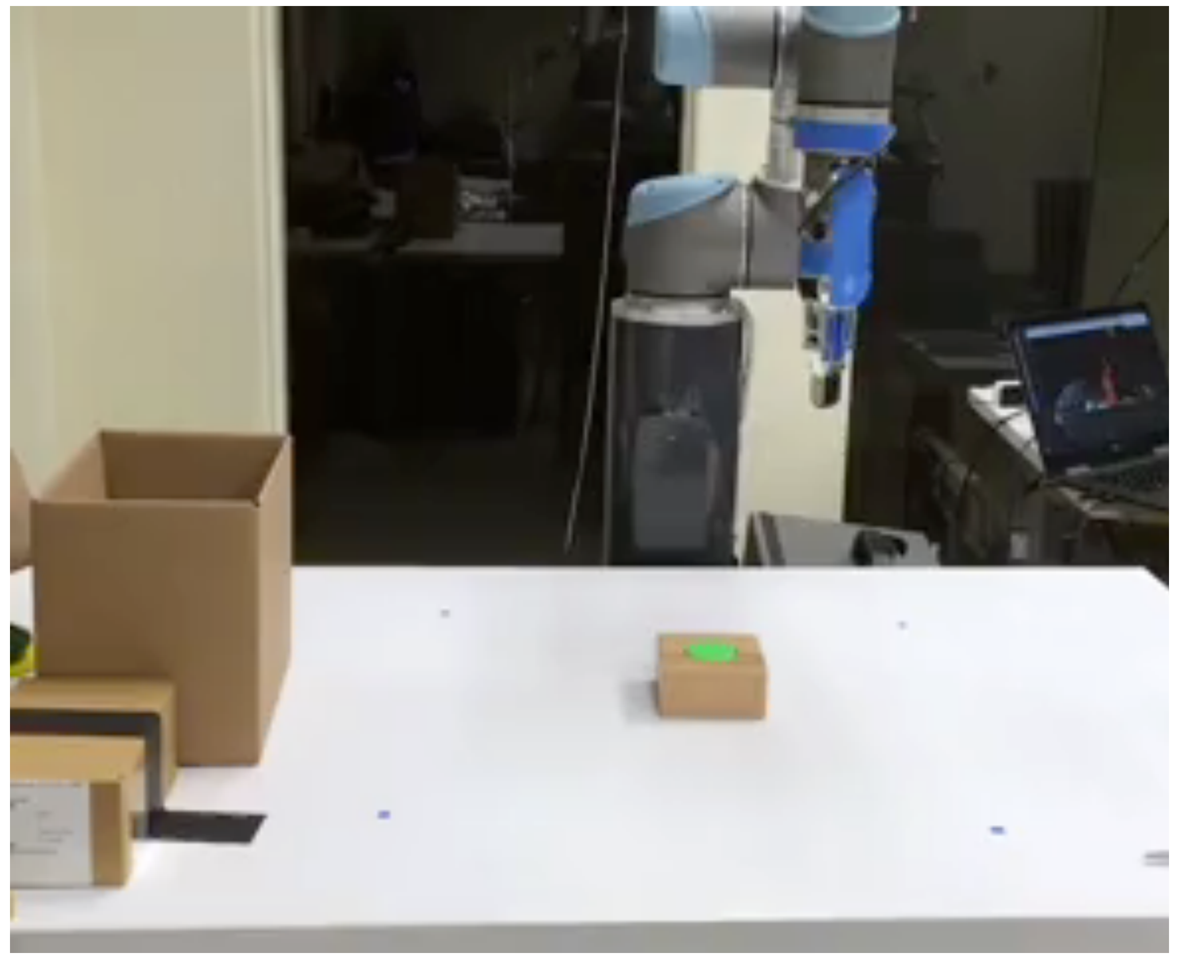


3D Scene Prior



Efficient exploration
+ Most useful observation

Active physical Interaction

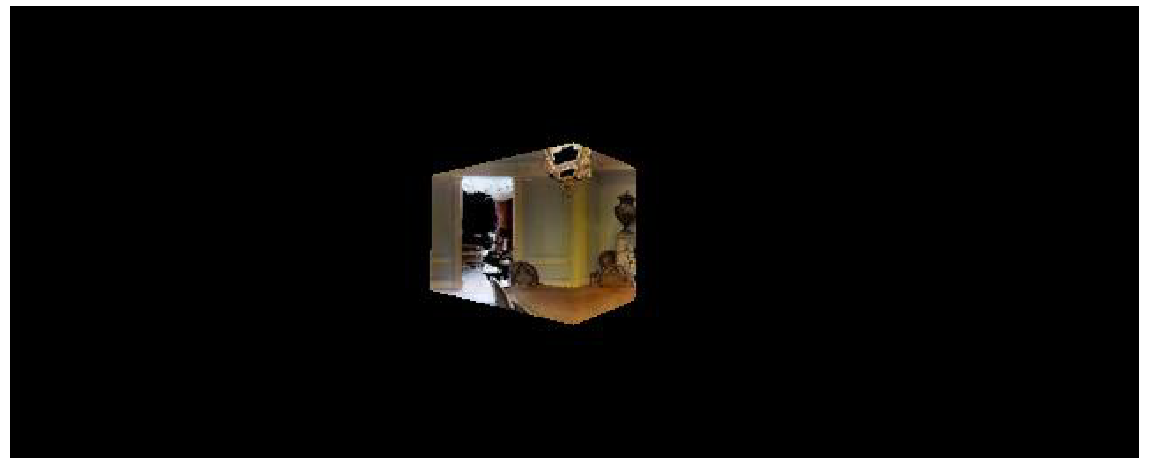


Actions: Poking,Grasping

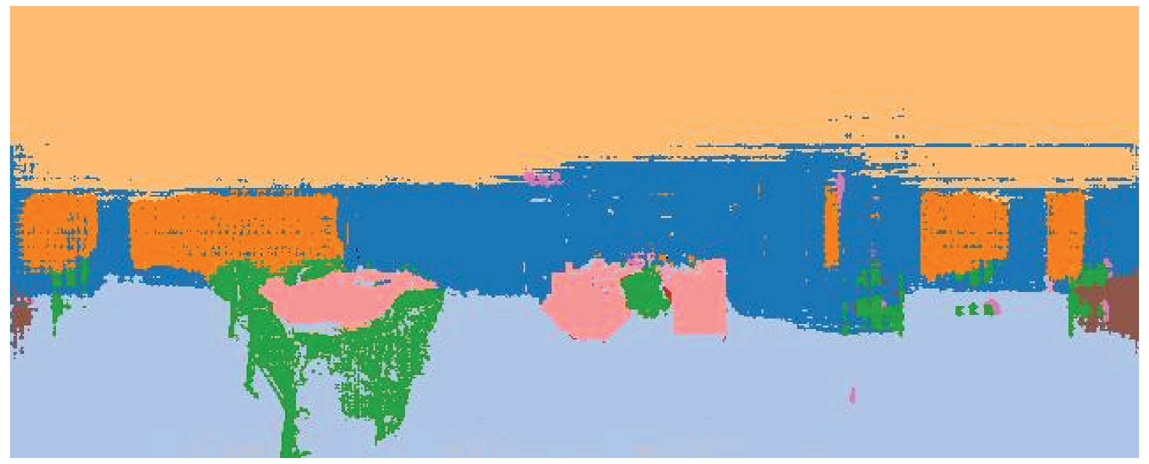
Physical properties:
Surface material
Friction coefficient

Richer Representation through Interaction

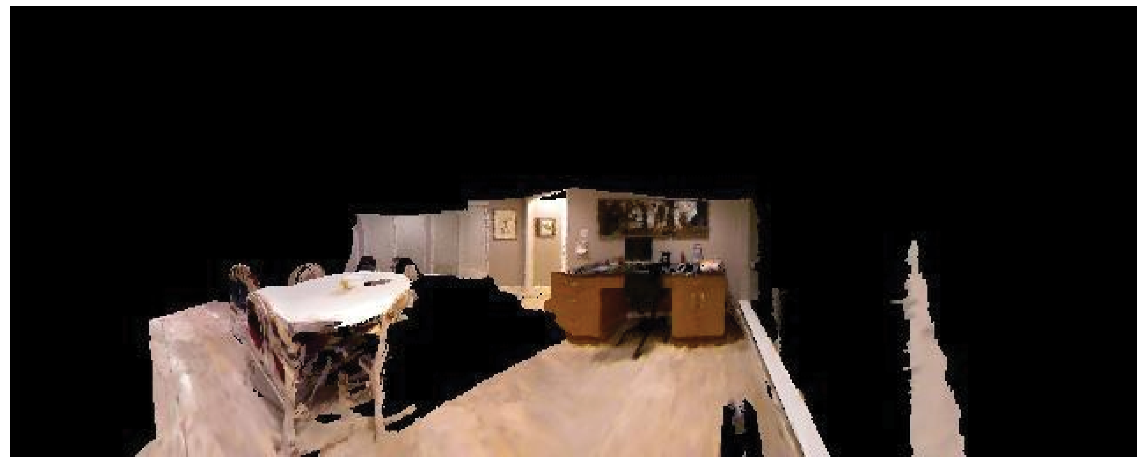
Active Exploration



Partial Observation

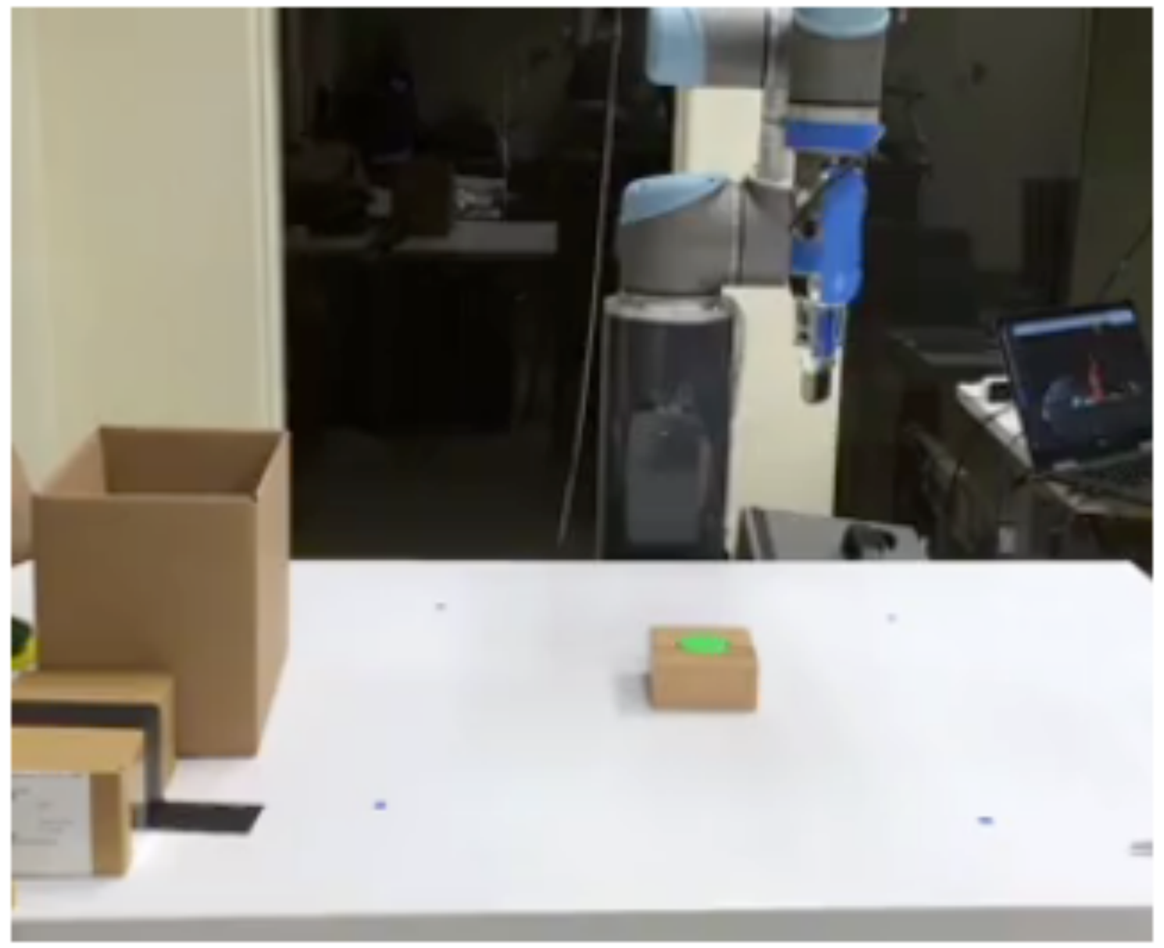


3D Scene Prior



Efficient Exploration
+ Most useful observation

Active physical Interaction



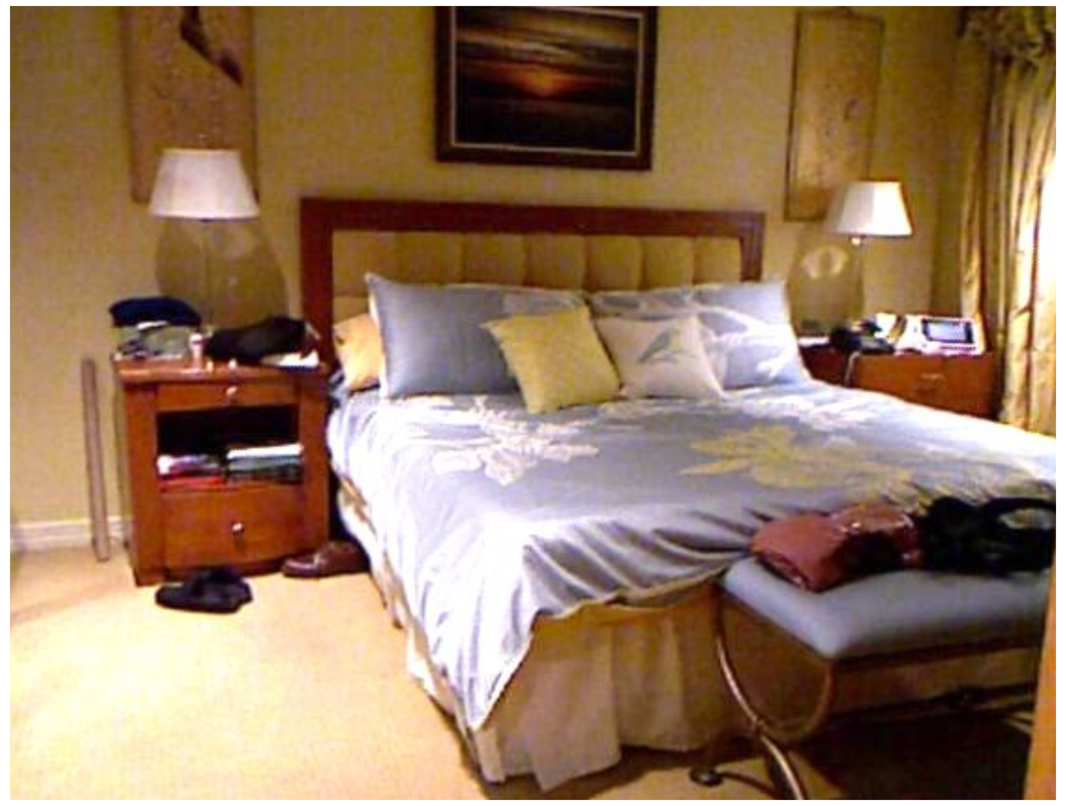
Actions: Pushing,Grasping

Physical properties:
Surface material
Friction coefficient

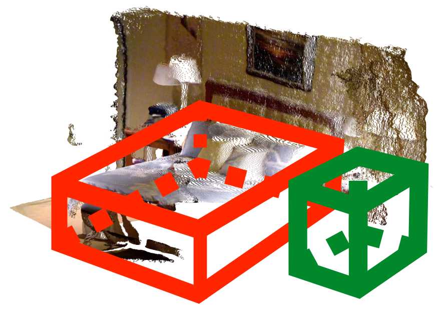
Actions: Tossing

Physical properties:
Mass distribution,
Aerodynamic

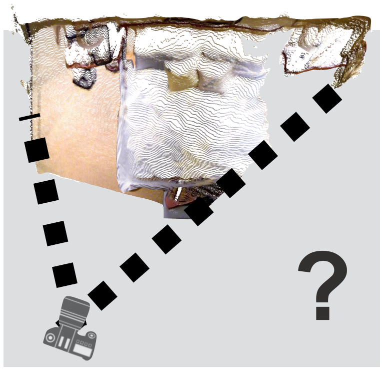
Comprehensive 3D Scene Understanding



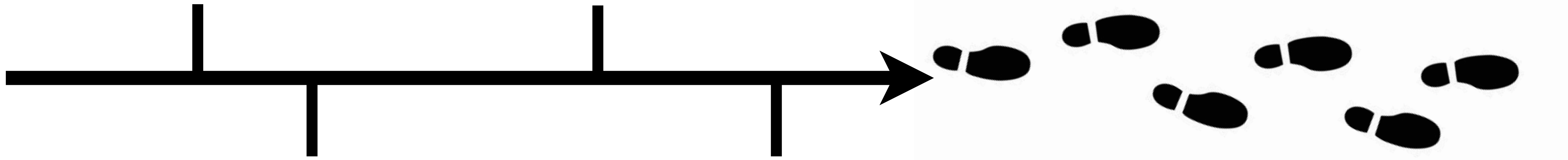
Passive Observers

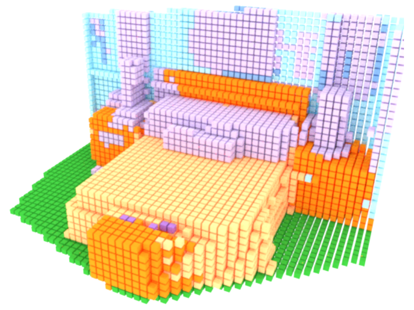


Amodal 3D

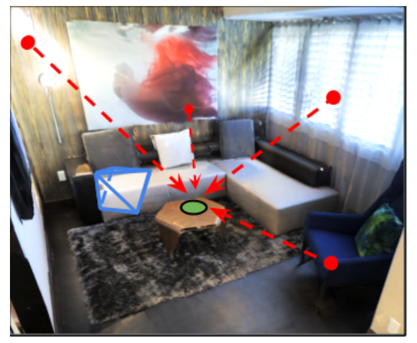


Beyond FoV



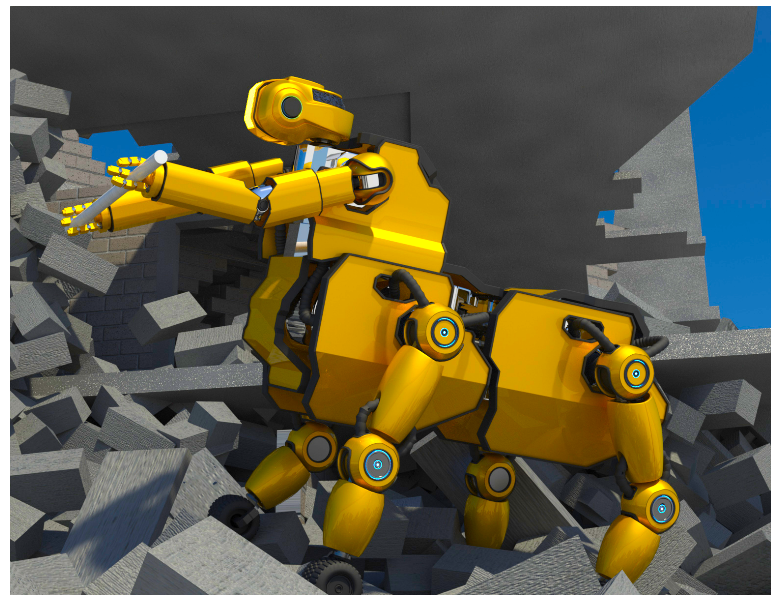


Higher Fidelity

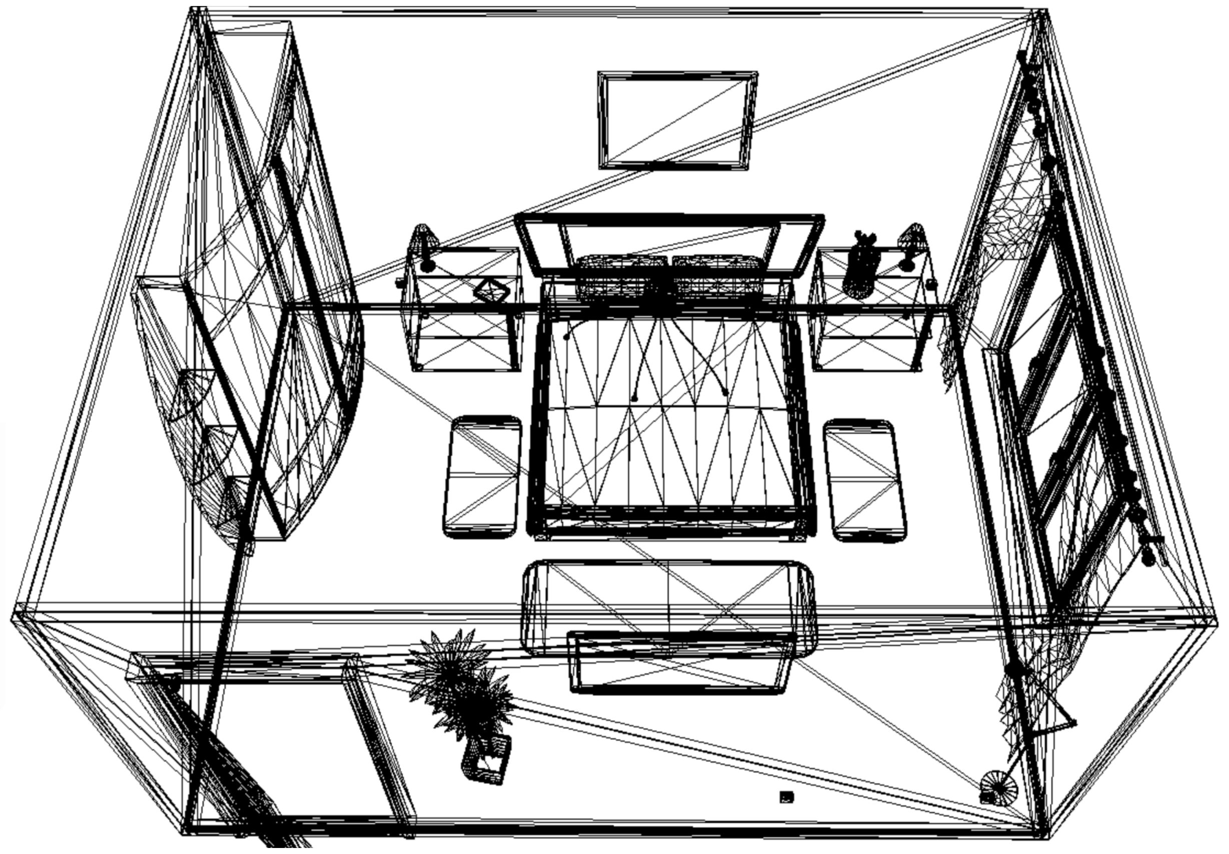


Beyond FoV Illumination

What's Next??



Active Explorer



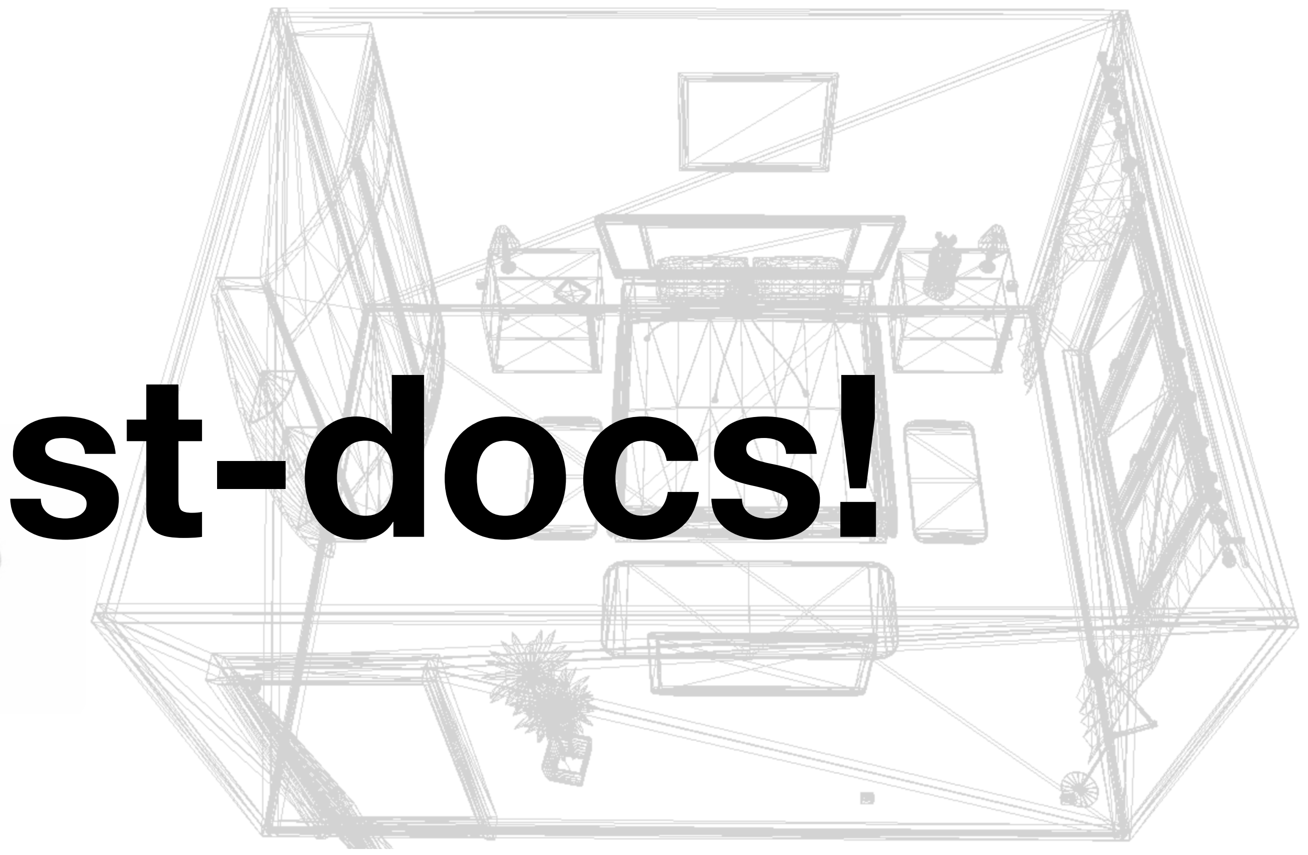
- Semantics Category
- 3D Location, Size
- Detailed Geometry
- Inter-Object Relationships
- Not Limited by FoV
- Action Affordances
- Phys. Properties
- Dynamics
- …

Comprehensive 3D Scene Understanding

We are looking for PhDs and Post-docs!

Amodal 3D
Beyond FoV
?
Higher Fidelity
Illumination
Passive Observers
Active Explorer

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK



- 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!