Aalto-yliopisto
Aalto-universitetet
Aalto University

UNIVERSITY OF TURKU

Semantic Mapping
From Segmentation to Scene Graphs

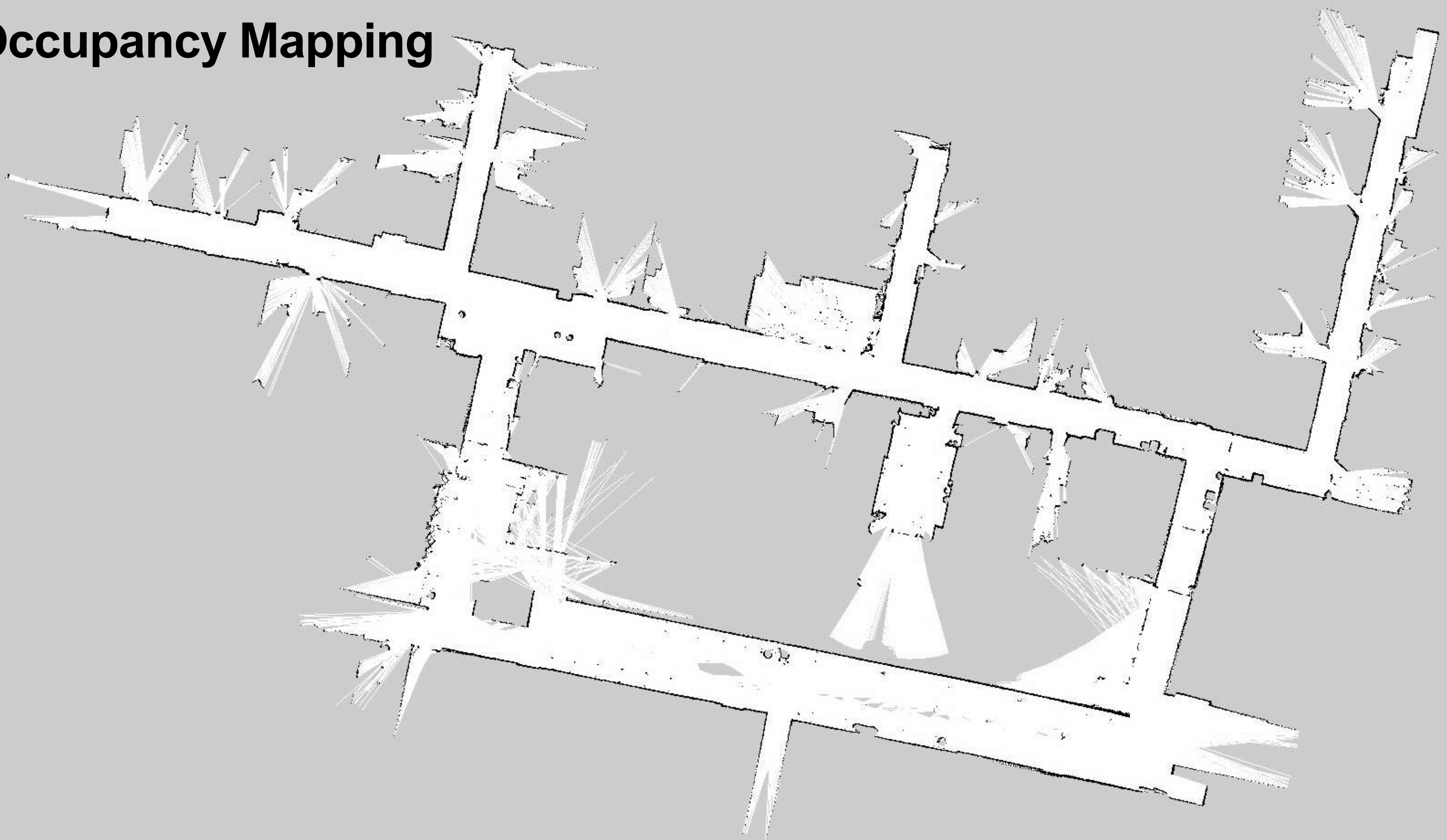
Francesco Verdoja
Academy Research Fellow
Dept. of Electrical Engineering and Automation

9.12.2025

Lecture outline

- Recap of occupancy mapping
- Semantics?
- Metric-semantic mapping
- Open-vocabulary semantic mapping
- Scene Graphs
- Adding time to semantic maps

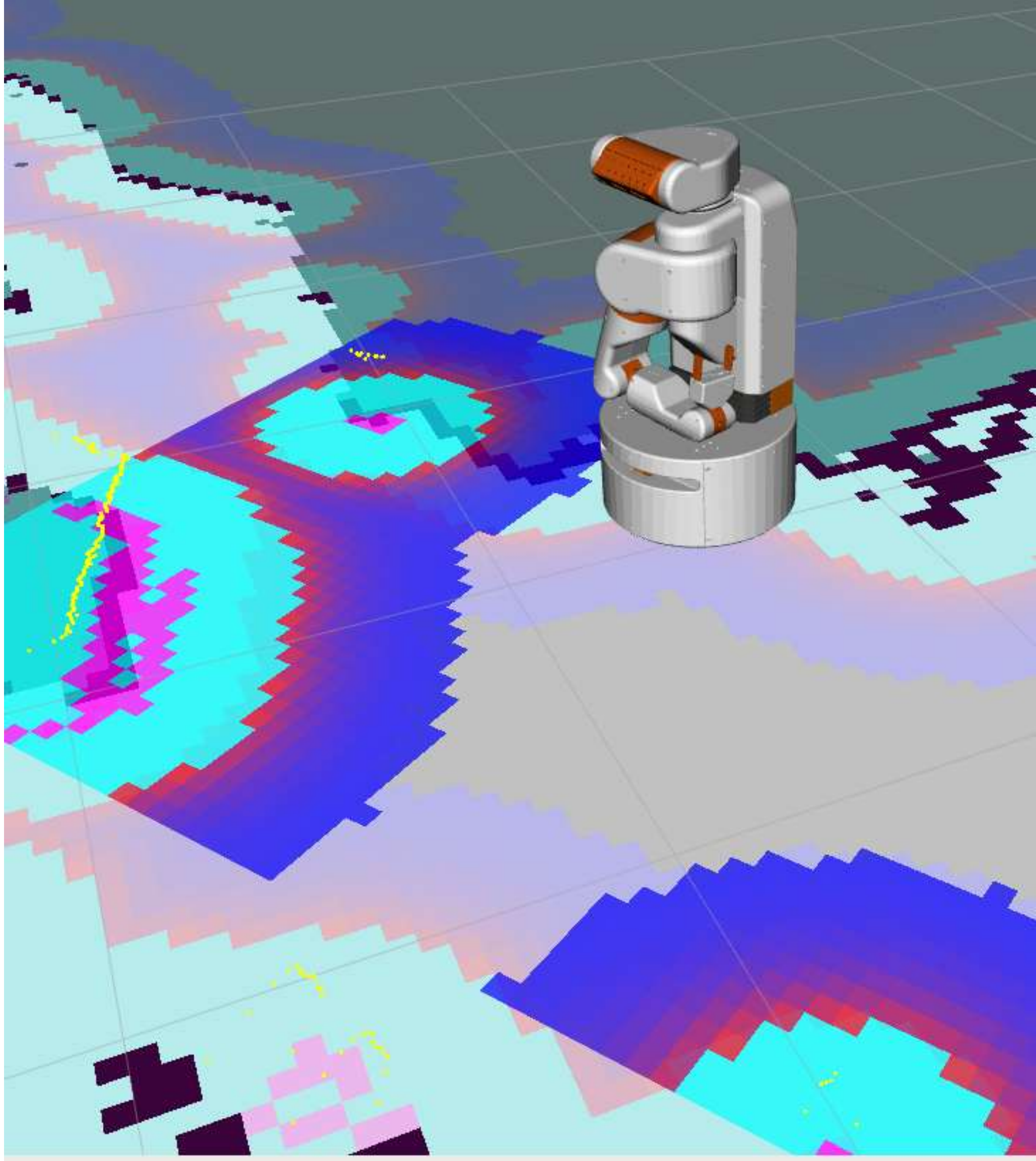
Occupancy Mapping



Capabilities supported by occupancy maps

- Global path planning given a goal point
- Local path planning / Obstacle avoidance
- Replanning around blockages
- Localization
- Navigation
- Docking
- Frontier-based exploration
- ...

A!



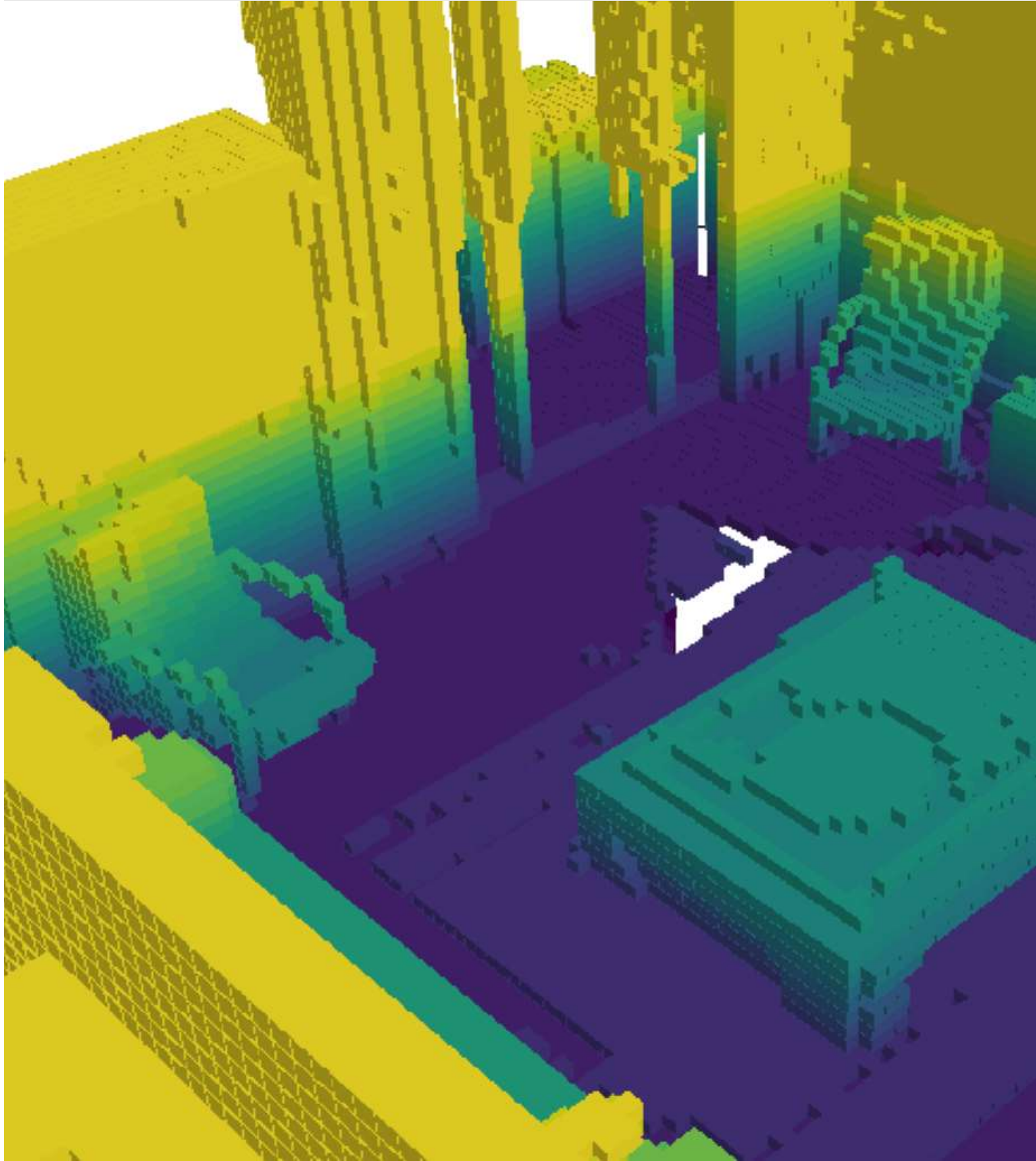
Applications

Limitations

“Move the chair closer to the table”

- Object instances?
- Object extents?
- Affordances?
- Grid independence?

Occupancy not enough for complex tasks (mobile manipulation, natural HRI)



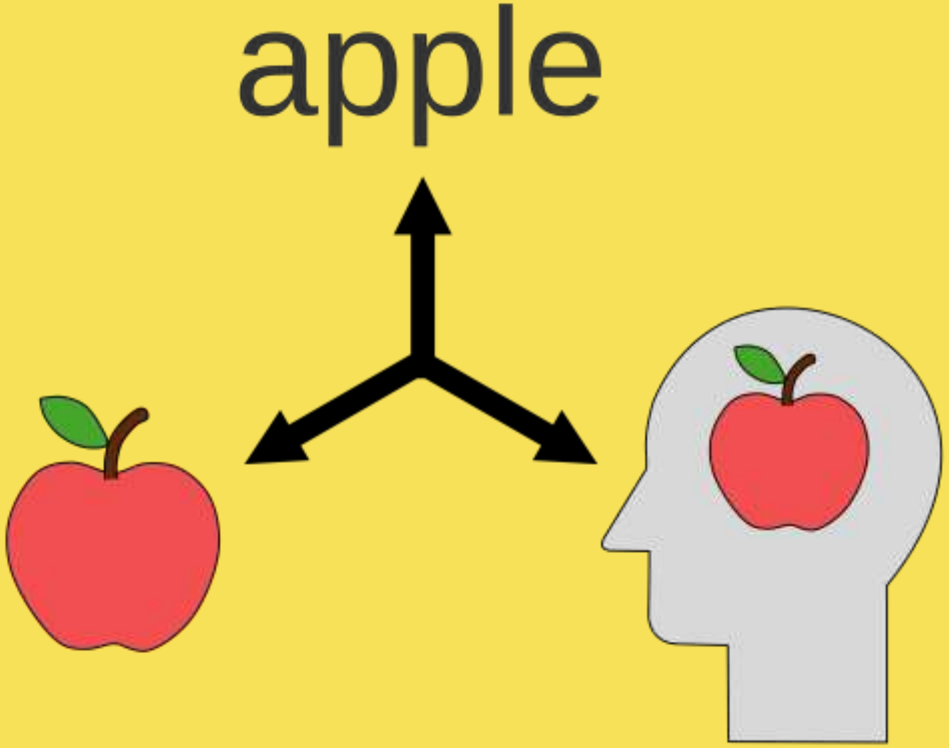
Semantics?

Semantics

- ***Meaning*** of things and words
- ***Grounding*** of symbols in reality

Enable decisions that depend on:

- Object identity (shelf vs wall)
- Function (charging stations)
- Affordances (door is *openable*)
- Human language (*“move the chair”*)



Semantic Mapping

Vocabulary

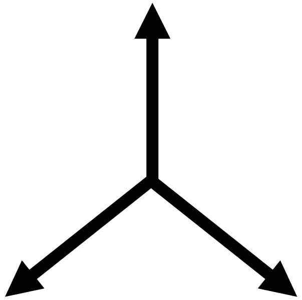
Chair [seat, move, …]

Table [carrier, …]

Mug [drink, move, …]

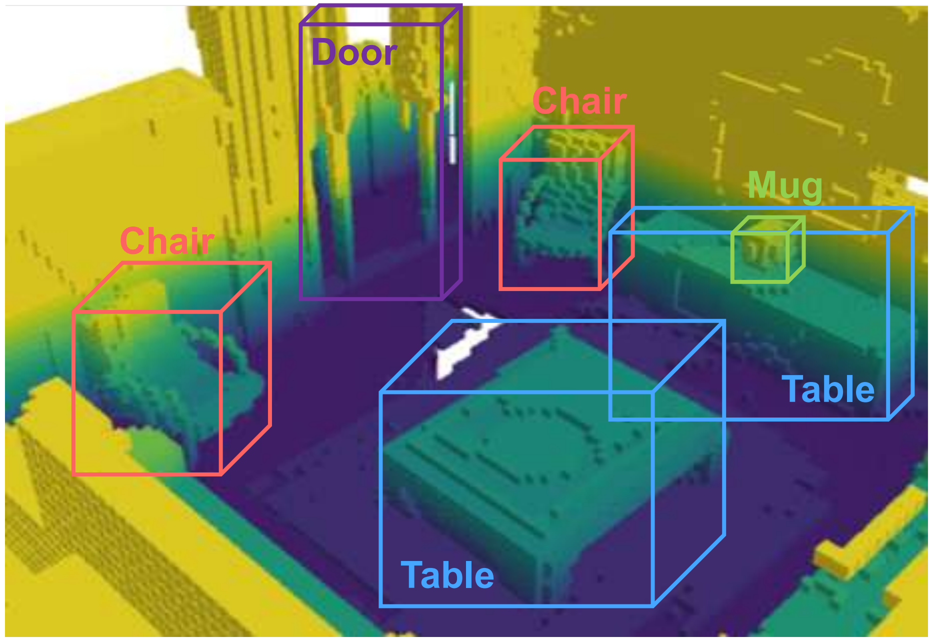
Door [open, close, pass, …]

…



Perception

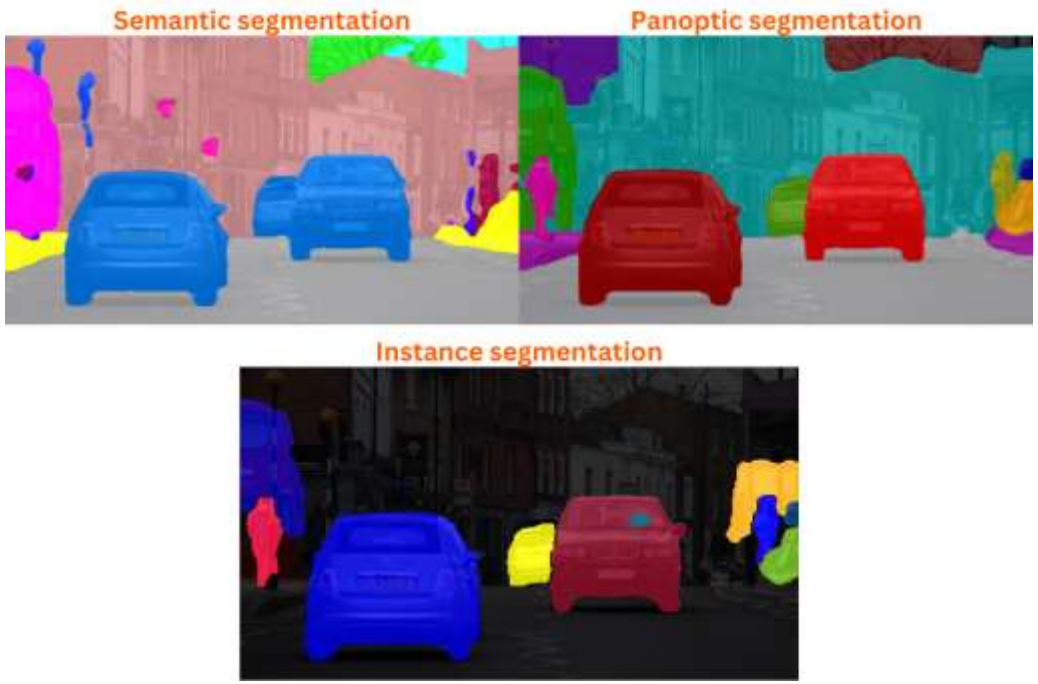
Mapping



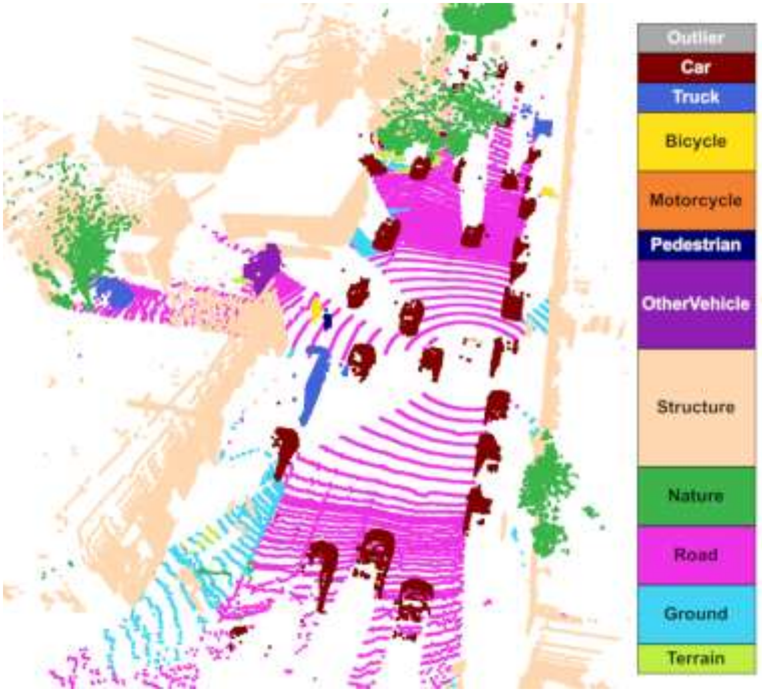
A!

Where do you get semantics?

RGB camera
Mask2Former, YOLO, Segment Anything…



LiDAR
RangeNet++, RandLA-net…



Requires raycasting!

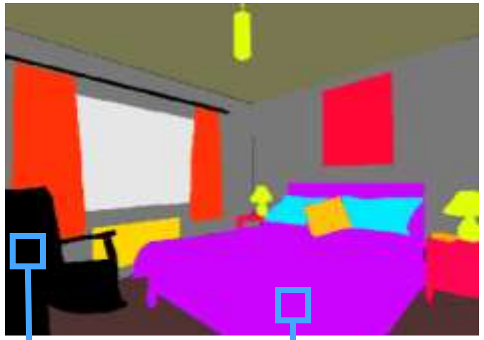
Metric-semantic mapping

Extending Occupancy Mapping

- Occupancy maps: O = {occupied, free}
- Semantic maps: C = {chair, table, door, mug, …}
- Both are **categorical distributions** $\mathrm{Cat}(K, \boldsymbol{p})$
  - $K > 0$ number of categories ($K = 2$ for occupancy map, *Bernoulli distribution*)
  - $\boldsymbol{p} = (p_1, p_2, \dots, p_\mathrm{K})$ probabilities of individual categories ($p_i \geq 0, \sum p_i = 1$)
  - Mode (i.e., most likely category): $i \mid p_i = \max(p_1, \dots, p_\mathrm{K})$

Semantic mapping as Bayesian inference

- ***For any map voxel*** $\mathrm{Cat}(K, \boldsymbol{v})$***:***
  $\boldsymbol{v} = (v_1, \dots, v_K)$ where $v_i \geq 0$ and $\sum v_i = 1$
- ***Measurement (one-hot):***
  $\mathbf{y} = (y_1, \dots, y_{\mathrm{K}})$, where $y_{\mathrm{i}} \in \{0,1\}$ and $\sum y_i = 1$
- ***Categorical likelihood:*** $p(\boldsymbol{y}|\boldsymbol{v}) = \prod v_i^{y_i}$



$\boldsymbol{y} = (0,0,1,0, \dots)$

chair

$\boldsymbol{y} = (0,1,0,0, \dots)$

bed

But how to find posterior $p(\boldsymbol{v}|\boldsymbol{y})$?

Dirichlet conjugate prior for categorical distributions

For any categorical distribution $\mathrm{Cat}(K, \boldsymbol{v})$:

- Given a **concentration hyperparameter** $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$
- **Dirichlet (conjugate) prior:** $p(\boldsymbol{v}|\boldsymbol{\alpha}) \sim \mathrm{Dir}(K, \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod v_i^{\alpha_i - 1}$
- $\boldsymbol{c} = (c_1, \ldots, c_K)$, number of observations of each category
- $\boldsymbol{\alpha}' = \boldsymbol{c} + \boldsymbol{\alpha} = (c_1 + \alpha_1, \ldots, c_K + \alpha_K)$, and $S(\boldsymbol{\alpha}') = \sum \alpha_i'$

Then:

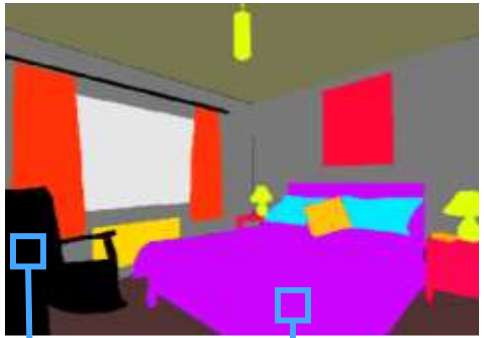
$$p(\boldsymbol{v}|\boldsymbol{c}, \boldsymbol{\alpha}) \sim \mathrm{Dir}(\boldsymbol{c} + \boldsymbol{\alpha}) \sim \mathrm{Dir}(\boldsymbol{\alpha}')$$

$$\mathbb{E}[v_i] = \frac{\alpha_i'}{S(\boldsymbol{\alpha}')} \qquad \mathbb{V}[v_i] = \frac{\alpha_i'(S(\boldsymbol{\alpha}') - \alpha_i')}{S(\boldsymbol{\alpha}')^2(S(\boldsymbol{\alpha}') + 1)}$$

A!

Back to semantic mapping

$\boldsymbol{\alpha} = (1,1,1,1,1,1,1,1,1,1)$



chair

$\boldsymbol{y} = (0,0,1,0,\dots)$

bed

$\boldsymbol{y} = (0,1,0,0,\dots)$

$\boldsymbol{\alpha}' = (1,1,2,1,\dots)$

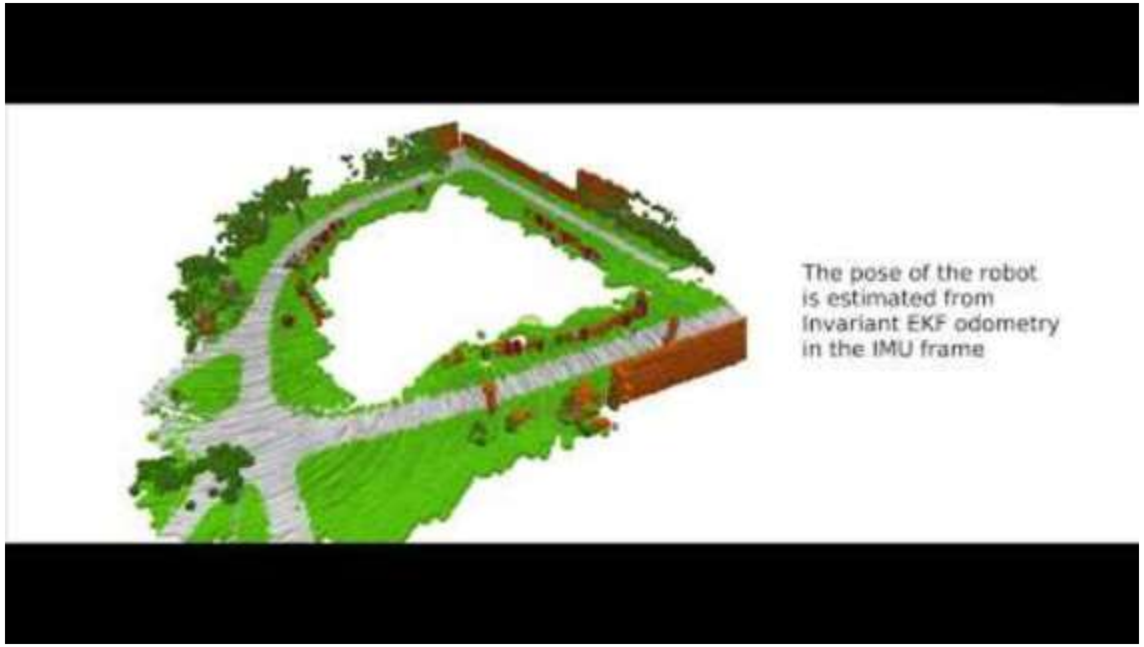
$\boldsymbol{\alpha}' = (1,2,1,1,\dots)$

$S(\boldsymbol{\alpha}') = 10$

$\boldsymbol{v} = (0.1, 0.1, 0.2, \dots)$

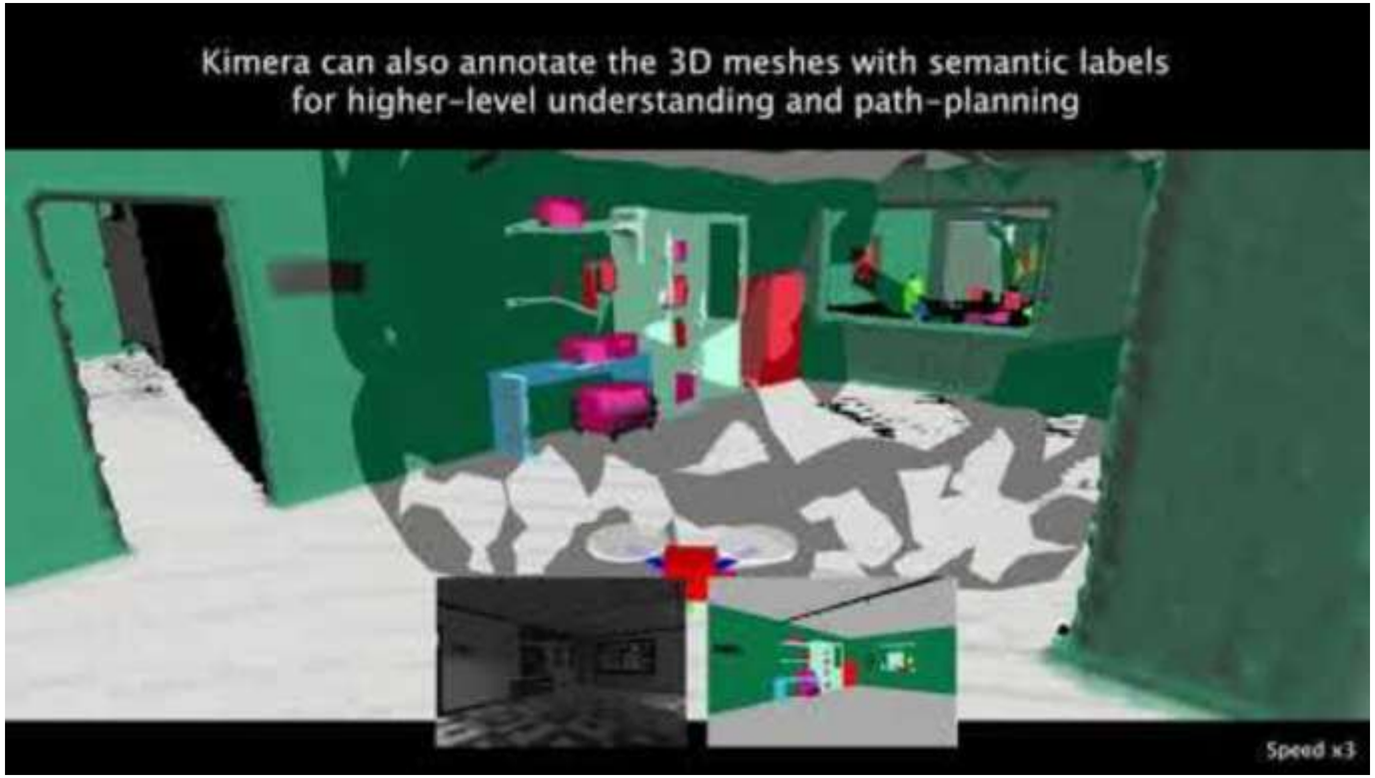
$\boldsymbol{v} = (0.1, 0.2, 0.1, \dots)$

$$\mathbb{E}[v_i] = \frac{\alpha_i'}{S(\boldsymbol{\alpha}')} \text{ and } \mathbb{E}[\boldsymbol{v}] = \operatorname{argmax}_i(v_i)$$



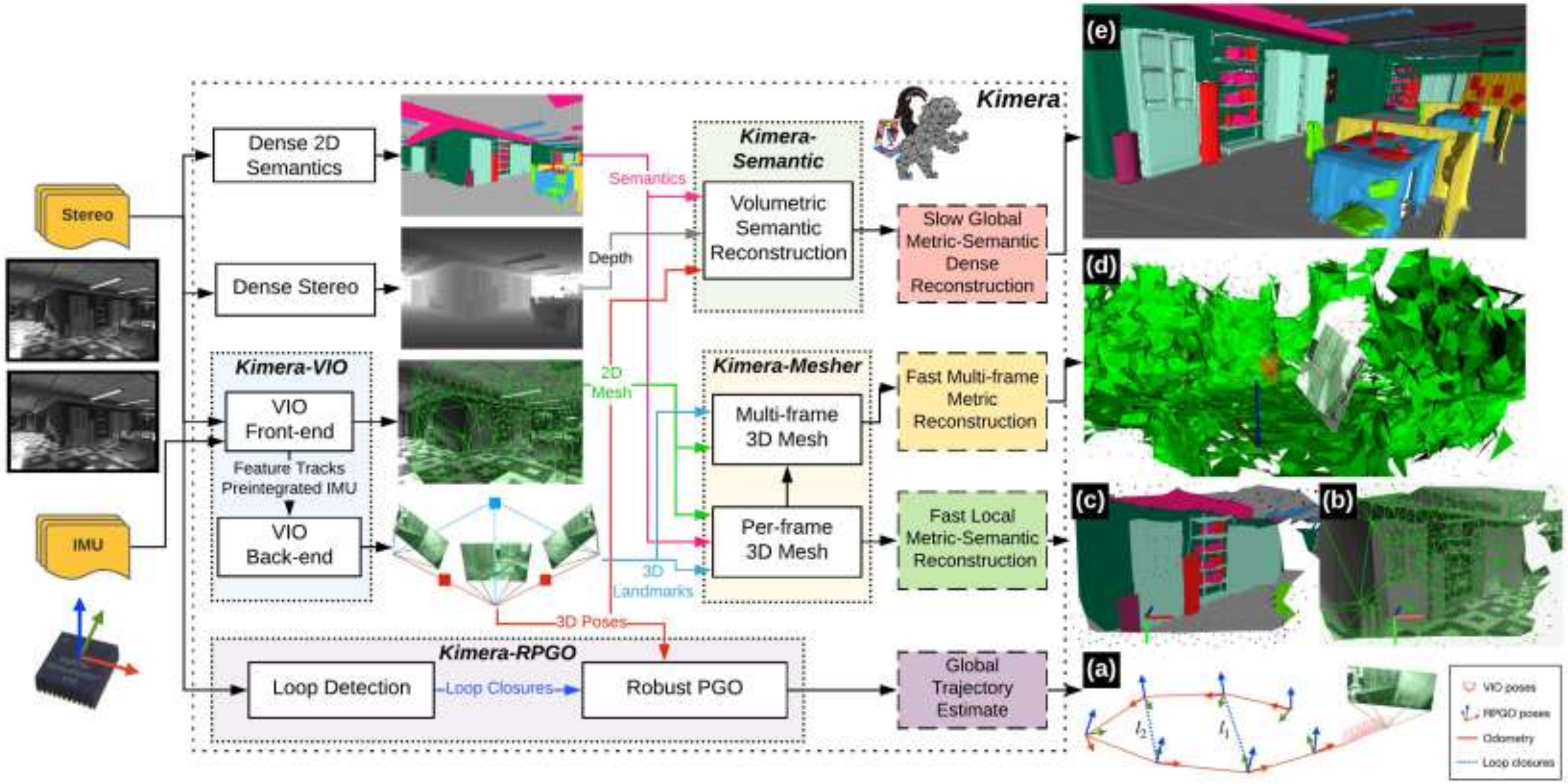
Gan, Lu, et al. "Bayesian spatial kernel smoothing for scalable dense semantic mapping." *IEEE Robotics and Automation Letters* 5.2 (2020): 790-797.

3D meshes instead of voxels (e.g., KIMERA)



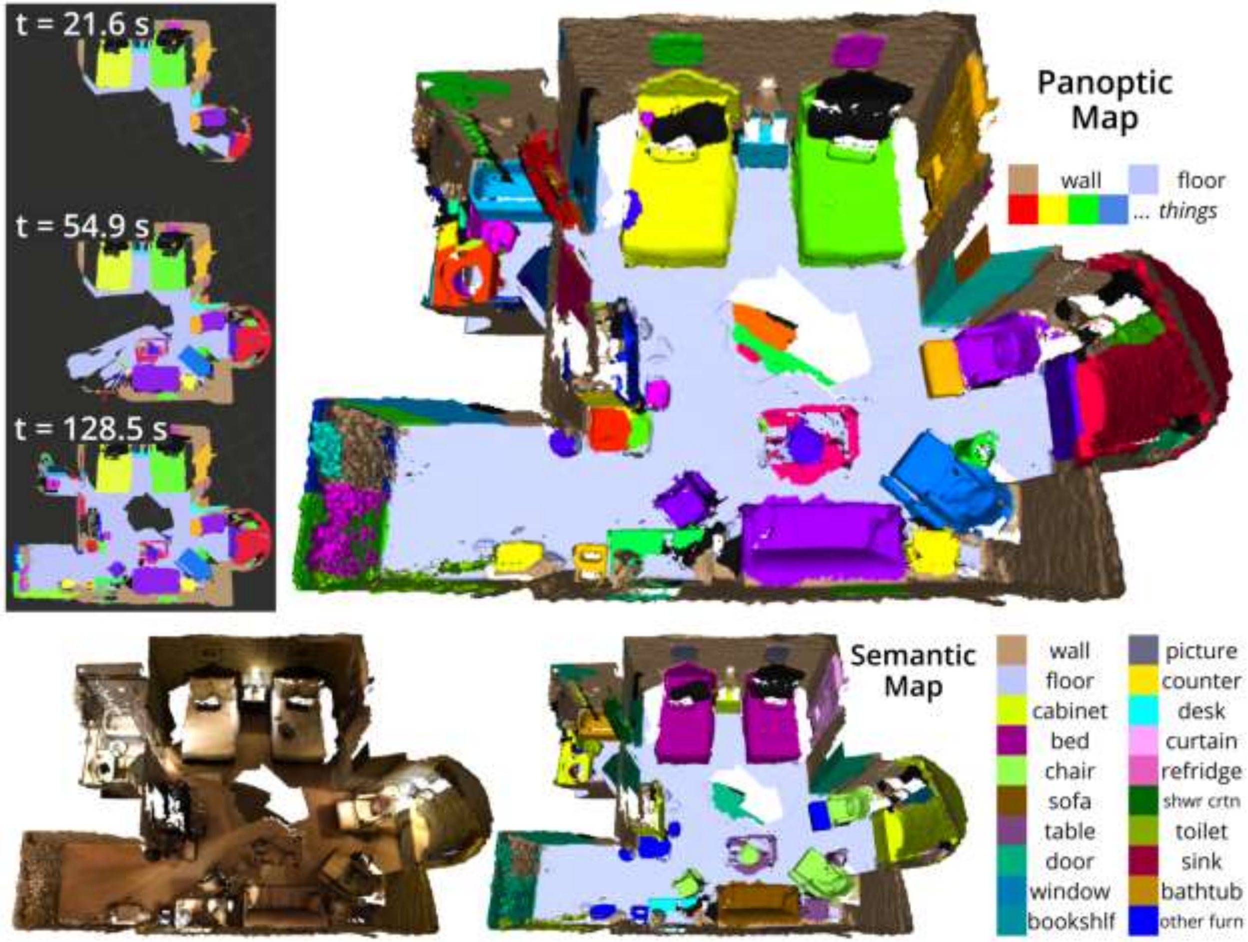
A. Rosinol, et al. “Kimera: From SLAM to spatial perception with 3D dynamic scene graphs,”
The Int. J. of Robotics Research, vol. 40, no. 12-14, pp. 1510–1546, 2021

A lot is borrowed from Visual SLAM



A!

Panoptic Maps for object instances

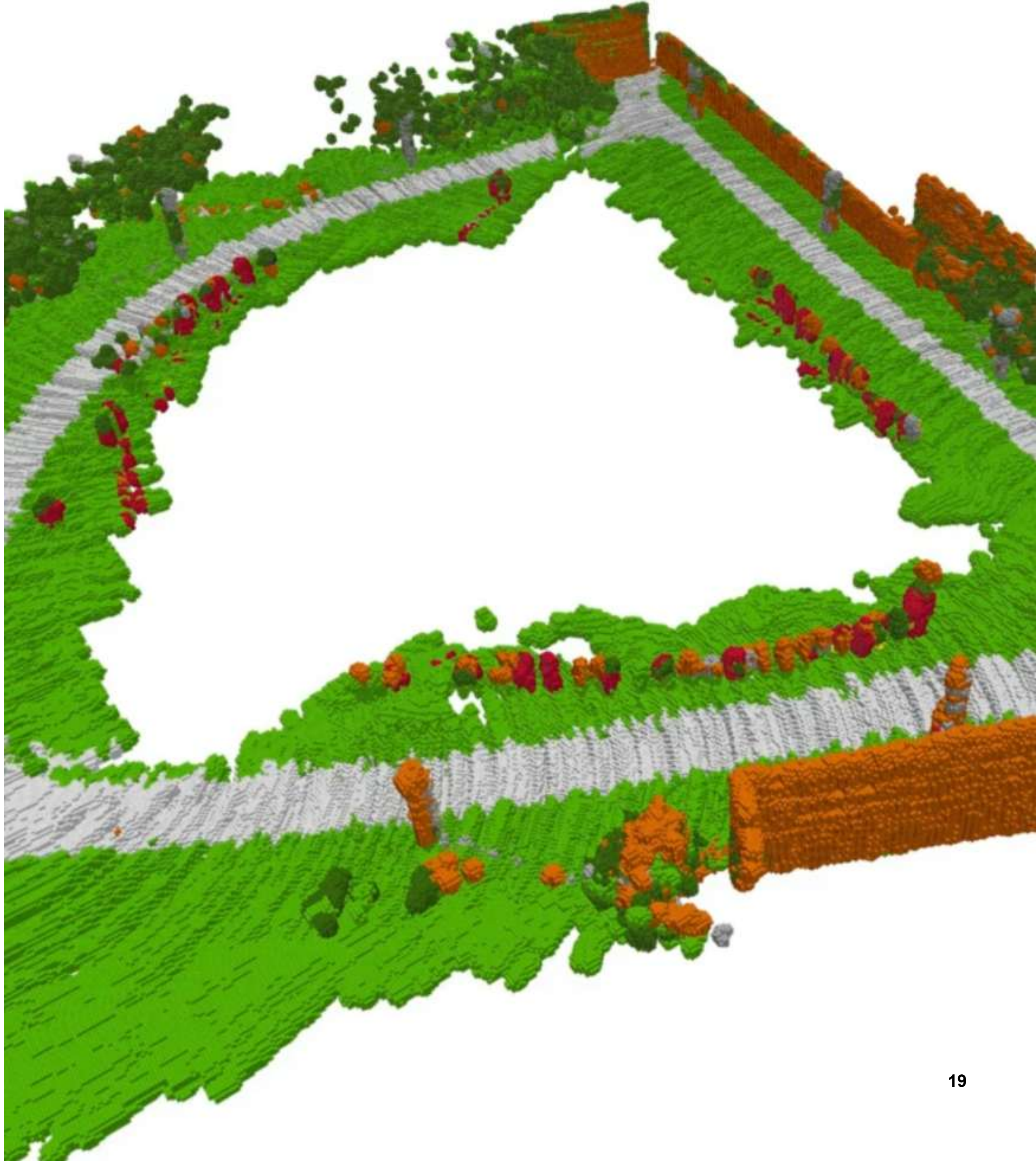


Narita, Gaku, et al. "Panopticfusion: Online volumetric semantic mapping at the level of stuff and things.« *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

A!

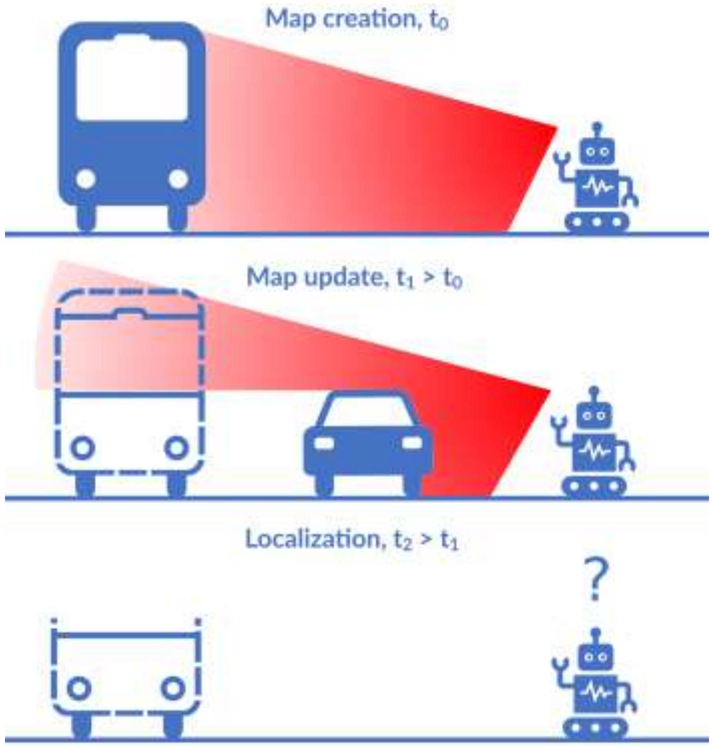
What have we fixed?

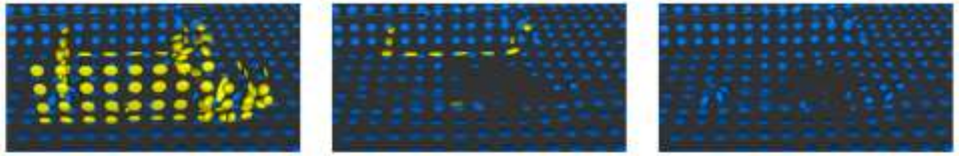
- Object instances? ◑
- Object extents? ✕
- Affordances? ◑
- Grid Independence? ✕



A!

Some efforts on addressing grid independence



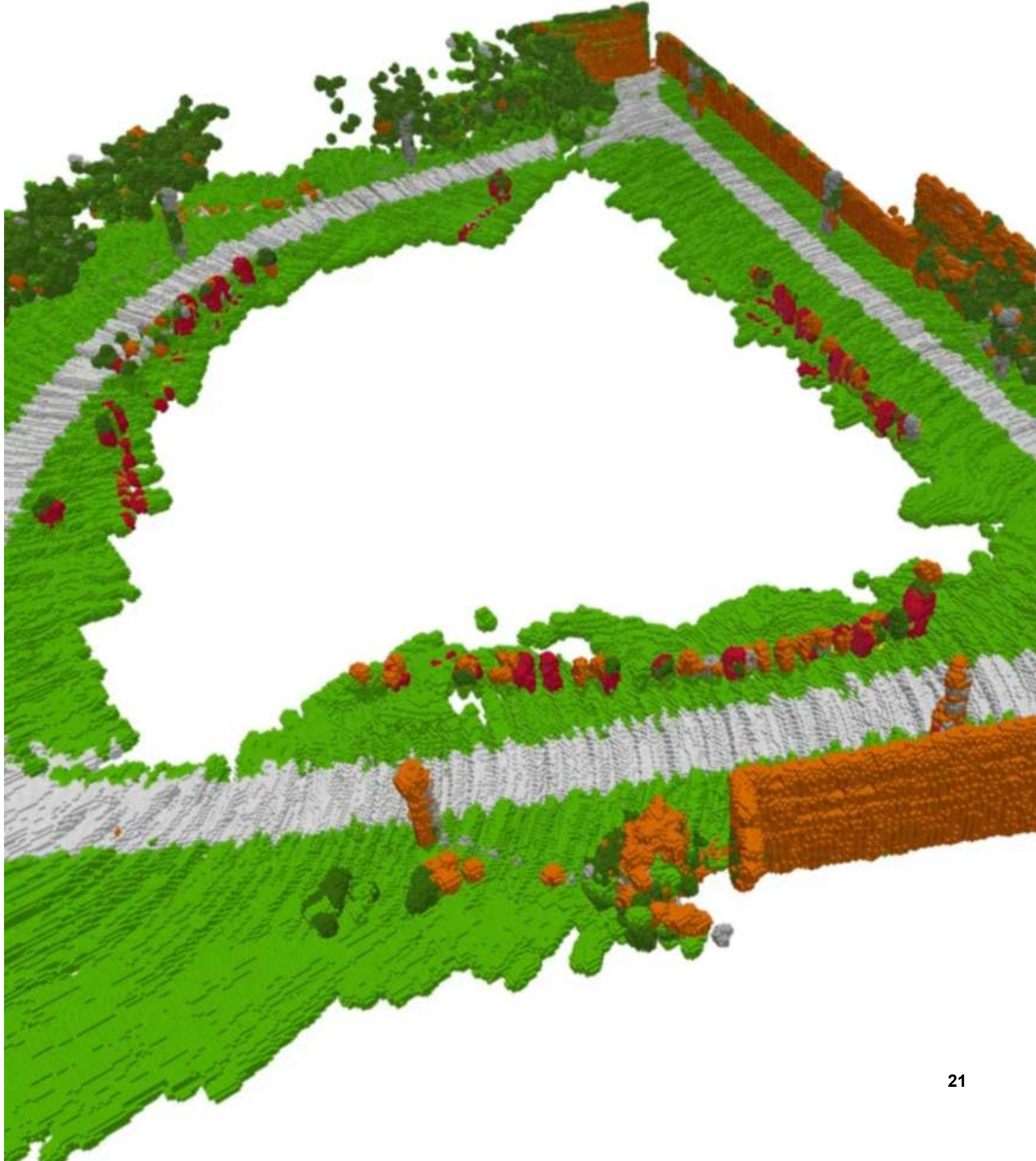


(a) Before the update (b) NDT-OM (c) Our method

Object-Oriented Grid Mapping in Dynamic Environments
Matti Pekkanen, Francesco Verdoja, and Ville Kyrki
2024 IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2024

What can we do?

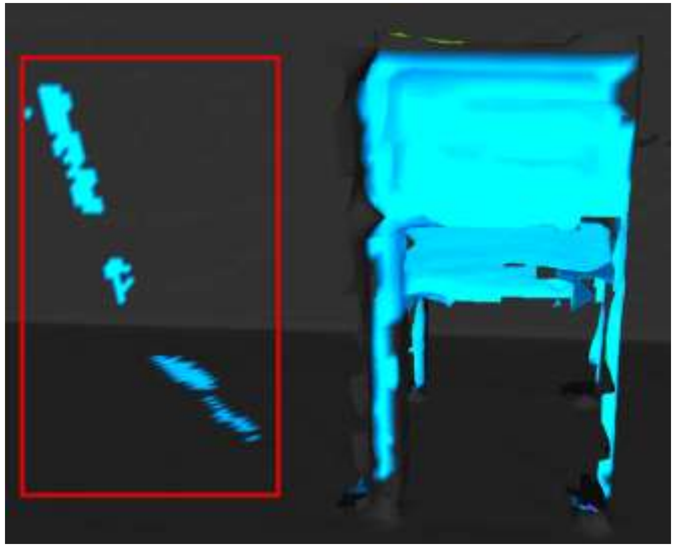
- Semantic-aware navigation
  - “stay on the road”
  - “stop at pedestrian crossings”
  - “never go closer than 2m to a tree”
  - “rest close to a wall”



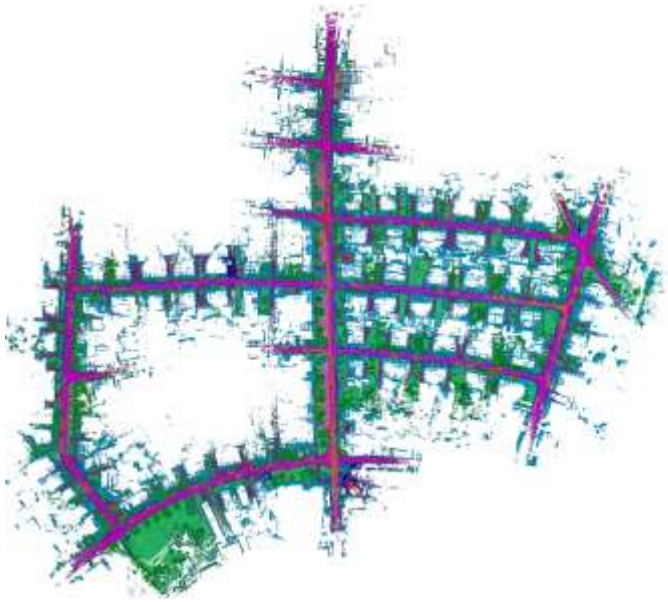
A!

New challenges

Class bleeding at boundaries (RGB+D calibration)



High memory footprint (submapping)



Complex map update and vocabolary extension

Beyond metric-semantic mapping

Open-vocabulary

Not limited to a closed set of predefined semantic labels

From voxels to concepts

Voxels are part of objects, rooms, and other semantic entities

From 3D to 4D+

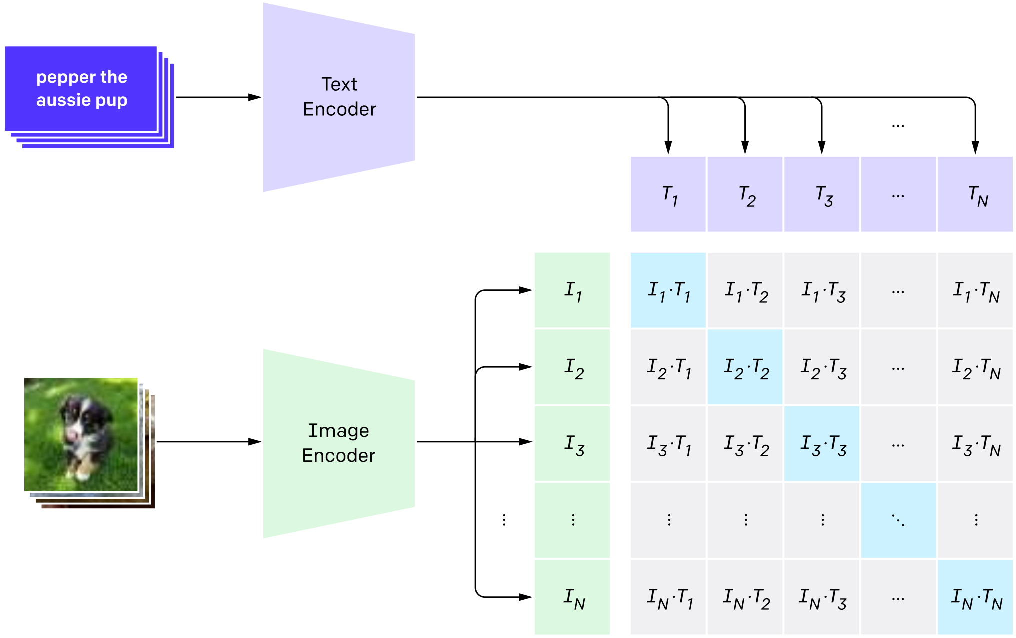
Reasoning and handling dynamic environments over time

A!

Open-vocabulary semantic maps

Visual-Language Models

- Coupled Transformer Neural Networks
  - Text: to N-dim embedding T
  - Images: to N-dim embeddings I
- Trained on (image, text caption) dataset
- Minimize distance between T and I for (image, caption) pair
- Maximize distance between T and I for non-pairs
- **CLIP** from OpenAI: 512-dim embedding



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.

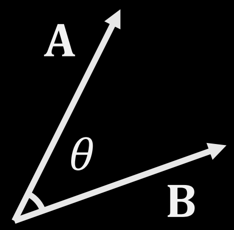
Embeddings and Cosine Similarity

Embeddings

- N-dim vectors $\mathbf{A} = (A_1, \ldots, A_N)$
- Often for VLMs they are unit-size, i.e., $\|\mathbf{A}\| = 1$

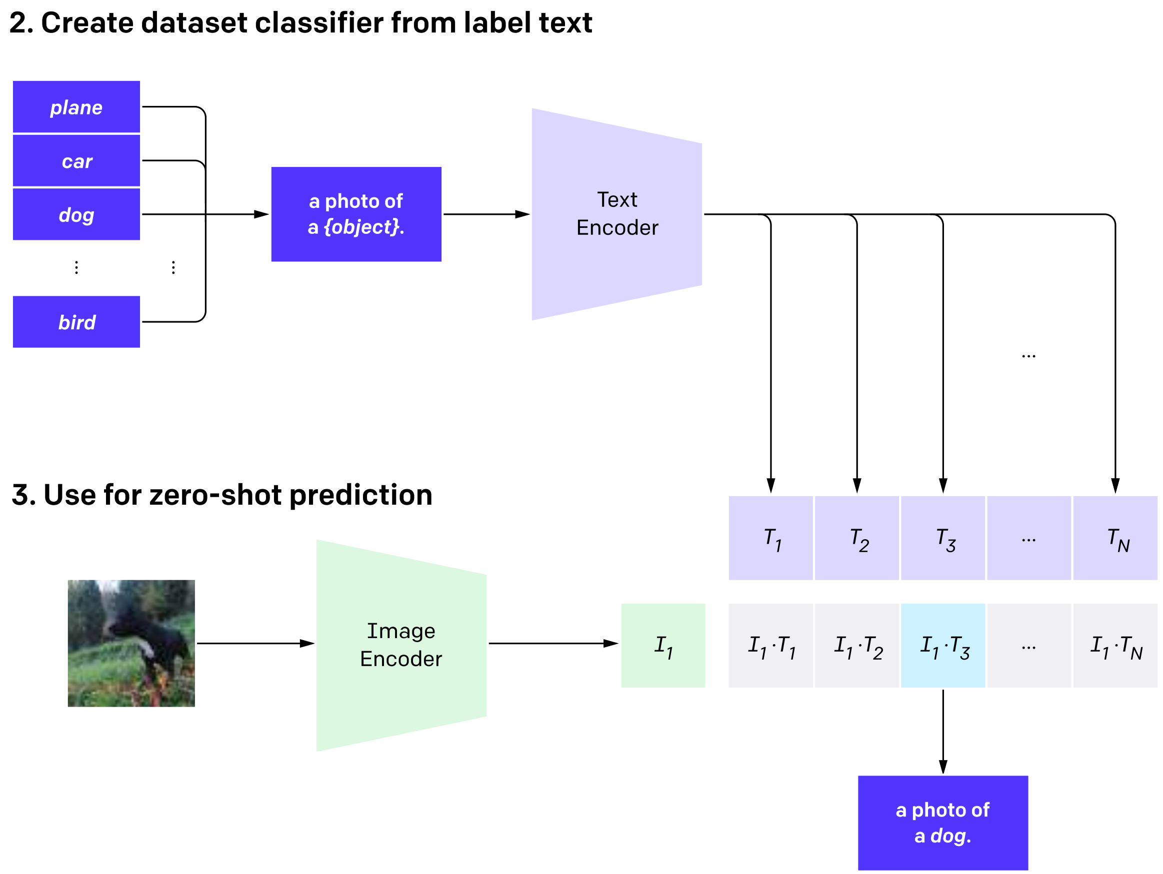
Cosine similarity

- $S_C(A, B) \coloneqq \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{\mathrm{N}} A_i B_i}{\sqrt{\sum_{i=1}^{\mathrm{N}} A_i^2} \cdot \sqrt{\sum_{i=1}^{\mathrm{N}} B_i^2}}$



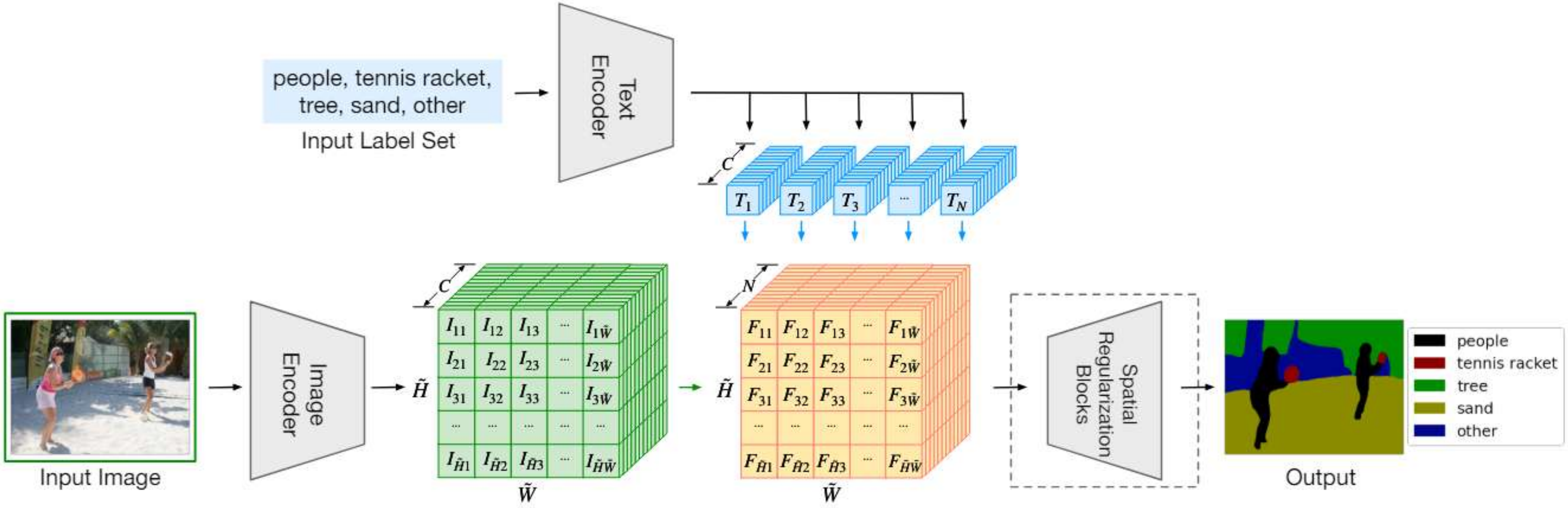
- $S_c \in [-1,1]$, with -1 opposite, +1 same, 0 orthogonal

Querying the model



A!

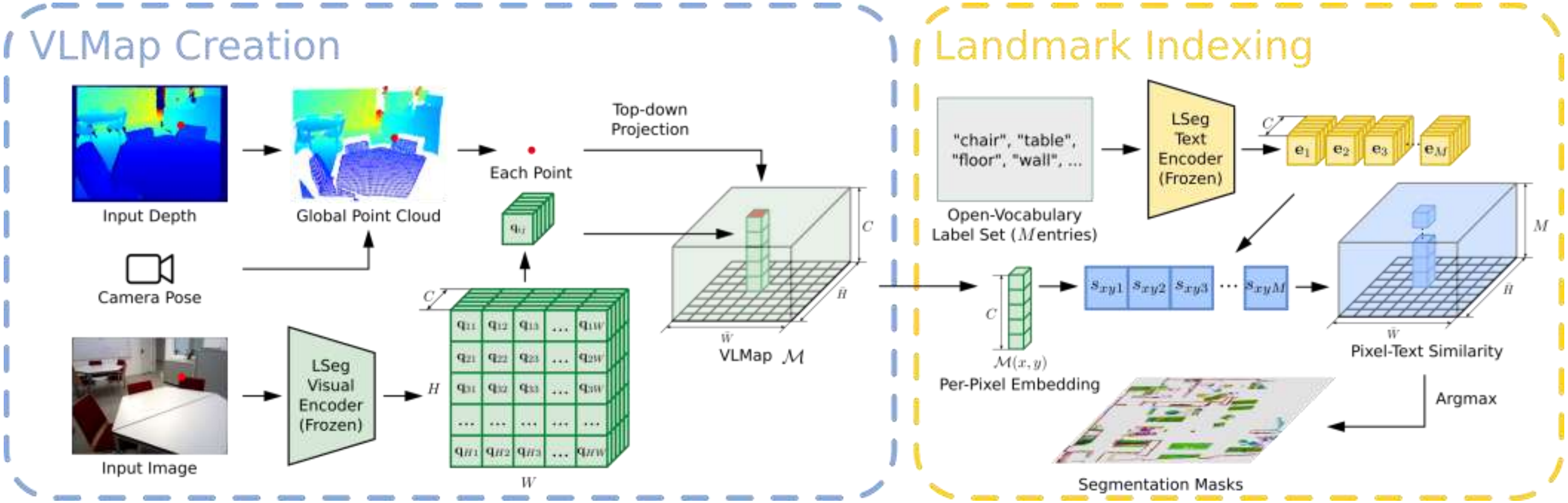
Pixel-level CLIP embeddings (e.g., LSeg)



Li, Boyi, et al. "Language-driven Semantic Segmentation."
2022 International Conference on Learning Representations (ICLR), 2022.

A!

Maps of Embeddings (e.g., VLMap)

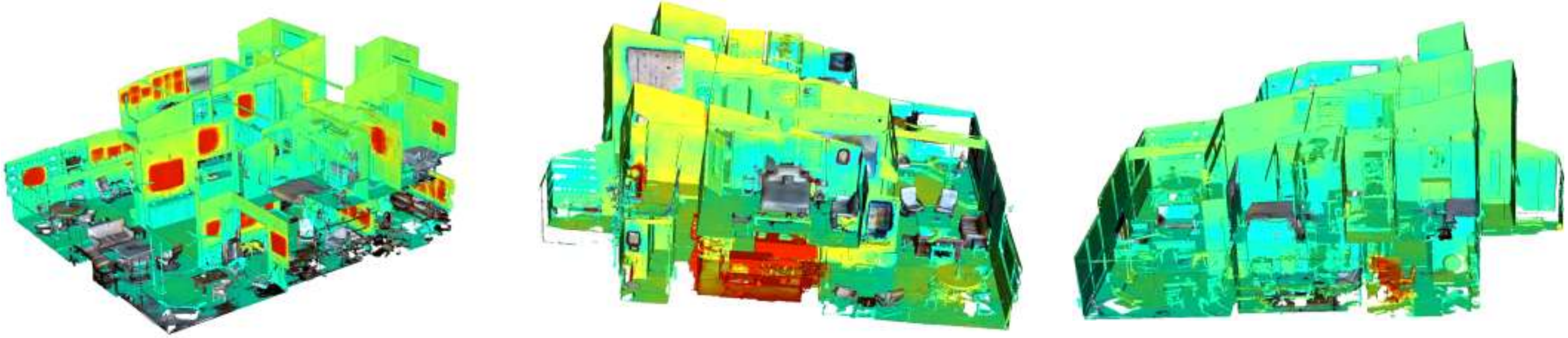


Huang, Chenguang, et al. "Visual Language Maps for Robot Navigation."
2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023.

A!

Open-vocabulary Querying

Painting Kitchen Work



Matti Pekkanen, Tsvetomila Mihaylova, Francesco Verdoja, and Ville Kyrki, “Do Visual-Language Grid Maps Capture Latent Semantics?” *2025 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2025.

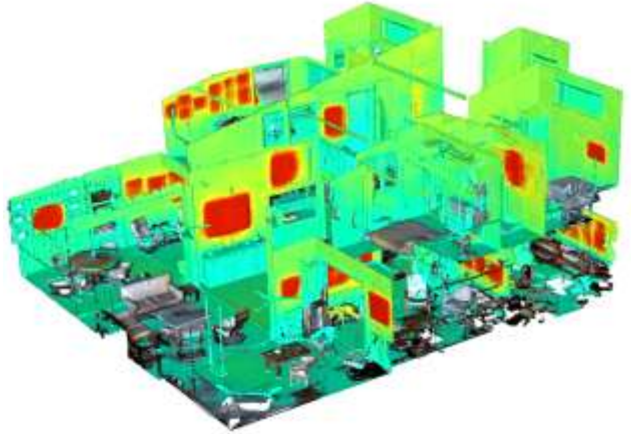
Robot interaction and planning (e.g., NLMap + SayCan)

Chen, Boyuan, et al. "Open-vocabulary Queryable Scene Representations for Real World Planning." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

Challenges

Quality is highly dependant on VLM performance

Maps are even larger (WxLxHx512)

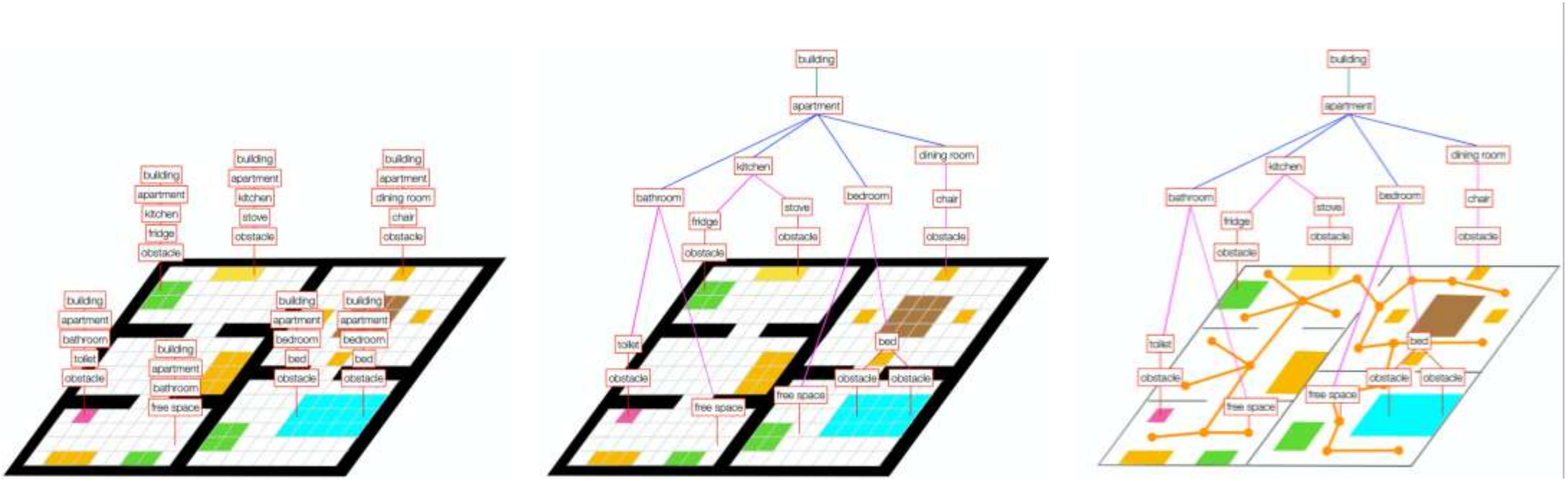


What happens for queries with missing target?

A!

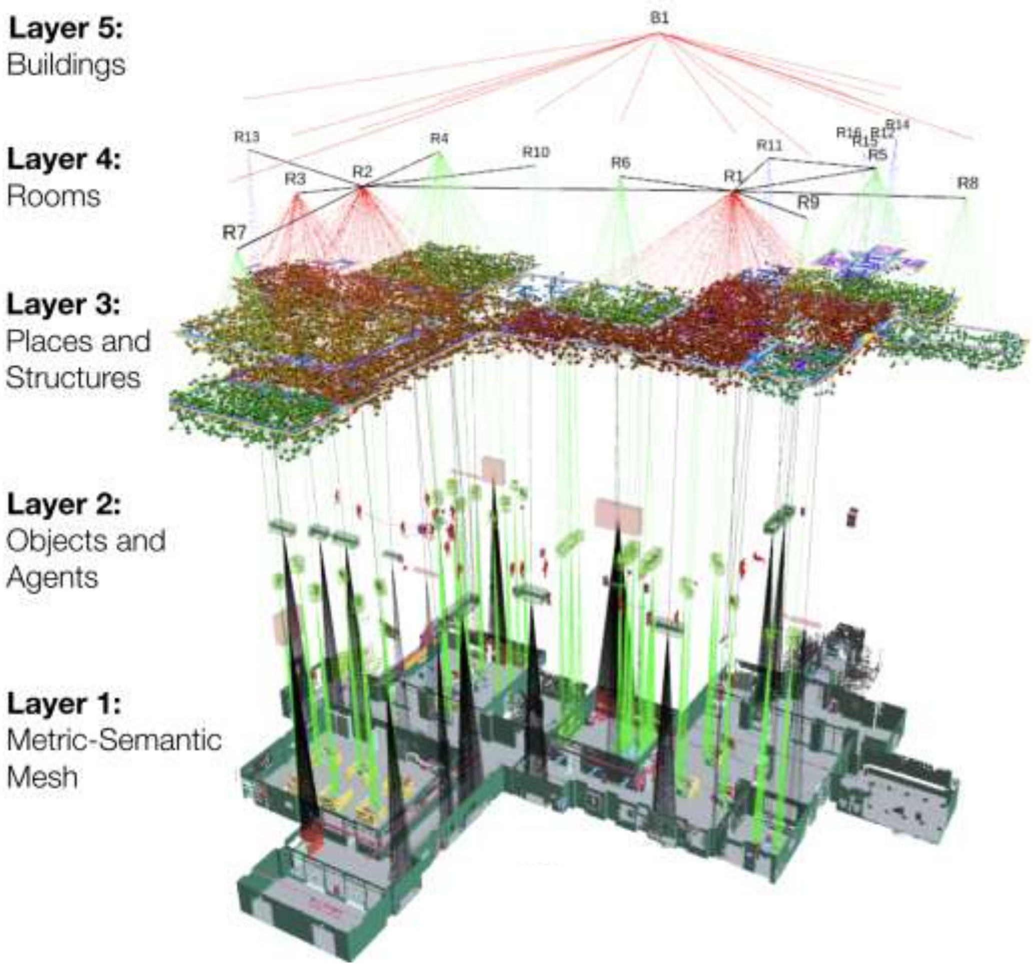
From voxels to concepts

The environment is hierarchical



3D Scene Graphs (3DSGs)

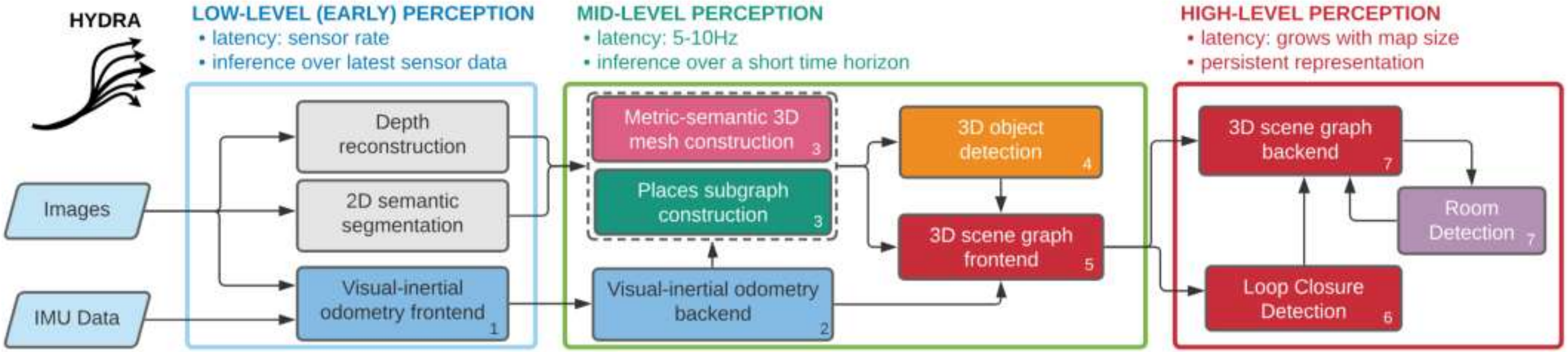
- **Hierarchical graph representation**
- **Objects, places, and rooms as nodes**
  - attributes (pose, shape, affordances)
  - connected to 3D mesh
  - Belong to semantic layers
- **Edges describe relations**
  - spatial (adjacency, inclusion, support)
  - functional (used-for, part-of)
  - …



A. Rosinol, et al. “Kimera: From SLAM to spatial perception with 3D dynamic scene graphs,” *The Int. J. of Robotics Research*, vol. 40, no. 12-14, pp. 1510–1546, 2021

A!

Building a 3D Scene Graph

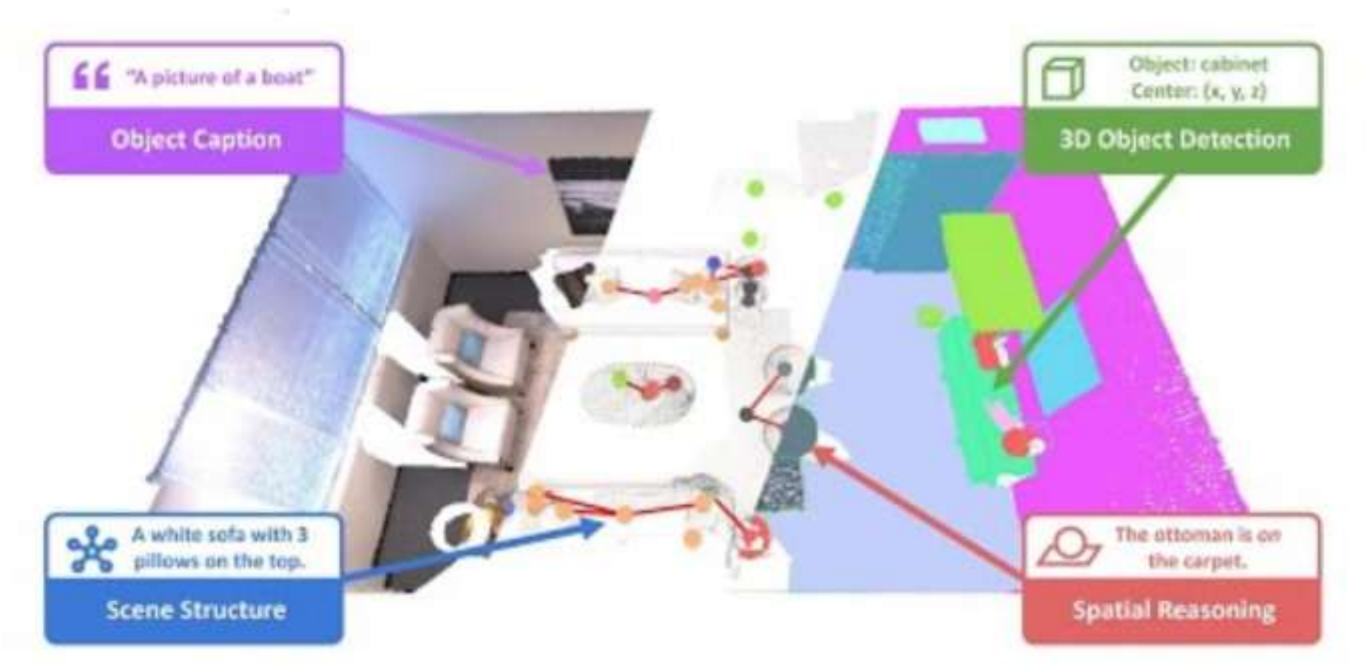


Hughes, Nathan, Yun Chang, and Luca Carlone. "Hydra: A Real-time Spatial Perception System for 3D Scene Graph Construction and Optimization." *Robotics: Science and Systems*. 2022.

Hydra in action

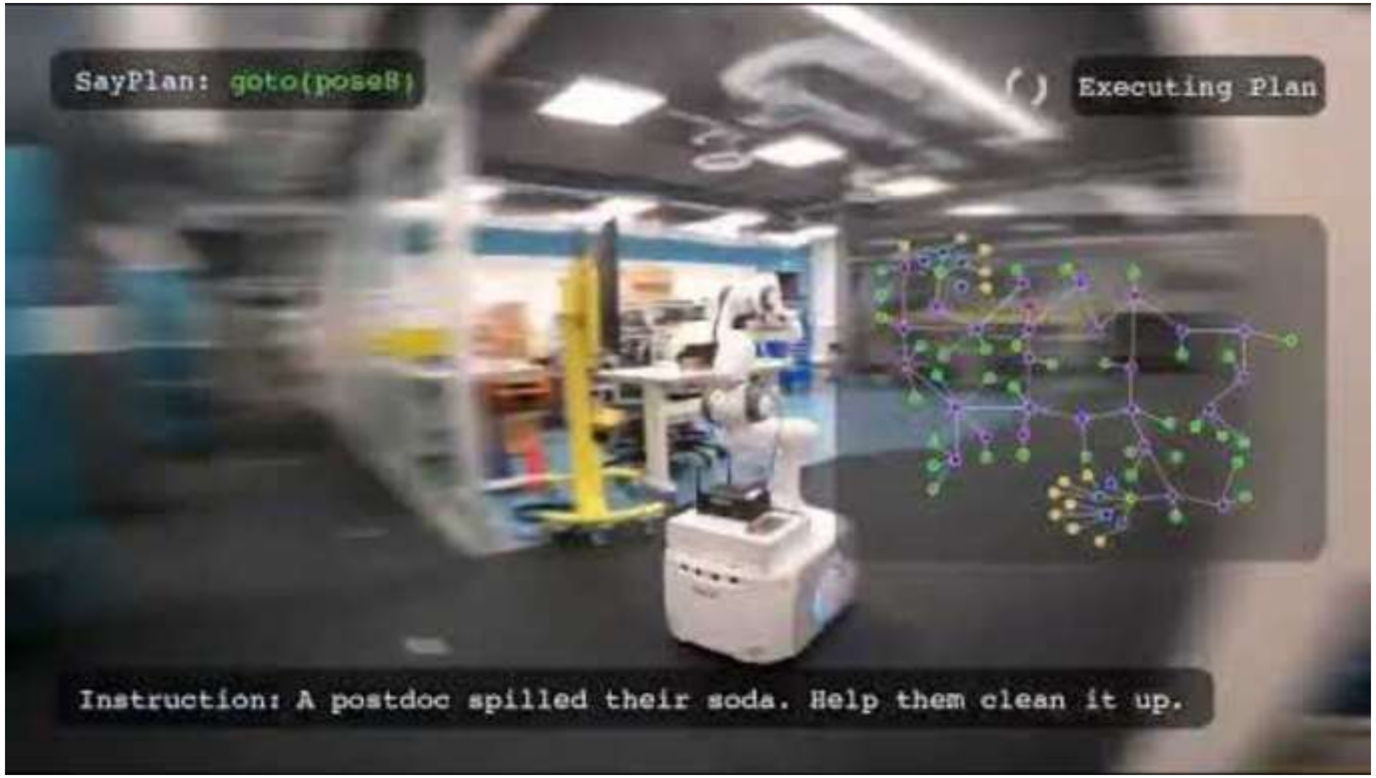
Aaron Ray, et al. "Task and Motion Planning in Hierarchical 3D Scene Graphs," *International Symposium of Robotics Research (ISRR)*, 2024

Scene graphs + embeddings (e.g., ConceptGraph)



Gu, Qiao, et al. "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning." *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

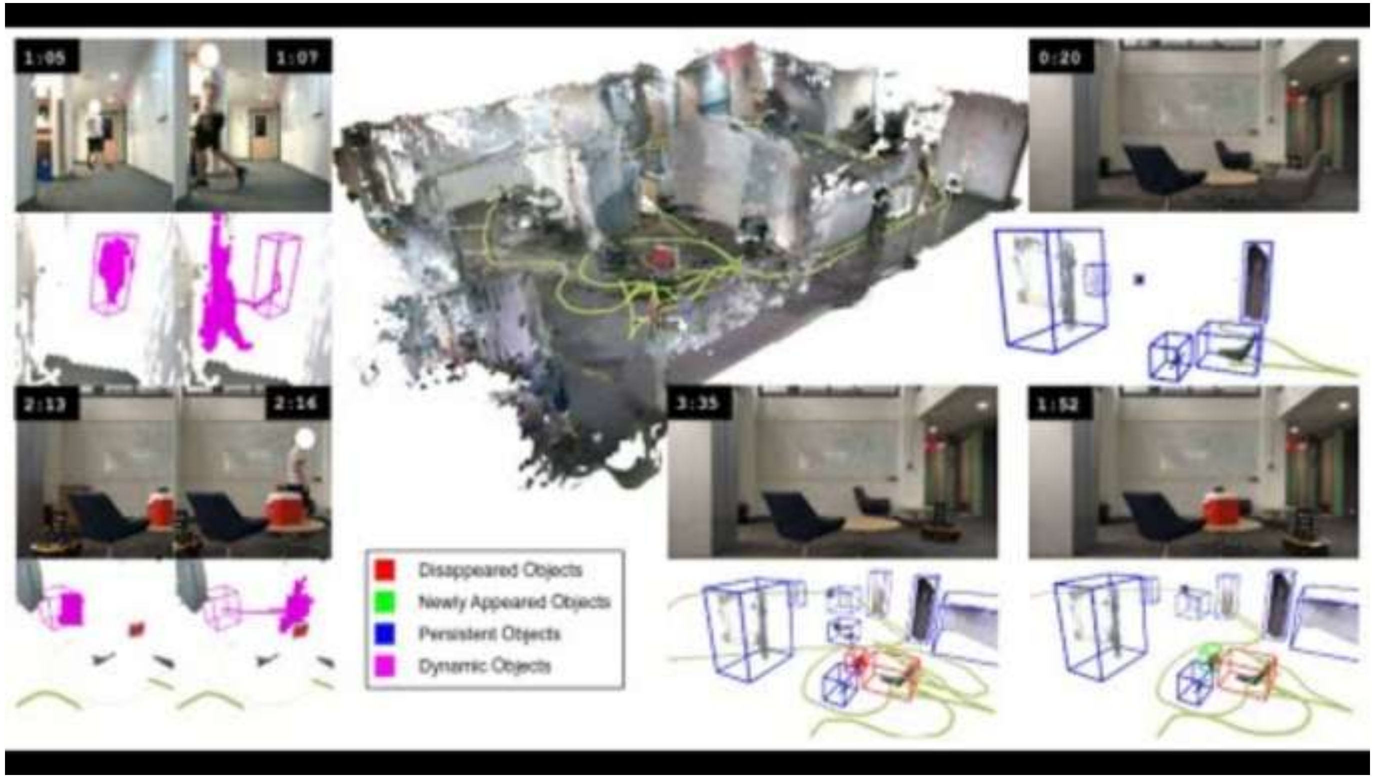
Using scene graphs in planning (e.g., SayPlan)



Rana, Krishan, et al. "SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning." *Conference on Robot Learning*. PMLR, 2023.

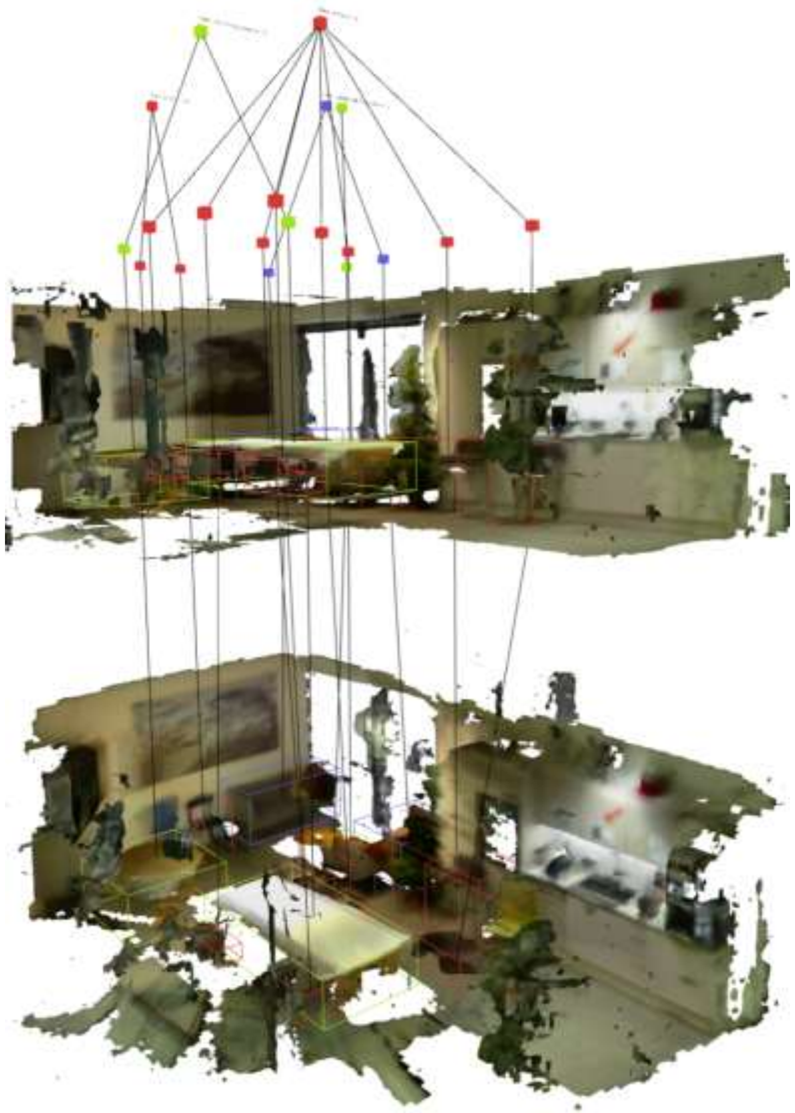
From 3D to 4D+

Dynamic Scene Graphs (e.g., Khronos)



Schmid, Lukas, et al. "Khronos: A Unified Approach for Spatio-Temporal Metric-Semantic SLAM in Dynamic Environments." *Robotics: Science and Systems*. 2024.

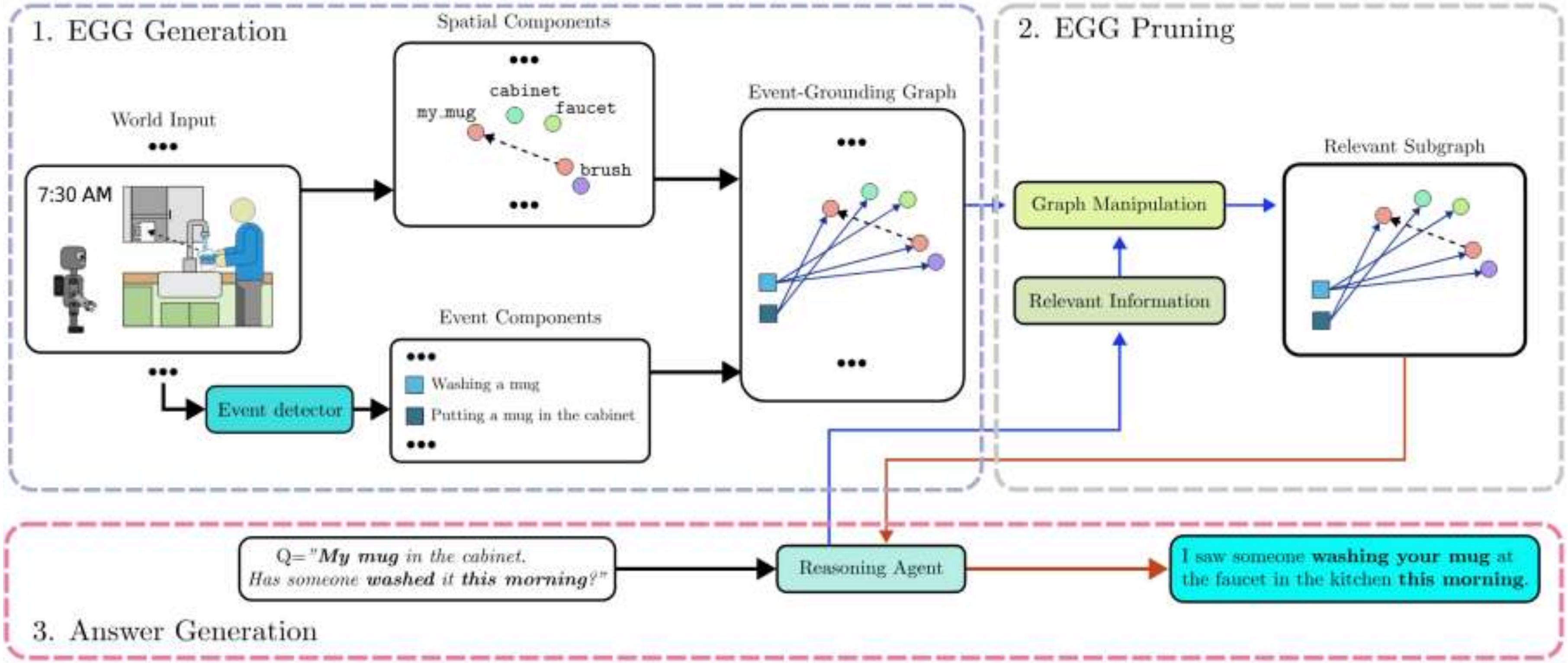
Updatable Scene Graphs (e.g., REACT)



Phuoc Nguyen, Francesco Verdoja, and Ville Kyrki
REACT: Real-time Efficient Attribute Clustering and Transfer for Updatable 3D Scene Graph
2025 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Oct 2025

A!

Event-Grounding Graphs (EGG)



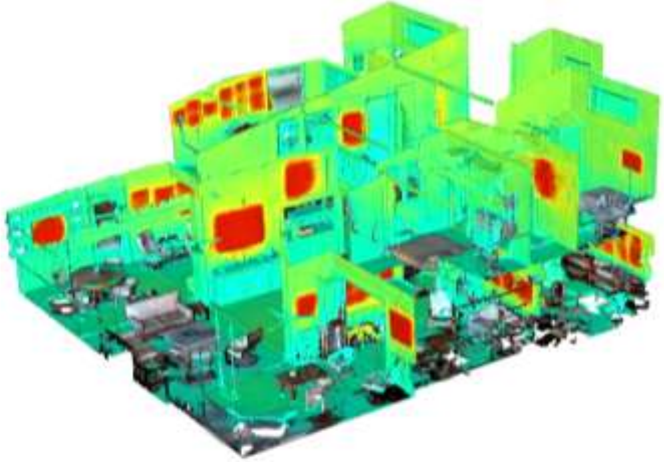
Nguyen, Phuoc, Francesco Verdoja, and Ville Kyrki. "Event-Grounding Graph: Unified Spatio-Temporal Scene Graph from Robotic Observations." *arXiv preprint arXiv:2510.18697* (2025), submitted to IEEE Robotics and Automation Letters (RA-L).

A!

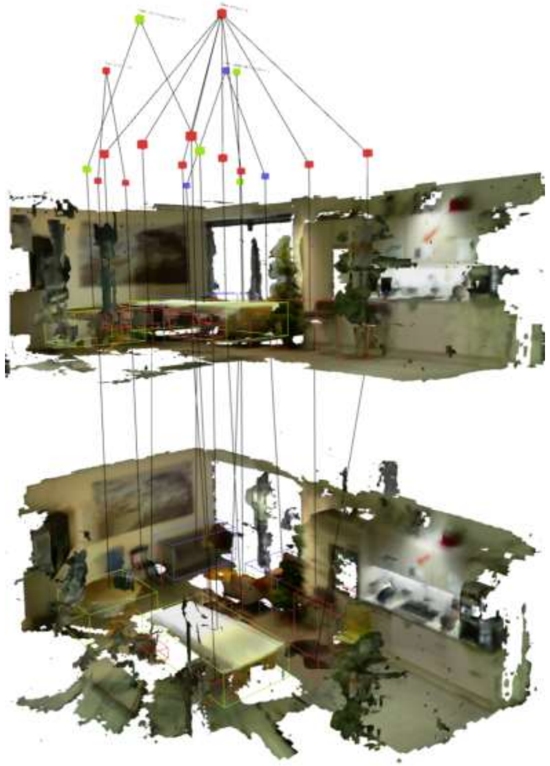
Takeaways

Semantic mapping is evolving rapidly

2020



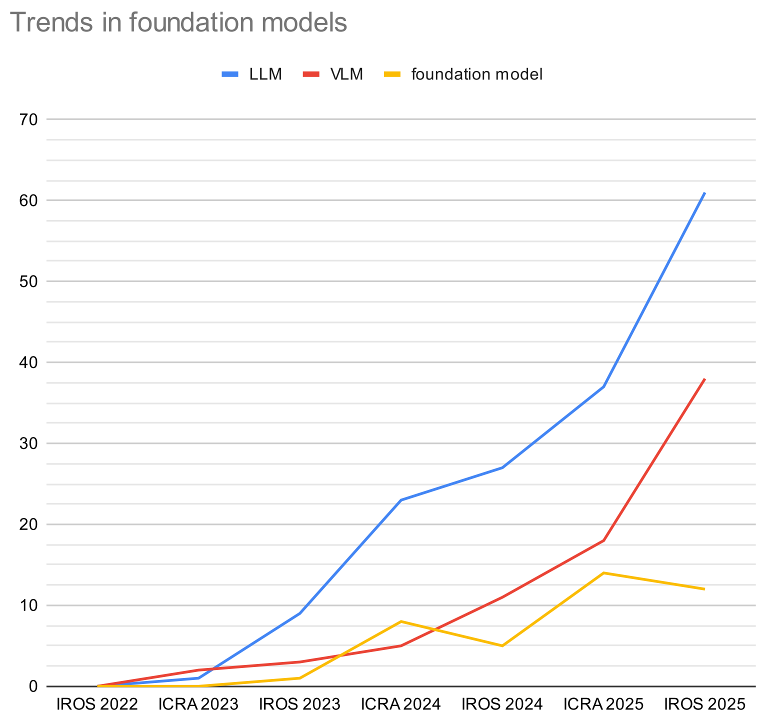
2023



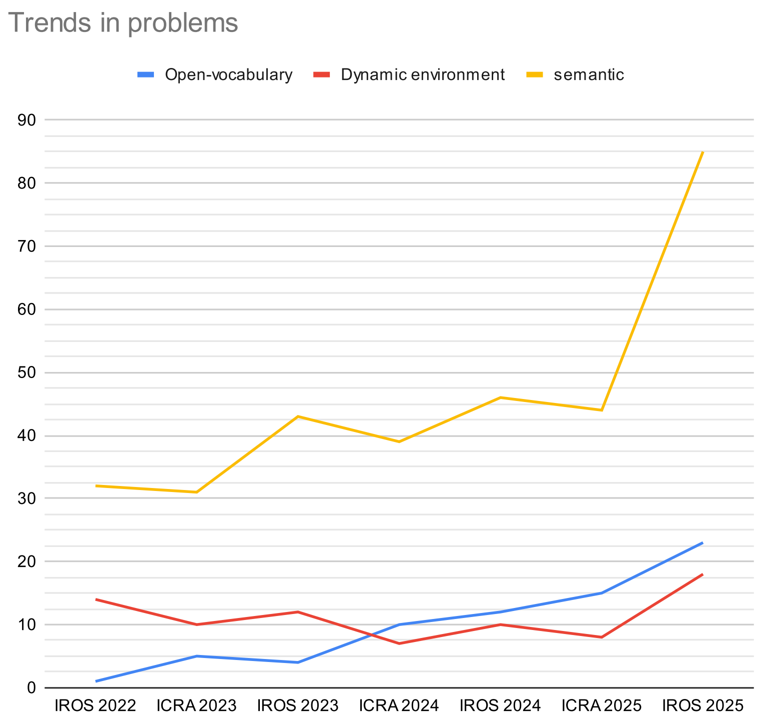
2025

A!

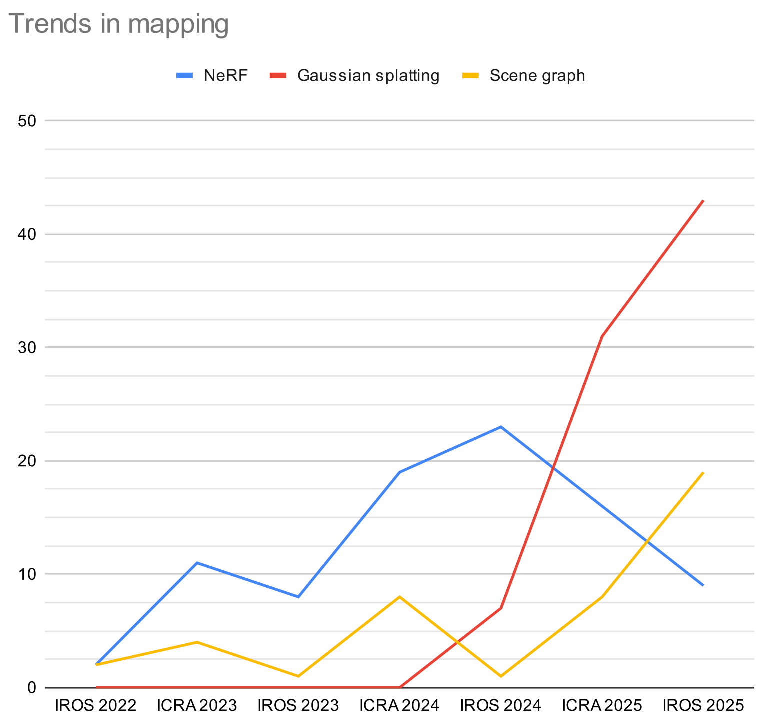
Trends



Trends in foundation models
LLM
VLM
foundation model
70
60
50
40
30
20
10
0
IROS 2022
ICRA 2023
IROS 2023
ICRA 2024
IROS 2024
ICRA 2025
IROS 2025



Trends in problems
Open-vocabulary
Dynamic environment
semantic
90
80
70
60
50
40
30
20
10
0
IROS 2022
ICRA 2023
IROS 2023
ICRA 2024
IROS 2024
ICRA 2025
IROS 2025



Trends in mapping
NeRF
Gaussian splatting
Scene graph
50
40
30
20
10
0
IROS 2022
ICRA 2023
IROS 2023
ICRA 2024
IROS 2024
ICRA 2025
IROS 2025

A!

Challenges and open problems

- **Lifelong mapping:** memory, forgetting, and map aging
- **Domain shift and generalization:** self-supervised and foundation models, VLMs, LLMs...
- **Multi-robot semantic mapping and map merging**
- **Task-specific maps:** sub-graph selection, planning domain generation
- **Dynamic scenes:** moving objects, time-dependent scene graphs

A!

Thank you

Francesco Verdoja
fverdoja.github.io

A!