PRINCETON
VISION GROUP

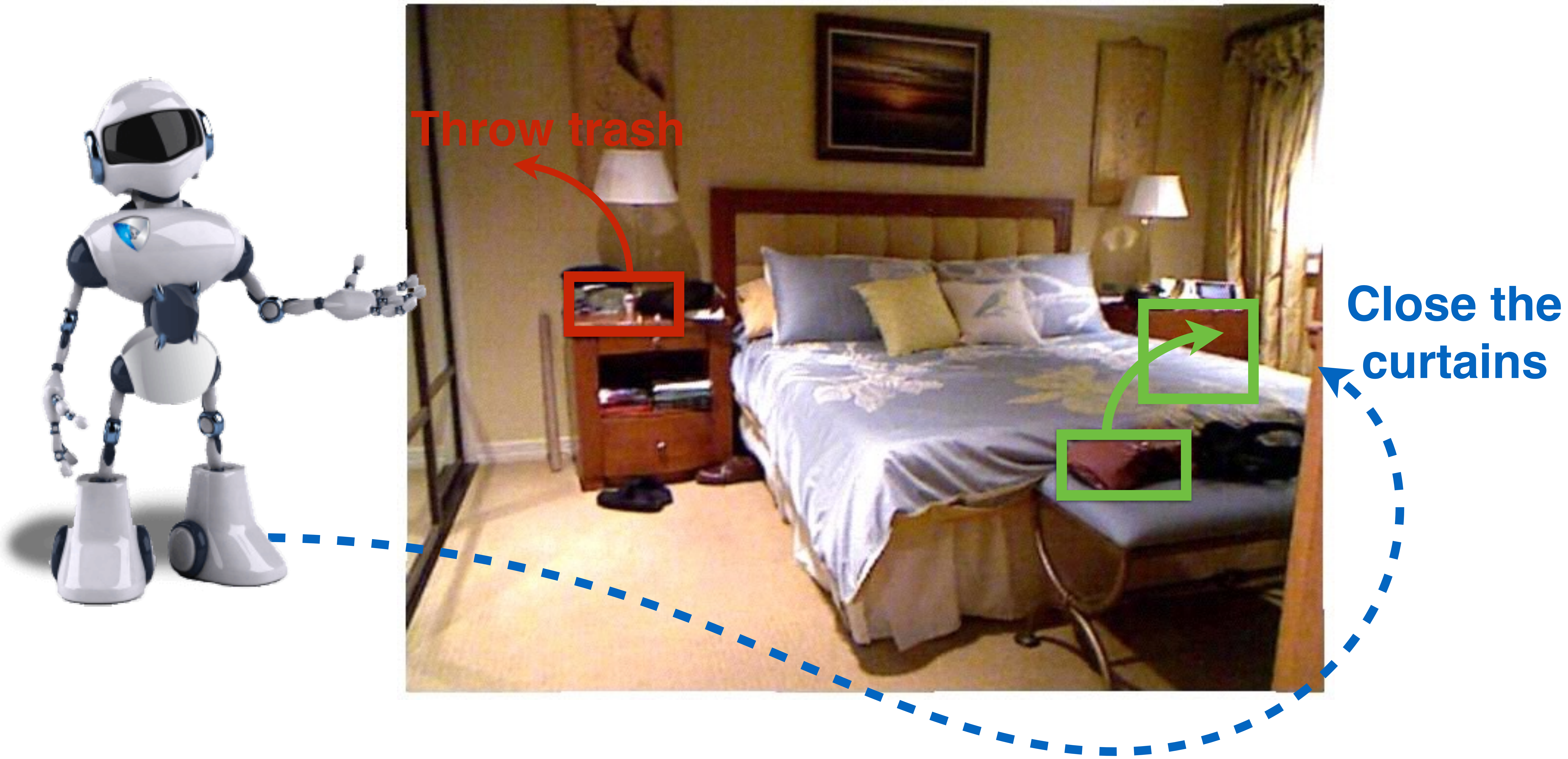
Semantic Scene Completion

from a Single Depth Image

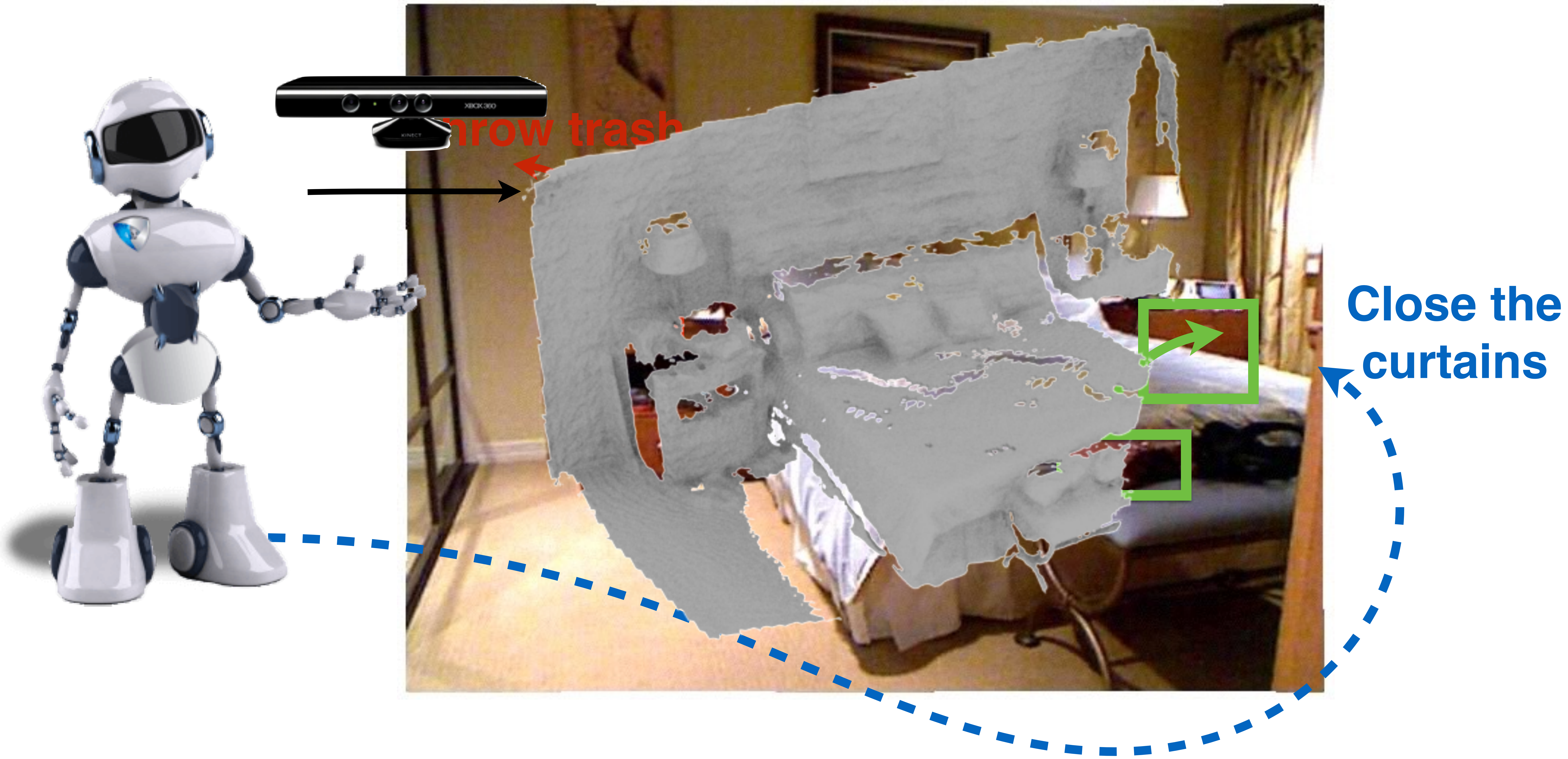
Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva and Thomas Funkhouser

Princeton University

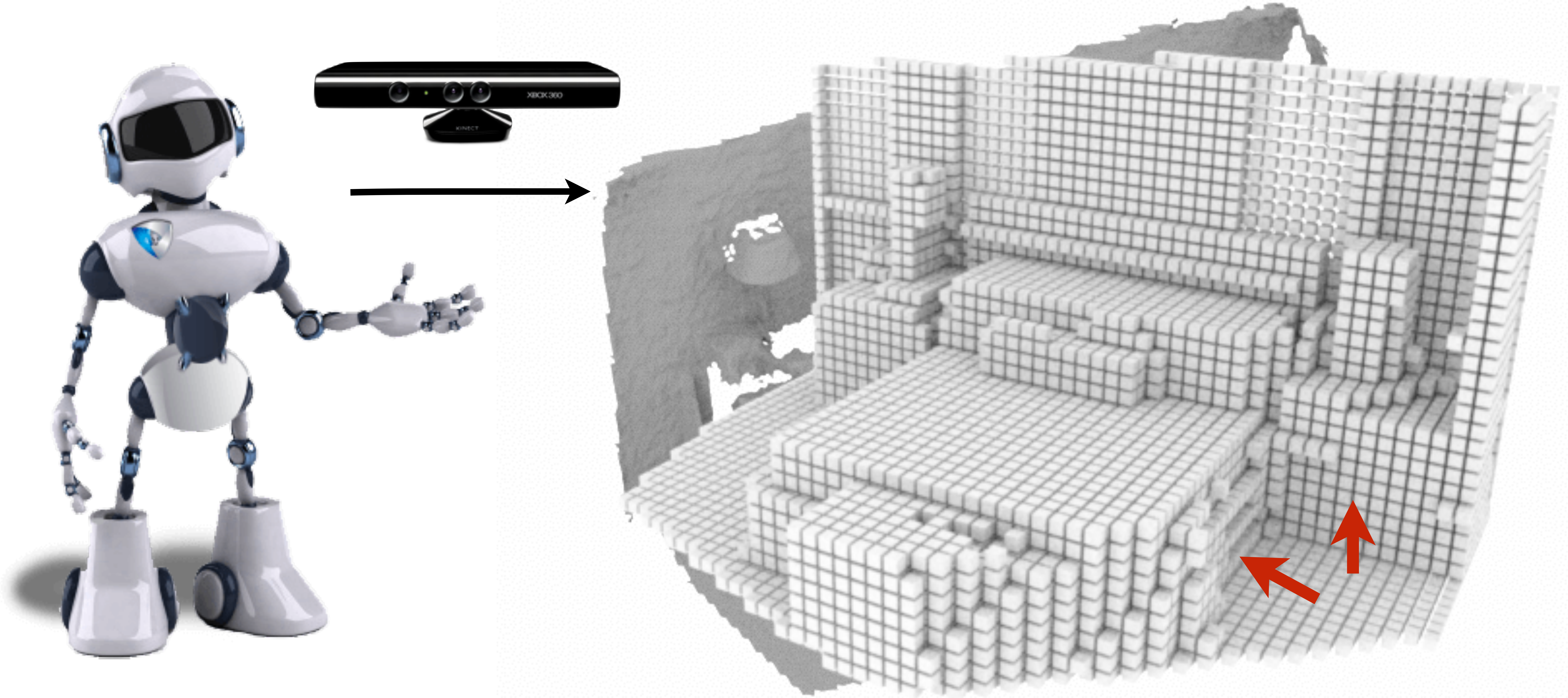
Motivation



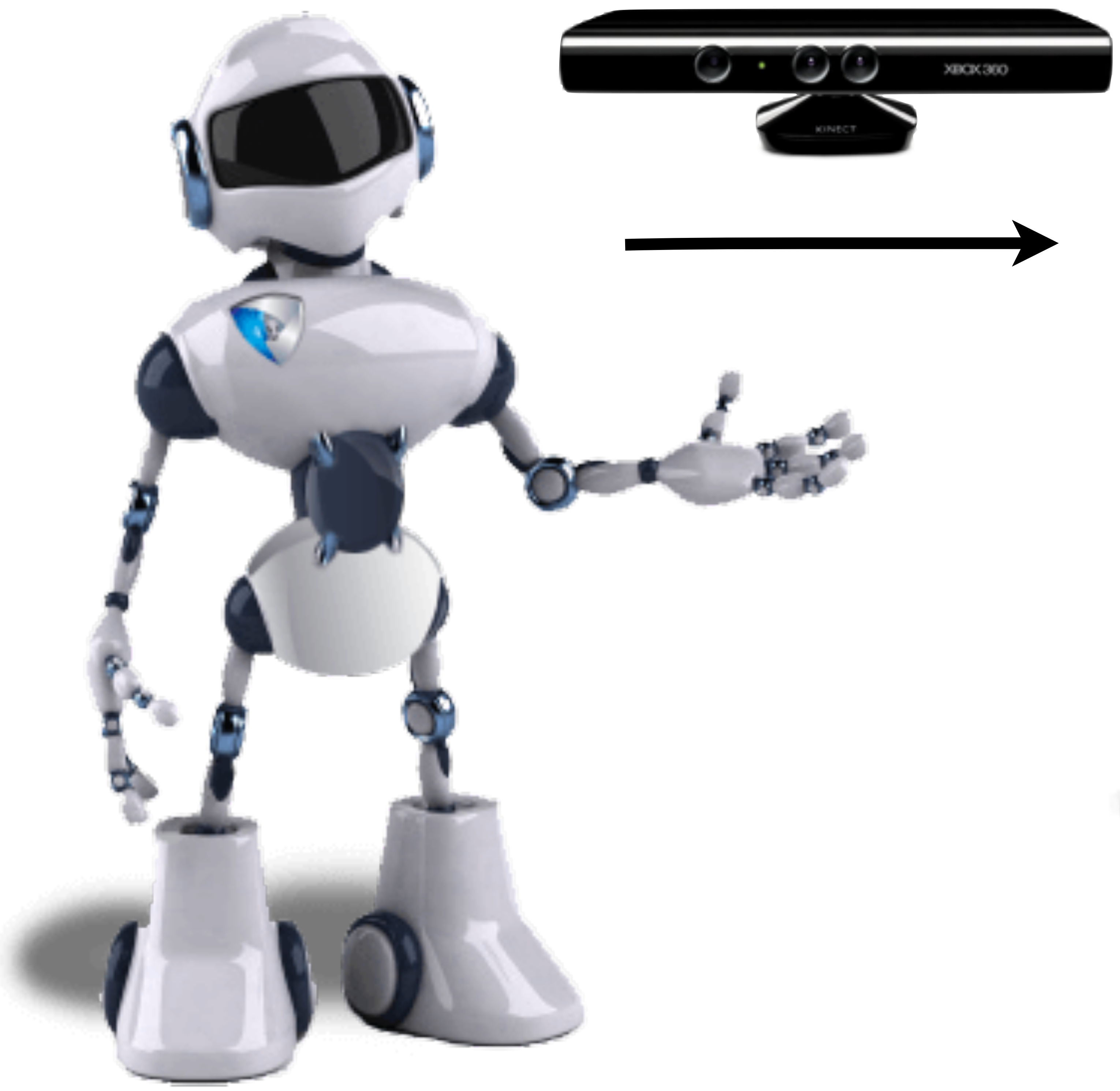
Partial observation

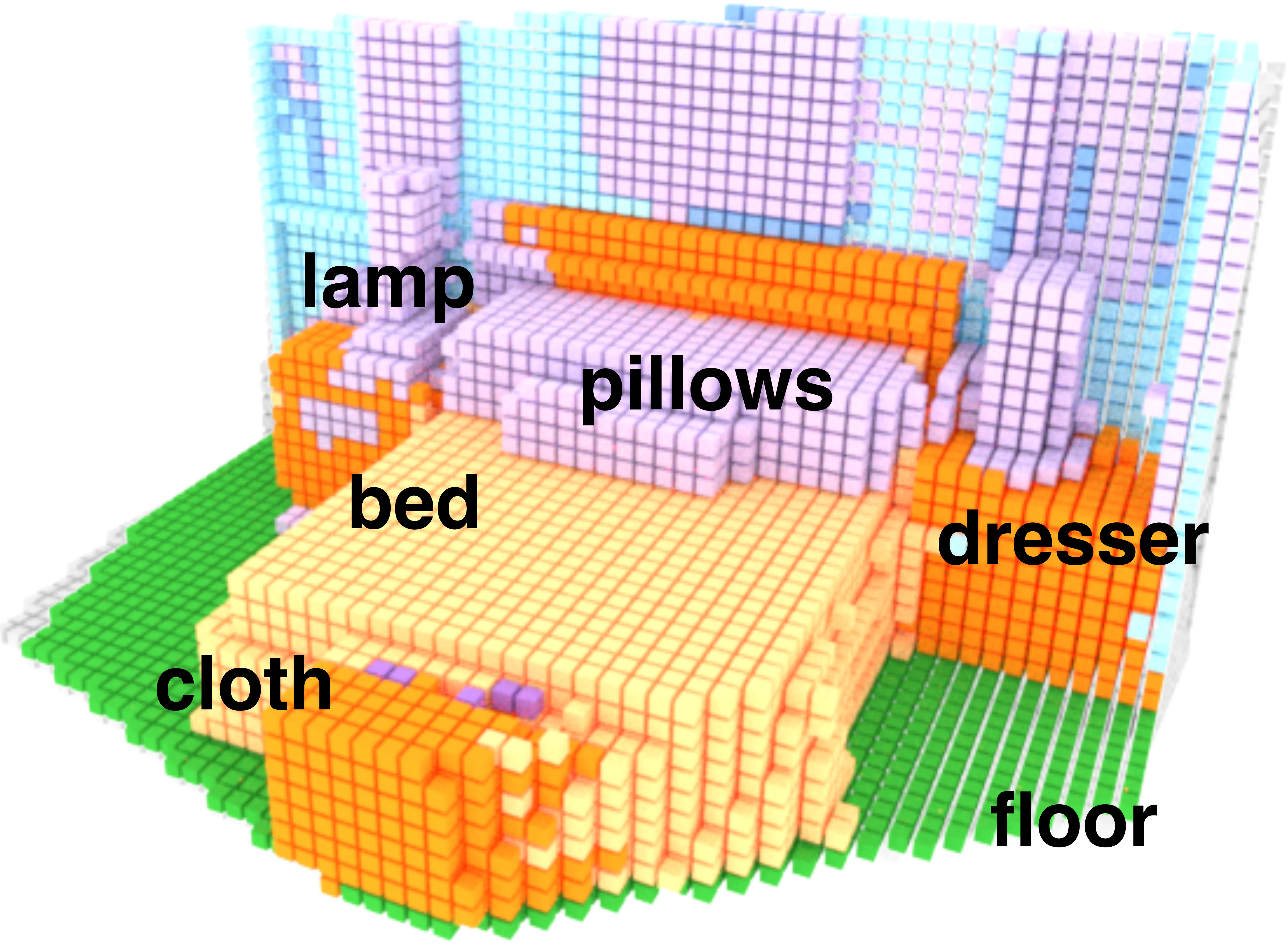


Complete 3D structure



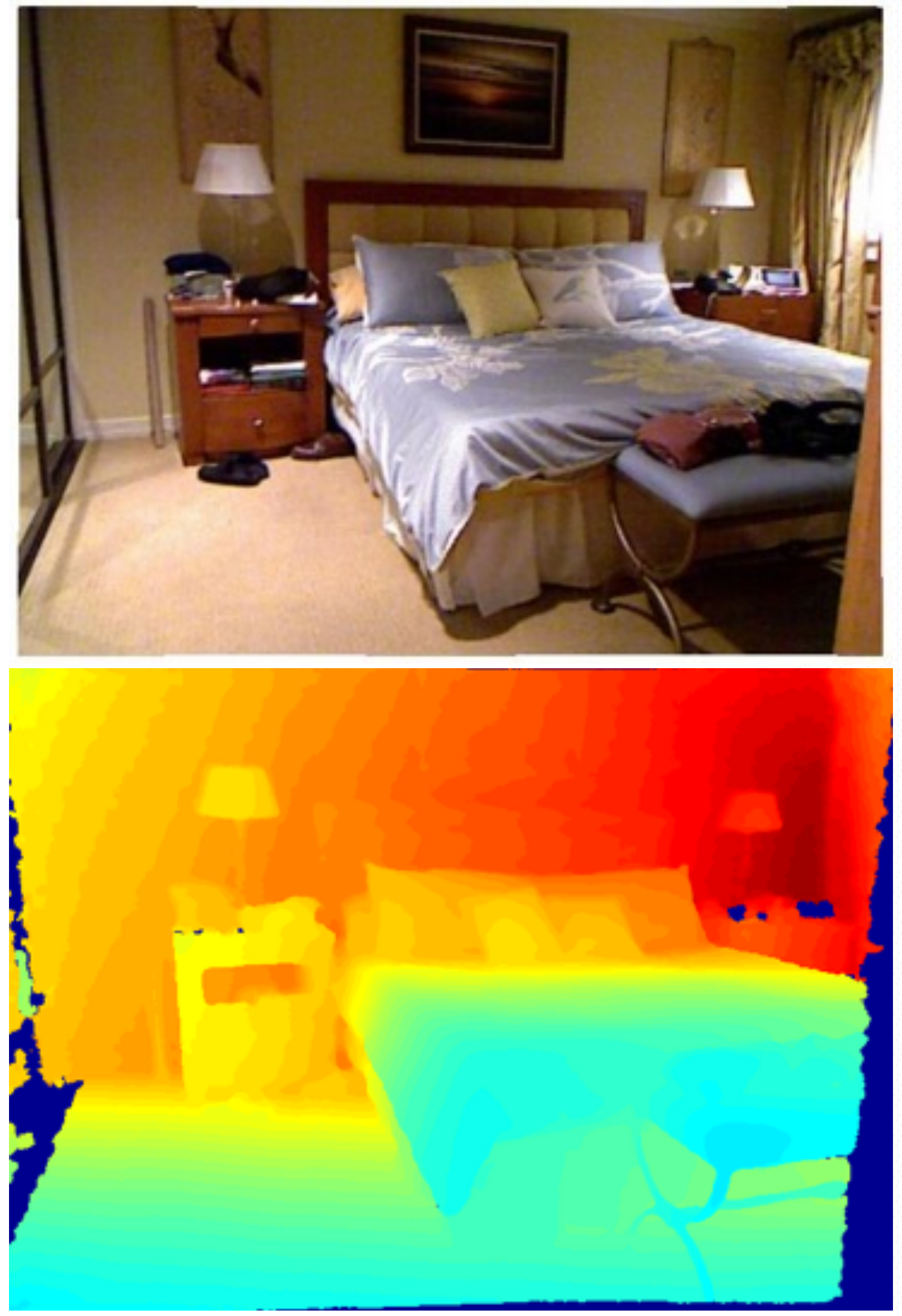
Semantic meaning

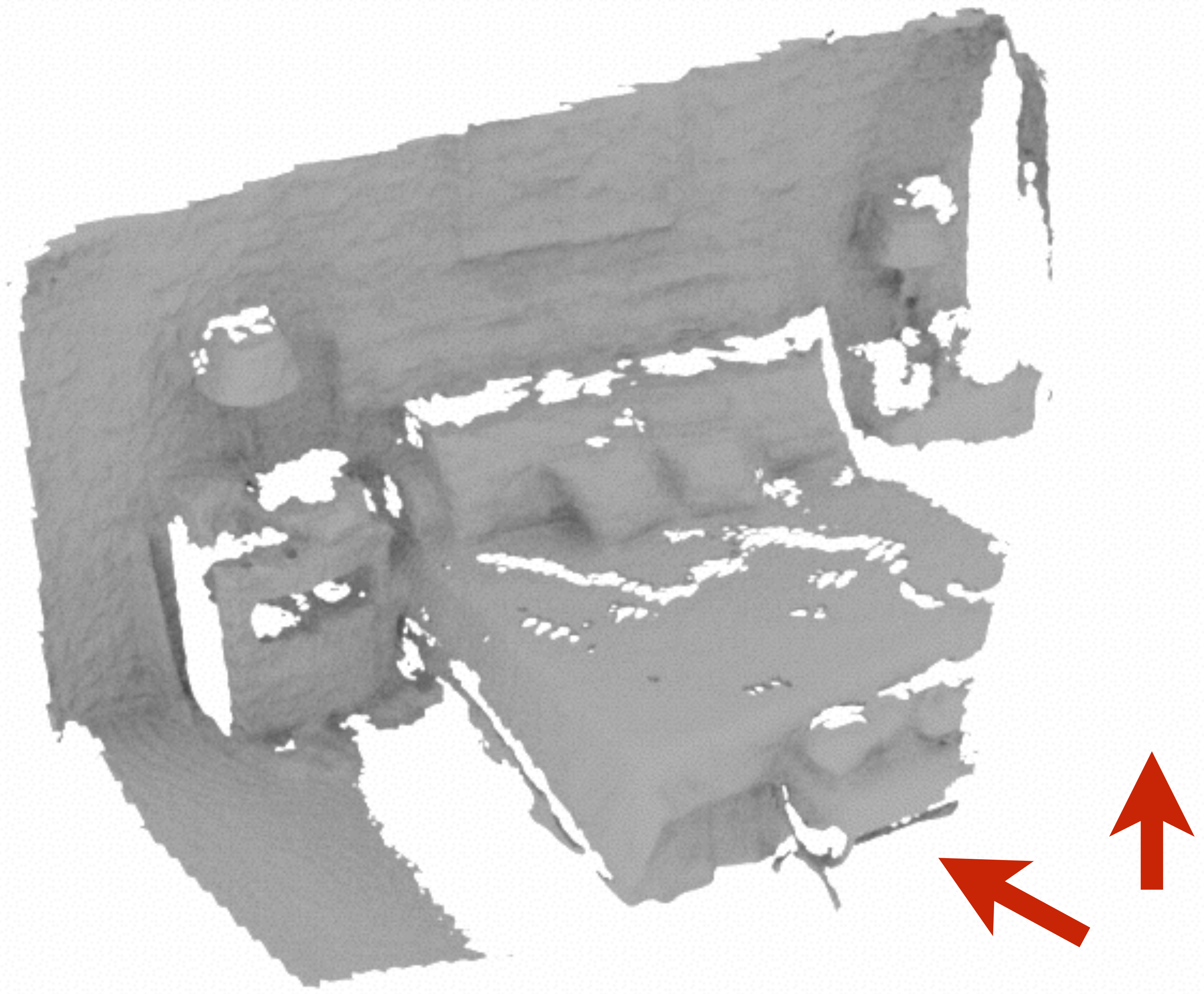


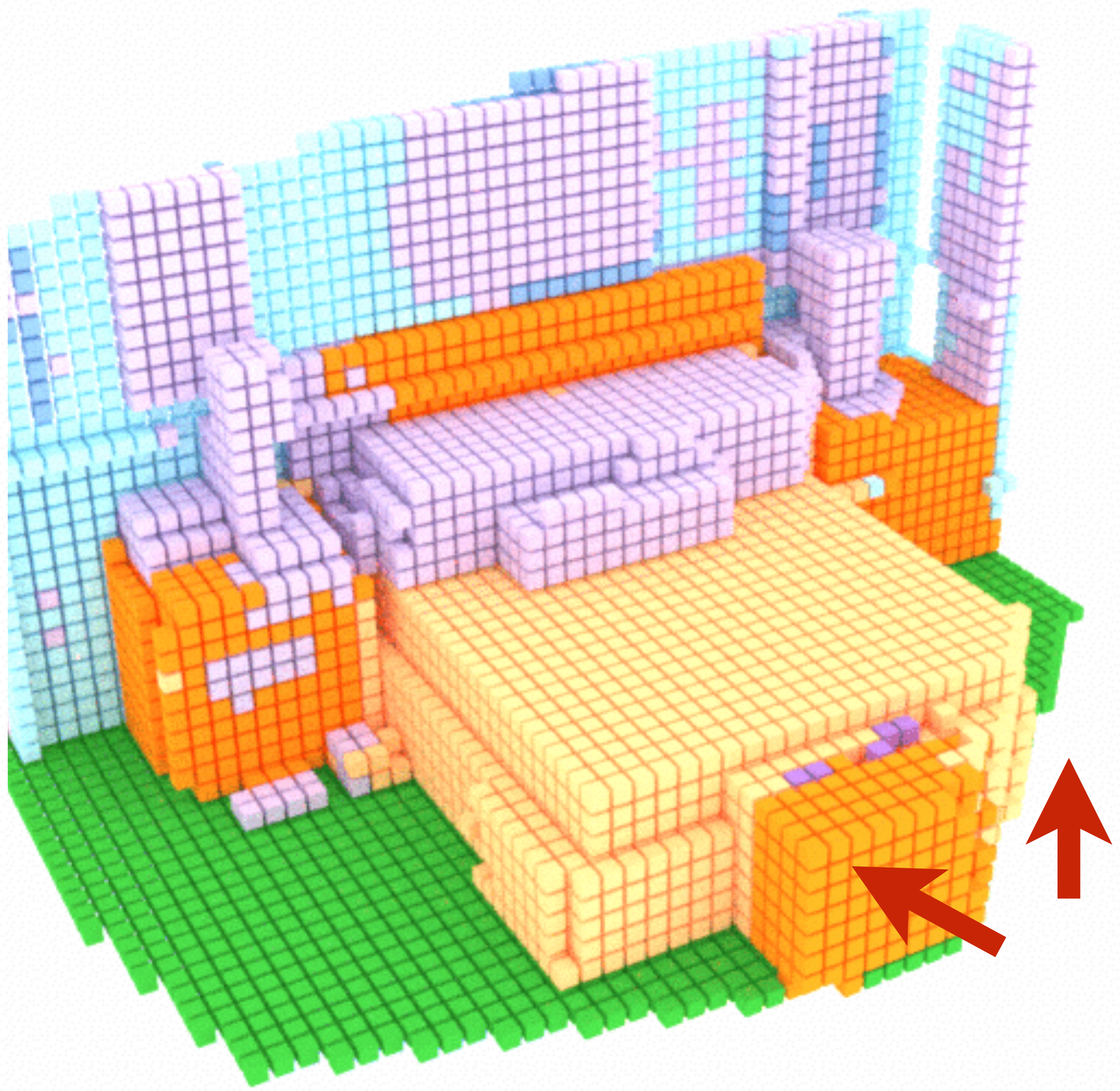


Semantic Scene Completion

Goal: volumetric occupancy + semantic



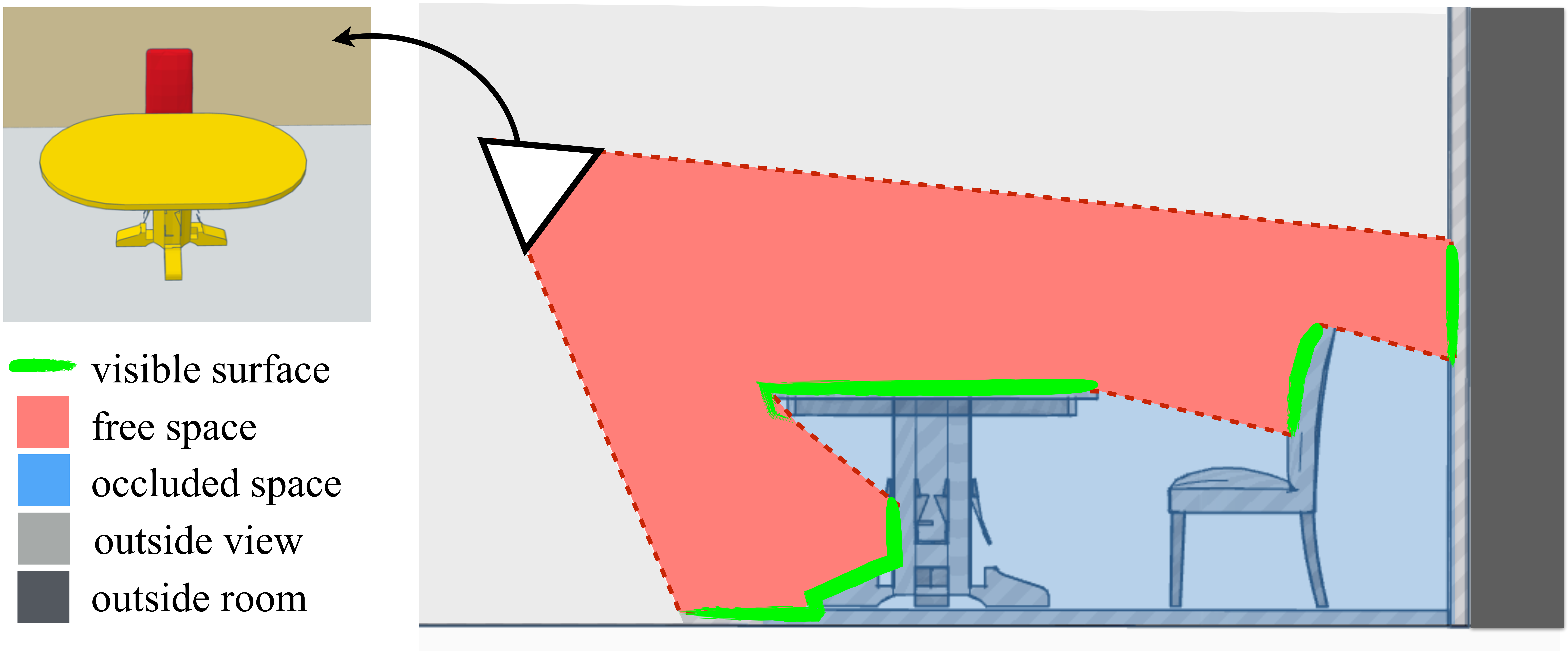




Input:
Single view depth map

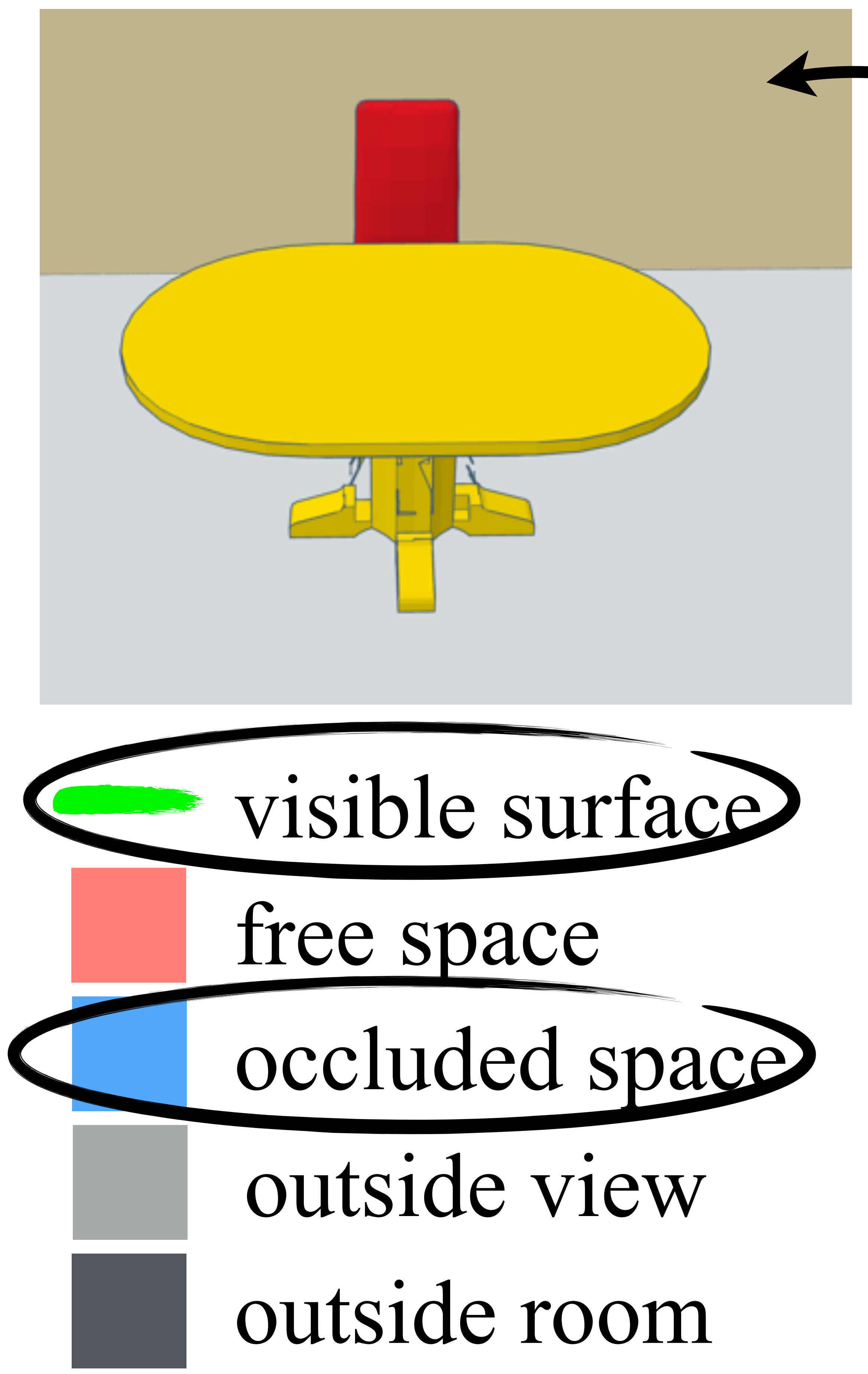
Output:
volumetric occupancy + semantic

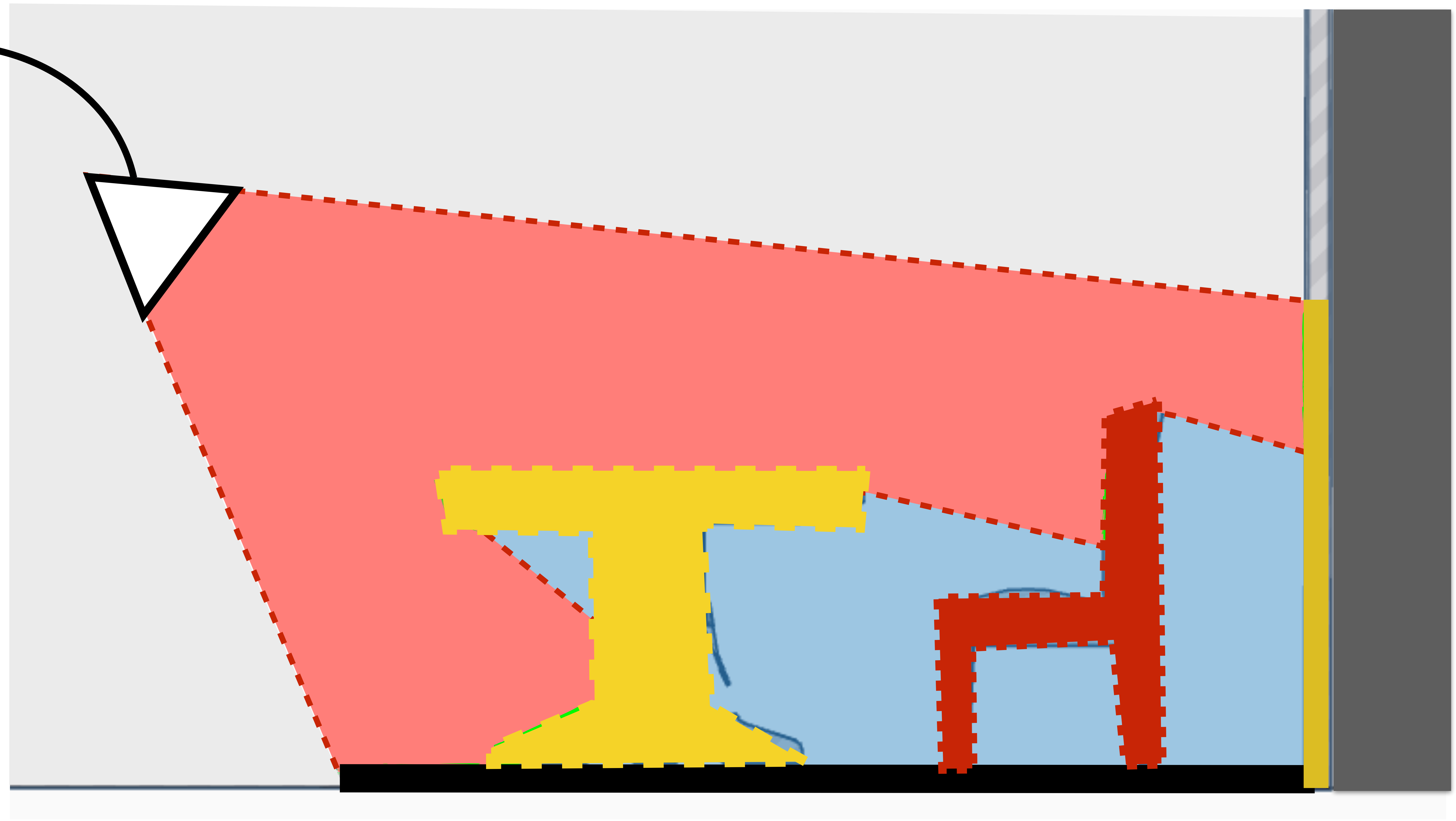
Problem definition



3D Scene

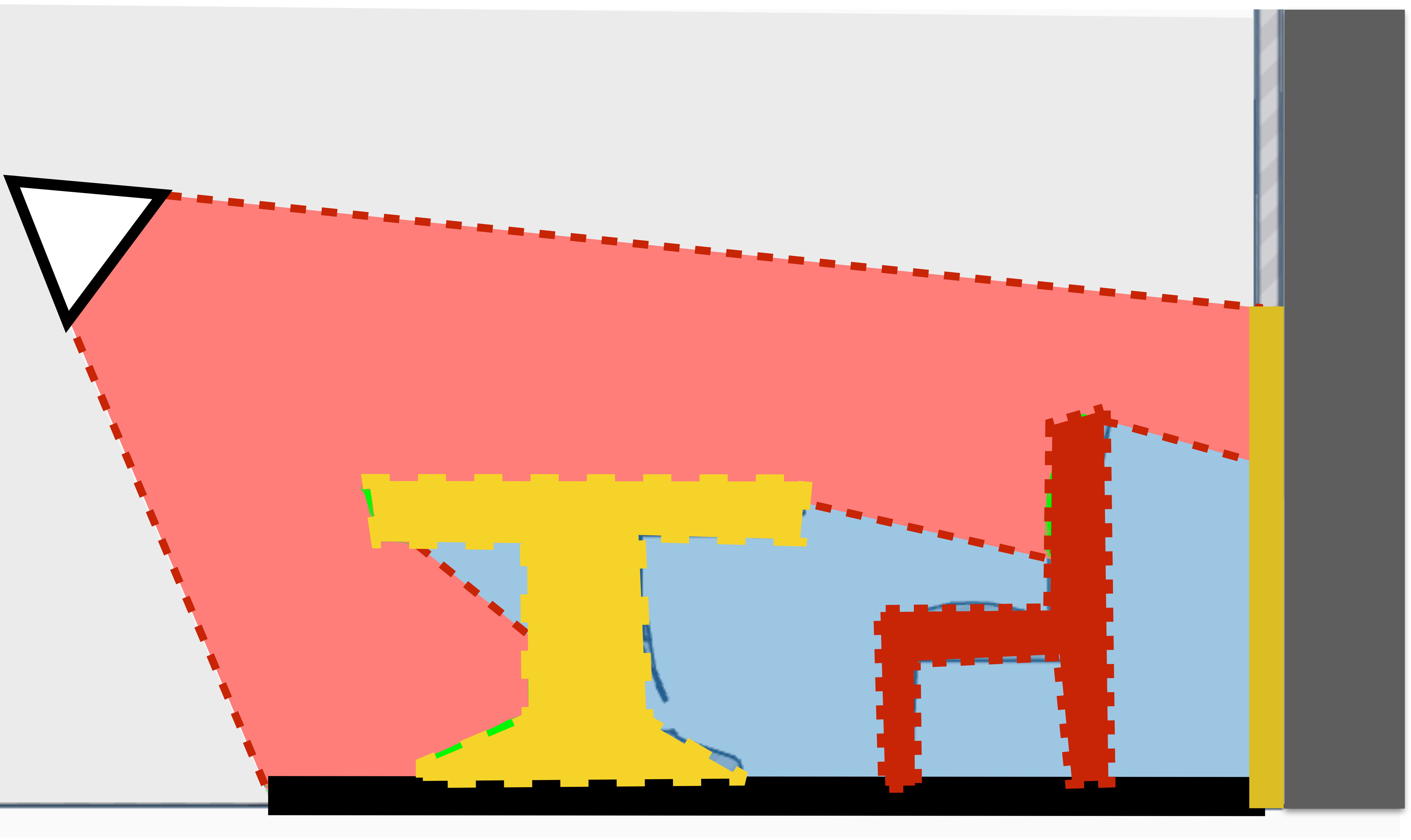
Problem definition





3D Scene

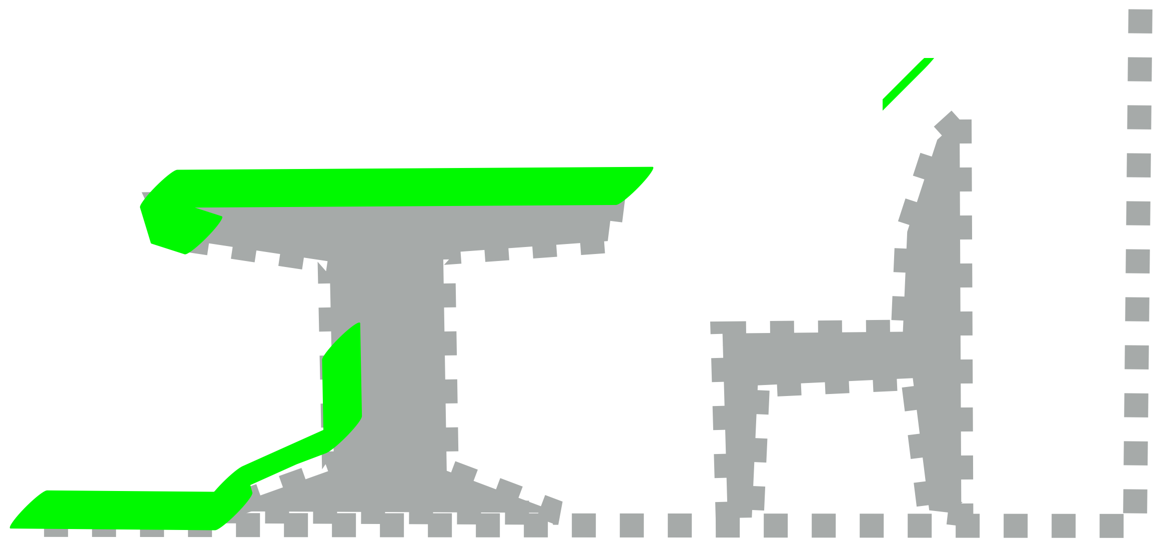
Prior work



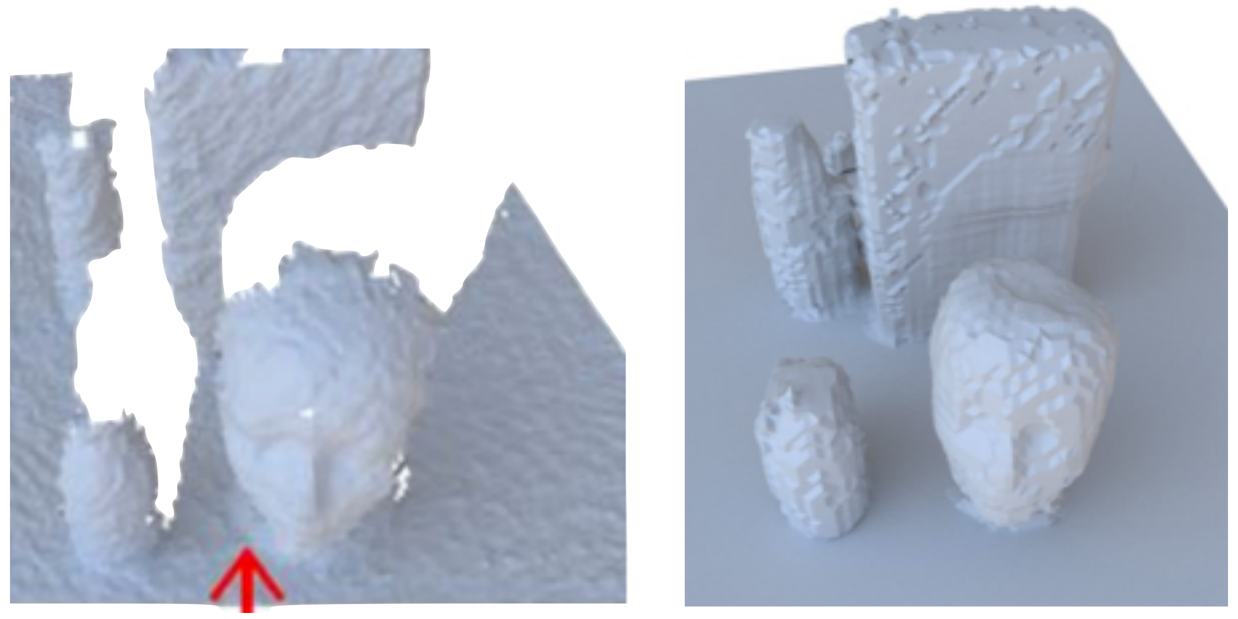
3D Scene

surface segmentation

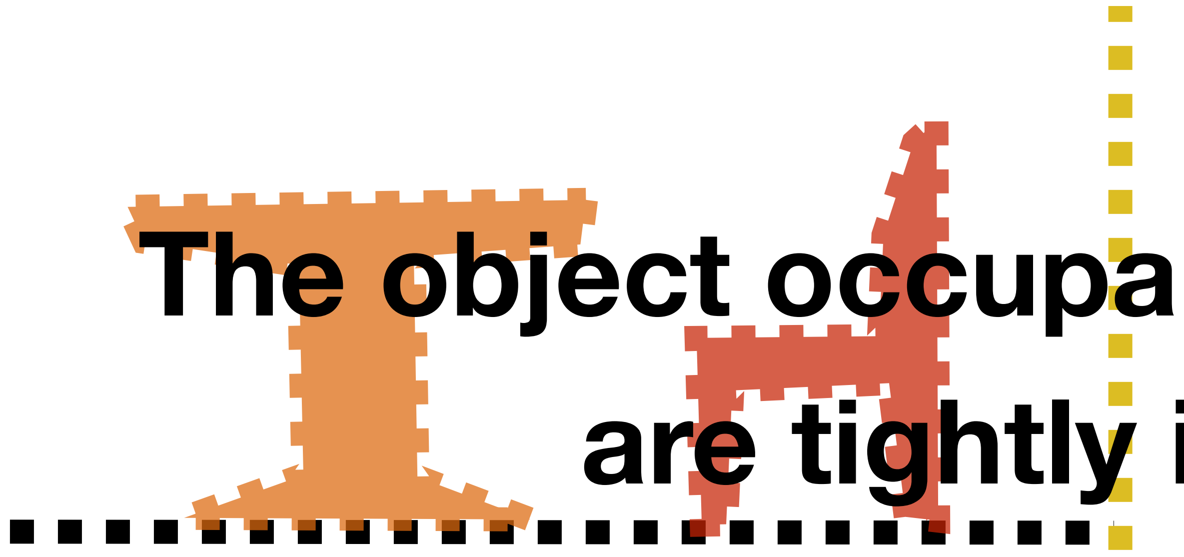
[Silberman *et al.*]



scene completion



[Firman *et al.*]



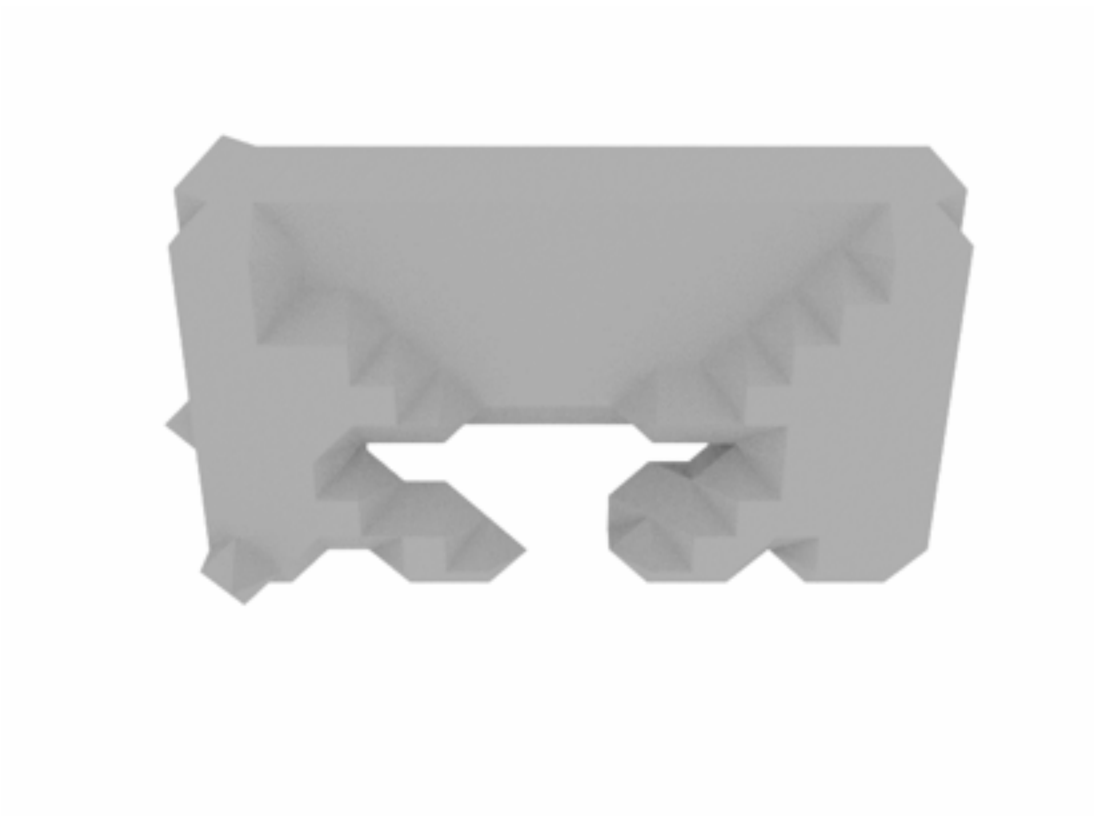
The object occupancy and the identity are tightly intertwined !

This paper

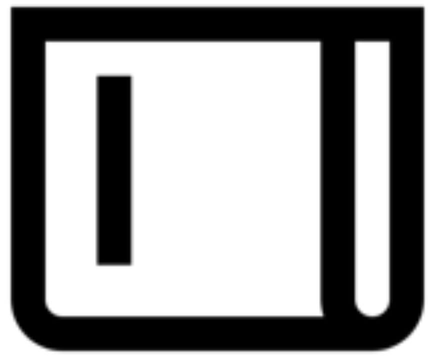
semantic scene completion

Object occupancy and semantic

Partial scan of common object



What is this?
What's the complete shape?

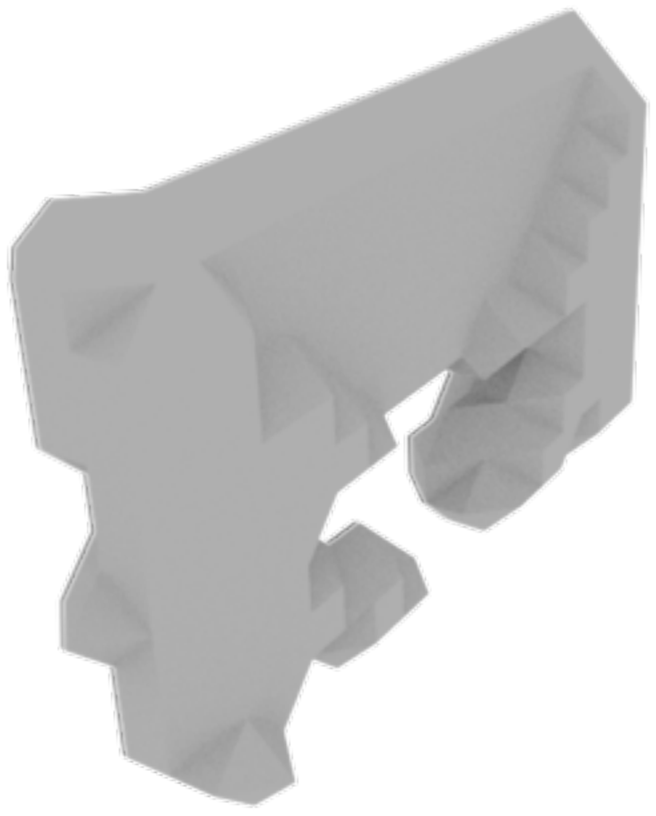


a book?

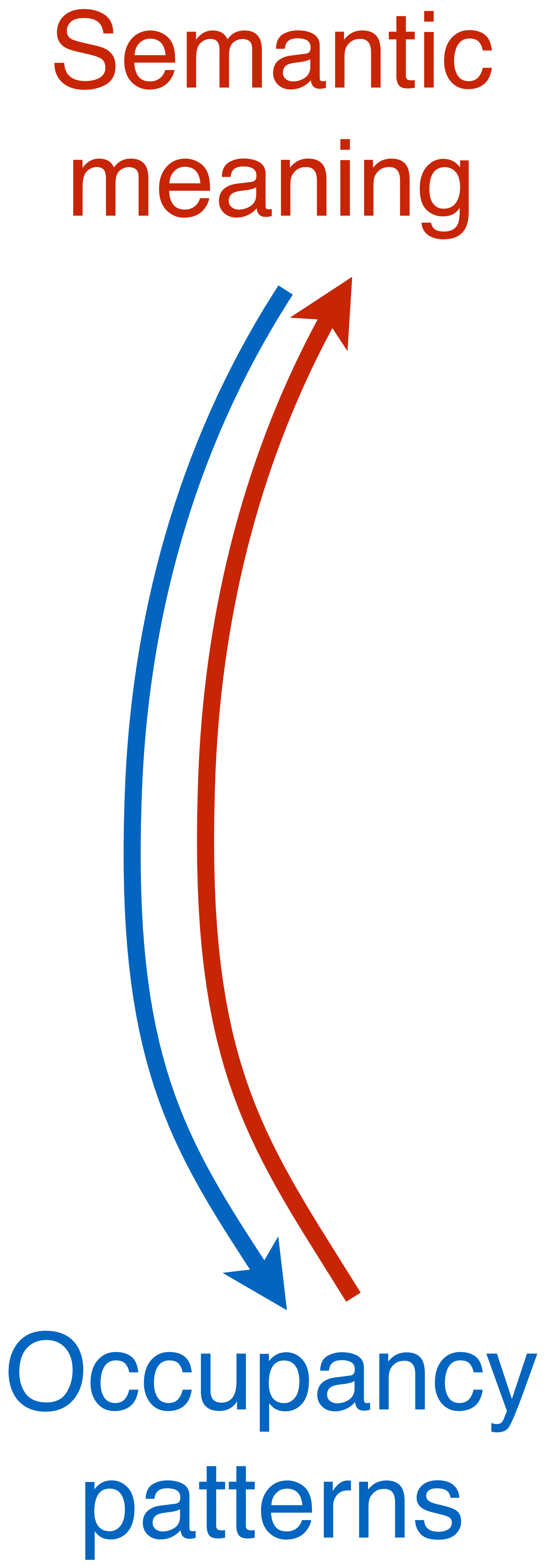
a TV?

a bed?

Object occupancy and semantic

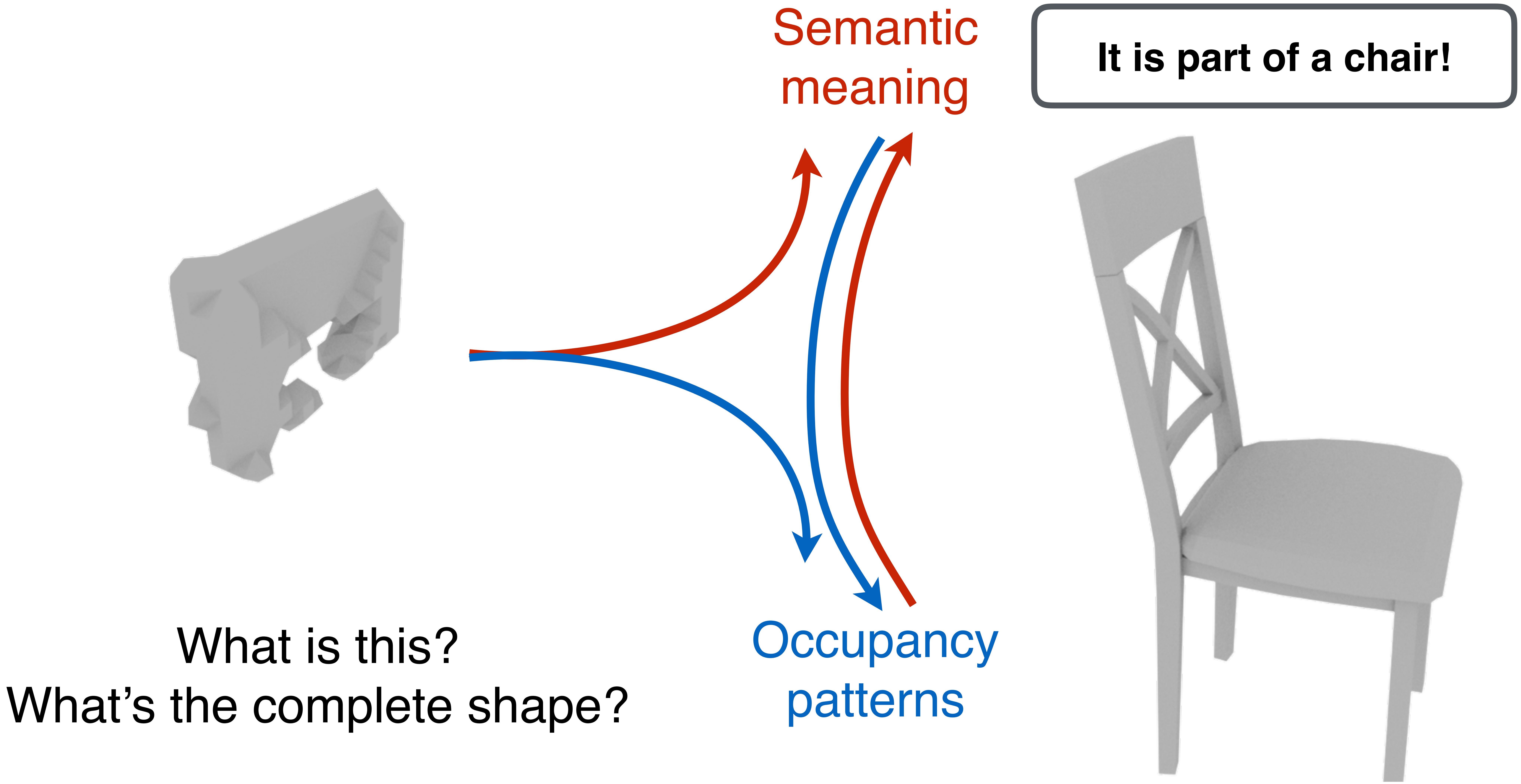


What is this?
What's the complete shape?

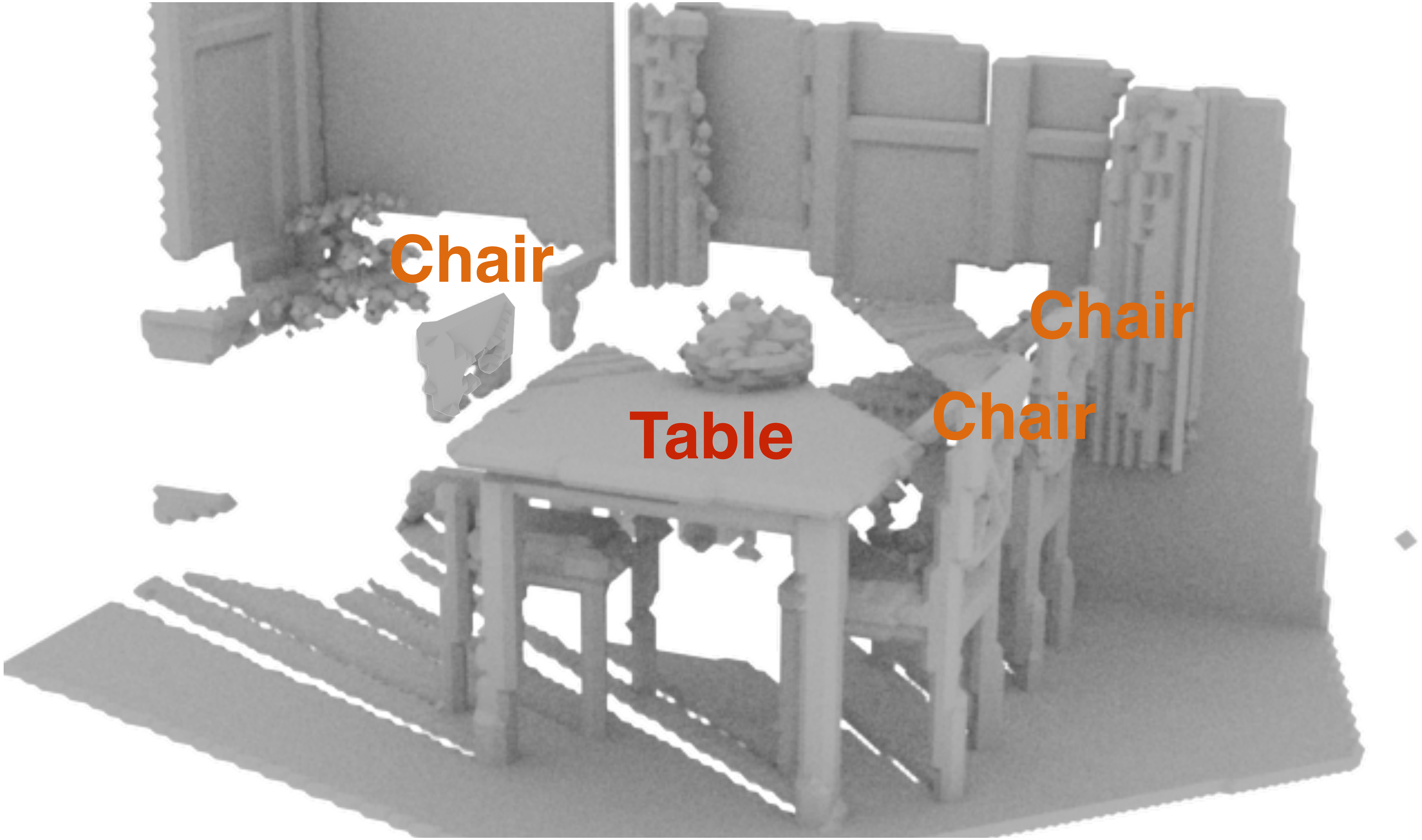


It is part of a chair!

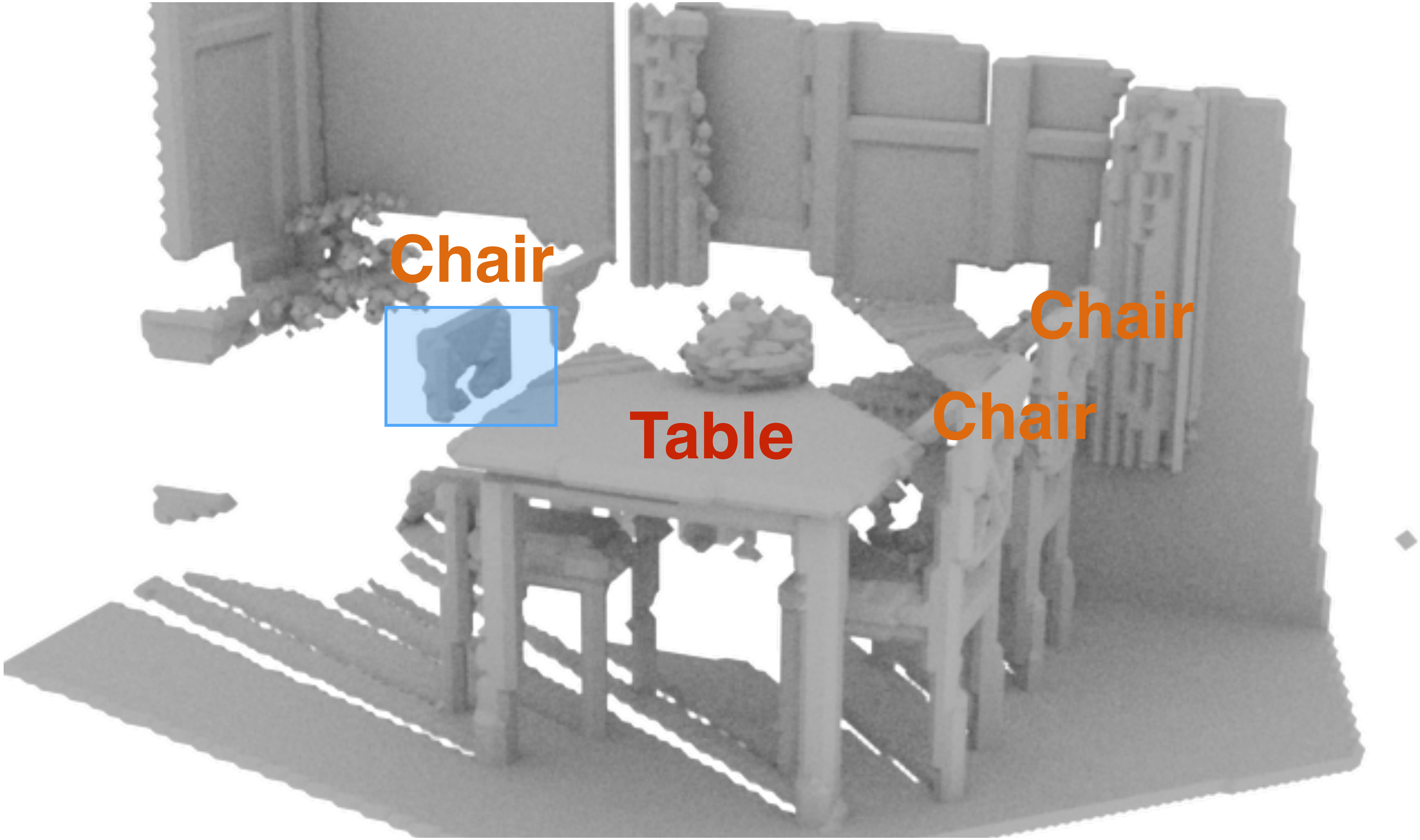
Object occupancy and semantic



3D context !

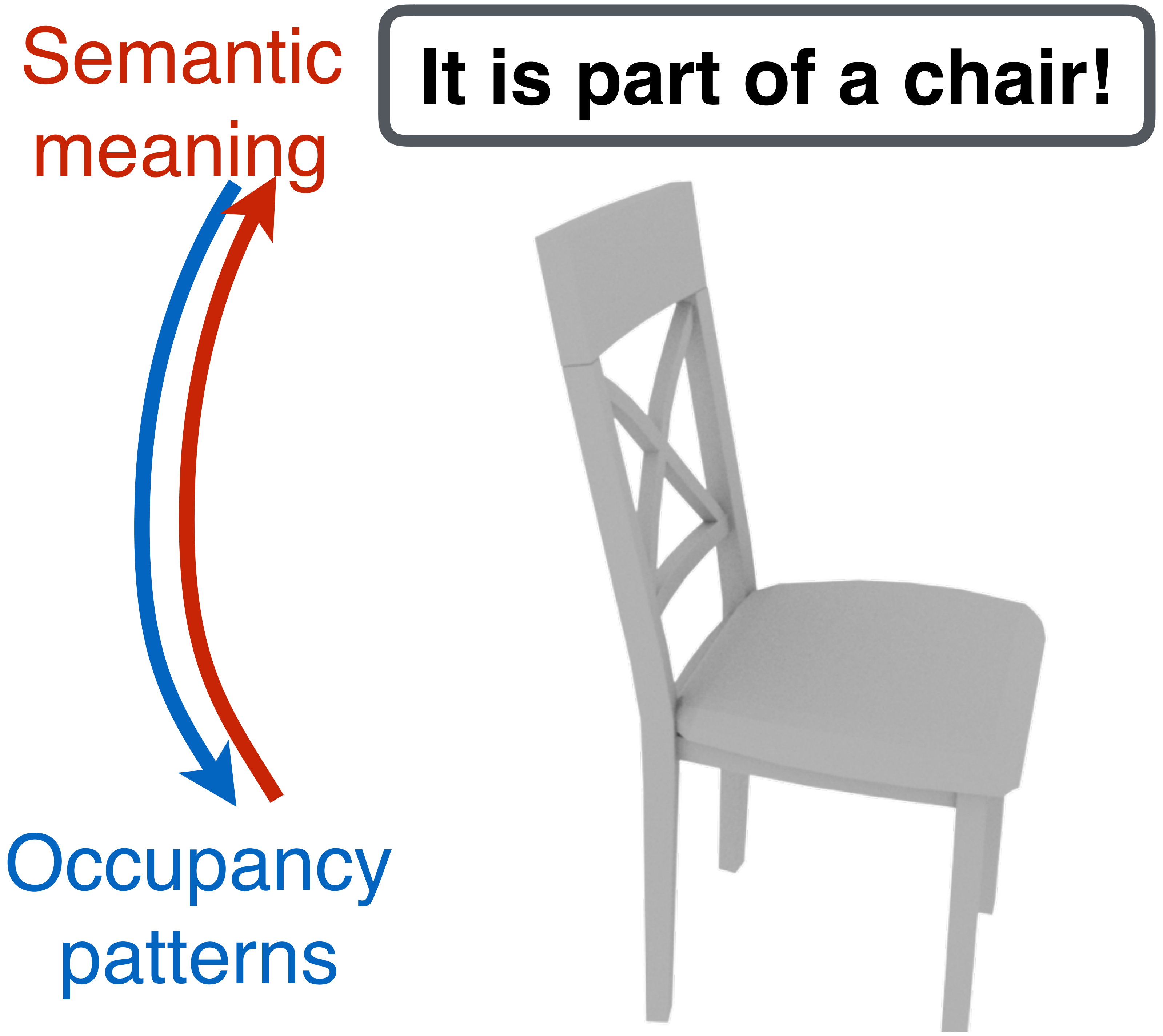


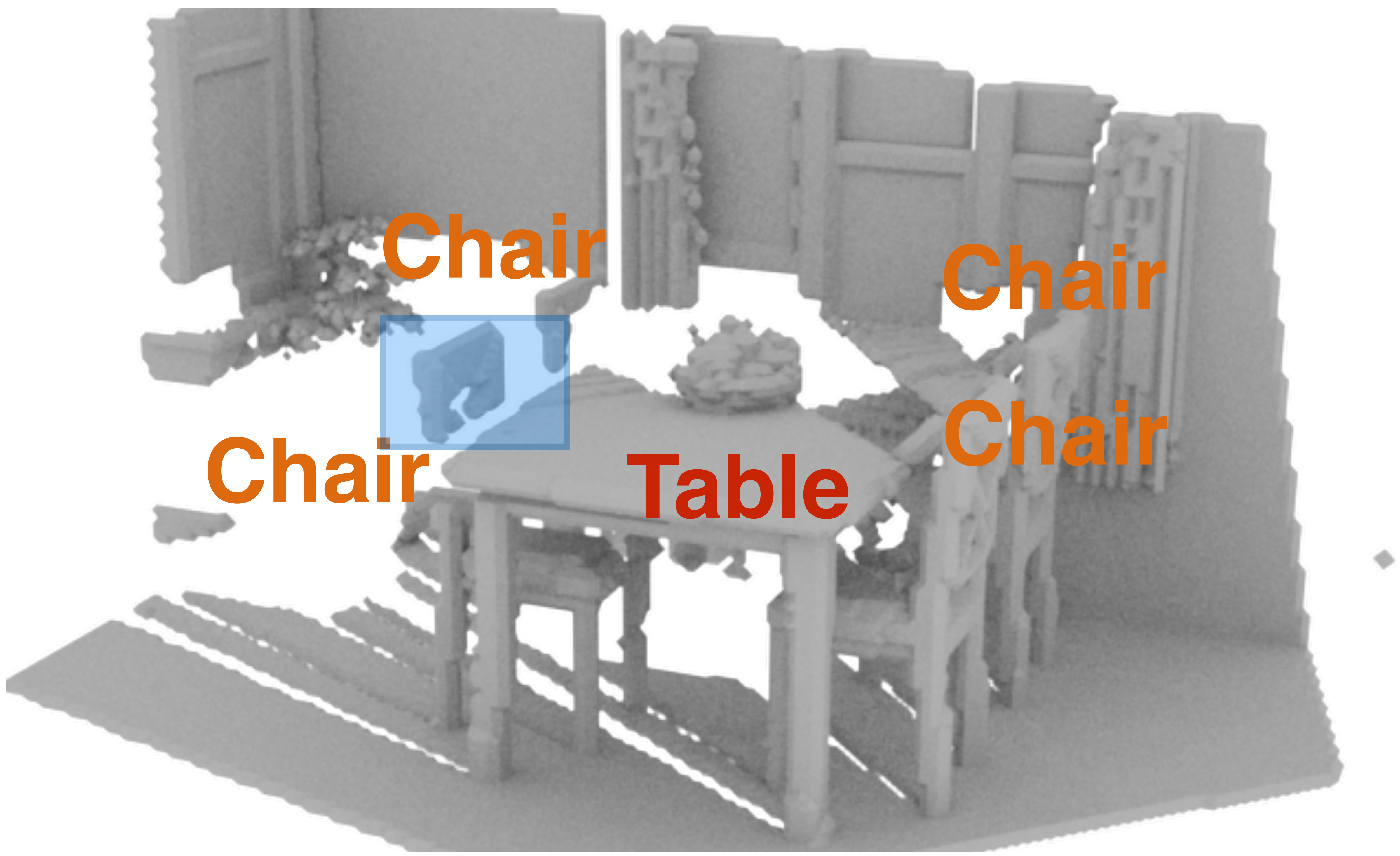
3D context with BIG receptive field!



Key ideas:

1. Object occupancy and the identity are tightly intertwined.
2. It is important to capture and understand 3D context with big receptive fields.



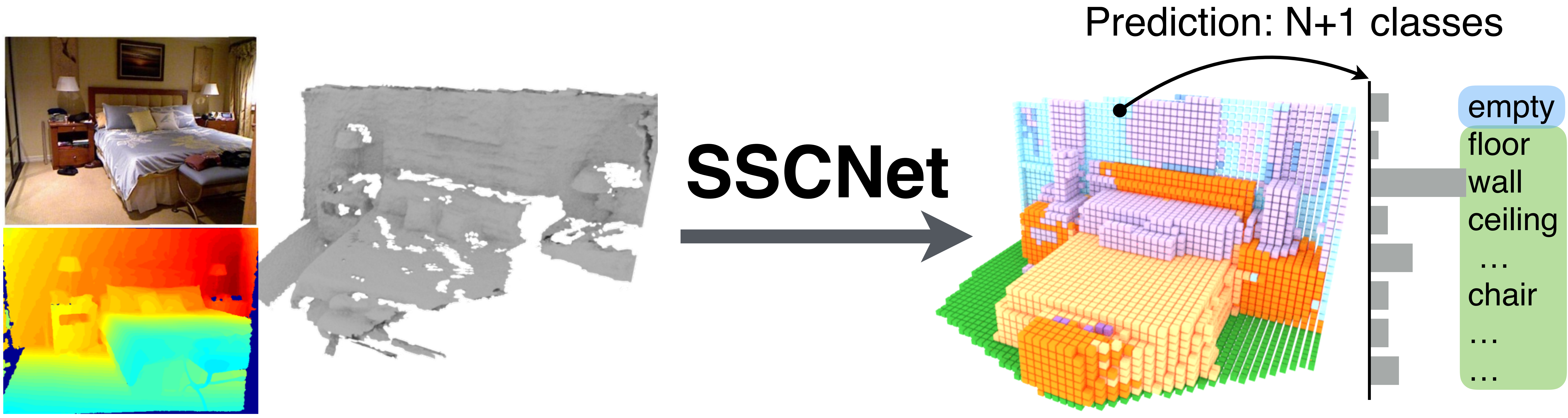


Semantic Scene Completion Network

Input: Single view depth map

Output: Semantic scene completion

Semantic Scene Completion Network

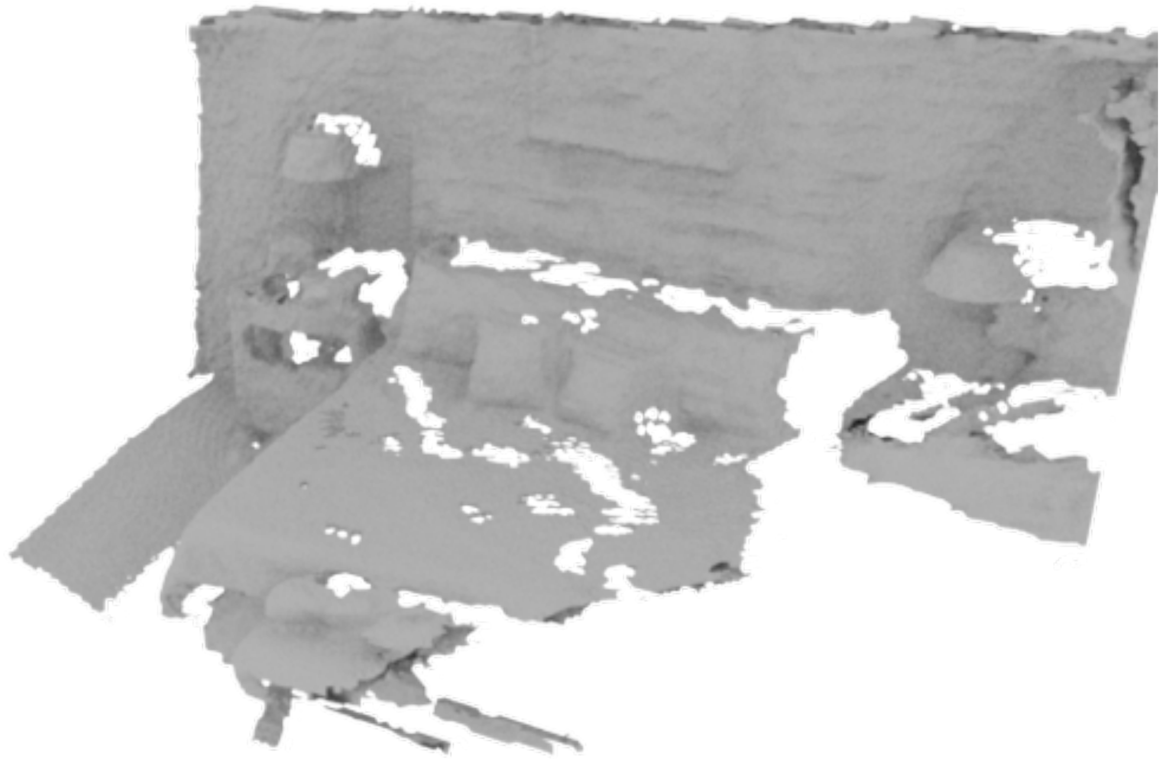


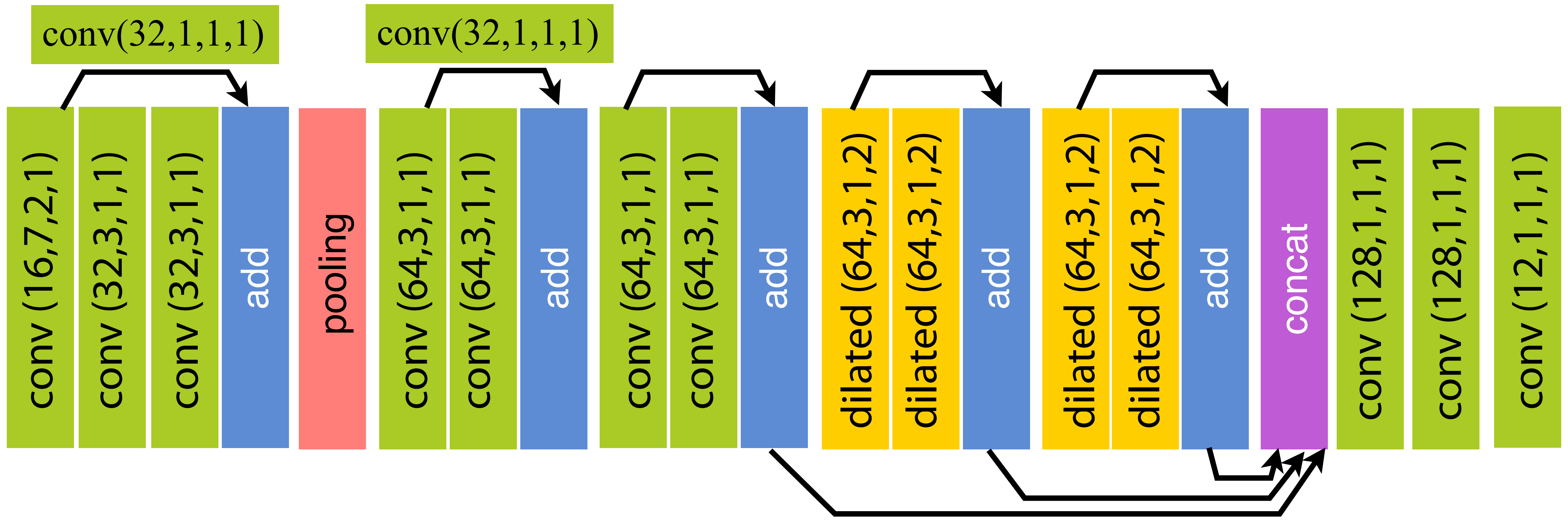
Input: Single view depth map

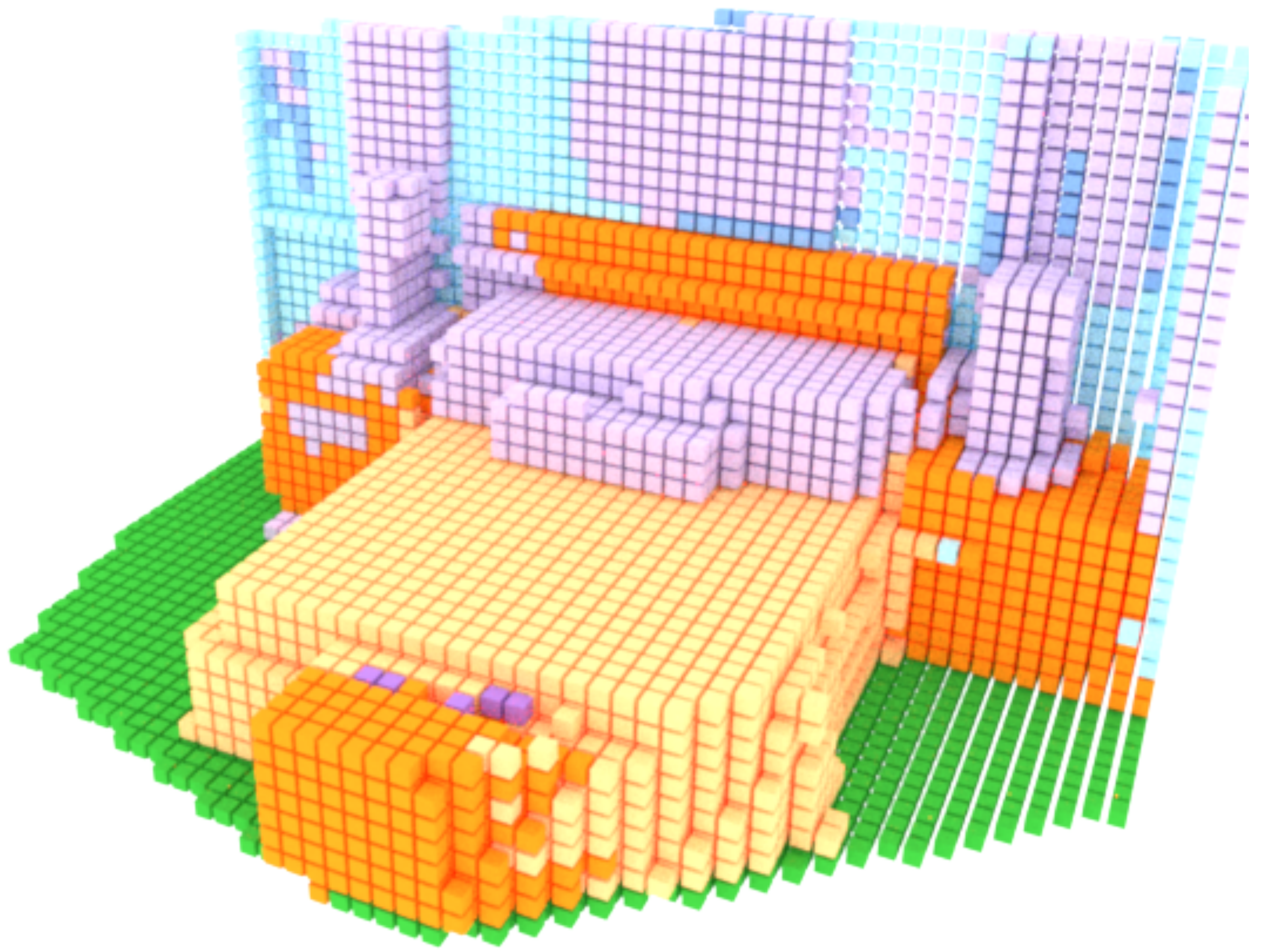
Output: Semantic scene completion

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

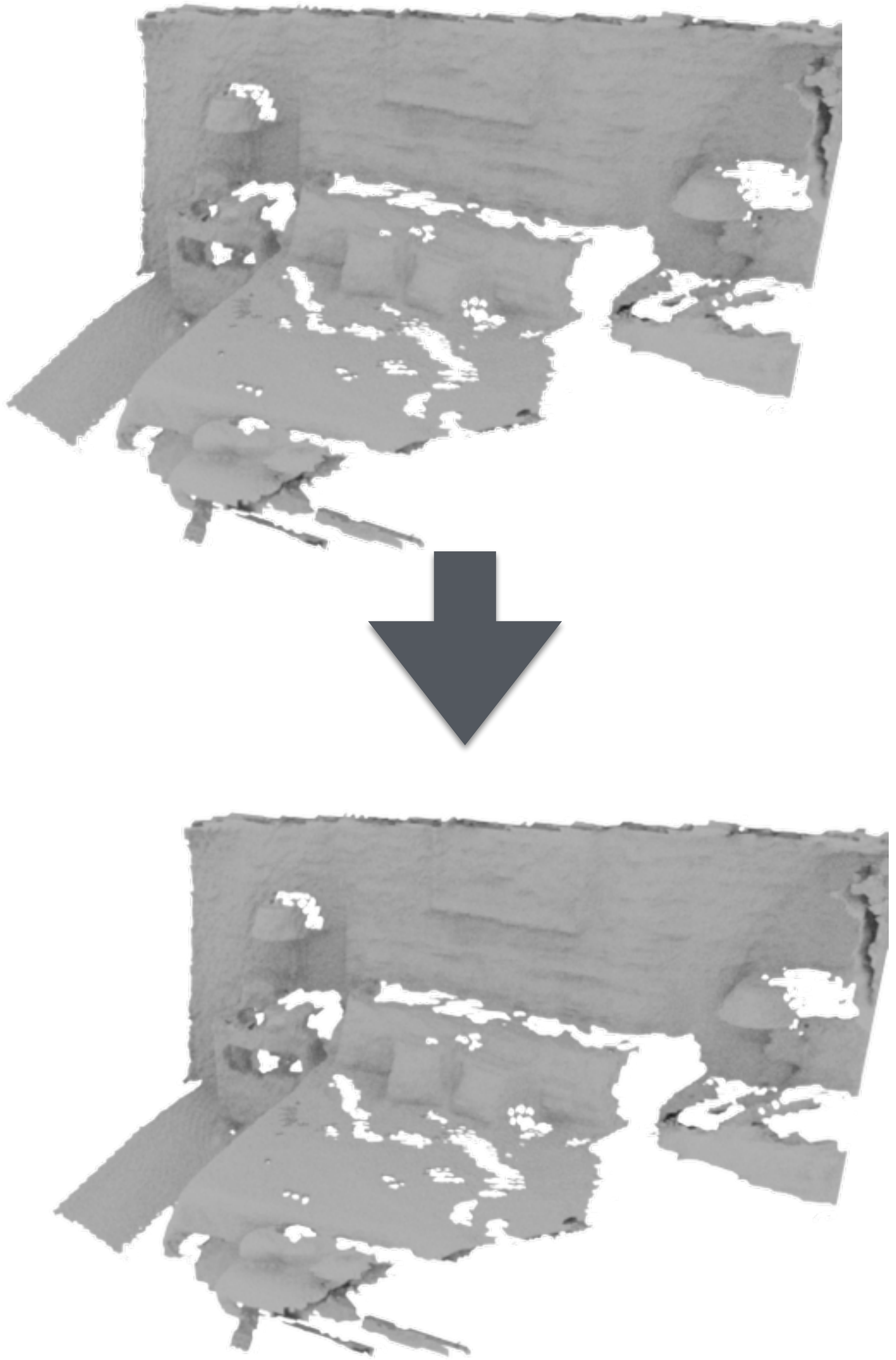
Semantic Scene Completion Network

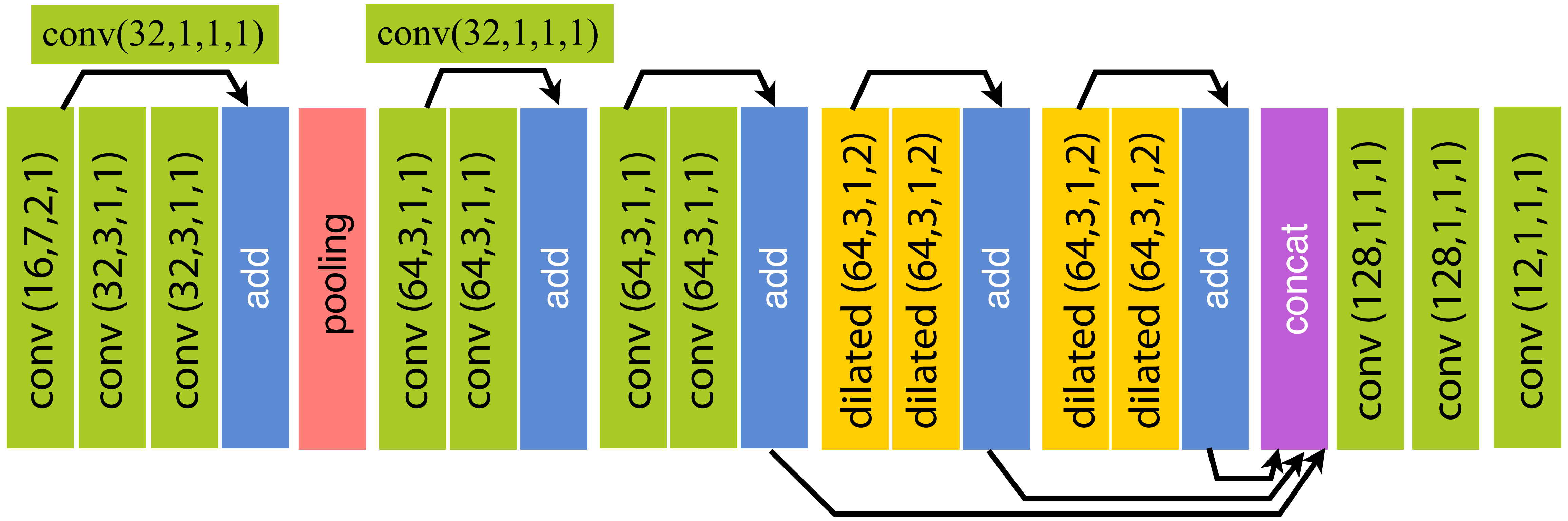


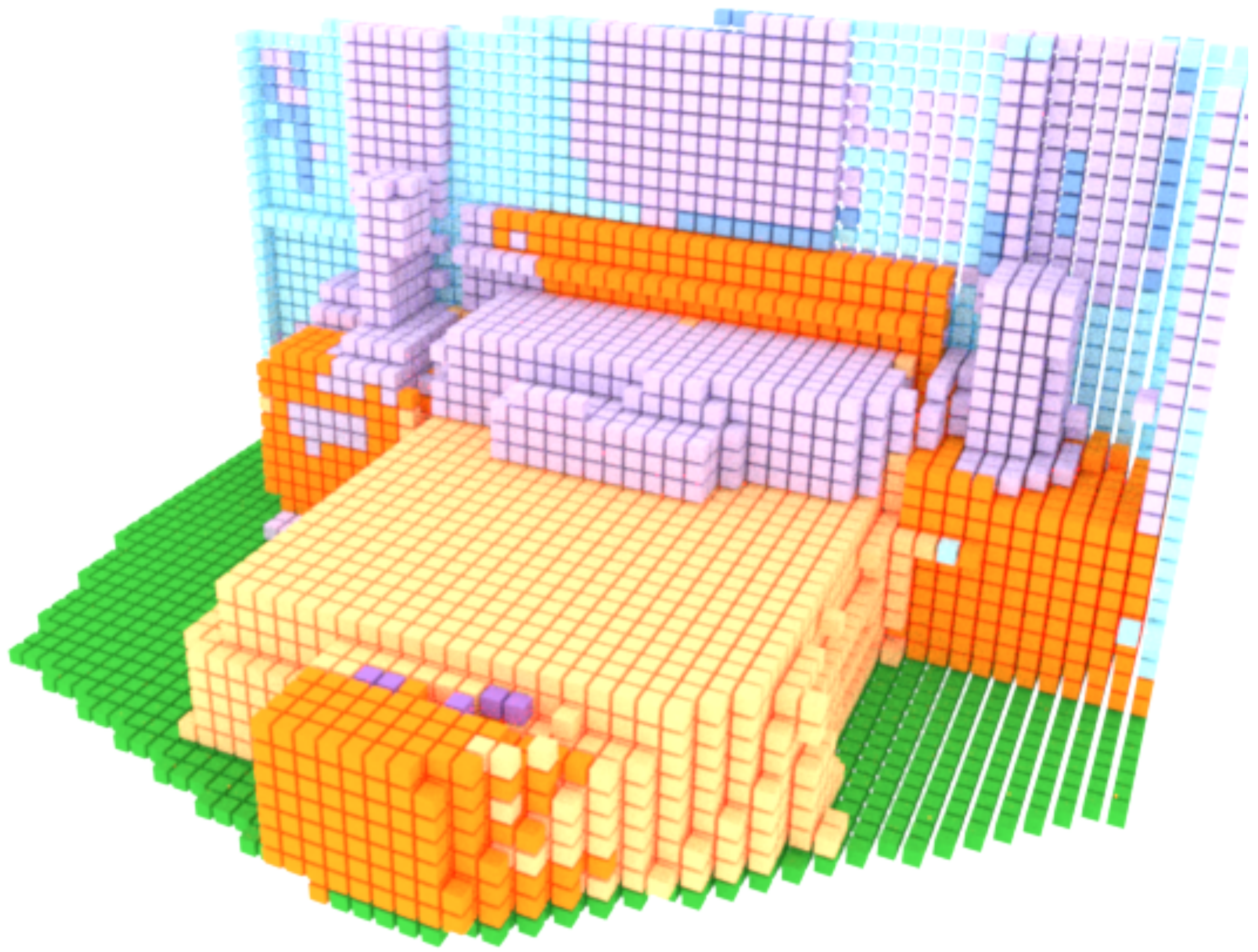




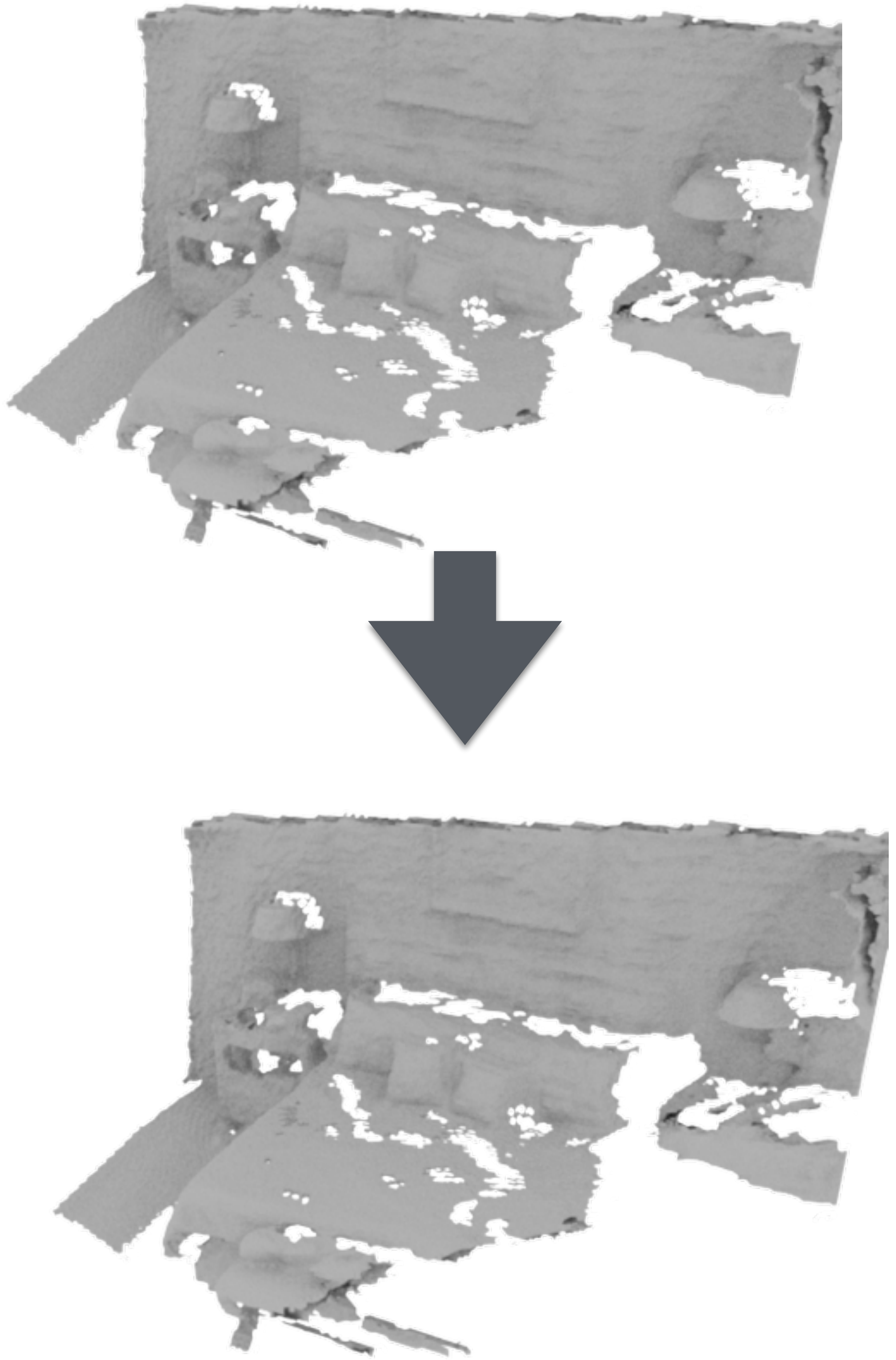
Semantic Scene Completion Network

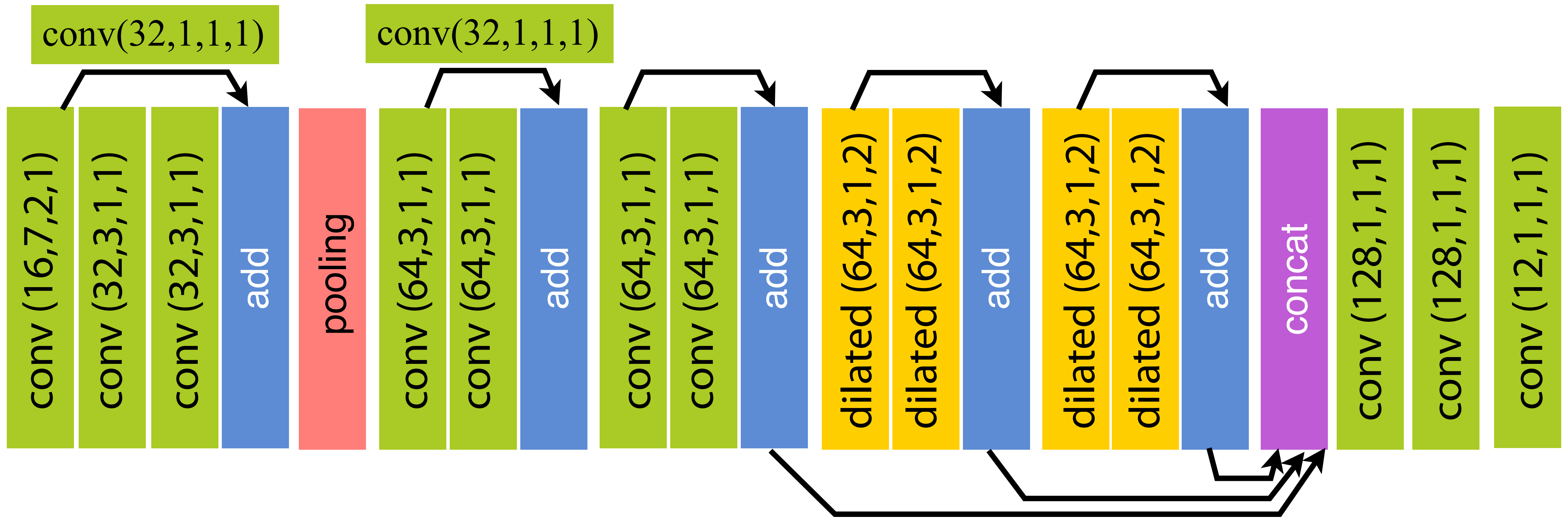


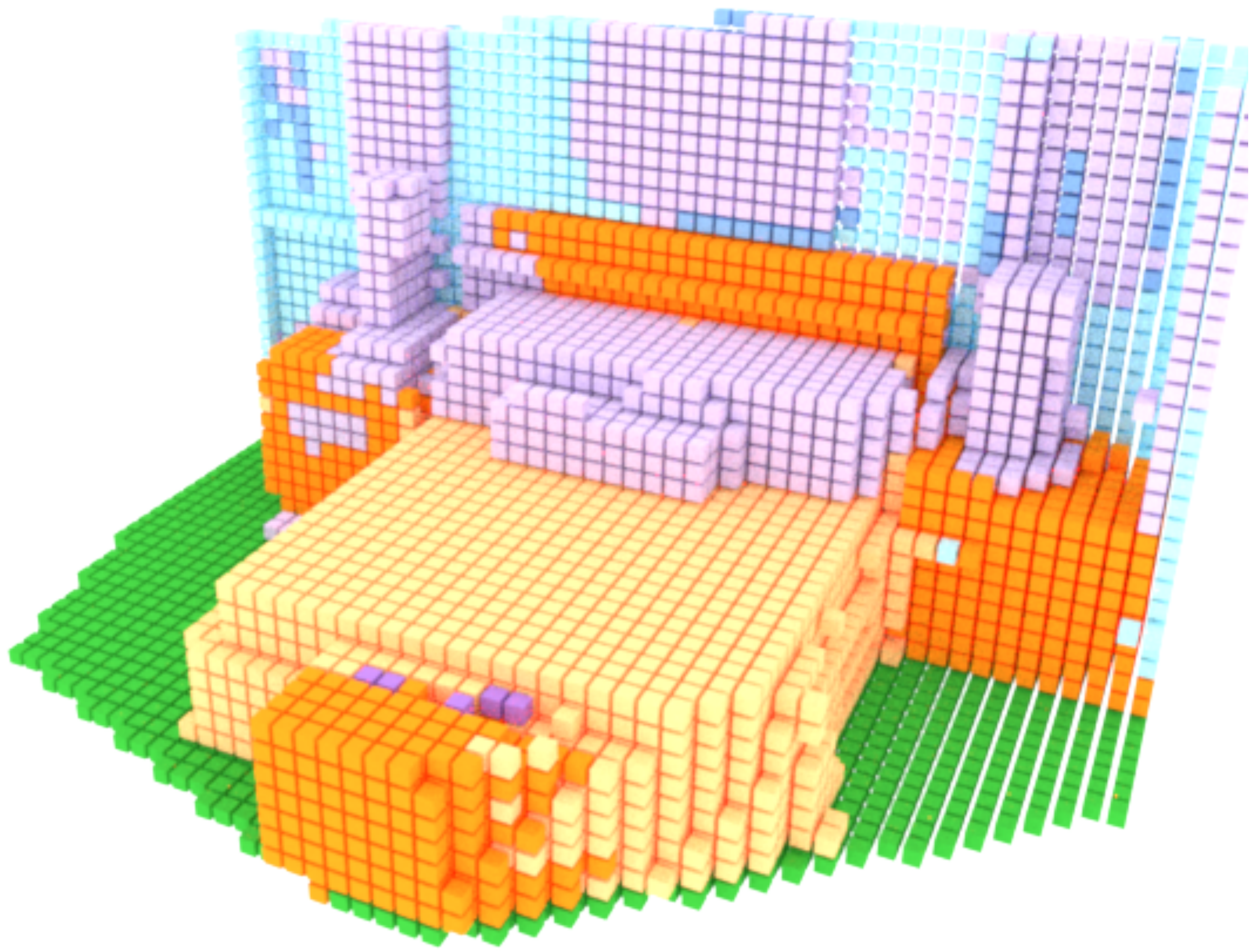




Semantic Scene Completion Network

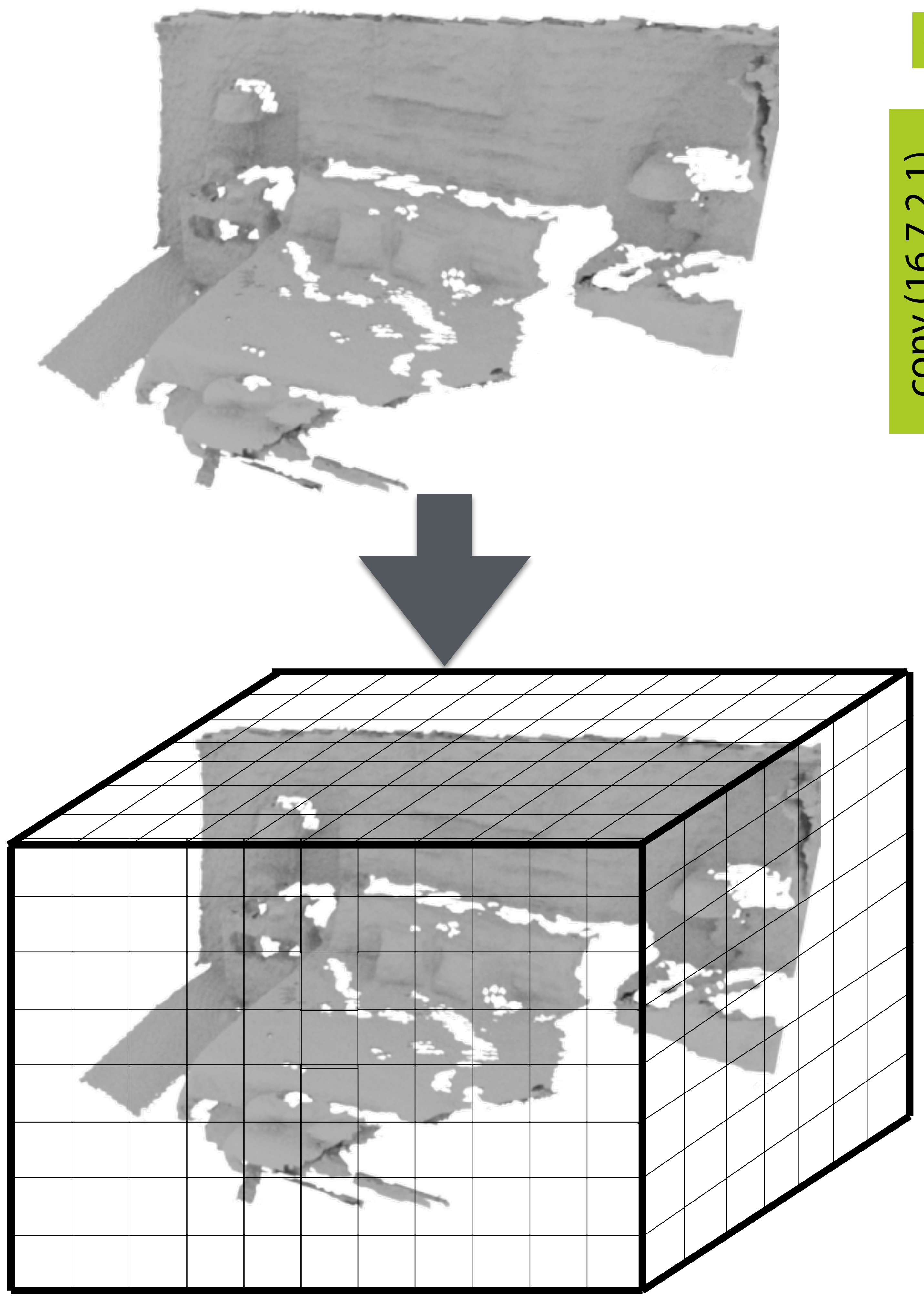


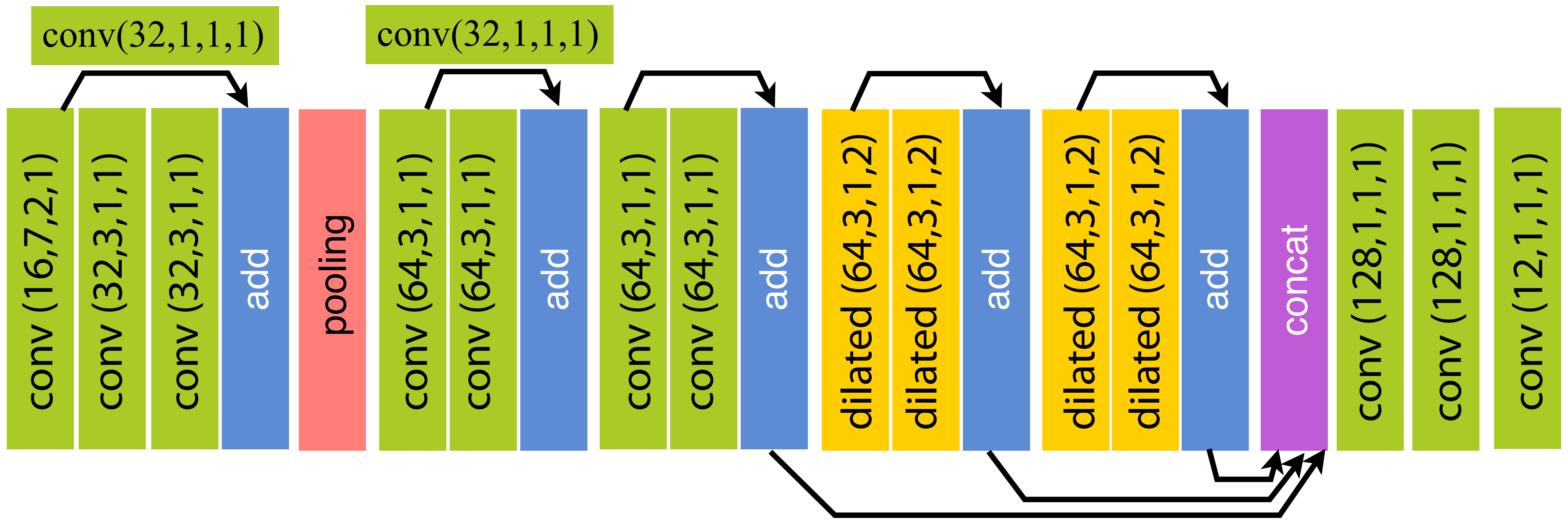


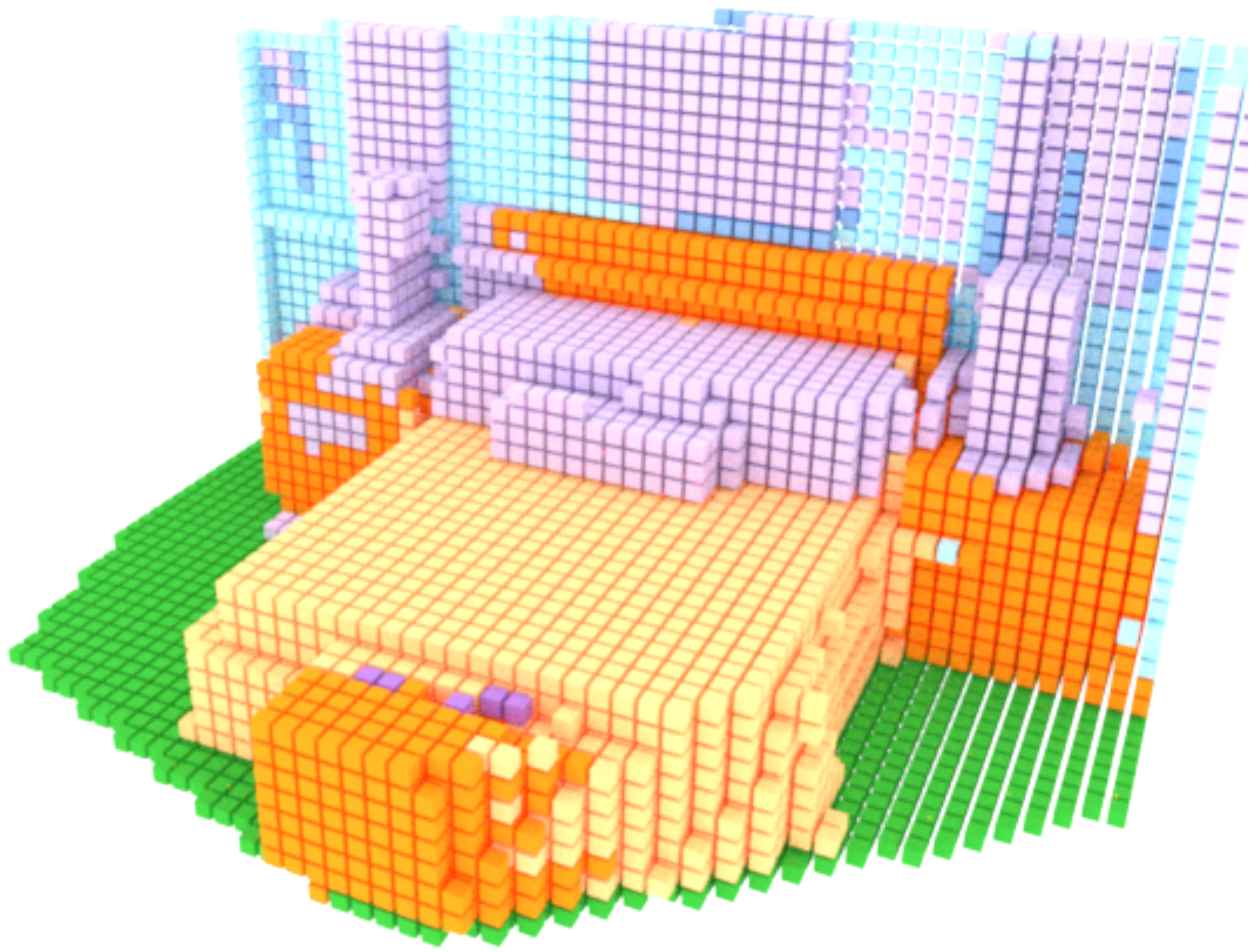


Encode 3D space using flipped TSDF

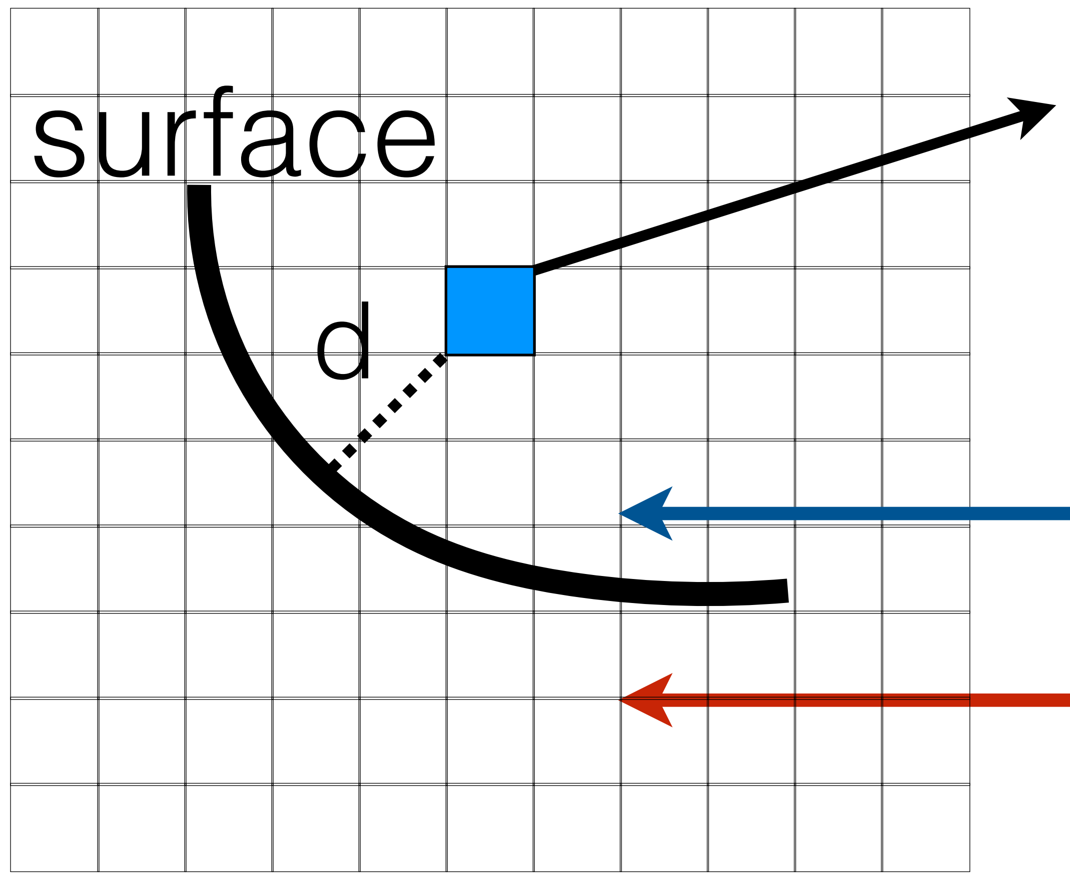
Semantic Scene Completion Network







Encode 3D space using flipped TSDF



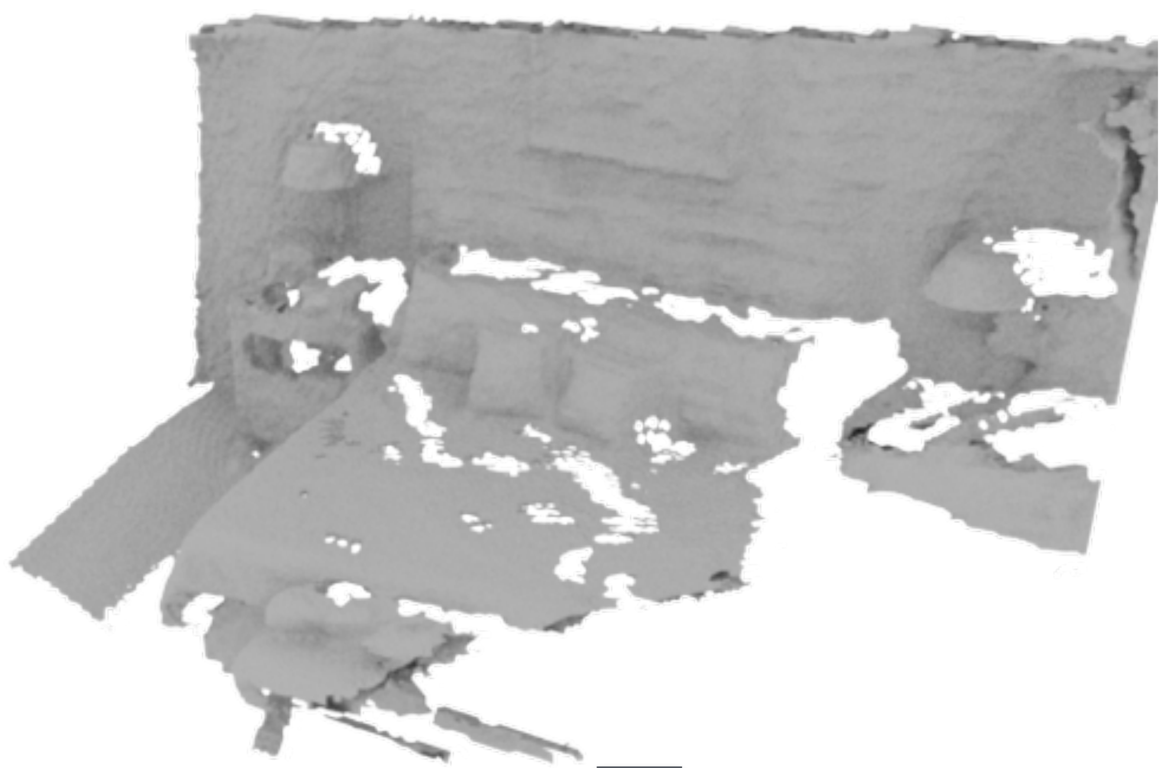
d = distance to the surface

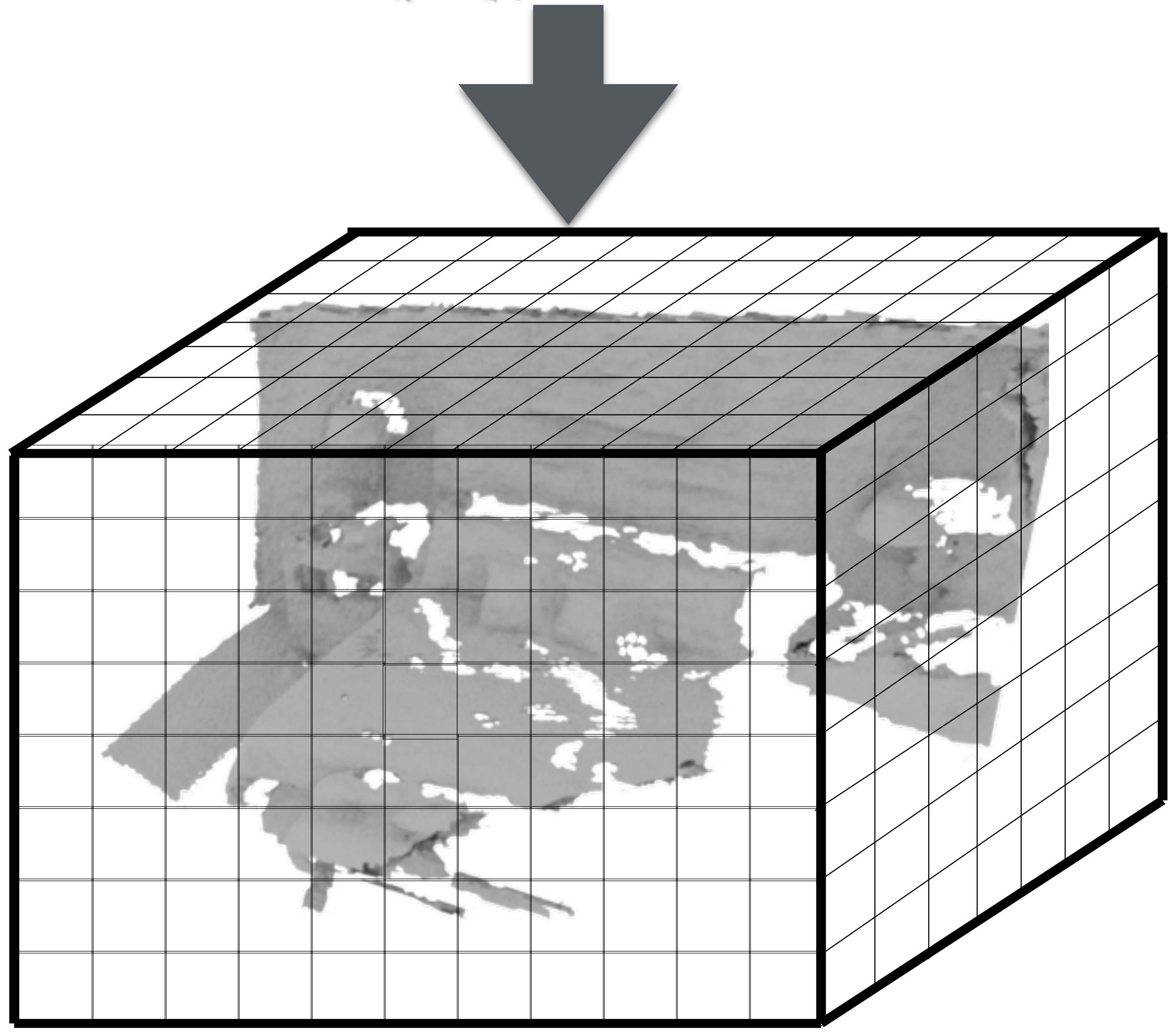
flipped TSDF = sign(1- min(1, d/dmax))

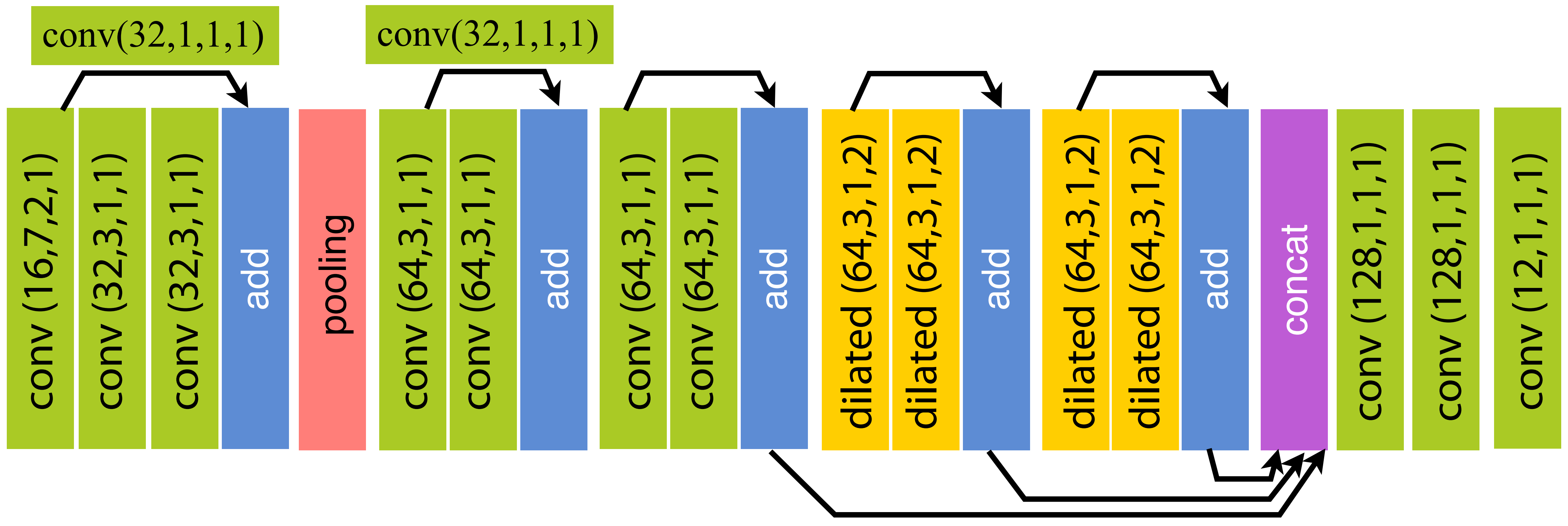
Occluded

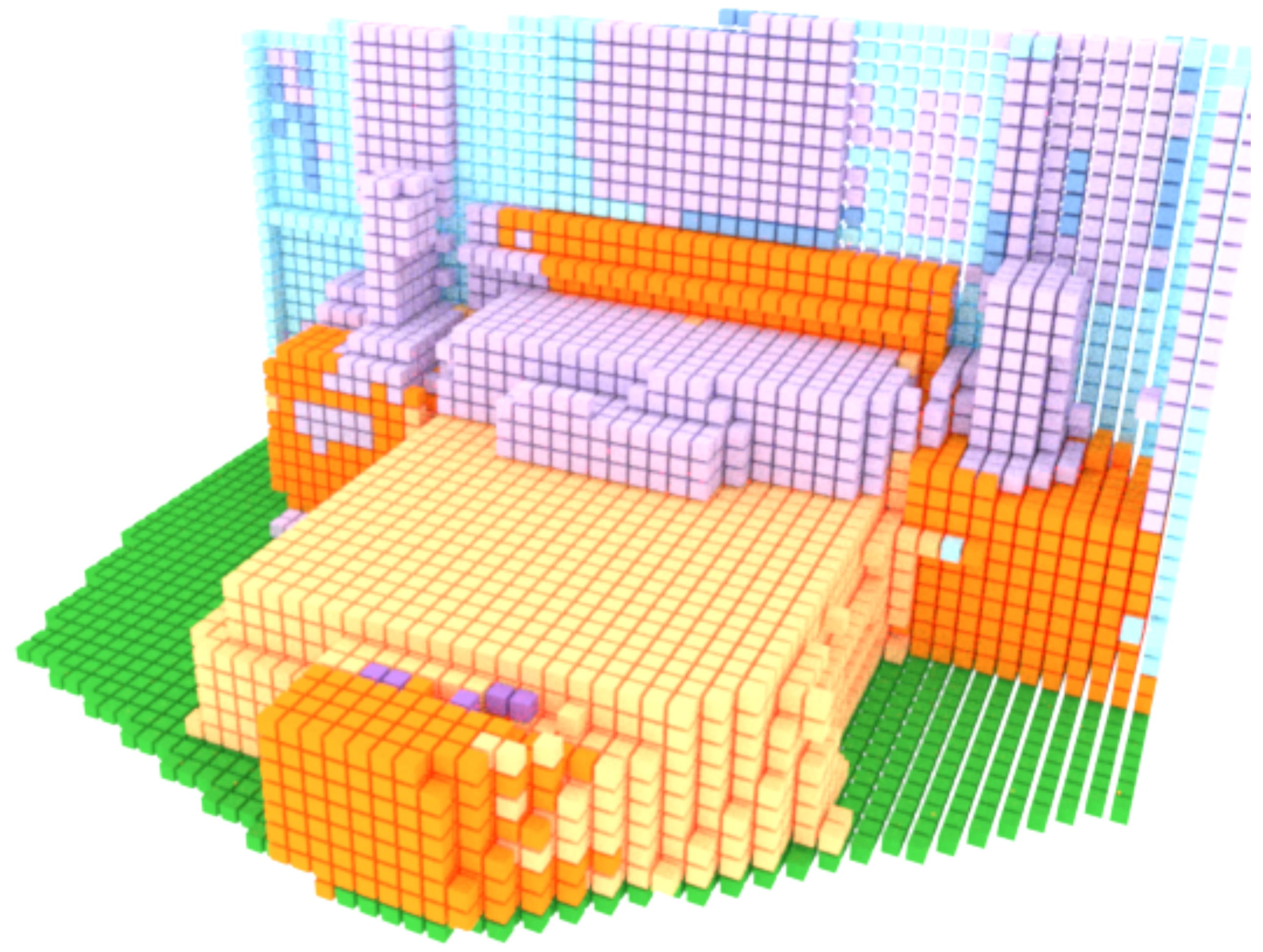
Free space

Semantic Scene Completion Network





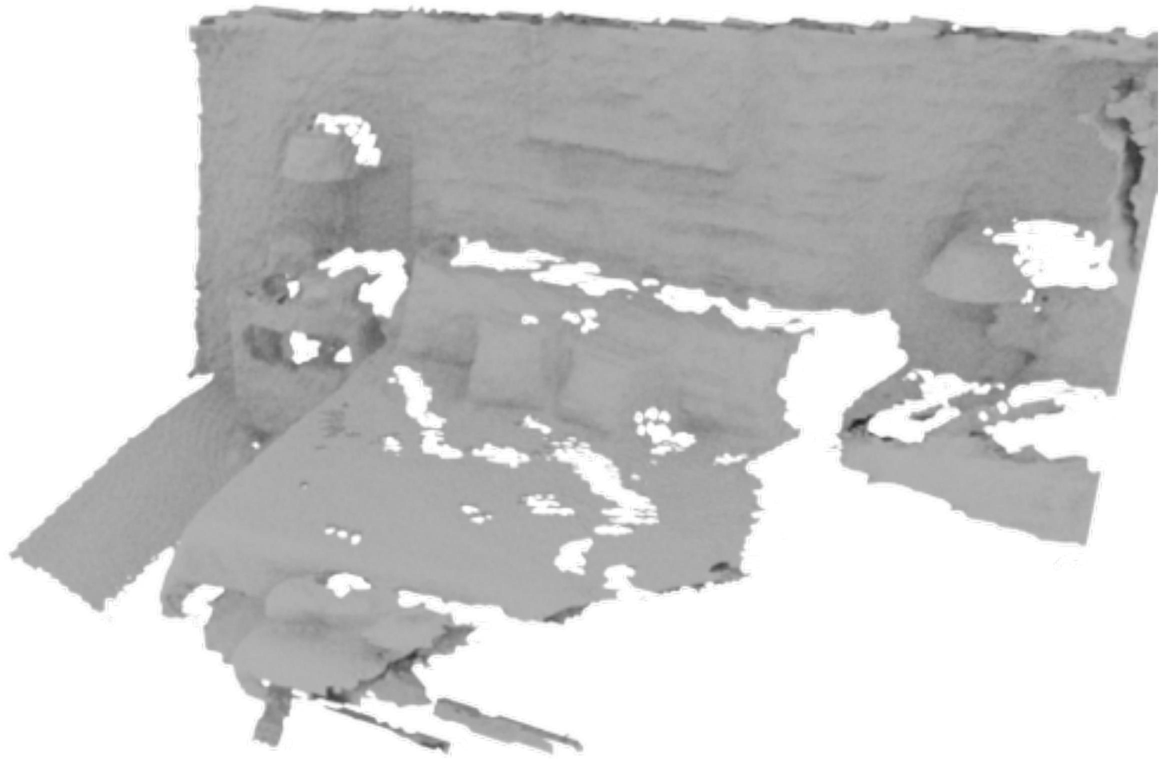


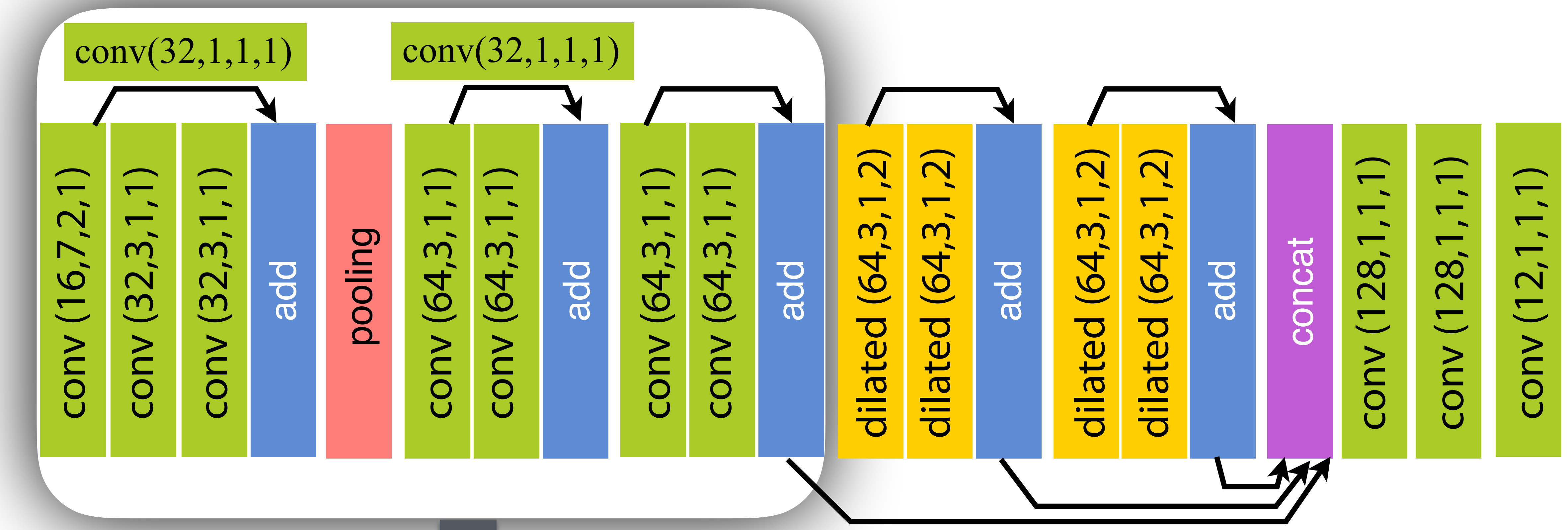


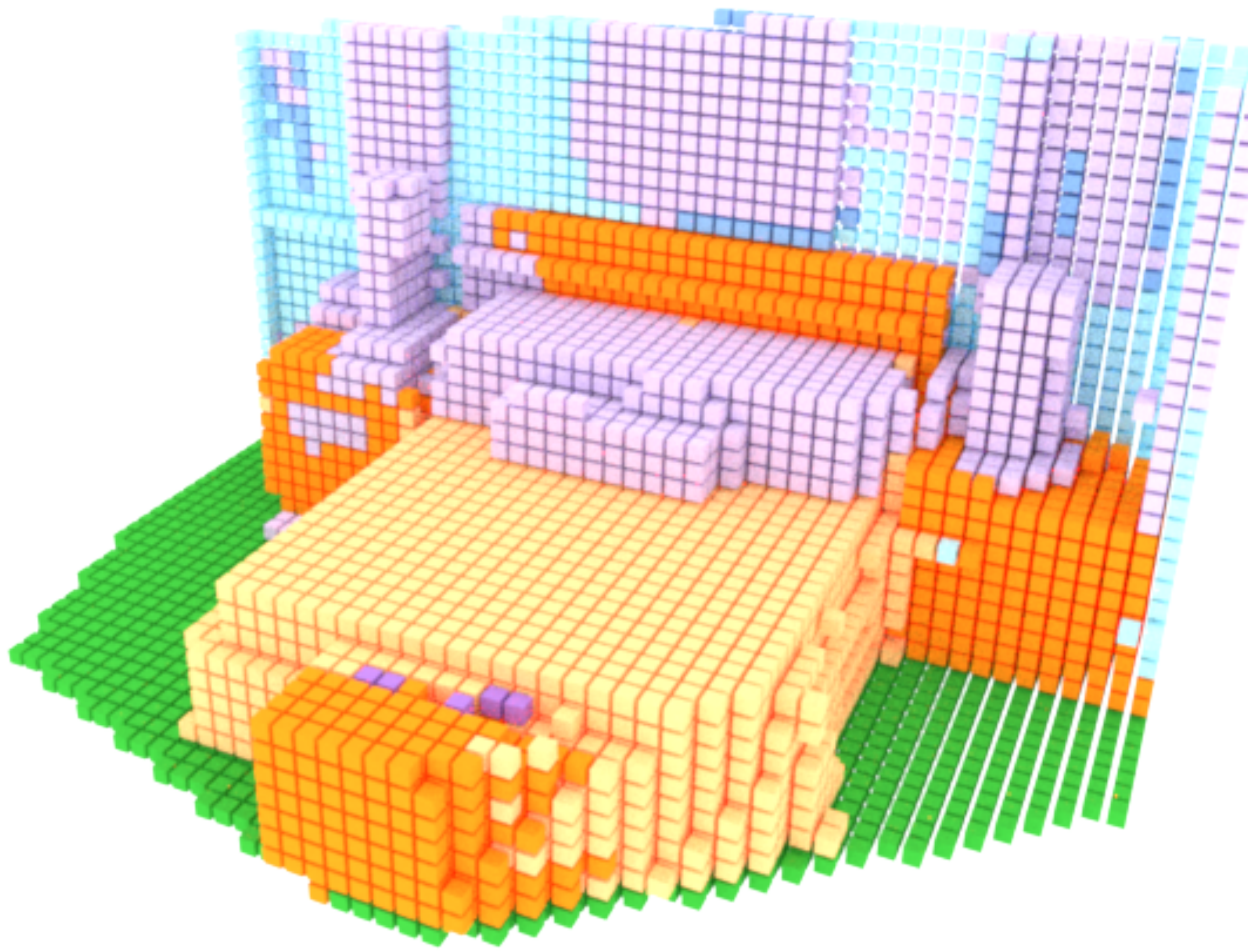
Compare to standard projective TSDF, flipTSDF:

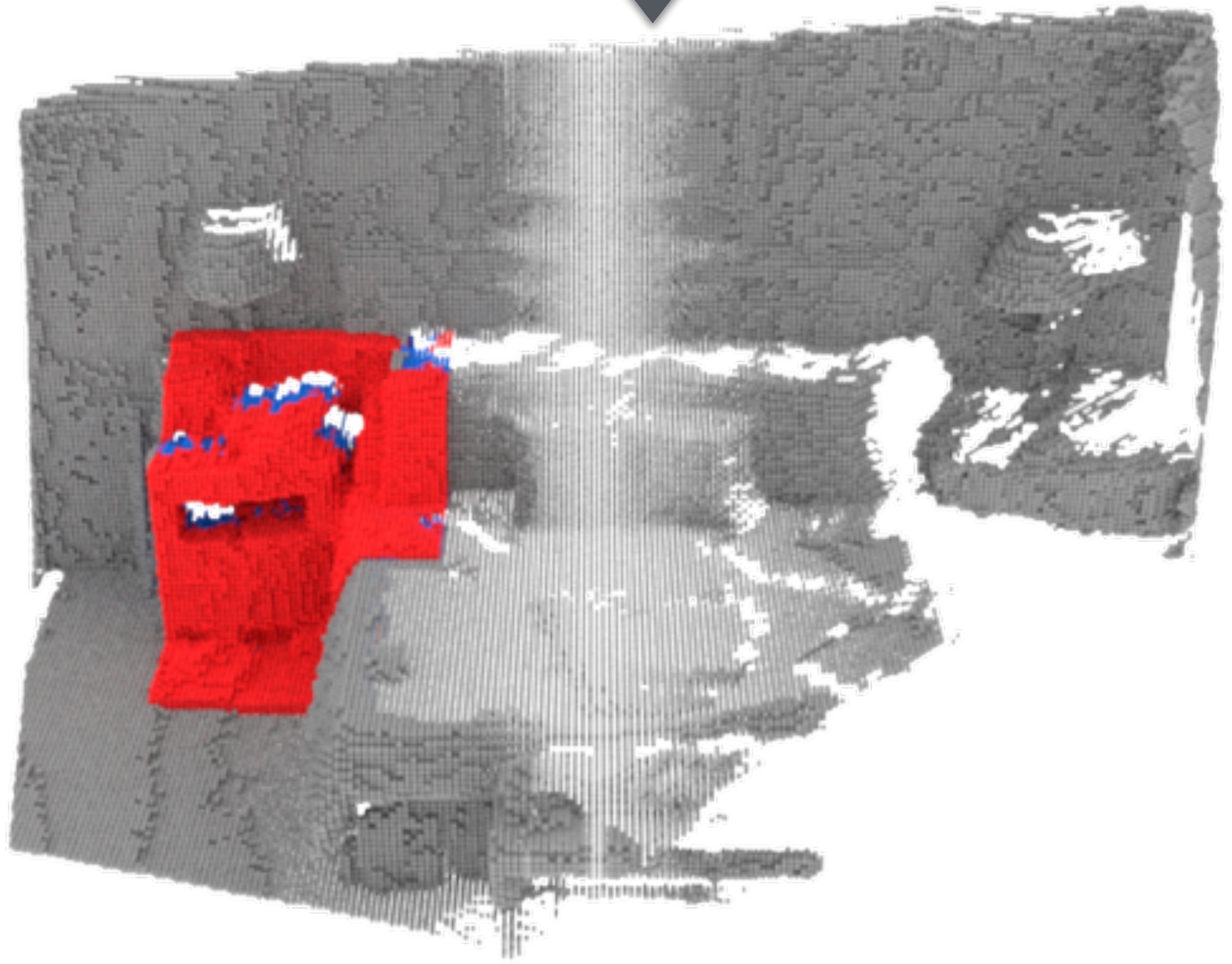
- has less viewpoint dependency
- concentrates the strongest gradient near surface

Semantic Scene Completion Network





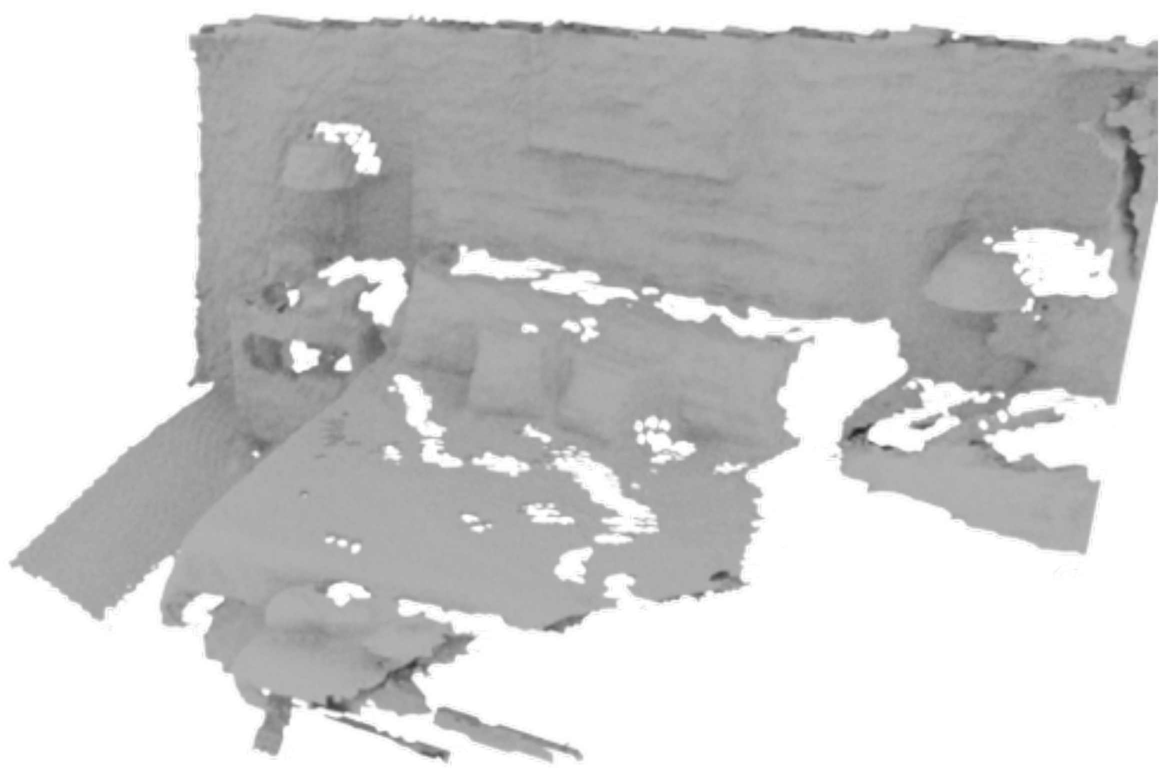


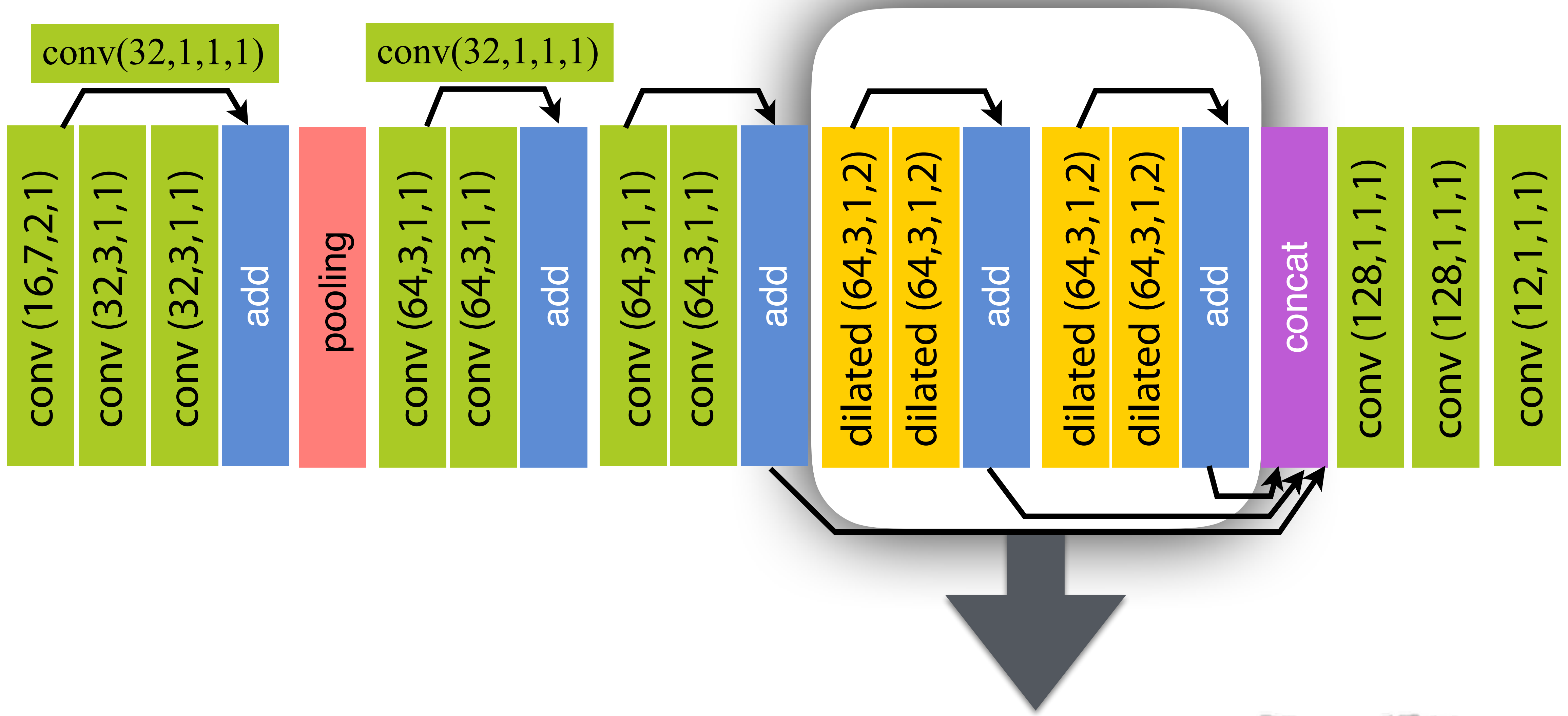


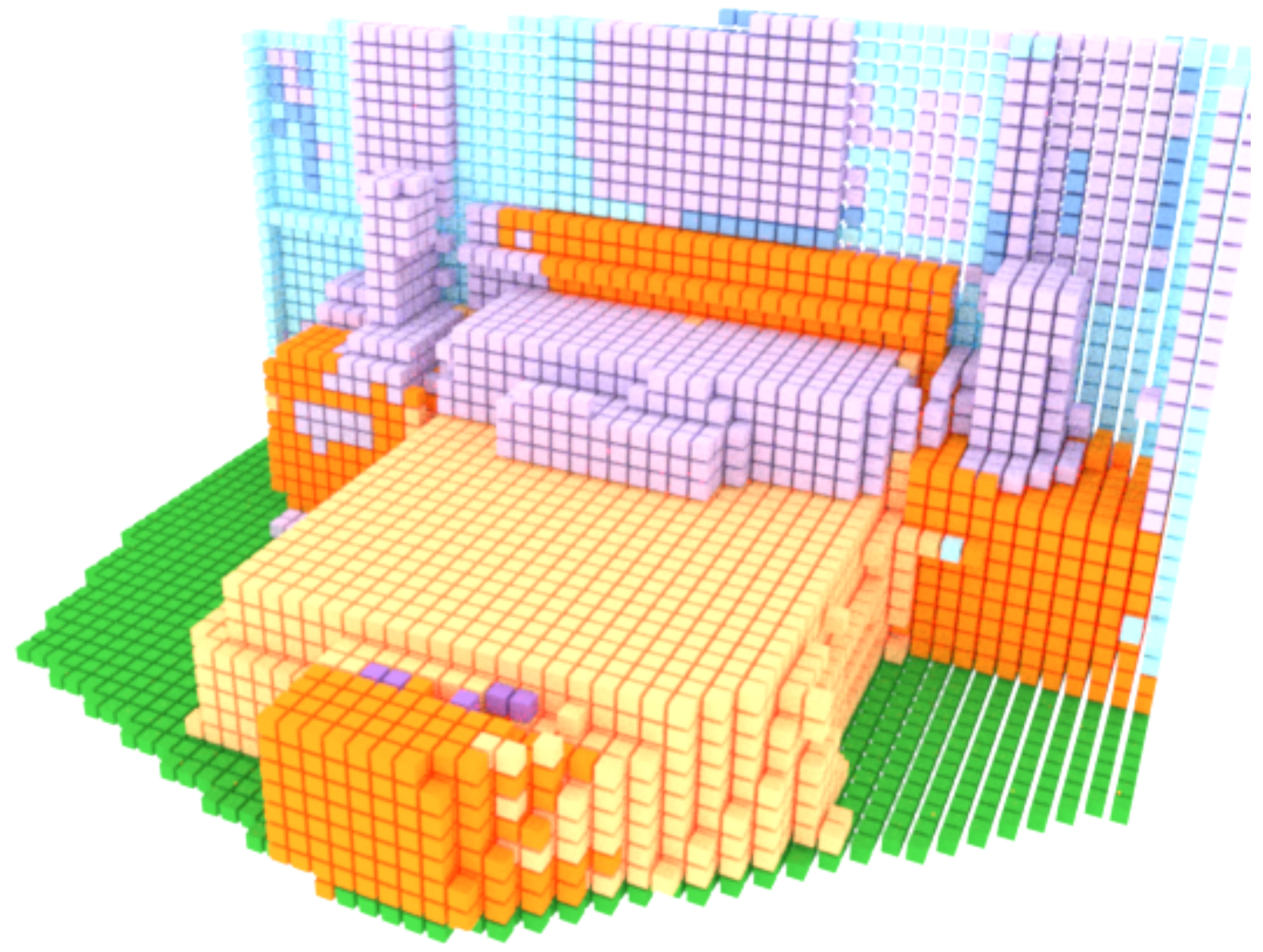
Local geometry

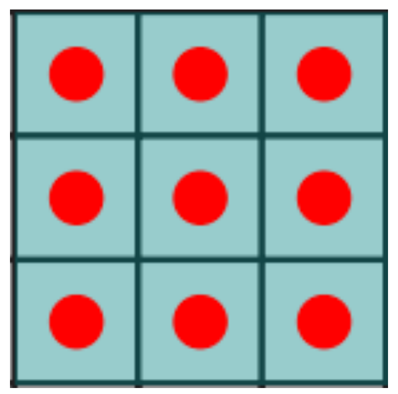
Receptive field: 0.98 m

Semantic Scene Completion Network

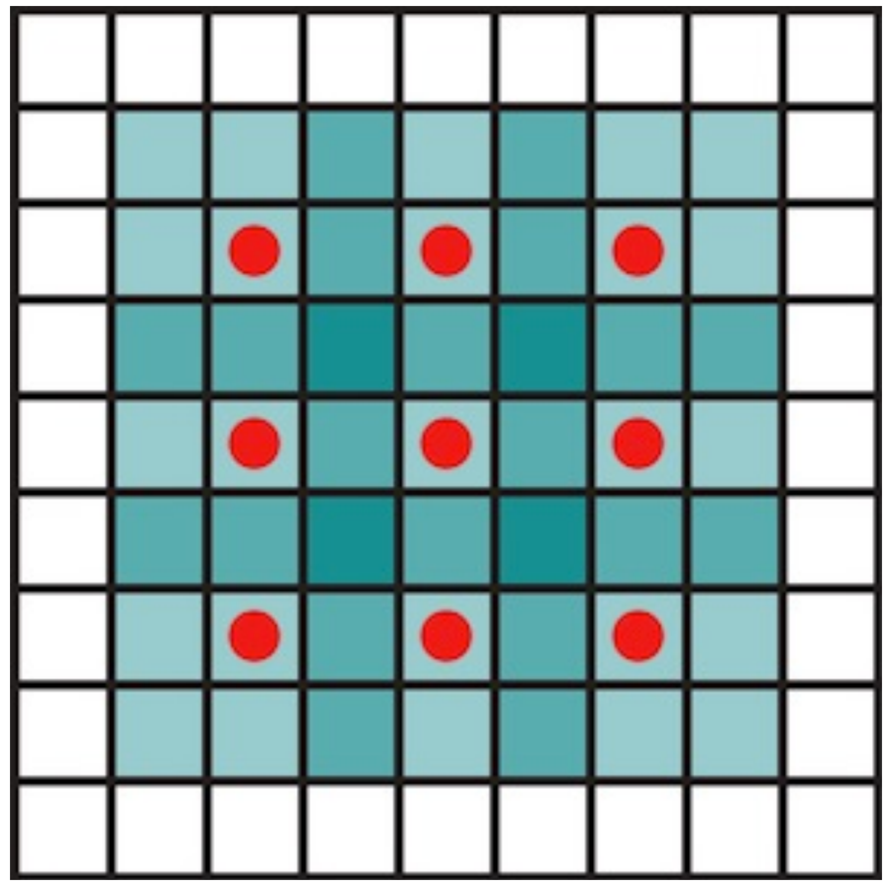




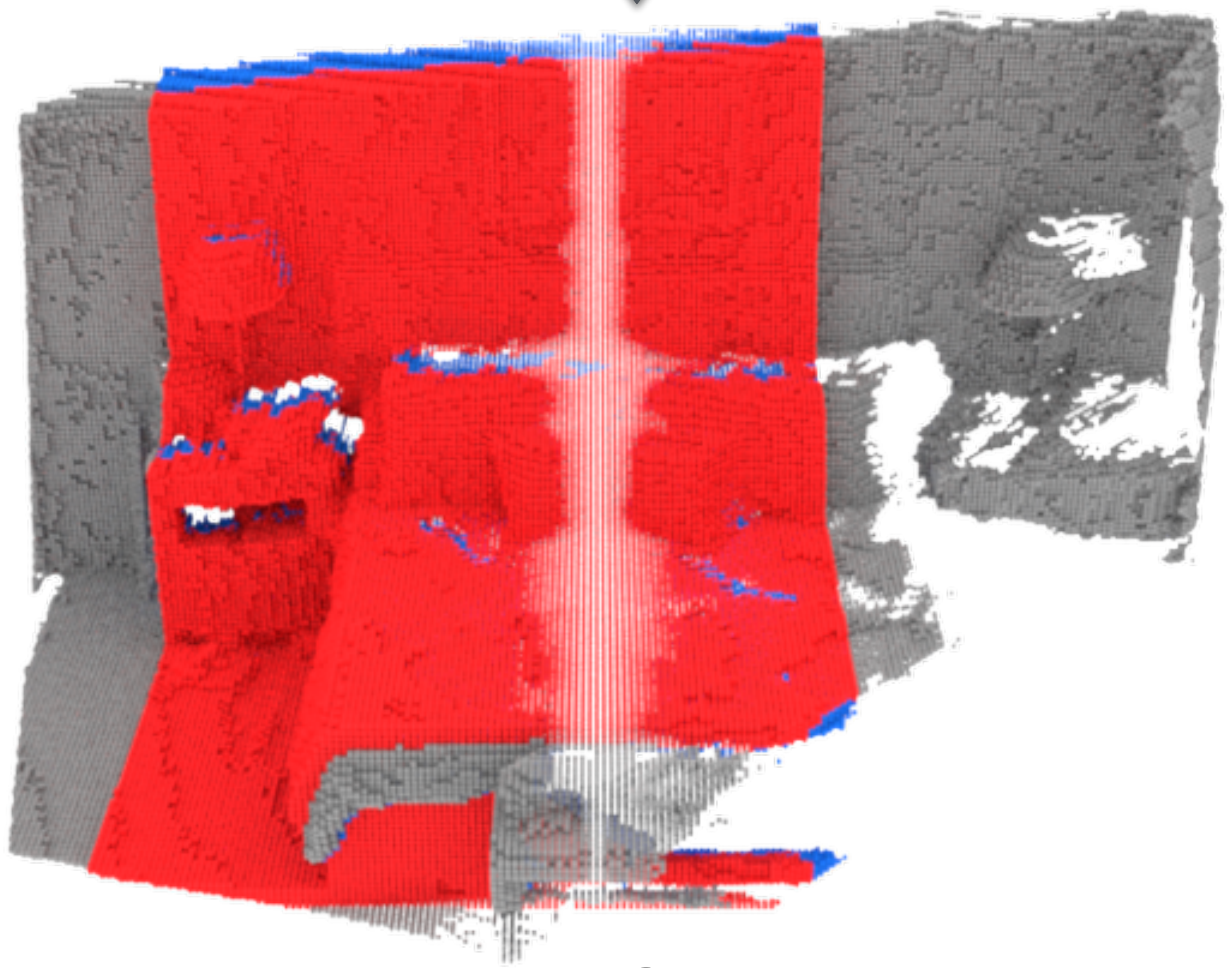




Normal kernel



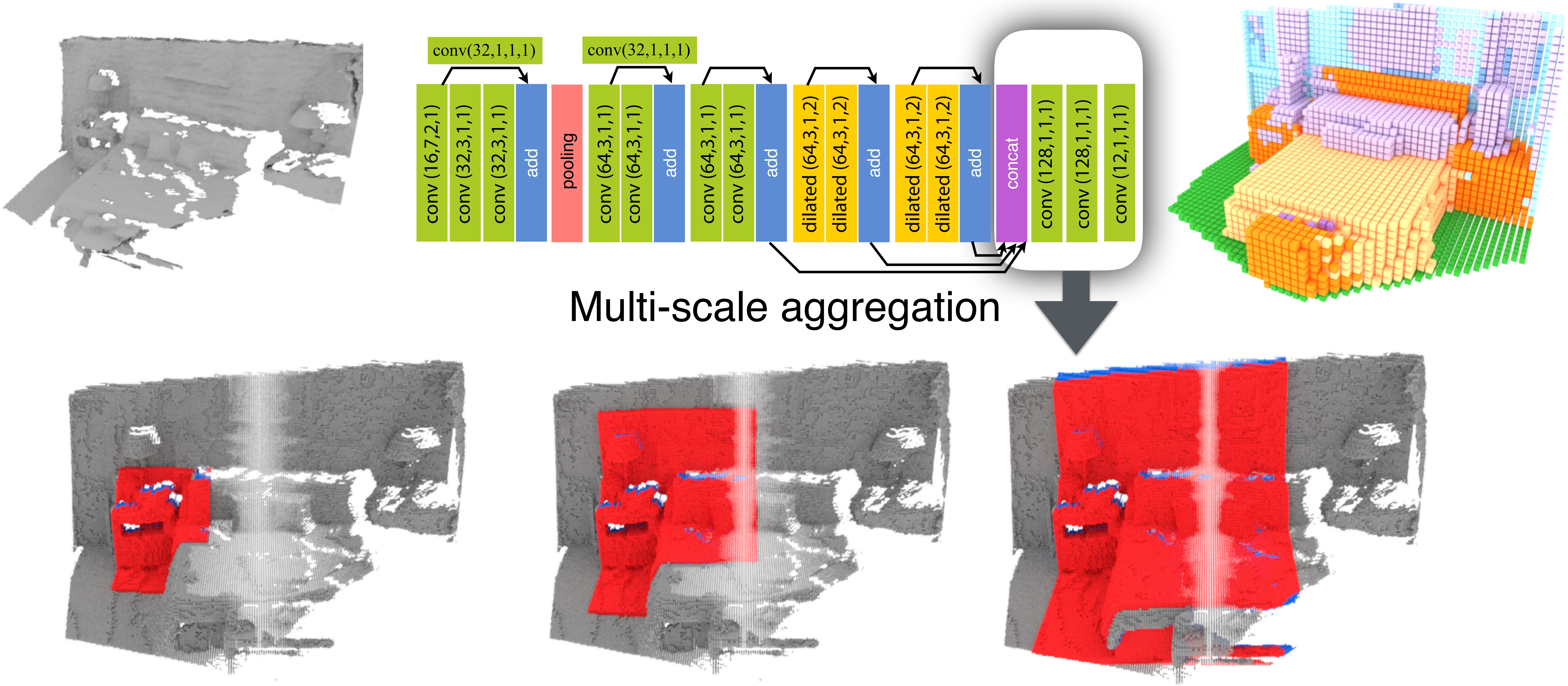
Dilation kernel



Receptive field: 2.26

Capturing higher-level 3D context by big receptive field

Semantic Scene Completion Network



Receptive field: 0.98 m

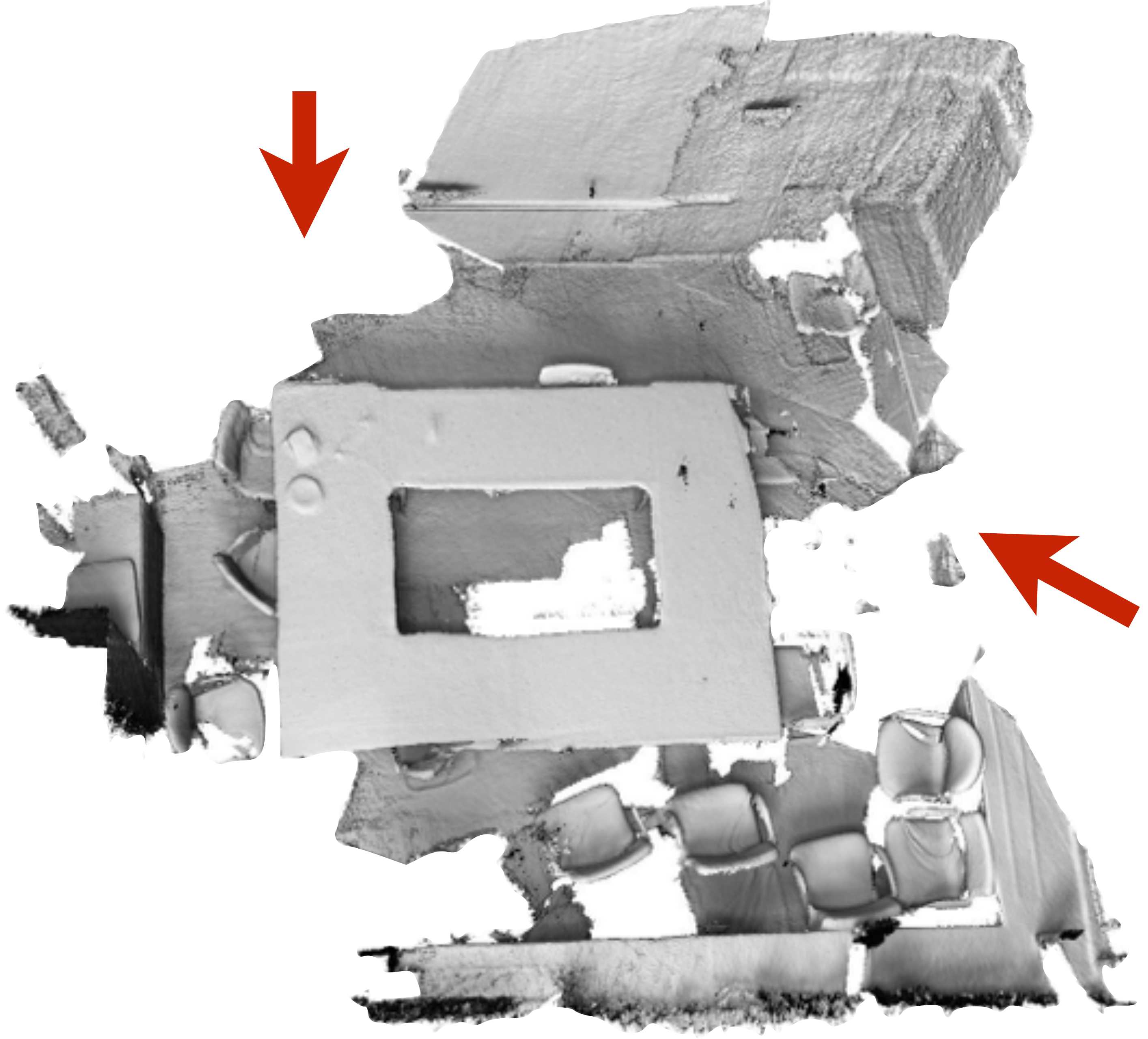
Receptive field:1.62 m

Receptive field: 2.26 m

How do we obtain training data ?

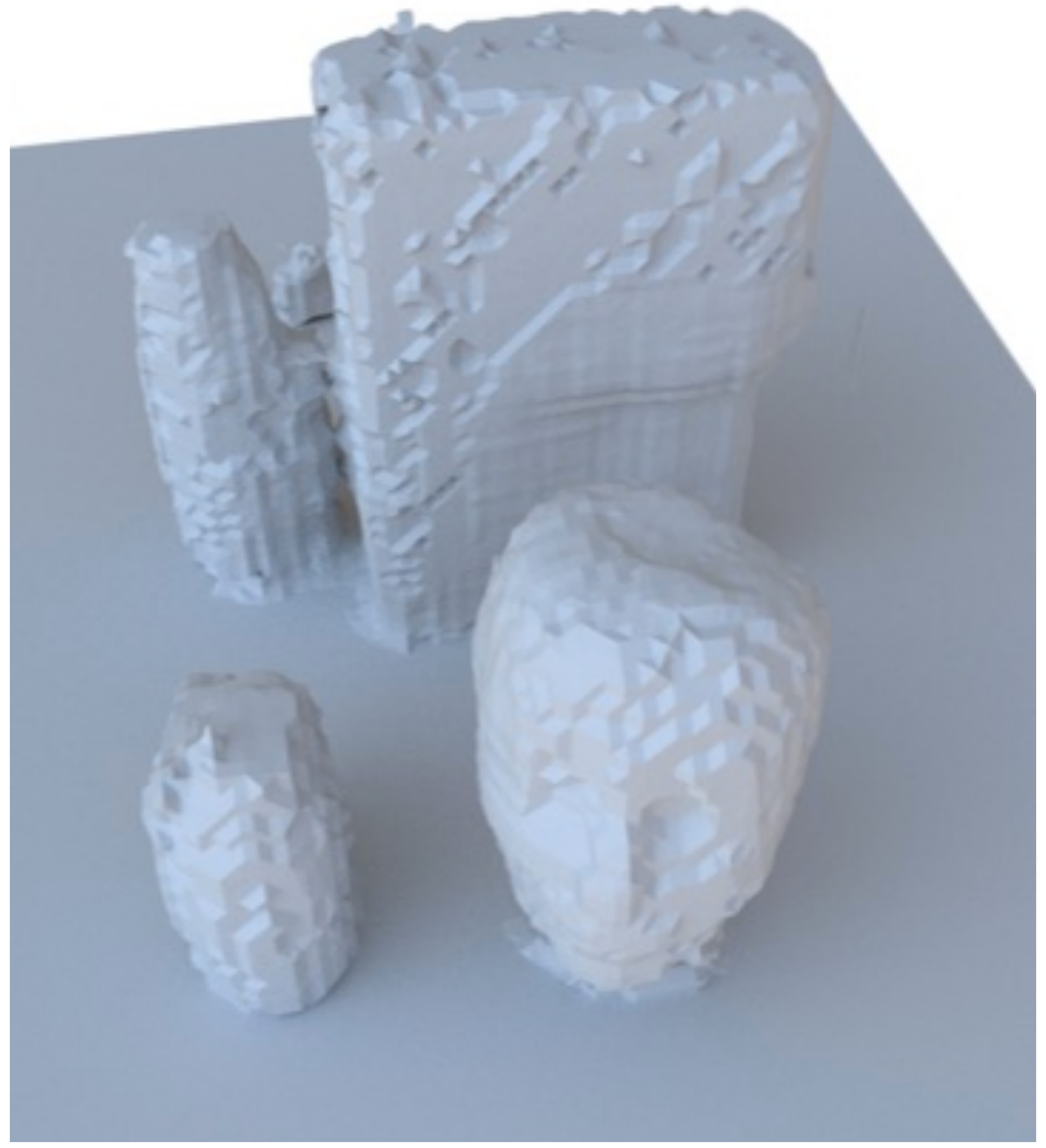
Only label visible surface

[Silberman et al.]



Partial observation

[xiao et al.]



Simple scenario

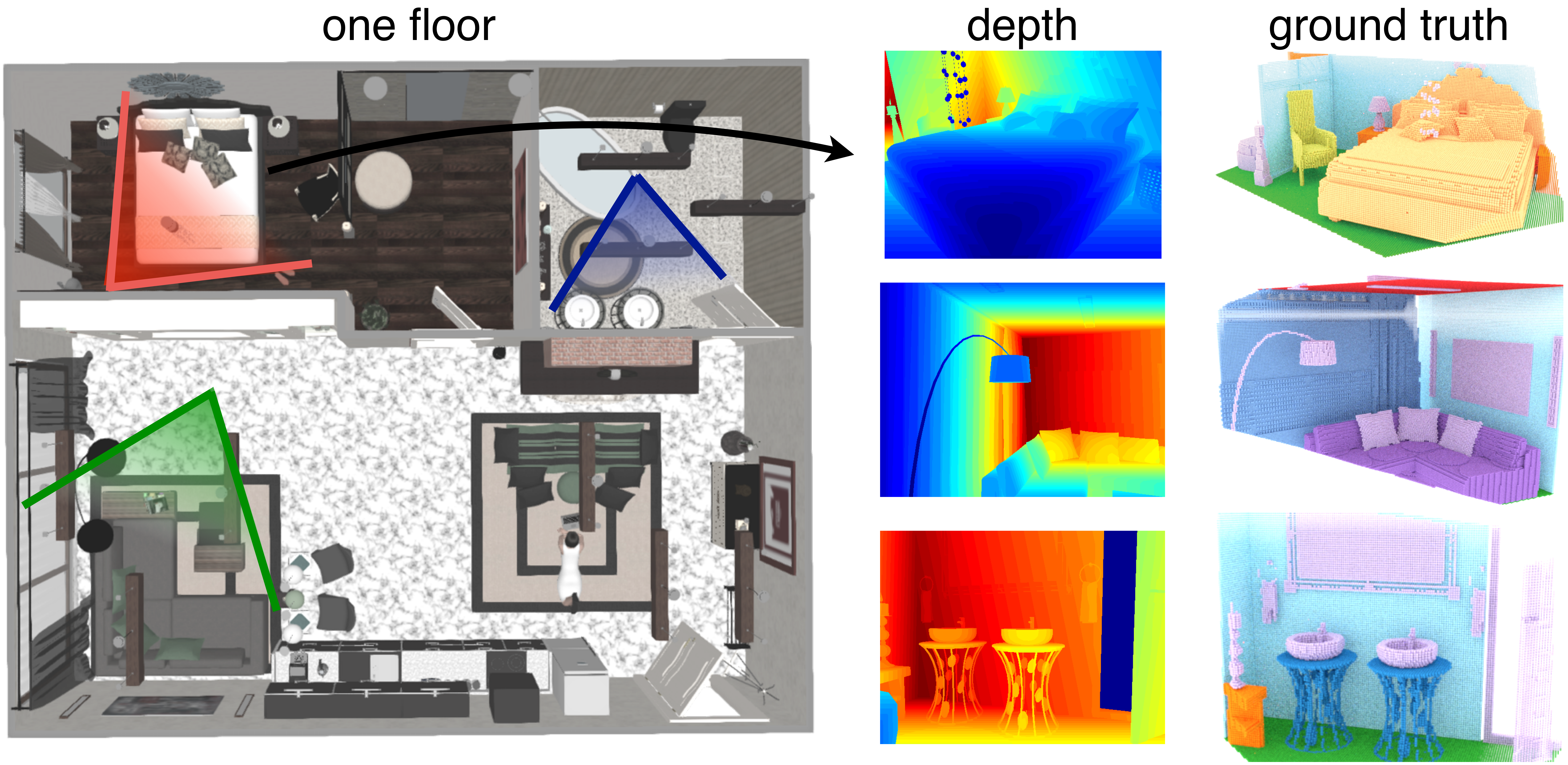
[Firman et al.]

No dense volumetric ground truth with semantic labels for the complete scene.

SUNCG dataset: over 40K houses

https://planner5d.com/

Synthesizing training data



Testing on real-world data

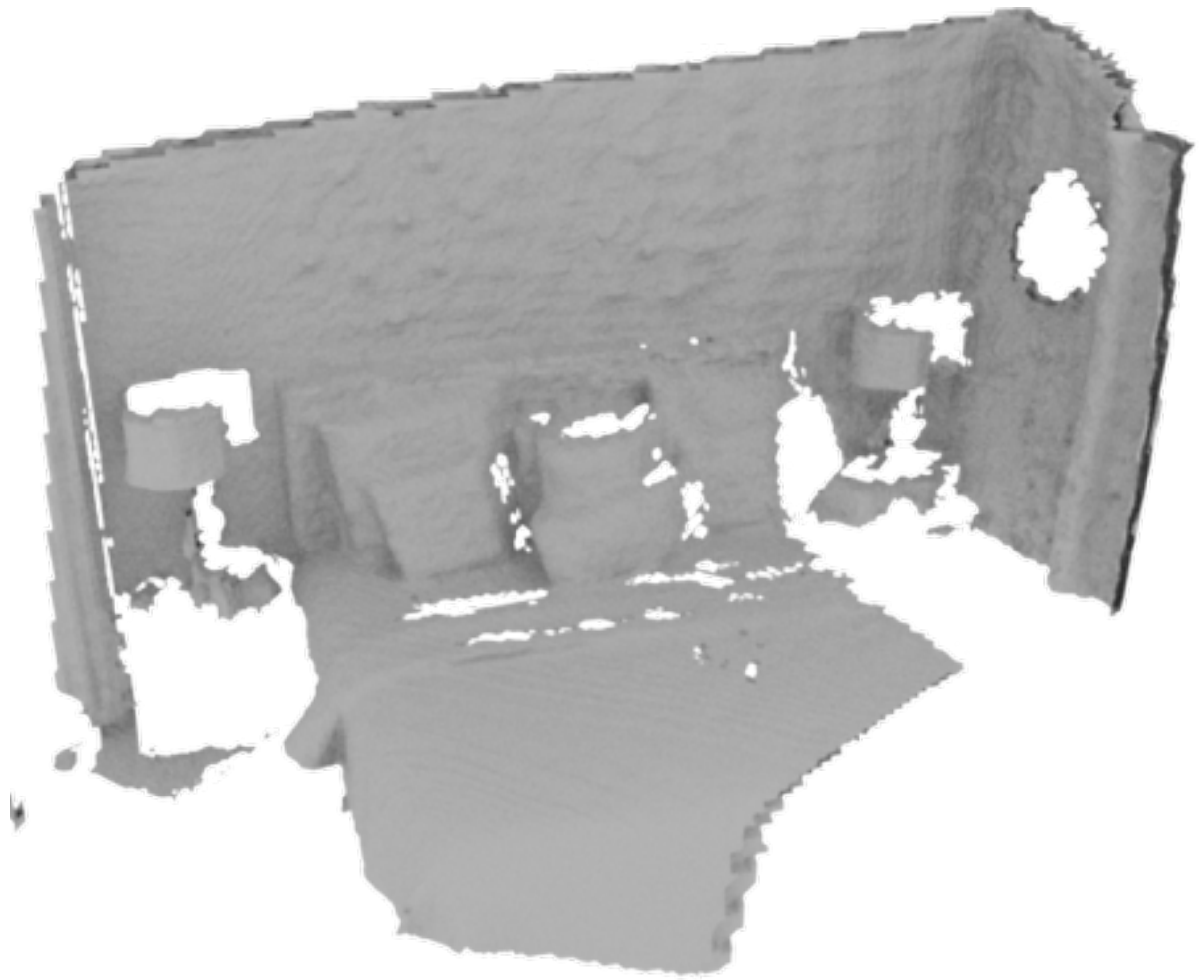
Training on SUNCG

Testing on NYU [1,2]

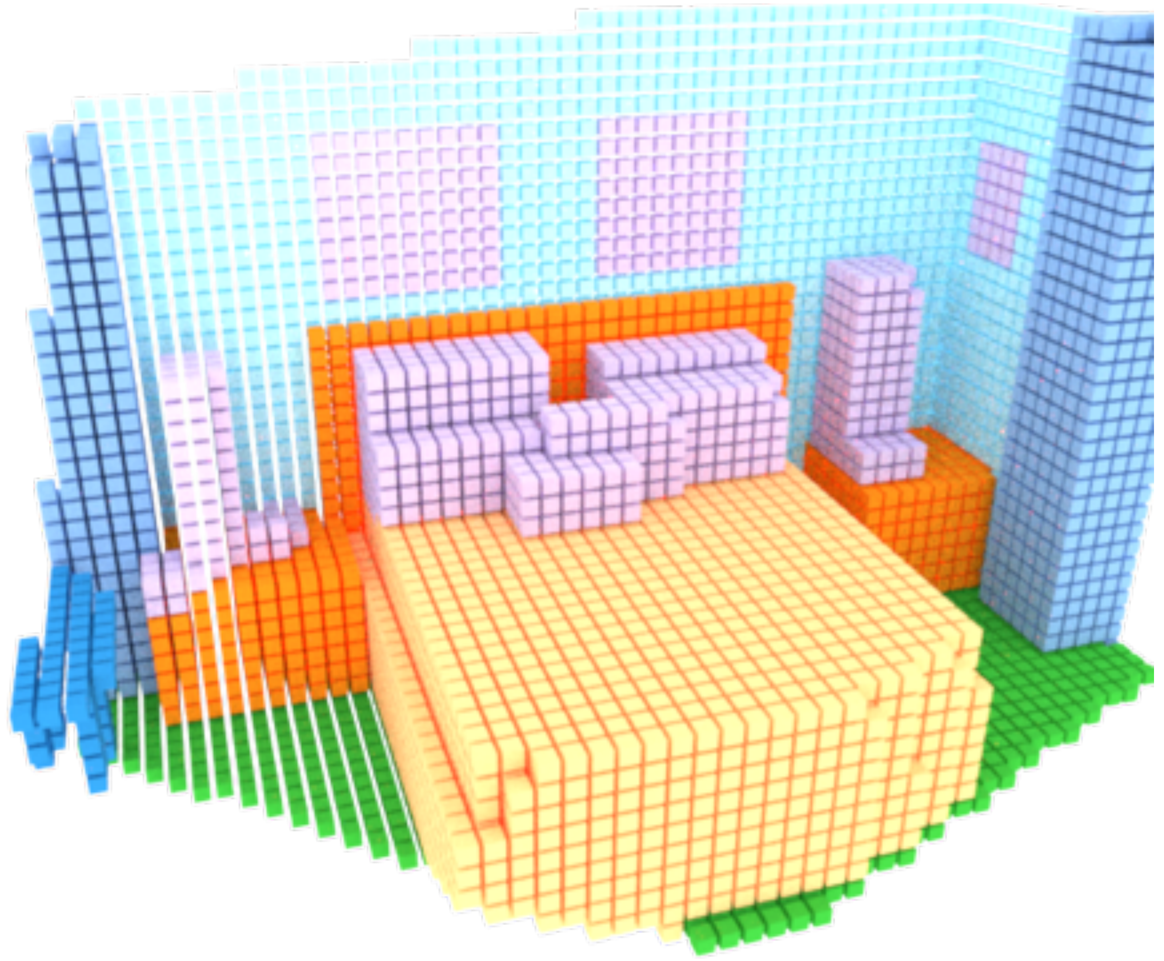
[1] **NYU depth v2:** Nathan Silberman, Pushmeet Kohli, Derek Hoiem, Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. ECCV 2012

[2] **Ground truth:** Ruiqi Guo, Derek Hoiem. Support surface prediction in indoor scenes. ICCV 2013

Comparison

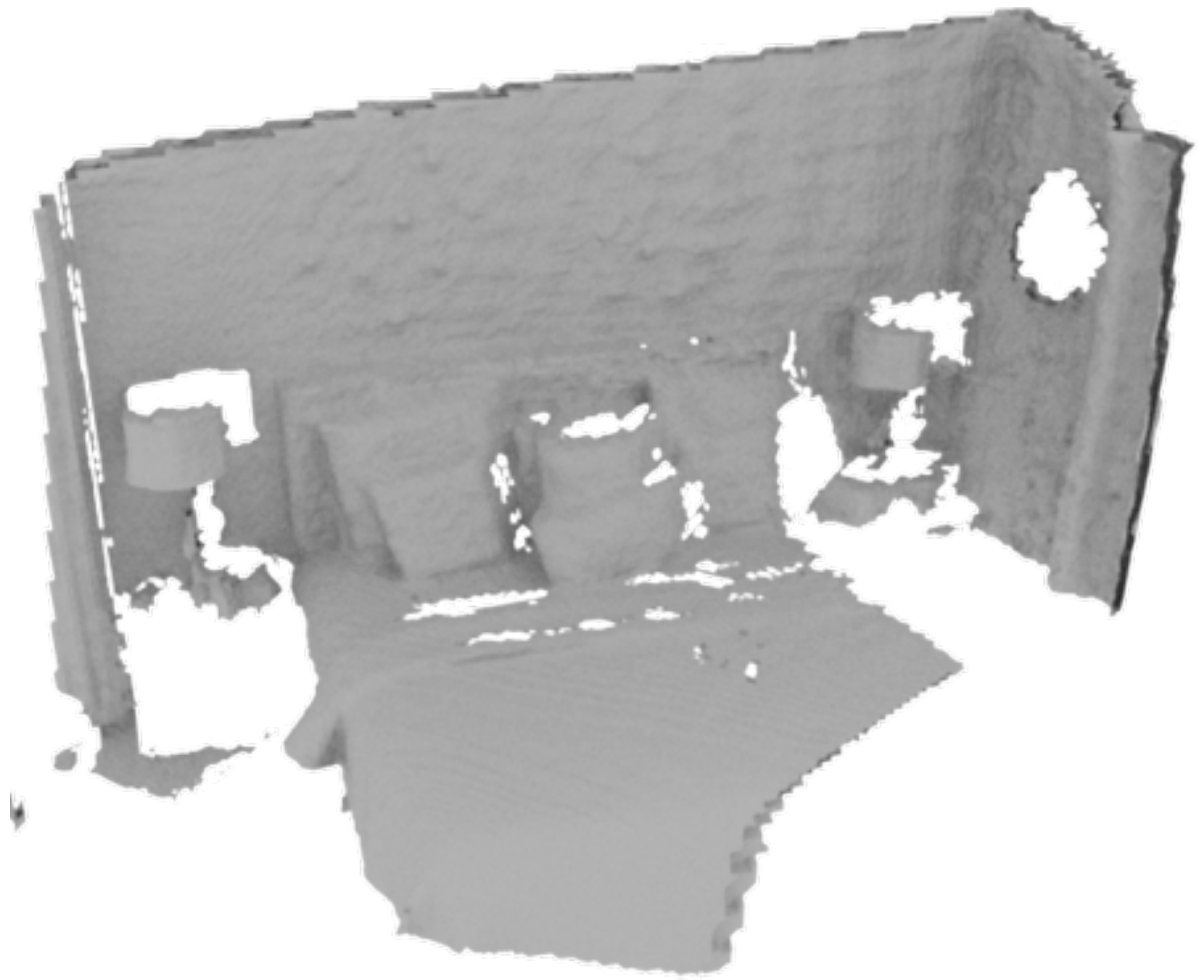


Observed Surface

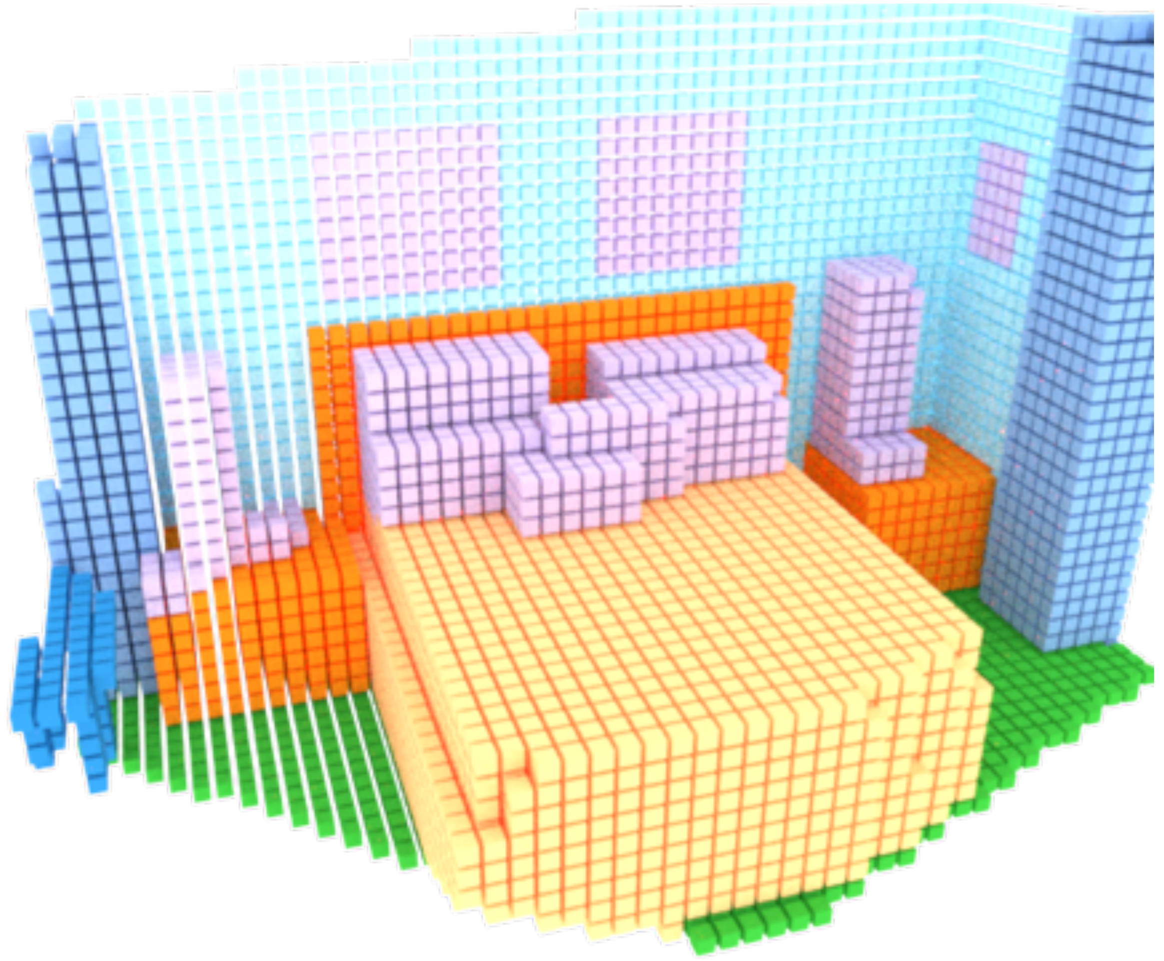


Ground Truth

Comparison

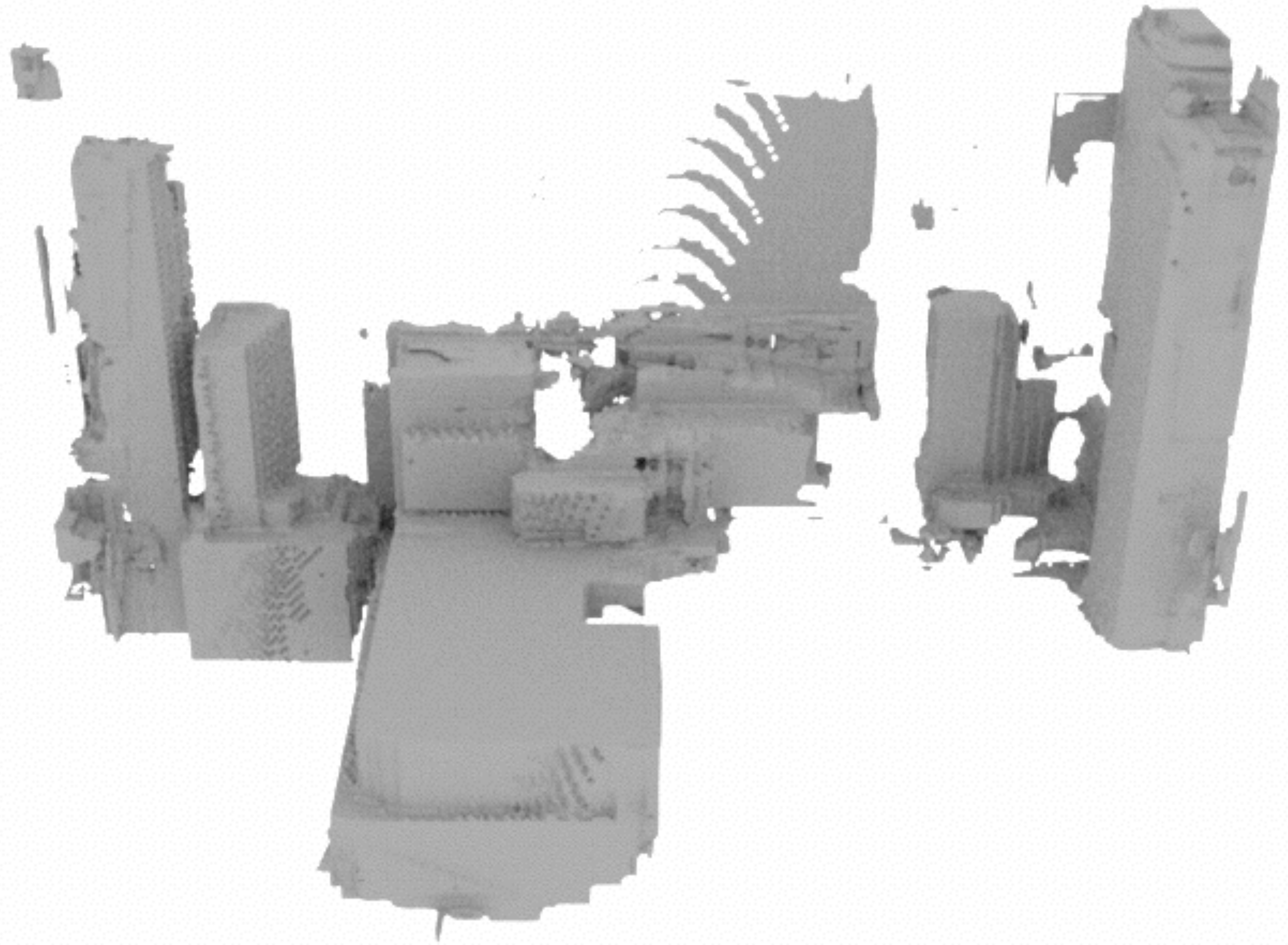


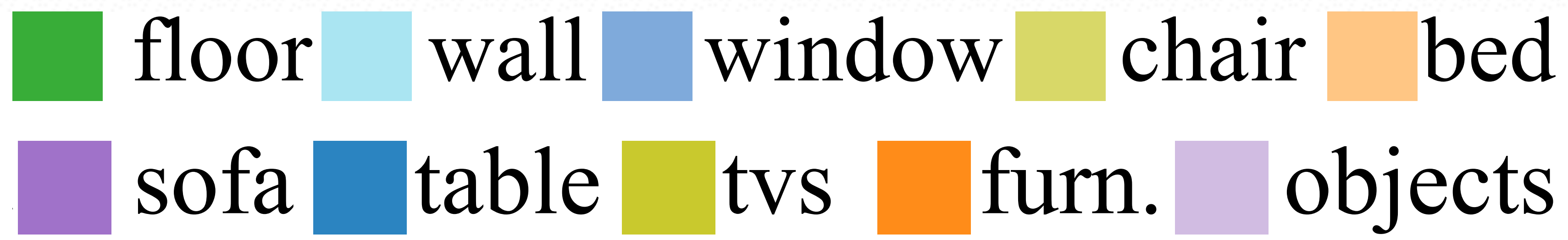
Observed Surface



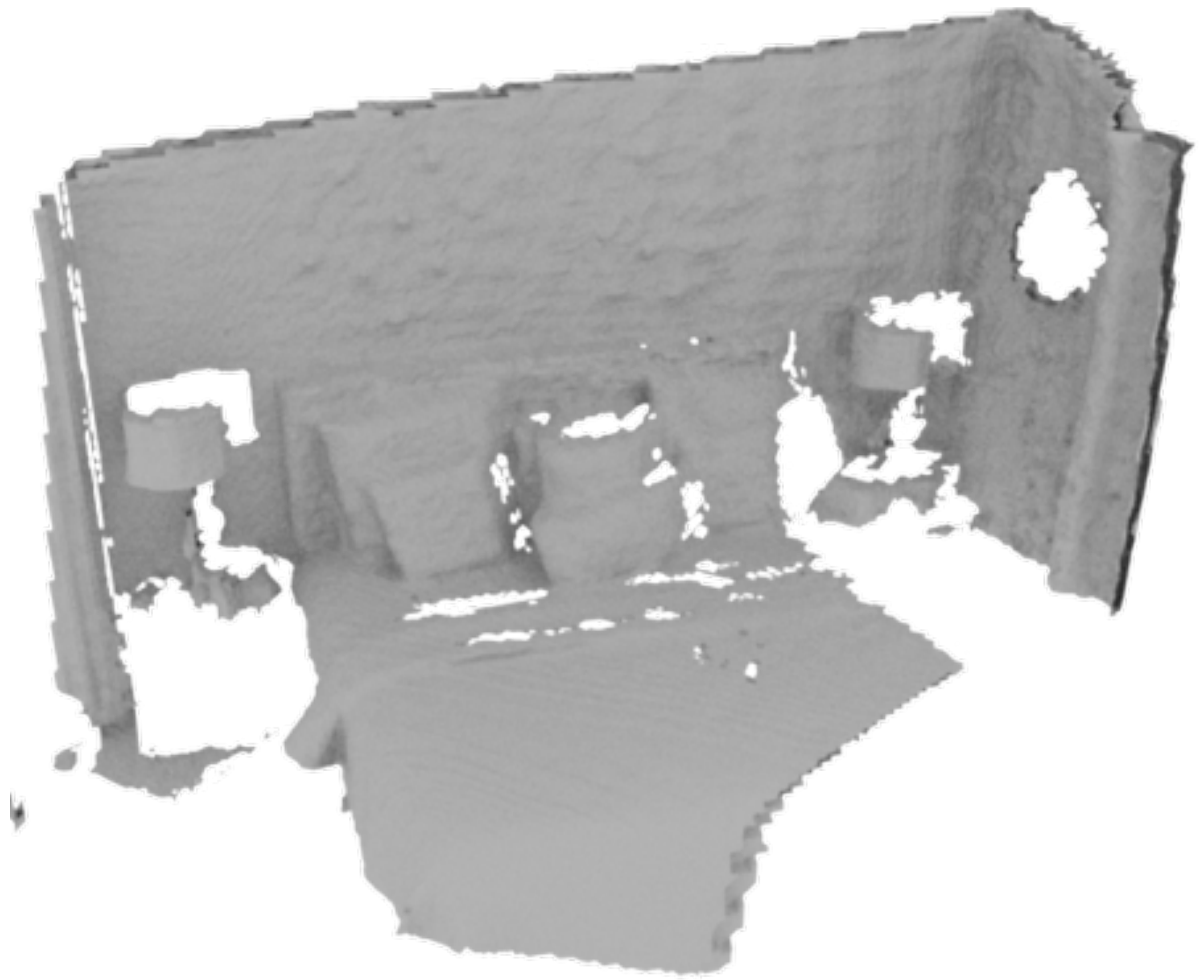
Ground Truth

Shape Completion
[Firman *et al.*]

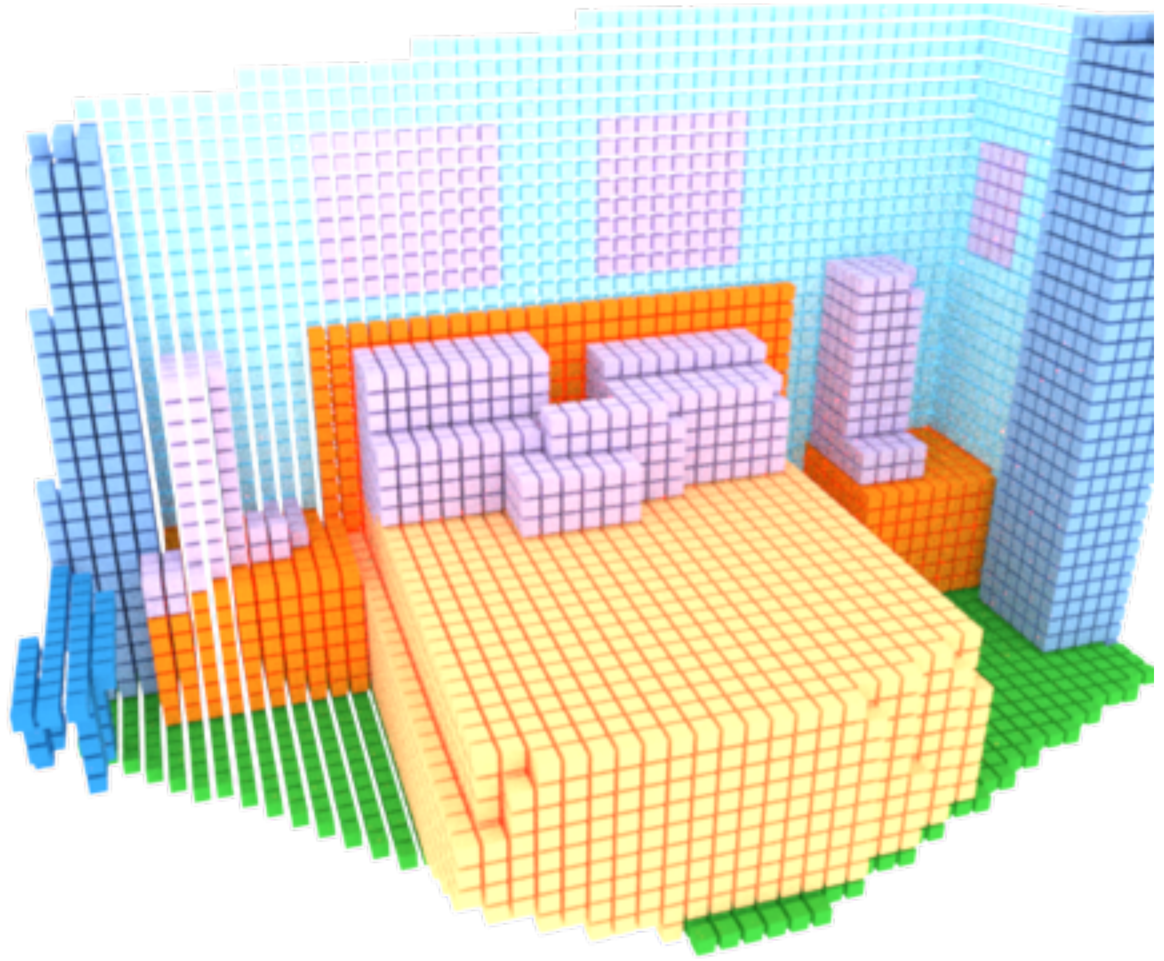




Comparison

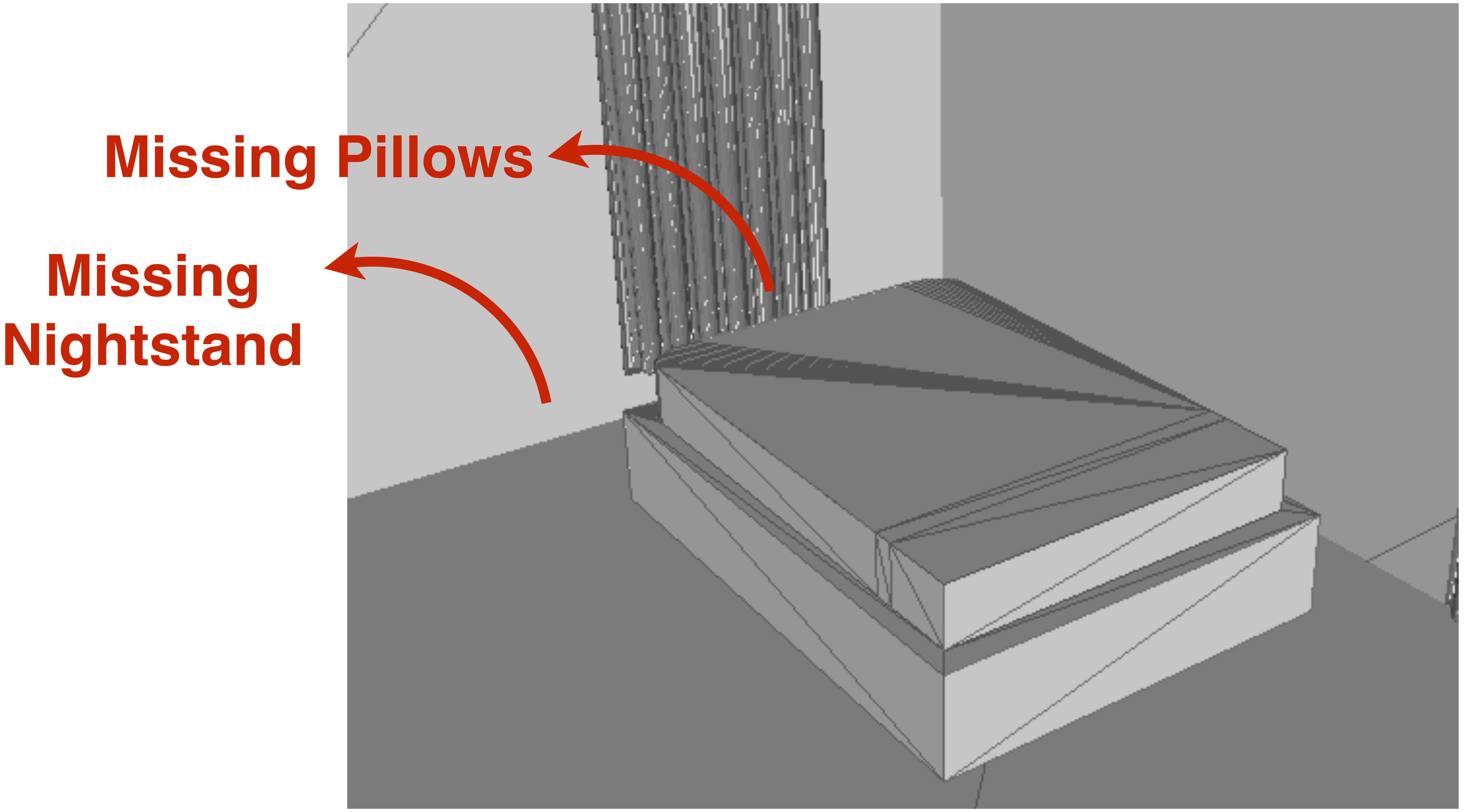


Observed Surface

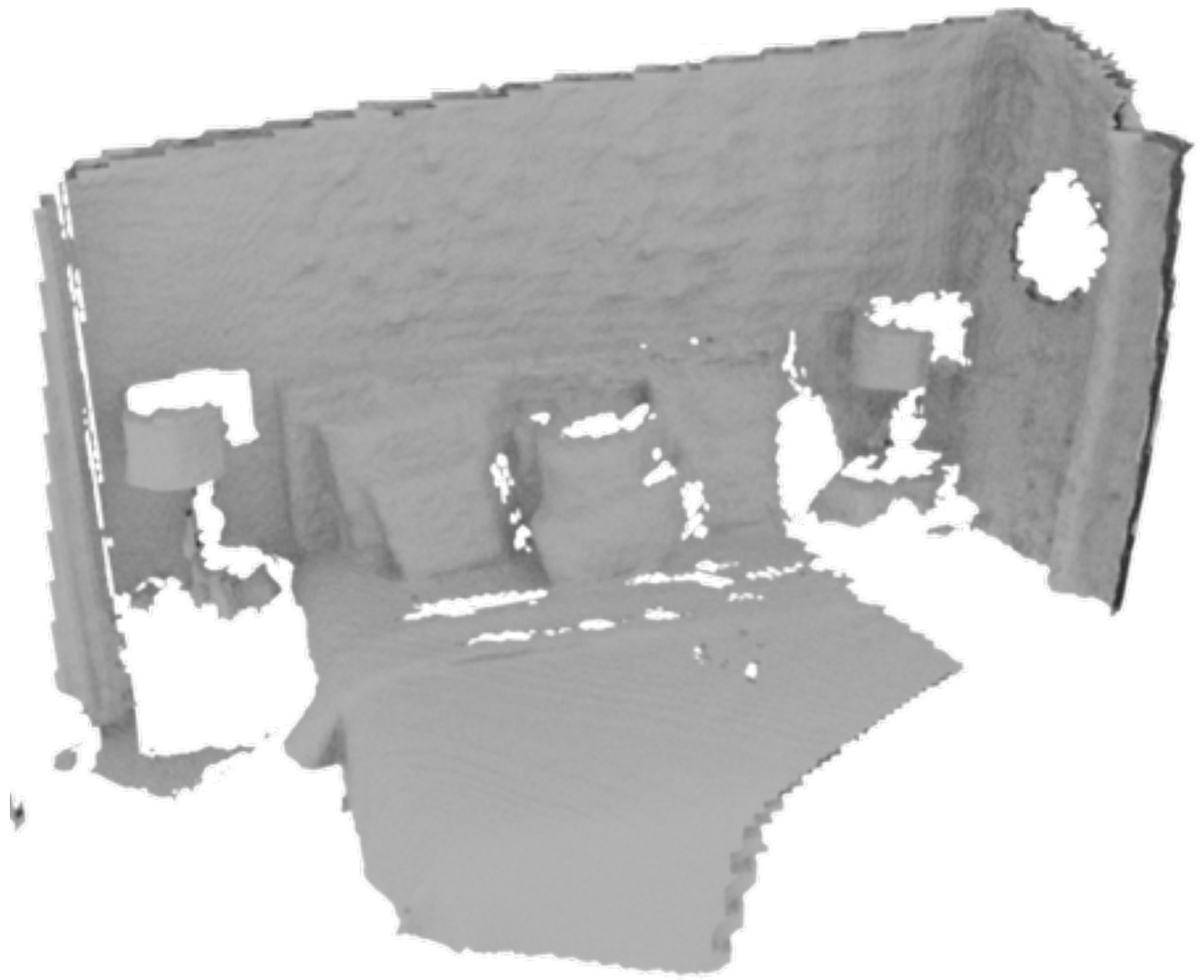


Ground Truth

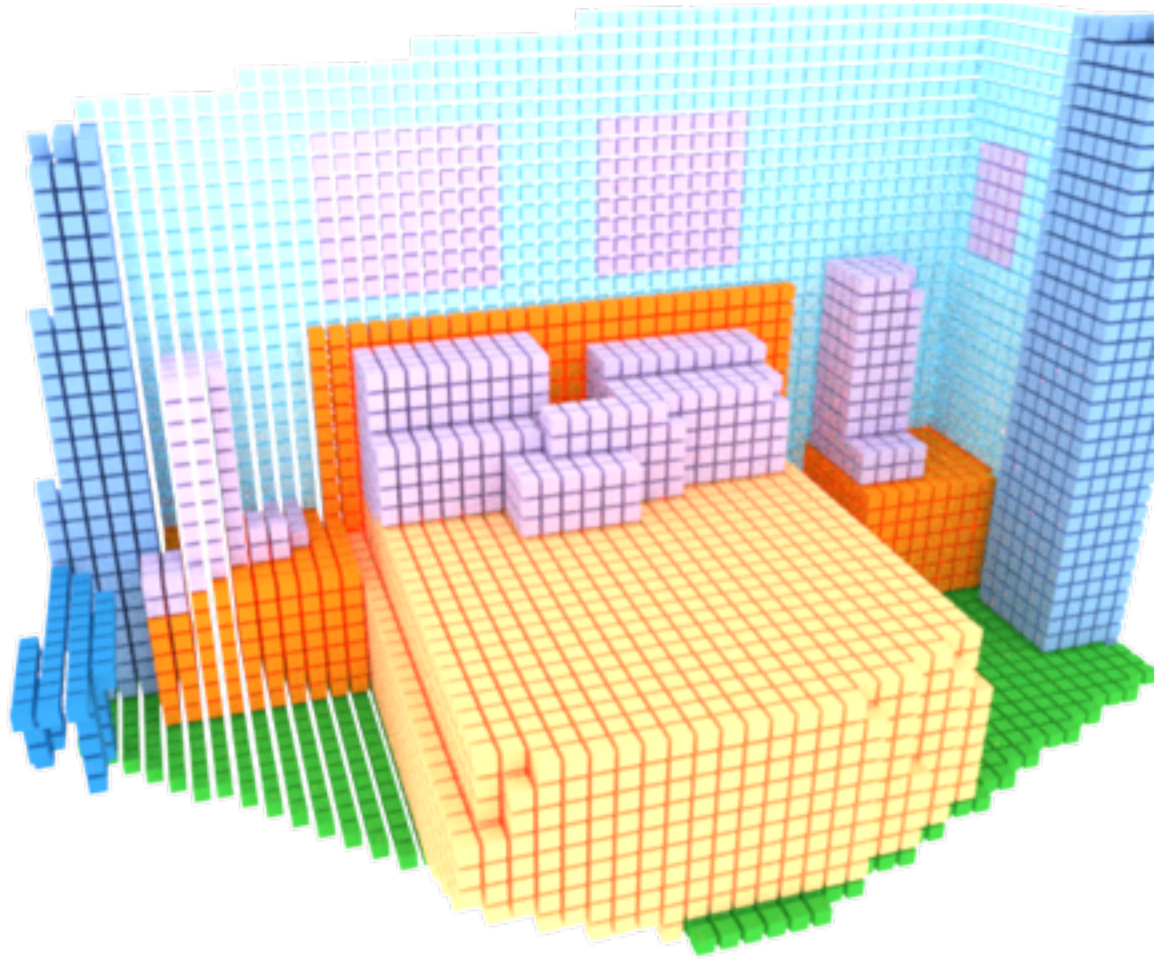
Model Retrieval+Fitting
[Geiger and Wang]



Comparison

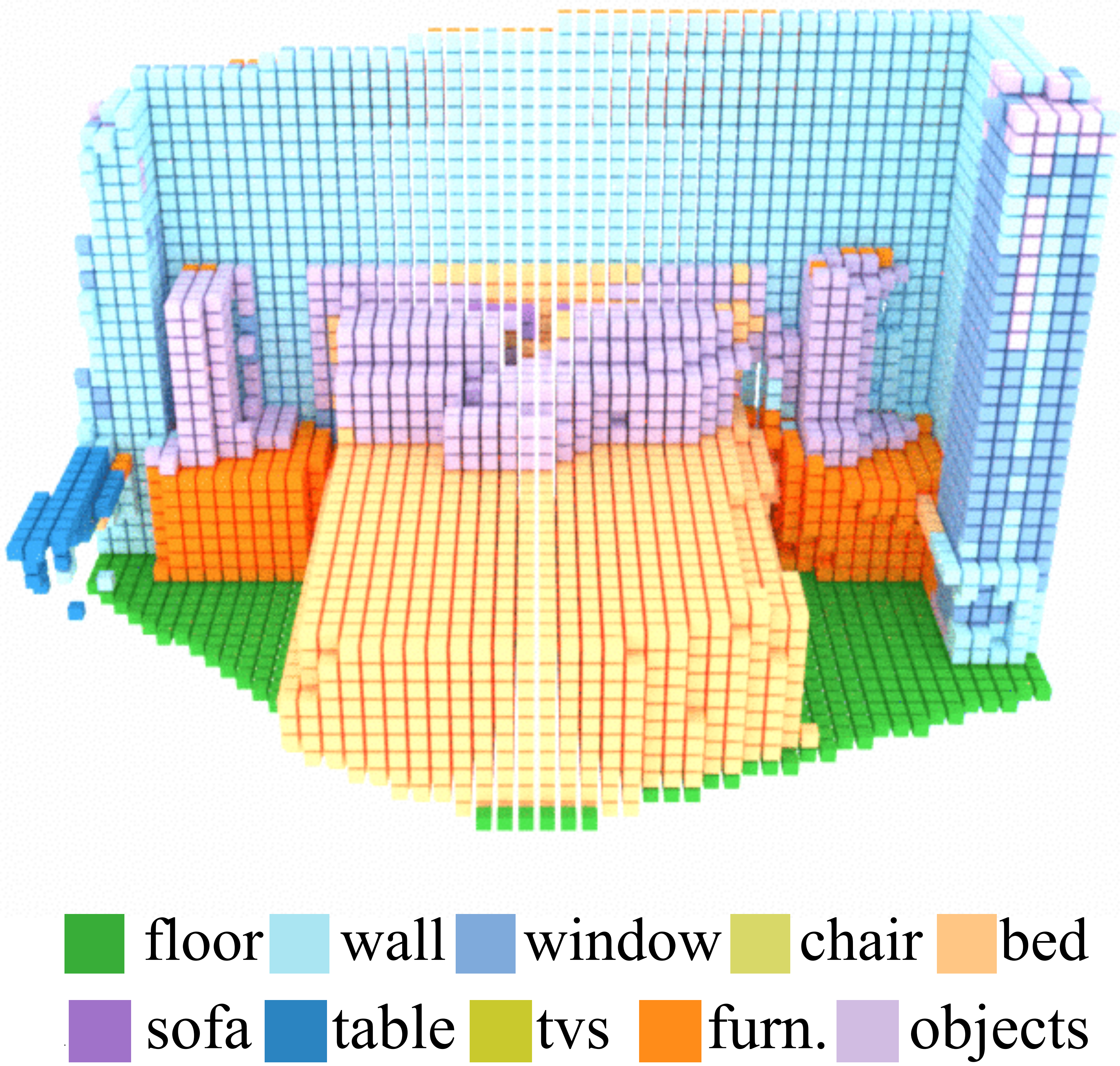


Observed Surface

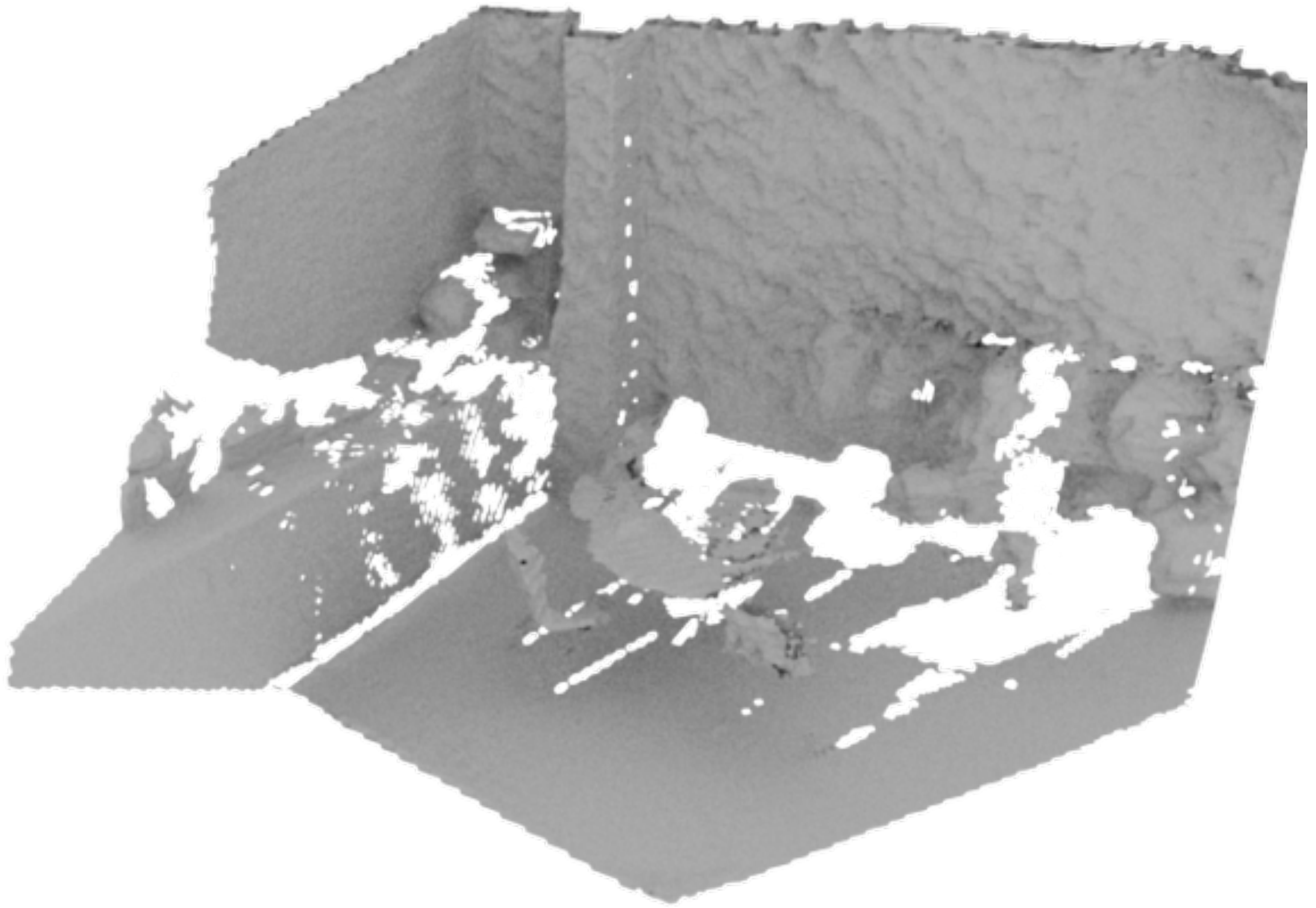


Ground Truth

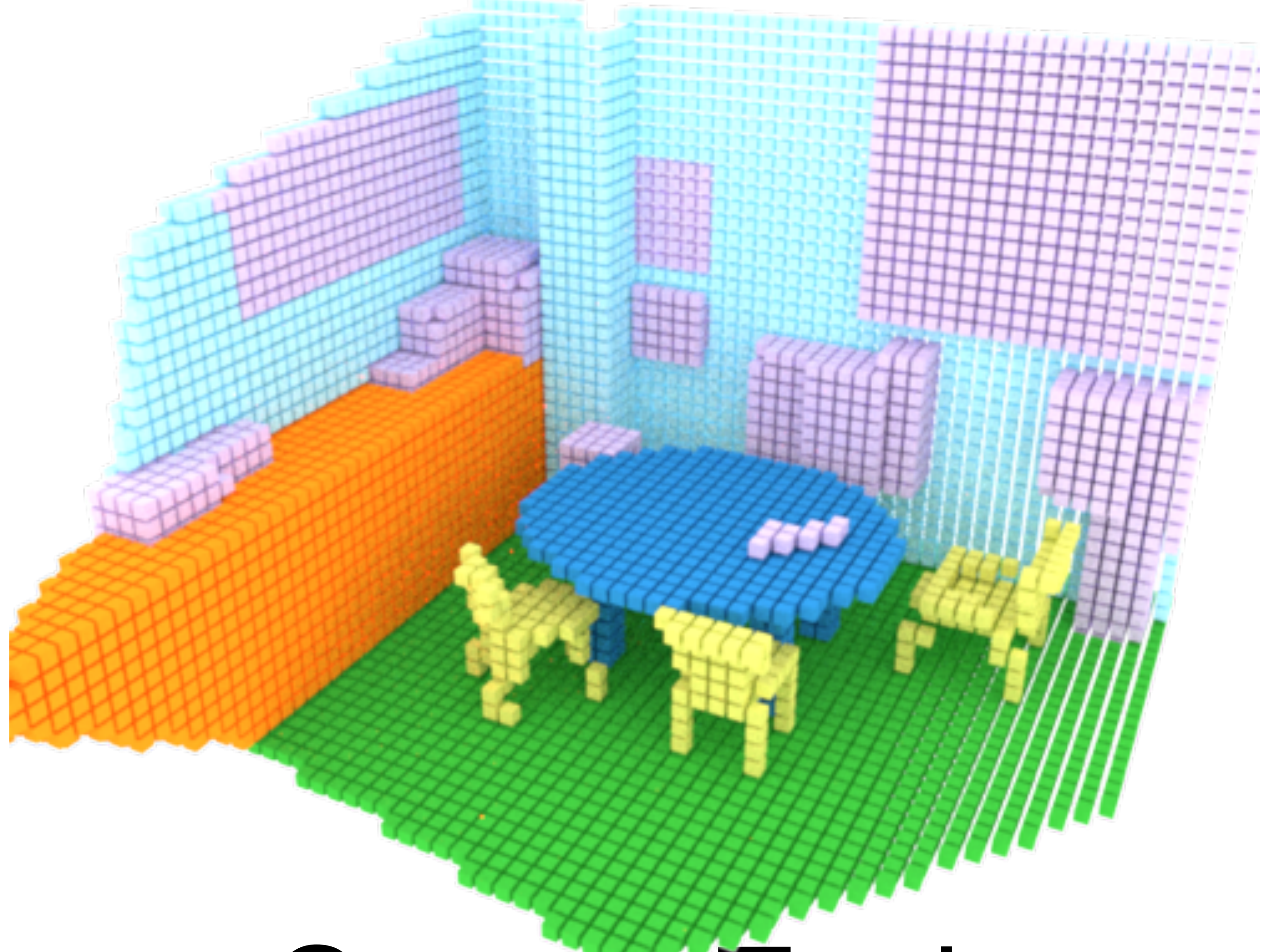
SSCNet



Comparison

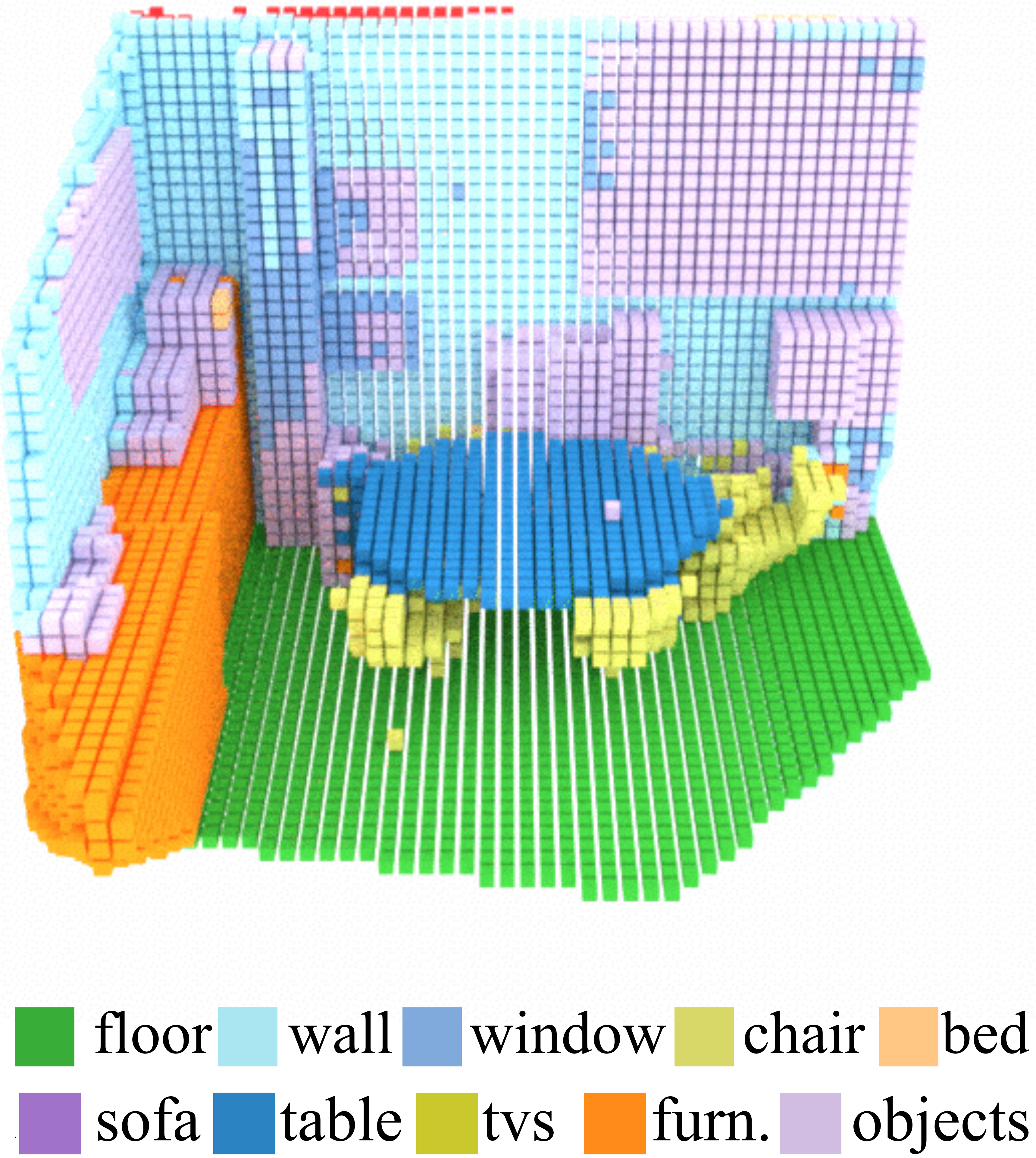


Observed Surface



Ground Truth

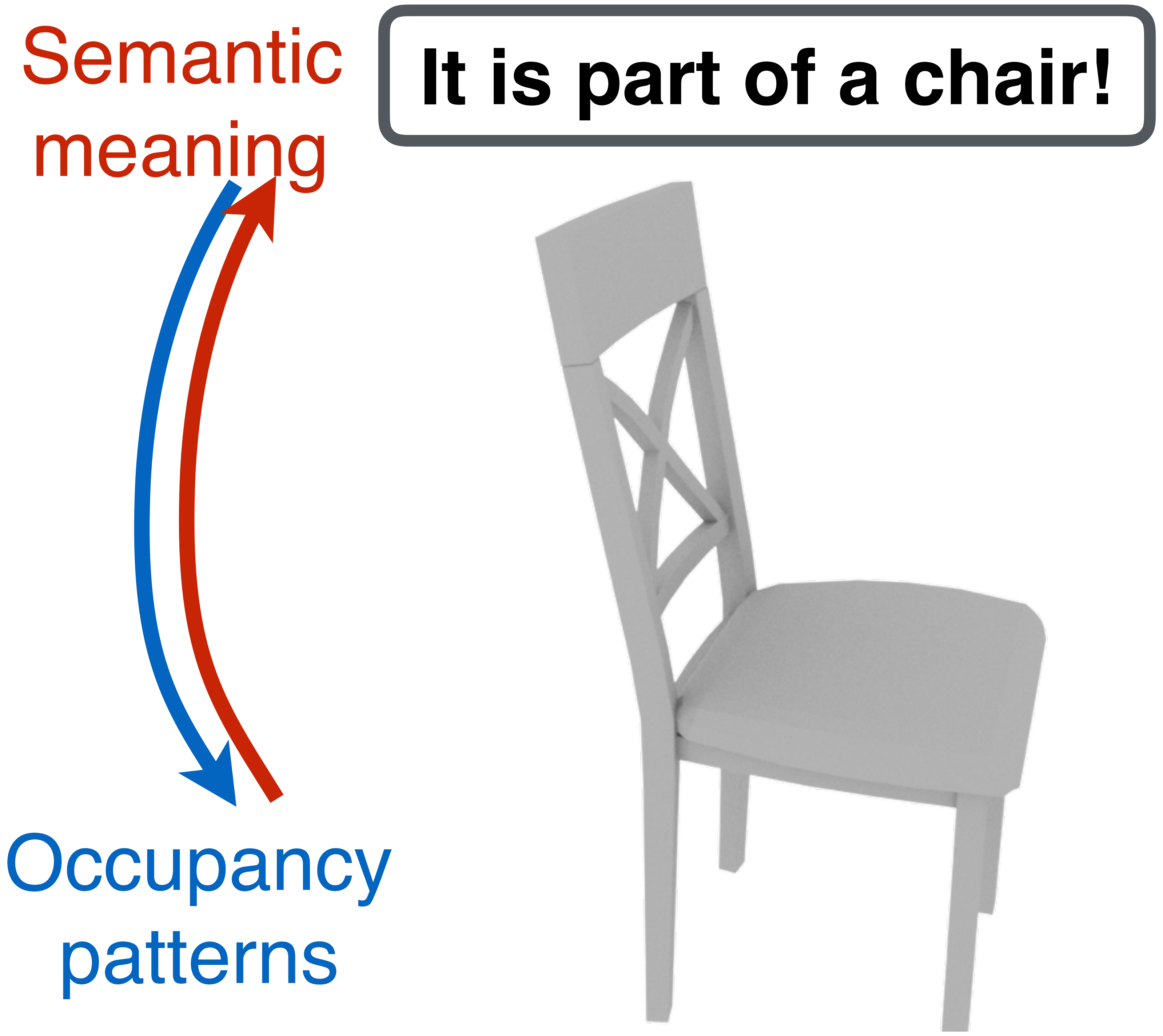
SSCNet

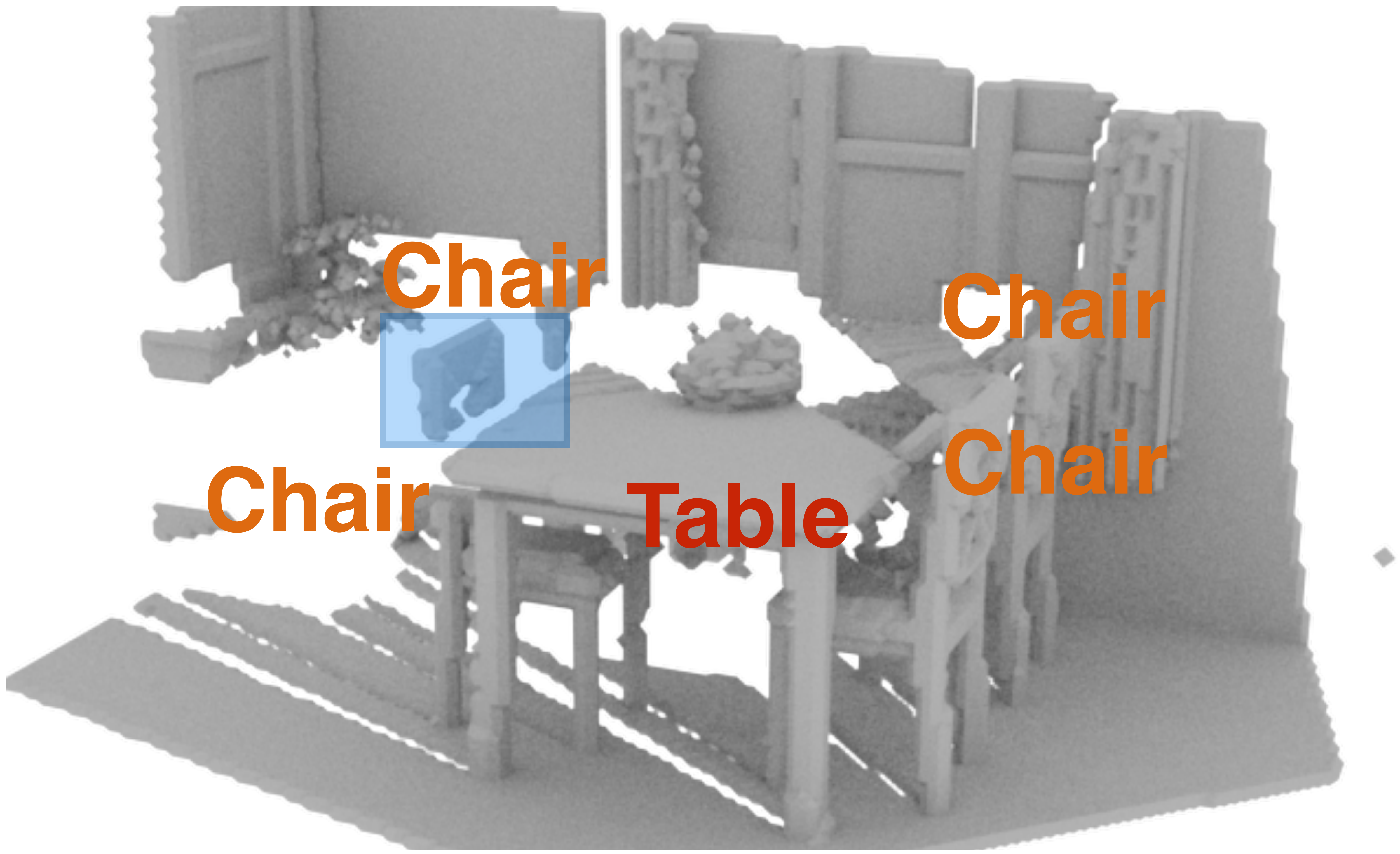


Analysis

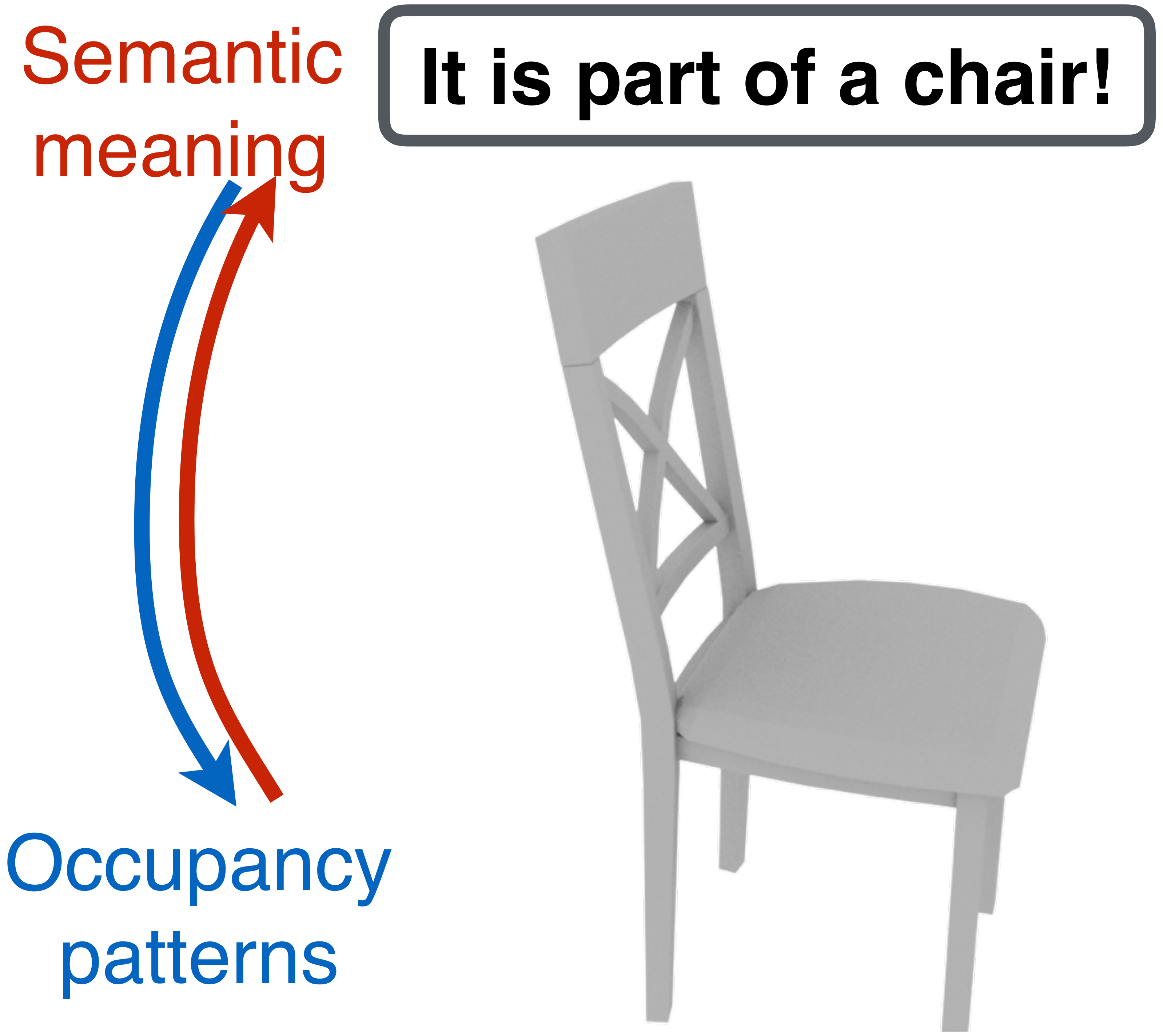
Key ideas:

1. Object occupancy and the identity are tightly intertwined.
2. It is important to capture and understand 3D context with big receptive fields.

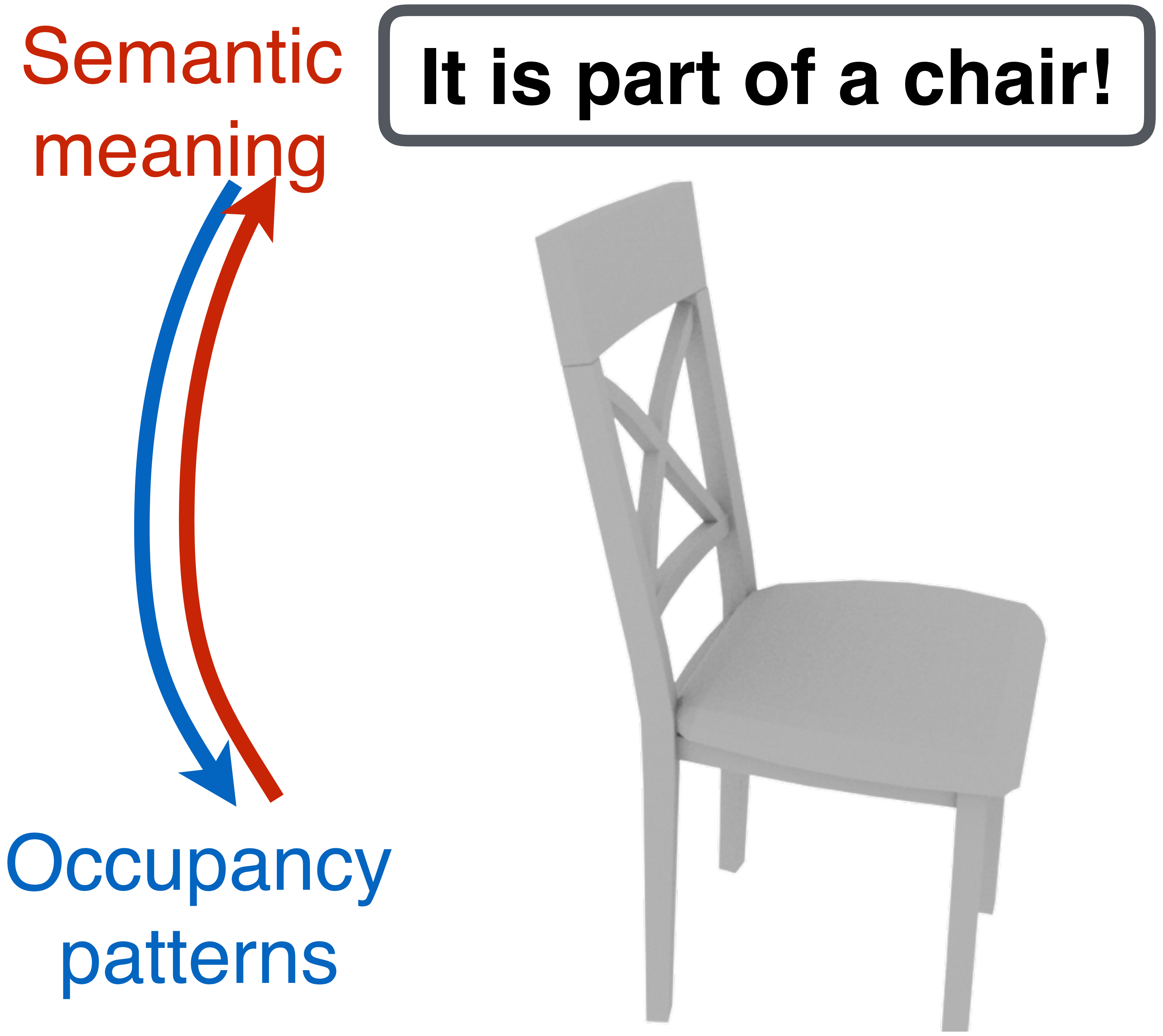




Does joint understanding help?



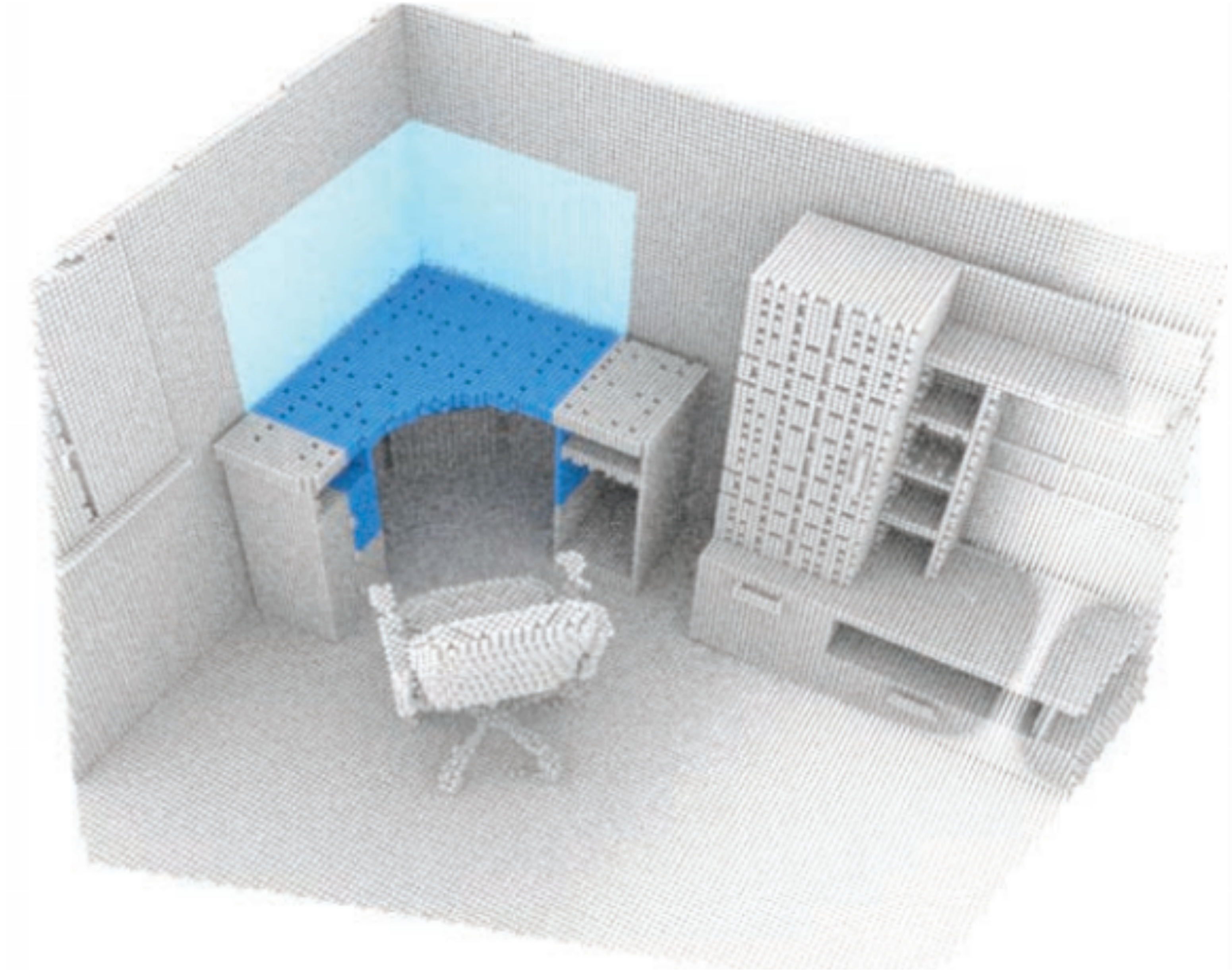
Does joint understanding help?



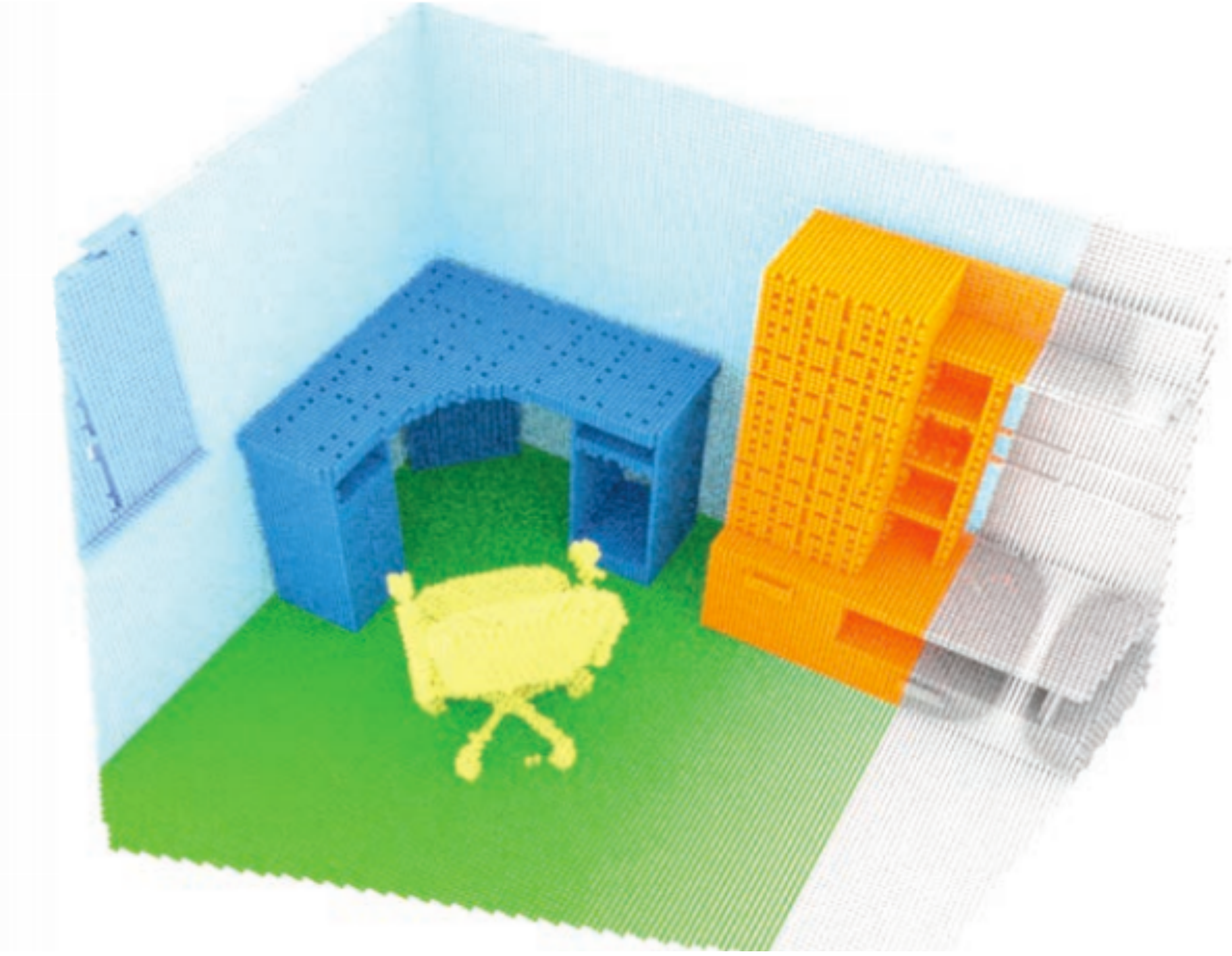
task	scene completion w/o semantics	semantic labeling w/o completion
completion only	64.8	-
semantic only	-	51.2
joint	73.0	54.2

Does a bigger receptive field help?

model/task	Basic	Basic + Dilation
semantic scene completion (IoU %)	38.0	44.3



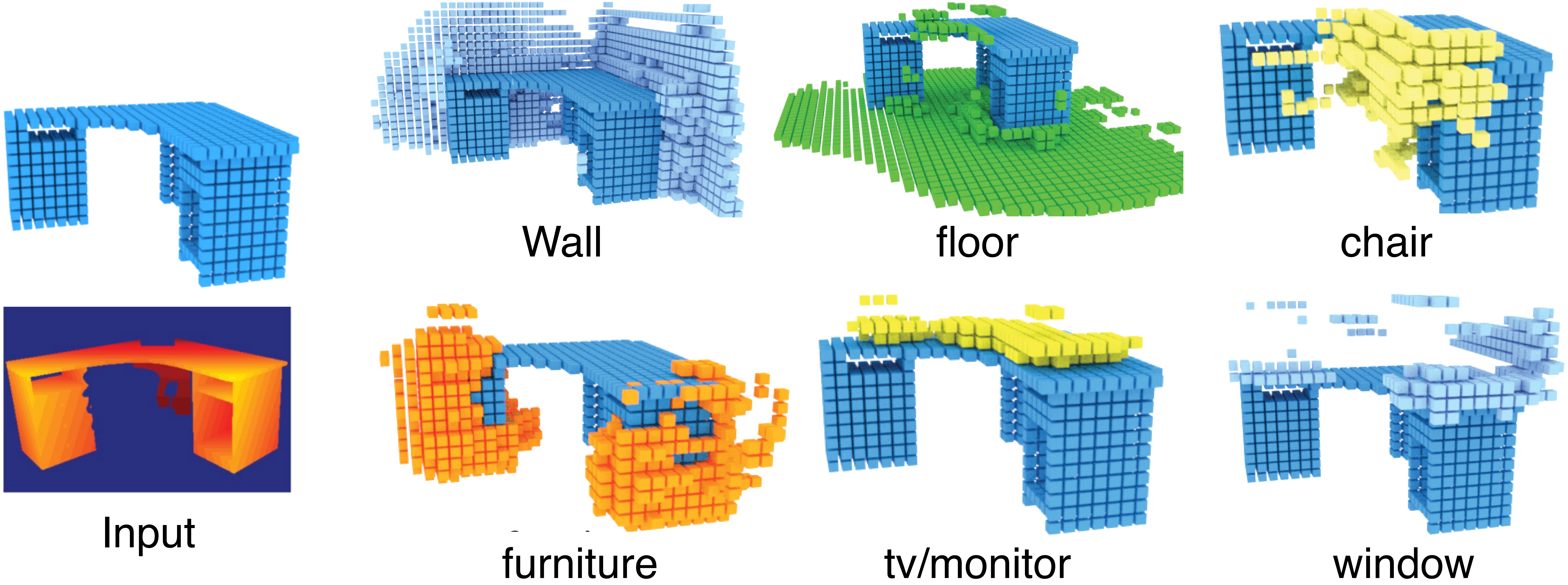
receptive field: 1 meter



receptive field: 2.26 meter

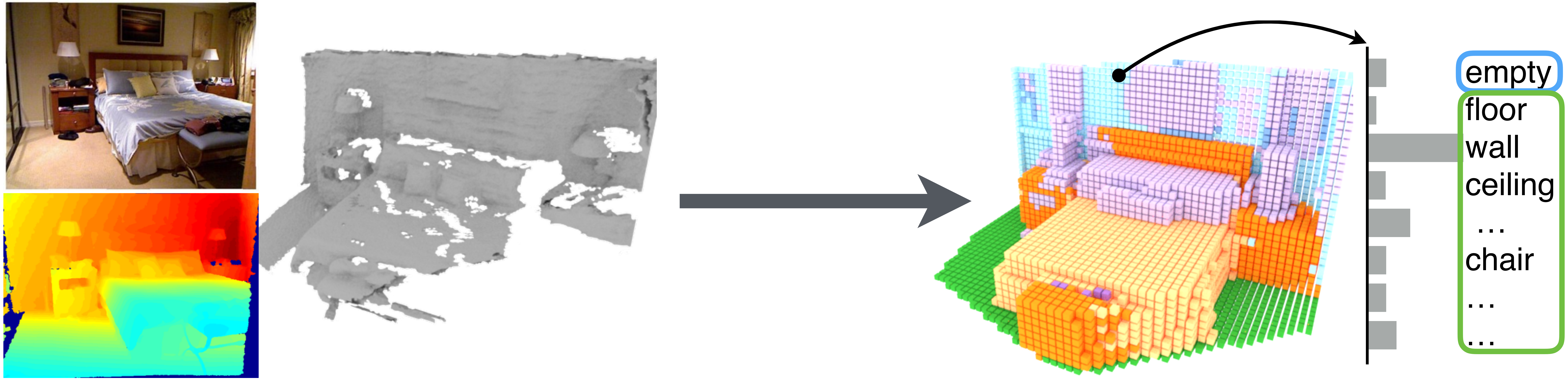
What 3D context does the network learn?

What 3D context does the network learn?



Conclusion

Conclusion

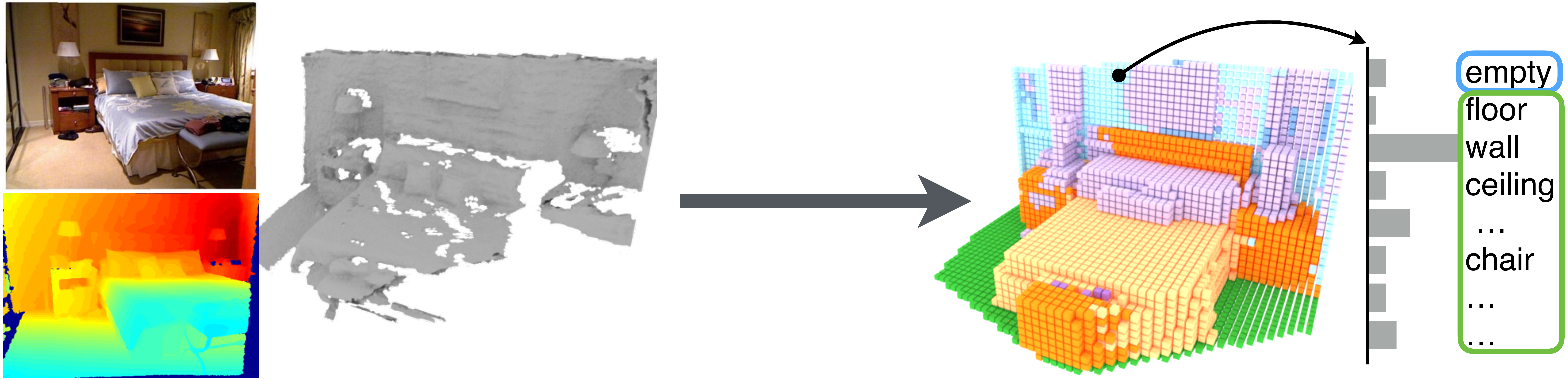


Semantic Scene Completion

- Semantic scene completion network, SSCNet
- A large-scale synthetic scene dataset, SUNCG

Code & Data: sscnet.cs.princeton.edu

Conclusion

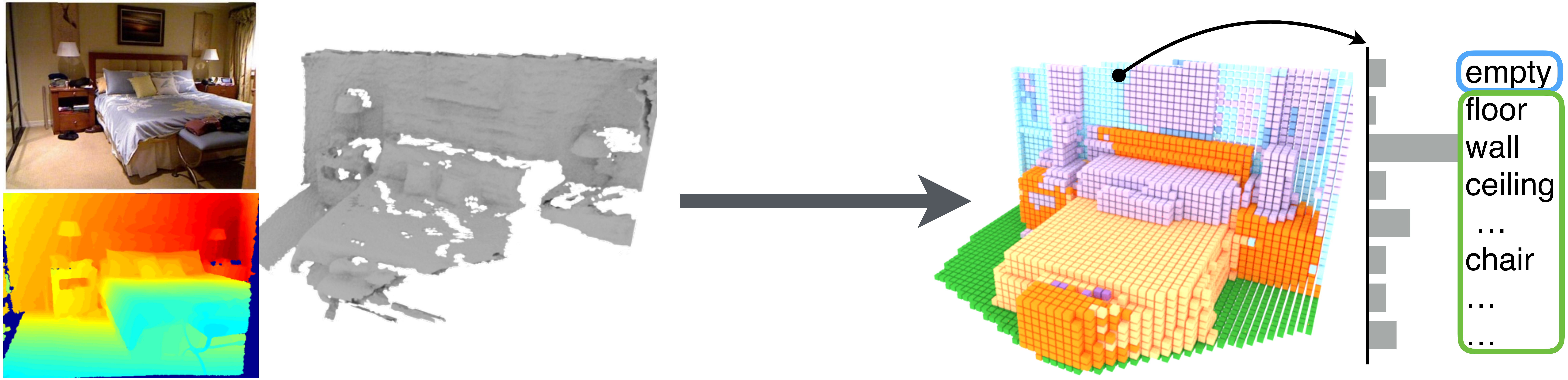


Semantic Scene Completion

- Semantic scene completion network, SSCNet
- A large-scale synthetic scene dataset, SUNCG

Code & Data: sscnet.cs.princeton.edu

Conclusion

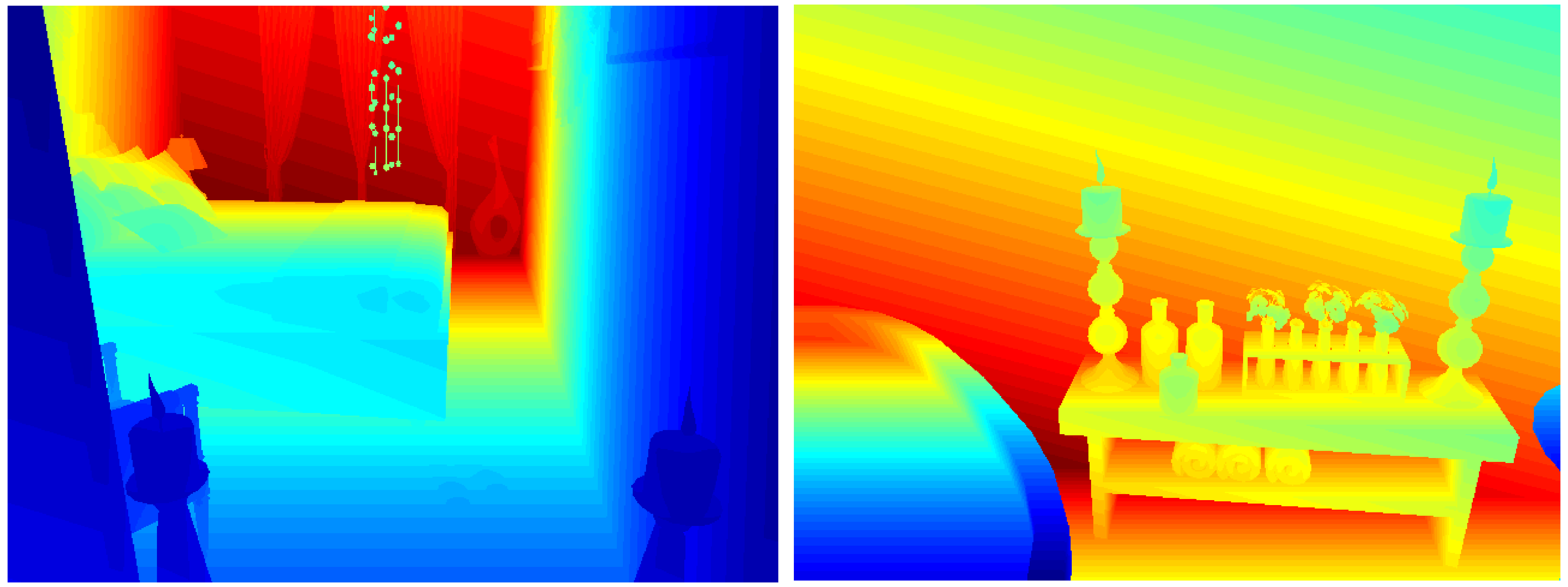


Semantic Scene Completion

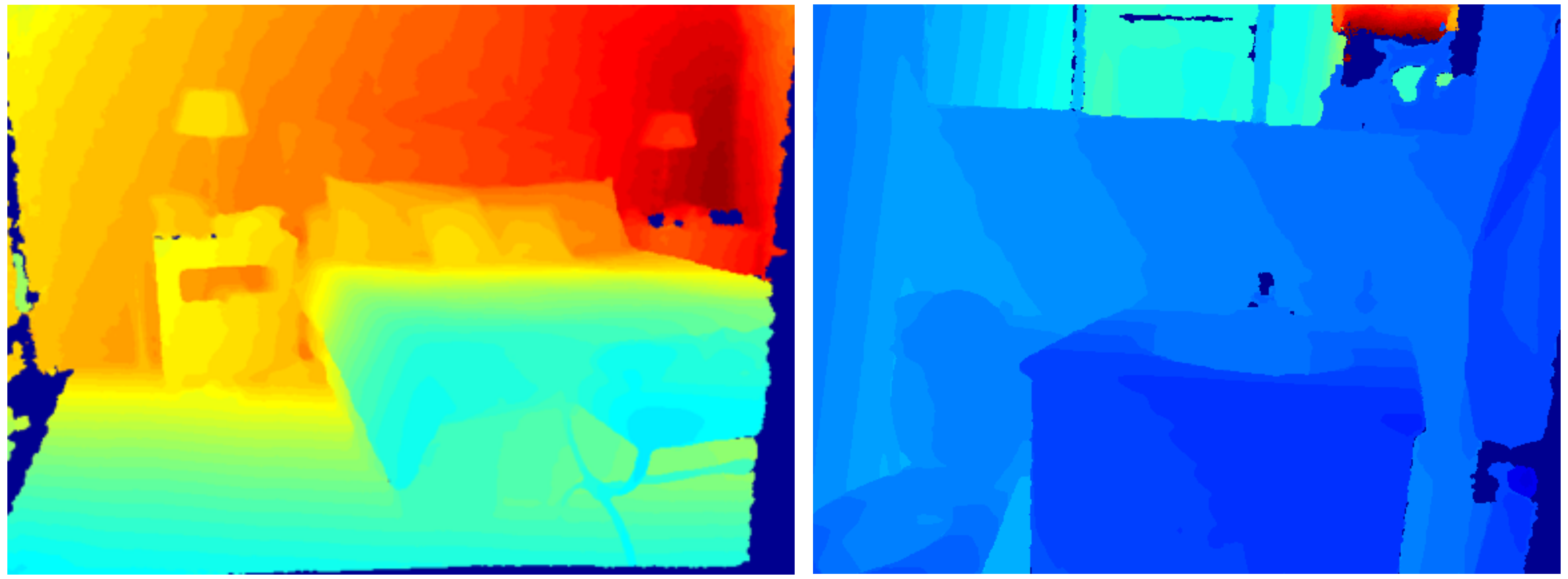
- Semantic scene completion network, SSCNet
- A large-scale synthetic scene dataset, SUNCG

Code & Data: sscnet.cs.princeton.edu

Does synthetic data help?



Rendered depth map

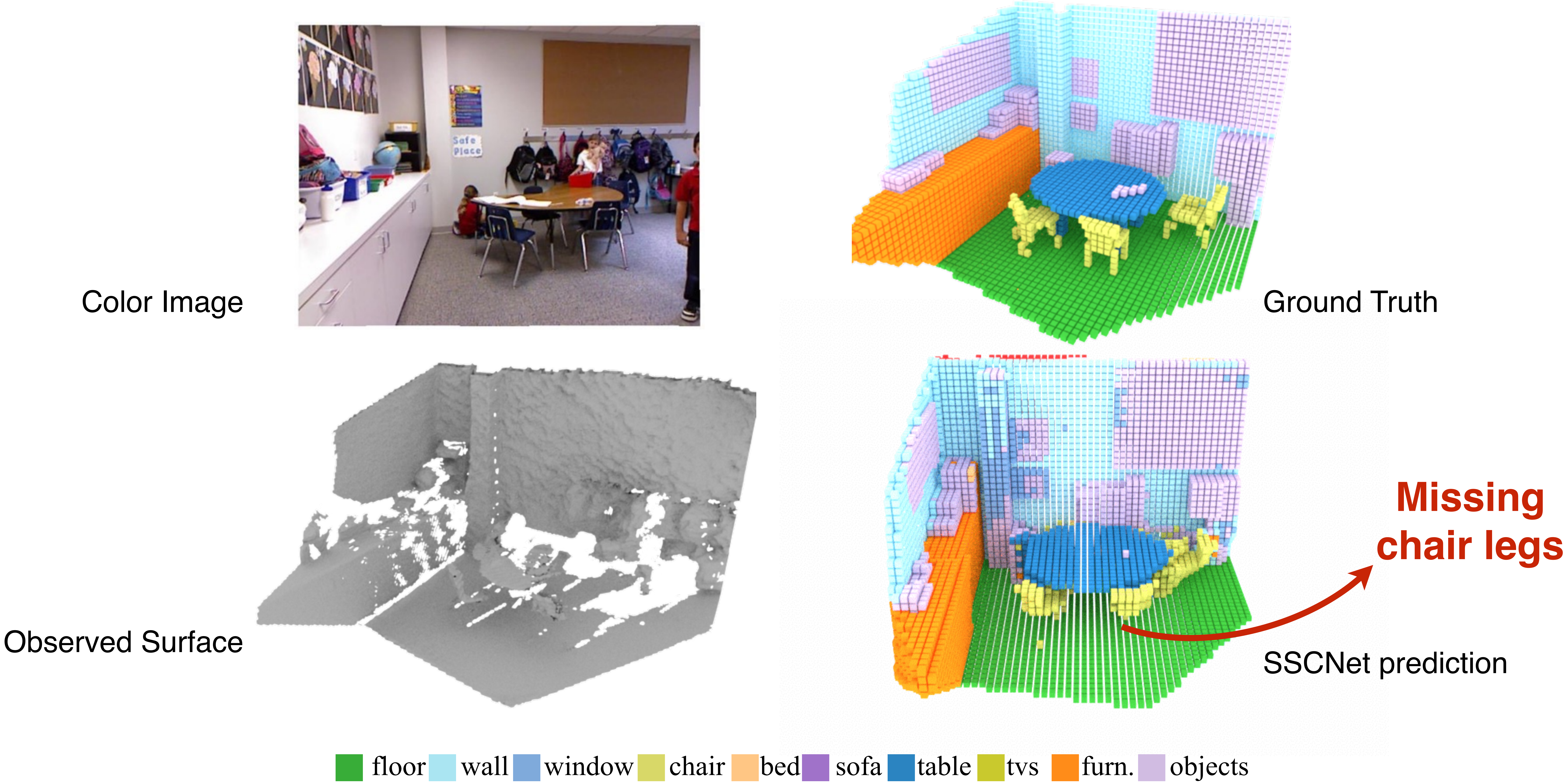


Kinect depth map

Test on NYU	NYU	SUNCG	SUNCG+NYU
semantic scene completion (IoU %)	24.7	20.2	30.5

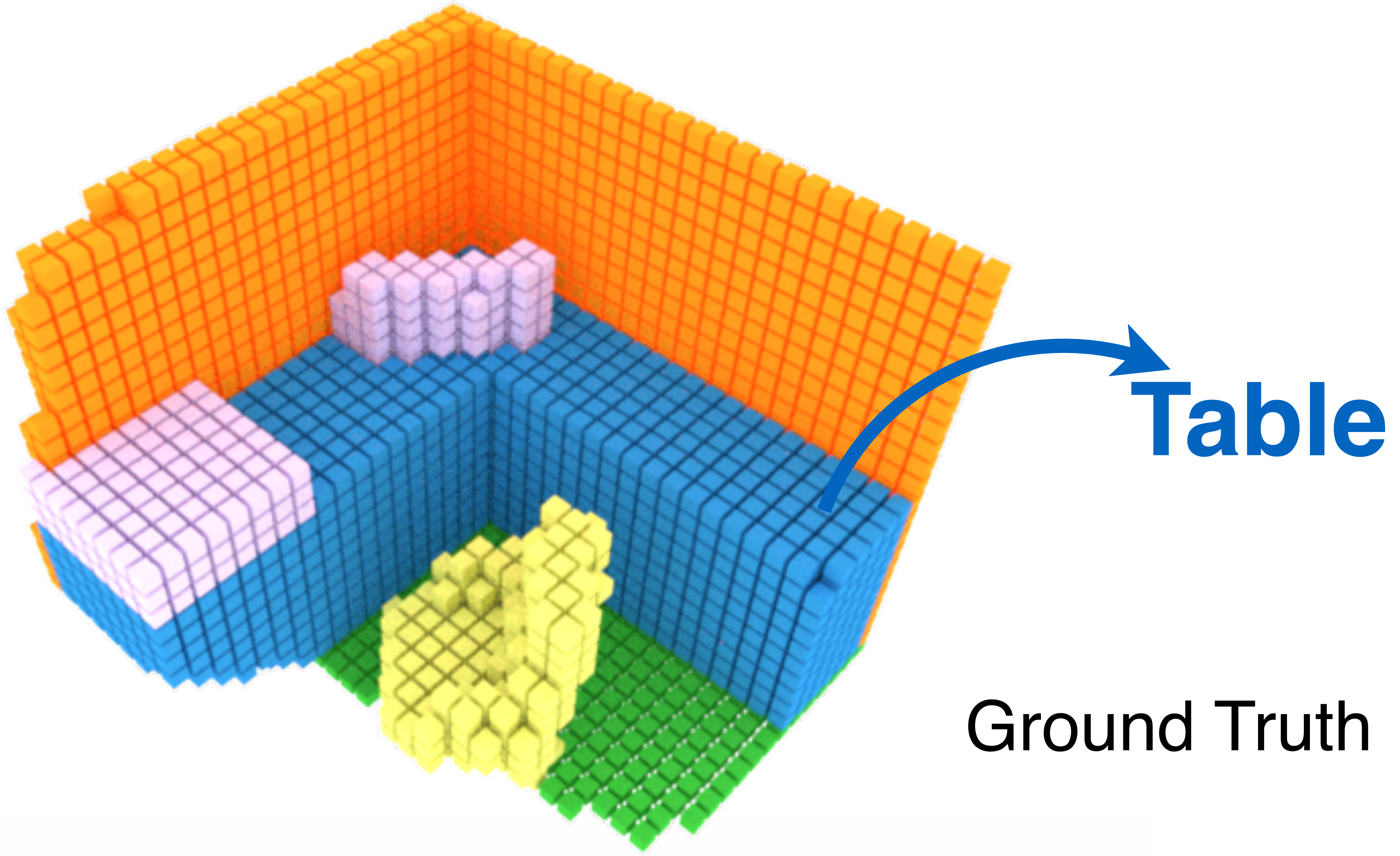
Failure cases

Failure cases: missing fine structures



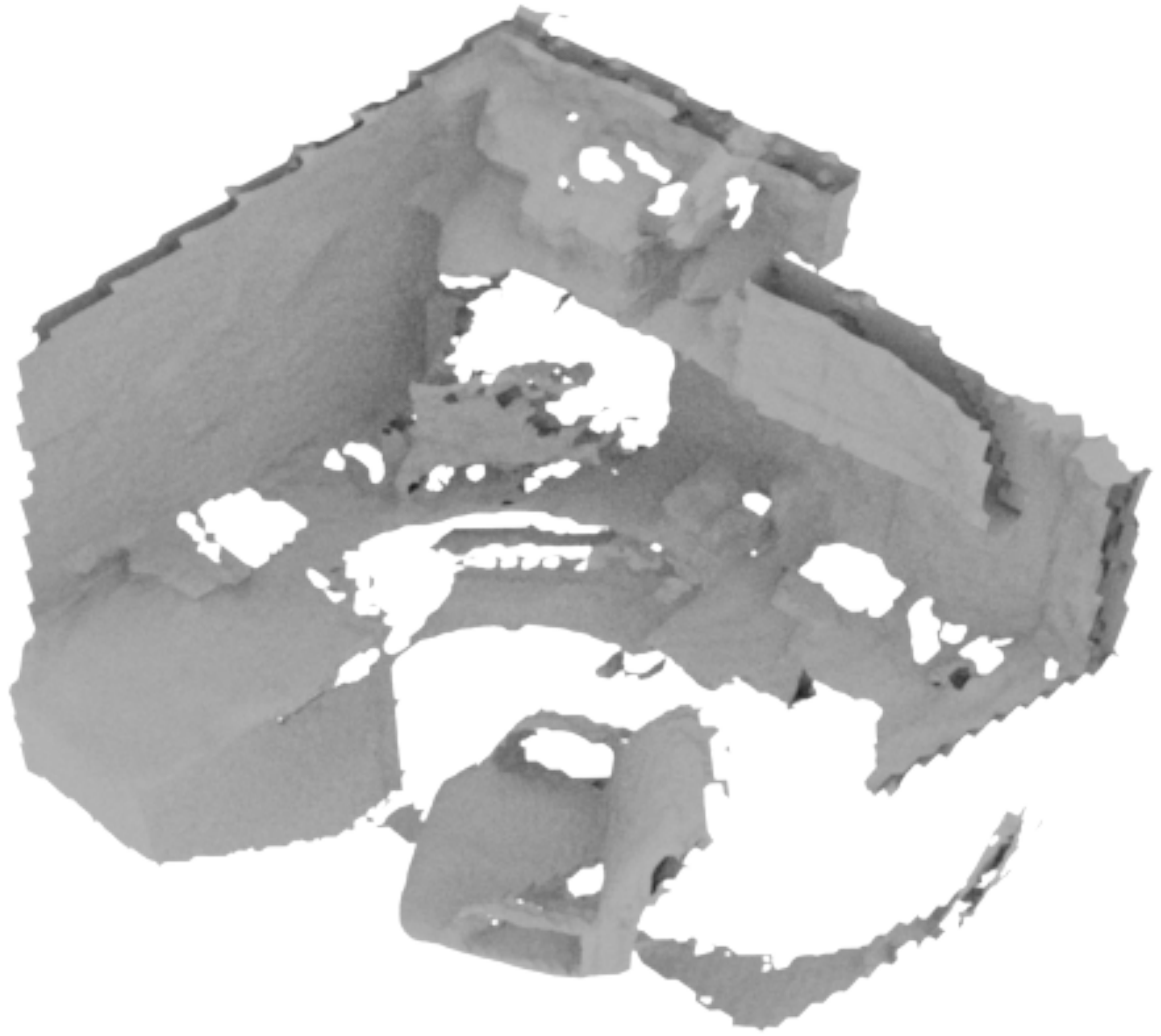
Failure cases: wrong object category

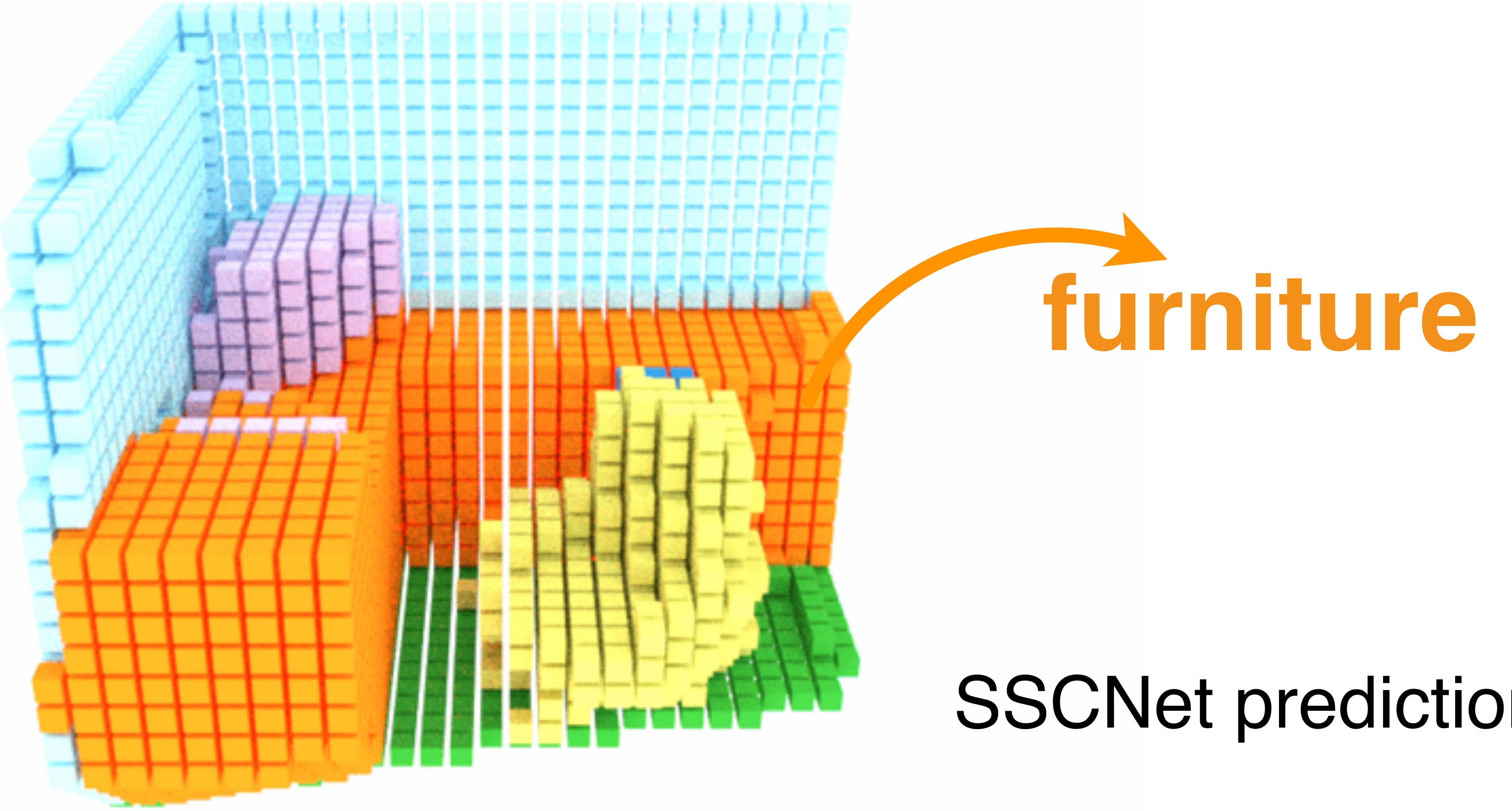
Color Image



Ground Truth

Observed Surface





SSCNet prediction

floor wall window chair bed sofa table tvs furn. objects