From VPS to SNAP

Beyond visual positioning: how localization turned out to provide efficient training of neural, semantic maps

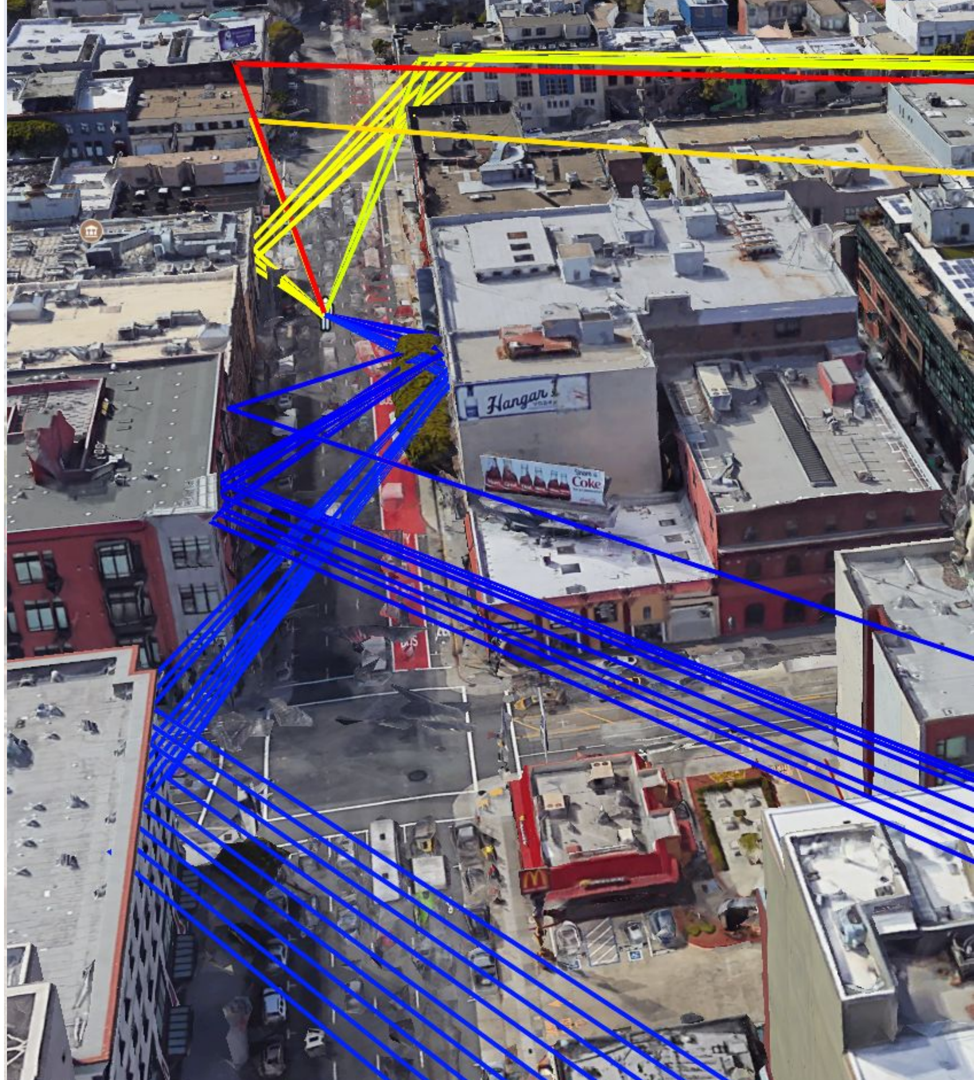
Eduard Trulls / Research Scientist at Google Zurich
ECCV'24 / Map-free Visual Relocalization Workshop

Google's **Visual Positioning Service** (VPS)

An image-based localization service available wherever we have StreetView

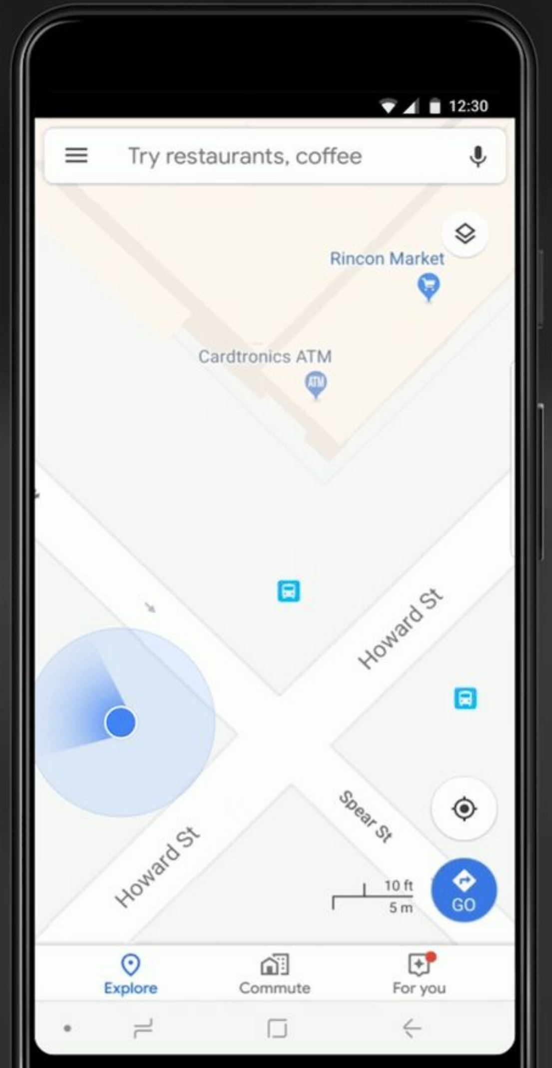
Outdoor localization

GPS suffers from reflections (multi-path). Compass is impacted by magnetic objects.



Improving the 'Blue Dot'

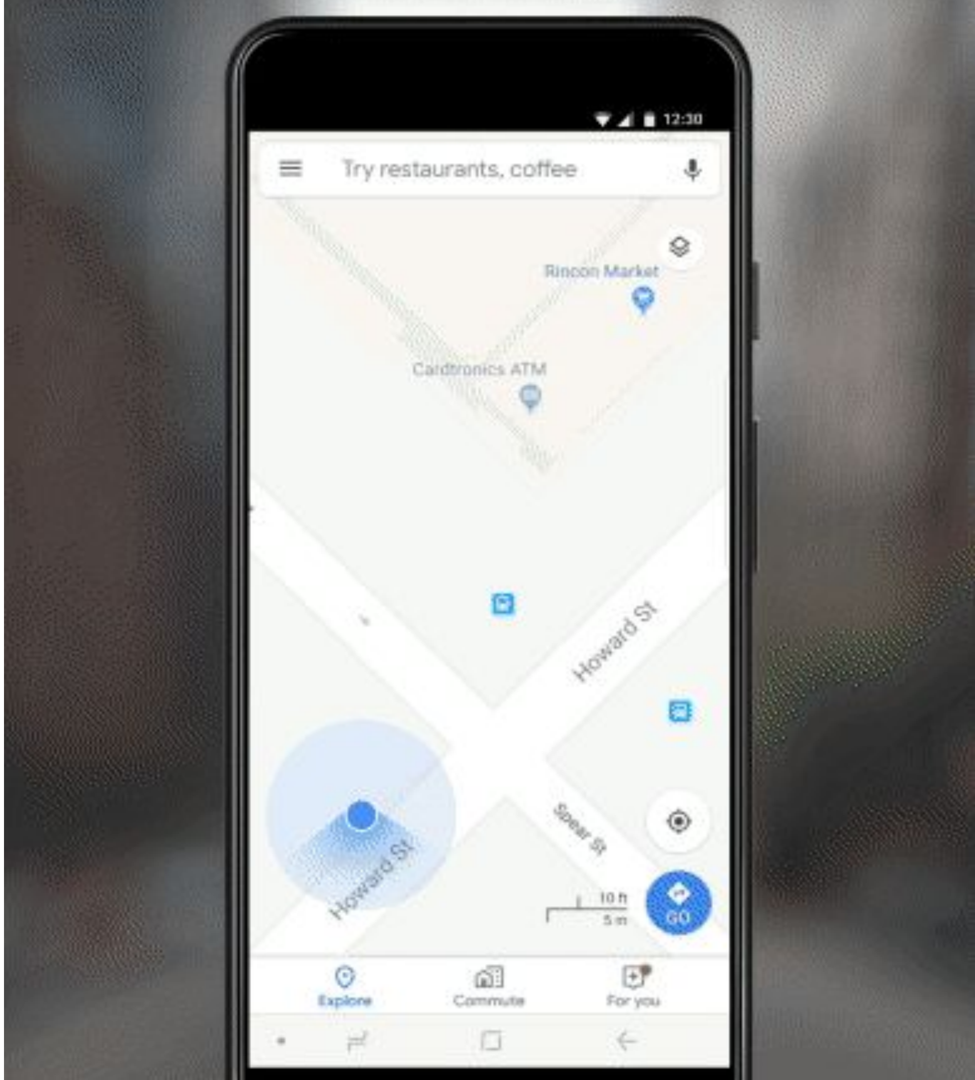
Image-based localization enables precise location and orientation



VPS enables large-scale AR

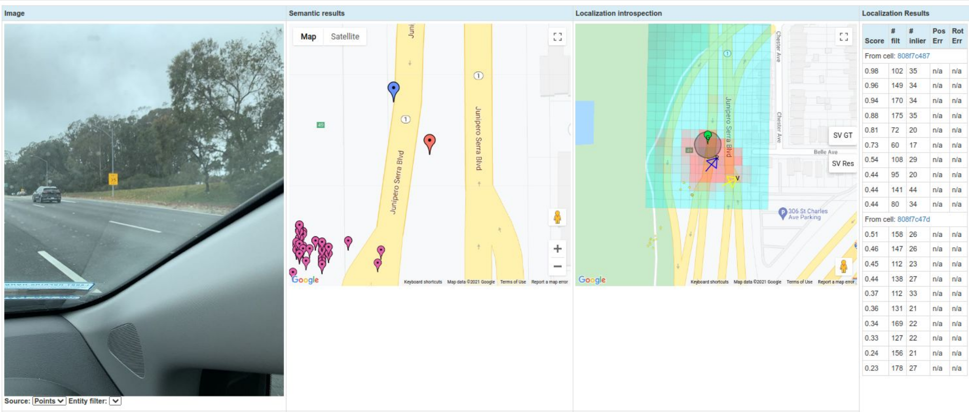
Sub-meter position and sub-deg orientation accuracy has drastic effect on AR use-cases

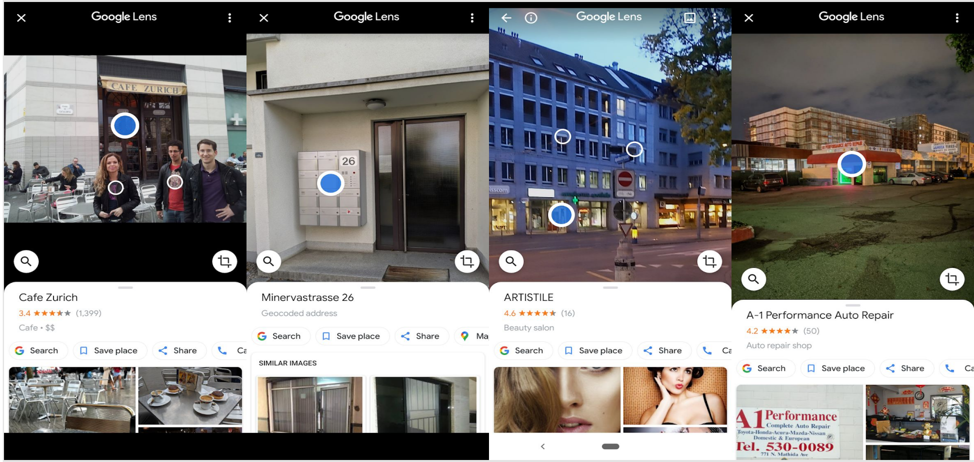
Try out yourself! See ***LiveView* walking navigation** in Google Maps



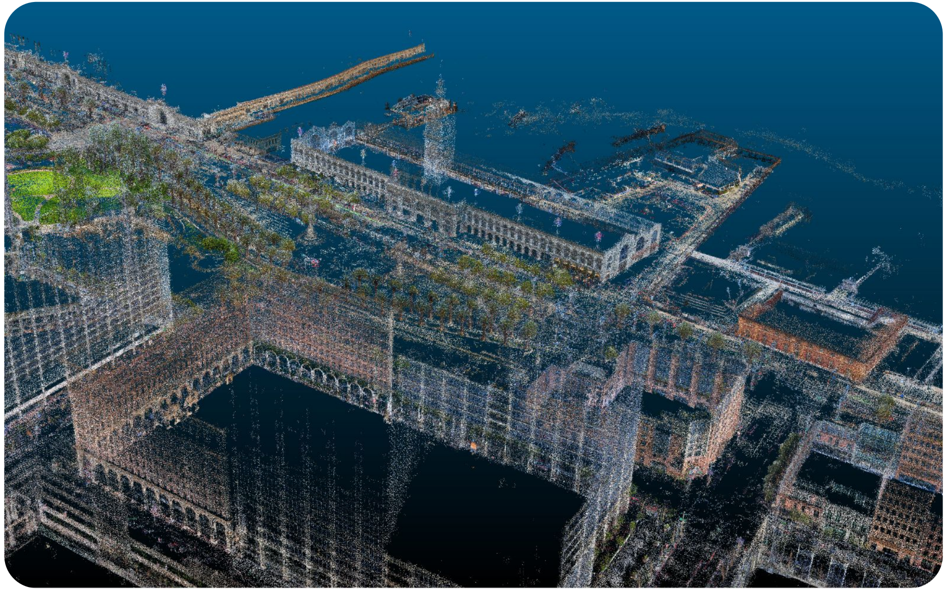
But it has many other use-cases!

VPS is also used to localize images from dashcams, monitor infrastructure, Google Lens, and user-contributed photos

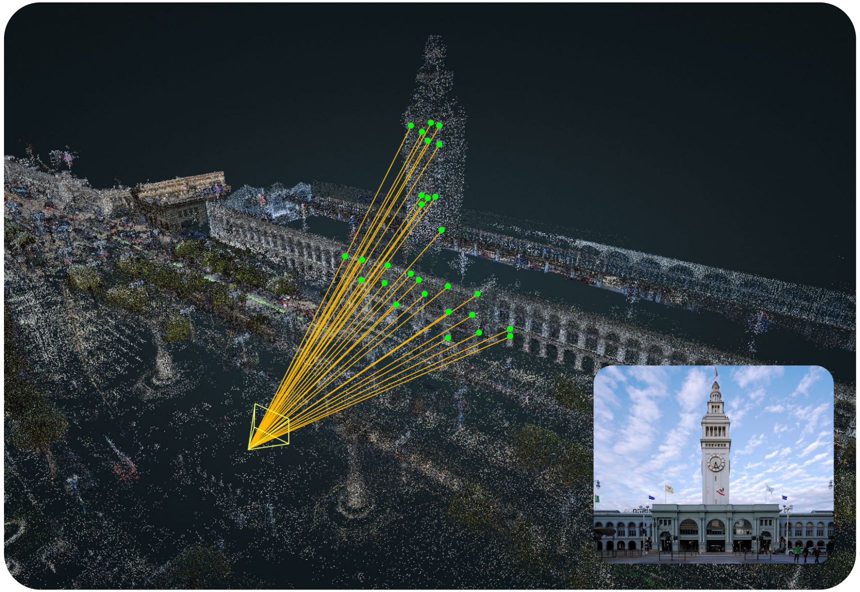




Still a traditional, structure-based method



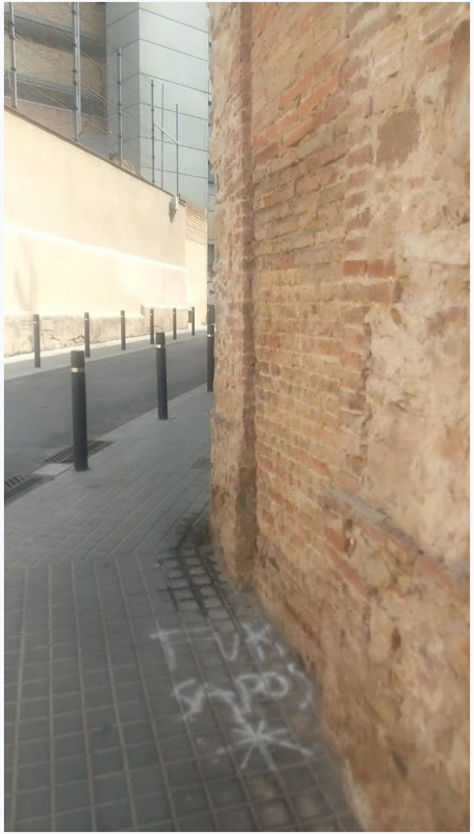
Large scale point-clouds from SV data are the foundation of VPS

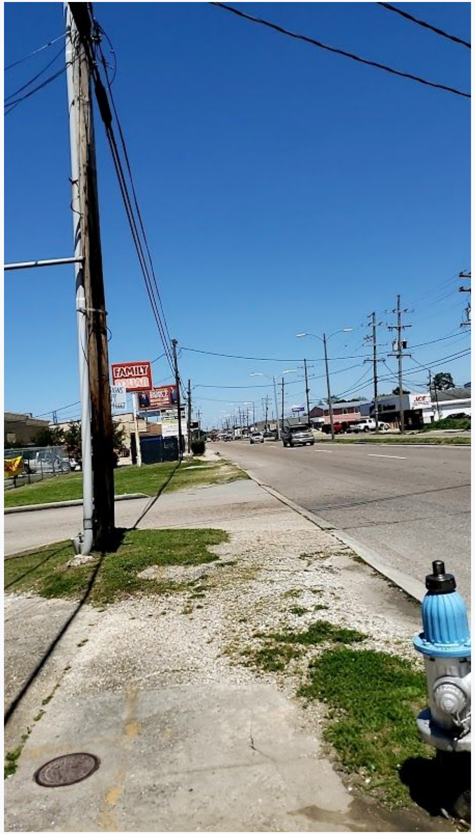


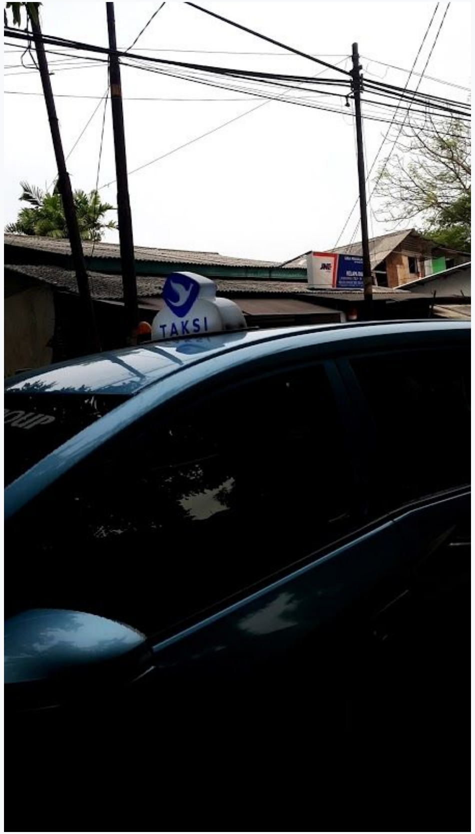
Queries are localized by matching points from the query image to the model

Details? Large-scale, real-time visual–inertial localization revisited (Simon Lynen et al, IJRR'20)

Challenging cases for VPS







SNAP: **S**elf-Supervised **N**eural M**ap**s





Paul-Edouard Sarlin
Google / ETH Zürich

Eduard Trulls
Google

Marc Pollefeys
ETH Zürich

Jan Hosang
Google

Simon Lynen
Google

SNAP: Self-Supervised Neural Maps for Visual Positioning and Semantic Understanding
Conference on Neural Information Processing Systems (NeurIPS), 2023

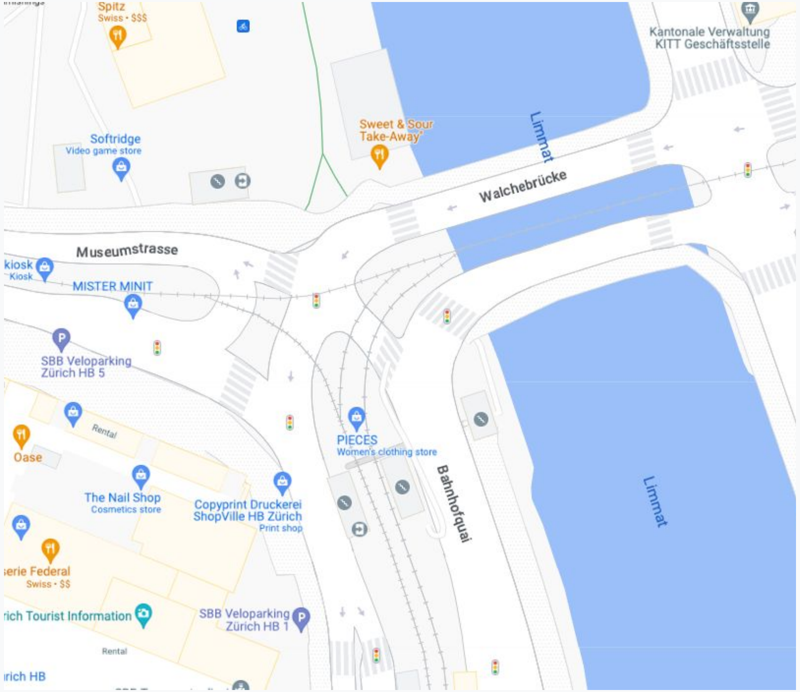
What makes a map useful for localization?

Abstract enough to be robust to changes

- Appearance, dynamic objects

While preserving **geometric & semantic information**

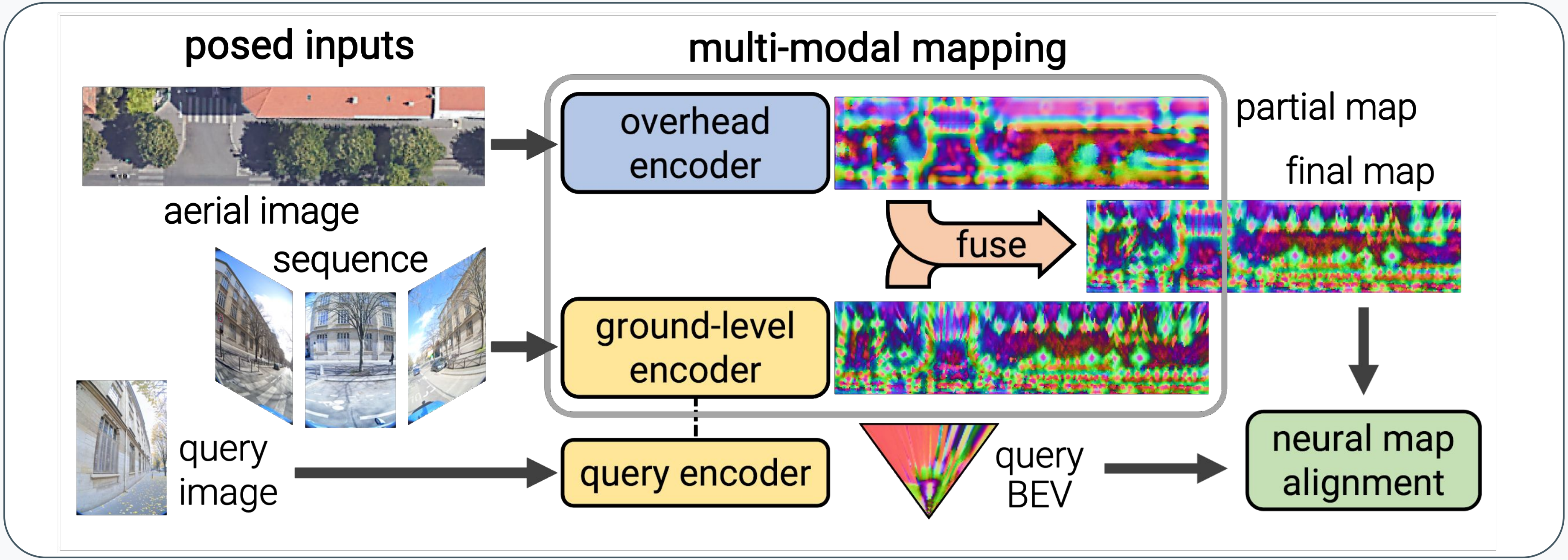
- *What distinctive objects and layout do I observe in the scene?*



Neither aerial, nor ground-level imagery itself makes for a good map

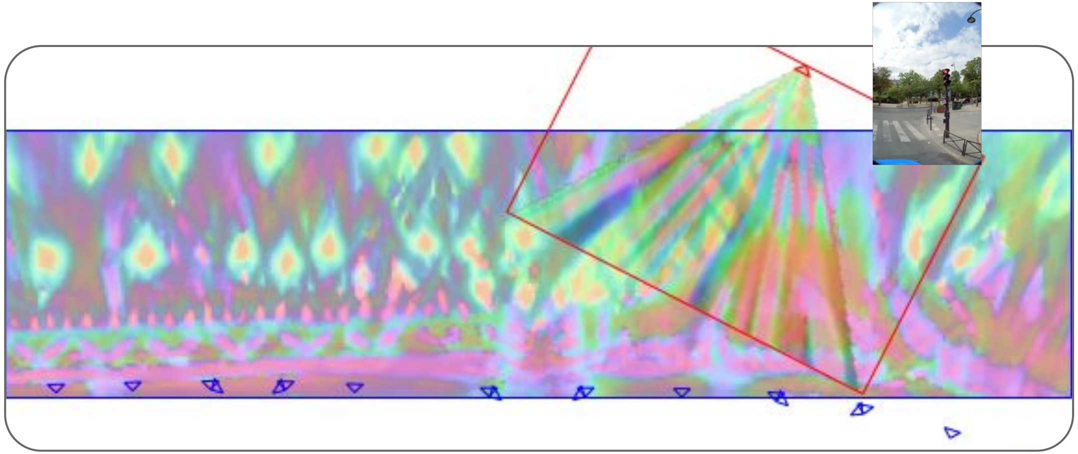
The right level of detail and abstraction is key

SNAP: **S**elf-Supervised **N**eural **Ma**ps



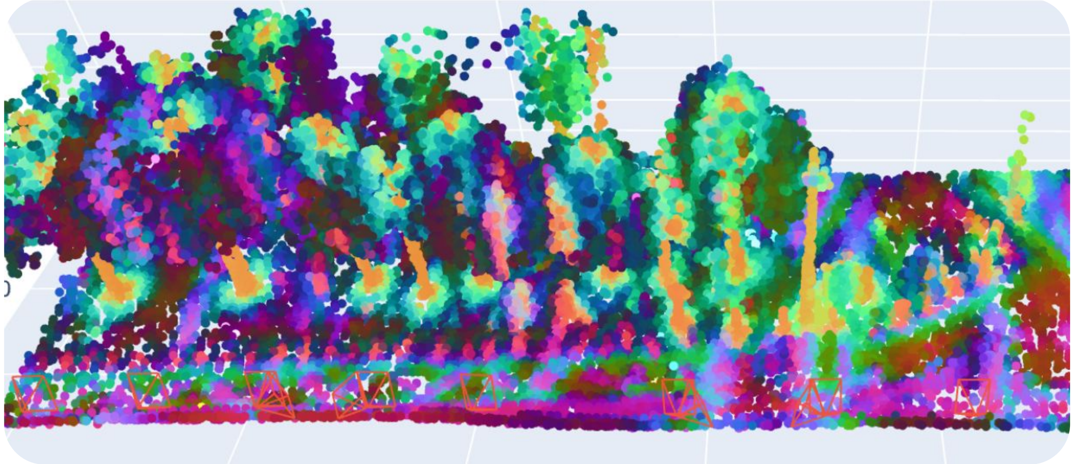
Aerial and ground-level images are complementary

How? SNAP is **trained to align** these neural maps, in a **contrastive fashion**



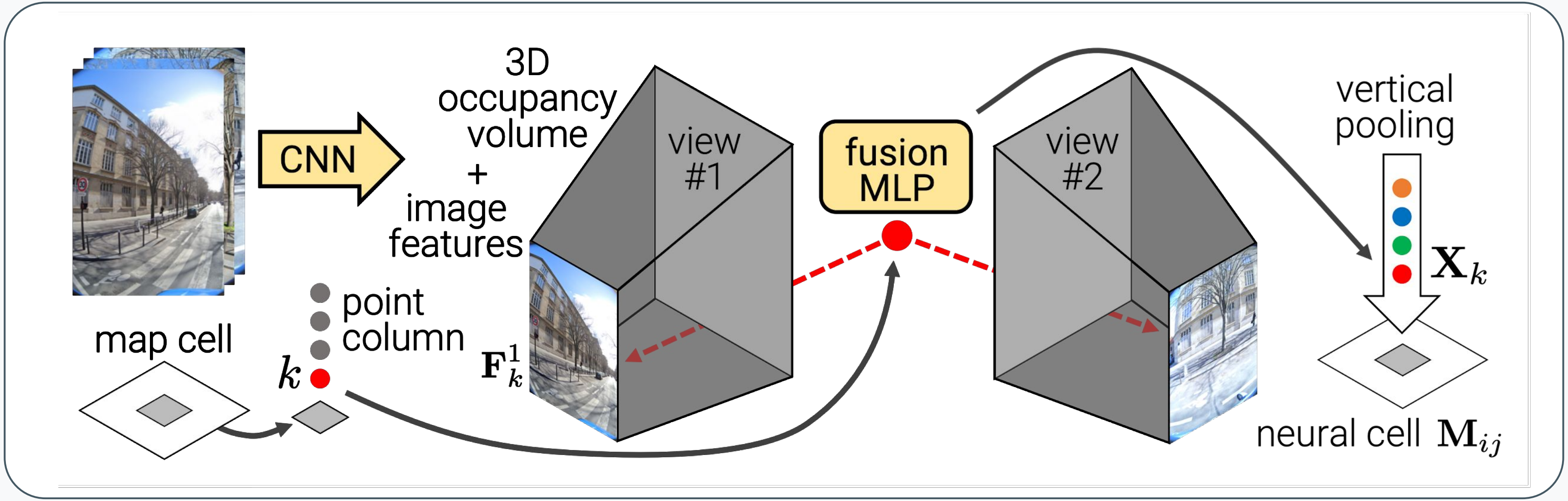
Localizing opposing views using SNAP

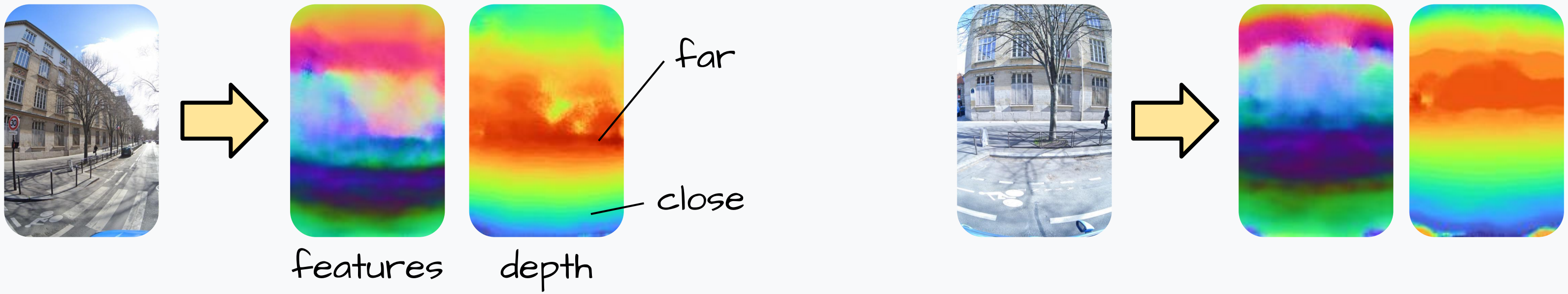
What happens? SNAP learns to discover objects **using only poses, without semantics**



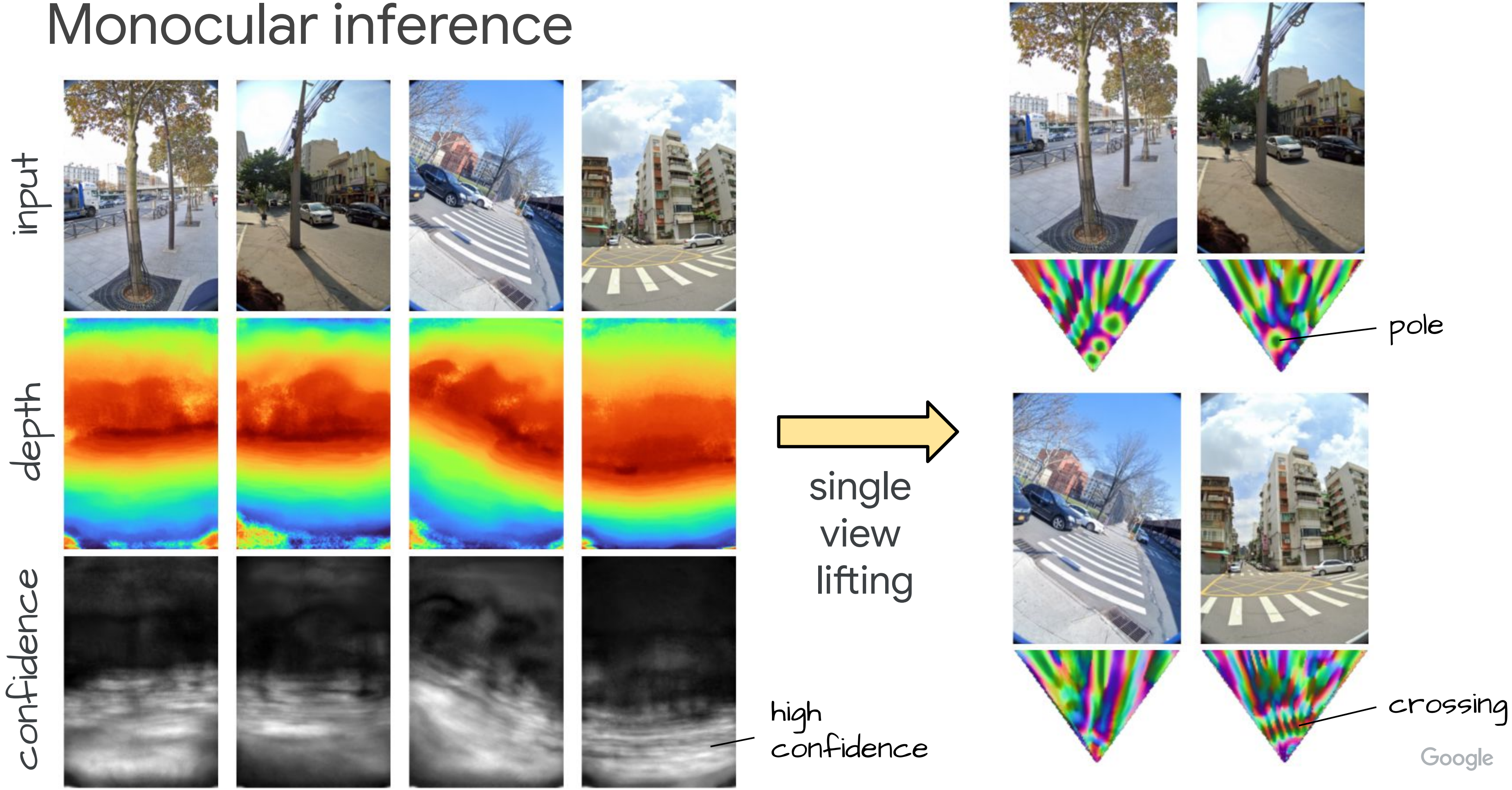
SNAP's neural map lifted to 3d using lidar

StreetView image encoder





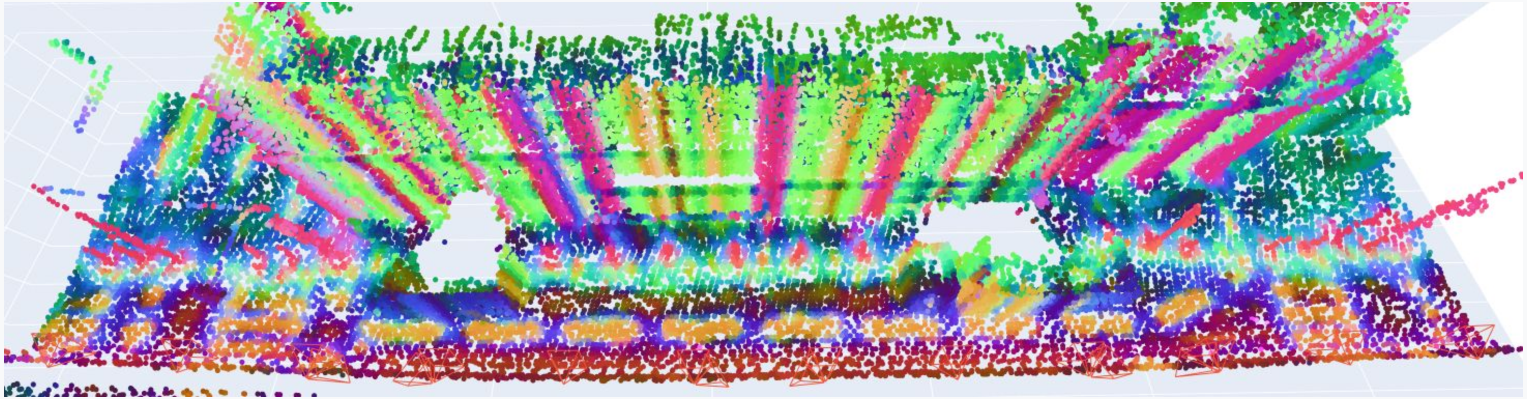
Monocular inference

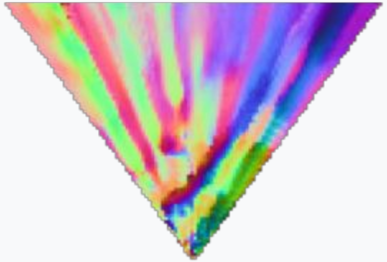


Learning from pose supervision

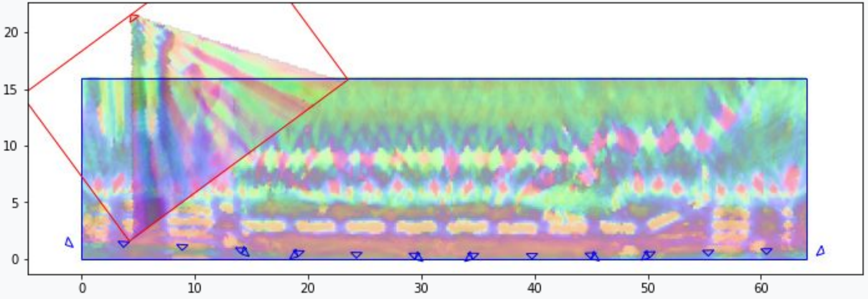
map
64x16m

query

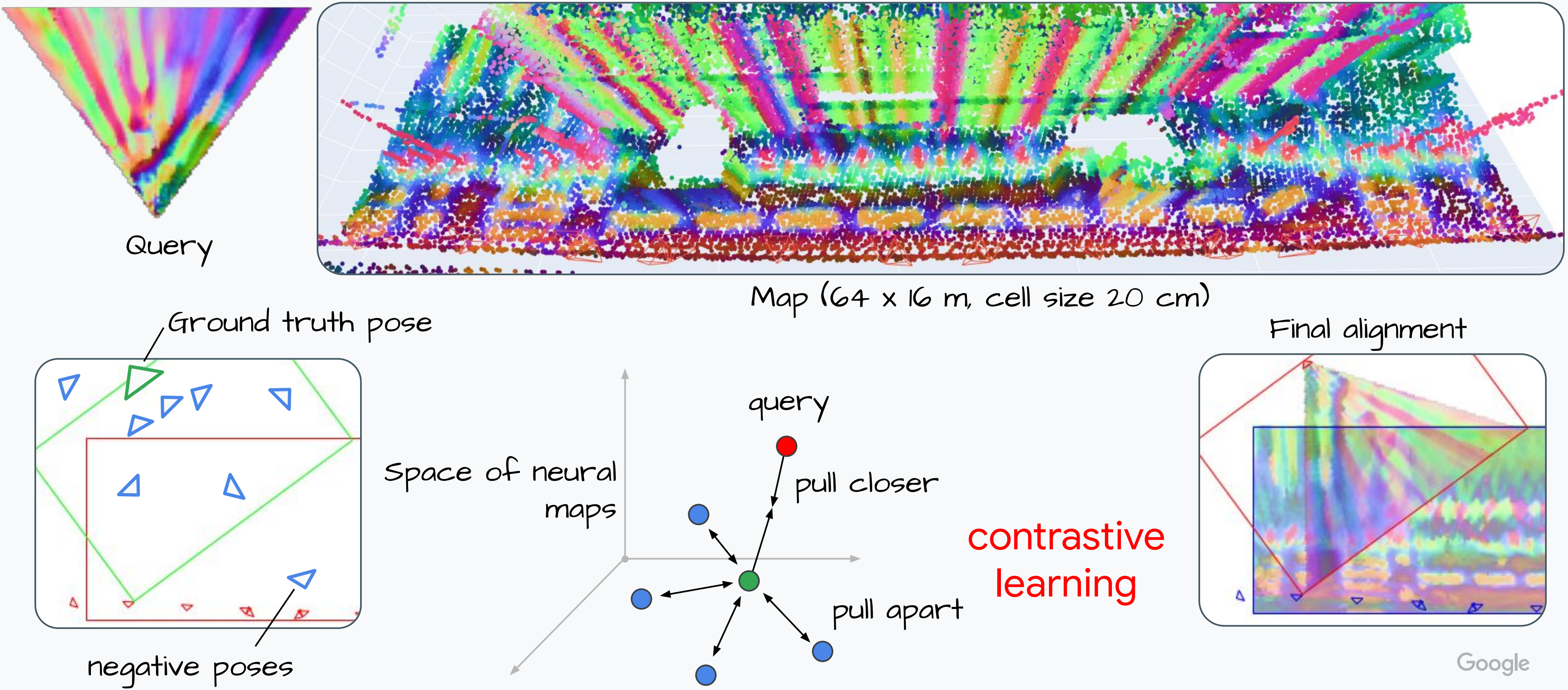




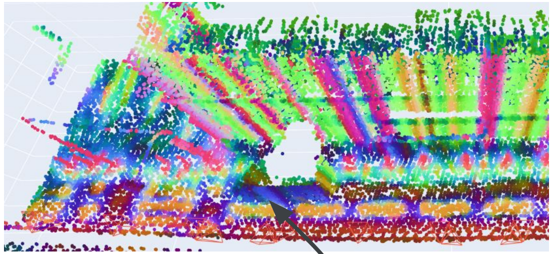
Δp=50cm, ΔR=0.5°



Learning from pose supervision

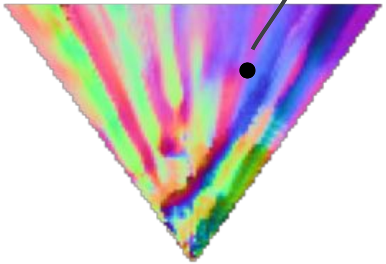


Sampling negative poses with RANSAC

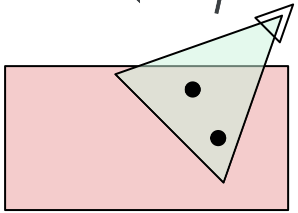


Map (lifted to 3d using lidar)

exhaustive matching

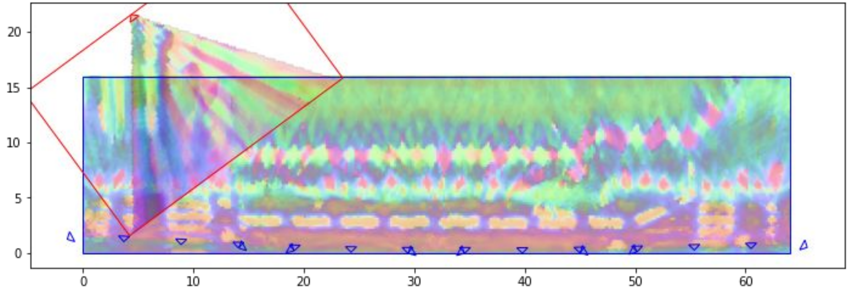


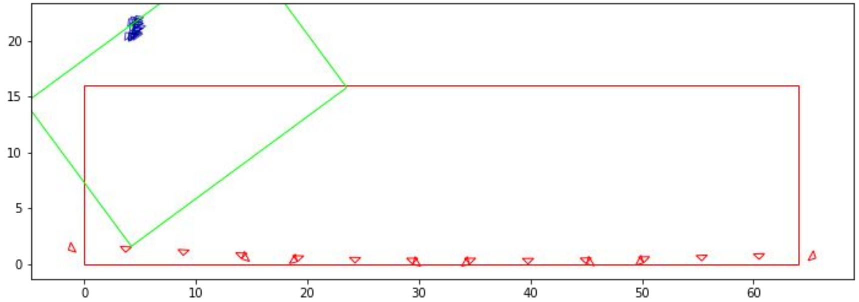
query



sample minimal set + 2-point solver

featuremetric pose voting



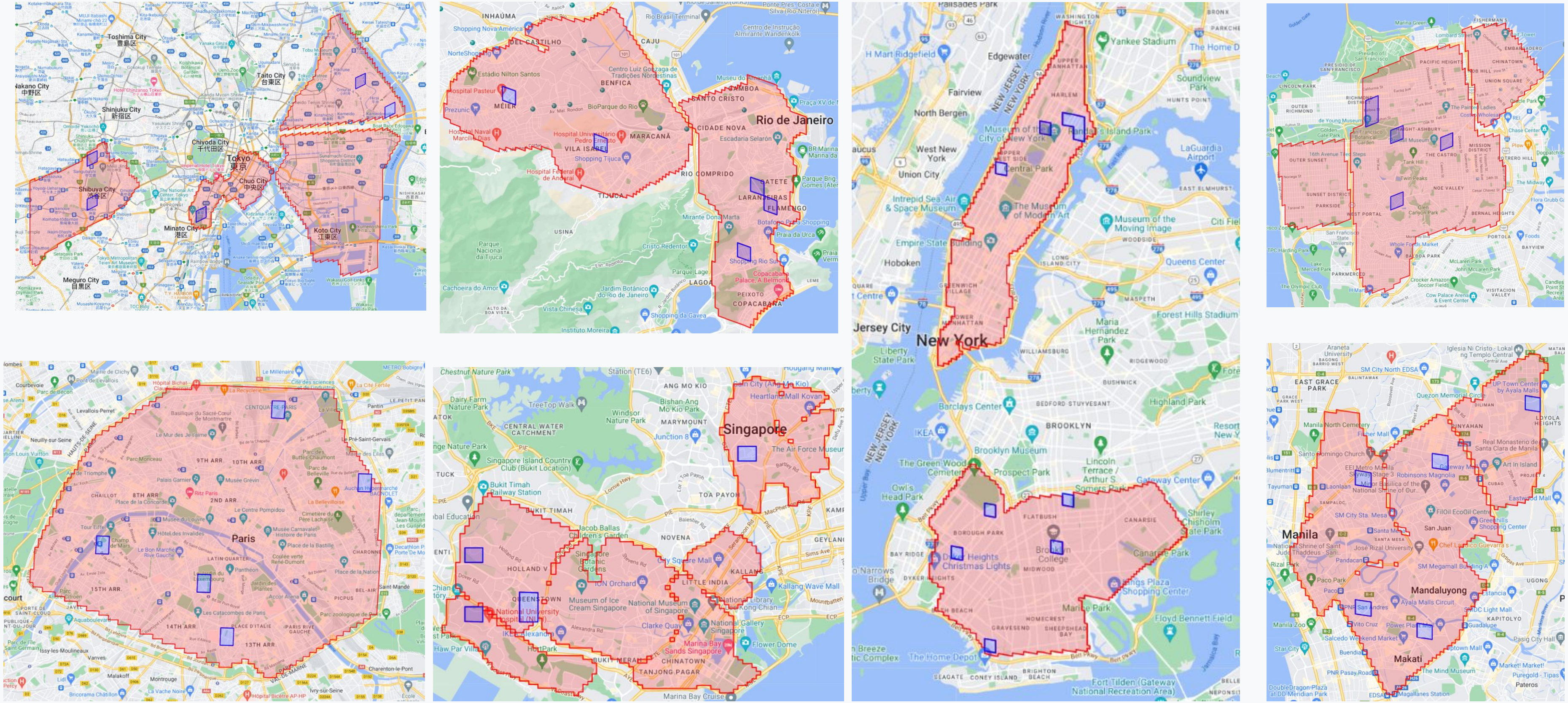


Softmax = distribution over poses

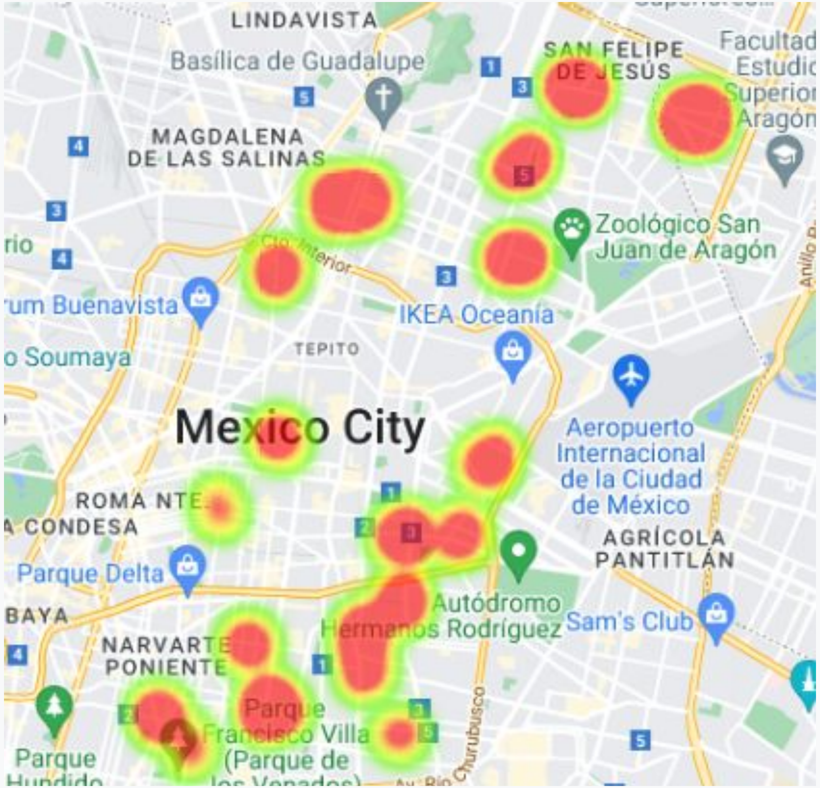
Google

Training: 11 cities in 5 continents

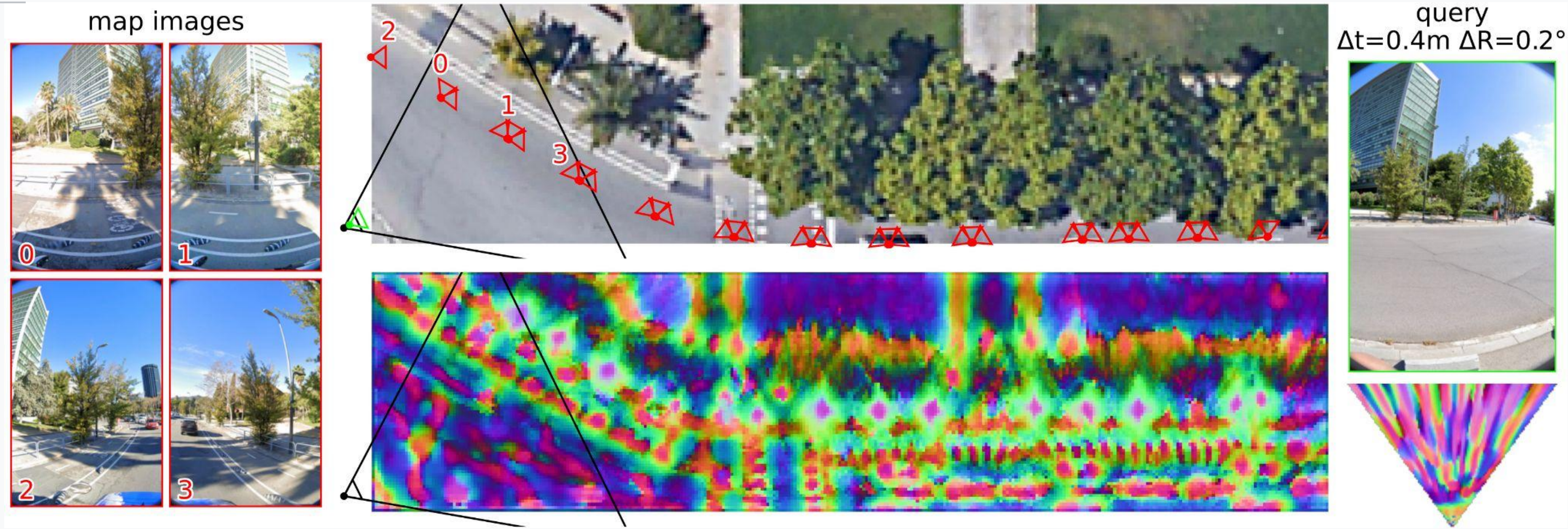
Blue: validation
Red: training



Test distribution: 6 cities

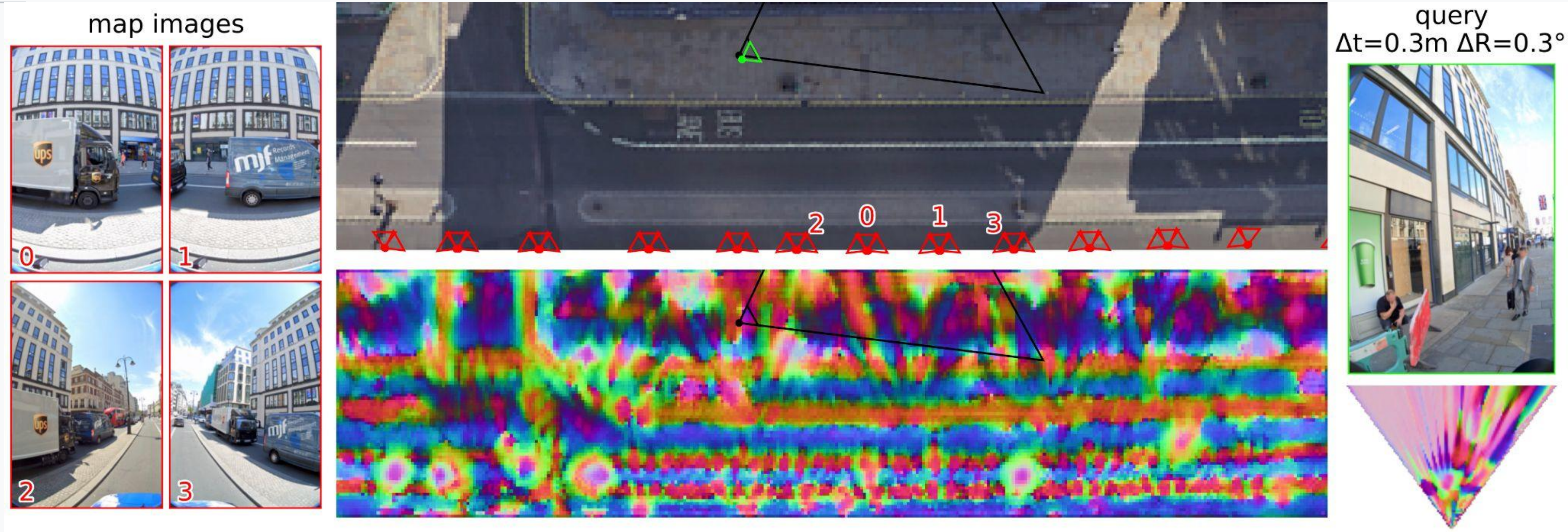


Localization examples

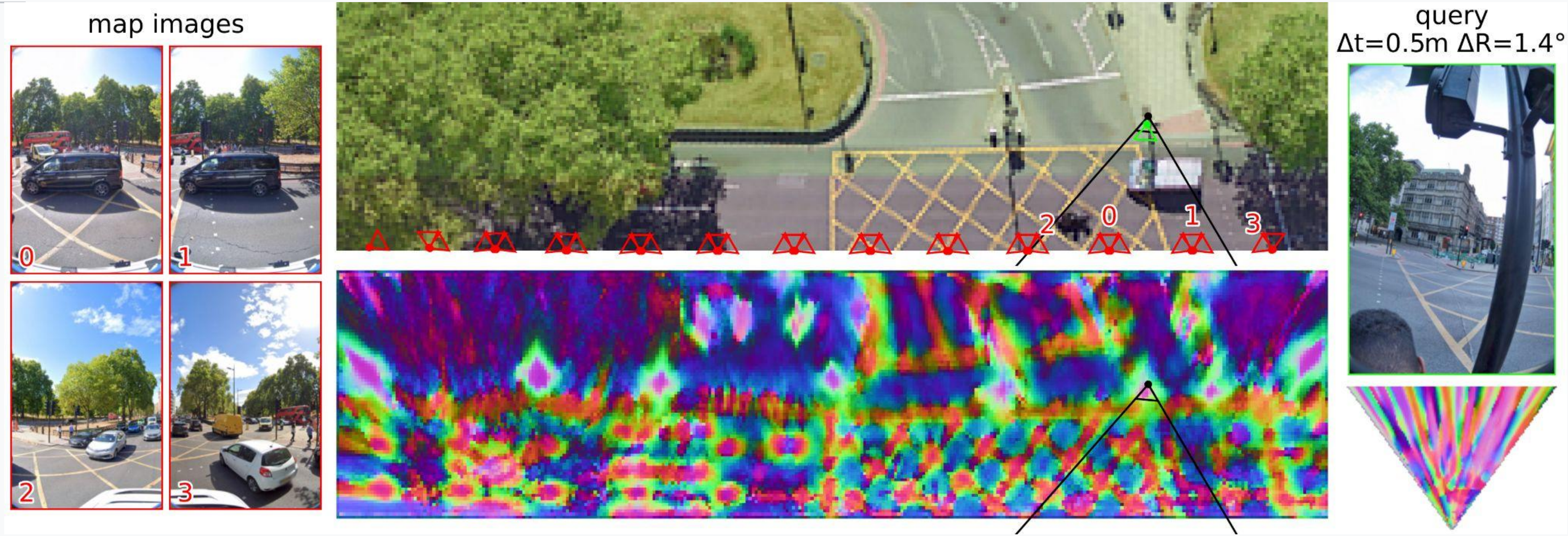


Google

Localization examples

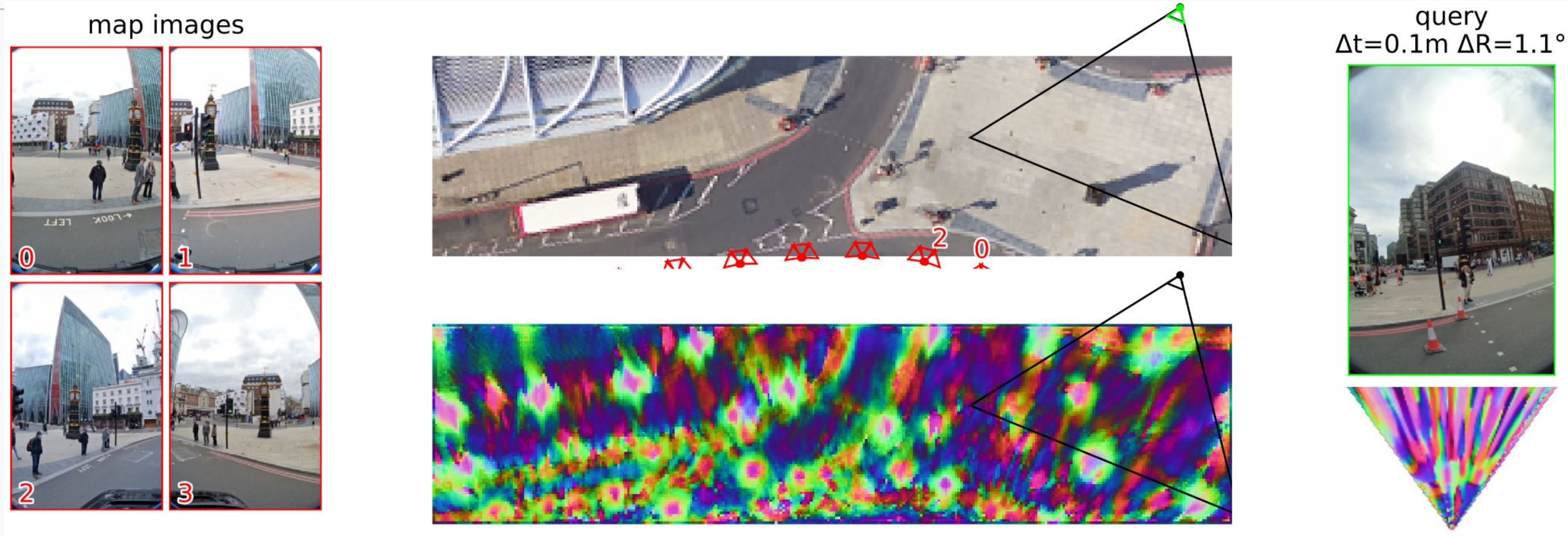


Localization examples

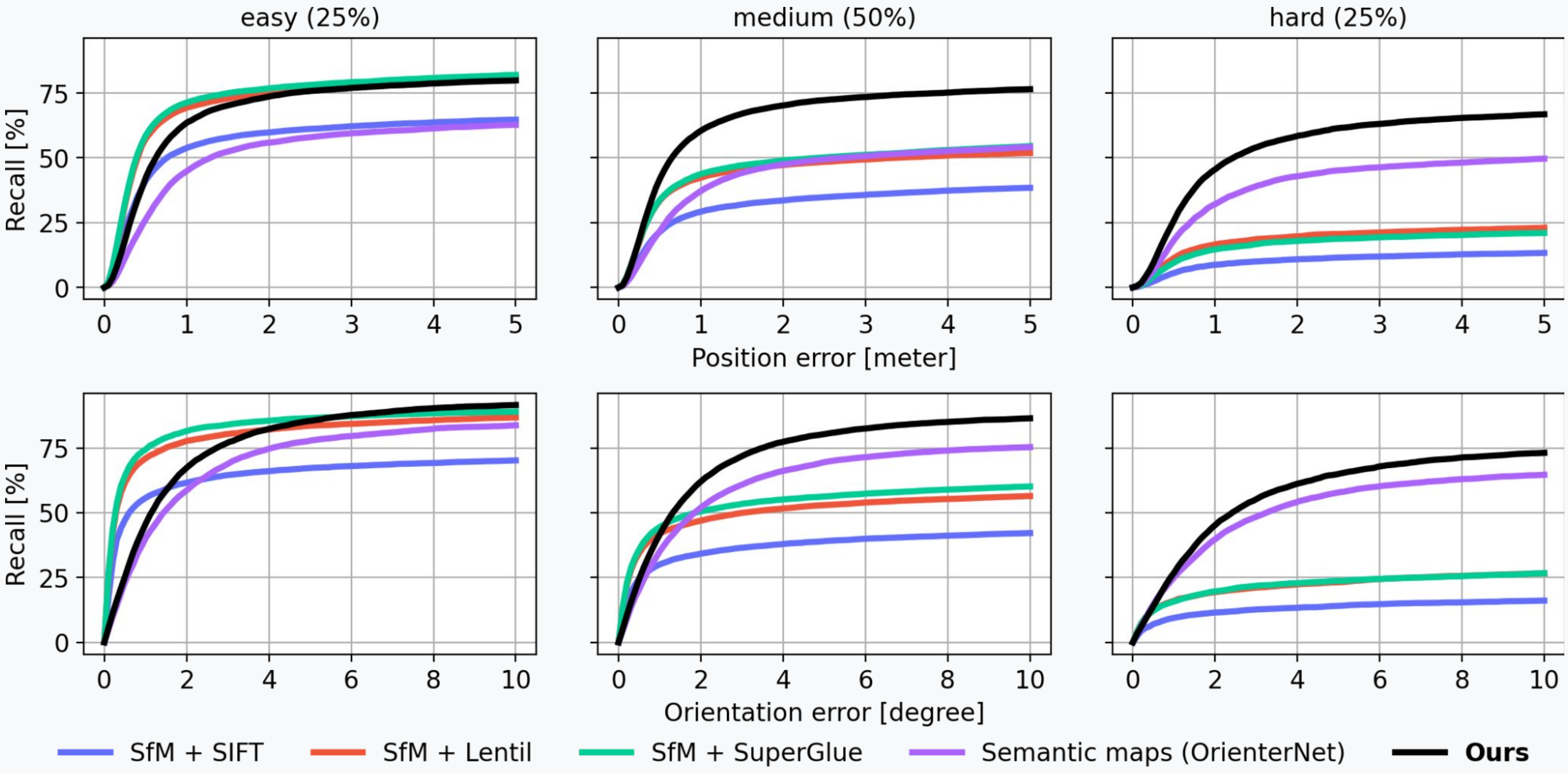


Google

Localization examples

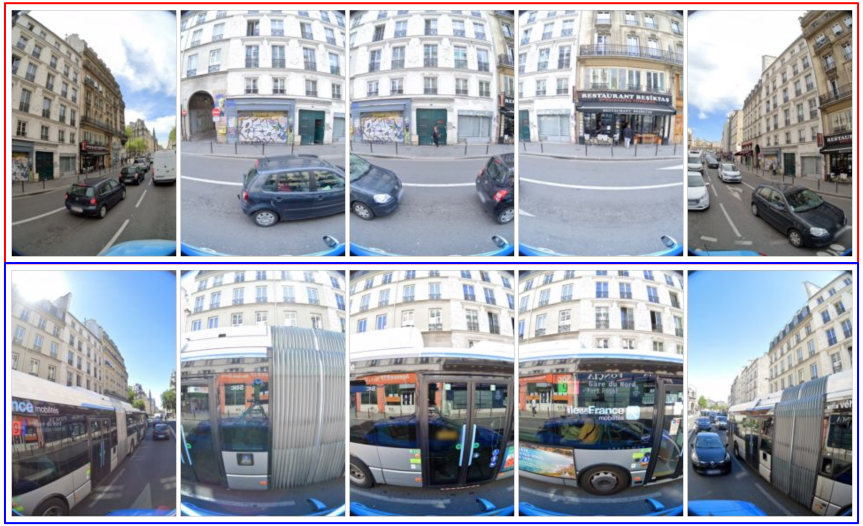


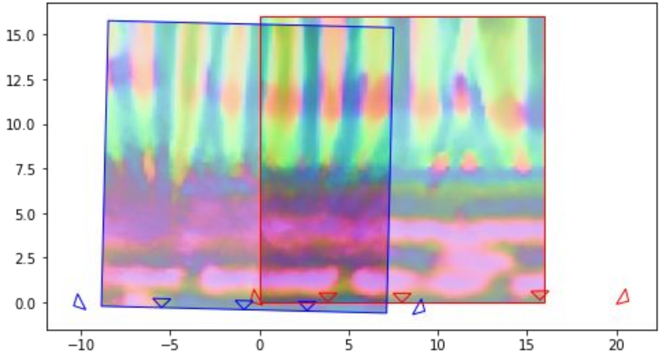
Comparison to other localization approaches



Google

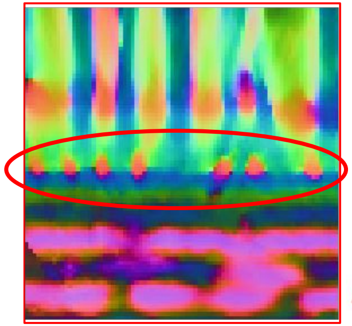
Sequence-to-Sequence

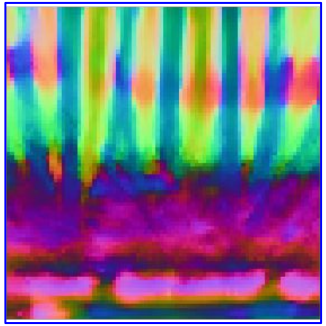


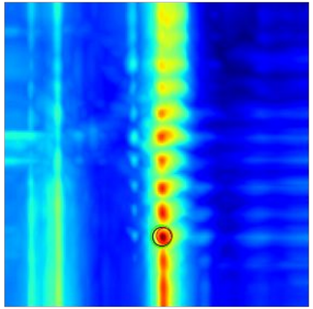


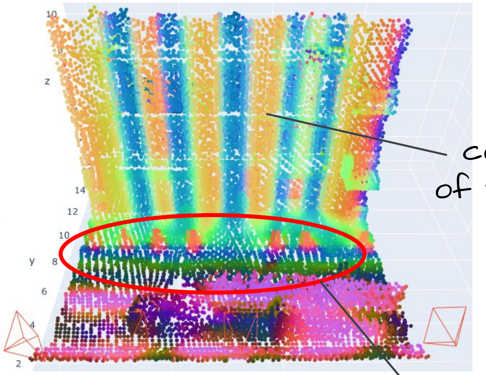
Δp=30cm
ΔR=0.05°

neural maps

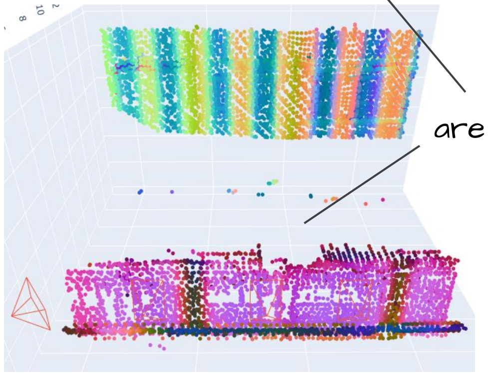








columns
of windows



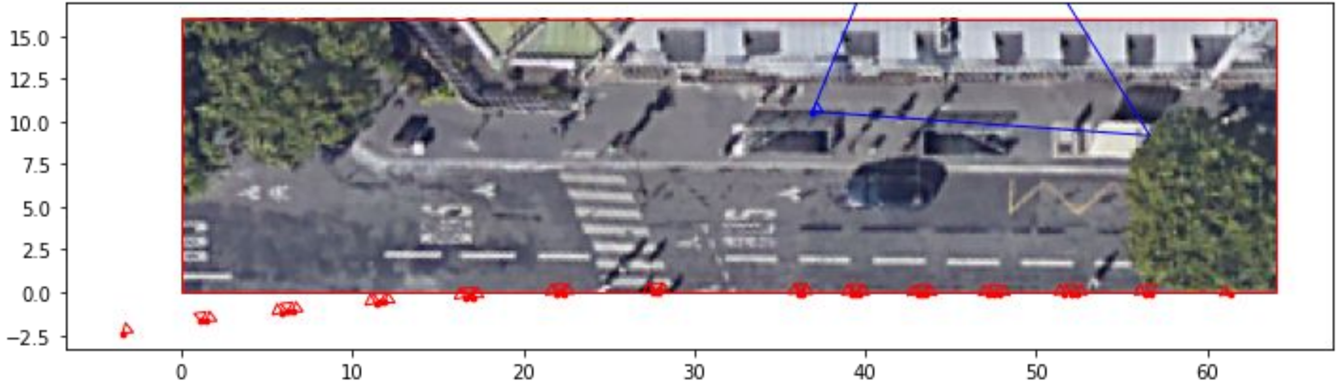
poles
are occluded

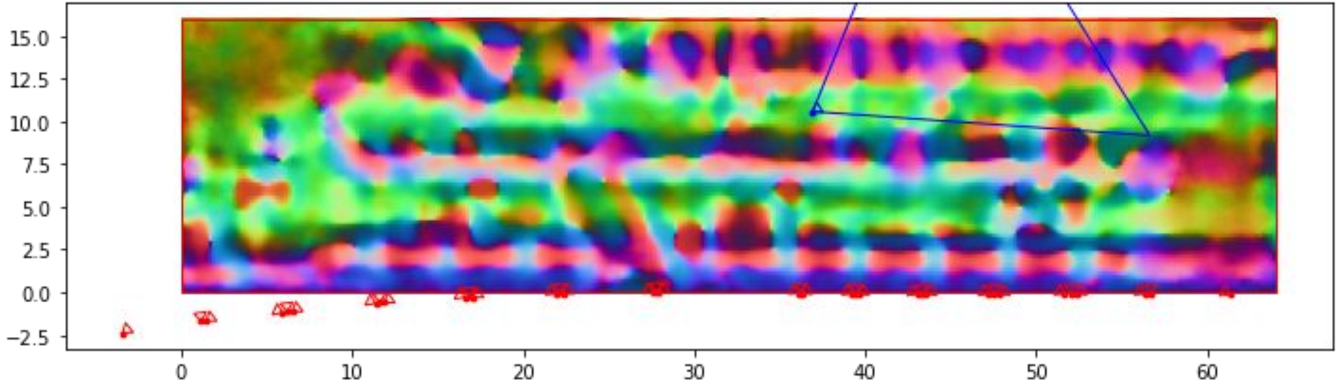
Feature-colored lidar points
(only for visualization)

The exhaustive likelihood
is multi-modal,
it captures the symmetry

Aerial-to-ground localization

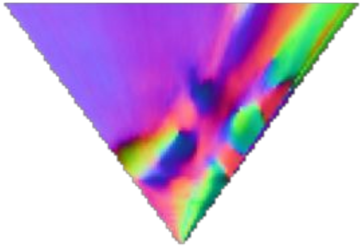
aerial input image





aerial map

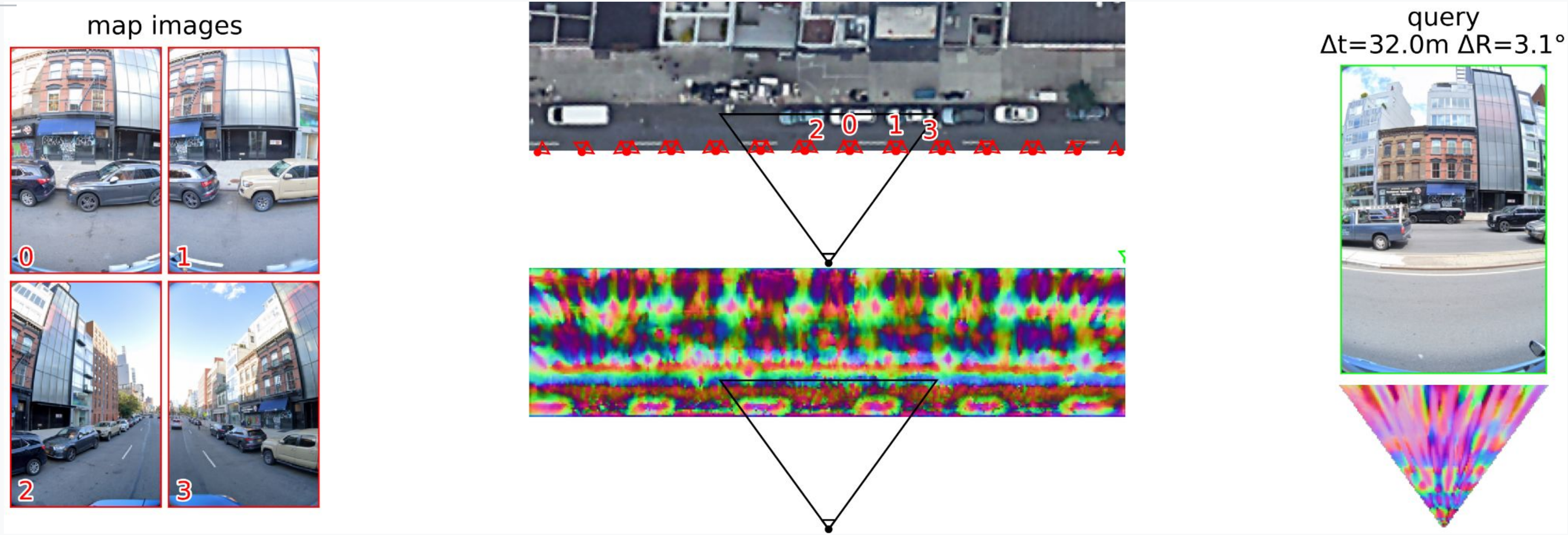
query image



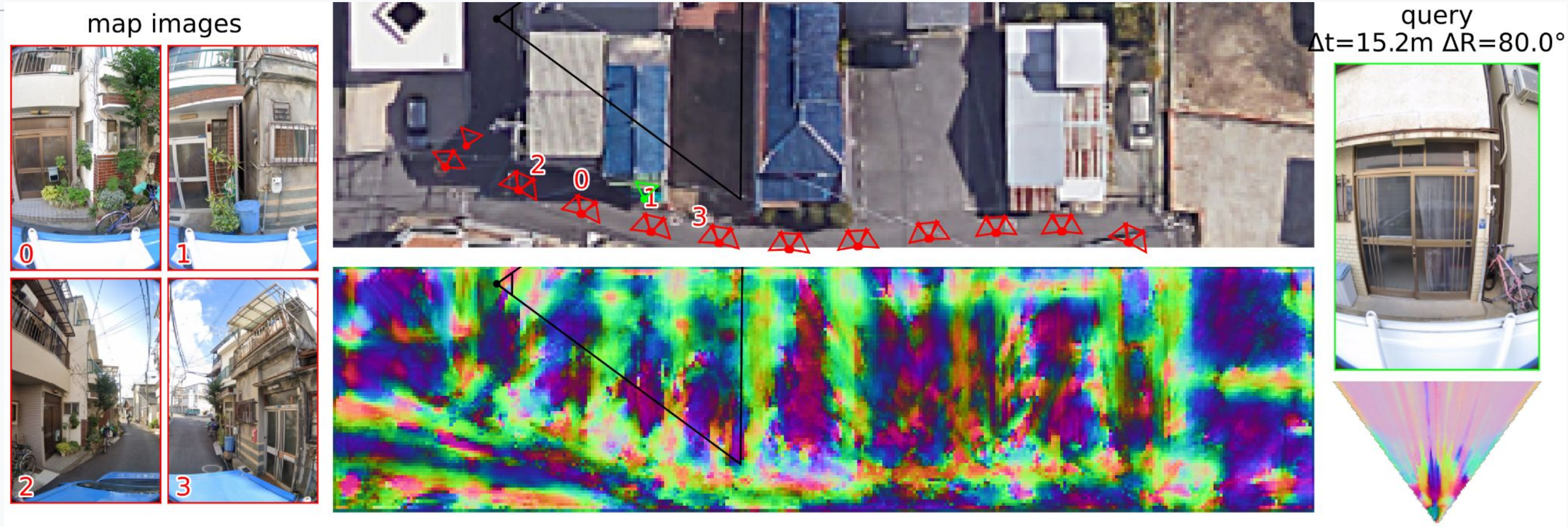
error: Δp=52cm, ΔR=0.7°

Google

Localization examples: failure case



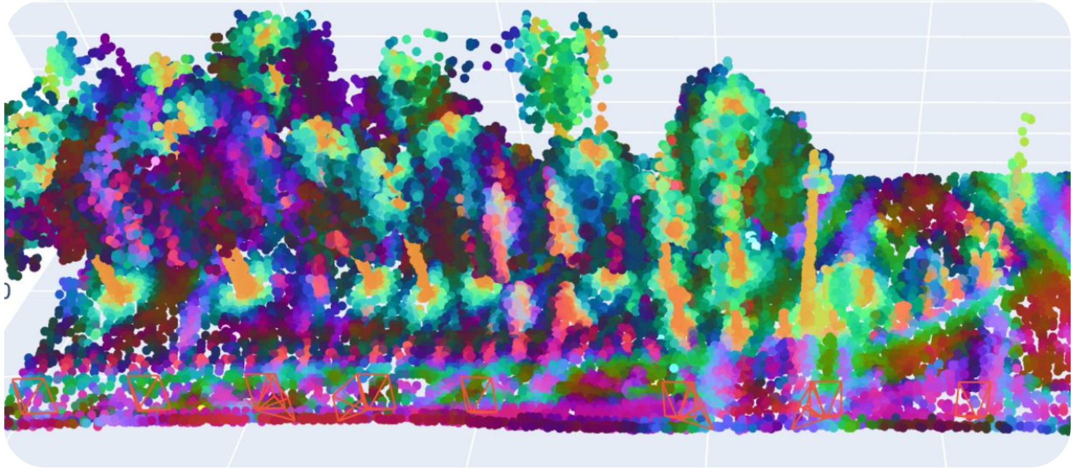
Localization examples: failure case



Beyond localization

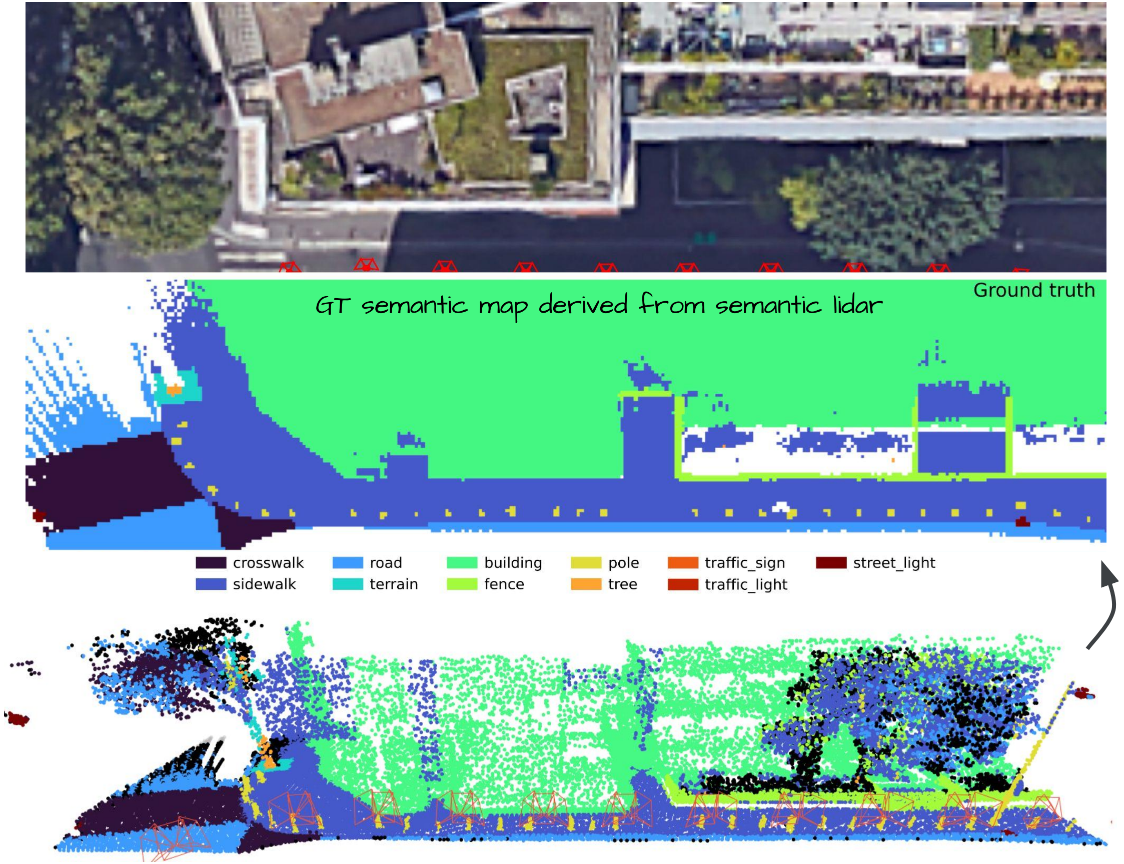
Self-supervised Neural Maps
for Visual Positioning
and Semantic Understanding

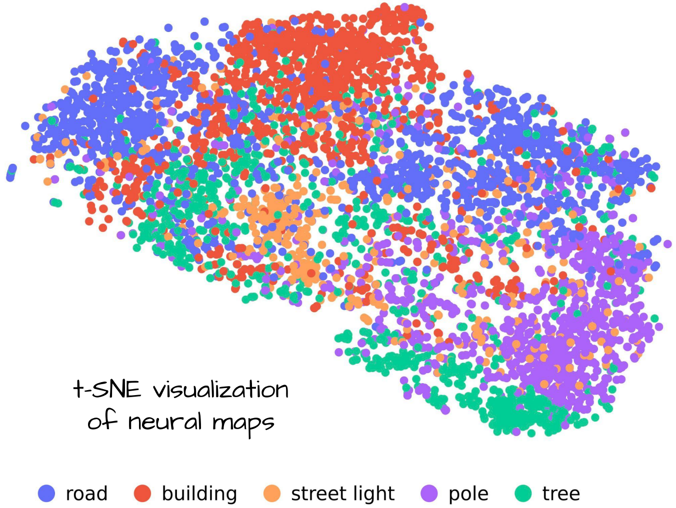
...while training only with poses!



SNAP's semantic map lifted to 3d using lidar

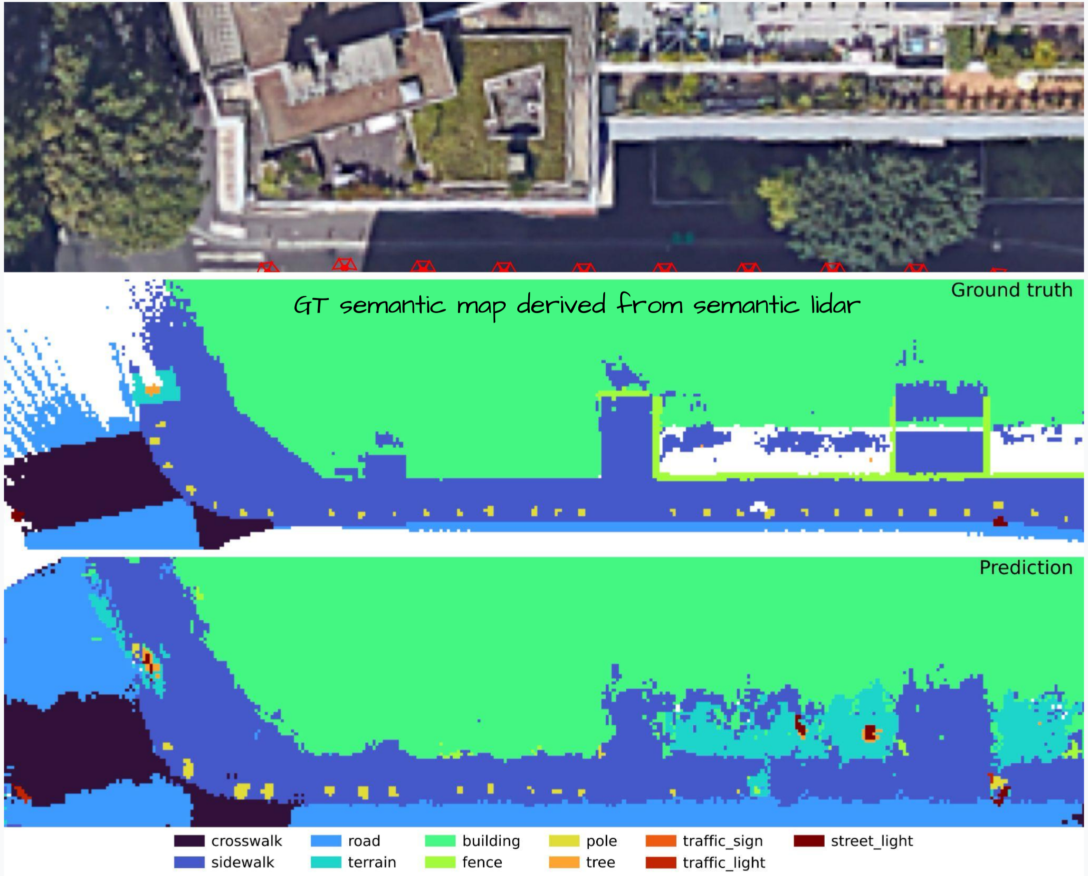
SNAP learns to discover objects

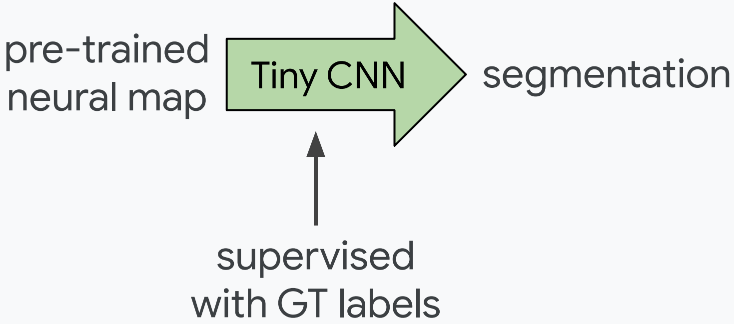


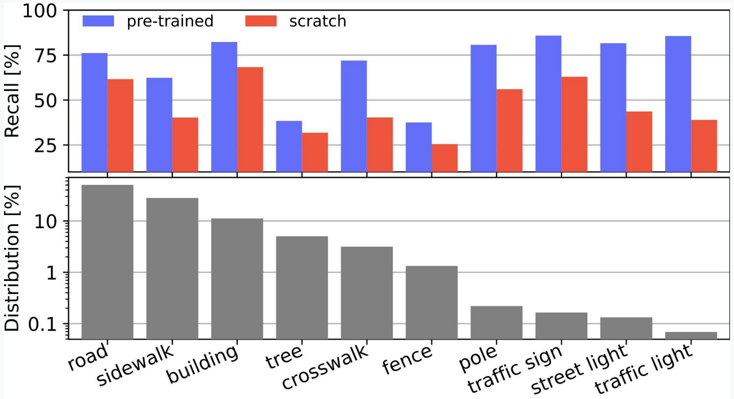


SNAP distinguishes trees vs poles **without any supervision**

Decoding explicit semantics



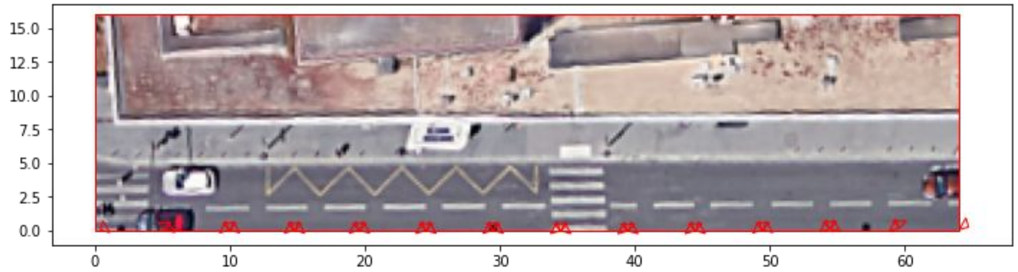




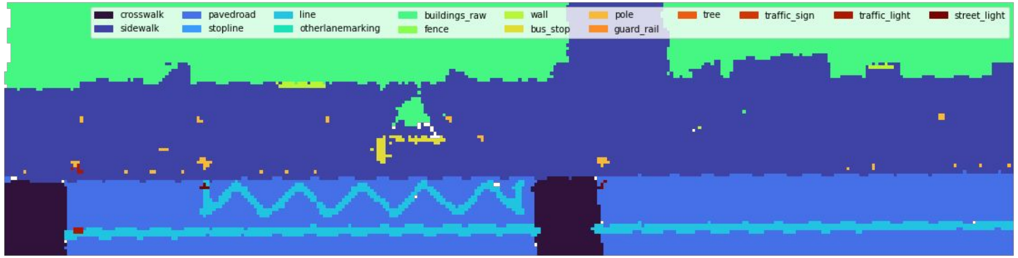
from only 2k labeled scenes

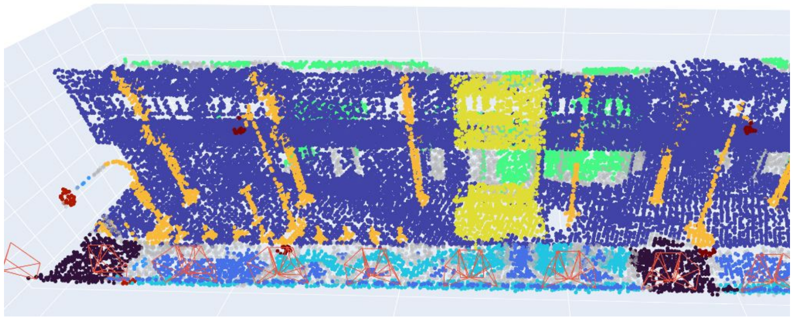
Google

Qualitative results

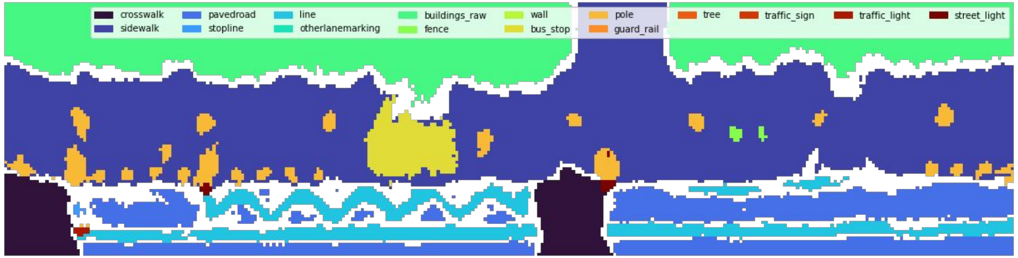


Ground truth





Prediction



Summary and open challenges

- Summary
  - SNAP learns **2D neural maps** directly from **posed multi-modal imagery**
  - Supervised with **only poses**, via contrastive learning
  - Localization serves as pre-training for **high-level semantics** without labels
- Limitations
  - Not as accurate for queries close to map images
  - Assumes known gravity and a location prior (3DOF not 6DOF)
  - Semantics are a good start, but true "foundation models" are still a few steps away
- What makes this possible?
  - A **unique corpus** with 200B+ posed StreetView images, co-registered with other modalities: aerial images, LiDAR, semantics, etc.
  - **We collaborate with universities and host interns!**
    - **Reach out! {trulls,slynen}@google.com.**
    - **Open "research internships" call @ Google careers website: October 25**