How do NeRF and CLIP advance 3D Scene Reconstruction and Understanding?

Songyou Peng

ETH Zurich and Max Planck Institute for Intelligent Systems

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Chinese University of Hong Kong (CUHK), Shenzhen

March 07, 2023

Who Am I?

- 4th Year PhD Student
  - Marc Pollefeys
  - Andreas Geiger
- Internships during PhD
  - 2021: Michael Zollhoefer
  - 2022: Tom Funkhouser
- Open to chat!

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

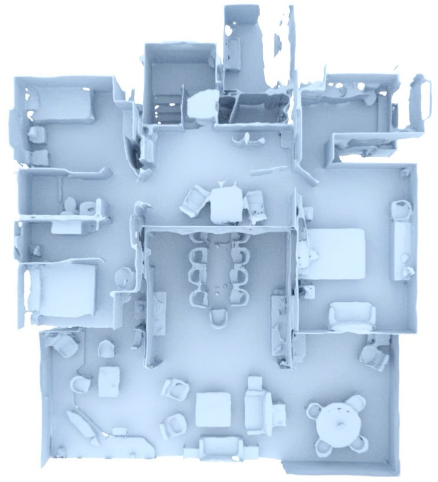
Meta

Google Research

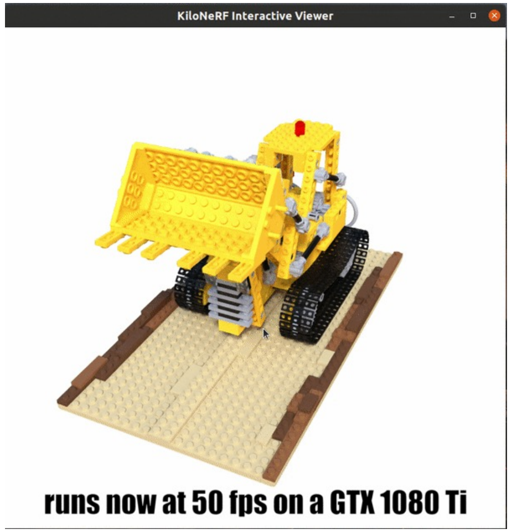
pengsongyou.github.io

My PhD Topics: Neural Scene Representations

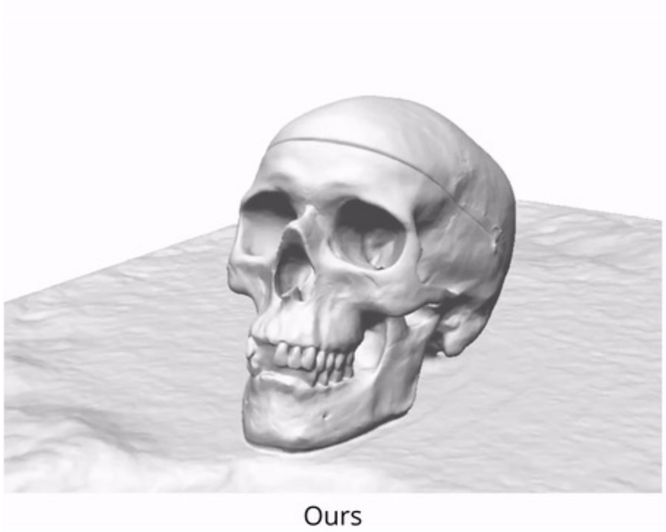
for 3D reconstruction, novel view synthesis, and SLAM



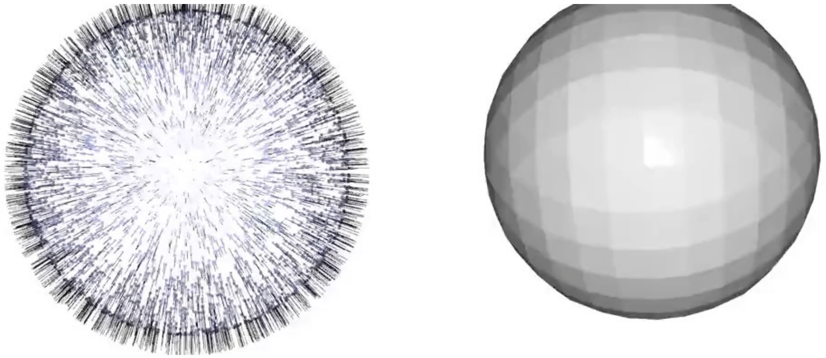
Convolutional Occupancy Networks
ECCV 2020 (Spotlight)



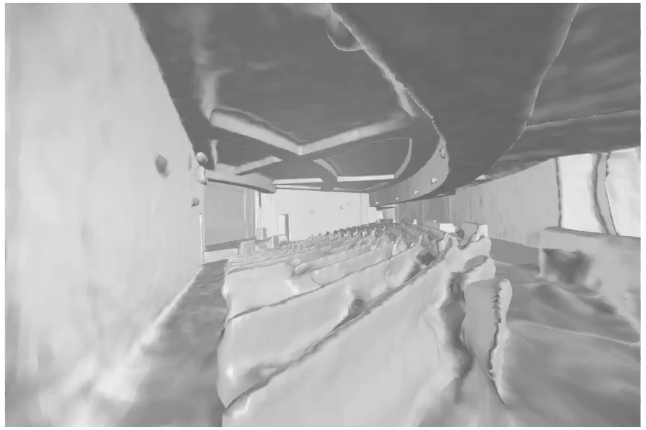
KiloNeRF
ICCV 2021



UNISURF
ICCV 2021 (Oral)



Shape As Points
NeurIPS 2021 (Oral)



Ours

MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022

How do NeRF and CLIP advance 3D Scene Reconstruction and Understanding?

How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?

How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?

NeRF is awesome!

Some problems still exist...

😢 Poor underlying geometry

😢 Camera poses needed

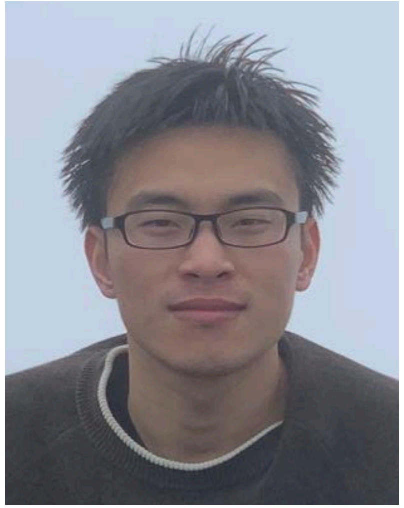
😊 MonoSDF

😊 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

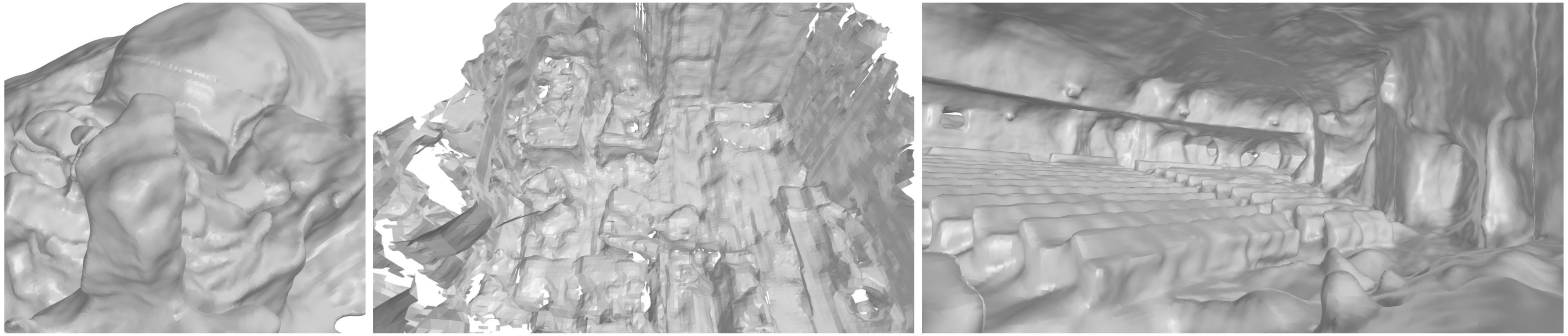
Neural Implicit Surfaces with Volume Rendering

[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

DTU (3 views) ScanNet (464 views) Tanks & Temples (298 views)

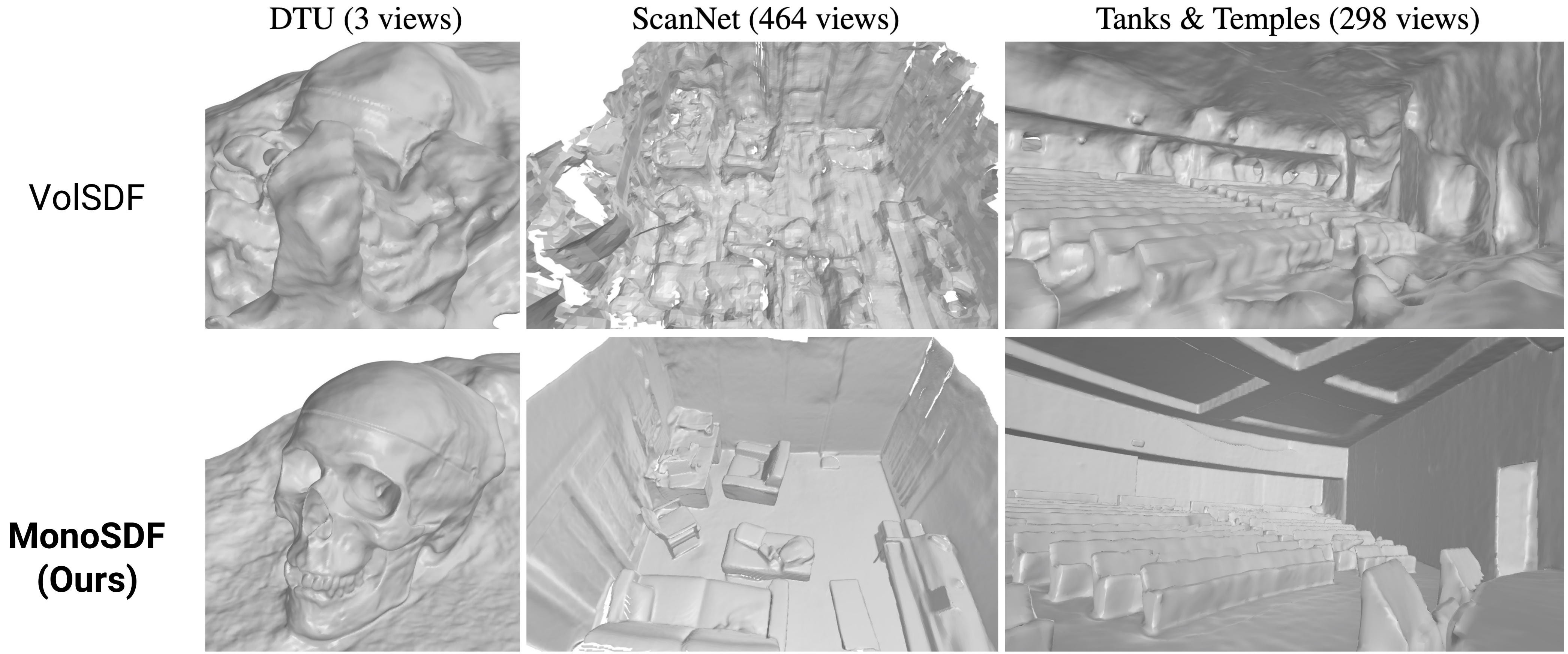
VolSDF



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

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

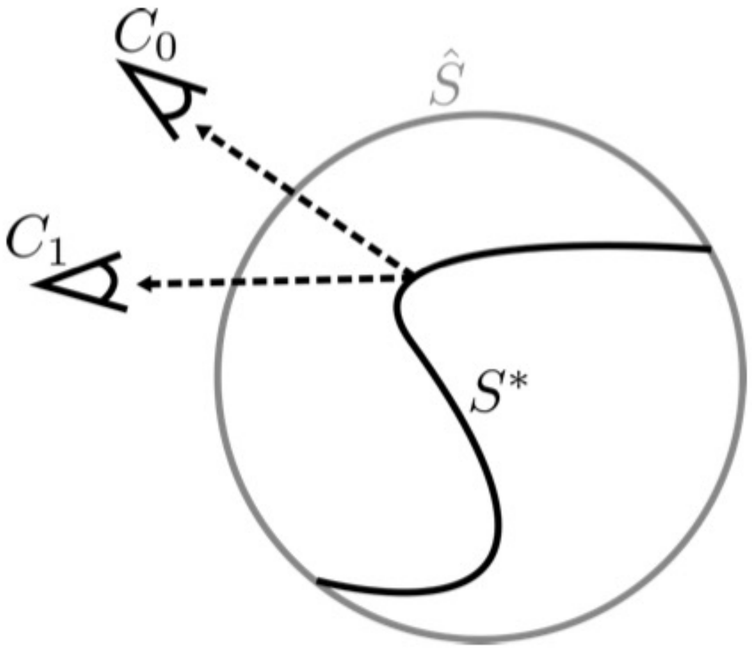
Neural Implicit Surfaces with Volume Rendering



- ✚ 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

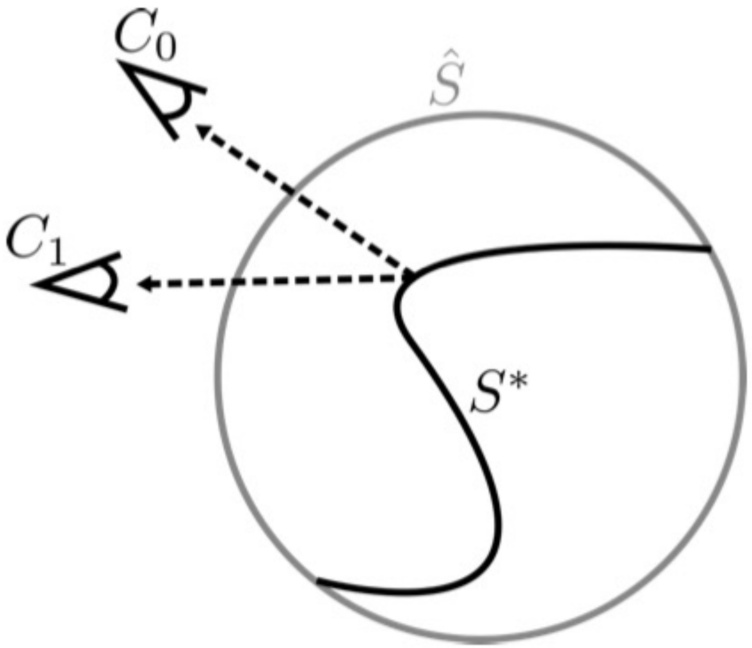
Shape-Appearance Ambiguity



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

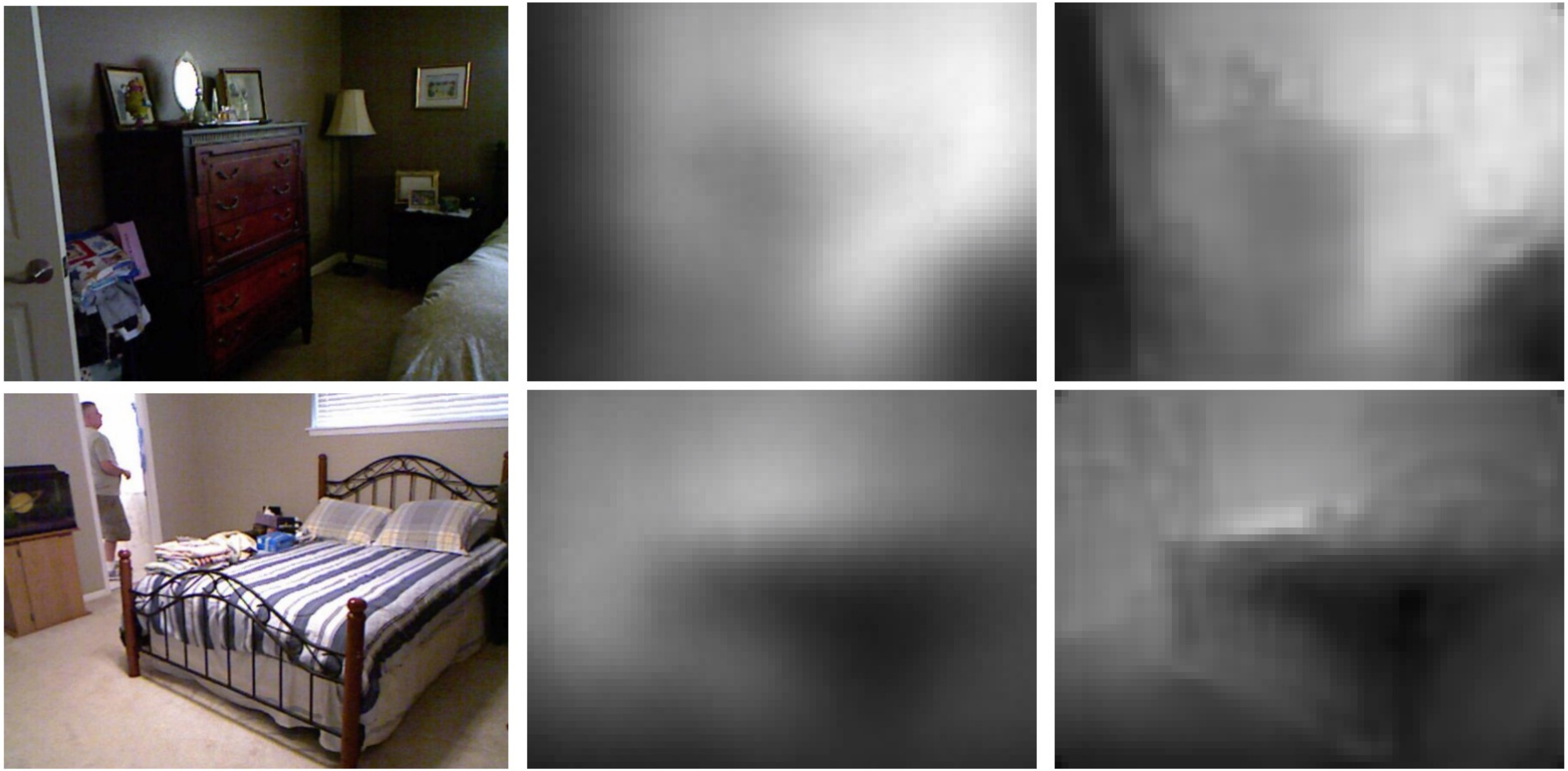


There exists an infinite number of photo-consistent explanations for input images!

⟹ Exploit monocular geometric priors

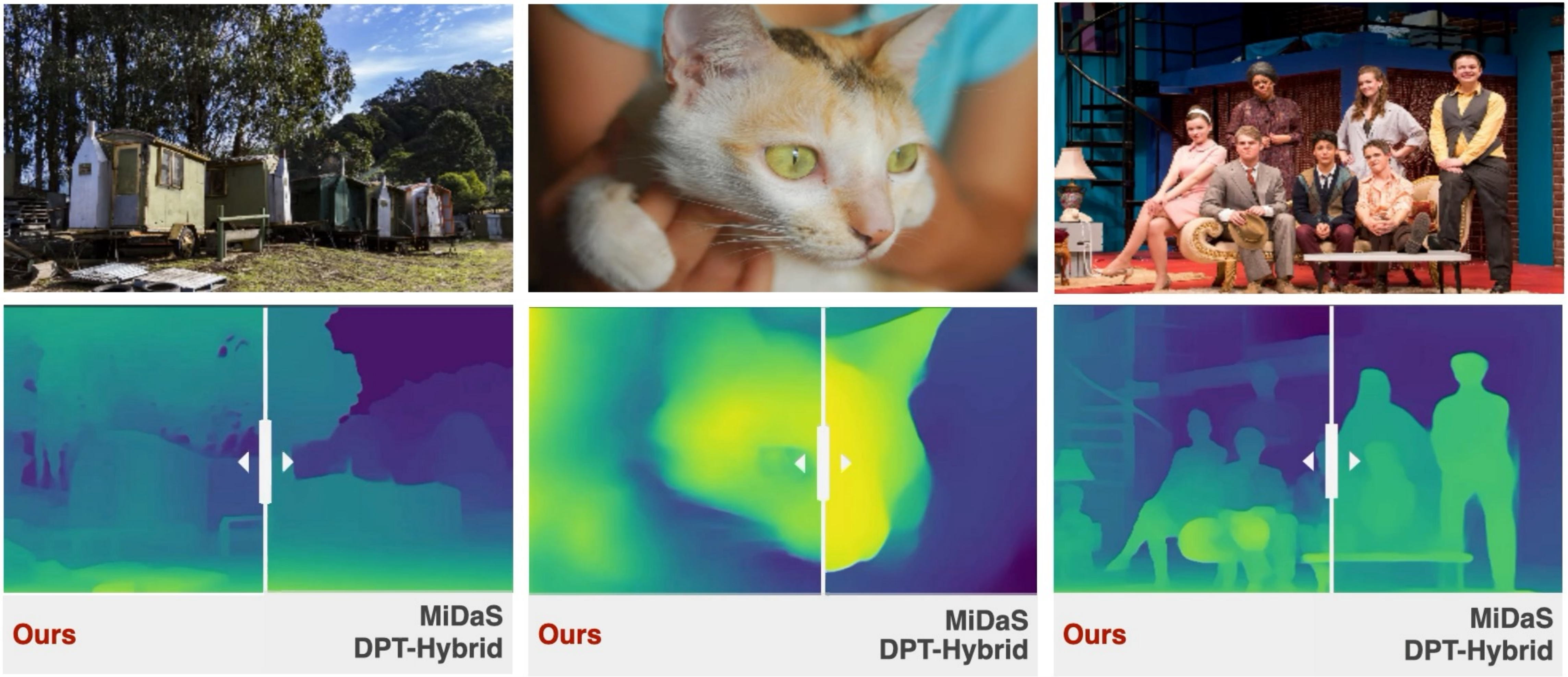
Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020

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

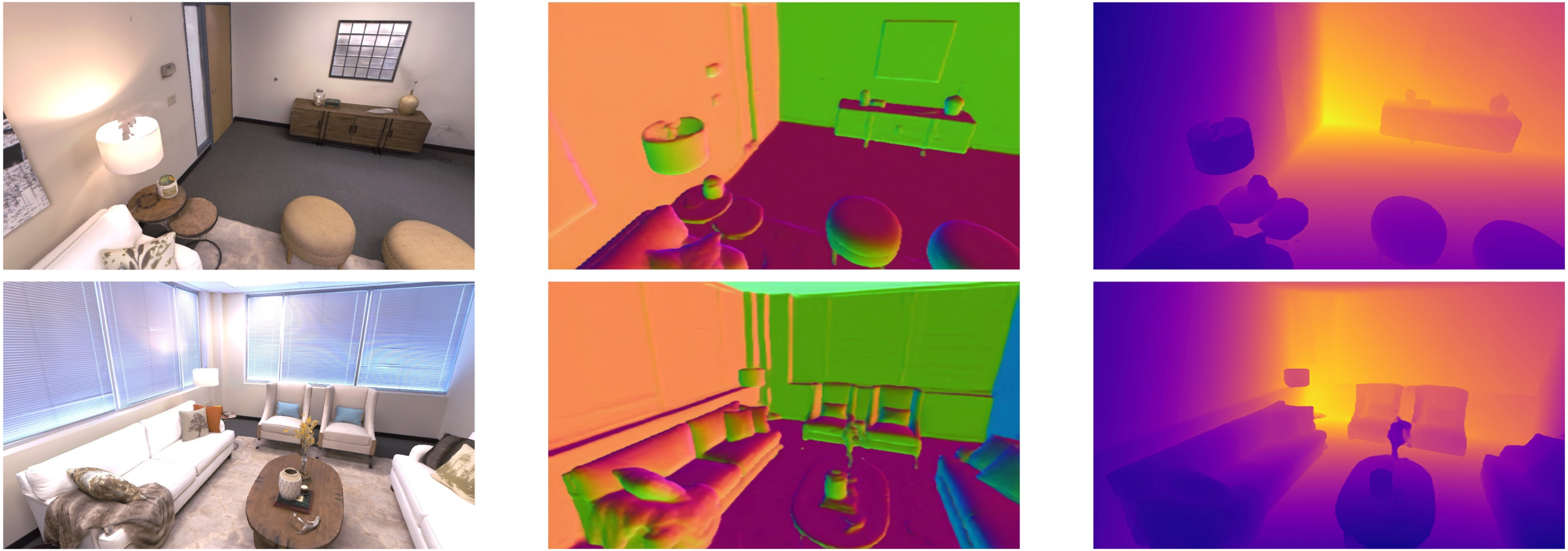
Omnidata



[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

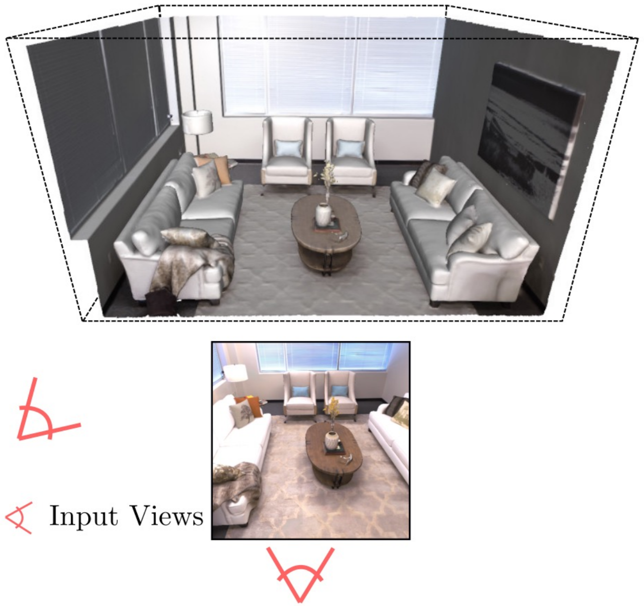


RGB Image | Omnidata Normal | Omnidata Depth

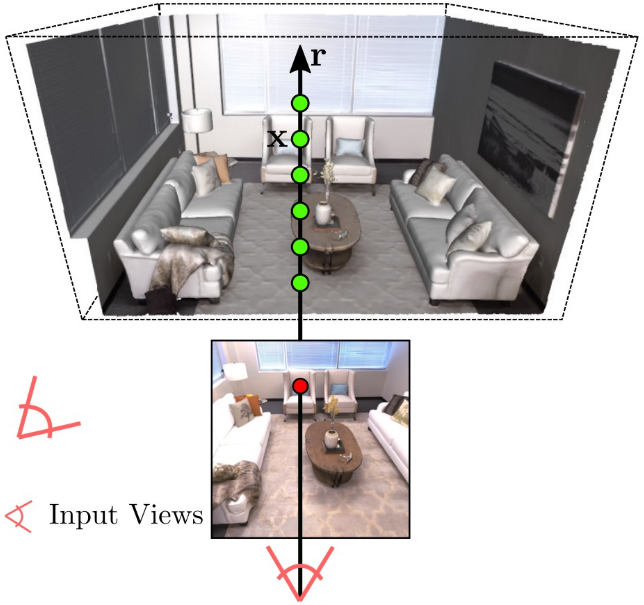
Eftekhar, Sax, Malik and Zamir: Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. ICCV, 2021.

MonoSDF

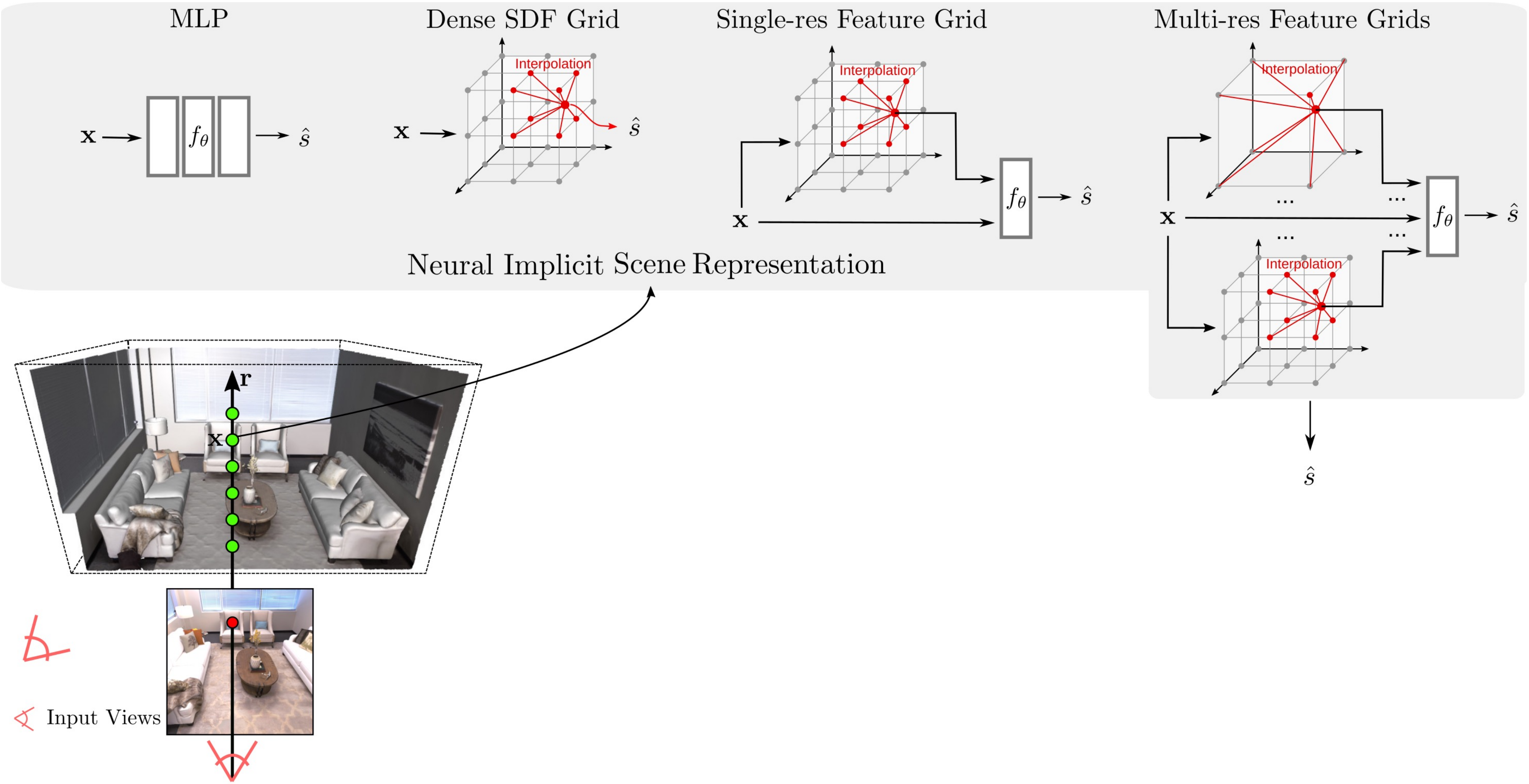
MonoSDF



MonoSDF

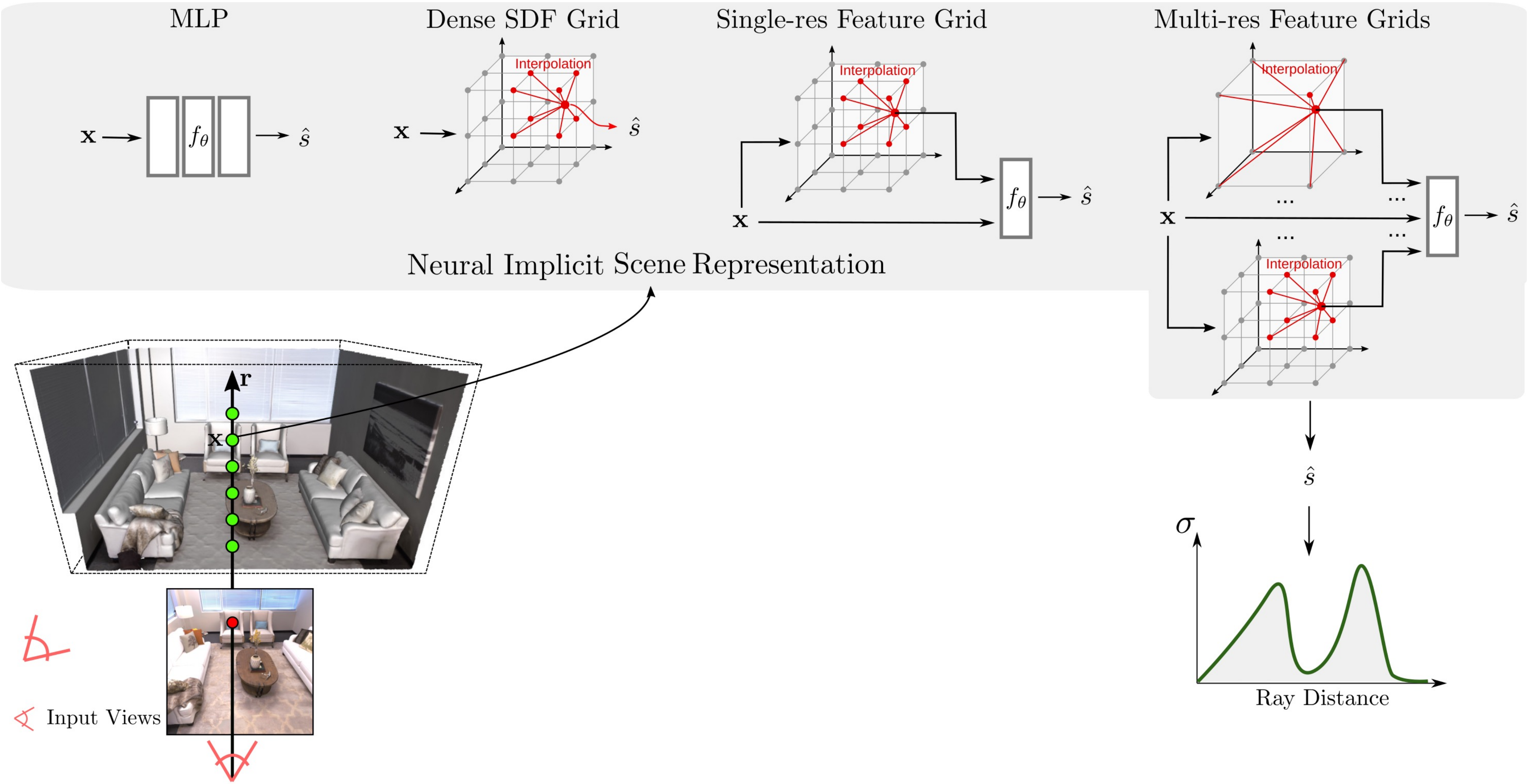


MonoSDF



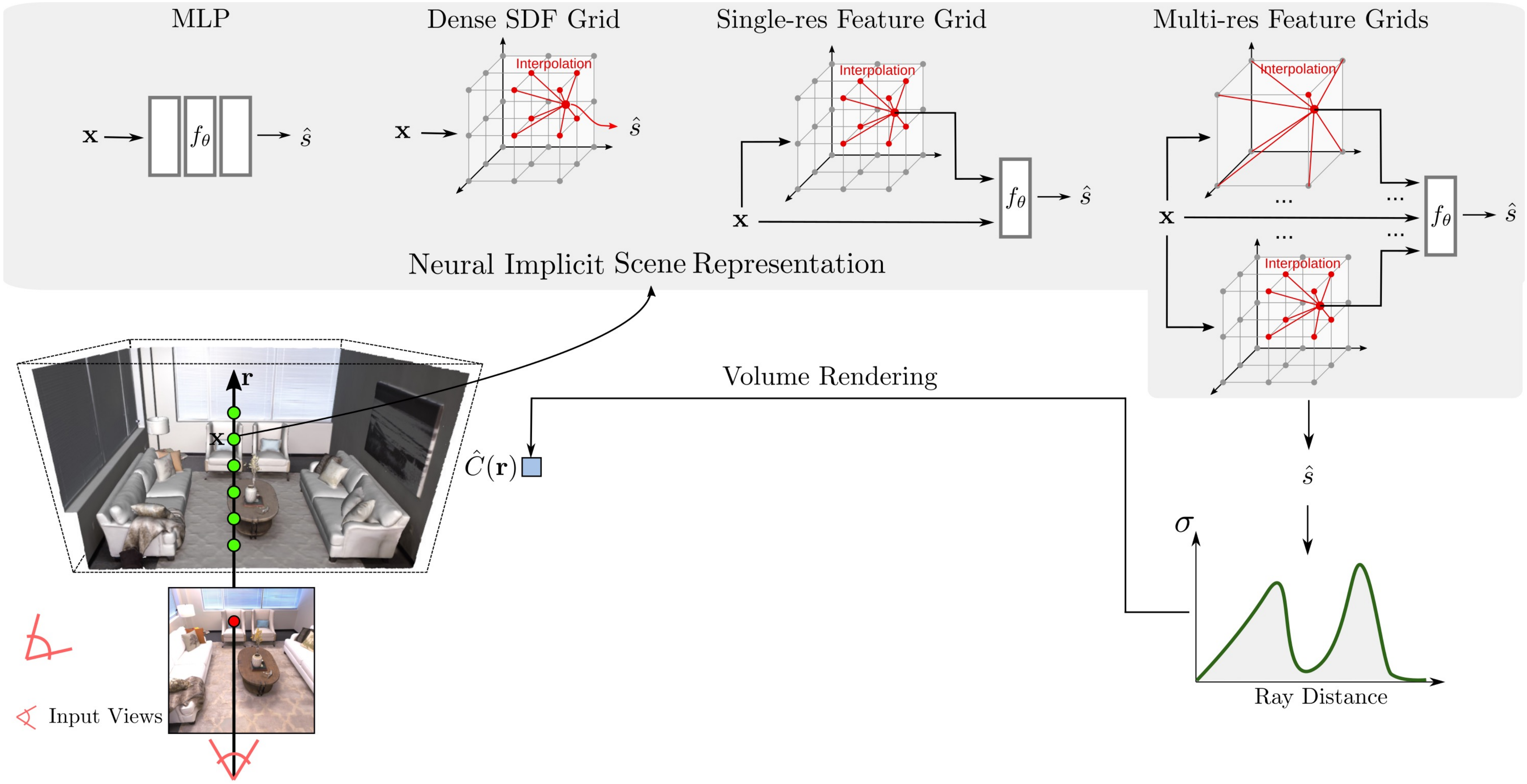
MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
Interpolation
Interpolation
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
$\mathbf{r}$
Input Views

MonoSDF



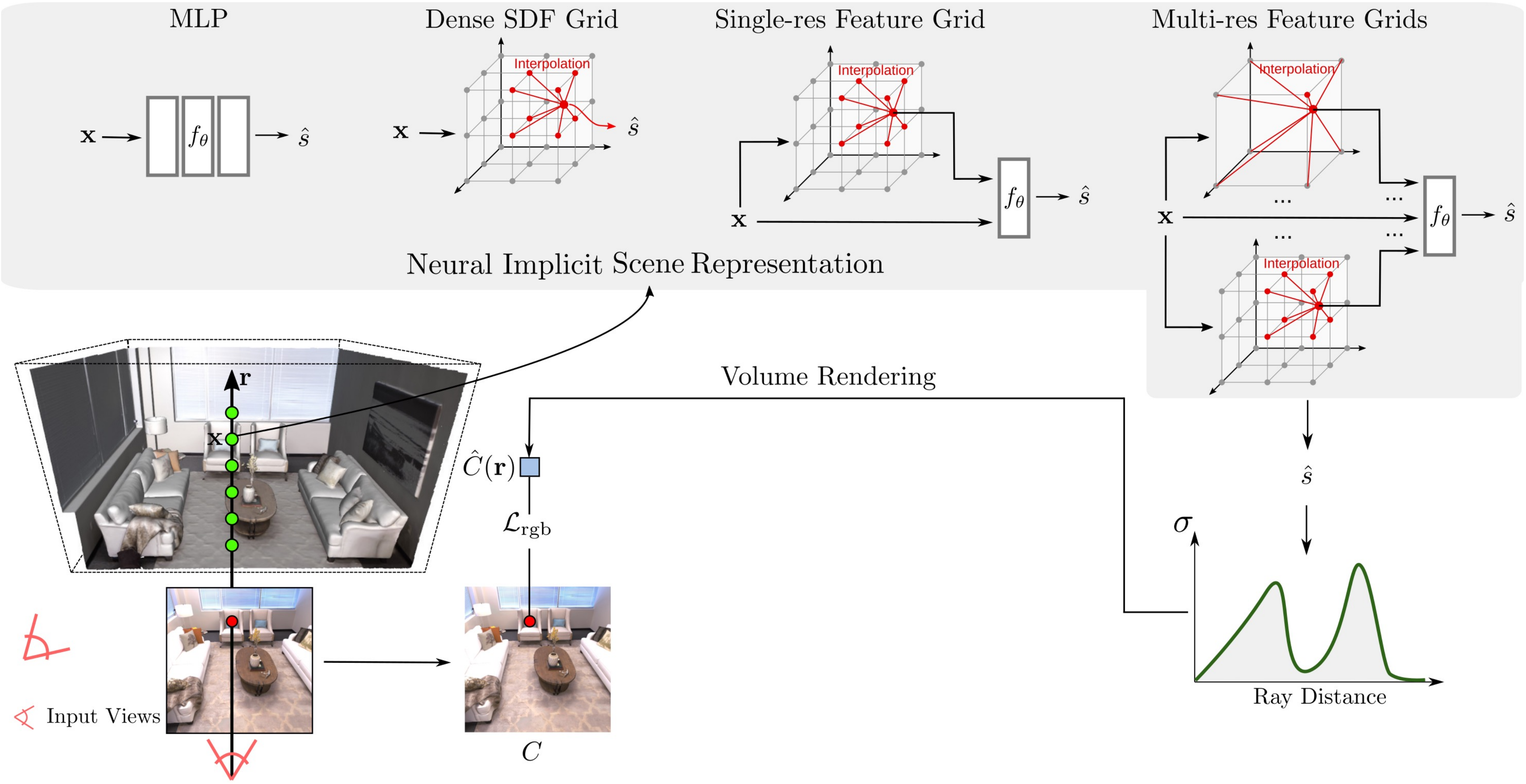
MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
x
f_θ
$\hat{s}$
Neural Implicit Scene Representation
r
Input Views
σ
Ray Distance

MonoSDF



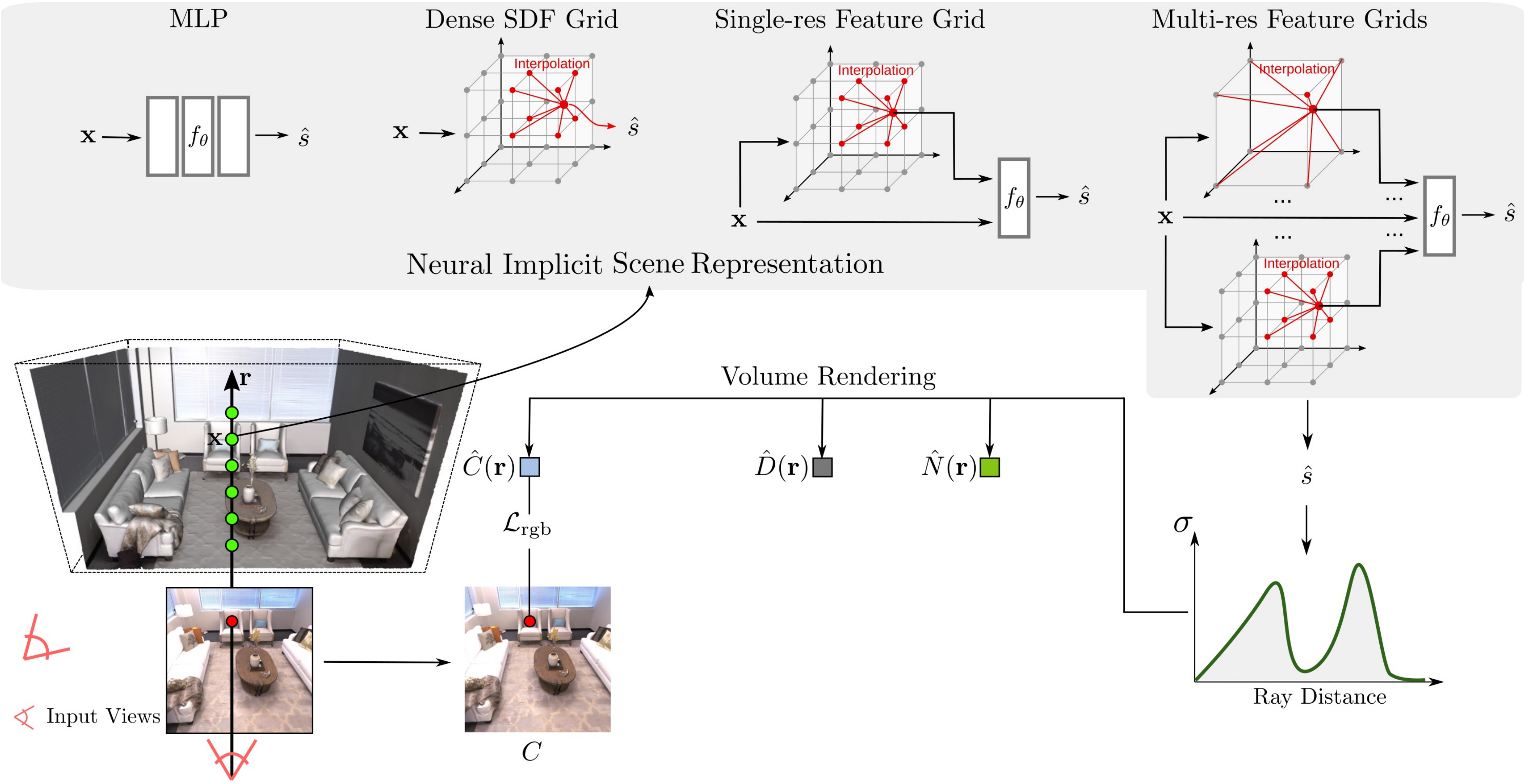
MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
Interpolation
Interpolation
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
Volume Rendering
$\hat{C}(\mathbf{r})$
$\mathbf{r}$
σ
Ray Distance
Input Views

MonoSDF

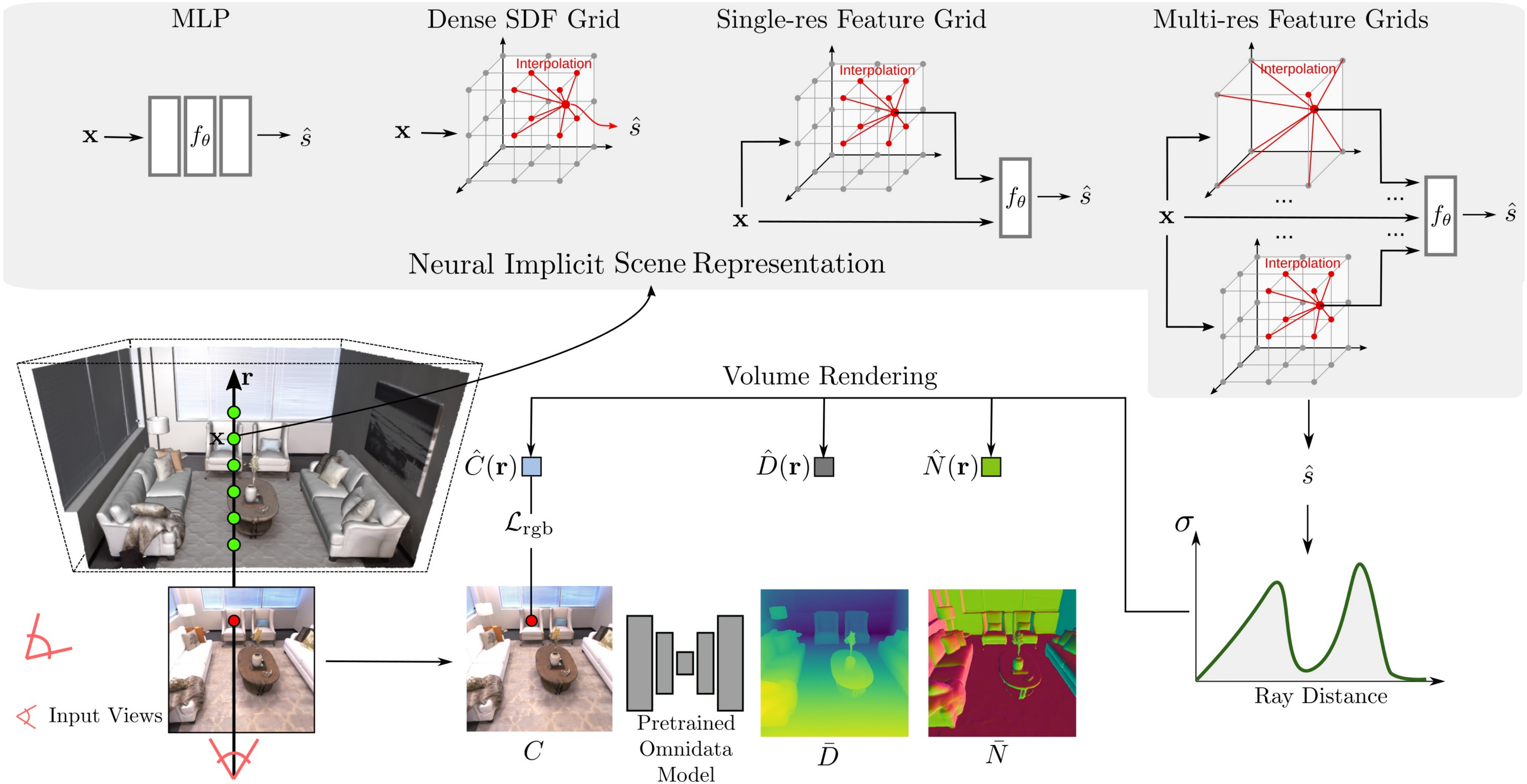


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
Volume Rendering
$\mathbf{r}$
$\hat{C}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
Input Views
σ
Ray Distance

MonoSDF

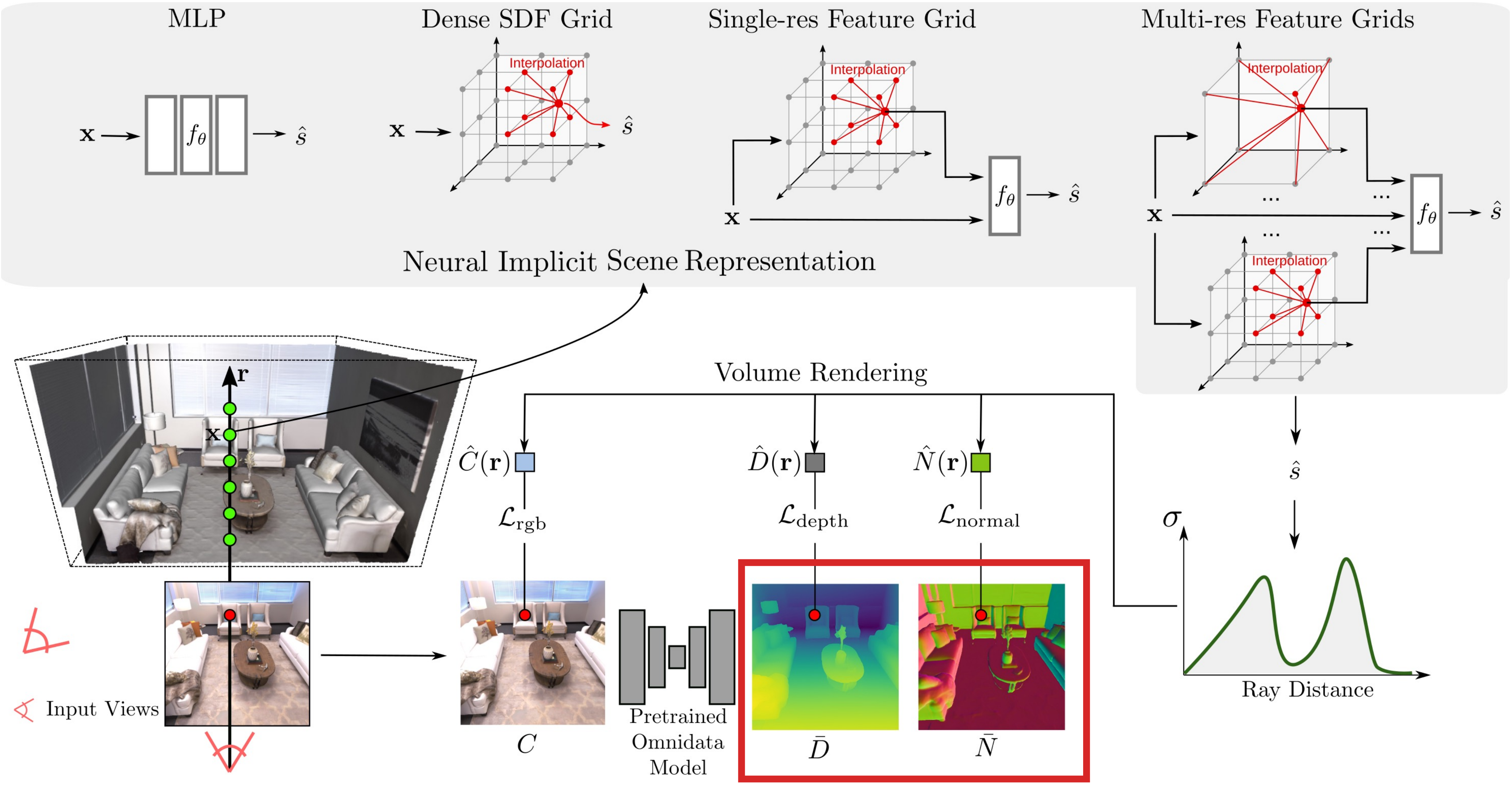

MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
$\mathbf{x}$
f_θ
$\hat{s}$
Interpolation
Neural Implicit Scene Representation
Volume Rendering
$\hat{C}(\mathbf{r})$
$\hat{D}(\mathbf{r})$
$\hat{N}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
$\mathbf{r}$
Input Views
σ
Ray Distance

MonoSDF



MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
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
$\bar{D}$
$\bar{N}$

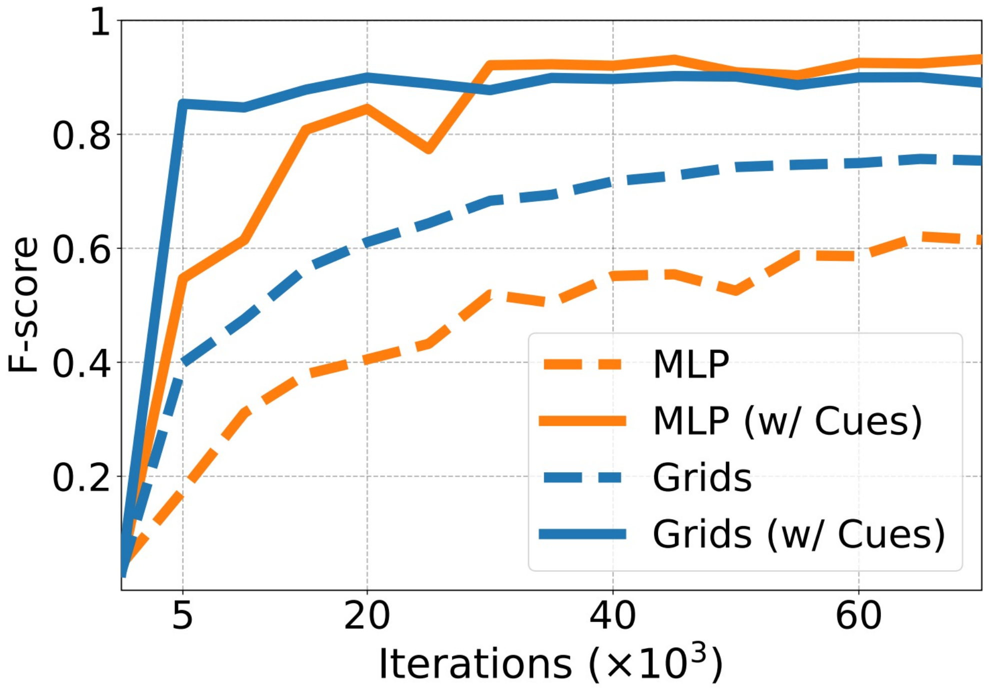
MonoSDF



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**



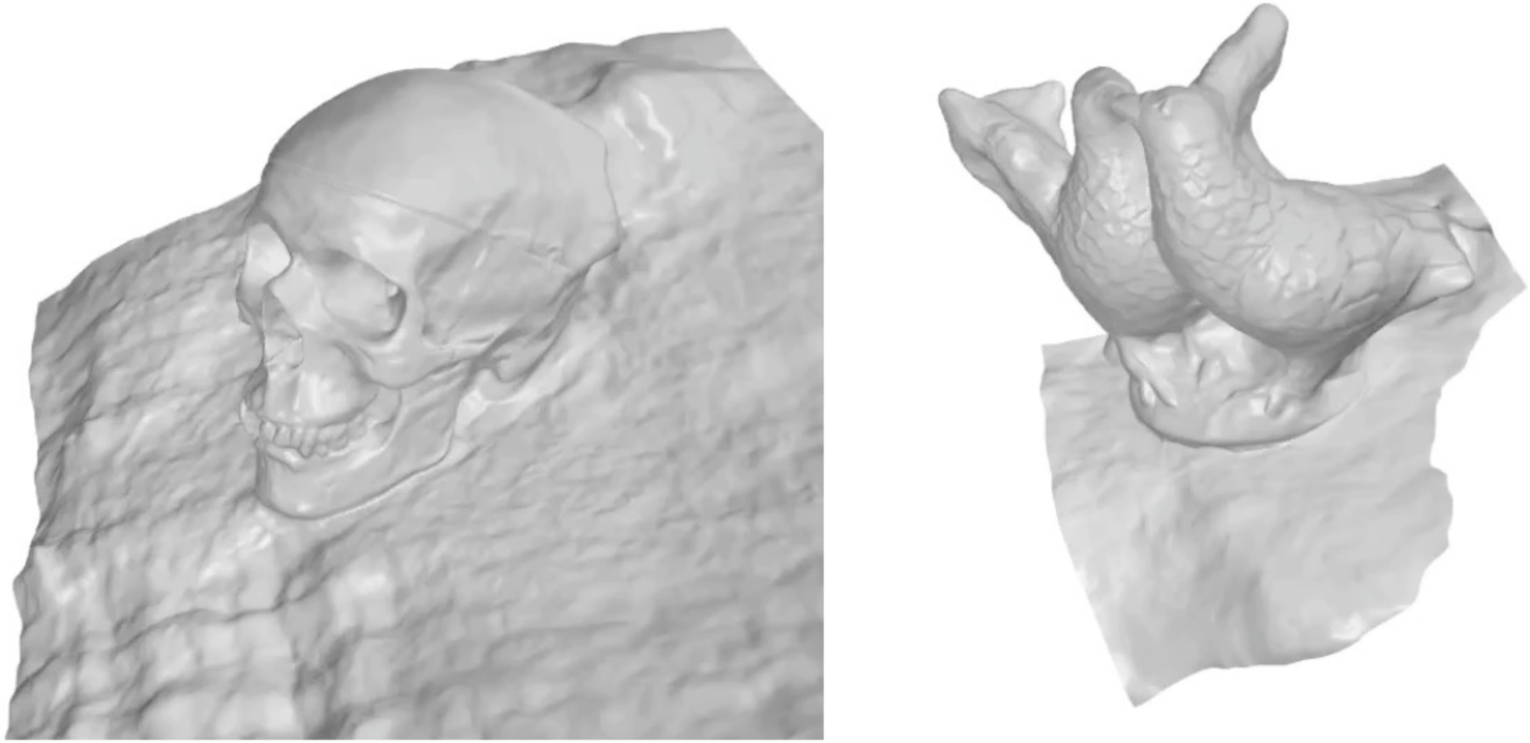
- Monocular cues improve reconstruction results significantly
- Combining **depth & normal** leads to best performance
- Monocular cues can improve **convergence speed**

Baseline Comparisons on ScanNet

Ours

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

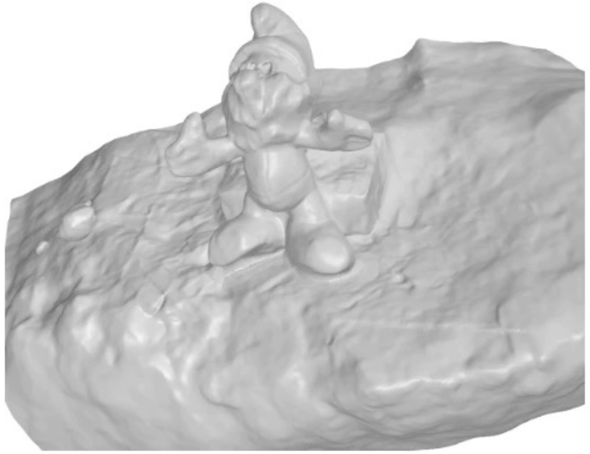
Baseline Comparisons on DTU (3-views)



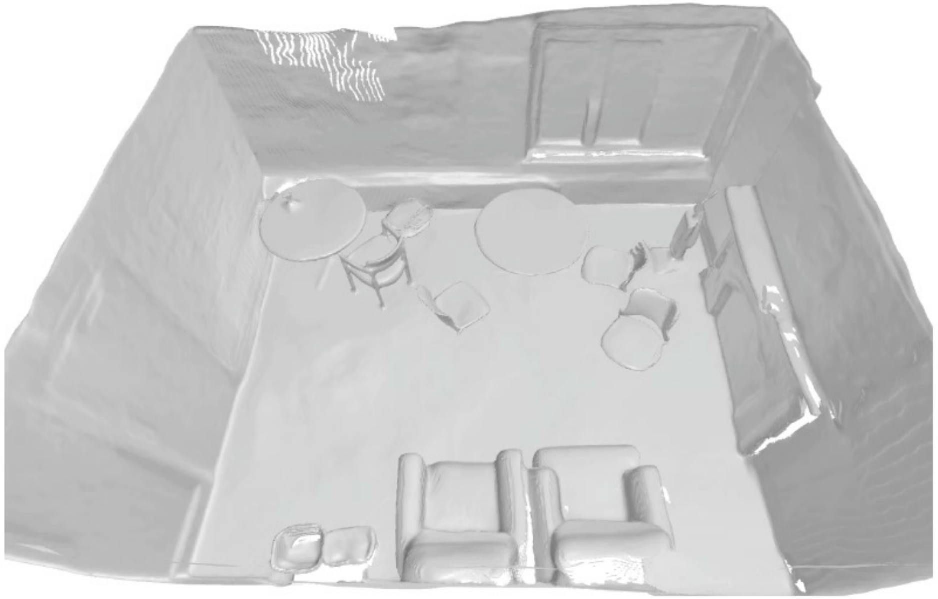
Ours

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 Haiwen & Dan's ICLR 2023 paper GOOD ☺
- Limitation: Still require camera poses given :(

NICE-SLAM

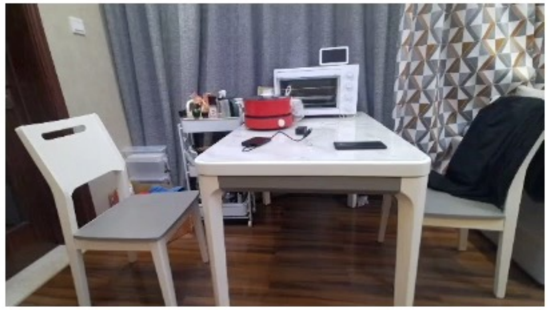
Neural Implicit Scalable Encoding for SLAM

CVPR 2022

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

* Equal Contributions

RGB-D Sequences



10x Speed

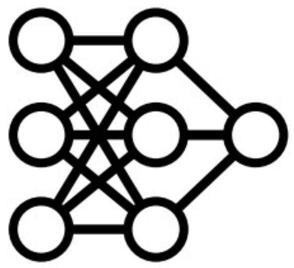
iMAP

[Sucar et al., ICCV'21]

First neural implicit-based online SLAM system

iMAP

[Sucar et al., ICCV'21]



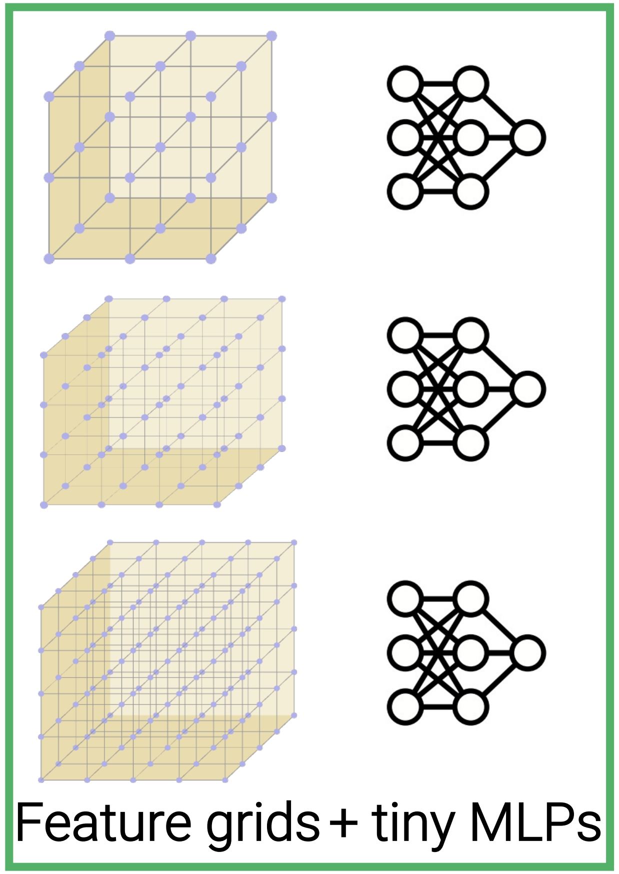
A single MLP

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

Predicted Poses

GT Poses

NICE-SLAM

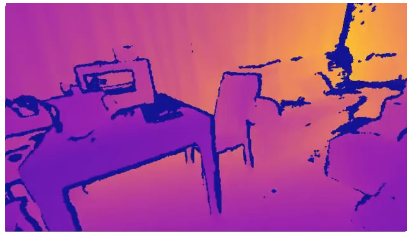


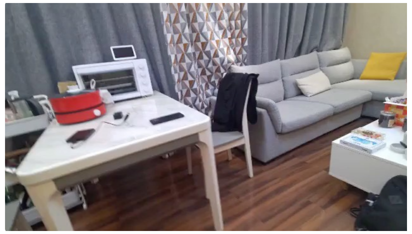
- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence

Predicted Poses
GT Poses

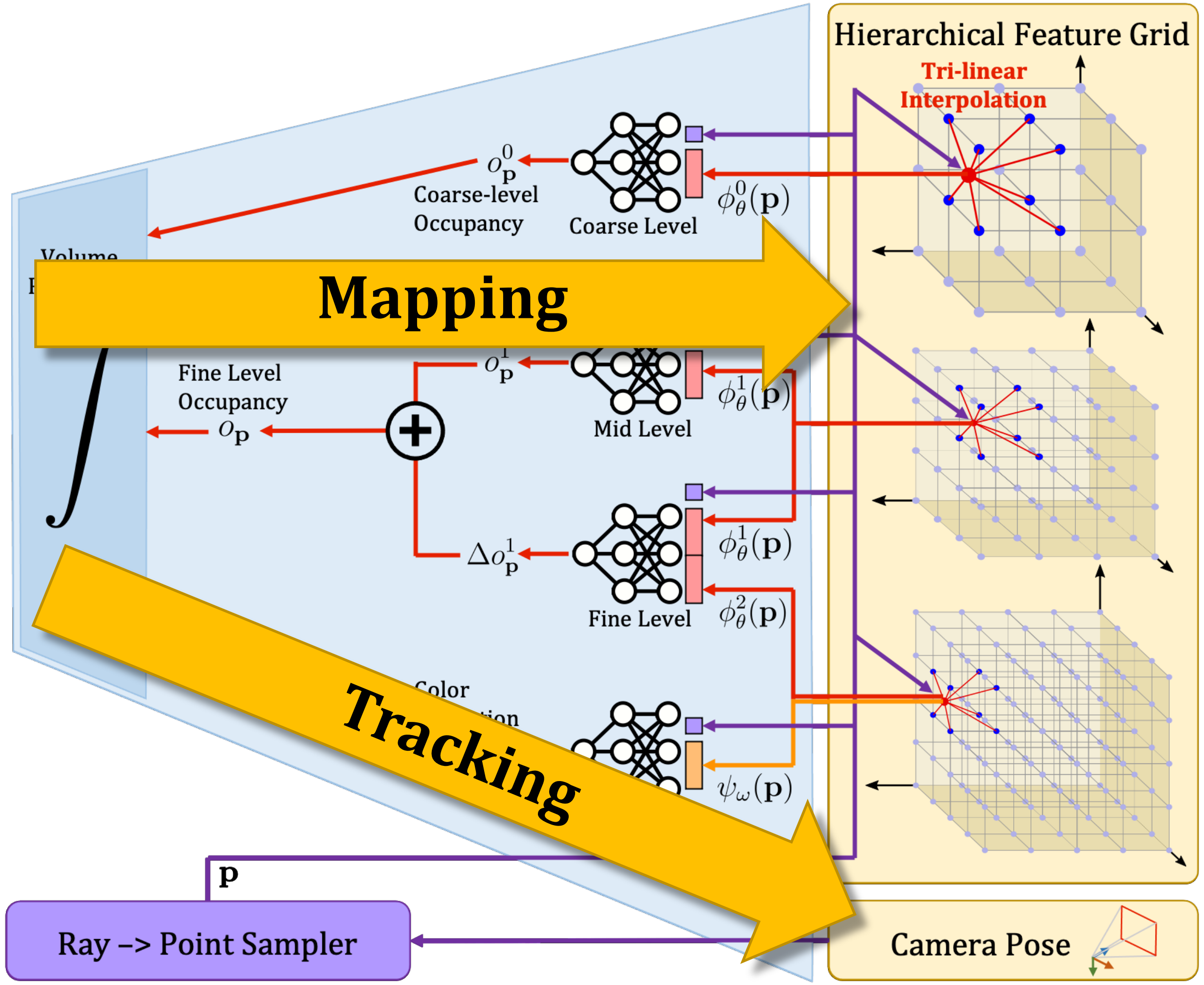
Pipeline

Input Depth





Input RGB



Results

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

4x Speed

Predicted Poses

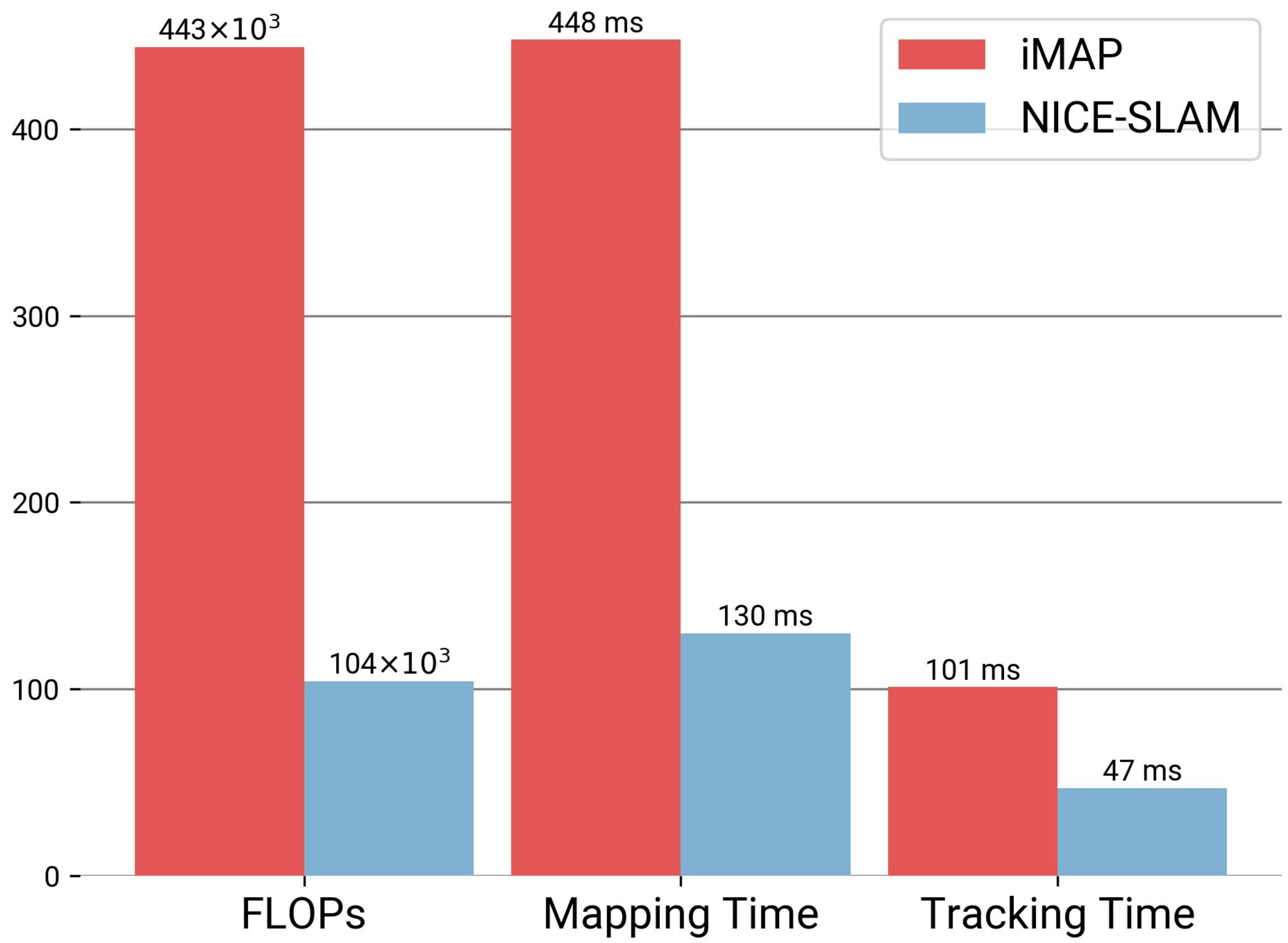
GT Poses

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

10x Speed



iMAP
NICE-SLAM
443×10^3
104×10^3
448 ms
130 ms
101 ms
47 ms
400
300
200
100
0
FLOPs
Mapping Time
Tracking Time

Take-home Message

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

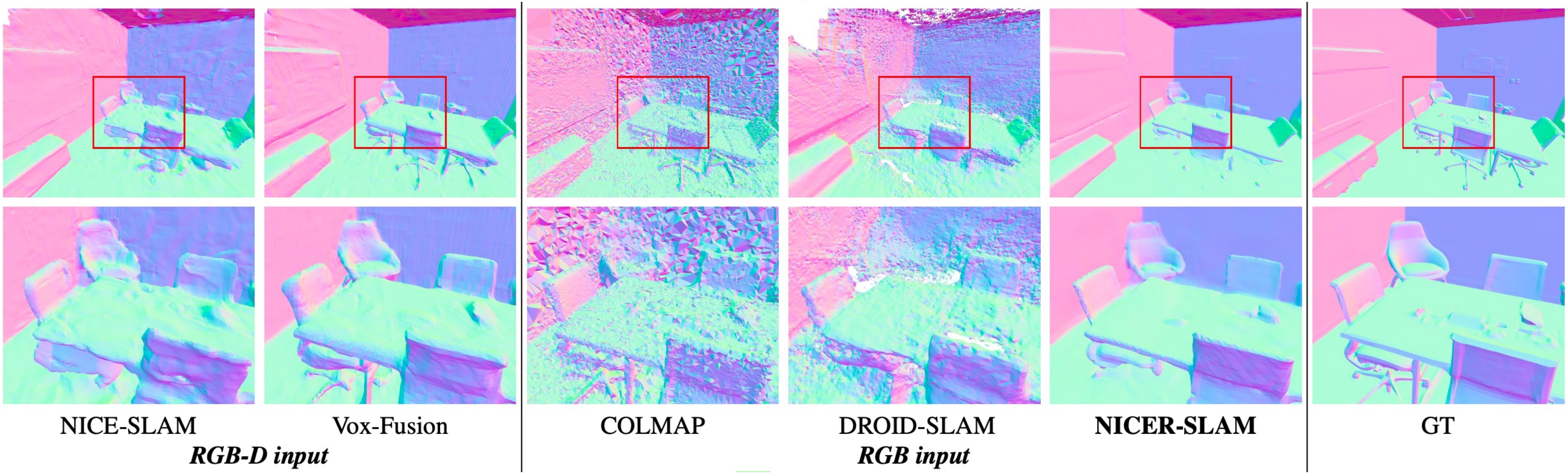
Limitations

- <u>Requires depths as input</u>
- 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

How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?

How does NeRF advance 3D Scene Reconstruction?

How does CLIP advance 3D Scene Understanding?

OpenScene

3D Scene Understanding with Open Vocabularies

CVPR 2023

Songyou Peng

Kyle Genova

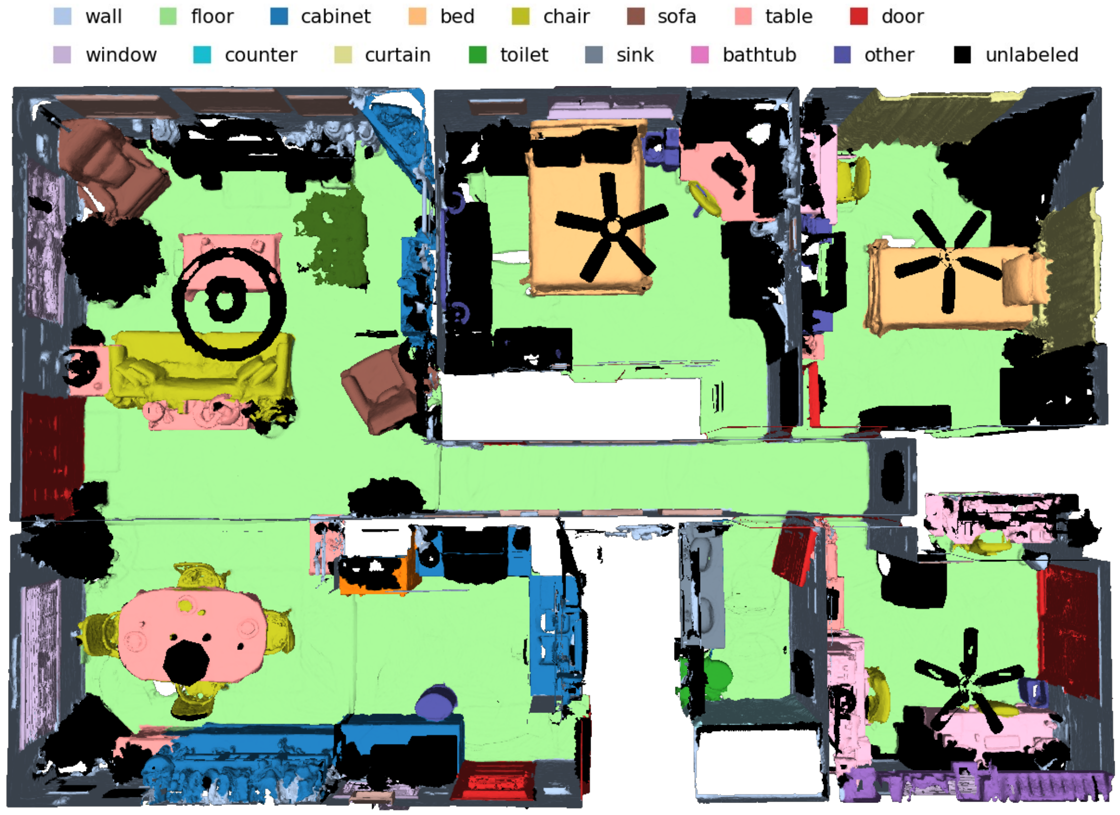
Chiyu “Max” Jiang

Andrea Tagliasacchi

Marc Pollefeys

Tom Funkhouser


Input 3D Geometry



Traditional Semantic Segmentation

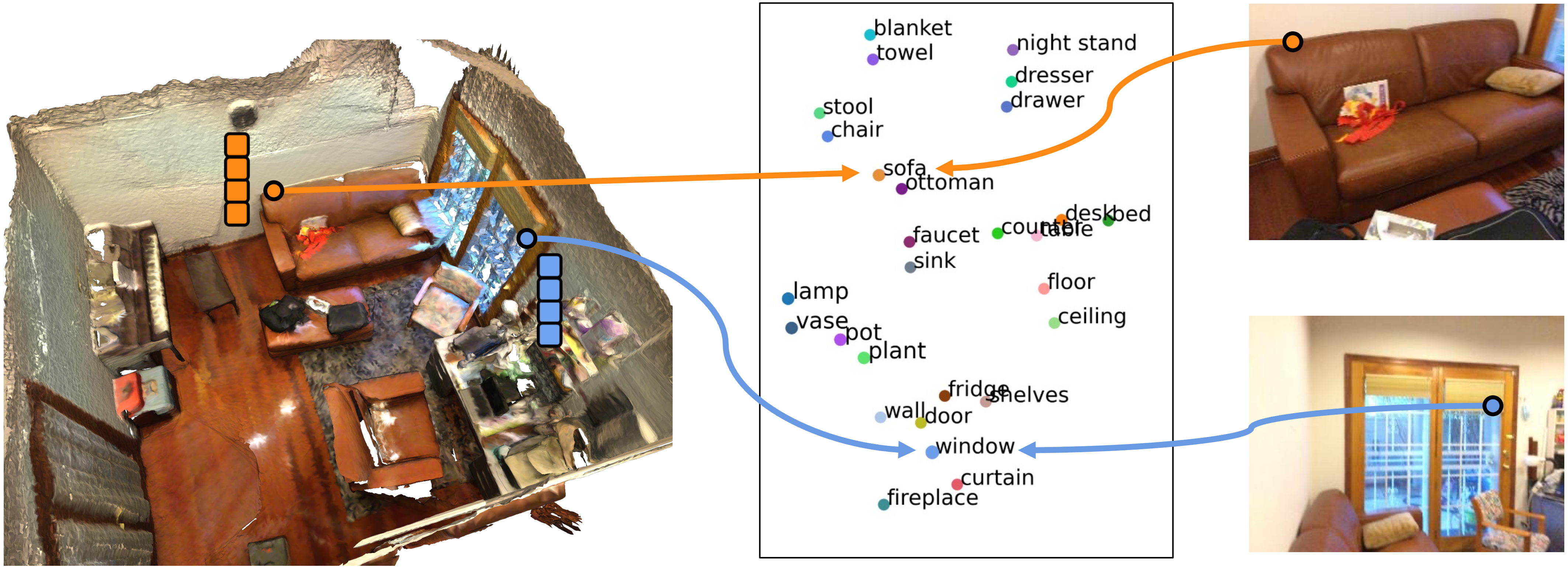
Only train and test on a few common classes

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

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

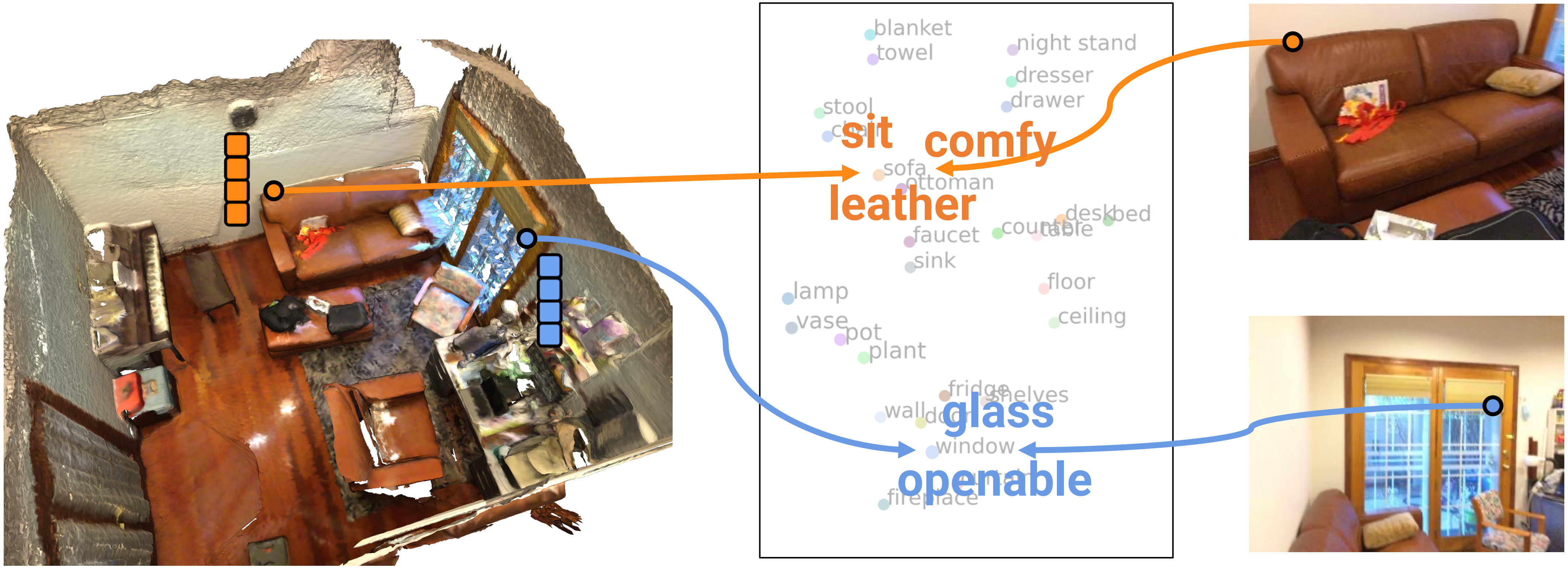


3D Geometry

CLIP Text Features
(visualize with T-SNE)

RGB Images

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



3D Geometry

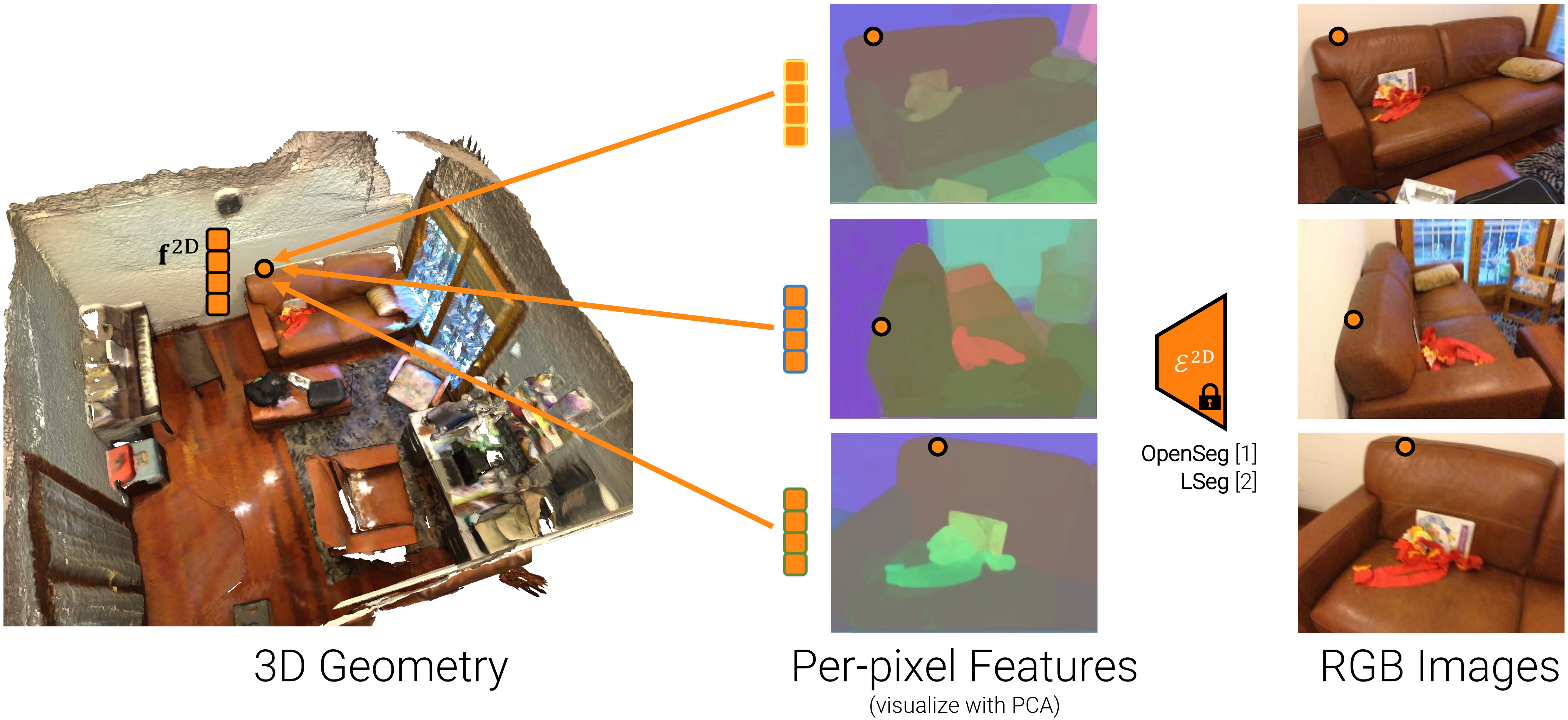
CLIP Text Features
(visualize with T-SNE)

RGB Images

Note: bold word embeddings are approximate

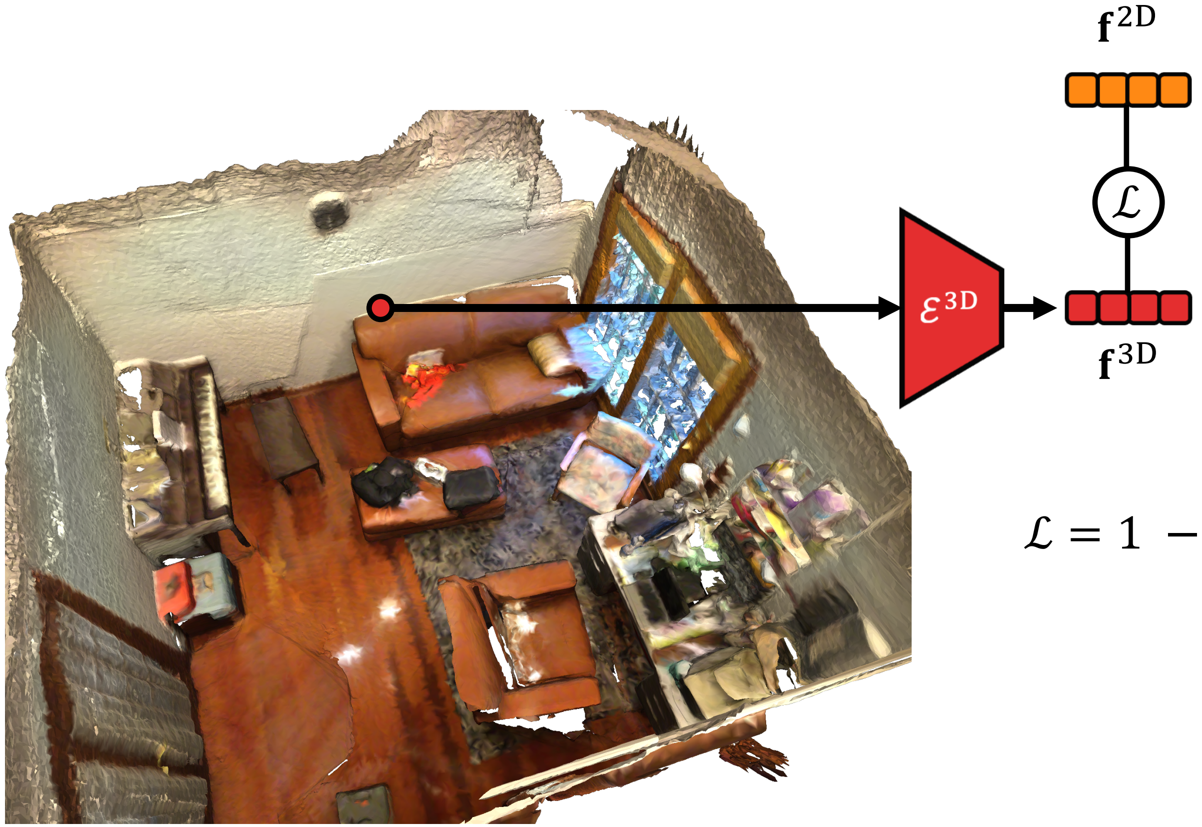
How to Learn Such Text-Image-3D Co-Embeddings?

Step 1: Multi-view Feature Fusion



[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

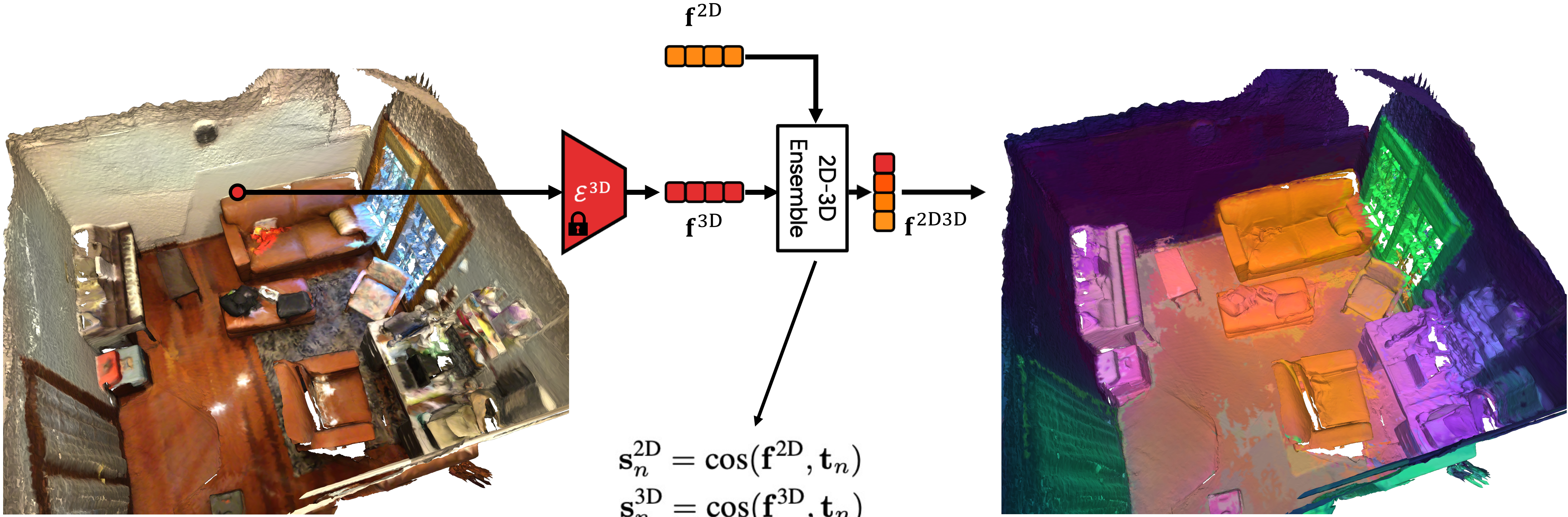
Step 2: 3D Distillation



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

3D Geometry

Step 3: 2D-3D Ensemble



3D Geometry

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

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)

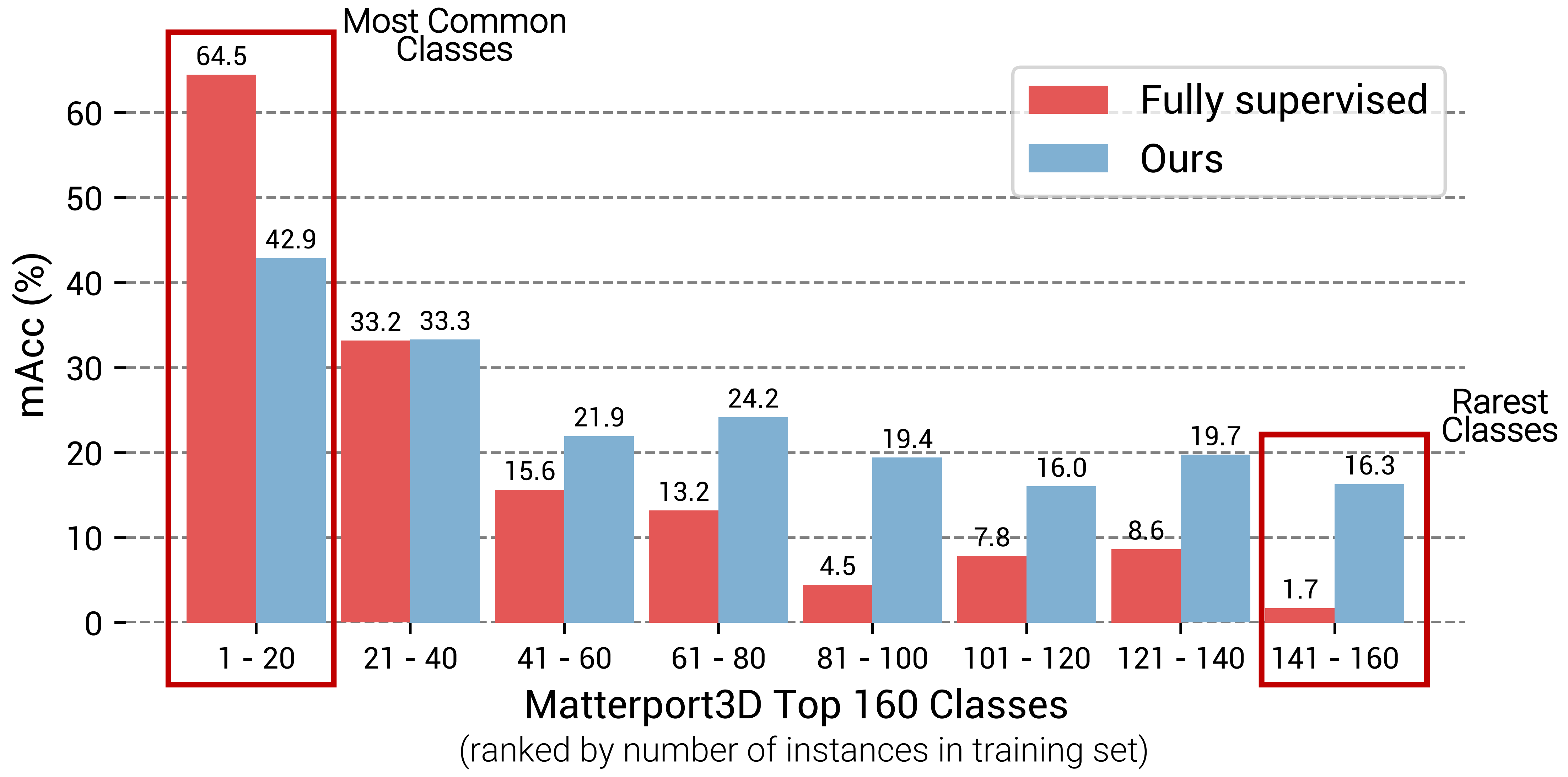
wall floor cabinet bed chair sofa table door window bookshelf picture counter desk curtain refrigerator shower curtain toilet sink bathtub other

Our Zero-shot 3D Segmentation

(160 classes)

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

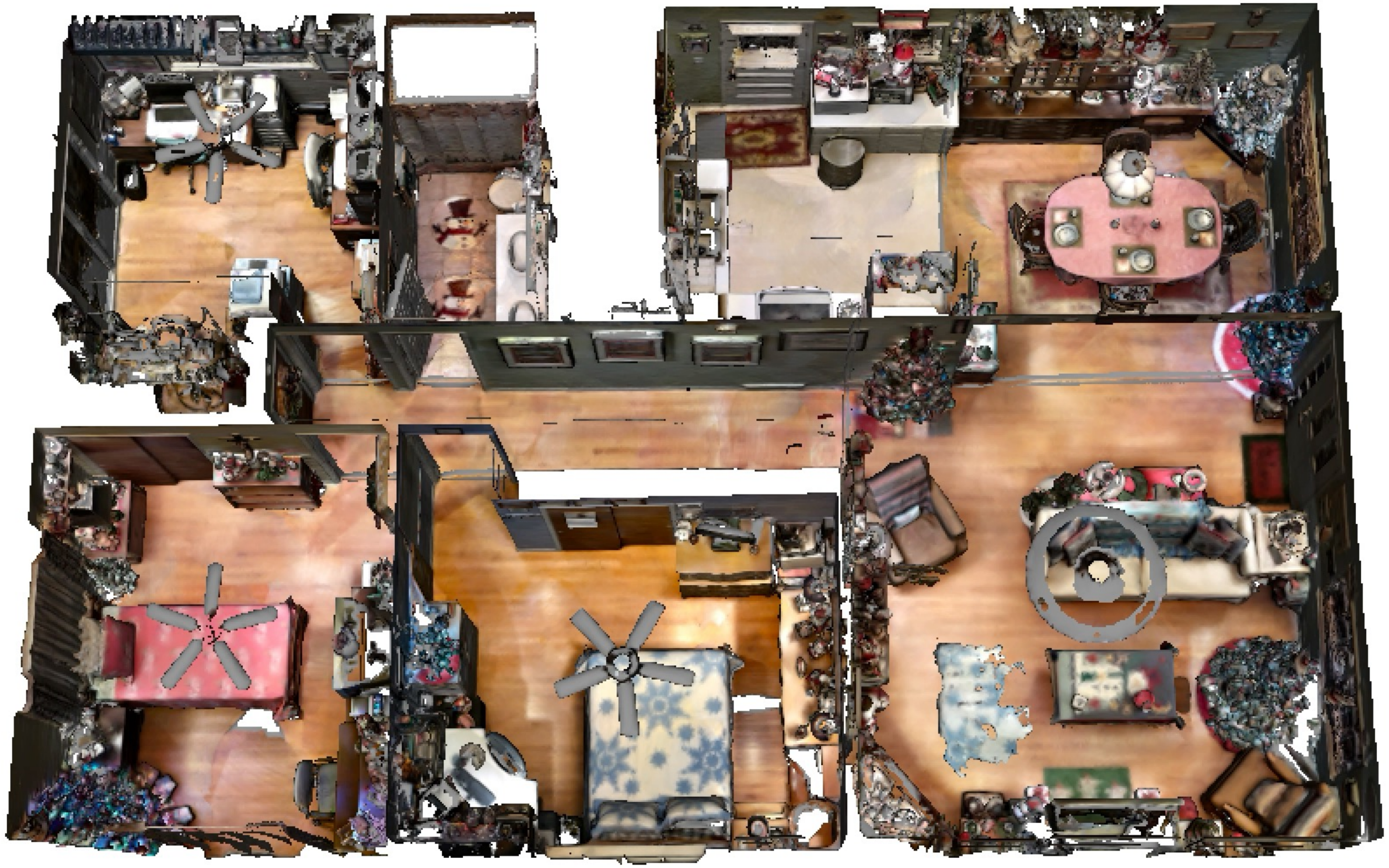
Comparison


Most Common
Classes
Rarest
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)

Interactive Demo

Open-vocabulary 3D Scene Exploration

Text queries:



Take-home Message

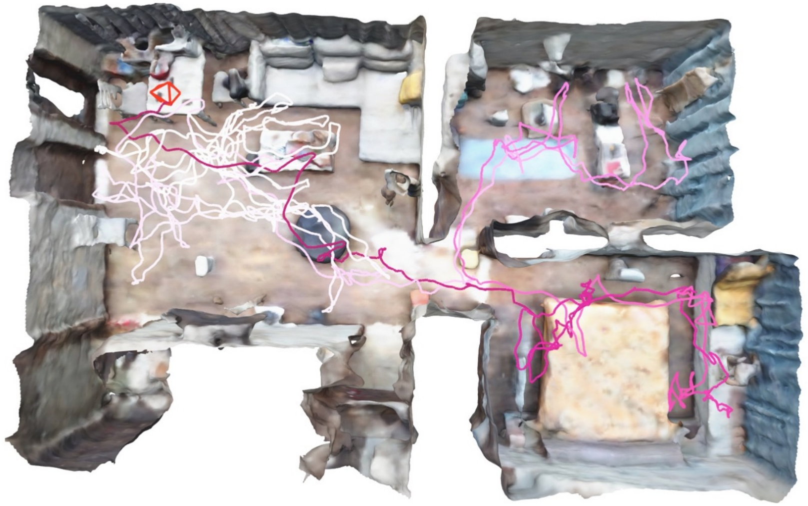
- We enable a wide range of applications by open-vocabulary queries

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

- The project can be improved in many aspects
  - Better feature fusion strategy than simple averaging
  - Combine CLIP features with NeRF/SLAM
    - concept-fusion.github.io

How does NeRF advance 3D Scene Reconstruction?

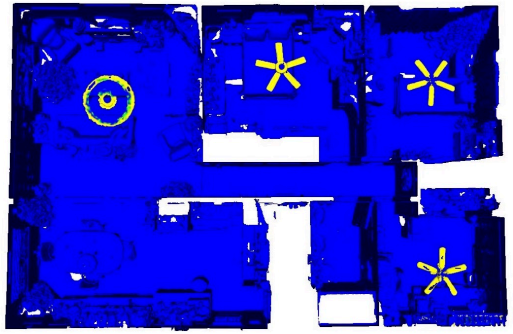
[NeurIPS'22] **MonoSDF**
github.com/autonomousvision/monosdf



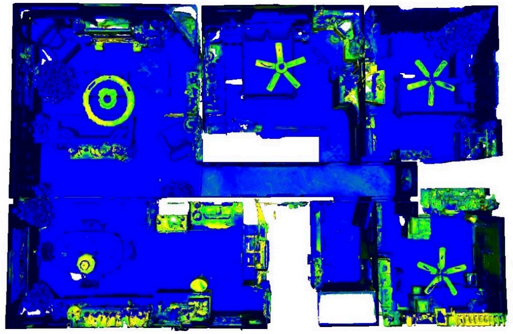
[CVPR'22] **NICE-SLAM**
pengsongyou.github.io/nice-slam

How does CLIP advance 3D Scene Understanding?

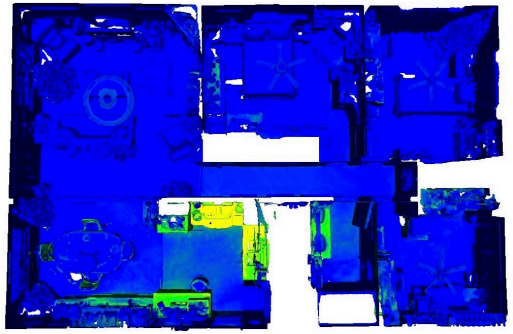
Input 3D Geometry



"fan" - Object



"metal" - Material



"kitchen" - Room Type

[CVPR'23] **OpenScene**
pengsongyou.github.io/openscene