2D Magic in a 3D World

Songyou Peng

Imperial College London

Mar 22, 2024

Who Am I?

- Senior Researcher — ETH zürich
- Incoming Research Scientist — Google Research

- Earned my PhD — ETH zürich
  - Marc Pollefeys
  - Andreas Geiger

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

- Internships during PhD
  - 2021: Michael Zollhoefer — Meta
  - 2022: Tom Funkhouser — Google Research

pengsongyou.github.io

Research Overview of My PhD

Learn to **Reconstruct** and **Understand** 3D World

ConvOccNet
ECCV 2020 (Spotlight)

MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

NICE-SLAM
CVPR 2022

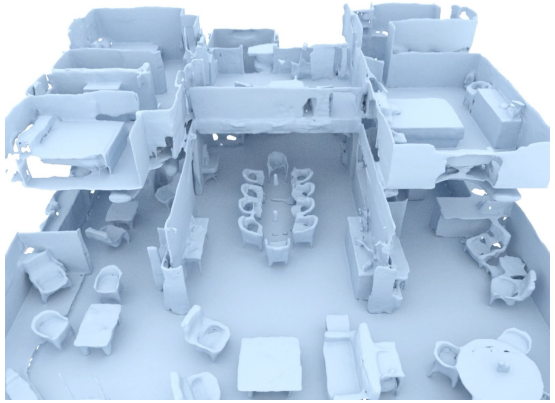
NICER-SLAM
3DV 2024 (Oral)

UNISURF
ICCV 2021 (Oral)

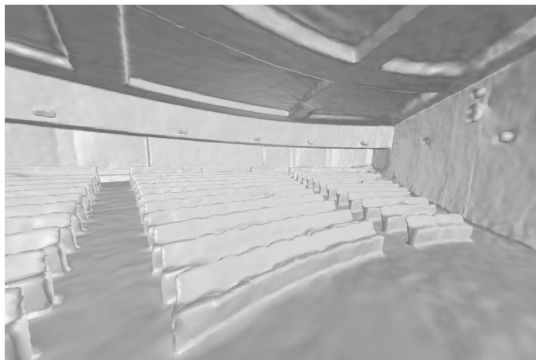
OpenScene
CVPR 2023

Research Overview of My PhD

Learn to Reconstruct and Understand 3D World

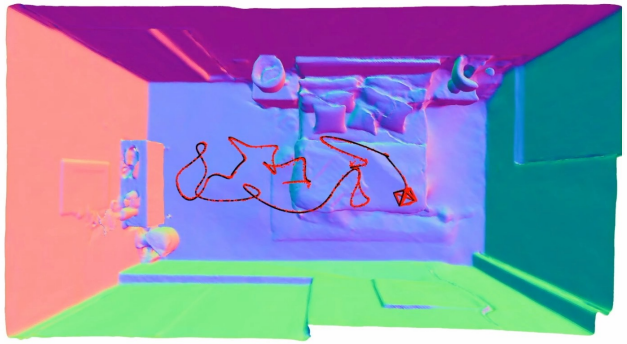


ConvOccNet
ECCV 2020 (Spotlight)



MonoSDF
NeurIPS 2022

NICE-SLAM
CVPR 2022



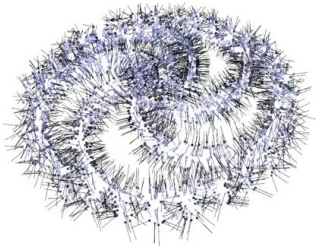
NICER-SLAM
3DV 2024 (Oral)

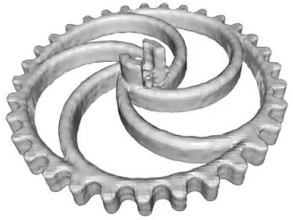
Topic #1:
Reconstruct Complex Scenes

Research Overview of My PhD

Learn to Reconstruct and Understand 3D World

Topic #2:
Fast Inference



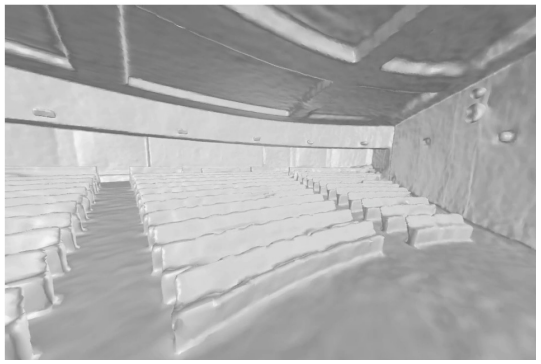


Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

Research Overview of My PhD

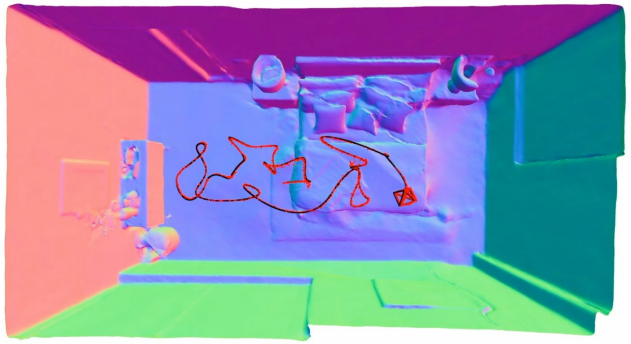
Learn to Reconstruct and Understand 3D World



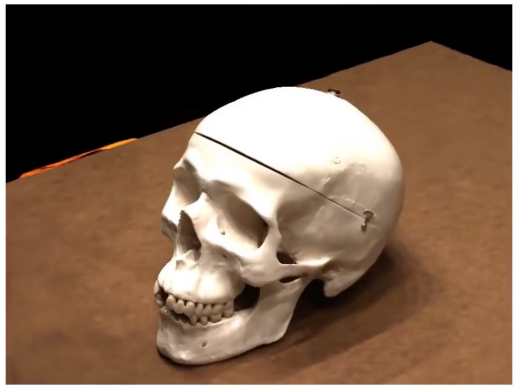
MonoSDF
NeurIPS 2022

Topic #3:
Reconstruct from 2D Observations

NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)



UNISURF
ICCV 2021 (Oral)

Research Overview of My PhD

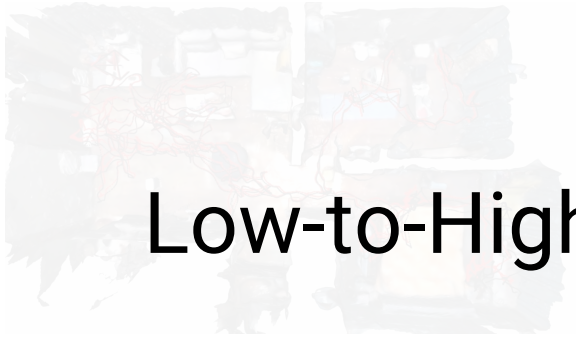
Learn to Reconstruct and Understand 3D World

ConvOccNet
ECCV 2020 (Spotlight)

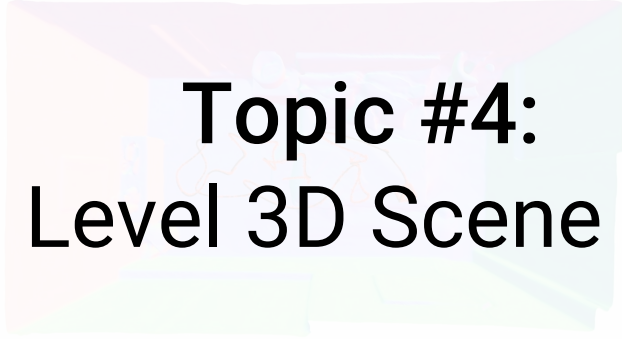
MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021



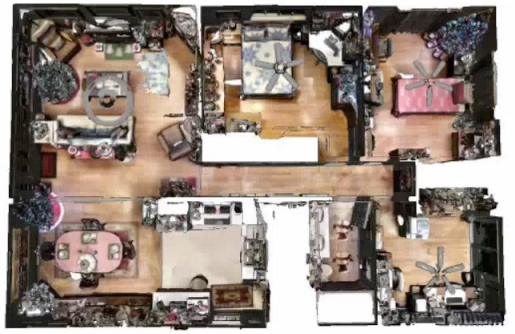
NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)

UNISURF
ICCV 2021 (Oral)

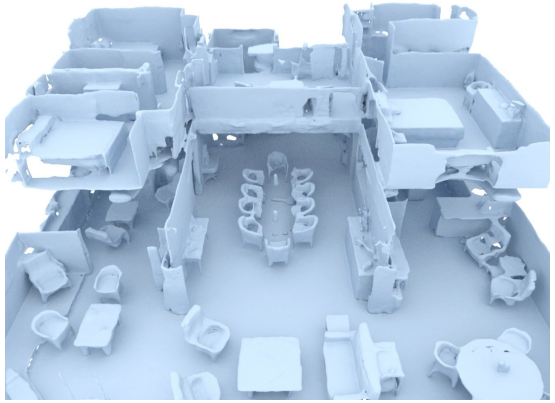
Topic #4:
Low-to-High Level 3D Scene Understanding



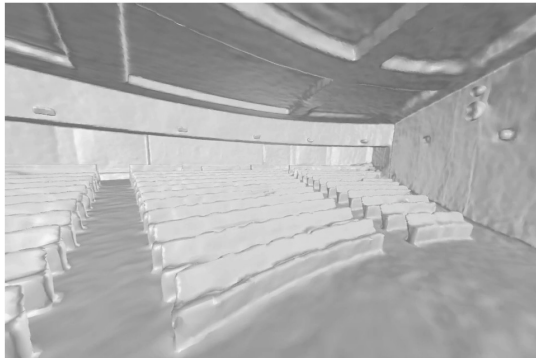
OpenScene
CVPR 2023

Research Overview of My PhD

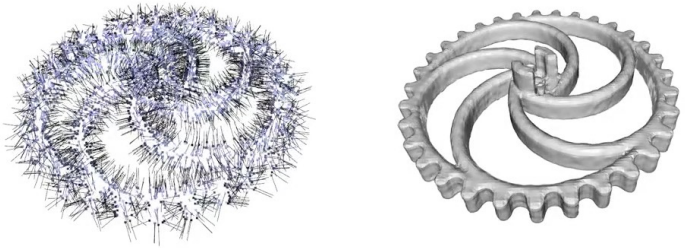
Learn to Reconstruct and Understand 3D World



ConvOccNet
ECCV 2020 (Spotlight)



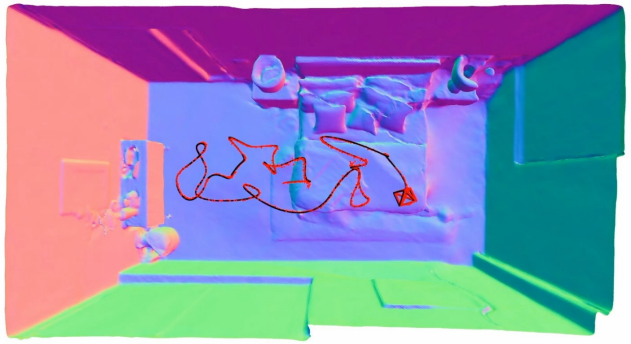
MonoSDF
NeurIPS 2022



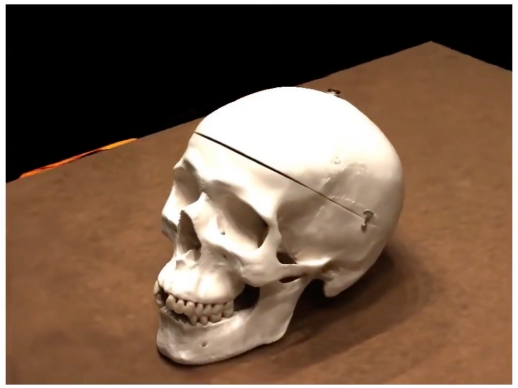
Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

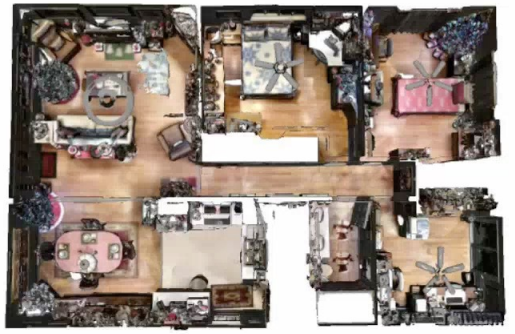
NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)



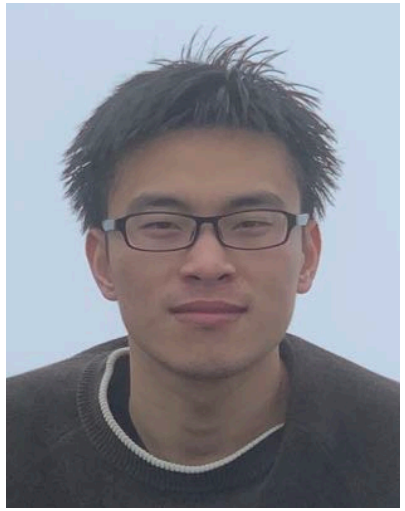
UNISURF
ICCV 2021 (Oral)



OpenScene
CVPR 2023

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

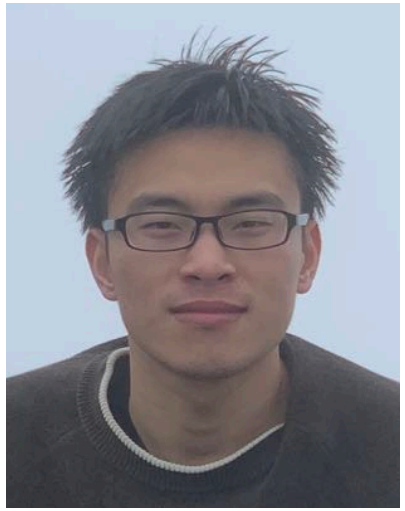
ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

NEURAL INFORMATION PROCESSING SYSTEMS

MonoSDF: Exploring **Monocular Geometric Cues** for Neural Implicit Surface Reconstruction


Zehao Yu


Songyou Peng


Michael Niemeyer


Torsten Sattler


Andreas Geiger

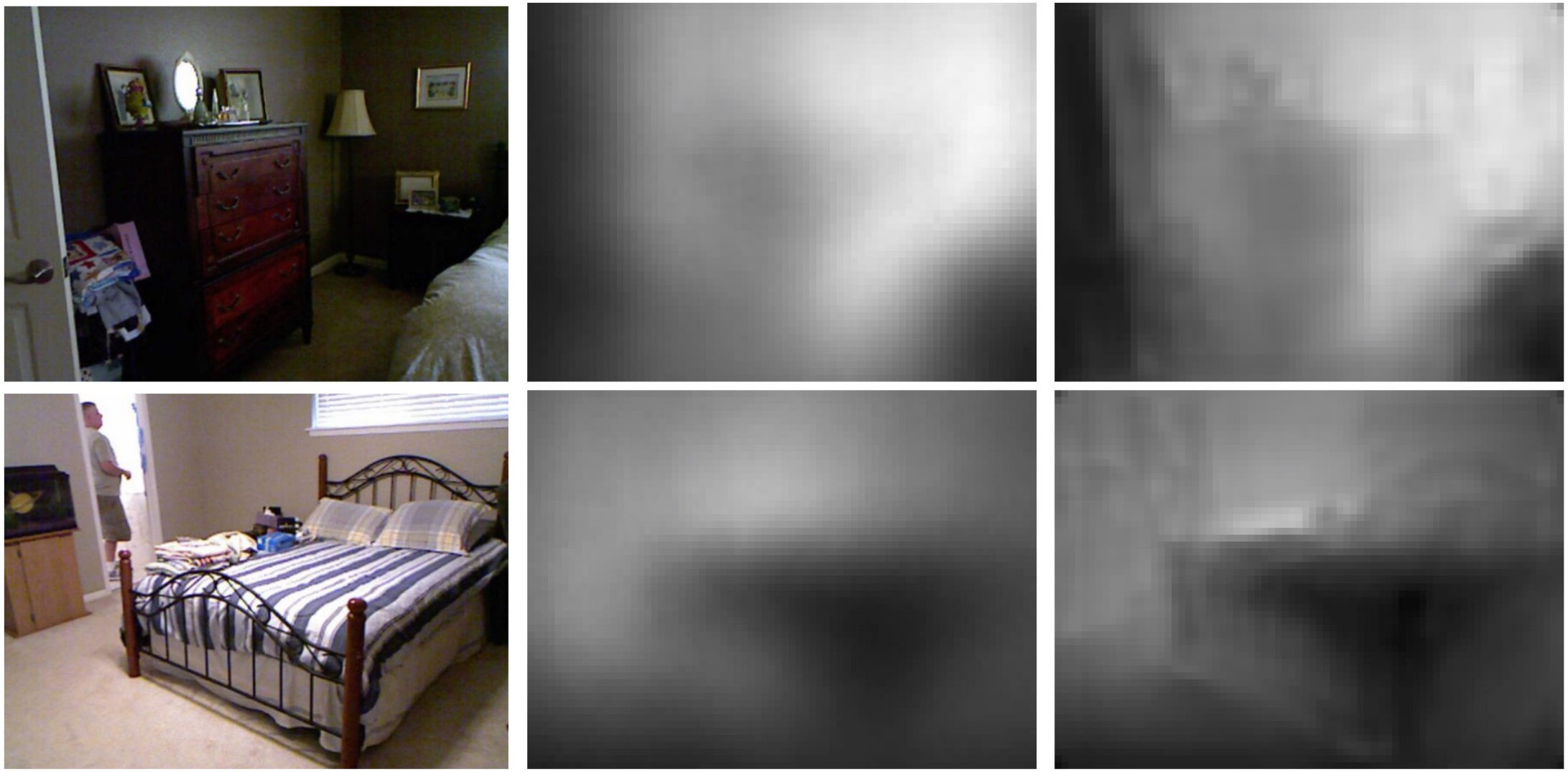
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

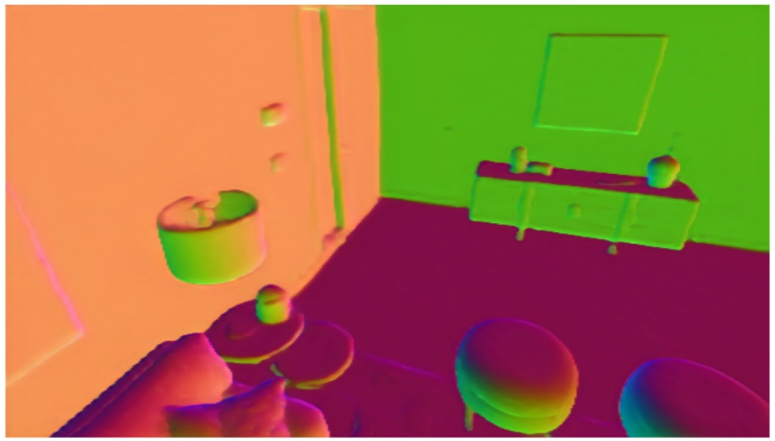
Depth Prediction from a Single Image

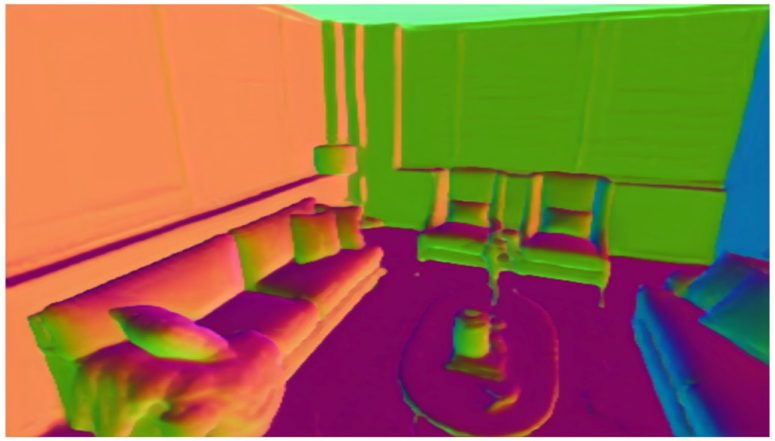


Eigen, Puhrsch and Fergus: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. NIPS, 2014

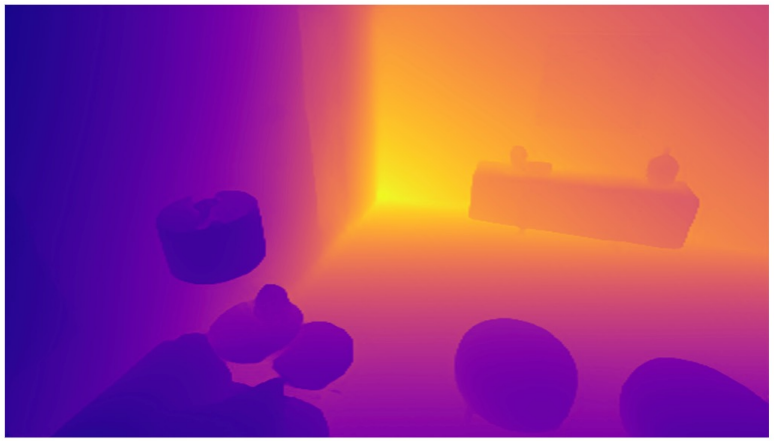
Omnidata

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.

2D Magic in a 3D World

Songyou Peng

Imperial College London

Mar 22, 2024

2D Magic in a 3D World

2D Magic in a 3D World

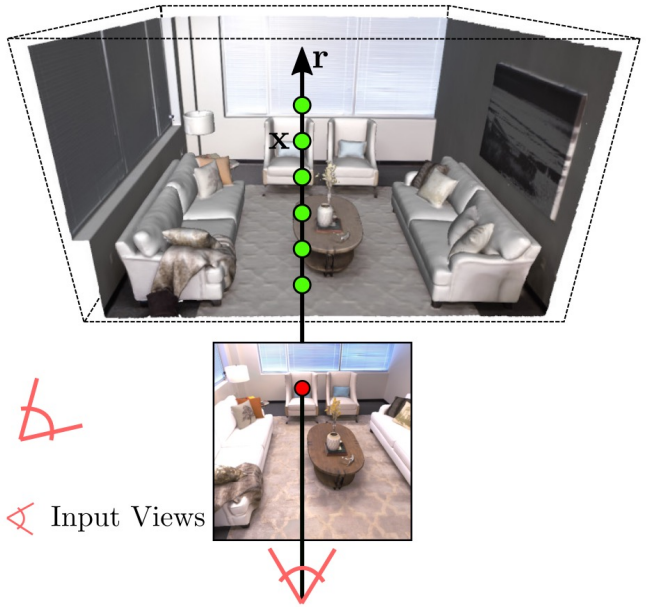
2D Monocular Cues Benefit 3D Reconstruction

3D Reconstruction Pipeline

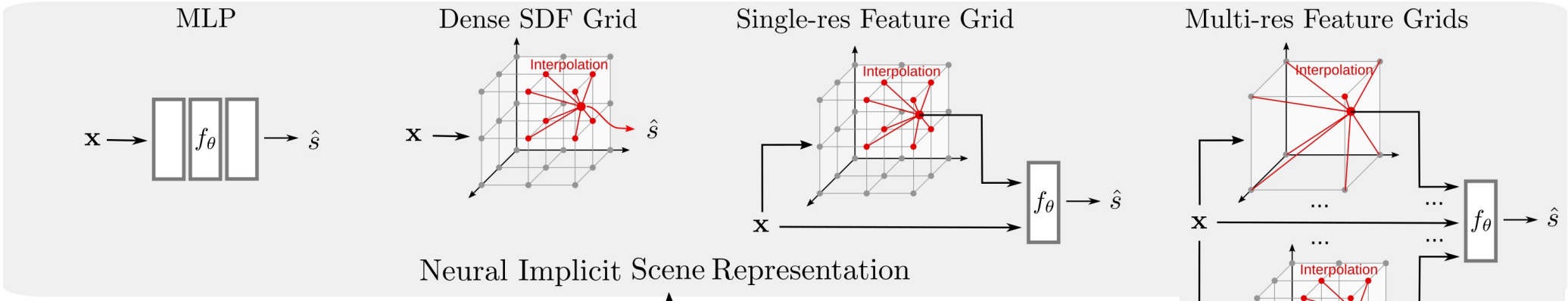
3D Reconstruction Pipeline

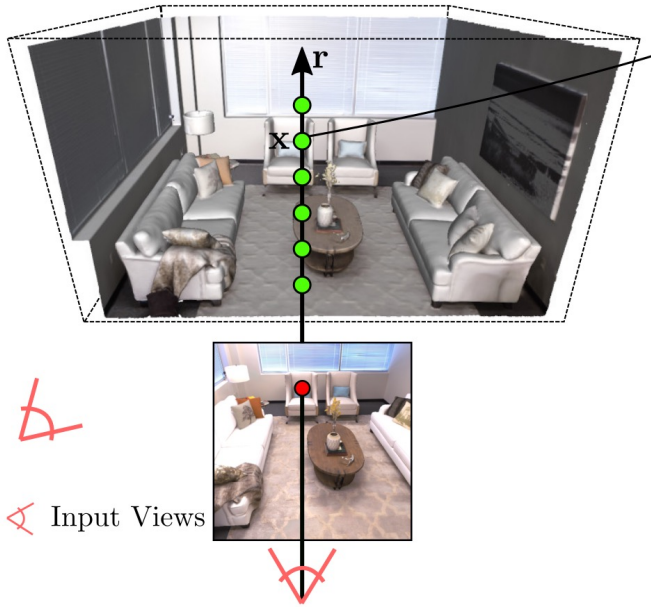
Input Views

3D Reconstruction Pipeline

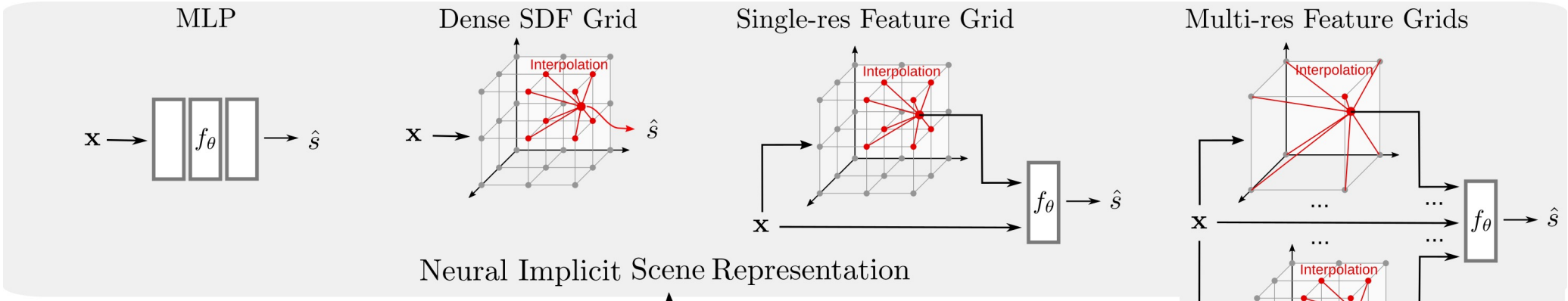


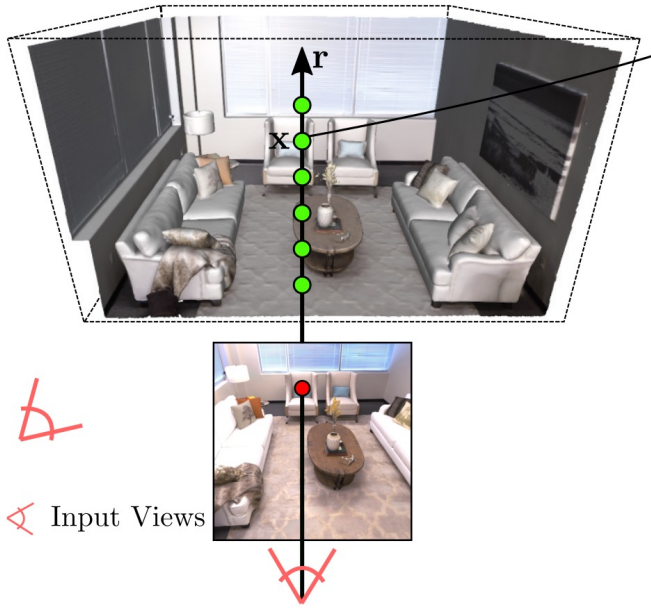
3D Reconstruction Pipeline

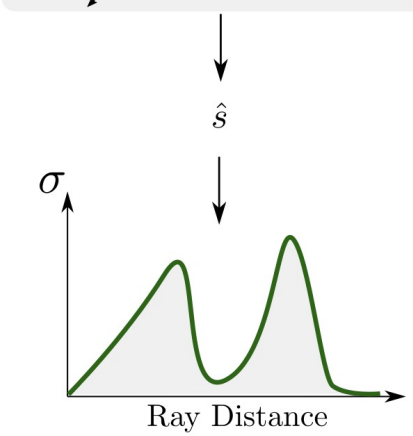




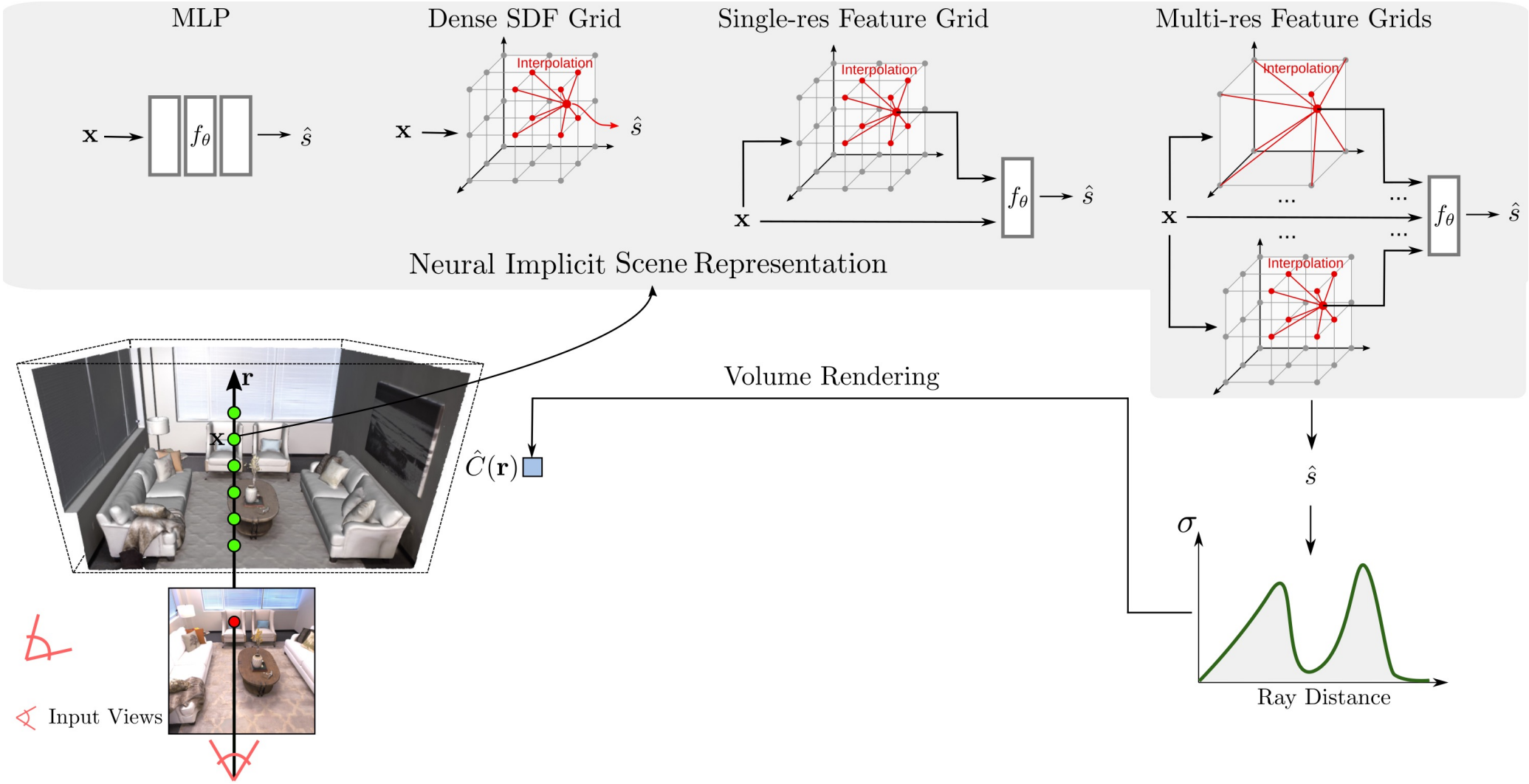
3D Reconstruction Pipeline





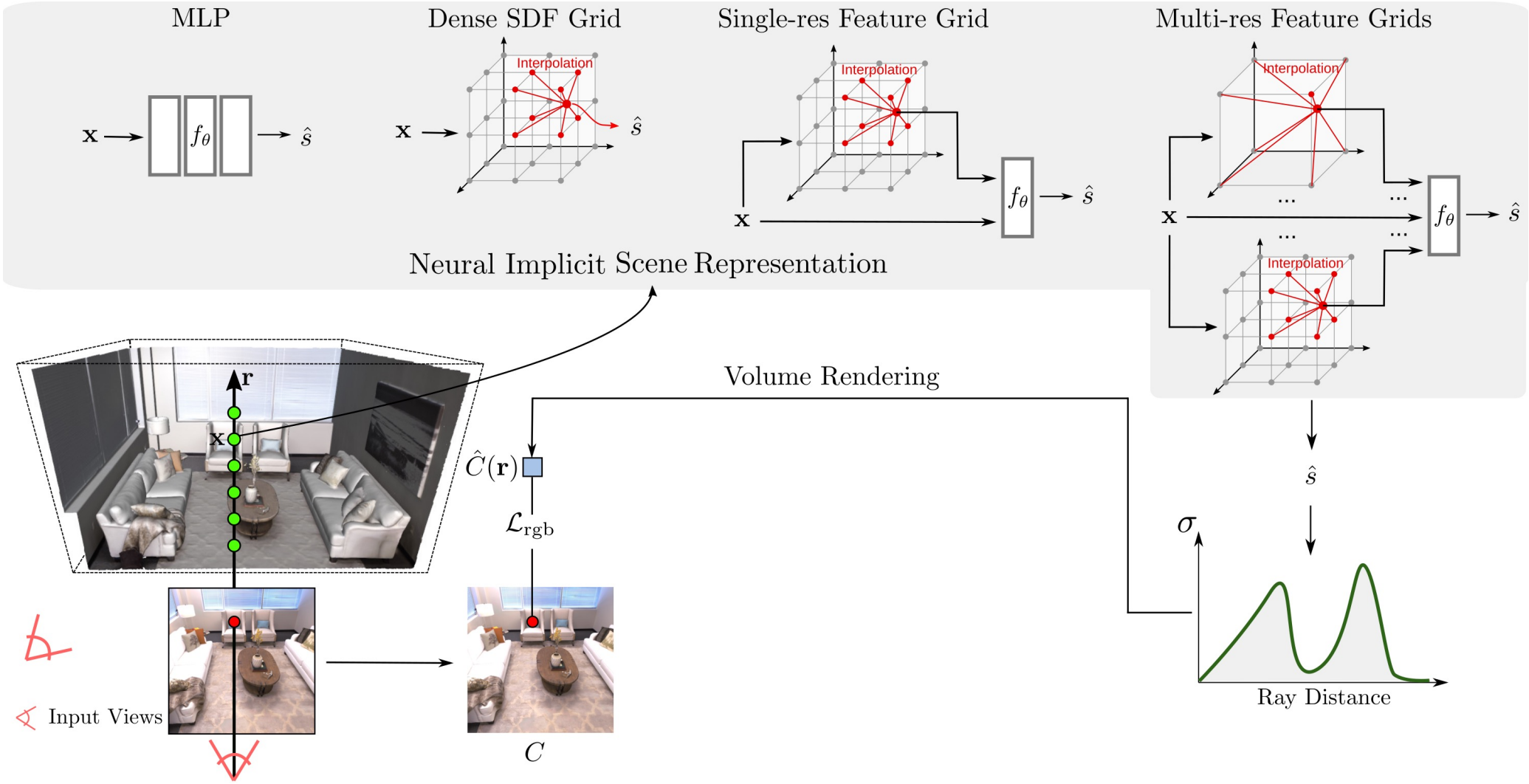


3D Reconstruction Pipeline



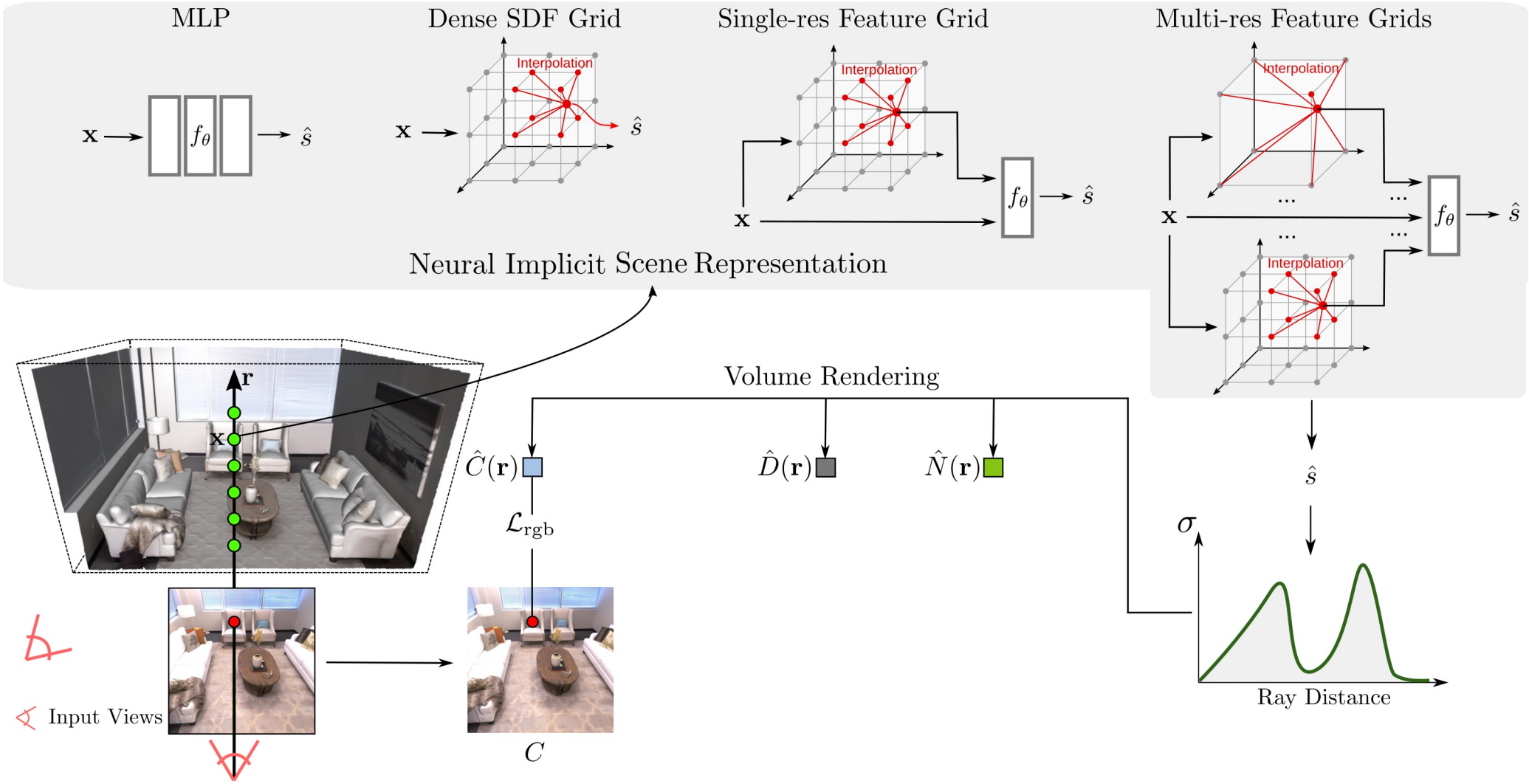
3D Reconstruction Pipeline

VolSDF/NeuS/UNISURF/Neuralangelo



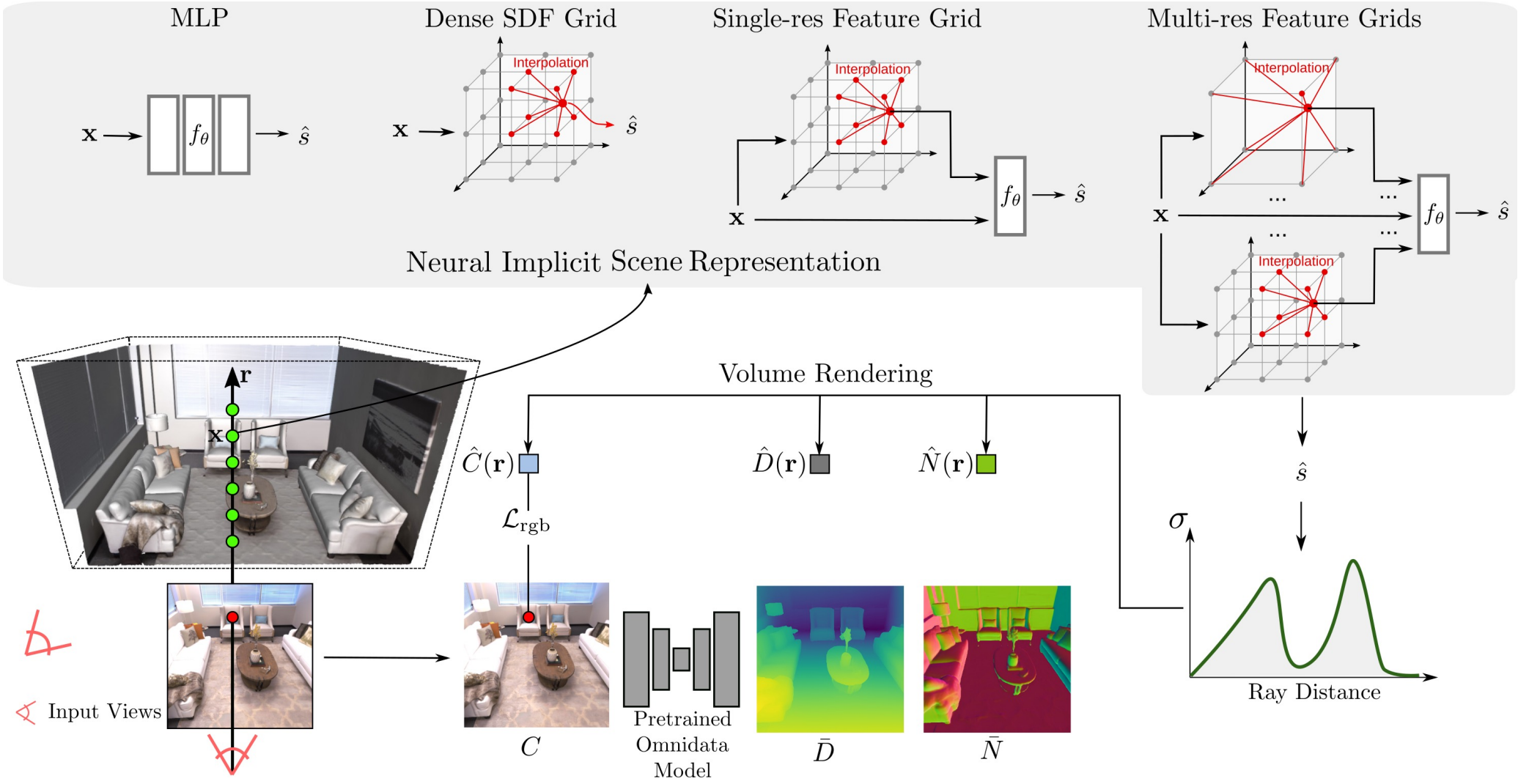
3D Reconstruction Pipeline

MonoSDF



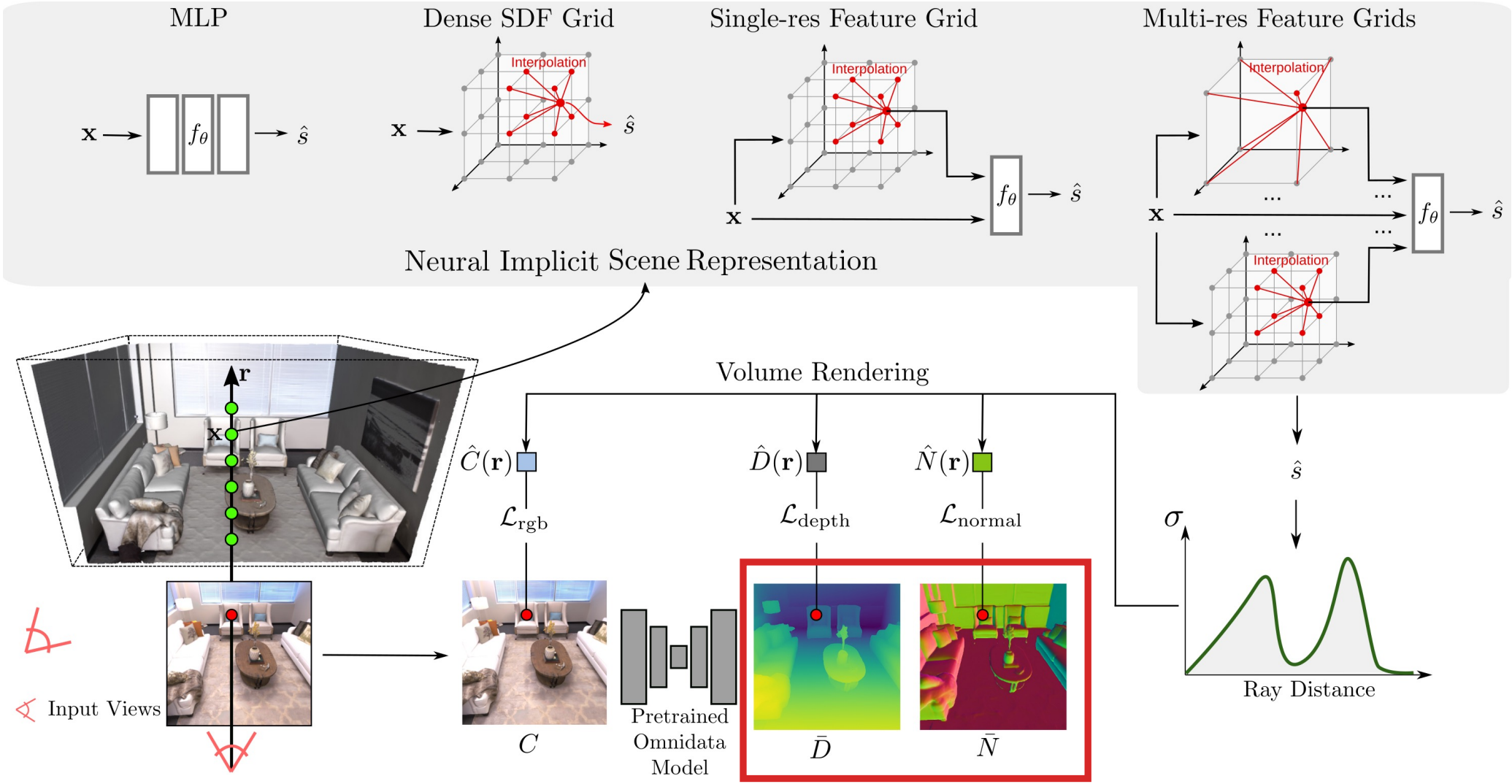
3D Reconstruction Pipeline

MonoSDF



3D Reconstruction Pipeline

MonoSDF



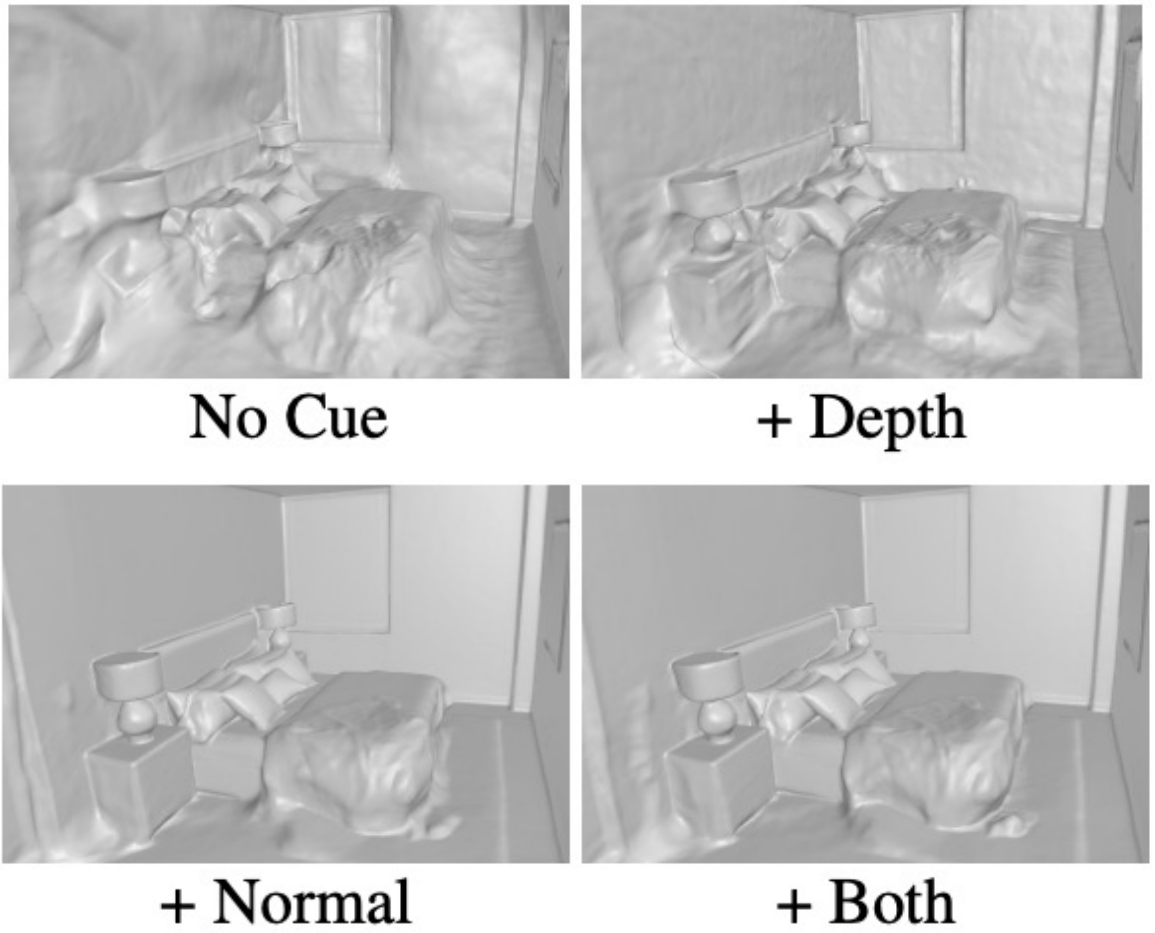
2D Magic in a 3D World

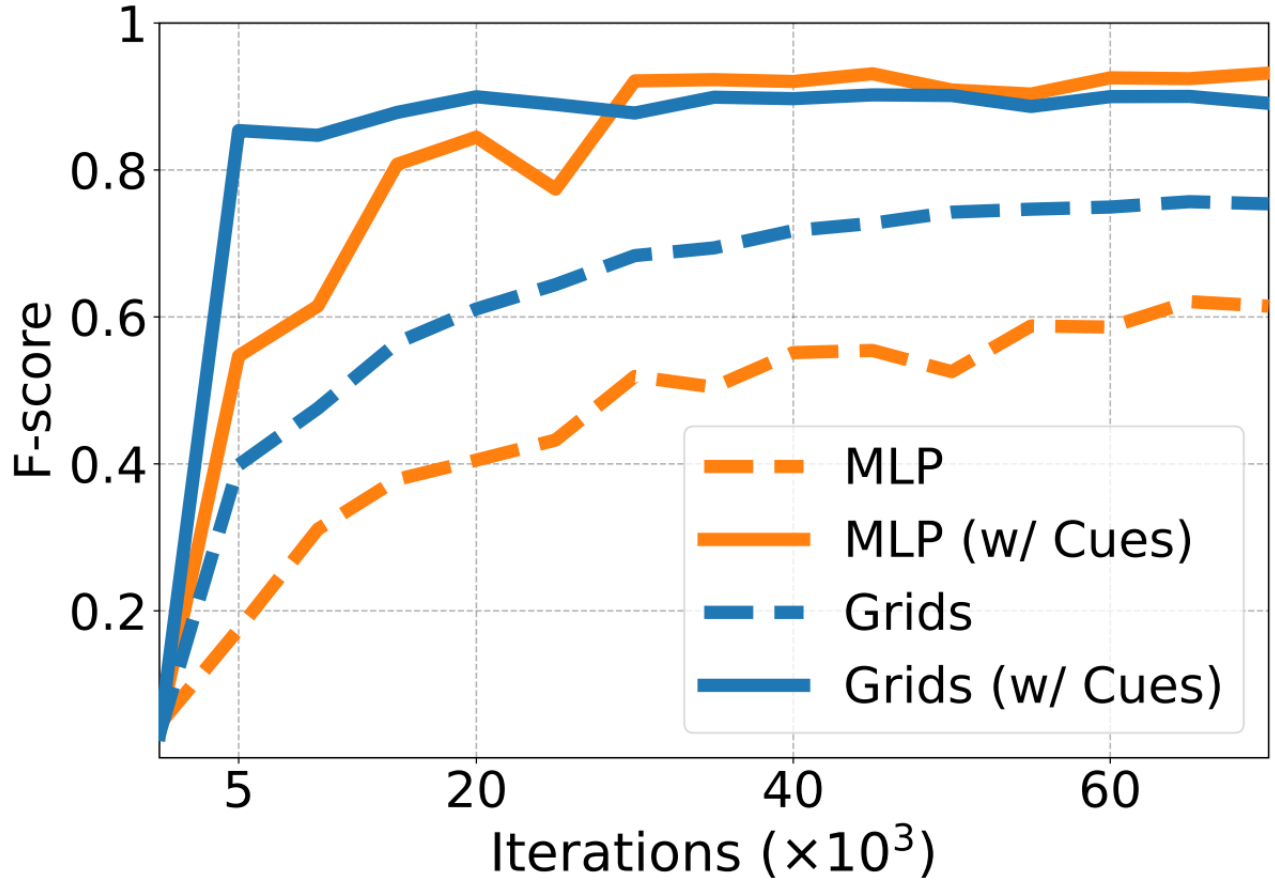
2D Monocular Cues Benefit 3D Reconstruction

Results

Baseline Comparison

Ours


No Cue
+ Depth
+ Normal
+ Both


F-score
1
0.8
0.6
0.4
0.2
5
20
40
60
Iterations (×10^3)
MLP
MLP (w/ Cues)
Grids
Grids (w/ Cues)

Large-Scale 3D Scene Reconstruction

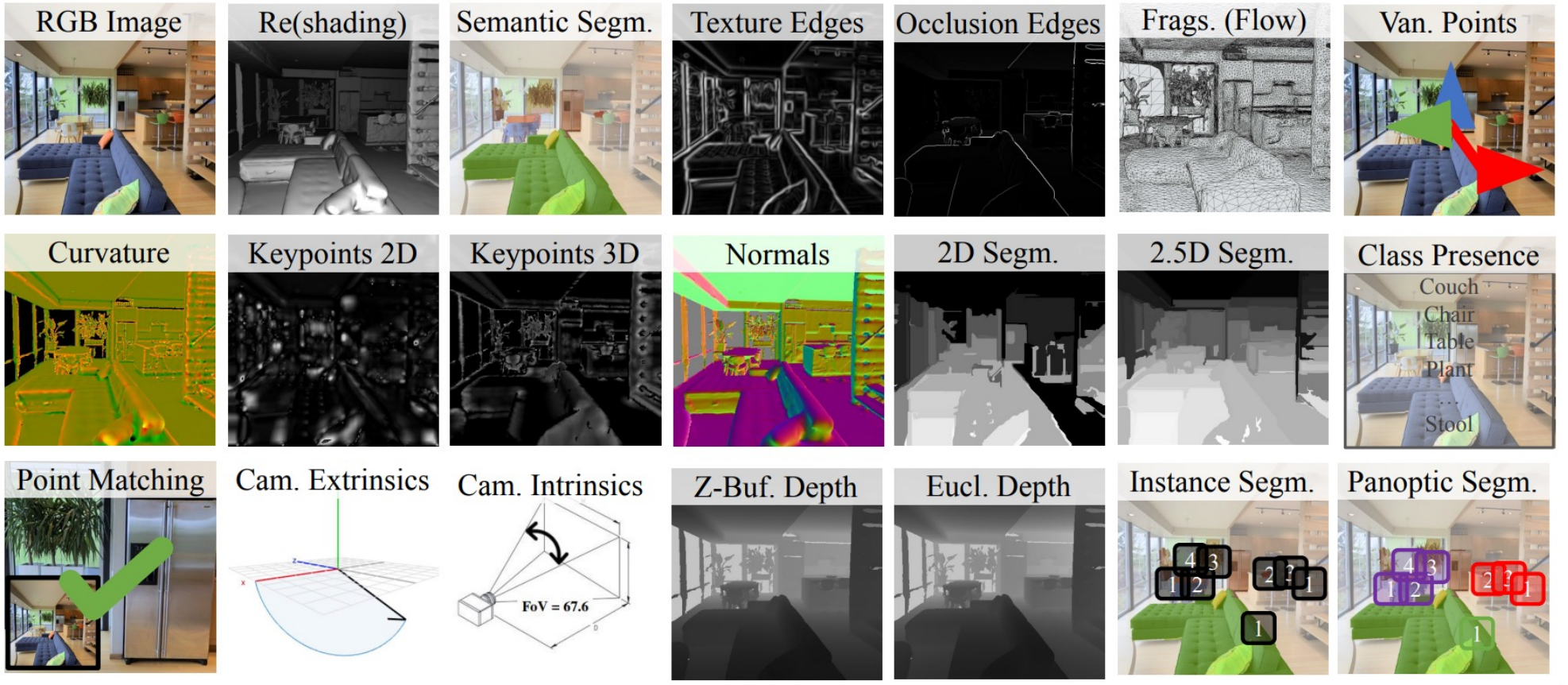
Tanks & Temples Dataset

2D Magic in a 3D World

2D Monocular Cues Benefit 3D Reconstruction

Omnidata

~**14M** Training Images



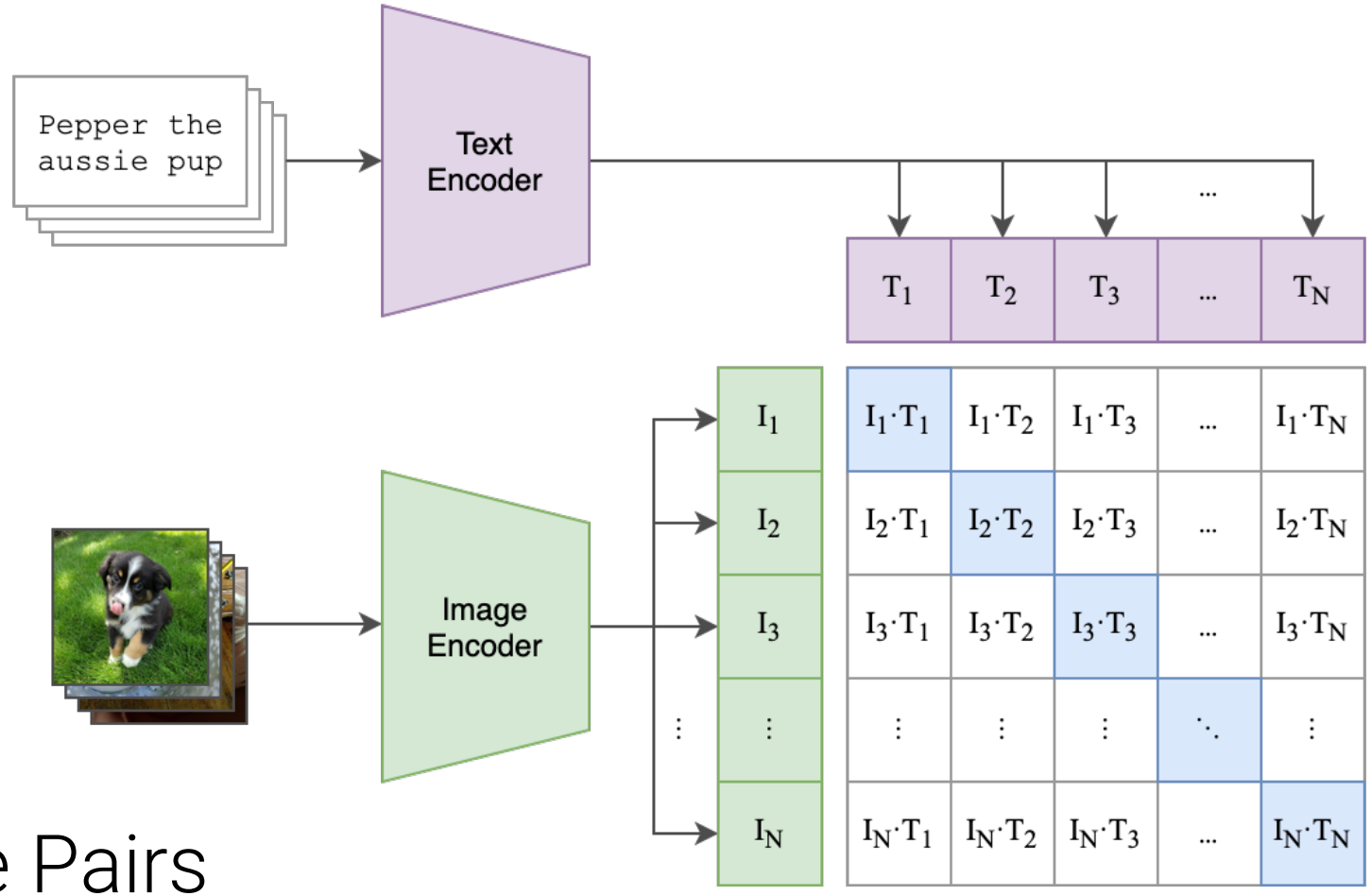
Eftekhar, Sax, Malik and Zamir: Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets from 3D Scans. ICCV, 2021.

What happened since 2021?

Text-Image Pretraining

CLIP

OpenAI



~**12.8B** Training Text-Image Pairs

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

Text-to-Image Generation

DALL·E 3

OpenAI

Imagen

Google Research

Stable Diffusion

stability.ai

L

SO **You**

Generate an image of “a nervous and nerdy guy presenting in front of many smart researchers at a lecture hall at Dyson Robotics Lab in Imperial College London”

ChatGPT

Betker et al.: Improving Image Generation with Better Captions. 2023
Saharia et al.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. NeurIPS 2022
Rombach et al.: High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022

2D Image Segmentation

SAM

Meta

~1B Mask + Human-in-the-loop

Kirillov et al.: Segment Anything. ICCV, 2023

2D Visual Features

DINO v2

Meta

1.2B Training Images

Oquab et al.: DINOv2: Learning Robust Visual Features without Supervision. TMLR, 2024

Text-to-Video Generation

Sora

OpenAI

??? B Training Videos?

Prompt: A movie trailer featuring the adventures of the 30 year old space man wearing a red wool knitted motorcycle helmet, blue sky, salt desert, cinematic style, shot on 35mm film, vivid colors.

2D Magic in a 3D World

Songyou Peng

Imperial College London

Mar 22, 2024

2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

[Generated by DALL·E 3]

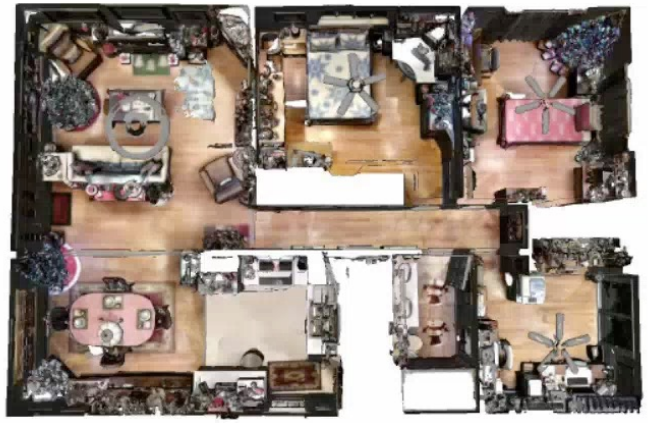
2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

3D Reconstruction

NeRF *On-the-go*
CVPR 2024

3D Scene Understanding



OpenScene
CVPR 2023

Segment3D
arXiv 2024

2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

3D Reconstruction

NeRF *On-the-go*
CVPR 2024

3D Scene Understanding

OpenScene
CVPR 2023

Segment3D
arXiv 2024

NeRF Is Awesome

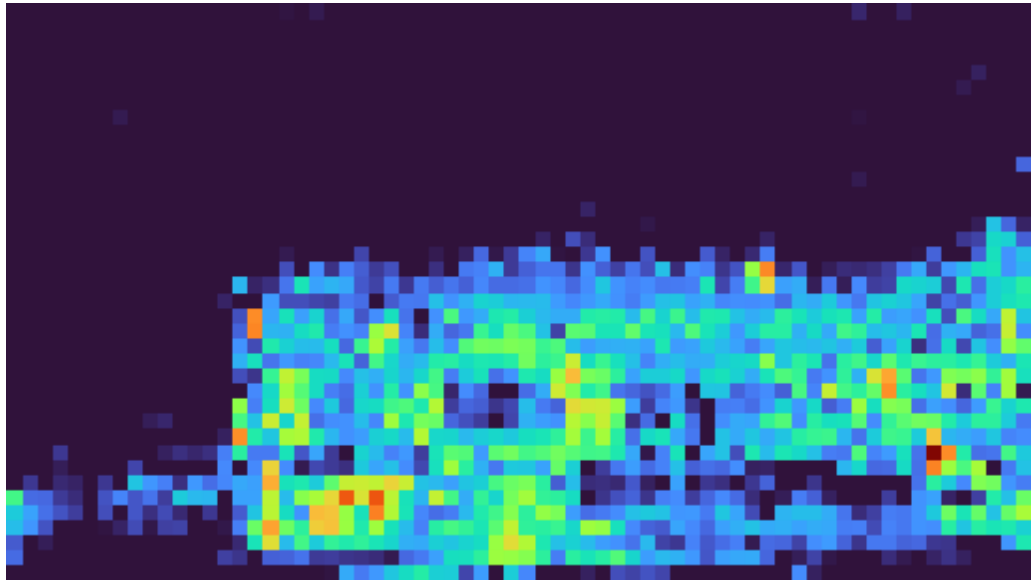

TOUR EIFFEL · CHAMP-DE-MARS · MUSÉE DU LOUVRE · NOTRE-DAME · MUSÉE D'ORSAY · OPERA GARNIER · CHAMPS-ELYSÉES · GRAND PALAIS · TROCADÉRO
BIGBUS PARIS · LESCARSROUGES
THE PARISIAN DEPARTMENT STORE
HOP-ON HOP-OFF

Motivation

How to obtain **distractor-free NeRFs** from **casually captured sequences**?

Uncertainty

Input RGB



Uncertainty Map

How to learn a good uncertainty map?

DINO v2

- A 2D foundation model producing **universal features**
- Preserve temporal-spatial consistency

Oquab et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv, 2023

How to Leverage the **2D Foundation Model** for **Distractor-free NeRF?**




NeRF *On-the-go*

Exploiting Uncertainty for Distractor-free NeRFs in the Wild


Weining Ren*


Zihan Zhu*

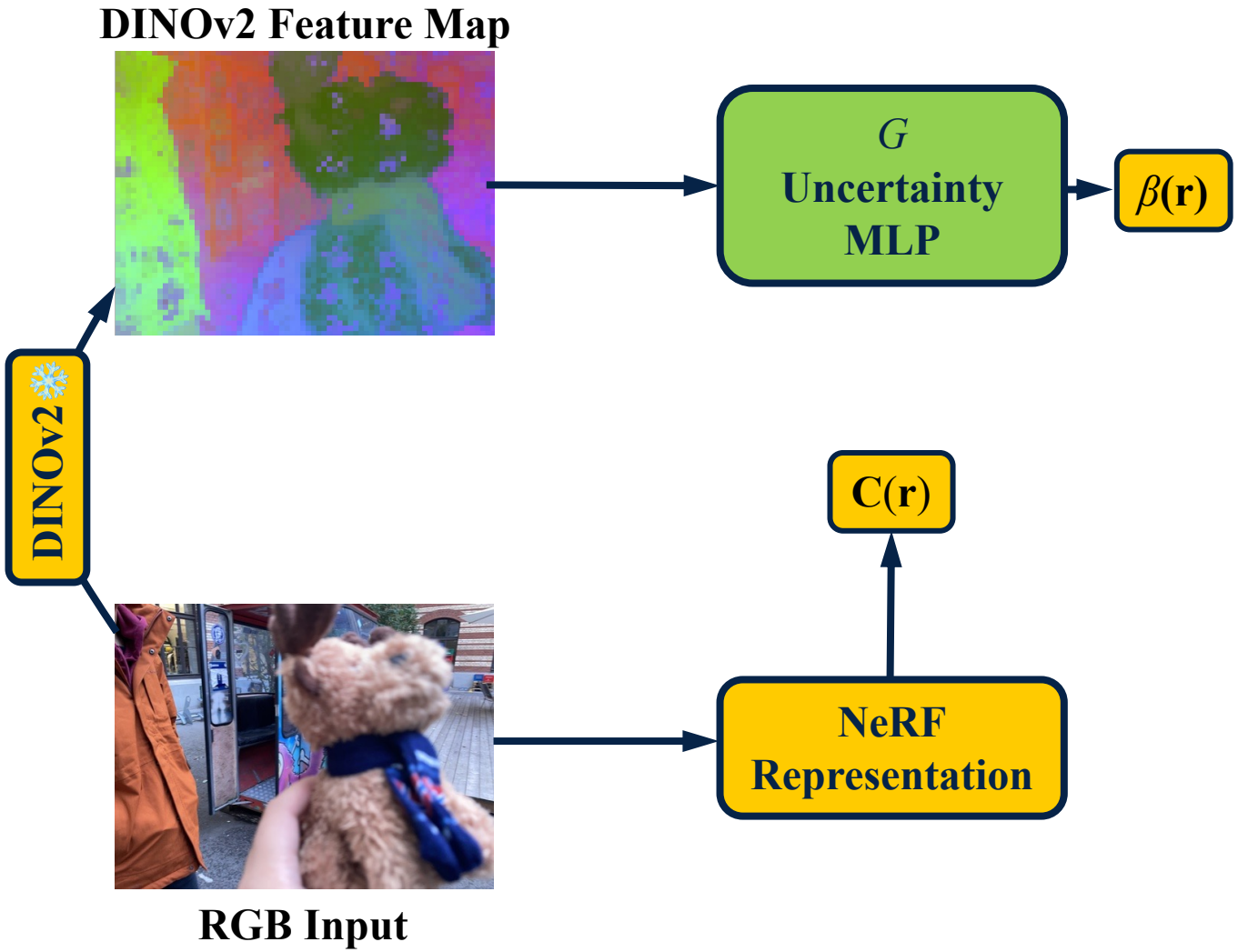

Boyang Sun


Julia Chen

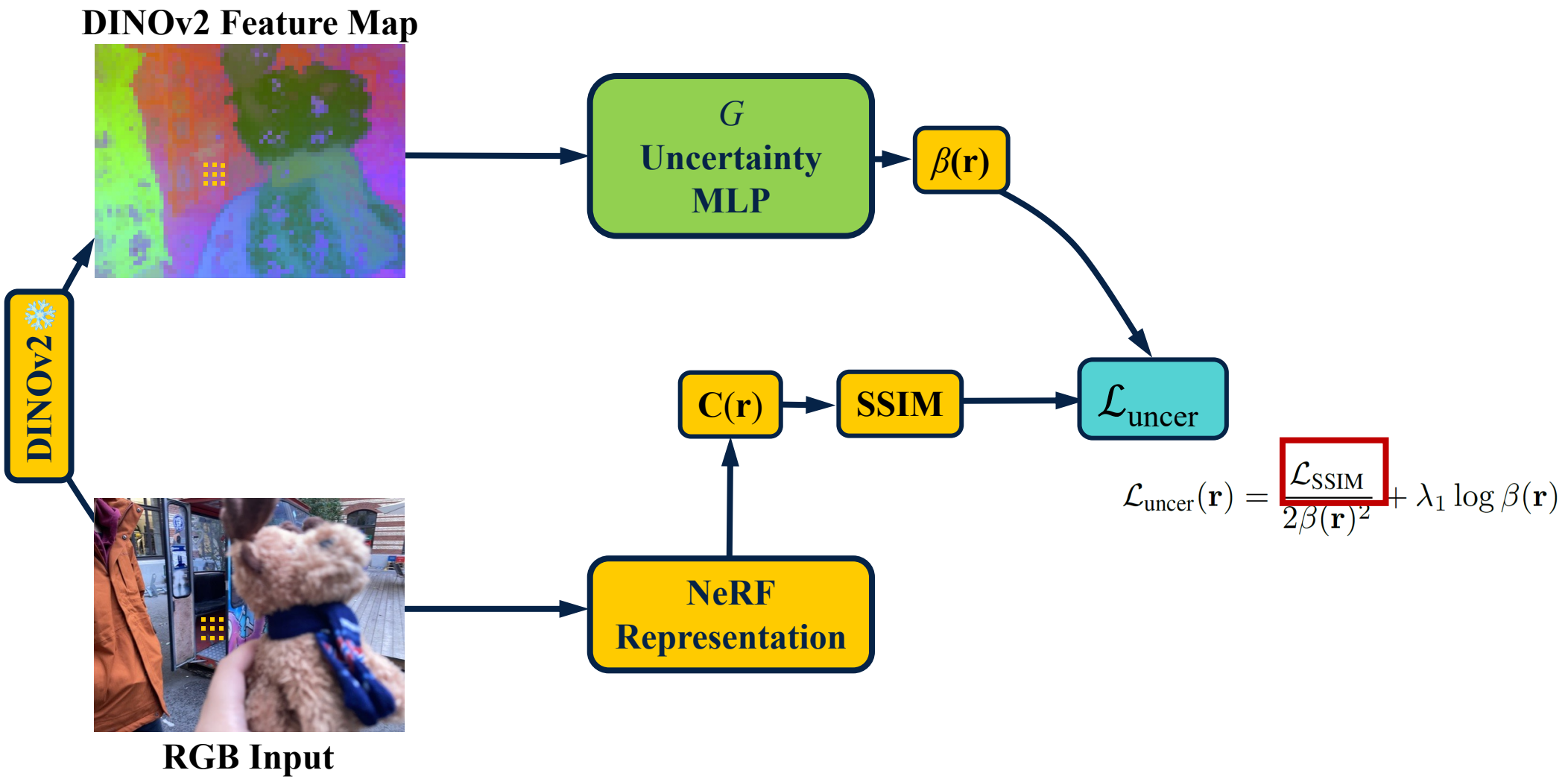

Marc Pollefeys


Songyou Peng

Pipeline

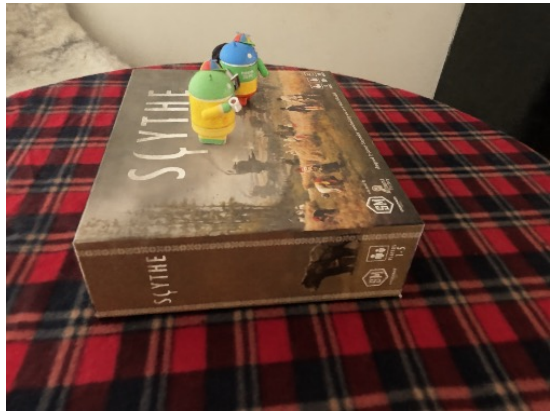

DINOv2 Feature Map
G
Uncertainty
MLP
β(r)
DINOv2
C(r)
NeRF
Representation
RGB Input

Pipeline



$$\mathcal{L}_{\text{uncer}}(\mathbf{r}) = \frac{\mathcal{L}_{\text{SSIM}}}{2\beta(\mathbf{r})^2} + \lambda_1 \log \beta(\mathbf{r})$$

To Learn the Uncertainty MLP…

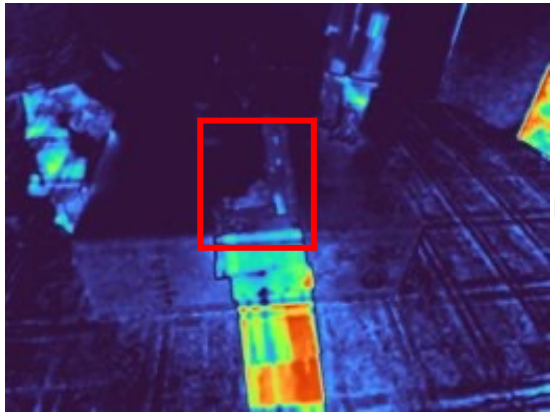


Rendered RGB

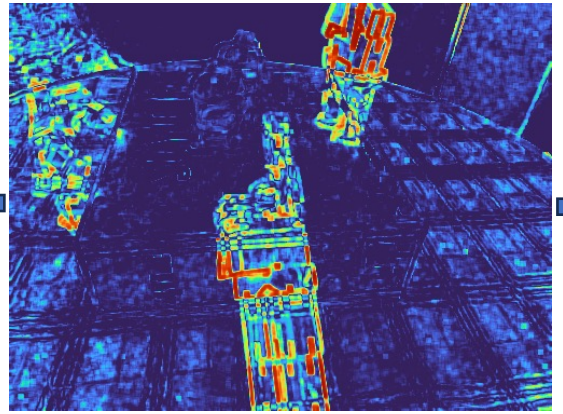
Train RGB

Why SSIM?

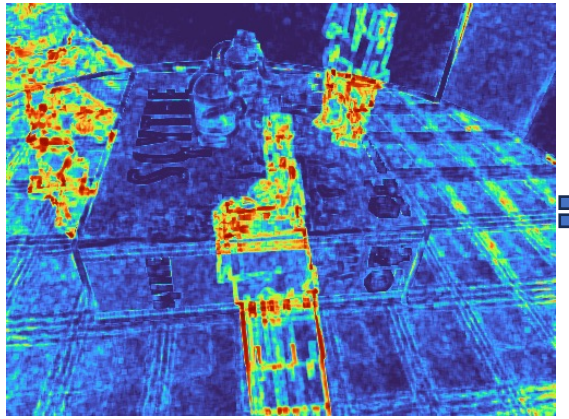
Leverage structure information when RGB is similar!



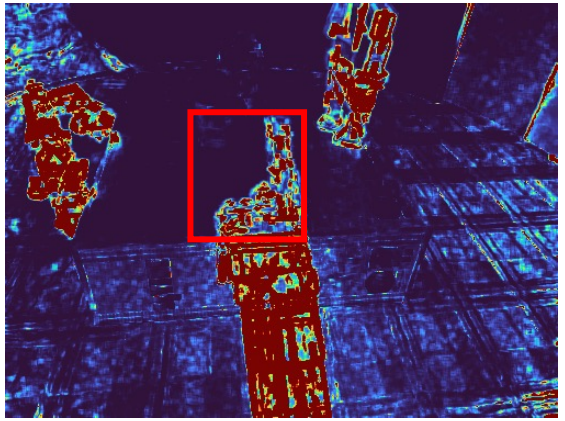
Luminance



Contrast

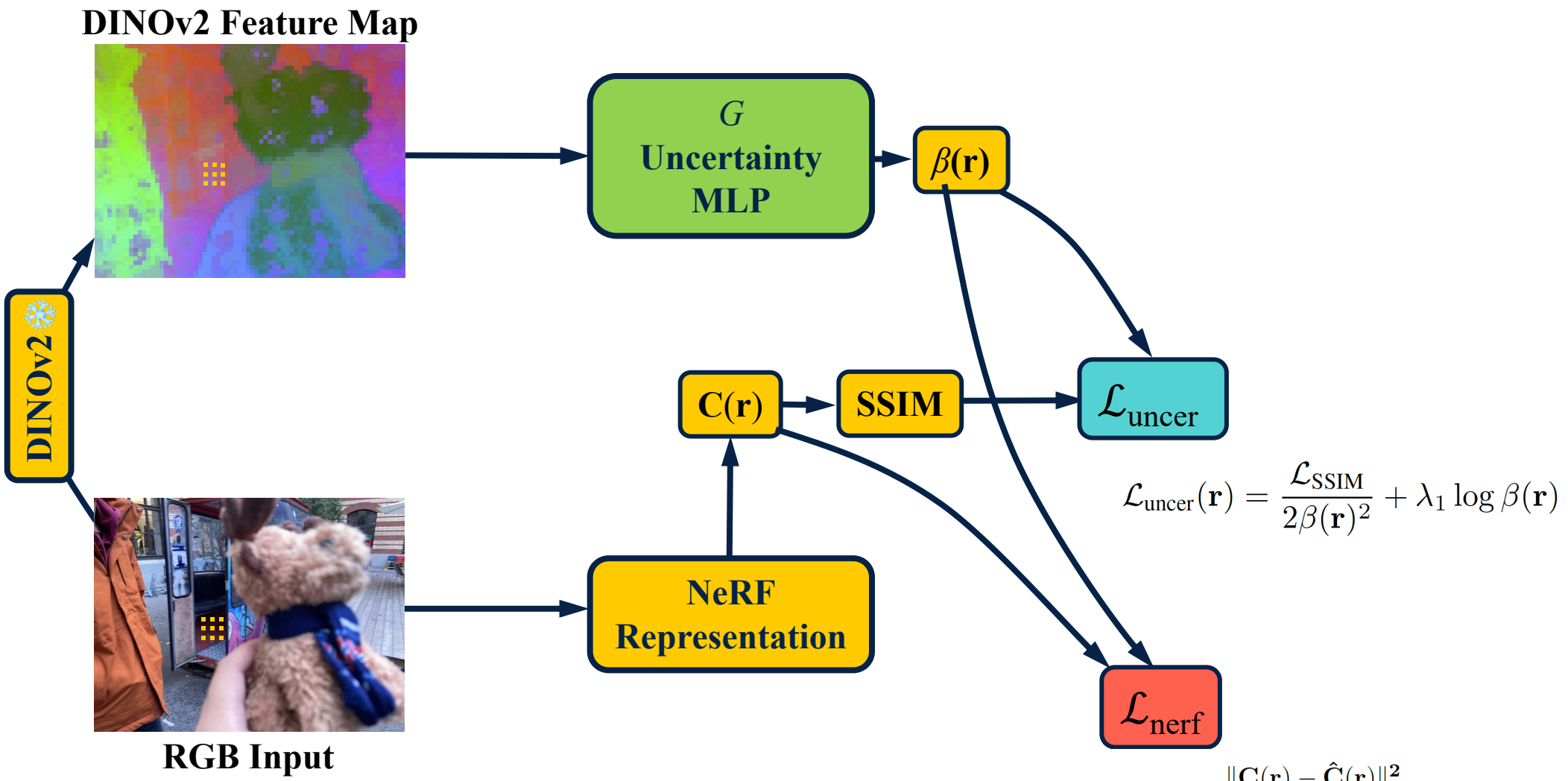


Structure

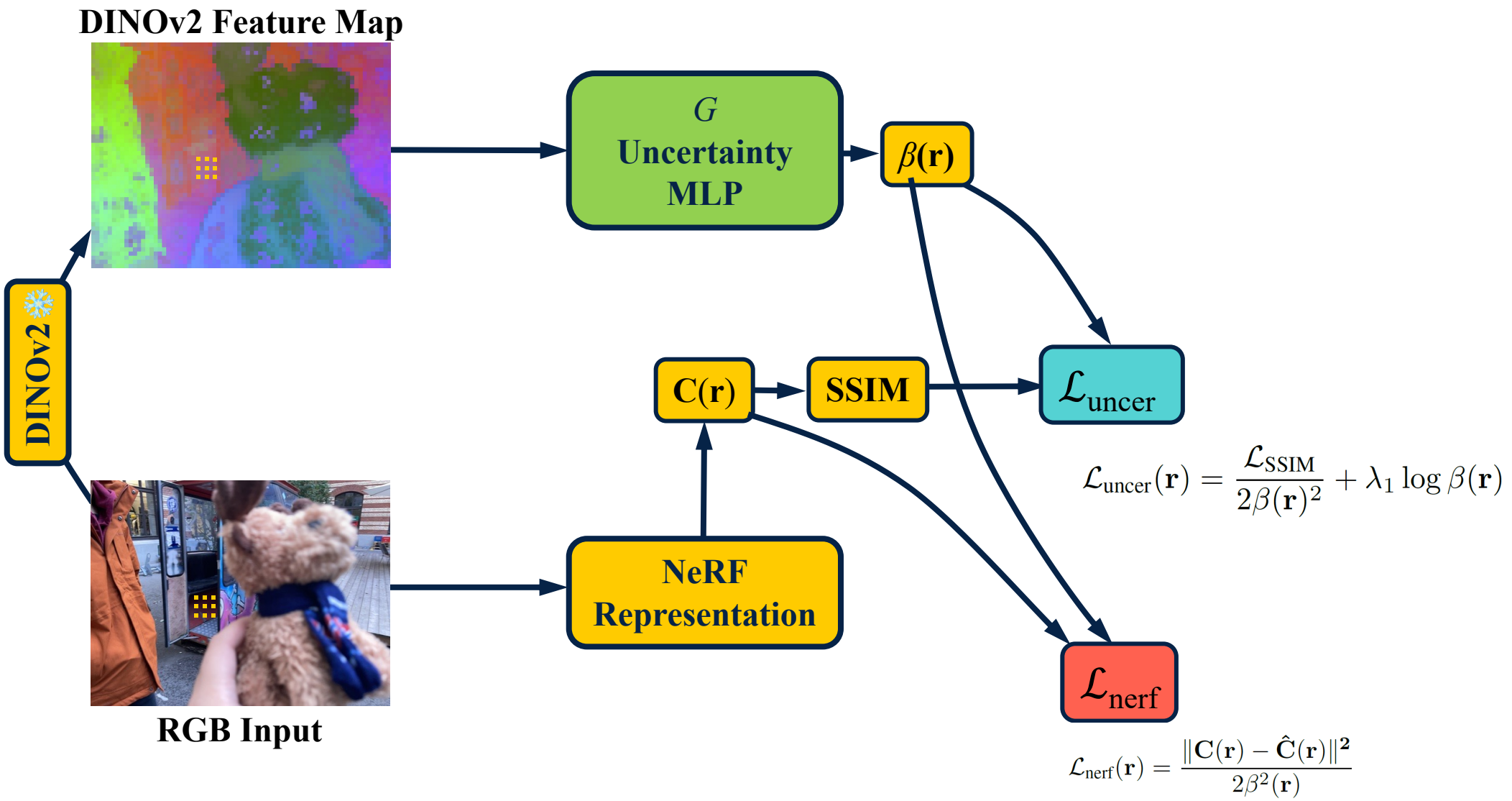


SSIM Error

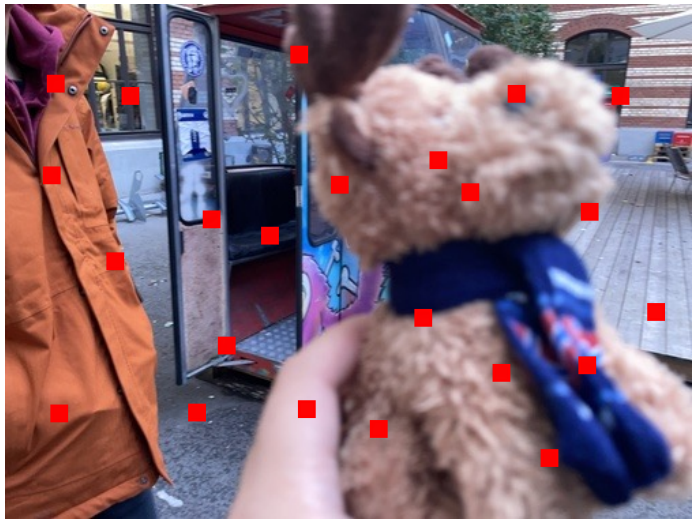
Pipeline


DINOv2 Feature Map
DINOv2
RGB Input
G
Uncertainty
MLP
β(r)
C(r)
SSIM
NeRF
Representation
$\mathcal{L}_{\text{uncer}}$
$\mathcal{L}_{\text{nerf}}$
$\mathcal{L}_{\text{uncer}}(\mathbf{r}) = \frac{\mathcal{L}_{\text{SSIM}}}{2\beta(\mathbf{r})^2} + \lambda_1 \log \beta(\mathbf{r})$
$\mathcal{L}_{\text{nerf}}(\mathbf{r}) = \frac{\|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|^2}{2\beta^2(\mathbf{r})}$

Pipeline


DINOv2 Feature Map
DINOv2
RGB Input
G
Uncertainty
MLP
β(r)
C(r)
SSIM
NeRF
Representation
$\mathcal{L}_{\text{uncer}}$
$\mathcal{L}_{\text{nerf}}$
$\mathcal{L}_{\text{uncer}}(\mathbf{r}) = \frac{\mathcal{L}_{\text{SSIM}}}{2\beta(\mathbf{r})^2} + \lambda_1 \log \beta(\mathbf{r})$
$\mathcal{L}_{\text{nerf}}(\mathbf{r}) = \frac{\|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|^2}{2\beta^2(\mathbf{r})}$

Sampling Strategy



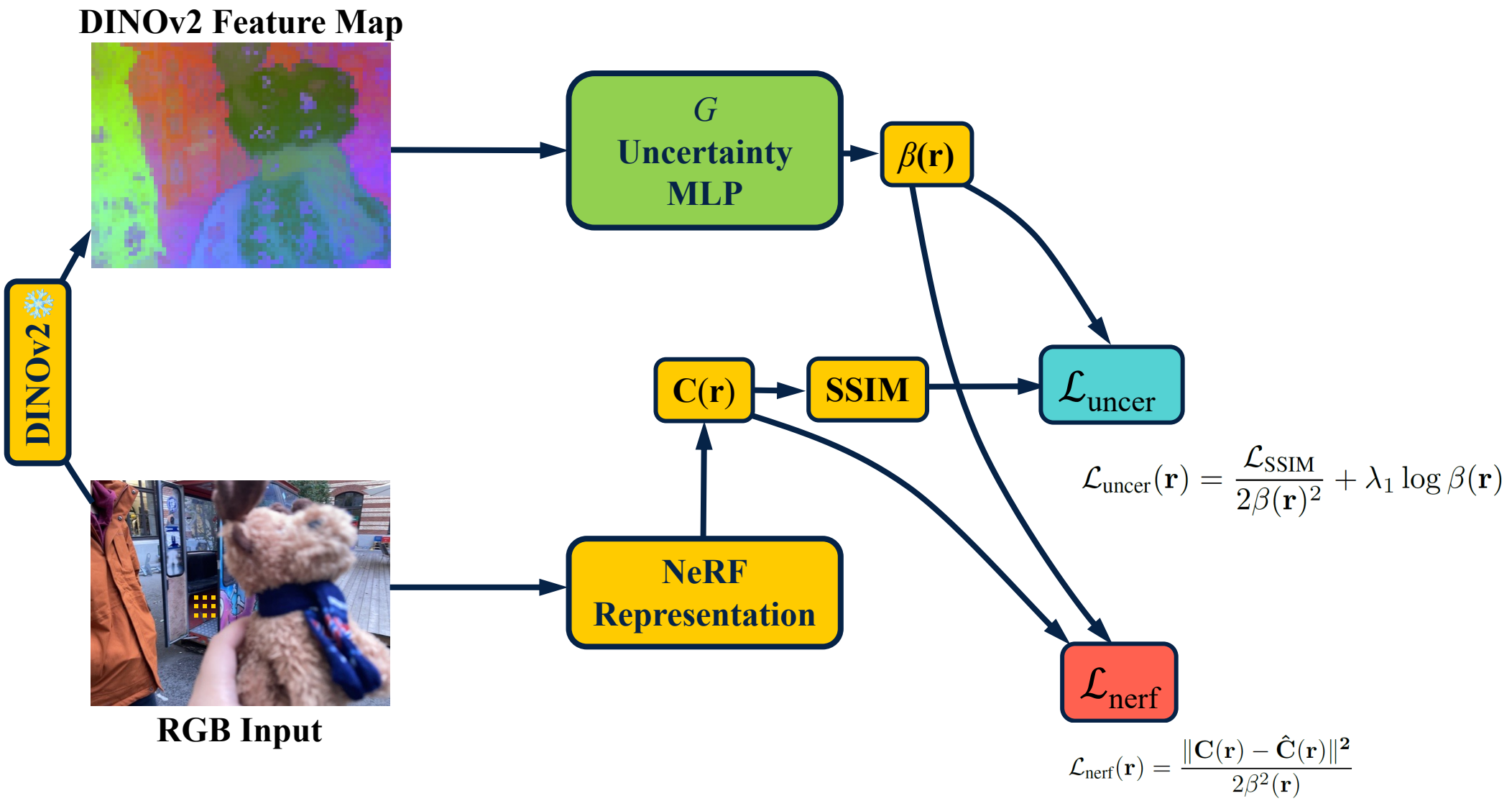
Random
(NeRF)

Patch
(RobustNeRF)

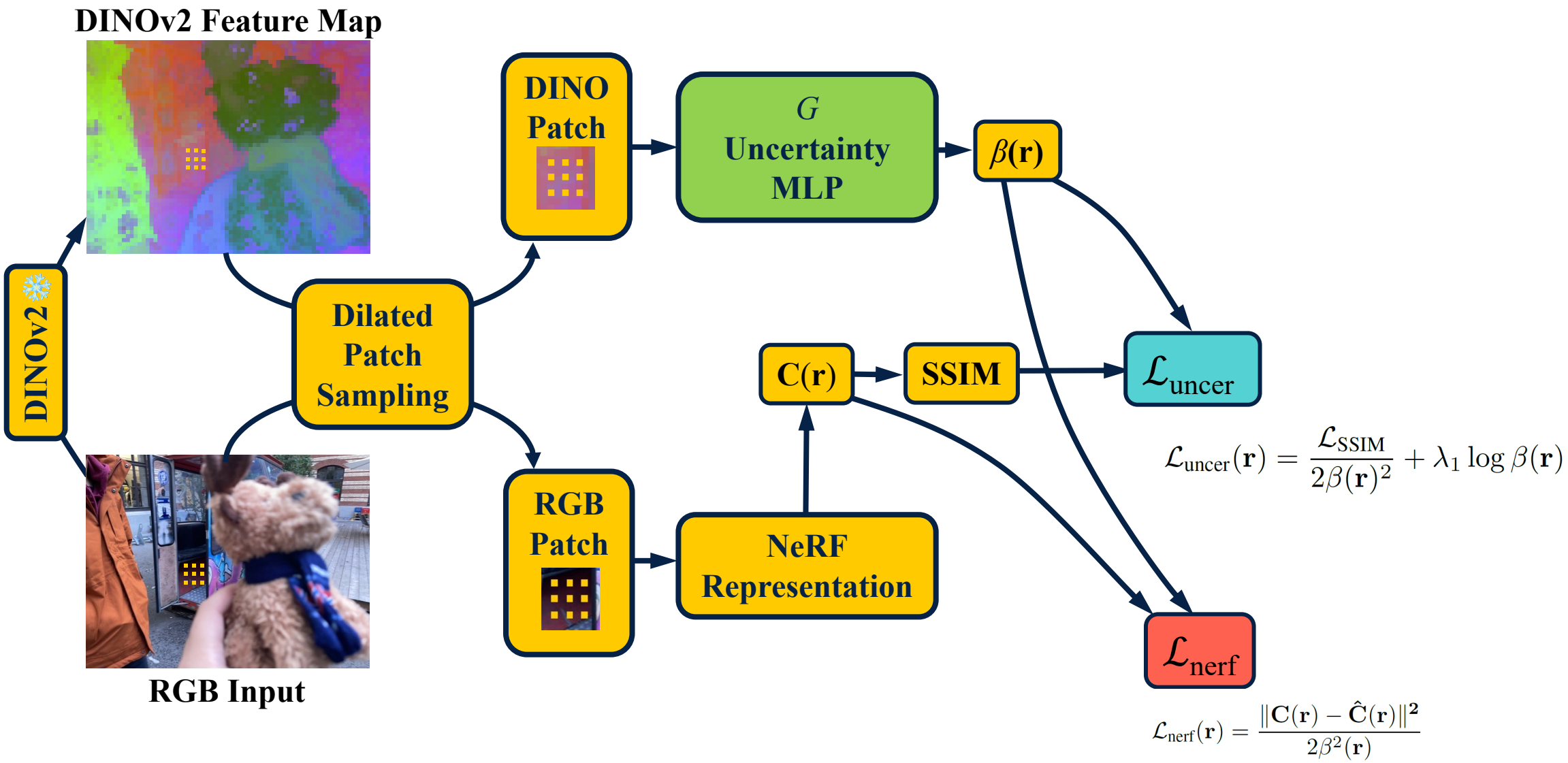
Dilated Patch
(Ours)

✓ **Larger Perceptive Field**: Improve optimization efficiency & reconstruction quality

Pipeline


DINOv2 Feature Map
DINOv2
RGB Input
G
Uncertainty
MLP
β(r)
C(r)
SSIM
NeRF
Representation
$\mathcal{L}_{\text{uncer}}$
$\mathcal{L}_{\text{nerf}}$
$\mathcal{L}_{\text{uncer}}(\mathbf{r}) = \frac{\mathcal{L}_{\text{SSIM}}}{2\beta(\mathbf{r})^2} + \lambda_1 \log \beta(\mathbf{r})$
$\mathcal{L}_{\text{nerf}}(\mathbf{r}) = \frac{\|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|^2}{2\beta^2(\mathbf{r})}$

Pipeline


DINOv2 Feature Map
DINOv2
Dilated Patch Sampling
DINO Patch
G Uncertainty MLP
β(r)
C(r)
SSIM
$\mathcal{L}_{\text{uncer}}$
RGB Patch
NeRF Representation
$\mathcal{L}_{\text{nerf}}$
RGB Input
$\mathcal{L}_{\text{uncer}}(\mathbf{r}) = \frac{\mathcal{L}_{\text{SSIM}}}{2\beta(\mathbf{r})^2} + \lambda_1 \log \beta(\mathbf{r})$
$\mathcal{L}_{\text{nerf}}(\mathbf{r}) = \frac{\|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|^2}{2\beta^2(\mathbf{r})}$

Results

On-the-go Dataset






Low Occlusion (5% ~ 10%)


Medium Occlusion (15% ~ 20%)


High Occlusion (~30%)

Occlusion Ratio: **Low**

`Statue` **- Input**

`Statue` - **Rendering Comparisons**

`Train Station` - **Input Images**

`Train Station` - **Rendering Comparisons**

Occlusion Ratio: **High**

`Patio-High` **- Input**

`Patio-High` - **Rendering Comparisons**

Analysis

Analysis - Efficiency

RobustNeRF

Sabour et al.: RobustNeRF: Ignoring Distractors with Robust Losses. CVPR, 2023 (Highlight)

Analysis - Efficiency

NeRF On-the-go

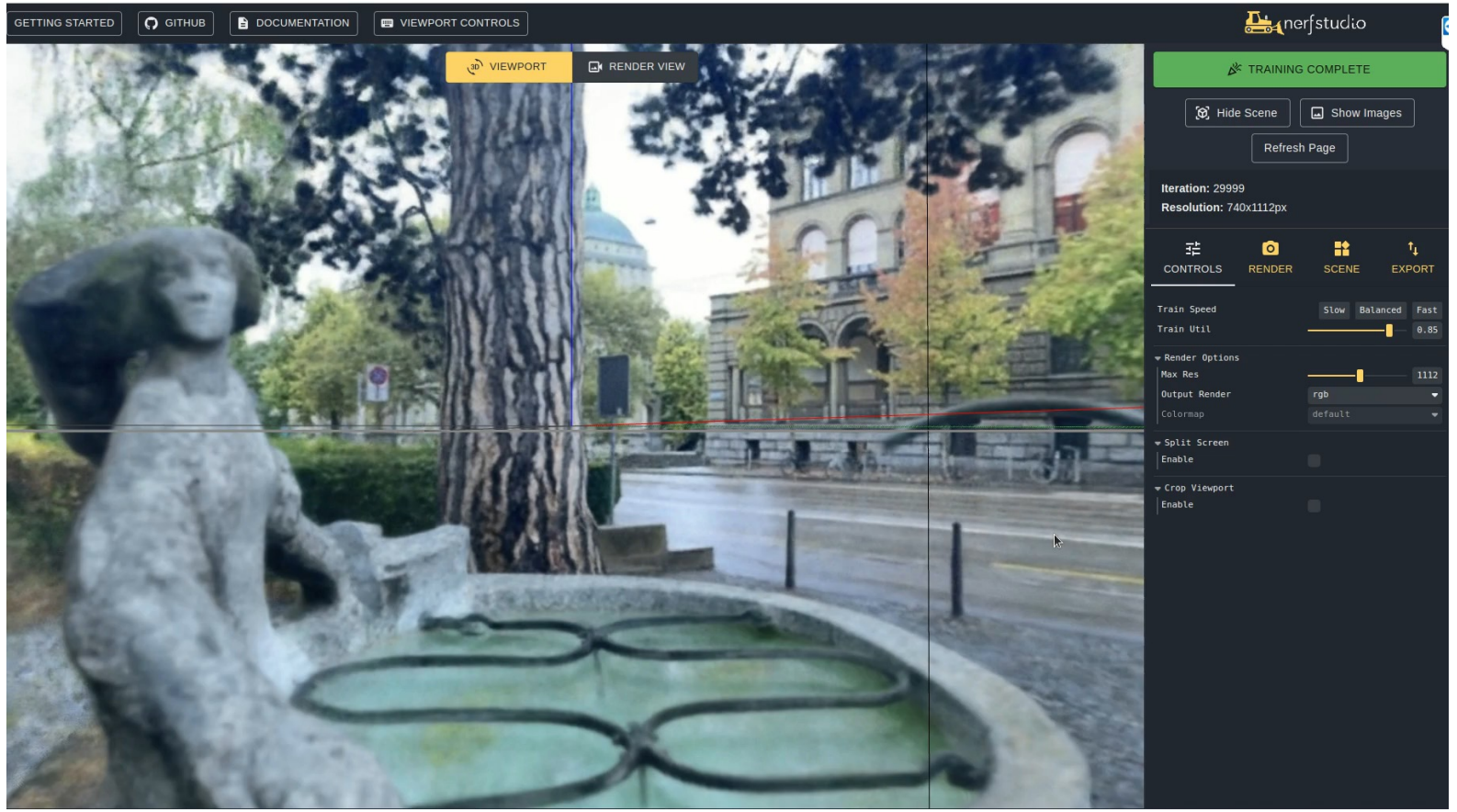
(Ours)

Analysis – Static Scene

RobustNeRF **Ours** MipNeRF 360 GT

Take-home Messages

- ***On-the-go*** module is plug-and-play for all NeRF methods
  - Integrated into NeRFStudio



Take-home Messages

- ***On-the-go*** module is plug-and-play for all NeRF methods
  - Integrated into NeRFStudio

- **2D foundation model** (DINOv2) rocks!

How to improve upon it?

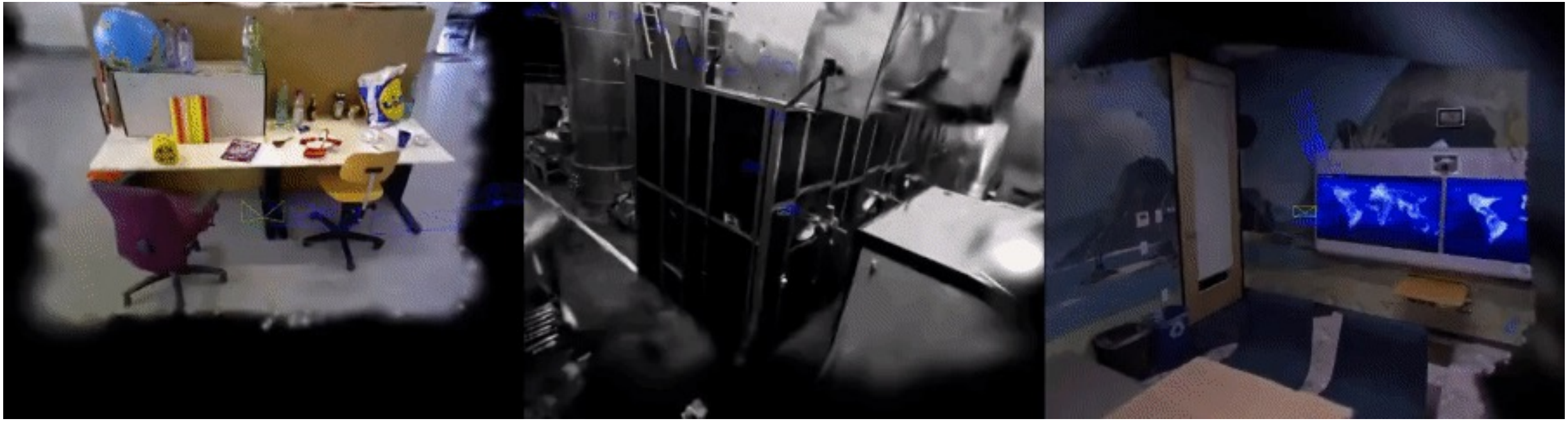
NeRF *On-the-go*

for VERY Large Urban Scenes

Zinuo You, Andreas Geiger, Anpei Chen: NeLF-Pro: Neural Light Field Probes. arXiv, 2023

NeRF *On-the-go*

Without COLMAP



Matsuki et al.: Gaussian Splatting SLAM. CVPR 2024

2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

3D Reconstruction

NeRF *On-the-go*
CVPR 2024

3D Scene Understanding

OpenScene
CVPR 2023

Segment3D
arXiv 2024

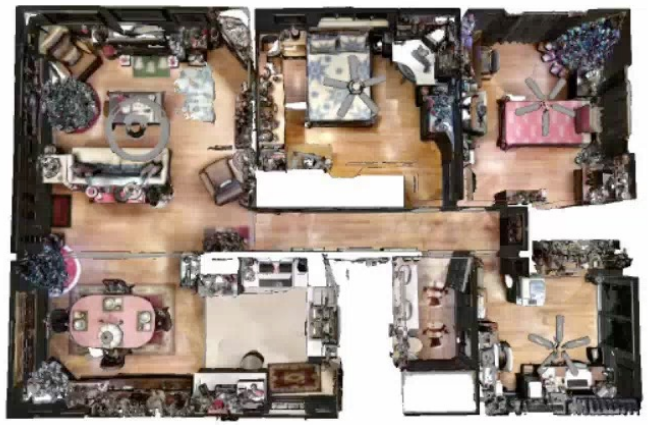
2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

3D Reconstruction

NeRF *On-the-go*
CVPR 2024

3D Scene Understanding



OpenScene
CVPR 2023

Segment3D
arXiv 2024

Input 3D Geometry

Input 3D Geometry

Traditional 3D Scene Understanding
(e.g. Semantic Segmentation)

Only **train** and **test** on a few common classes

3D Scene Understanding Tasks w/o Labels

Input 3D Geometry

- Affordance prediction

3D Scene Understanding Tasks w/o Labels

- Affordance prediction

Example: “where can I sit?”

3D Scene Understanding Tasks w/o Labels

Input 3D Geometry

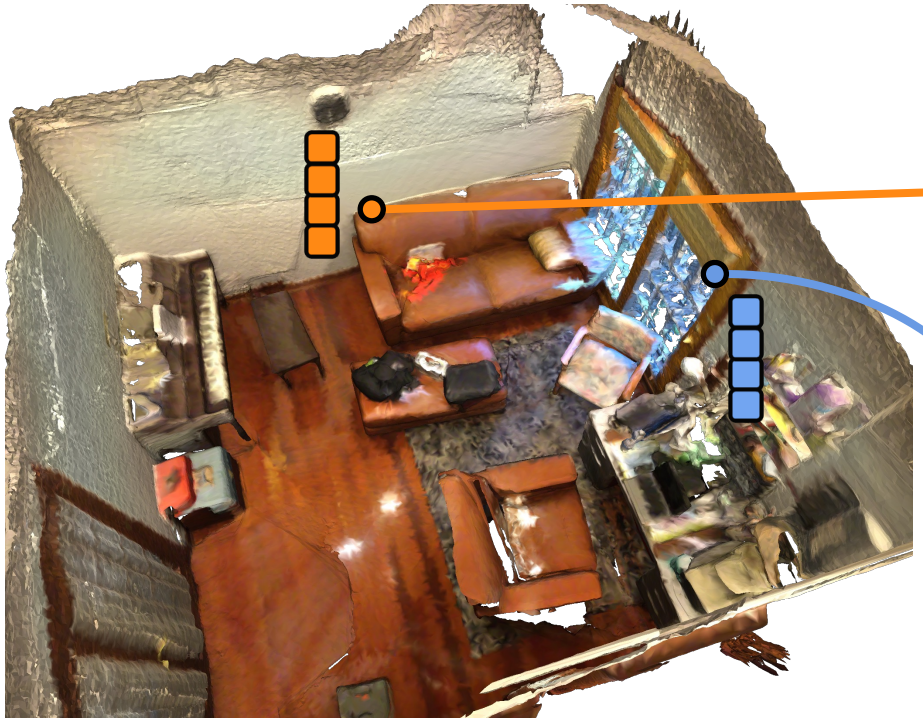
- Affordance prediction
- Material identification
- Physical property estimation
- Rare object retrieval
- Activity site prediction
- Fine-grained semantic segmentation
- Many more…

How to have a single model for all these 3D tasks
without any labeled 3D data?

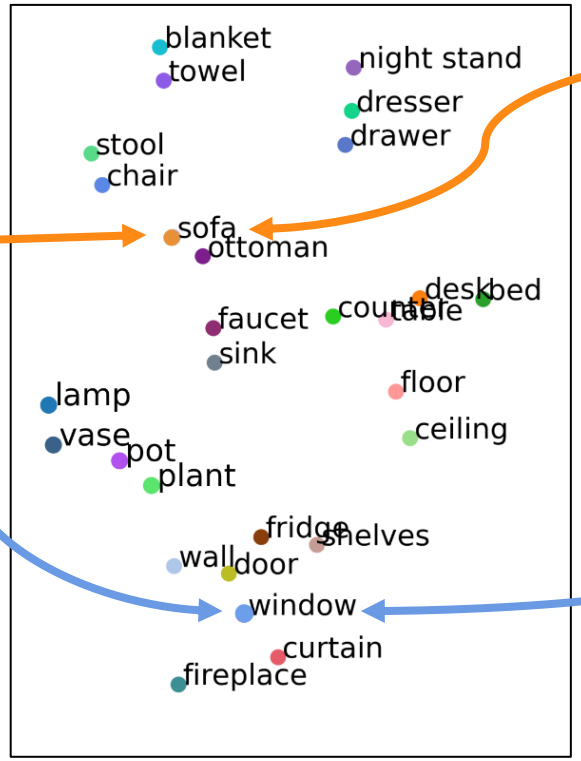
Leverage **2D foundation models**

Key Idea

Co-embed 3D Features with CLIP Features



3D Geometry



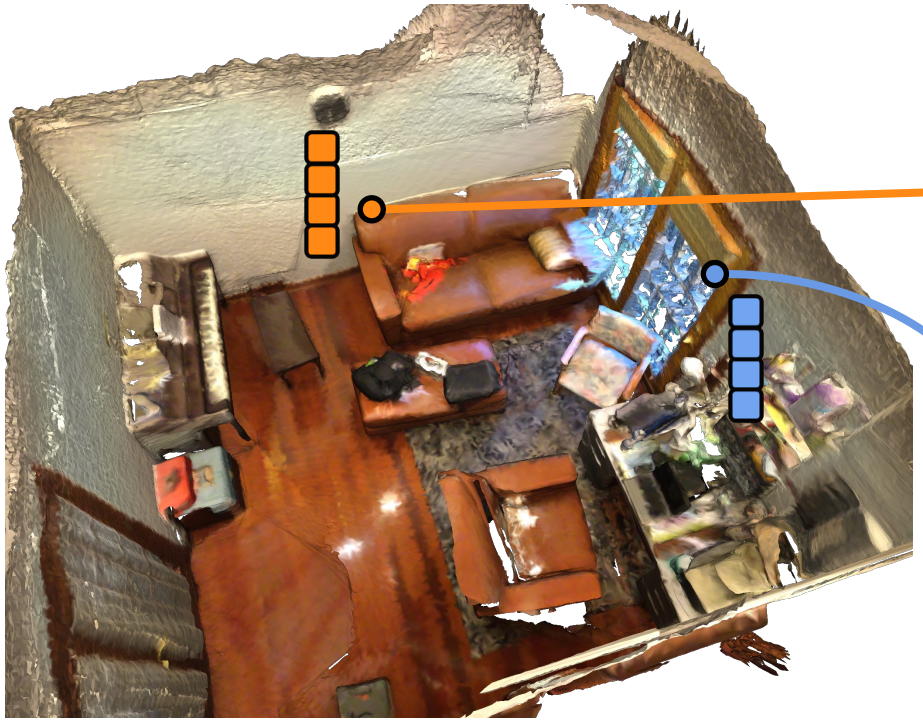
CLIP Text Features

(visualize with T-SNE)

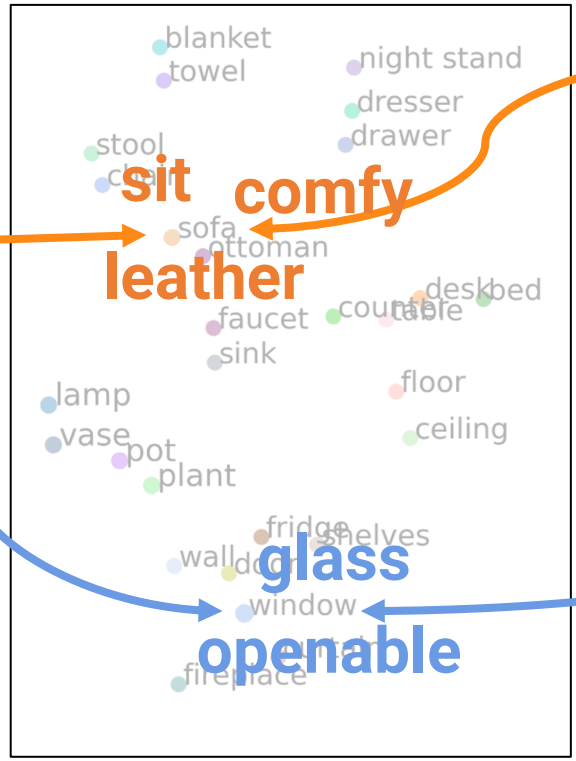
RGB Images

Key Idea

Co-embed 3D Features with CLIP Features



3D Geometry



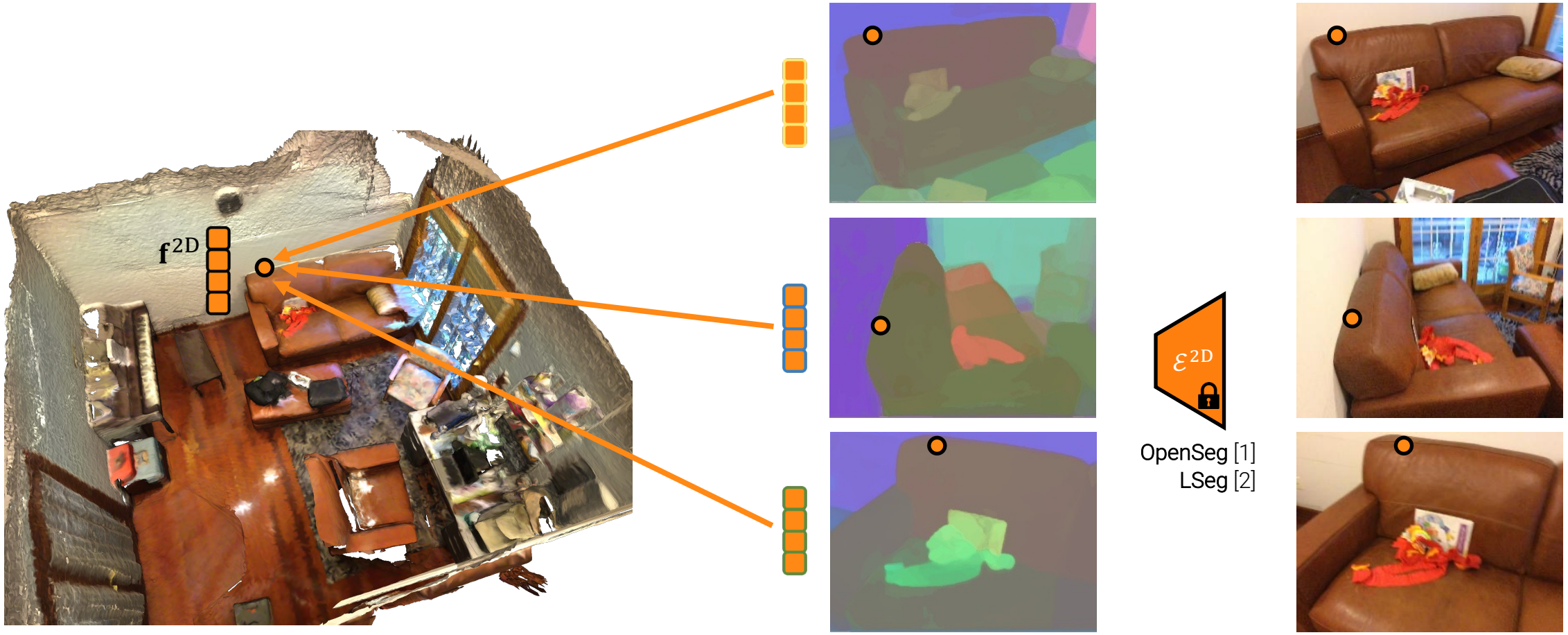
CLIP Text Features

(visualize with T-SNE)

RGB Images

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

Step 1: Multi-view Feature Fusion



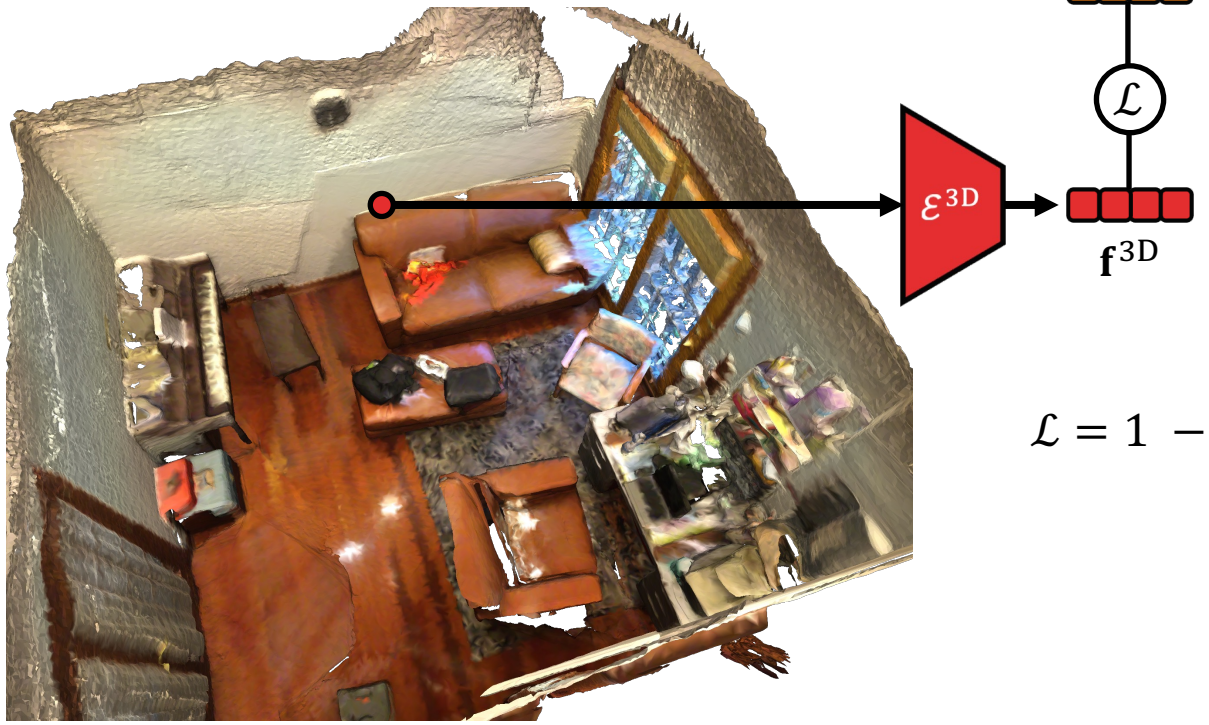
3D Geometry

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

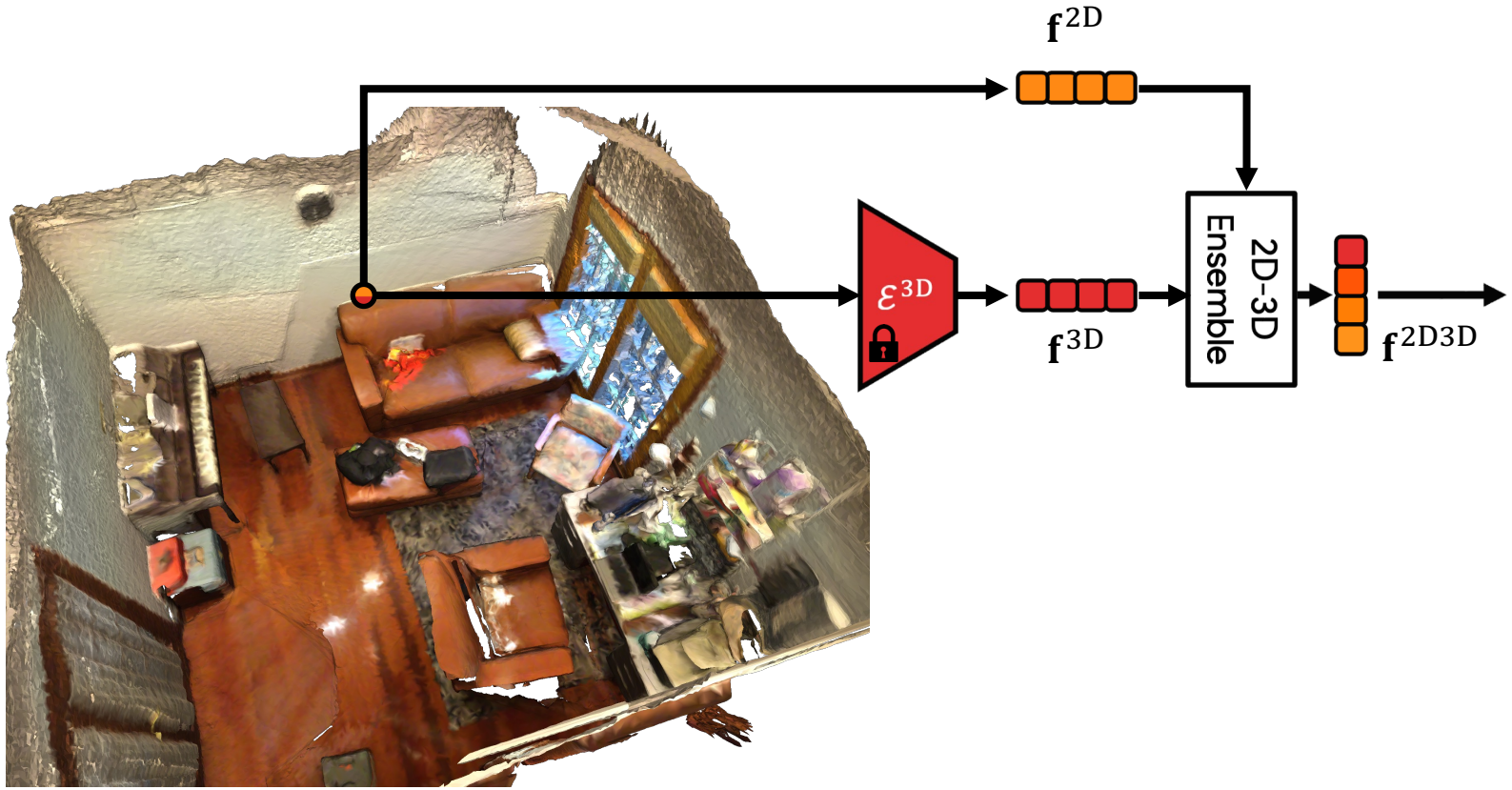
Step 2: 3D Feature Distillation



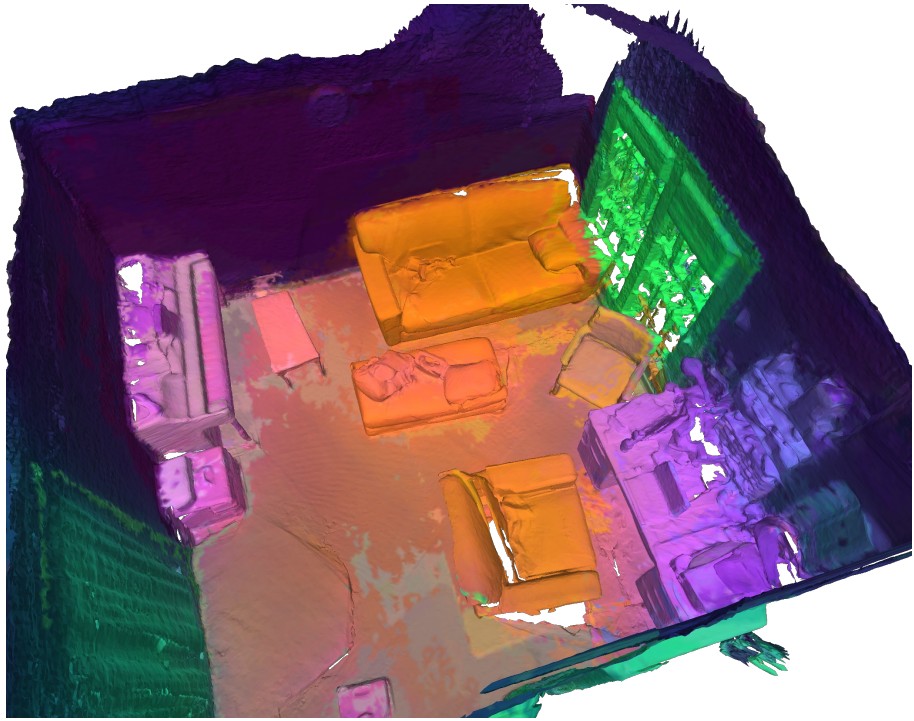
$$\mathcal{L} = 1 - \cos(\mathbf{f}^{2D}, \mathbf{f}^{3D})$$

3D Geometry

Inference: 2D-3D Ensemble



3D Geometry



2D-3D Ensemble Features

(visualize with PCA)

Open-Vocabulary, Zero-shot

3D Semantic Segmentation

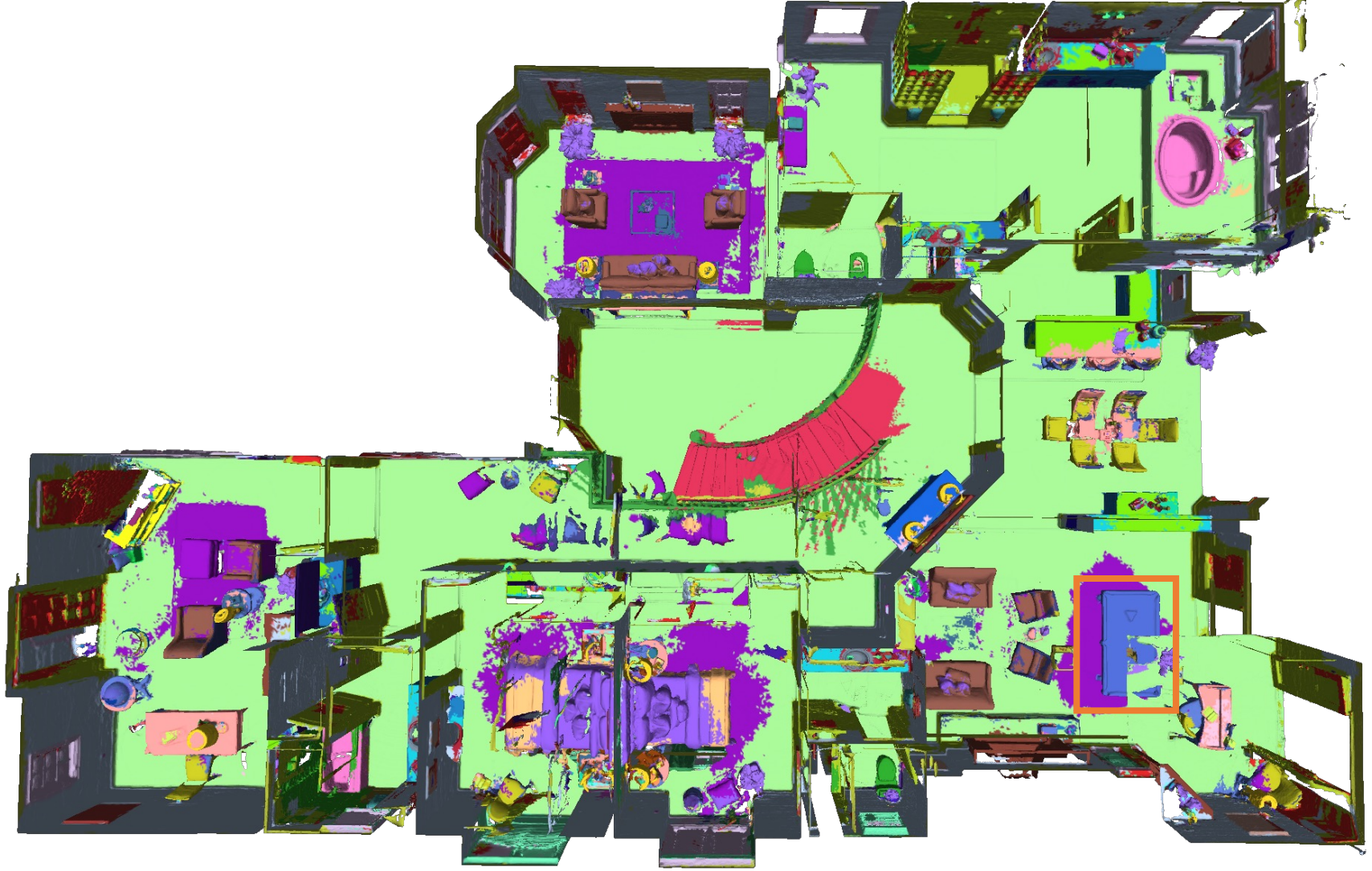
Input 3D Geometry

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

(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

Image-based 3D Scene Query

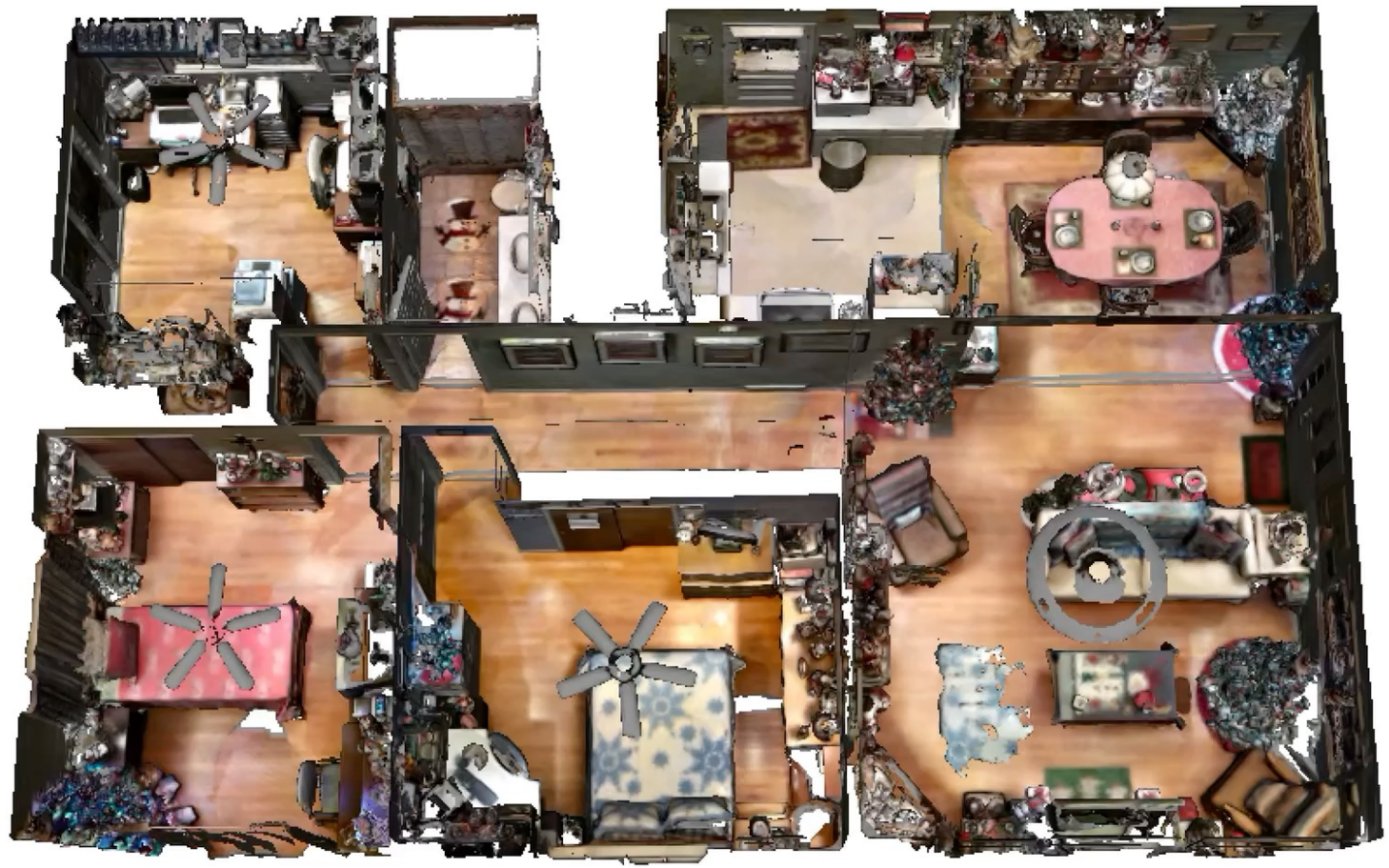
Image Queries

Given 3D Geometry

Interactive Demo

Open-vocabulary 3D Scene Exploration

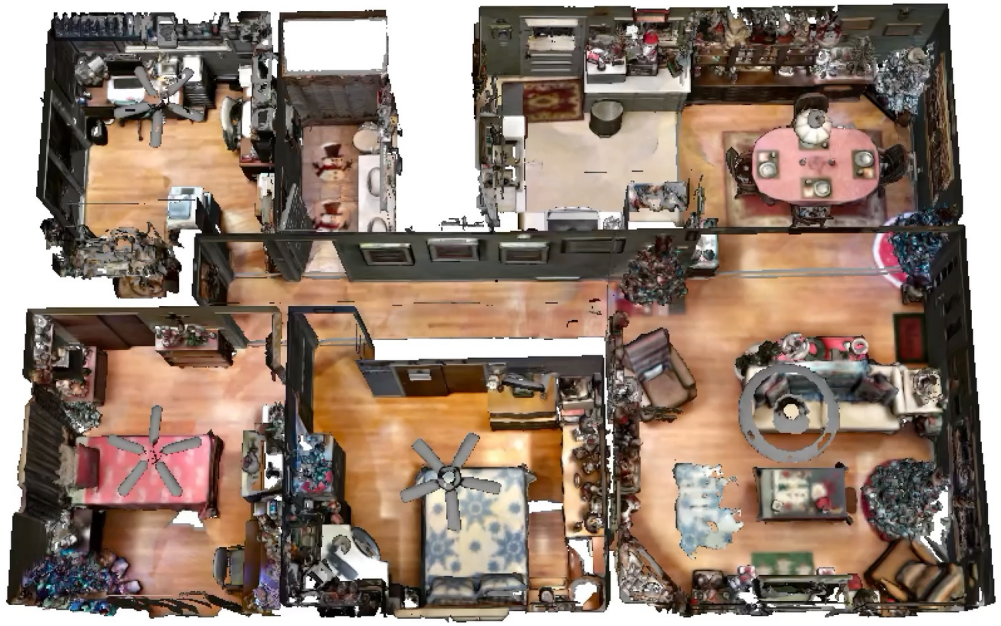
Text queries:



Take-home Messages

- ✚ Open up a **wide range of applications** by leveraging large 2D vision-language models
- ✚ Inspire future works to shift to open-vocabulary tasks
- ➖ Segmentation quality is quite limited

OpenScene



Accurately segment and understand 3D scenes is essential!

Peng et al.: OpenScene: 3D Scene Understanding with Open Vocabularies. CVPR, 2023

Motivation

Instance Segmentation Methods Requires 3D Manual Labels

Input 3D Scene

3D Semantic Instances

😭 **Expensive** and **challenging** to annotate 3D masks

😭 Perform poorly in scenes **out of training distribution**

Schult et al.: Mask3D: Mask Transformer for 3D Instance Segmentation. ICRA, 2023

Motivation

2D Foundation Model Rocks!

🤩 SAM exhibits **extraordinary ability to generalize**

🤔 Only applicable to **2D data**

Kirillov et al.: Segment Anything. ICCV, 2023 (Best Paper Honorable Mention)

How to obtain accurate 3D segmentation
without any manual 3D labels?

Leverage **2D foundation models**

Segment3D

Learning Fine-Grained Class-Agnostic 3D Segmentation without Manual Labels


Rui Huang


Songyou Peng


Ayça Takmaz


Federico Tombari


Marc Pollefeys

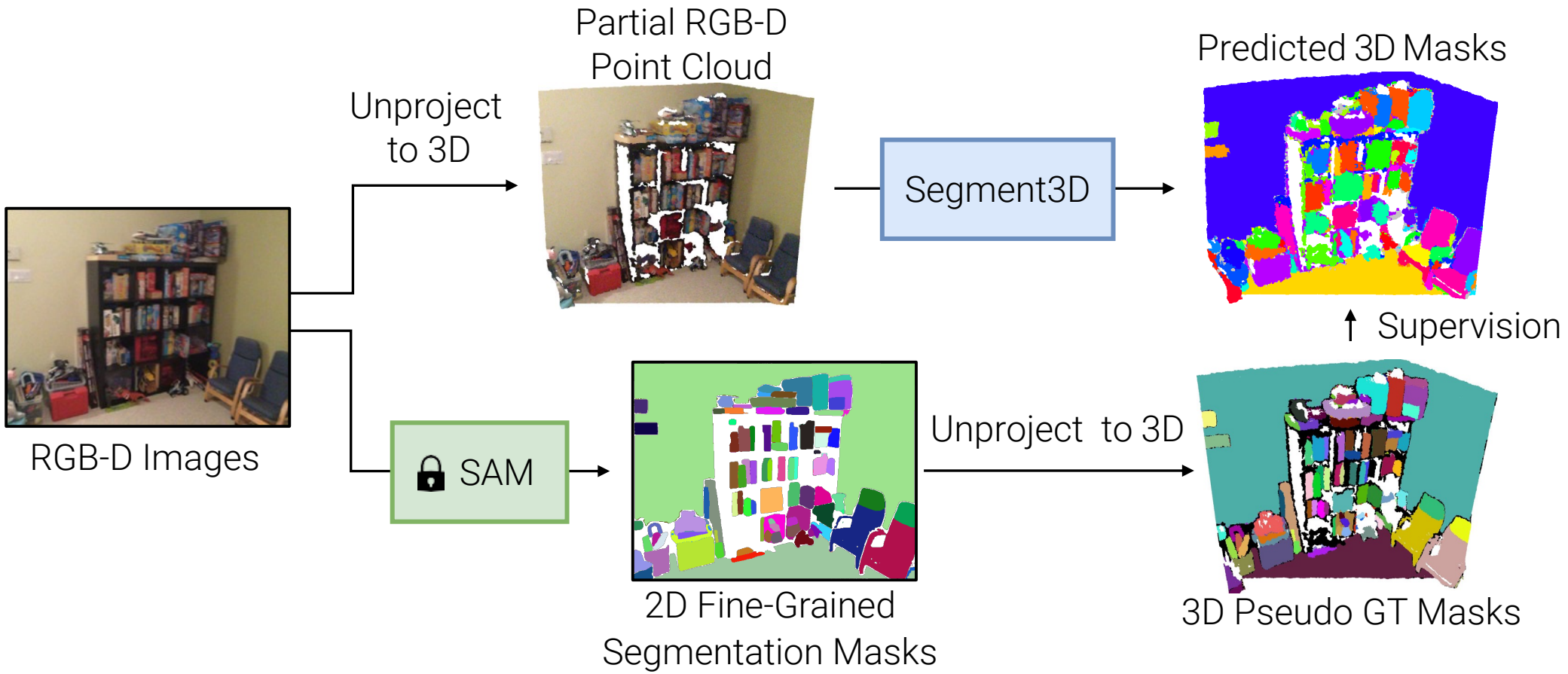

Shiji Song


Gao Huang


Francis Engelmann

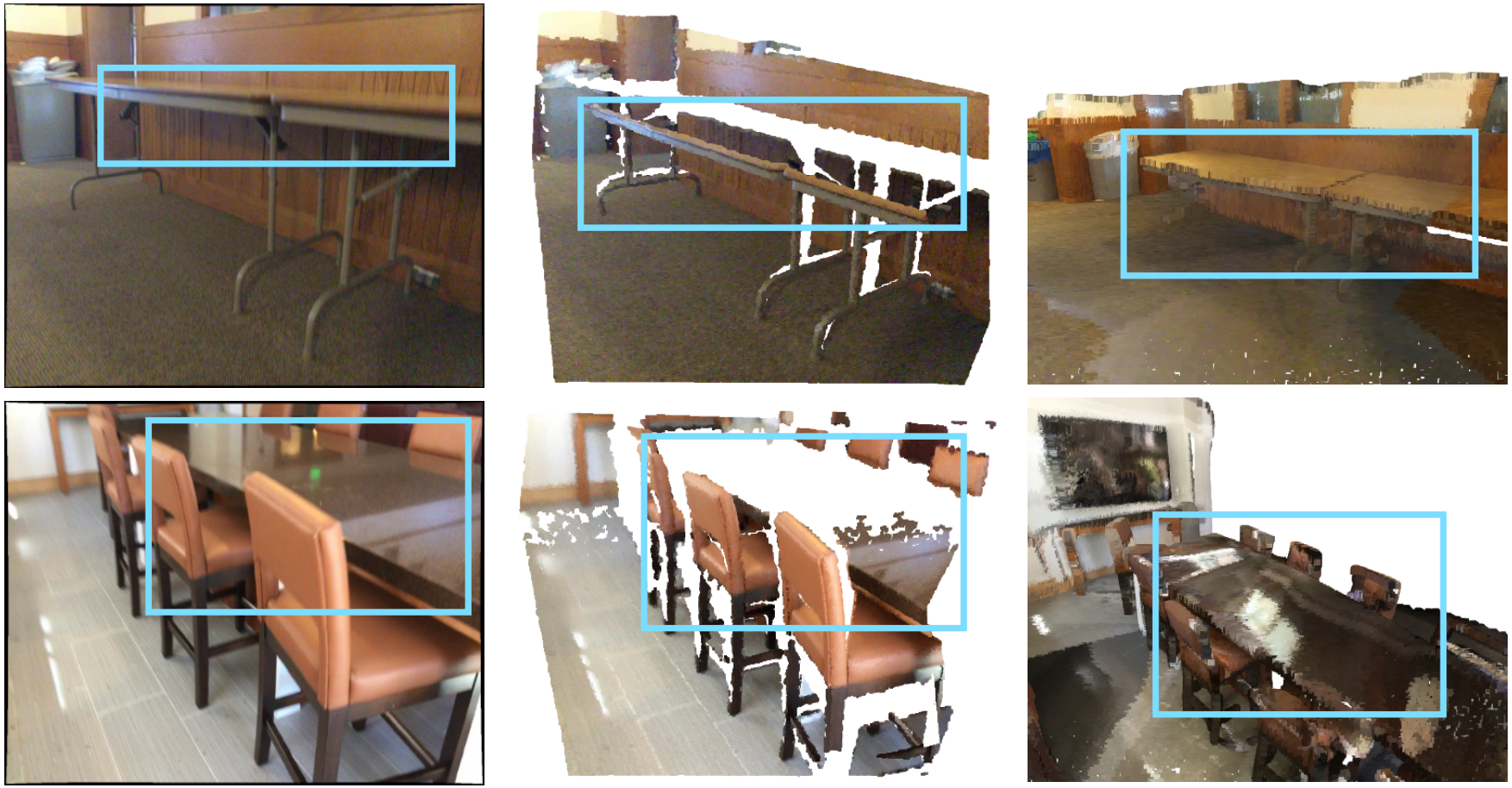
Segment3D

Stage 1: Pre-training on Partial Point Clouds



Segment3D

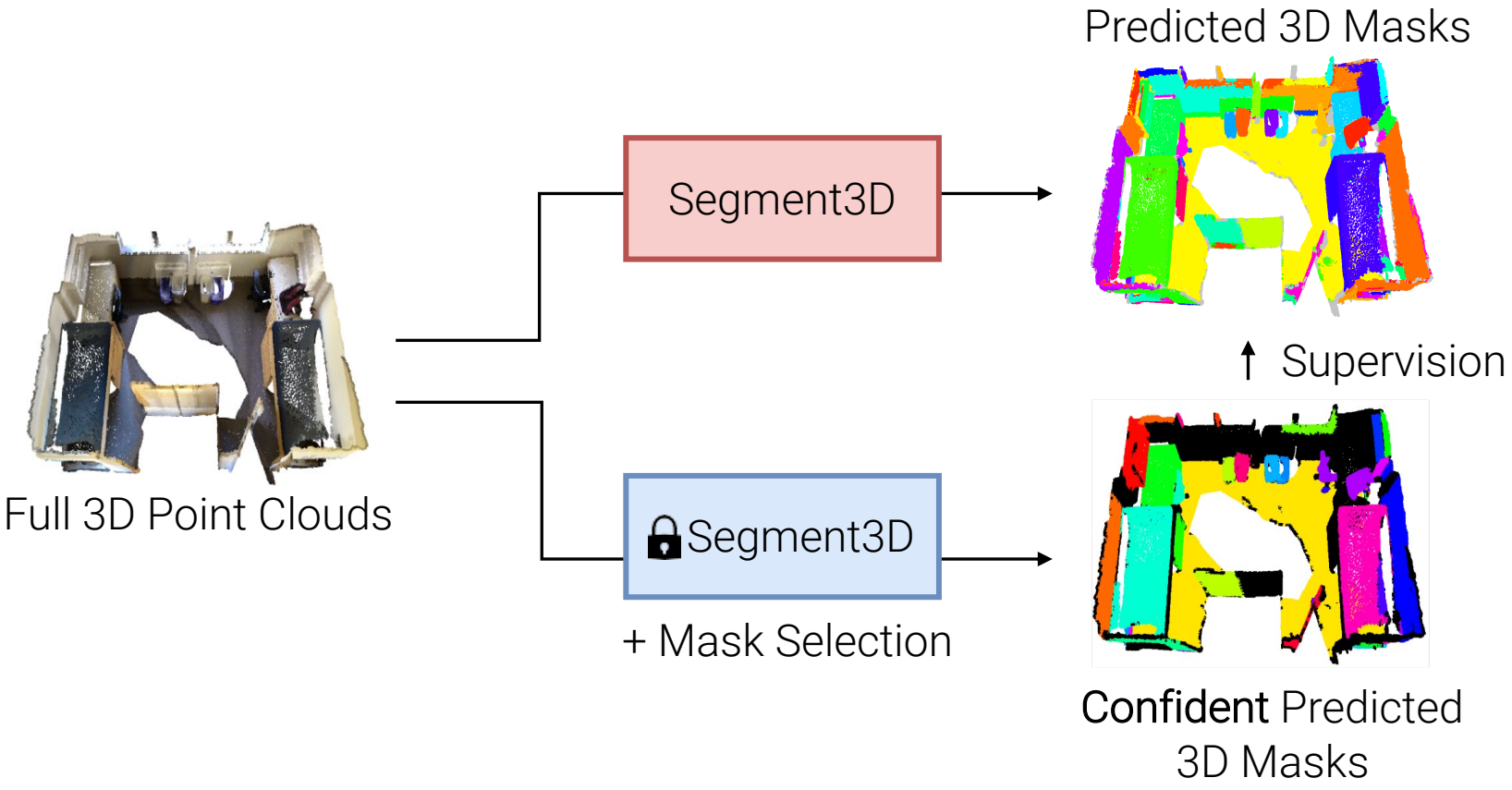
Domain Gap Between Partial and Full Point Clouds



RGB Images | Partial Point Clouds | Full Point Clouds

Segment3D

Stage 2: Fine-tune on Full Point Clouds



No manual labels are needed at all!

Results

Class-Agnostic 3D Segmentation

ScanNet++ Validation Set

Input Point Clouds

Mask3D [1]

[1] Schult et al.: Mask3D: Mask Transformer for 3D Instance Segmentation. ICRA, 2023

Class-Agnostic 3D Segmentation

ScanNet++ Validation Set

Input Point Clouds

Segment3D (Ours)

Class-Agnostic 3D Segmentation

ScanNet++ Validation Set

Segment3D (Ours)

GT

Class-Agnostic 3D Segmentation

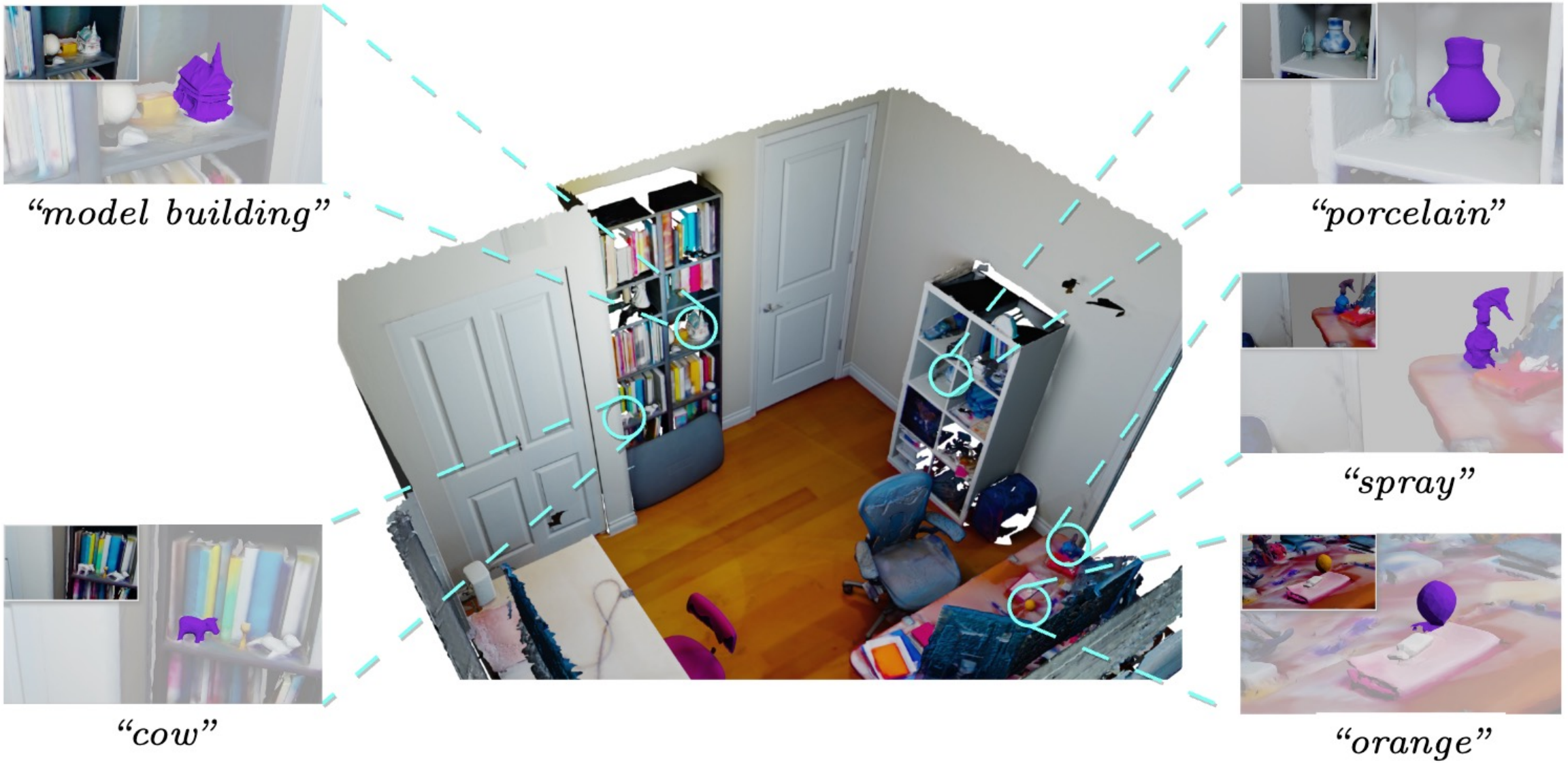
ScanNet++ Validation Set

Model	Avg. Inference Time/s	*without post-processing*			*with post-processing*		
		AP	**AP$_{50}$**	**AP$_{25}$**	**AP**	**AP$_{50}$**	**AP$_{25}$**
Mask3D [41]	0.7	8.7	15.5	27.2	14.3	21.3	29.9
SAM3D [50]	386.7	3.9	9.3	22.1	8.4	16.1	30.0
Felzenszwalb *et al.* [16]	12.6	5.8	11.6	27.2	–	–	–
Segment3D (**Ours**)	0.7	**13.0**	**23.8**	**38.3**	**20.2**	**30.9**	**42.7**

Effect of Two-Stage Training

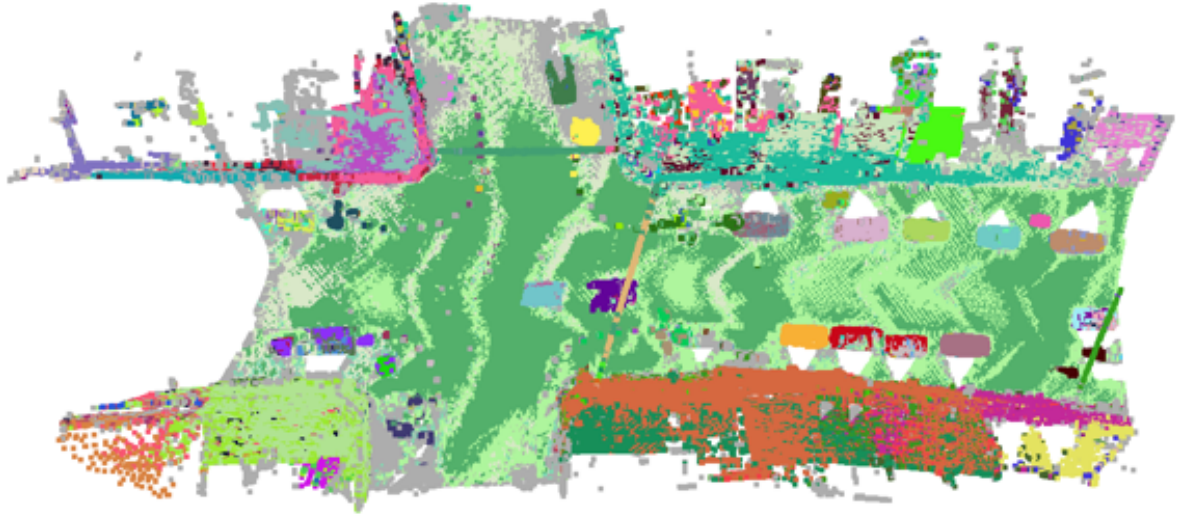
Training Stages	AP	AP_{50}	AP_{25}
Pre-Training (Stage 1)	6.8	14.2	30.6
+ Fine-Tuning (Stage 2)	**13.0** (+6.2)	**23.8** (+9.6)	**38.3** (+7.7)

Open-Vocabulary Segmentation



Outdoor Scenes In-the-Wild

Mask3D



Segment3D

Take-home Messages

- No 3D manual labels are used for training at all!

- **2D foundation model** (SAM) rocks!

Future work

- Unify as a single-stage pipeline
- Single pipeline for open-vocabulary 3D segmentation

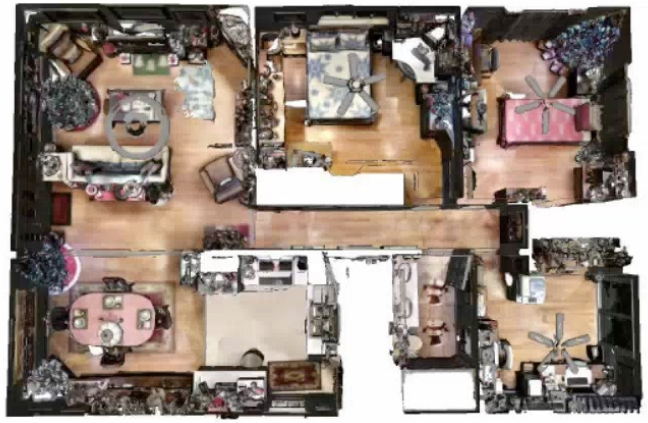
2D Magic in a 3D World

2D Foundation Models for 3D Vision Tasks

3D Reconstruction

NeRF *On-the-go*
CVPR 2024

3D Scene Understanding



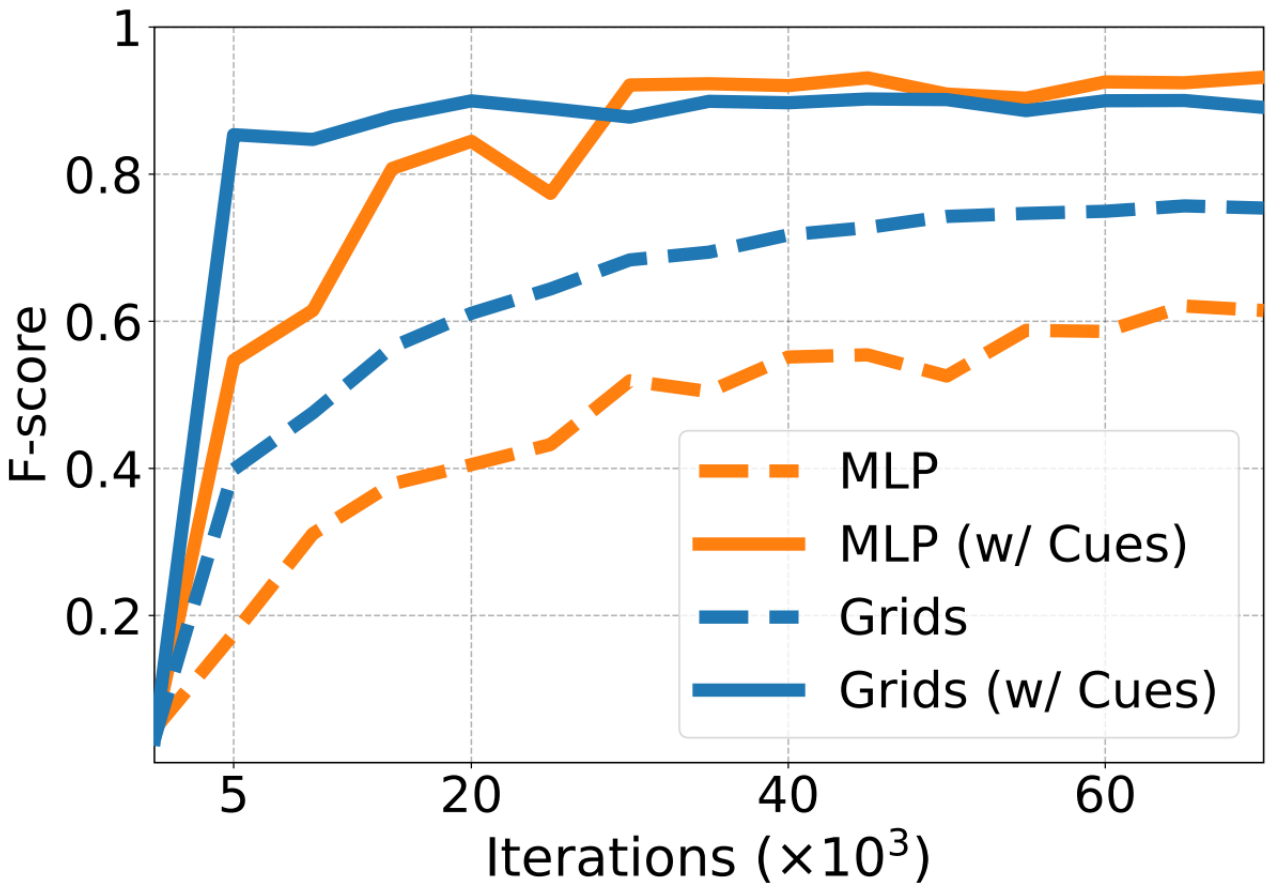
OpenScene
CVPR 2023

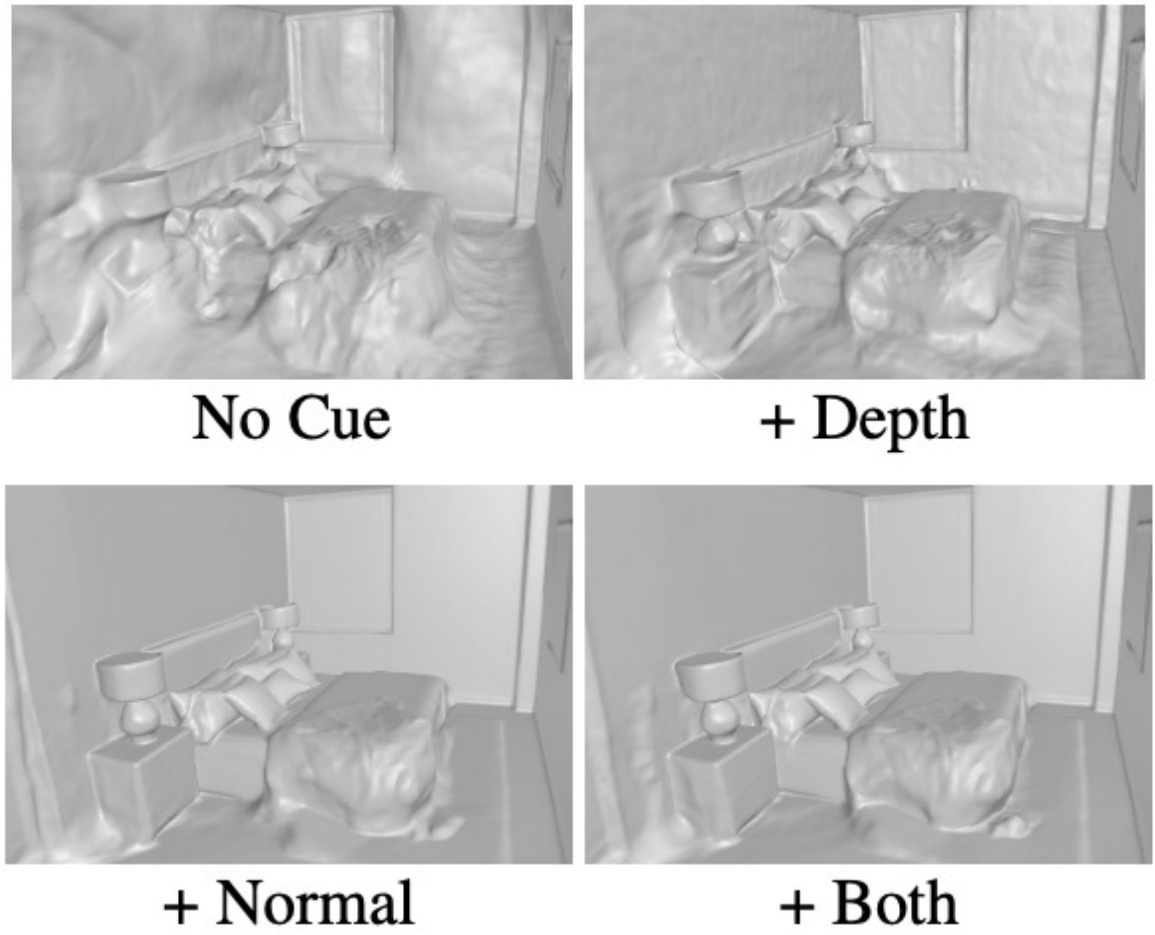
Segment3D
arXiv 2024

This talk focuses on how to **leverage**
2D foundation models for 3D tasks

So, what is next?

Monocular Cues Help!





Performance is bounded by the monocular cues!

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

Current Interests

Next-Generation Monocular Predictor

Marigold: Stable Diffusion-Based Monocular Predictor

Ke, Obukhov, Huang, Metzger, Daudt, Schindler: Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. CVPR, 2024

Current Interests

Next-Generation Monocular Predictor

- Other **modalities** like surface normals, uncertainty, etc…
- Inference **speed**
- **Video depth** predictor (temporal-consistent)
- **Metric depths**
- …

Ke, Obukhov, Huang, Metzger, Daudt, Schindler: Marigold: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. arXiv, 2023

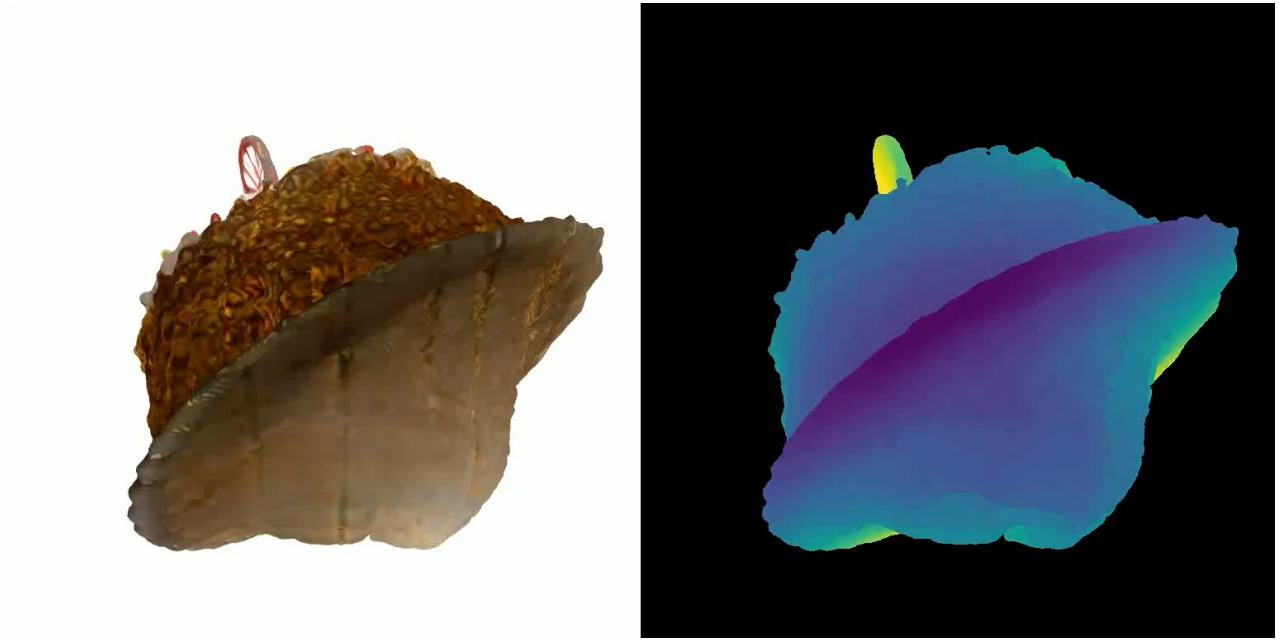
This talk focuses on how to **leverage** **2D foundation models** for 3D tasks

Current Interests

3D Foundation Models

DreamFusion
[ICLR'23]

1.5 Hours



Large Reconstruction Models
[ICLR'24]

5 Seconds!

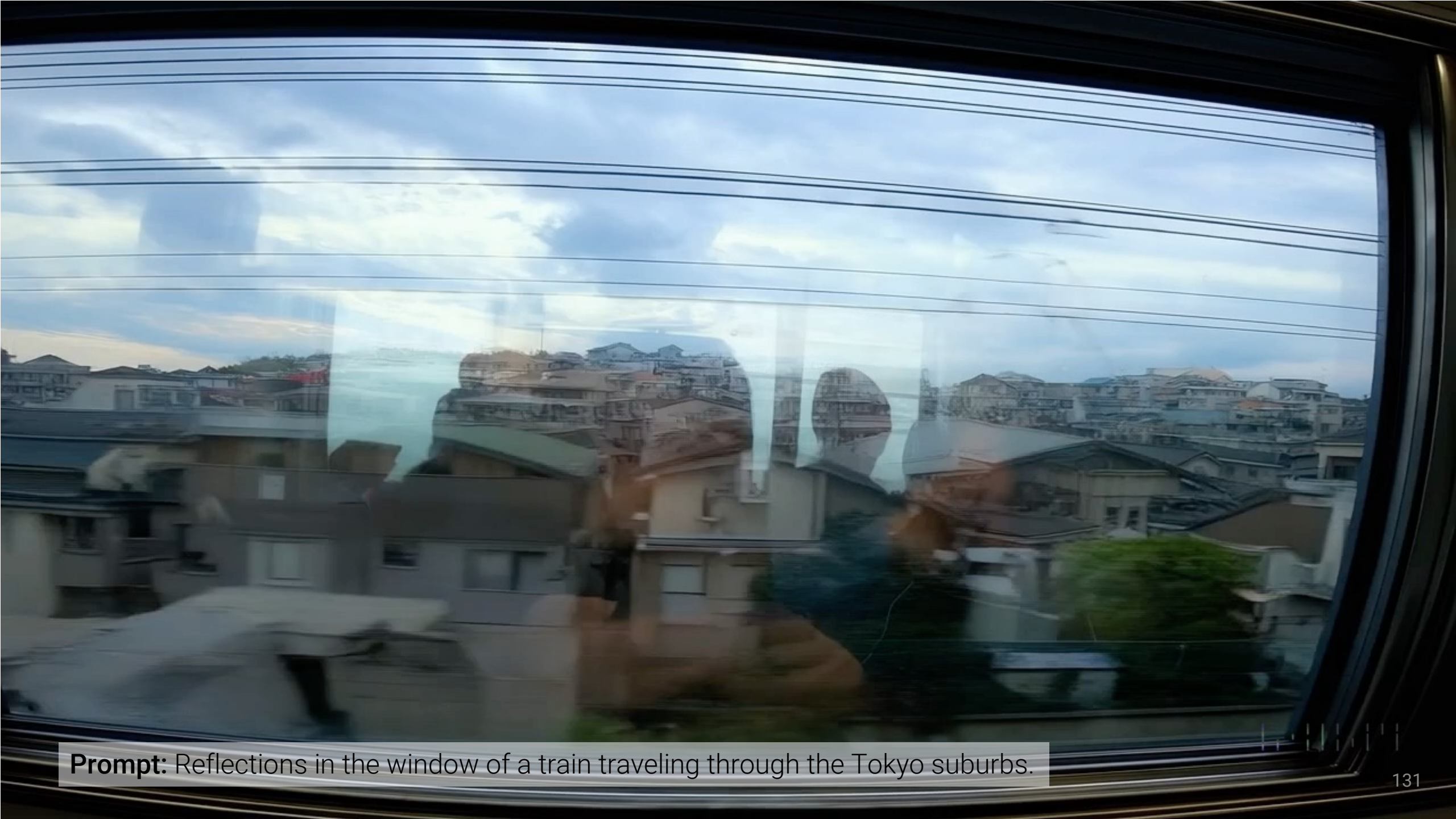
So far, only object-level 3D foundation models

Current Interests

3D Foundation Models

- How to leverage internet-level in-the-wild data?
- How to train in a *smarter* way?
  - Multi-view generator + sparse-view reconstruction?
  - How to better leverage geometry under the data-driven pipeline?

Then… Sora is here



Prompt: Reflections in the window of a train traveling through the Tokyo suburbs.

Prompt: Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up from it’s tires,……

Open Questions…

- Do we really need a 3D foundation model?
- How to inject 3D to help video foundation models?
- How to leverage Sora for 3D generation?
- How to leverage Sora to model dynamic+static scenes?
- Controllability/Local editability in generative models
- In this era, how to *survive* as a PhD student in universities?

2D Magic in a 3D World

Songyou Peng

pengsongyou.github.io

Q&A