ETH zürich [1]

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Meta [2]

REALITY LABS

OrienterNet

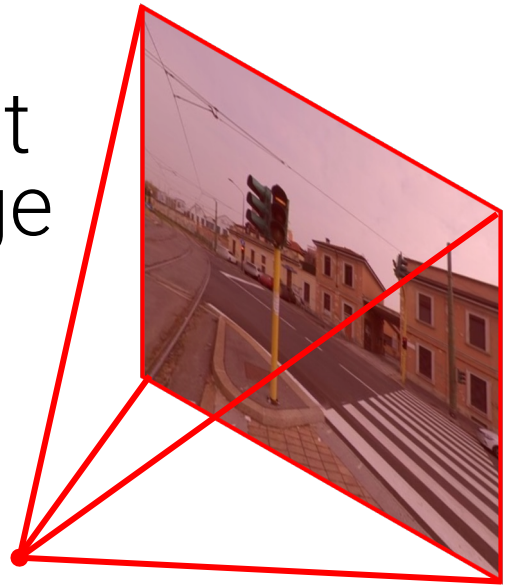
Visual Localization in 2D Public Maps with Neural Matching

Paul-Edouard Sarlin[1] Daniel DeTone[2] Tsun-Yi Yang[2] Armen Avetisyan[2]
Julian Straub[2] Tomasz Malisiewicz[2] Samuel Rota Bulo[2]
Richard Newcombe[2] Peter Kontschieder[2] Vasileios Balntas[2]

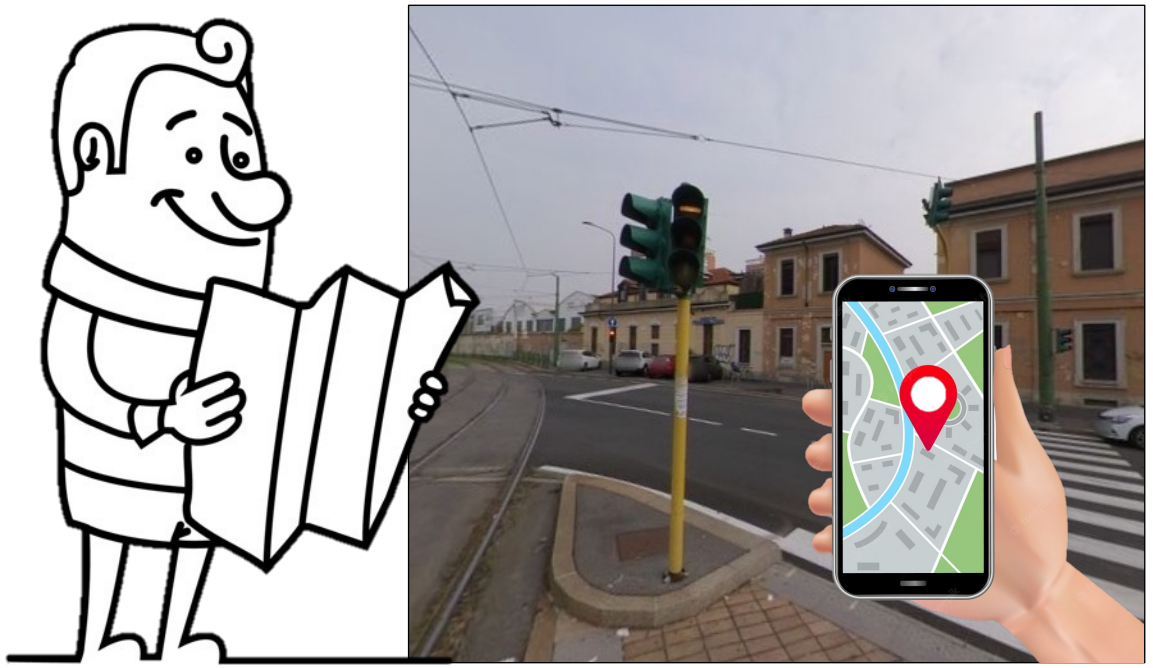
psarlin.com/orienternet

Humans use simple 2D maps

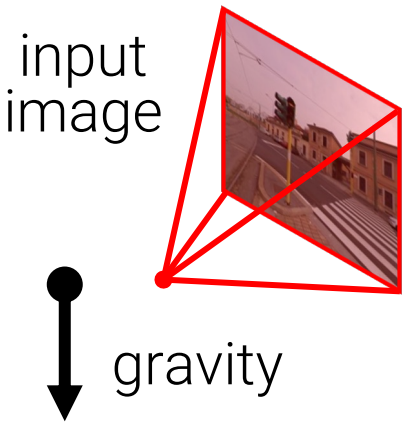
input image



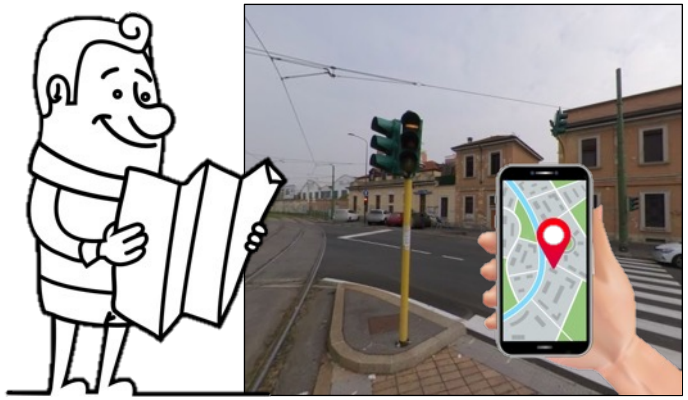
where am I?



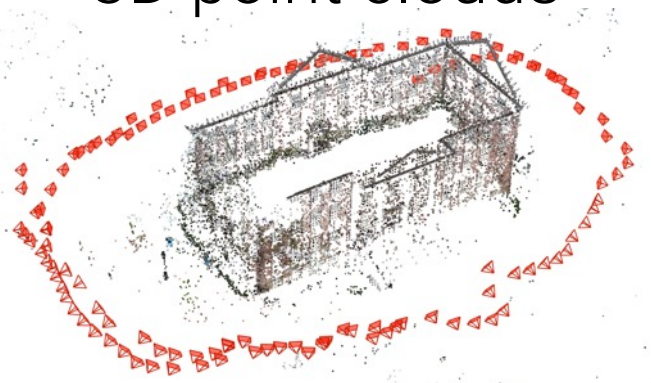
inputs



Humans use 2D maps

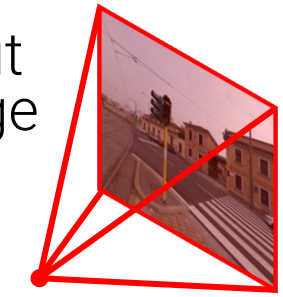


Existing algorithms:
3D point clouds



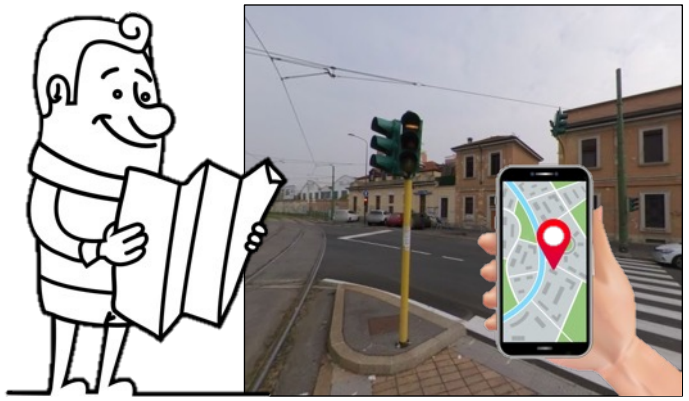
inputs

input image

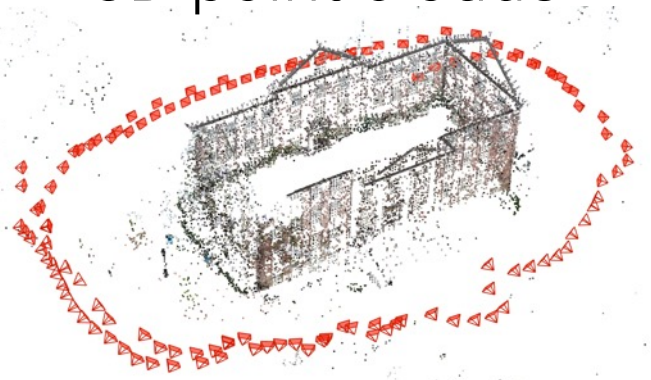


gravity

Humans use 2D maps

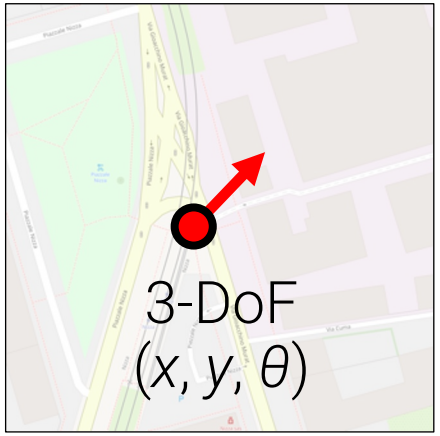


Existing algorithms: 3D point clouds



OpenStreetMap

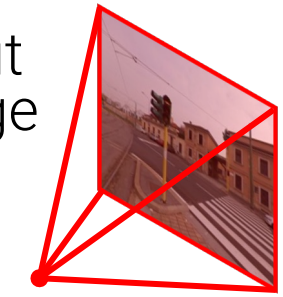
OrienterNet



camera pose

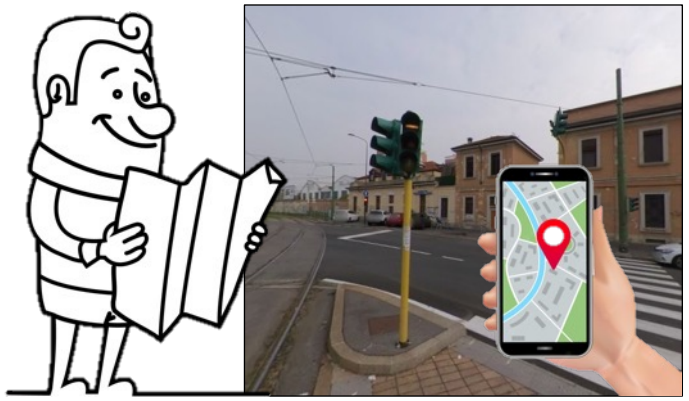
inputs

input image

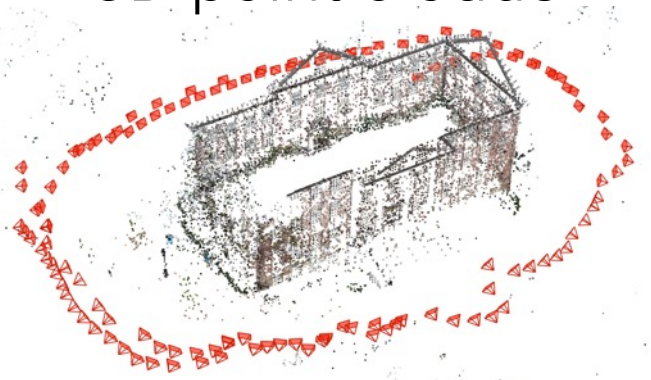


gravity

Humans use 2D maps

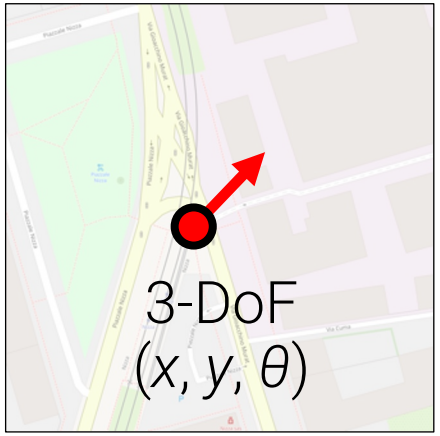


Existing algorithms: 3D point clouds



OpenStreetMap

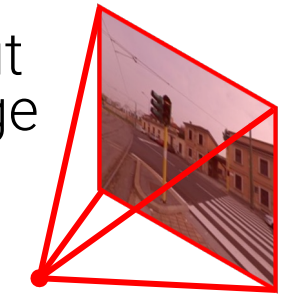
OrienterNet



camera pose

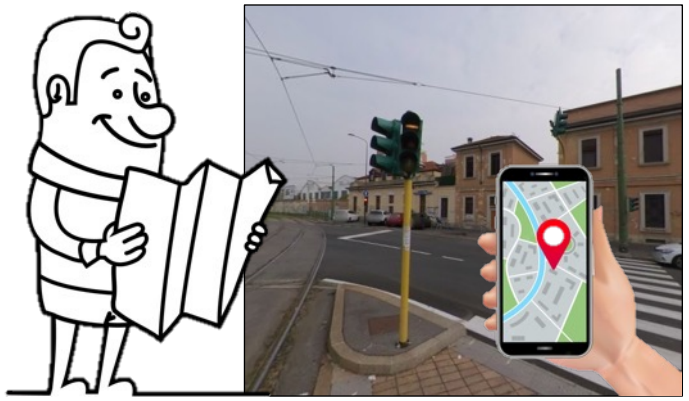
inputs

input image

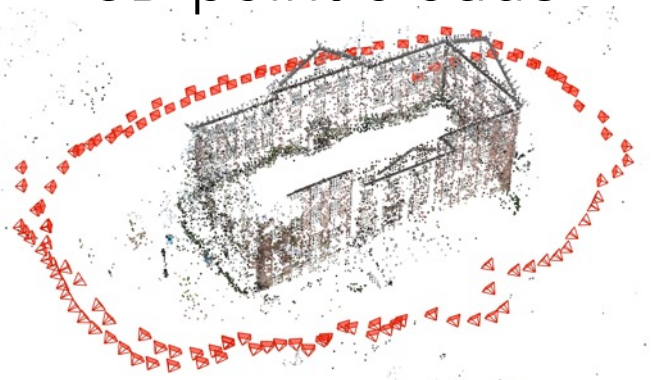


gravity

Humans use 2D maps



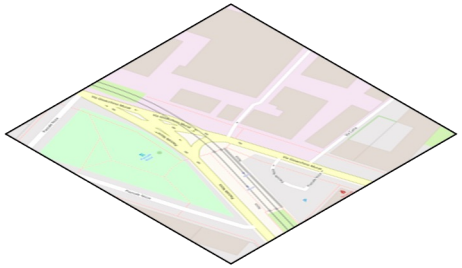
Existing algorithms: 3D point clouds

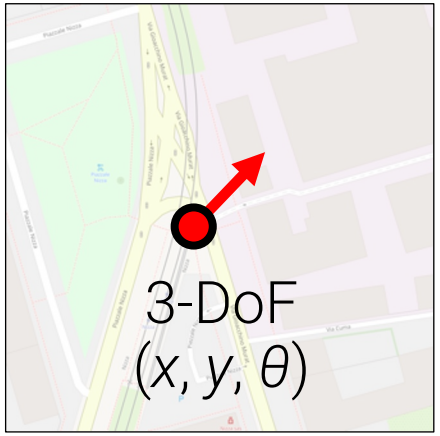


OpenStreetMap

OrienterNet

neural map matching

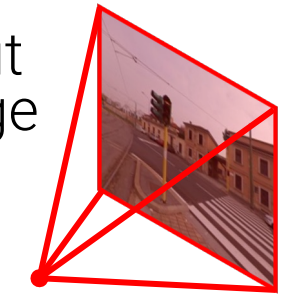




camera pose

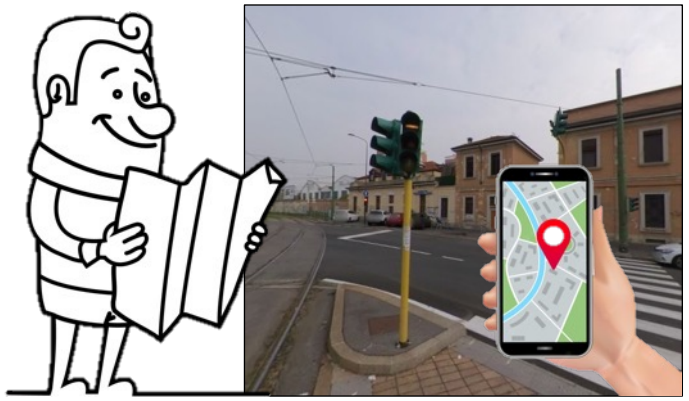
inputs

input image

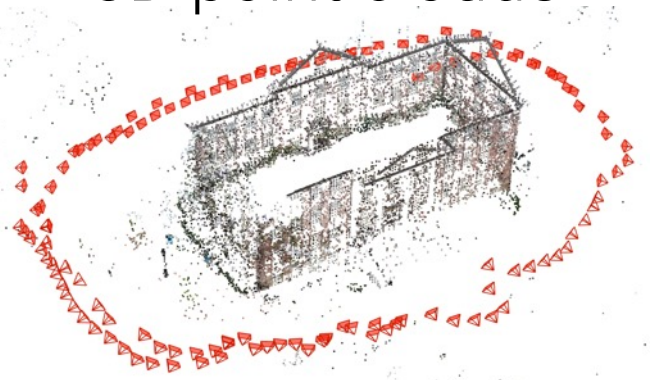


gravity

Humans use 2D maps



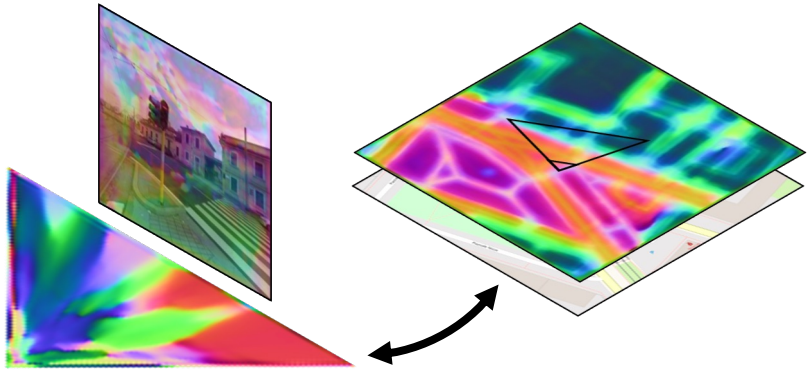
Existing algorithms: 3D point clouds

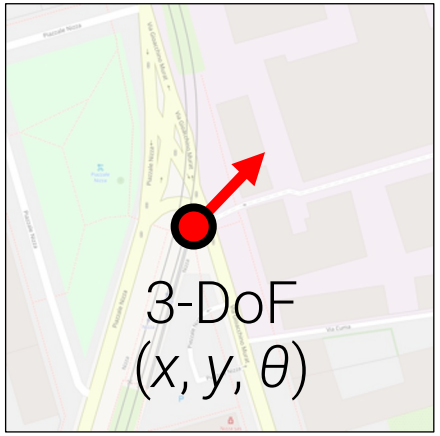


OrienterNet

neural map matching

OpenStreetMap

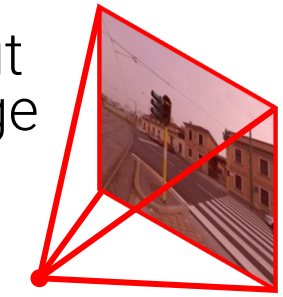




camera pose

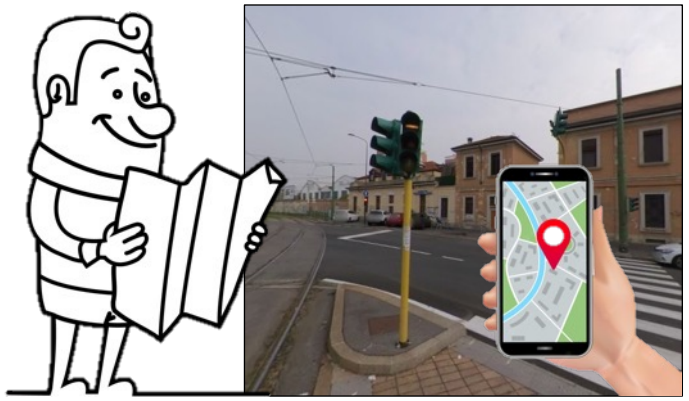
inputs

input image

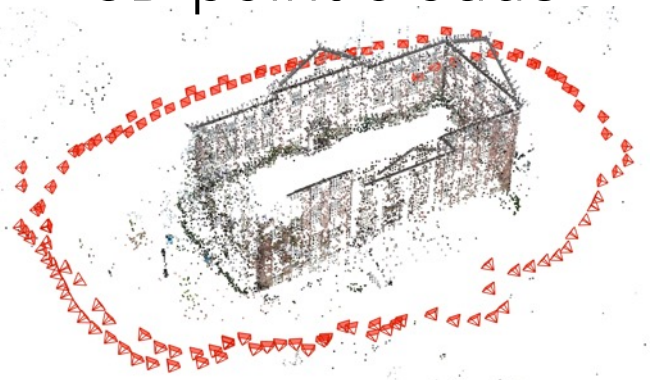


gravity

Humans use 2D maps

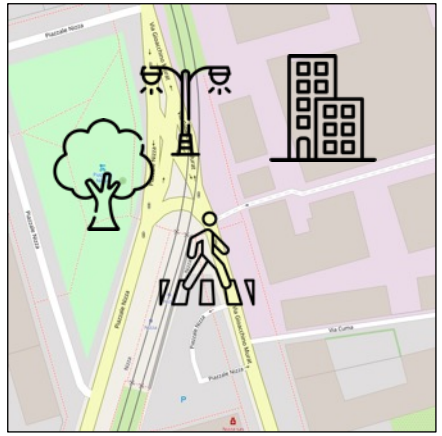


Existing algorithms: 3D point clouds

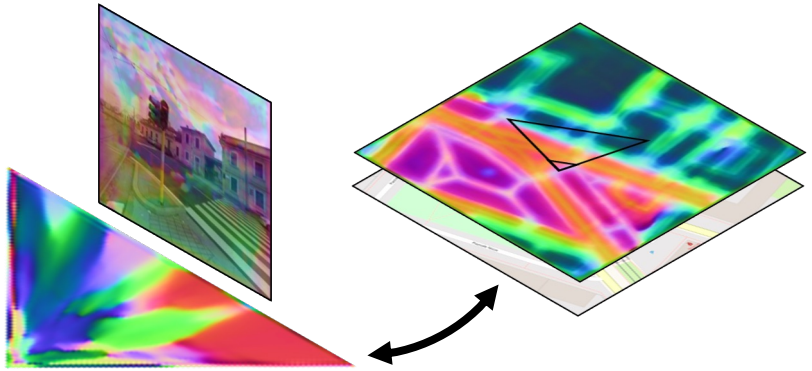


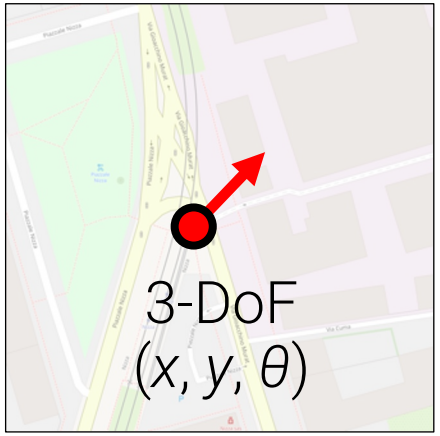
OrienterNet

neural map matching



OpenStreetMap





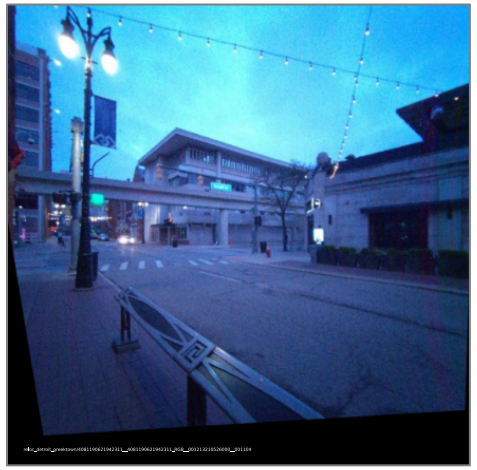
camera pose

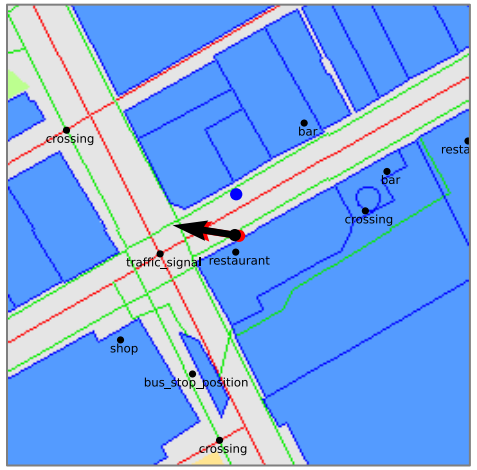
Zero-shot generalization

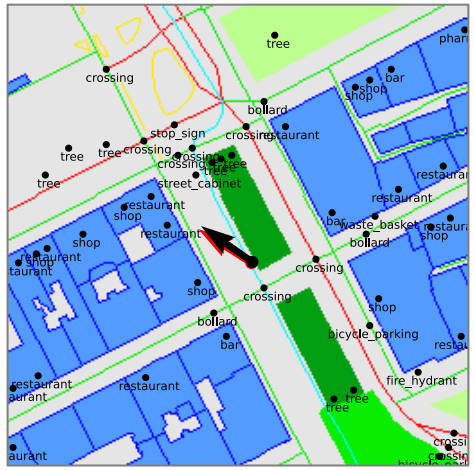
Aria

Mapillary

KITTI





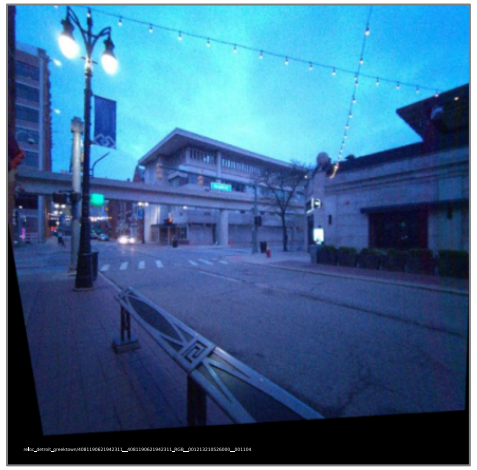


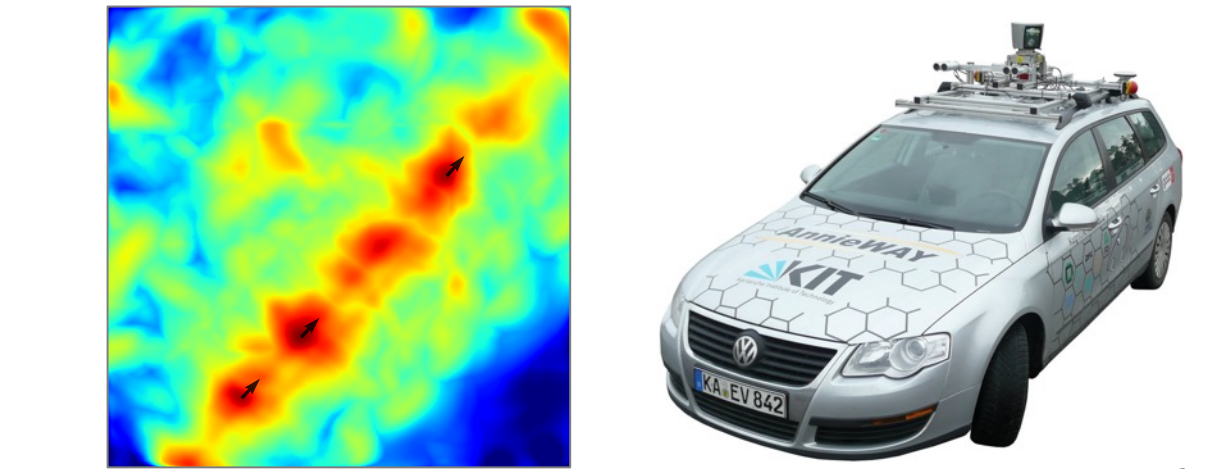
Zero-shot generalization

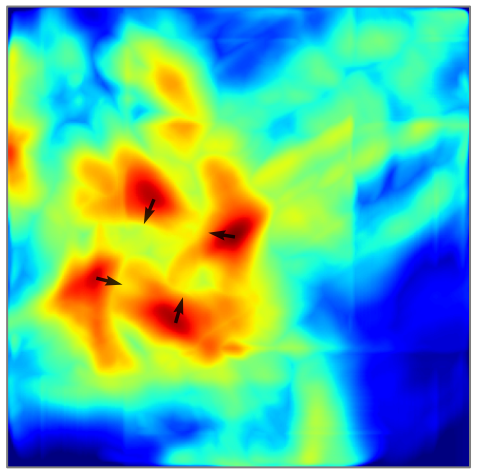
Aria

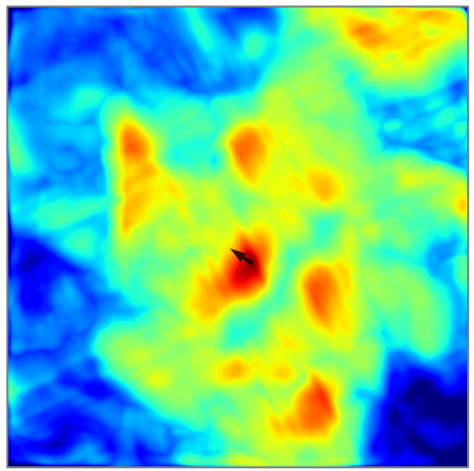
Mapillary

KITTI









In this video

What are current approaches?

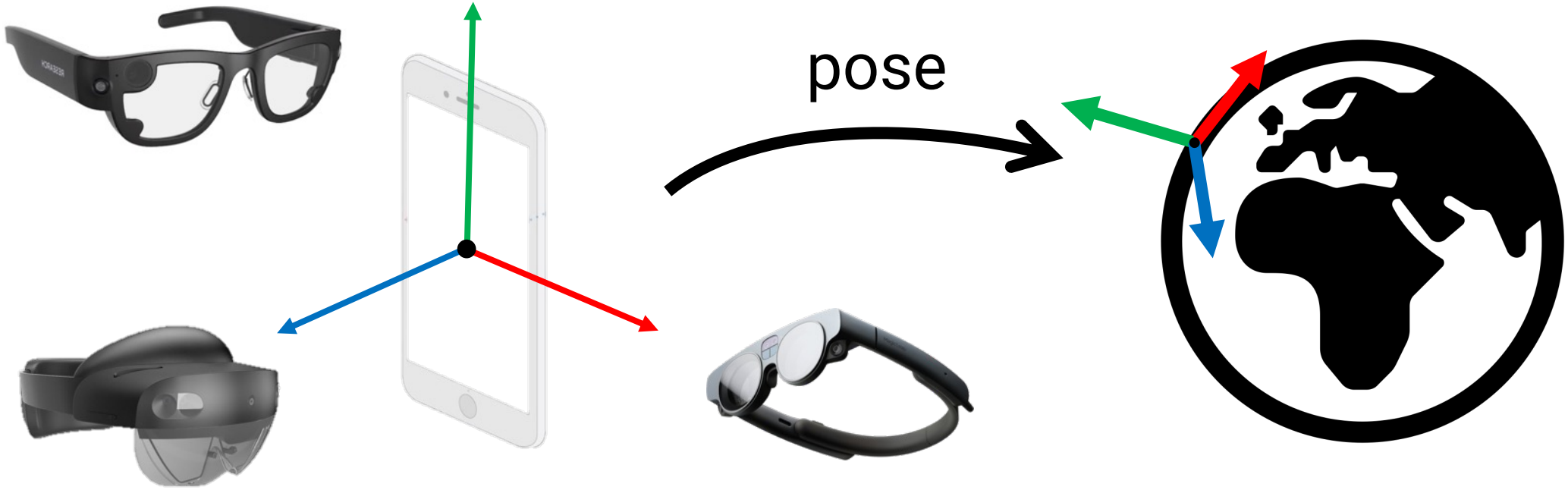
How does OrienterNet work?

How well does it work?

Positioning

Recover the 6-DoF pose of the device

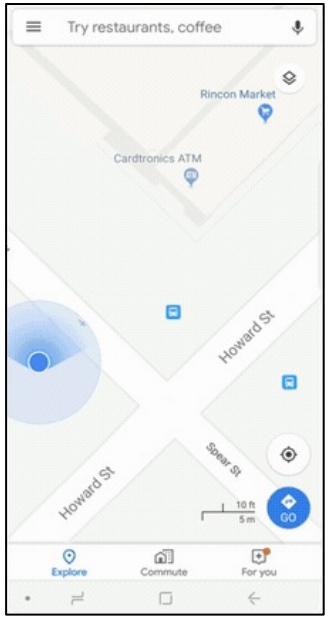
- 3D translation + rotation
- global reference frame

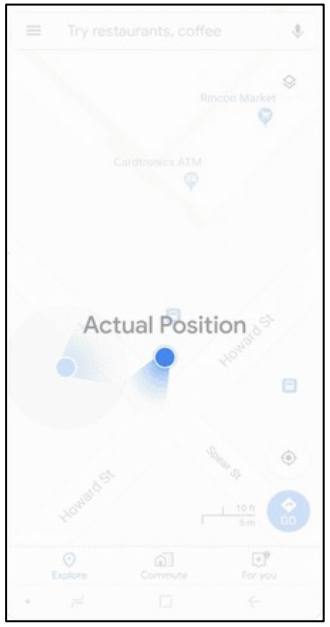


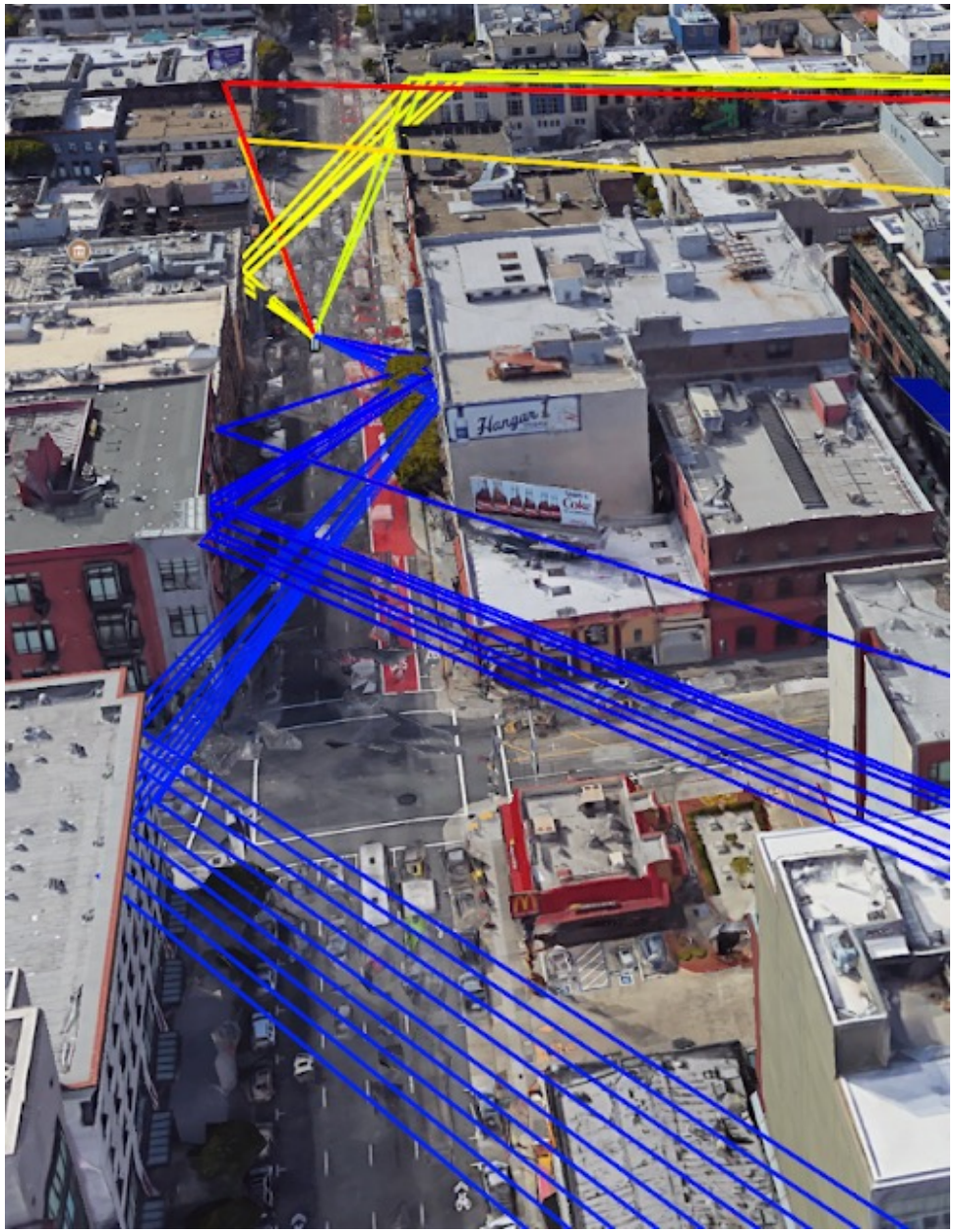
Meta Aria, Microsoft HoloLens, Magic Leap 2

GPS+compass is not enough

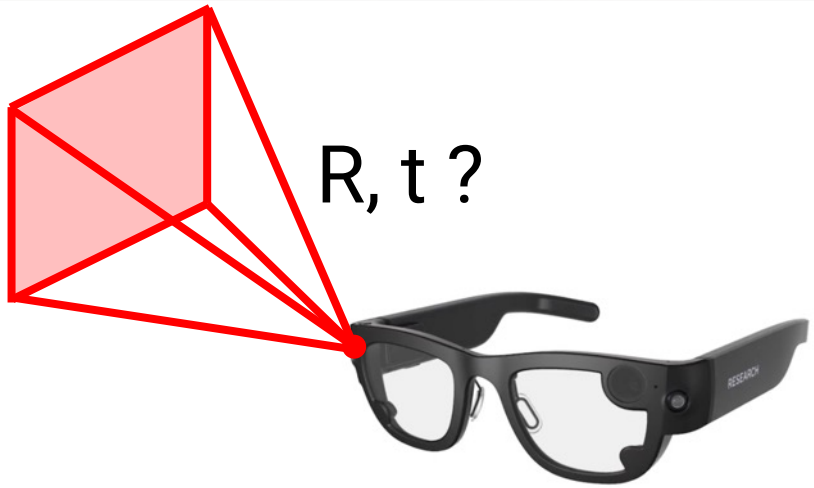
- Low accuracy
- Only 3 DoF
- Commonly unreliable: urban canyon, metal structures



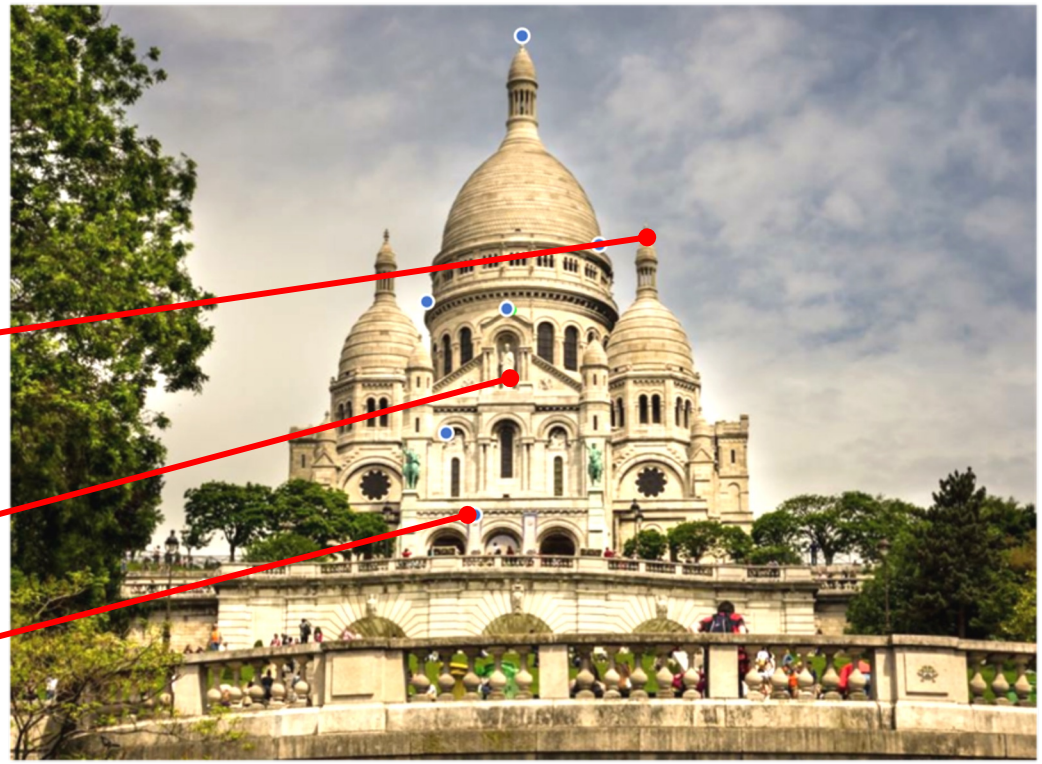


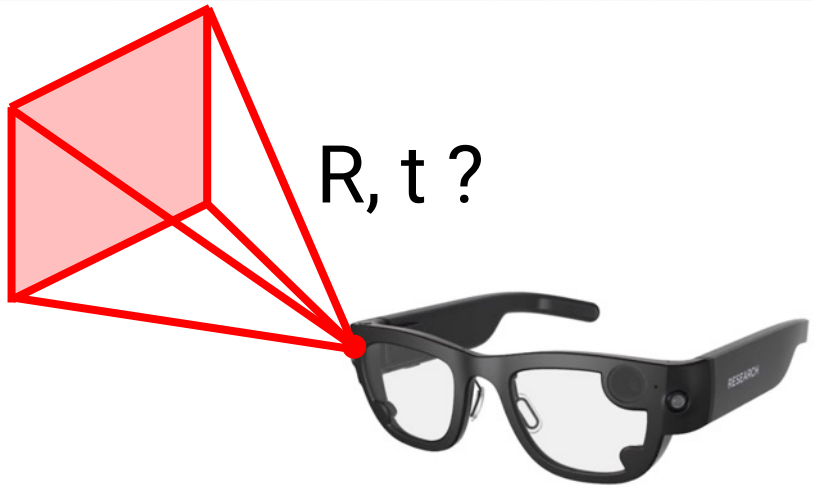


Google Maps

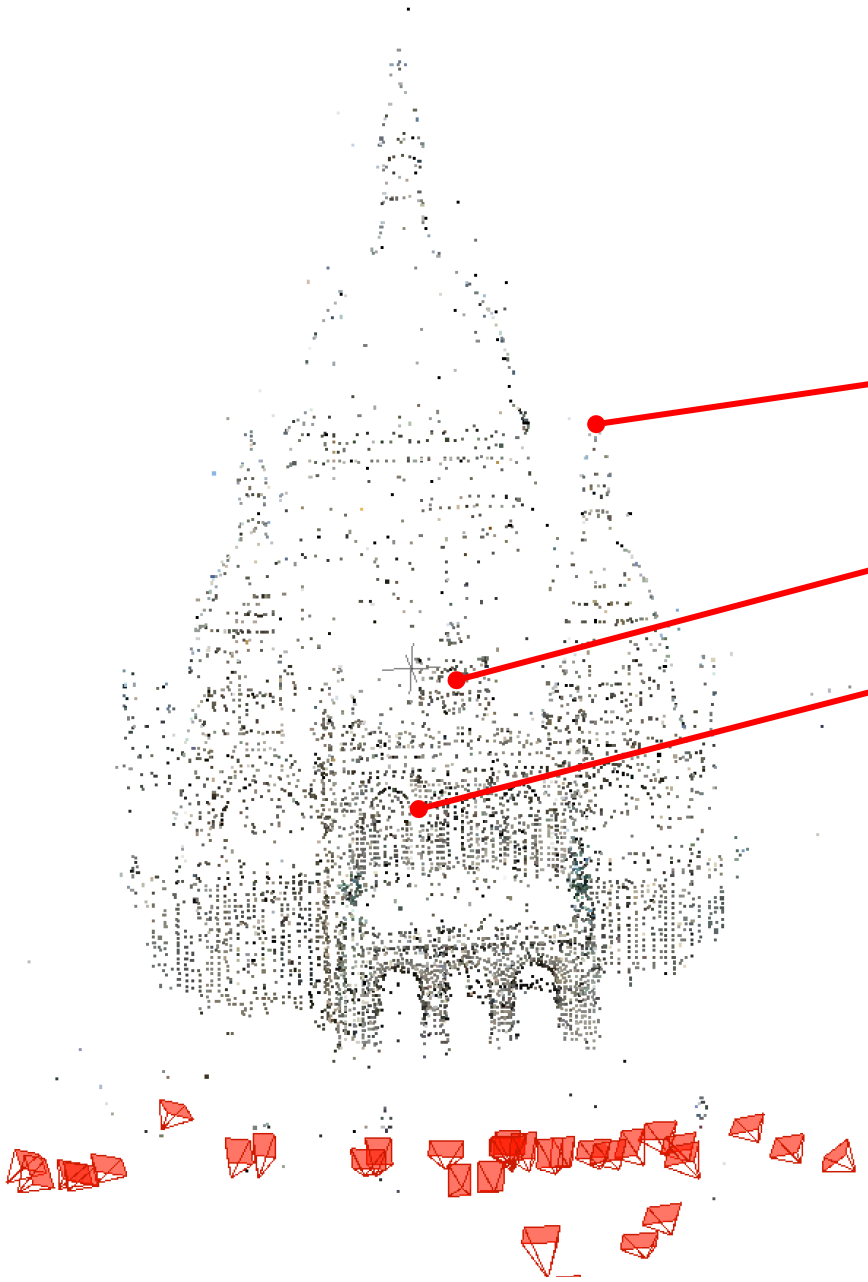


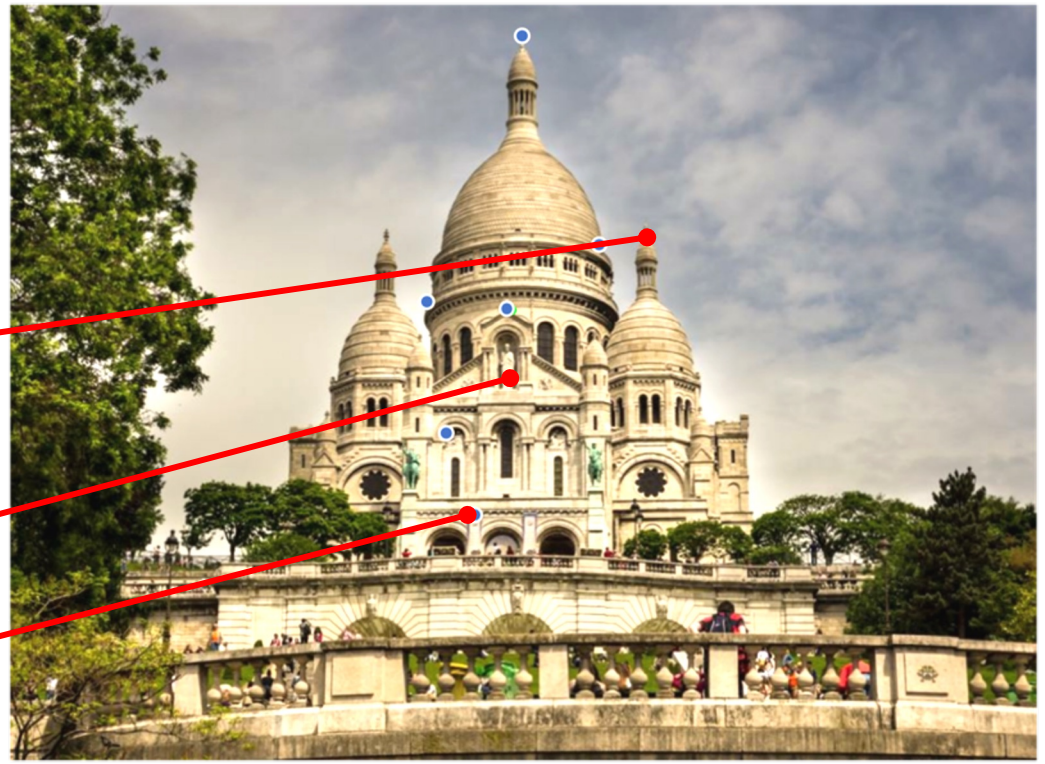
6-DoF Localization

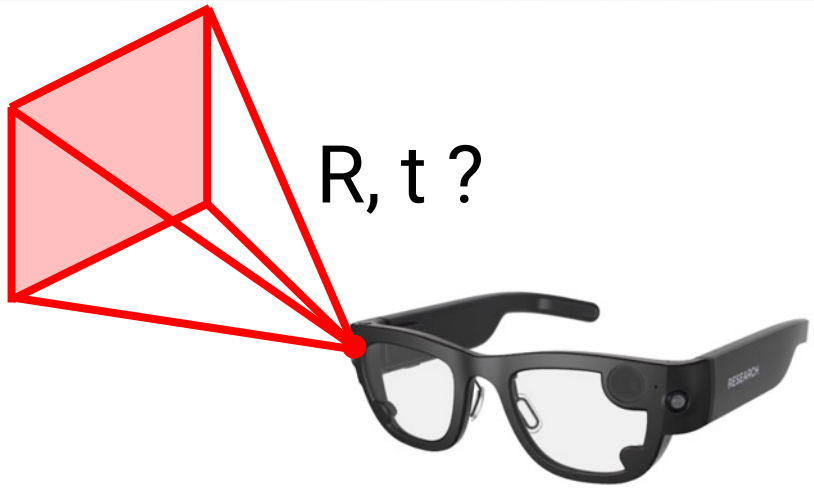




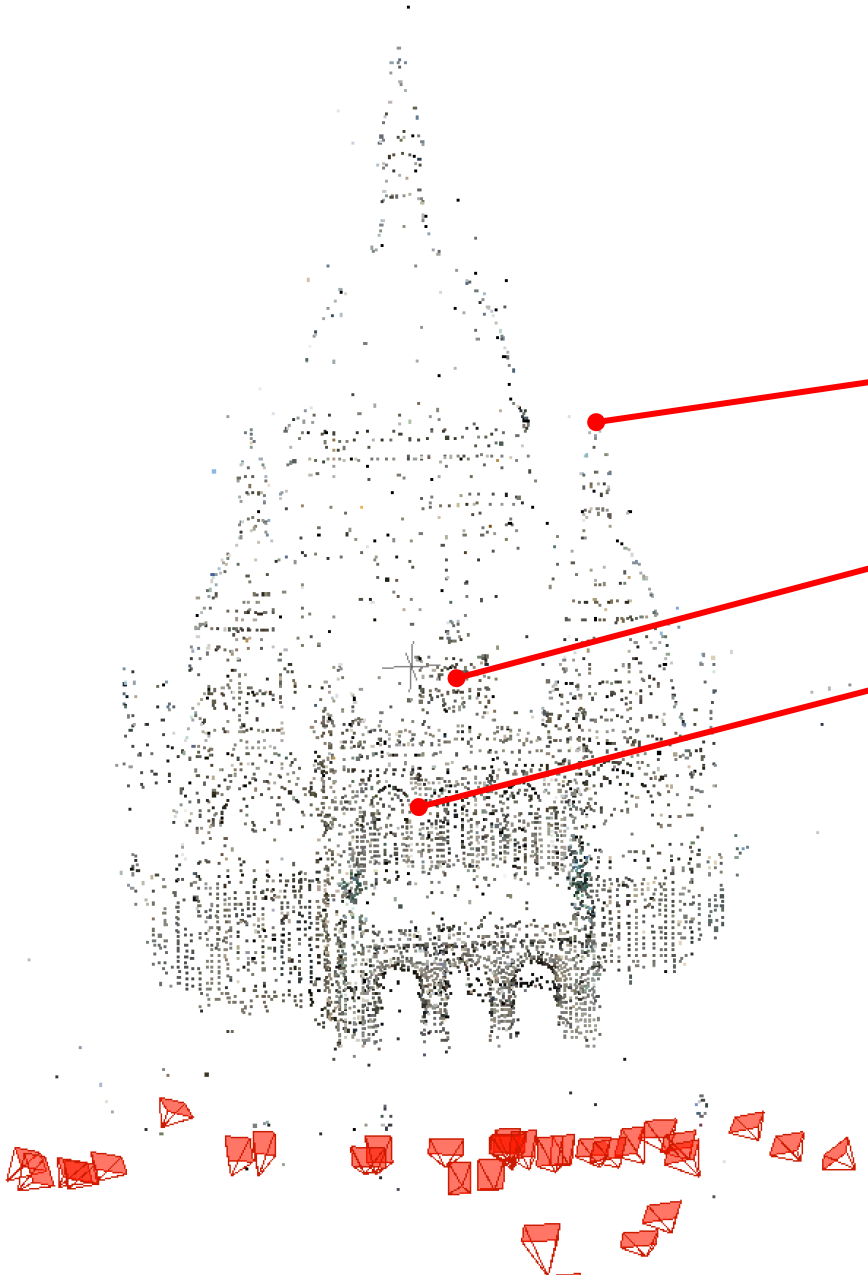
6-DoF Localization

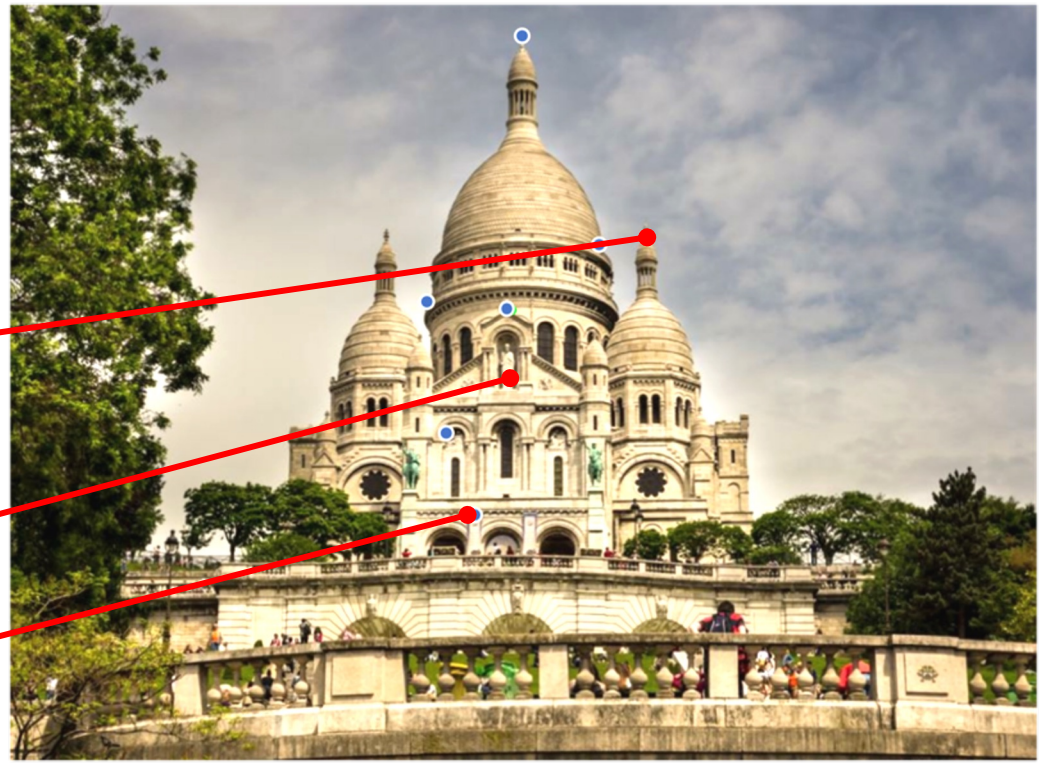


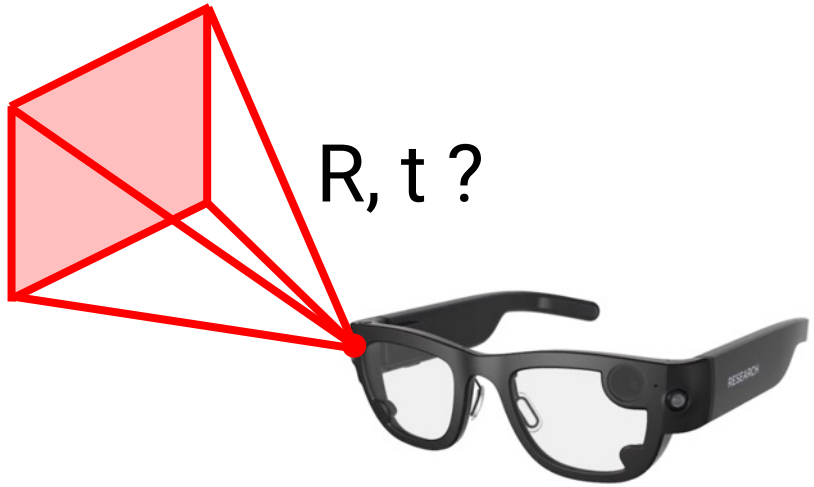




6-DoF Localization

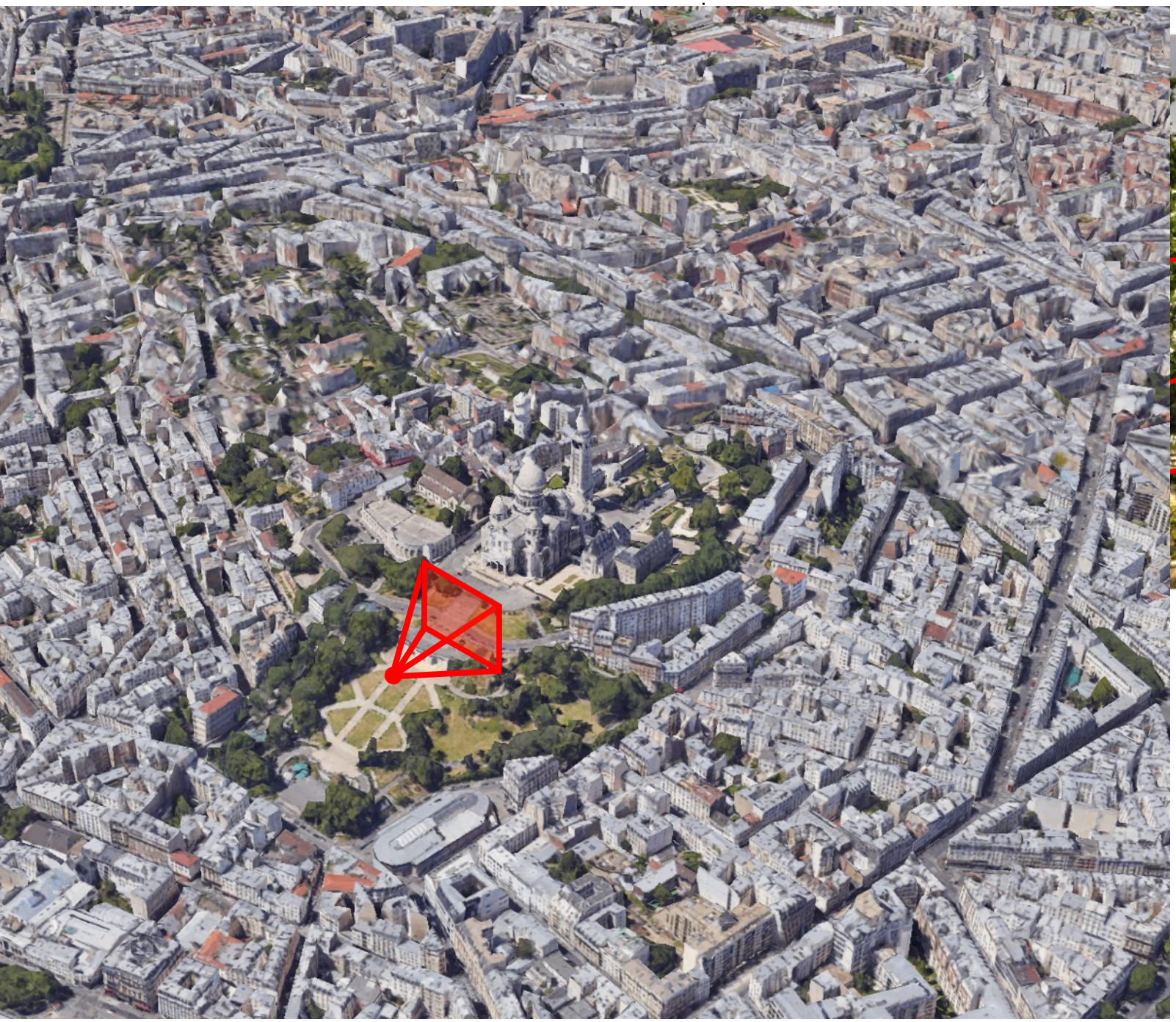


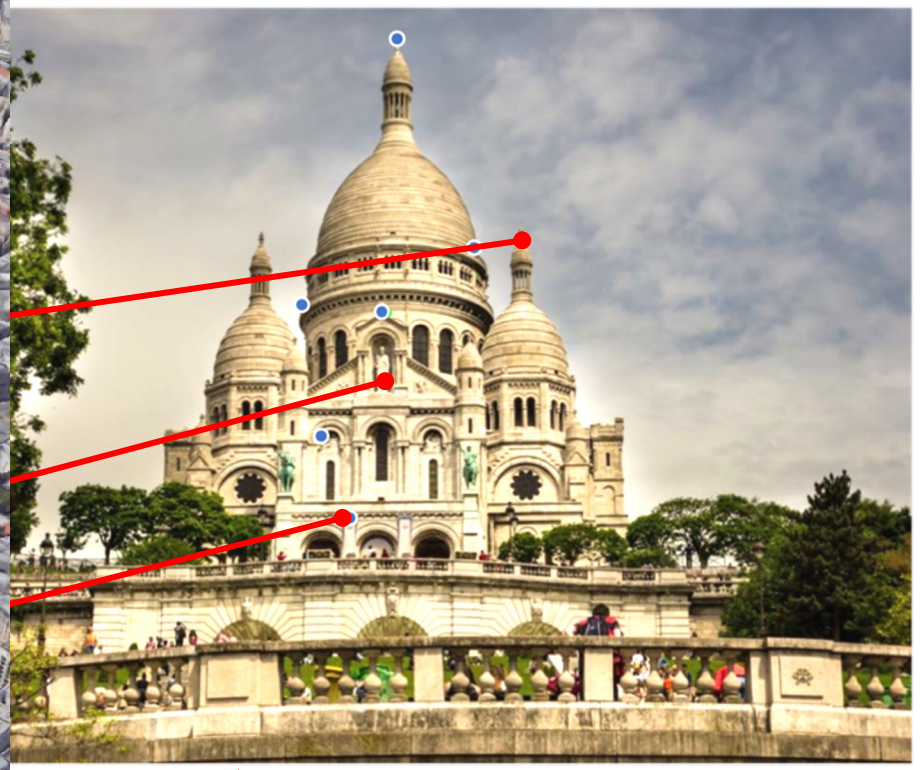


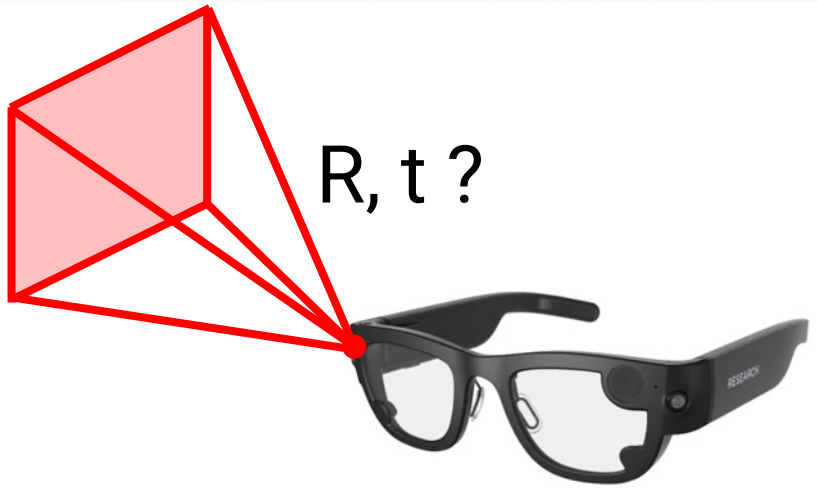


Structure-from-Motion

6-DoF Localization







6-DoF Localization

Limitations of 3D maps

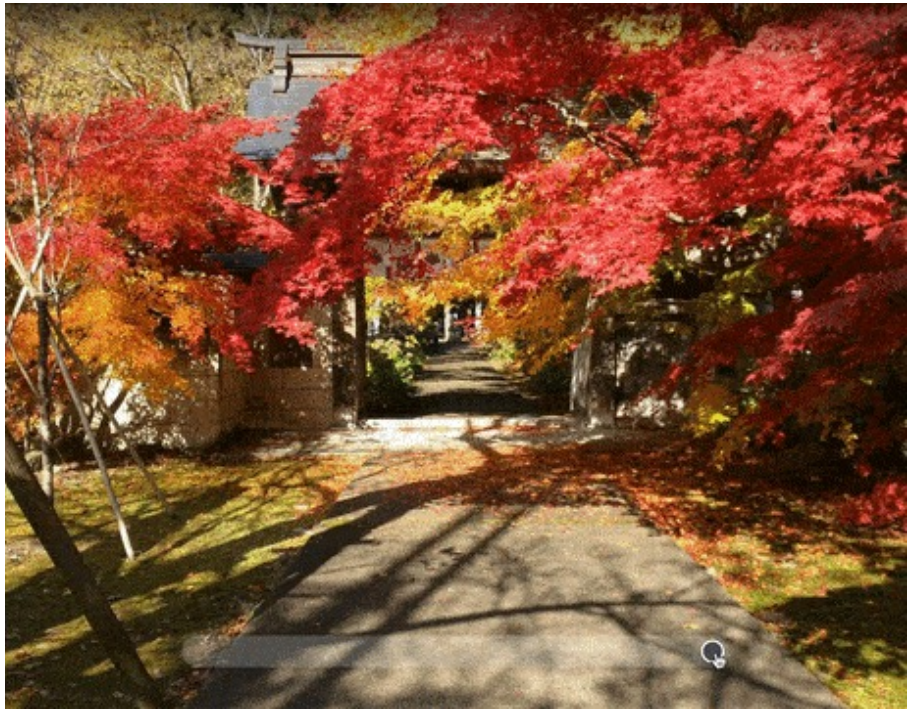
Build & update

Mapping fleet
Frequent updates

Google StreetView

Storage

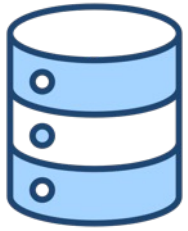
Privacy



Mapillary

Limitations of 3D maps

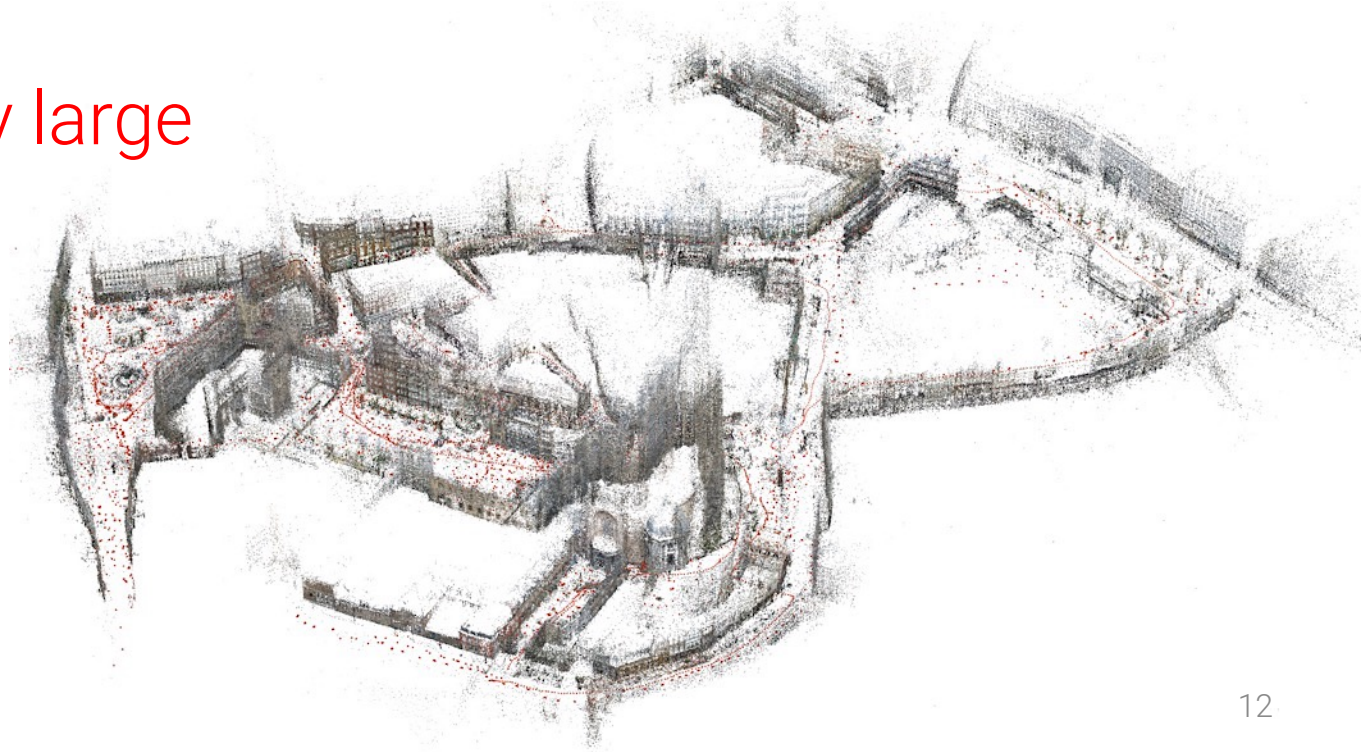
Build
& update



Storage

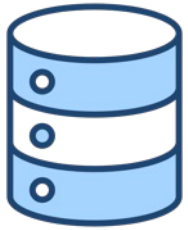
Very large

Privacy



Limitations of 3D maps

Build
& update

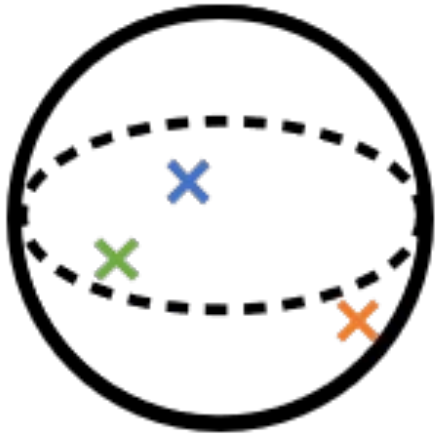


Storage

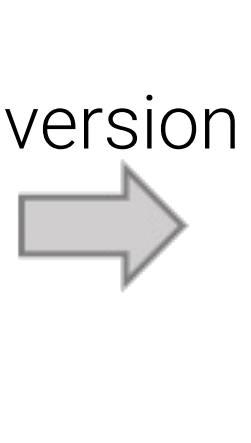
Privacy

keypoints

descriptors



inversion



reconstruction

Mihai Dusmanu

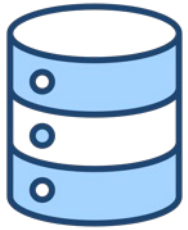
Risk of inversion

Limitations of 3D maps

Build
& update

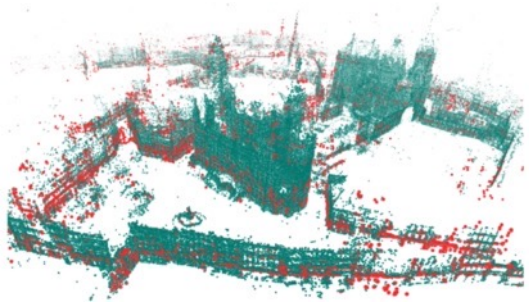
Mapping fleet
Frequent updates

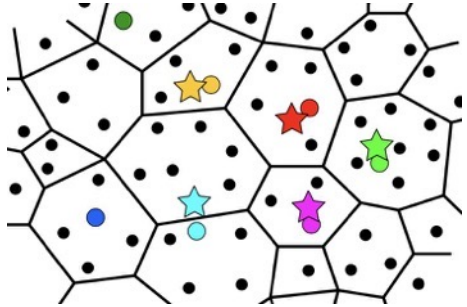
Compression
& Quantization



Storage

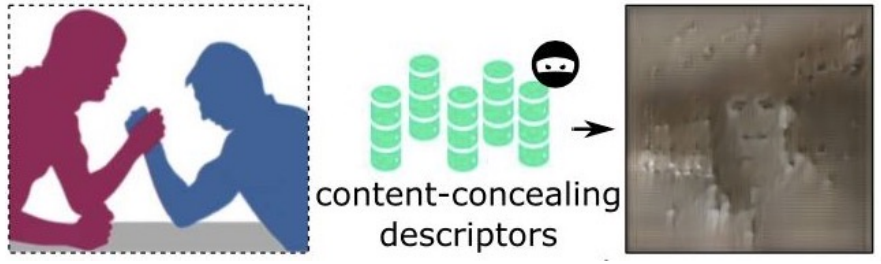
Very large





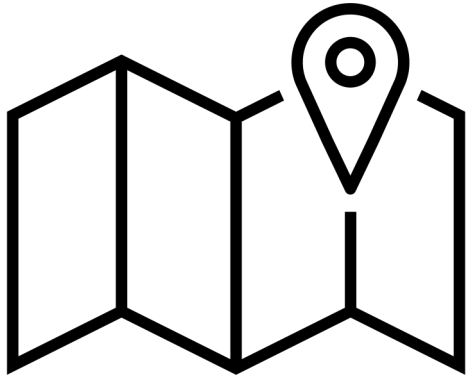
Privacy

Risk of inversion

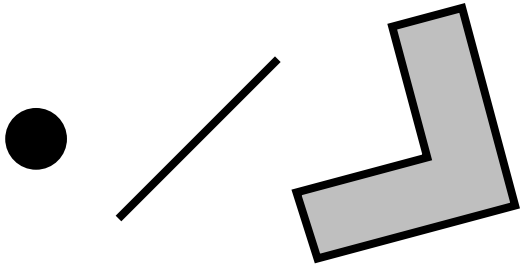


Privacy-preserving descriptors

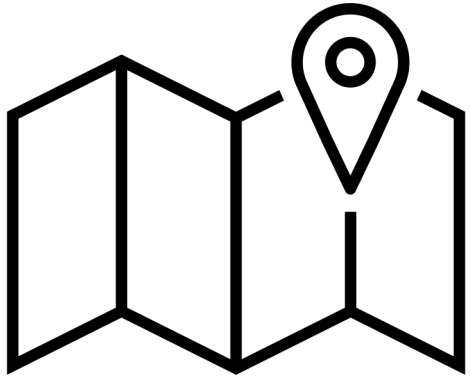
Semantic 2D maps



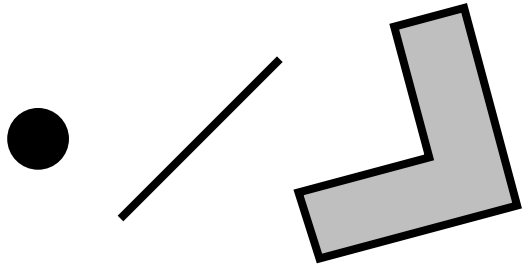
Planimetric



Semantic 2D maps

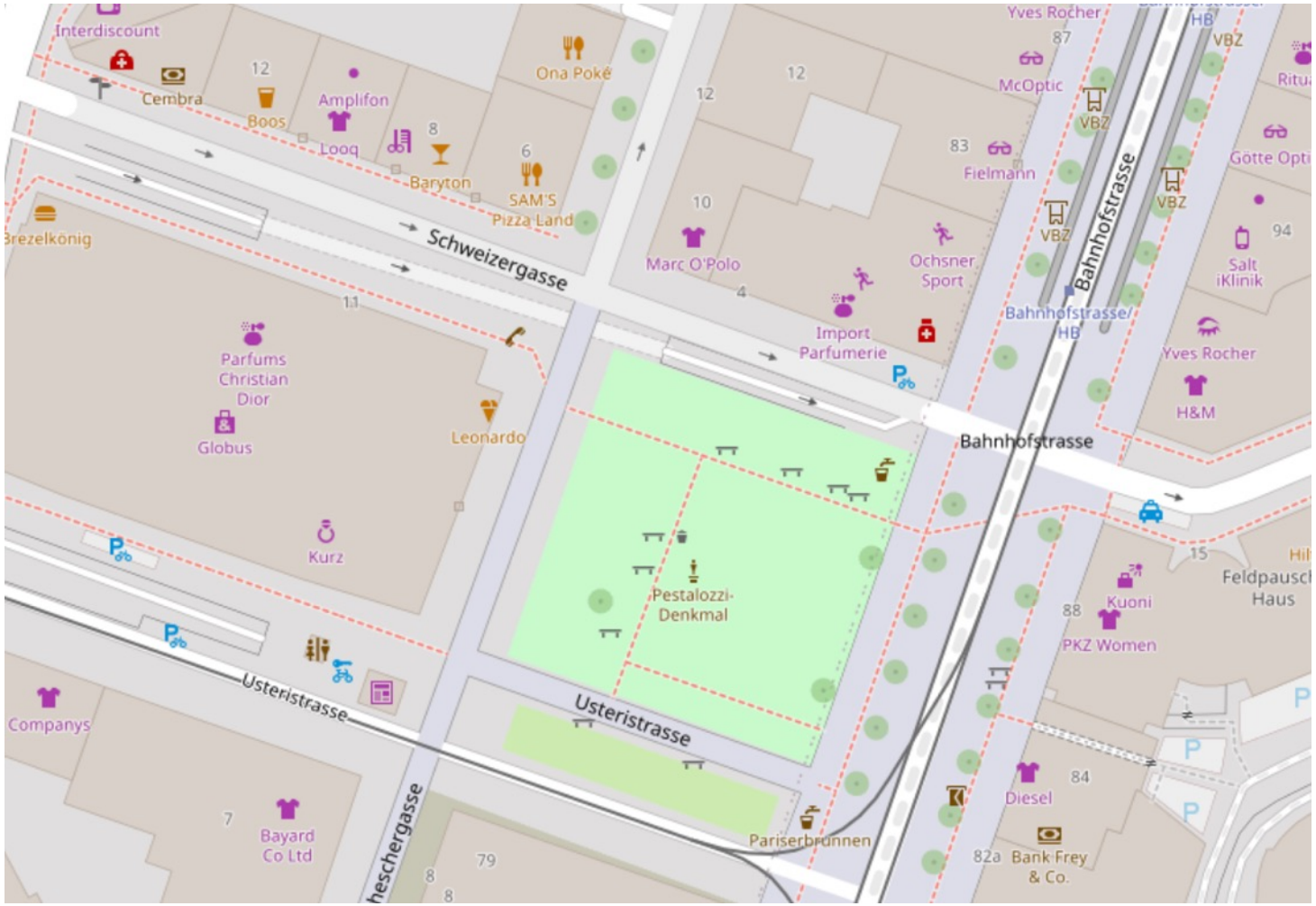


Planimetric

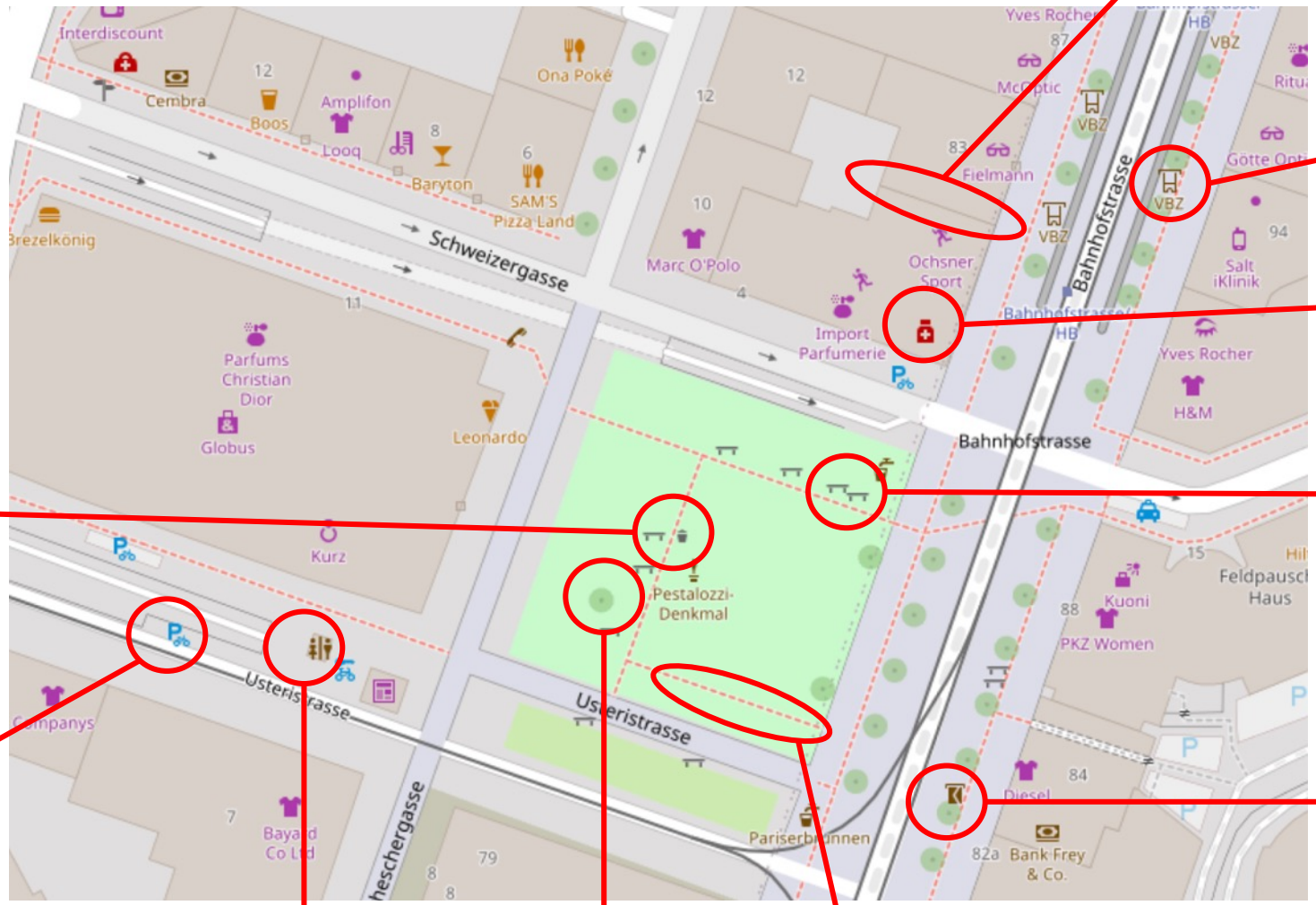


OpenStreetMap

Rich semantics



Rich semantics



building boundary

ticket machine

pharmacy

bench

post box

pedestrian path

tree

restrooms

bike parking

trash can

Rich semantics

building boundary

ticket machine

pharmacy

bench

post box

pedestrian path

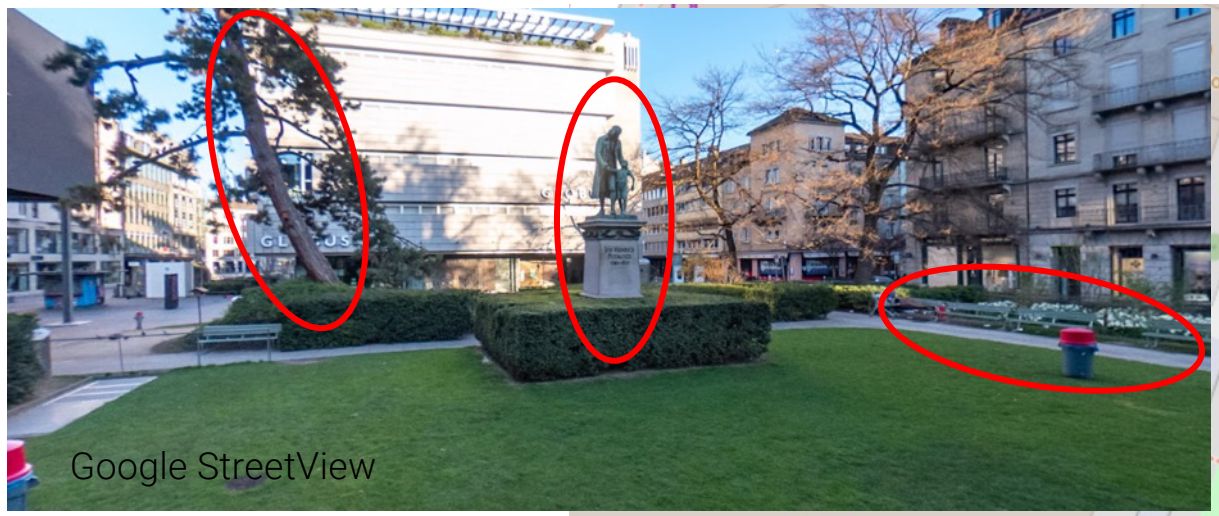
tree

restrooms

bike parking

trash can

Rich semantics



building boundary

ticket machine

pharmacy

bench

post box

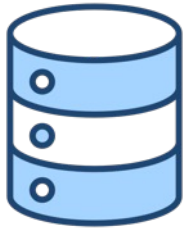
pedestrian path

tree

restrooms

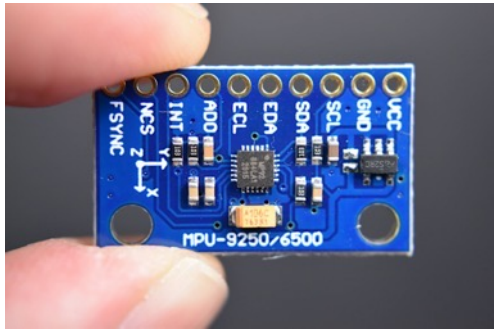
bike parking

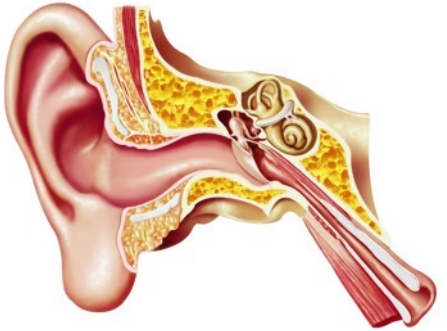
trash can

	3D maps	2D maps
 Build & update	Mapping fleet Frequent updates	Public No appearance updates
 Storage	Very large	Compact Transfer on-device
 Privacy	Risk of inversion	No private info

Simplifying assumptions

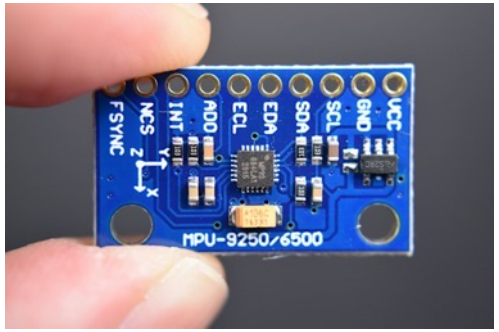
- Known gravity direction

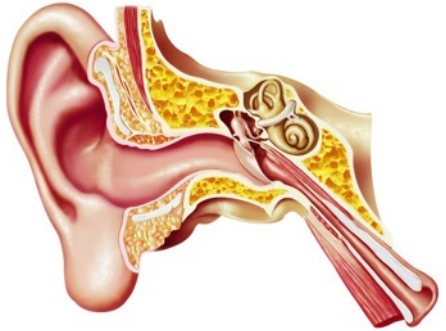




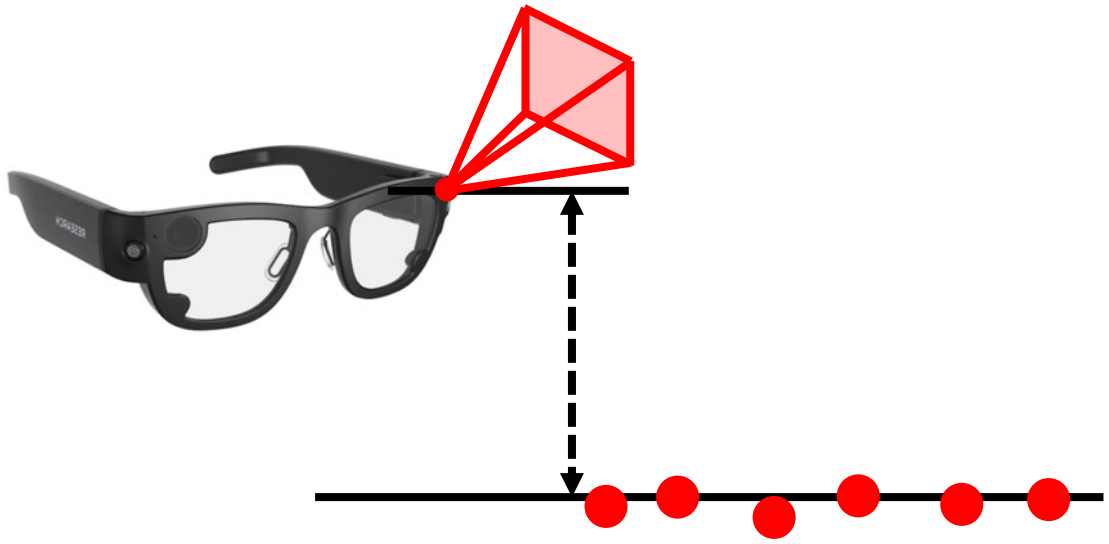
Simplifying assumptions

- Known gravity direction



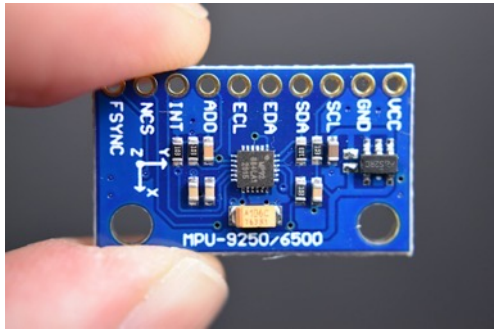


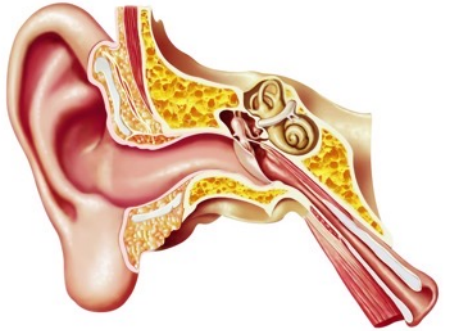
- Unnecessary vertical position



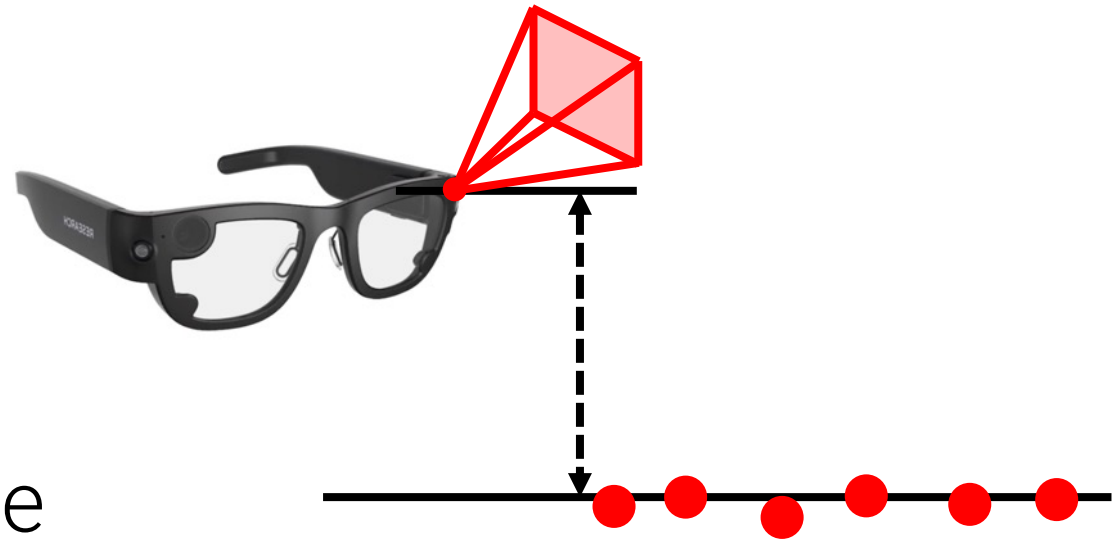
Simplifying assumptions

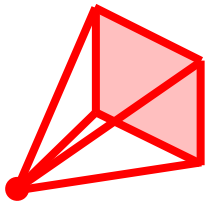
- Known gravity direction

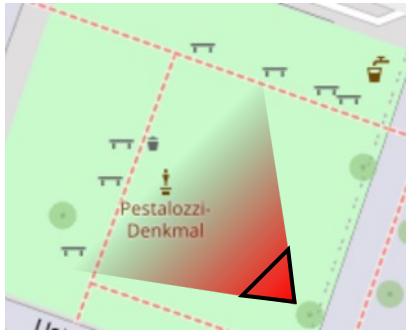




- Unnecessary vertical position





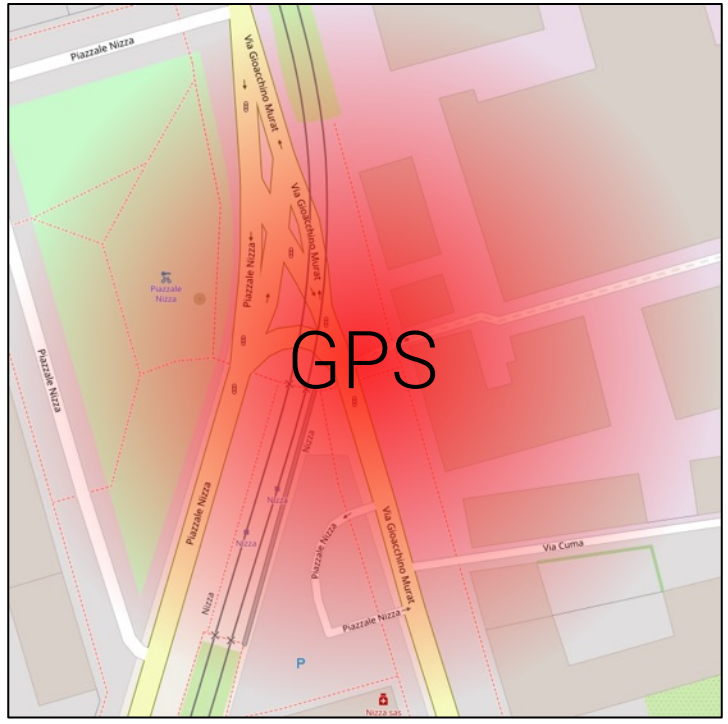


3-DoF pose
(x, y, θ)

Problem setup

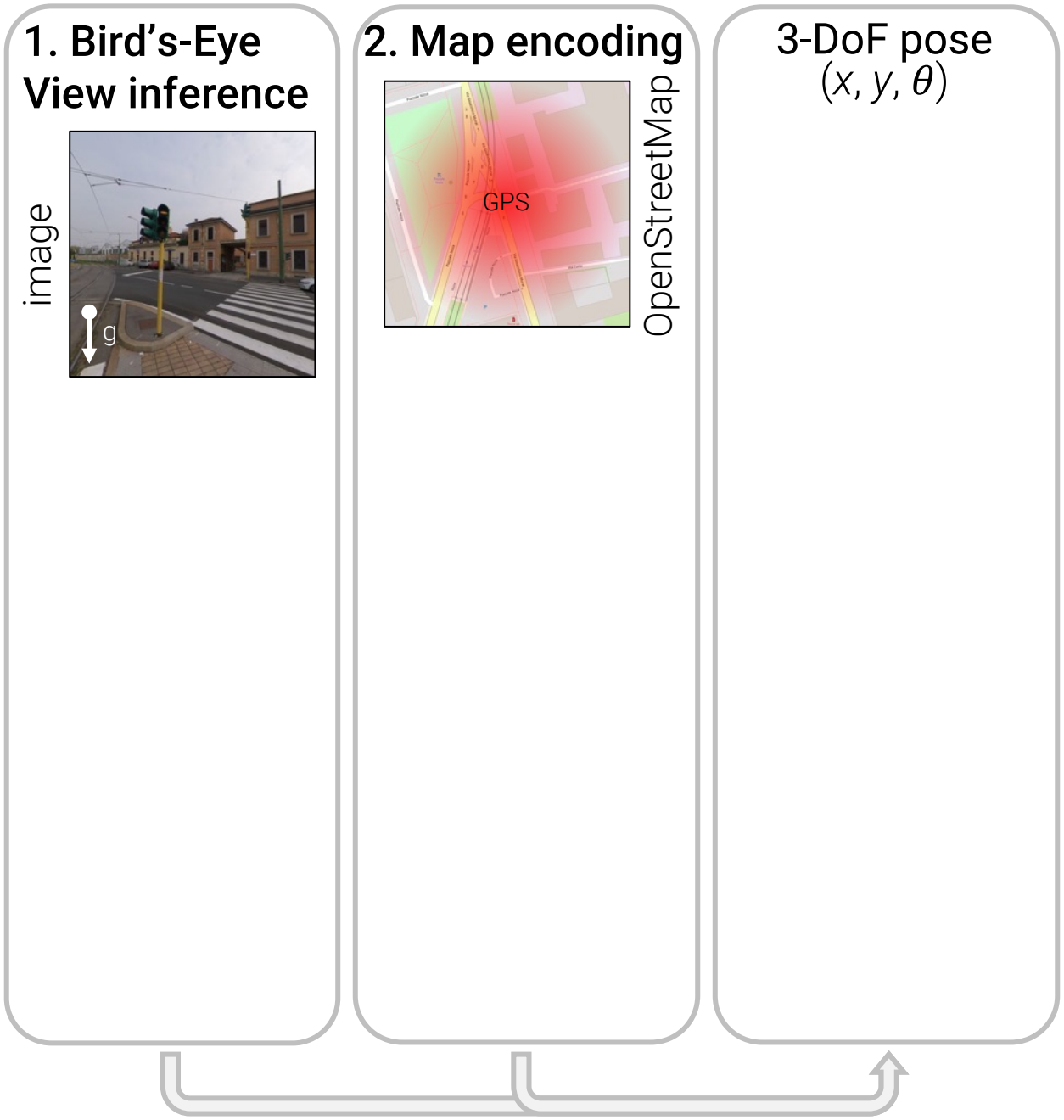
image
+ gravity

128m x 128m

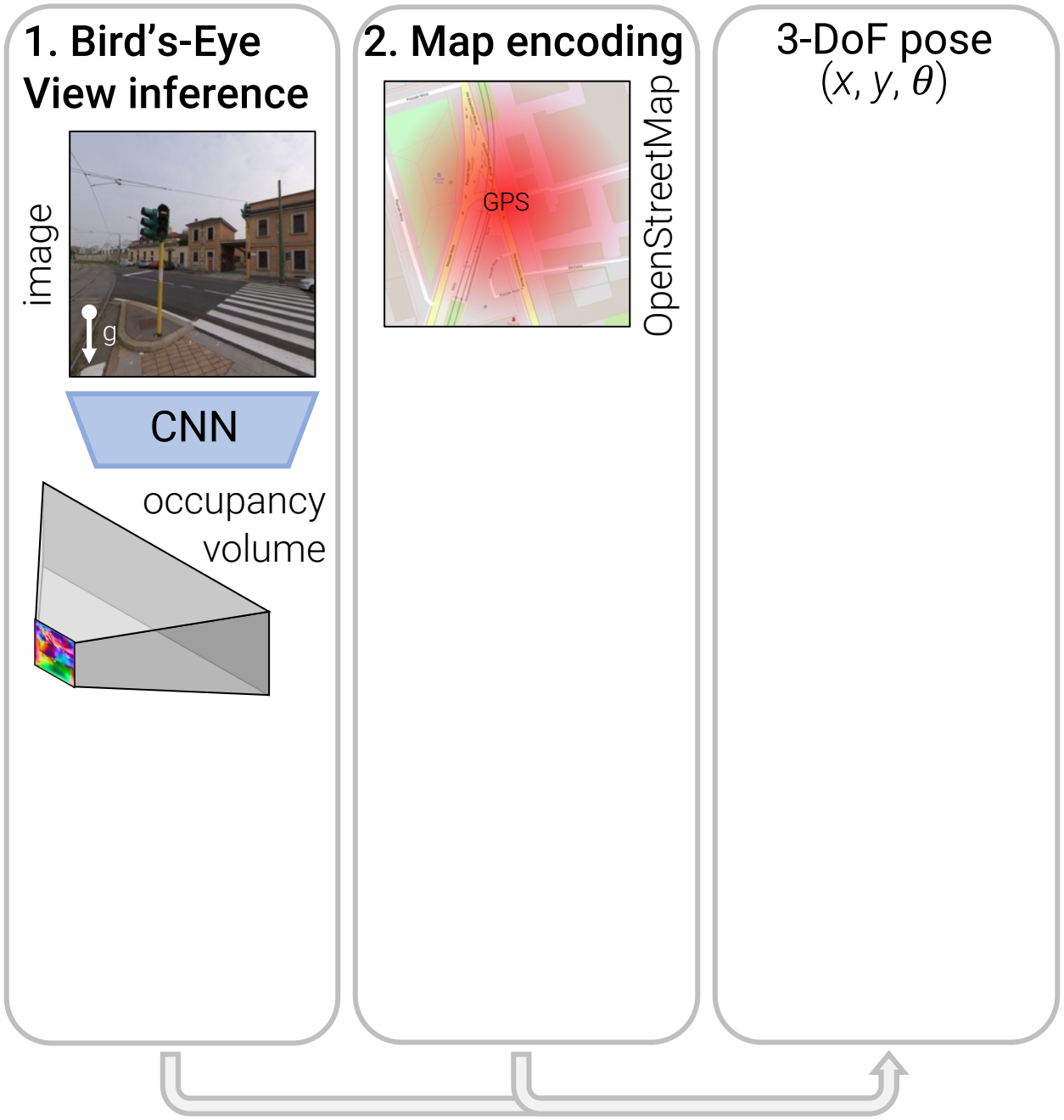


OpenStreetMap

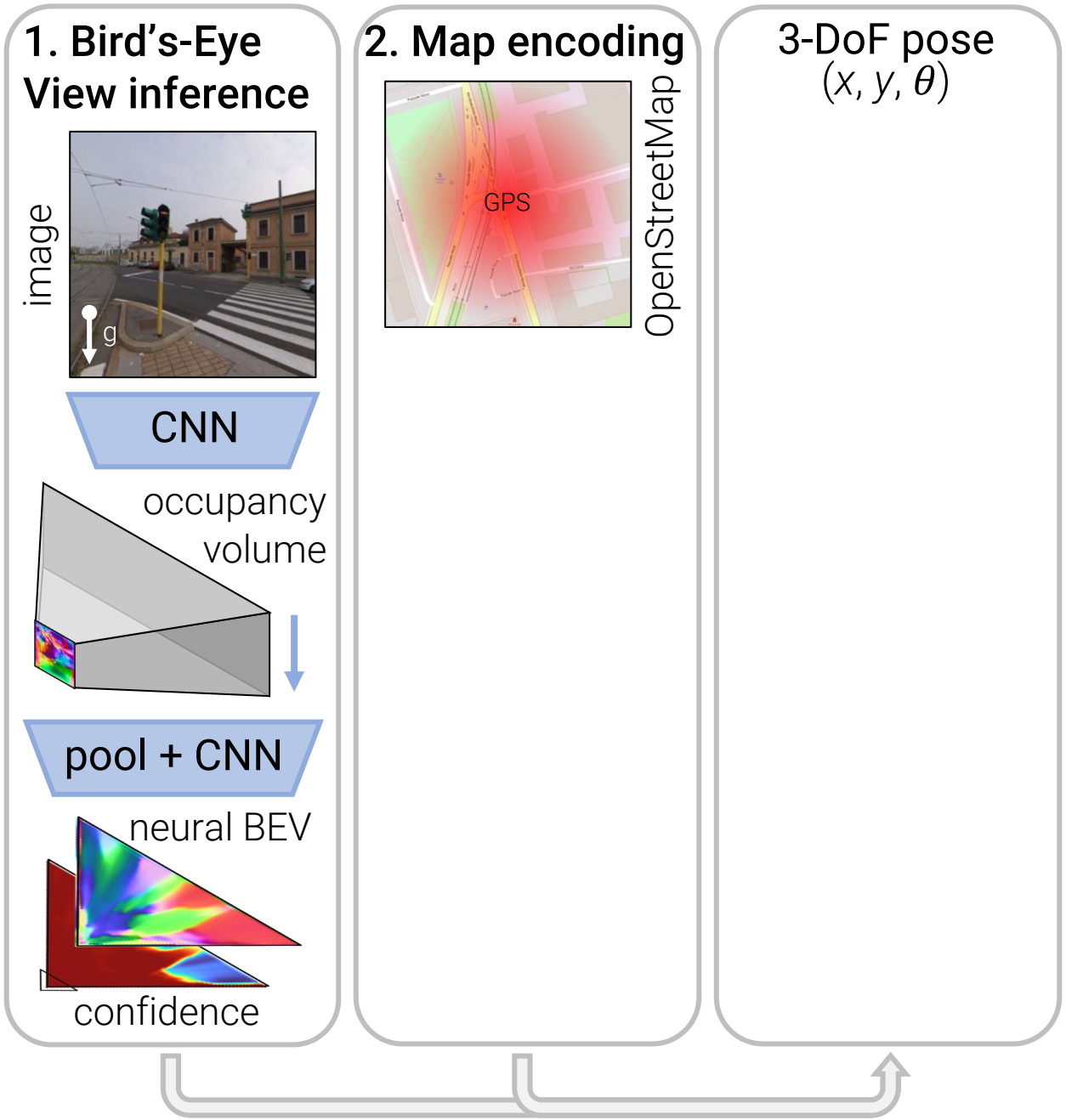
The OrienterNet architecture



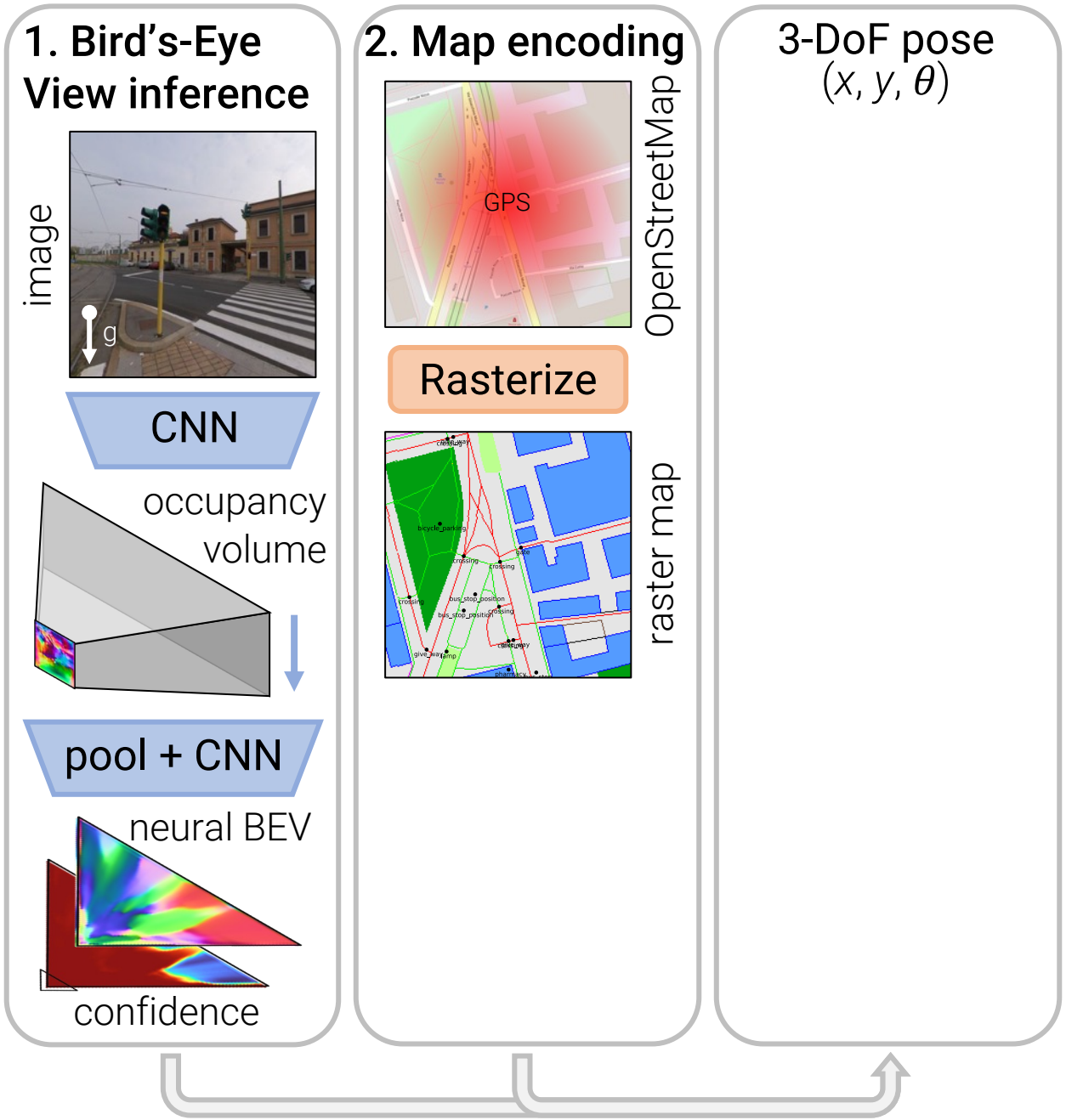
The OrienterNet architecture



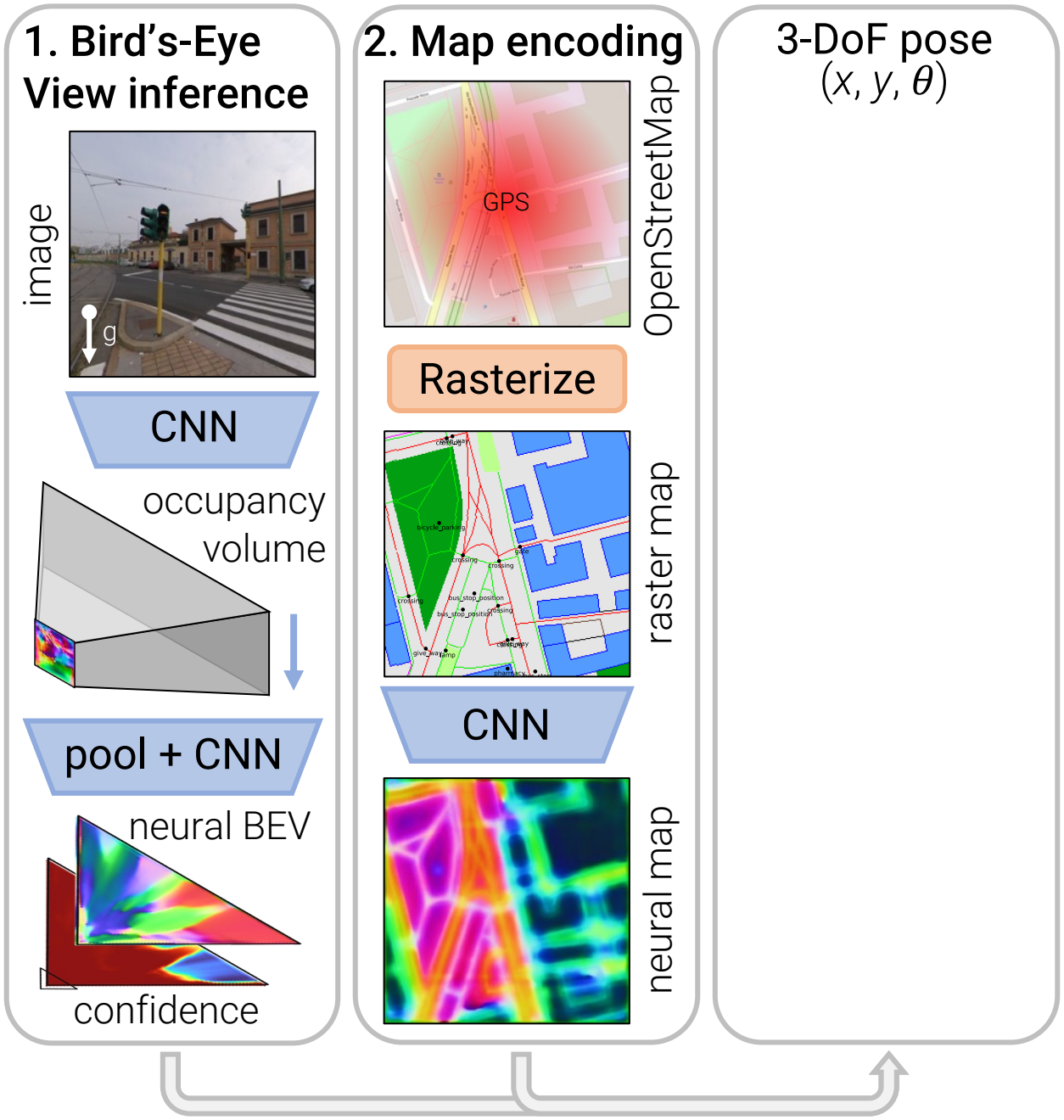
The OrienterNet architecture



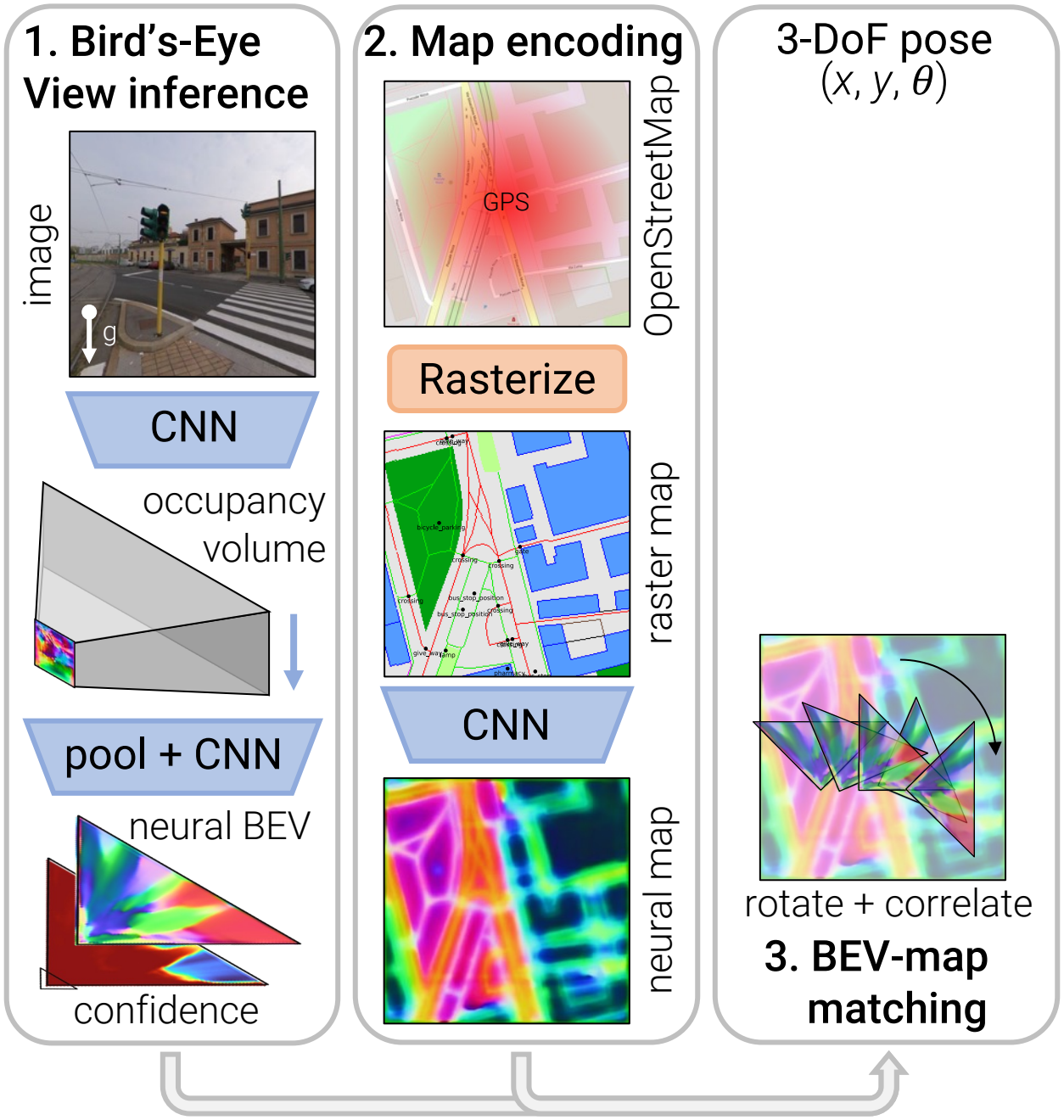
The OrienterNet architecture



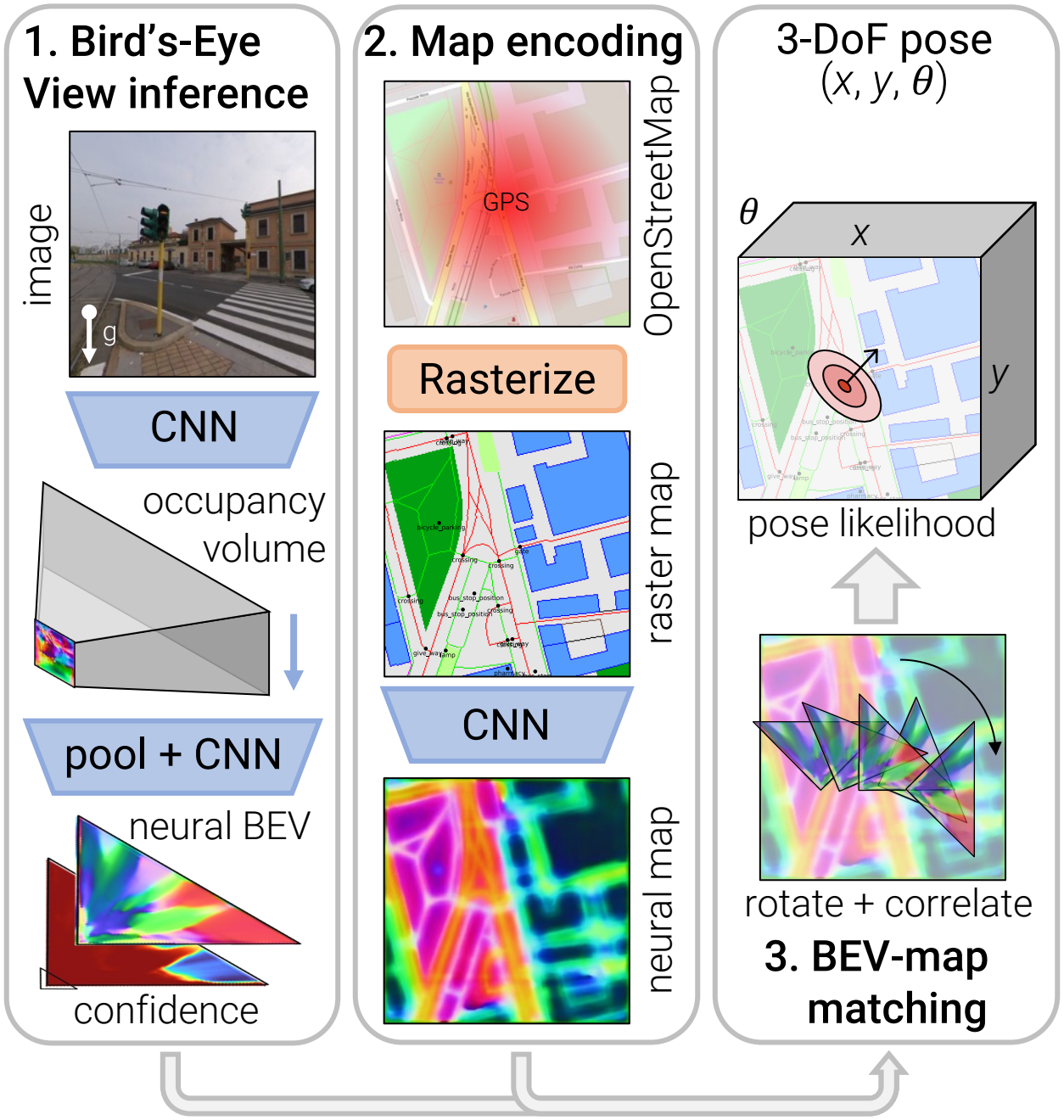
The OrienterNet architecture



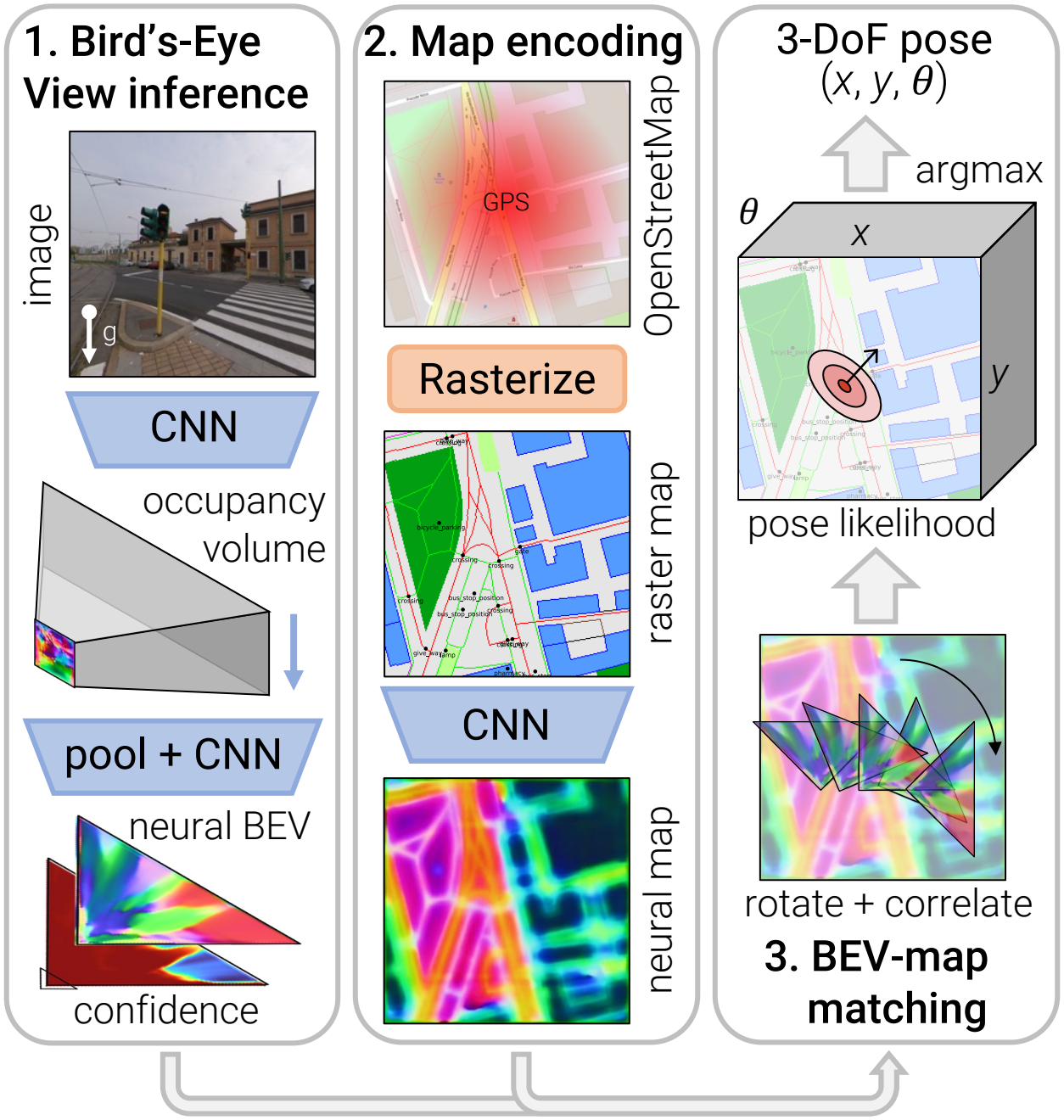
The OrienterNet architecture



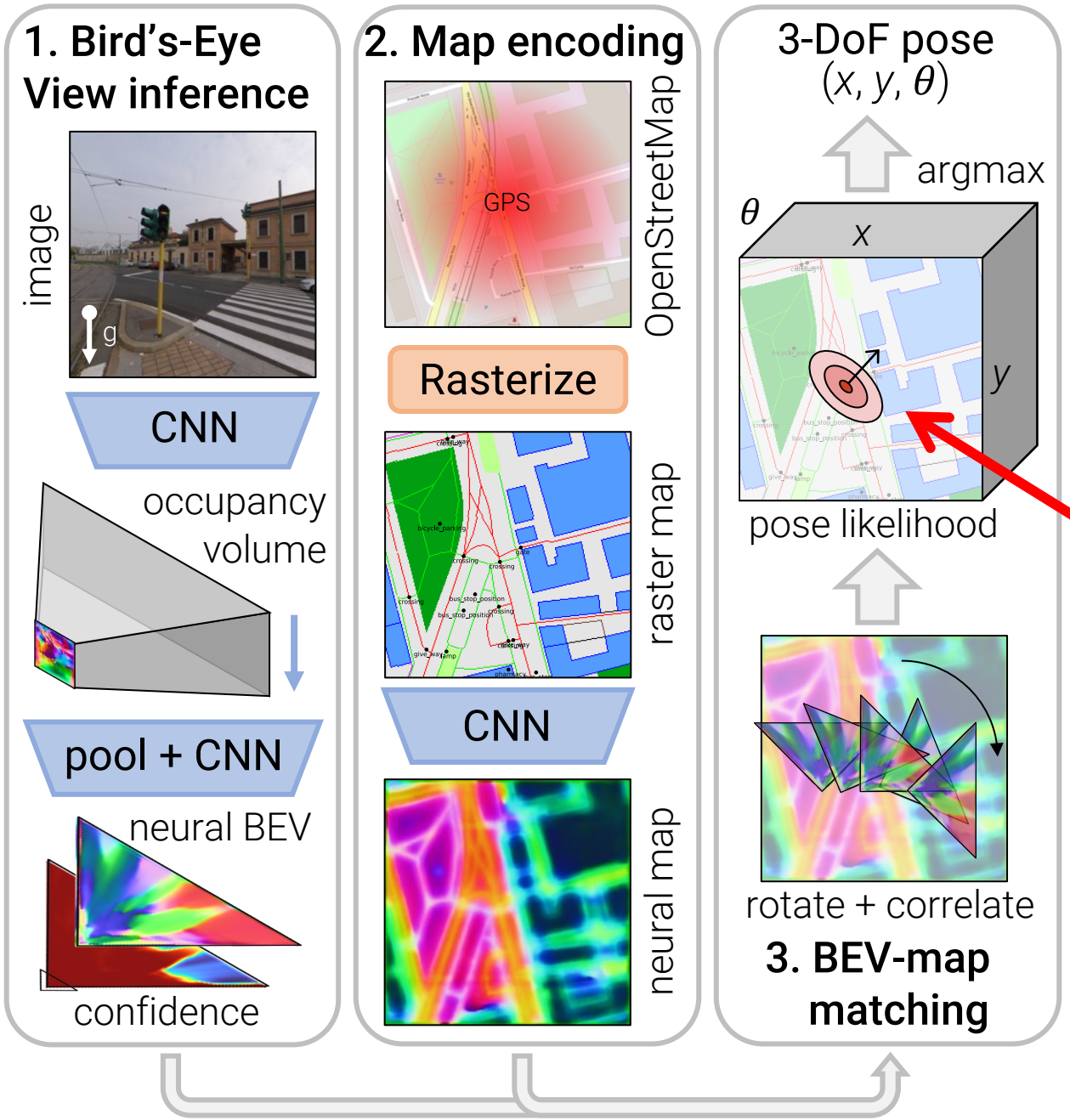
The OrienterNet architecture



The OrienterNet architecture



The OrienterNet architecture



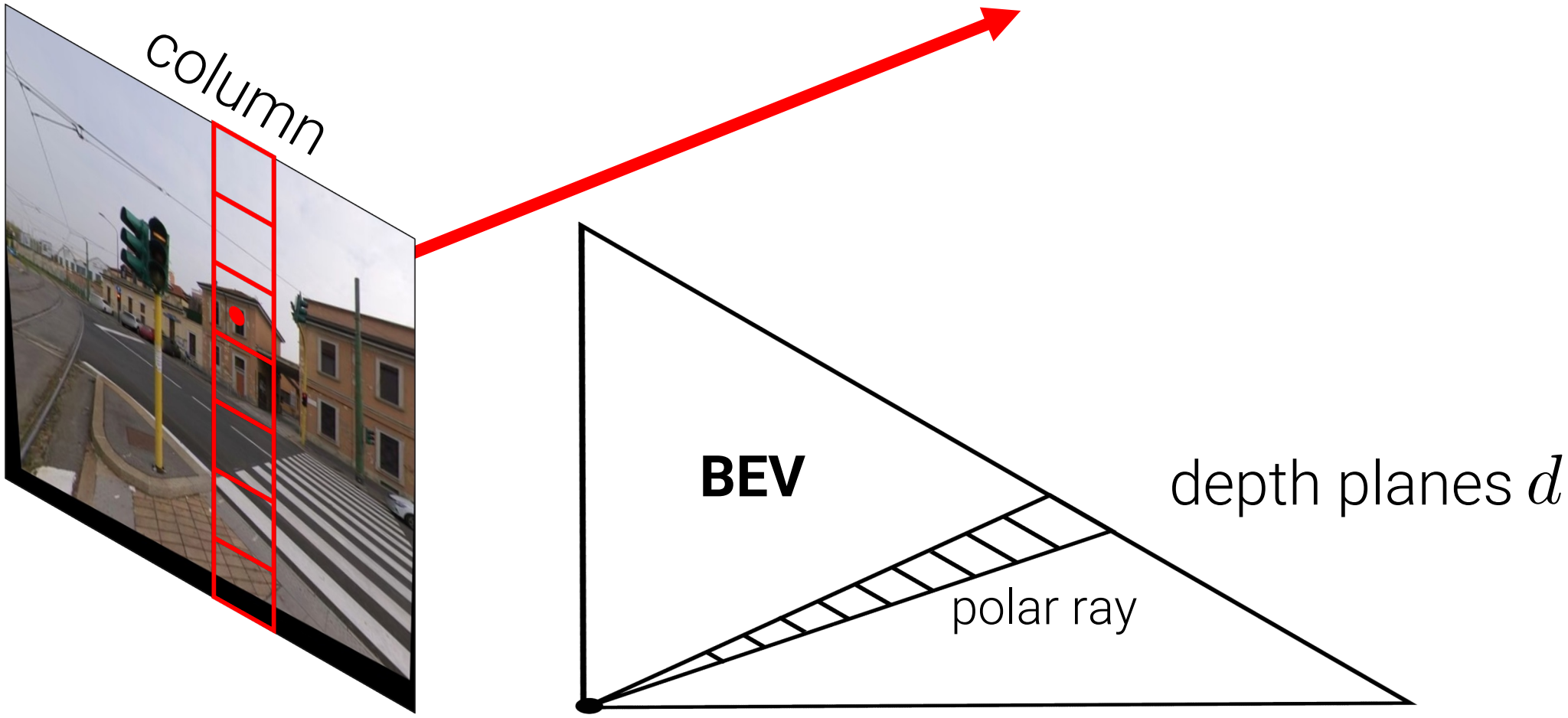
Maximize likelihood of GT pose

1. Bird’s Eye View inference

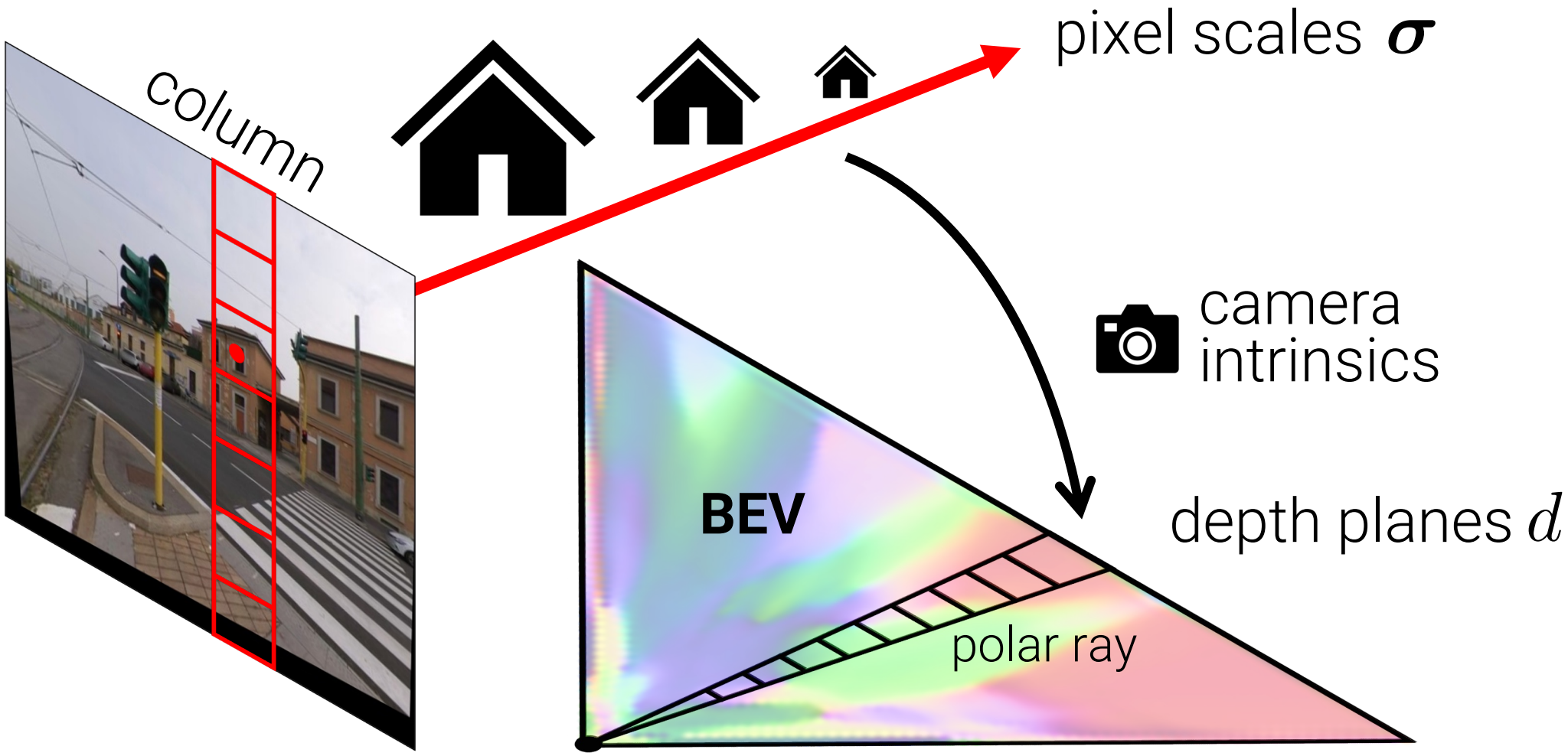
1. Bird’s Eye View inference

gravity-aligned

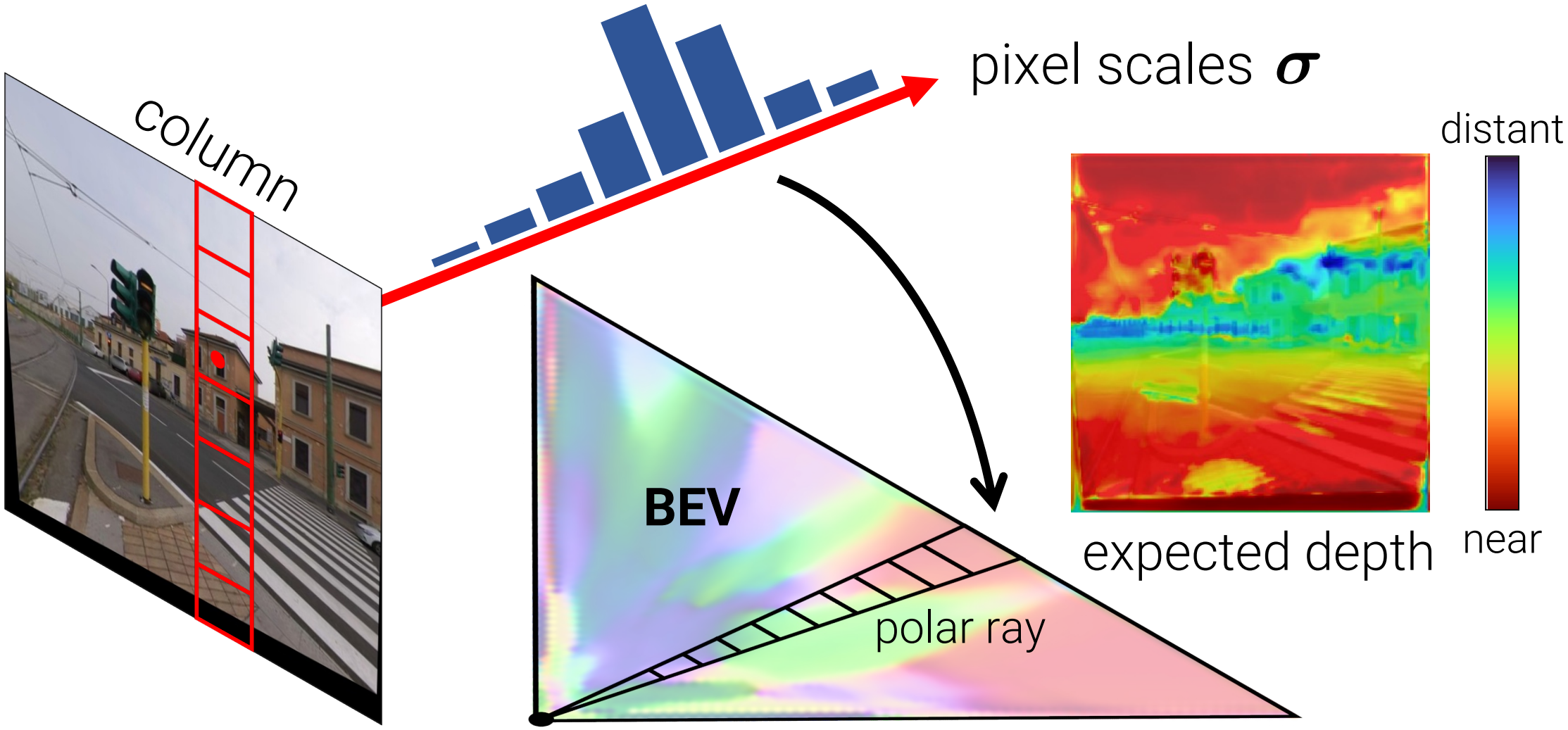
1. Bird’s Eye View inference



1. Bird's Eye View inference



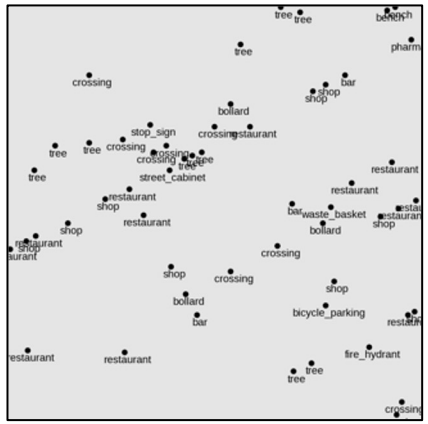
1. Bird's Eye View inference



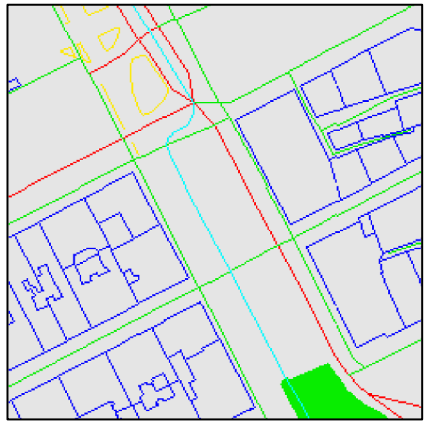
2. Map encoding

OpenStreetMap
vector elements

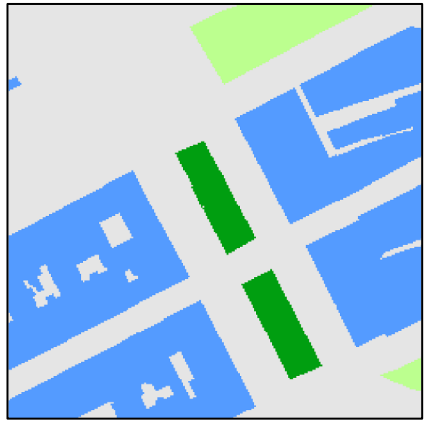
rasterize



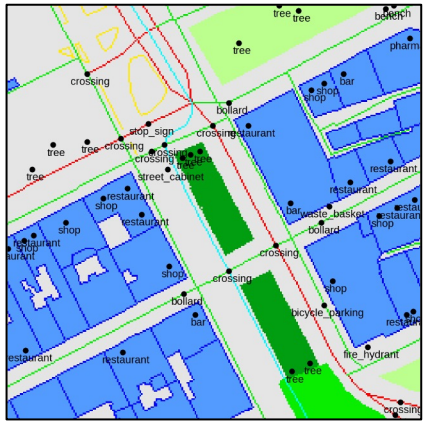
nodes

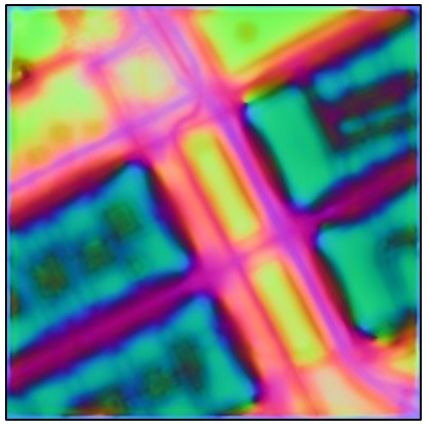


lines



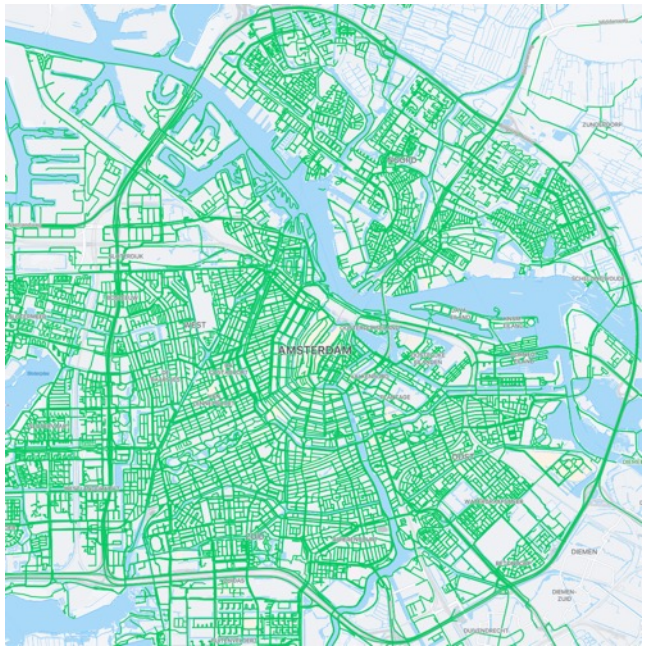
areas

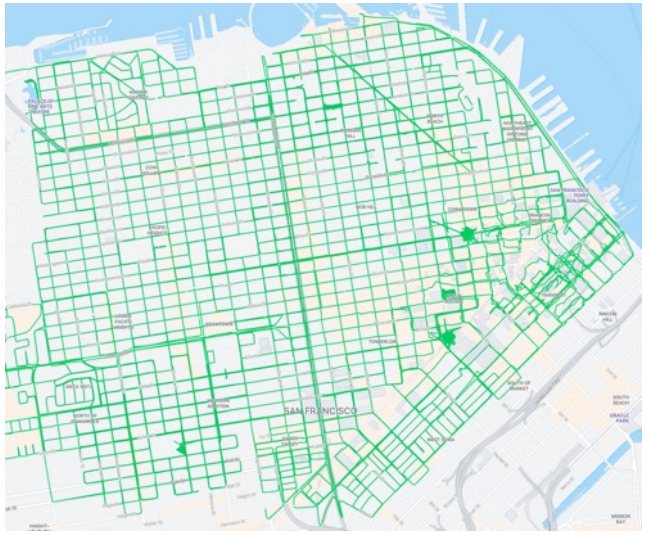




Training a single strong model

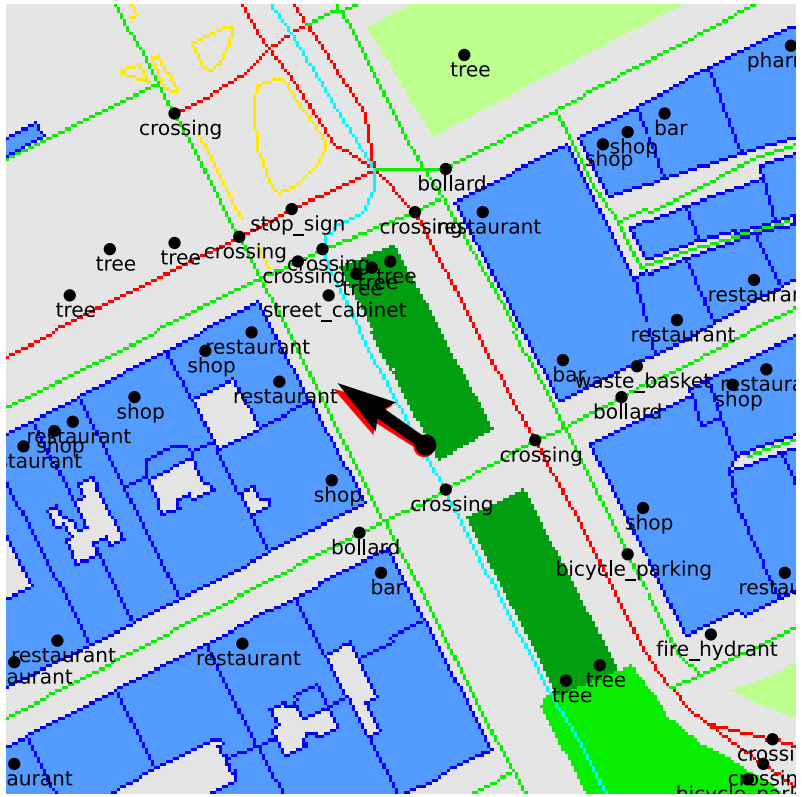
- Publicly-available data from Mapillary
- 760k images from 12 cities across Europe & US
- Hand-held, car, bike



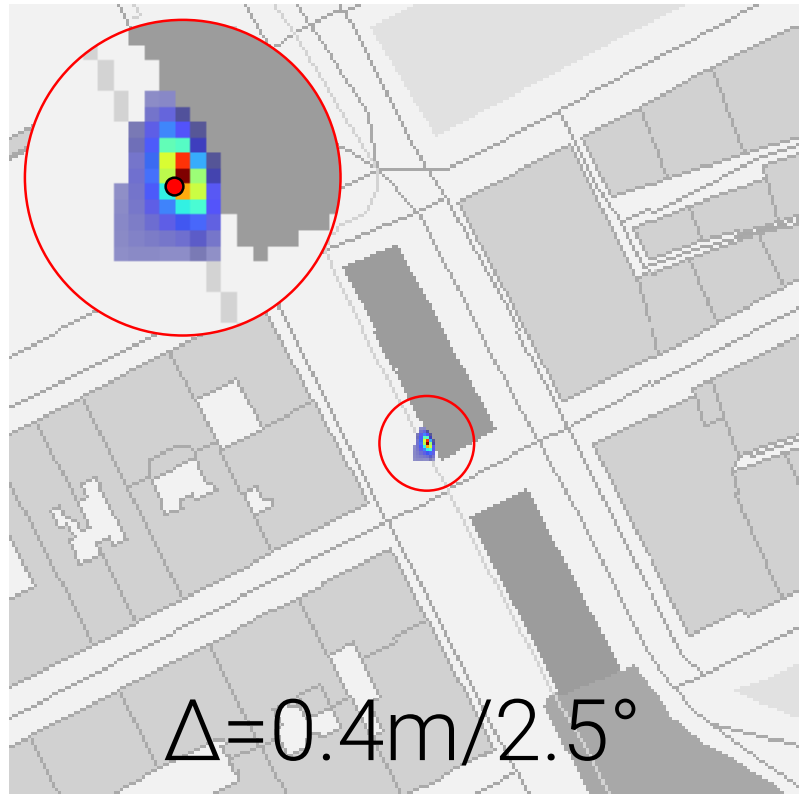


input image

raster map



likelihood



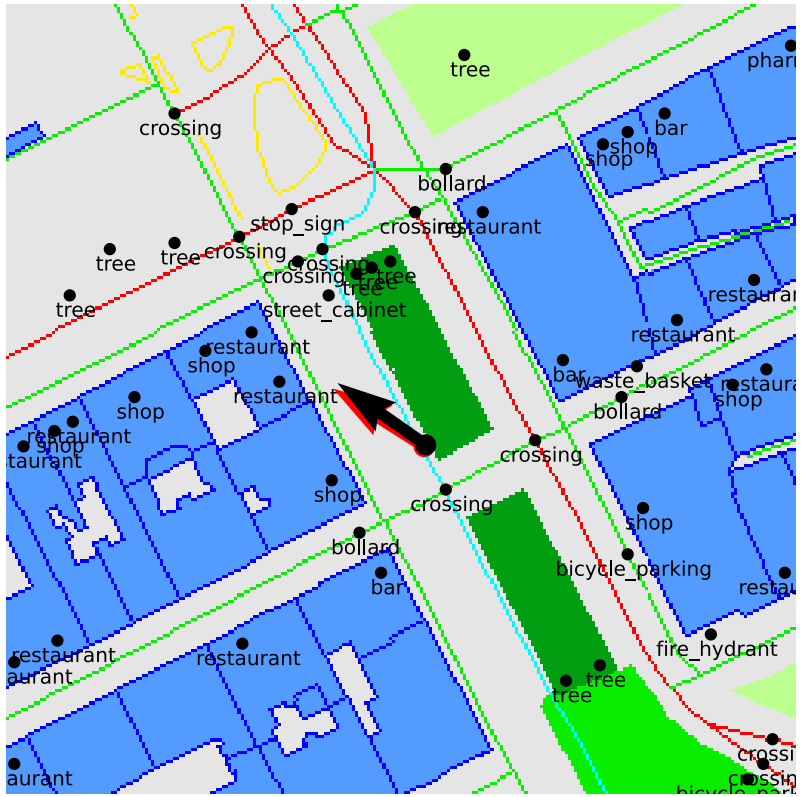
ground truth

prediction

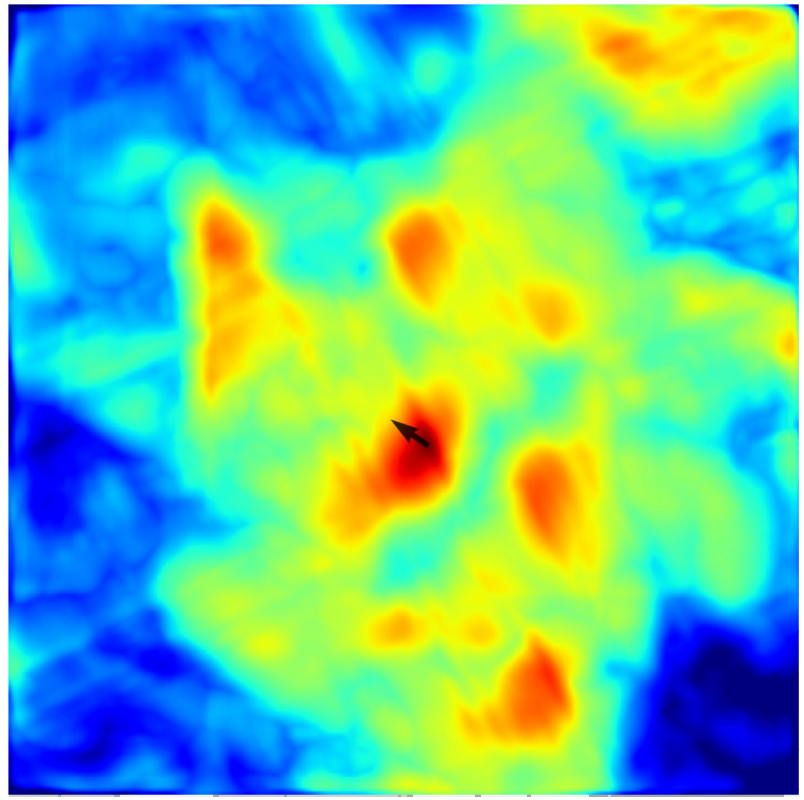
building area and outline, road, footway, cycleway, grass, park, playground, parking, fence

input image

raster map



likelihood

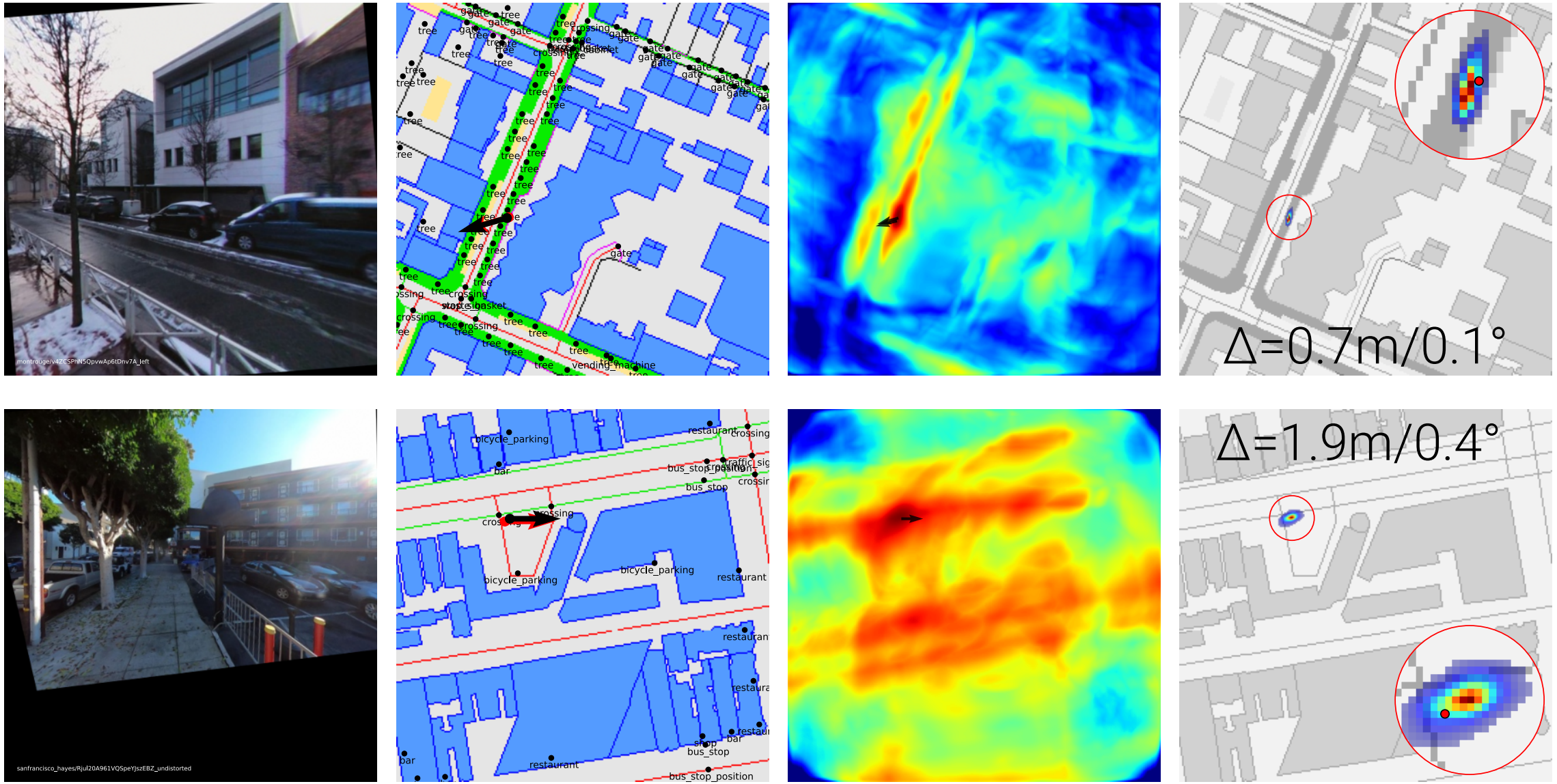


ground truth

prediction

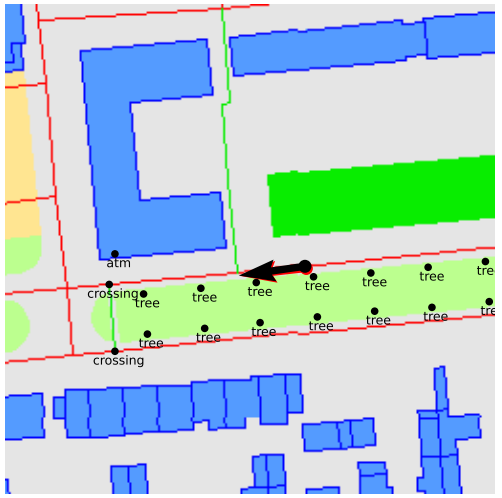
building area and outline, road, footway, cycleway, grass, park, playground, parking, fence

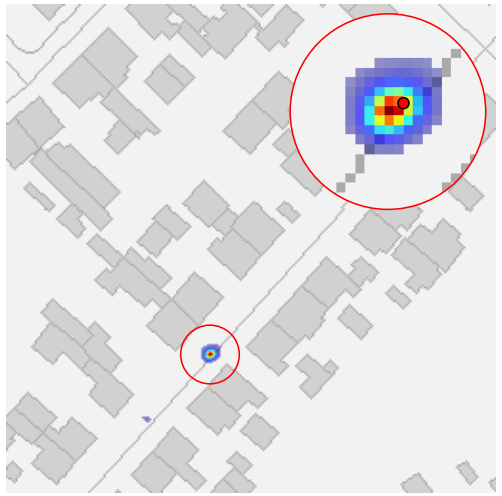
input image raster map log-likelihood likelihood



building area and outline, road, footway, cycleway, grass, park, playground, parking, fence

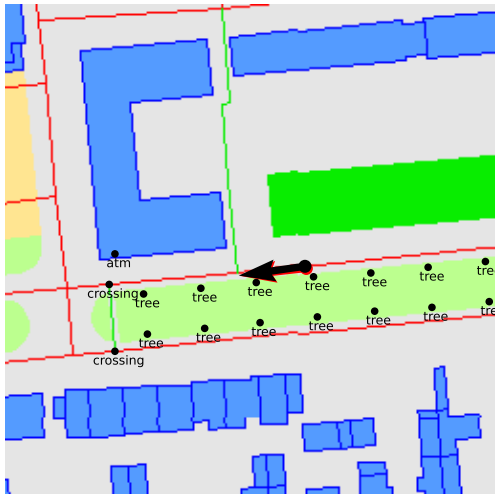
Driving data – KITTI

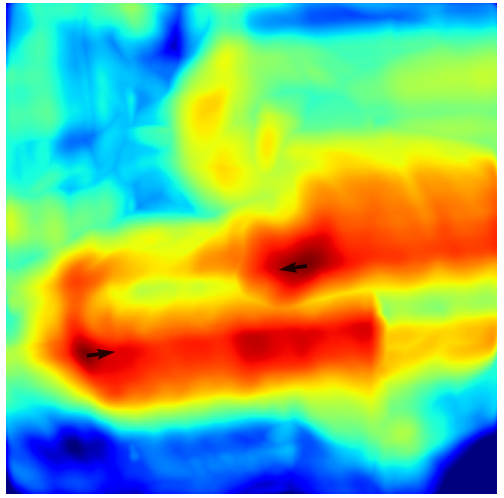


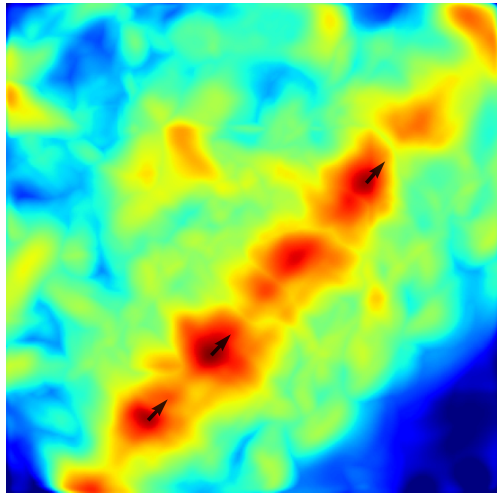


building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

Driving data – KITTI

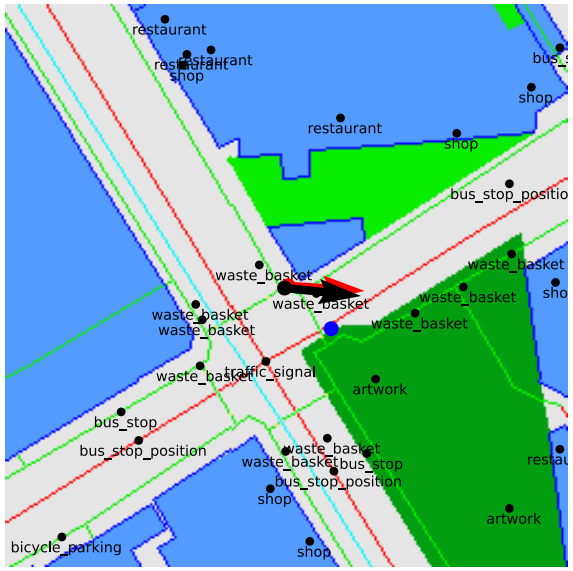


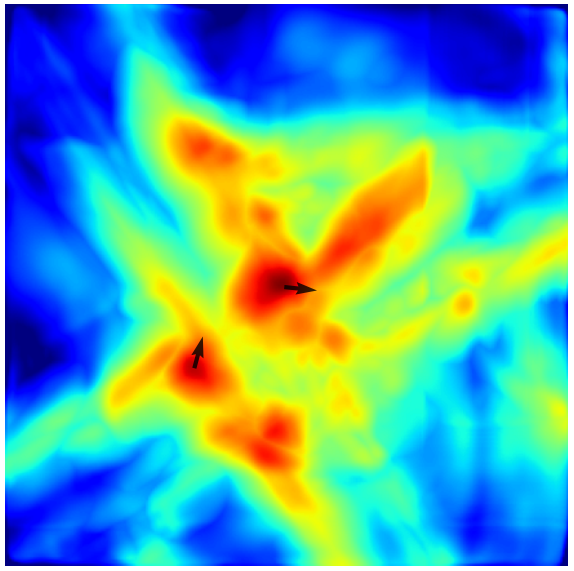


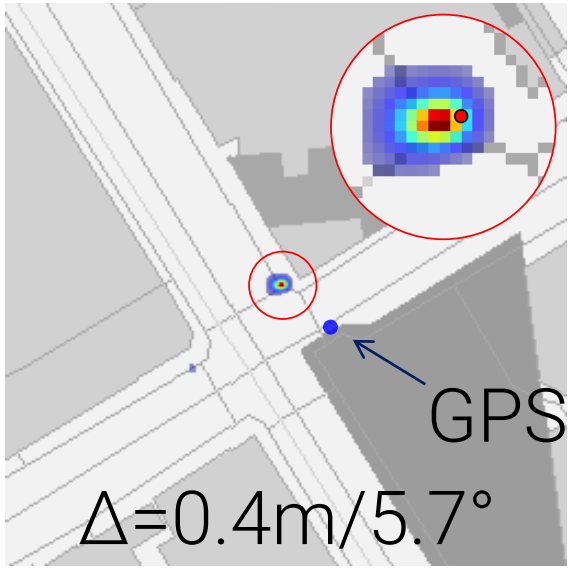


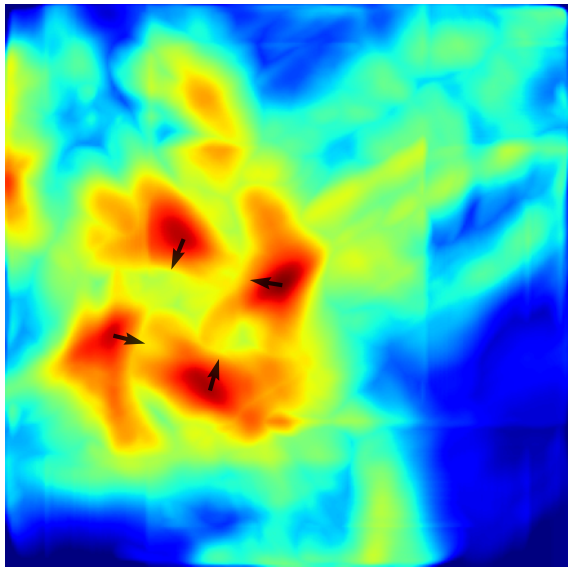
building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

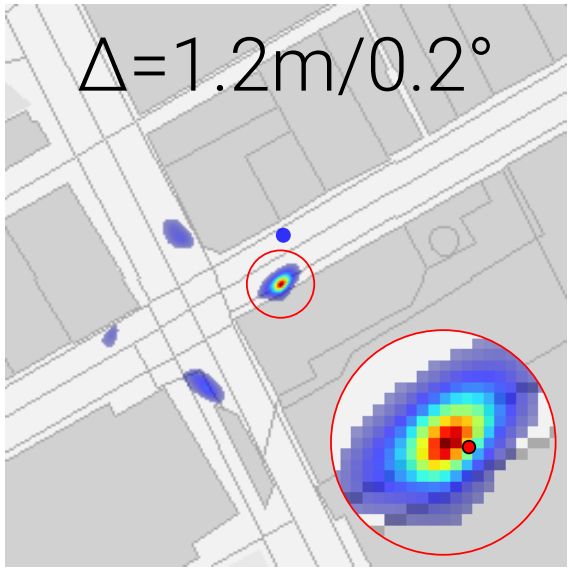
AR data – Aria glasses





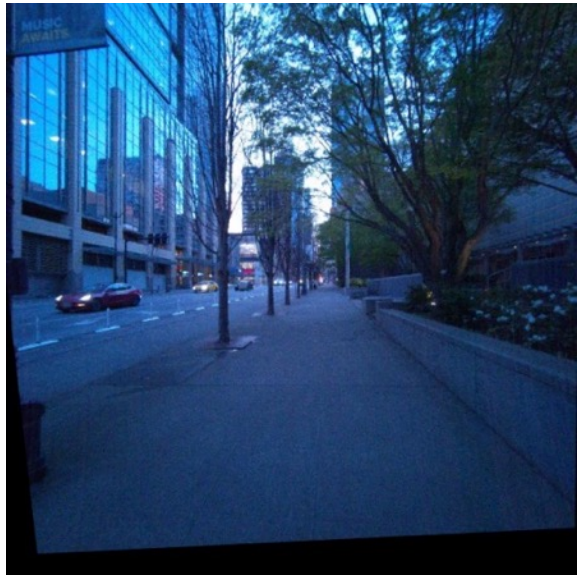


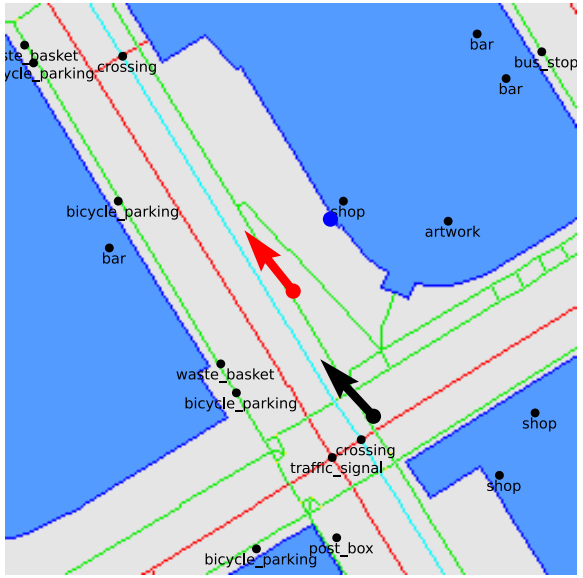


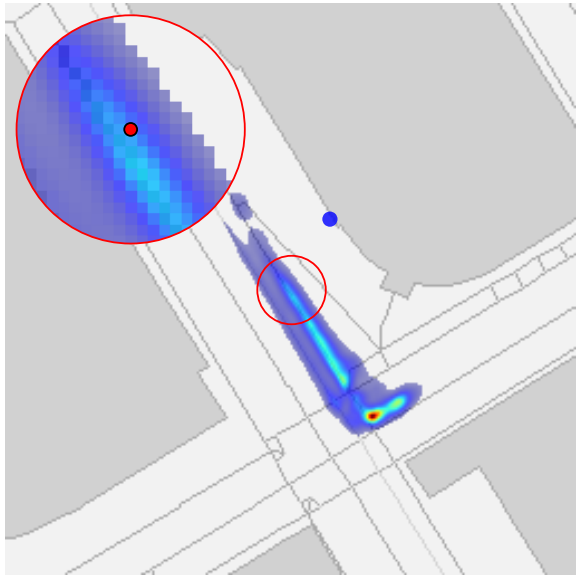


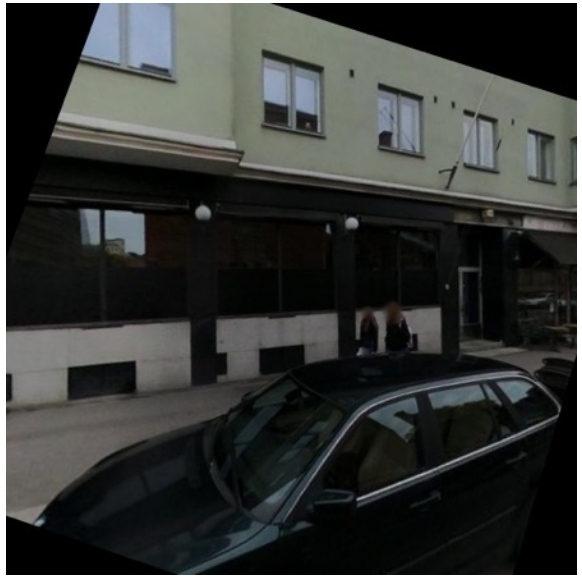
building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

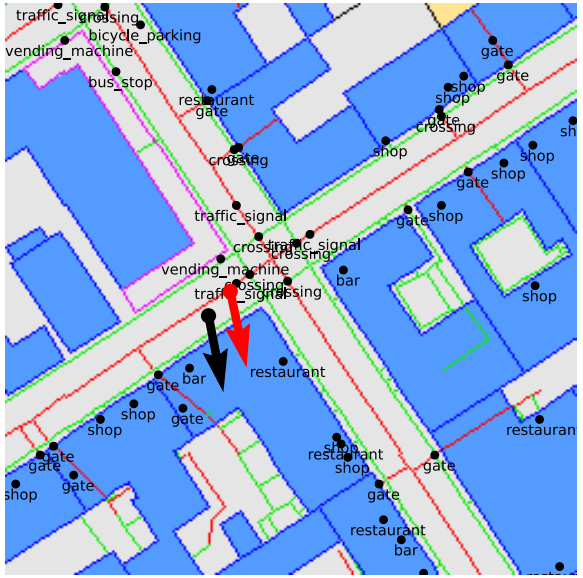
Failure cases

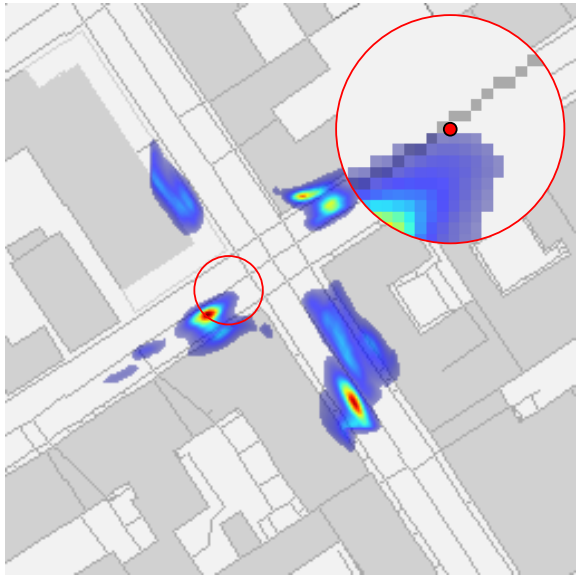




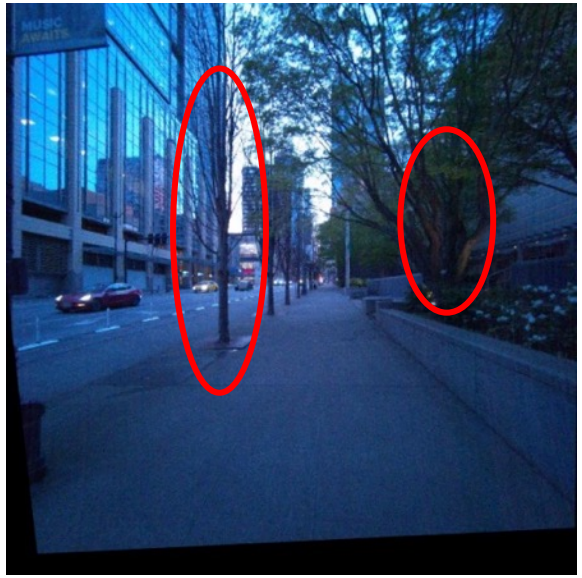


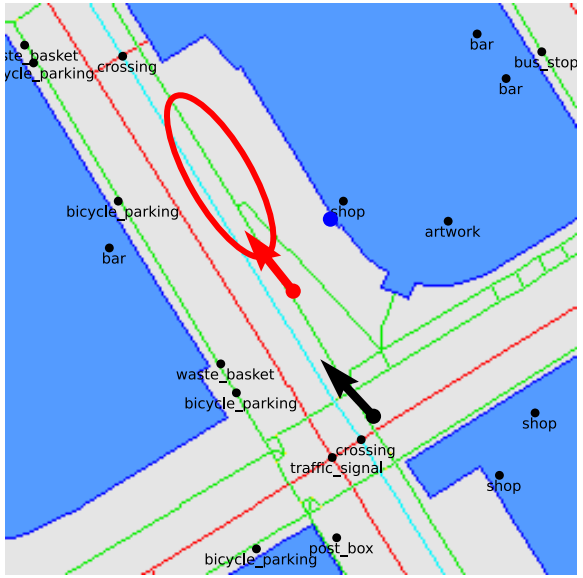


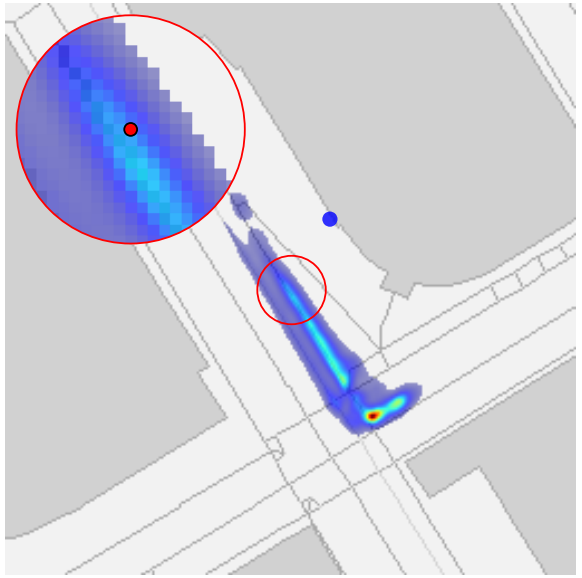


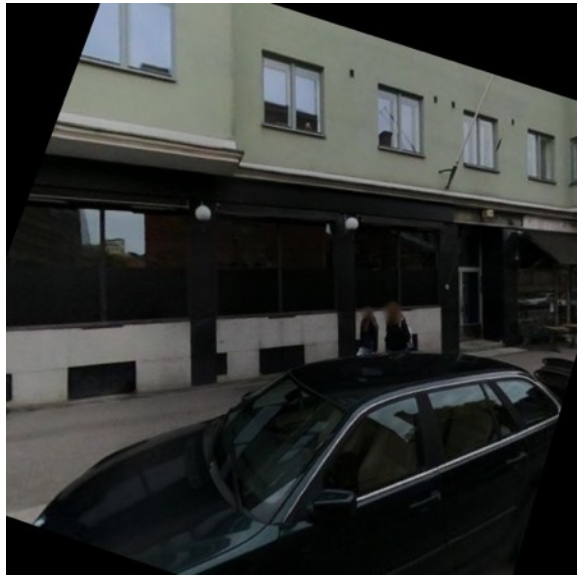


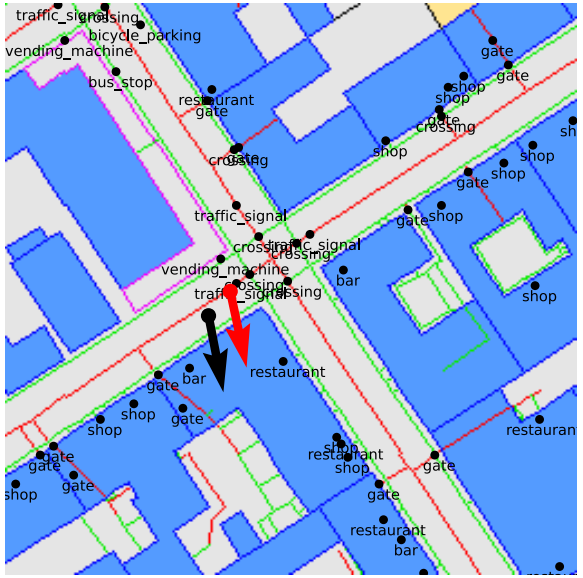
Failure cases

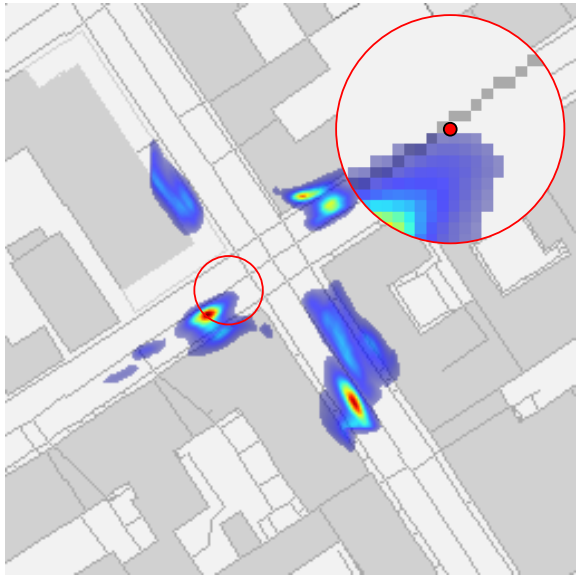












Sequence localization

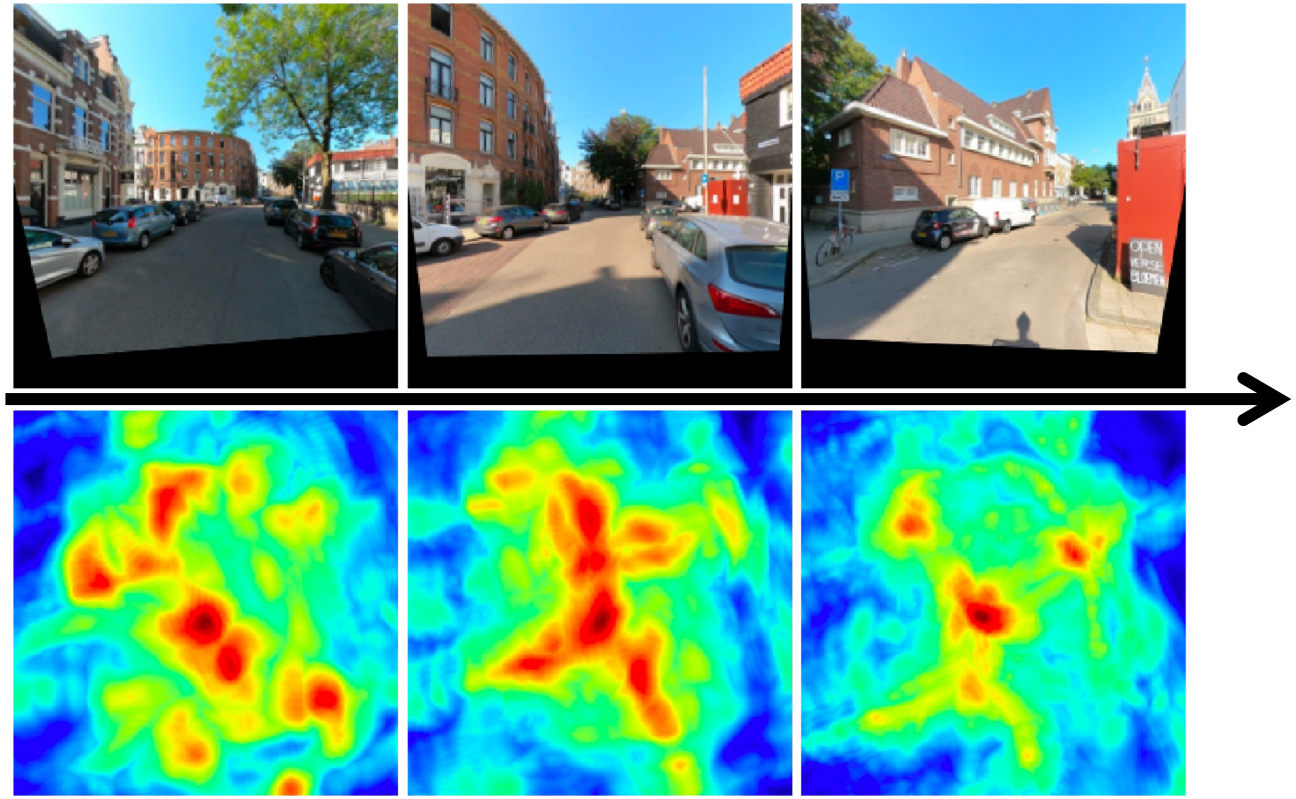
Fuse successive predictions assuming known relative poses

$$P(\boldsymbol{\xi}_i|\{\mathbf{I}_j\}, \text{map}) = \prod_k P(\boldsymbol{\xi}_i \oplus \hat{\boldsymbol{\xi}}_{ij}|\mathbf{I}_j, \text{map})$$

input image

time

log likelihood



Sequence localization

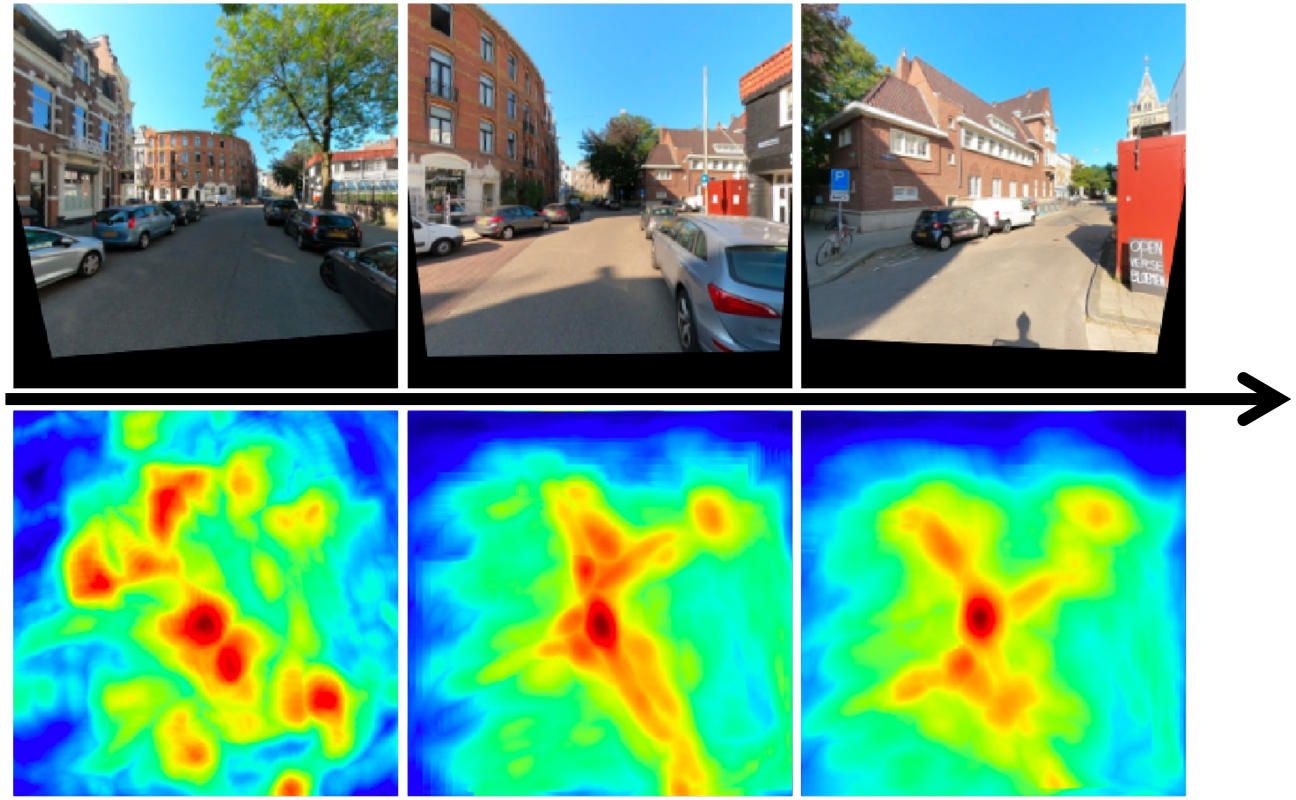
Fuse successive predictions assuming known relative poses

$$P(\boldsymbol{\xi}_i|\{\mathbf{I}_j\}, \text{map}) = \prod_k P(\boldsymbol{\xi}_i \oplus \hat{\boldsymbol{\xi}}_{ij}|\mathbf{I}_j, \text{map})$$

input image

time

log likelihood



Sequence localization

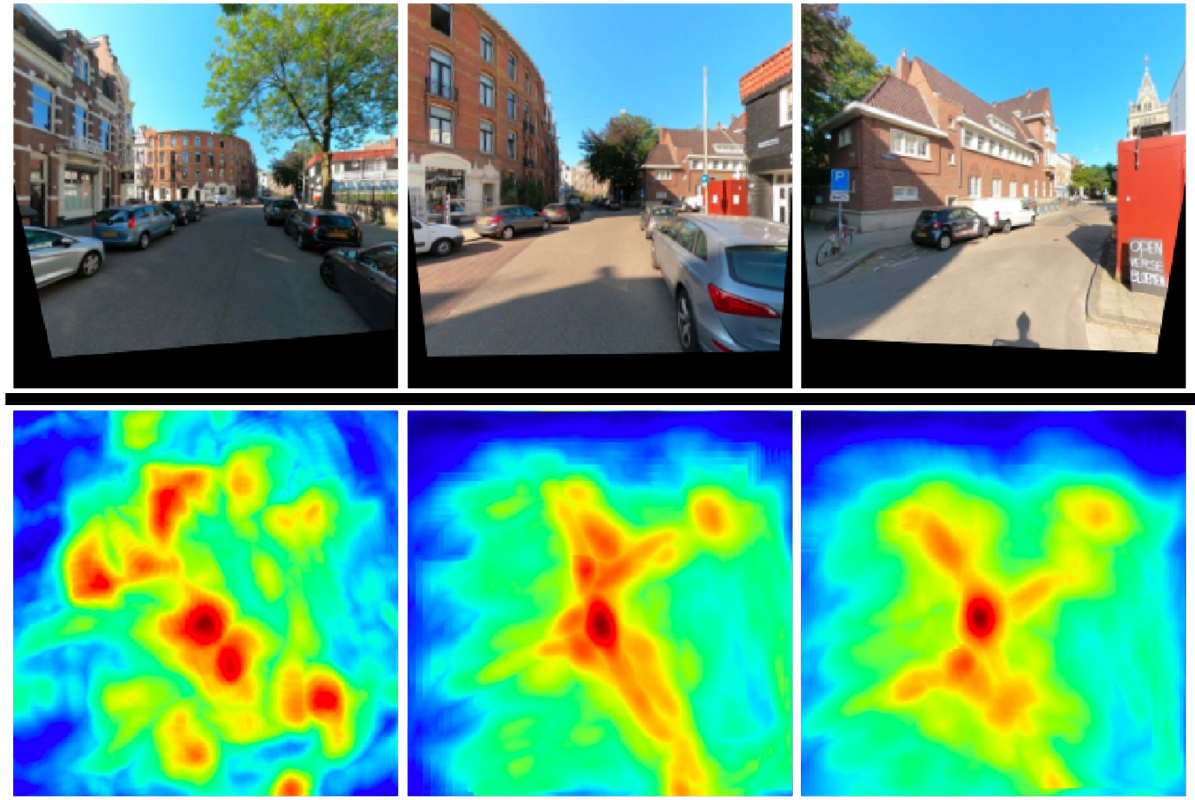
Fuse successive predictions assuming known relative poses

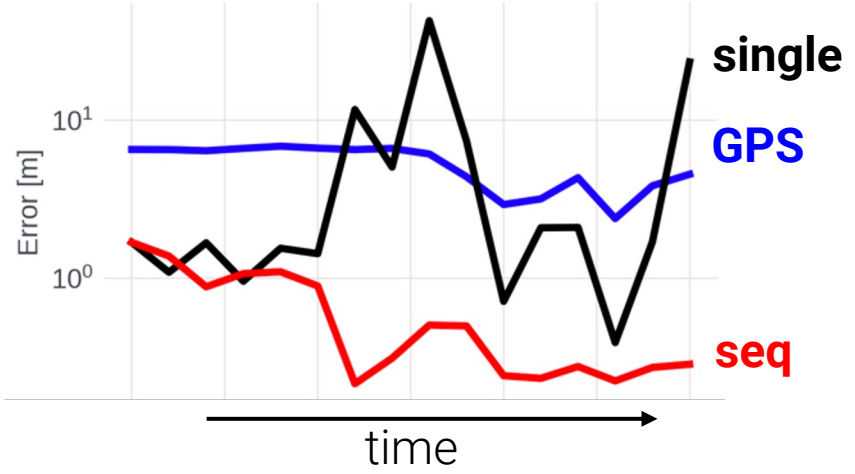
$$P(\boldsymbol{\xi}_i|\{\mathbf{I}_j\}, \text{map}) = \prod_k P(\boldsymbol{\xi}_i \oplus \hat{\boldsymbol{\xi}}_{ij}|\mathbf{I}_j, \text{map})$$

input image

time

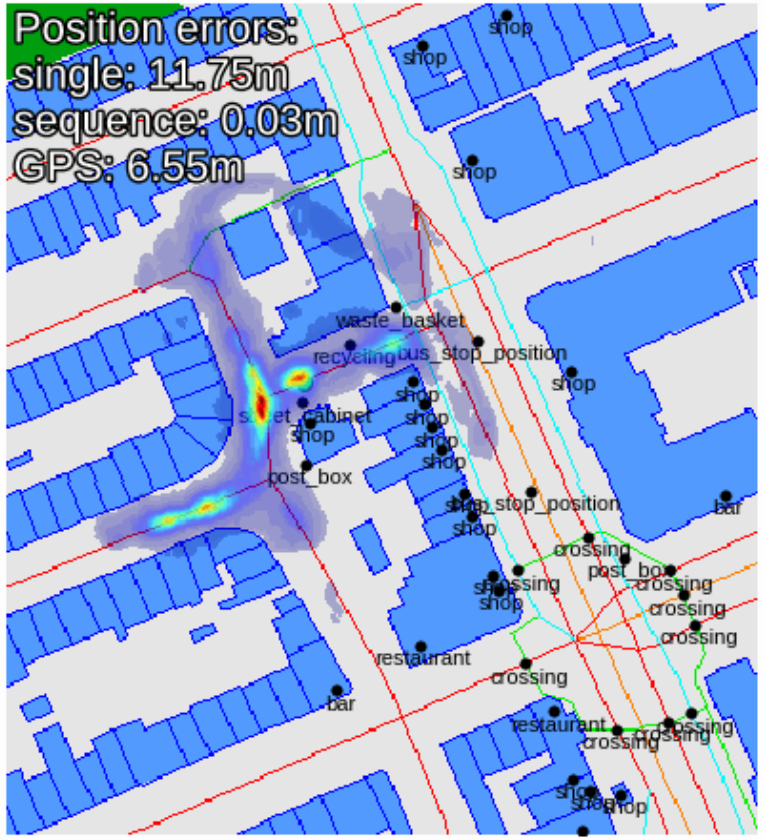
log likelihood



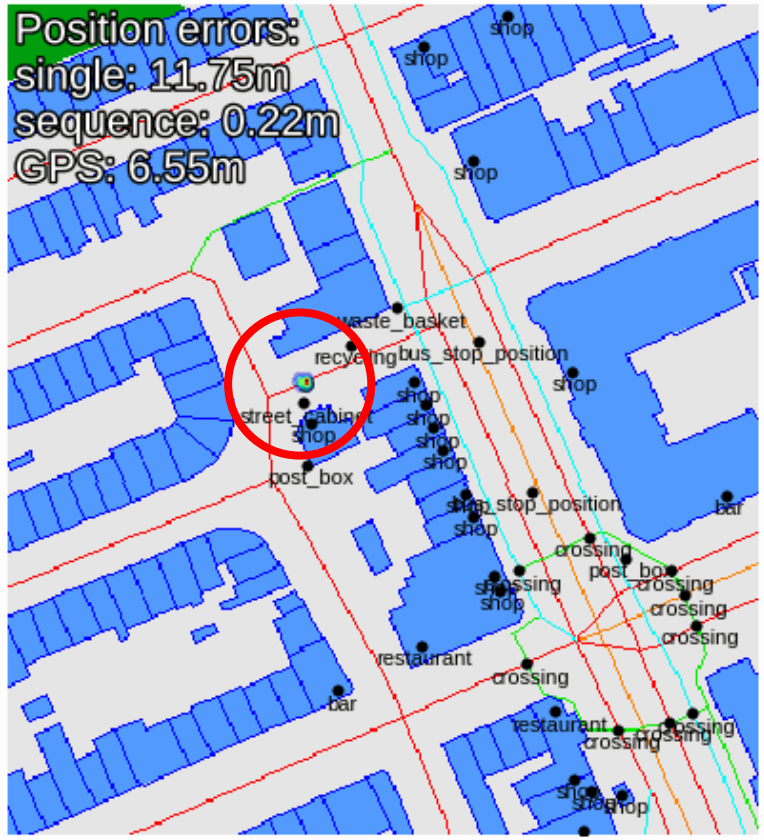


Sequence localization

input image



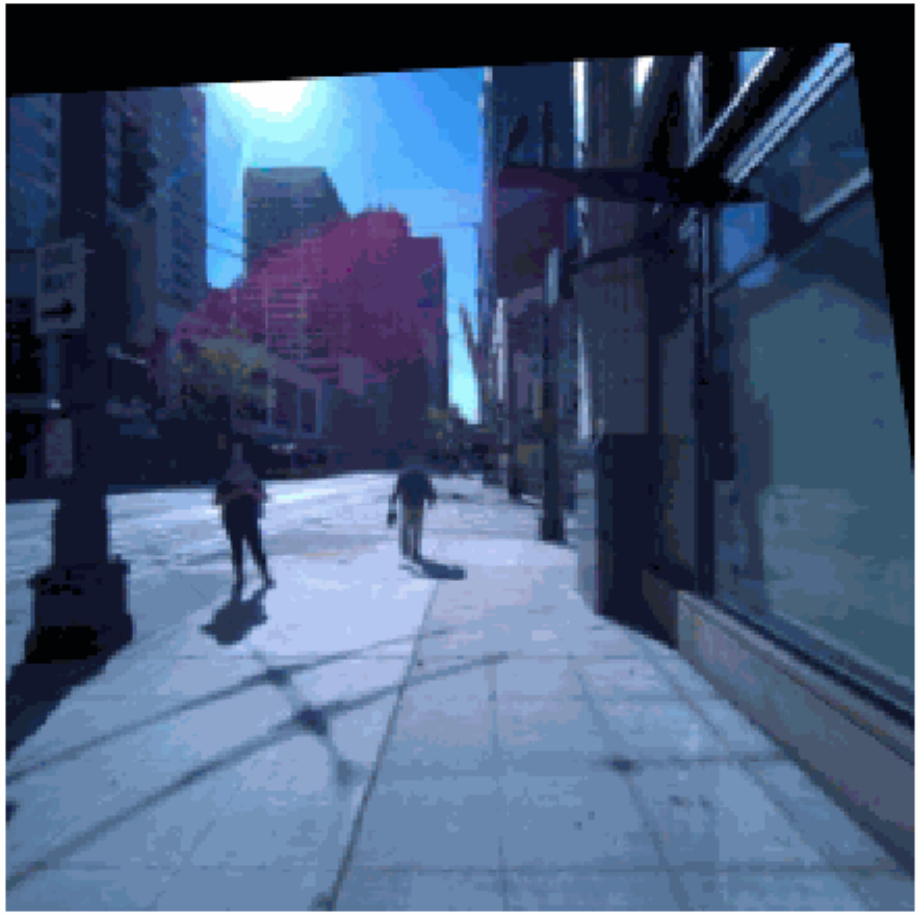
single-frame likelihood



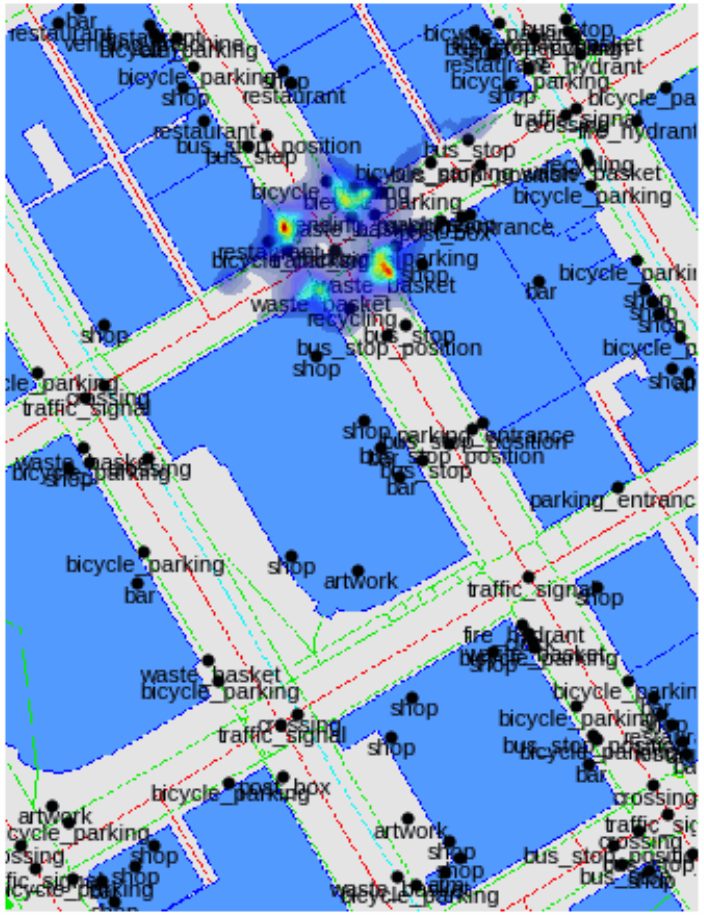
sequence likelihood

ground truth position

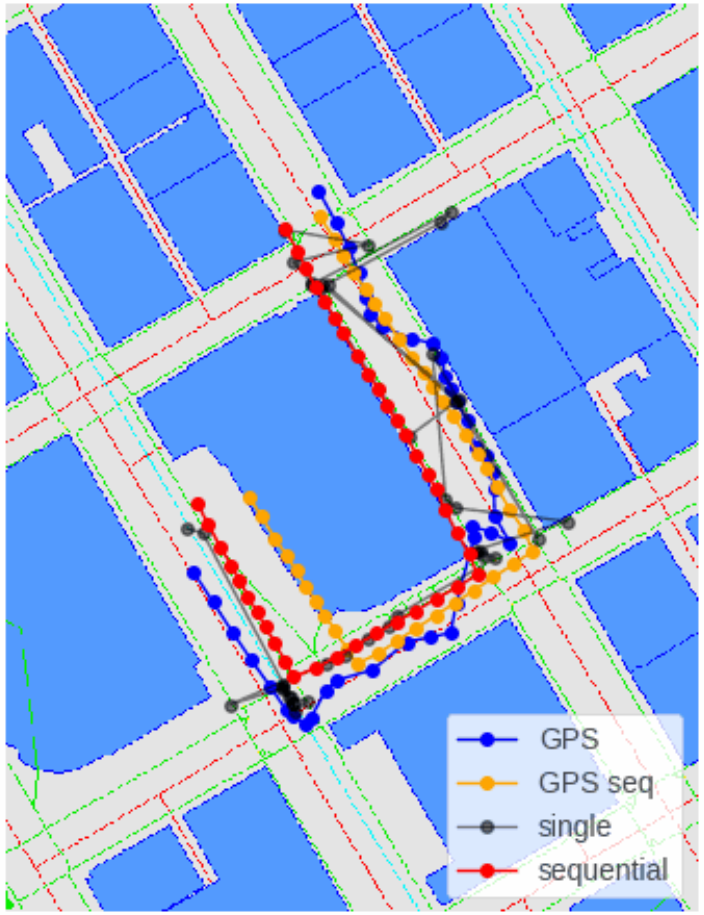
Sequence localization – Aria Seattle



input image

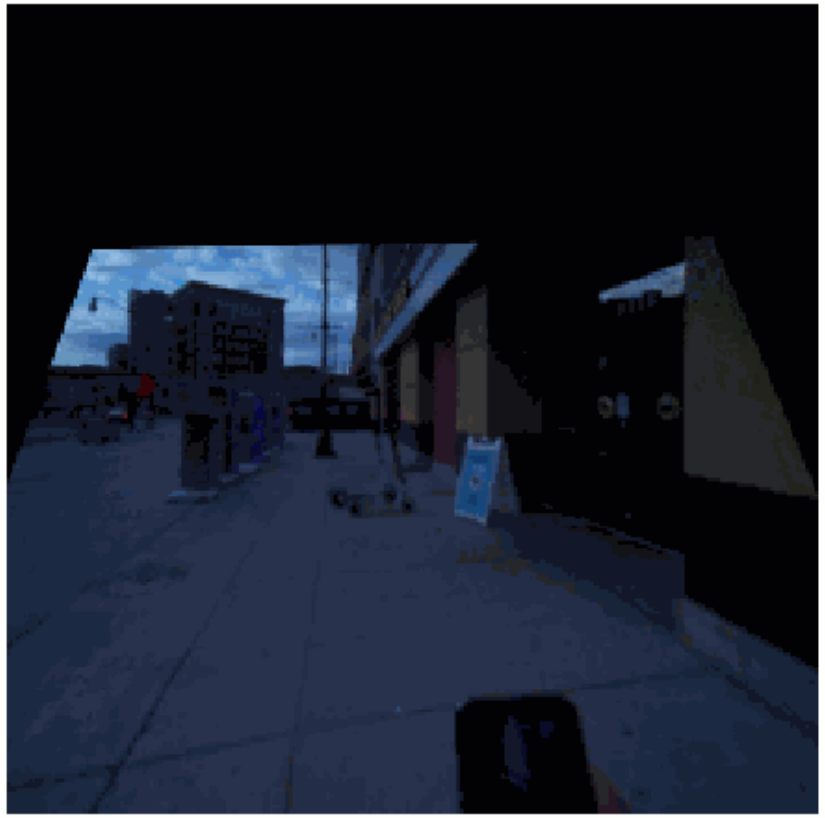


single-frame likelihood

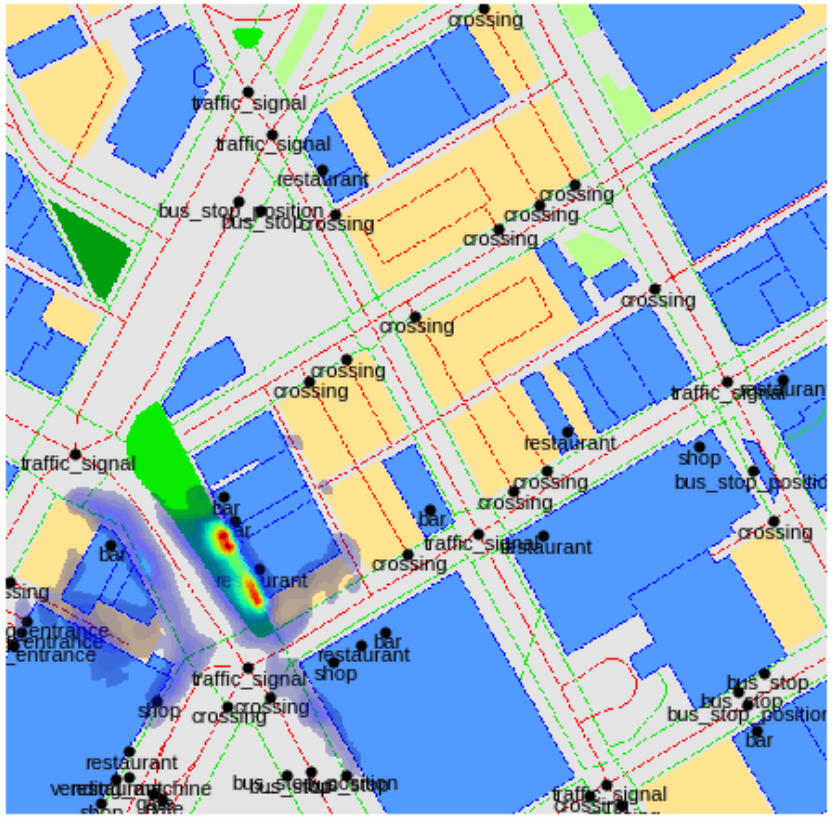


final trajectories

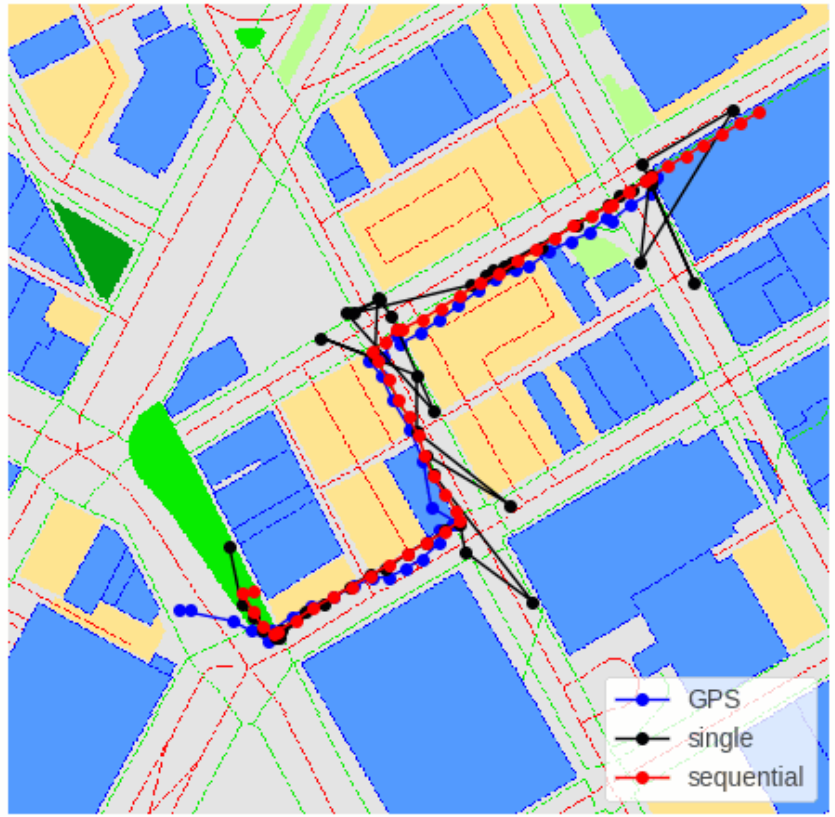
Sequence localization – Aria Detroit



input image



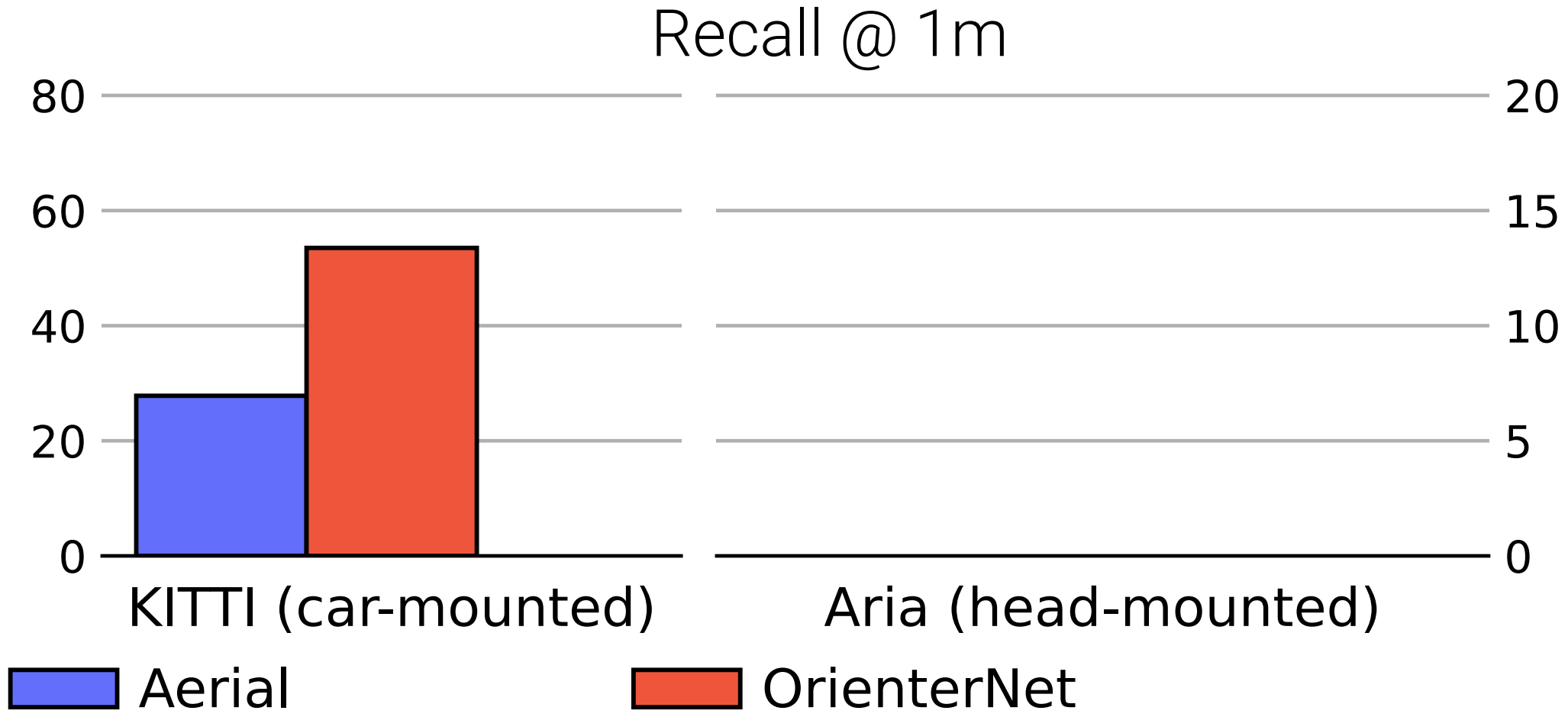
single-frame likelihood



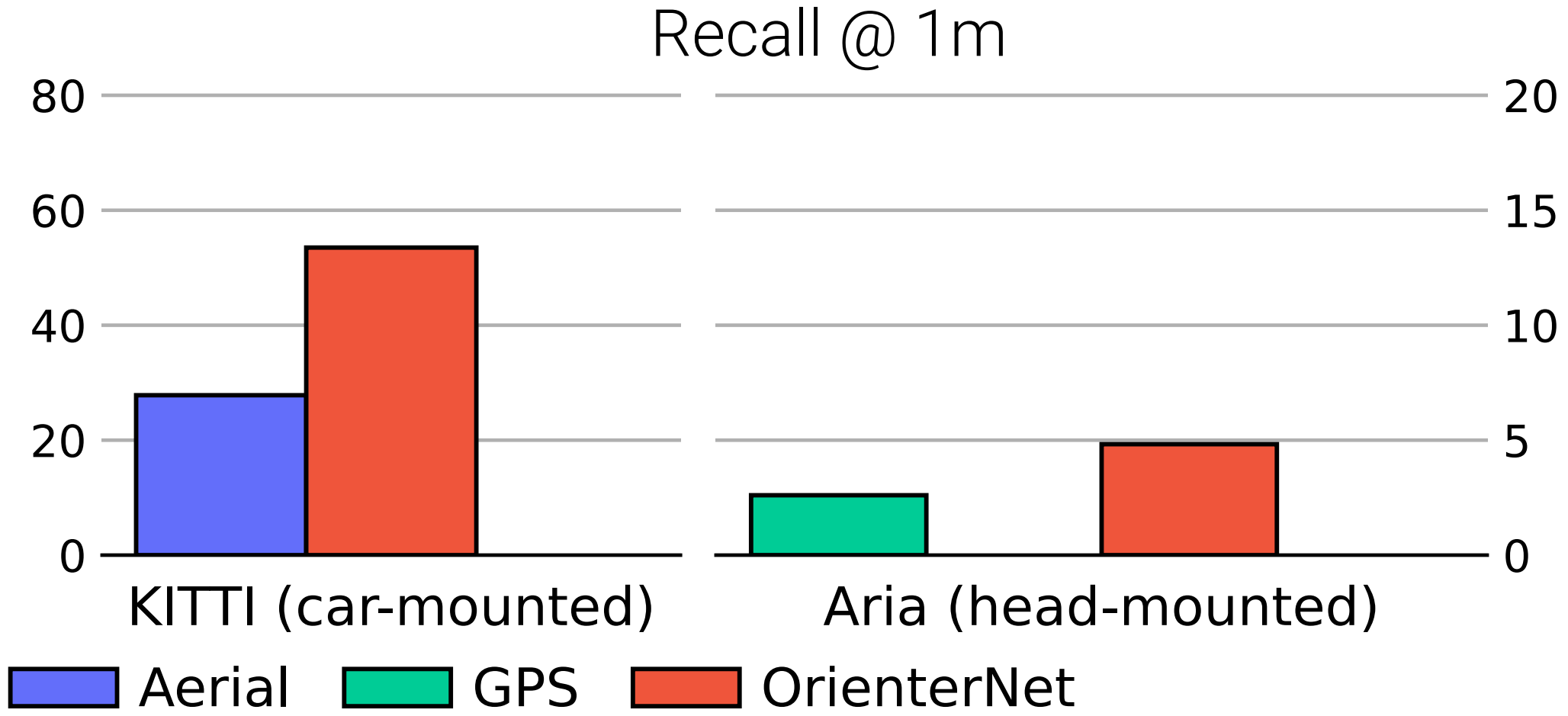
final trajectories

Quantitative results

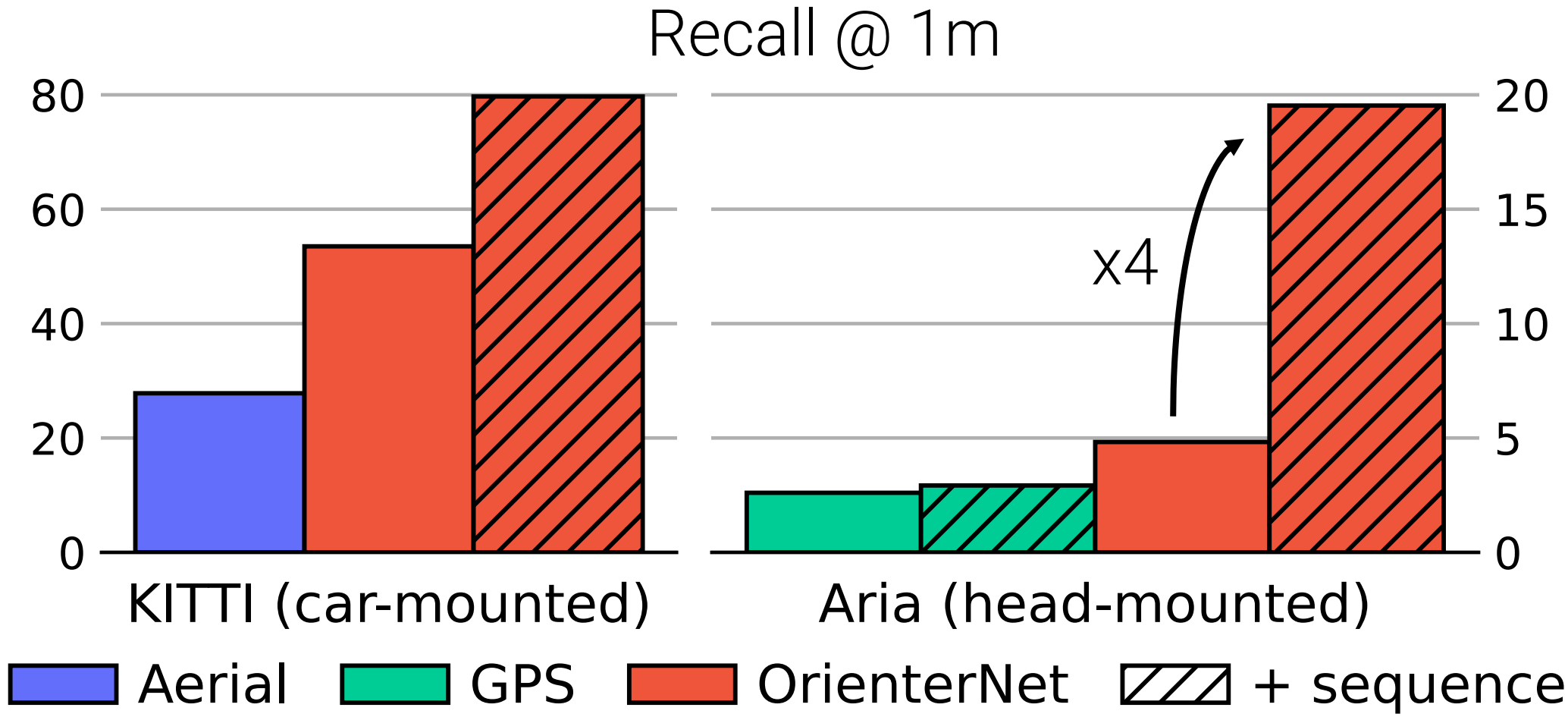
Quantitative results



Quantitative results



Quantitative results



OrienterNet

Visual Localization in 2D Public Maps with Neural Matching

Paul-Edouard Sarlin[1] Daniel DeTone[2] Tsun-Yi Yang[2] Armen Avetisyan[2]
Julian Straub[2] Tomasz Malisiewicz[2] Samuel Rota Bulo[2]
Richard Newcombe[2] Peter Kontschieder[2] Vasileios Balntas[2]

psarlin.com/orienternet