Large-Scale 3D Scene Reconstruction with NeRF

Songyou Peng

ETH Zurich and Max Planck Institute for Intelligent Systems

Stanford Computational Imaging Lab

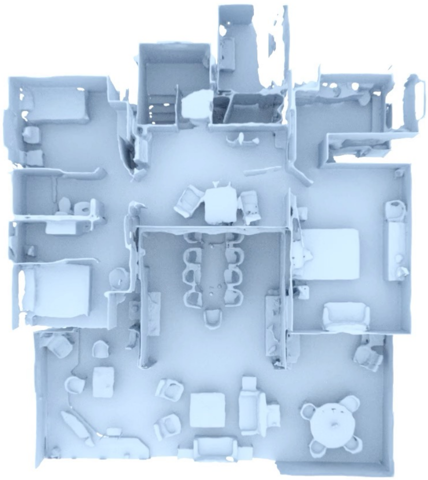
Oct 26, 2022

Who Am I?

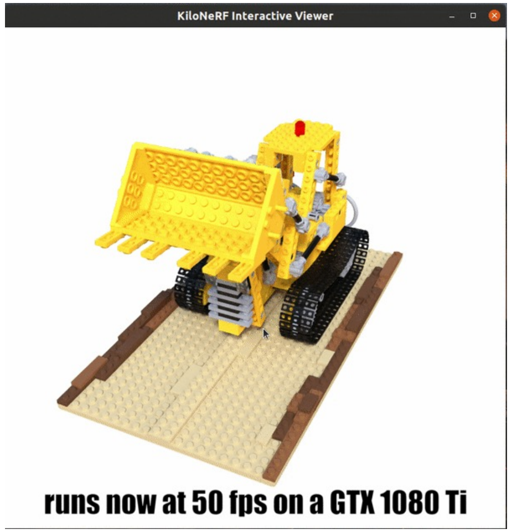
- **PhD Student since 2019.09**
  - Marc Pollefeys
  - Andreas Geiger
- **Internships during PhD**
  - 2021: Michael Zollhoefer
  - Ongoing: Tom Funkhouser
- Open to 1:1 chat!

pengsongyou.github.io

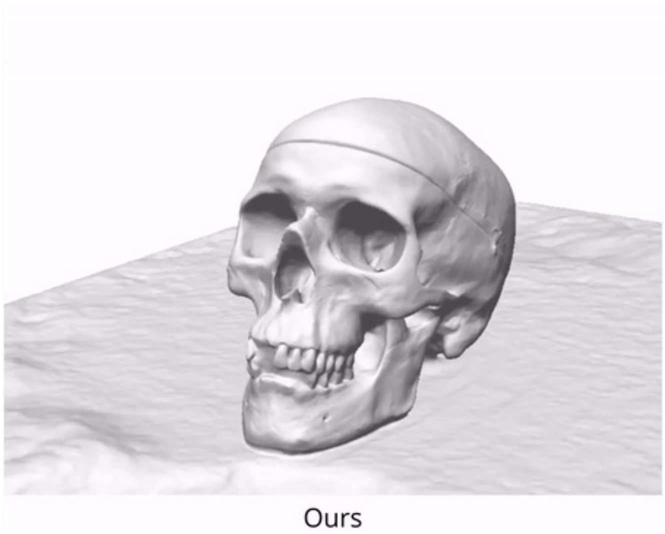
My PhD Topics: Neural Scene Representations
for 3D reconstruction, novel view synthesis, and SLAM



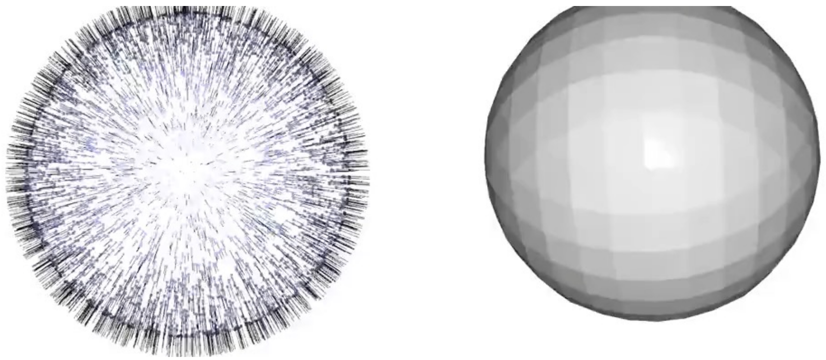
Convolutional Occupancy Networks
ECCV 2020 (Spotlight)



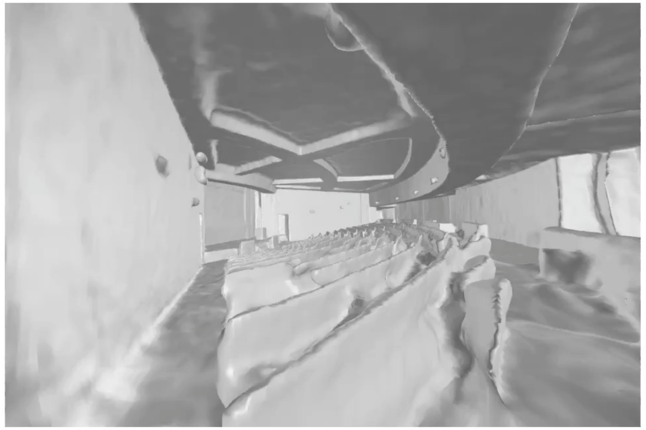
KiloNeRF
ICCV 2021



UNISURF
ICCV 2021 (Oral)



Shape As Points
NeurIPS 2021 (Oral)



Ours

MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022

NeRF is awesome!



Some problems still exist…

😢 Poor underlying geometry

😢 Camera poses needed

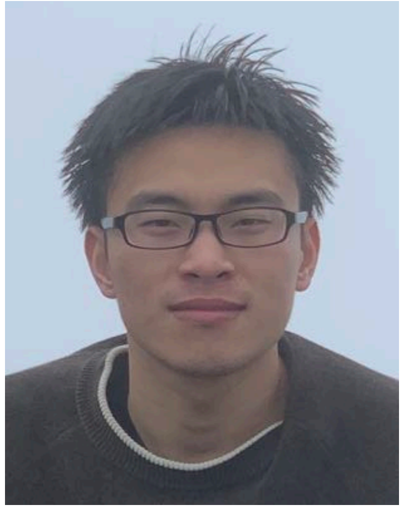
😊 MonoSDF

😊 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

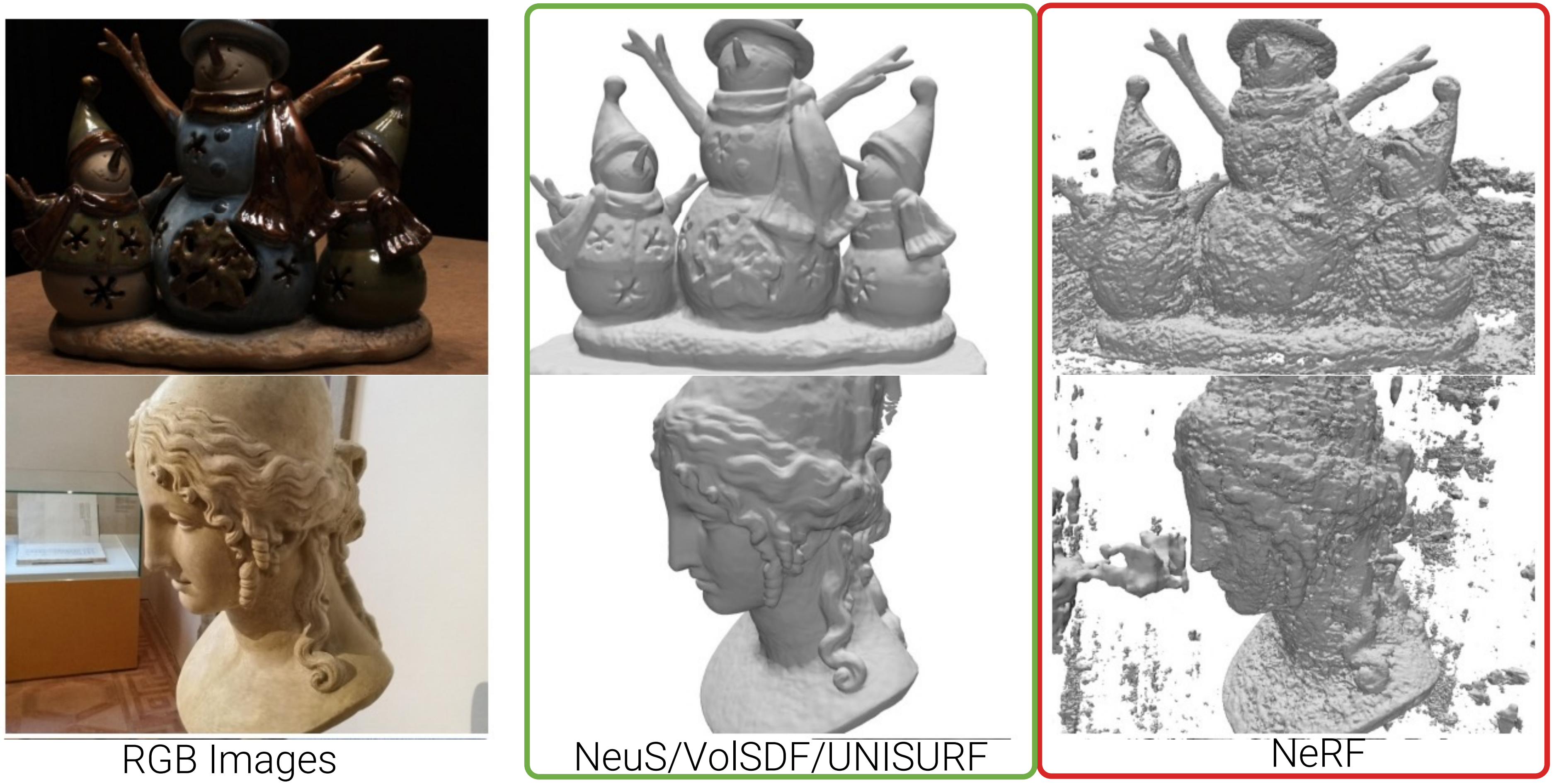
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

Neural Implicit Surfaces with Volume Rendering

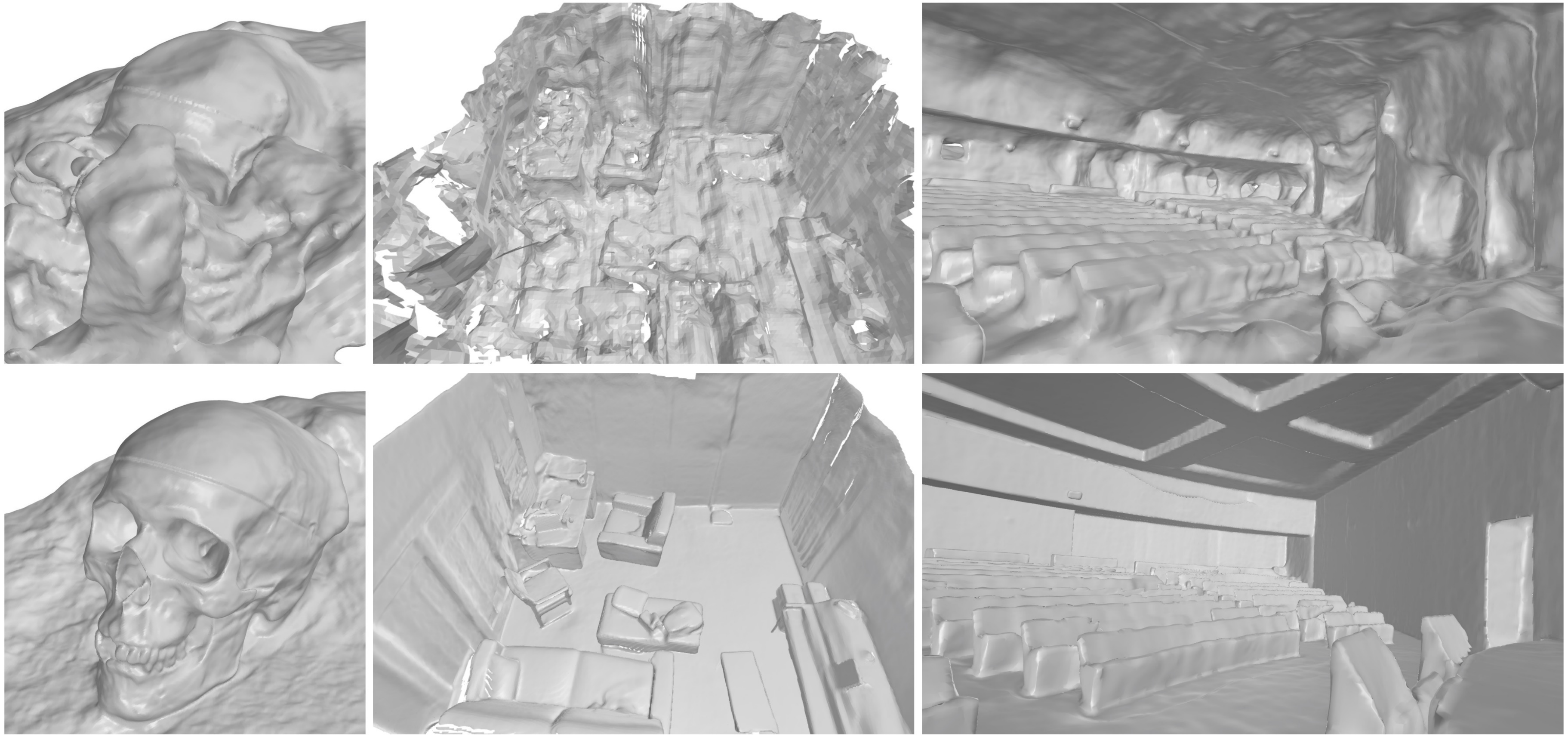


[1] Oechsle, Peng, Geiger: UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction. ICCV, 2021
[2] Wang, Liu, Liu, Theobalt, Komura, Wang: NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. NeurIPS, 2021
[3] Yariv, Gu, Kasten,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

Neural Implicit Surfaces with Volume Rendering

VolSDF

MonoSDF

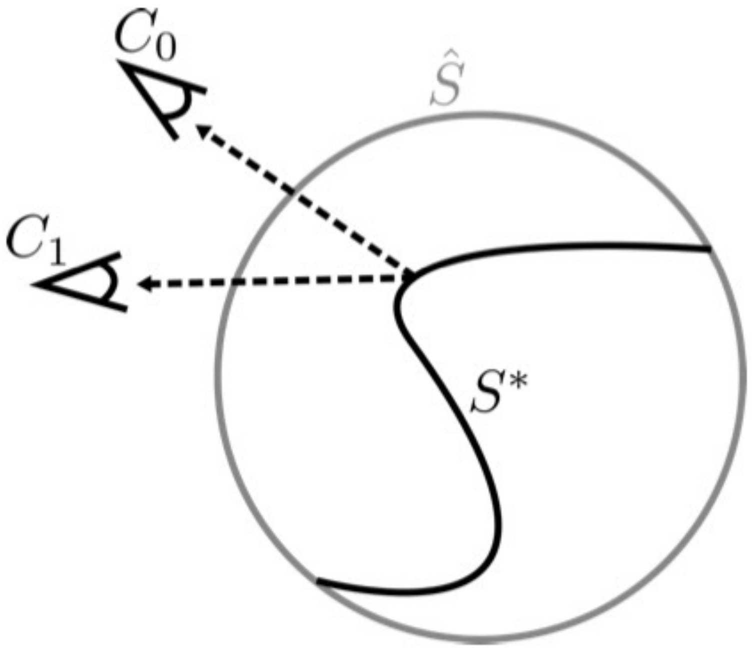


DTU (3 views) | ScanNet (464 views) | Tanks & Temples (298 views)

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

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

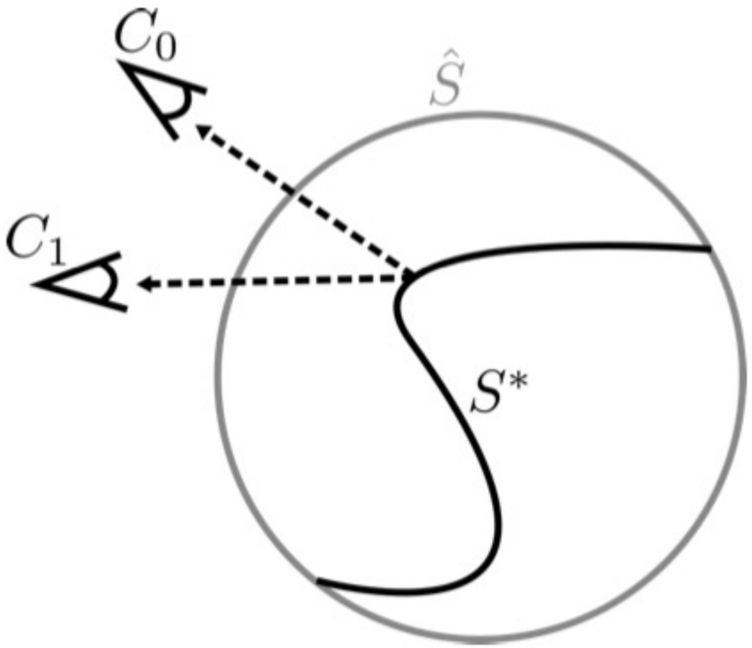
Shape-Appearance Ambiguity



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

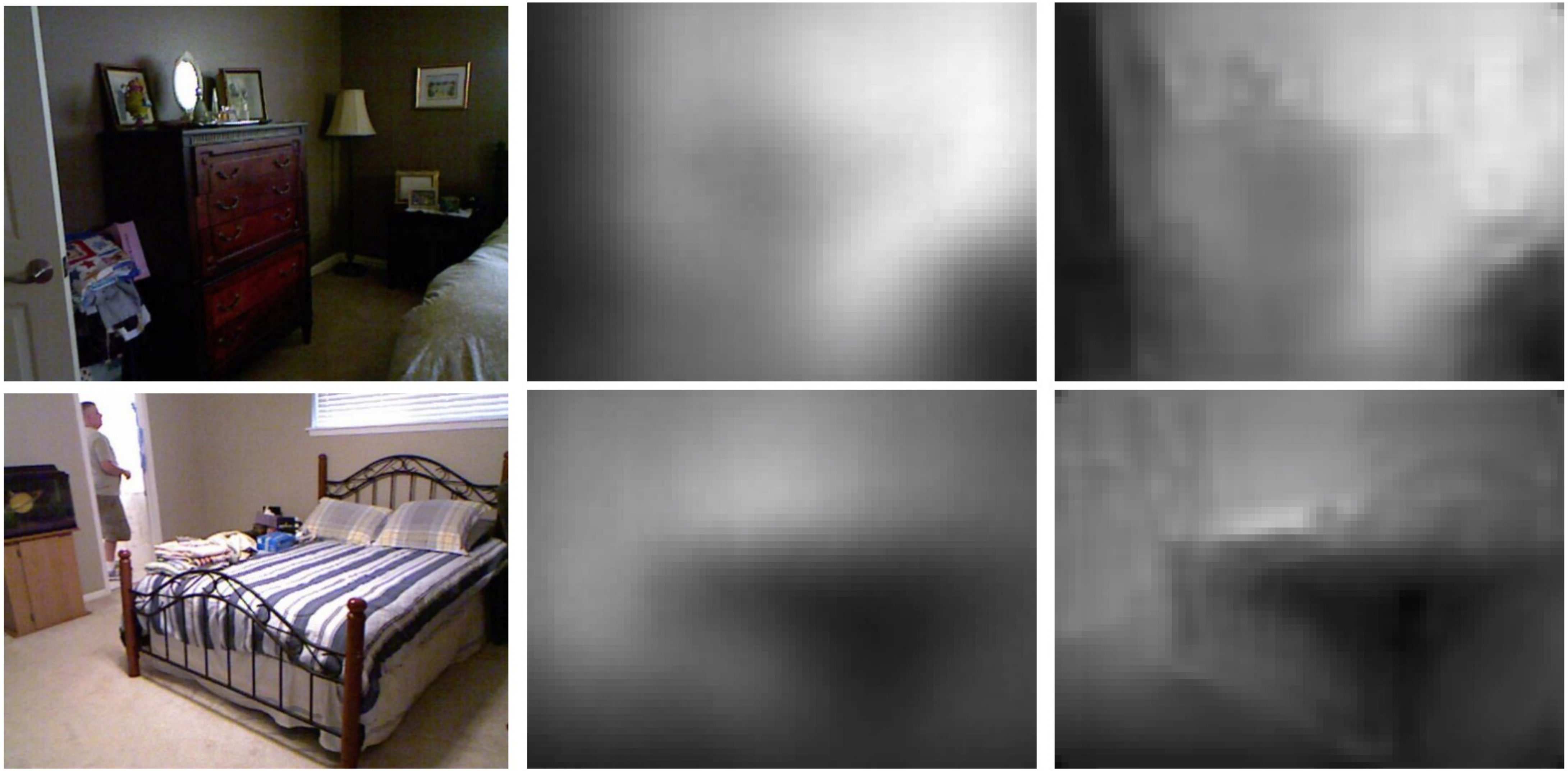


There exists an infinite number of photo-consistent explanations for input images!

⟹ Exploit monocular geometric priors

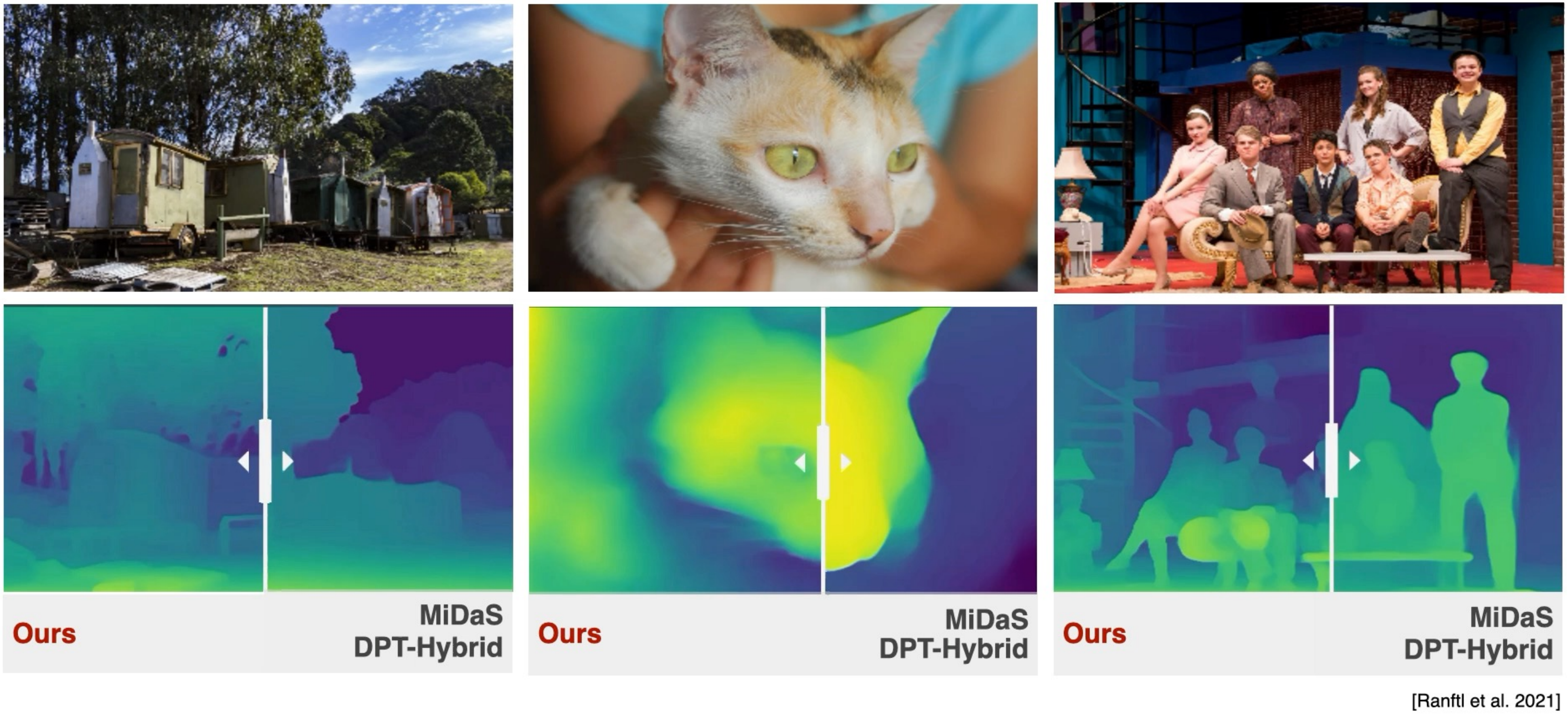
Zhang, Riegler, Snavely, Koltun: NeRF++: Analyzing and Improving Neural Radiance Fields. ArXiv, 2020

Depth Map Prediction from a Single Image



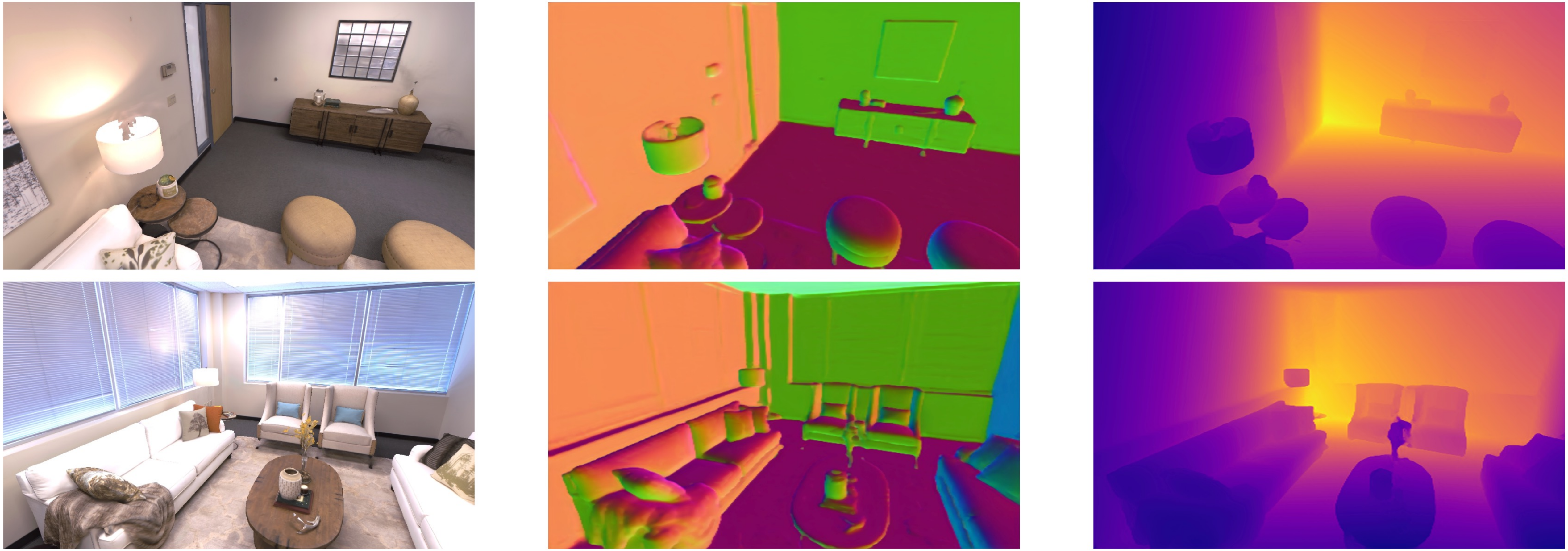
Eigen, Puhrsch and Fergus: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. NIPS, 2014

Omnidata



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

Omnidata

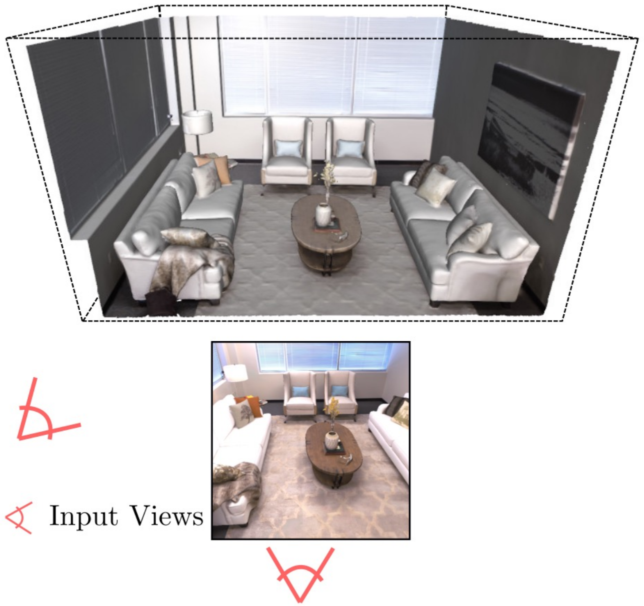


RGB Image | Omnidata Normal | Omnidata Depth

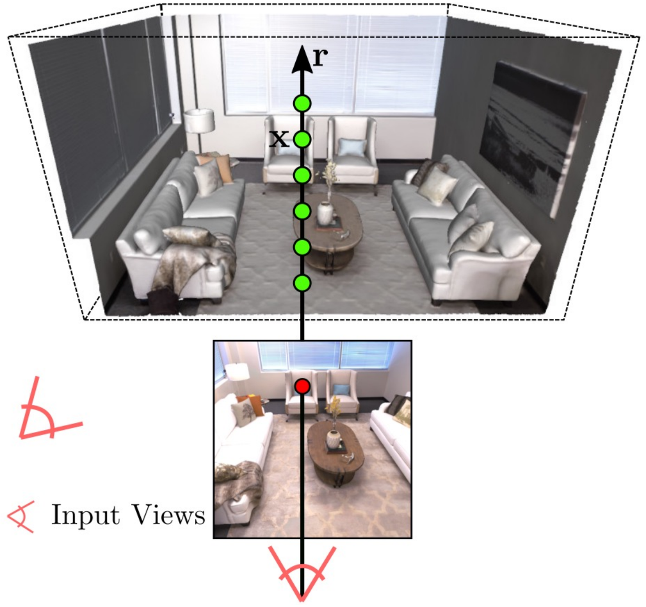
Eftekhar, Sax, Malik and Zamir: Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. ICCV, 2021.

MonoSDF

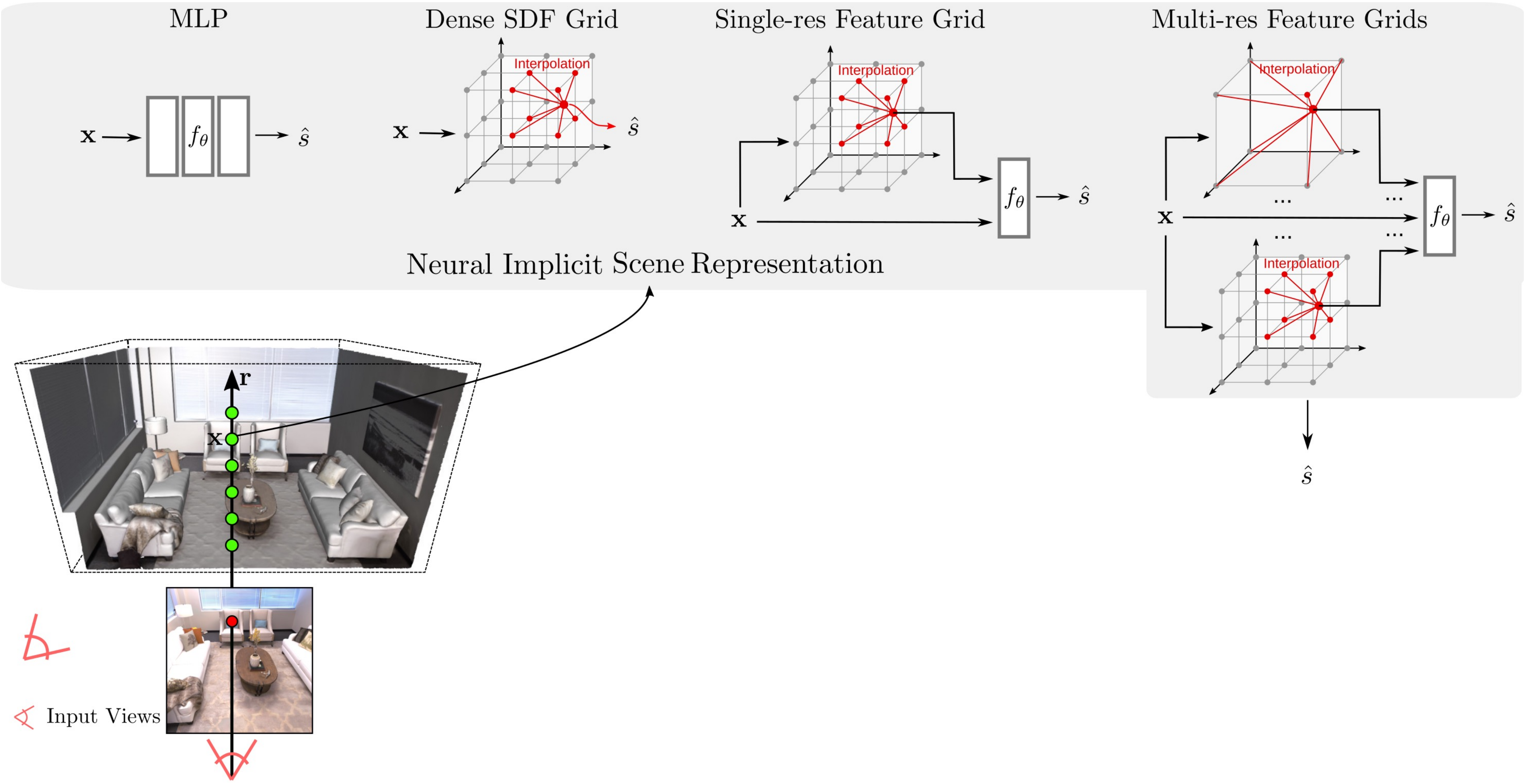
MonoSDF



MonoSDF

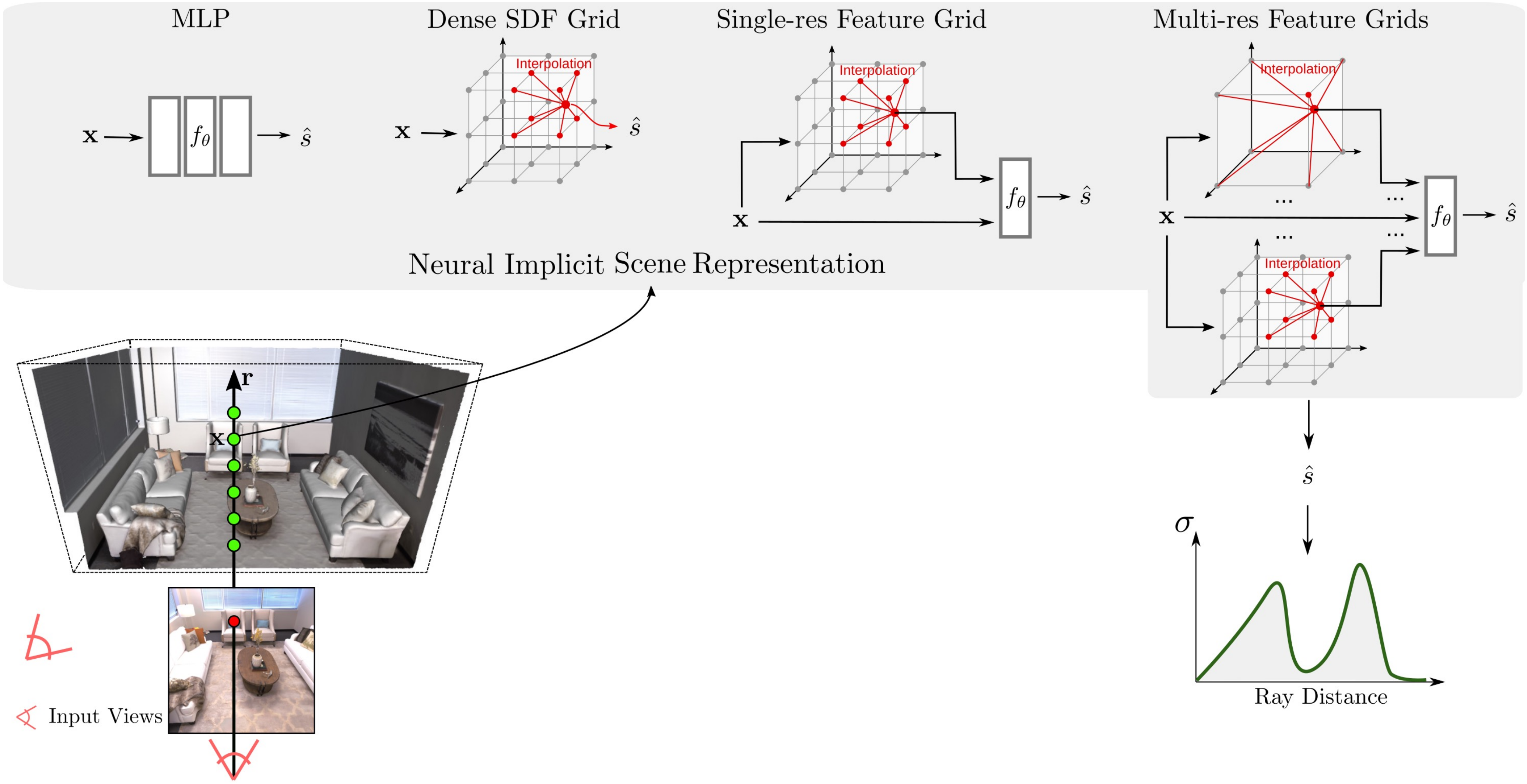


MonoSDF

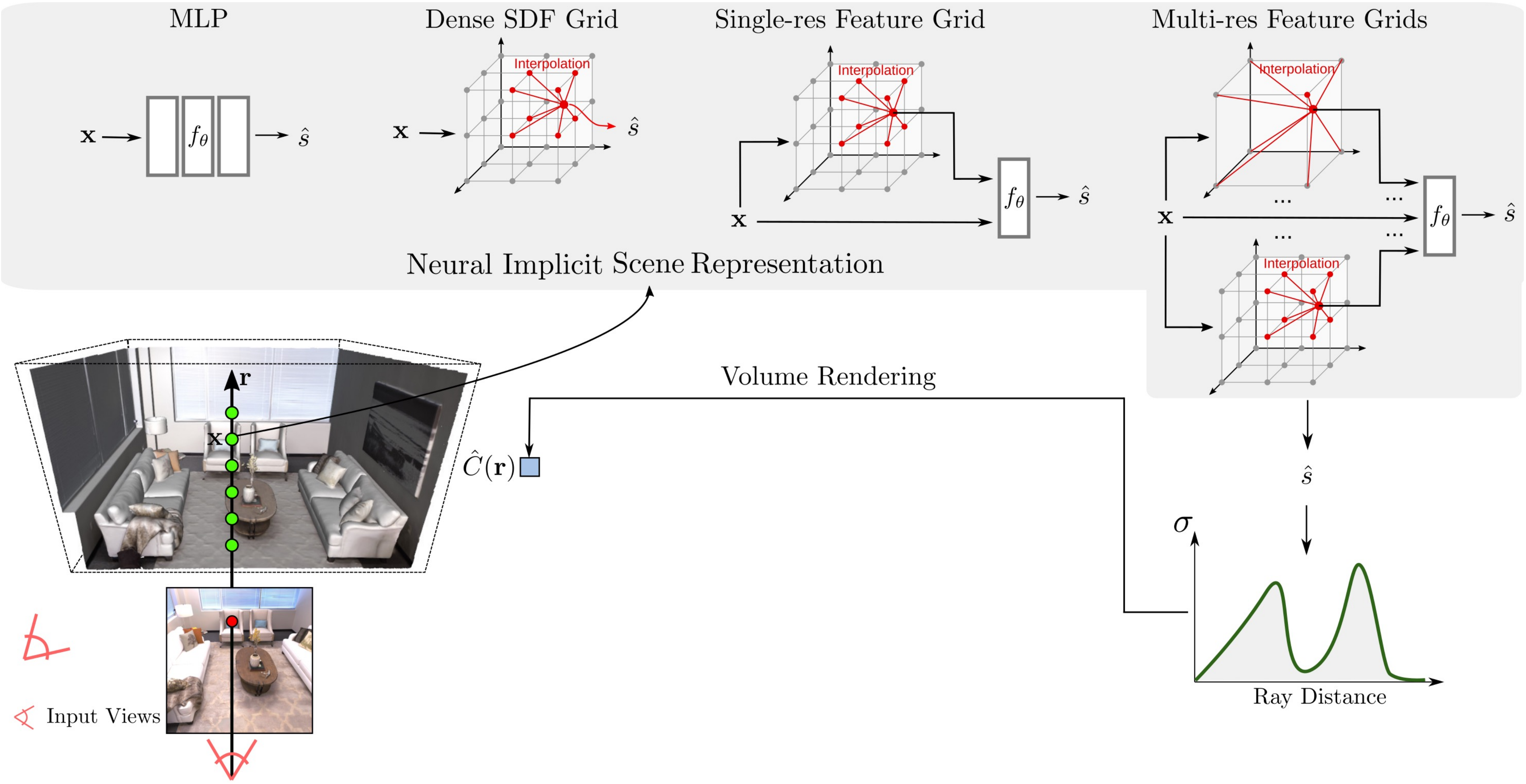


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
Interpolation
Interpolation
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
$\mathbf{r}$
Input Views

MonoSDF

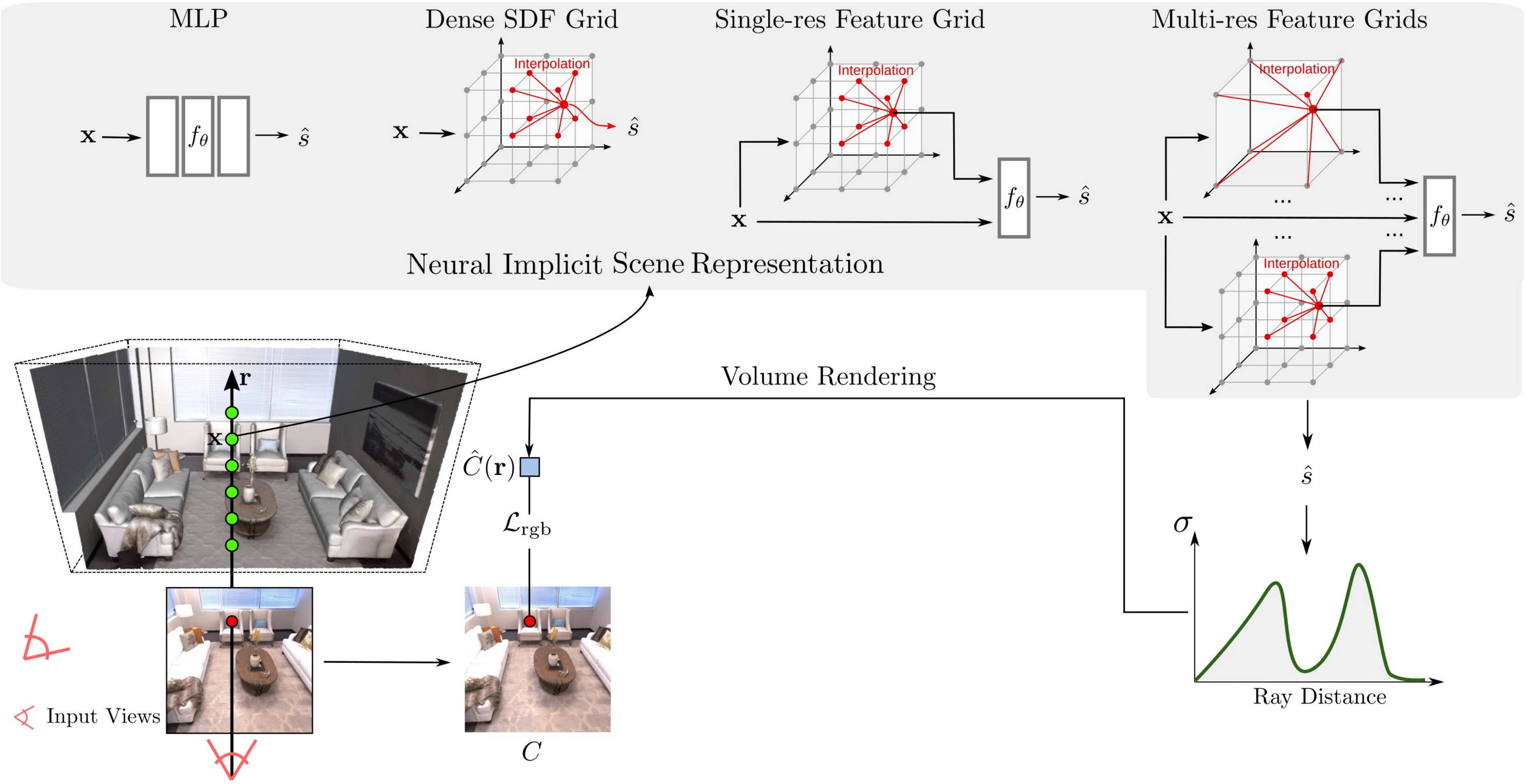

MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
$\mathbf{r}$
Input Views
σ
Ray Distance

MonoSDF



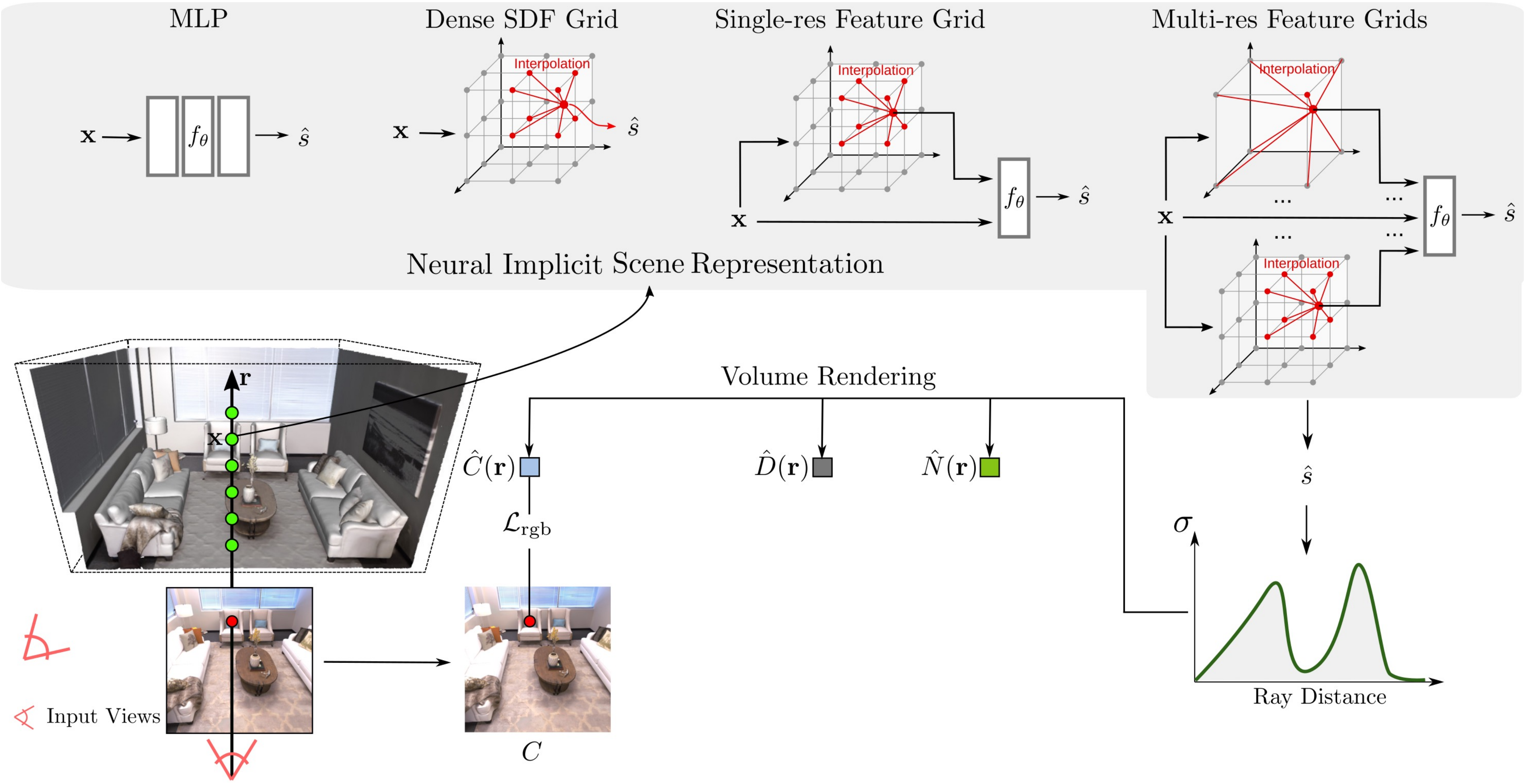
MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
$\mathbf{x}$
f_θ
$\hat{s}$
Neural Implicit Scene Representation
Volume Rendering
$\mathbf{r}$
$\hat{C}(\mathbf{r})$
Input Views
σ
Ray Distance

MonoSDF



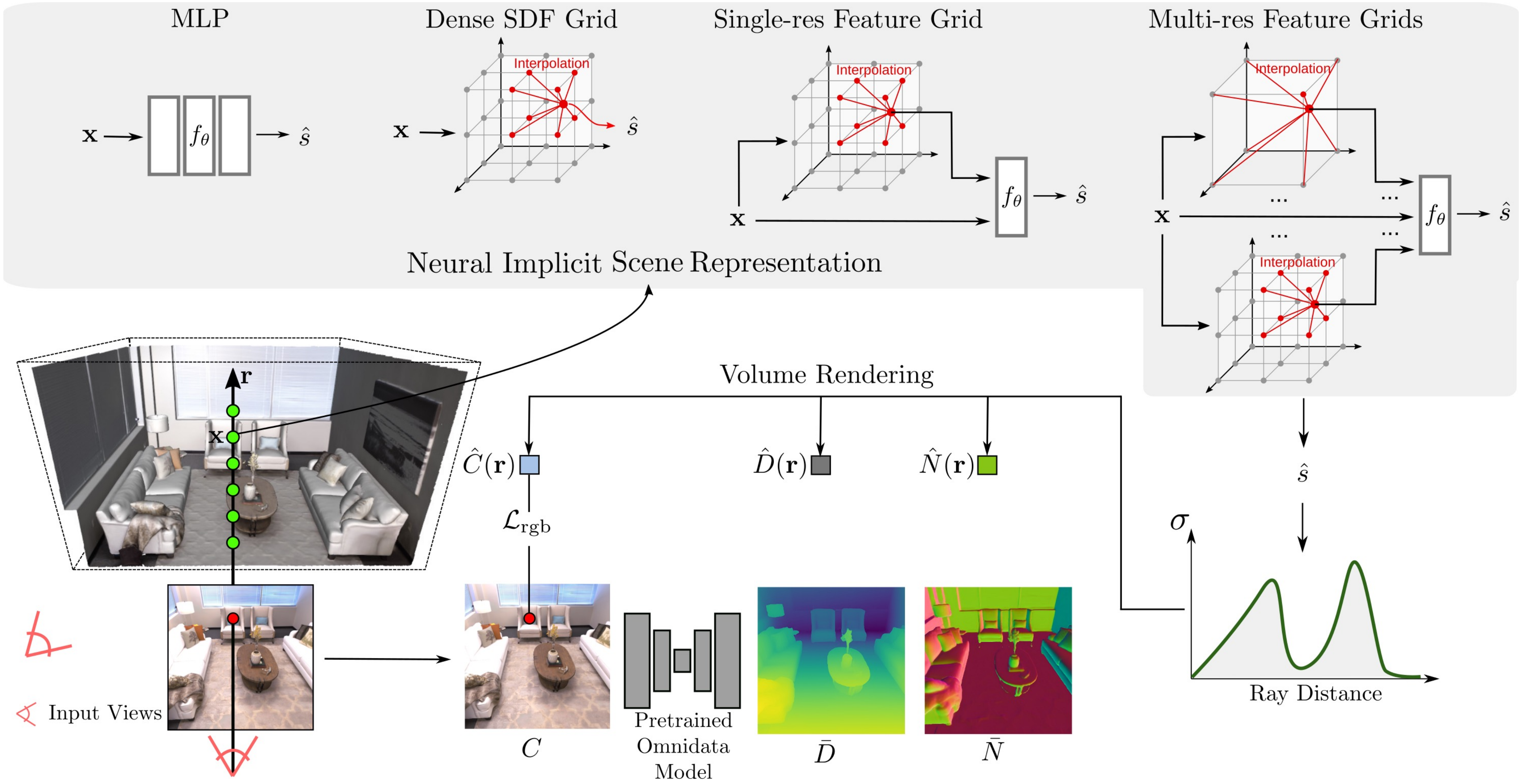
MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
x
f_θ
$\hat{s}$
...
Neural Implicit Scene Representation
Volume Rendering
r
$\hat{C}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
Input Views
σ
Ray Distance

MonoSDF

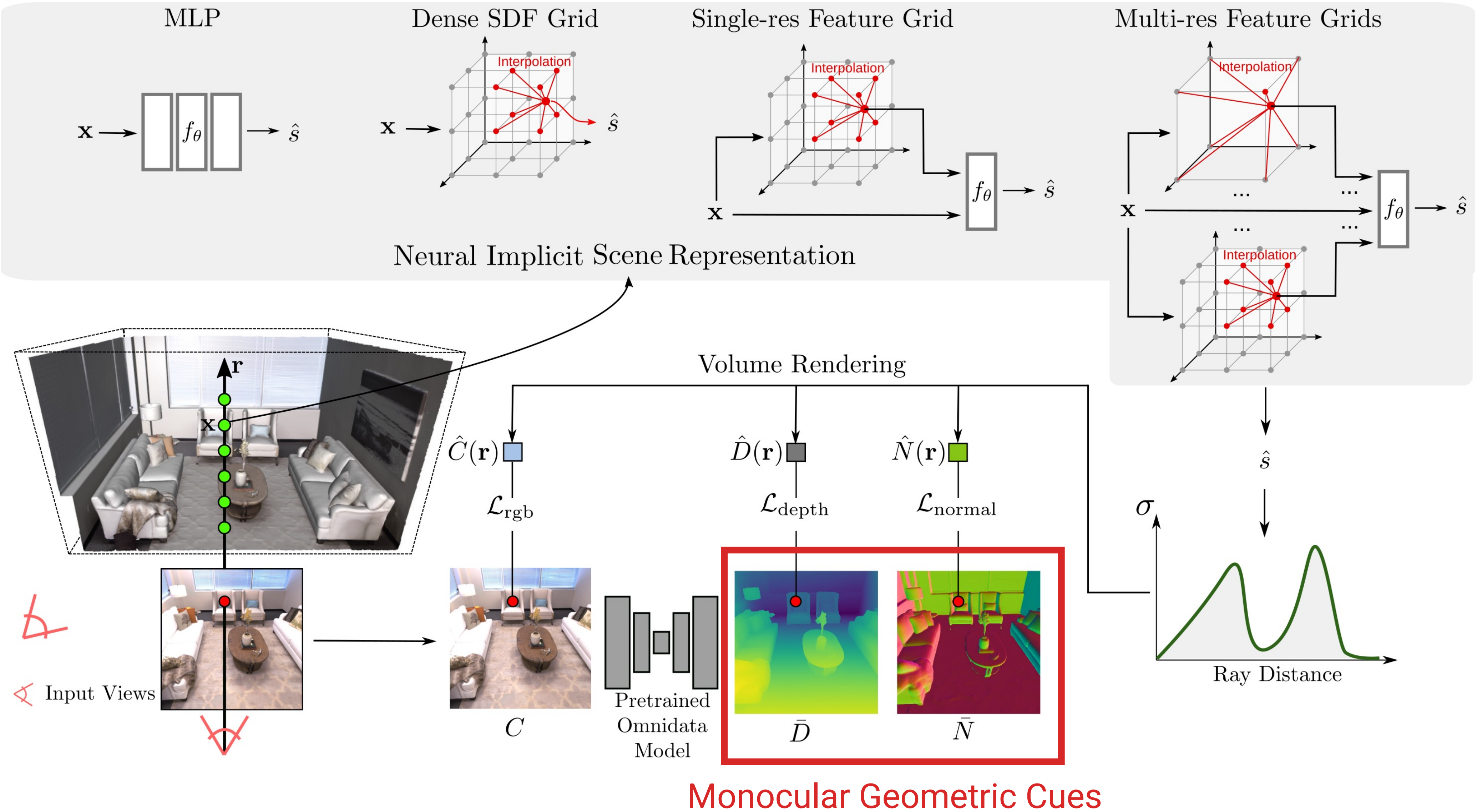


MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
Neural Implicit Scene Representation
Volume Rendering
$\hat{C}(\mathbf{r})$
$\hat{D}(\mathbf{r})$
$\hat{N}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
C
σ
Ray Distance
Input Views
$\mathbf{x}$
$\mathbf{r}$
f_θ
$\hat{s}$

MonoSDF

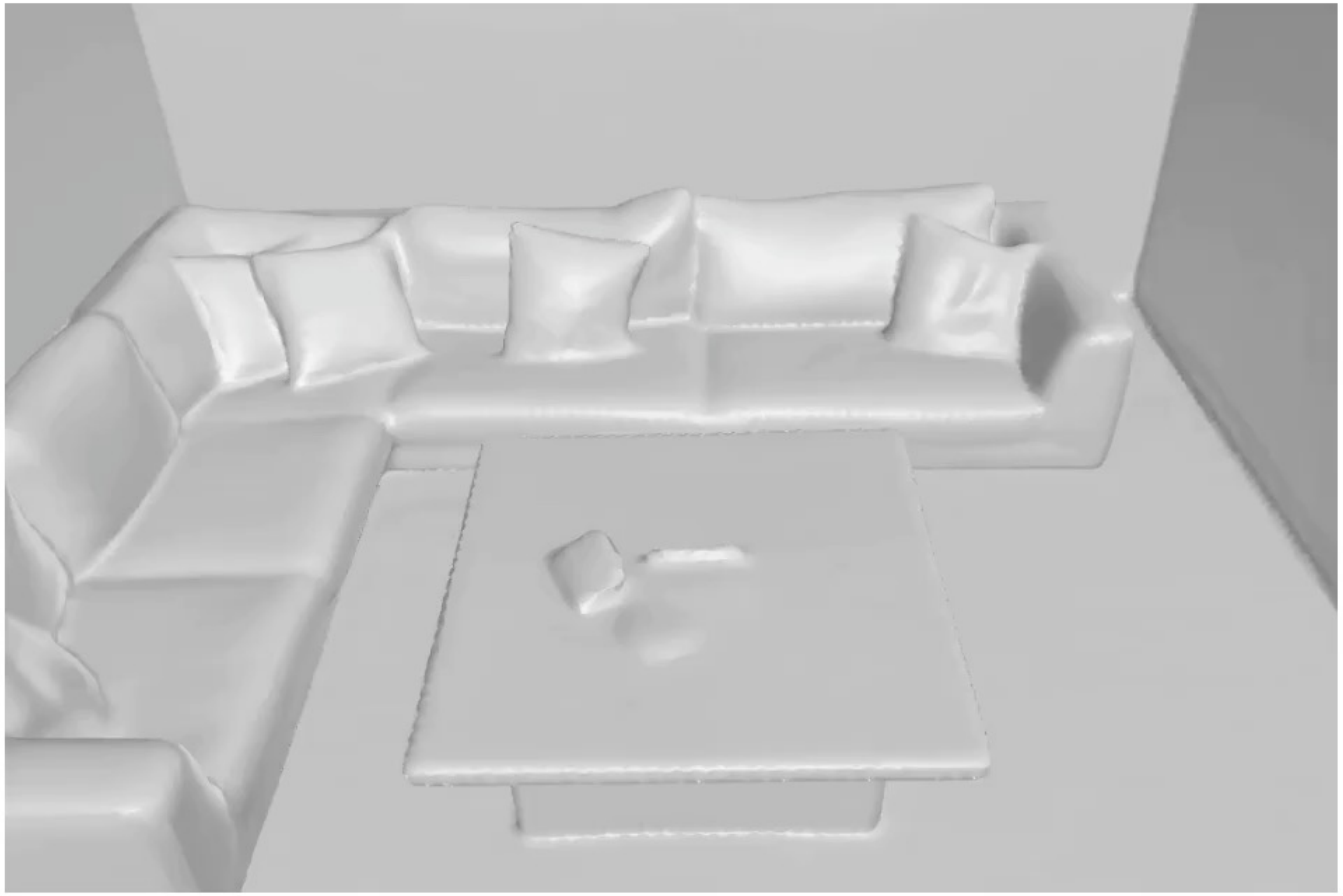

MLP
Dense SDF Grid
Single-res Feature Grid
Multi-res Feature Grids
Interpolation
x
f_θ
$\hat{s}$
Neural Implicit Scene Representation
Volume Rendering
$\hat{C}(\mathbf{r})$
$\hat{D}(\mathbf{r})$
$\hat{N}(\mathbf{r})$
$\mathcal{L}_{\text{rgb}}$
r
Input Views
C
Pretrained Omnidata Model
$\bar{D}$
$\bar{N}$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

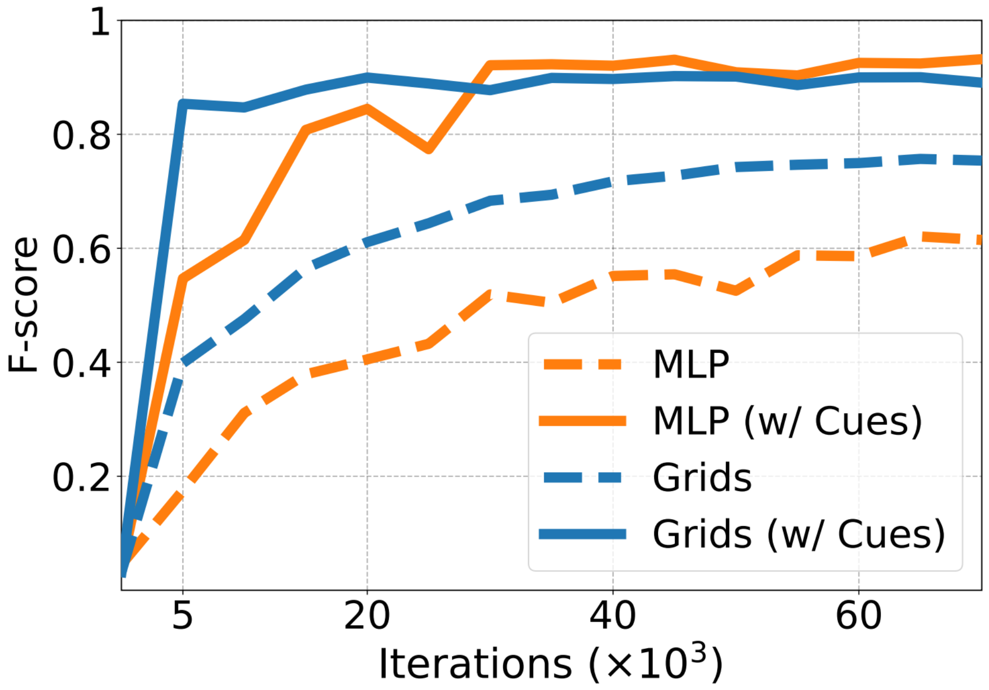
Ablation Study



Depth & Normal 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**



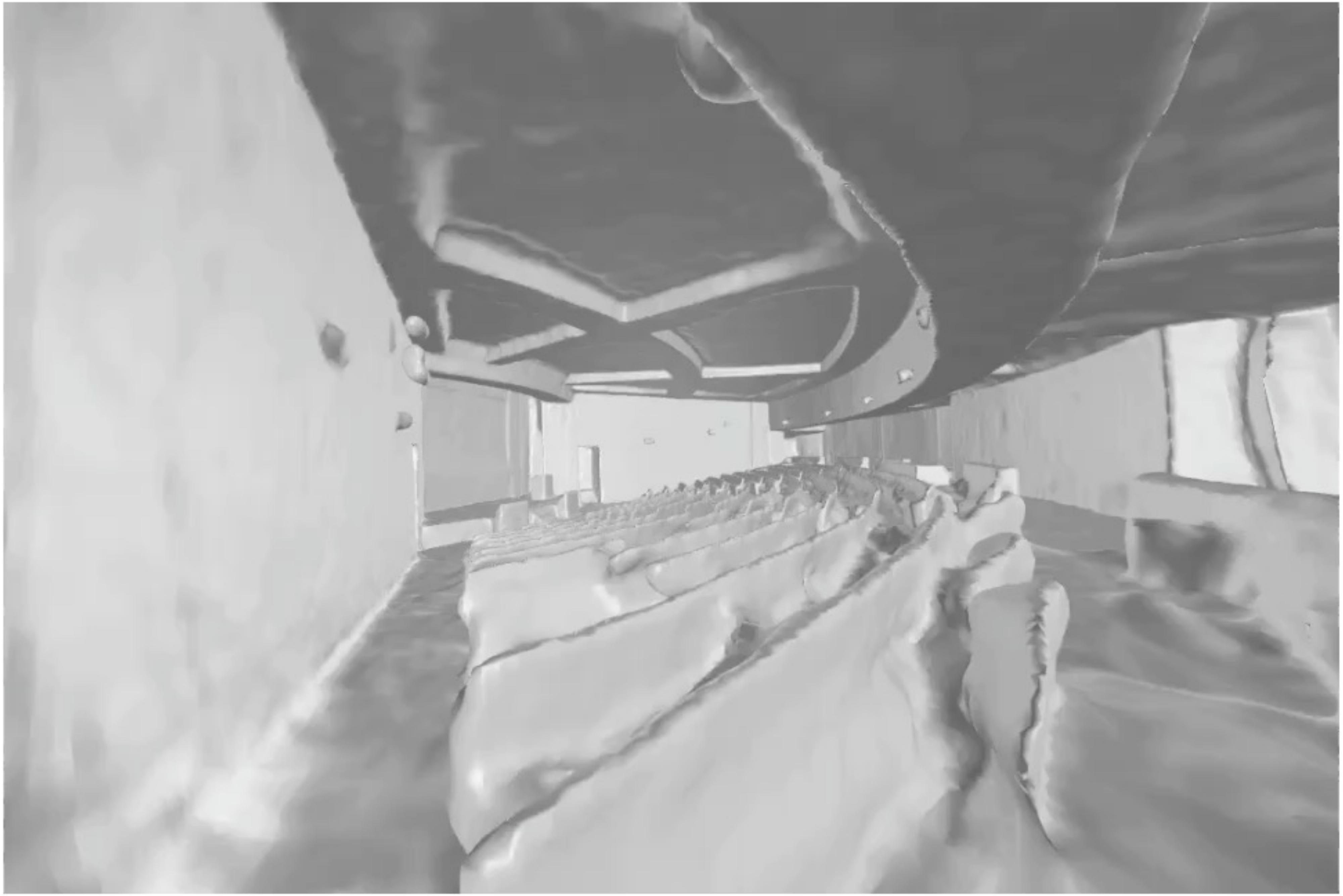
! Monocular cues improve reconstruction results significantly

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

! Monocular cues can improve **convergence speed**

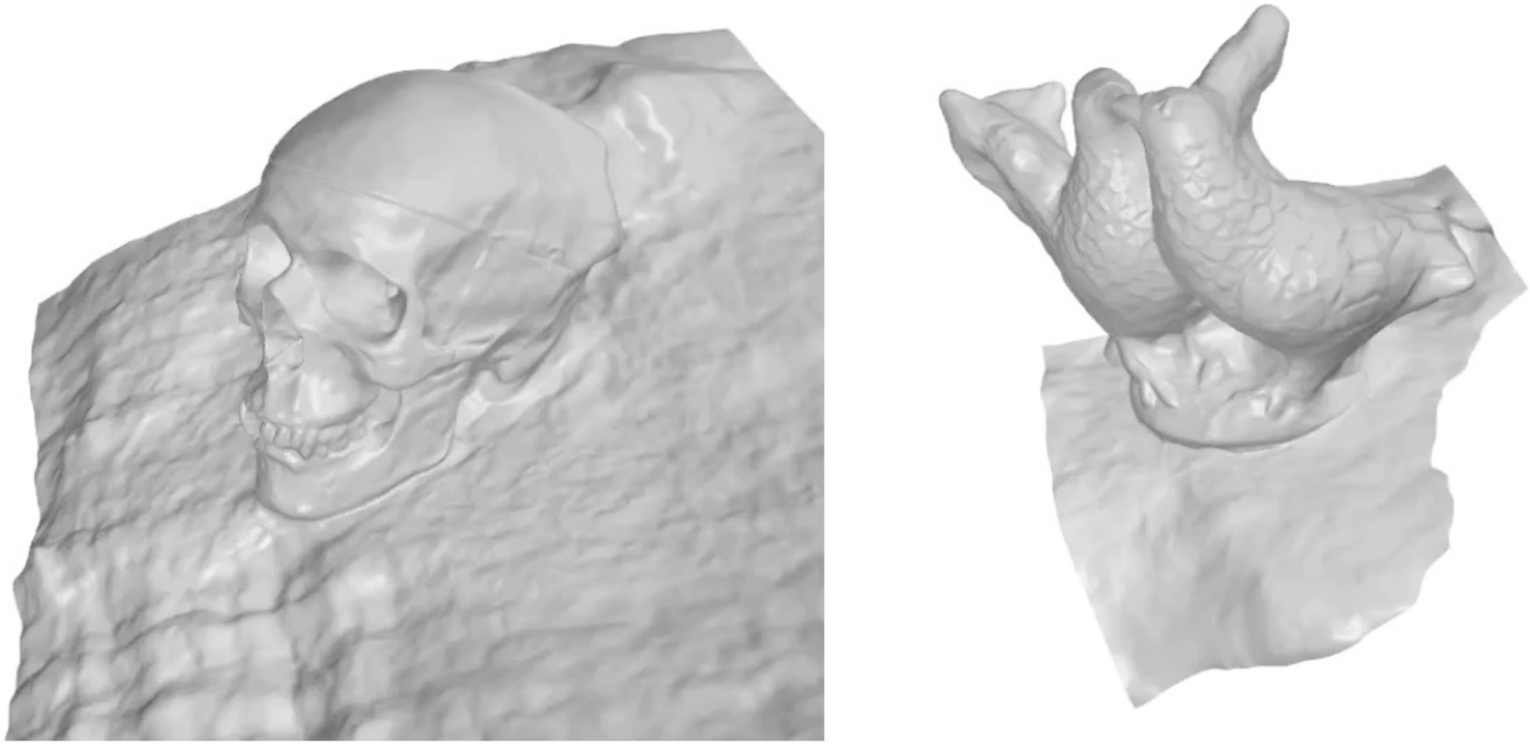
Baseline Comparisons on ScanNet

Ours



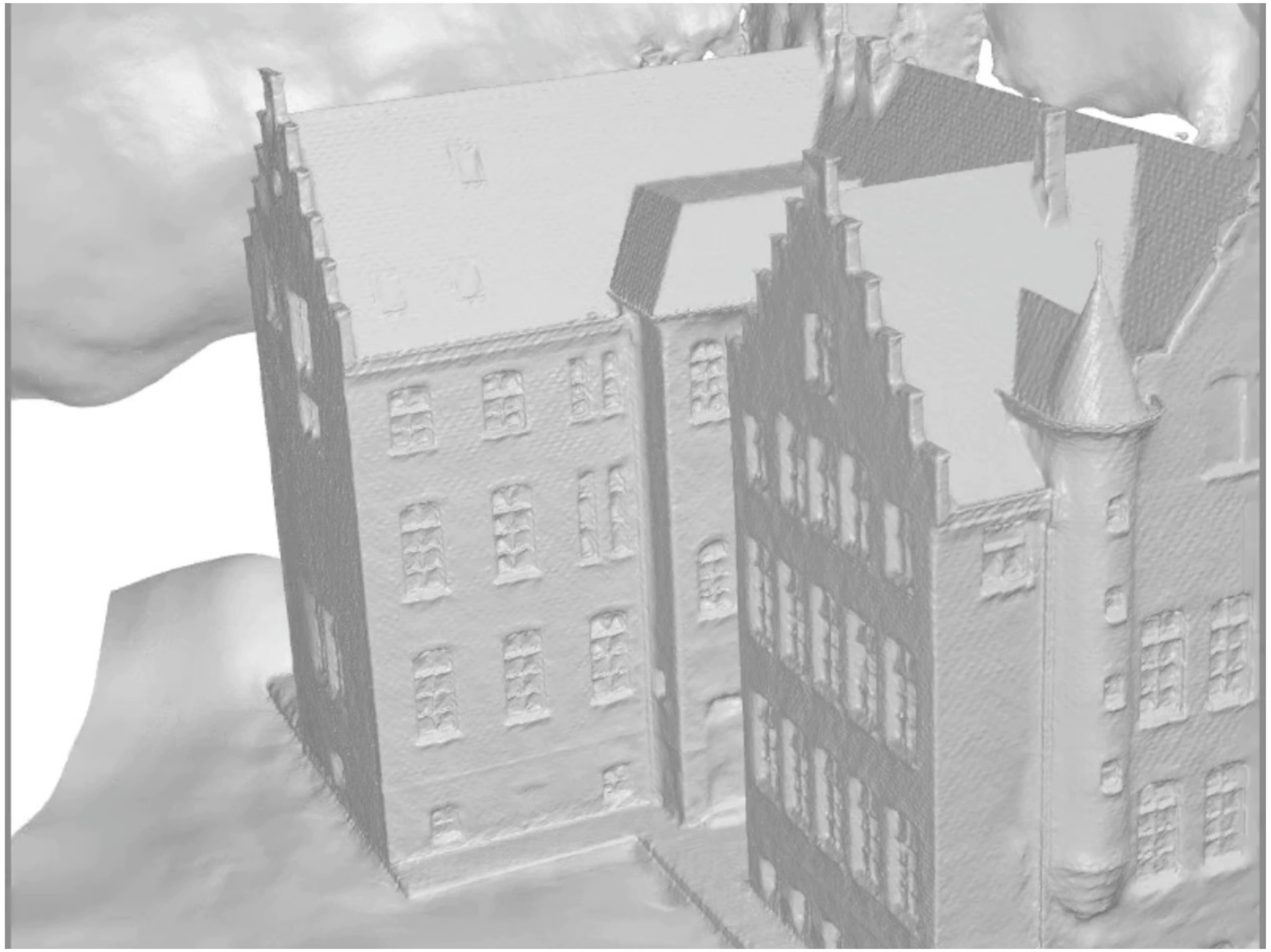
Ours

Baseline Comparisons on DTU (3-views)



Ours

Baseline Comparisons on DTU (all views)

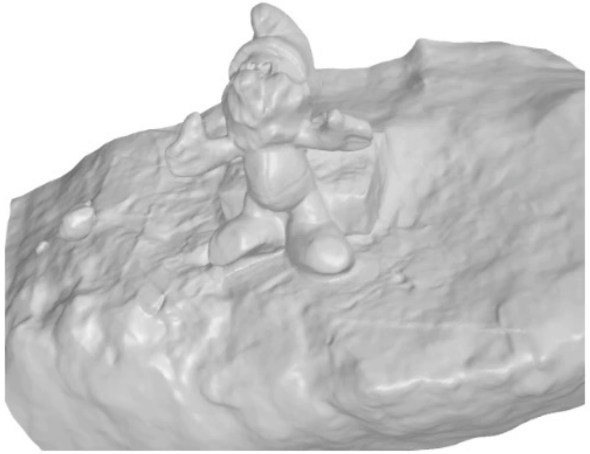


Ours (Grids)

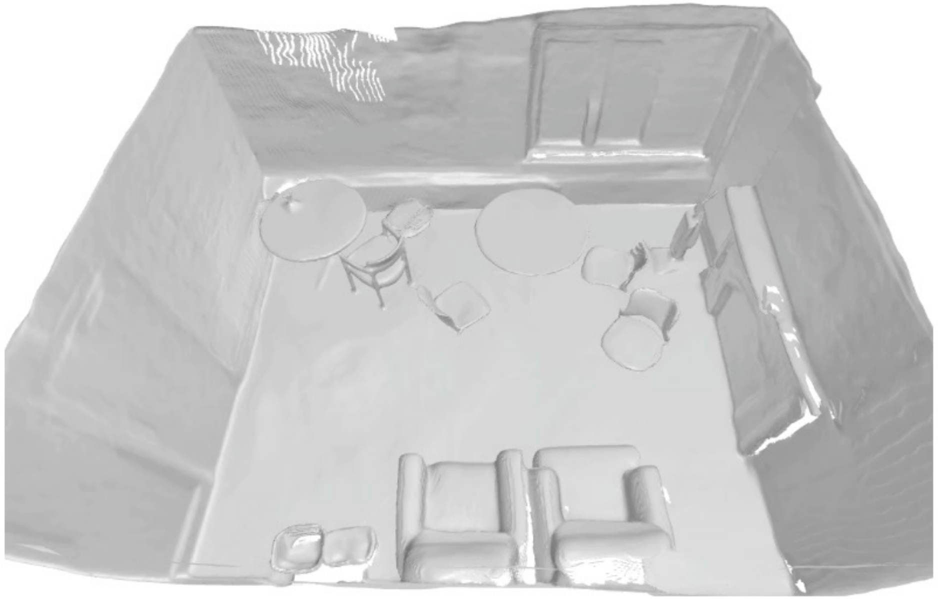
Multi-Res. Feature Grids with <u>High-Res. Cues</u>

Take-home Message

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



DTU (3 views)



ScanNet

Tanks and Temples

- Monocular cues improve reconstruction results and speed up optimization
- Analysis and investigate multiple scene representations
- Limitation: Still require camera poses given :(

NICE-SLAM

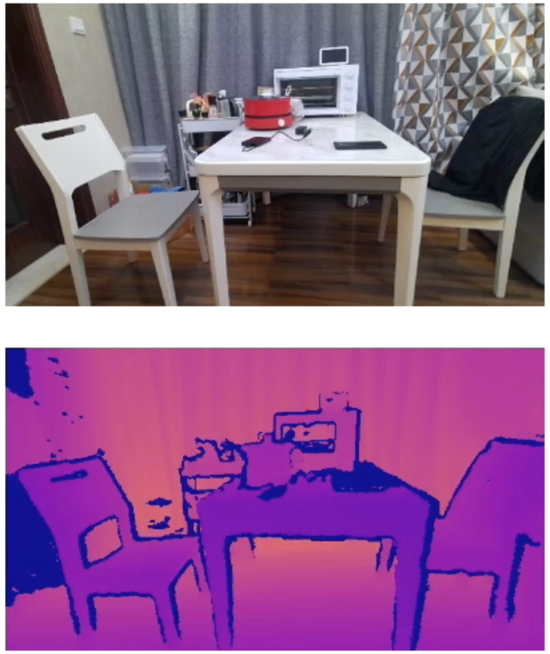
Neural Implicit Scalable Encoding for SLAM

CVPR 2022

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

* Equal Contributions

RGB-D Sequences



40x Speed

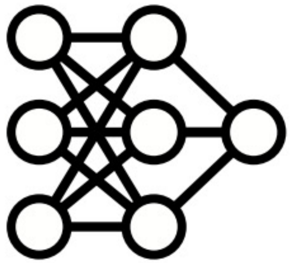
iMAP

[Sucar et al., ICCV'21]

First neural implicit-based online SLAM system

iMAP

[Sucar et al., ICCV'21]



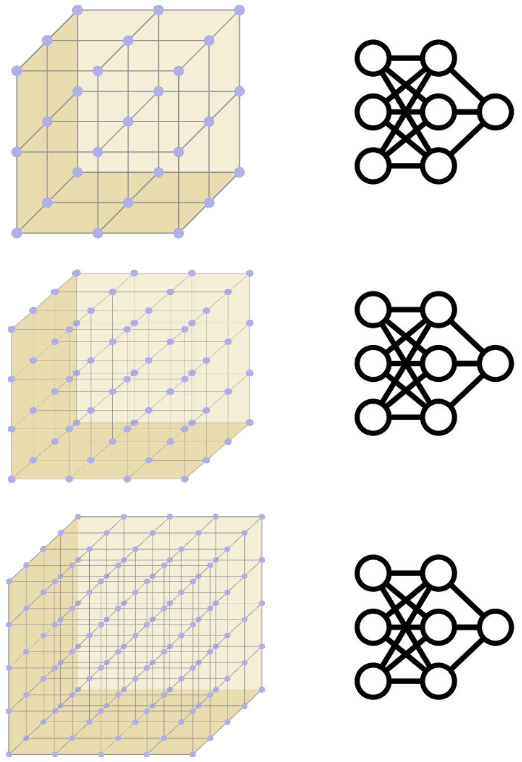
A single MLP

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

Predicted Poses

GT Poses

NICE-SLAM

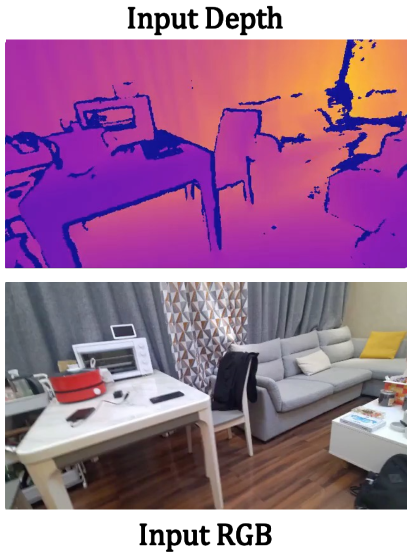


Feature grids + tiny MLPs

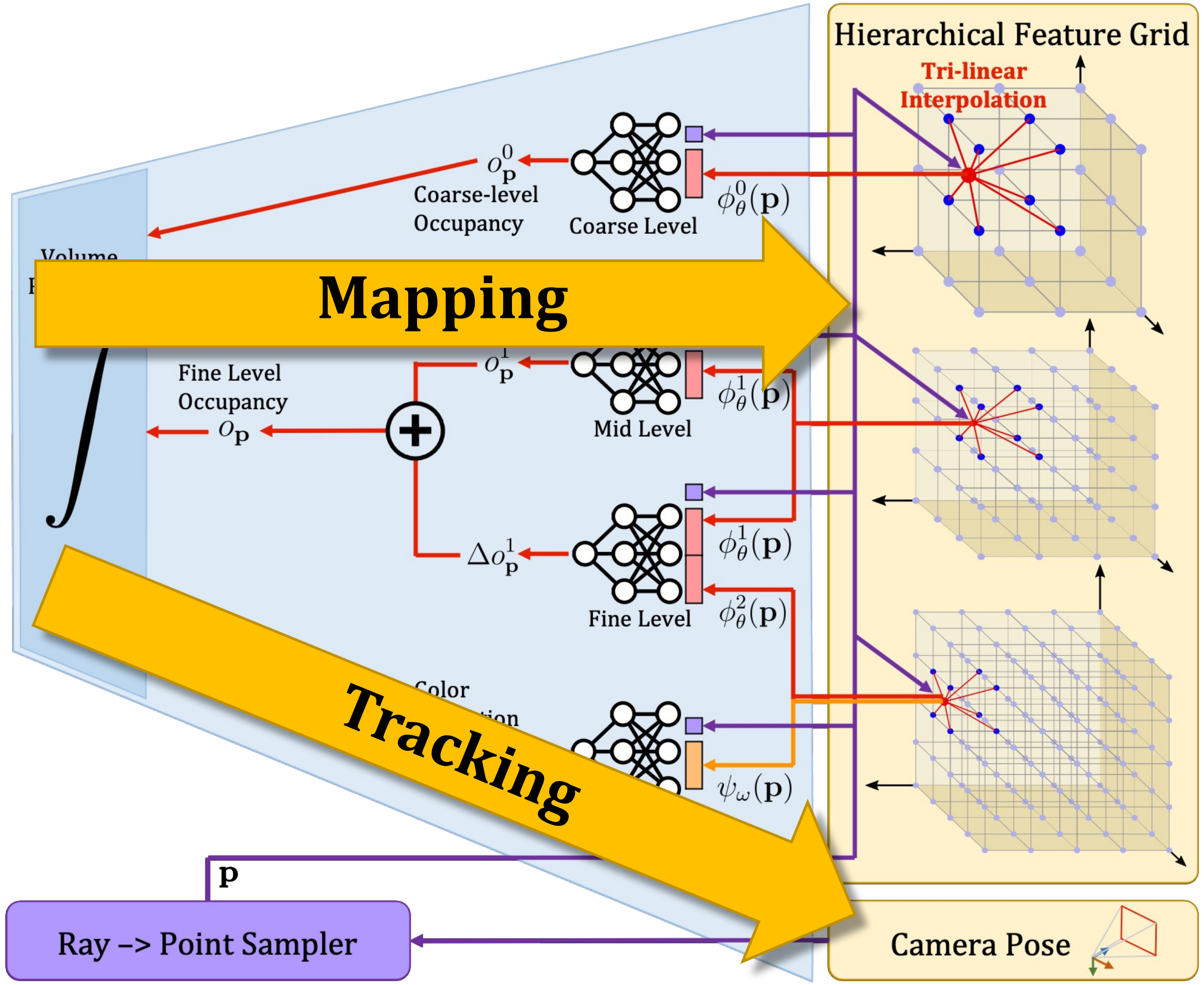
- Applicable to **large-scale scenes**
- Local update → **No forgetting problem**
- **Fast** convergence

Predicted Poses
GT Poses

Pipeline



Input Depth
Input RGB



Hierarchical Feature Grid
Tri-linear Interpolation
$o^0_{\mathbf{p}}$
Coarse-level Occupancy
Coarse Level
$\phi^0_\theta(\mathbf{p})$
Mapping
$o^1_{\mathbf{p}}$
Fine Level Occupancy
$o_{\mathbf{p}}$
Mid Level
$\phi^1_\theta(\mathbf{p})$
$\Delta o^1_{\mathbf{p}}$
$\phi^1_\theta(\mathbf{p})$
Fine Level
$\phi^2_\theta(\mathbf{p})$
Tracking
$\psi_\omega(\mathbf{p})$
$\mathbf{p}$
Ray –> Point Sampler
Camera Pose

Results

iMAP*

(our re-implementation of iMAP)

NICE-SLAM

4x Speed

Predicted Poses

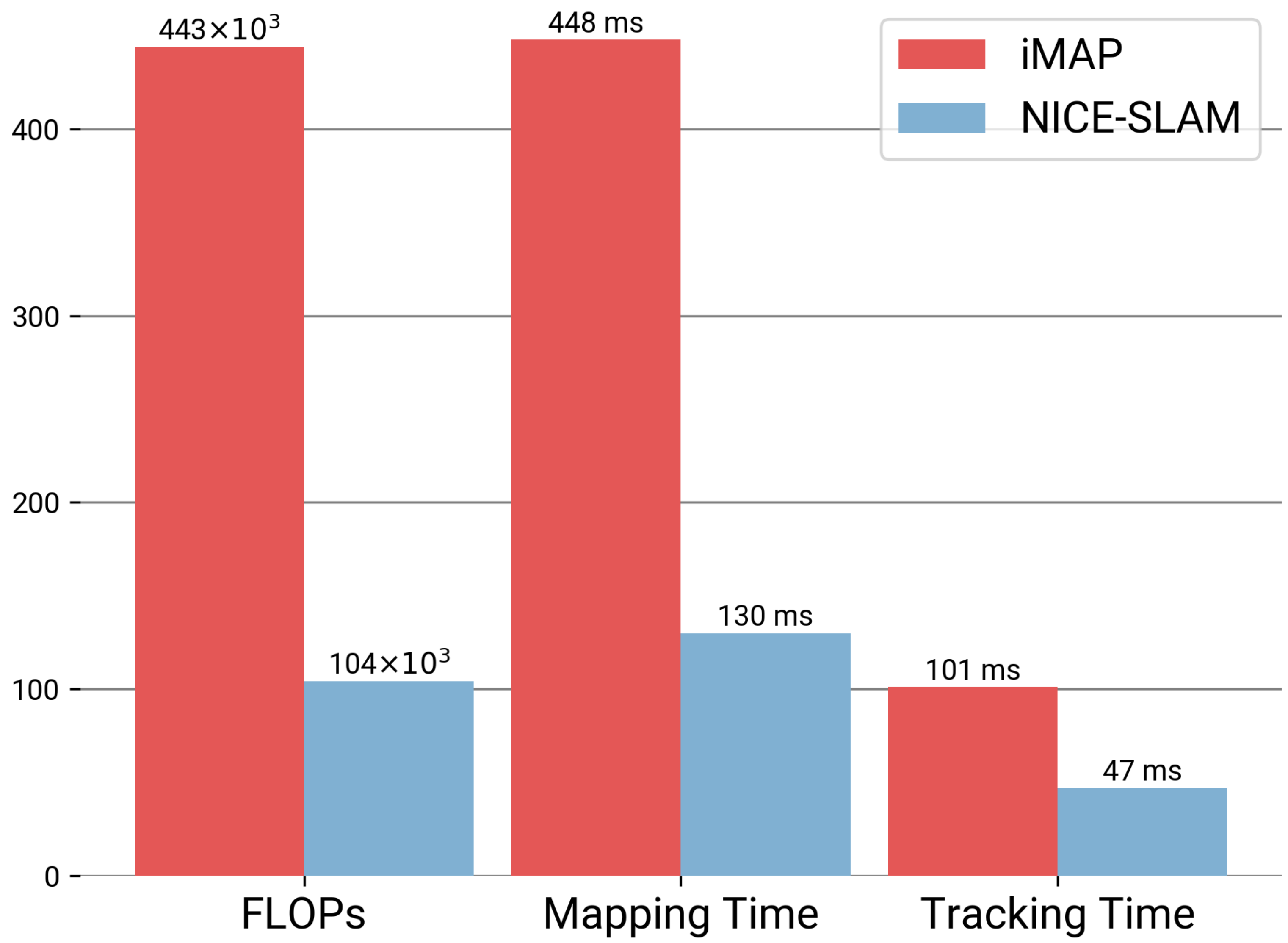
GT Poses

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 [TOG'22]

Limitations

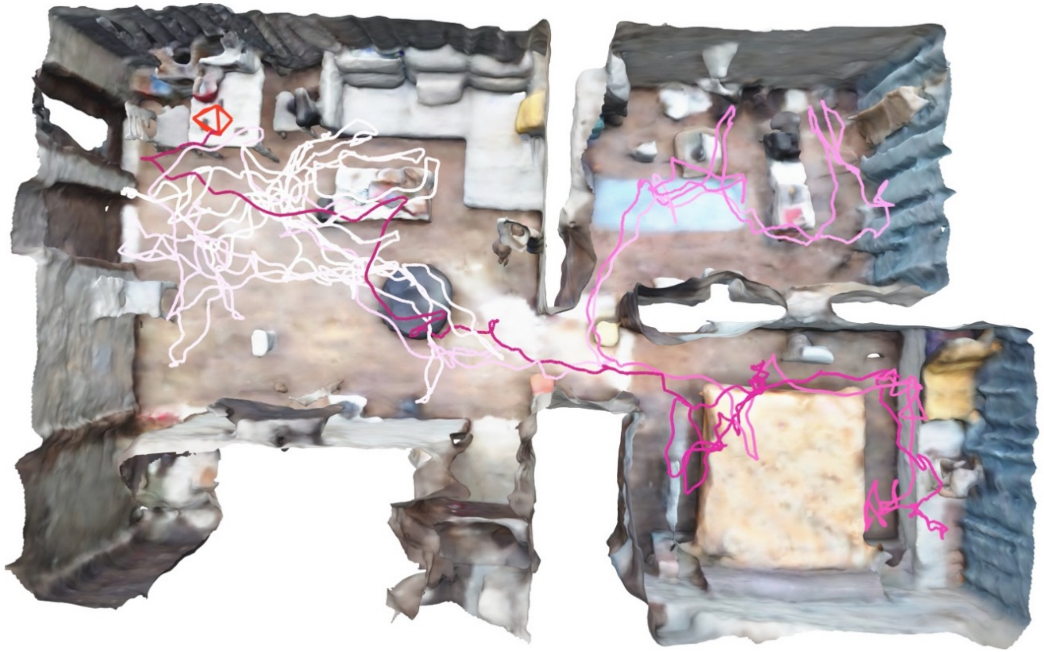
- Requires depths as input
- Only bounded scenes
- Still not real-time

Final Remarks

- NeRF-based multi-view surface reconstruction still has rooms to improve
- A completely COLMAP-free NeRF pipeline?
- What is THE representation?

Large-scale Scene Reconstruction with NeRF

MonoSDF
github.com/autonomousvision/monosdf



NICE-SLAM
github.com/cvg/nice-slam

Thank you!