Learning Neural Scene Representations for 3D Reconstruction and Understanding

Songyou Peng

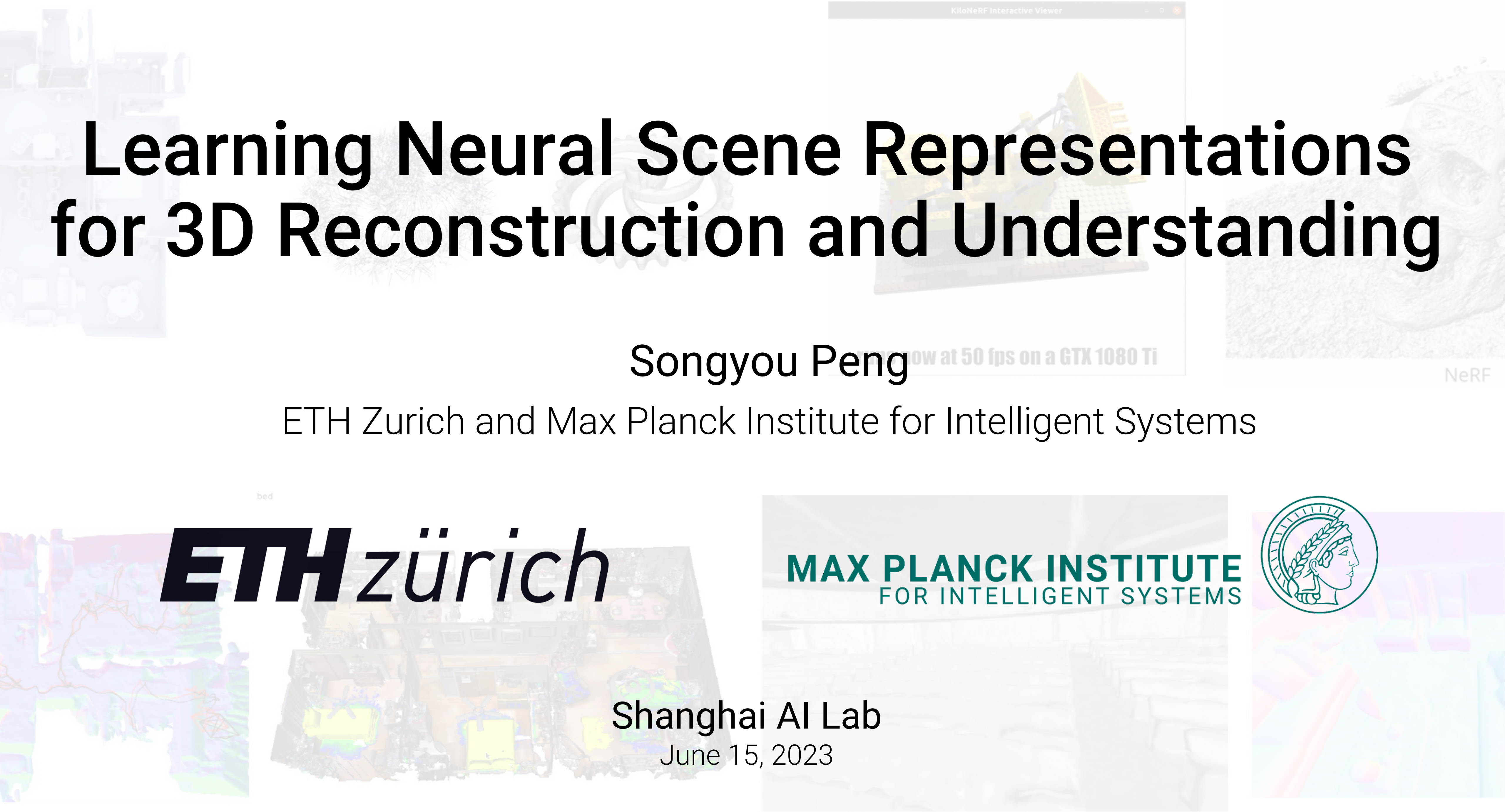
ETH Zurich and Max Planck Institute for Intelligent Systems

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Shanghai AI Lab

June 15, 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

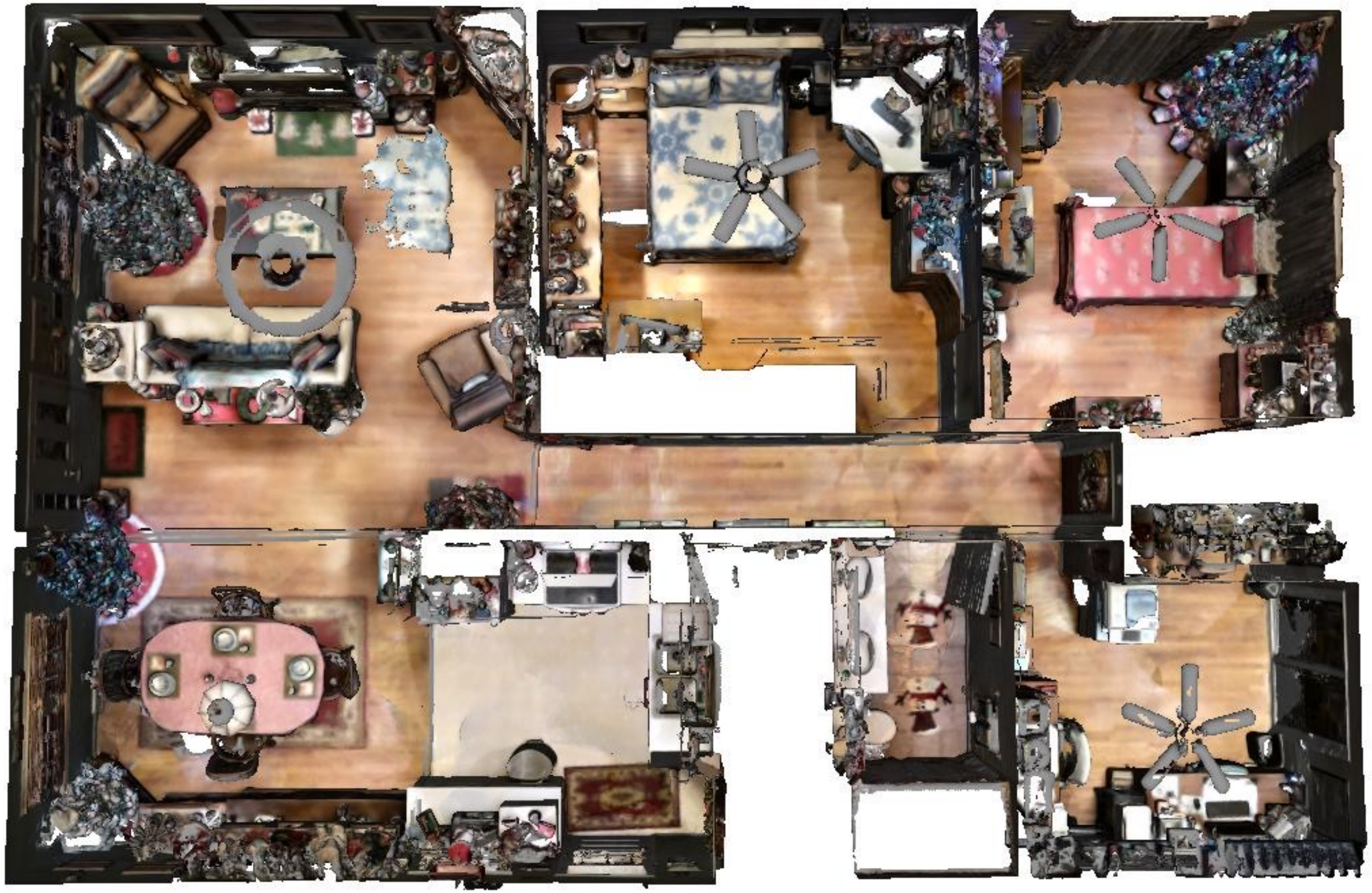
Meta

Google Research

pengsongyou.github.io

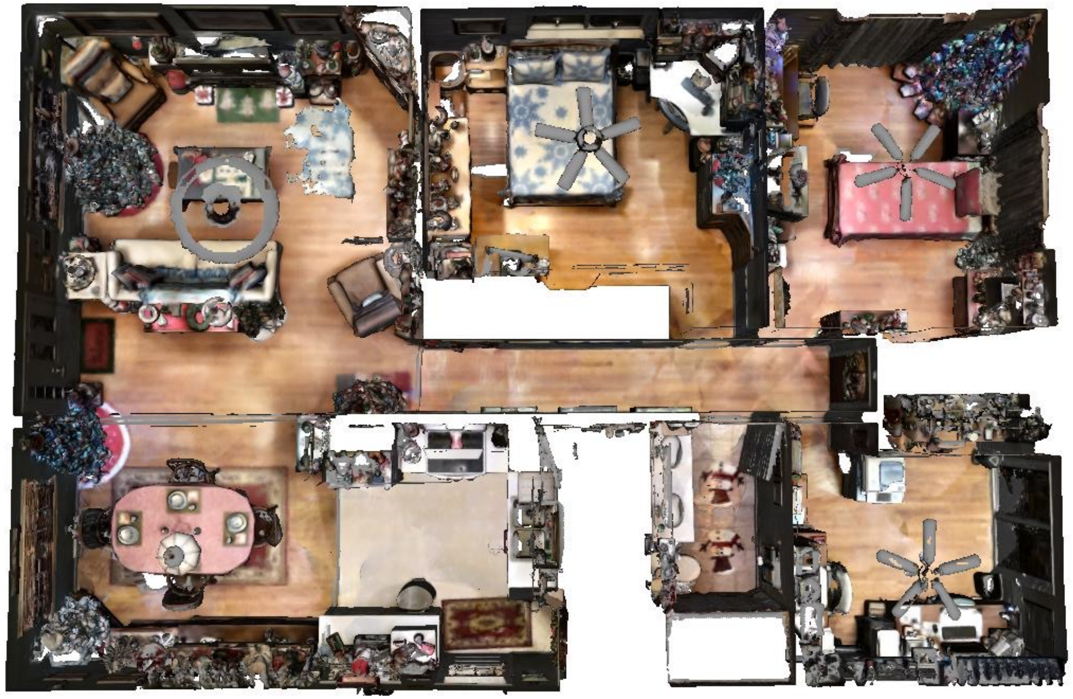
Motivation

Input Images

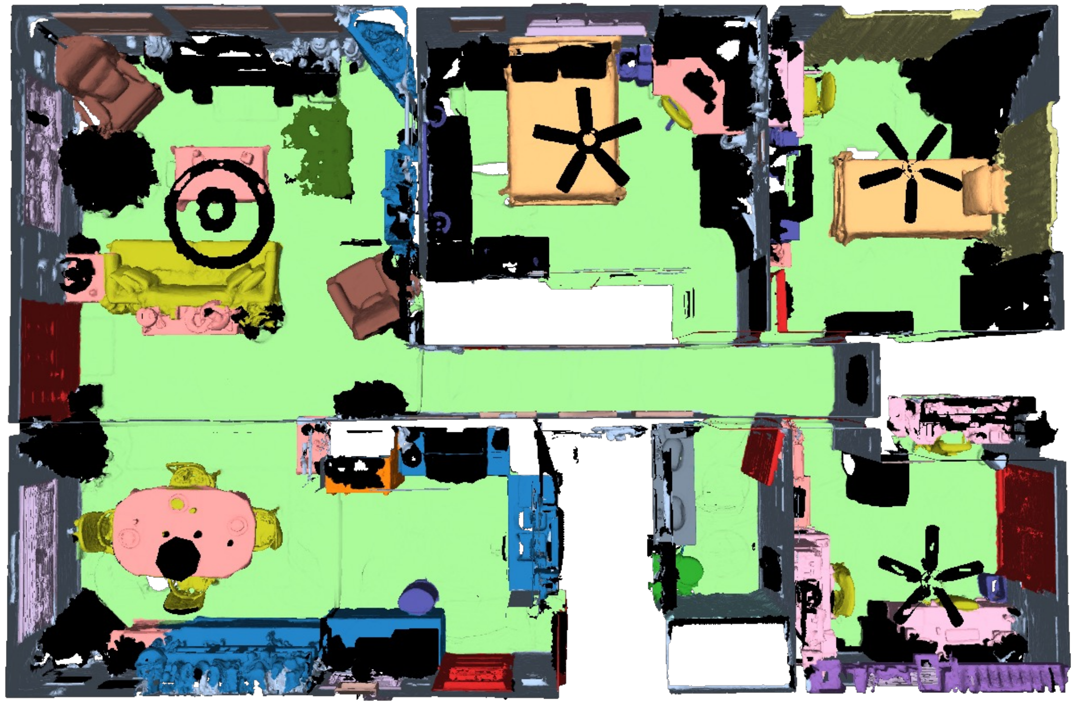


3D Reconstruction

Motivation

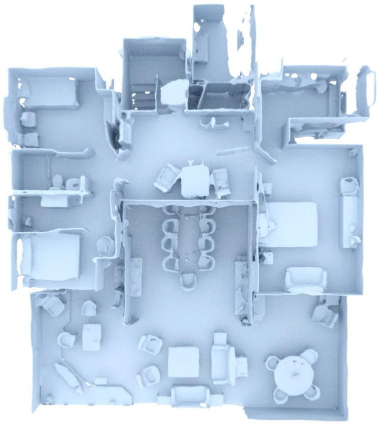


3D Reconstruction

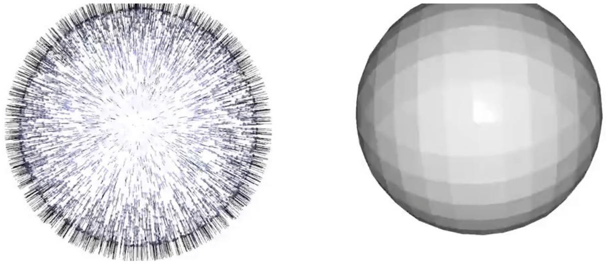


3D Scene Understanding

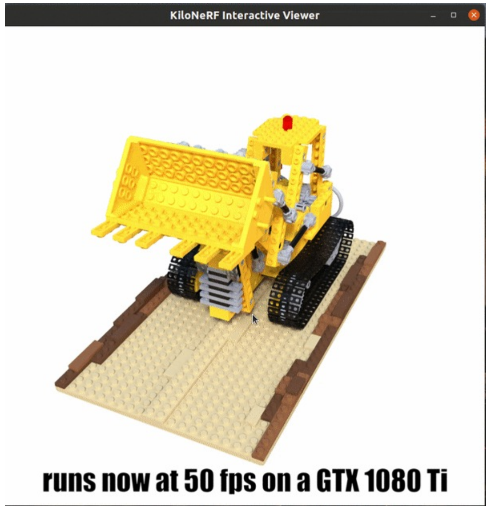
My PhD Topics: Neural Scene Representations
for 3D reconstruction and 3D scene understanding



Convolutional Occupancy Nets
ECCV 2020 (Spotlight)

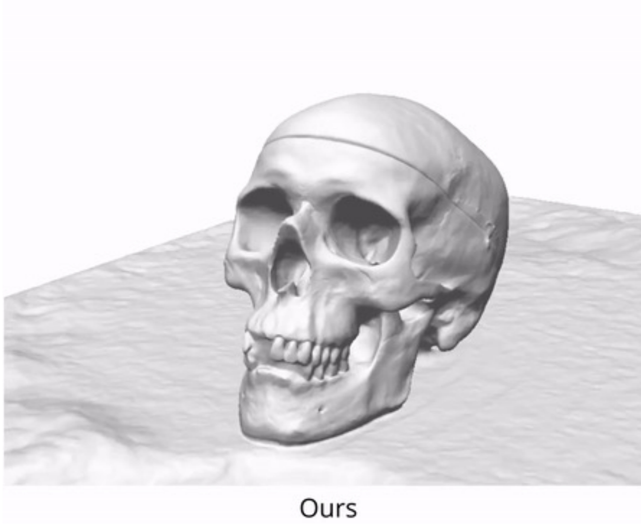


Shape As Points
NeurIPS 2021 (Oral)

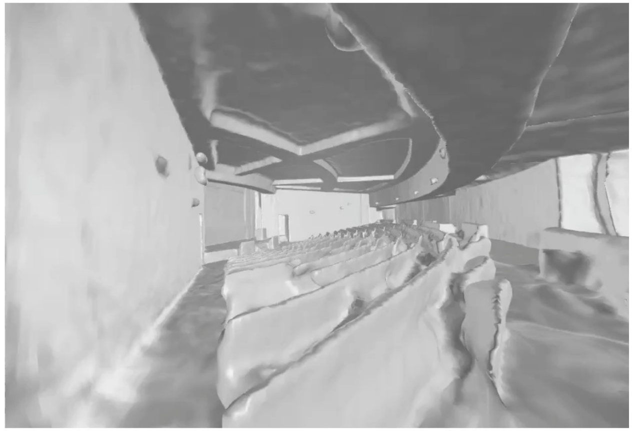


runs now at 50 fps on a GTX 1080 Ti

KiloNeRF
ICCV 2021



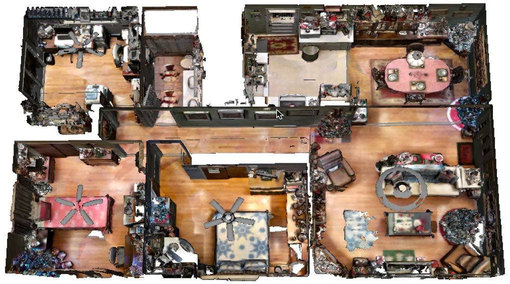
UNISURF
ICCV 2021 (Oral)



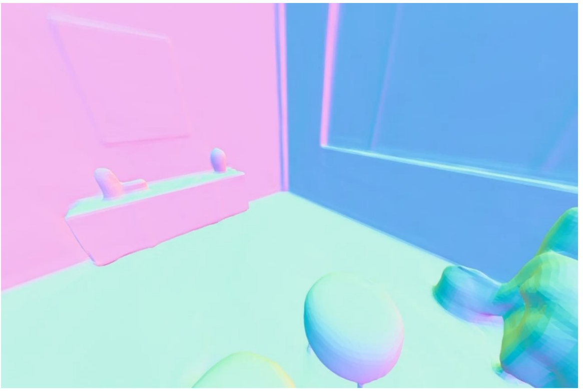
Ours
MonoSDF
NeurIPS 2022

floor

NICE-SLAM
CVPR 2022

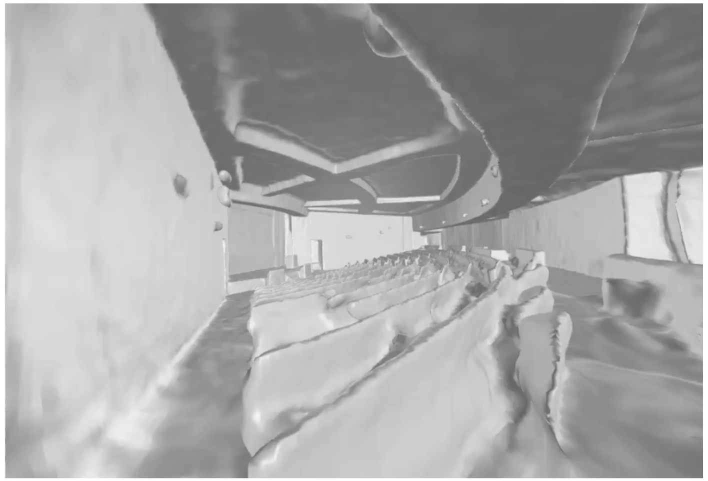


OpenScene
CVPR 2023



NICER-SLAM
arXiv 2023

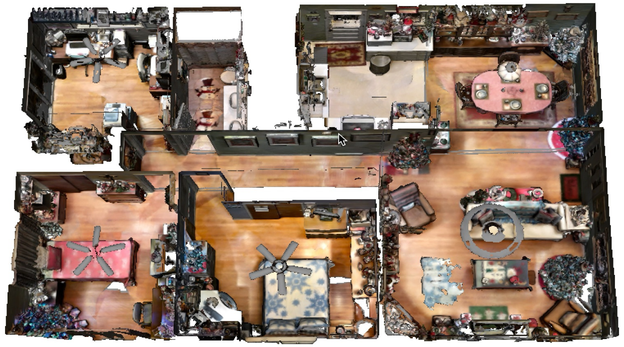
My PhD Topics: Neural Scene Representations
for 3D reconstruction and 3D scene understanding



Ours

MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022



OpenScene
CVPR 2023

NeRF is awesome!

Some existing problems…

😢 Poor underlying geometry

😢 Camera poses needed

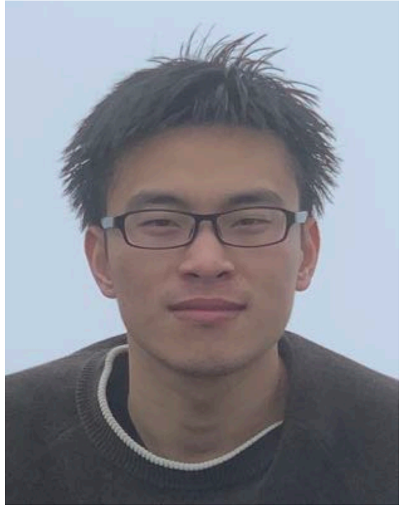
😊 MonoSDF

😊 NICE-SLAM

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

NEURAL INFORMATION PROCESSING SYSTEMS

MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction



Zehao Yu

Songyou Peng

Michael Niemeyer

Torsten Sattler

Andreas Geiger

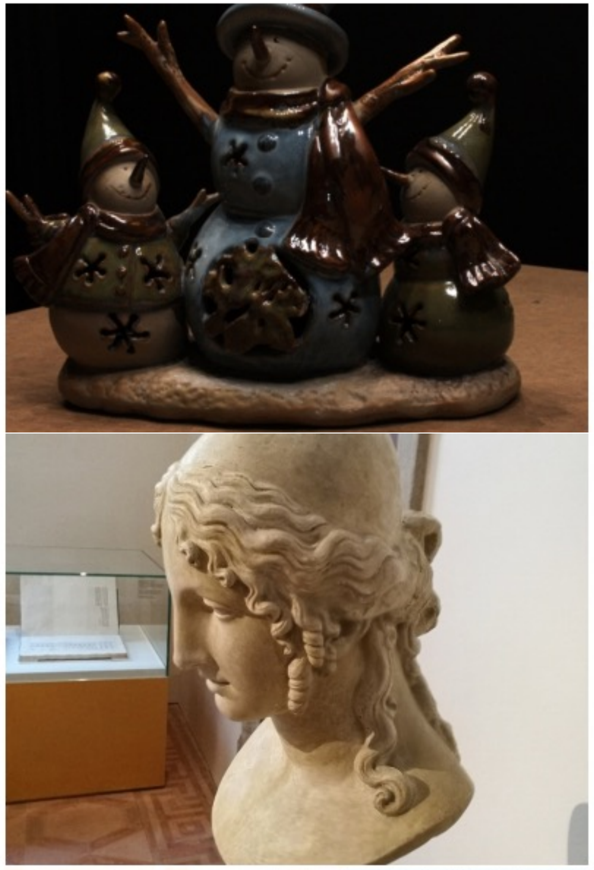
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ETH zürich

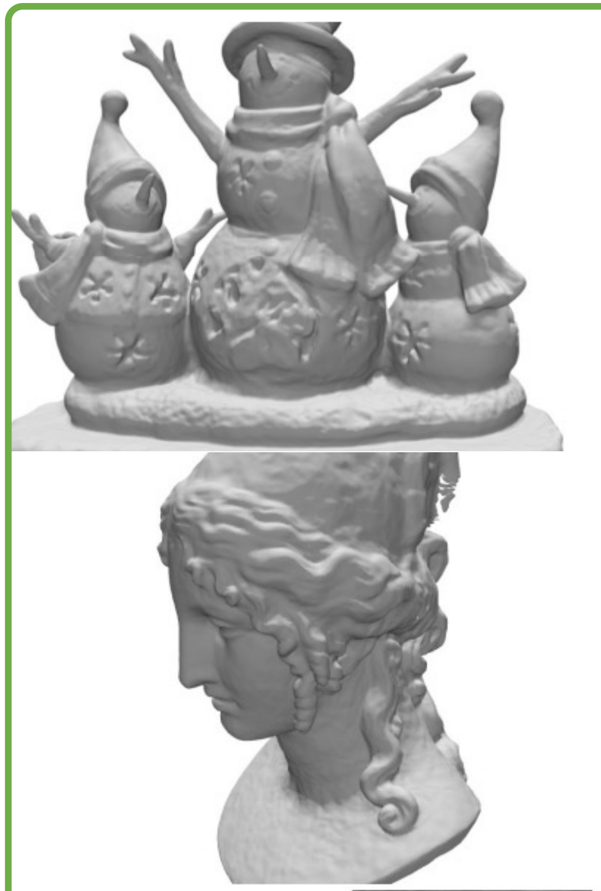
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

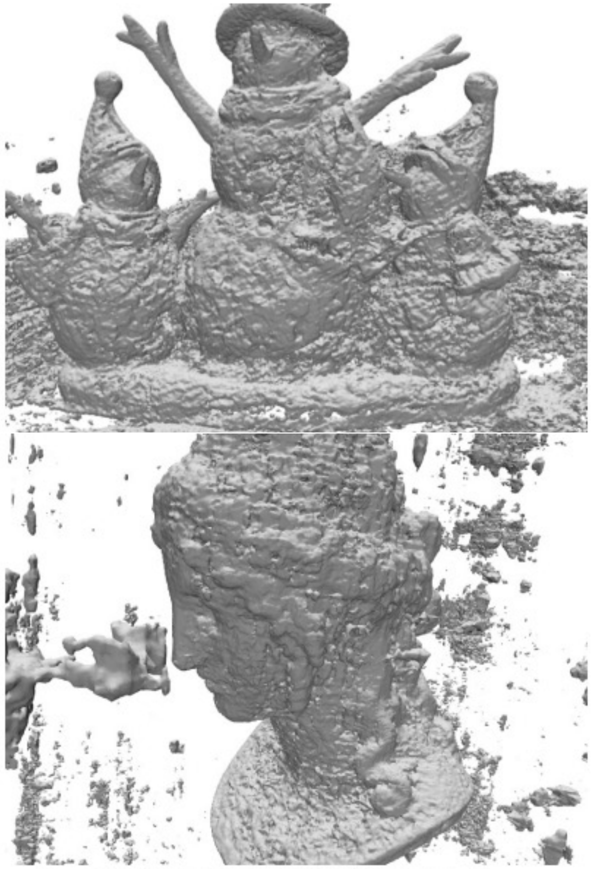
Neural Implicit Surfaces with Volume Rendering



RGB Images



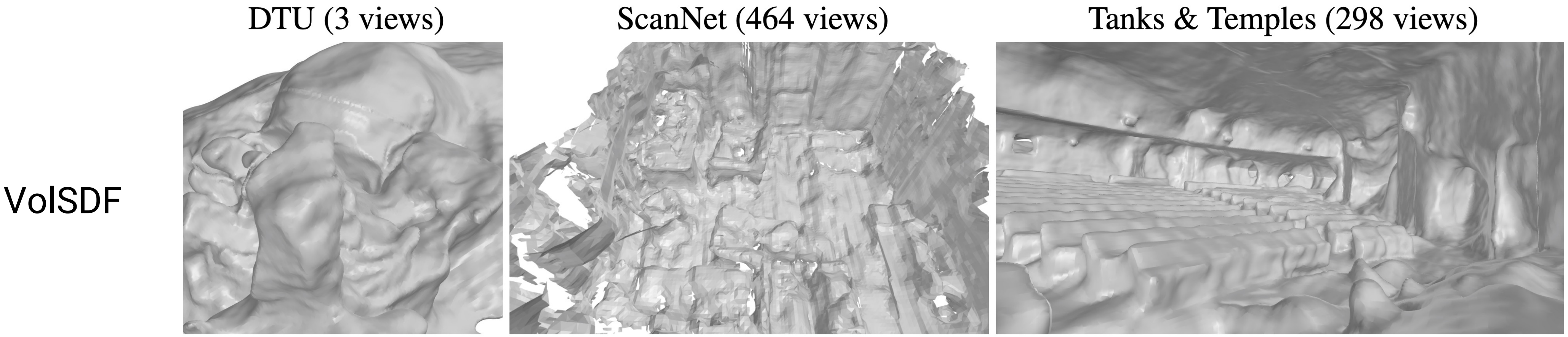
VolSDF/NeuS/UNISURF



NeRF

[1] Oechsle, Peng, Geiger: UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. ICCV, 2021
[2] Wang, Liu, Liu, Theobalt, Komura, Wang: NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. NeurIPS, 2021
[3] Yariv, Gu, Kasten, Lipman: Volume rendering of neural implicit surfaces. NeurIPS, 2021

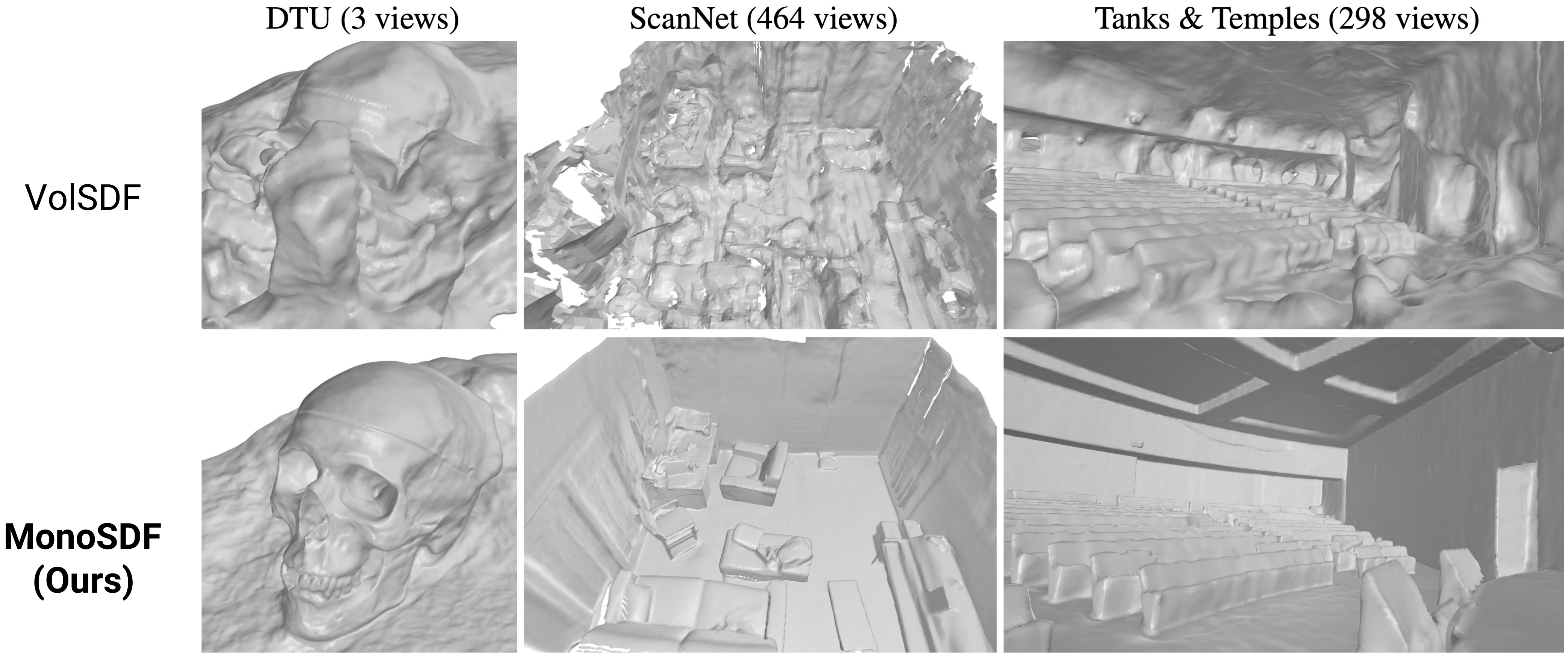
Neural Implicit Surfaces with Volume Rendering



- Fails with sparse input views
- Poor results in large-scale indoor scenes

Yariv, Gu, Kasten, Lipman: Volume rendering of neural implicit surfaces. NeurIPS, 2021

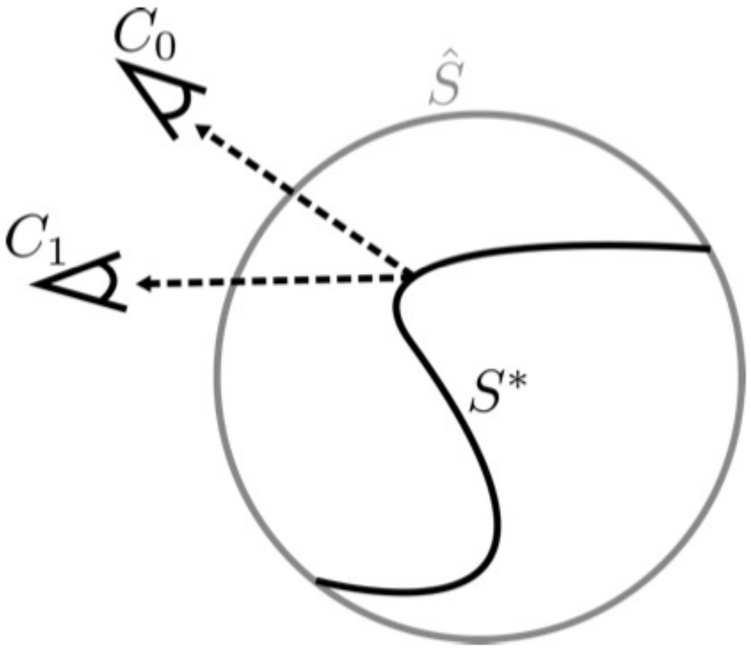
Neural Implicit Surfaces with Volume Rendering



- ➕ Manage to reconstruct with sparse views
- ➕ Nice 3D reconstruction in large-scale indoor scenes

Yu, Peng, Niemeyer, Sattler, Geiger: MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction. NeurIPS, 2022

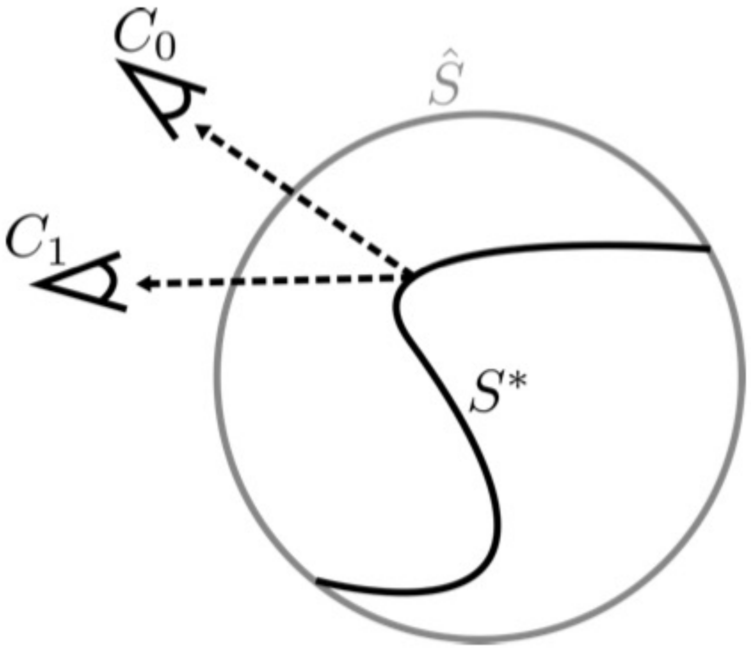
Shape-Appearance Ambiguity



There exists an infinite number of photo-consistent explanations for input images!

Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020

Shape-Appearance Ambiguity

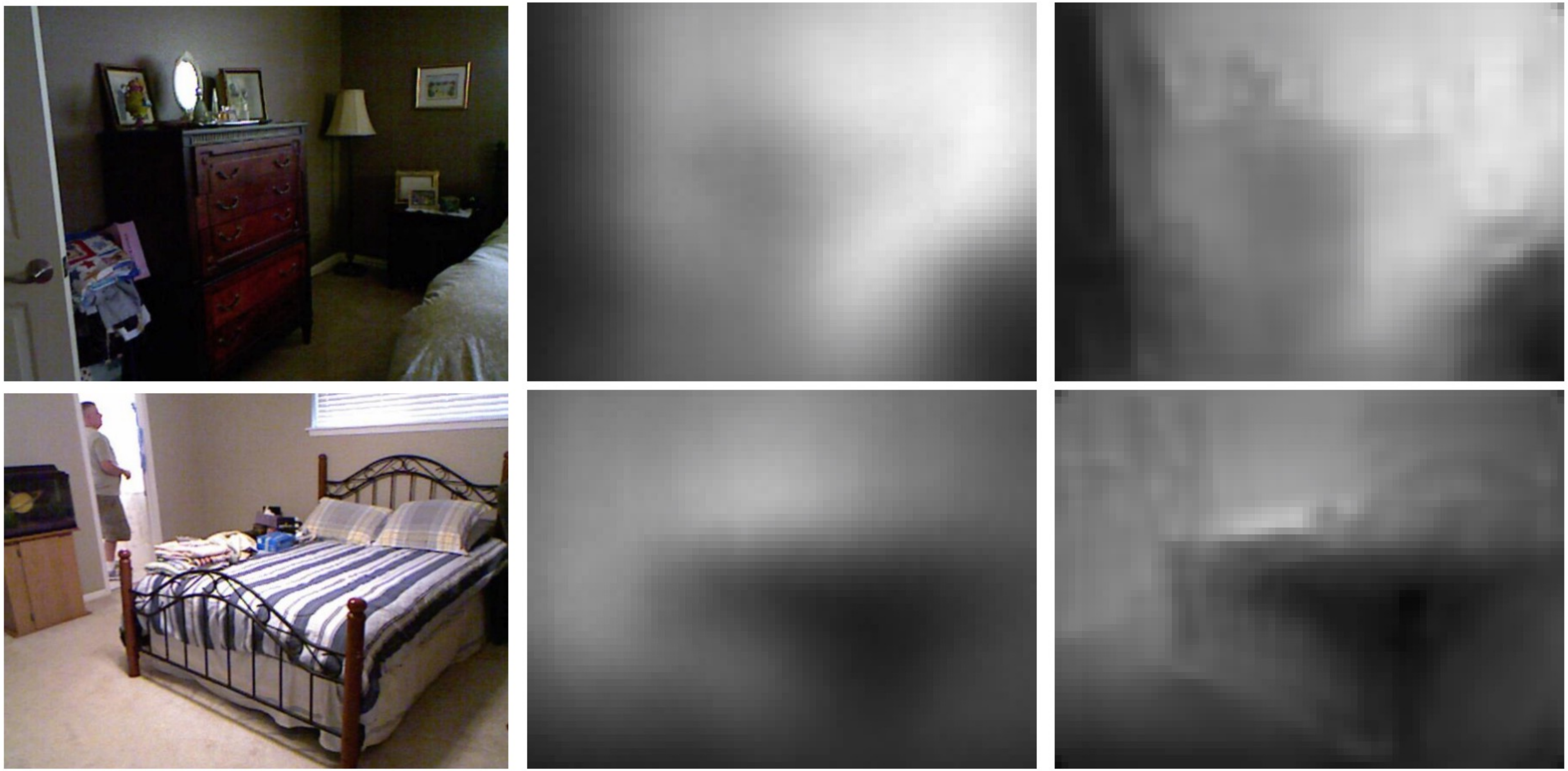


There exists an infinite number of photo-consistent explanations for input images!

⇒ Exploit monocular geometric priors

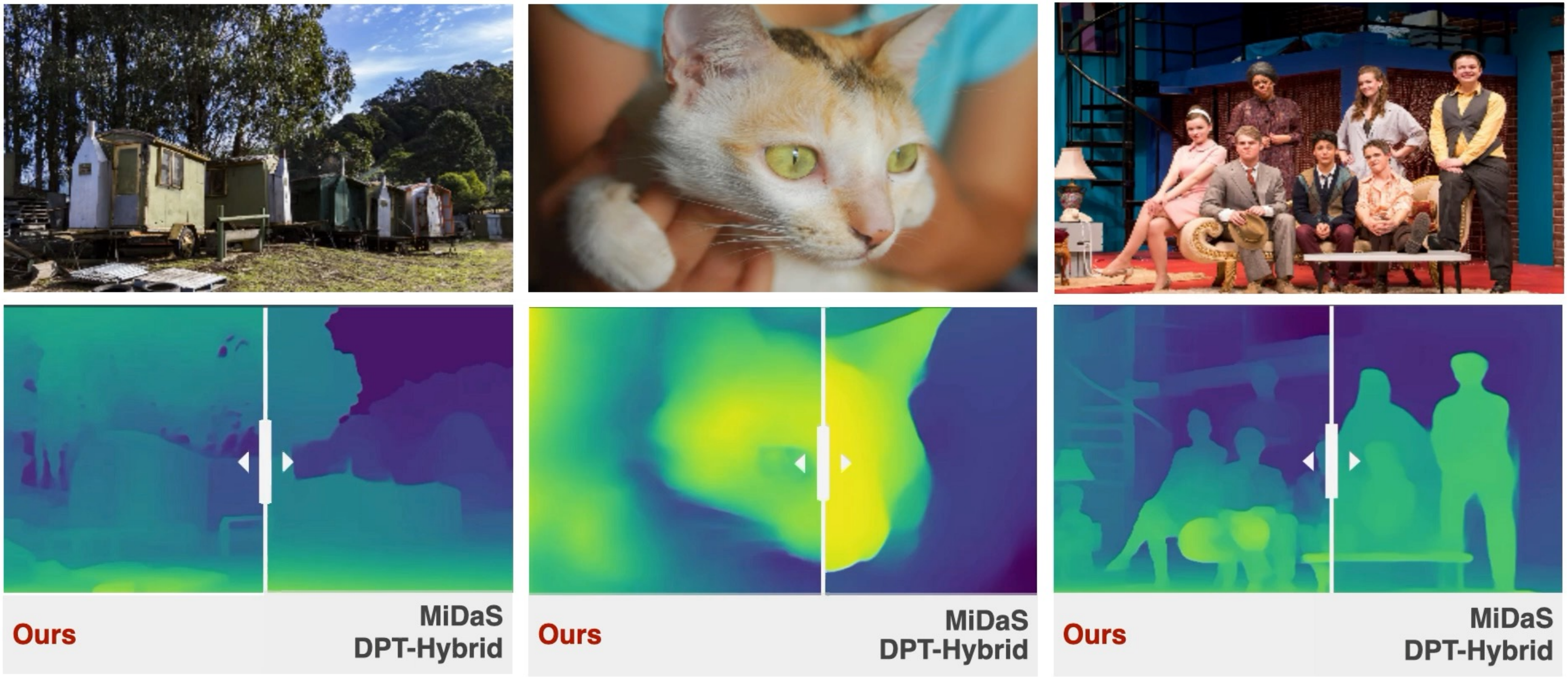
Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020

Depth Map Prediction from a Single Image



Eigen, Puhrsch and Fergus: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. NIPS, 2014

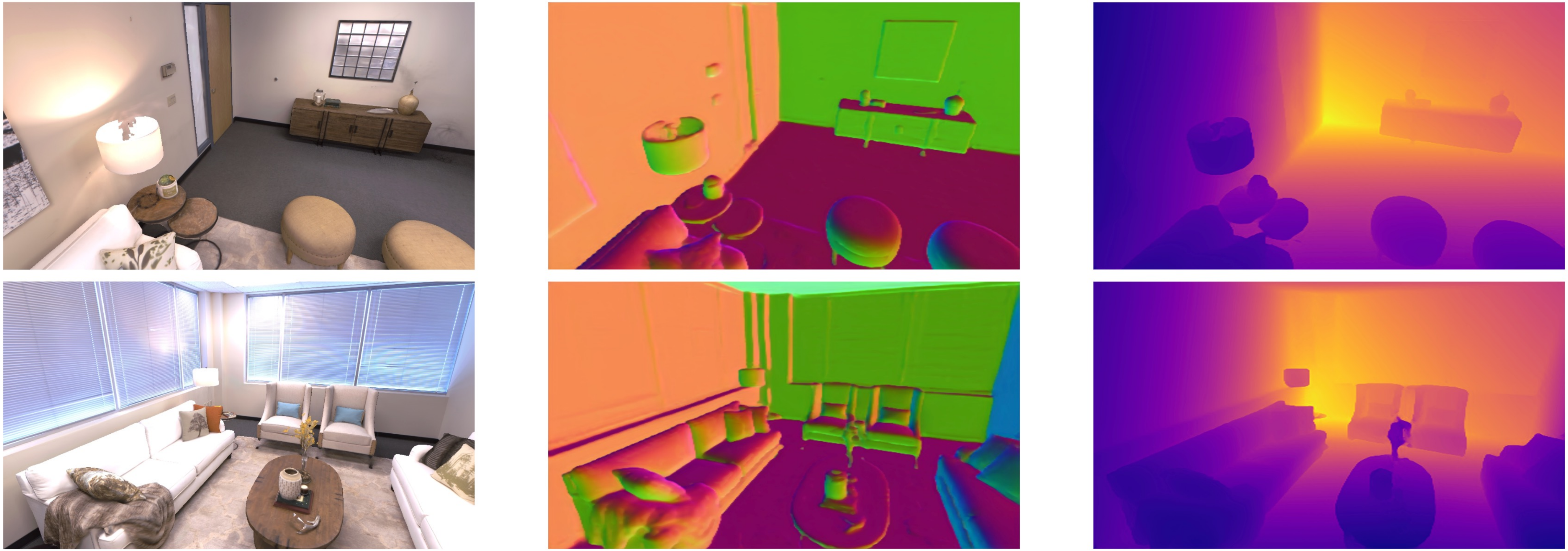
Omnidata



[Ranftl et al. 2021]

Eftekhar, Sax, Malik and Zamir: Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. ICCV, 2021.

Omnidata



RGB Image

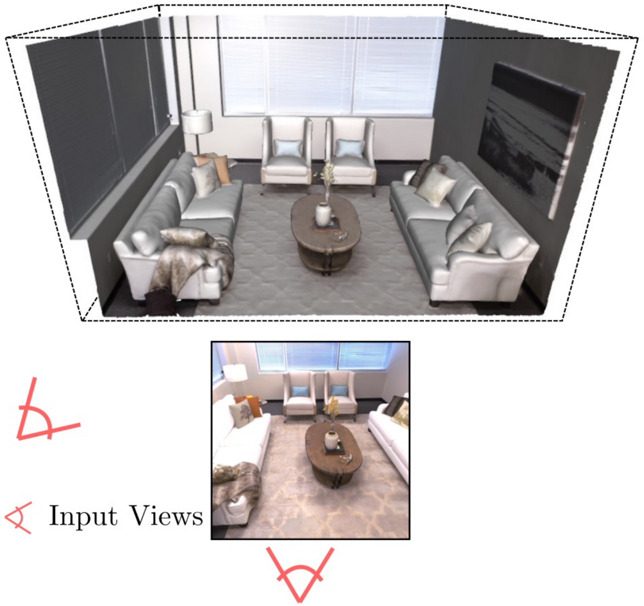
Omnidata Normal

Omnidata Depth

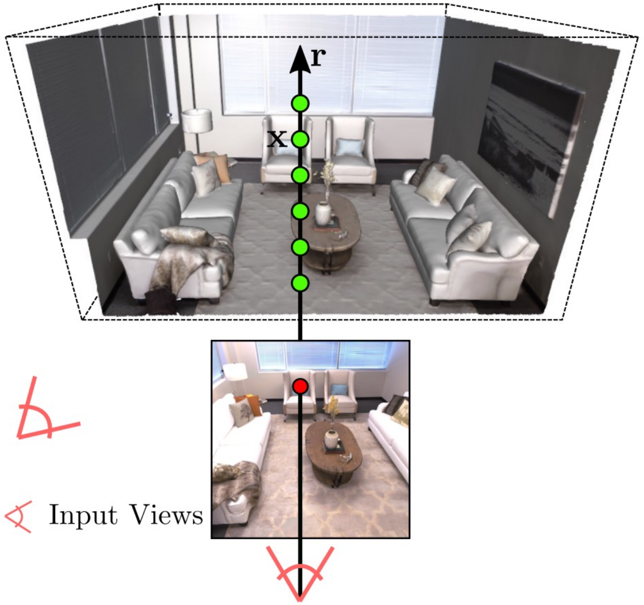
Eftekhar, Sax, Malik and Zamir: Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. ICCV, 2021.

MonoSDF

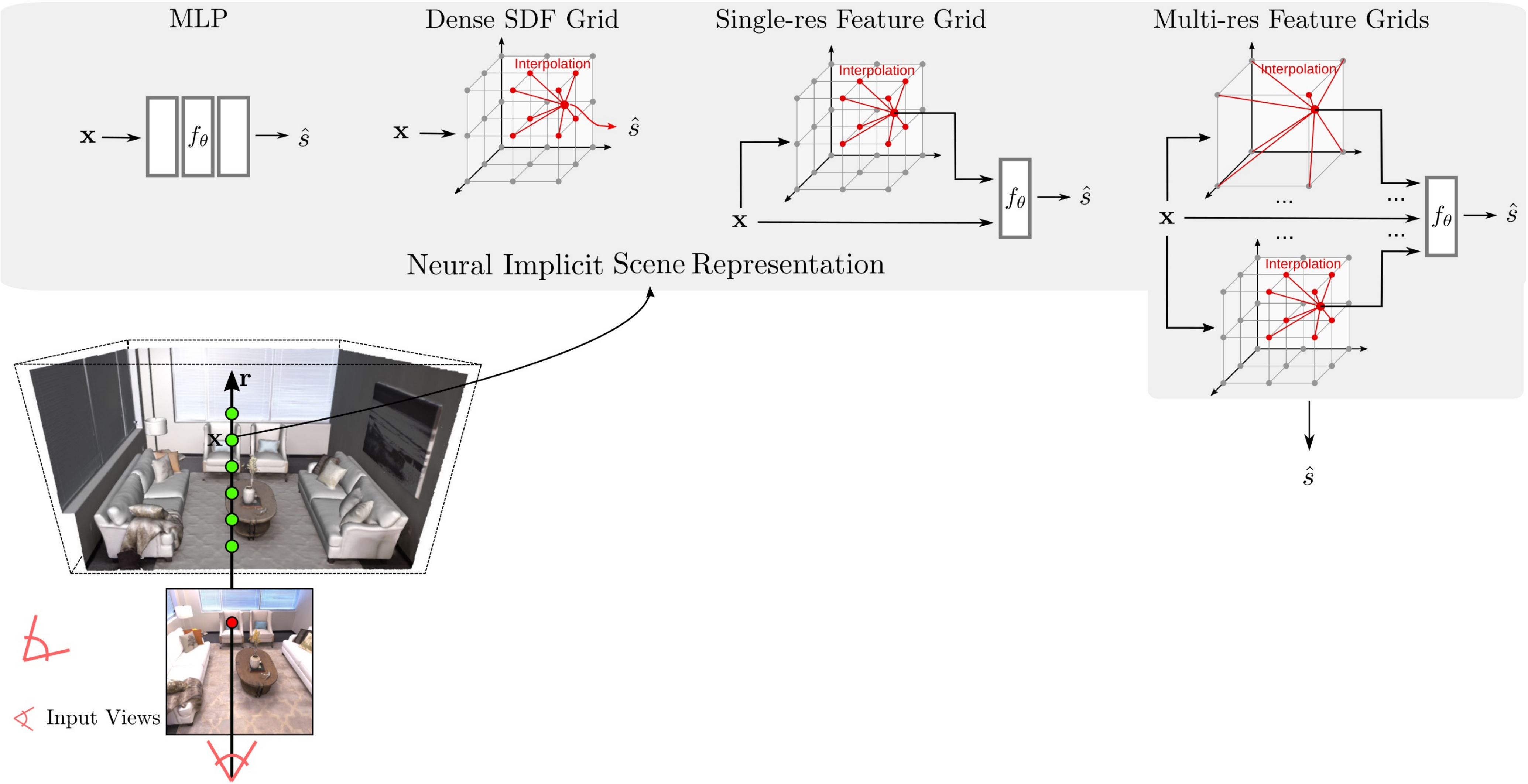
MonoSDF



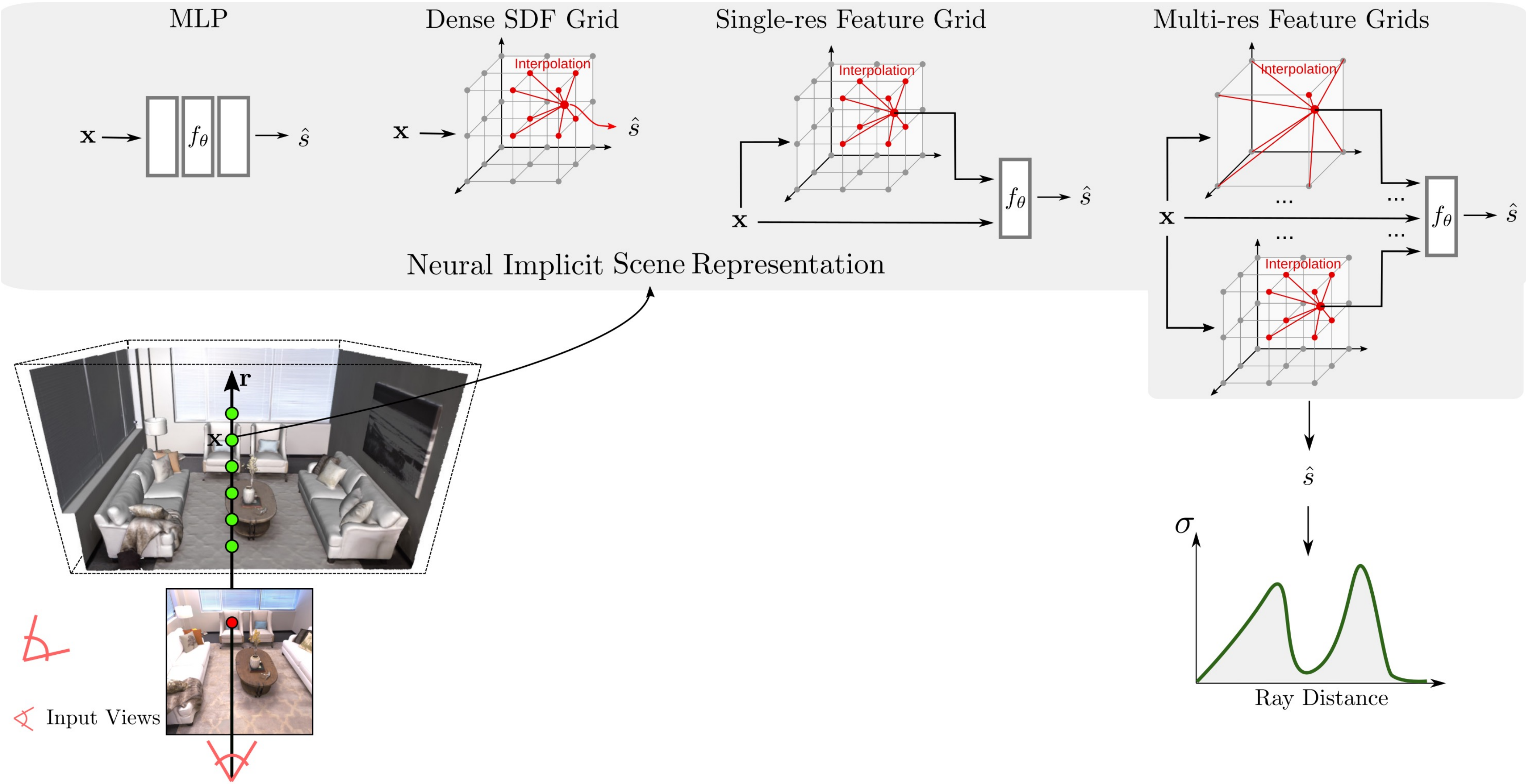
MonoSDF



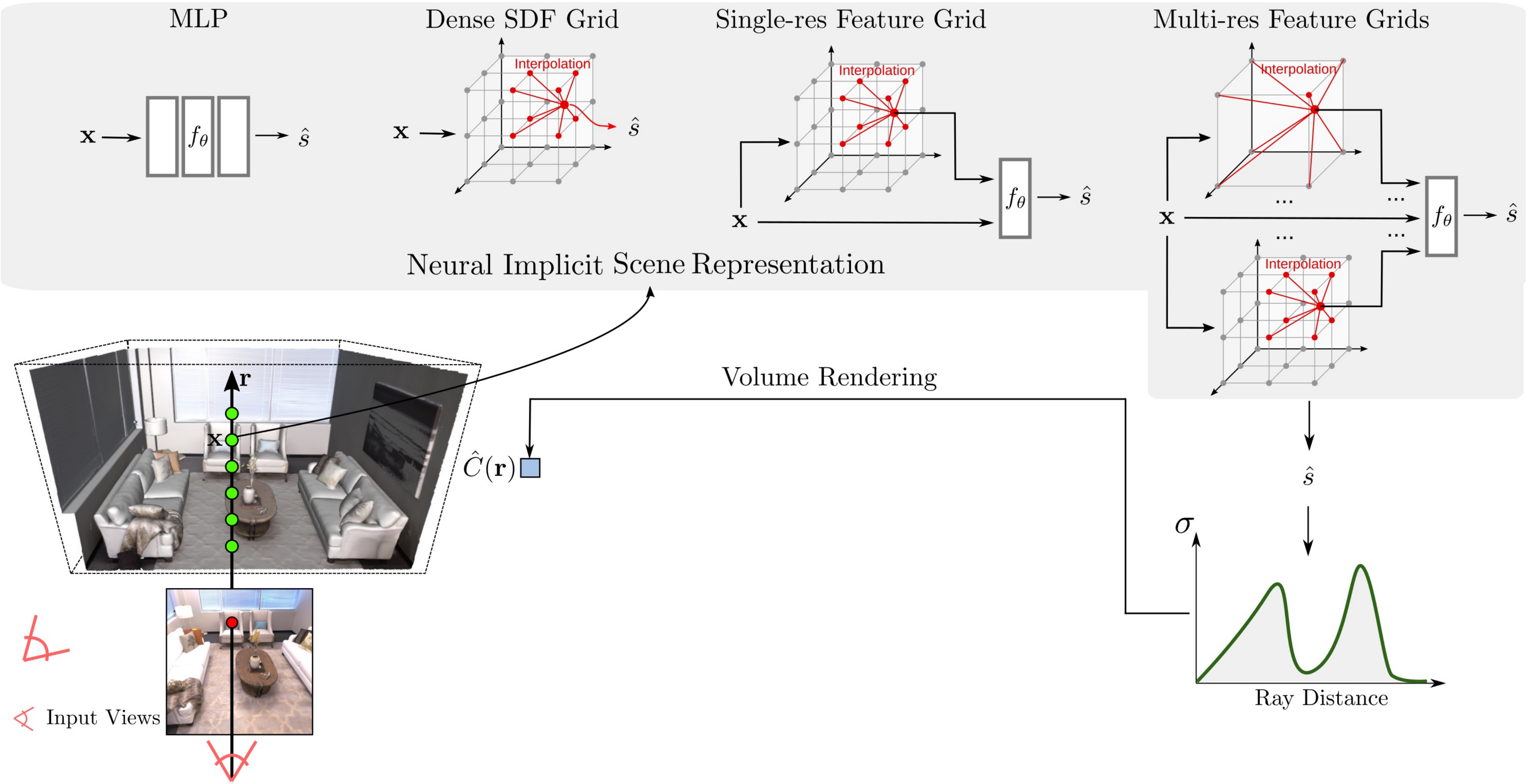
MonoSDF


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
x
f_θ
$\hat{s}$
...
Neural Implicit Scene Representation
r
Input Views

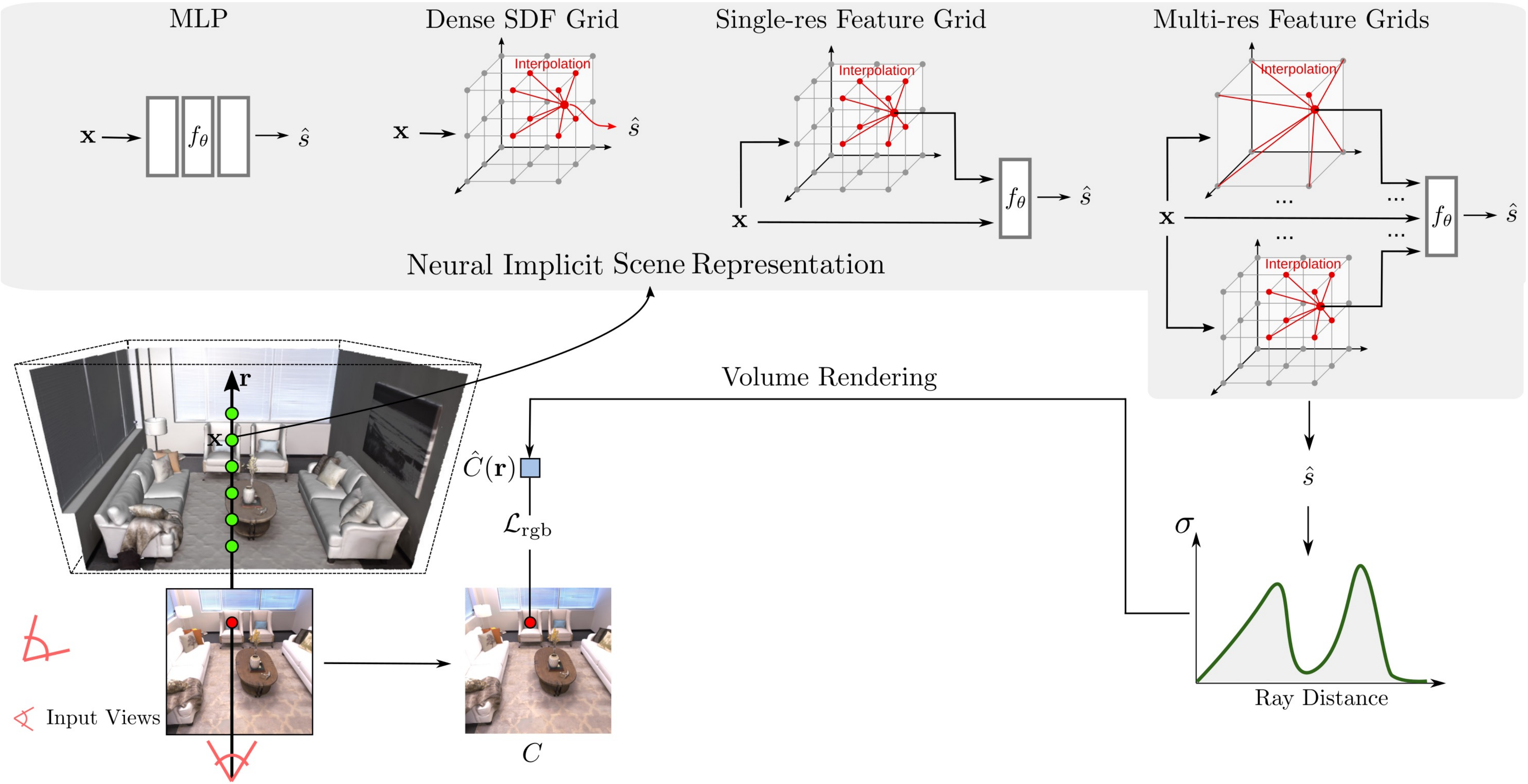
MonoSDF


MLP
$\mathbf{x}$
f_θ
$\hat{s}$
Dense SDF Grid
Interpolation
$\mathbf{x}$
$\hat{s}$
Single-res Feature Grid
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Multi-res Feature Grids
Interpolation
...
$\mathbf{x}$
f_θ
$\hat{s}$
Interpolation
Neural Implicit Scene Representation
$\mathbf{r}$
$\mathbf{x}$
Input Views
$\hat{s}$
σ
Ray Distance

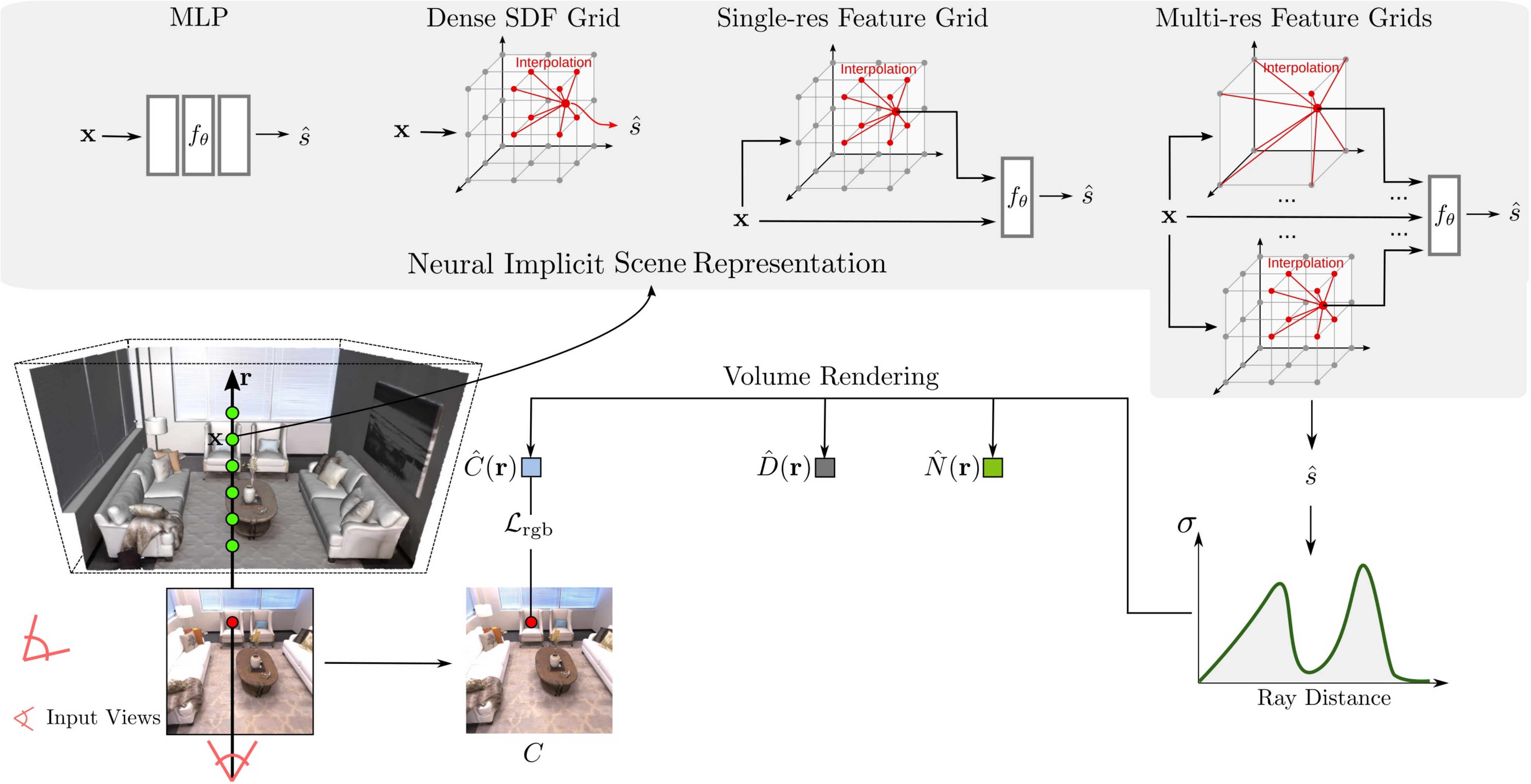
MonoSDF


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
$\mathbf{x}$
f_θ
$\hat{s}$
Interpolation
...
Neural Implicit Scene Representation
Volume Rendering
$\mathbf{r}$
$\hat{C}(\mathbf{r})$
σ
Ray Distance
Input Views

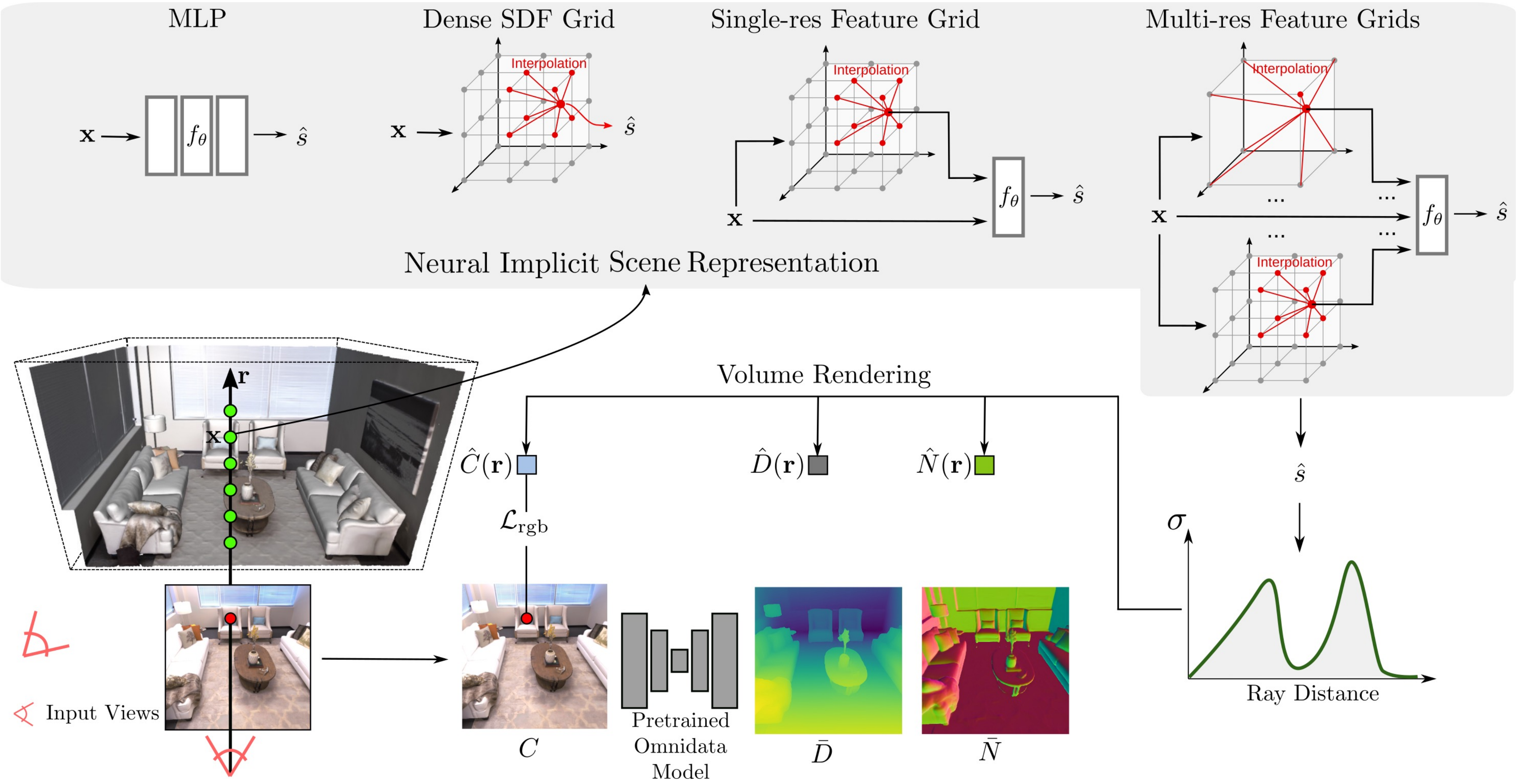
MonoSDF


MLP
$\mathbf{x}$
f_θ
$\hat{s}$
Dense SDF Grid
Interpolation
$\mathbf{x}$
$\hat{s}$
Single-res Feature Grid
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Multi-res Feature Grids
Interpolation
...
$\mathbf{x}$
f_θ
$\hat{s}$
Interpolation
Neural Implicit Scene Representation
Volume Rendering
$\mathbf{r}$
$\mathbf{x}$
$\hat{C}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
Input Views
$\hat{s}$
σ
Ray Distance

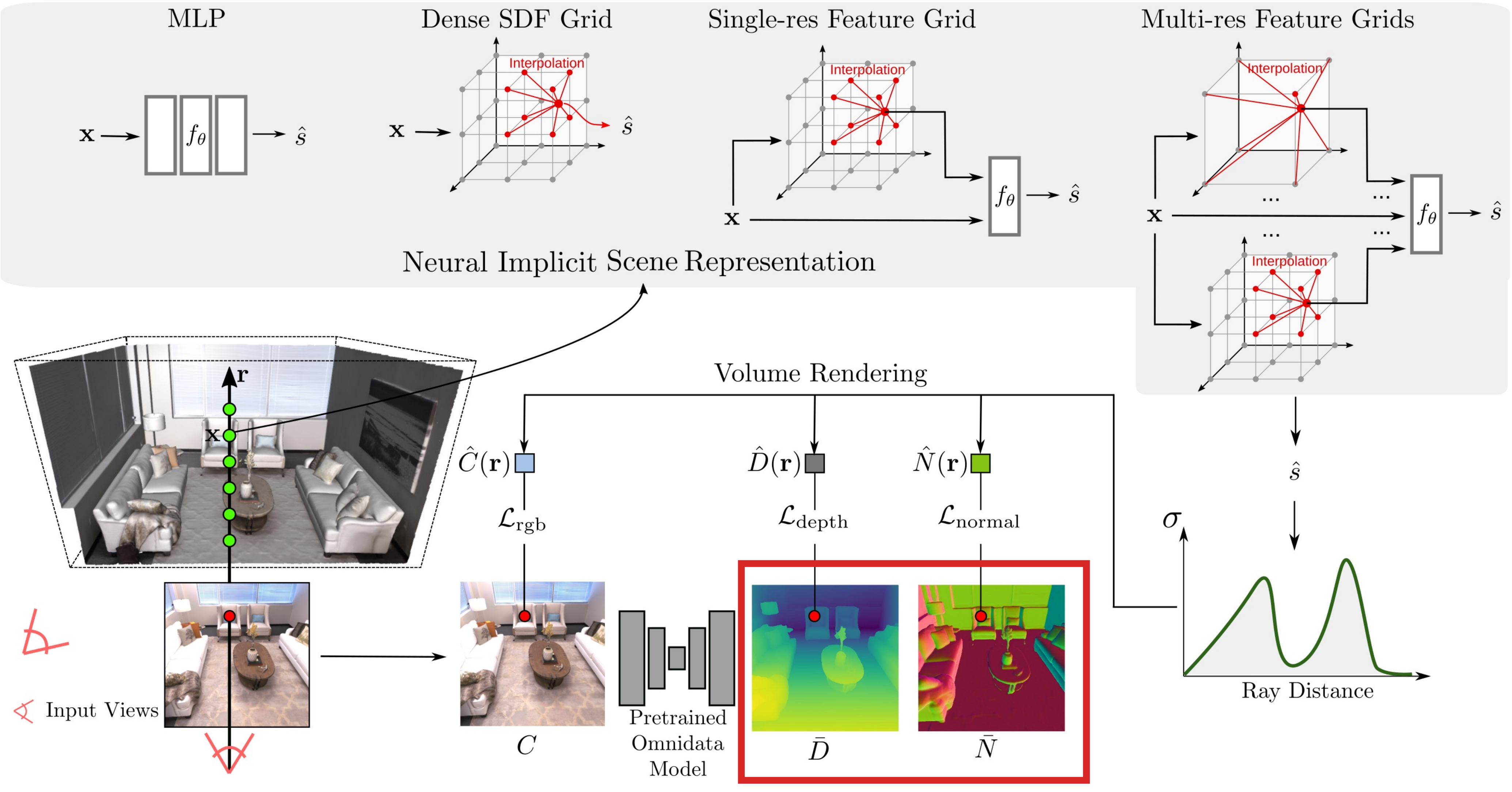
MonoSDF


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
...
Neural Implicit Scene Representation
$\mathbf{r}$
Volume Rendering
$\hat{C}(\mathbf{r})$
$\hat{D}(\mathbf{r})$
$\hat{N}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
σ
Ray Distance
Input Views

MonoSDF


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
$\mathbf{x}$
f_θ
$\hat{s}$
Interpolation
...
Neural Implicit Scene Representation
Volume Rendering
$\mathbf{r}$
$\hat{C}(\mathbf{r})$
$\hat{D}(\mathbf{r})$
$\hat{N}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
σ
Ray Distance
Input Views
C
Pretrained Omnidata Model
$\bar{D}$
$\bar{N}$

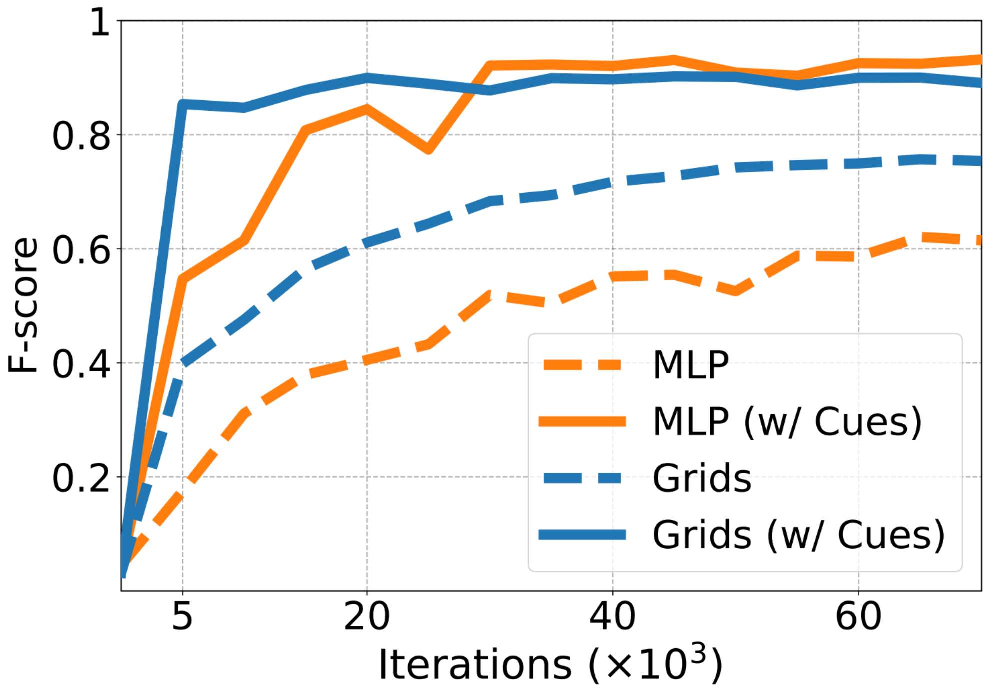
MonoSDF



Monocular Geometric Cues

Ablation Study

		Normal C.↑	Chamfer-L_1 ↓	F-score ↑
MLP	No Cues	86.48	6.75	66.88
	Only Depth	90.56	4.26	76.42
	Only Normal	91.35	3.19	85.84
	Both Cues	**92.11**	**2.94**	**86.18**
Multi-Res. Grids	No Cues	87.95	5.03	78.38
	Only Depth	90.87	3.75	80.32
	Only Normal	89.90	3.61	81.28
	Both Cues	**90.93**	**3.23**	**85.91**

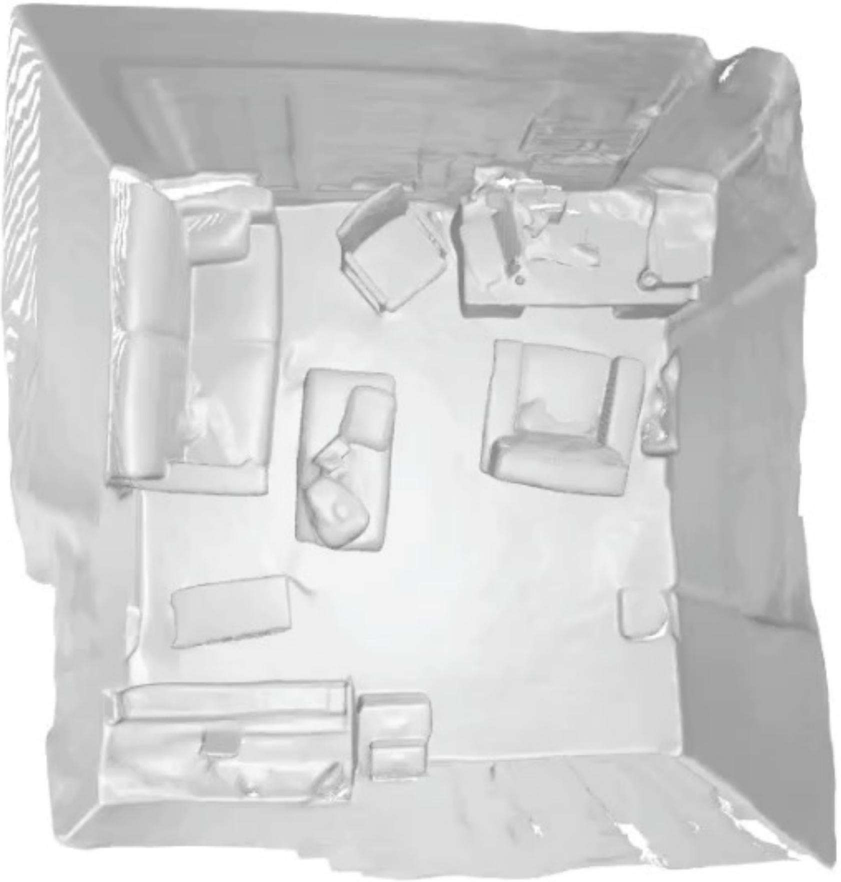


! Monocular cues improve reconstruction results significantly

! Combining **depth & normal** leads to best performance

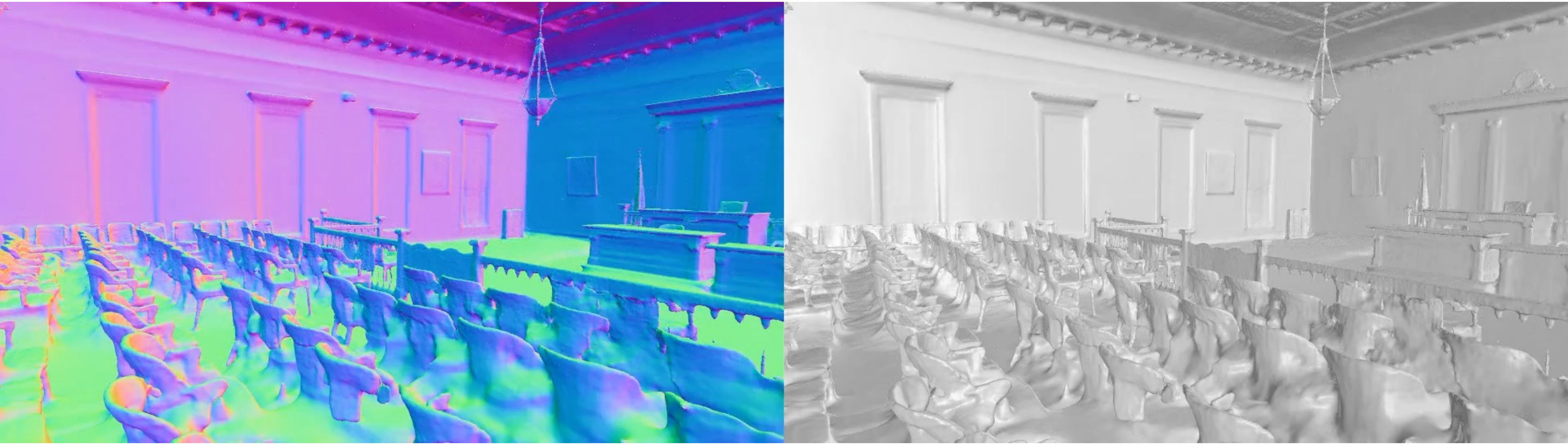
! Monocular cues can improve **convergence speed**

Baseline Comparisons on ScanNet

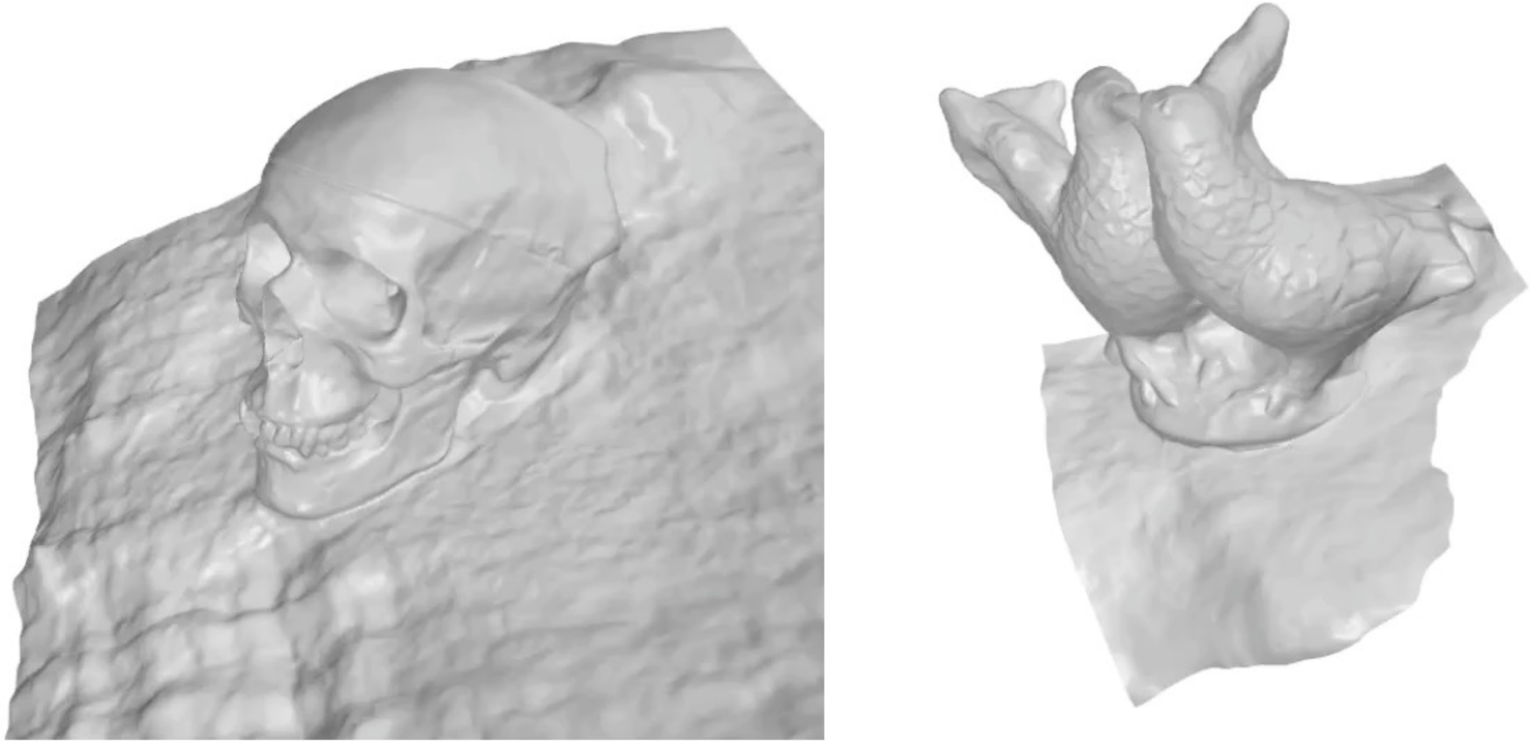


Ours

Multi-Res. Feature Grids with High-Res. Cues



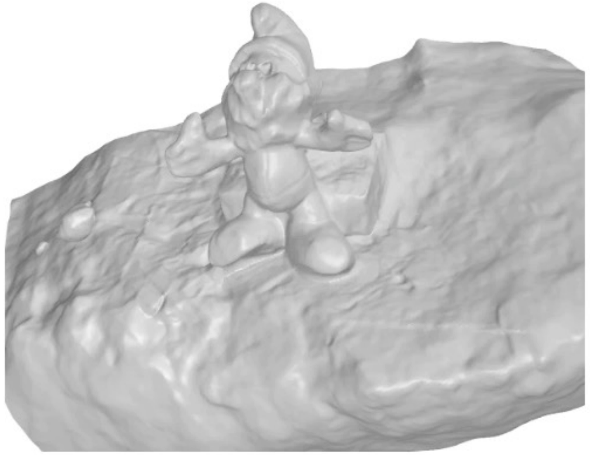
Baseline Comparisons on DTU (3-views)



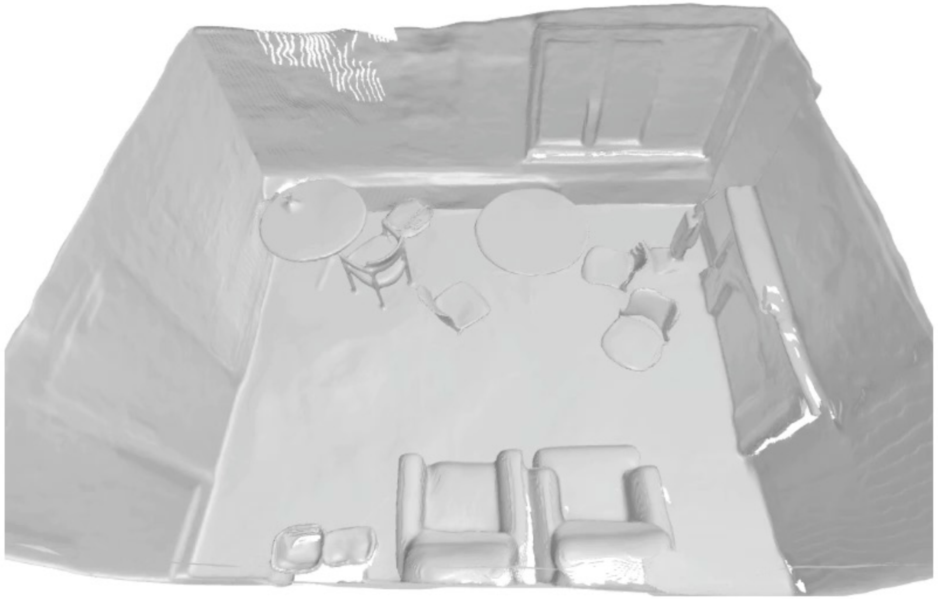
Ours

Take-home Message

https://niujinshuchong.github.io/monosdf/



DTU (3 views)



ScanNet

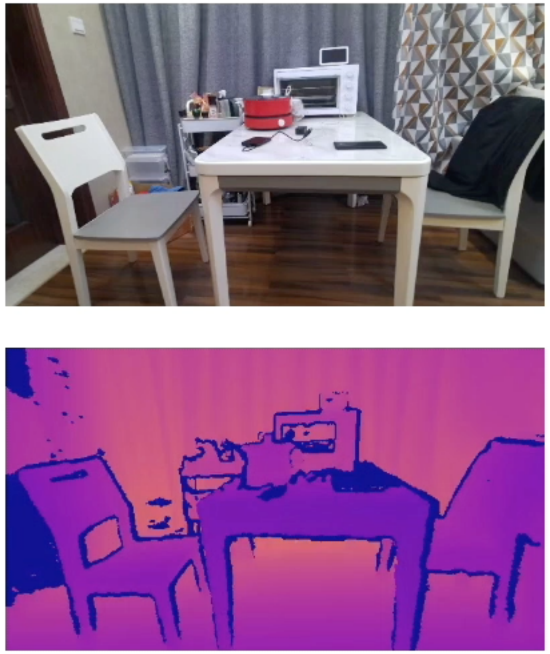
Tanks and Temples

! **Monocular cues** improve reconstruction results and speed up optimization

! Inspire applications in other fields [GOOD, ICLR 2023]

! Limitation: Still require camera poses given :(

RGB-D Sequences



40x Speed

NICE-SLAM

Neural Implicit Scalable Encoding for SLAM

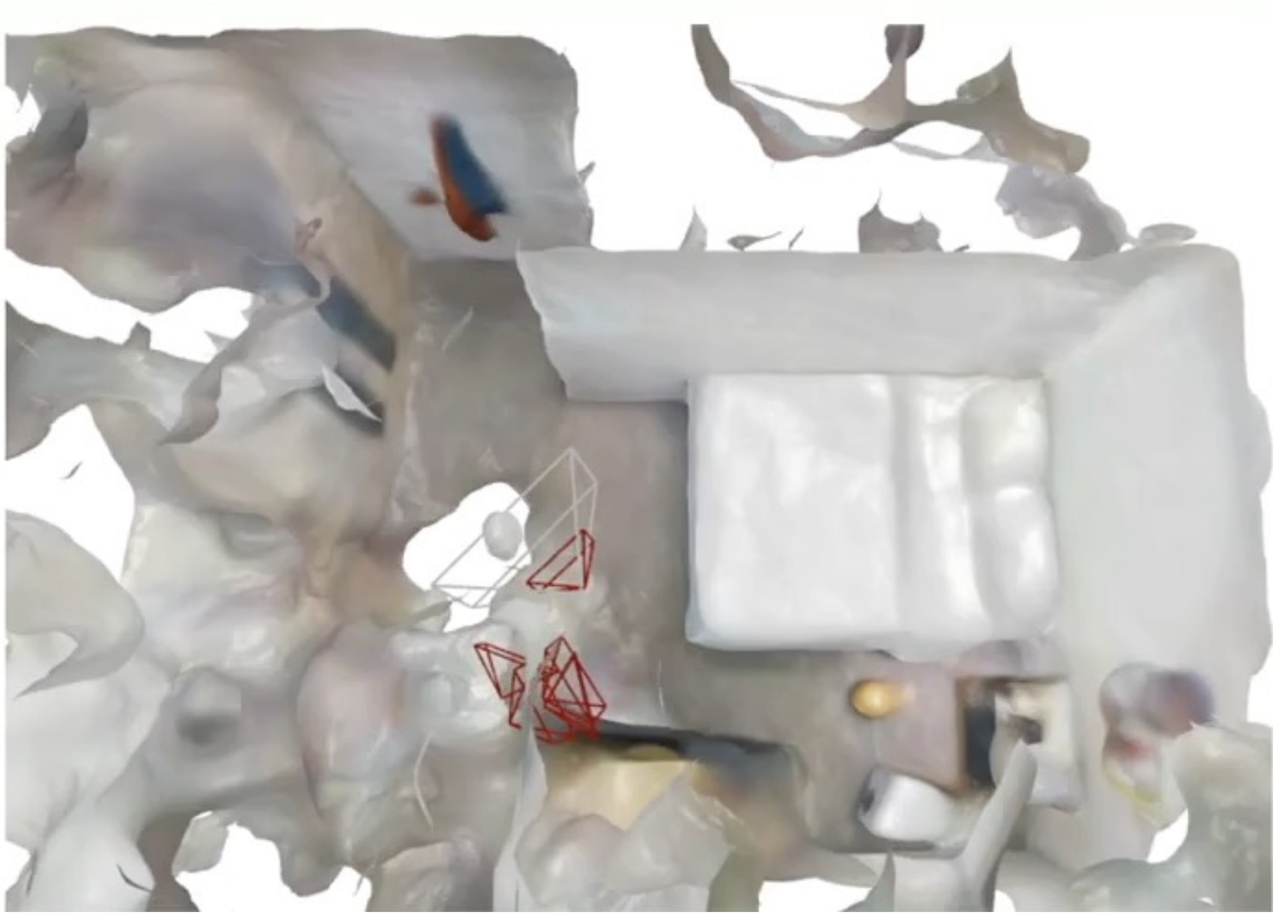
CVPR 2022

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

* Equal Contributions

iMAP

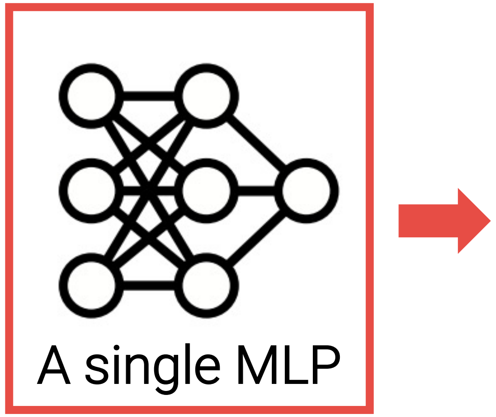
[Sucar et al., ICCV’21]



First neural implicit-based online SLAM system

iMAP

[Sucar et al., ICCV’21]



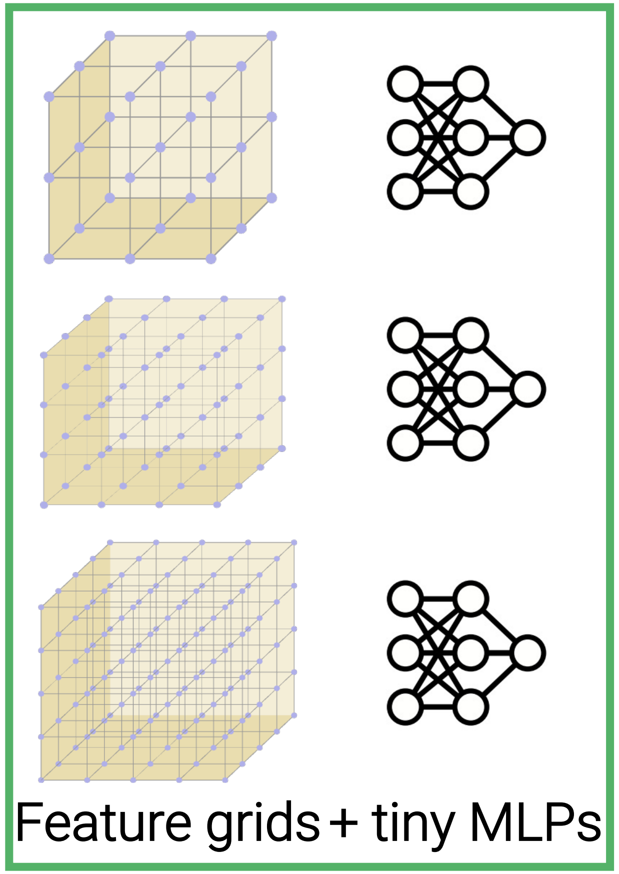
A single MLP

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

🟥 Predicted Poses

▬ GT Poses

NICE-SLAM



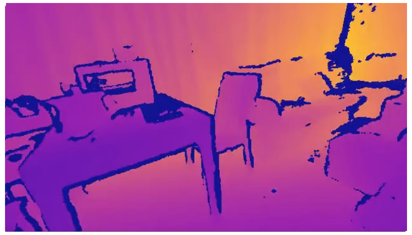
- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence

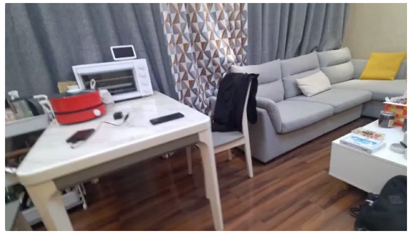
— Predicted Poses

— GT Poses

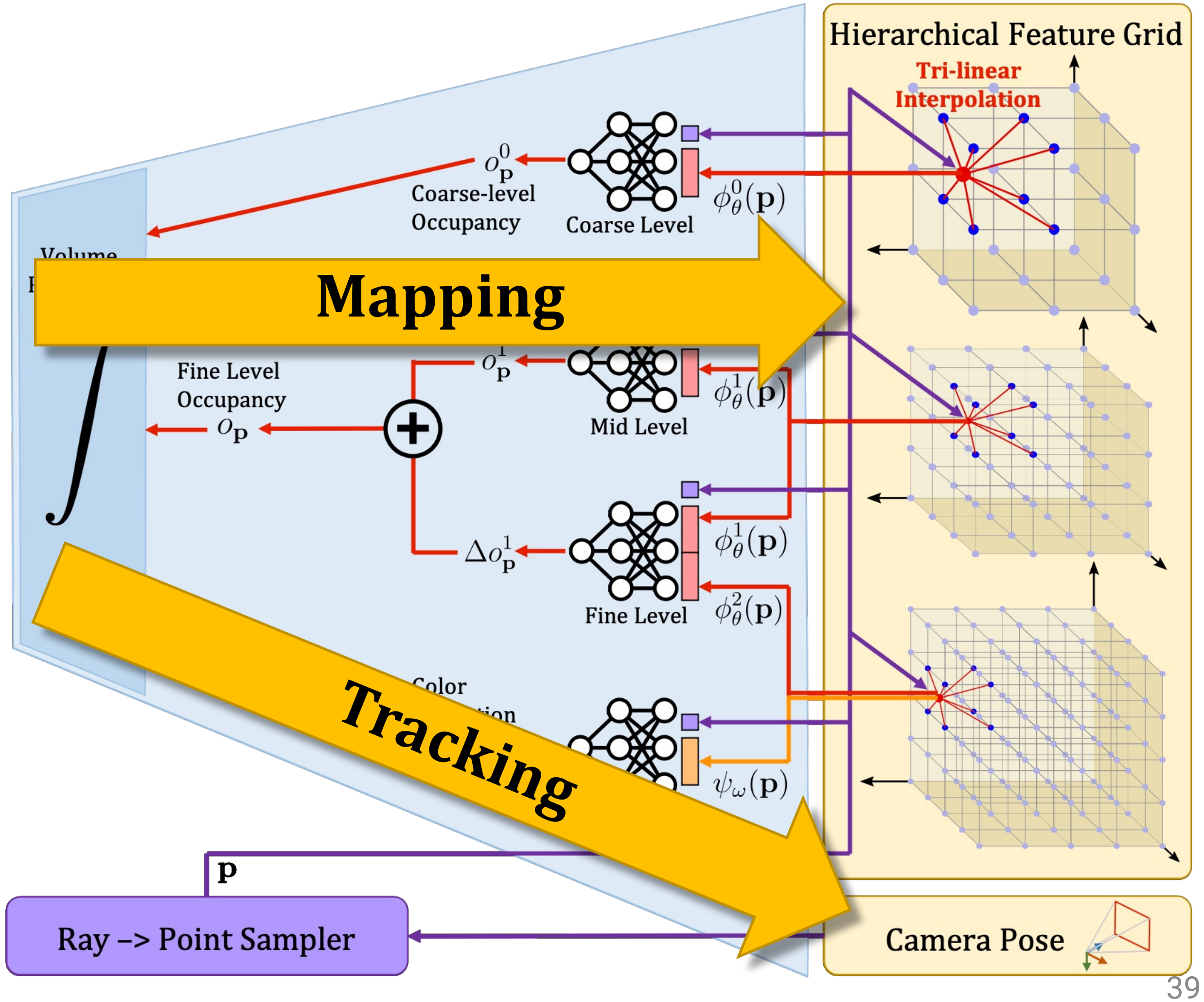
Pipeline

Input Depth





Input RGB



Results

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

4x Speed

— Predicted Poses

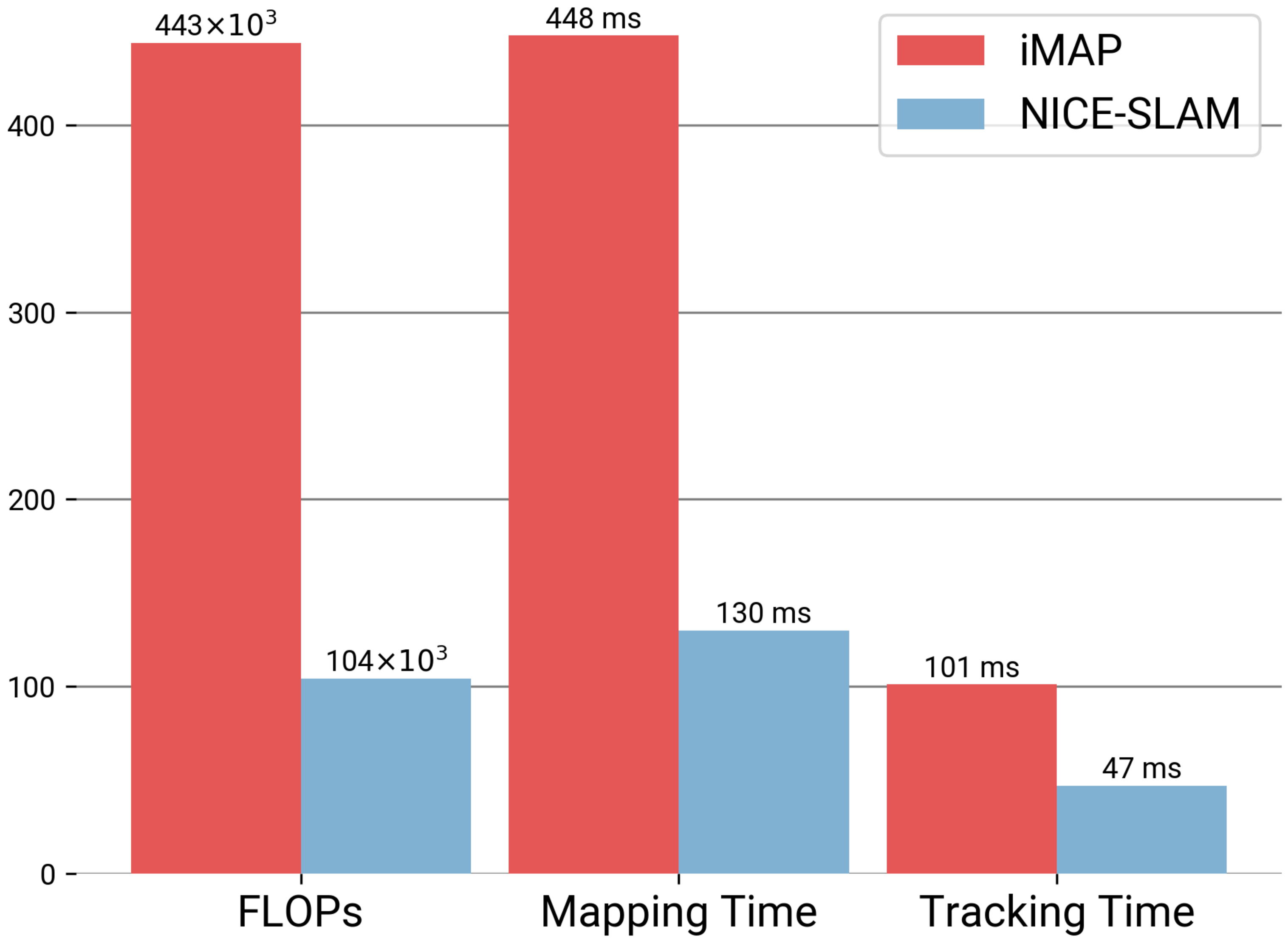
— GT Poses

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]

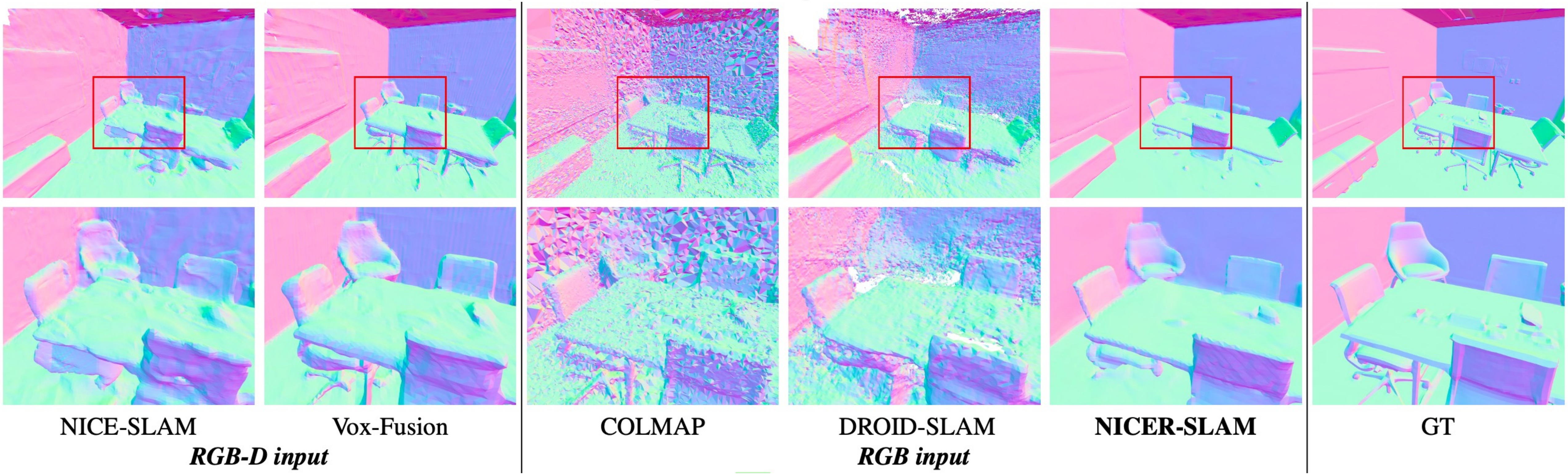
Limitations

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

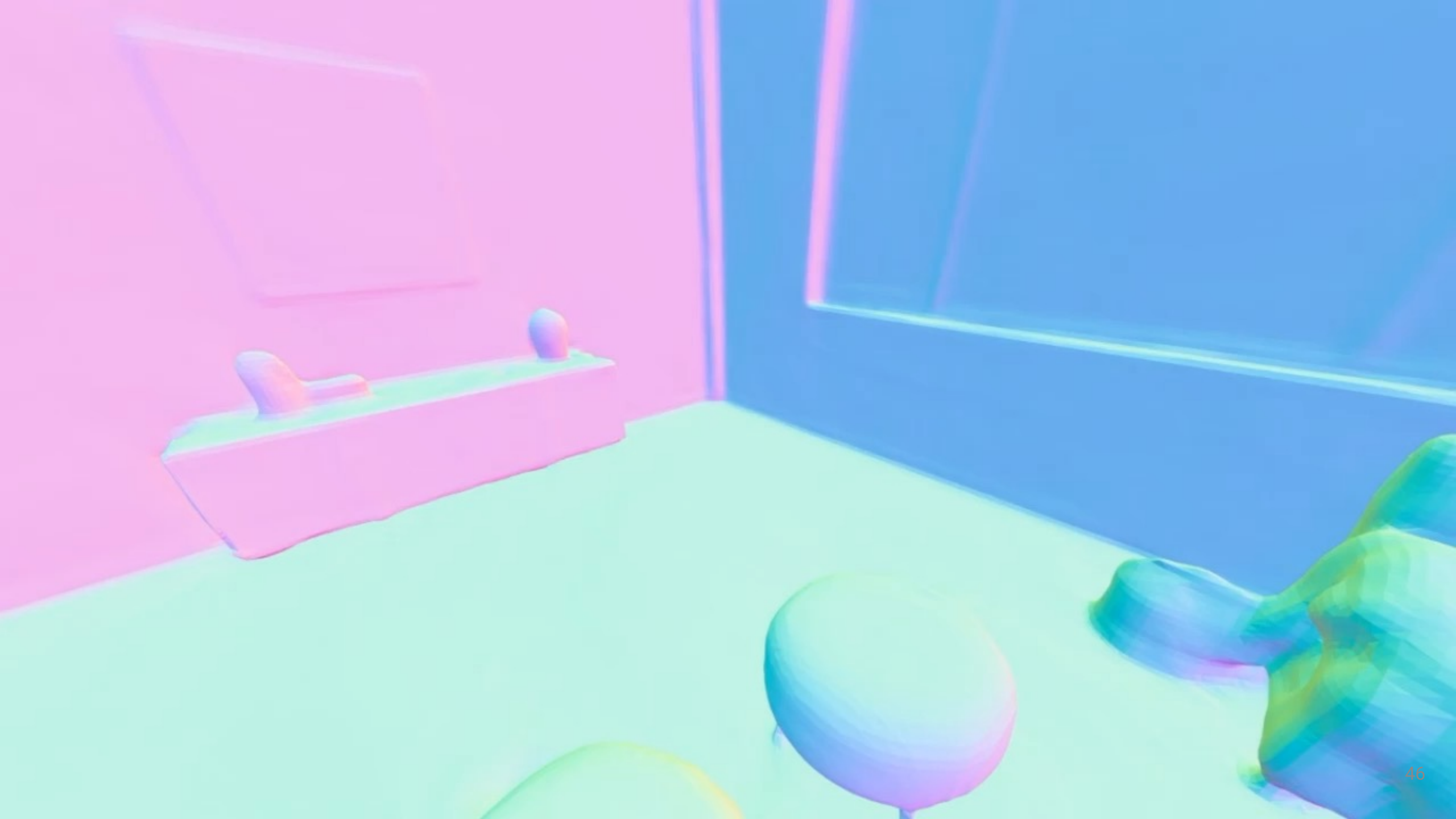
NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM

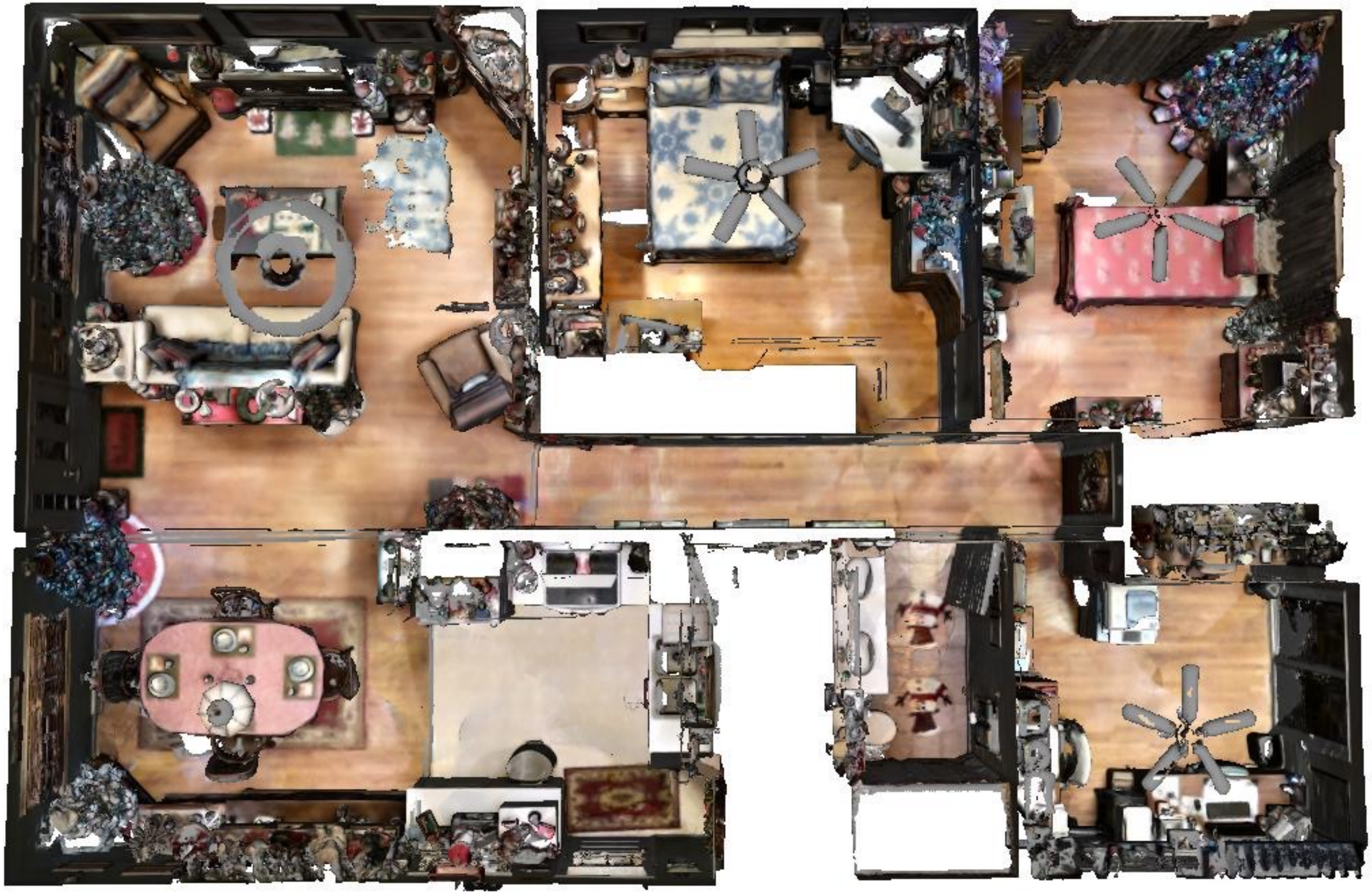
Zihan Zhu[1*] Songyou Peng[1,2*] Viktor Larsson[3] Zhaopeng Cui[4]
Martin R. Oswald[1,5] Andreas Geiger[6] Marc Pollefeys[1,7]

[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

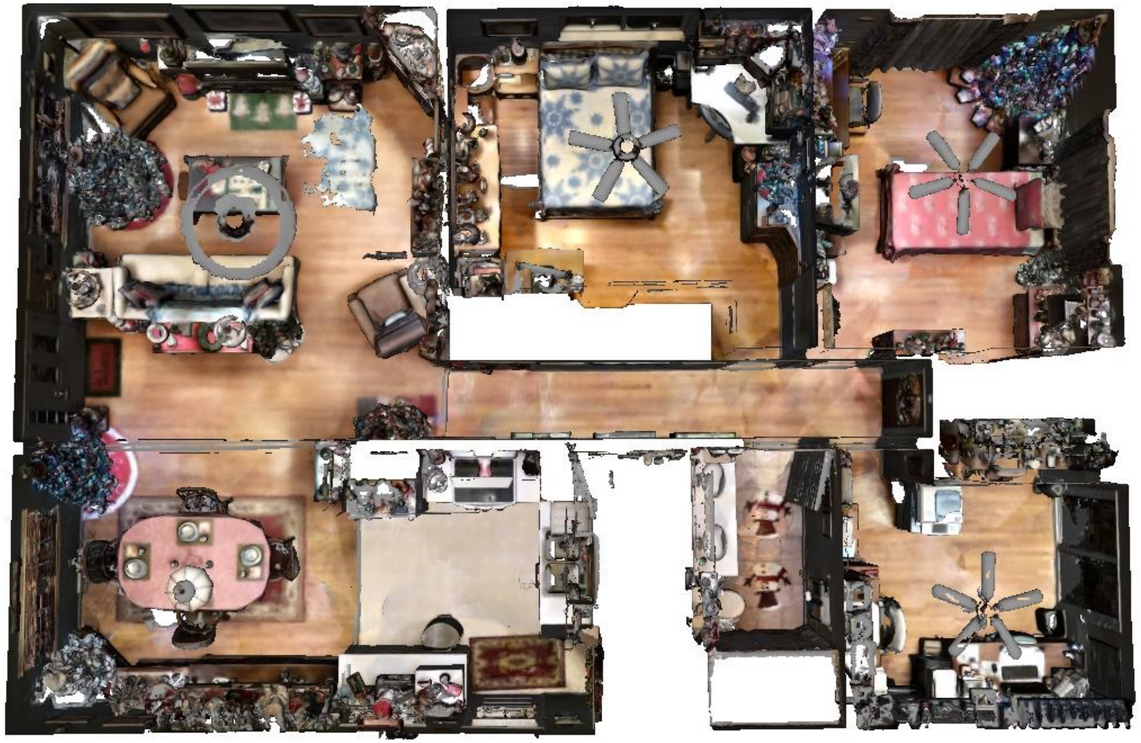


https://arxiv.org/abs/2302.03594

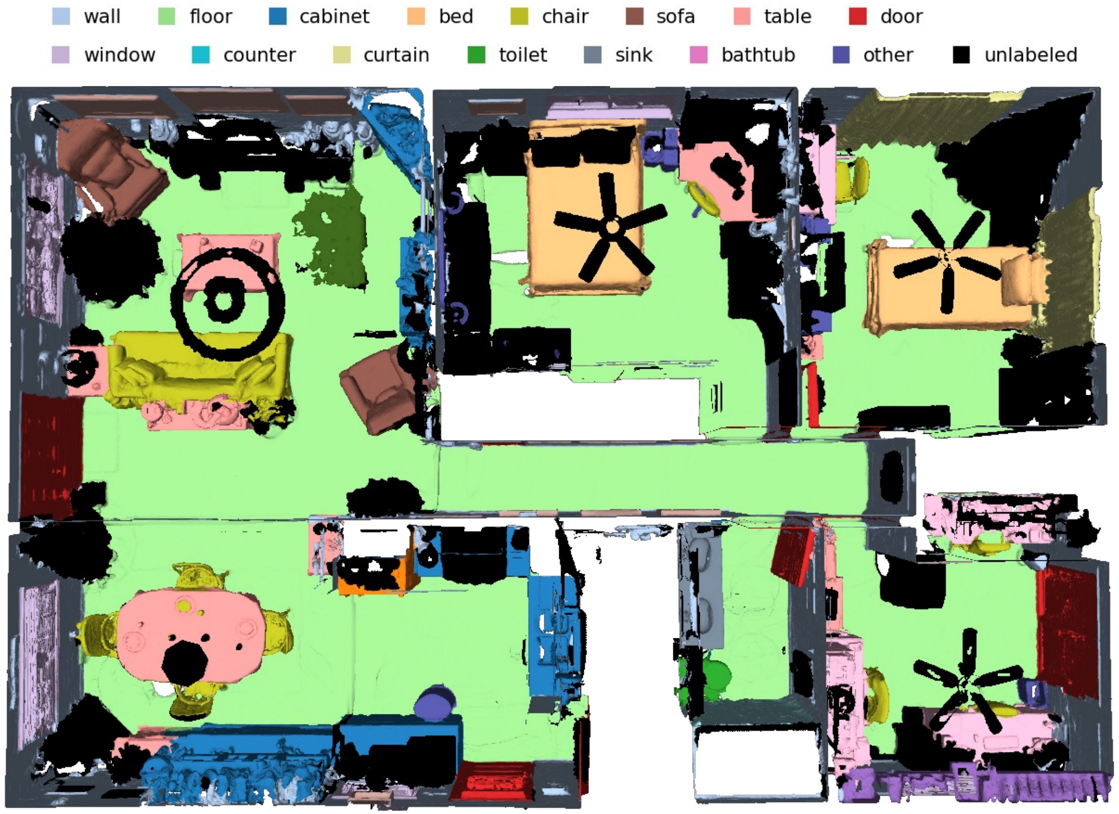




Input 3D Geometry



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

Google ETH zürich SFU

OpenScene

3D Scene Understanding with Open Vocabularies

CVPR 2023

Songyou Peng

Kyle Genova

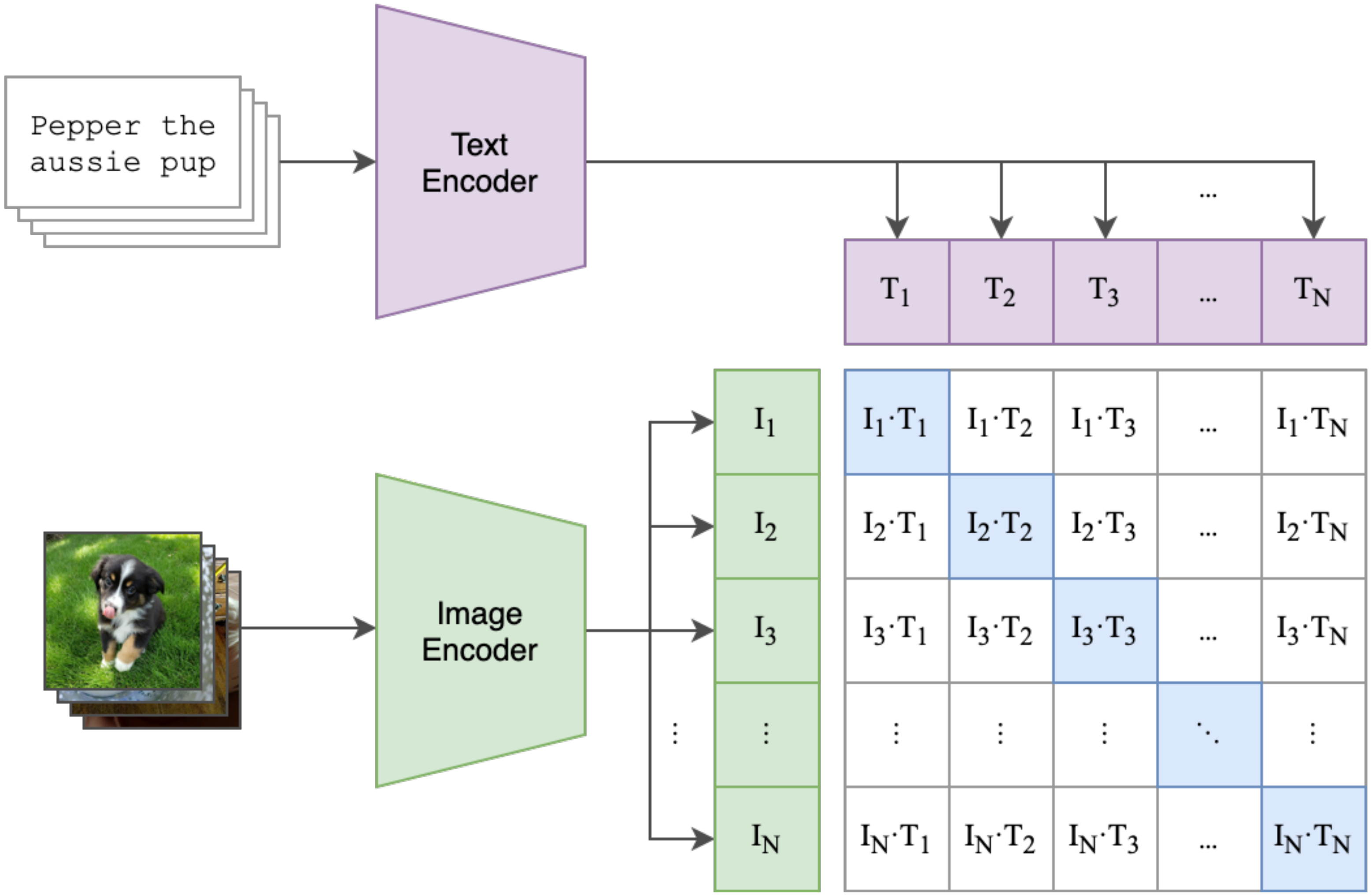
Chiyu “Max” Jiang

Andrea Tagliasacchi

Marc Pollefeys

Tom Funkhouser

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

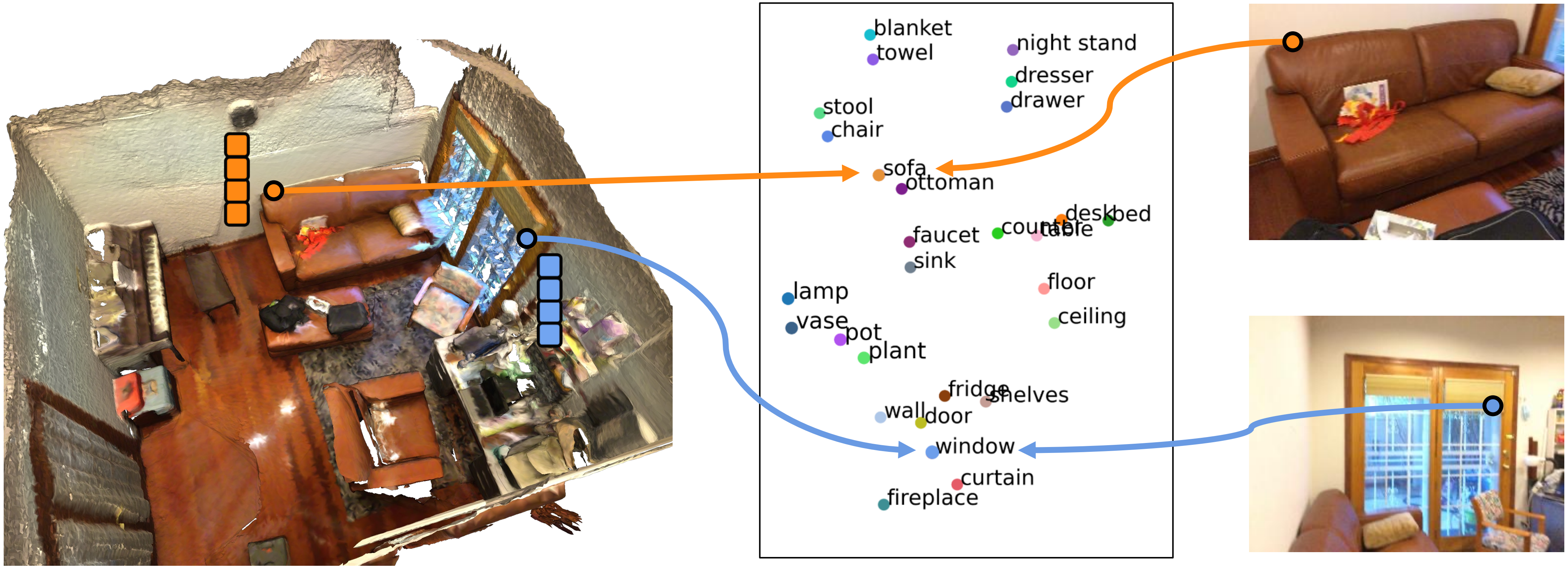


$$
\begin{array}{c|ccccc}
 & T_1 & T_2 & T_3 & \dots & T_N \\
\hline
I_1 & I_1 \cdot T_1 & I_1 \cdot T_2 & I_1 \cdot T_3 & \dots & I_1 \cdot T_N \\
I_2 & I_2 \cdot T_1 & I_2 \cdot T_2 & I_2 \cdot T_3 & \dots & I_2 \cdot T_N \\
I_3 & I_3 \cdot T_1 & I_3 \cdot T_2 & I_3 \cdot T_3 & \dots & I_3 \cdot T_N \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
I_N & I_N \cdot T_1 & I_N \cdot T_2 & I_N \cdot T_3 & \dots & I_N \cdot T_N
\end{array}
$$

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



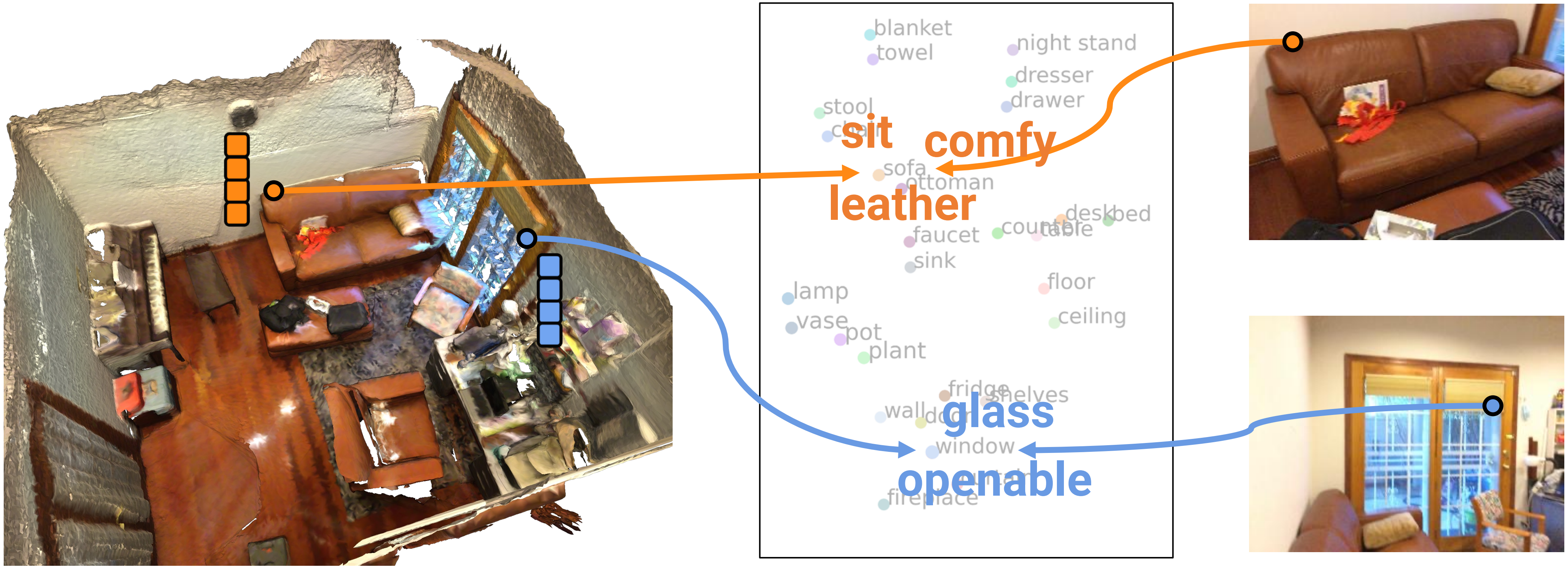
3D Geometry

CLIP Text Features

(visualize with T-SNE)

RGB Images

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



3D Geometry

CLIP Text Features

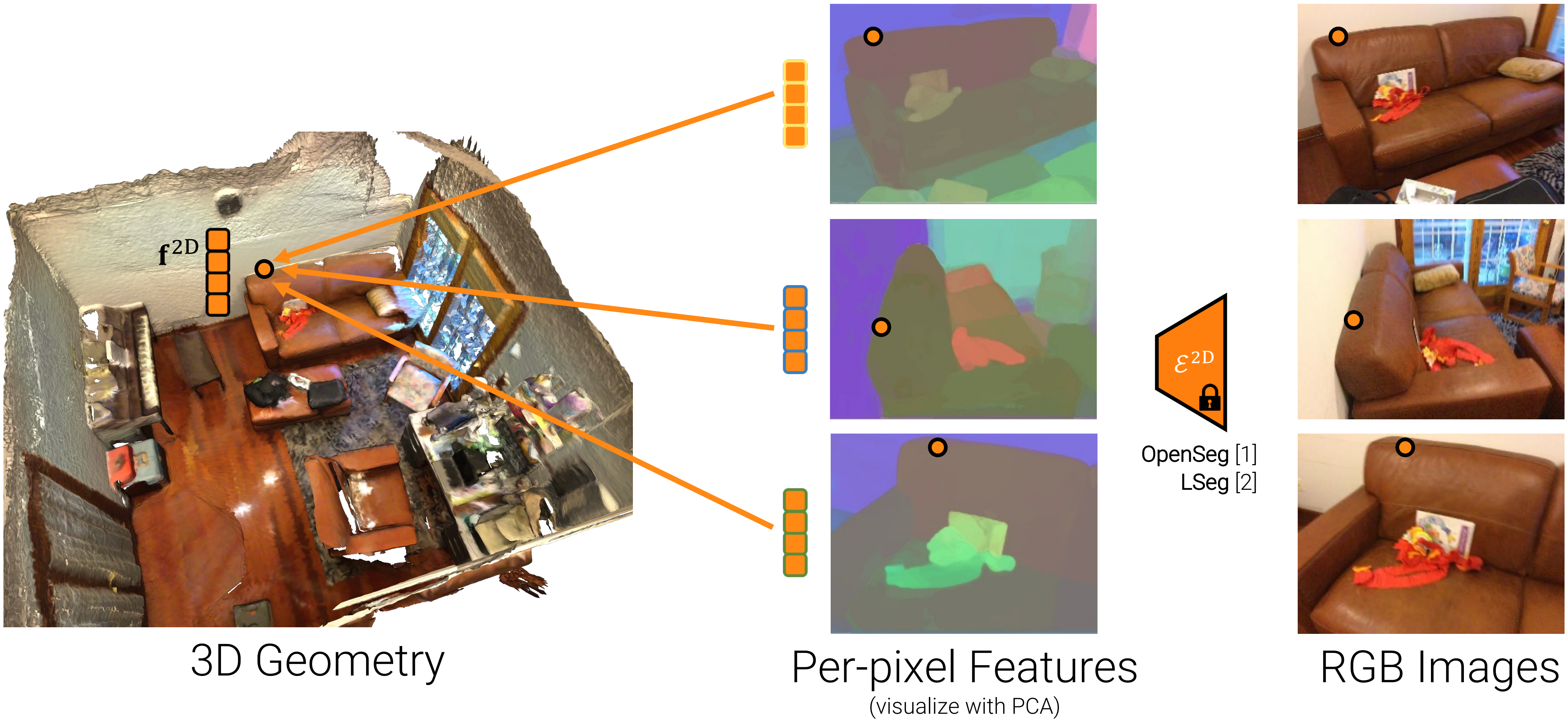
(visualize with T-SNE)

RGB Images

Note: bold word embeddings are approximate

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

Step 1: Multi-view Feature Fusion



3D Geometry

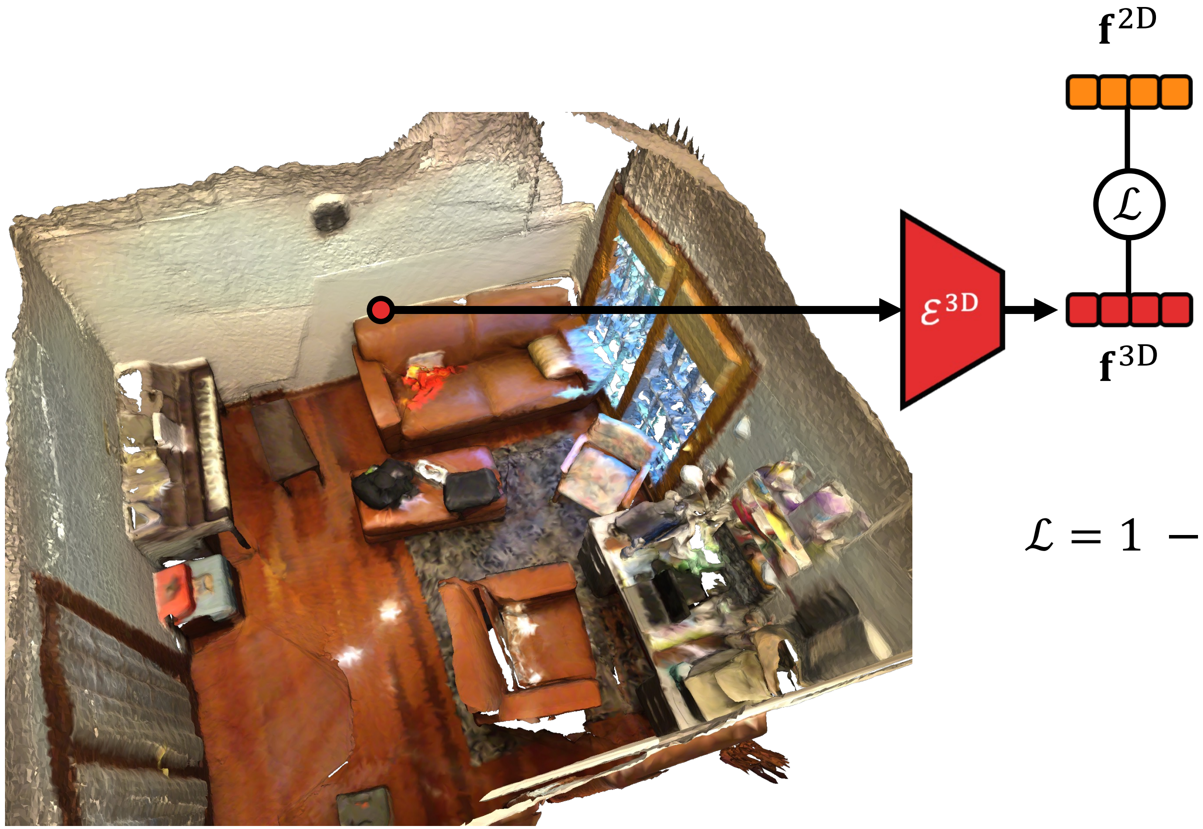
Per-pixel Features
(visualize with PCA)

OpenSeg [1]
LSeg [2]

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

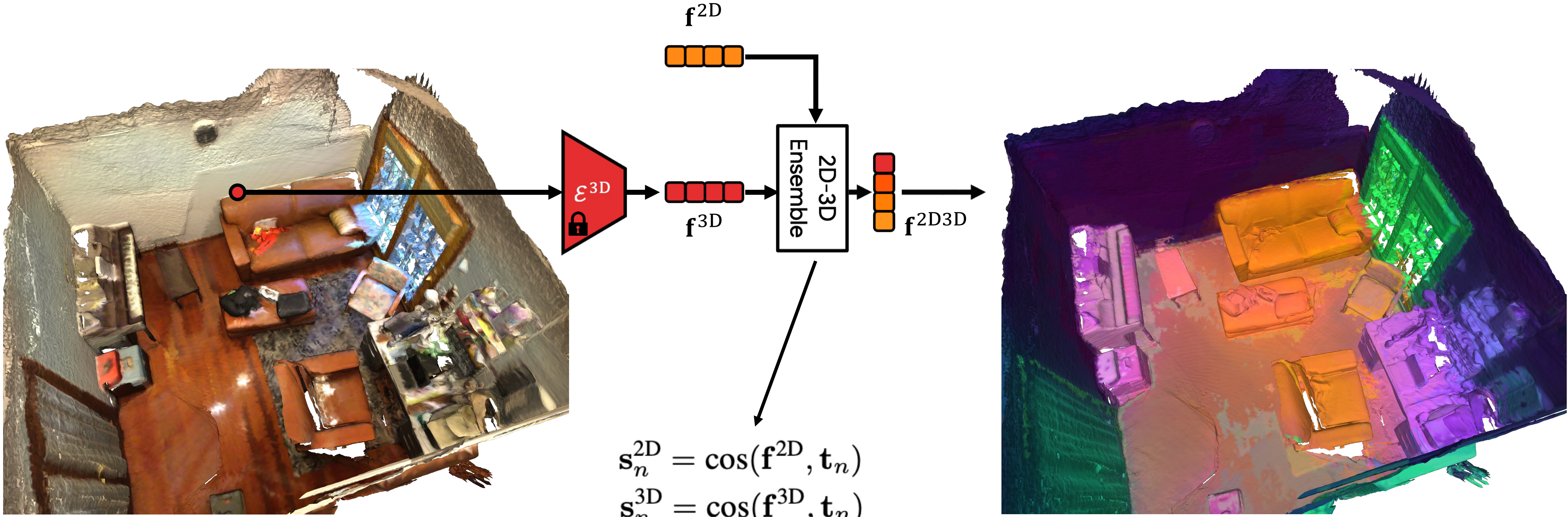
Step 2: 3D Distillation



3D Geometry

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

Step 3: 2D-3D Ensemble


$\mathbf{f}^{2\mathrm{D}}$
$\mathcal{E}^{3\mathrm{D}}$
$\mathbf{f}^{3\mathrm{D}}$
2D-3D Ensemble
$\mathbf{f}^{2\mathrm{D3D}}$

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

3D Geometry

2D-3D Ensemble Features

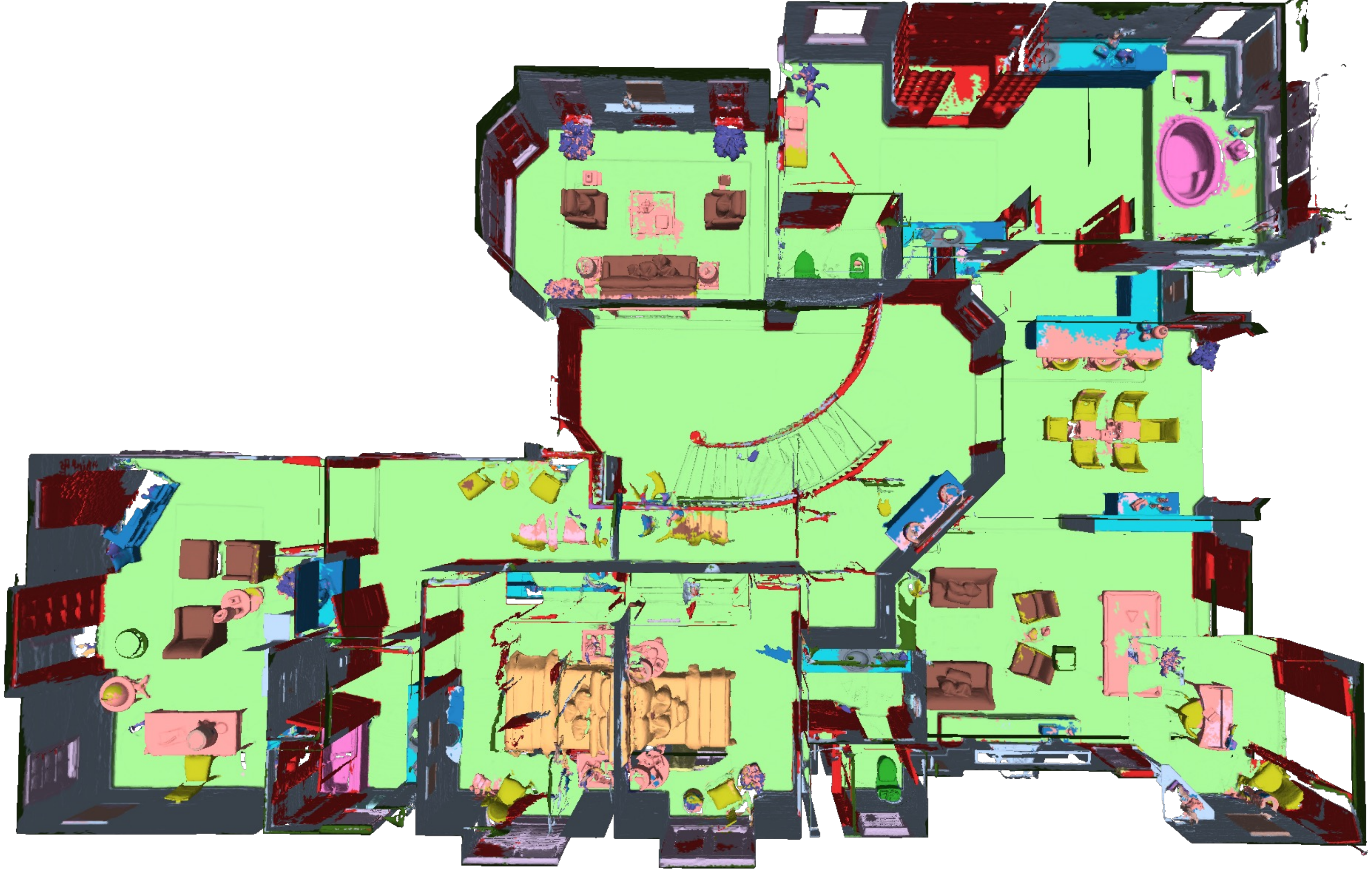
(visualize with PCA)

Choose the feature with
the highest max score among all prompts

Open-Vocabulary, Zero-shot

3D Semantic Segmentation

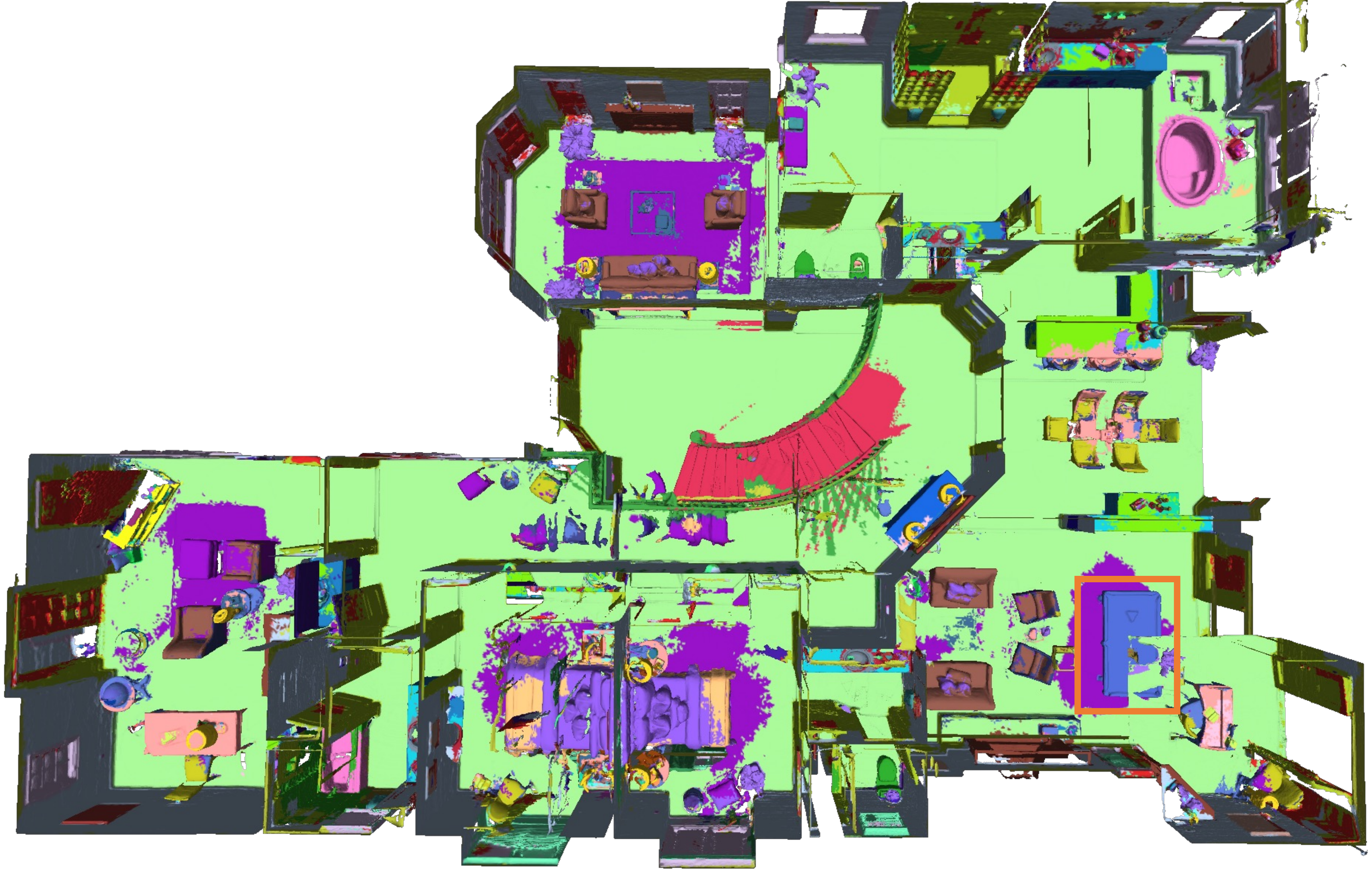
Input 3D Geometry



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

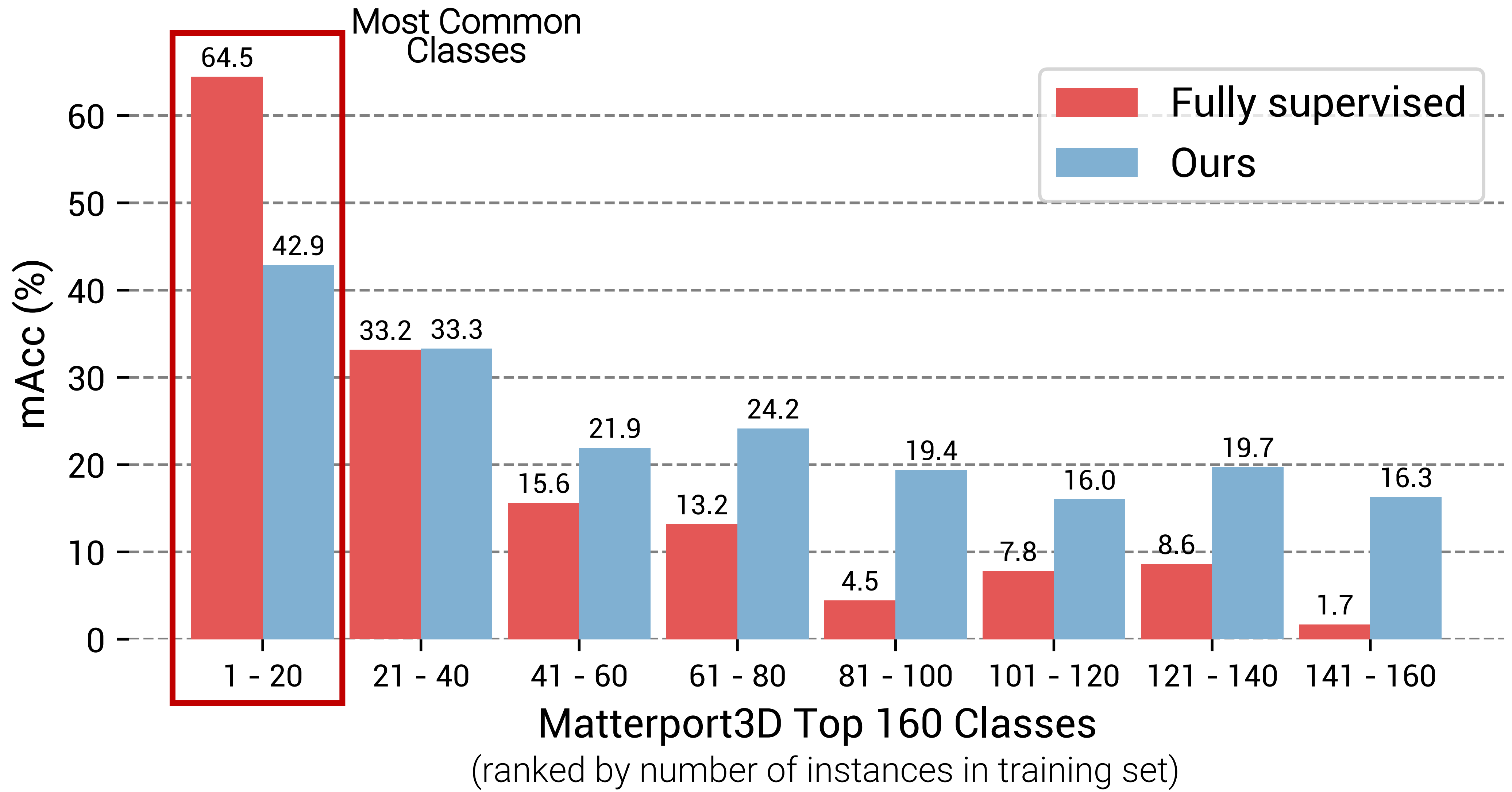


Our Zero-shot 3D Segmentation

(160 classes)

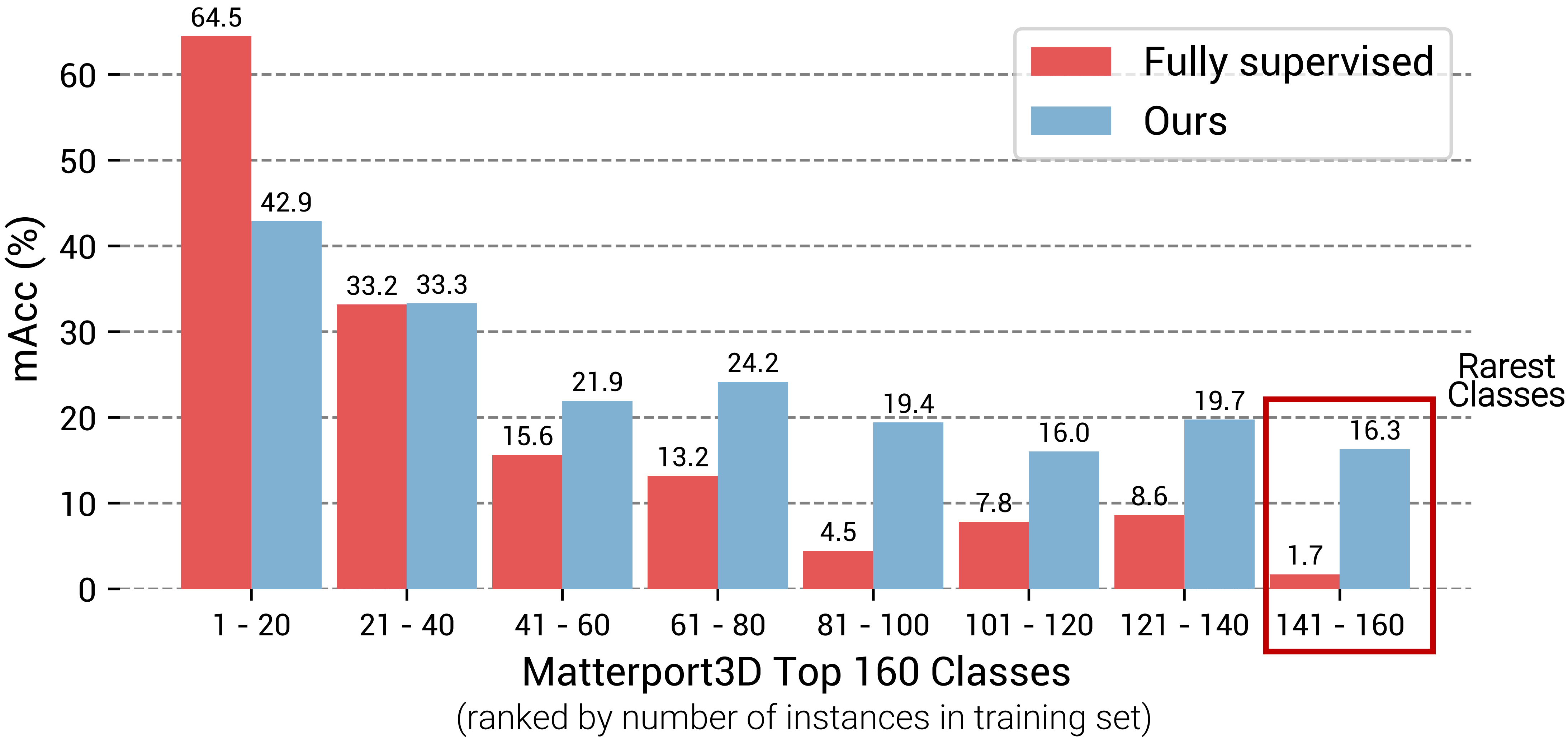
- wall
- door
- ceiling
- floor
- picture
- window
- chair
- pillow
- cabinet
- curtain
- table
- plant
- mirror
- towel
- sink
- shelves
- bed
- night stand
- toilet
- column
- banister
- stairs
- stool
- vase
- pot
- desk
- box
- coffee table
- counter
- bench
- garbage bin
- fireplace
- bathtub
- 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
Rarest Classes
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)

Ablation

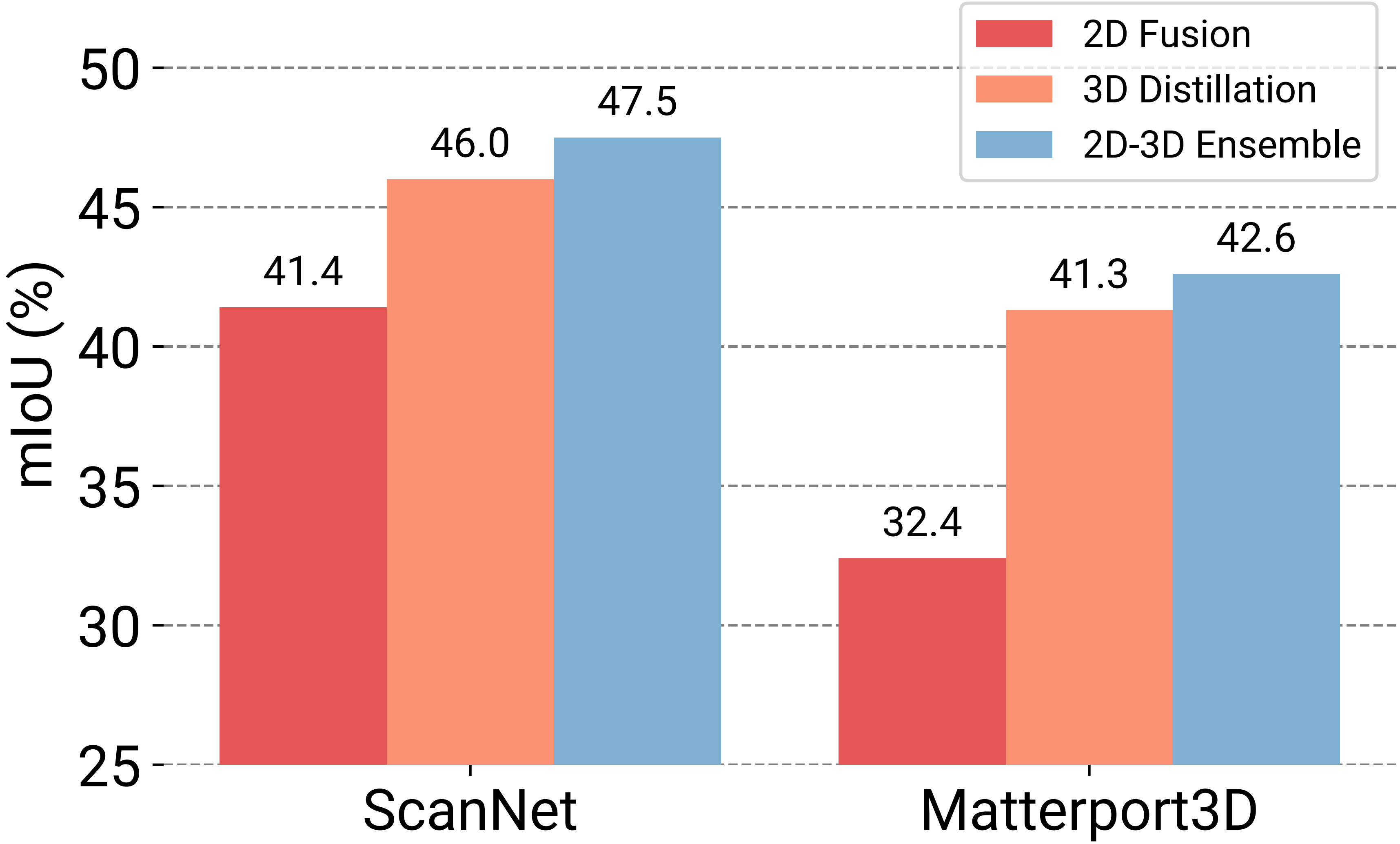

mIoU (%)
50
45
40
35
30
25
2D Fusion
3D Distillation
2D-3D Ensemble
41.4
46.0
47.5
32.4
41.3
42.6
ScanNet
Matterport3D

Image-based 3D Scene Query

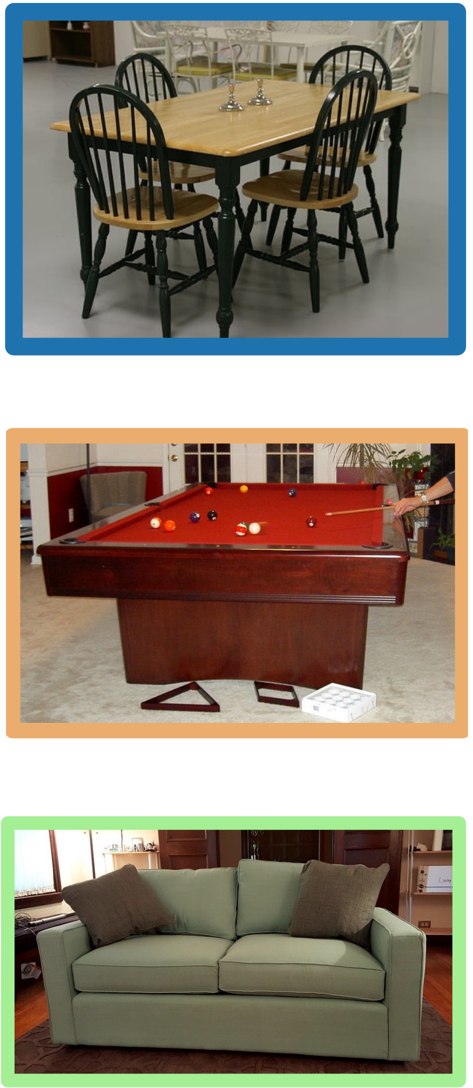
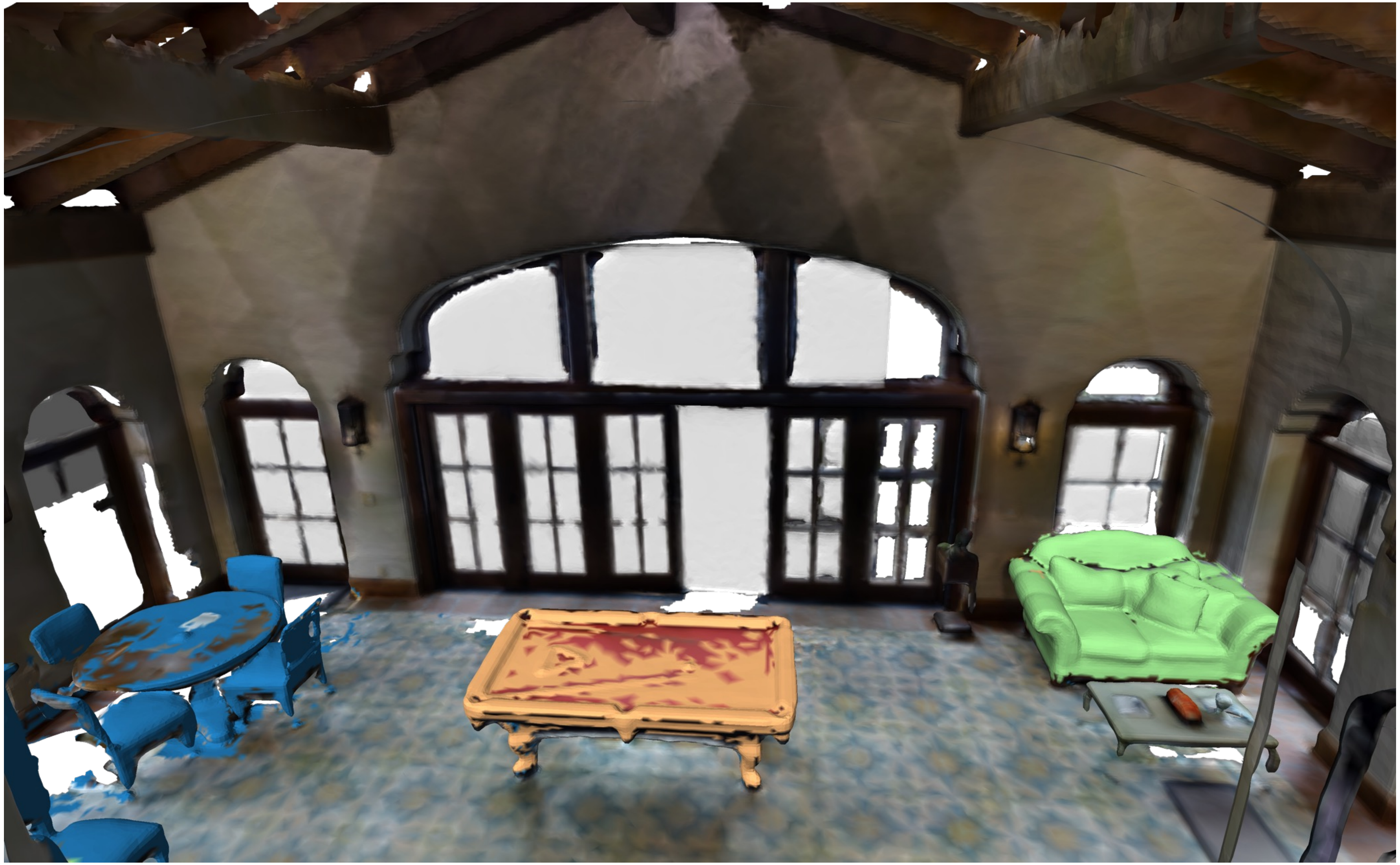


Image Queries

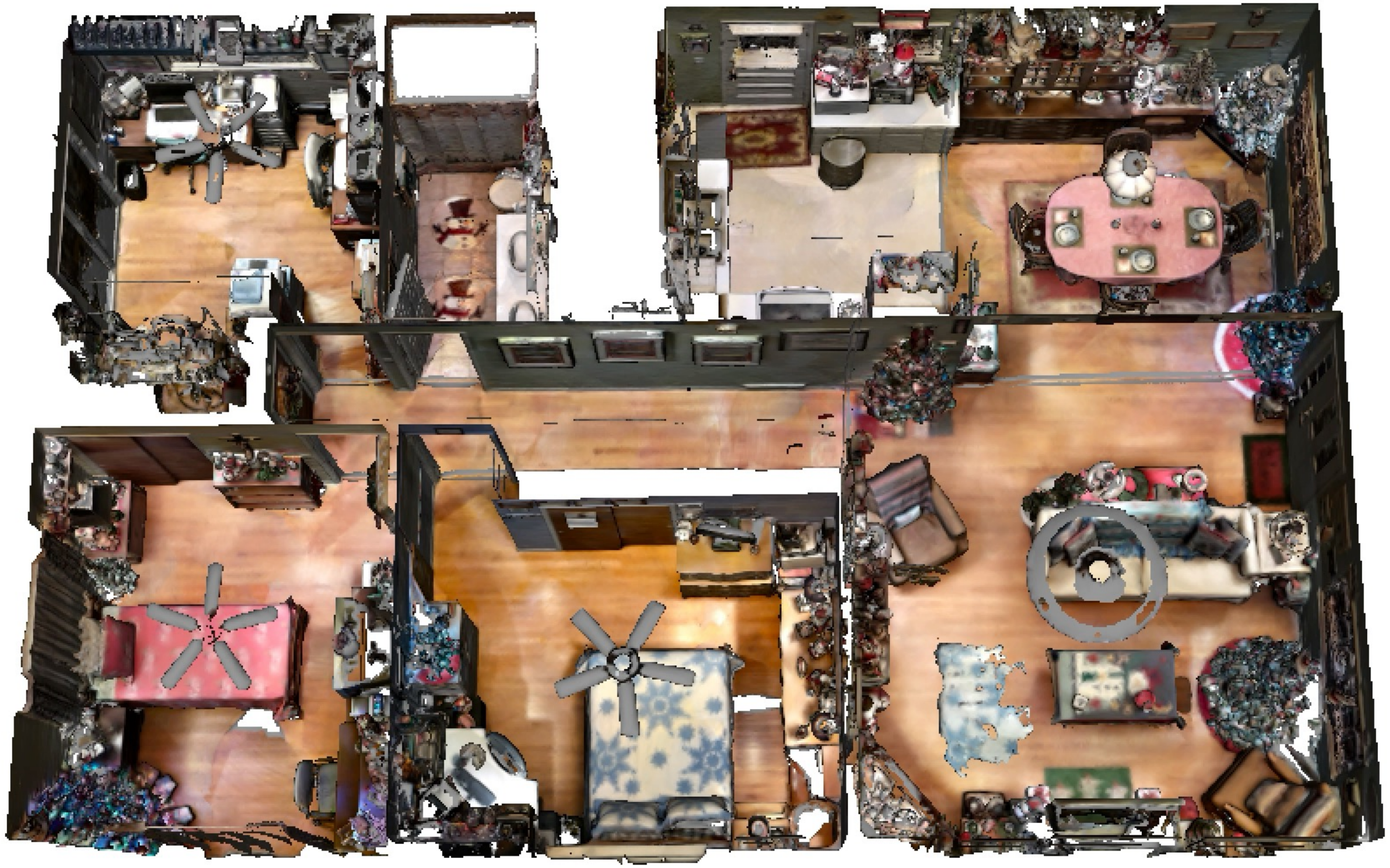


Given 3D Geometry

Interactive Demo

Open-vocabulary 3D Scene Exploration

Text queries: □



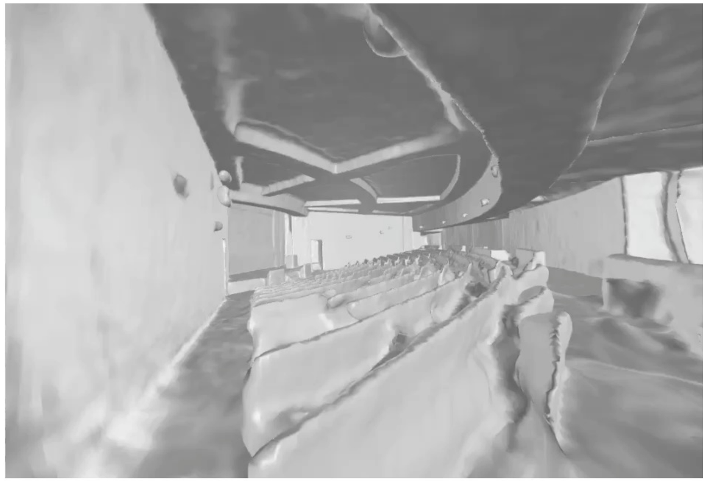
Take-home Message

- We enable a **wide range of applications** by open-vocabulary queries

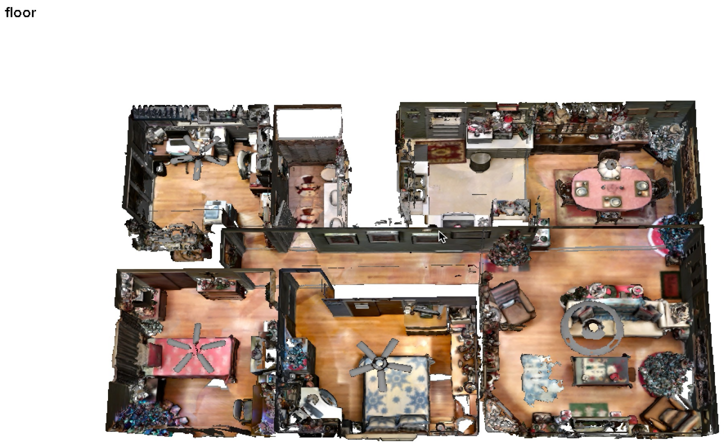
- This can hopefully influence how people train 3D scene understanding systems in the future

- Our real-time demo already shows the **possibility to directly apply to AR/VR**

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



Ours



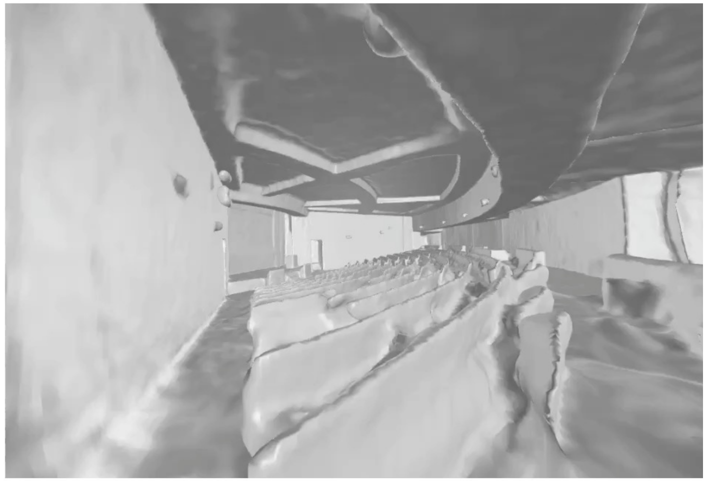
MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022

OpenScene
CVPR 2023

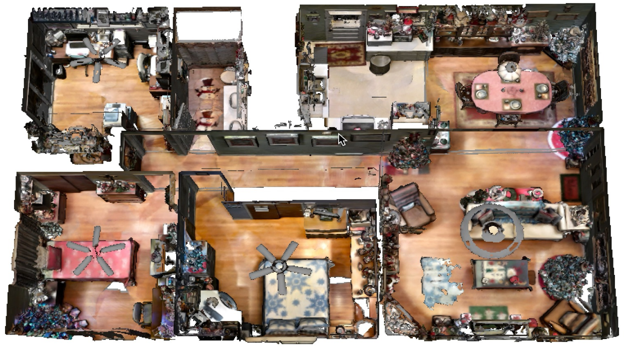
Learning Neural Scene Representations for 3D Reconstruction and Understanding

Songyou Peng



Ours

floor



MonoSDF
NeurIPS 2022
niujinshuchong.github.io/monosdf/

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

OpenScene
CVPR 2023
pengsongyou.github.io/openscene

Thank you!