Open-Vocabulary
3D Scene Understanding
towards Embodied Manipulation

Francis Engelmann, Computer Vision and Geometry Group, ETH Zurich
ETH AI Center Postdoctoral Fellow | June 7th, 2024

ETH zürich

ETH AI CENTER

D INFK

Human-Centric AI for the Greater Good

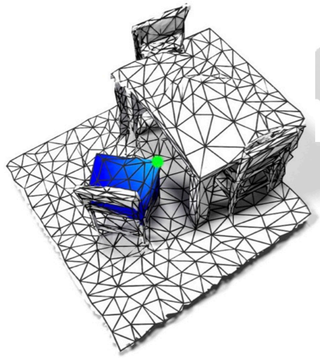
Augmenting human capabilities

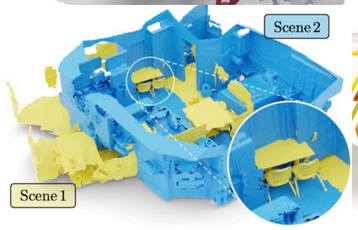
Enhancing our daily lives

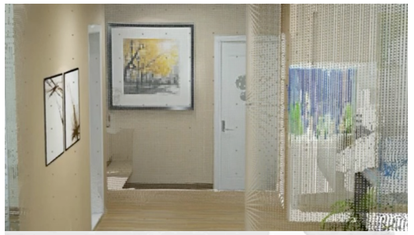
Blending the physical and virtual world

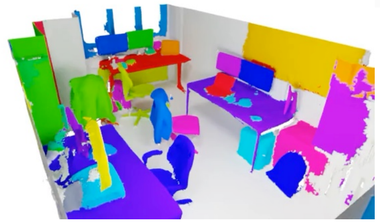

Boston Dynamics
ETH zürich
Computer Vision and Geometry Lab

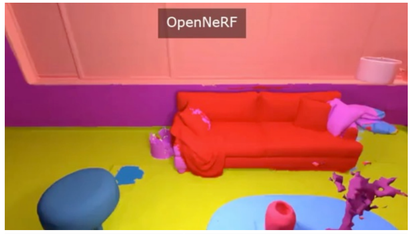
Towards 3D Scene Understanding

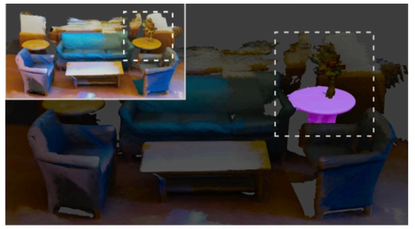












"side table with a flower vase on it"

Deep Learning Models for 3D Scenes

[ICCVW'17] Spotlight [ECCVW'18] [GCPR'19] [ICRA'20] [CVPR'20] Oral [CVPR'20]

Label-efficient Learning

[3DV'21] Oral [ECCV'22] Oral [ICCV'23] [3DV'24] Spotlight [arXiv'24]

Transformers for 3D

[ICRA'23] [ICCV'23] [CVPR'23]

Open-Vocabulary 3D Scene Understanding

[NeurIPS'23] [CVPR'24] Oral [ICLR'24]

3D Scene Understanding

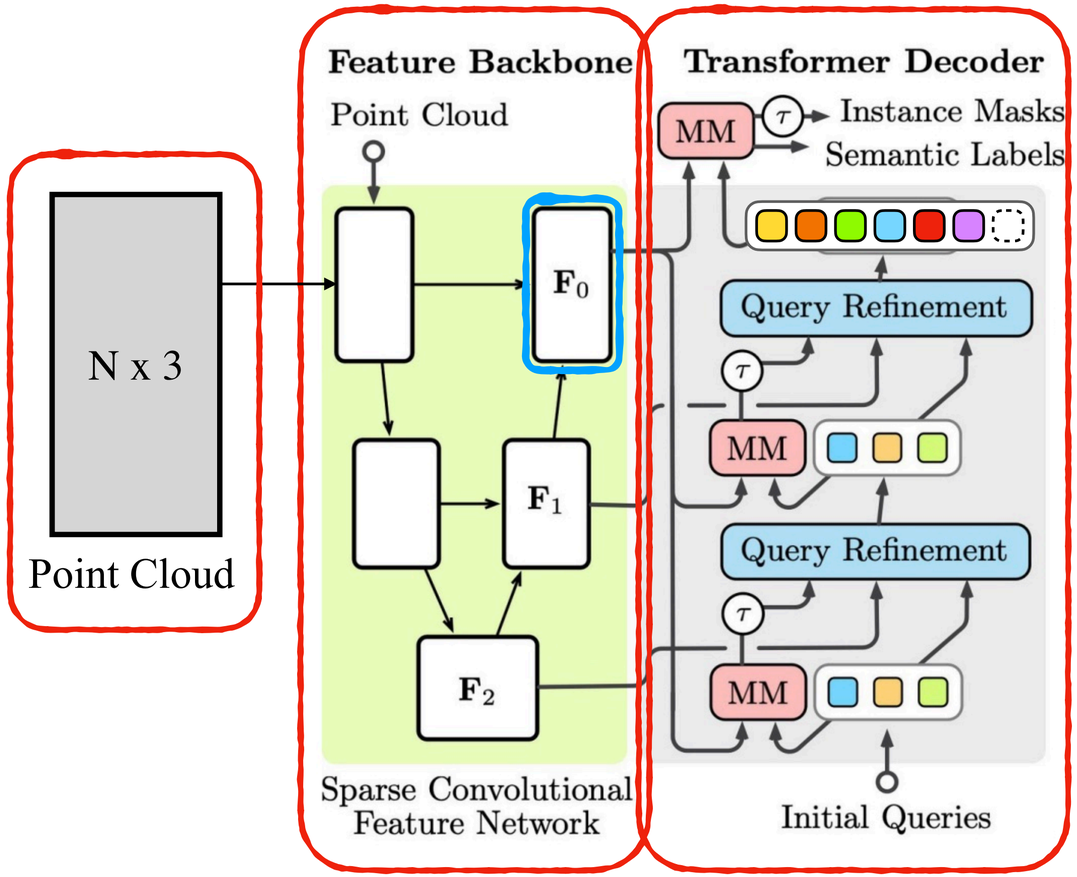
Exemplary Task: 3D Semantic Instance Segmentation

Input: **3D Scan**

Output: **Semantic Instance Masks**

3D Semantic Segmentation

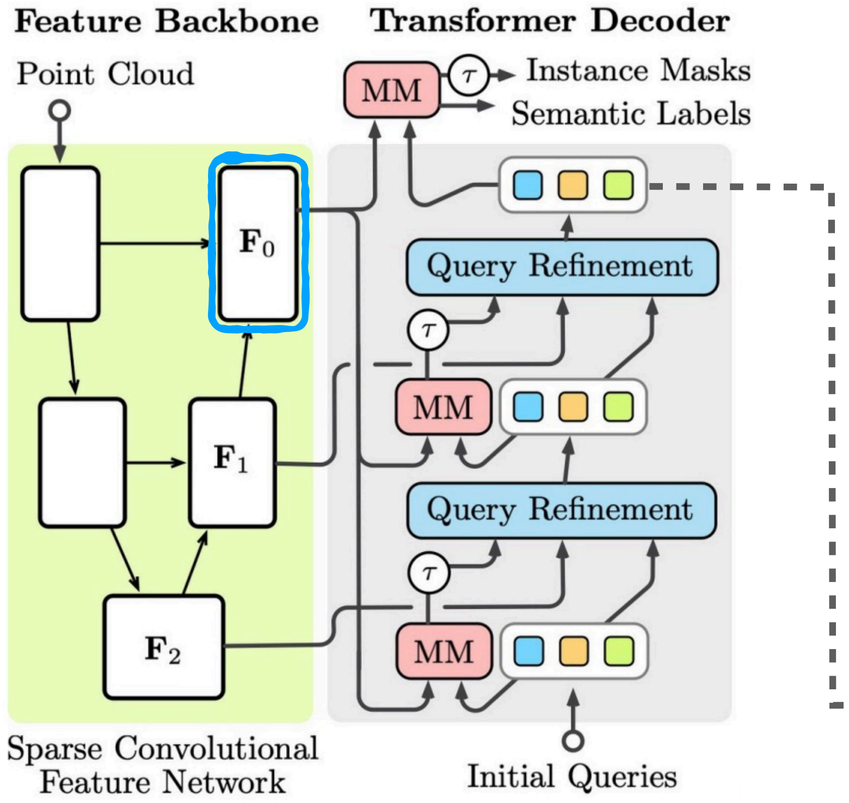
Mask Transformer for 3D Instance Segmentation [1]



[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Semantic Segmentation

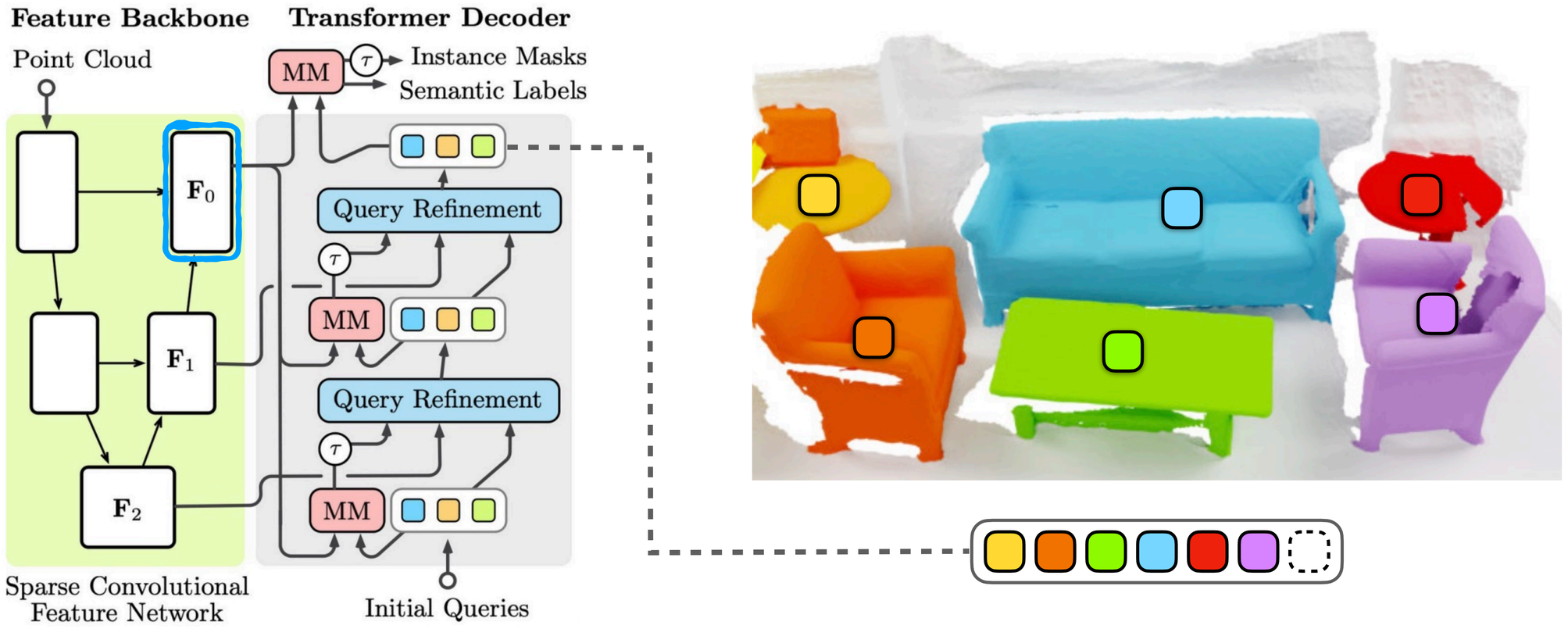
Mask Transformer for 3D Instance Segmentation [1]



[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Semantic Segmentation

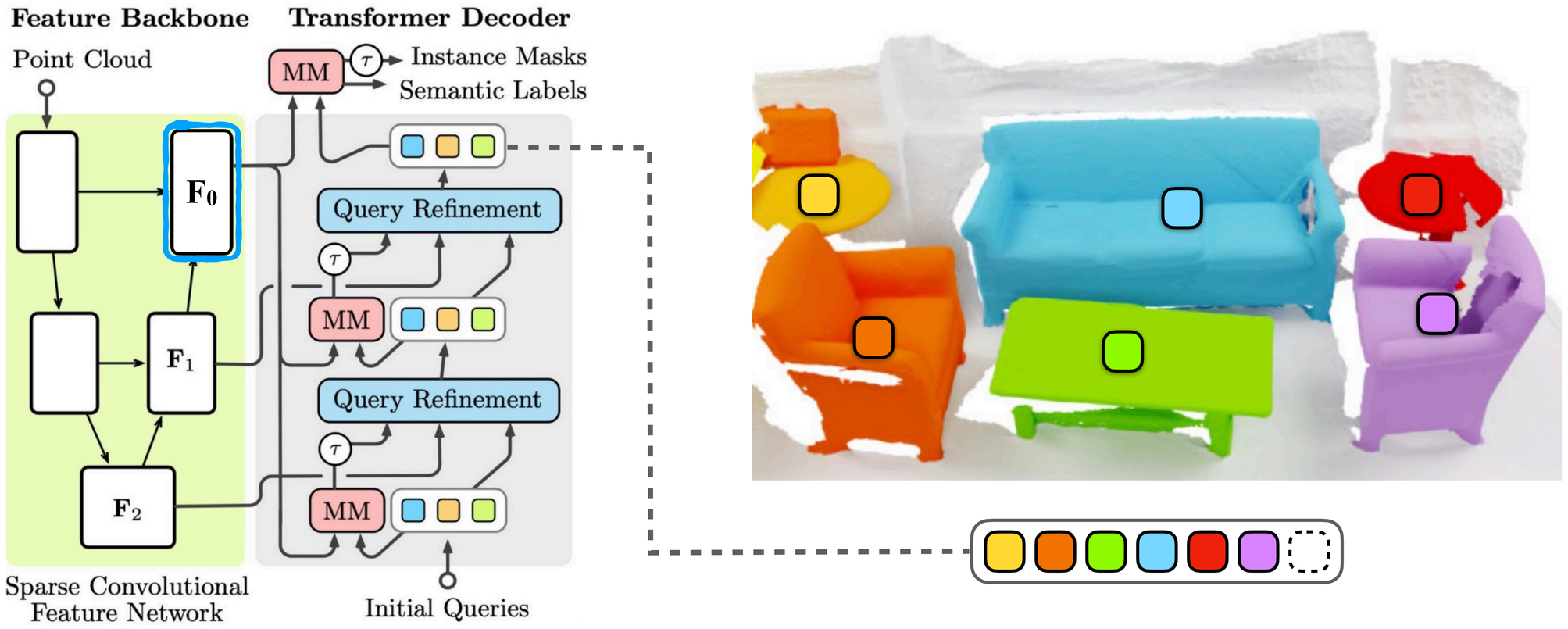
Mask Transformer for 3D Instance Segmentation [1]



[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Semantic Segmentation

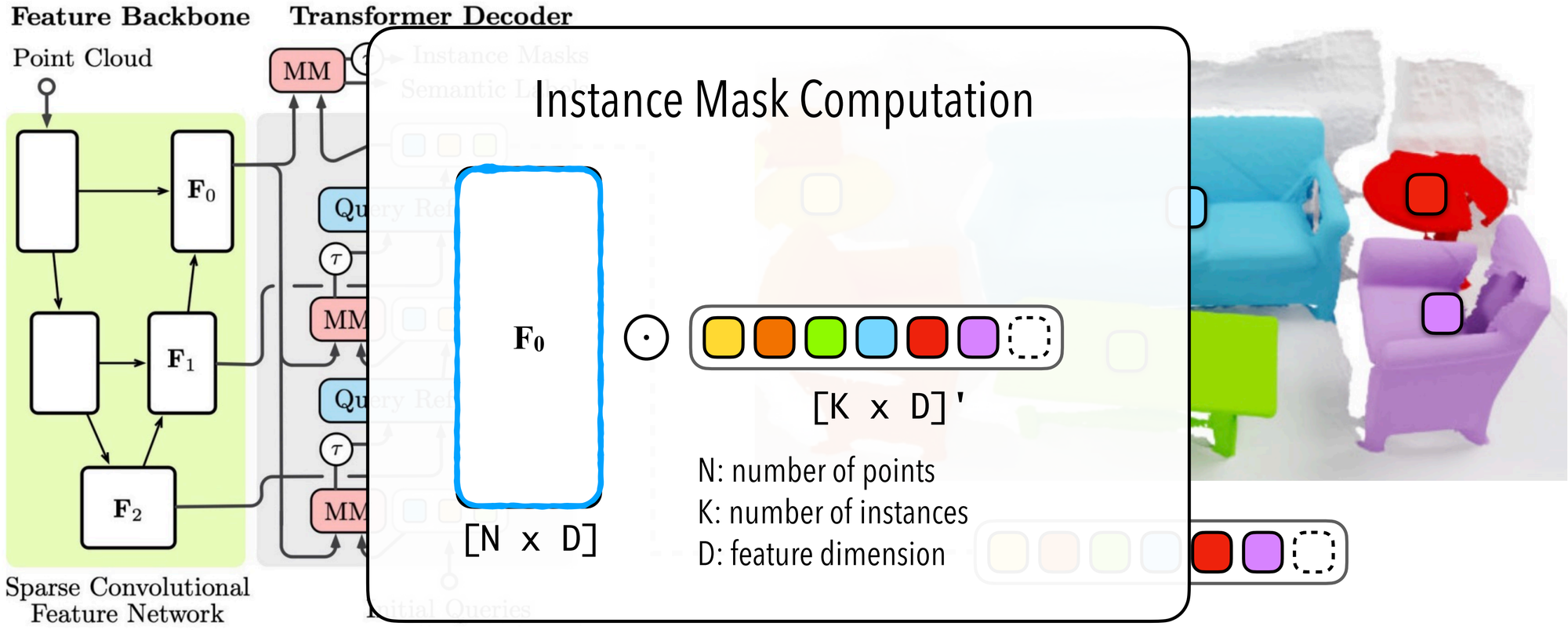
Mask Transformer for 3D Instance Segmentation [1]



[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Semantic Segmentation

Mask Transformer for 3D Instance Segmentation [1]



[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Semantic Segmentation

Mask3D: Mask Transformer for 3D Instance Segmentation

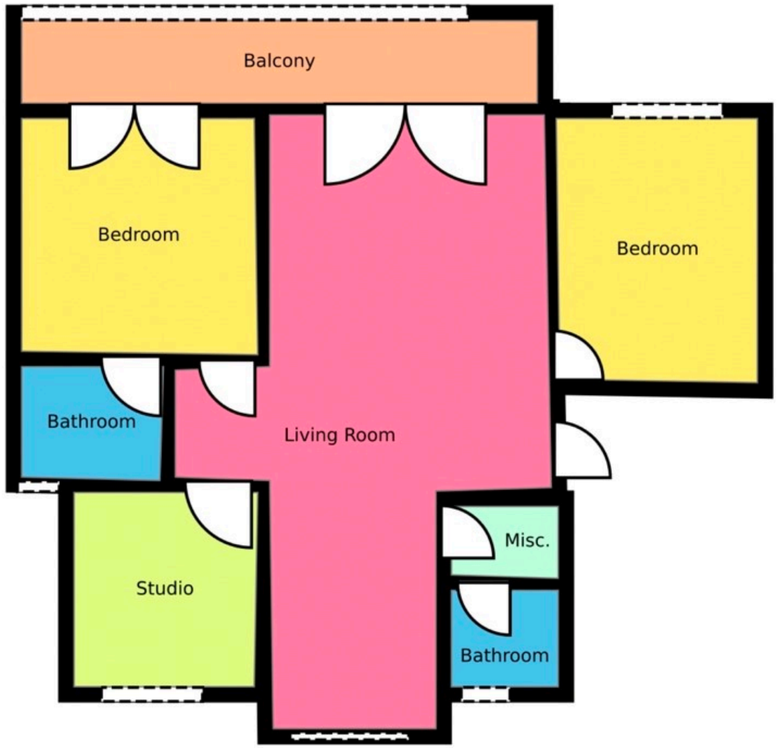
Online Demo

mask3d demo

DEMO

Floorplan Reconstruction from 3D Scans

RoomFormer [1]



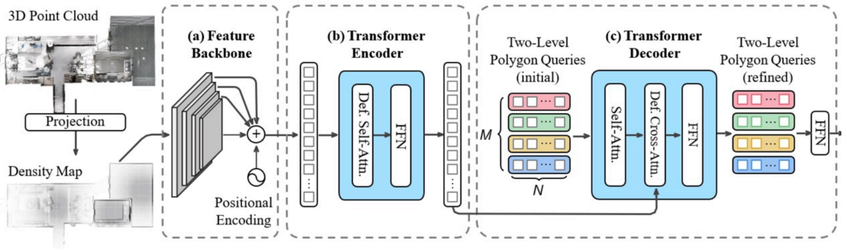
Input: **3D Point Cloud**

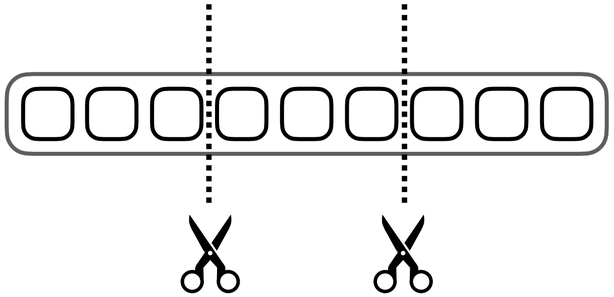
Output: **Vectorized 2D Floorplan**

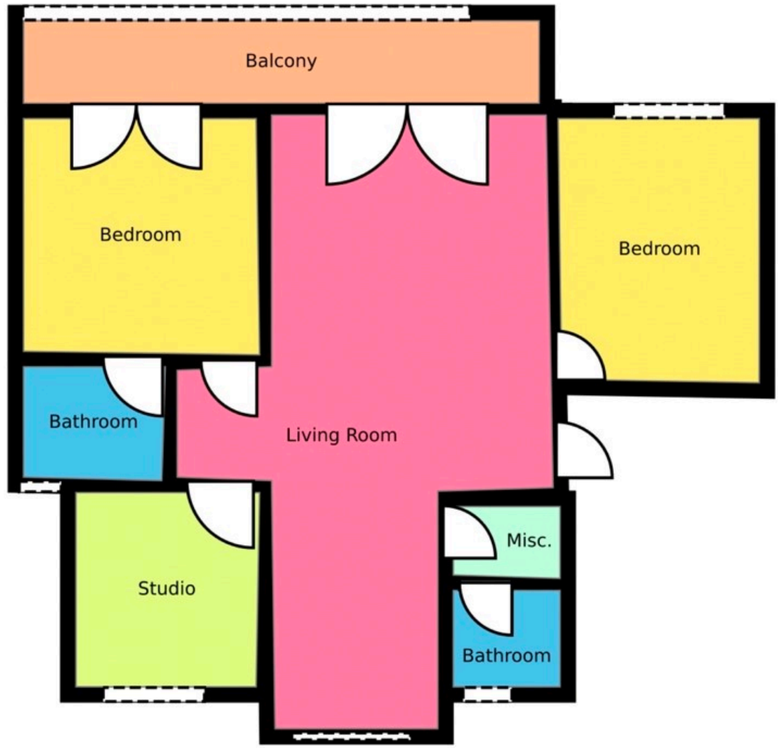
[1] Yue et al. "Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries" CVPR'23

Floorplan Reconstruction from 3D Scans

RoomFormer [1]







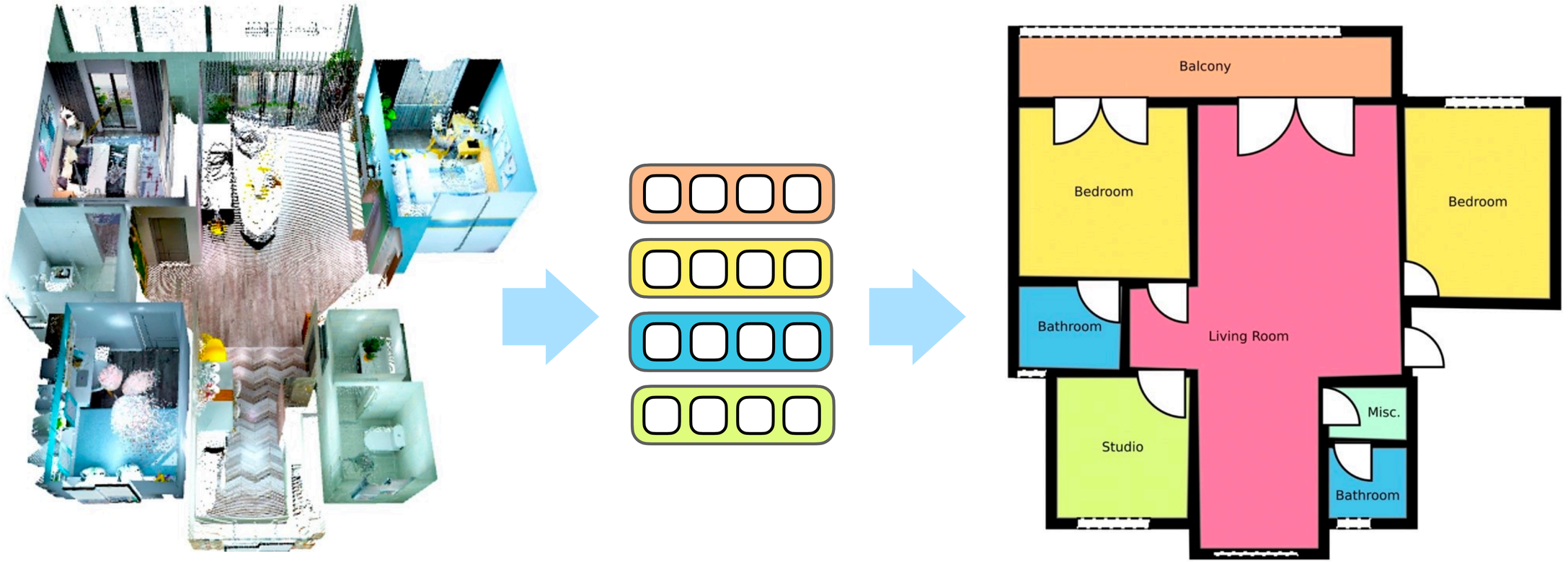
Input: **3D Point Cloud**

Output: **Vectorized 2D Floorplan**

[1] Yue et al. "Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries" CVPR'23

Floorplan Reconstruction from 3D Scans

RoomFormer [1]



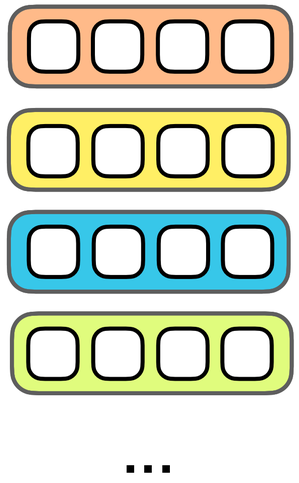
Input: **3D Point Cloud**

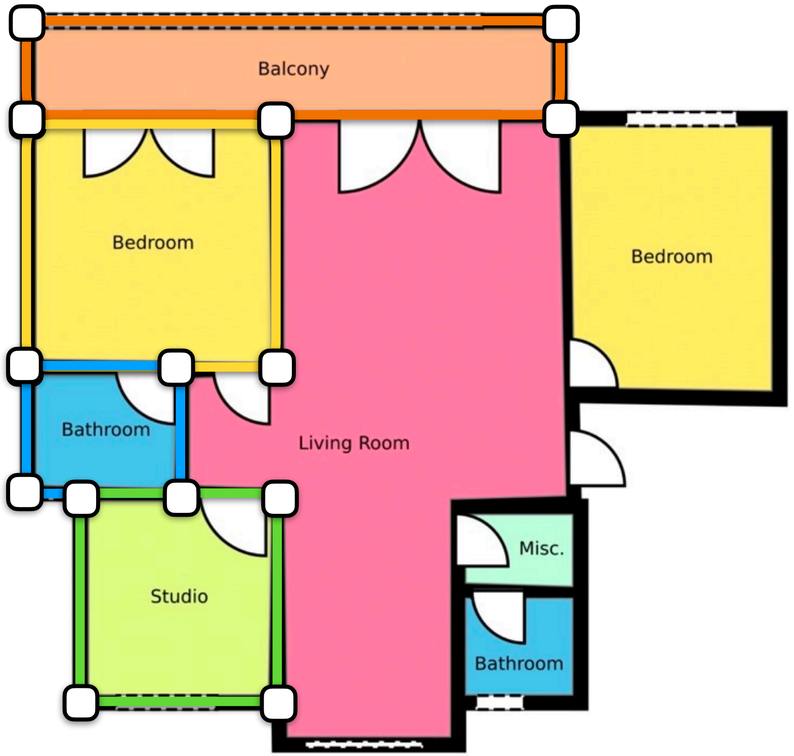
Output: **Vectorized 2D Floorplan**

[1] Yue et al. "Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries" CVPR'23

Floorplan Reconstruction from 3D Scans

RoomFormer [1] representation: **Floorplan as set of polygons**





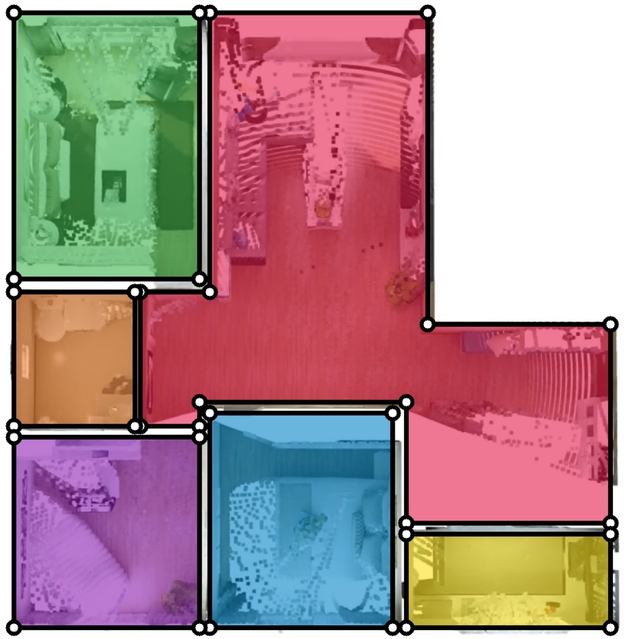
Input: **3D Point Cloud**

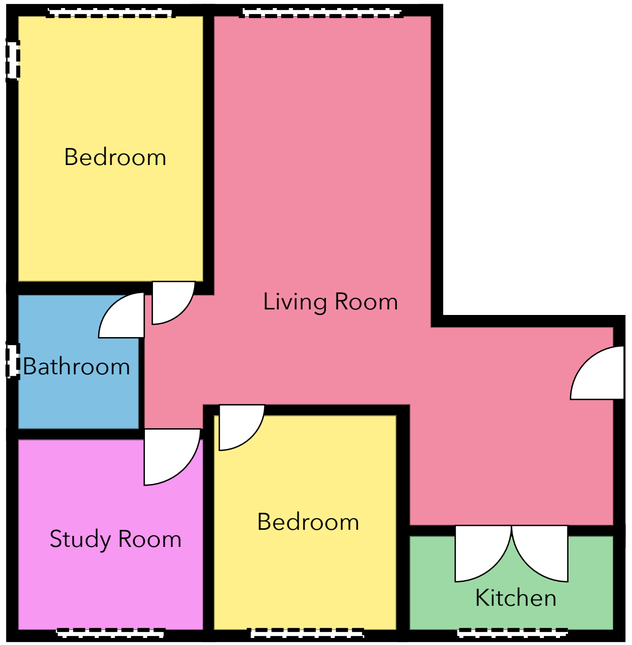
Output: **Vectorized 2D Floorplan**

[1] Yue et al. "Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries" CVPR'23

Floorplan Reconstruction from 3D Scans

RoomFormer [1]





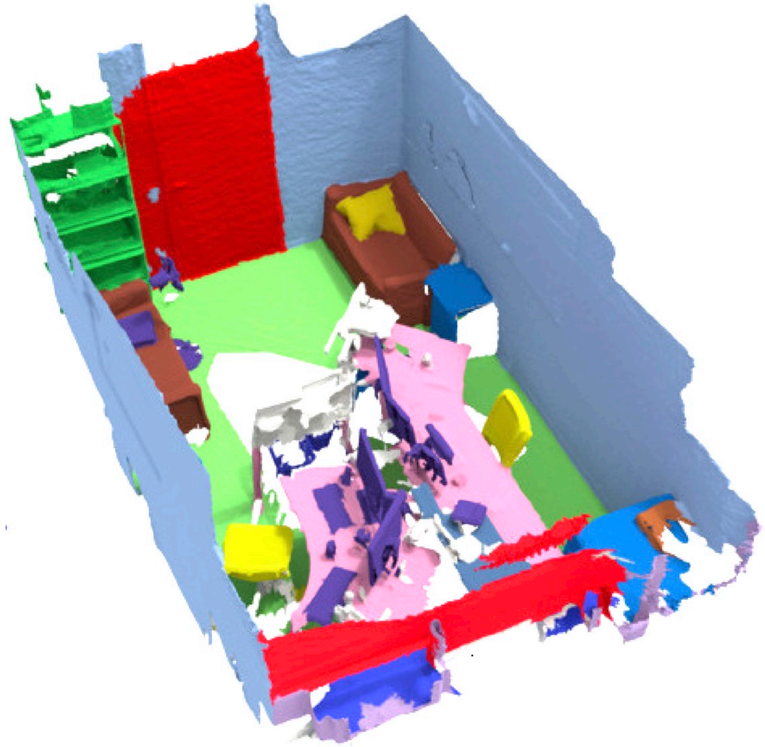
Input: **3D Scan**

Output: **2D Floorplan**

Additionally: **Semantic elements**
(Room types, doors, windows)

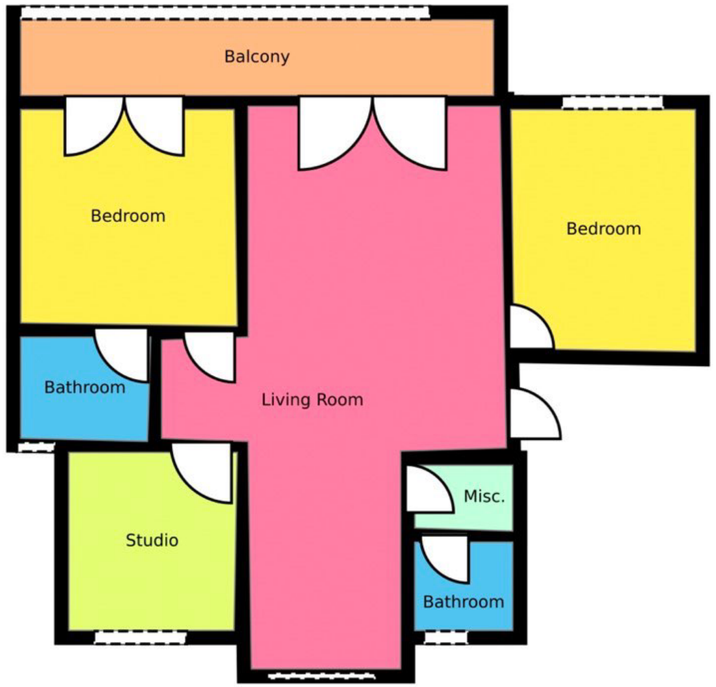
[1] Yue et al. "Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries" CVPR'23

3D Scene Understanding



3D Scene Segmentation

"Which objects are in the scene?"



Vectorized Floorplans

"Structural scene elements?"

Scene Objects ✔

Structural Elements ✔

What is missing ?

3D Segmentation of Humans

Human-Body Part Segmentation

Input: **3D Point Cloud**

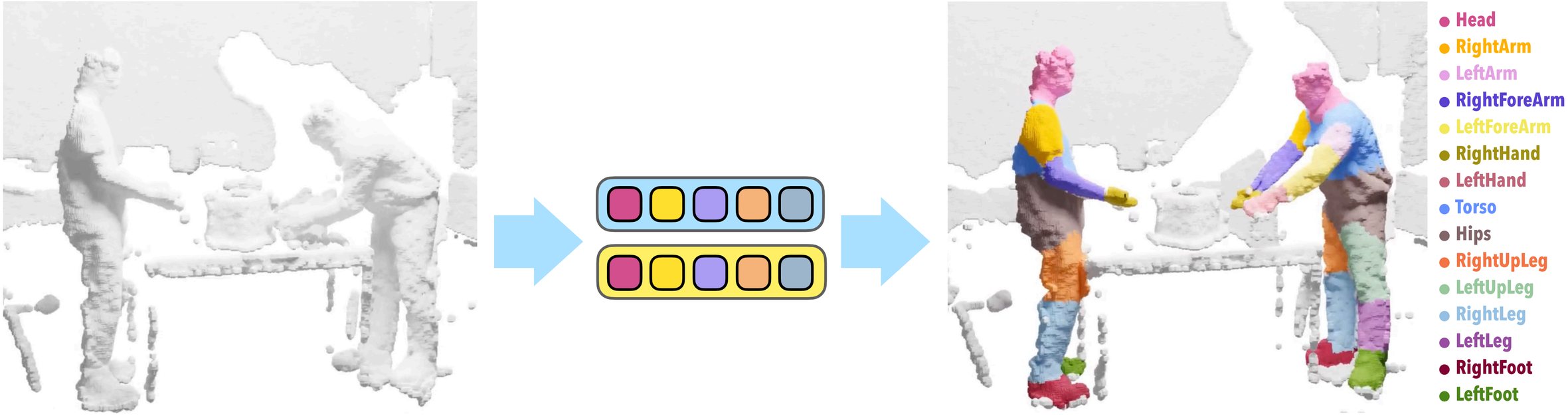
Output: **Multi-Human Body-Parts**

- Head
- RightArm
- LeftArm
- RightForeArm
- LeftForeArm
- RightHand
- LeftHand
- Torso
- Hips
- RightUpLeg
- LeftUpLeg
- RightLeg
- LeftLeg
- RightFoot
- LeftFoot

[1] Takmaz et al. "Human3D: 3D Segmentation of Humans in Point Clouds with Synthetic Data" ICCV'23

3D Segmentation of Humans

Human-Body Part Segmentation



Input: **3D Point Cloud**

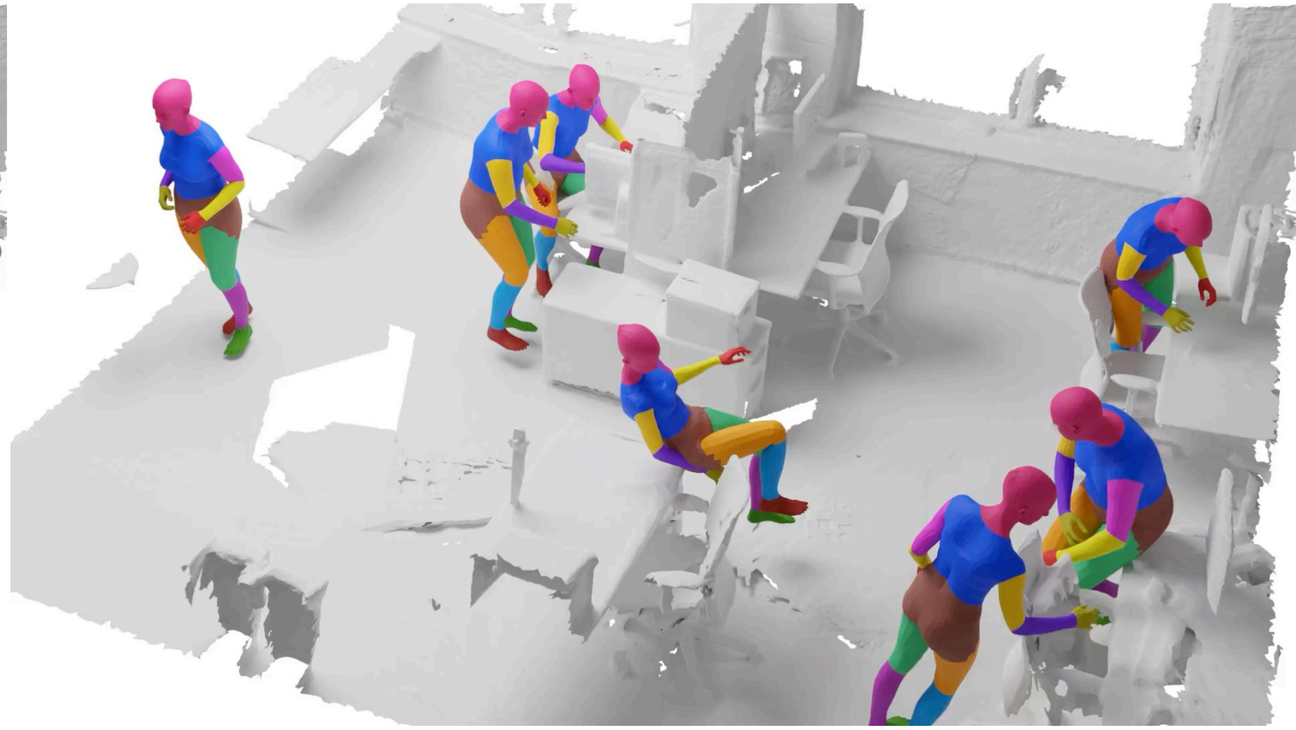
Output: **Multi-Human Body-Parts**

[1] Takmaz et al. "Human3D: 3D Segmentation of Humans in Point Clouds with Synthetic Data" ICCV'23

3D Segmentation of Humans

Synthetic Training Data

Synthesized Human Instances

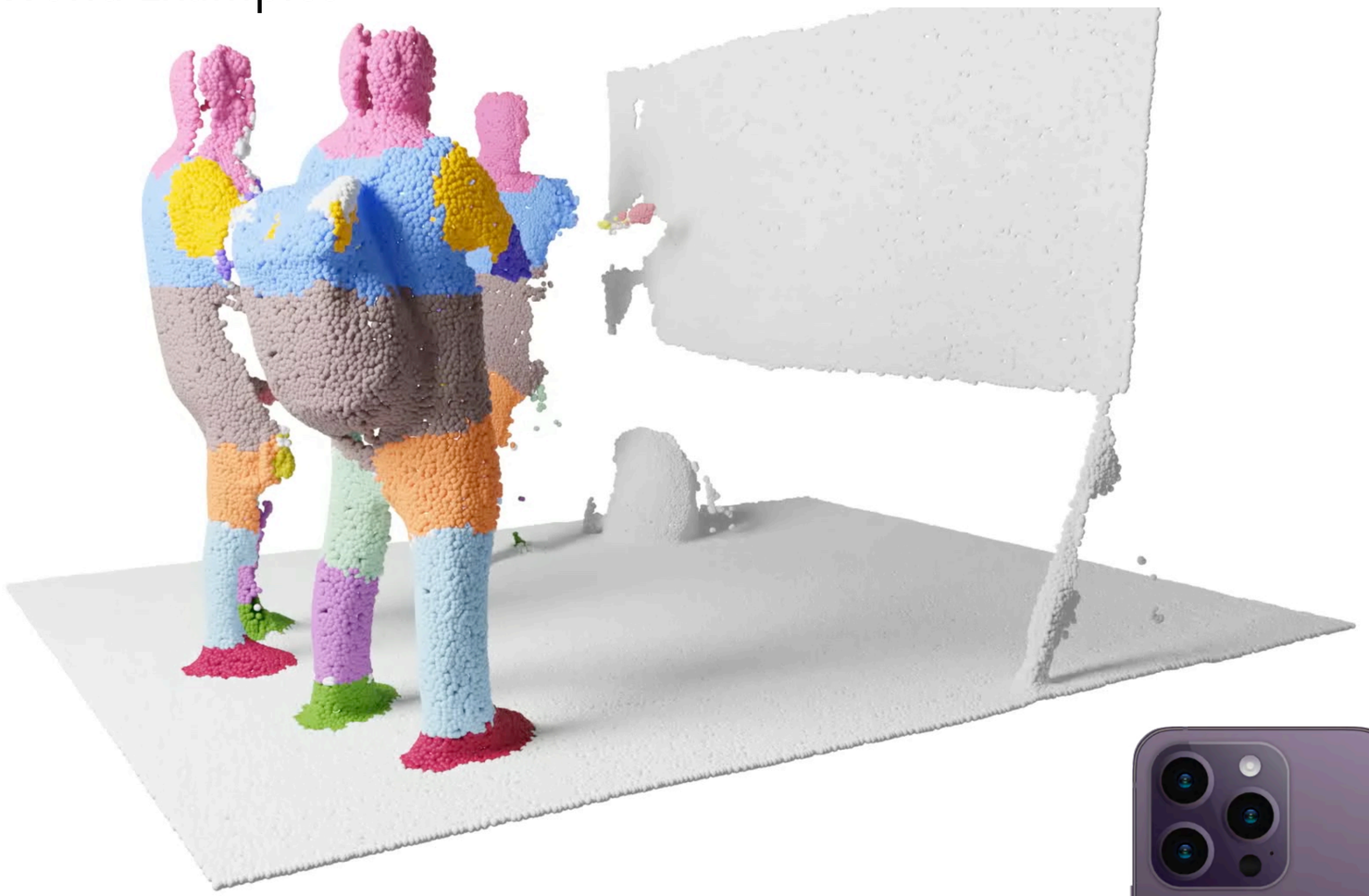


Synthesized Human Body Parts

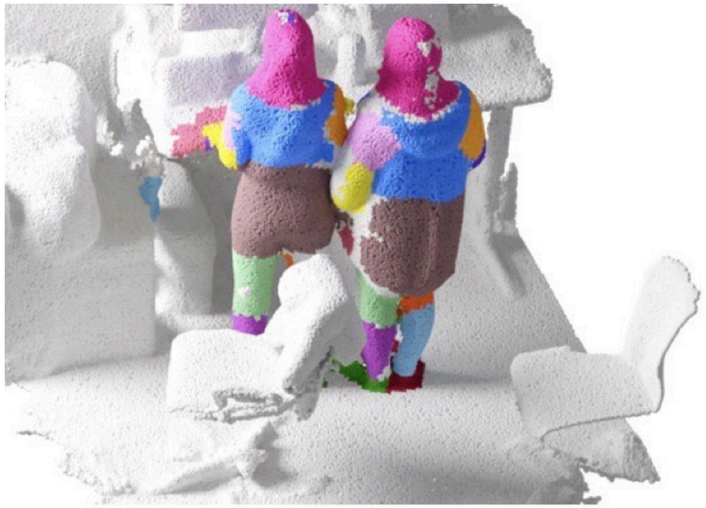
[1] Takmaz et al. "Human3D: 3D Segmentation of Humans in Point Clouds with Synthetic Data" ICCV'23

3D Segmentation of Humans

Real-World Examples

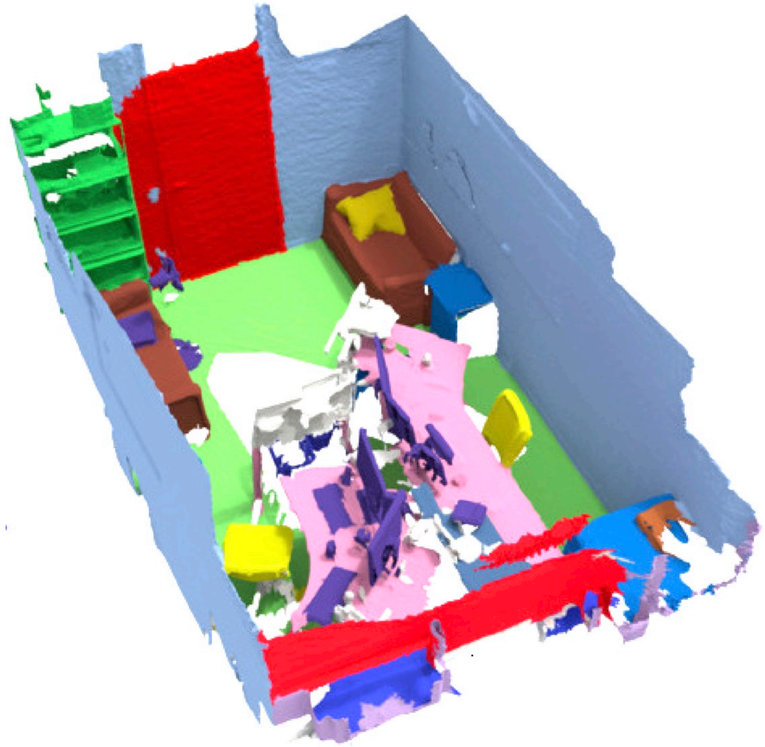


Mobile Scanner



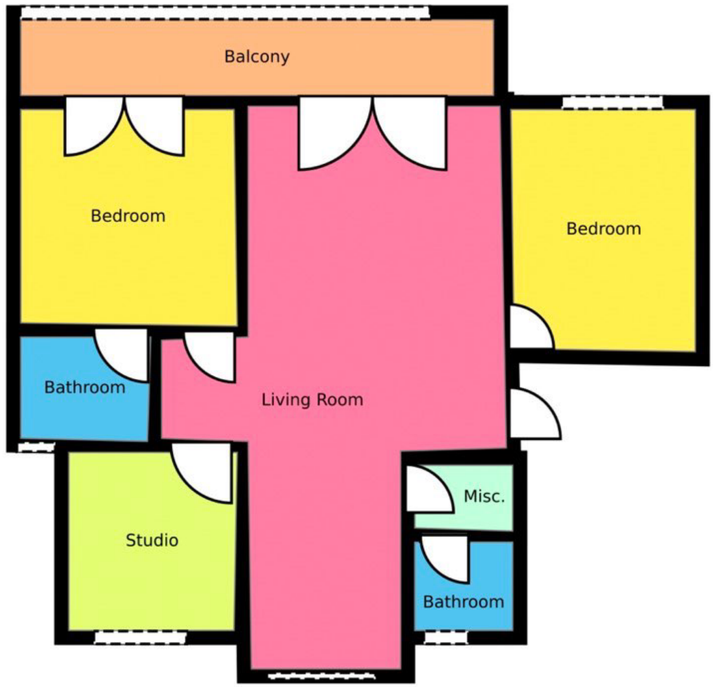
3D Scene Understanding

Tasks: From an input 3D scan, the goal is to obtain ...



3D Scene Segmentation

"Which objects are in the scene?"



Vectorized Floorplans

"Structural scene elements?"

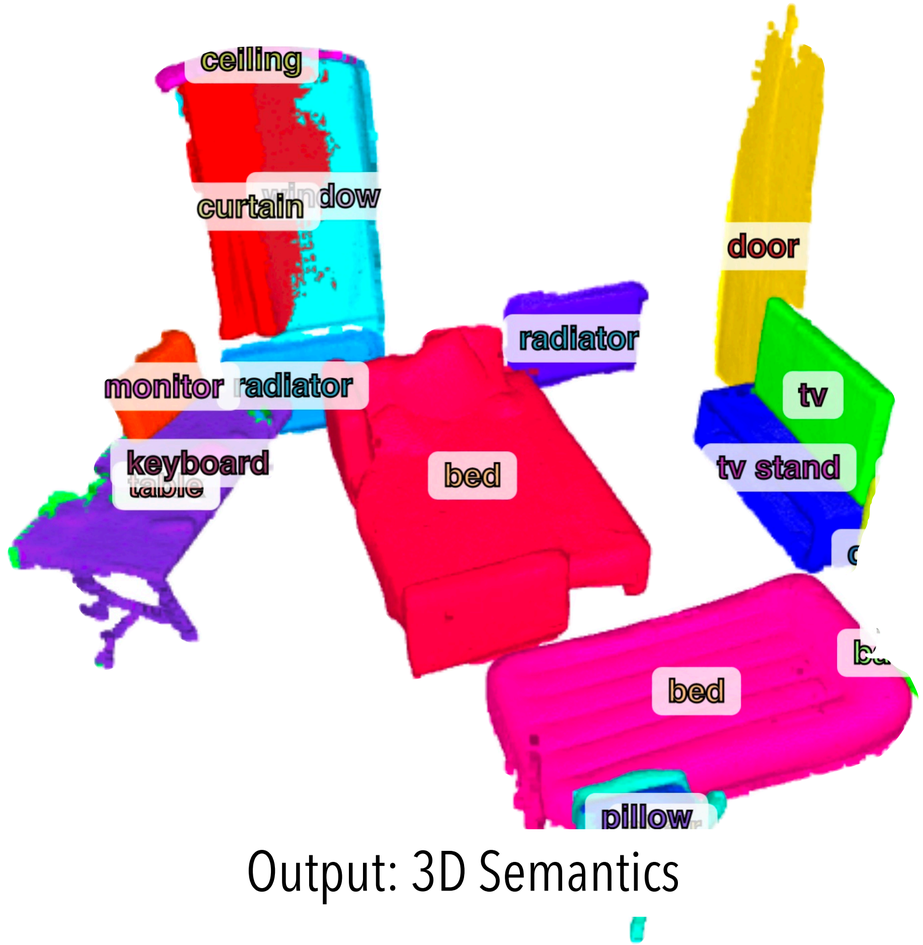
Human Part Segmentation

"Human-scene interactions?"

3D Scene Understanding *In-the-Wild*

Current models work well for a large variety of tasks ...

Input: 3D Point Cloud



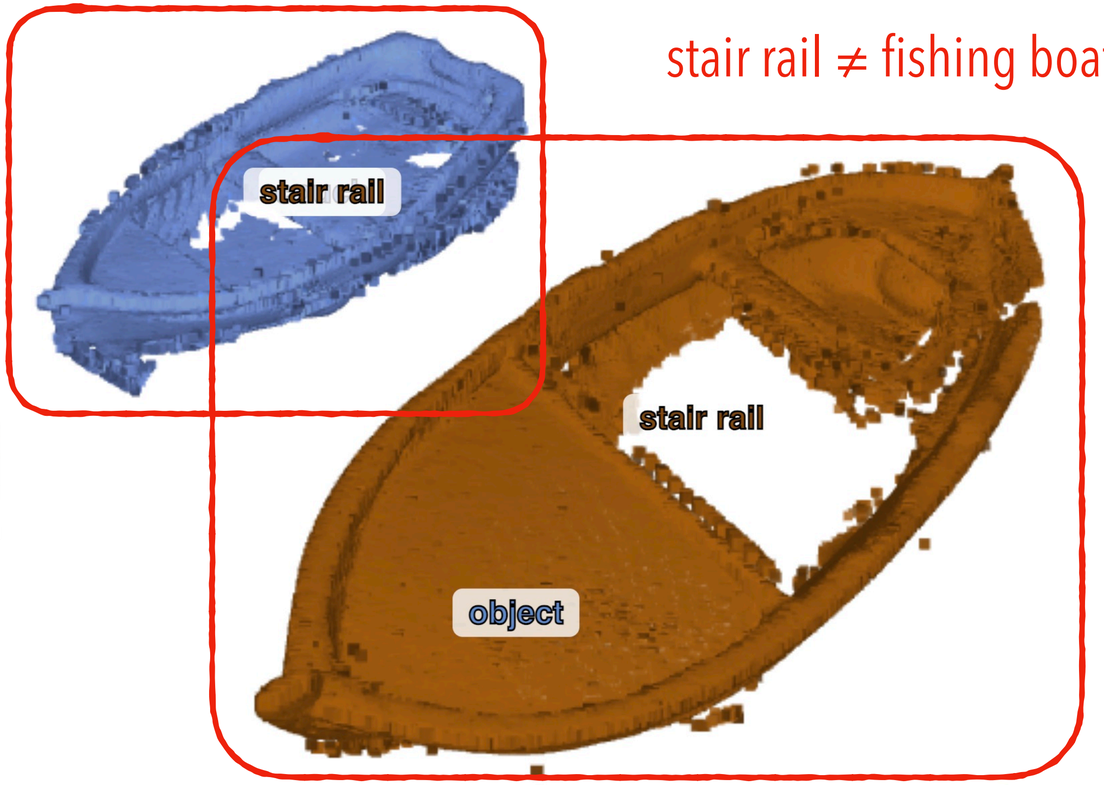
Output: 3D Semantics

[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

3D Scene Understanding *In-the-Wild*

... but **limited to a <u>predefined</u> closed set of classes!**

Input: 3D Point Cloud

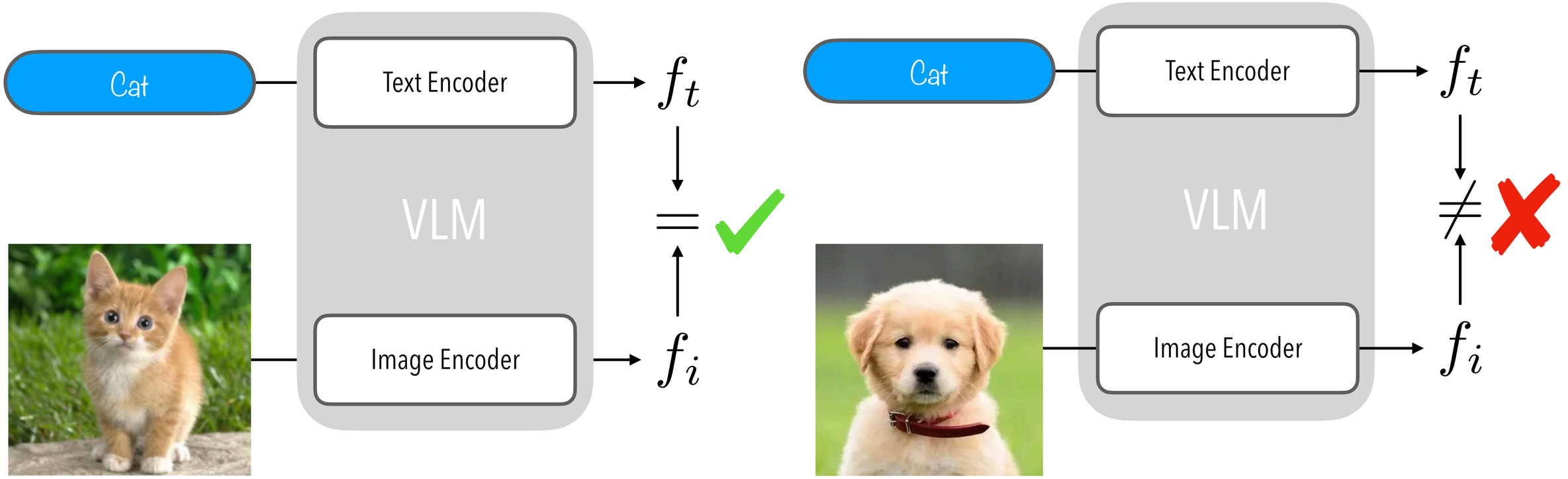


Output: 3D Semantics

[1] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

How can we achieve Open-Vocabulary 3D Scene Understanding?

Large Visual Language Model (VLM) *for example CLIP [1]*



[1] Radford et al. "*Learning Transferable Visual Models From Natural Language Supervision*" ICML'21

Open-Vocabulary 3D Scene Understanding

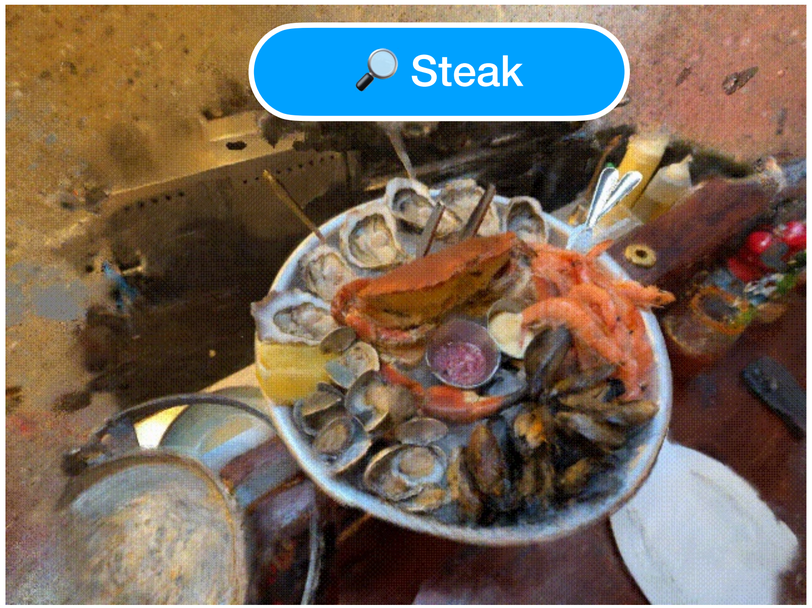
Use Visual-Language-Model (VLM) to query scene.

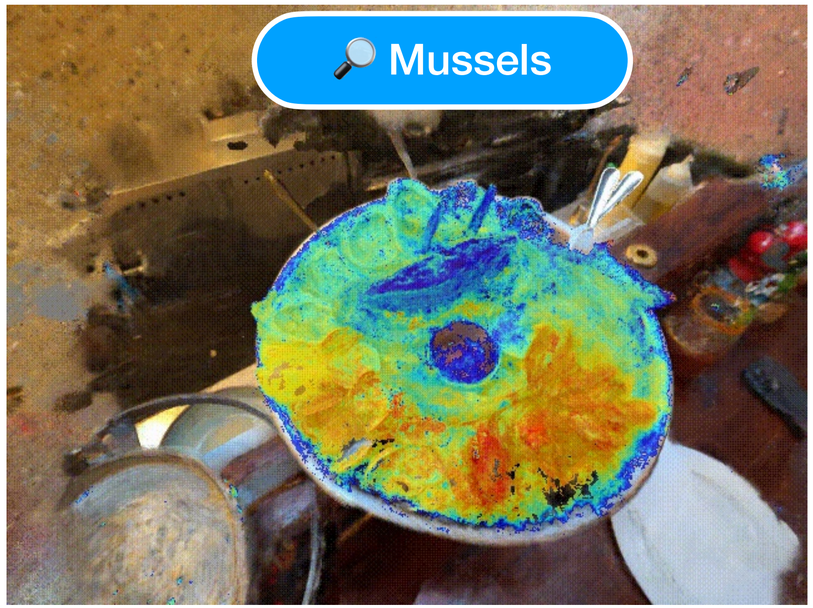
Similarity Score

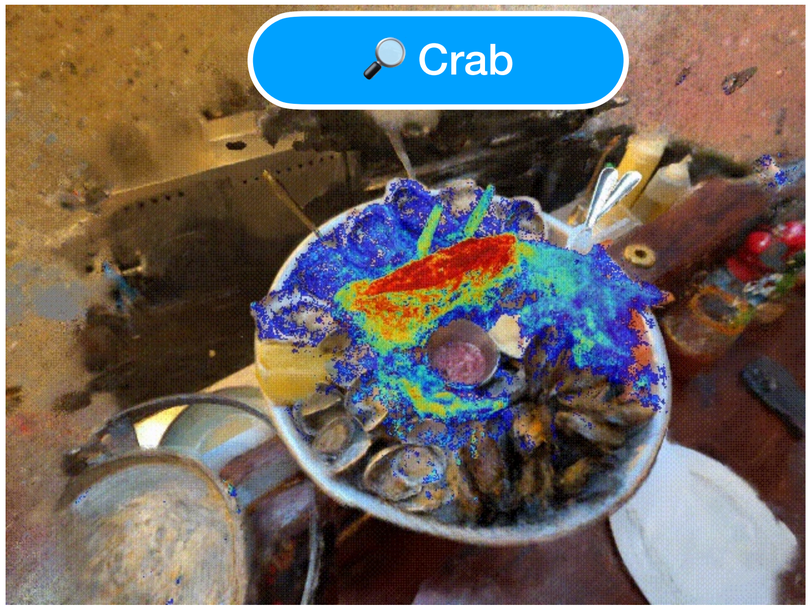
Dissimilar — Similar

Mechanism for zero-short image segmentation:

1. Compute CLIP [1] encoding of text query and per-pixel CLIP features via OpenSeg [2]
2. Get response from dot-product of normalized encodings







[1] Radford et al. "Learning Transferable Visual Models From Natural Language Supervision" ICML'21
[2] Ghiasi et al. "Scaling open-vocabulary image segmentation with image-level labels" ECCV'22
[3] Engelmann et al. "OpenNeRF" ICLR'24

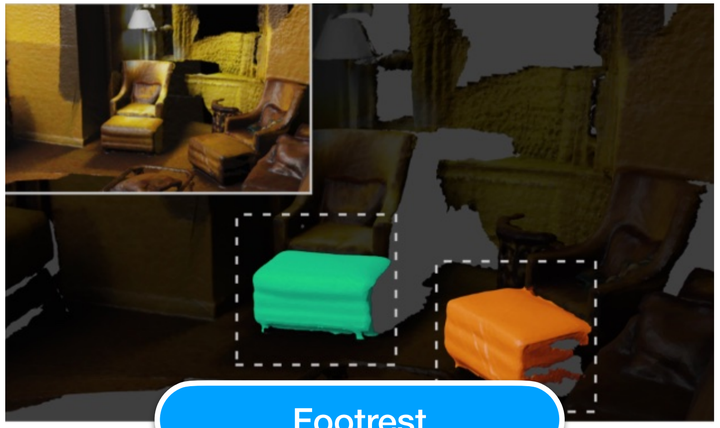
Searching 3D Scenes

OpenMask3D [1]

Input: 3D Scene Representation + Search Query

🔎 Search Query

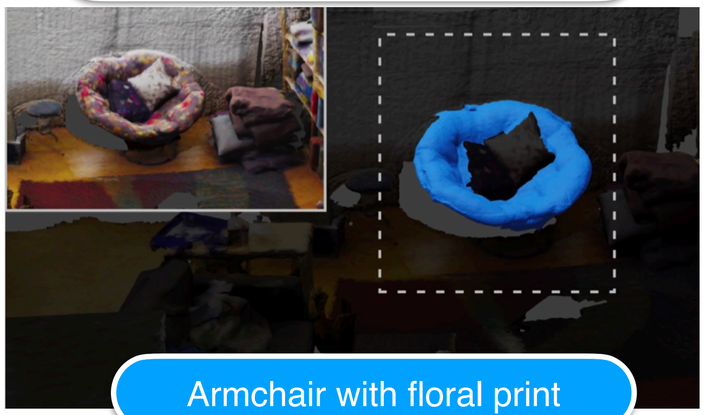
Output: 3D instance masks corresponding to search query



Footrest

Side table with a flower vase

A comfy seat

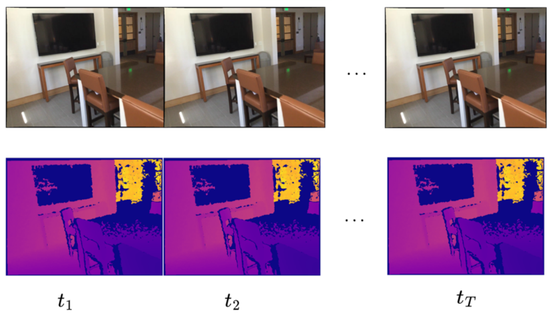


Armchair with floral print

OpenMask3D: Open-Vocabulary 3D Instance Segmentation

3D Scene Representation

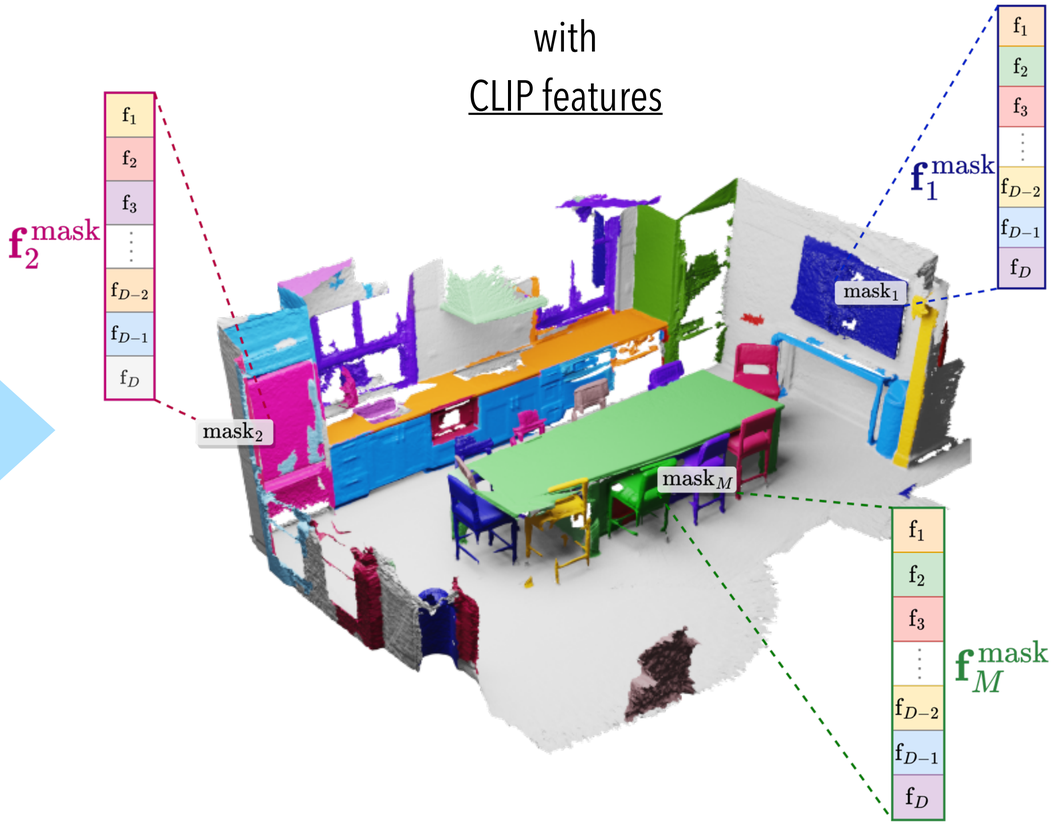
3D Reconstruction



Posed RGB-D Sequence

OpenMask3D

Instance Masks
with
CLIP features



OpenMask3D: Open-Vocabulary 3D Instance Segmentation

How to obtain the instance masks?

3D Reconstruction

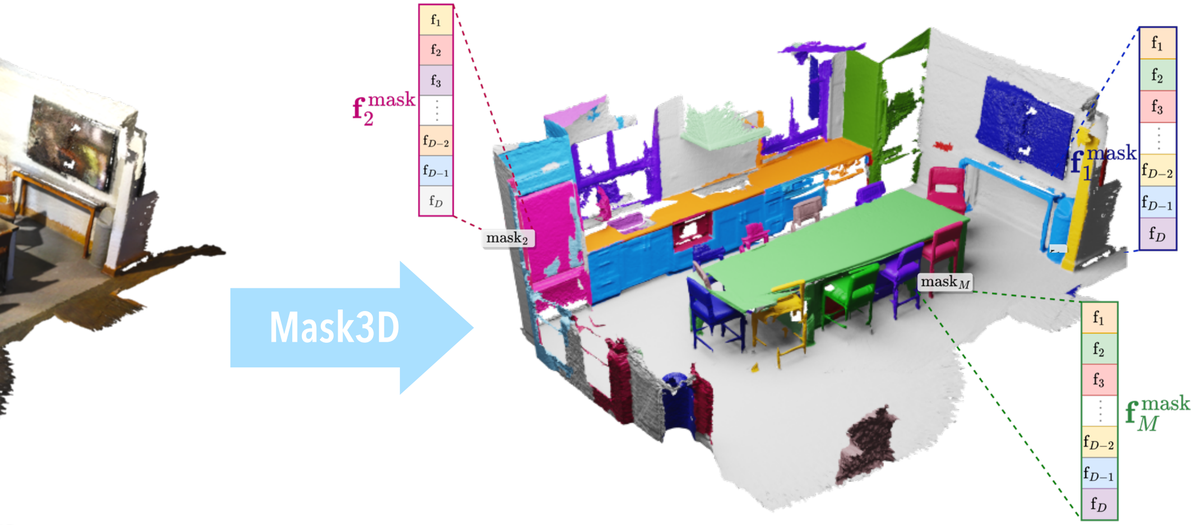
Mask3D

3D Instance Masks Proposals
(class-agnostic)

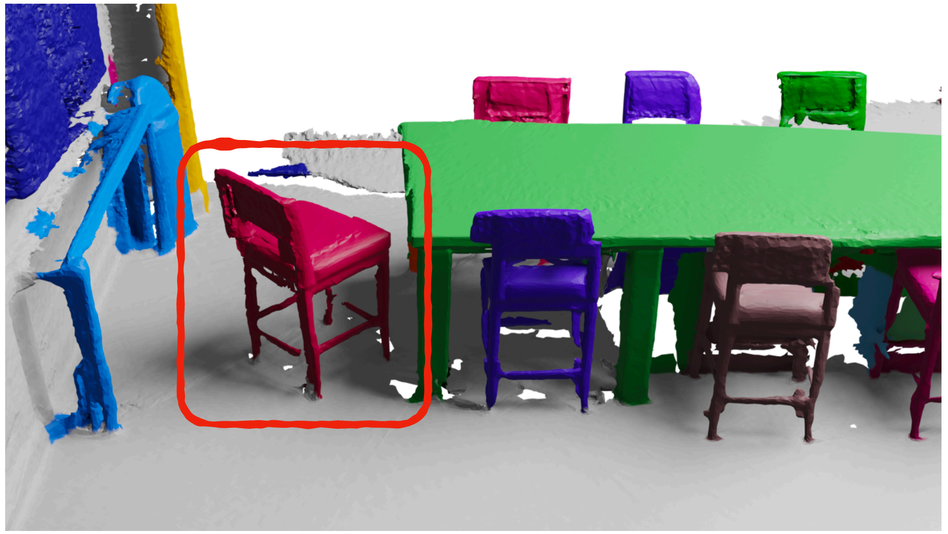
[1] Takmaz et al. "OpenMask3D: Open-Vocabulary 3D Instance Segmentation" NeurIPS'23
[2] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

OpenMask3D: Open-Vocabulary 3D Instance Segmentation

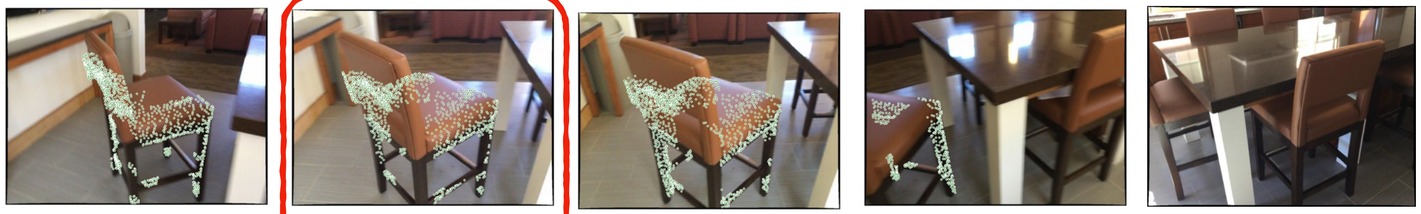
How to obtain the per-mask CLIP features?



3D Instance Masks Proposals
(class-agnostic)



Project 3D mask to 2D views



Visibility score: 100% 90% 94% 30% 0%

1. Compute tight bounding box via SAM.
2. Compute multi-scale CLIP features.
3. Average over multiple scales & views (top k views).

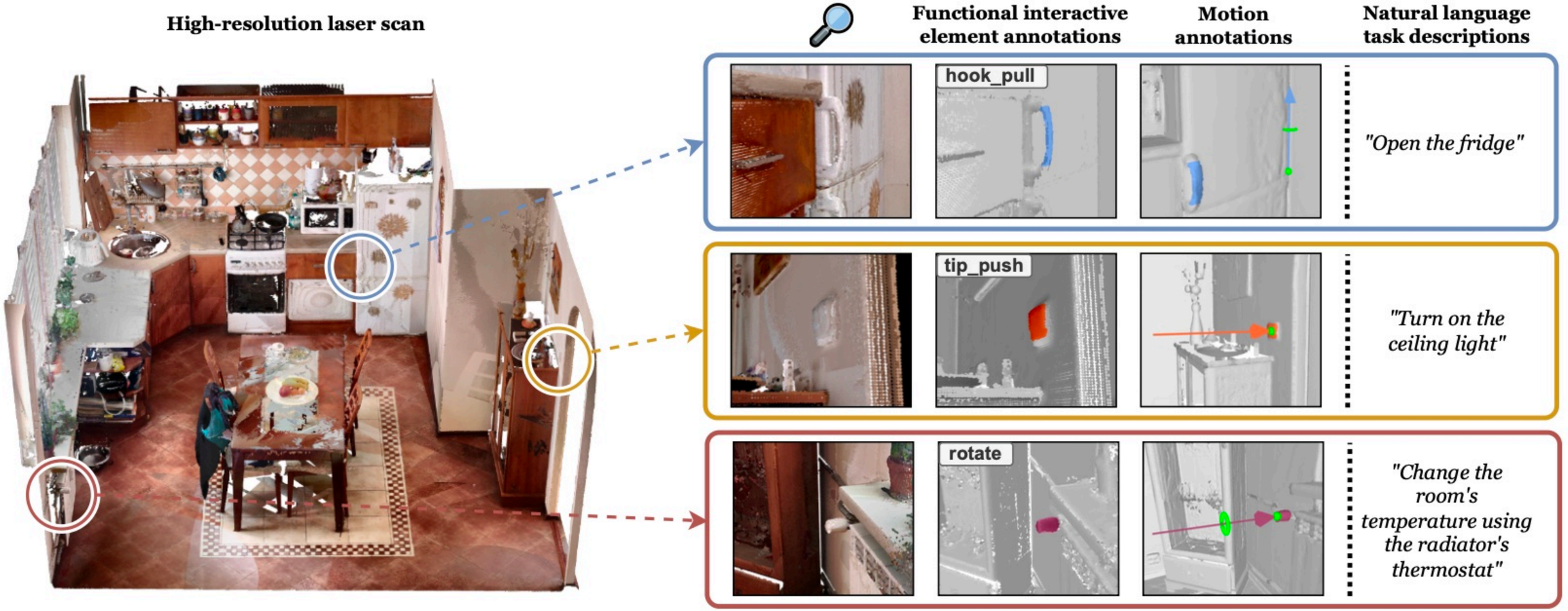
[1] Takmaz et al. "OpenMask3D: Open-Vocabulary 3D Instance Segmentation" NeurIPS'23
[2] Schult et al. "Mask3D: Mask Transformer for 3D Instance Segmentation" ICRA'23

DEMO

What is missing?

3D Interactive and Functional Elements

Beyond object-level scene understanding

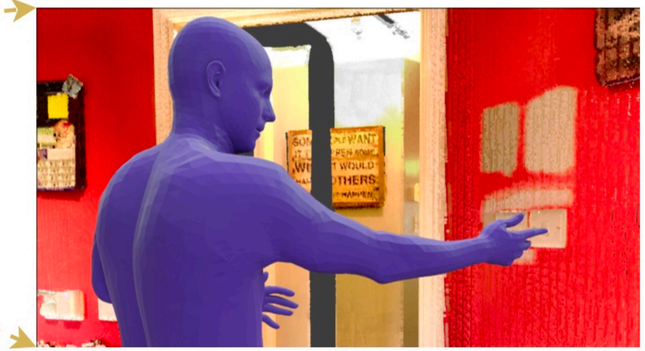


[1] Delitzas et al. "SceneFun3D: Fine-grained Functionality and Affordance Understanding in 3D Scenes" CVPR'24

3D Interactive and Functional Elements

Beyond object-level scene understanding: Down-stream tasks

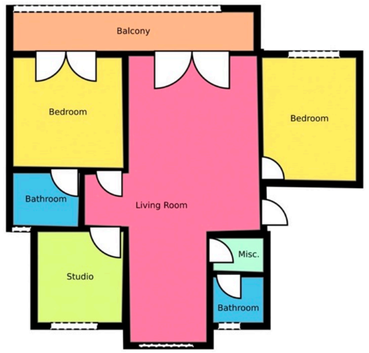
"Open the door"

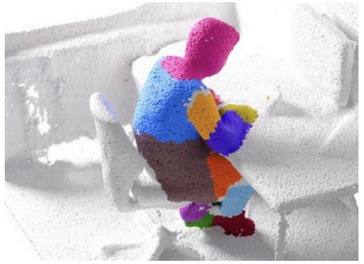


"Turn off the ceiling light"

[1] Delitzas et al. "SceneFun3D: Fine-grained Functionality and Affordance Understanding in 3D Scenes" CVPR'24

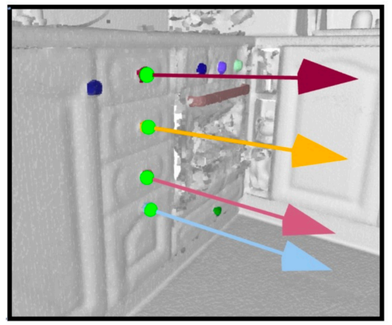
Next Steps: High-Fidelity Functional 3D Spaces





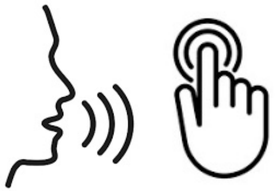
Multi-Level Perception

Vectorized Functional, Causal, Semantic 3D/4D Scene Representation

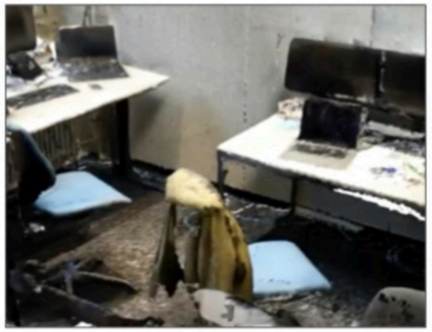


Learn common-sense functionalities from Human-Scene Interactions and foundation models

Ego-centric Vision *Ambient AI* *Speech* *Touch*



Fine-grained 3D Scene Generation with Functionalities



Collaborators


Marc Pollefeys


Federico Tombari


Siyu Tang


Konrad Schindler

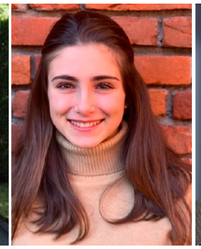

Or Litany

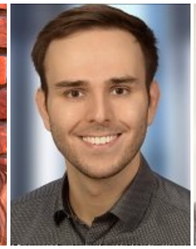

Gerard Pons-Moll

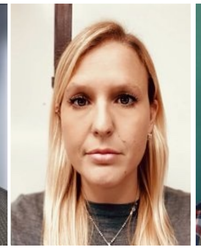

Bastian Leibe

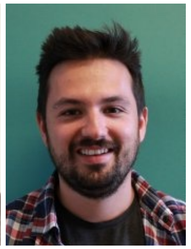

Alex Delitzas

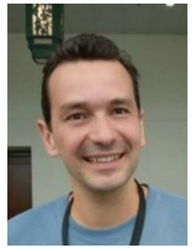

Ayca Takmaz


Elisabetta Fedele


Jonas Schult


Zuria Bauer


Michael Niemeyer


Vitto Ferrari


Yang Miao


Dora Kontogianni


Olga Vysotzka


Yuanwen Yue

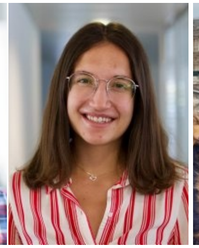

Songyou Peng


Silvan Weder


Despi Paschalidou


Alexey Nekrasov

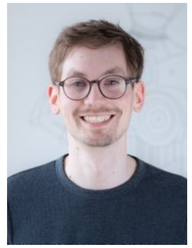

Kostas Rematas


Irem Kaftan


Idil Zulfikar

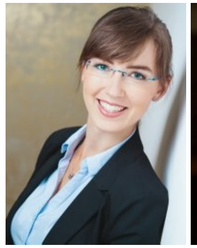

Oliver Lemke

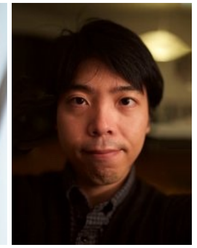

Julian Chibane


Hermann Blum


Daniel Barath


Johanna Wald


Cathrin Elich


Keisuke Tateno


Xi Wang


GREEK WATERS PILOT

Open-Vocabulary
3D Scene Understanding
towards Embodied Manipulation

Francis Engelmann, Computer Vision and Geometry Group, ETH Zurich
ETH AI Center Postdoctoral Fellow | June 7th, 2024

ETH zürich

ETH AI CENTER

D INFK