Massachusetts Institute of Technology

3D Dynamic Scene Graphs

Actionable Spatial Perception with Places, Objects, and Humans

Antoni Rosinol*, Arjun Gupta, Marcus Abate, Jingnan Shi, Luca Carlone

*arosinol@mit.edu

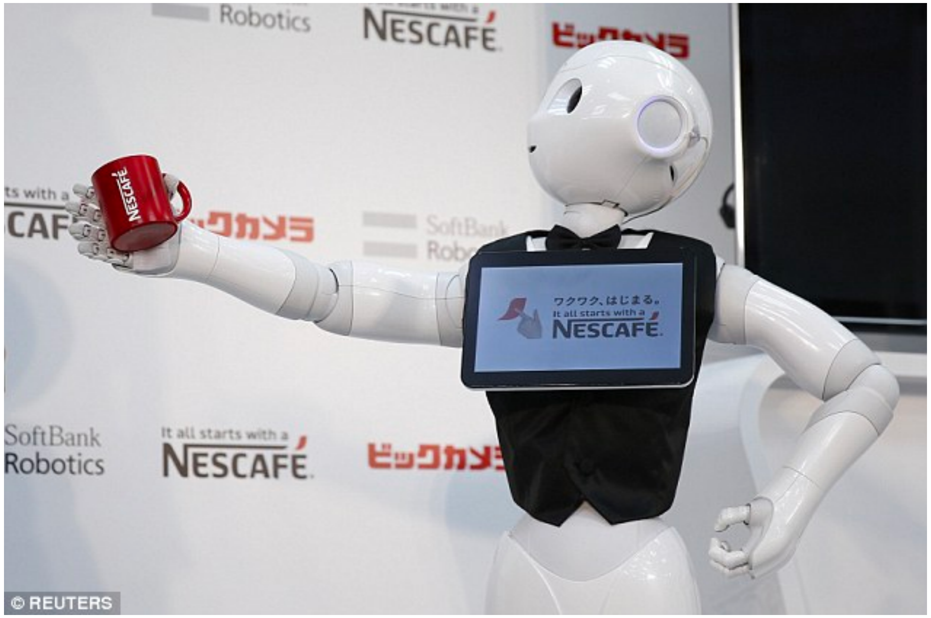
LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

dcist

"la Caixa"

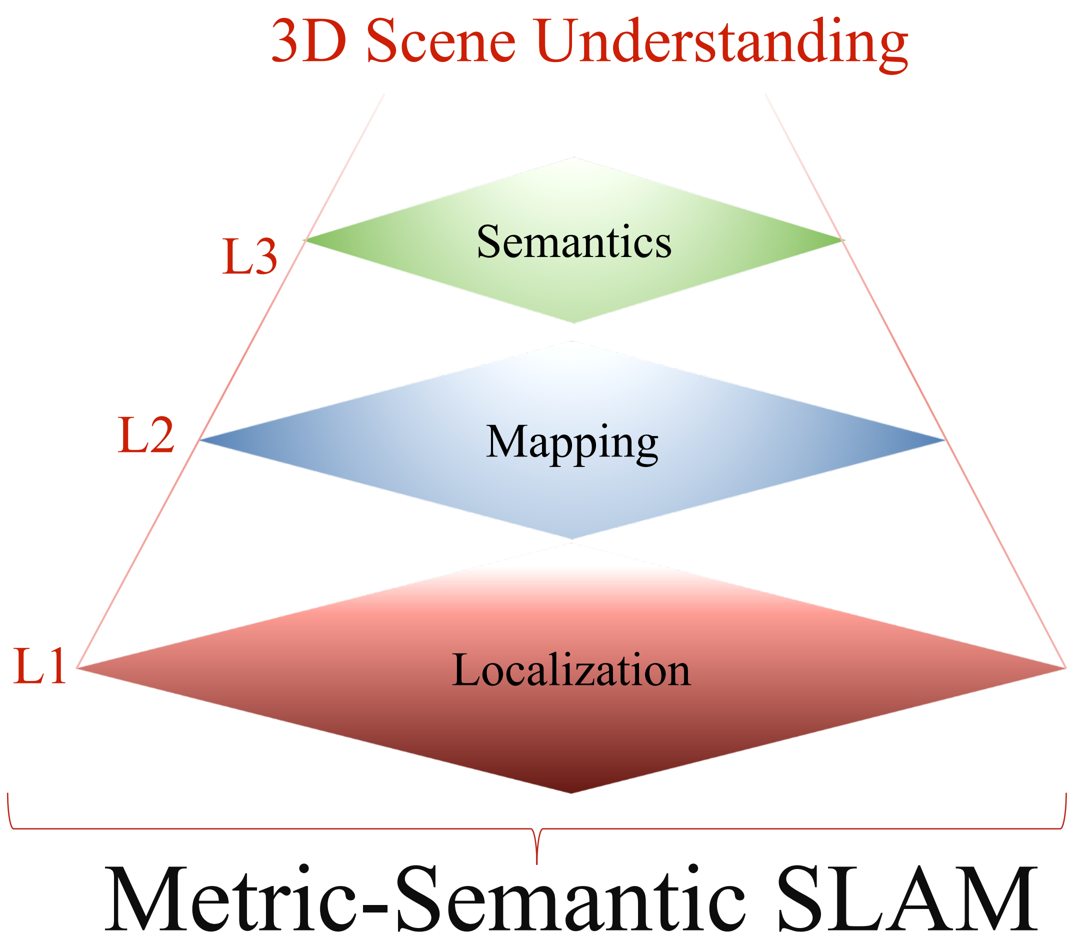
Motivation

Fully autonomous systems should operate given high-level tasks and figure out the necessary low-level tasks.



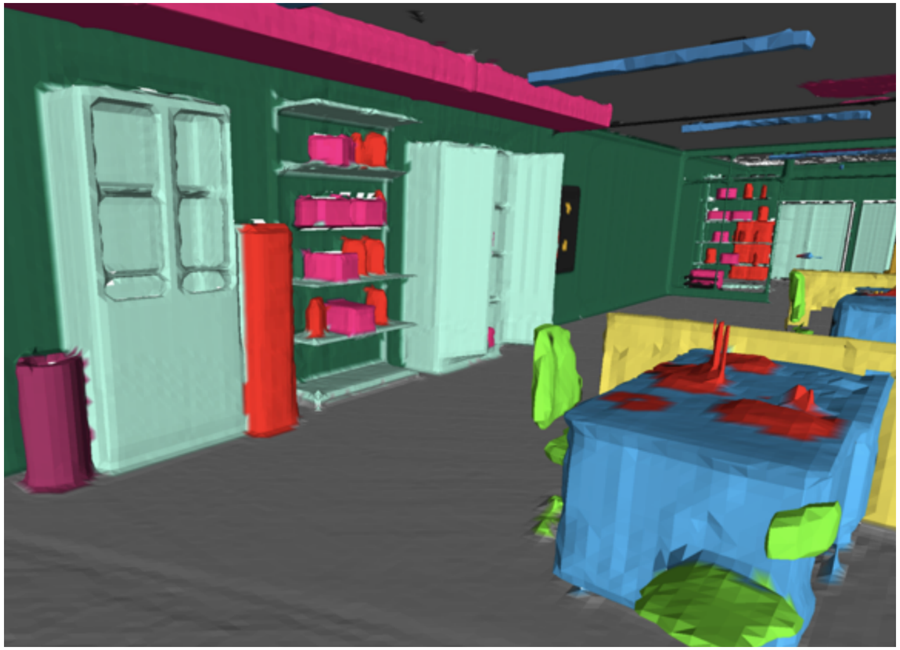
Bottleneck: 3D Scene Understanding

What does a robot need to accomplish high-level tasks?

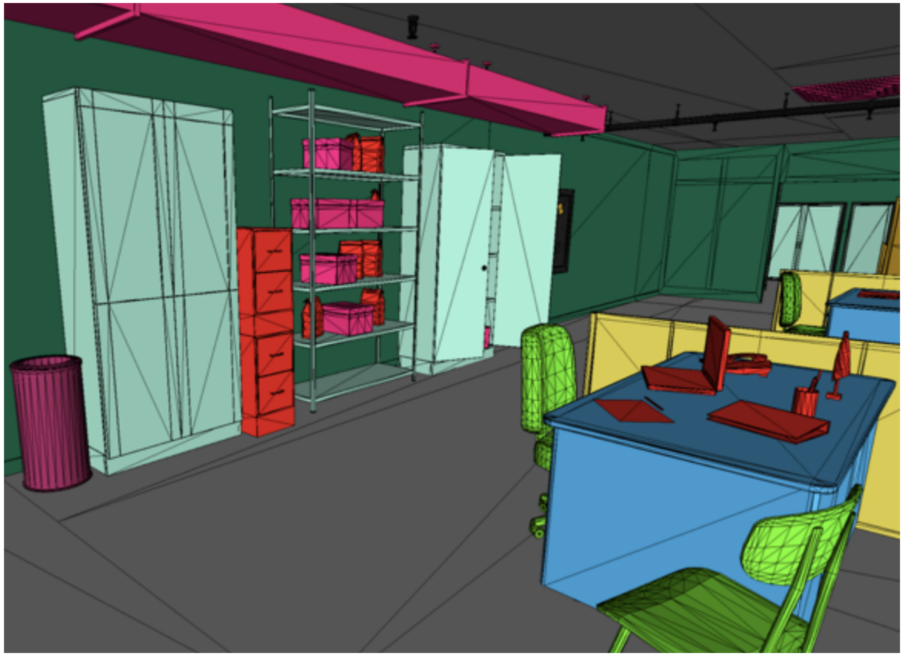


Kimera: Real-Time Metric-Semantic SLAM [1]

- Accurate and Robust State Estimation: state-of-the-art VIO
- Faithfull metric-semantic reconstruction
- Real-Time 100ms per frame (CPU-only)



Estimated

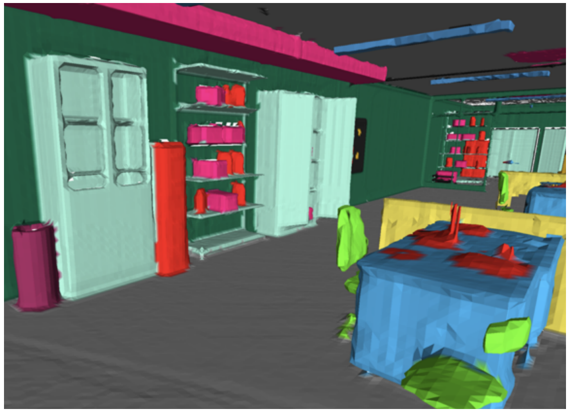


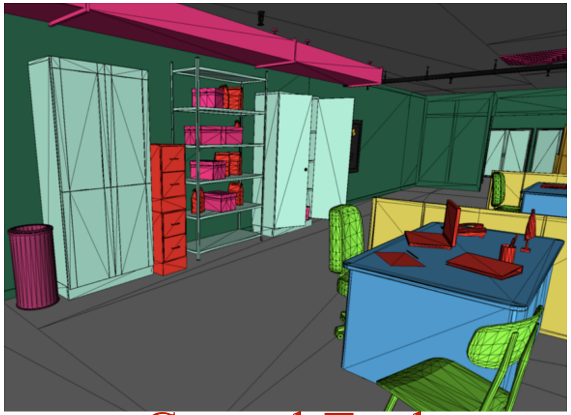
Ground-Truth

[1] Rosinol, Antoni and Abate, Marcus and Chang, Yun and Carlone, Luca.
"Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping", ICRA 2020

Problem

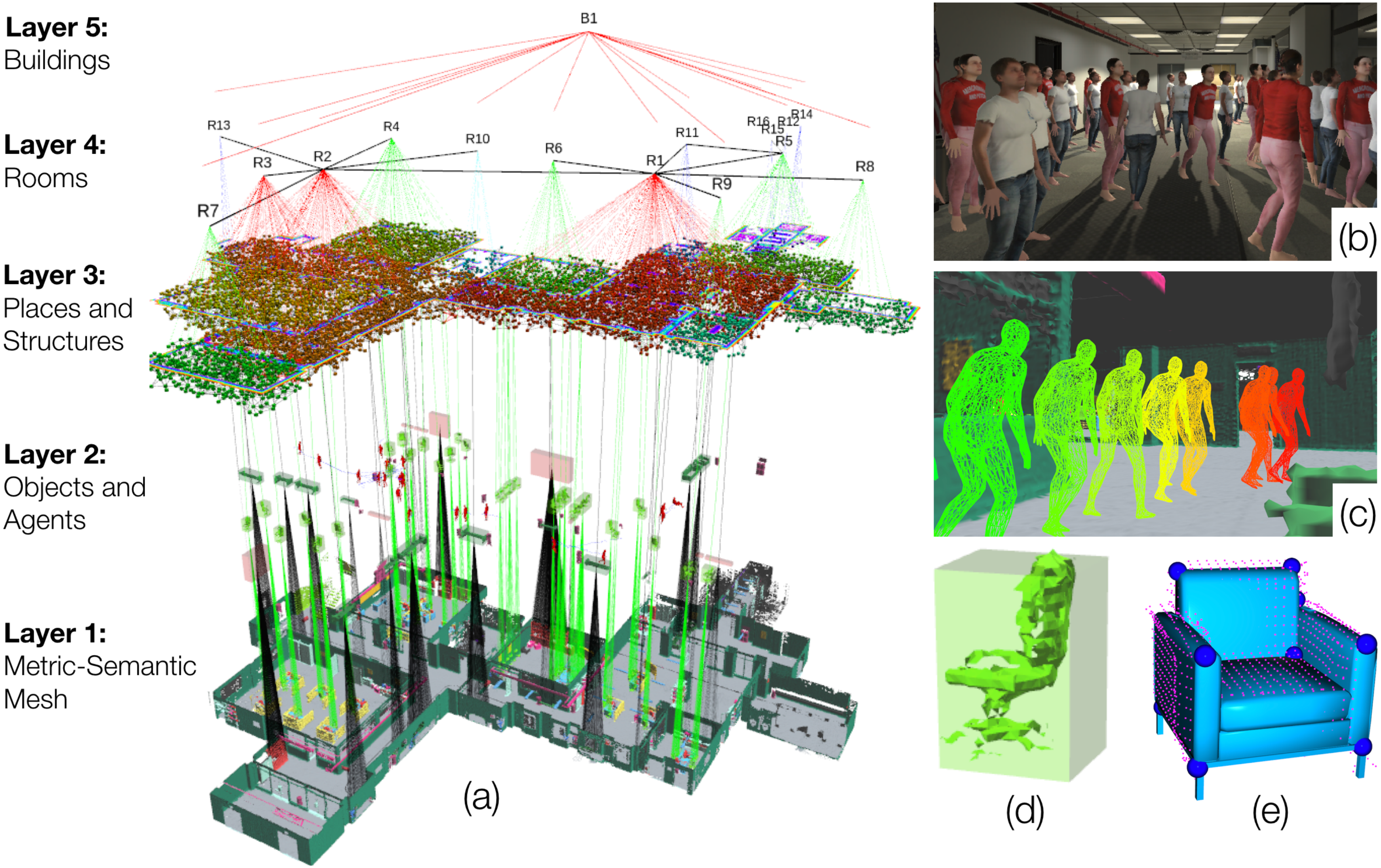
- Raw 3D semantic mesh is **not** actionable:
  - Obstacle Avoidance and Planning:
    - Not readily usable for path planning: `go to the kitchen`
  - Human-Robot Interaction:
    - 3D model readable for both humans and robots
    - Difficult to answer queries: `how many chairs are there?`
  - Long-term Autonomy:
    - Compact representation
    - Different levels of Abstractions
    - Forget/retain relevant information


Estimated


Ground-Truth

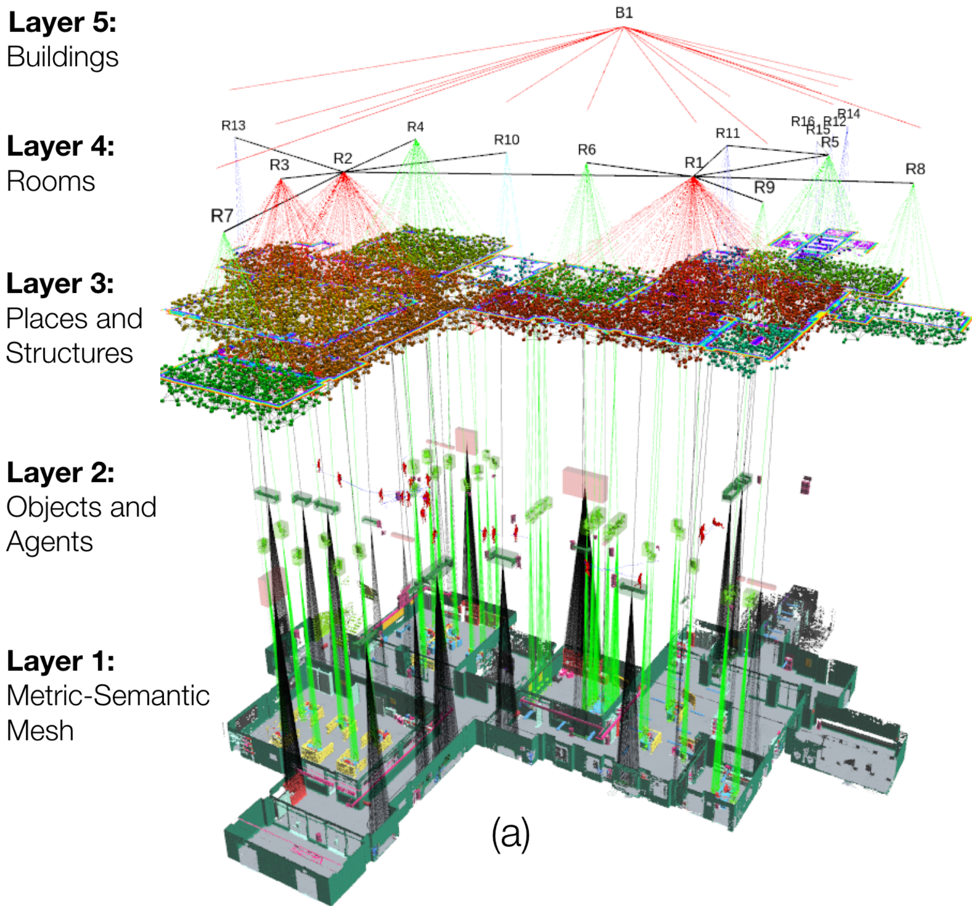
3D Dynamic Scene-Graphs

3D Dynamic Scene-Graphs (DSGs)



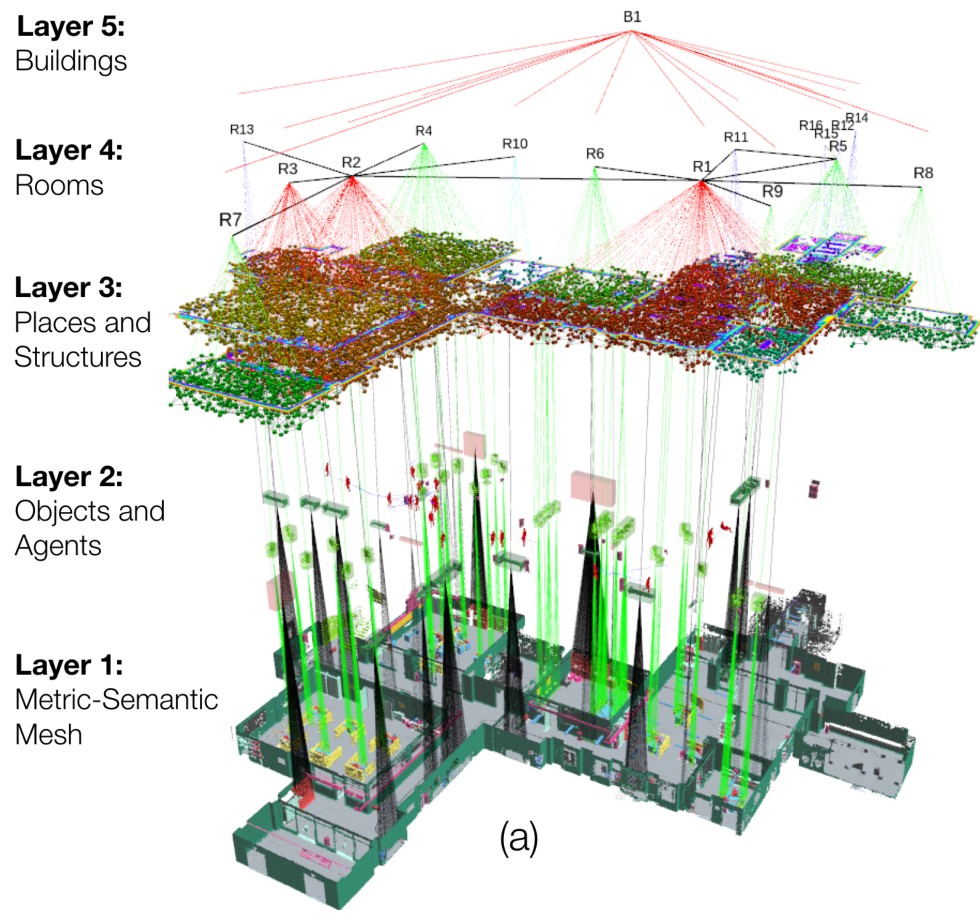
Layers

- Layer 1: Metric-Semantic 3D Mesh
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings



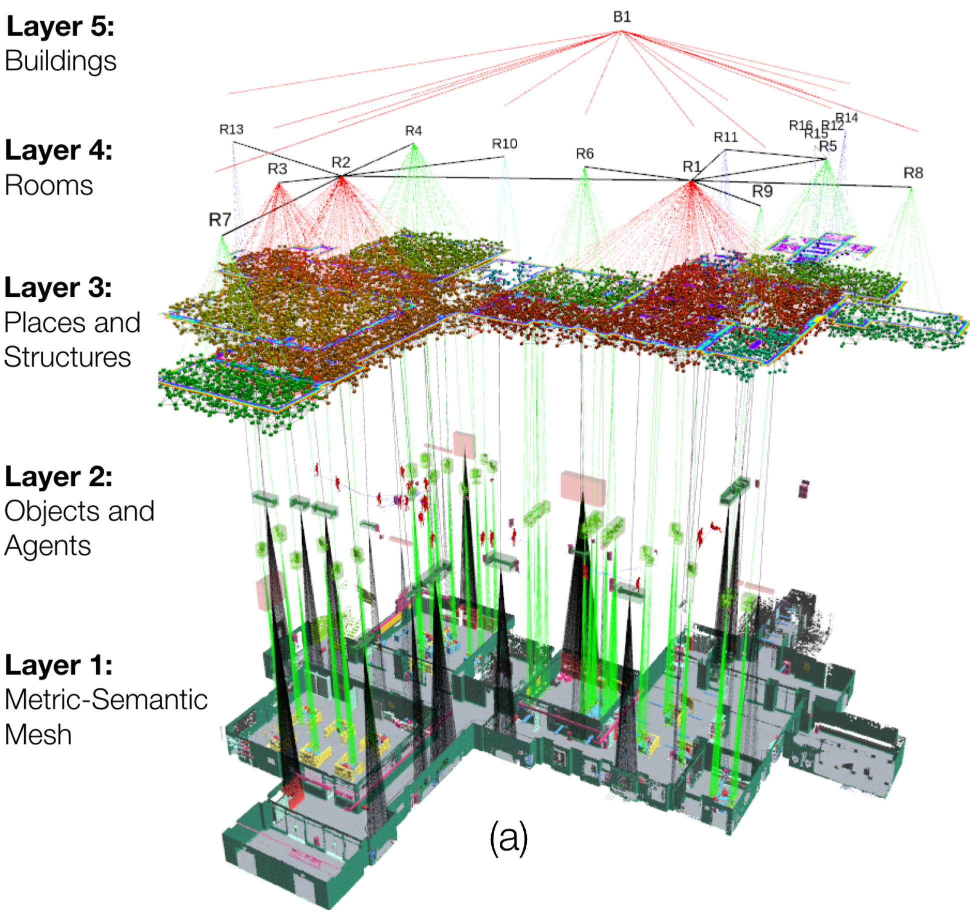
Layers

- Layer 1: Metric-Semantic 3D Mesh (Kimera)
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings



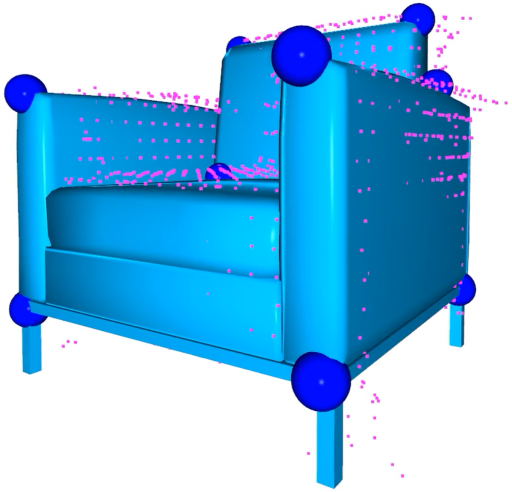
Layers

- Layer 1: Metric-Semantic 3D Mesh
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings



Layer 2: Objects and Agents

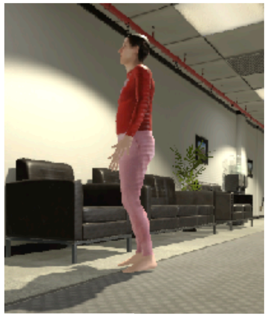
- Object Attributes:
  - 3D Centroid, bounding box, semantic label, and instance id.
- Object instance extraction:
  1. Extract portions of the mesh with a semantic label.
  2. Clustering to extract instances (assumes 3D objects' instances are not touching!)
  3. Calculate centroid and bounding-box.
- We distinguish between:
  - Known objects: for which we have a CAD model, and
  - Unknown objects: no prior 3D model
- Known object instance fitting:
  1. Extract 3D keypoints (spheres in blue)
  2. Match all 3D keypoints from estimate and CAD model (=> outliers)
  3. Use TEASER++[1] to remove outliers and fit CAD model.



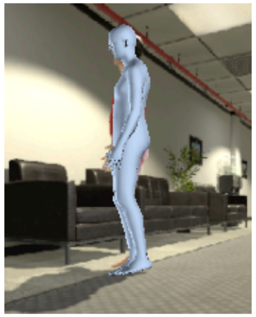
[1] Yang, Heng and Shi, Jingnan and Carlone, Luca. Teaser: Fast and certifiable point cloud registration. https://arxiv.org/abs/2001.07715

Layer 2: Objects and Agents

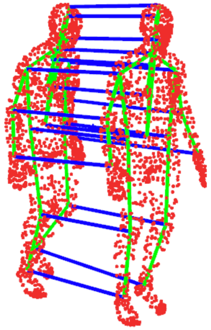
- Agents: dynamic entities in the environment: vehicles, humans, robots...
  - We model Agents by:
    i. 3D Pose Graph*: describing their trajectory over time
    ii. 3D Mesh Model: describing their (non-rigid) shape
    iii. Semantic class: human, robot, ...
- Human Agents:
  1. Detection:
     1. Extract bounding box of image from semantic segmentation
     2. Estimate 3D mesh model (SMPL) of human using [1].
  2. Tracking:
     1. Incrementally build pose-graph with motion model
     2. Remove outliers and/or incorrect data associations by enforcing joint consistency (blue segments in (c))



(a) Image



(b) Detection



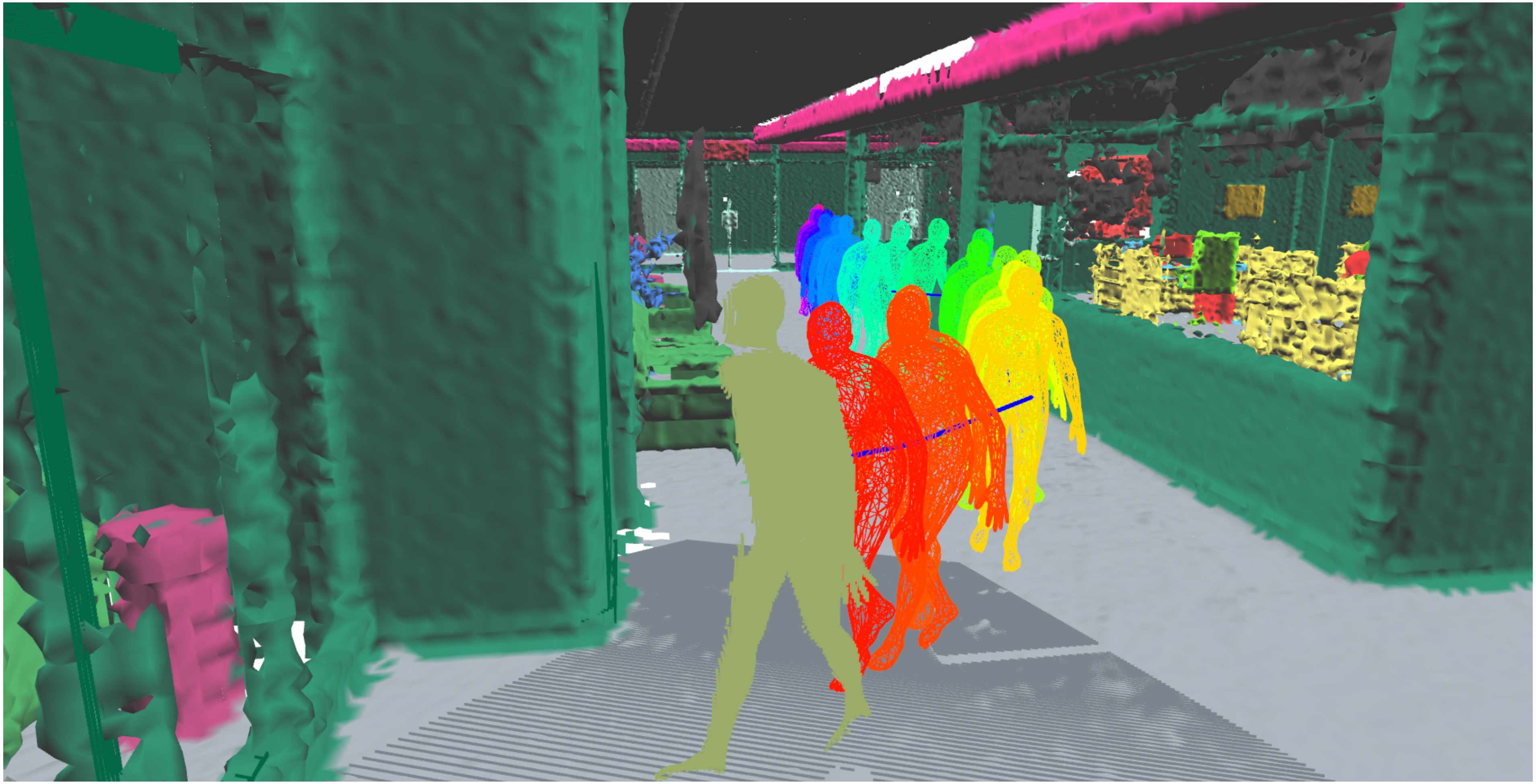
(C) Tracking

* A pose graph is a collection of time-stamped 3D poses where edges model pairwise relative measurements

[1] Kolotouros, Nikos and Pavlakos, Georgios and Daniilidis, Kostas . Convolutional mesh regression for single-image human shape reconstruction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

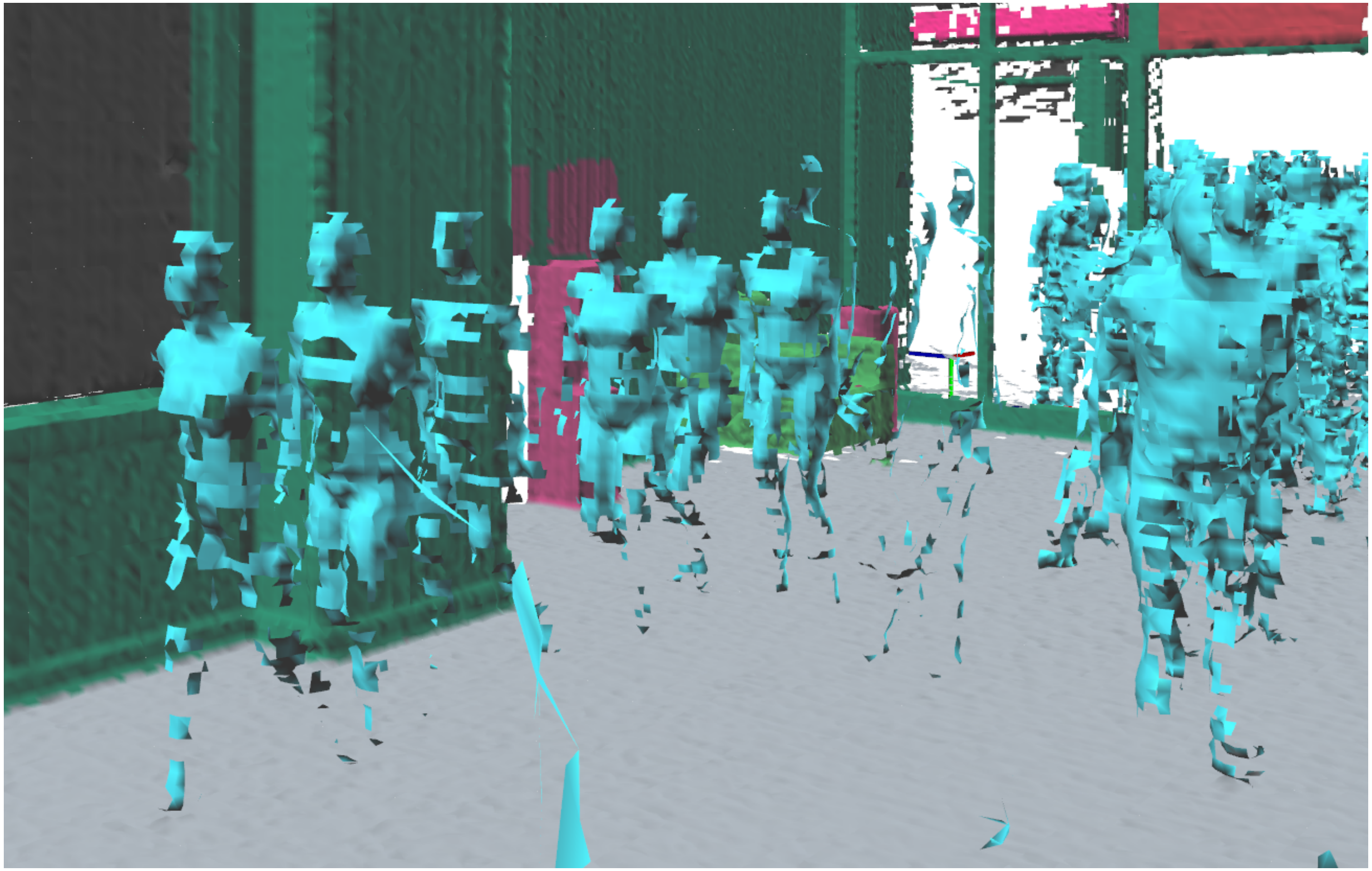
Layer 2: Objects and Agents

- Human Agent Tracking:
  - Blue trajectory: corresponds to the built pose-graph
  - Rainbow human mesh: associated detections with pose-graph vertices.



Layer 2: Objects and Agents

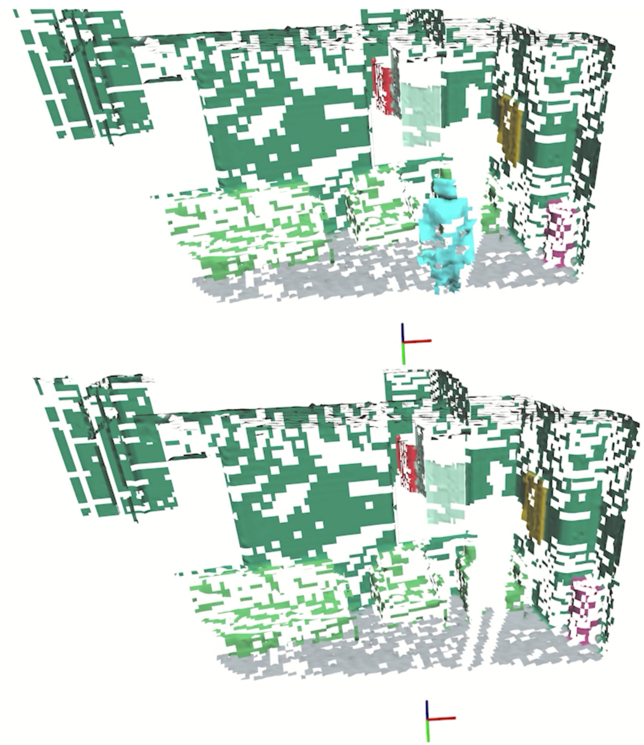
- Dynamic Masking:
  - Non-static agents can corrupt 3D reconstruction: we avoid integrating dynamic agents in 3D metric-semantic mesh.



Layer 2: Objects and Agents

- Localization: KLT-IMU + 2-point RANSAC
- Mapping: Dynamic Masking
  - Avoid integrating dynamic agents in 3D metric-semantic mesh.

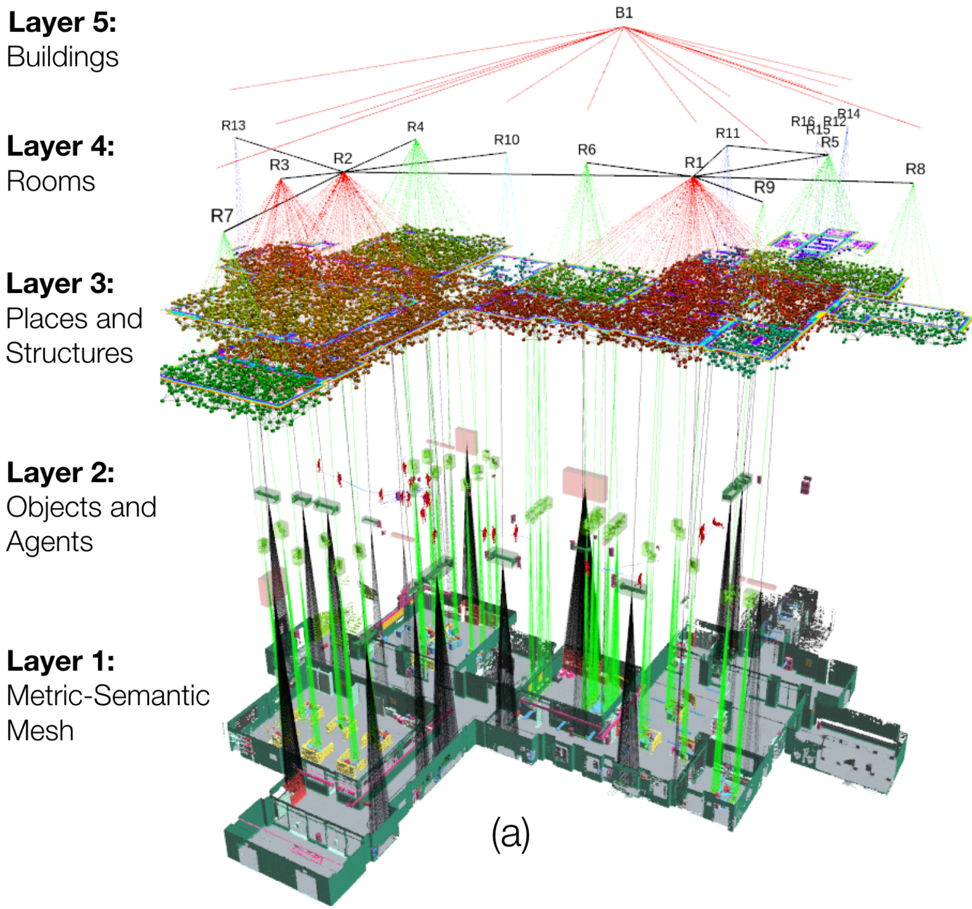
RGB Frame



We extend Kimera to mask dynamic objects in the mesh and use IMU-aware feature tracking, increasing robustness in crowded scenes

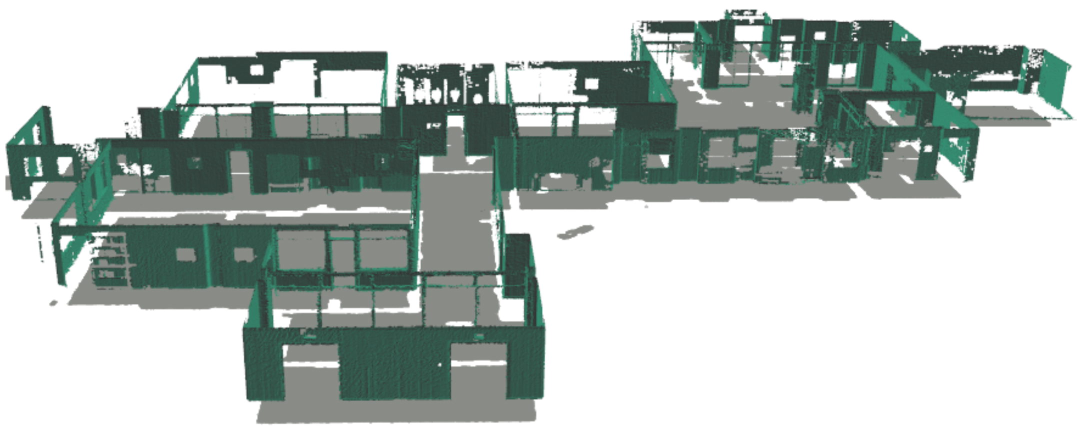
Layers

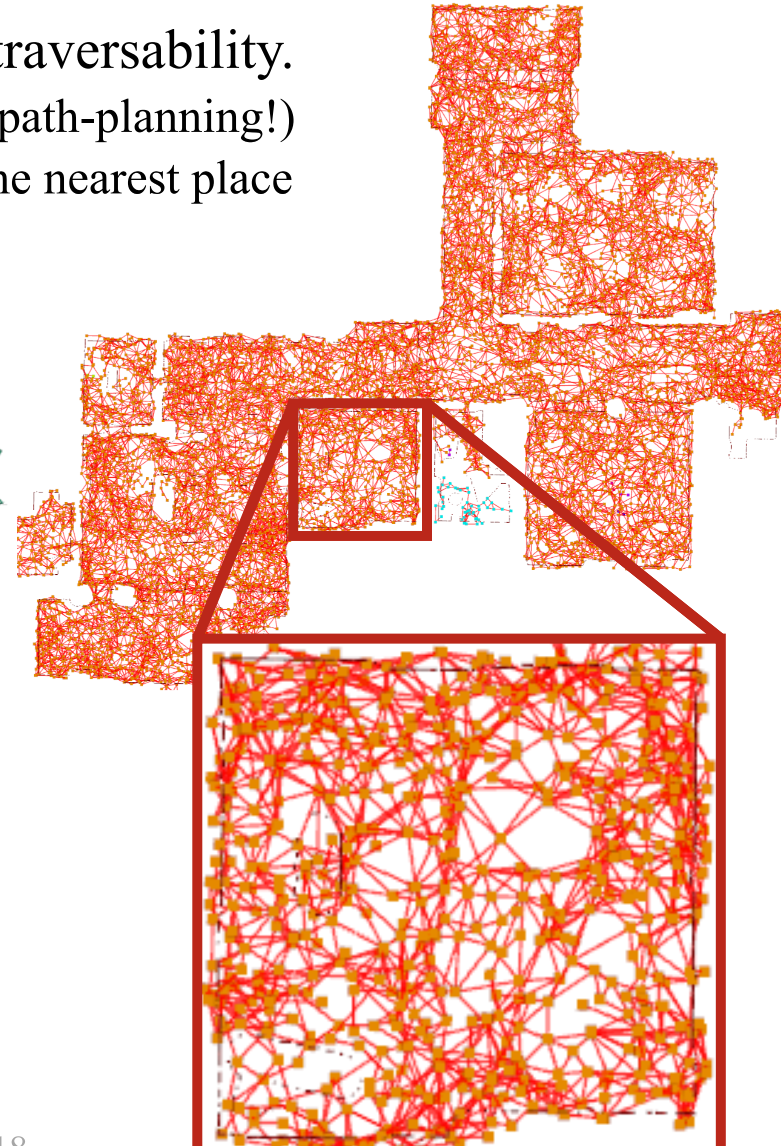
- Layer 1: Metric-Semantic 3D Mesh
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings



Layer 3: Places and Structures

- Places: free-space locations, edges represent traversability.
  - Modelled as a topological map (readily usable for path-planning!)
  - Each object and agent in Layer 2 is connected to the nearest place
- Structures:
  - Walls, floor, ceiling, pillars…

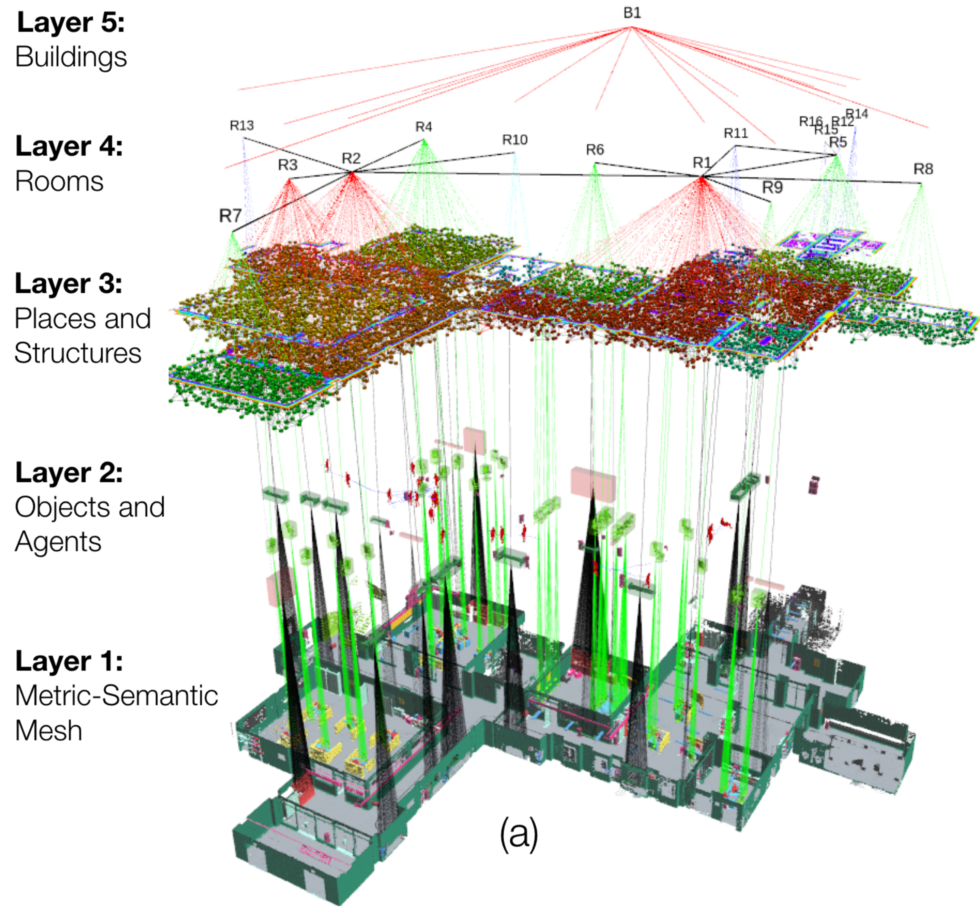




[1] H Oleynikova, Z Taylor, R Siegwart, J Nieto.
Sparse 3d topological graphs for micro-aerial vehicle planning, IROS 2018.

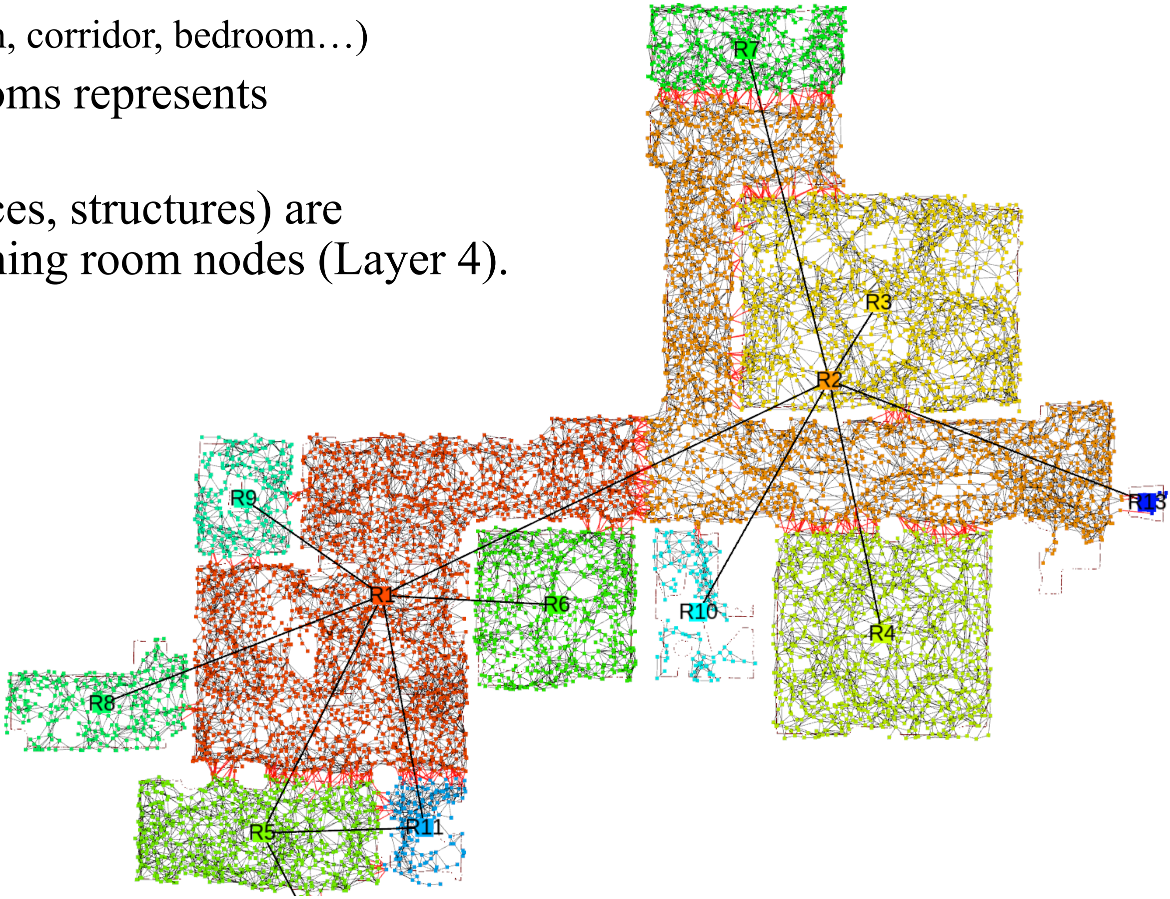
Layers

- Layer 1: Metric-Semantic 3D Mesh
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings



Layer 4: Rooms

- Rooms: as well as corridors, halls …
  - Attributes:
    i. 3D pose
    ii. Bounding box
    iii. Semantic class (kitchen, corridor, bedroom…)
  - Connectivity between rooms represents traversability
  - Elements in Layer 3 (places, structures) are connected to their containing room nodes (Layer 4).



Layer 4: Rooms

- Rooms detection:
  1. A 2D slice of the 3D ESDF (Euclidean Signed Distance Function) below the detected ceiling is constant almost everywhere except near walls. Fig. (a).
  2. Truncate 2D ESDF to obtain disconnected sections corresponding to rooms. Fig. (b).
  3. Label nodes that fall inside a disconnected ESDF section with one room label (this only labels a subset of all nodes)
  4. Using topology of the Places graph, infer the rest of room labels using majority voting.

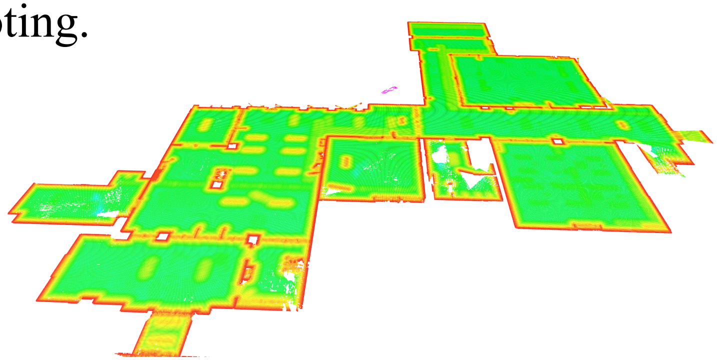


Fig. (a) 2D slice of 3D ESDF

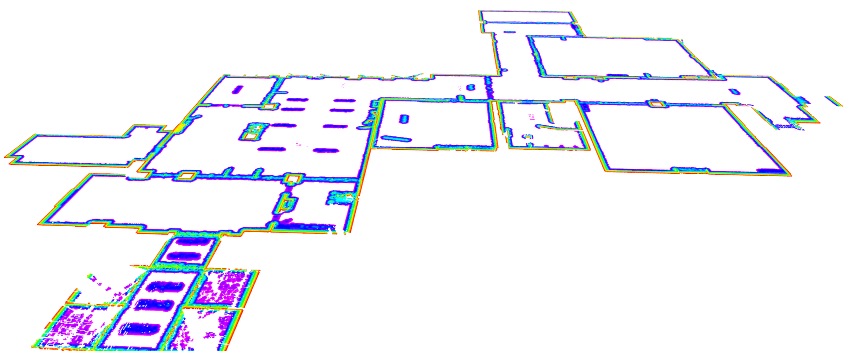
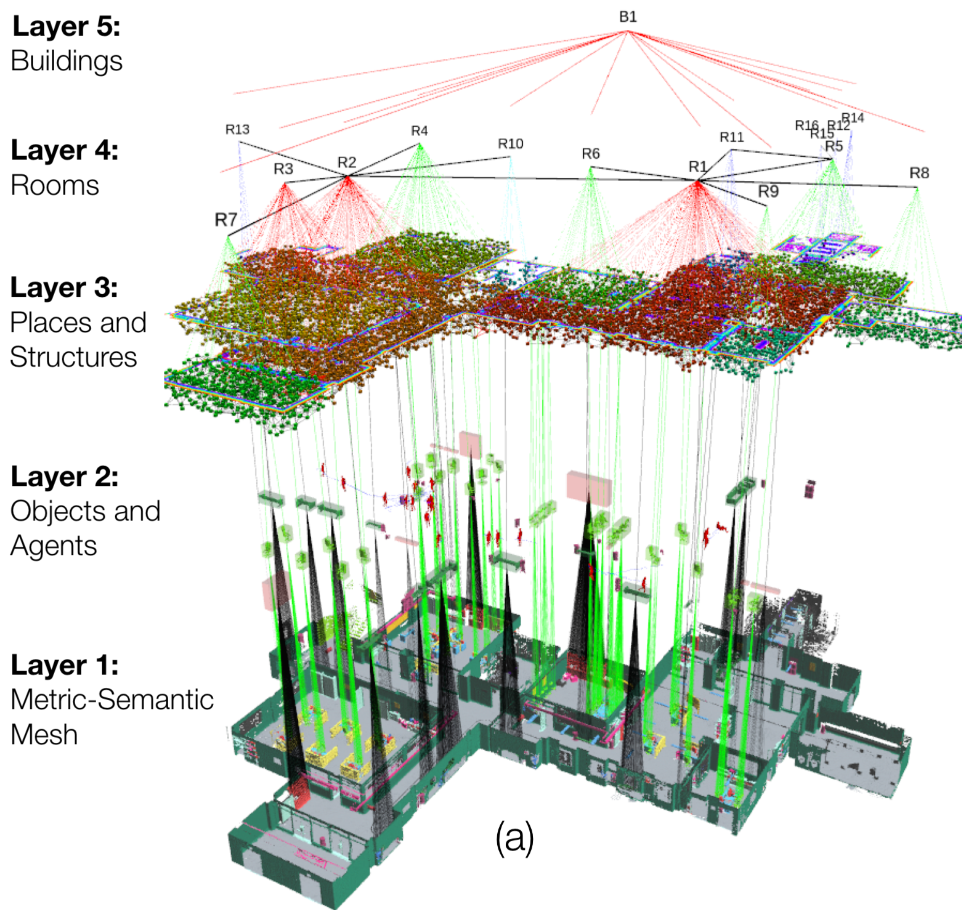


Fig. (b) Truncated 2D ESDF

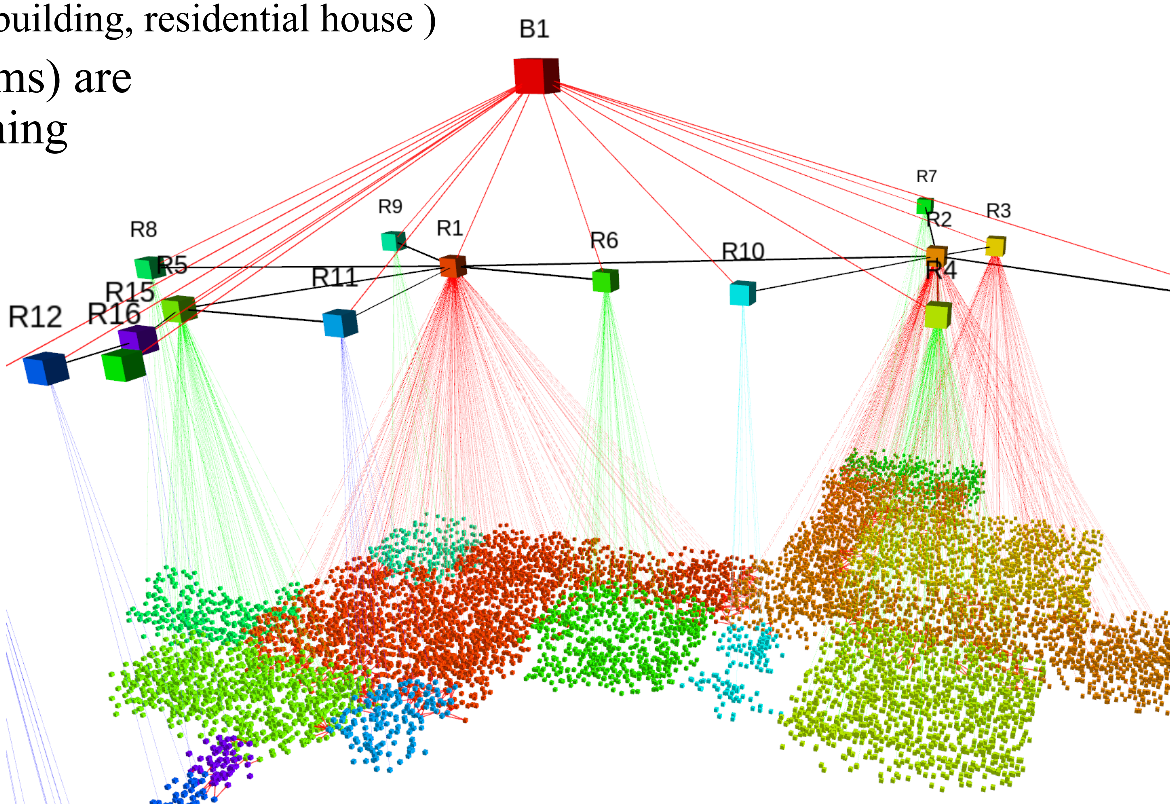
Layers

- Layer 1: Metric-Semantic 3D Mesh
- Layer 2: Objects and Agents
- Layer 3: Places and Structures
- Layer 4: Rooms
- Layer 5: Buildings

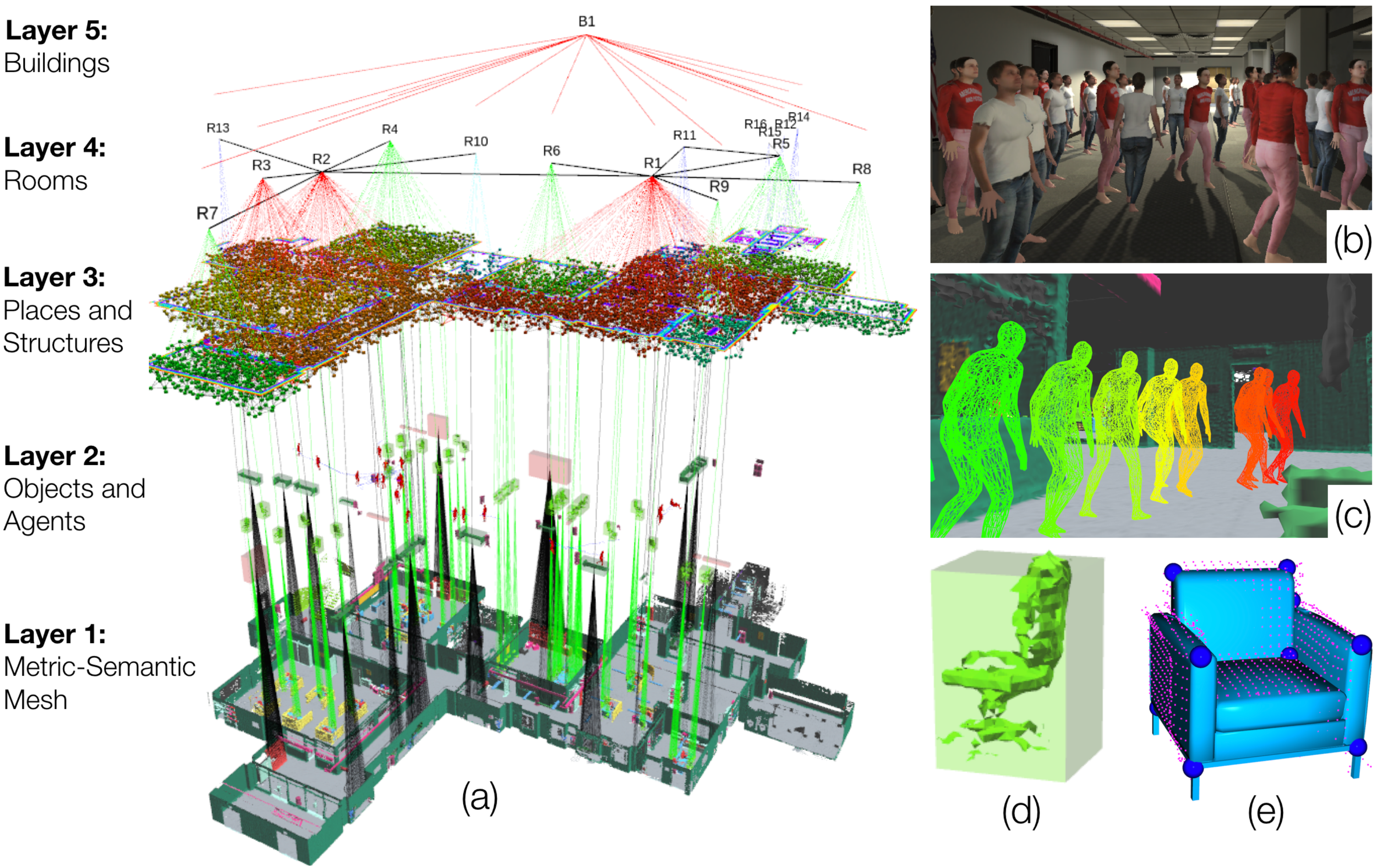


Layer 5: Buildings

- Buildings
  - Attributes:
    i. 3D pose
    ii. Bounding box
    iii. Semantic class (office building, residential house)
  - Elements in Layer 4 (rooms) are connected to their containing building (Layer 5).



3D Dynamic Scene-Graphs



Thank you!