Neural Scene Representations for 3D Reconstruction and Scene Understanding

Songyou Peng

PhD Supervisors
Marc Pollefeys & Andreas Geiger

External Examiners
Leonidas Guibas & Vincent Sitzmann

ETH zürich

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

Intelligent systems interact with **3D environments**

3D Reconstruction
Create digital twins from real scenes

3D Scene Understanding
Analyze the scene digitally

Video Credit: YouTube - Real time archviz apartment

Key Challenges

Reconstruct and Understand 3D Environments

- Reconstruct 3D scenes **at scale**
- Reconstruct 3D scenes **at speed**
- Reconstruct purely **from 2D observations**

Key Challenges

Reconstruct and Understand 3D Environments

- Reconstruct 3D scenes **at scale**
- Reconstruct 3D scenes **at speed**
- Reconstruct purely **from 2D observations**

- Understand **arbitrary concepts** in a 3D scene
- Learn to understand **without labeled 3D data**

Research Overview of My PhD

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

ConvOccNet
ECCV 2020 (Spotlight)

MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

NICE-SLAM
CVPR 2022

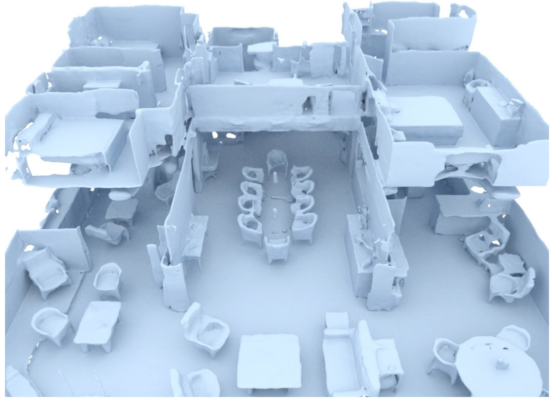
NICER-SLAM
3DV 2024 (Oral)

UNISURF
ICCV 2021 (Oral)

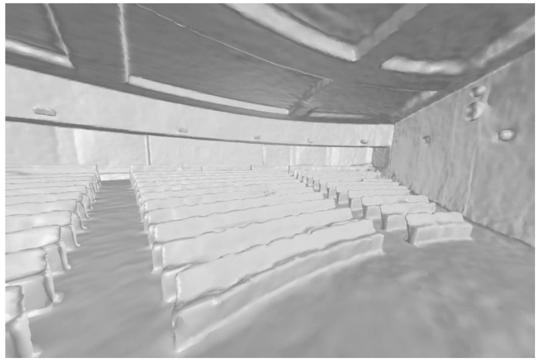
OpenScene
CVPR 2023

Research Overview of My PhD

Learn to Reconstruct and Understand 3D Environments



ConvOccNet
ECCV 2020 (Spotlight)



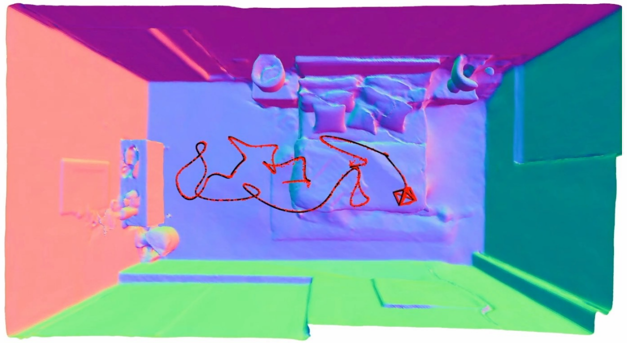
MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

KiloNeRF
ICCV 2021

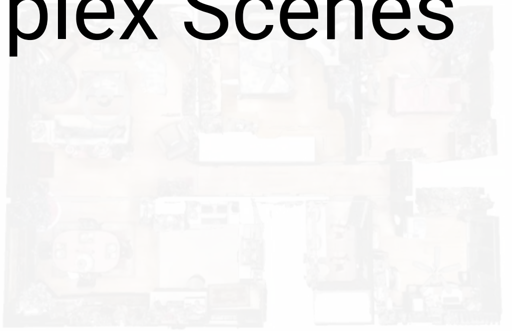
Topic #1:
Reconstruct Complex Scenes

NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)

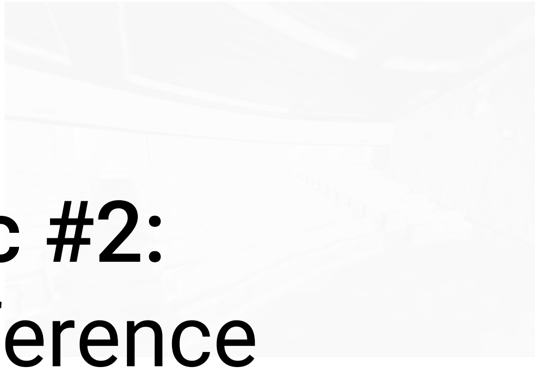
UNISURF
ICCV 2021 (Oral)



OpenScene
CVPR 2023

Research Overview of My PhD

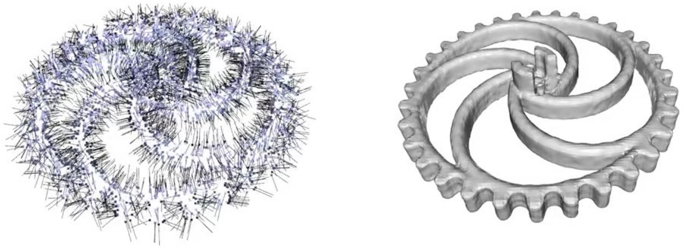
Learn to Reconstruct and Understand 3D Environments

Topic #2:
Fast Inference

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)

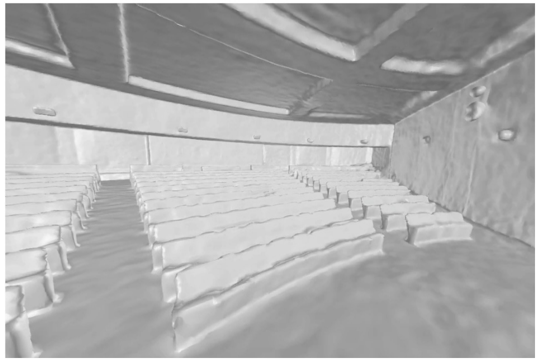
UNISURF
ICCV 2021 (Oral)

OpenScene
CVPR 2023

Research Overview of My PhD

Learn to Reconstruct and Understand 3D Environments

ConvOccNet
ECCV 2020 (Spotlight)

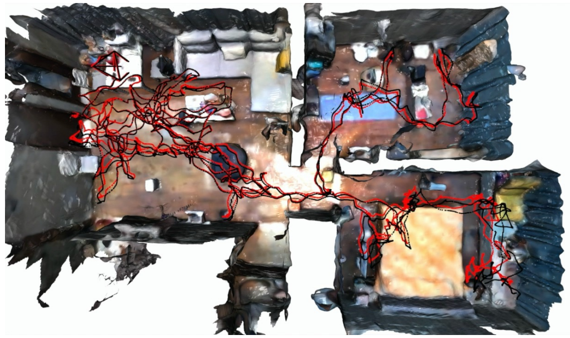


MonoSDF
NeurIPS 2022

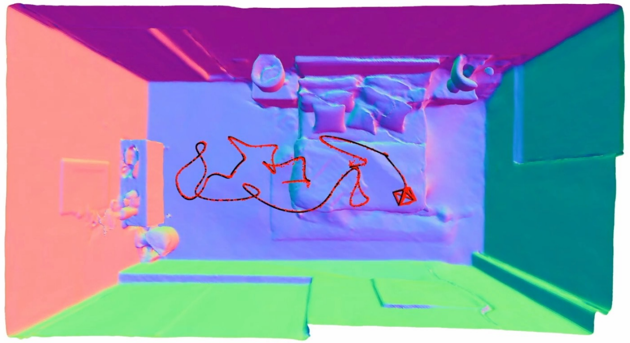
Topic #3:
Reconstruct from 2D Observations

Shape As Points
NeurIPS 2021 (Oral)

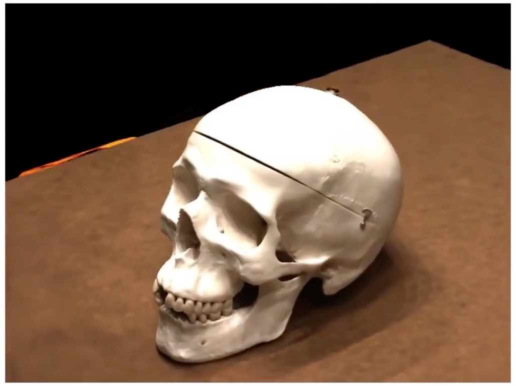
KiloNeRF
ICCV 2021



NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)



UNISURF
ICCV 2021 (Oral)

OpenScene
CVPR 2023

Research Overview of My PhD

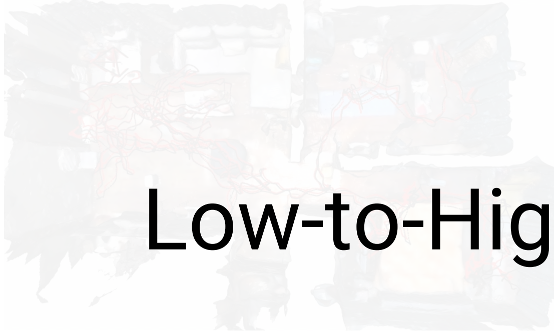
Learn to Reconstruct and Understand 3D Environments

ConvOccNet
ECCV 2020 (Spotlight)

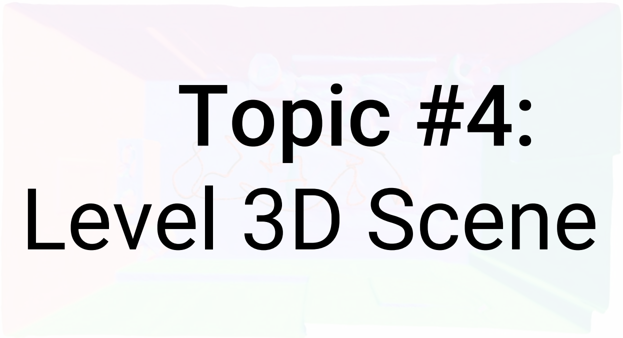
MonoSDF
NeurIPS 2022

Shape As Points
NeurIPS 2021 (Oral)

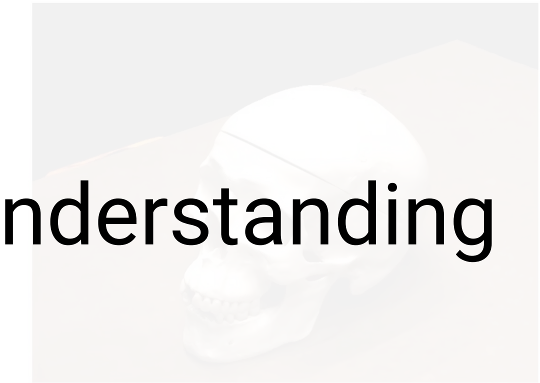
KiloNeRF
ICCV 2021



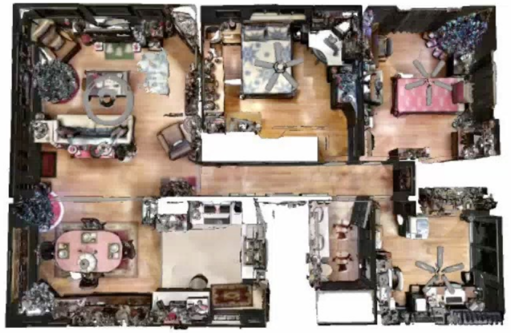
NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)



UNISURF
ICCV 2021 (Oral)

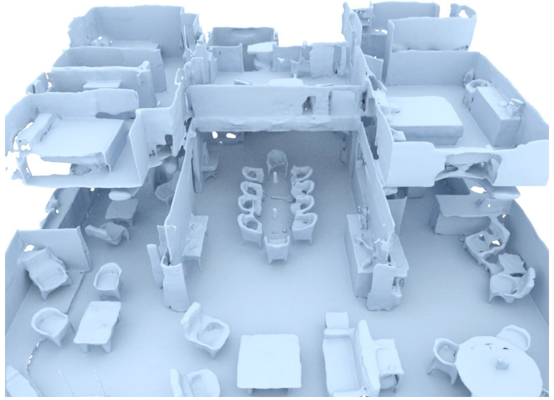


OpenScene
CVPR 2023

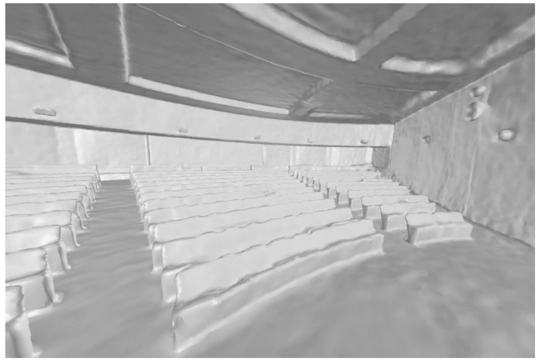
Topic #4:
Low-to-High Level 3D Scene Understanding

Research Overview of My PhD

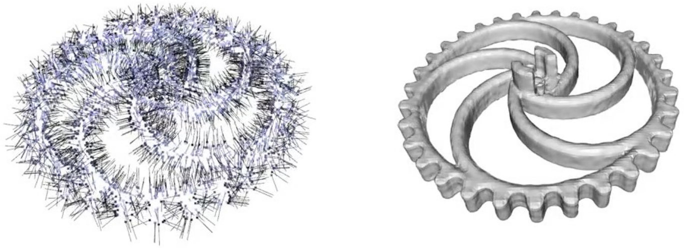
Learn to Reconstruct and Understand 3D Environments



ConvOccNet
ECCV 2020 (Spotlight)

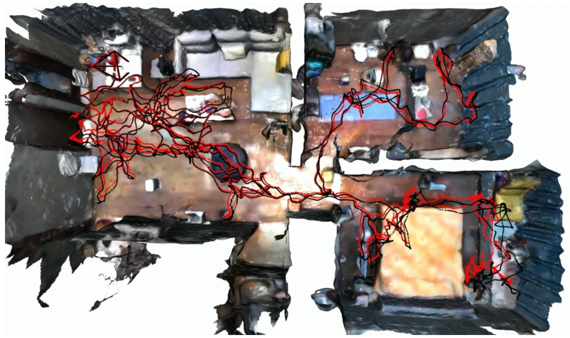


MonoSDF
NeurIPS 2022

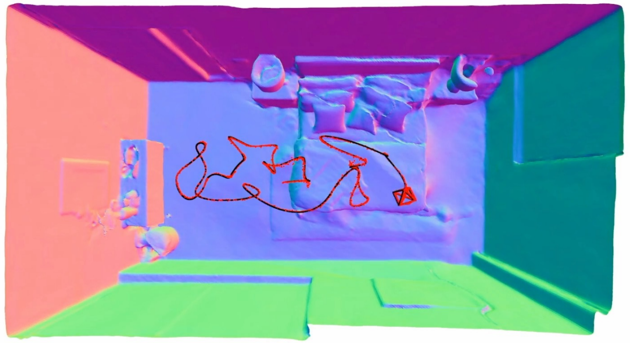


Shape As Points
NeurIPS 2021 (Oral)

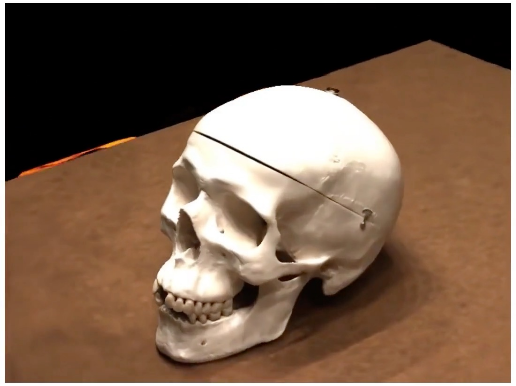
KiloNeRF
ICCV 2021



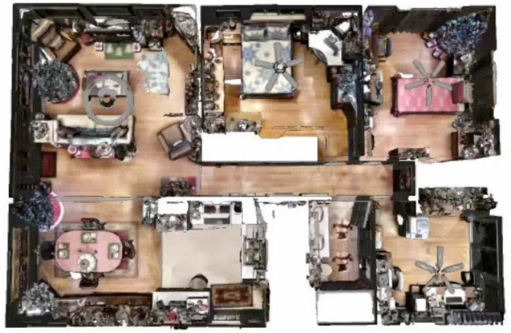
NICE-SLAM
CVPR 2022



NICER-SLAM
3DV 2024 (Oral)



UNISURF
ICCV 2021 (Oral)



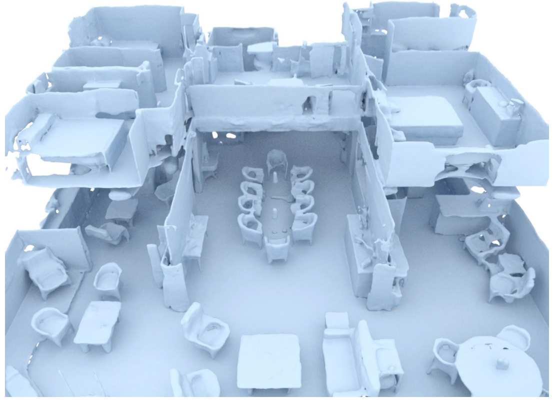
OpenScene
CVPR 2023

This Thesis

Develop 3D Neural Scene Representations

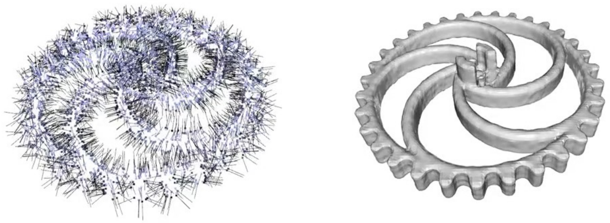
for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes



ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference



Shape As Points
NeurIPS 2021 (Oral)

3. From 2D Observations

NICE-SLAM
CVPR 2022

4. Arbitrary Queries

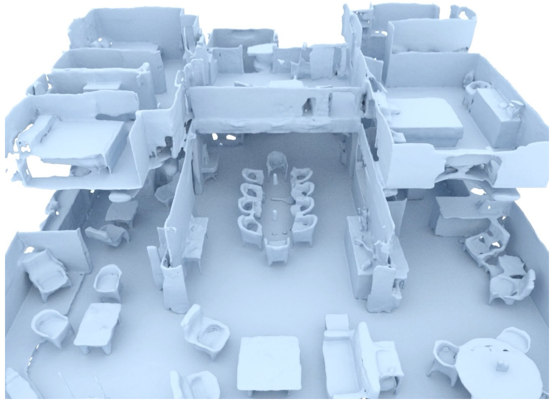
OpenScene
CVPR 2023

This Thesis

Develop 3D Neural Scene Representations

for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes



ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference

Shape As Points
NeurIPS 2021 (Oral)

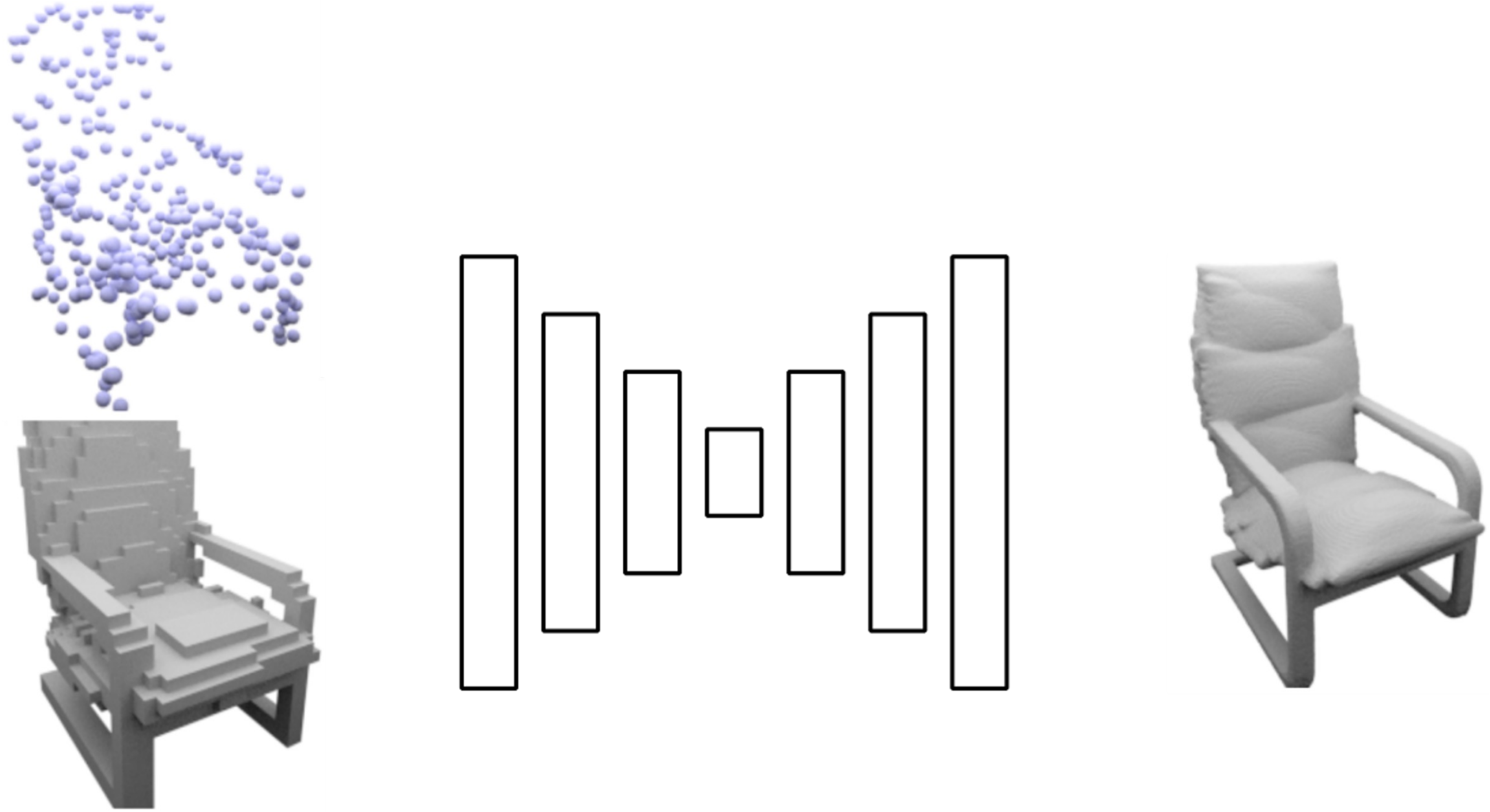
3. From 2D Observations

NICE-SLAM
CVPR 2022

4. Arbitrary Queries

OpenScene
CVPR 2023

Learning-based 3D Reconstruction

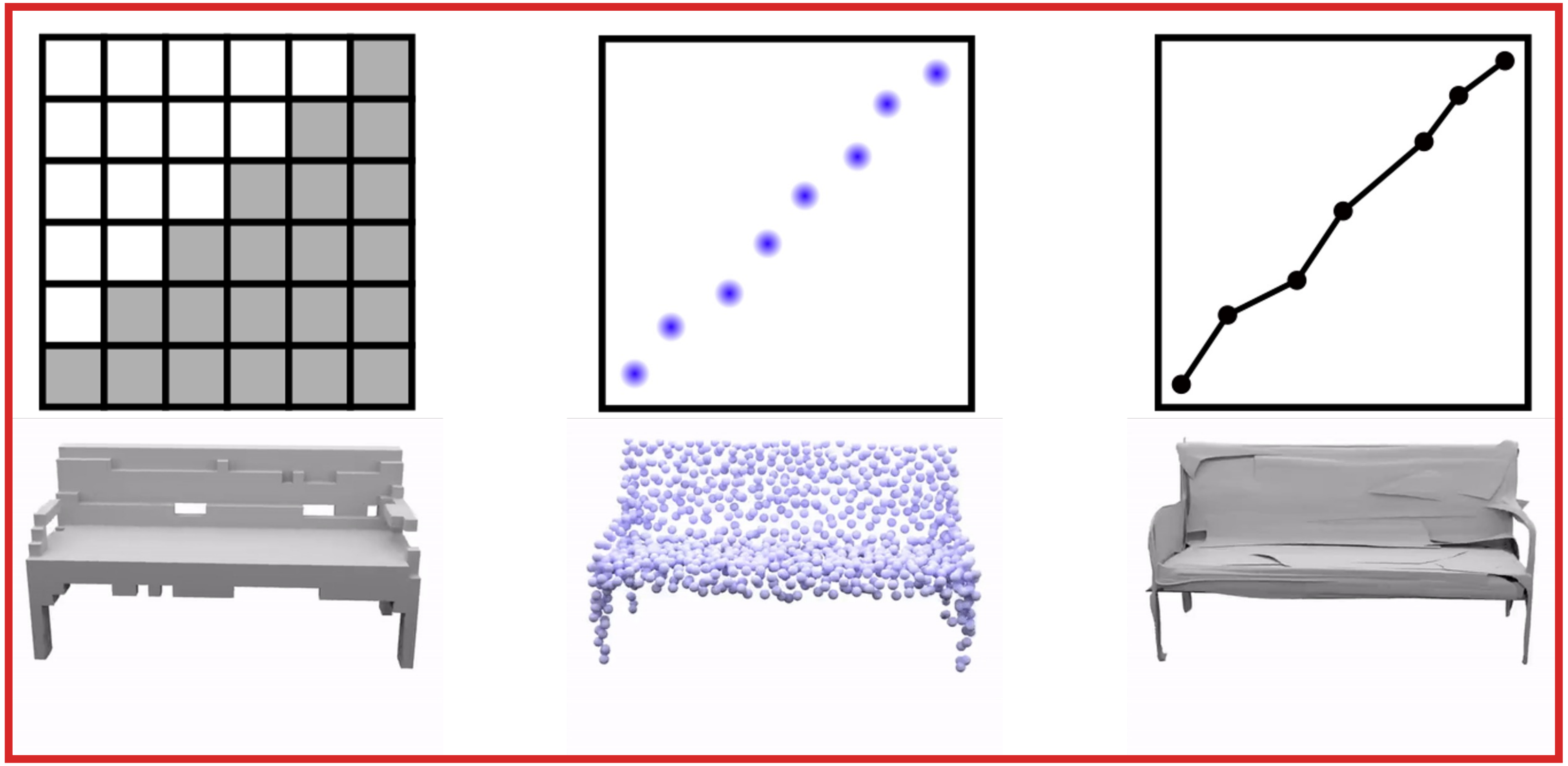


Input   Neural Network   3D Reconstruction

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

3D Representations

Traditional Explicit Representations



Discretization

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

3 Seminal Papers at the Same CVPR!

Neural Implicit Representations

Occupancy Networks: Learning 3D Reconstruction in Function Space

Lars Mescheder[1] Michael Oechsle[1,2] Michael Niemeyer[1] Sebastian Nowozin[3†] Andreas Geiger[1]
[1]Autonomous Vision Group, MPI for Intelligent Systems and University of Tübingen
[2]ETAS GmbH, Stuttgart
[3]Google AI Berlin

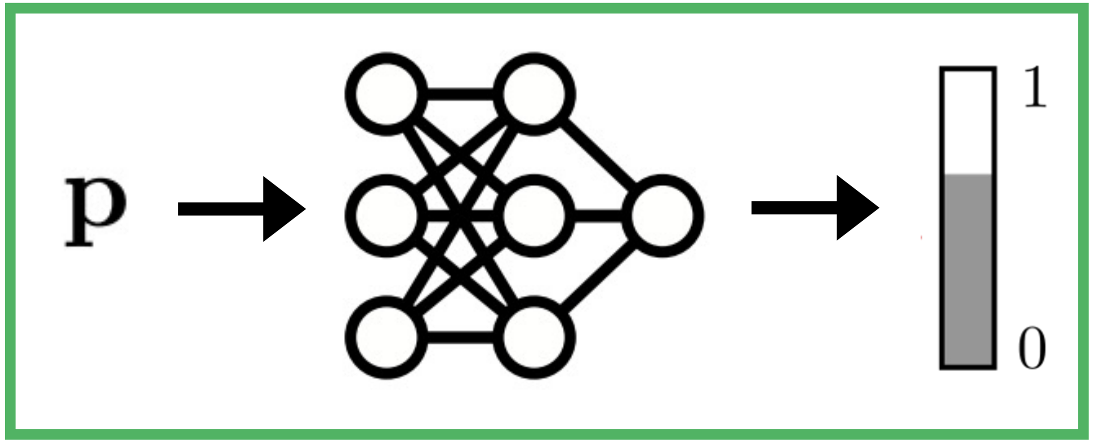
DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation

Jeong Joon Park[1,3†] Peter Florence [2,3†] Julian Straub[3] Richard Newcombe[3] Steven Lovegrove[3]
[1]University of Washington [2]Massachusetts Institute of Technology [3]Facebook Reality Labs

Learning Implicit Fields for Generative Shape Modeling

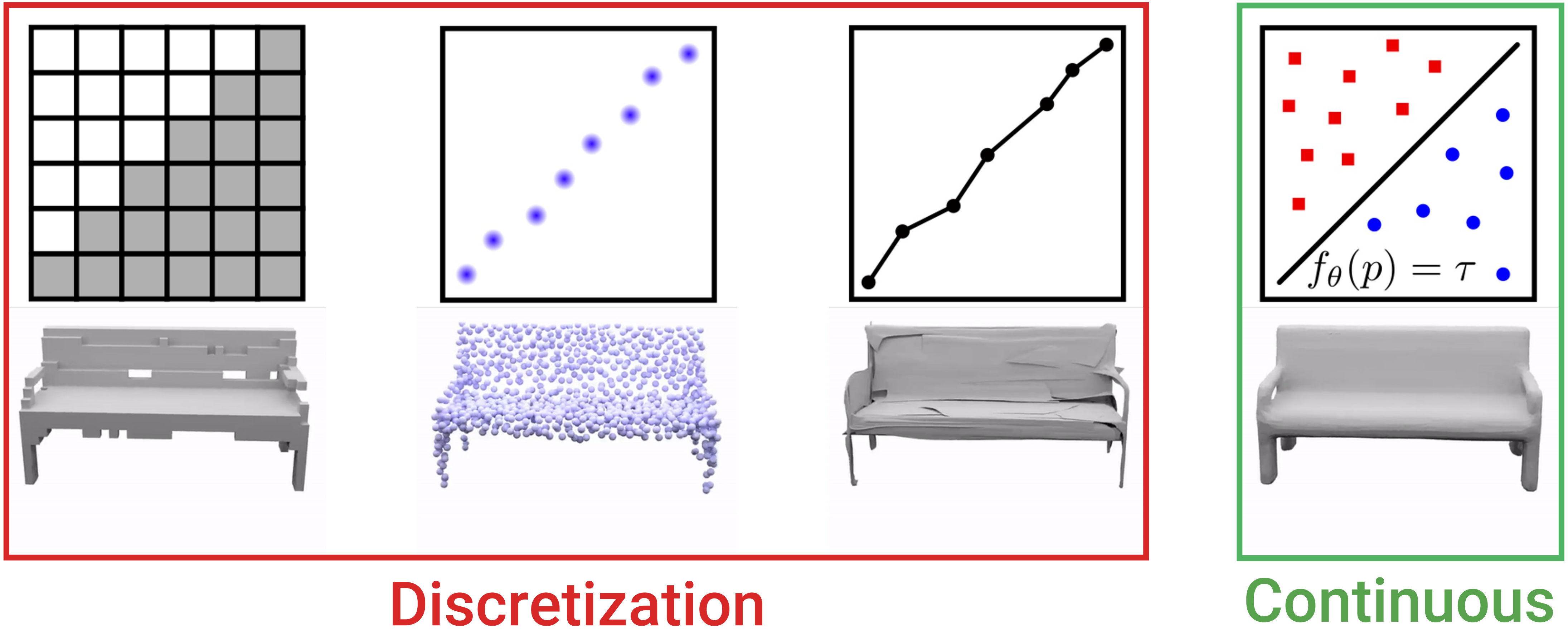
Zhiqin Chen
Simon Fraser University
zhiqinc@sfu.ca

Hao Zhang
Simon Fraser University
haoz@sfu.ca



3D Representations

Neural Implicit Representations

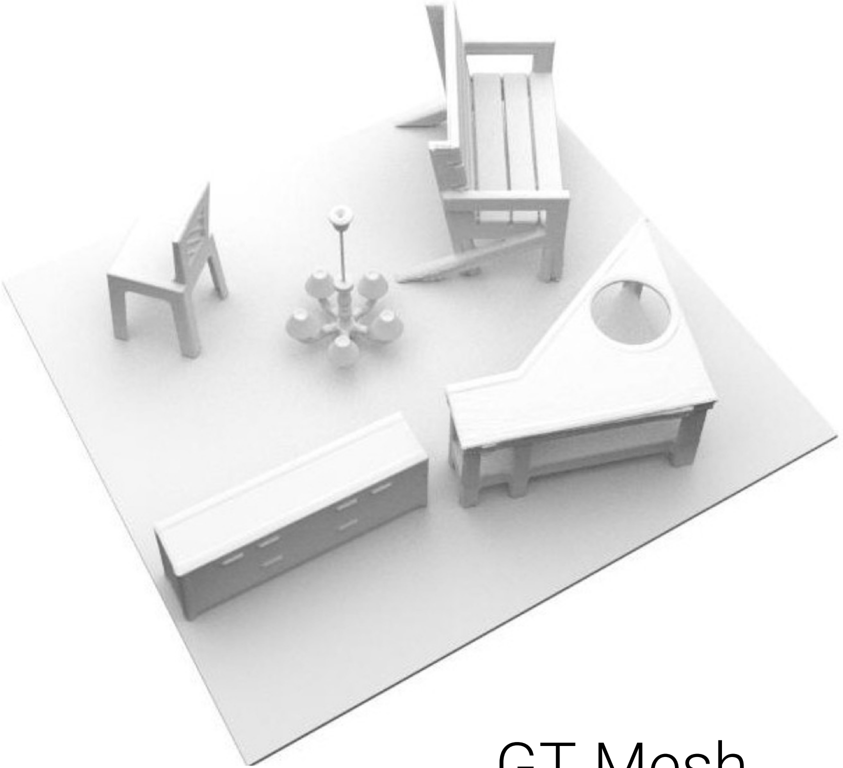


Mescheder, Oechsle, Niemeyer, Nowozin and Geiger: Occupancy Networks: Learning 3D Reconstruction in Function Space. CVPR, 2019

Limitations

Neural Implicit Representations

Works well for **simple objects**, but poorly on **complex scenes**



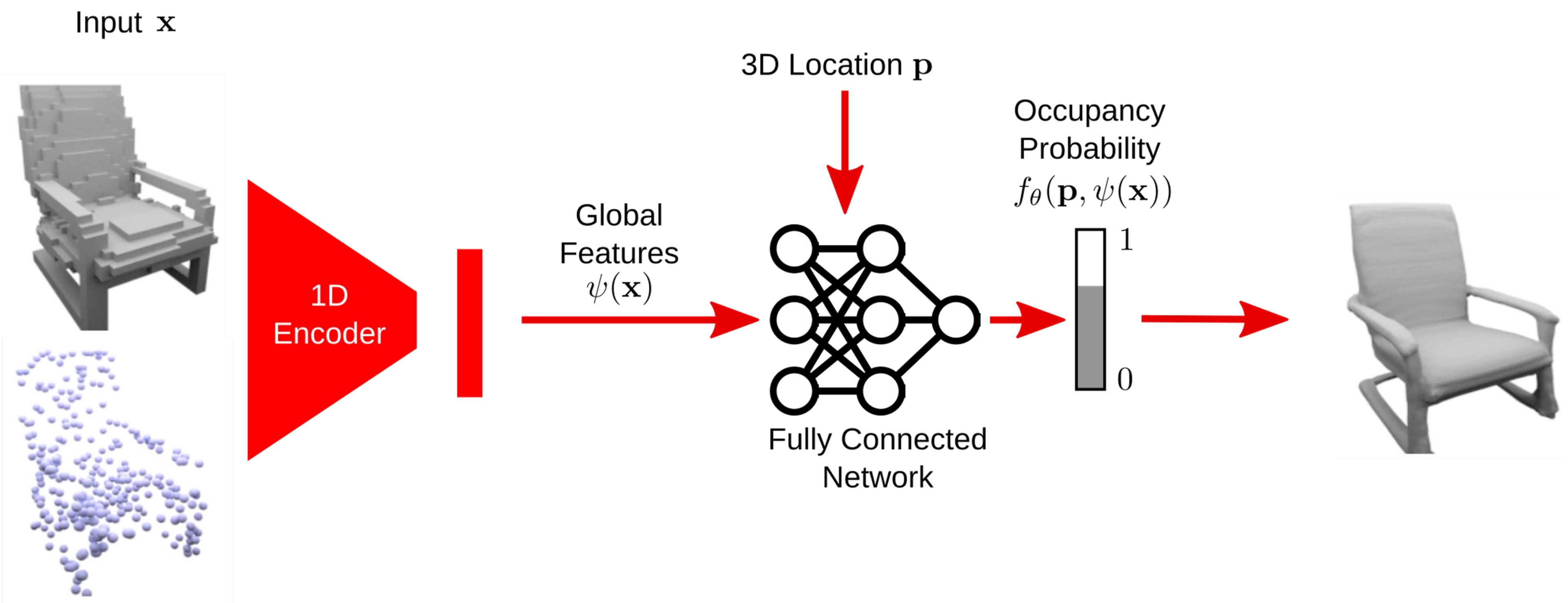
OccNet

GT Mesh

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

Limitations

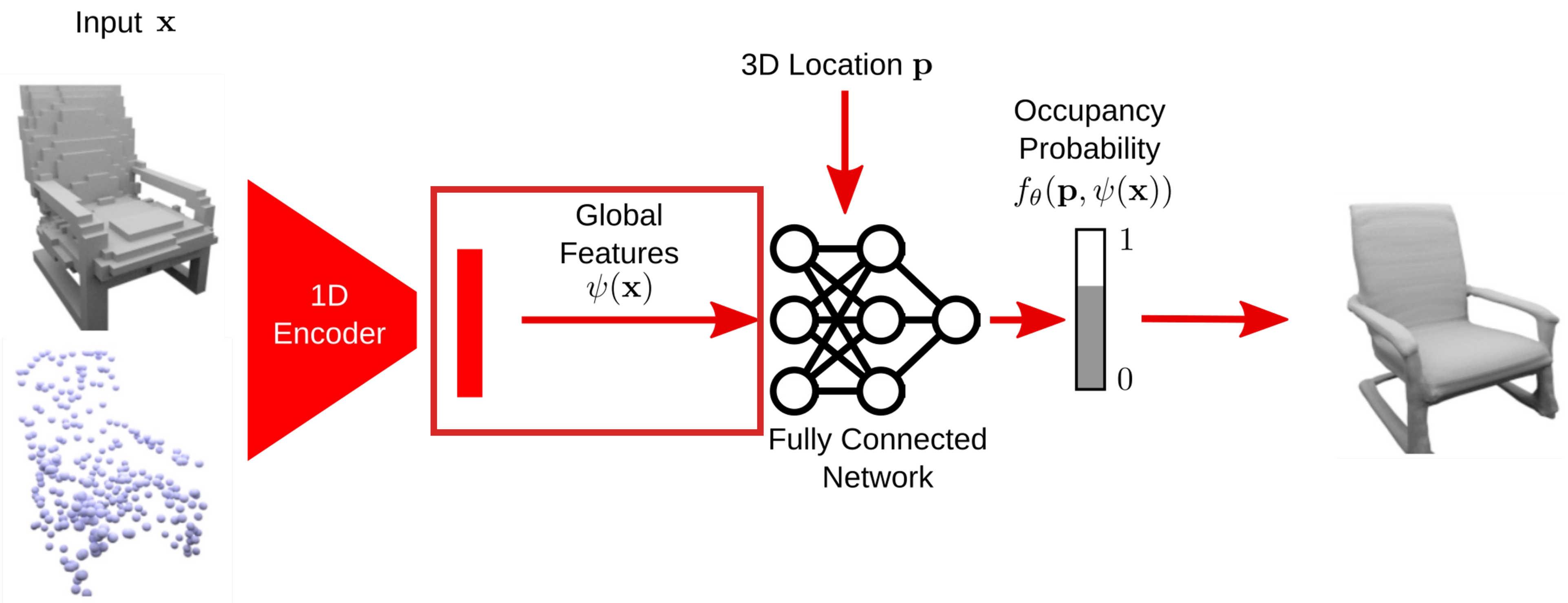
Neural Implicit Representations



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

Limitations

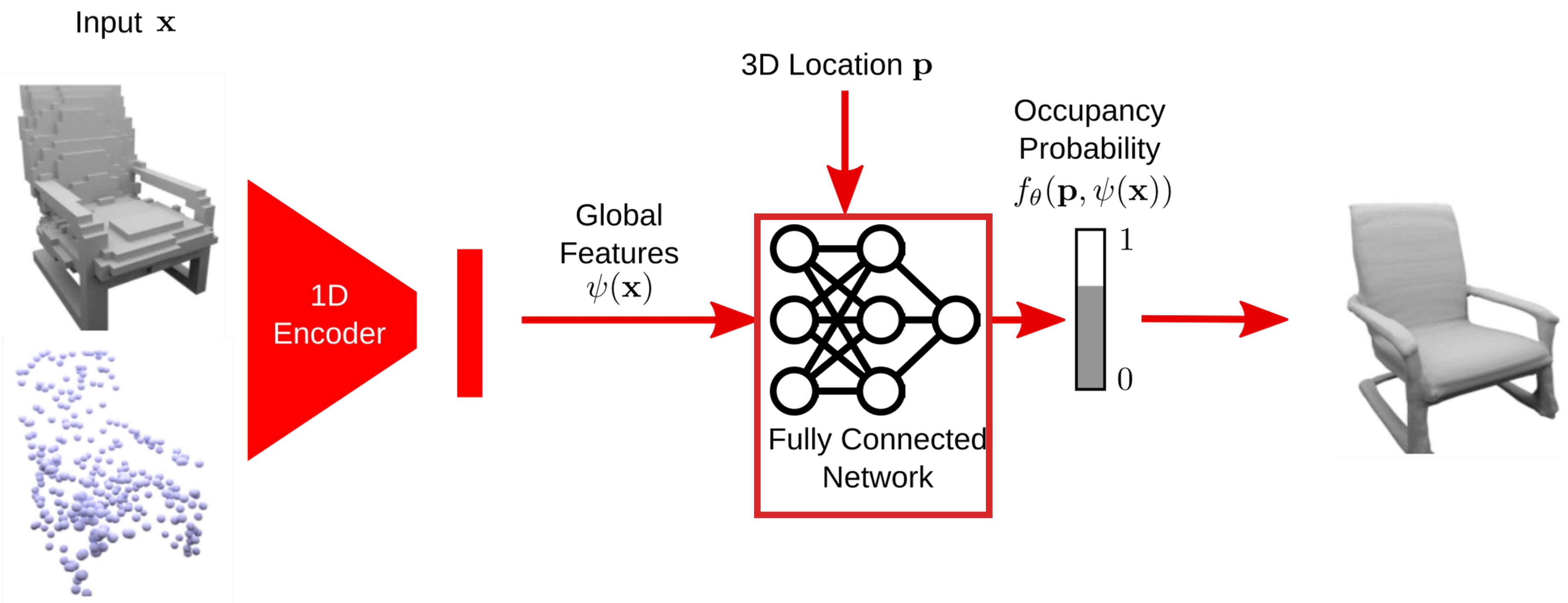
Neural Implicit Representations



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

Limitations

Neural Implicit Representations



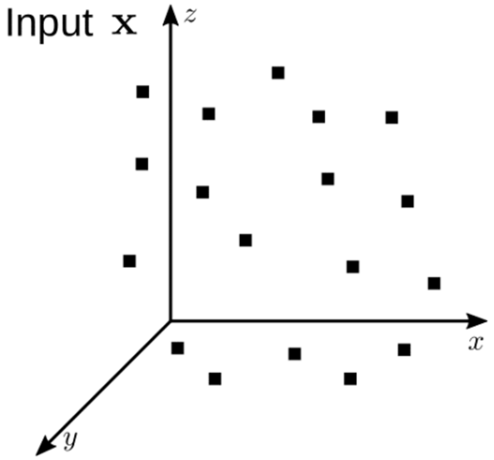
- Global latent code ⇒ **overly smooth geometry**
- Fully-connected architecture ⇒ **no translation equivariance**

How to reconstruct large-scale 3D scenes with

neural implicit representations?

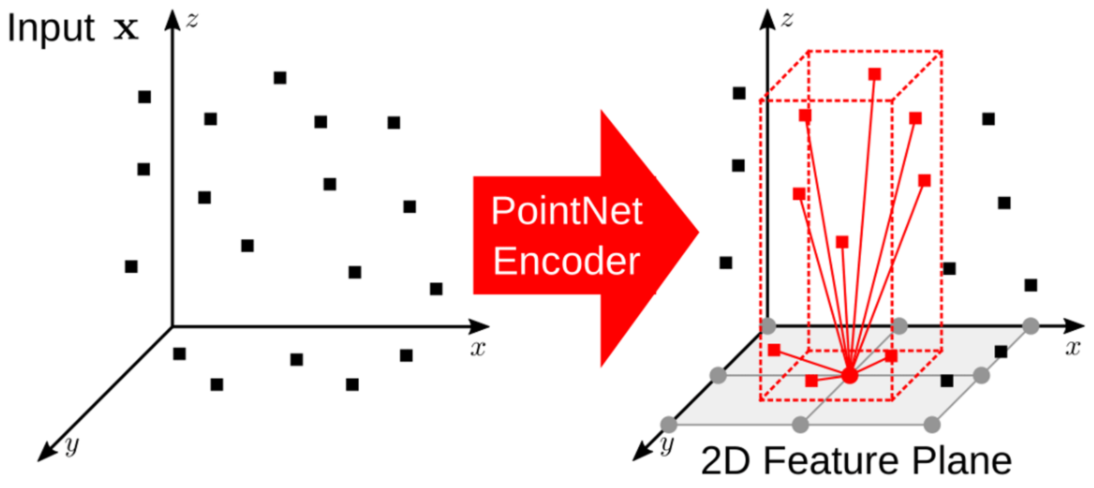
Main Idea

Convolutional Occupancy Networks



Main Idea

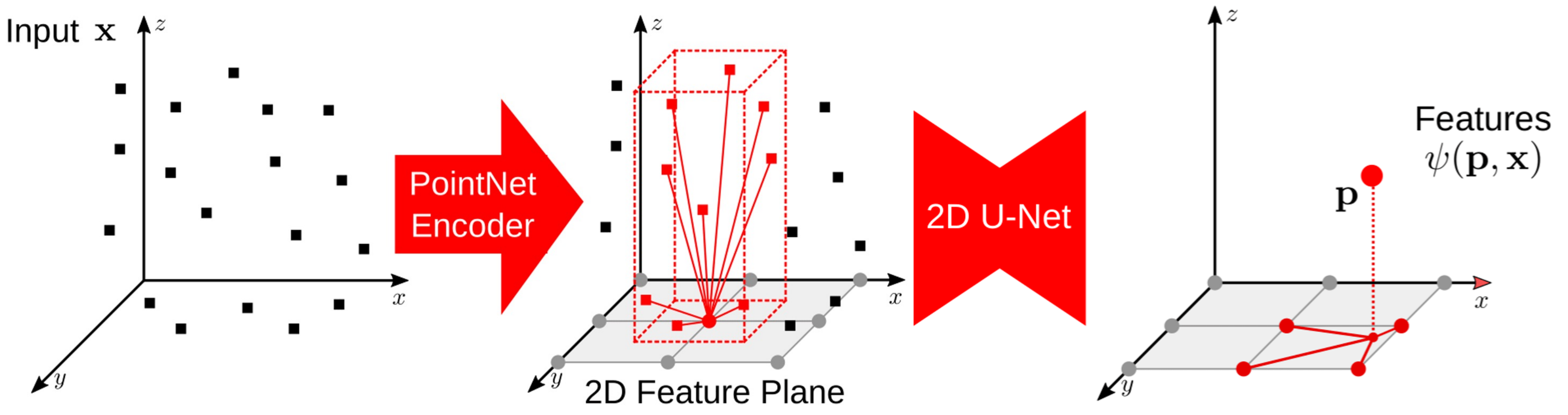
Convolutional Occupancy Networks



- **2D Plane Encoder**: Project point features onto the canonical plane

Main Idea

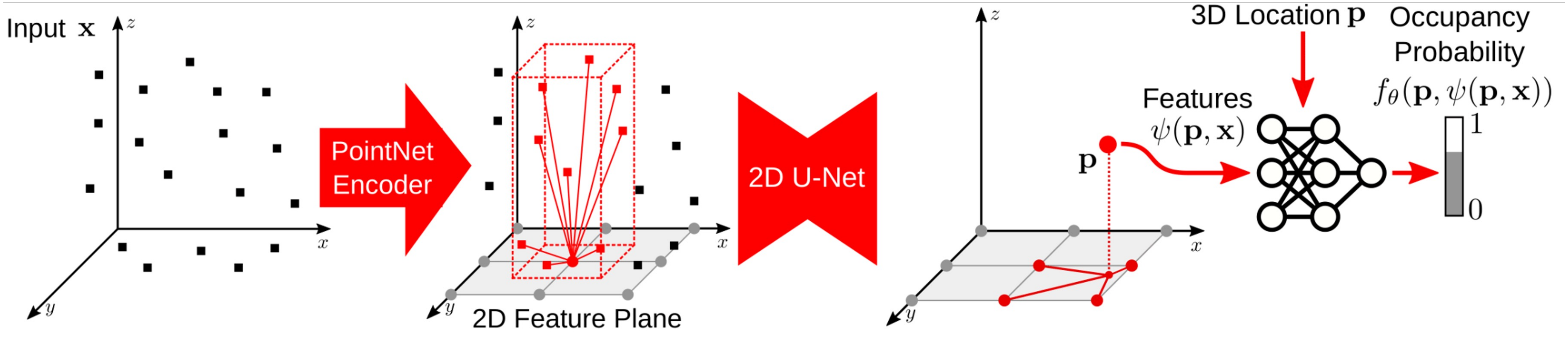
Convolutional Occupancy Networks



- **2D Plane Encoder**: Project point features onto the canonical plane
- **2D Plane Decoder**: Processed by UNet, query features via bilinear interpolation

Main Idea

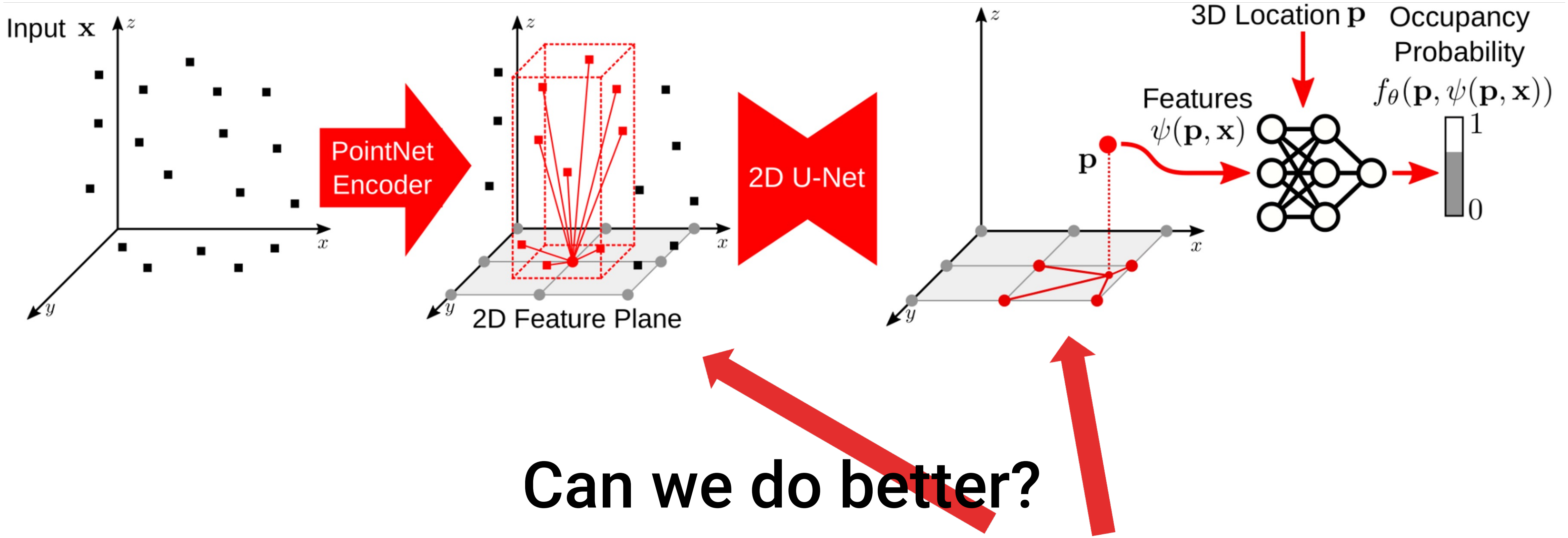
Convolutional Occupancy Networks



- **2D Plane Encoder**: Project point features onto the canonical plane
- **2D Plane Decoder**: Processed by UNet, query features via bilinear interpolation
- **Occupancy Net**: Shallow MLP $f_\theta(\cdot)$

Main Idea

Convolutional Occupancy Networks

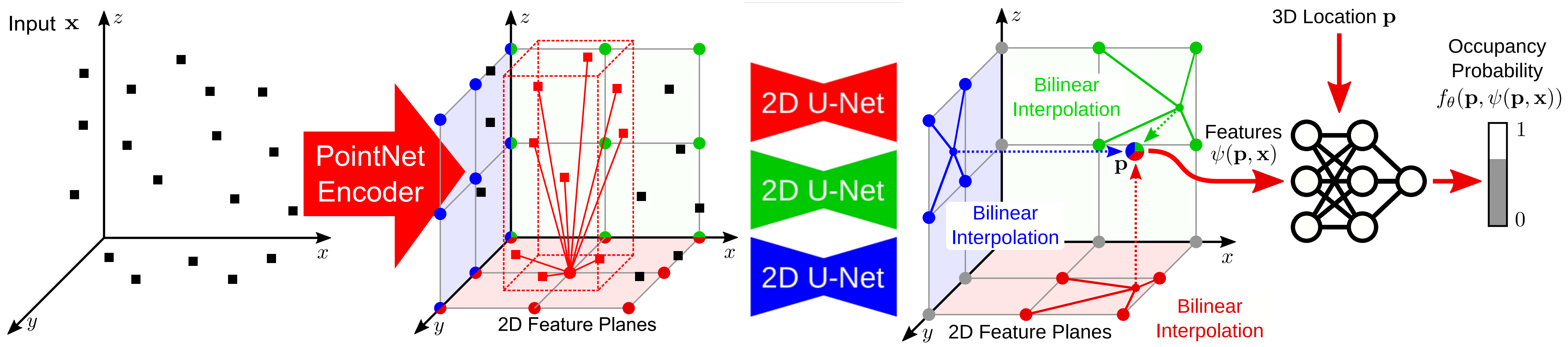


Can we do better?

For now, features only on the ground plane…

Main Idea – "Tri-plane"

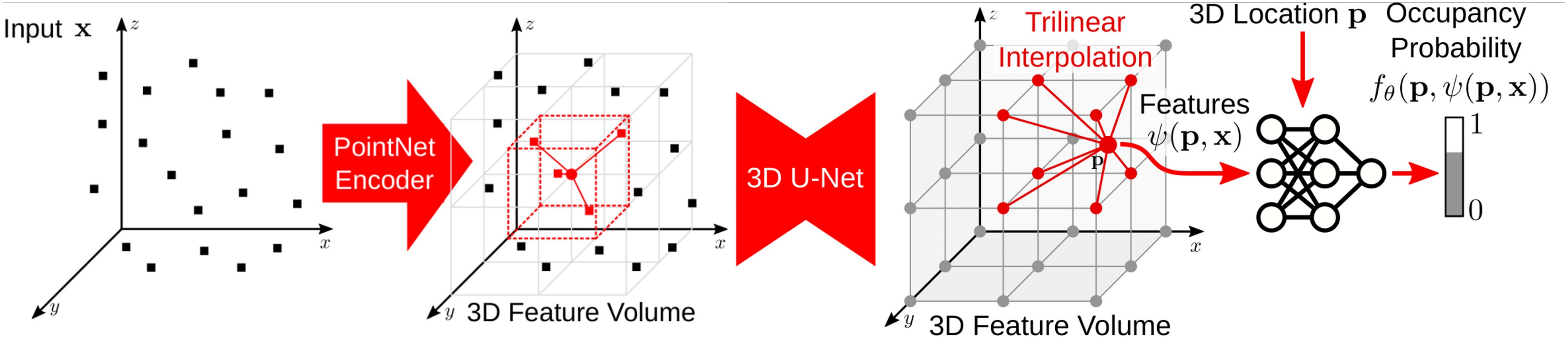
Convolutional Occupancy Networks



Project features on X, Y, Z **canonical planes**

Main Idea – 3D Volume

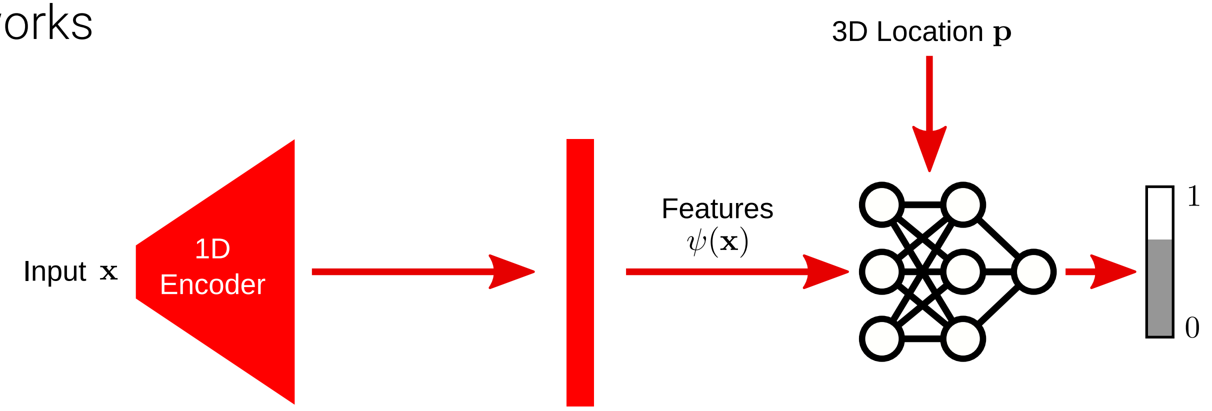
Convolutional Occupancy Networks



Encode local information into a **3D feature volume**

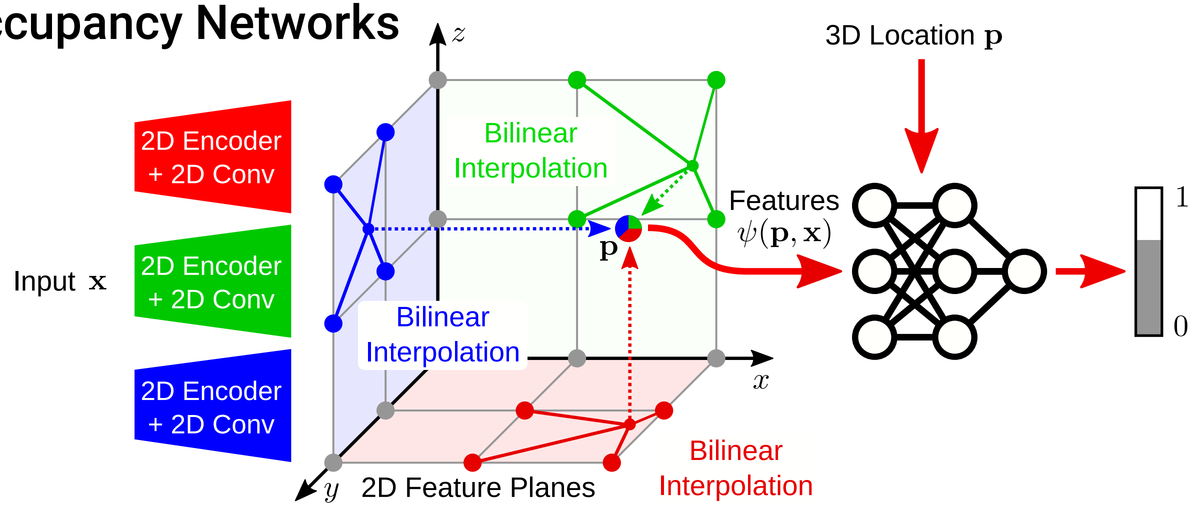
Comparison

Occupancy Networks



- global feature
- heavy FC network
- no translation equivariance

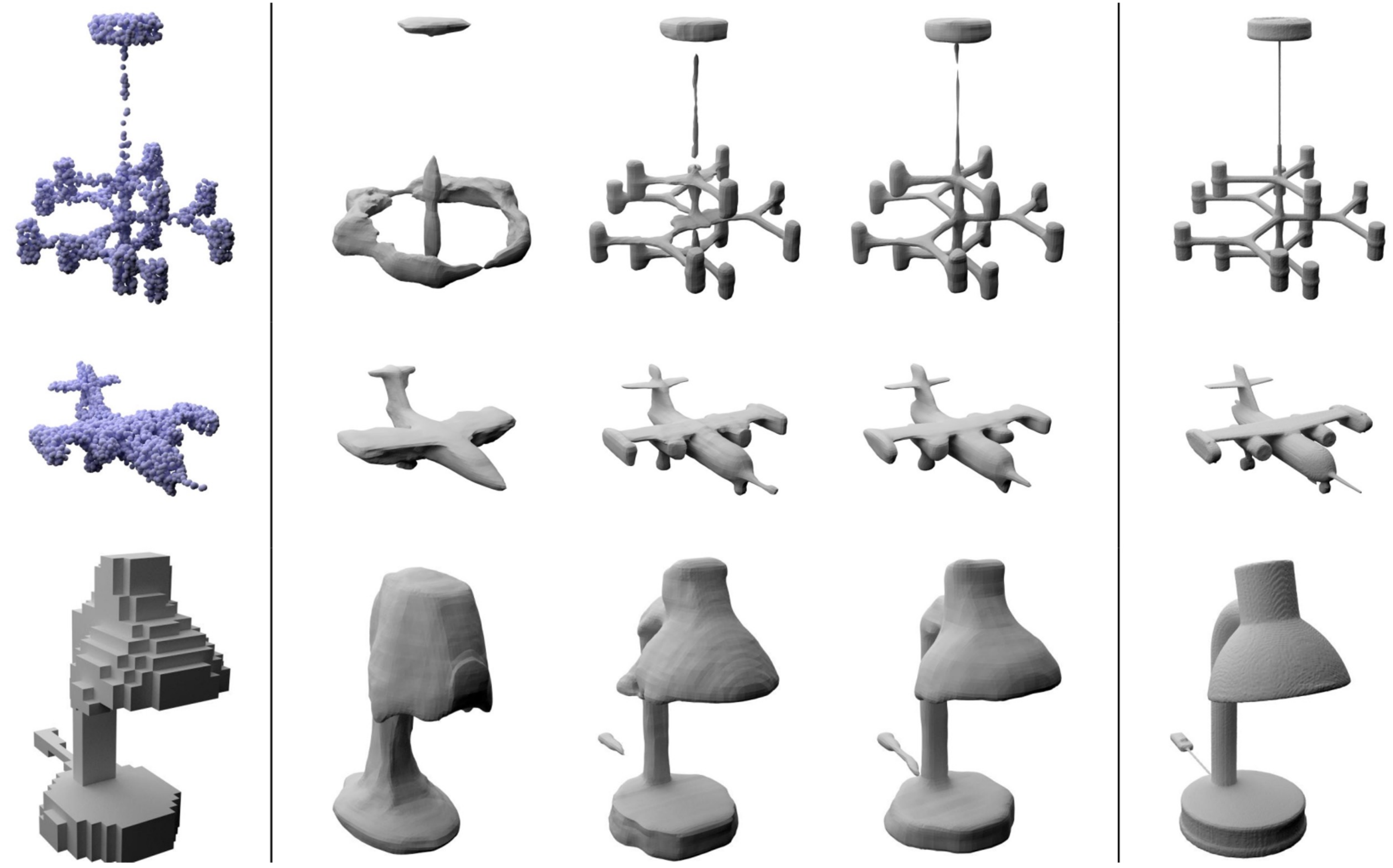
Convolutional Occupancy Networks



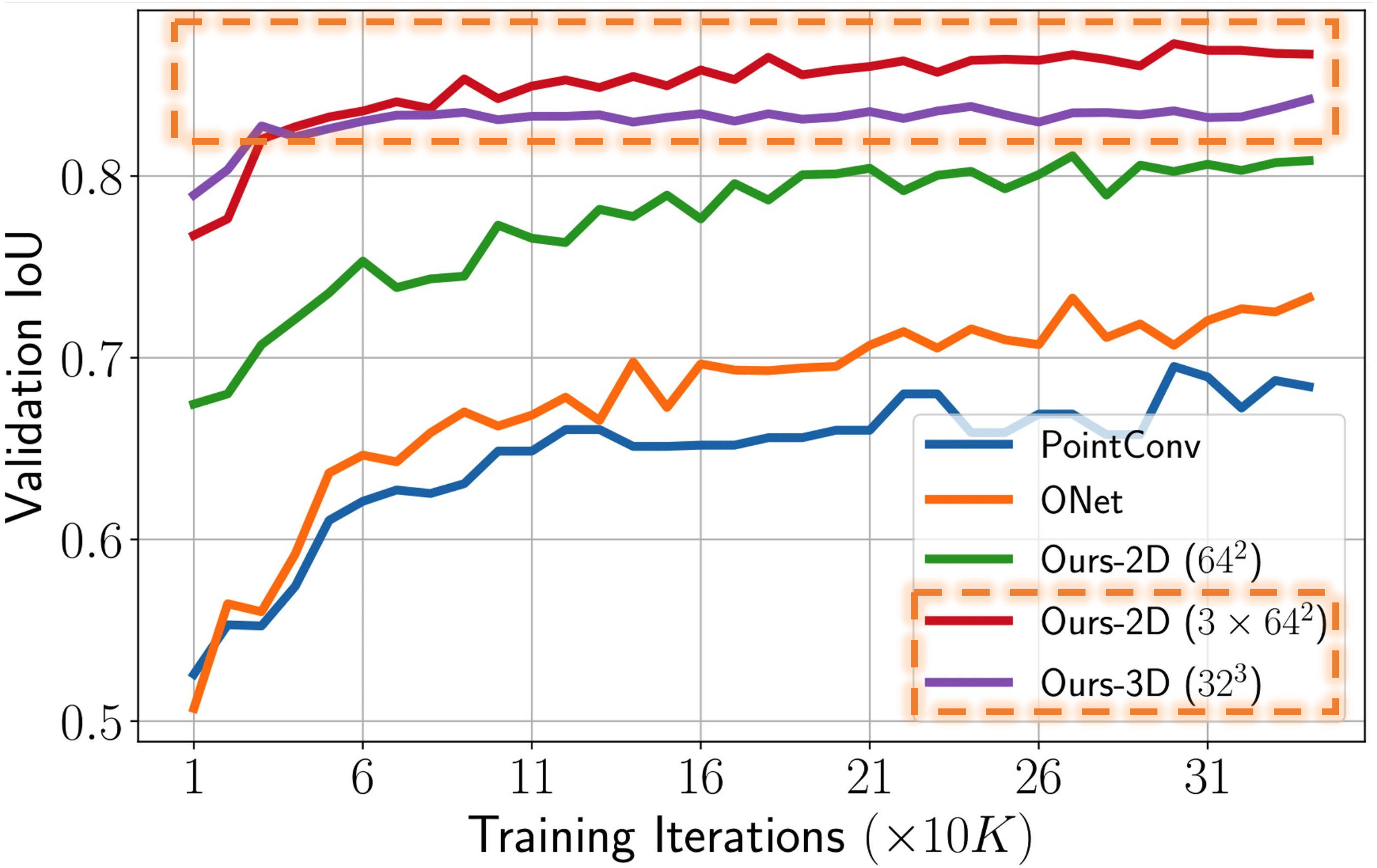
+ local feature
+ shallow FC network
+ translation equivariance

Results

Object-Level Reconstruction

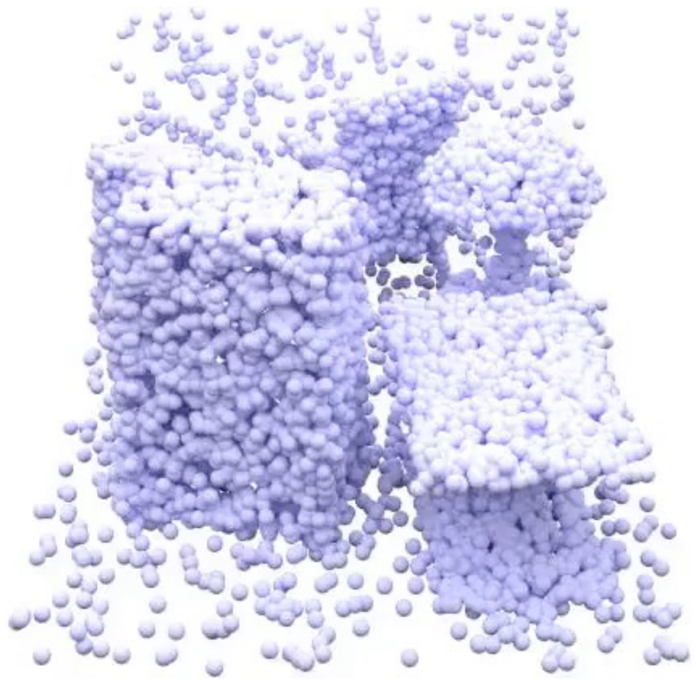


Training Speed



Scene-Level Reconstruction

Train and evaluate on synthetic room

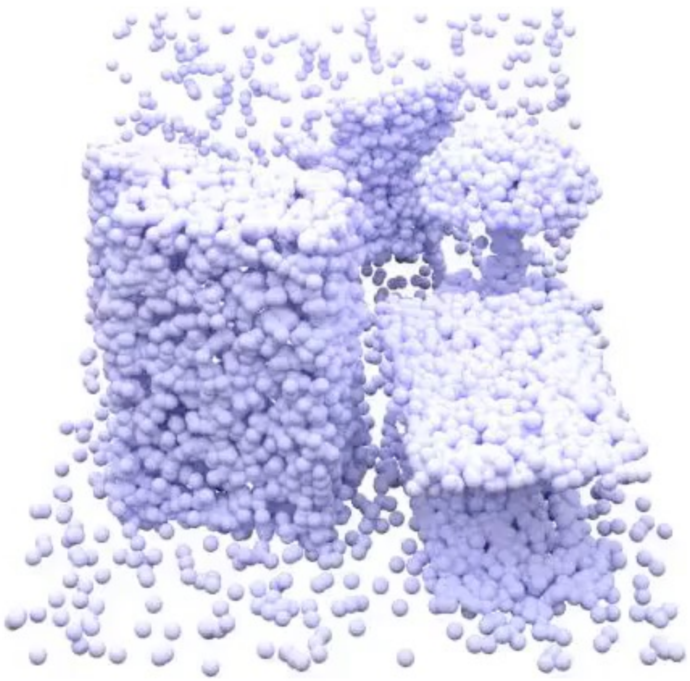


Input

GT Mesh

Scene-Level Reconstruction

OccNet fails on room-level reconstruction



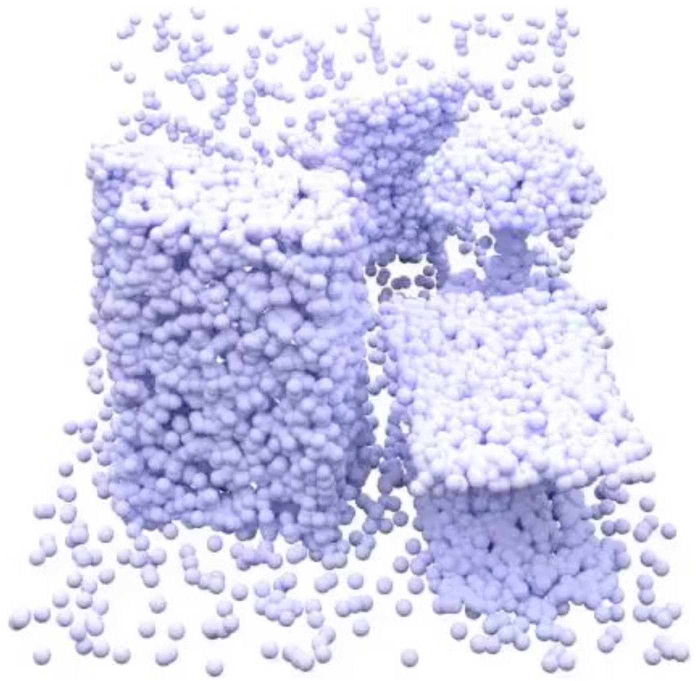
Input

OccNet

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

Scene-Level Reconstruction

SPSR requires surface normal, output is noisy



Input

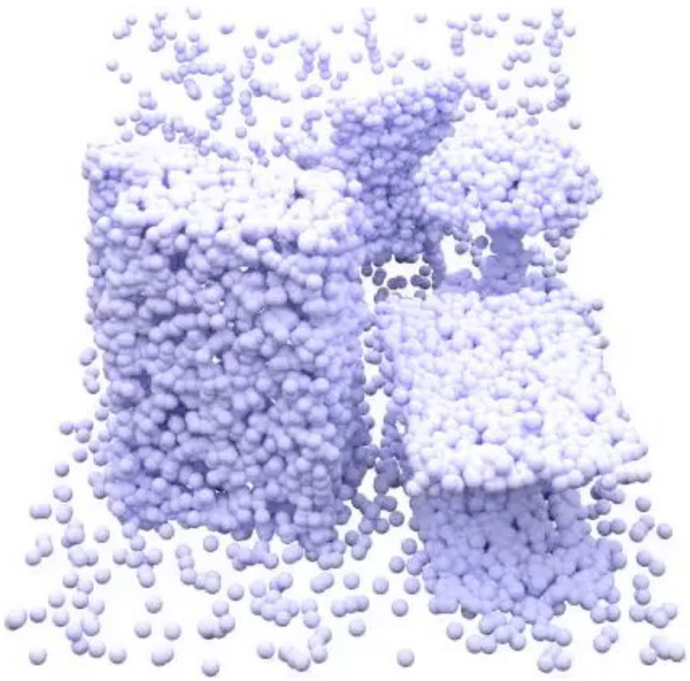
SPSR

(Screened Poisson Surface Reconstruction)

Kazhdan and Hoppe: Screened Poisson Surface Reconstruction. ToG, 2013

Scene-Level Reconstruction

Ours preserves better details



Input

Ours

Peng, Niemeyer, Mescheder, Pollefeys, and Geiger: Convolutional Occupancy Networks. ECCV, 2020

Scene-Level Reconstruction

OccNet

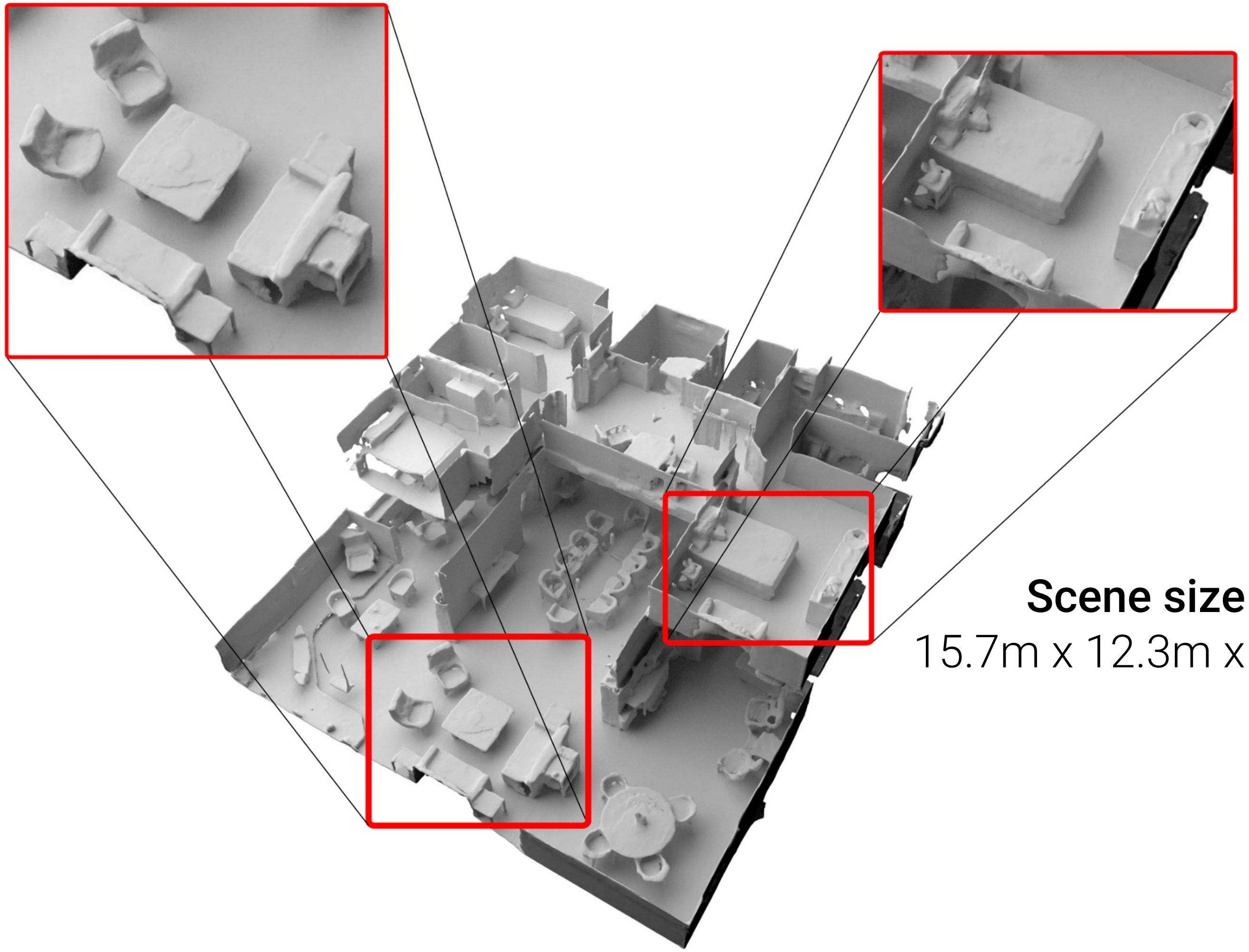
SPSR

Ours

Large-Scale Reconstruction

Reconstruct a big house in Matterport3D

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

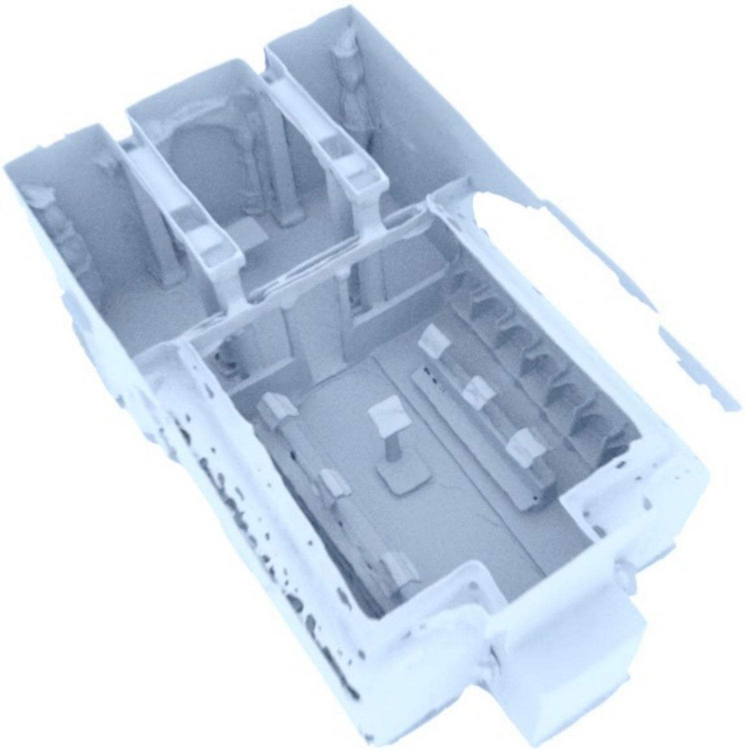


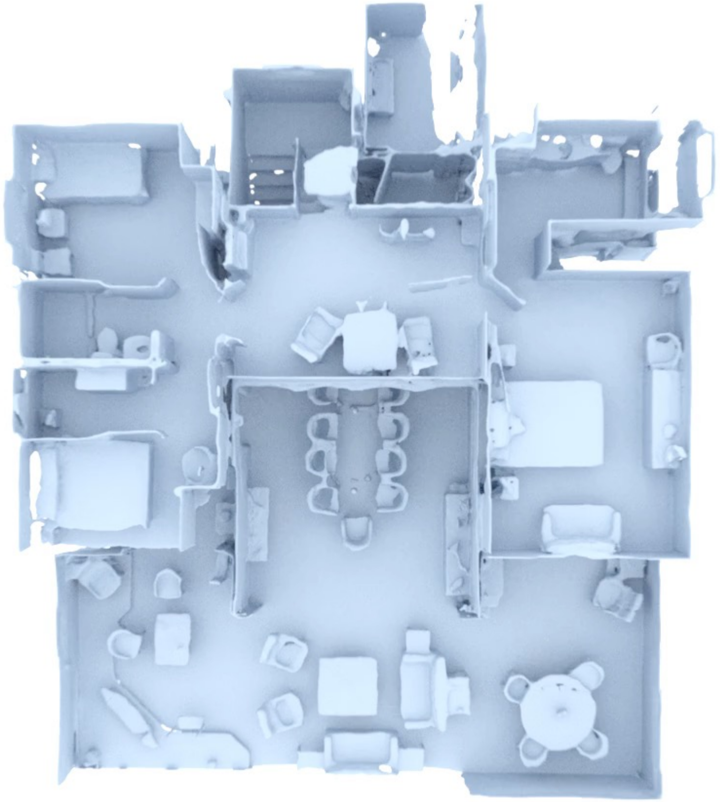
Scene size
15.7m x 12.3m x 4.5m

Our reconstruction output

Large-Scale Reconstruction

Reconstruct large-scale scenes in Matterport3D





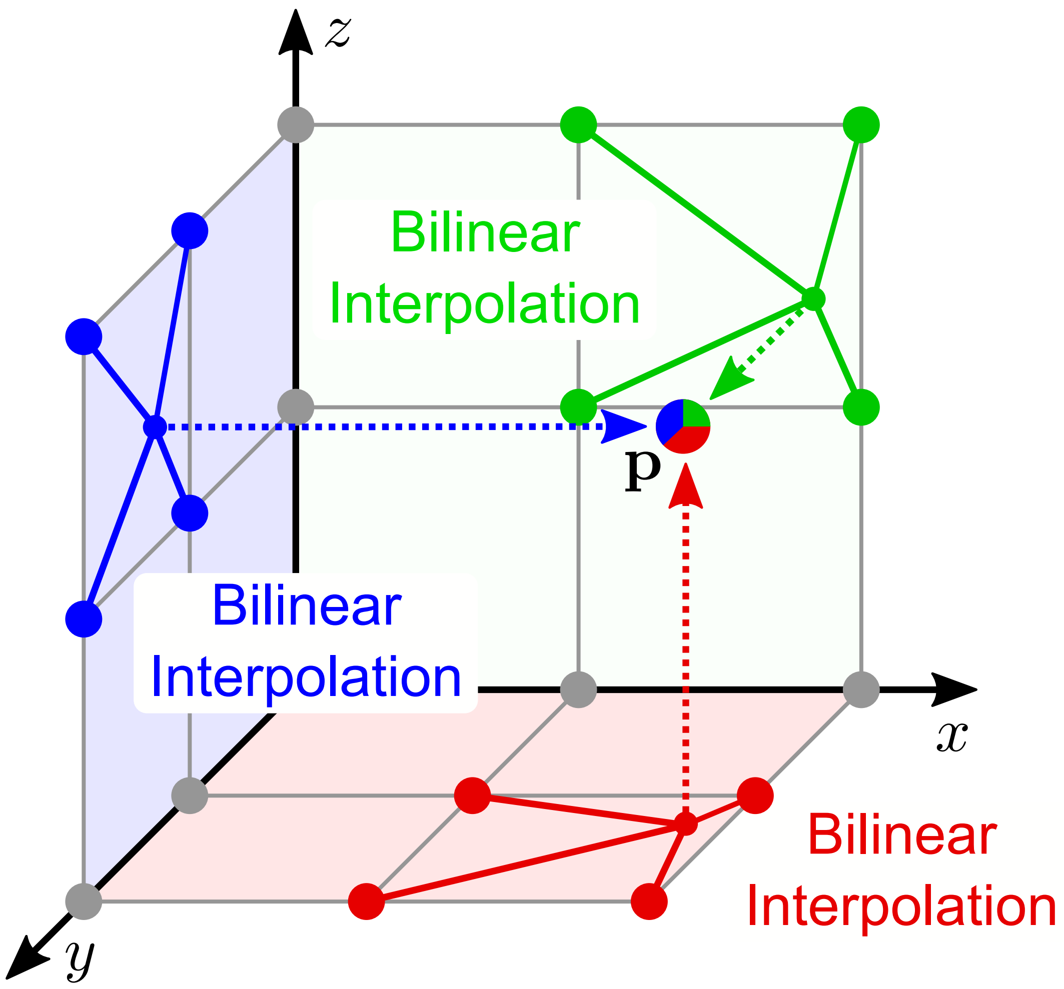
ConvOccNet - TL;DR

- Three **hybrid representations** for neural fields

  a) Ground plane   b) **Tri-plane**   c) 3D volume

- CNN's **translation equivariance** rocks
- **Synthetic-to-real** generalization

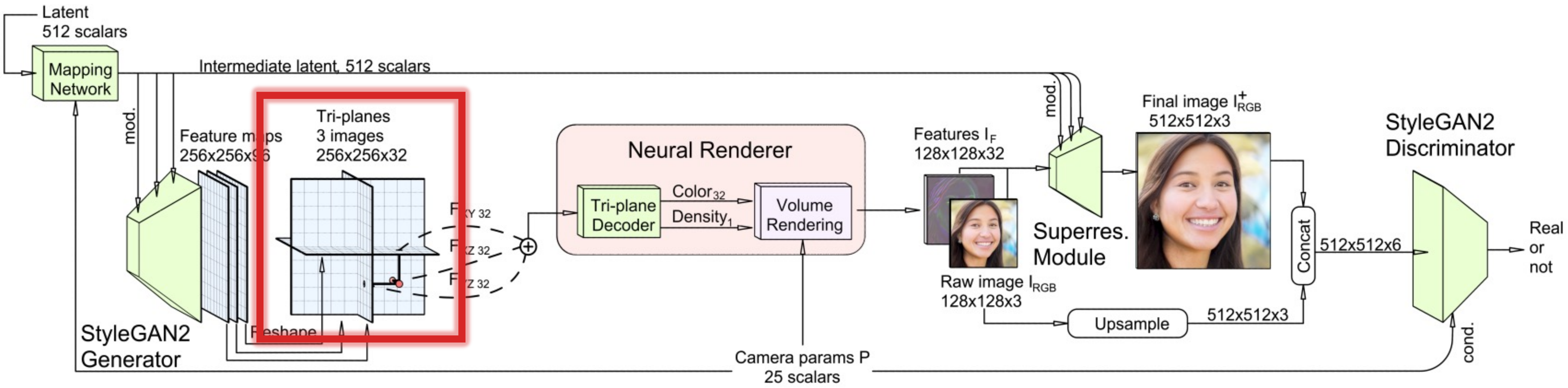
"Tri-plane" Representations



Reviewer 2: *"What is the point of having that 3-plane representation?"*

“Tri-plane” Representations

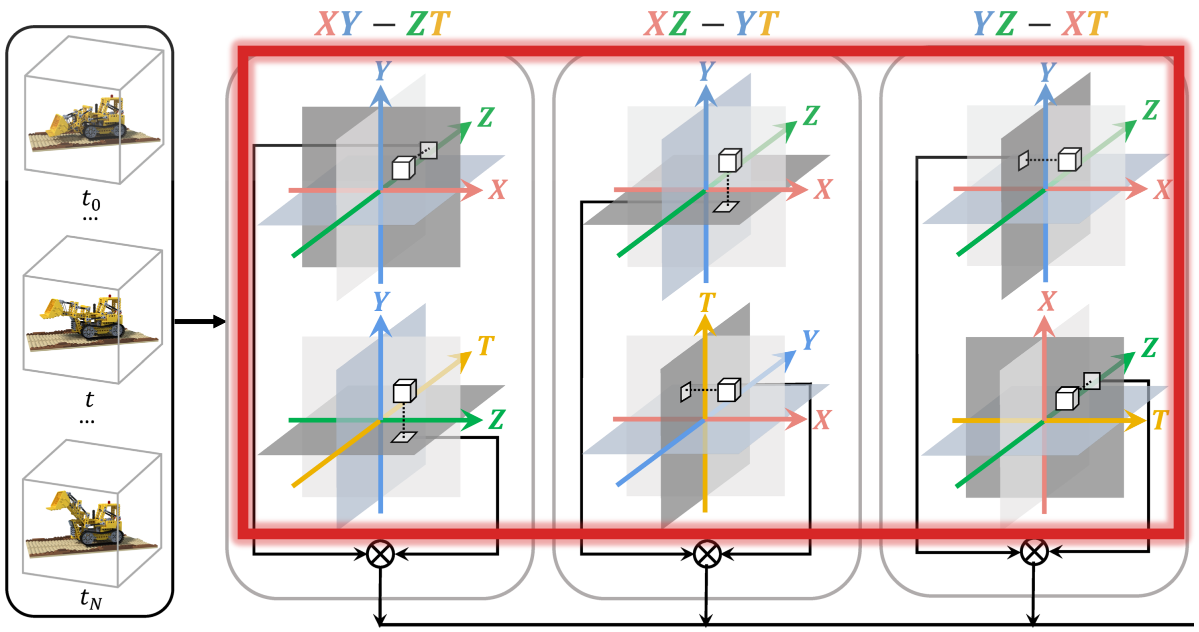
High Fidelity 3D-Aware View Synthesis



Chan et al.: Efficient Geometry-aware 3D Generative Adversarial Networks. CVPR 2022

"Tri-plane" Representations

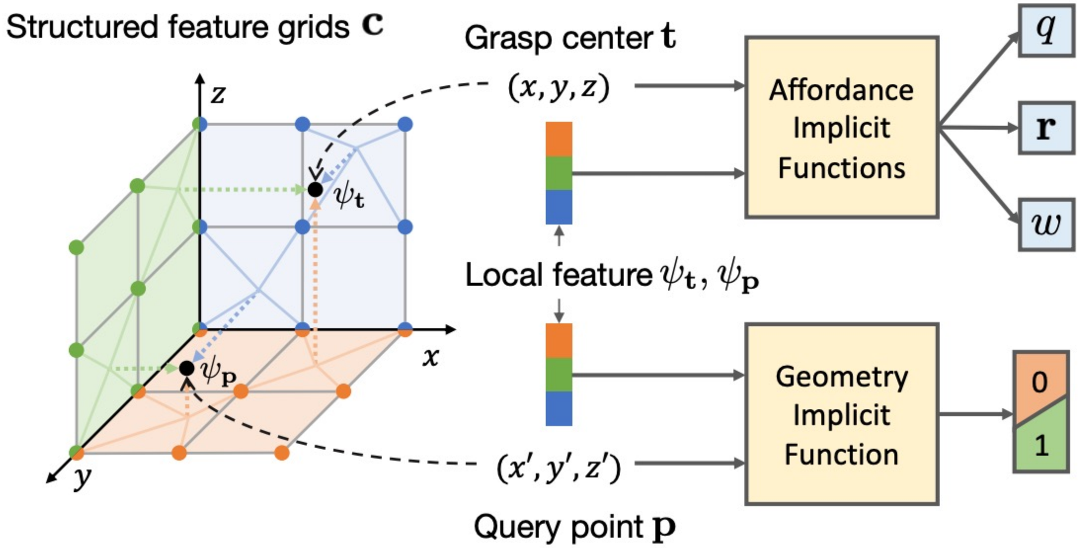
Efficient 4D View Synthesis



Cao and Johnson: HexPlane: A Fast Representation for Dynamic Scenes. CVPR 2023

“Tri-plane” Representations

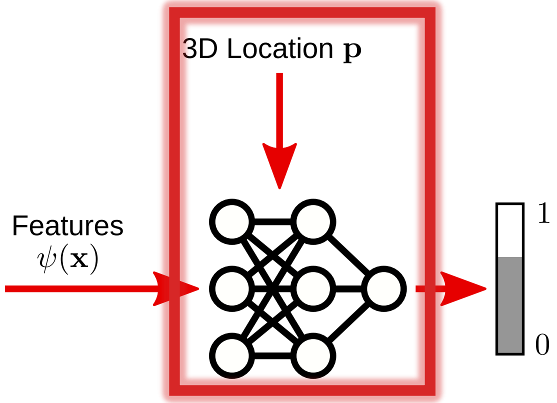
Robot Grasping



Jiang et al.: Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations. RSS 2021

ConvOccNet - Limitations

— Very slow inference



For a grid of 128^3, > 2 million MLP forward passes !

Shape As Points: A Differentiable Poisson Solver

Songyou Peng[1,2] **Chiyu "Max" Jiang**[*†] **Yiyi Liao**[2,3†] **Michael Niemeyer**[2,3]
Marc Pollefeys[1,4] **Andreas Geiger**[2,3]

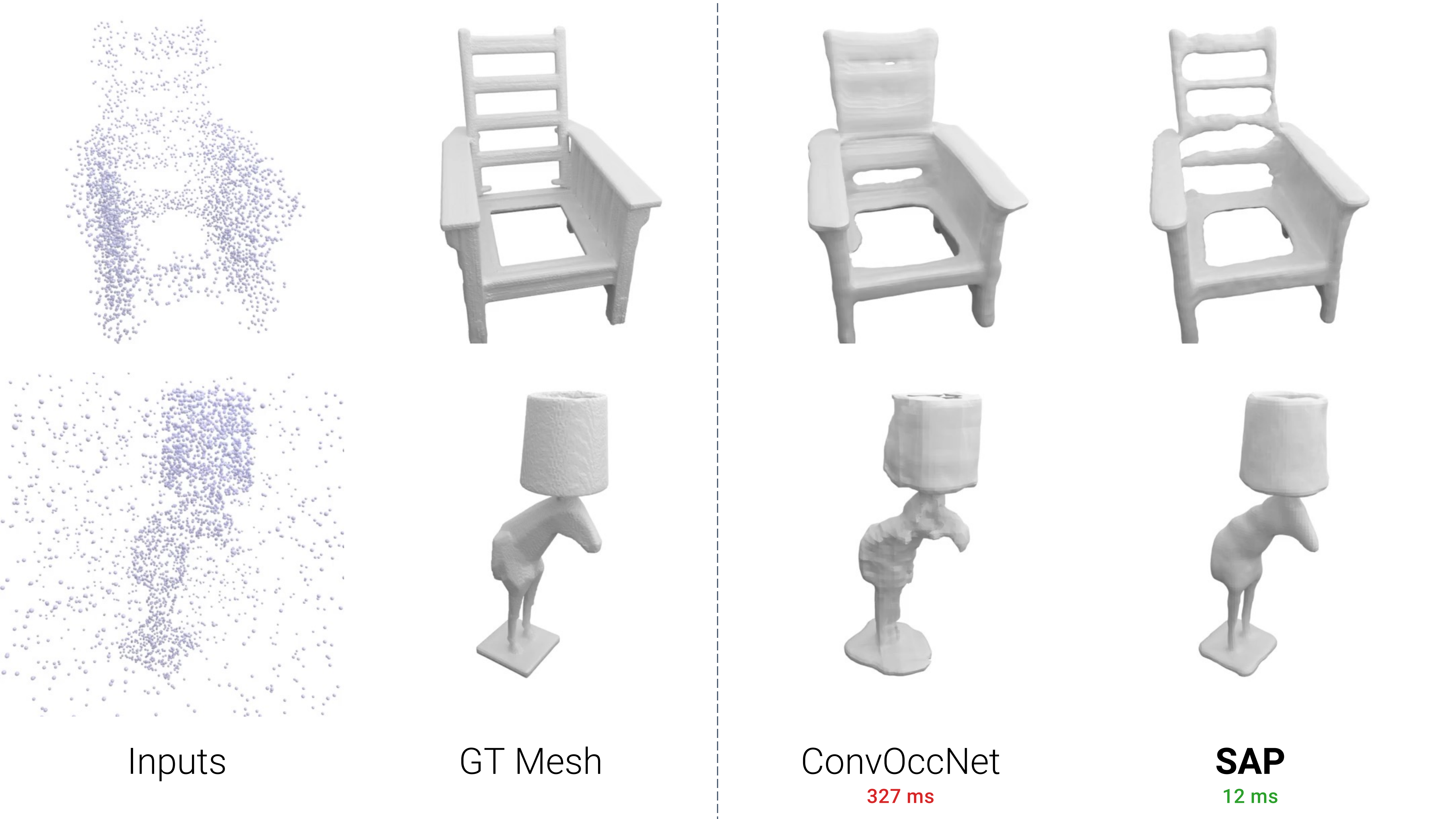
[1]ETH Zurich [2]Max Planck Institute for Intelligent Systems, Tübingen
[3]University of Tübingen [4]Microsoft

Shape As Points

A Differentiable Point-to-Mesh Layer

No network evaluation, fast!

Peng, Jiang, Liao, Niemeyer, Pollefeys, and Geiger: Shape As Points: A Differentiable Poisson Solver. NeurIPS 2021


Inputs
GT Mesh
ConvOccNet
327 ms
SAP
12 ms

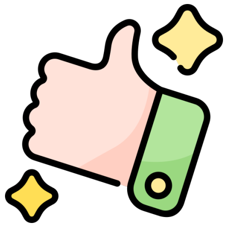
ConvOccNet - Limitations

- **Very slow inference**
  - 😊 We have **SAP**
- **Only reconstruct from 3D noisy point clouds**

Can we **online** reconstruct purely **from 2D observations**?

ConvOccNet - Limitations

- Very slow inference
  - 😊 We have **SAP**
- Only reconstruct from 3D noisy point clouds



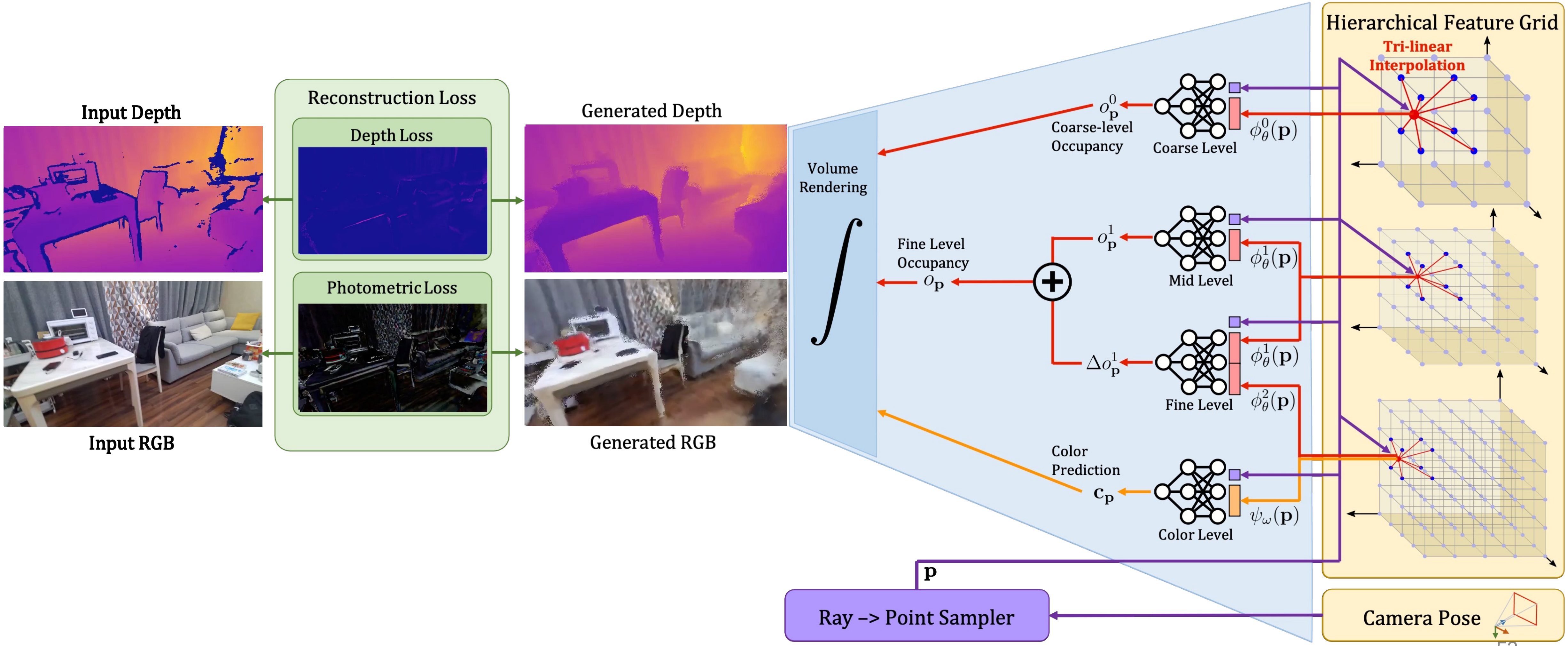
NICE-SLAM: Neural Implicit Scalable Encoding for SLAM

Zihan Zhu[1,2*] Songyou Peng[2,4*] Viktor Larsson[3] Weiwei Xu[1] Hujun Bao[1]
Zhaopeng Cui[1†] Martin R. Oswald[2,5] Marc Pollefeys[2,6]

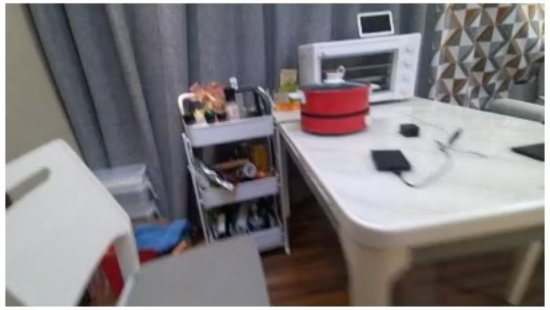
[1]State Key Lab of CAD&CG, Zhejiang University [2]ETH Zurich [3]Lund University
[4]MPI for Intelligent Systems, Tübingen [5]University of Amsterdam [6]Microsoft

NICE-SLAM

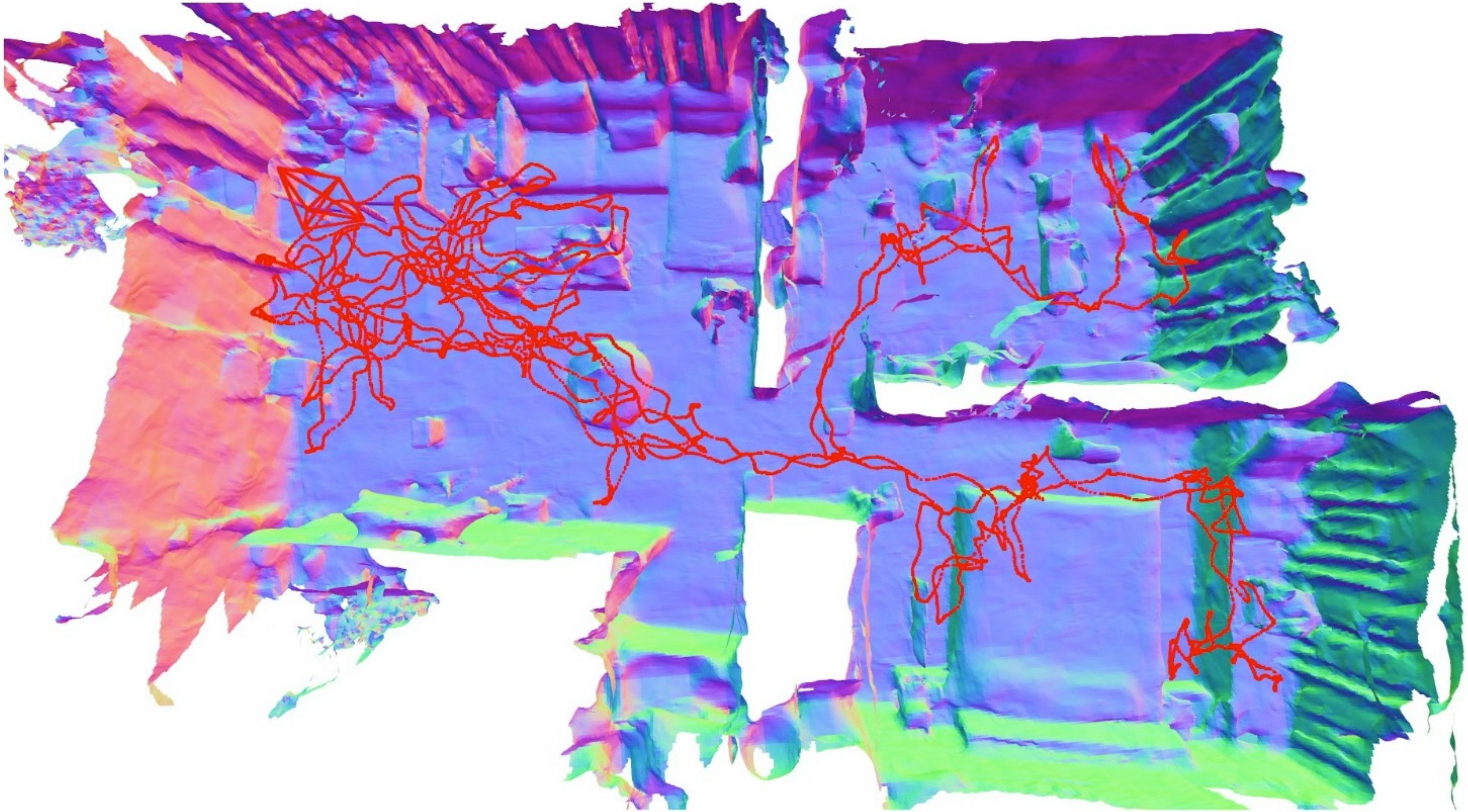
Neural Implicit Scalable Encoding for SLAM



RGB-D Sequences



40x Speed

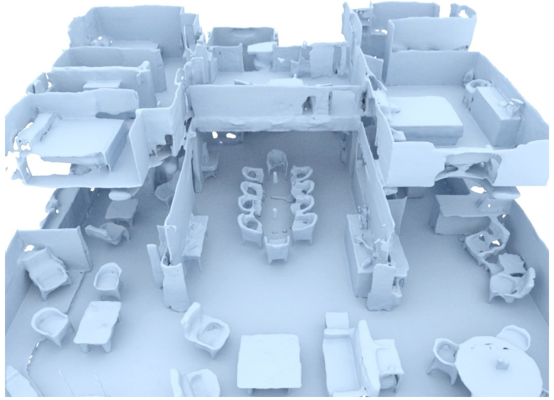


This Thesis

Develop 3D Neural Scene Representations

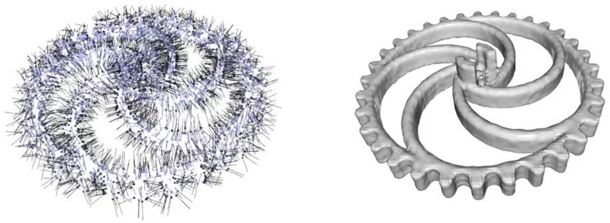
for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes



ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference



Shape As Points
NeurIPS 2021 (Oral)

3. From 2D Observations

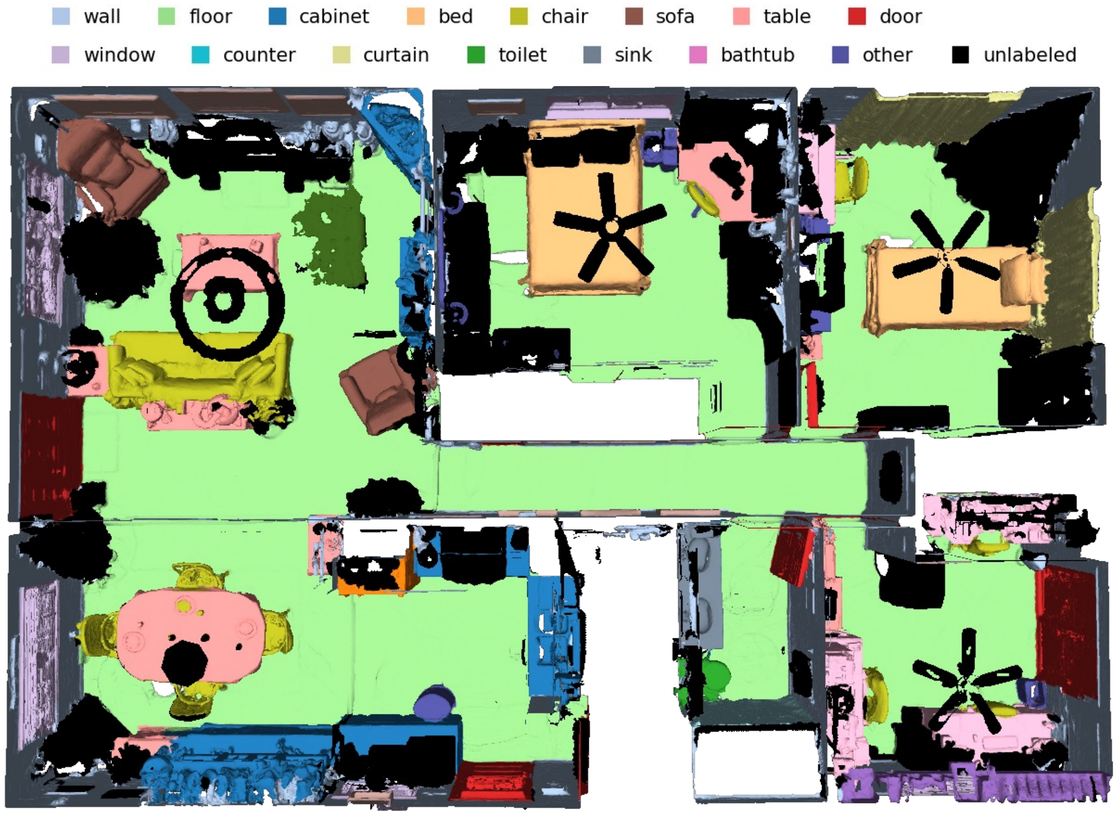
NICE-SLAM
CVPR 2022

4. Arbitrary Queries

OpenScene
CVPR 2023

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

Example: "where can I sit?"

- Affordance prediction

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 learn a scene representation to handle all these tasks **without labeled 3D data**?

This Thesis

Develop 3D Neural Scene Representations for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes

ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference

Shape As Points
NeurIPS 2021 (Oral)

3. From 2D Observations

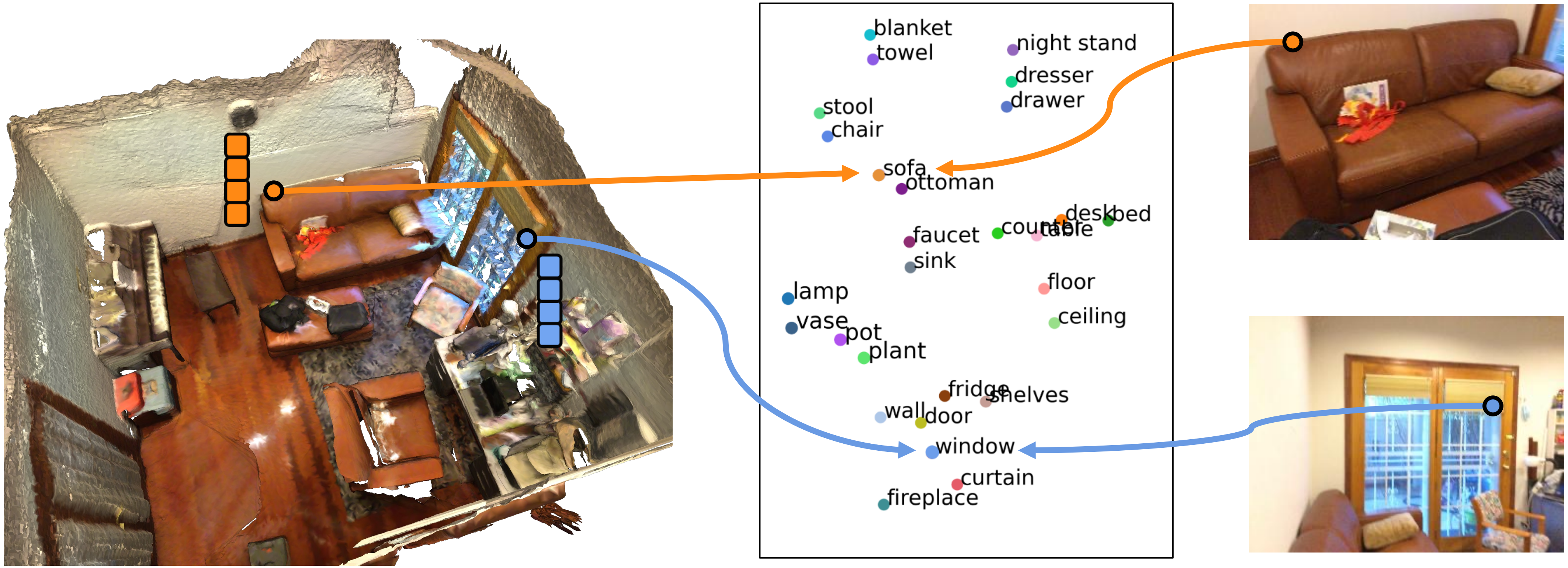
NICE-SLAM
CVPR 2022

4. Arbitrary Queries

OpenScene
CVPR 2023

Key Idea

Co-embed 3D Features with CLIP Features



3D Geometry

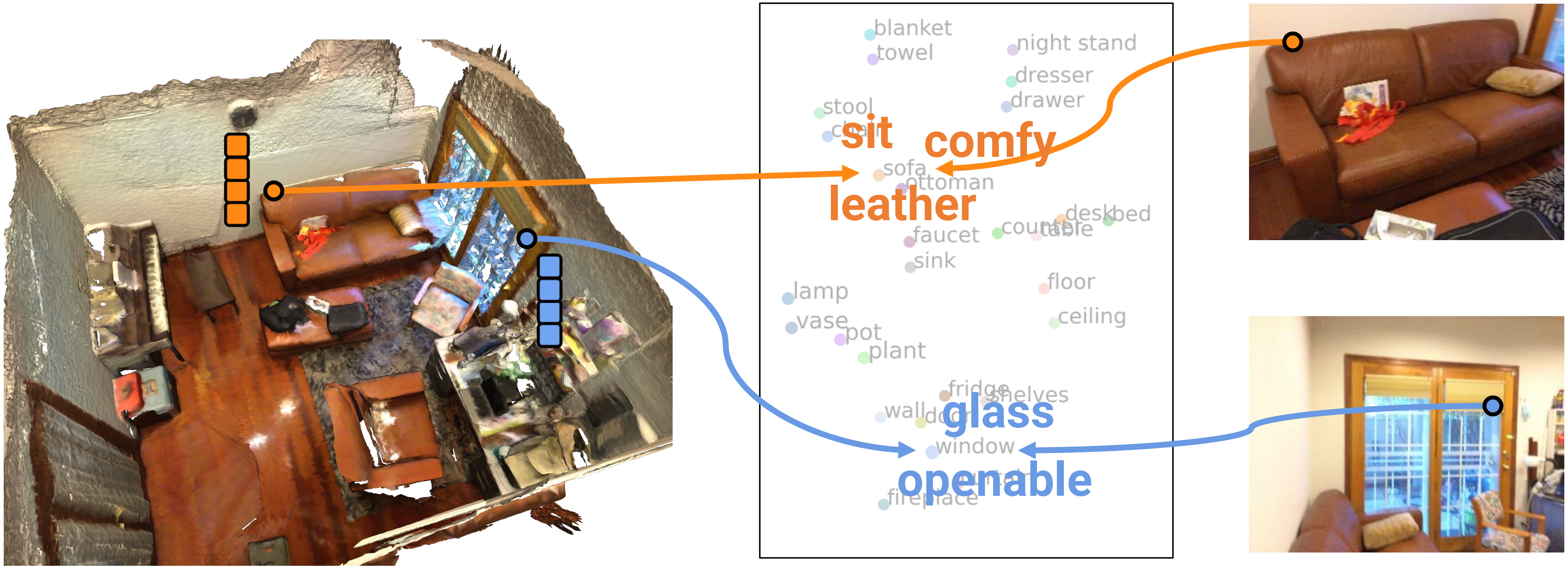
CLIP Text Features

(visualize with T-SNE)

RGB Images

Key Idea

Co-embed 3D Features with CLIP Features



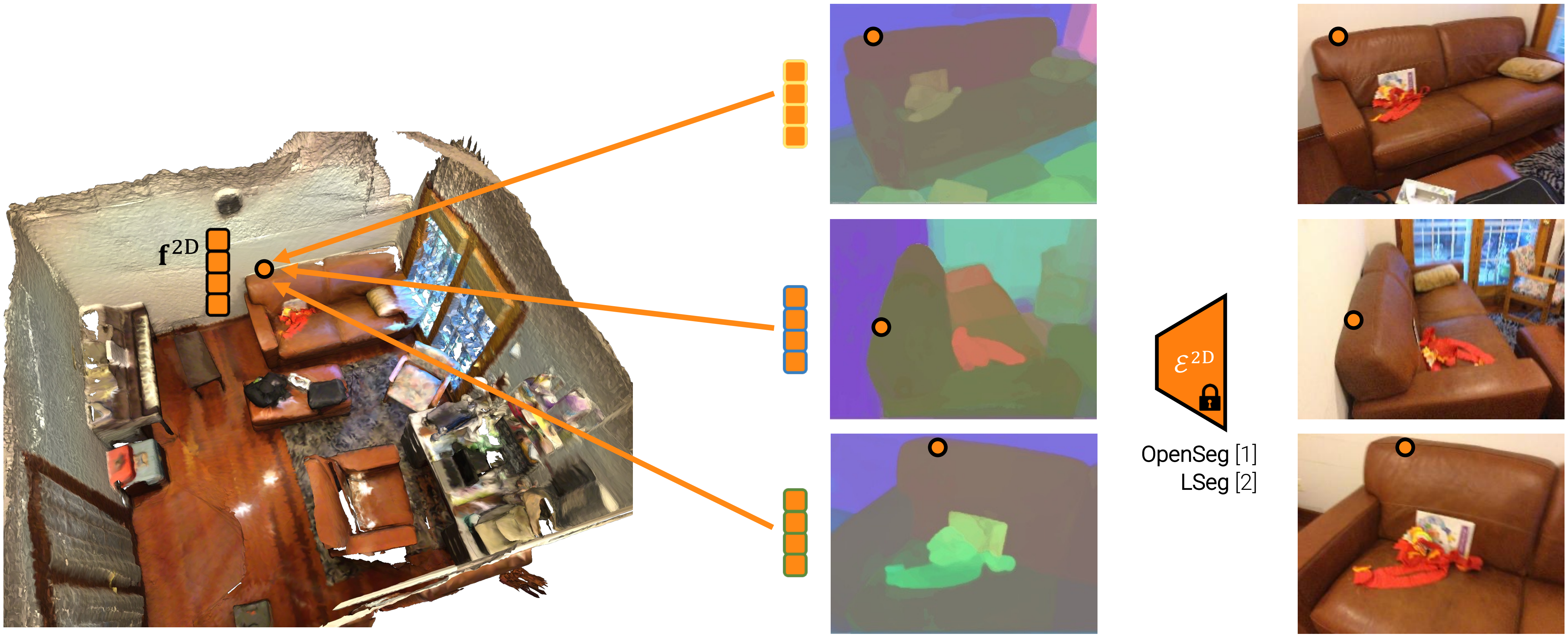
3D Geometry

CLIP Text Features
(visualize with T-SNE)

RGB Images

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

Step 1: Multi-view Feature Fusion



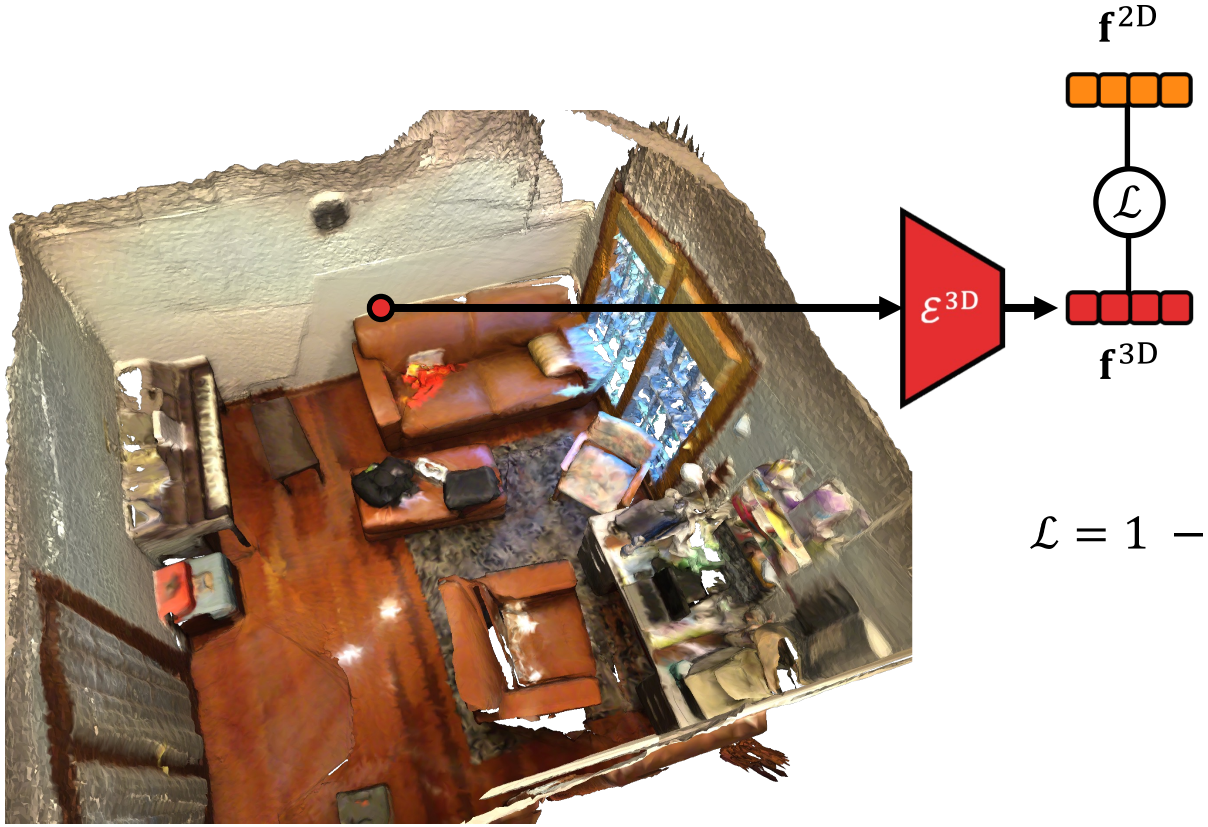
3D Geometry

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

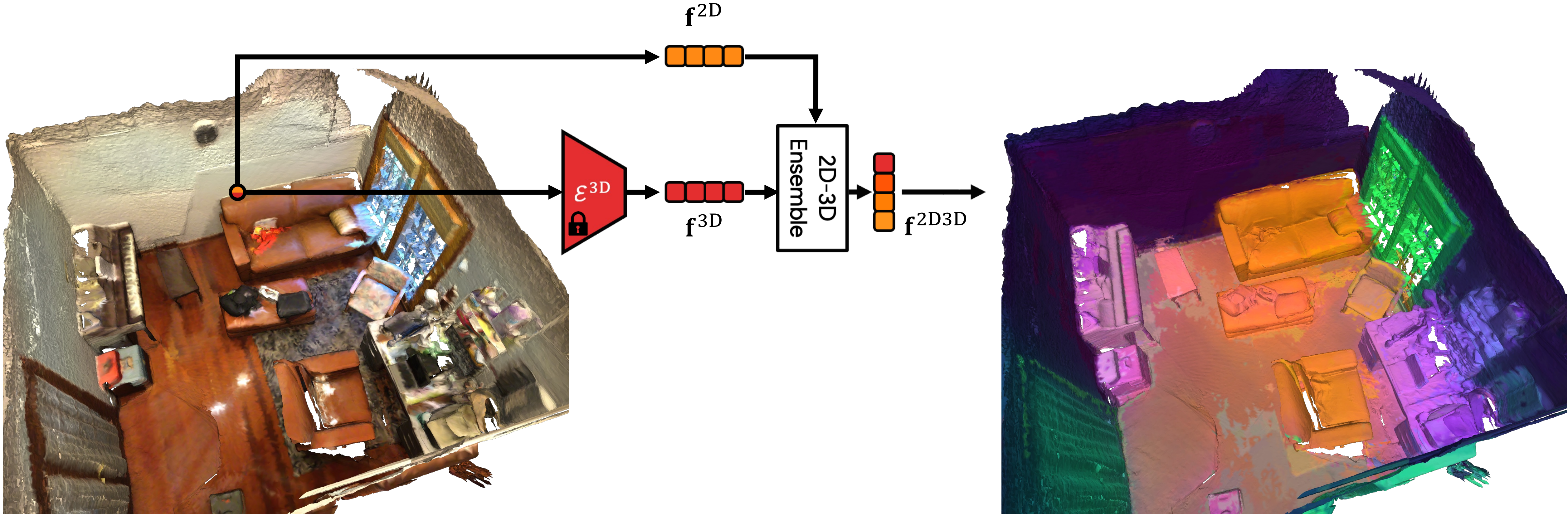
Step 2: 3D Feature Distillation



3D Geometry

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

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
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

Image-based 3D Scene Query



Image Queries

Given 3D Geometry

Interactive Demo

Open-vocabulary 3D Scene Exploration

Text queries: glass

OpenScene - TL;DR

- ➕ Open up a **wide range of applications** by leveraging large vision-language models
- ➕ Inspire future works to shift to open-vocabulary tasks
- ➖ Currently all power comes from 2D foundation models

This Thesis

Develop Neural Scene Representations

for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes

ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference

Shape As Points
NeurIPS 2021 (Oral)

3. From 2D Observations

NICE-SLAM
CVPR 2022

4. Arbitrary Queries

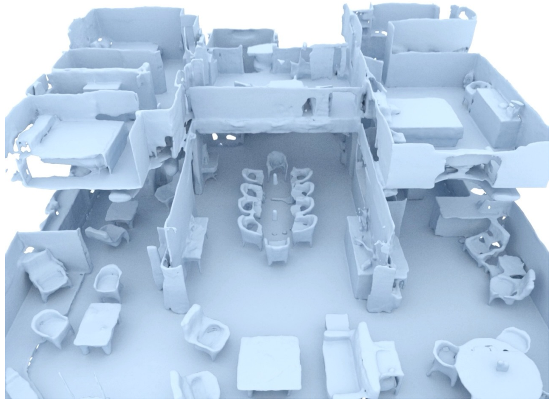
OpenScene
CVPR 2023

This Thesis

Develop Neural Scene Representations

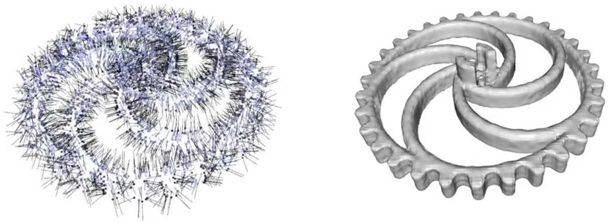
for 3D Reconstruction and 3D Scene Understanding

1. Complex Scenes



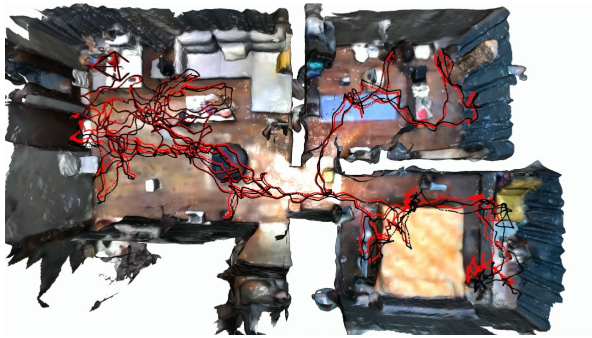
ConvOccNet
ECCV 2020 (Spotlight)

2. Fast Inference



Shape As Points
NeurIPS 2021 (Oral)

3. From 2D Observations



NICE-SLAM
CVPR 2022

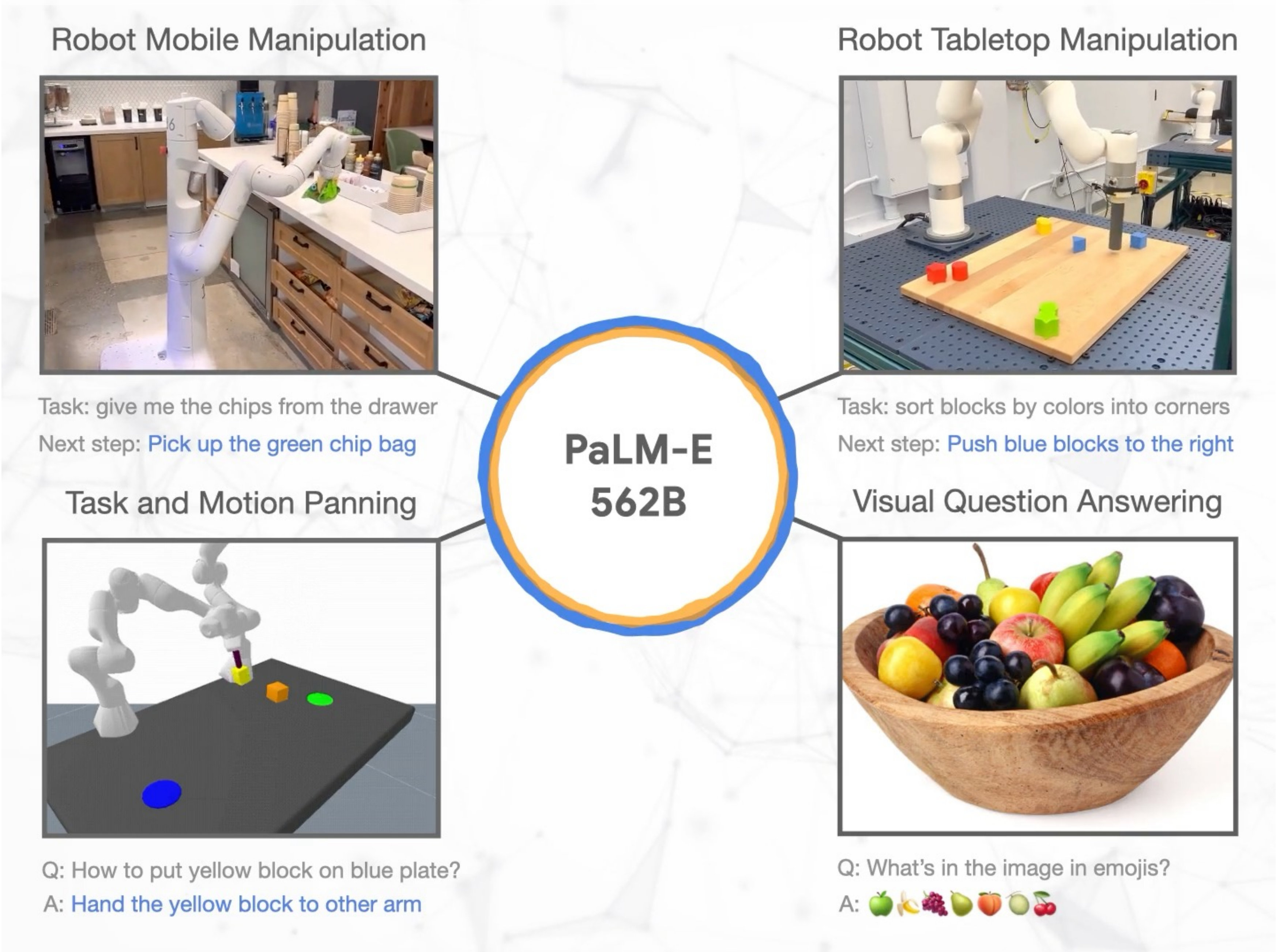
4. Arbitrary Queries

OpenScene
CVPR 2023

What is Next?

Practicality

Leverage Foundation Models for Everything



Video Credit:
PALM-E

Acknowledgements

Supervisors


Marc Pollefeys


Andreas Geiger

External Examiners


Leo Guibas


Vincent Sitzmann

Collaborators


Michael Niemeyer


Lars Mescheder


Michael Oechsle


Yiyi Liao


Chiyu “Max” Jiang


Christian Reiser


Zihan Zhu


Zhaopeng Cui


Shaohui Liu


Viktor Larsson


Martin R. Oswald


Zehao Yu


Tom Funkhouser


Kyle Genova

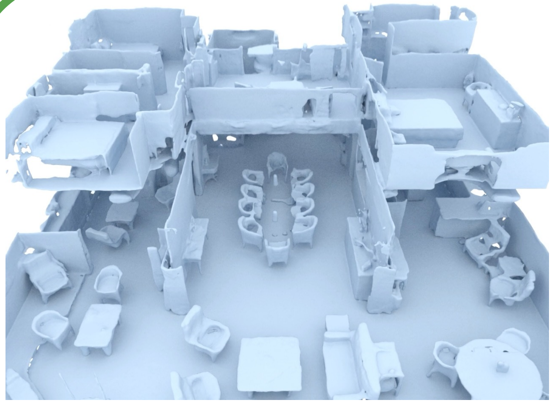

Andrea Tagliasacchi


Shengqu Cai


Thank you all for such a wonderful journey!
ICCV23
PARIS
COMPUTER SOCIETY
Google

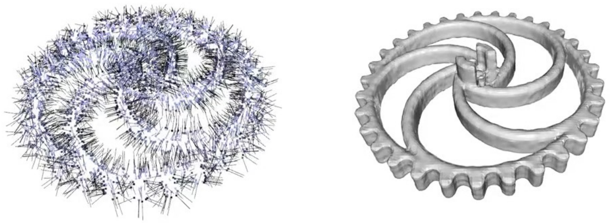
Neural Scene Representations for 3D Reconstruction and Scene Understanding

Songyou Peng



ConvOccNet

pengsongyou.github.io/conv_onet



Shape As Points

pengsongyou.github.io/sap

NICE-SLAM

pengsongyou.github.io/nice-slam

OpenScene

pengsongyou.github.io/openscene

ETH zürich

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS