Learning to Reconstruct and Understand the 3D World

Songyou Peng

ETH Zurich and Max Planck Institute for Intelligent Systems

Microsoft Mixed Reality & AI Lab – Zurich

May 31, 2023

Who Am I?

- **Final-year PhD Student**
  - Marc Pollefeys
  - Andreas Geiger
- **Internships during PhD**
  - 2021: Michael Zollhoefer
  - 2022: Tom Funkhouser
- Before PhD, worked in Singapore, and interned at INRIA and TUM

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Meta

Google Research

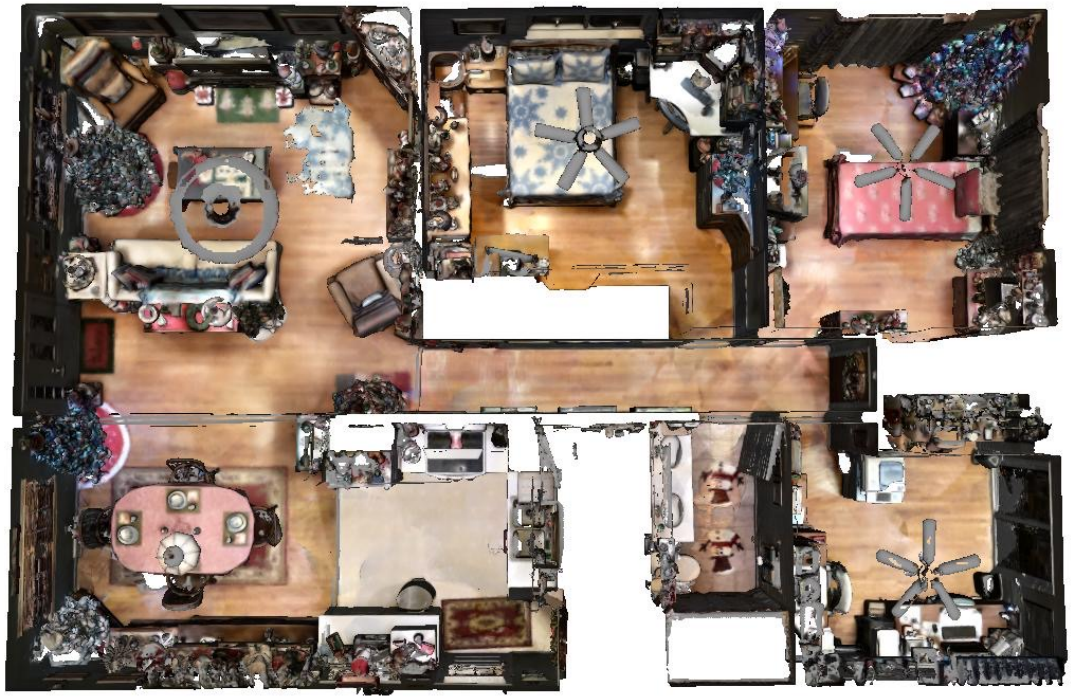
pengsongyou.github.io

Motivation

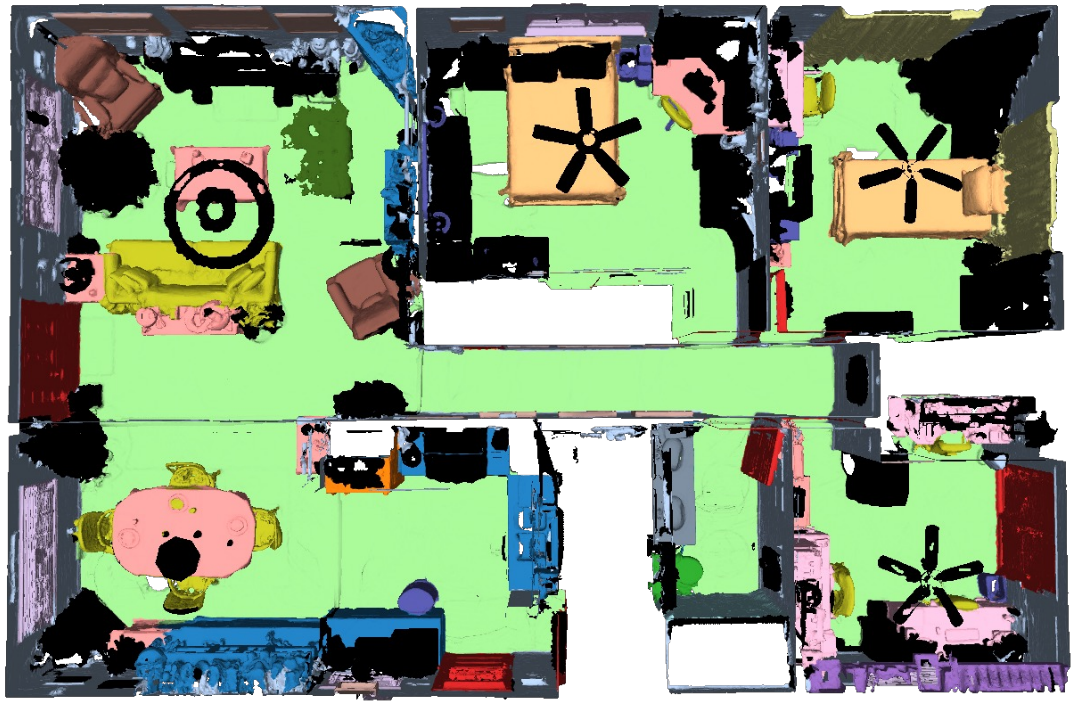
Input Images

3D Reconstruction

Motivation

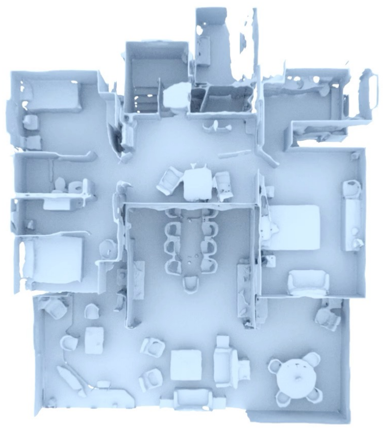


3D Reconstruction

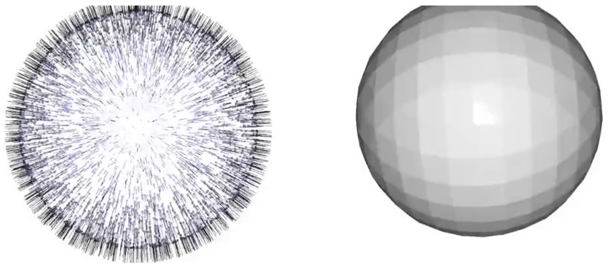


3D Scene Understanding

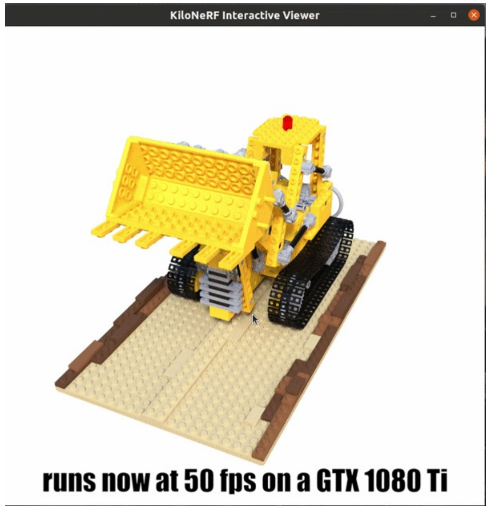
My PhD Topics: Neural Scene Representations
for 3D reconstruction and 3D scene understanding



Convolutional Occupancy Nets
ECCV 2020 (Spotlight)

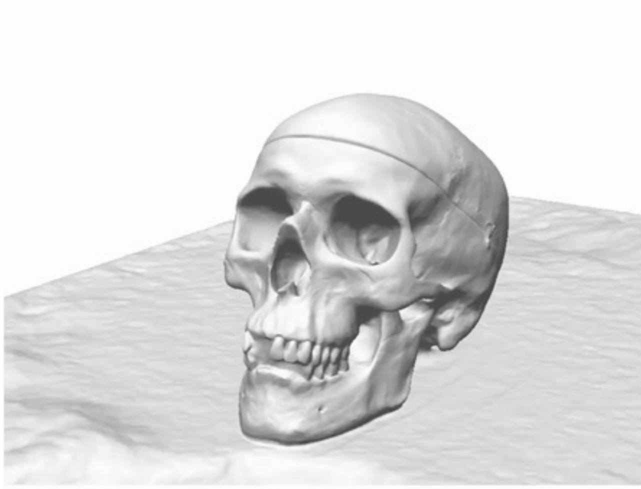


Shape As Points
NeurIPS 2021 (Oral)



runs now at 50 fps on a GTX 1080 Ti

KiloNeRF
ICCV 2021

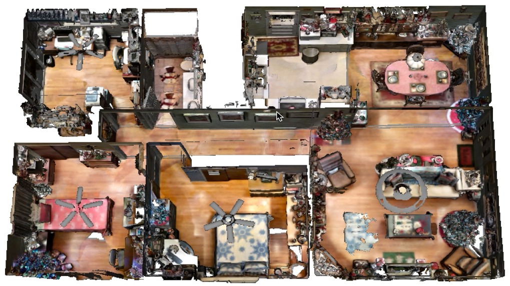


Ours

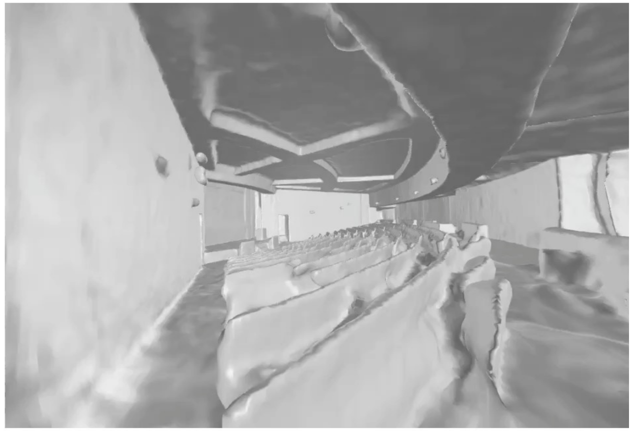
UNISURF
ICCV 2021 (Oral)

floor

NICE-SLAM
CVPR 2022

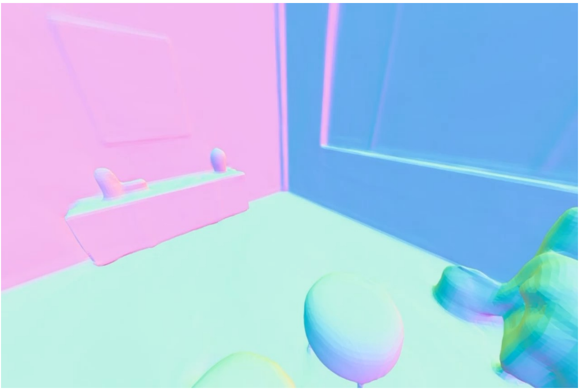


OpenScene
CVPR 2023



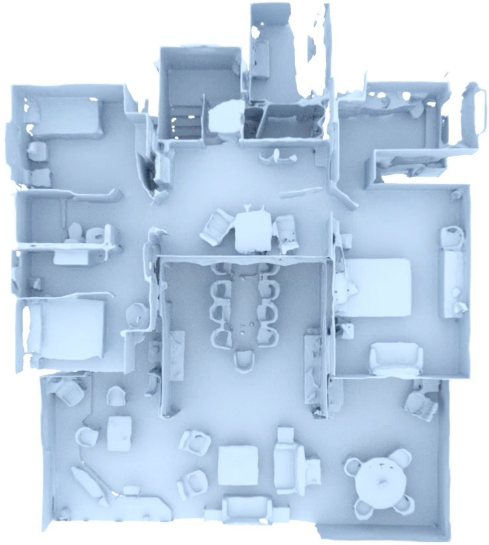
Ours

MonoSDF
NeurIPS 2022



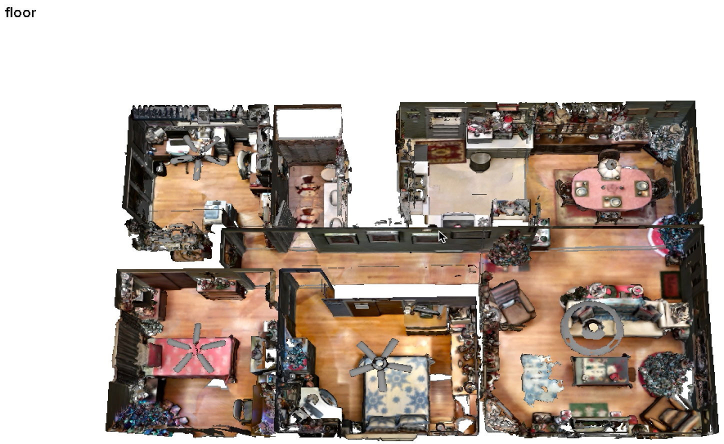
NICER-SLAM
arXiv 2023

My PhD Topics: Neural Scene Representations
for 3D reconstruction and 3D scene understanding



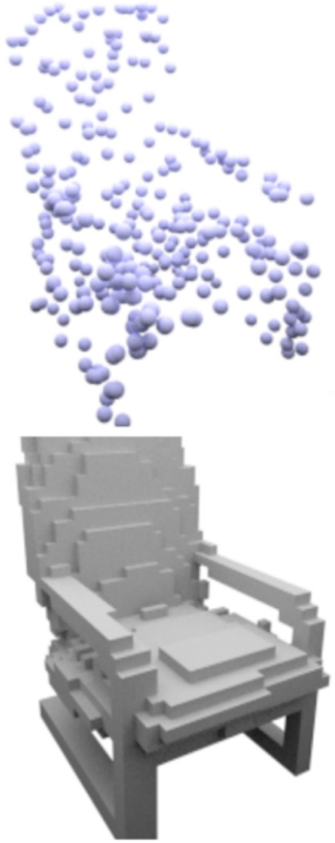
Convolutional Occupancy Networks
ECCV 2020 (Spotlight)

NICE-SLAM
CVPR 2022

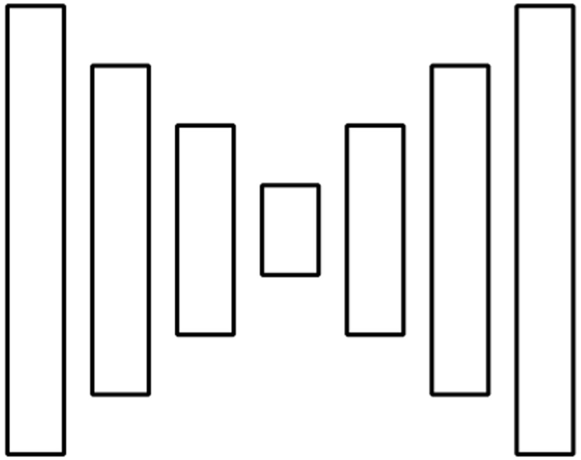


OpenScene
CVPR 2023

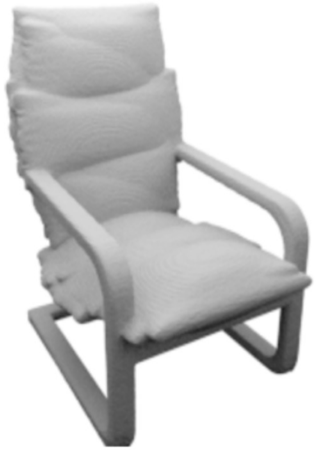
Learning-based 3D Surface Reconstruction



Input



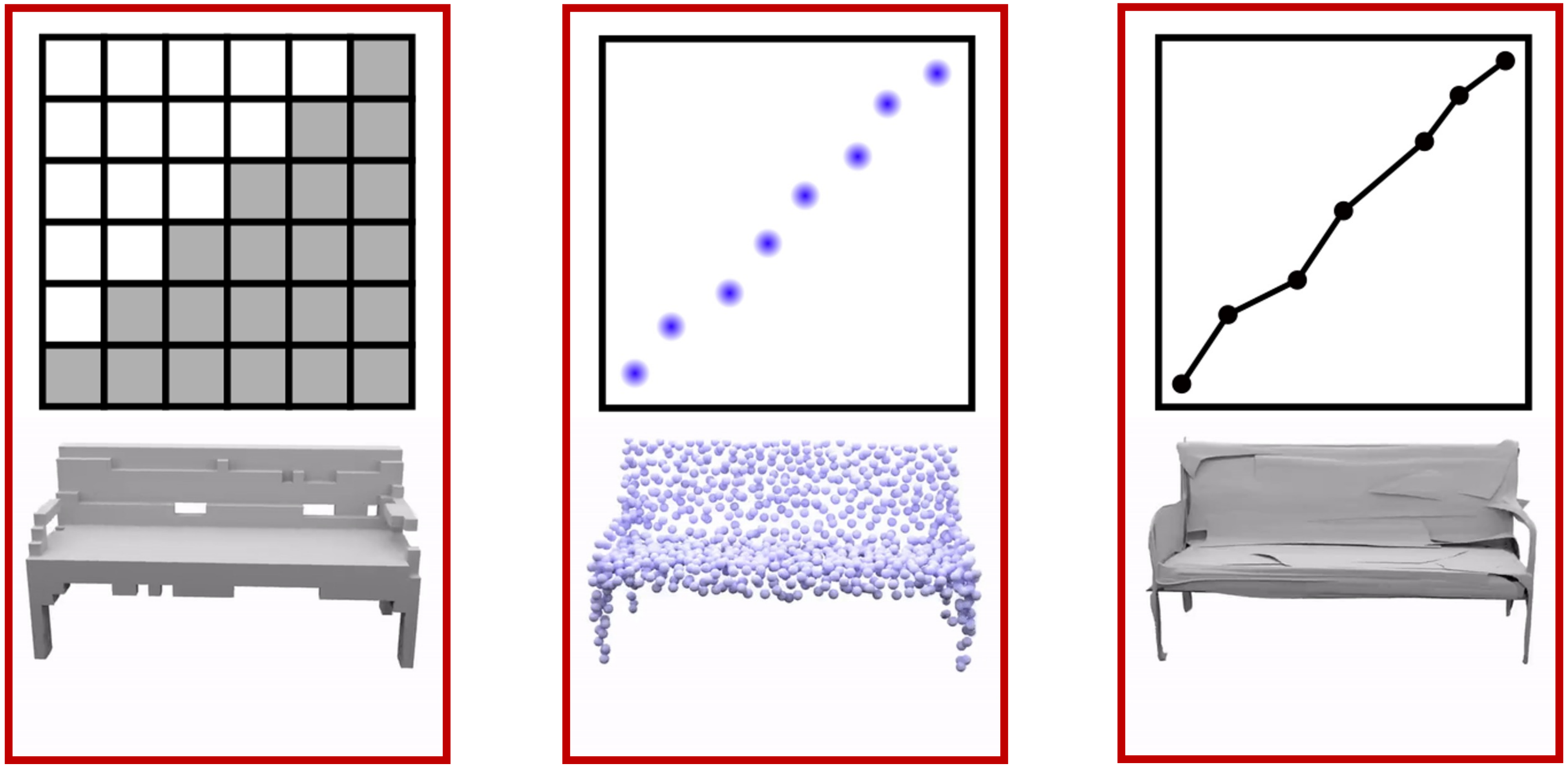
Neural Network



3D
Reconstruction

What is a good **3D representation**?

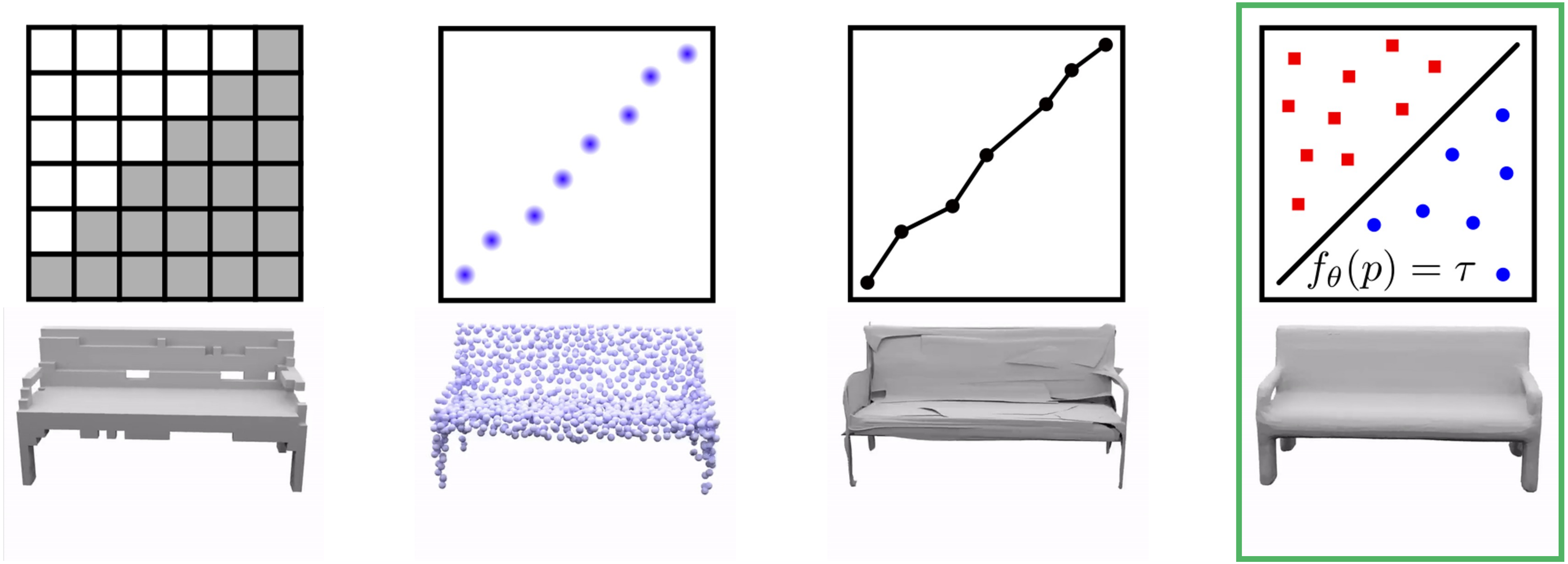
3D Representations



- Traditional Explicit Representations ⇒ **Discrete**

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

3D Representations

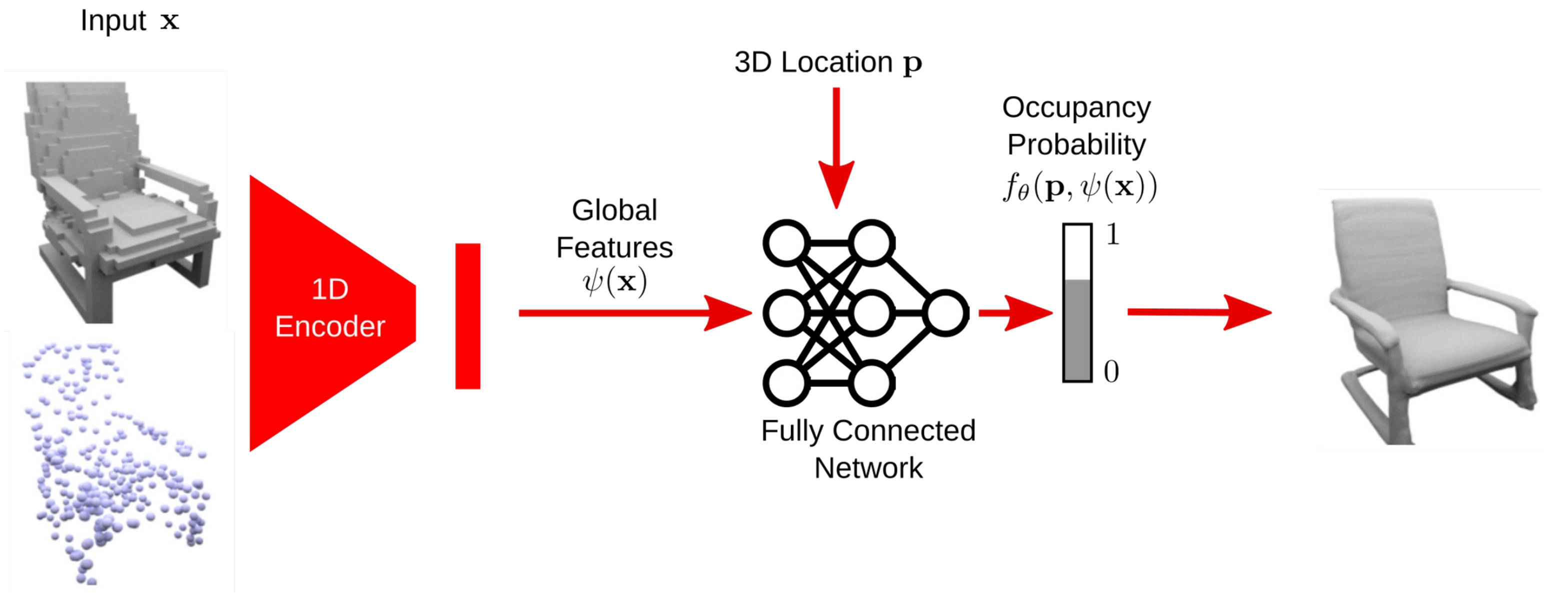


- Traditional Explicit Representations ⇒ **Discrete**
- Implicit Neural Representation ⇒ **Continuous**

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

Limitations

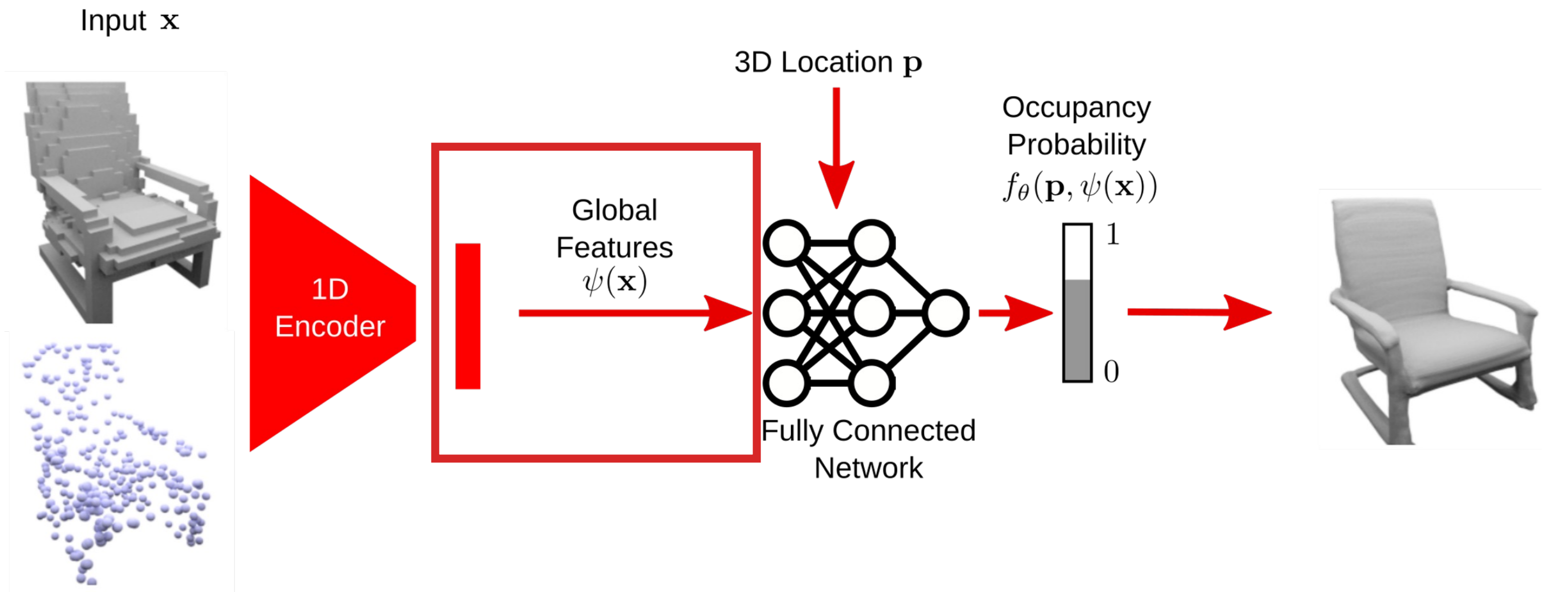
Structure of neural implicit representations:


Input $\mathbf{x}$
1D Encoder
Global Features $\psi(\mathbf{x})$
3D Location $\mathbf{p}$
Fully Connected Network
Occupancy Probability $f_\theta(\mathbf{p}, \psi(\mathbf{x}))$
1
0

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

Limitations

Structure of neural implicit representations:

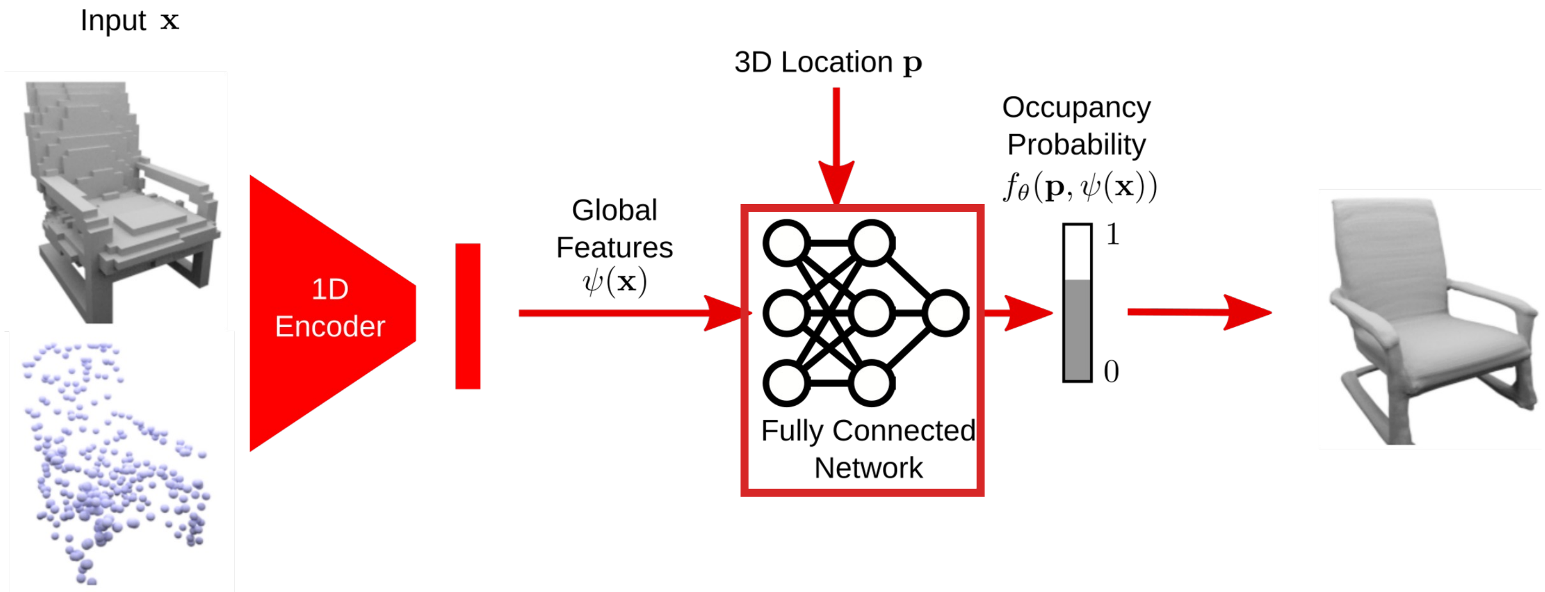

Input $\mathbf{x}$
1D Encoder
Global Features $\psi(\mathbf{x})$
3D Location $\mathbf{p}$
Fully Connected Network
Occupancy Probability $f_\theta(\mathbf{p}, \psi(\mathbf{x}))$
1
0

- Global latent code ⇒ **overly smooth geometry**

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

Limitations

Structure of neural implicit representations:



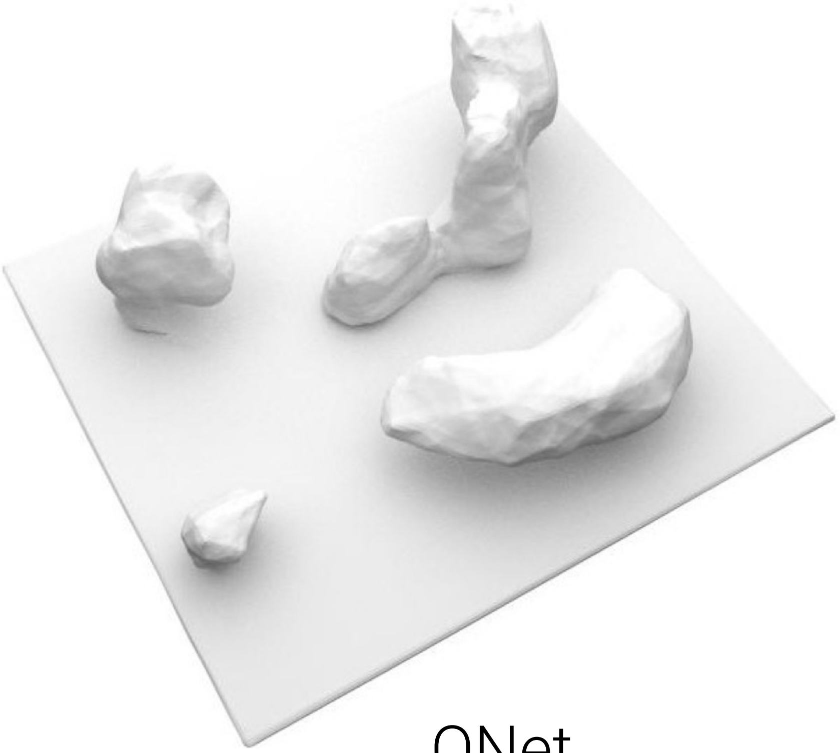
Input $\mathbf{x}$
1D Encoder
Global Features $\psi(\mathbf{x})$
3D Location $\mathbf{p}$
Fully Connected Network
Occupancy Probability $f_\theta(\mathbf{p}, \psi(\mathbf{x}))$
1
0

- Global latent code ⇒ **overly smooth geometry**
- Fully-connected architecture ⇒ **no translation equivariance**

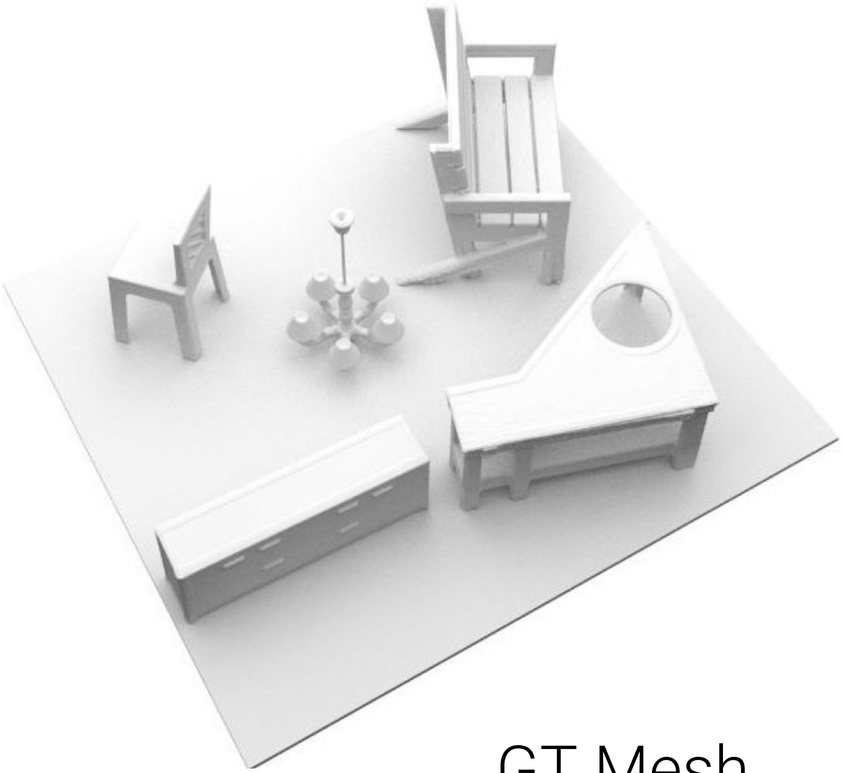
Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

Limitations

Implicit models work well for **simple objects** but poorly on **complex scenes**:



ONet



GT Mesh

Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

How to reconstruct large-scale 3D scenes with **neural implicit representations**?

ETH zürich

ECCV'20 ONLINE 23-28 AUGUST 2020

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Convolutional Occupancy Networks

Songyou Peng

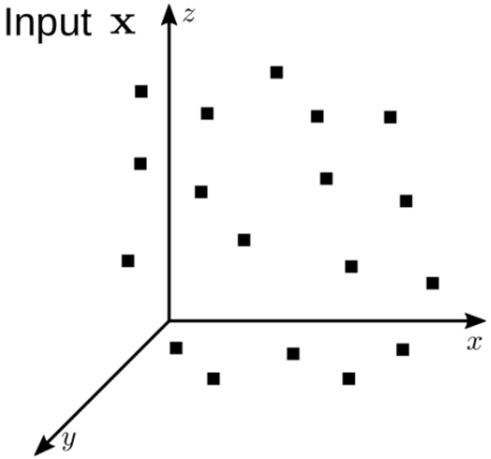
Michael Niemeyer

Lars Mescheder

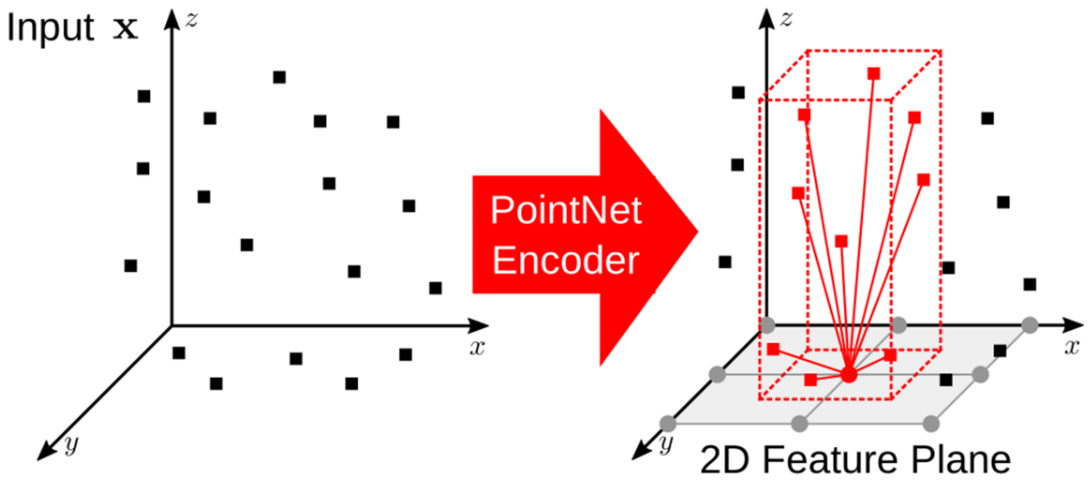
Marc Pollefeys

Andreas Geiger

Main Idea

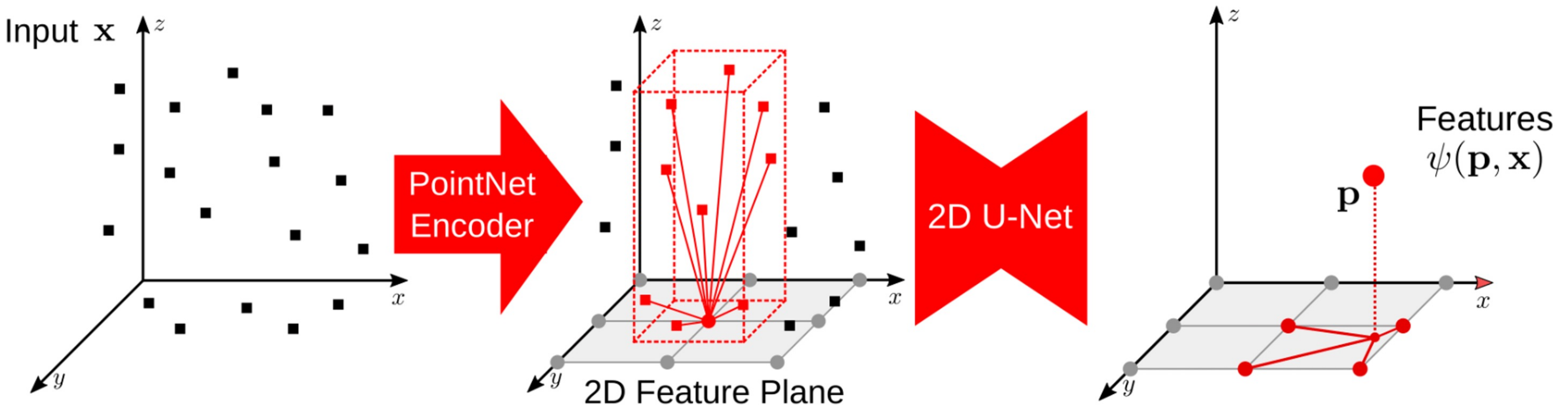


Main Idea



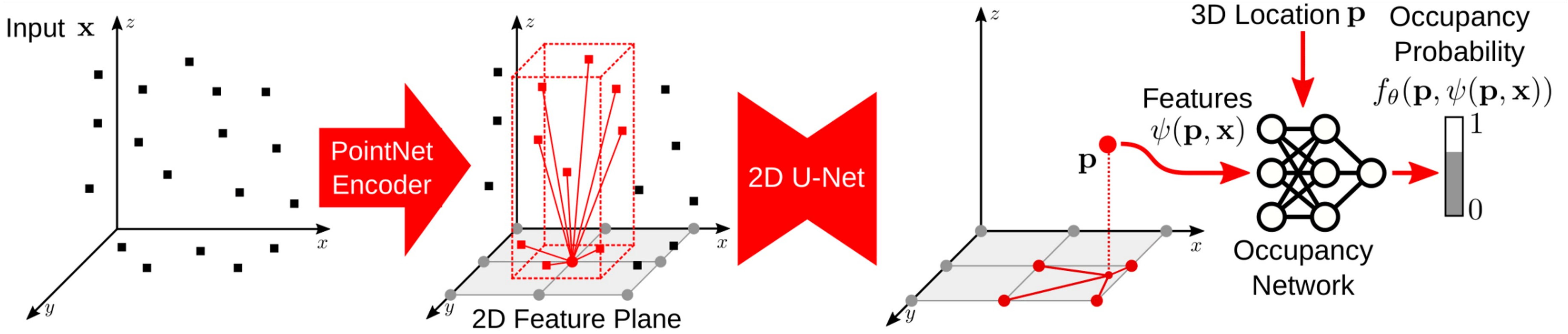
- **2D Plane Encoder**: Use a local PointNet to process input, project onto canonical plane

Main Idea



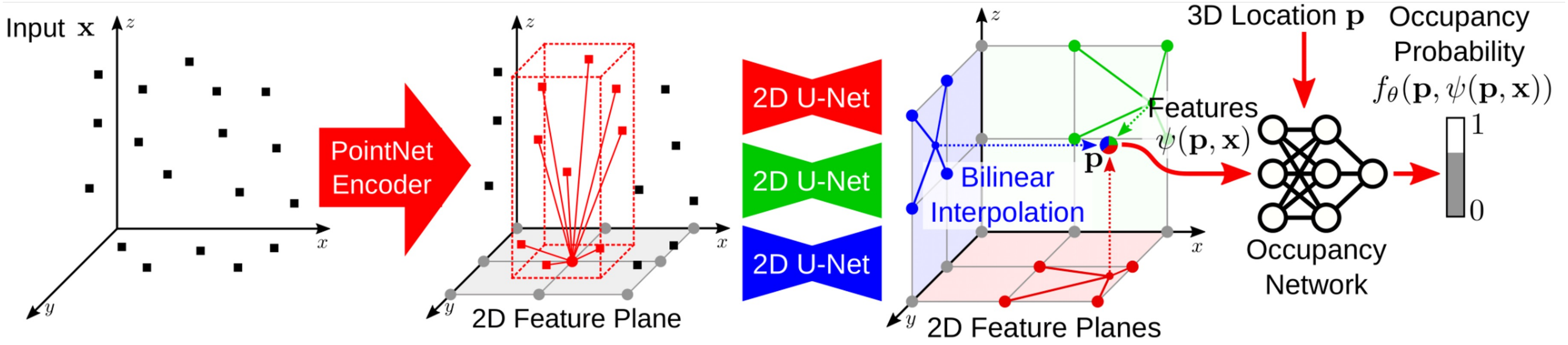
- **2D Plane Encoder**: Use a local PointNet to process input, project onto canonical plane
- **2D Plane Decoder**: Processed by U-Net, query features via bilinear interpolation

Main Idea



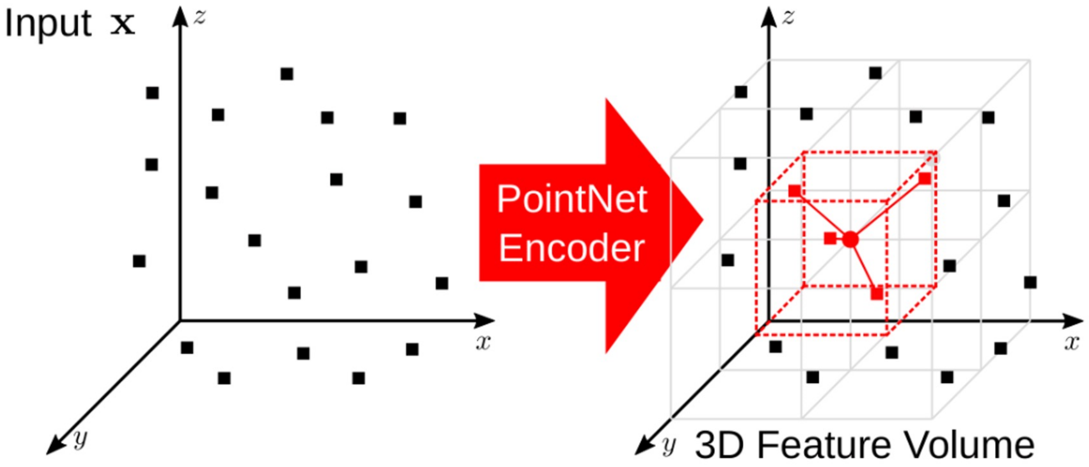
- **2D Plane Encoder**: Use a local PointNet to process input, project onto canonical plane
- **2D Plane Decoder**: Processed by U-Net, query features via bilinear interpolation
- **Occupancy Readout**: Shallow occupancy network $f_\theta(\cdot)$

Main Idea



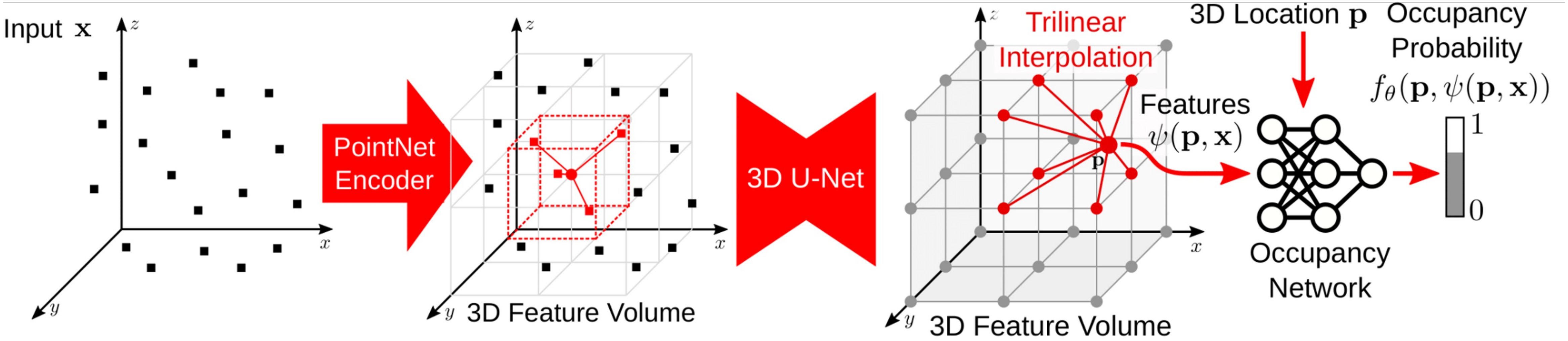
- **2D Plane Encoder**: Use a local PointNet to process input, project onto **3-canonical planes**
- **2D Plane Decoder**: Processed by U-Net, query features via bilinear interpolation
- **Occupancy Readout**: Shallow occupancy network $f_\theta(\cdot)$

Main Idea – 3D



- **3D Volume Encoder**: Use a local PointNet to process input, volumetric feature encoding

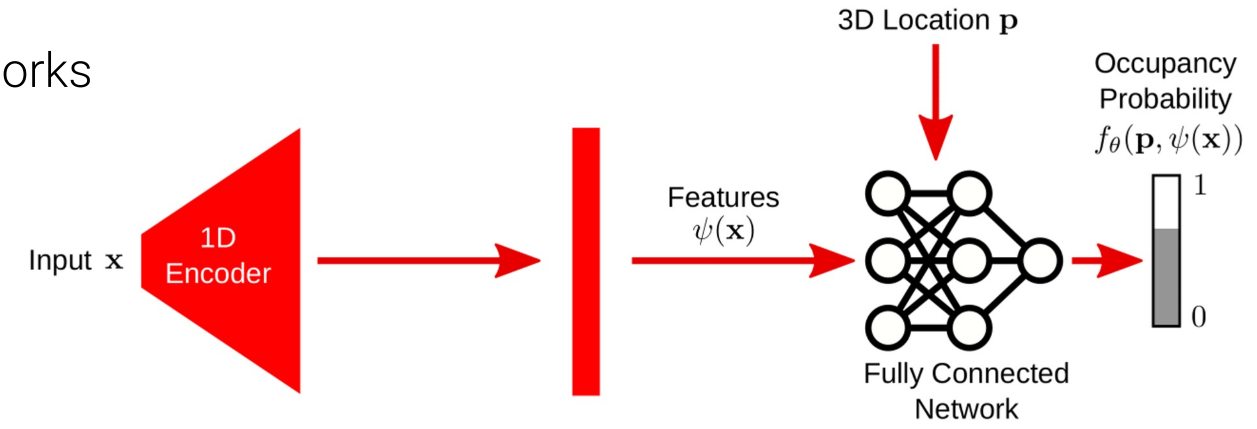
Main Idea – 3D



- **3D Volume Encoder**: Use a local PointNet to process input, volumetric feature encoding
- **3D Volume Decoder**: Processed by 3D U-Net, query features via trilinear interpolation
- **Occupancy Readout**: Shallow occupancy network $f_\theta(\cdot)$

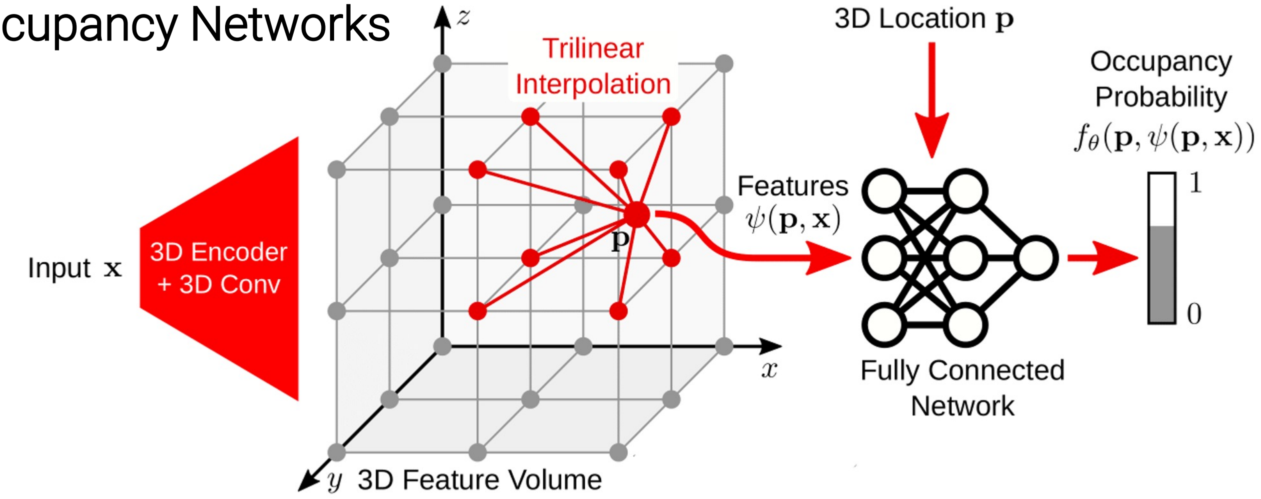
Comparison

Occupancy Networks


Input $\mathbf{x}$
1D Encoder
Features $\psi(\mathbf{x})$
3D Location $\mathbf{p}$
Occupancy Probability $f_\theta(\mathbf{p}, \psi(\mathbf{x}))$
1
0
Fully Connected Network

- − global feature
- − heavy FC network
- − no translation equivariance

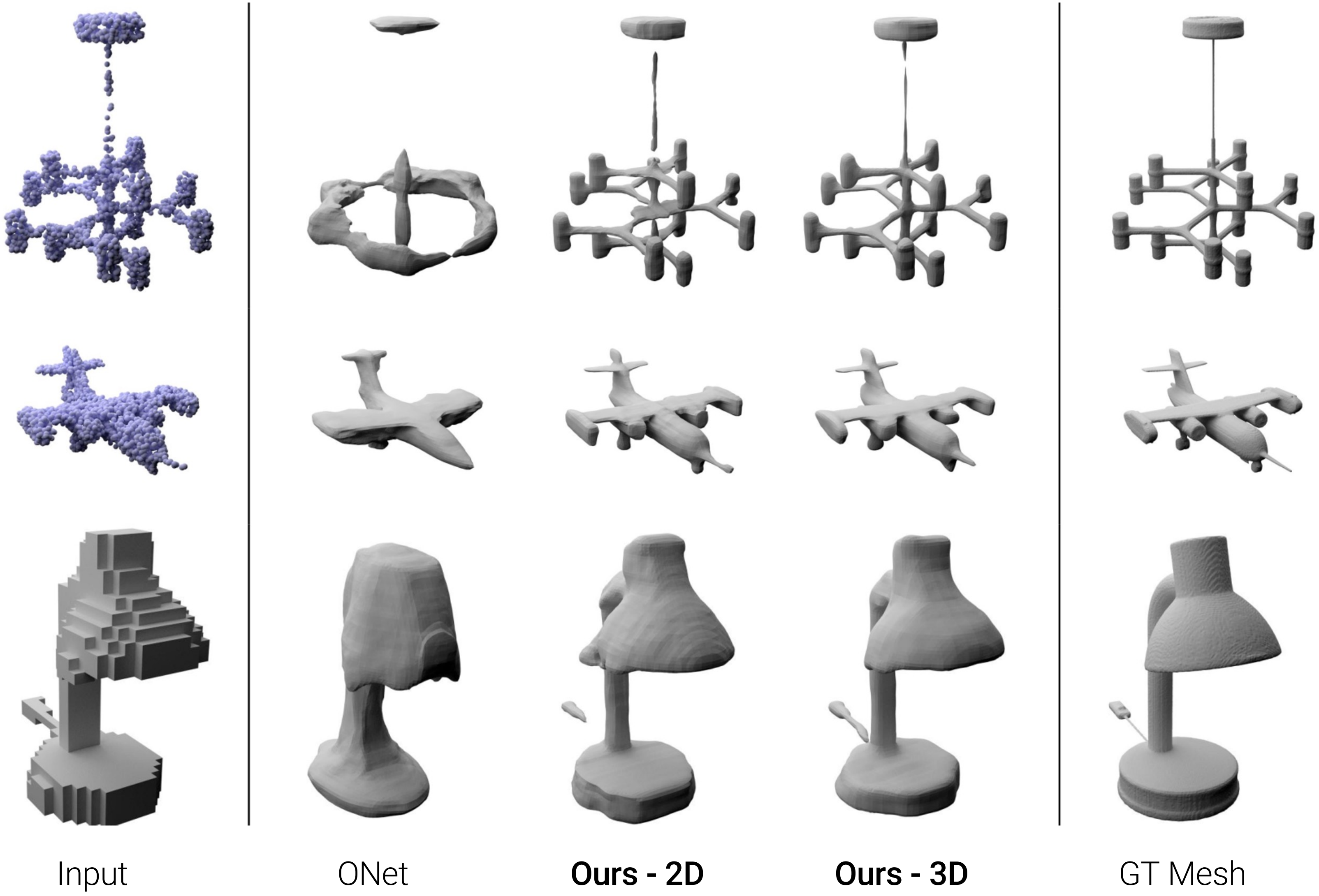
Convolutional Occupancy Networks



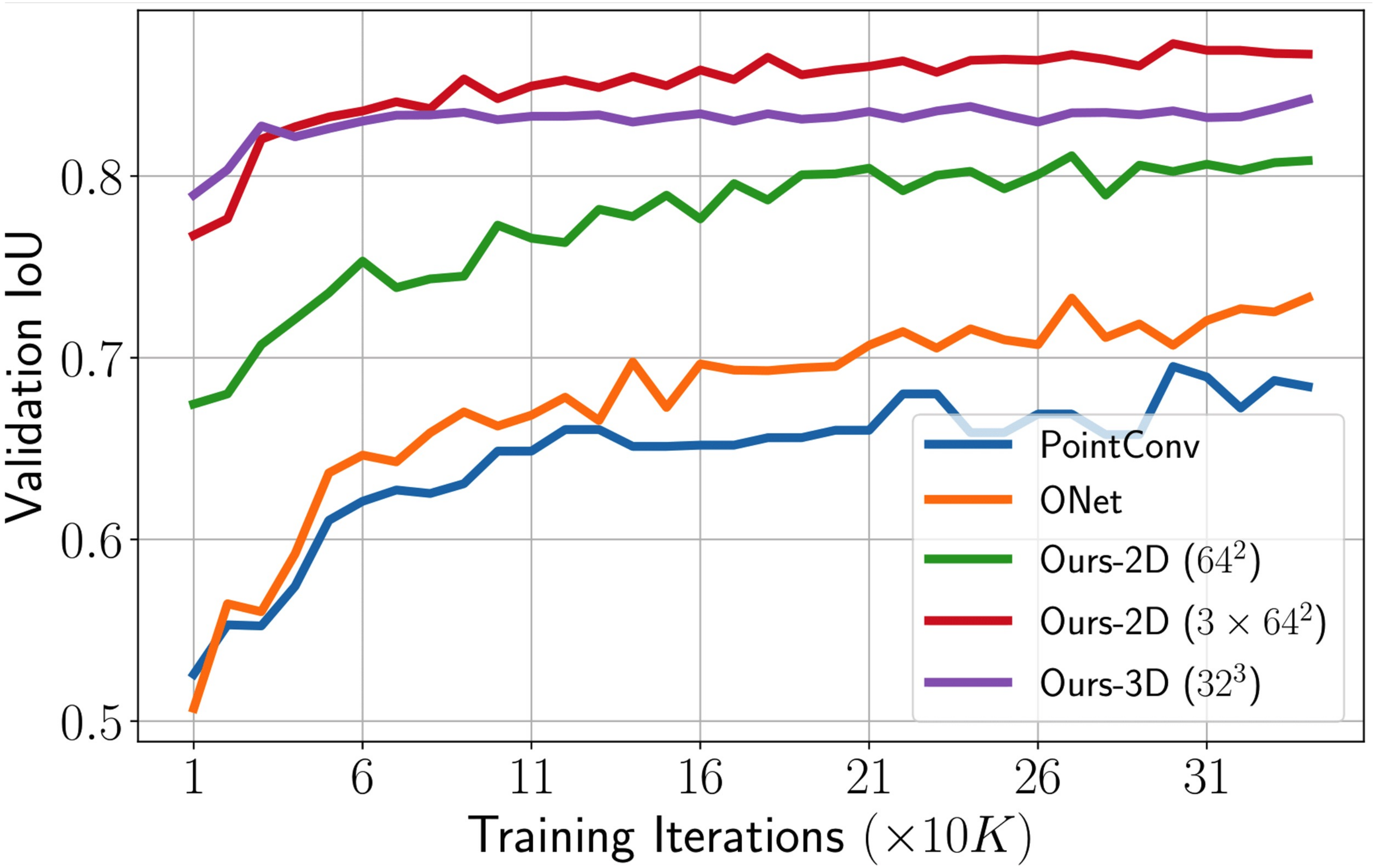
- + local feature
- + shallow FC network
- + translation equivariance

Results

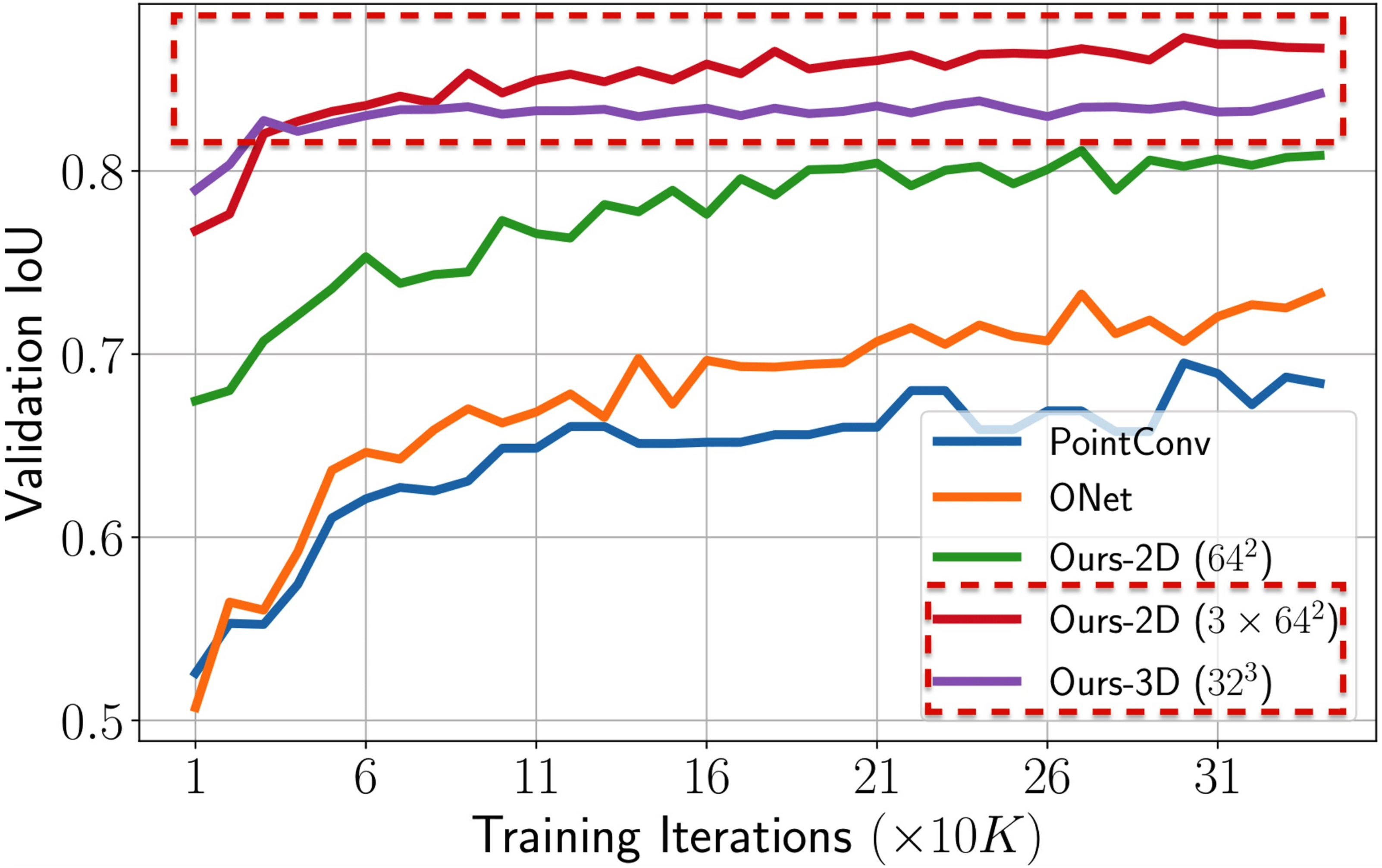
Object-Level Reconstruction



Training Speed

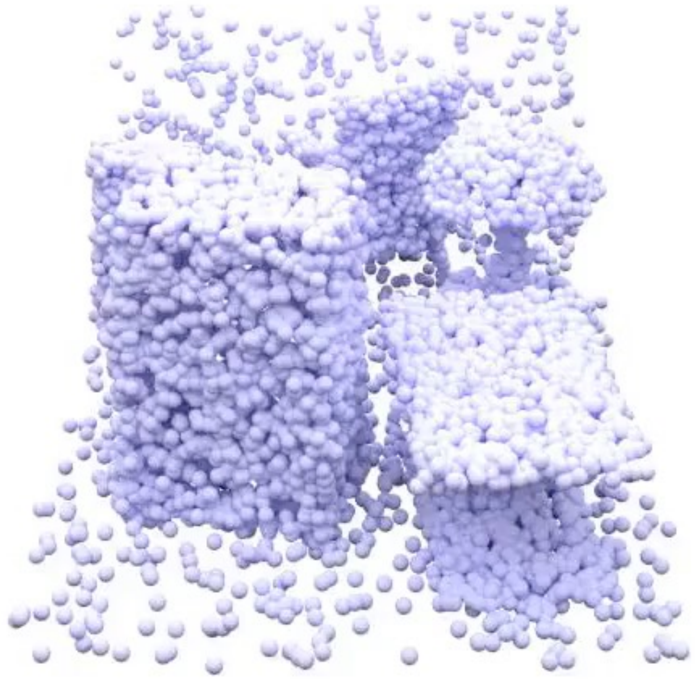


Training Speed

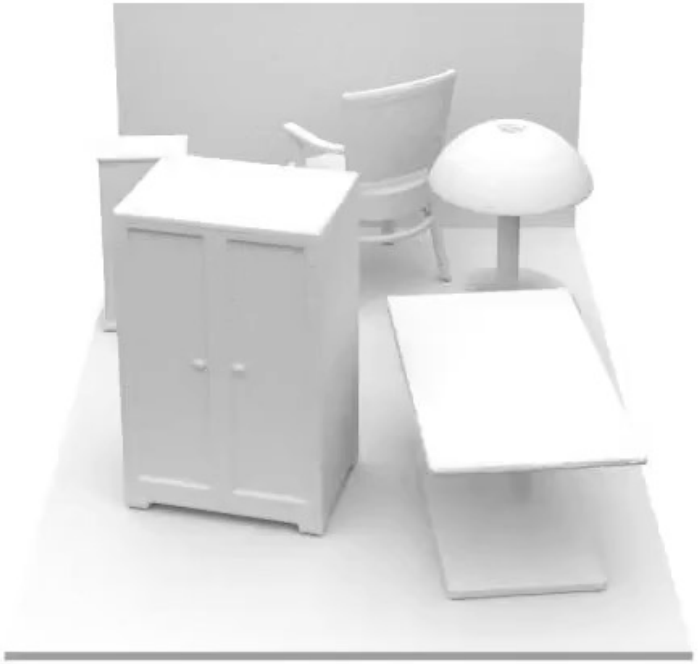


Scene-Level Reconstruction: Synthetic

- Trained and evaluated on synthetic rooms



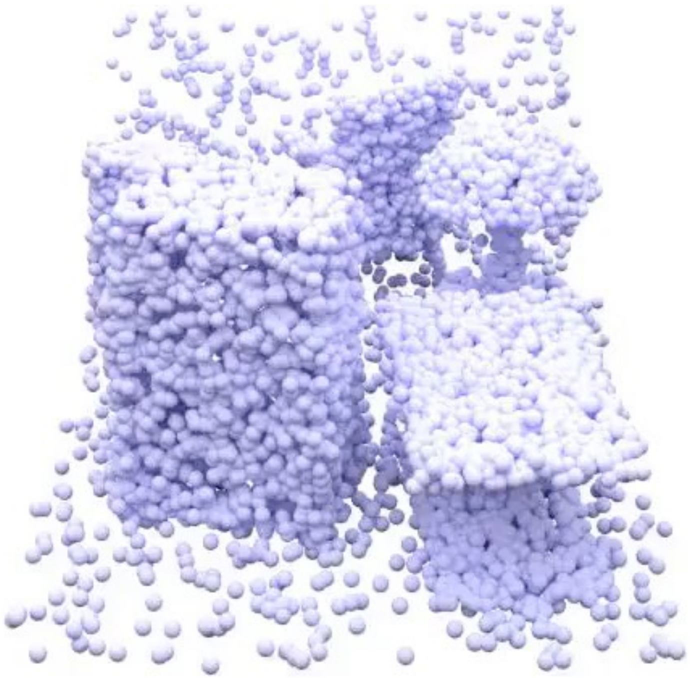
Input



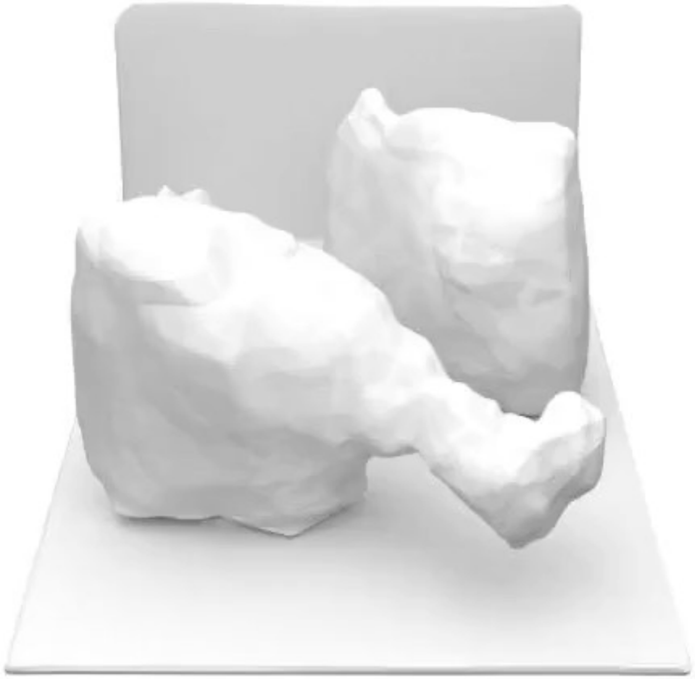
GT Mesh

Scene-Level Reconstruction: Synthetic

- ONet fails on room-level reconstruction



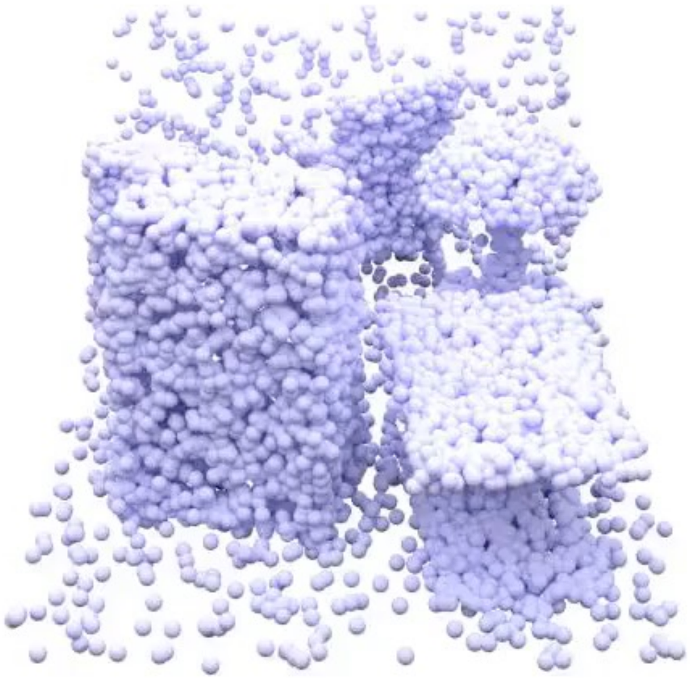
Input



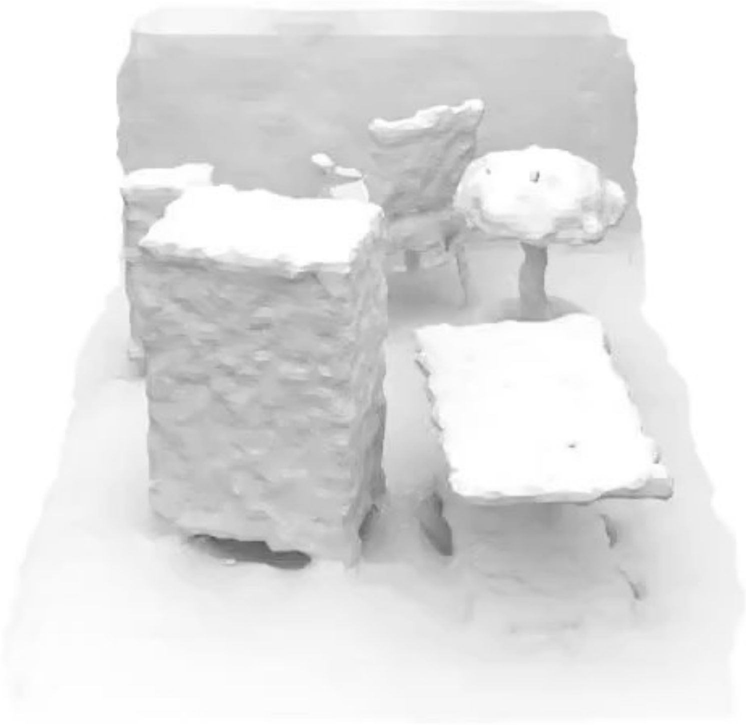
ONet

Scene-Level Reconstruction: Synthetic

- SPSR requires surface normals, output is noisy



Input

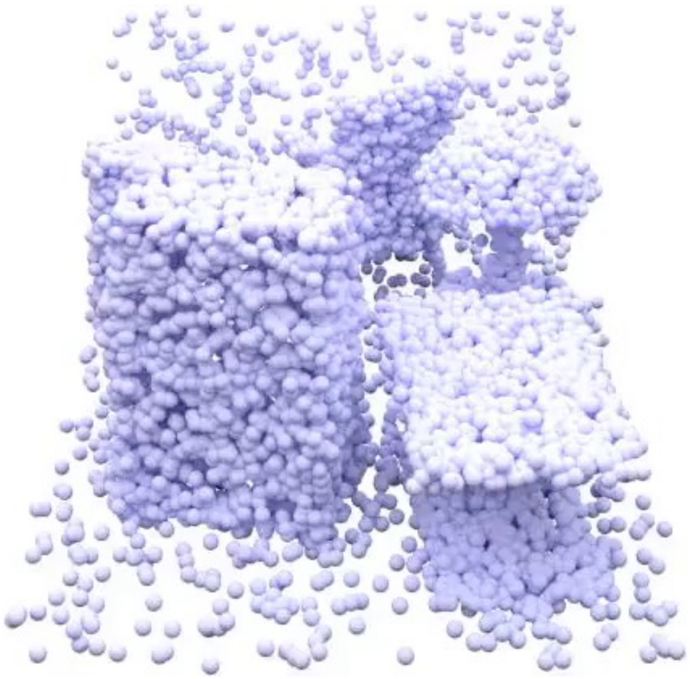


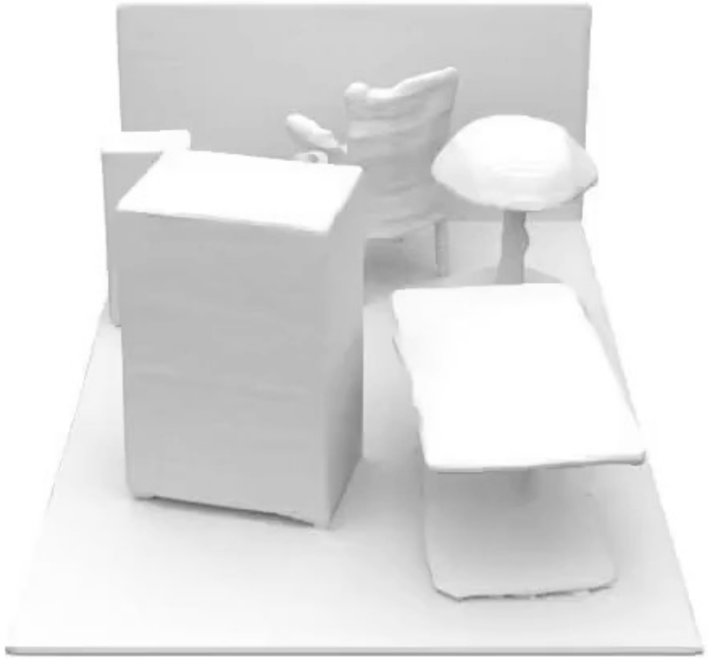
SPSR

(Screened Poisson Surface Reconstruction)

Scene-Level Reconstruction: Synthetic

- Our method preserves better details





Input

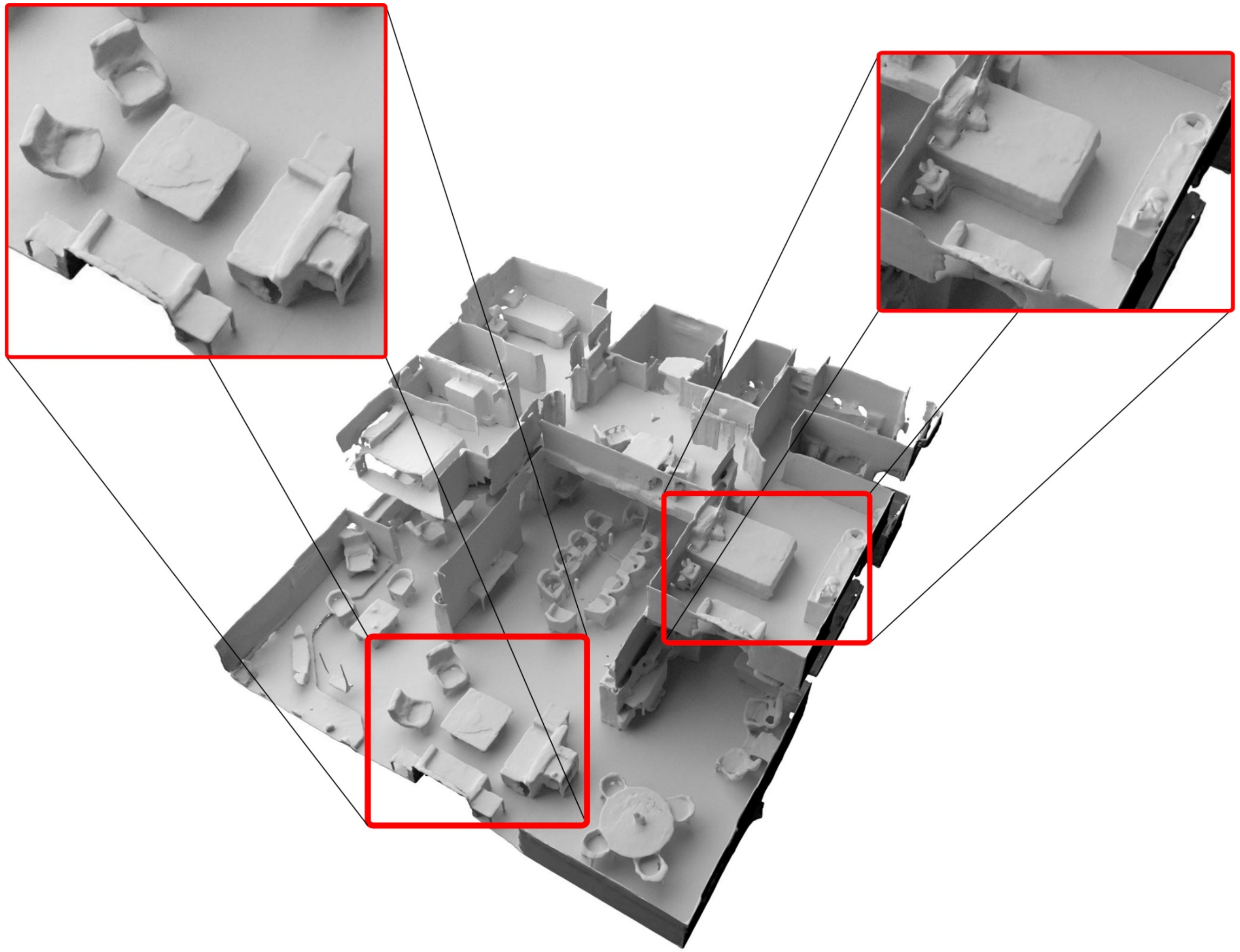
Ours

Large-Scale Reconstruction

Scene size: 15.7m x 12.3m x 4.5m

Results on Matterport3D

- Fully convolutional model
- Trained on synthetic crops
- Sliding-window evaluation
- Scale to any scene size



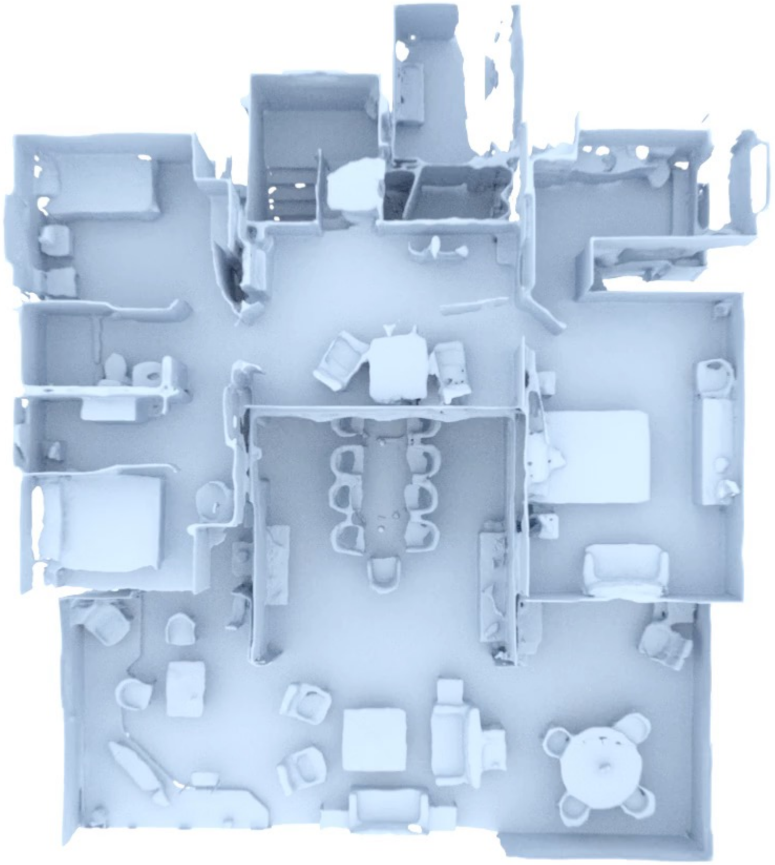
Our reconstruction output

Large-Scale Reconstruction

Scene size: 15.7m x 12.3m x 4.5m

Results on Matterport3D

- Fully convolutional model
- Trained on synthetic crops
- Sliding-window evaluation
- Scale to any scene size



Our reconstruction output

Take-home Messages

- Introduce 3 different expressive hybrid representations for neural fields
- CNN's translation equivariance enables to reconstruct large scenes
- The "**tri-plane**" representation became VERY popular
    - Especially in the **NeRF era**, see e.g. EG3D [CVPR'21], TensoRF [ECCV'22]

Limitations

- Not rotational equivariance

NeRF is awesome!

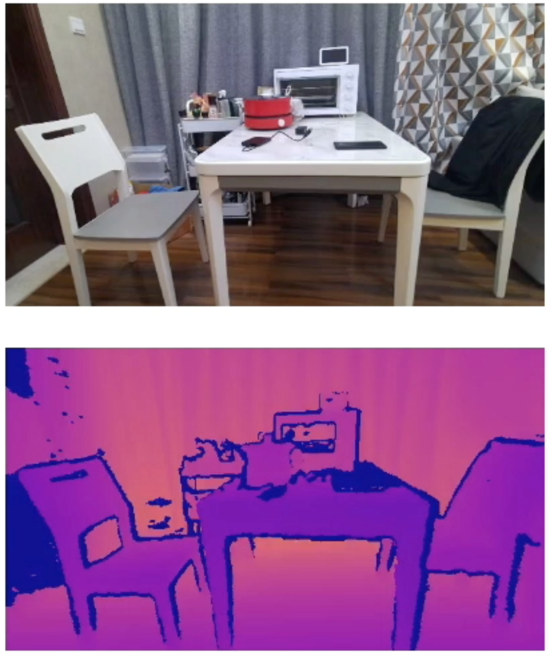
Some existing problems…

😢 Poor underlying geometry

😢 Camera poses needed

Mildenhall*, Srinivasan*, Tancik* et al: NeRF : Representing Scenes as Neural Radiance Fields for View Synthesis. ECCV 2020

RGB-D Sequences



40x Speed

NICE-SLAM

Neural Implicit Scalable Encoding for SLAM

CVPR 2022

Zihan Zhu* Songyou Peng* Viktor Larsson Weiwei Xu Hujun Bao
Zhaopeng Cui Martin R. Oswald Marc Pollefeys

* Equal Contributions

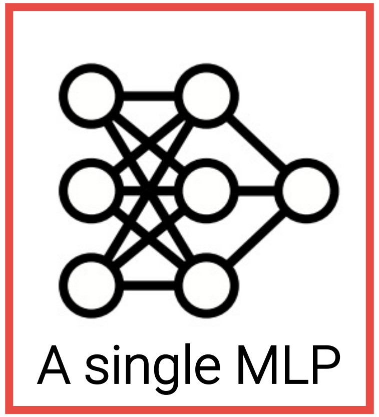
iMAP

[Sucar et al., ICCV’21]

First neural implicit-based online SLAM system

iMAP

[Sucar et al., ICCV’21]

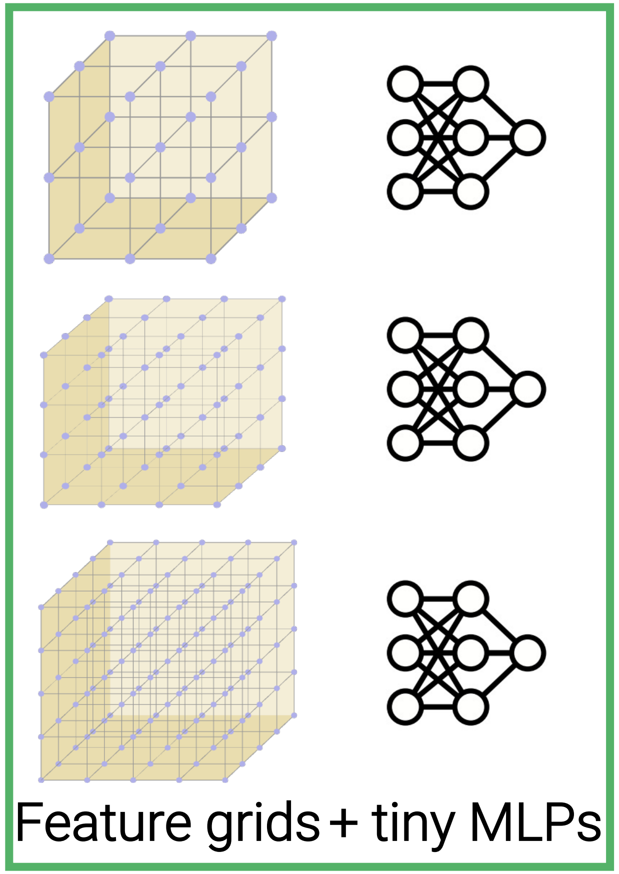


- Fail when scaling up to larger scenes
- Global update → Catastrophic forgetting
- Slow convergence

— Predicted Poses

— GT Poses

NICE-SLAM

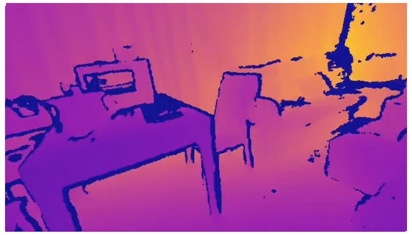


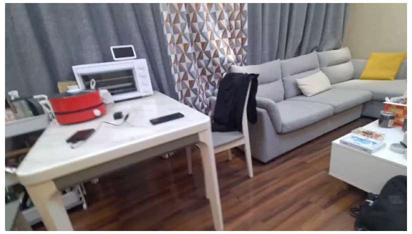
- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence

🟥 Predicted Poses
— GT Poses

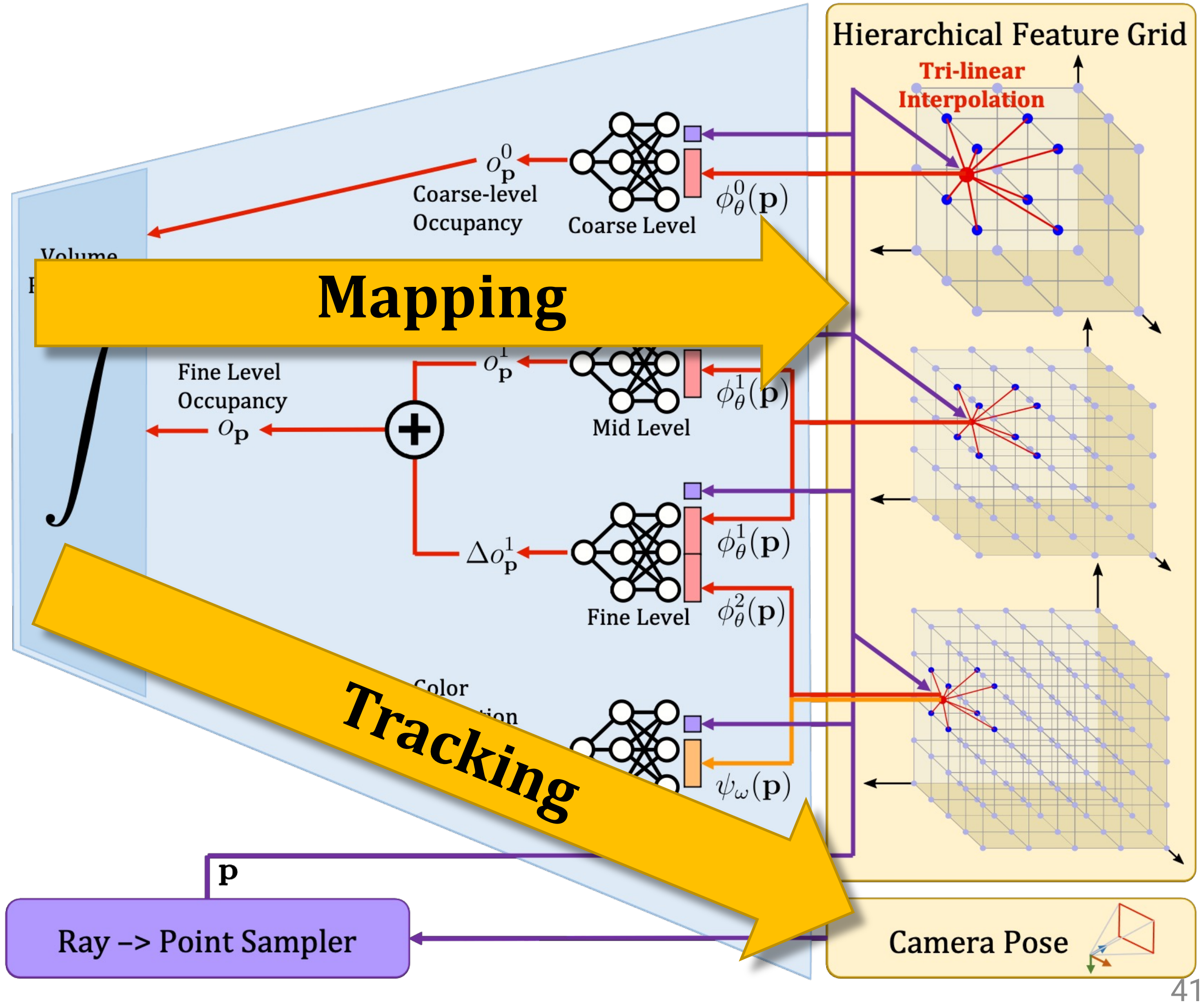
Pipeline

Input Depth





Input RGB



Results

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

4x Speed

— Predicted Poses

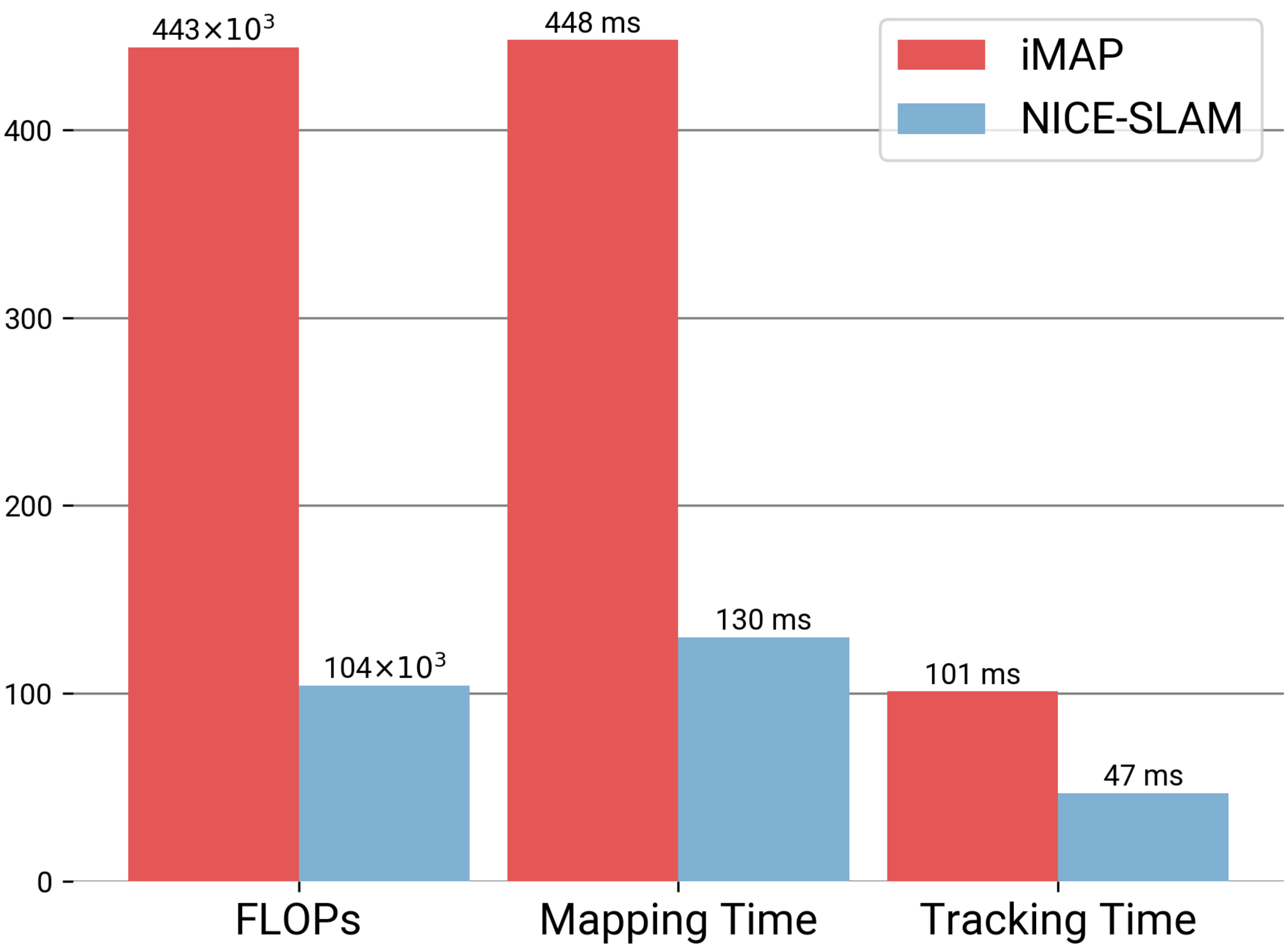
— GT Poses

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

10x Speed



Note: Runtime evaluation setting from iMAP paper, not the best-performing setting

Take-home Message

- A NICE NeRF-based SLAM system for indoor scenes
- Hierarchical feature grids + a tiny MLP **seems to be a trend**!
    - Instant-NGP [SIGGRAPH’22 Best Paper]

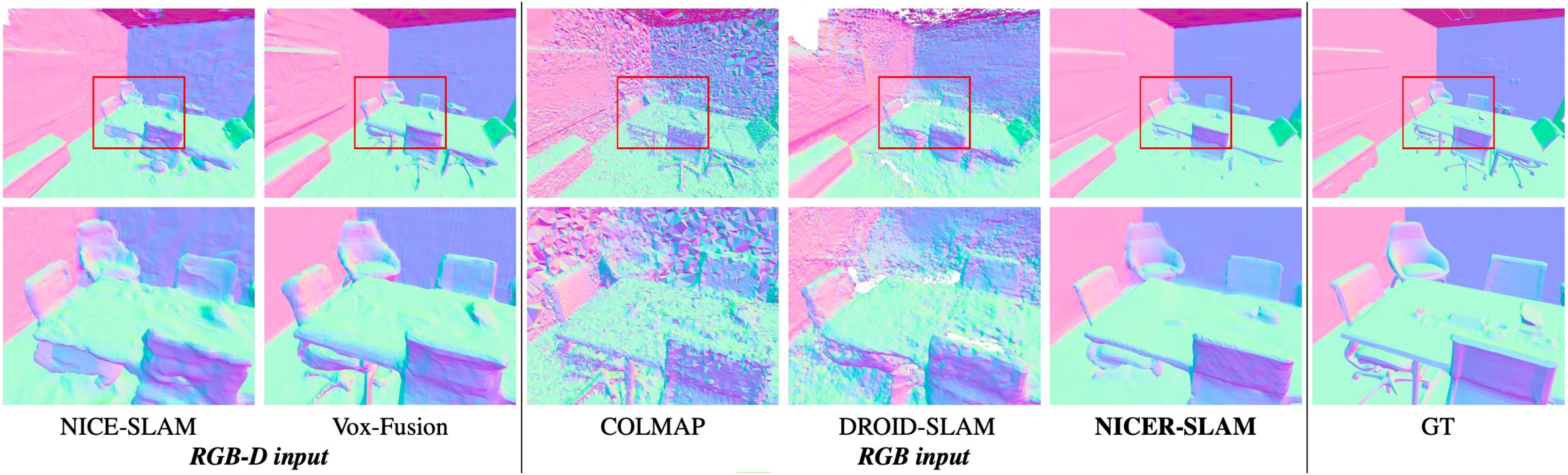
Limitations

- **Requires depths as input**
- Only bounded scenes
- Still not real-time

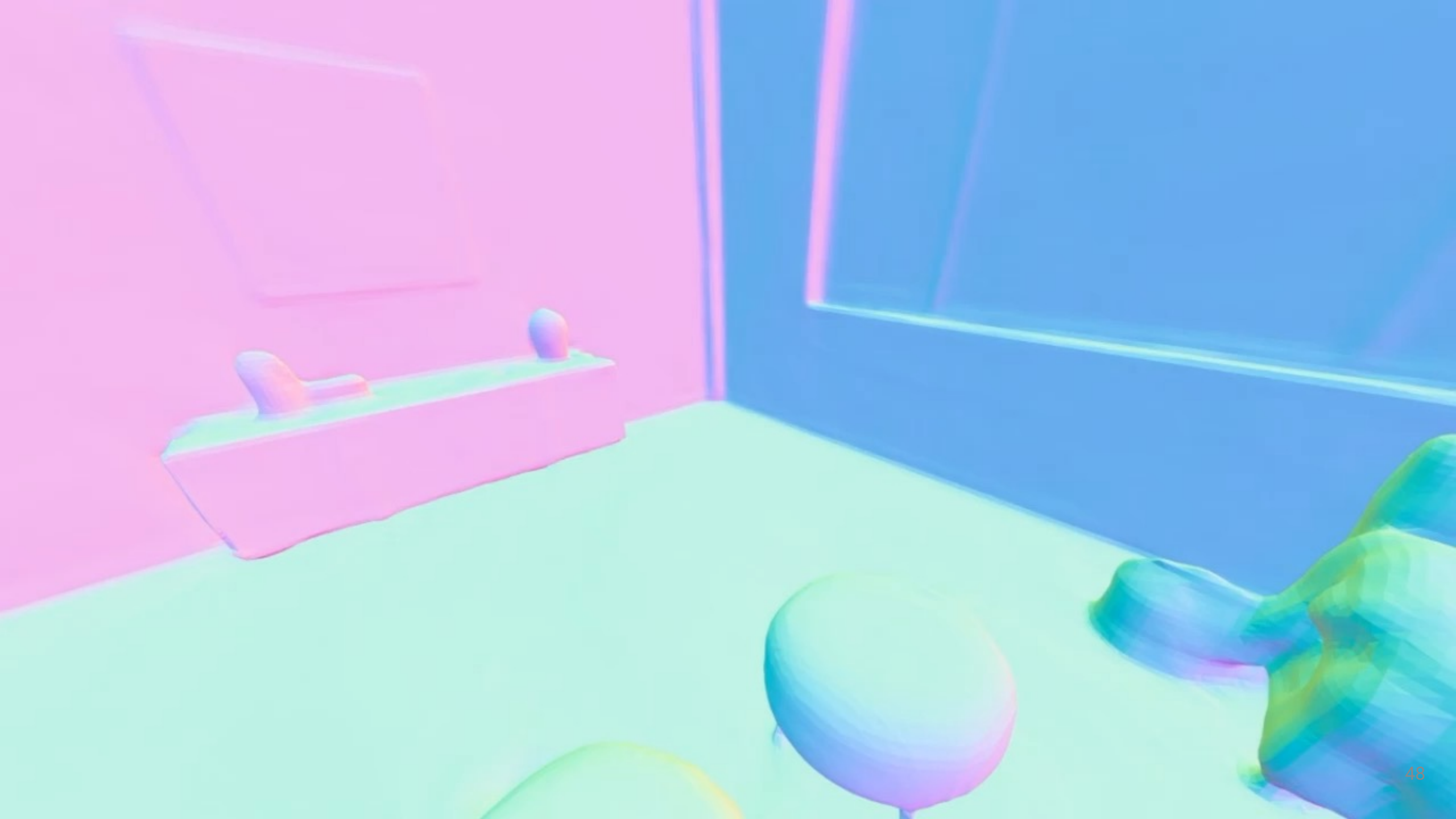
NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM

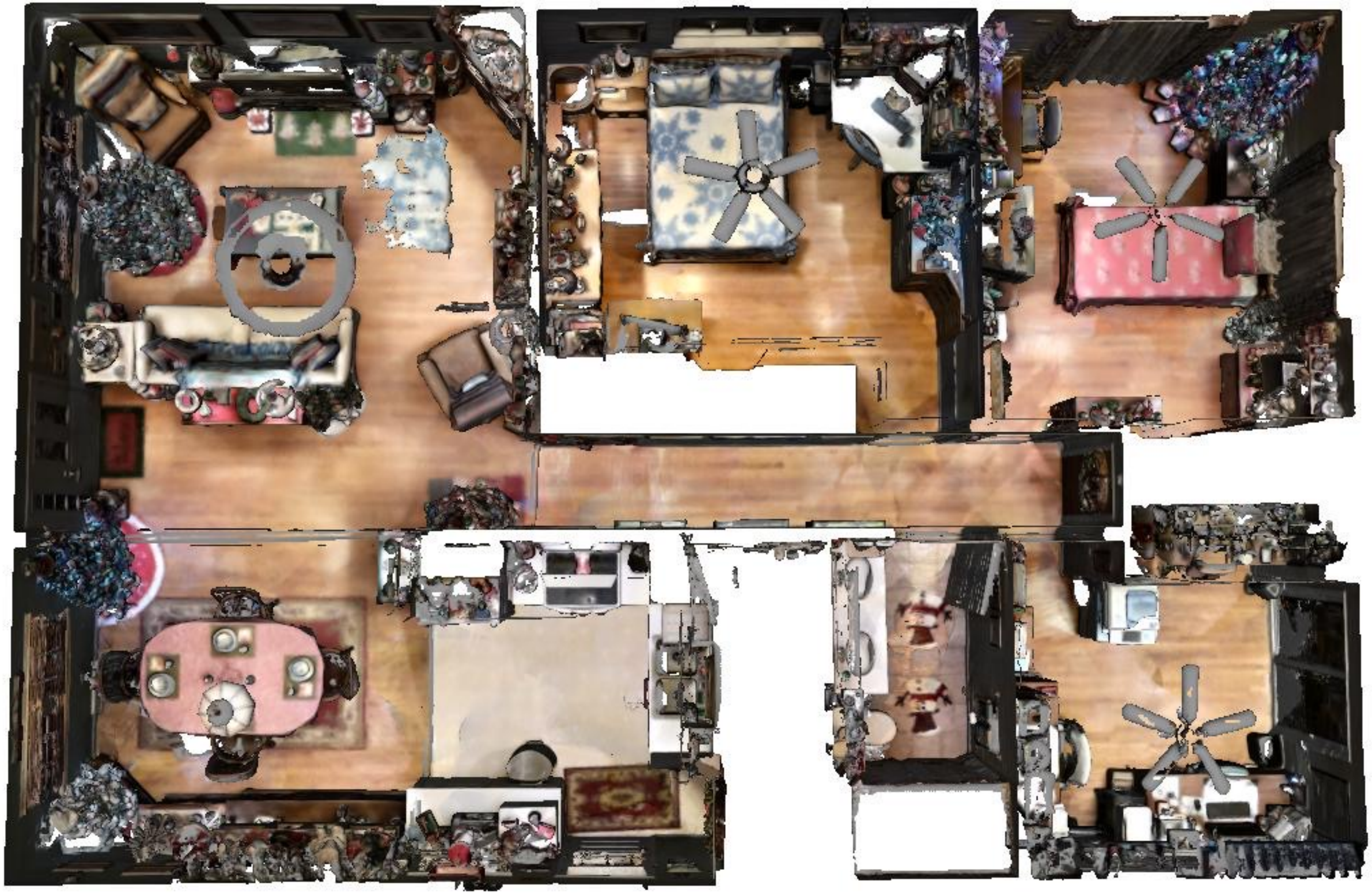
Zihan Zhu[1*] Songyou Peng[1,2*] Viktor Larsson[3] Zhaopeng Cui[4]
Martin R. Oswald[1,5] Andreas Geiger[6] Marc Pollefeys[1,7]

[1]ETH Zürich [2]MPI for Intelligent Systems, Tübingen [3]Lund University
[4]State Key Lab of CAD&CG, Zhejiang University [5]University of Amsterdam
[6]University of Tübingen, Tübingen AI Center [7]Microsoft

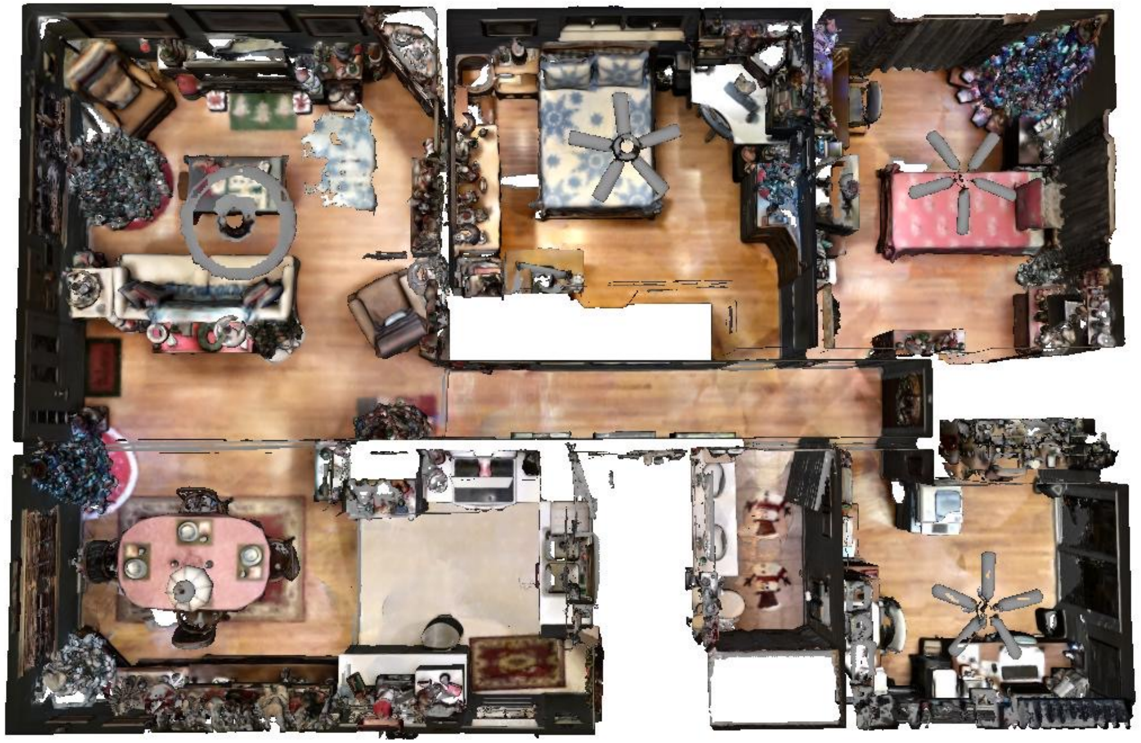


https://arxiv.org/abs/2302.03594

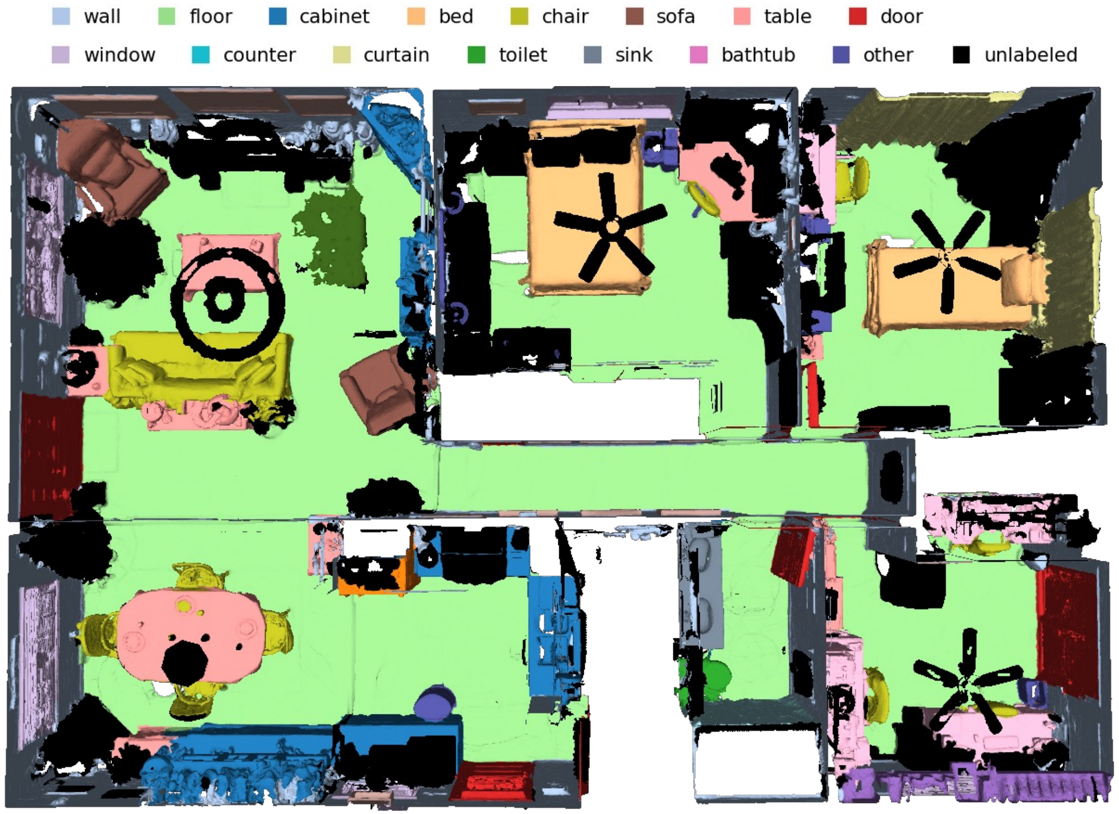




Input 3D Geometry



Input 3D Geometry



Traditional Semantic Segmentation

Only train and test on a few common classes

Input 3D Geometry

- Affordance prediction
- Material identification
- Physical property estimation
- Rare object retrieval
- Activity site prediction
- Fine-grained semantic segmentation
- Many more…

3D Scene Understanding Tasks w/o Labels

OpenScene

3D Scene Understanding with Open Vocabularies

CVPR 2023

Songyou Peng

Kyle Genova

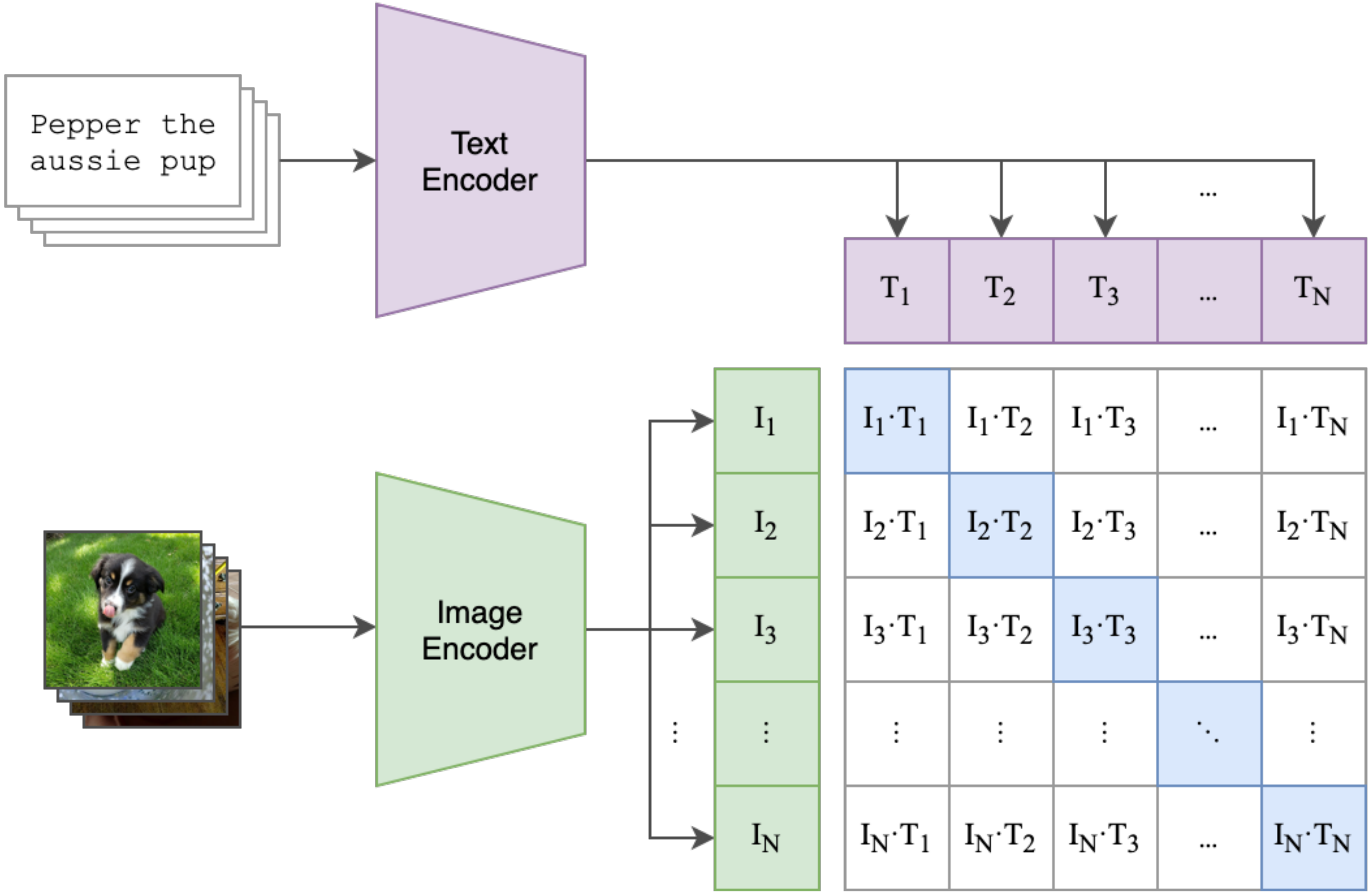
Chiyu "Max" Jiang

Andrea Tagliasacchi

Marc Pollefeys

Tom Funkhouser

Key Idea: Co-embed 3D features with CLIP features

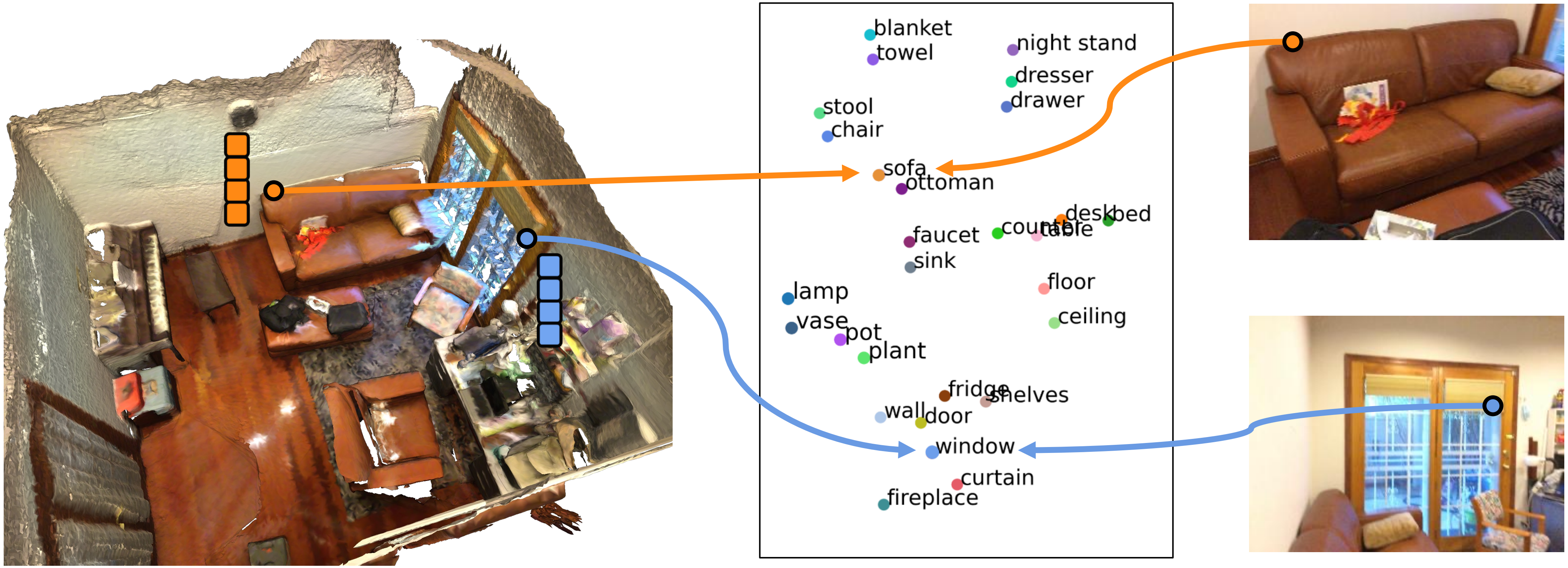


	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

CLIP: Contrastive Language-Image Pre-Training

Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

Key Idea: Co-embed 3D features with CLIP features



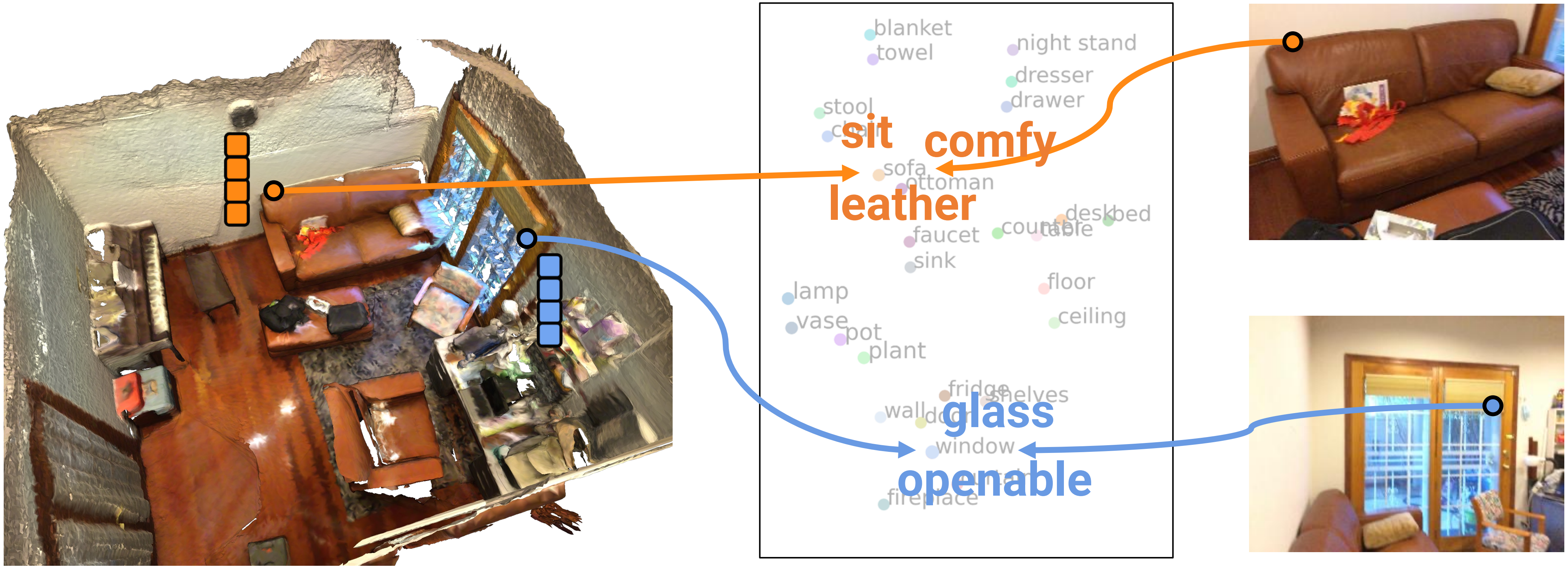
3D Geometry

CLIP Text Features

(visualize with T-SNE)

RGB Images

Key Idea: Co-embed 3D features with CLIP features



3D Geometry

CLIP Text Features

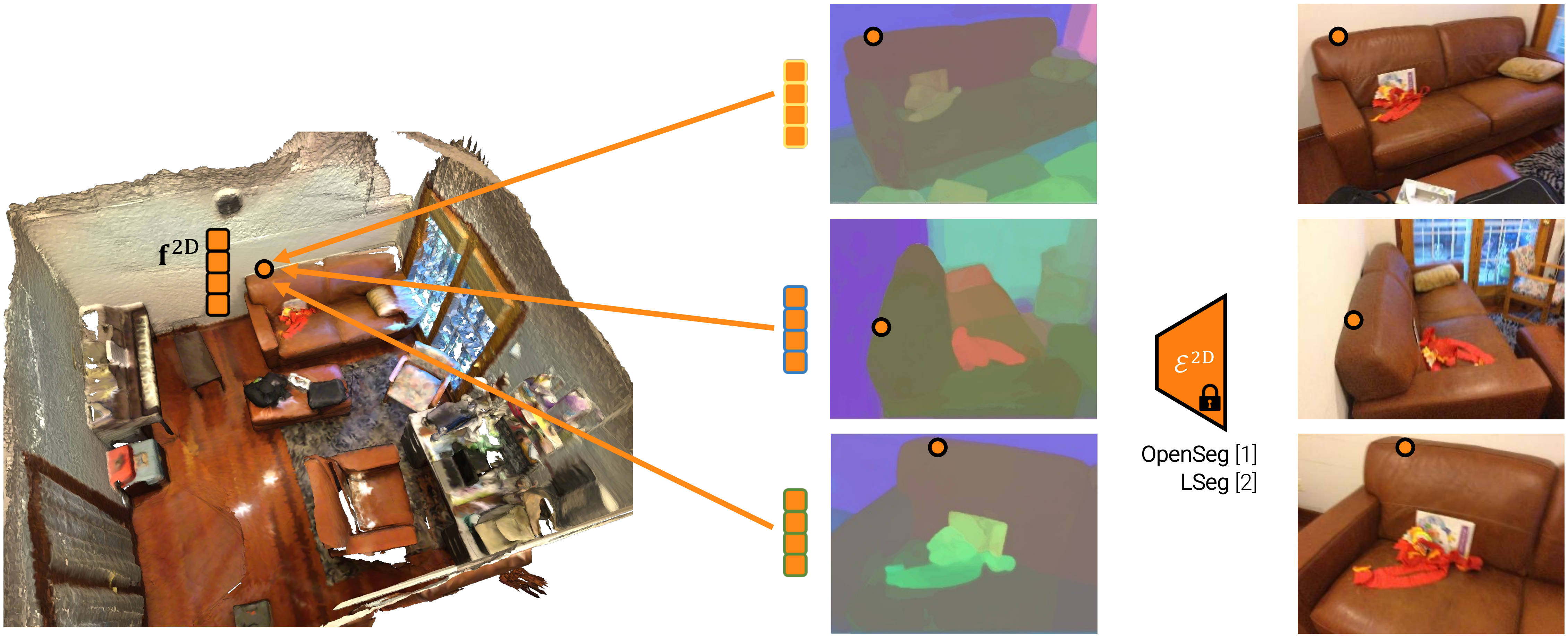
(visualize with T-SNE)

RGB Images

Note: bold word embeddings are approximate

How to Learn Such Text-Image-3D Co-Embeddings?

Step 1: Multi-view Feature Fusion



OpenSeg [1]
LSeg [2]

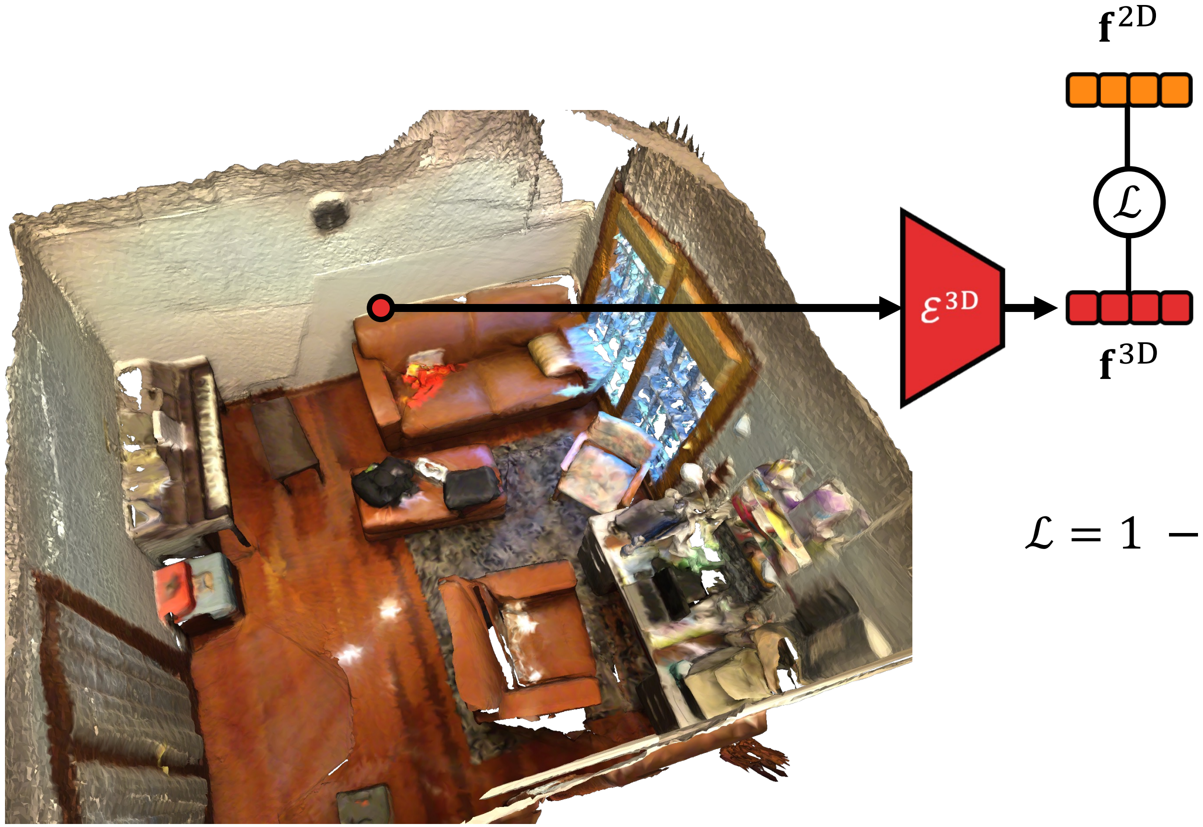
3D Geometry

Per-pixel Features
(visualize with PCA)

RGB Images

[1] Ghiasi, Gu, Cui, Lin: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. ECCV 2022
[2] Li, Weinberger, Belongie, Koltun, Ranftl: Language-driven Semantic Segmentation. ICLR 2022

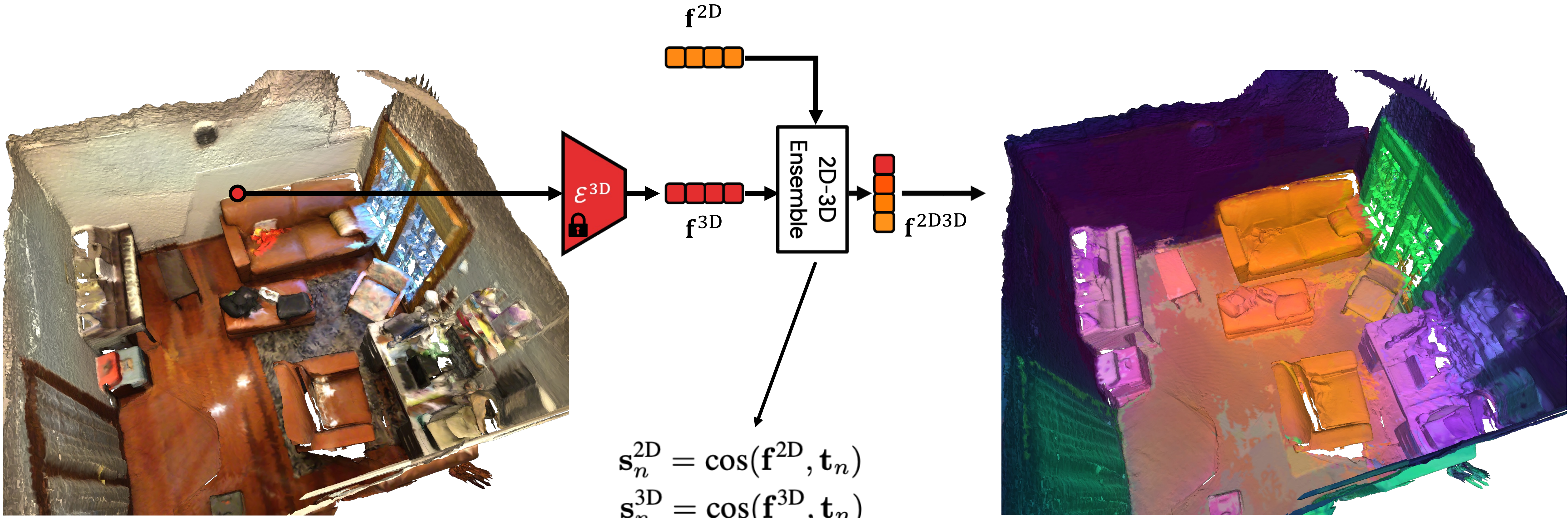
Step 2: 3D Distillation


$\mathbf{f}^{\mathrm{2D}}$
$\mathcal{L}$
$\mathcal{E}^{\mathrm{3D}}$
$\mathbf{f}^{\mathrm{3D}}$

3D Geometry

$$\mathcal{L} = 1 - \cos(\mathbf{f}^{\mathrm{2D}} - \mathbf{f}^{\mathrm{3D}})$$

Step 3: 2D-3D Ensemble


$\mathbf{f}^{2\mathrm{D}}$
$\mathcal{E}^{3\mathrm{D}}$
$\mathbf{f}^{3\mathrm{D}}$
2D-3D Ensemble
$\mathbf{f}^{2\mathrm{D}3\mathrm{D}}$

$$\mathbf{s}_n^{2\mathrm{D}} = \cos(\mathbf{f}^{2\mathrm{D}}, \mathbf{t}_n)$$
$$\mathbf{s}_n^{3\mathrm{D}} = \cos(\mathbf{f}^{3\mathrm{D}}, \mathbf{t}_n)$$

3D Geometry

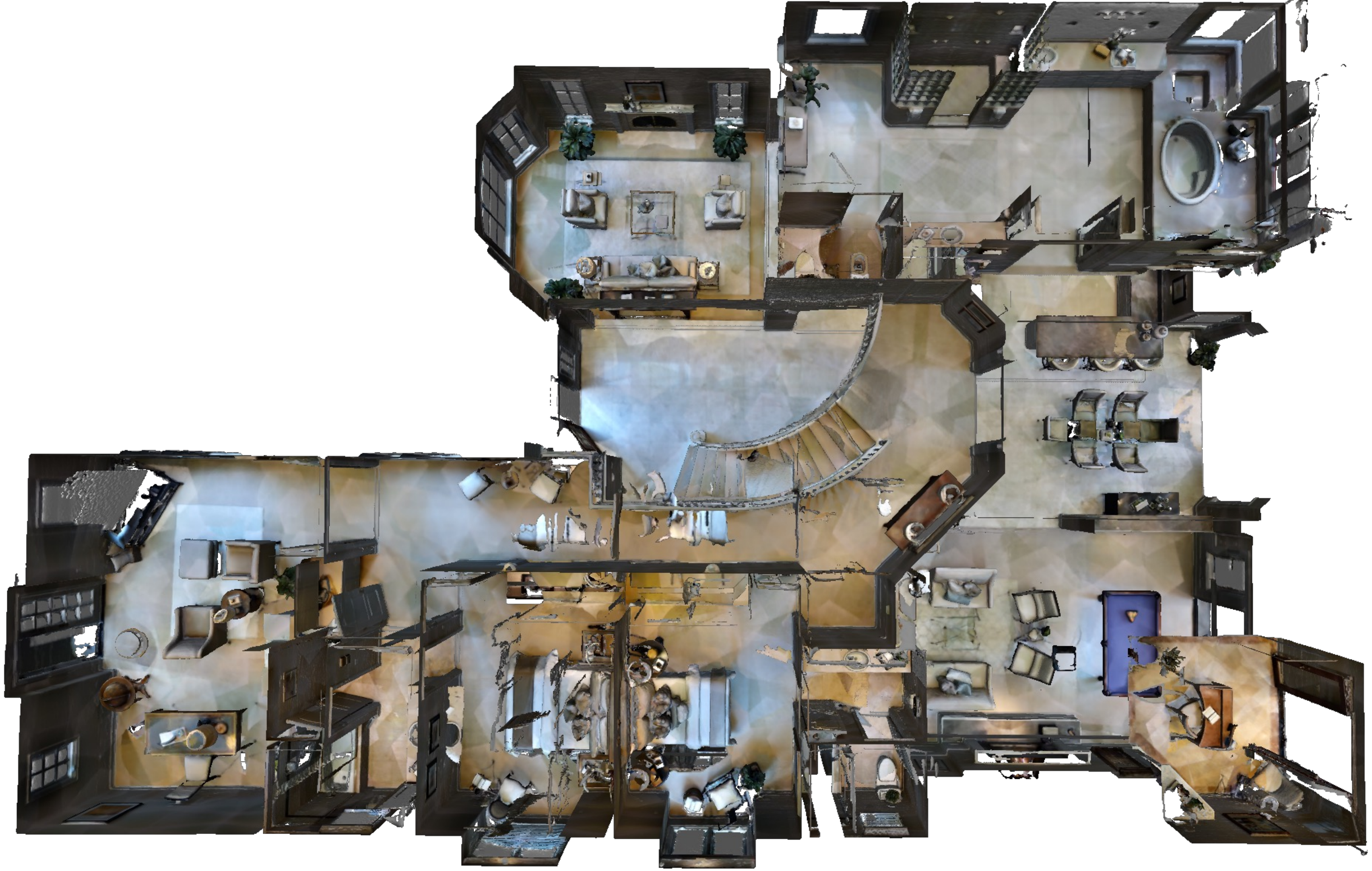
2D-3D Ensemble Features

(visualize with PCA)

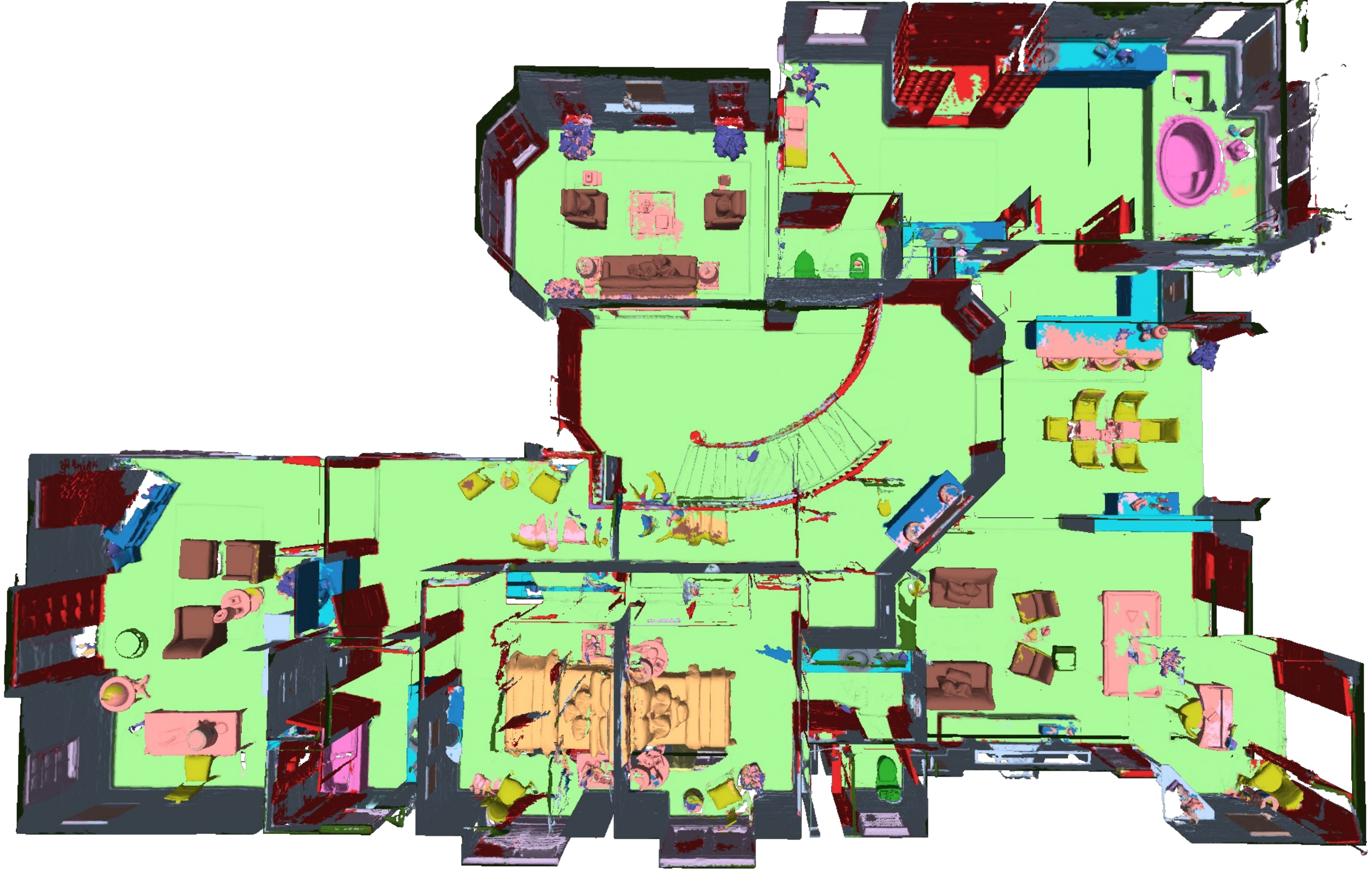
Choose the feature with
the highest max score among all prompts

Open-Vocabulary, Zero-shot

3D Semantic Segmentation



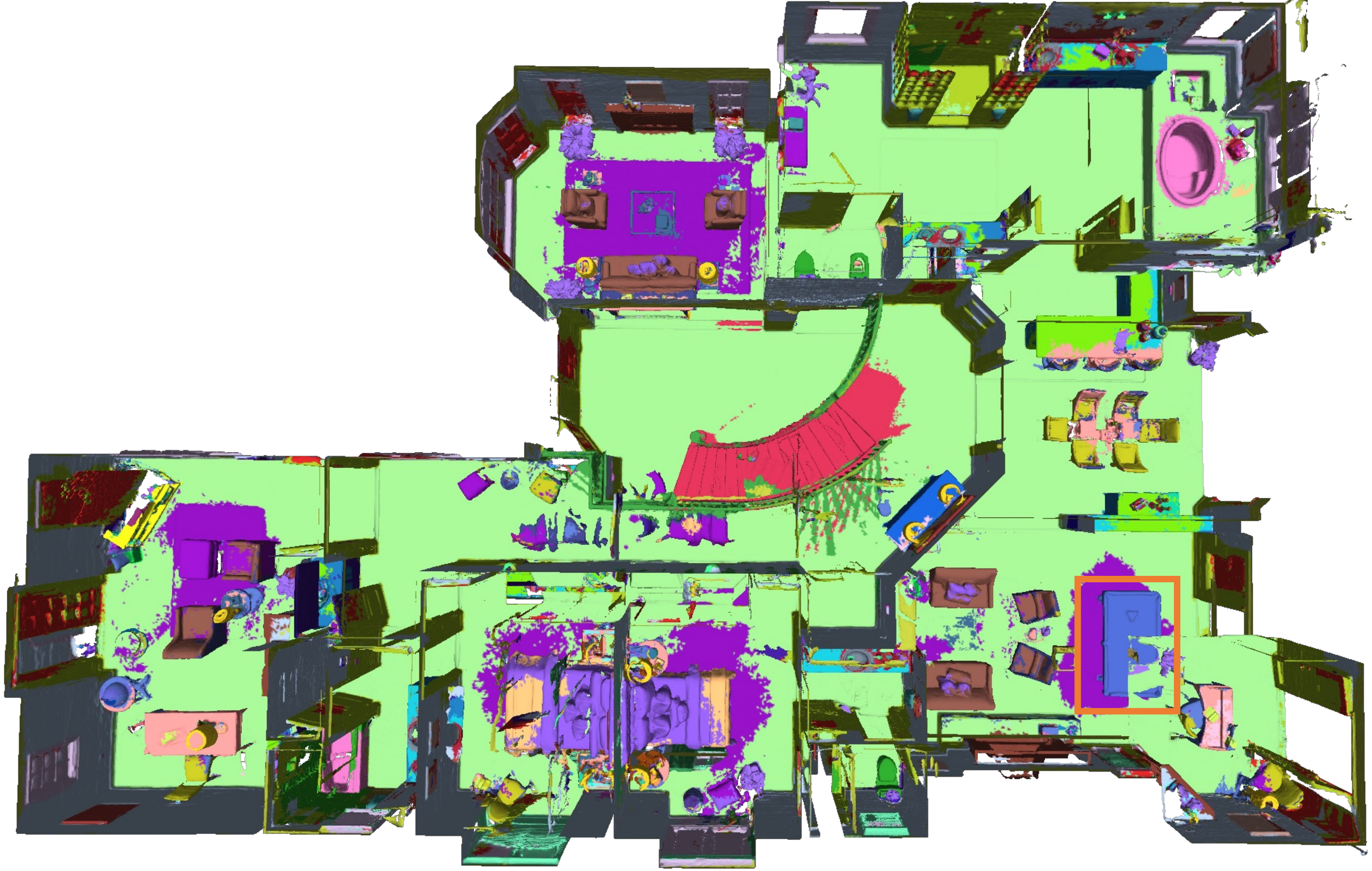
Input 3D Geometry



Our Zero-shot 3D Segmentation

(20 classes)

■ wall ■ floor ■ cabinet ■ bed ■ chair ■ sofa ■ table ■ door ■ window ■ bookshelf ■ picture ■ counter ■ desk ■ curtain ■ refrigerator ■ shower curtain ■ toilet ■ sink ■ bathtub ■ other

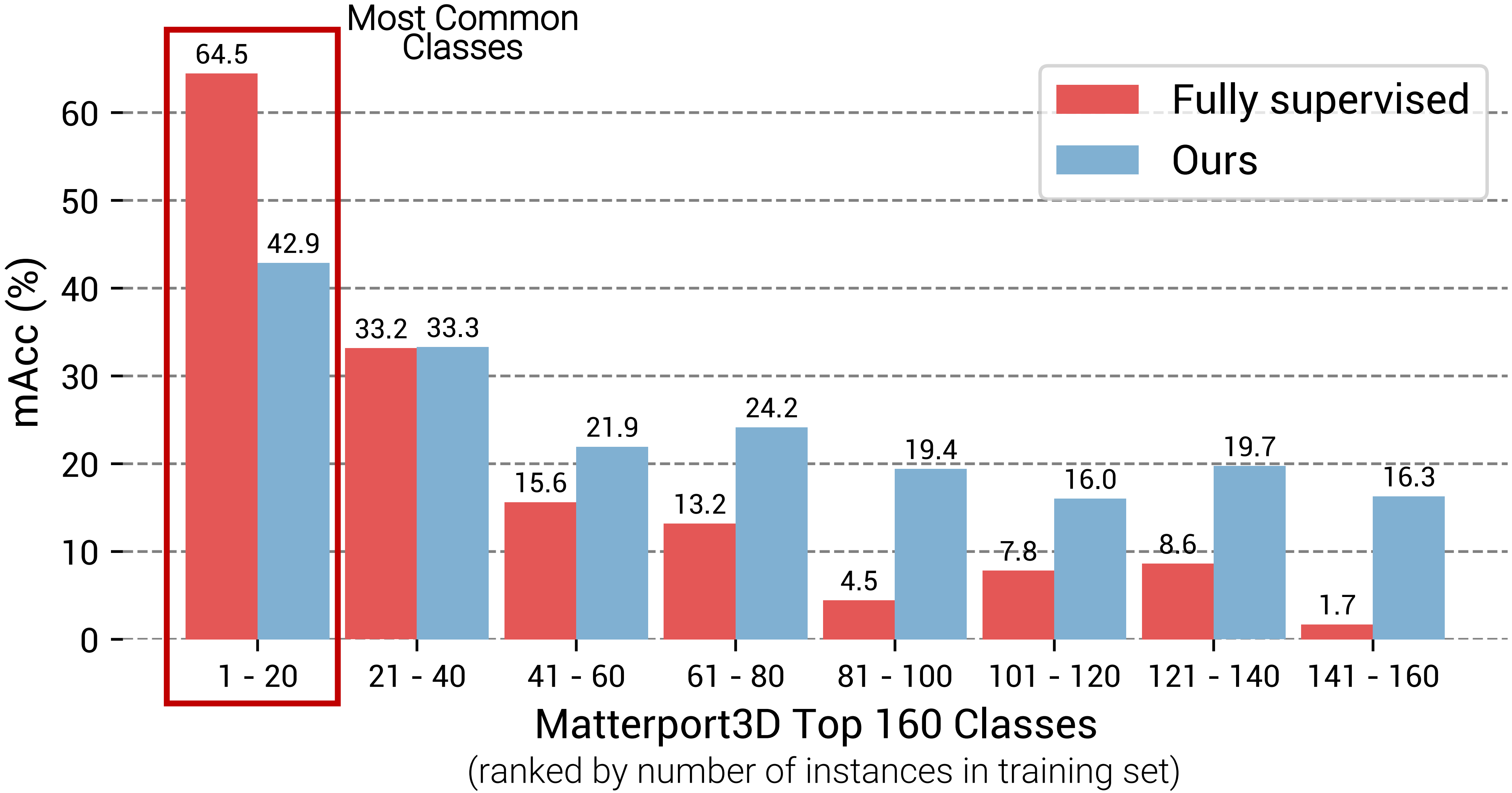


Our Zero-shot 3D Segmentation

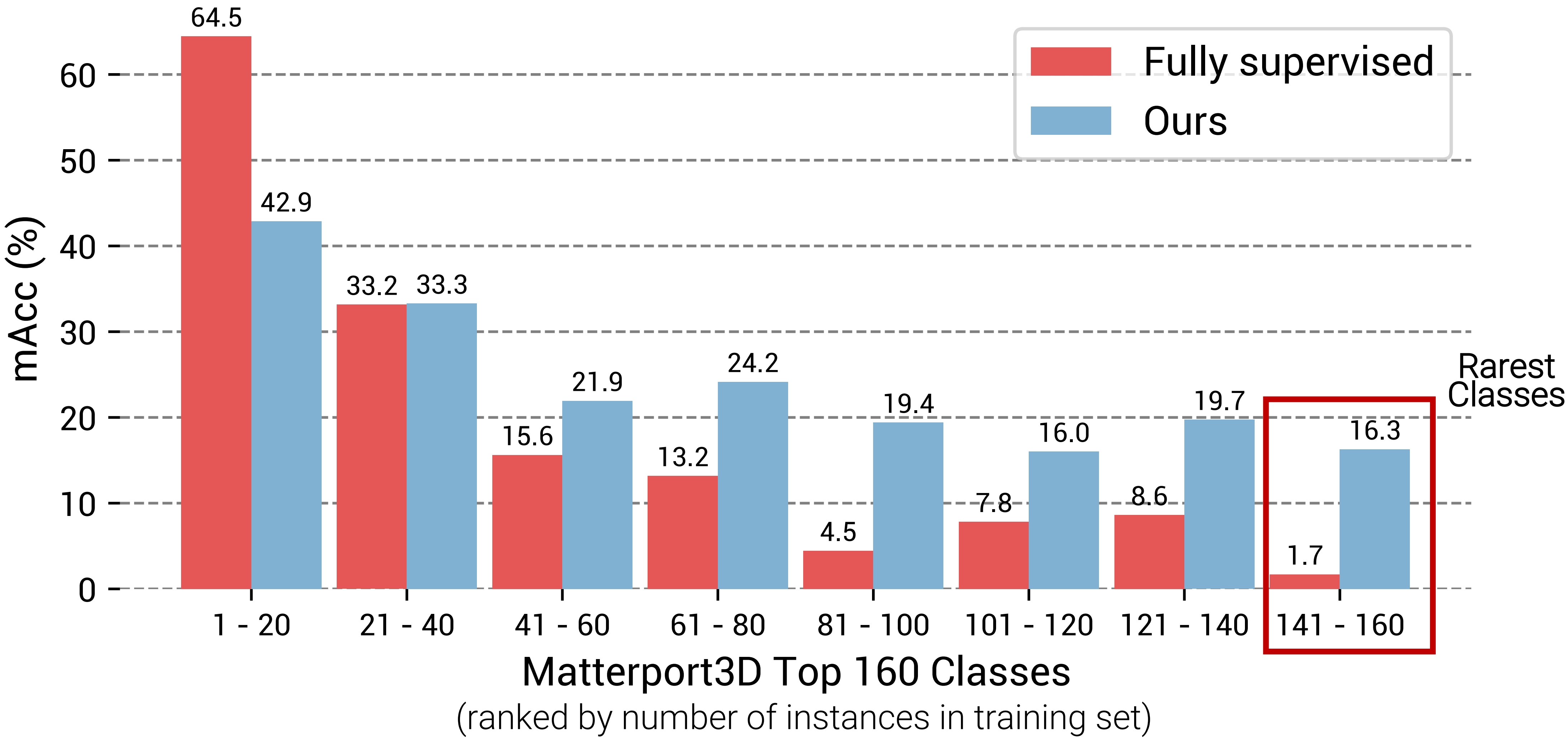
(160 classes)

wall
door
ceiling
floor
picture
window
chair
pillow
cabinet
curtain
table
plant
mirror
towel
sink
shelves
bed
night stand
toilet
column
banister
stairs
stool
vase
pot
desk
box
coffee table
counter
bench
garbage bin
fireplace
bathtub
book
air vent
faucet
photo
toilet paper
fan
railing
dresser
rug
ottoman
bottle
refridgerator
bookshelf
wardrobe
pipe
stand
drawer
container
light switch
purse
door way
basket
chandelier
clock
stove
washing machine
shower curtain
bin
chest
microwave
blinds
tissue box
tv stand
shoe
heater
headboard
bucket
candle
flower pot
furniture
air conditioner
fire extinguisher
curtain rod
printer
telephone
blanket
handle
soap
thermostat
radiator
kitchen island
paper towel
sheet
glass
dishwasher
cup
ladder
garage door
piano
board
rope
ball
excercise equipment
hanger
candlestick
light
scale
bag
display case
toilet paper holder
tray
urn
decorative plate
pool table
jacket
bottle of soap
water cooler
tea pot
stuffed animal
paper towel dispenser
lamp shade
car
toilet brush
drum
whiteboard
range hood
candelabra
toy
foot rest
soap dish
cleaner
computer
knob
paper
projector

Comparison


Most Common Classes
Fully supervised
Ours
mAcc (%)
0
10
20
30
40
50
60
64.5
42.9
33.2
33.3
15.6
21.9
13.2
24.2
4.5
19.4
7.8
16.0
8.6
19.7
1.7
16.3
1 - 20
21 - 40
41 - 60
61 - 80
81 - 100
101 - 120
121 - 140
141 - 160
Matterport3D Top 160 Classes
(ranked by number of instances in training set)

Comparison



🟥 Fully supervised
🟦 Ours
mAcc (%)
0
10
20
30
40
50
60
64.5
42.9
33.2
33.3
15.6
21.9
13.2
24.2
4.5
19.4
7.8
16.0
8.6
19.7
1.7
16.3
1 - 20
21 - 40
41 - 60
61 - 80
81 - 100
101 - 120
121 - 140
141 - 160
Rarest Classes
Matterport3D Top 160 Classes
(ranked by number of instances in training set)

Ablation

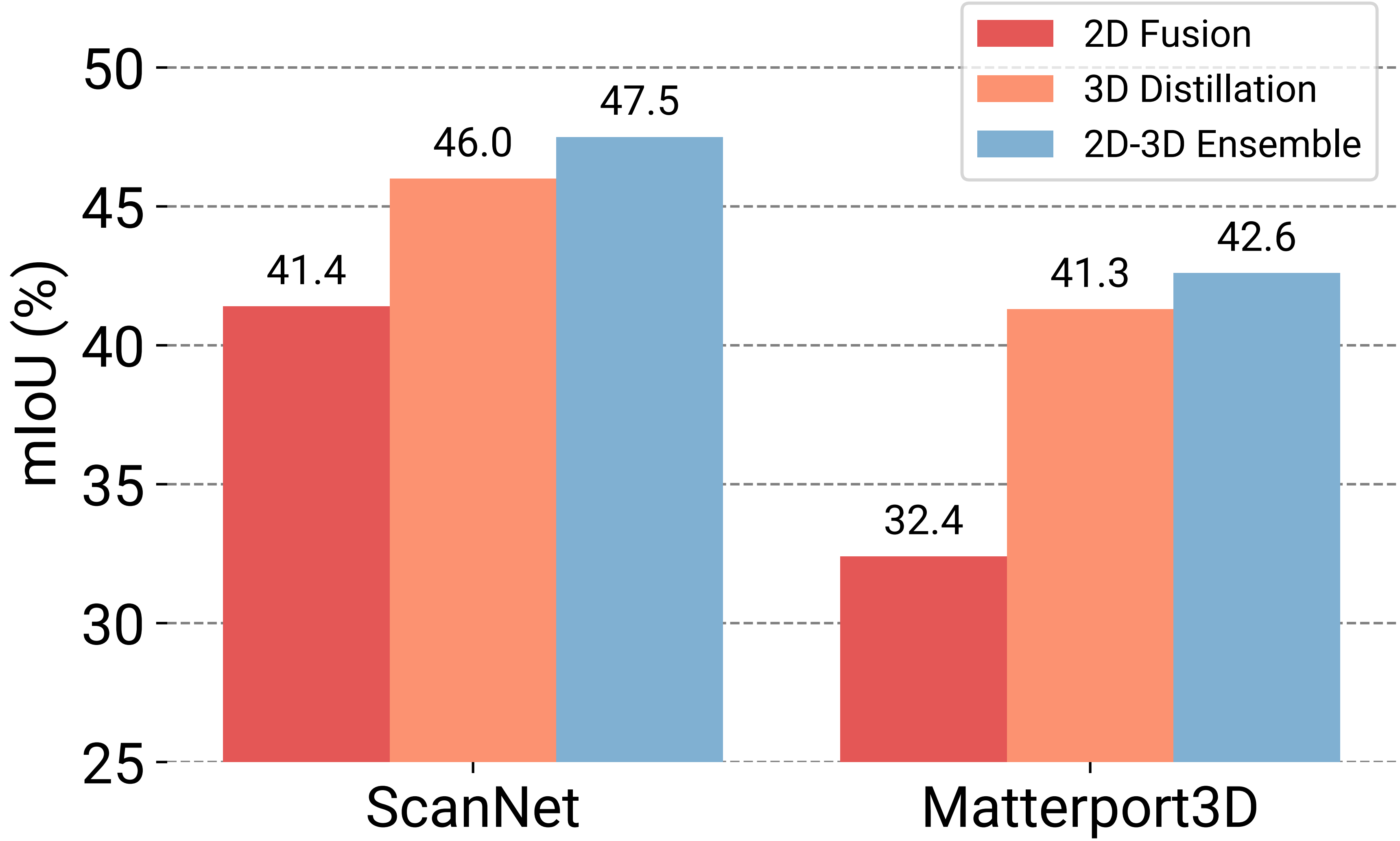

2D Fusion
3D Distillation
2D-3D Ensemble
mIoU (%)
50
45
40
35
30
25
41.4
46.0
47.5
32.4
41.3
42.6
ScanNet
Matterport3D

Image-based 3D Scene Query

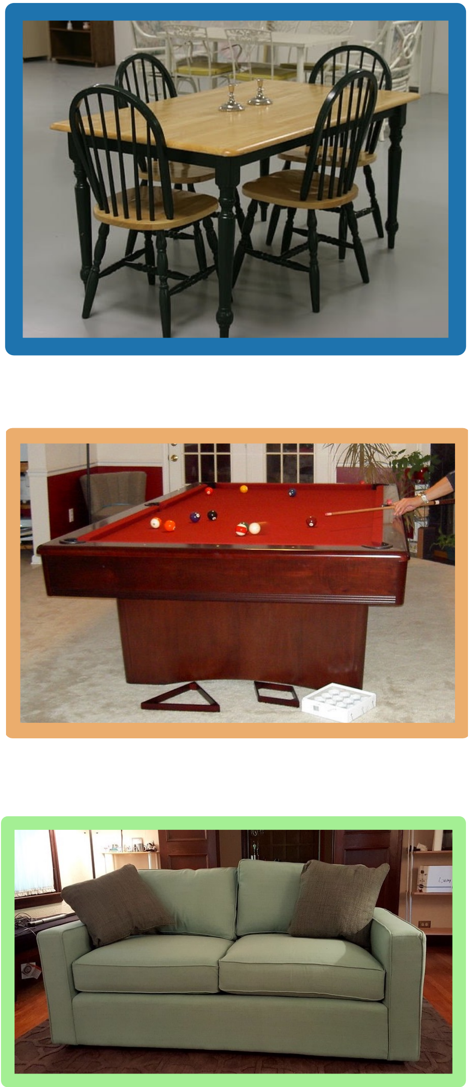
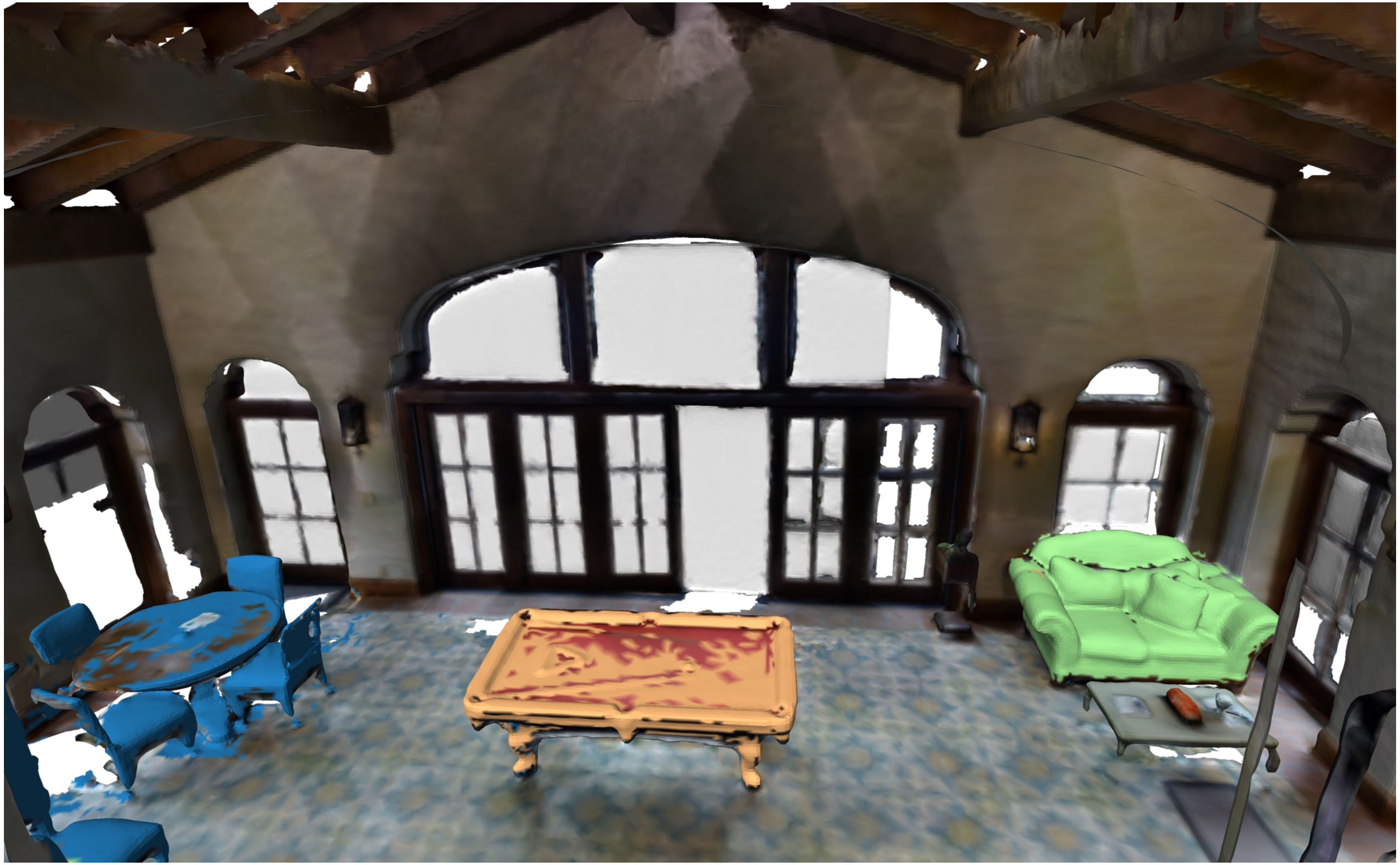


Image Queries

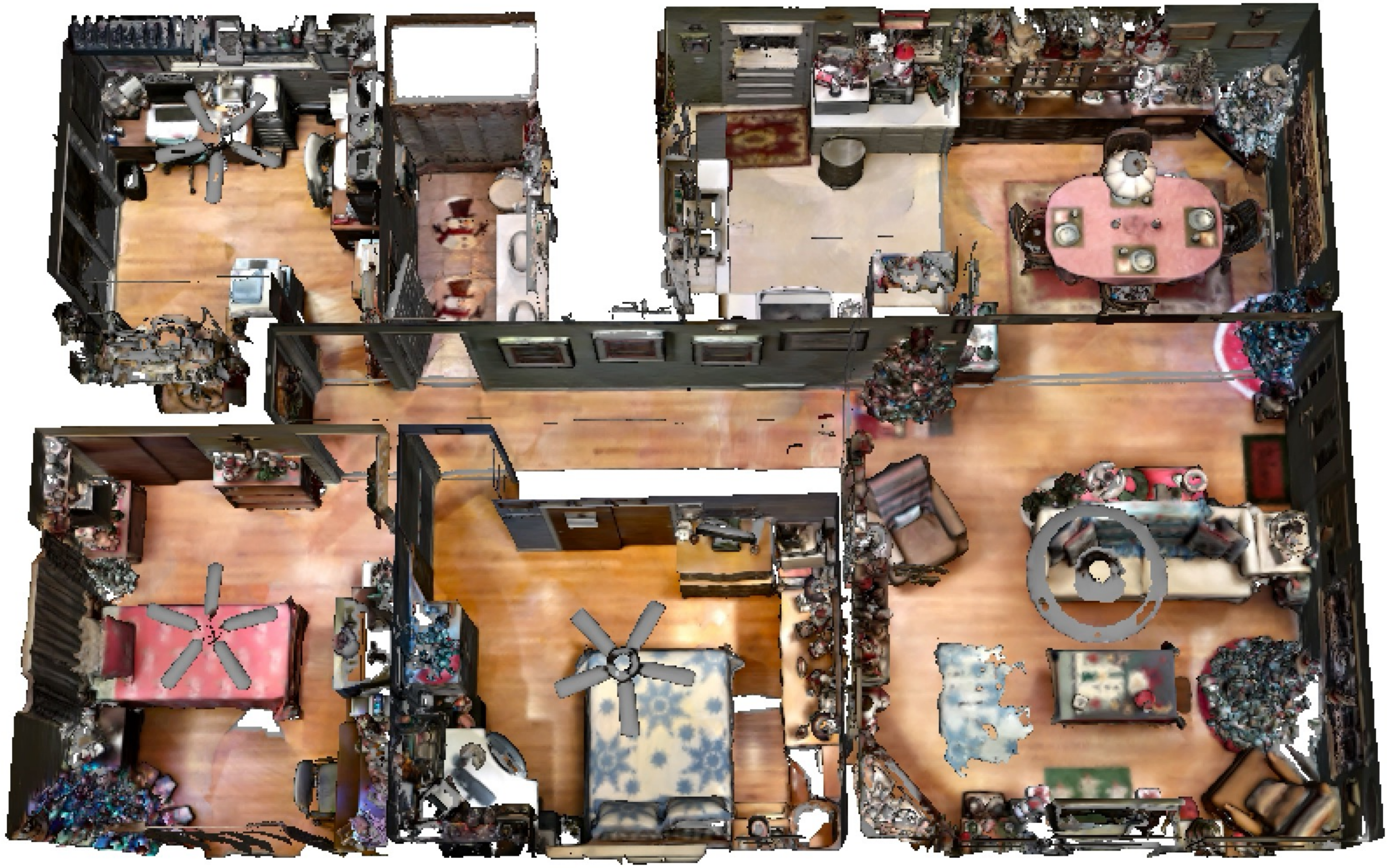


Given 3D Geometry

Interactive Demo

Open-vocabulary 3D Scene Exploration

Text queries:

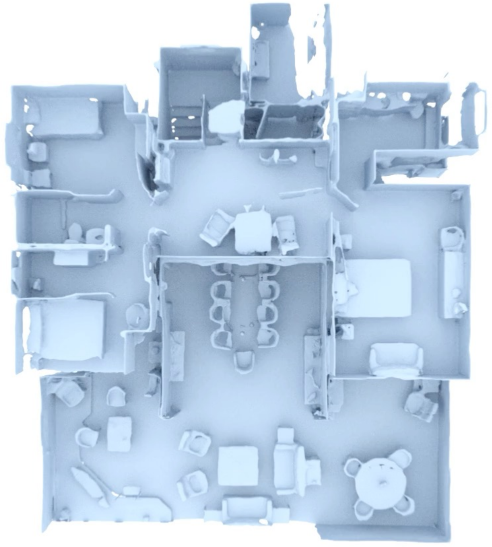


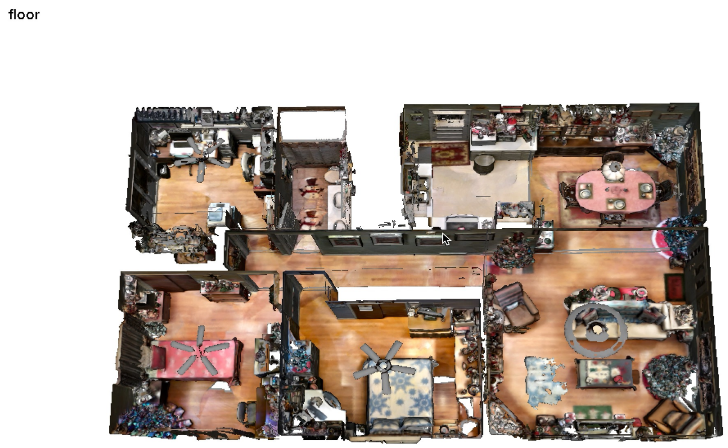
Take-home Message

- We enable a **wide range of applications** by open-vocabulary queries
- This can hopefully influence how people train 3D scene understanding systems in the future
- Our real-time demo already shows the **possibility to directly apply to AR/VR**

Learn to Reconstruct and Understand the 3D World

Songyou Peng





Convolutional Occupancy Networks
ECCV 2020 (Spotlight)
pengsongyou.github.io/conv_onet

NICE-SLAM
CVPR 2022
pengsongyou.github.io/nice-slam

OpenScene
CVPR 2023
pengsongyou.github.io/openscene

Thank you!