# 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]

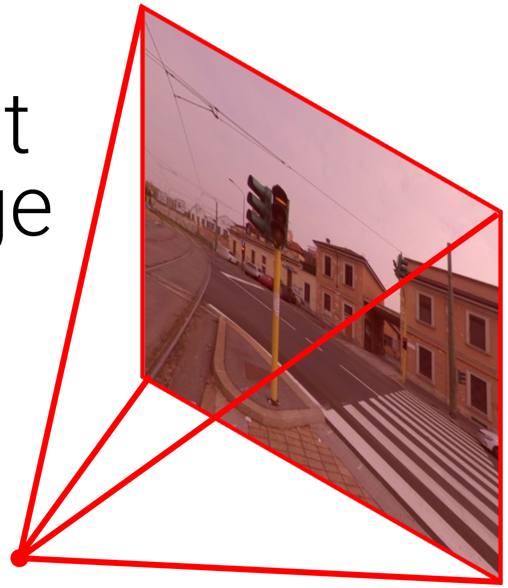
1 ETH zürich
2 Meta Reality Labs

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

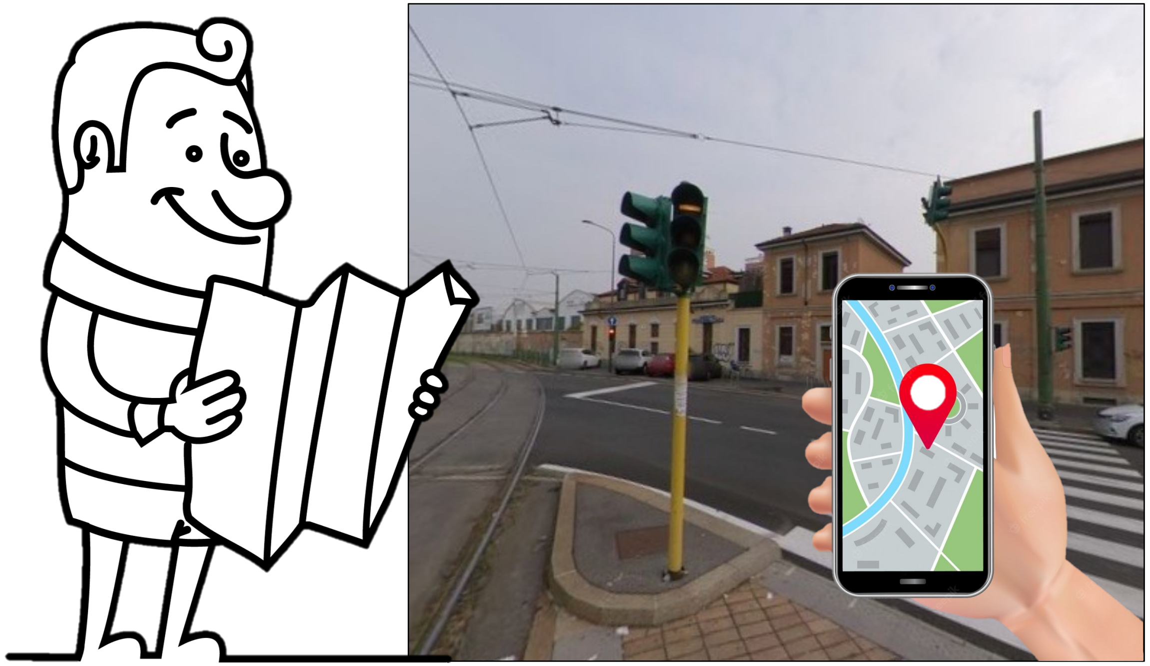
CVPR 2023 psarlin.com/orienternet Poster THU-PM-098

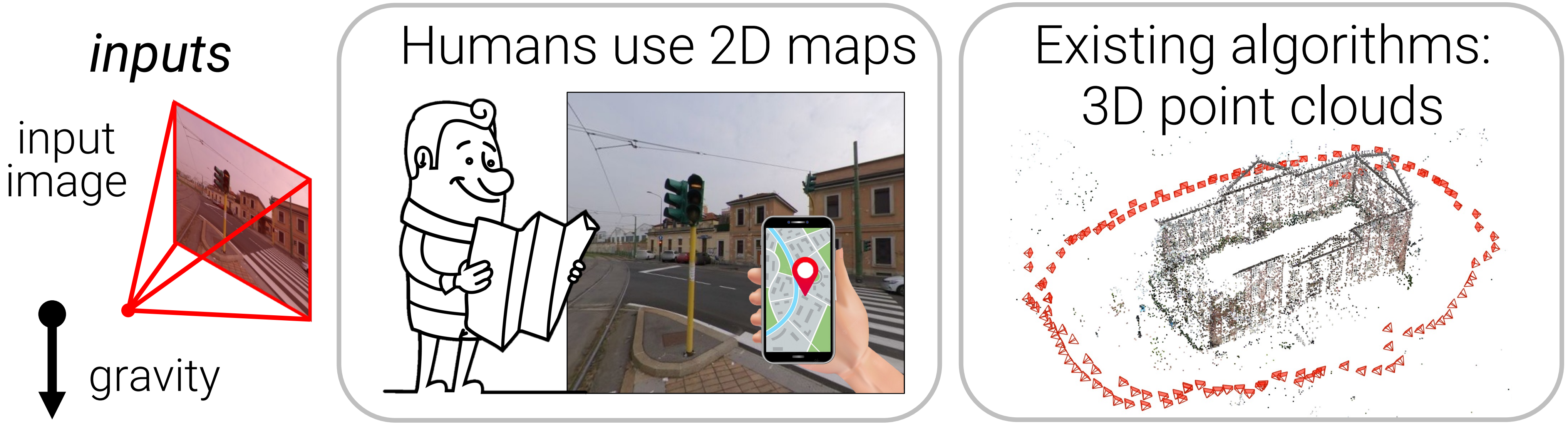
# Humans use simple 2D maps

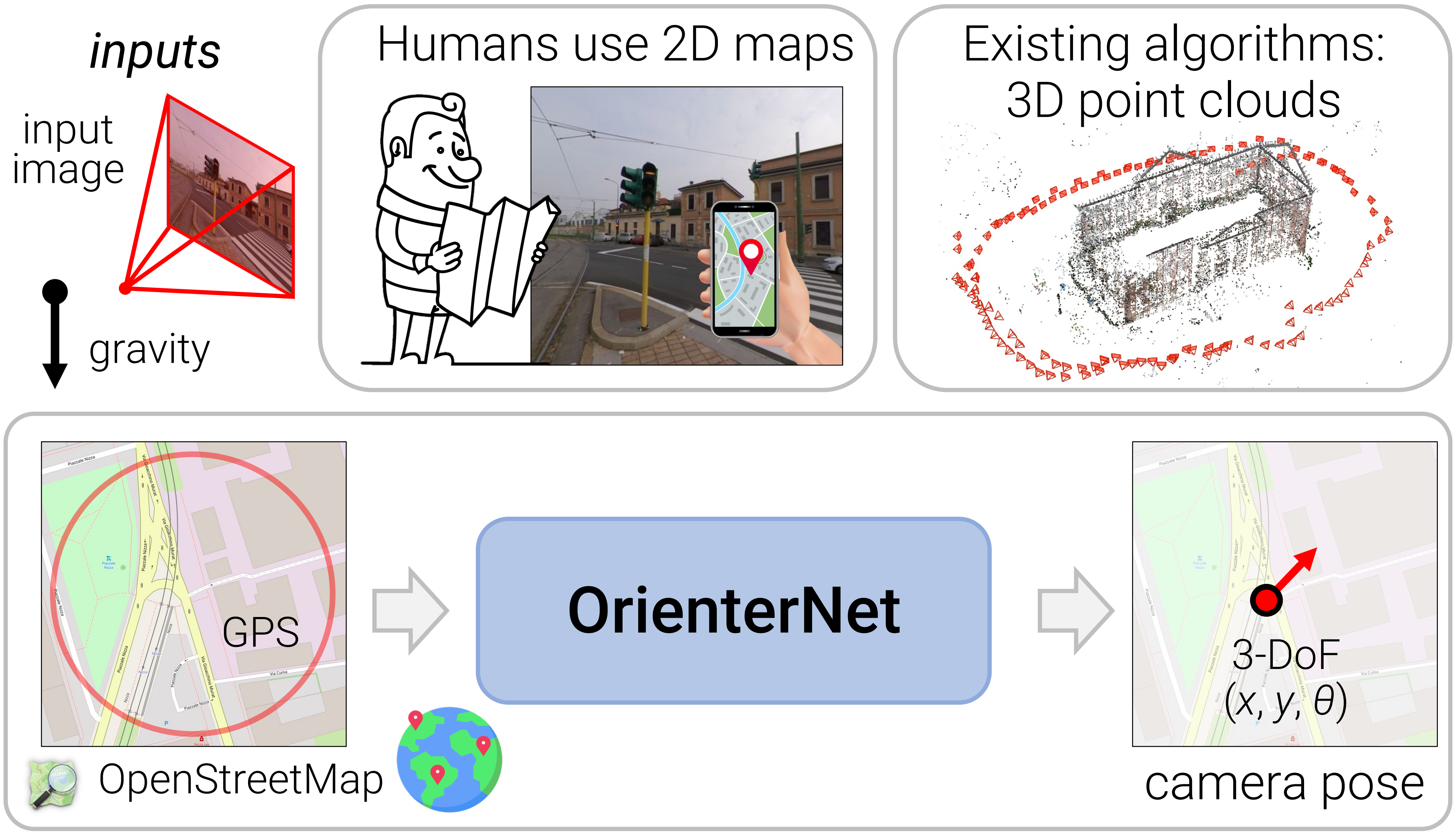
input
image

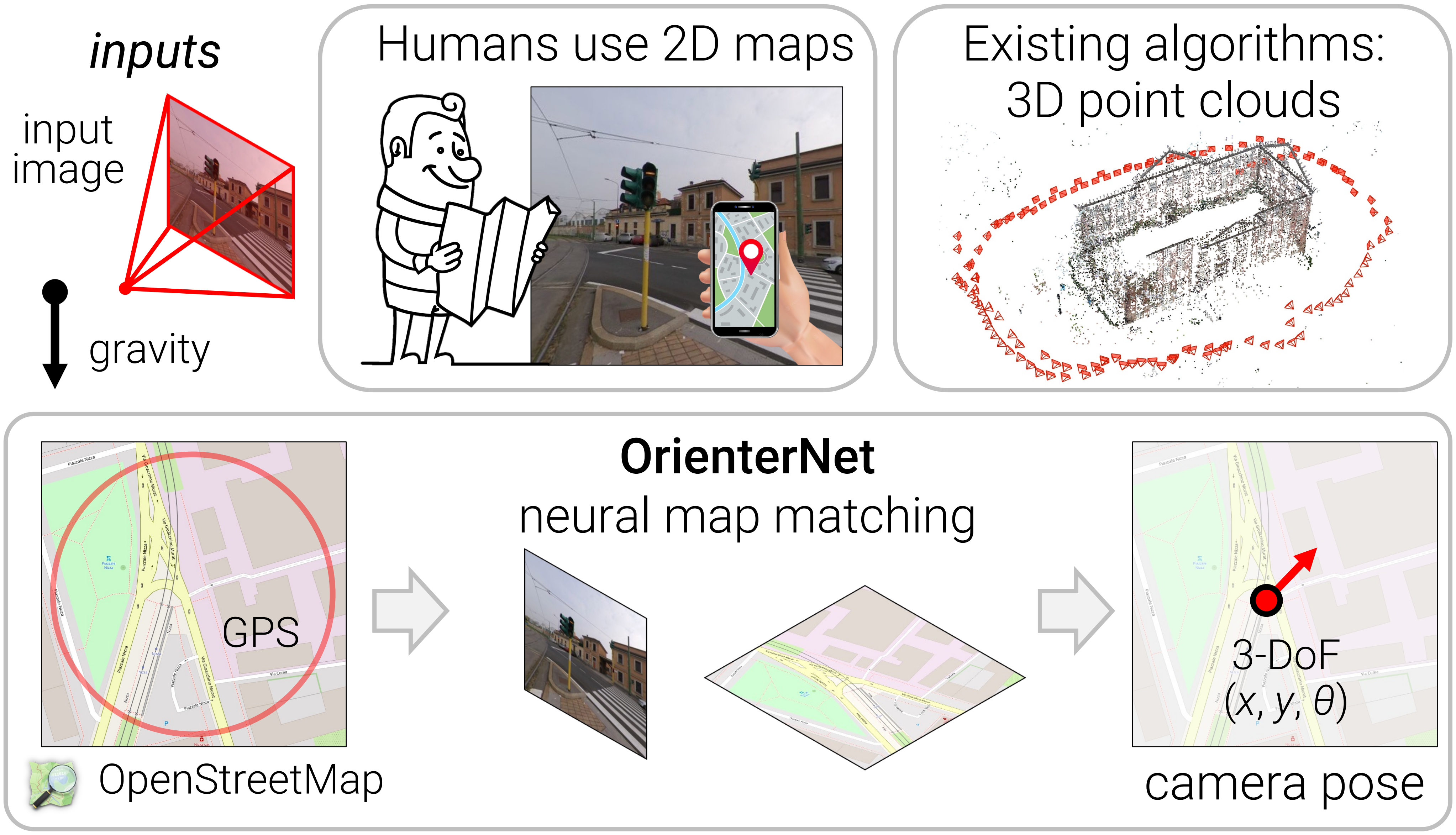


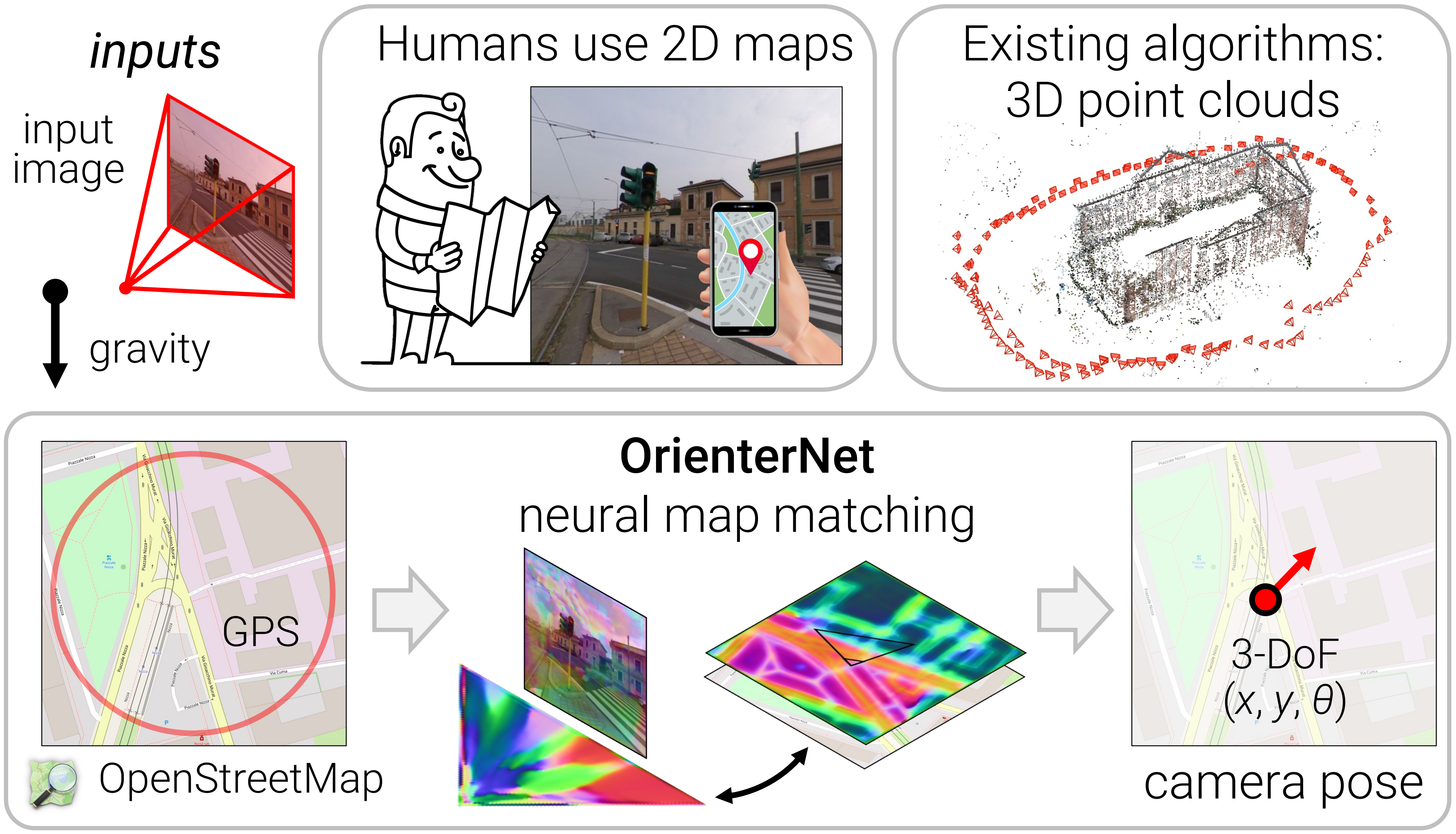
where am I?



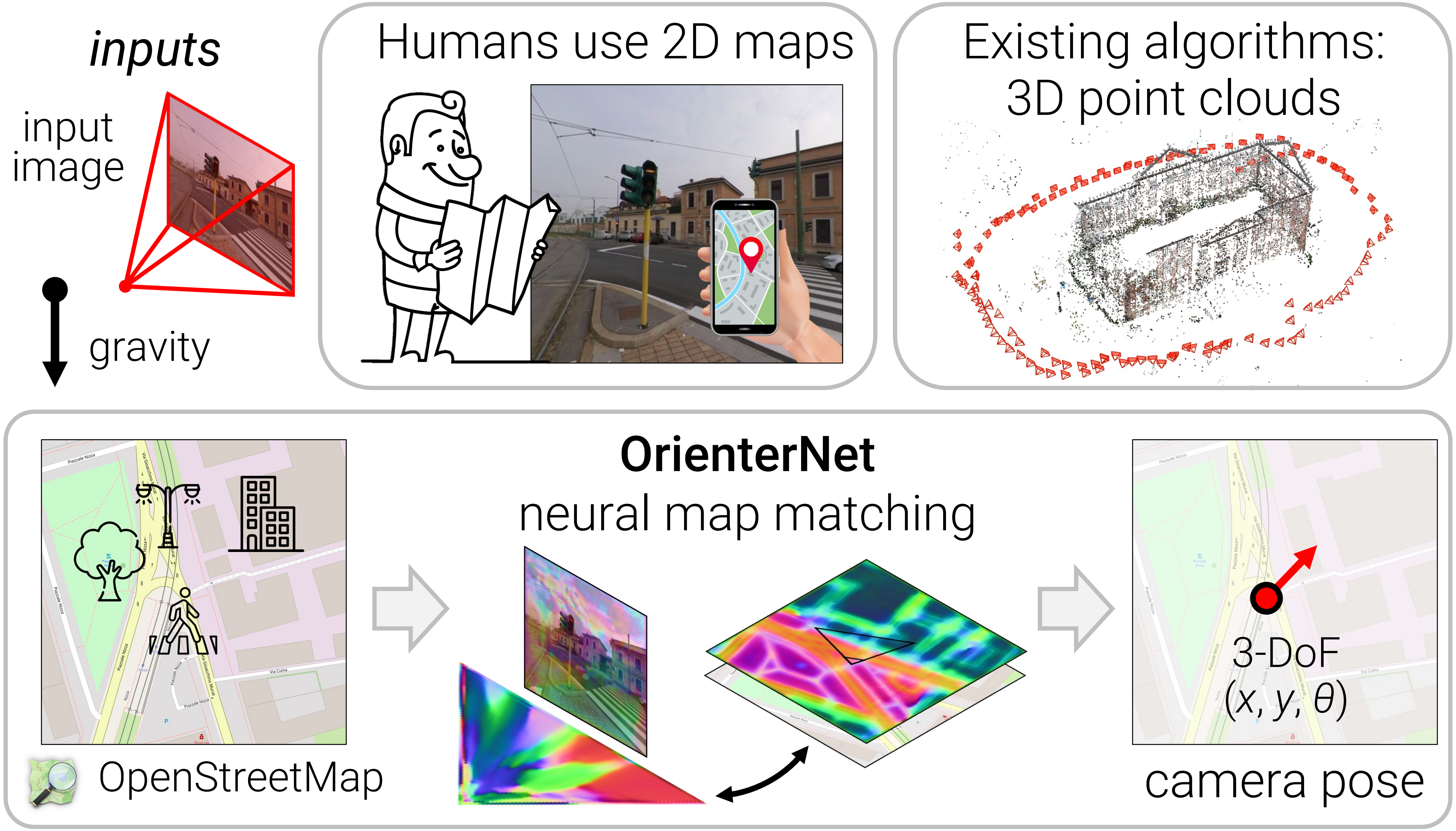

inputs
input image
gravity
Humans use 2D maps
Existing algorithms:
3D point clouds


inputs
input image
gravity
Humans use 2D maps
Existing algorithms: 3D point clouds
GPS
OpenStreetMap
OrienterNet
3-DoF
(x, y, θ)
camera pose


inputs
input image
gravity
Humans use 2D maps
Existing algorithms: 3D point clouds
OrienterNet
neural map matching
GPS
OpenStreetMap
3-DoF
(x, y, θ)
camera pose



inputs
input image
gravity
Humans use 2D maps
Existing algorithms: 3D point clouds
GPS
OpenStreetMap
OrienterNet
neural map matching
3-DoF
(x, y, θ)
camera pose

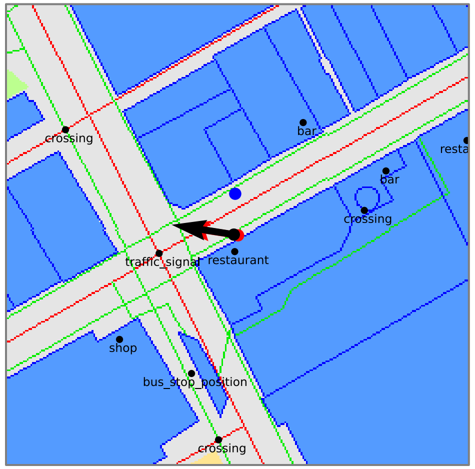

inputs
input image
gravity
Humans use 2D maps
Existing algorithms: 3D point clouds
OpenStreetMap
OrienterNet
neural map matching
3-DoF
(x, y, θ)
camera pose

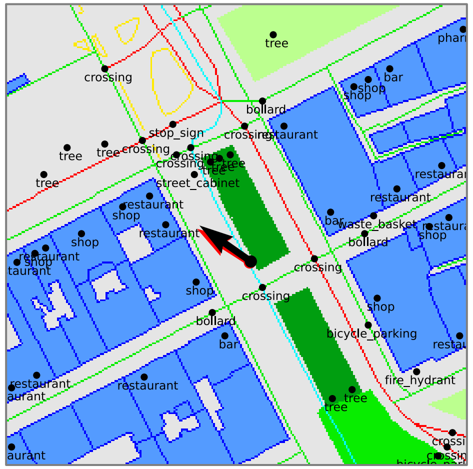
# Zero-shot generalization

Aria

Mapillary

KITTI



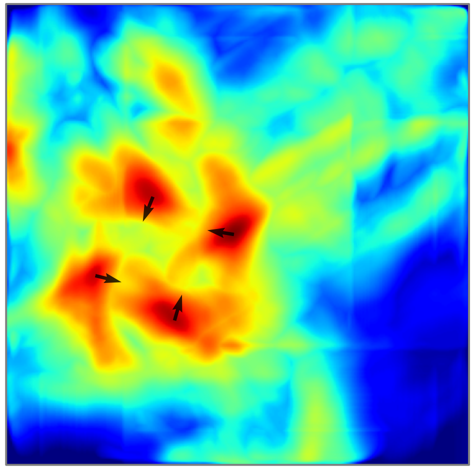



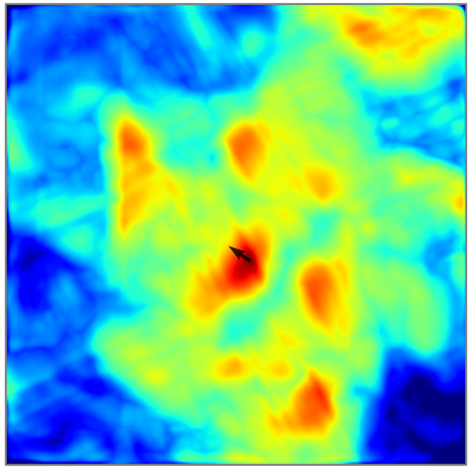

# Zero-shot generalization

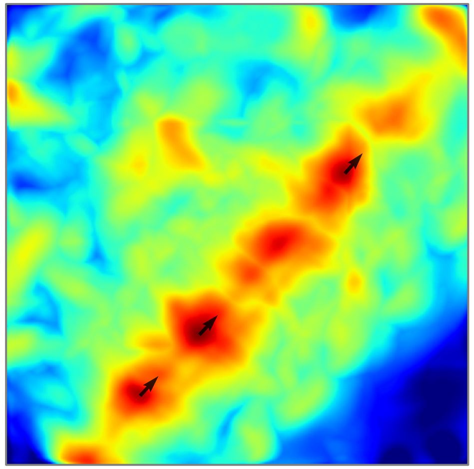
Aria 

Mapillary 

KITTI



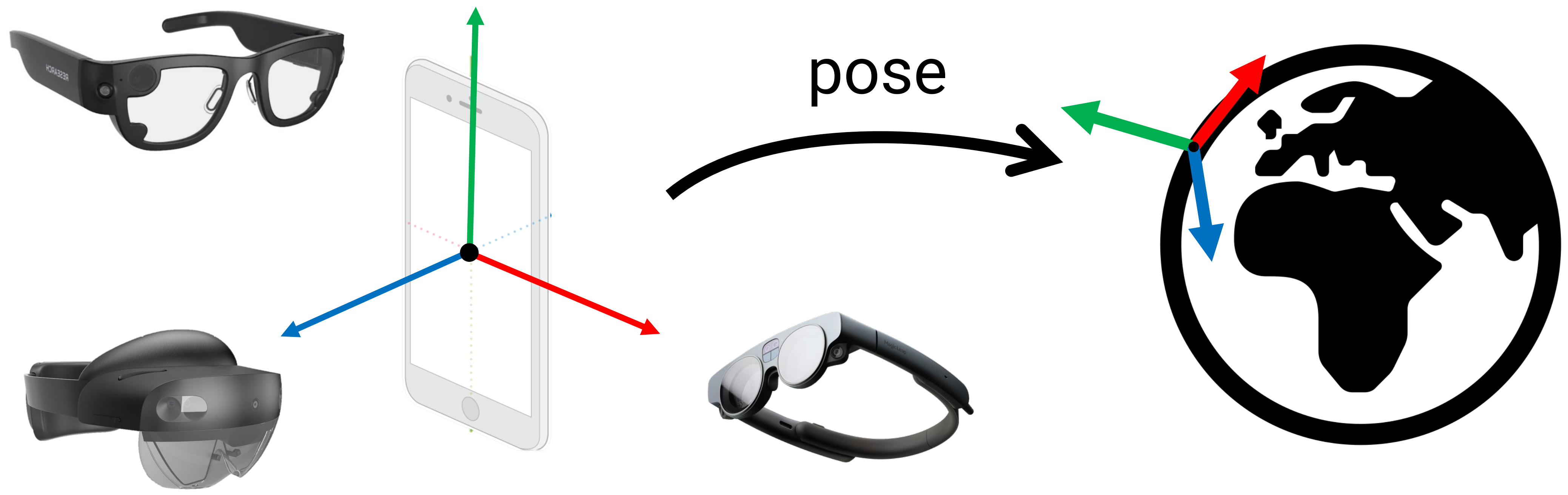




# Positioning

Recover the 6-DoF pose of the device

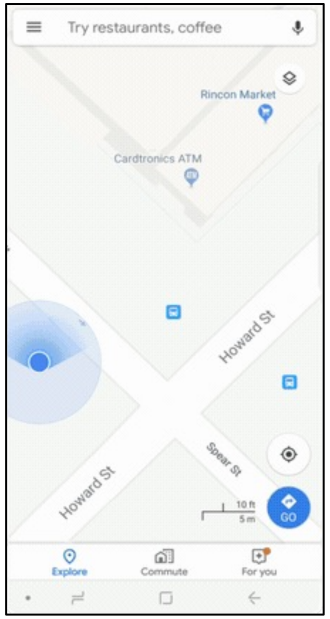
- 3D translation + rotation
- global reference frame

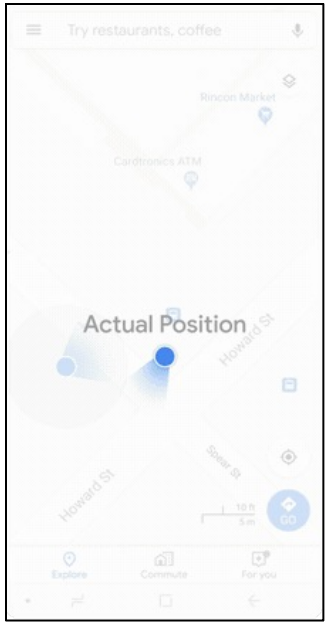


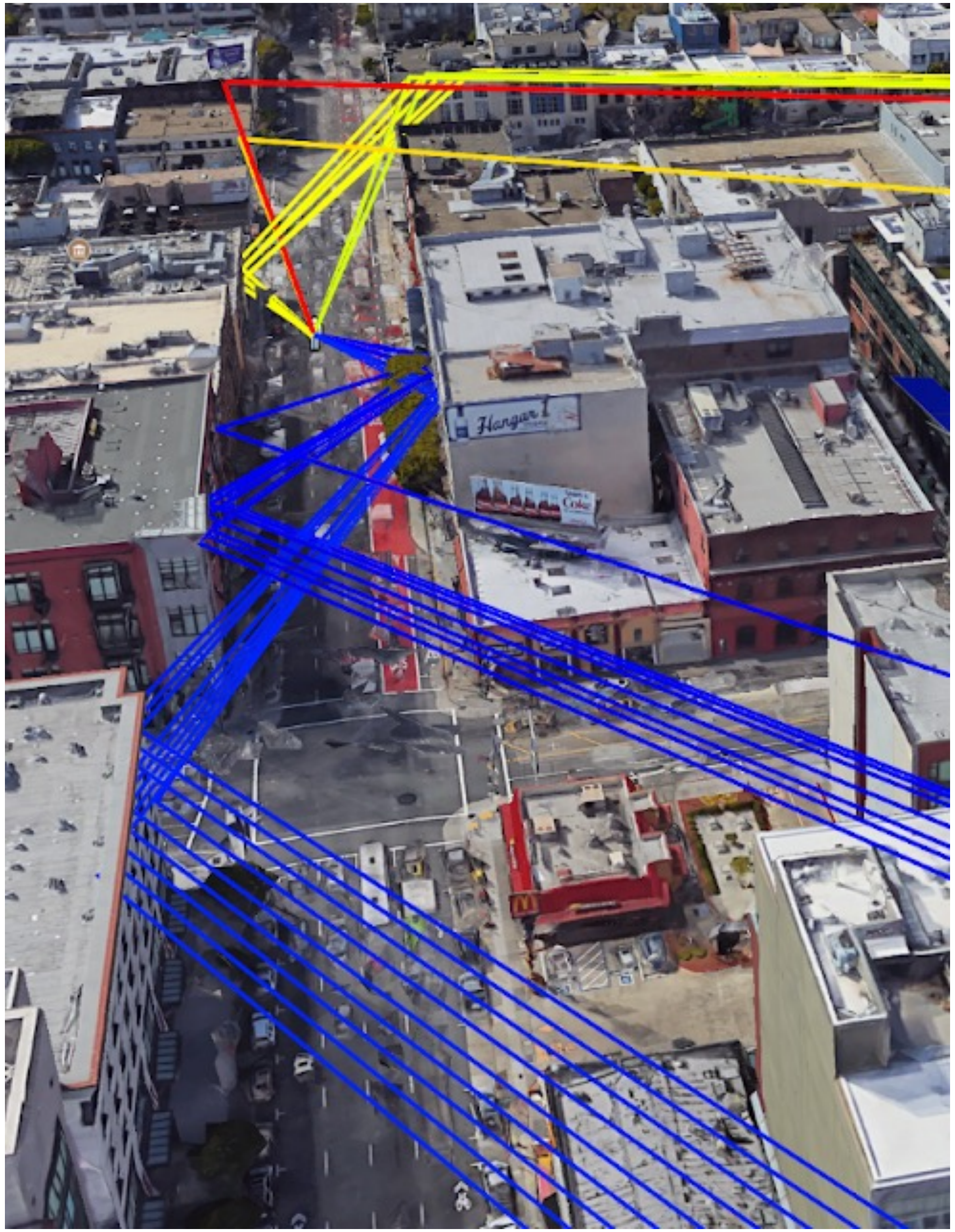

Meta Aria, Microsoft HoloLens, Magic Leap 2

# GPS+compass is not enough

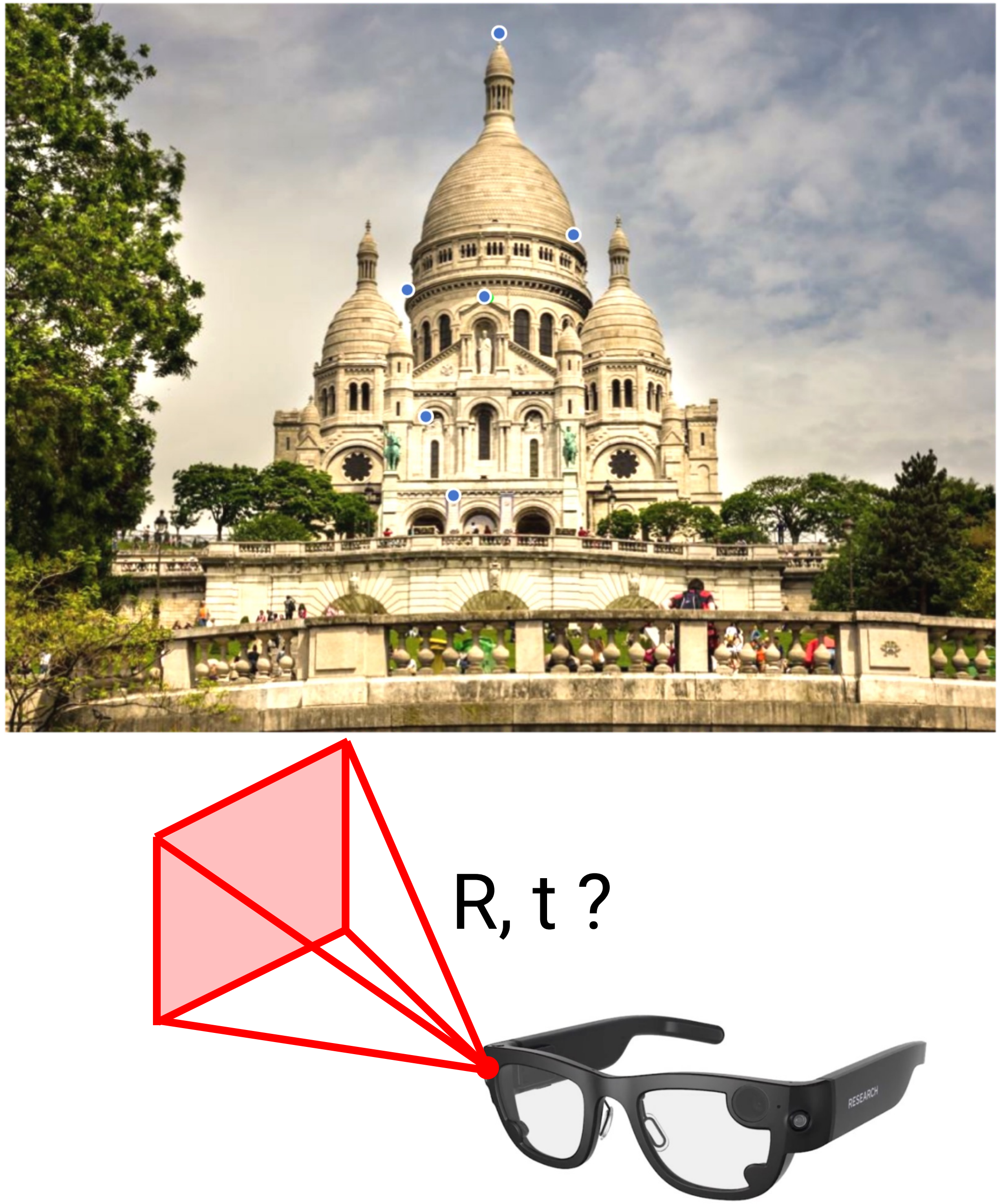
- Low accuracy
- Only 3 DoF
- Commonly unreliable: urban canyon, metal structures



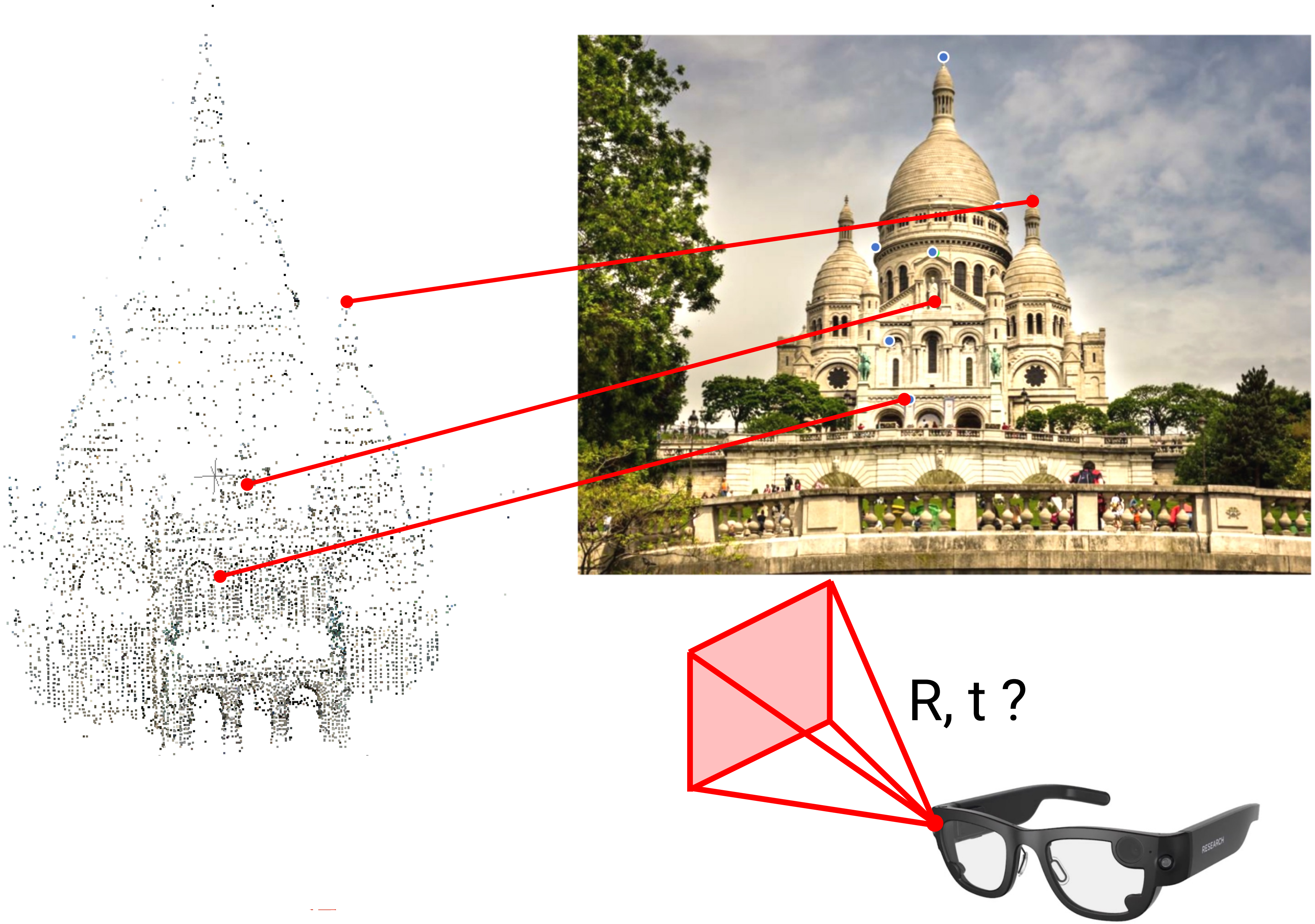



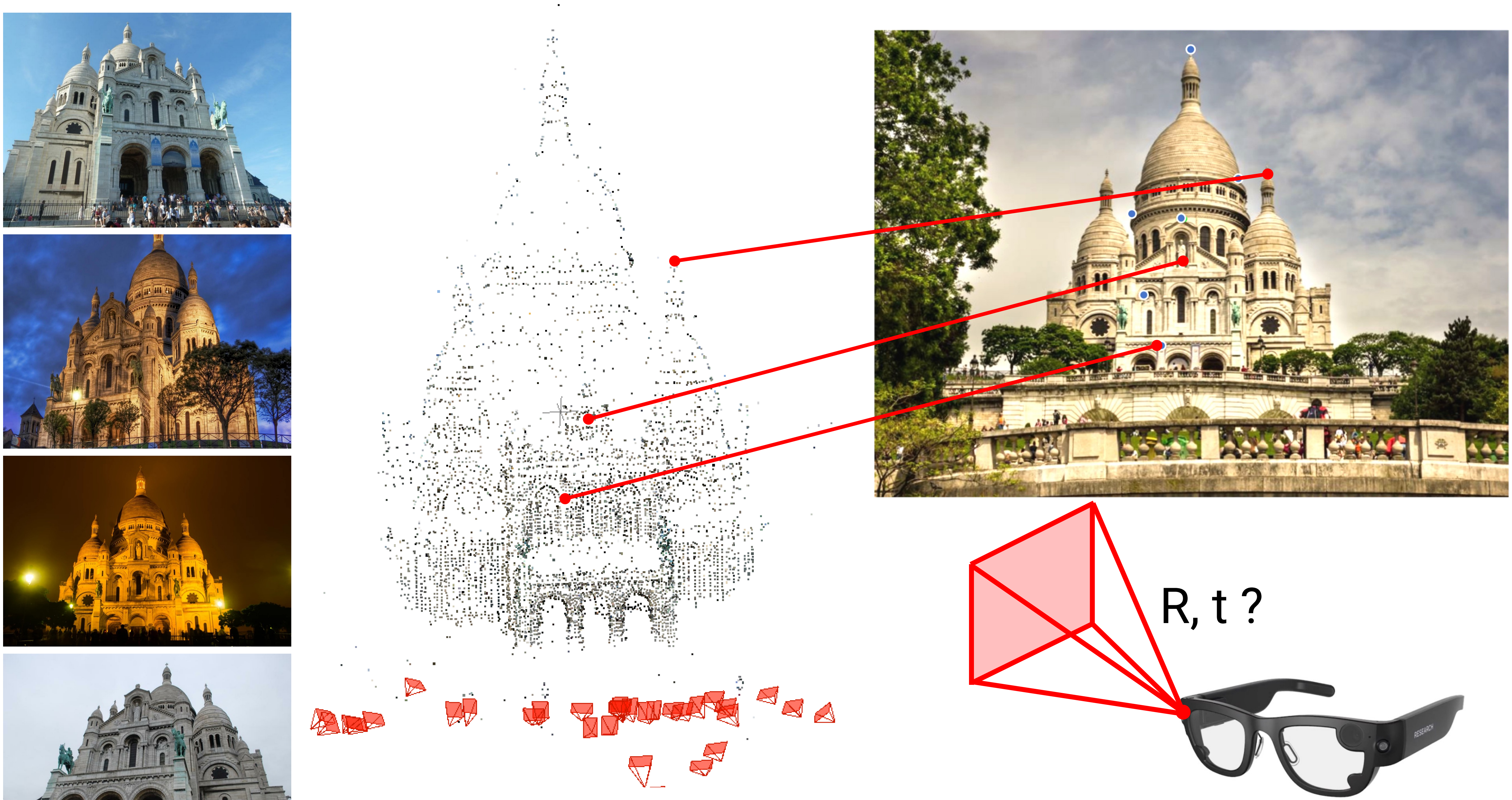
Google Maps



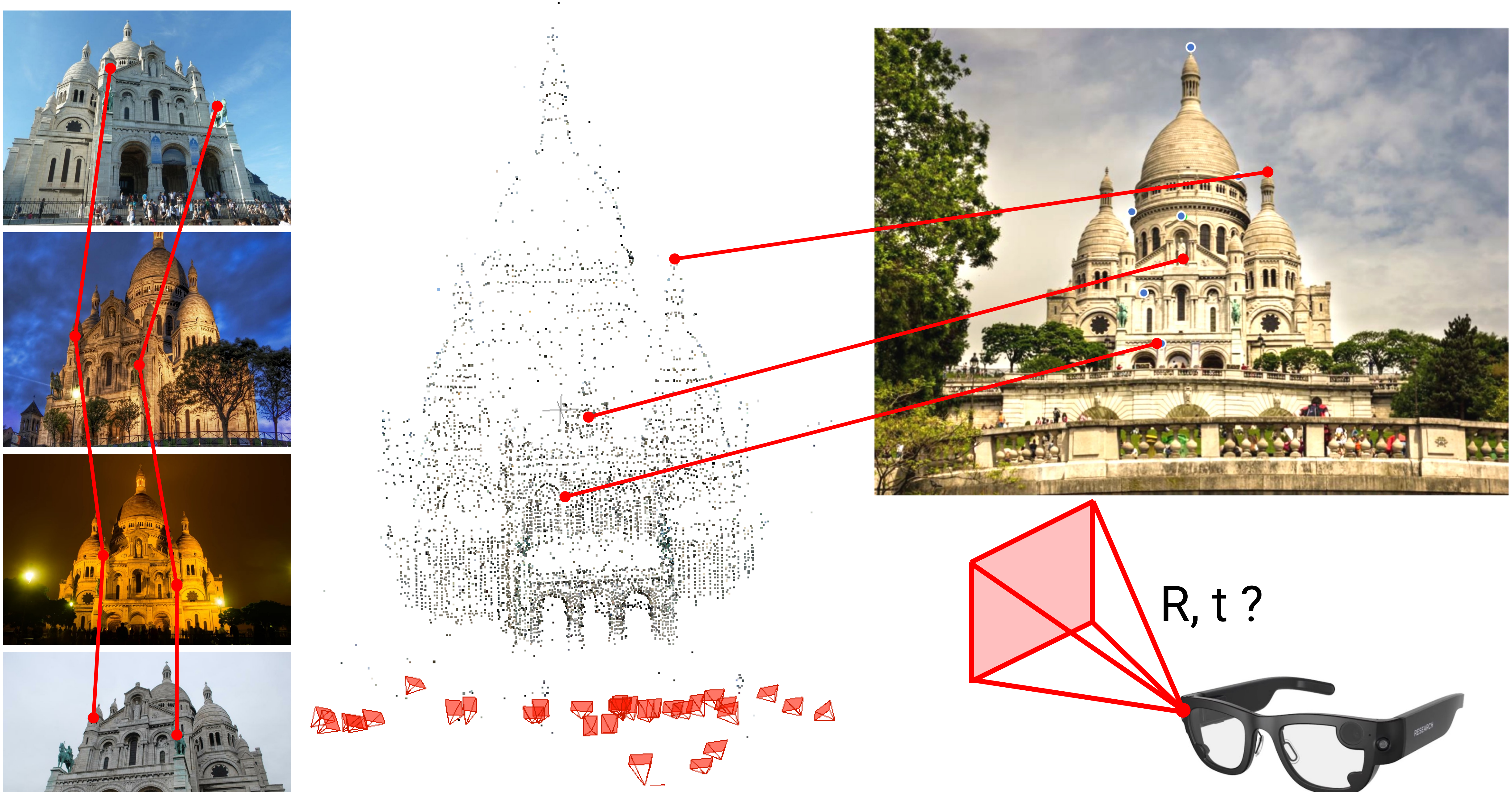

# 6-DoF Localization




6-DoF Localization

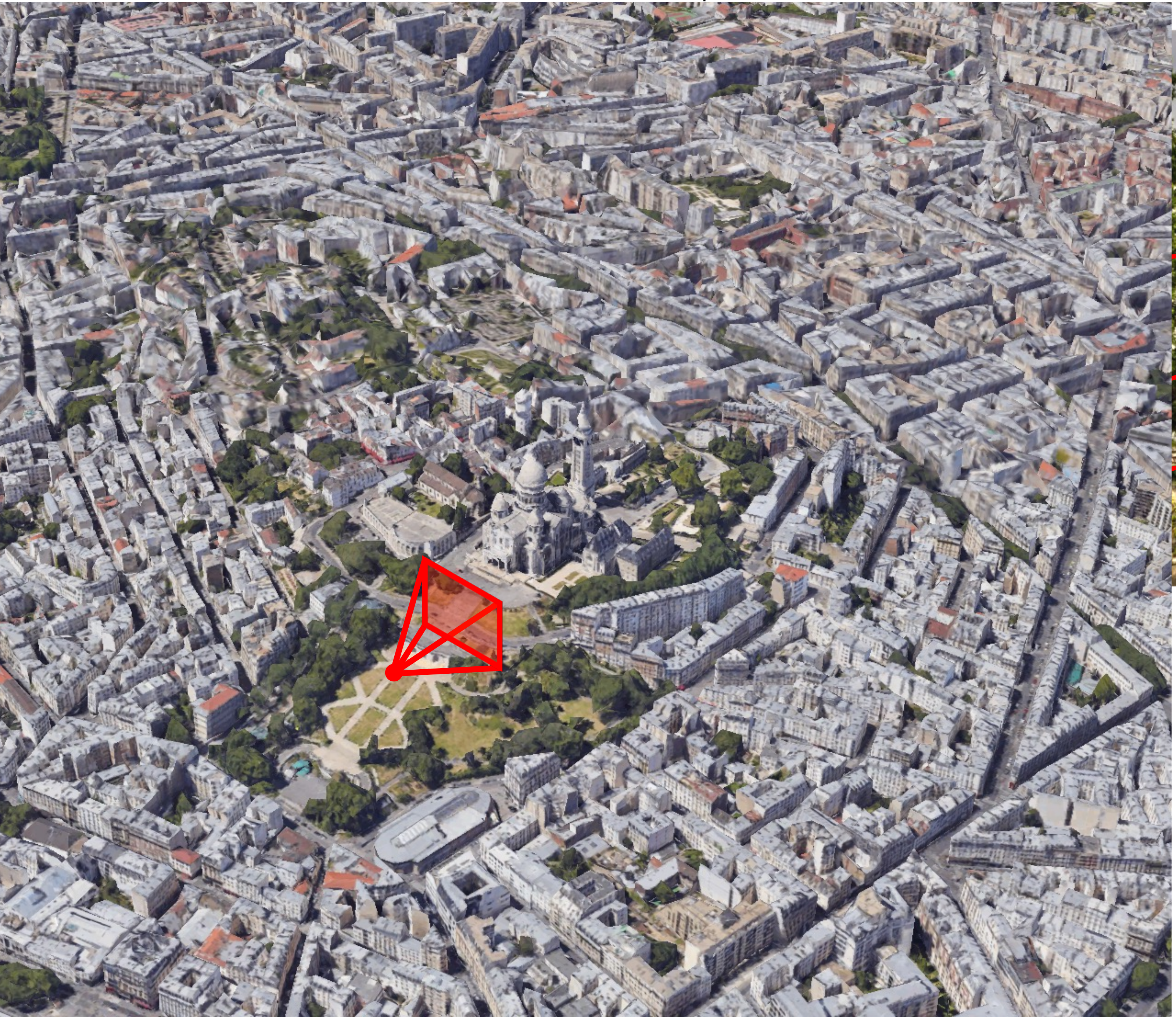


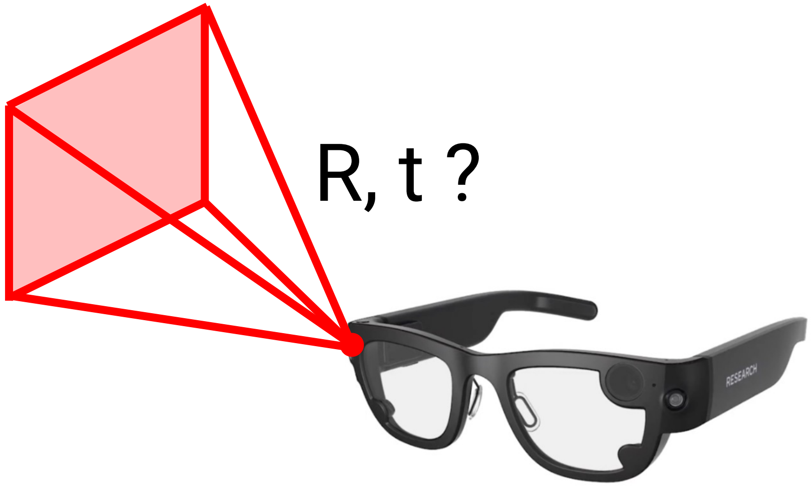

6-DoF Localization




Structure-from-Motion

6-DoF Localization






6-DoF Localization

# Limitations of 3D maps

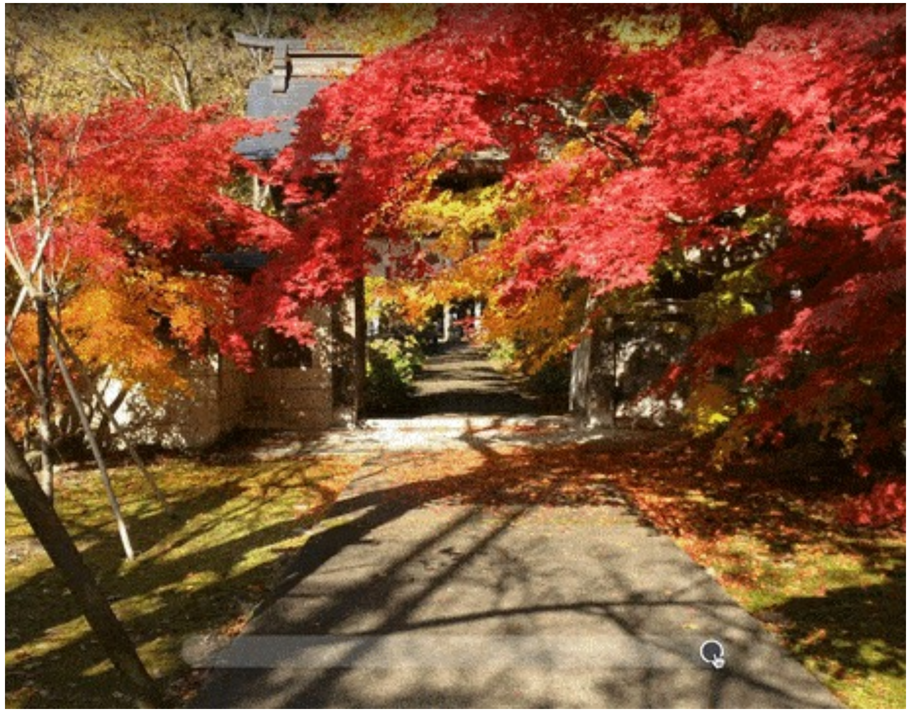
**Build & update**

Storage

Privacy

Mapping fleet
Frequent updates


Google StreetView



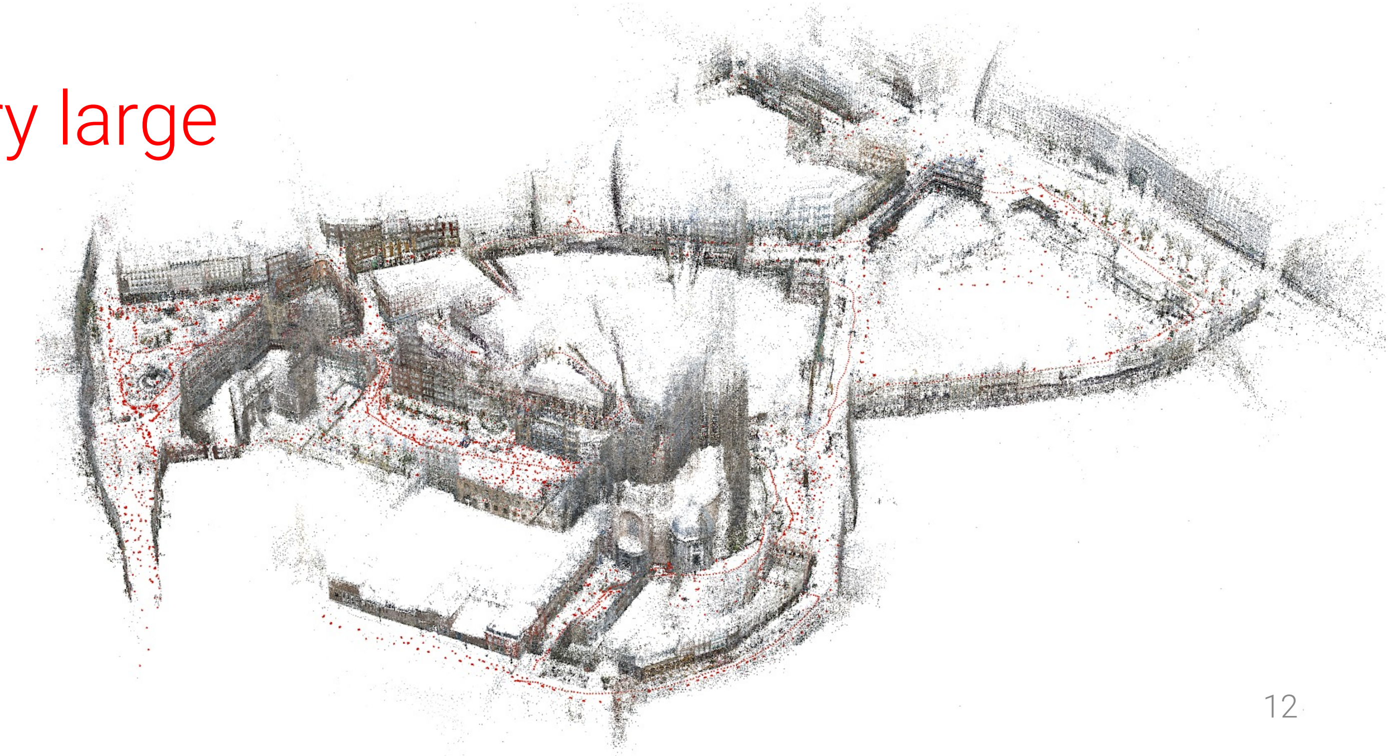
Mapillary

# Limitations of 3D maps

Build
& update

**Storage**

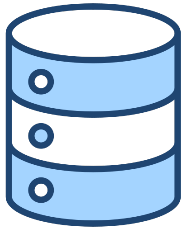
Very large



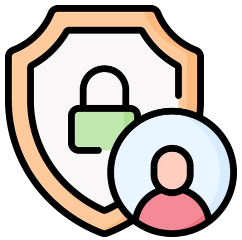
Privacy

# Limitations of 3D maps

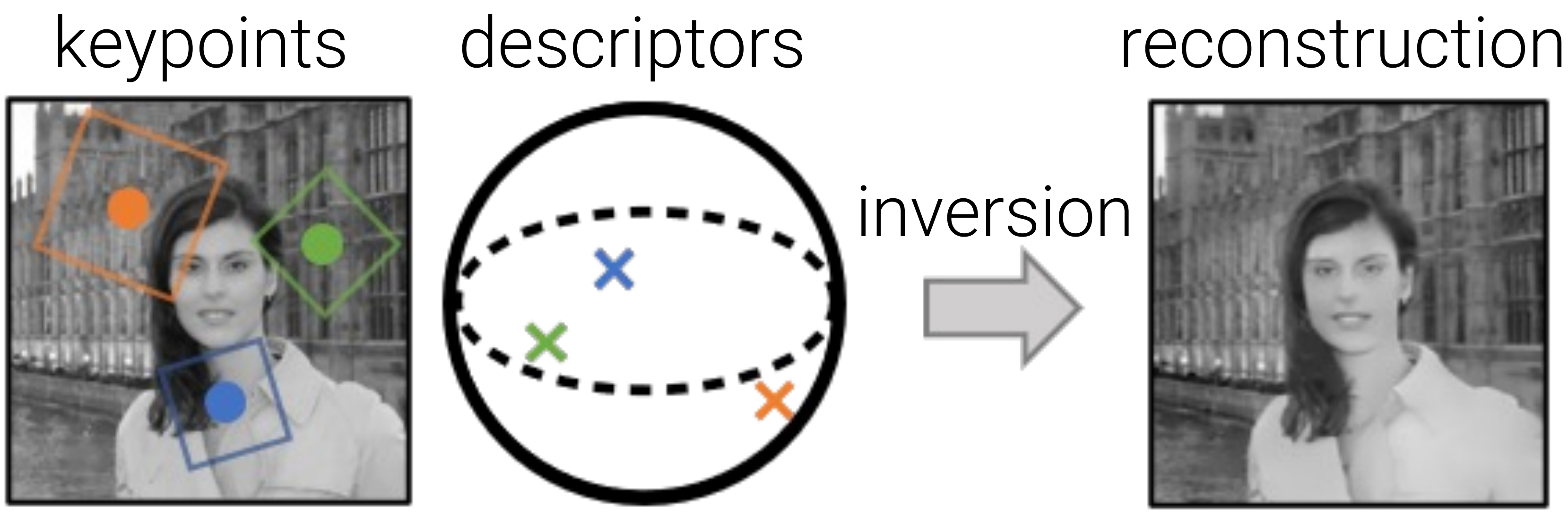
Build
& update



Storage



**Privacy**




Mihai Dusmanu

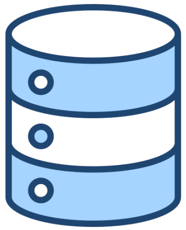
Risk of inversion

# Limitations of 3D maps

Build
& update

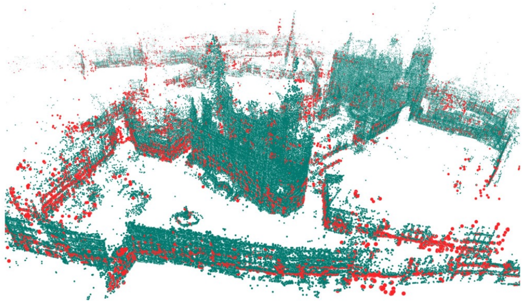
Mapping fleet
Frequent updates

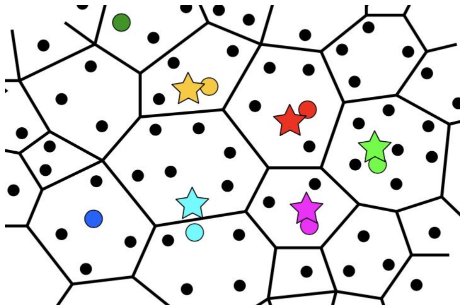
Compression
& Quantization

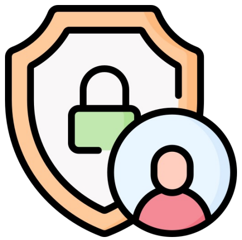


Storage

Very large





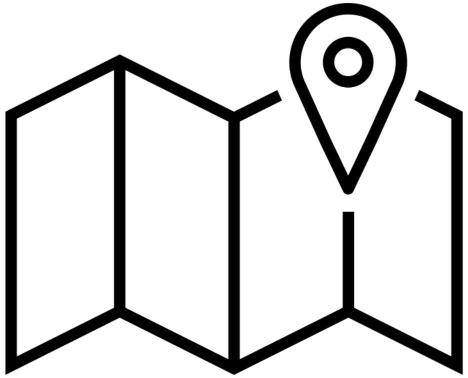


Privacy

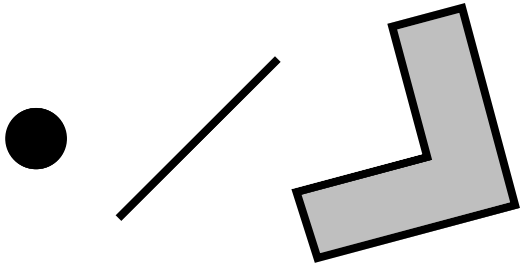
Risk of inversion


Privacy-preserving descriptors

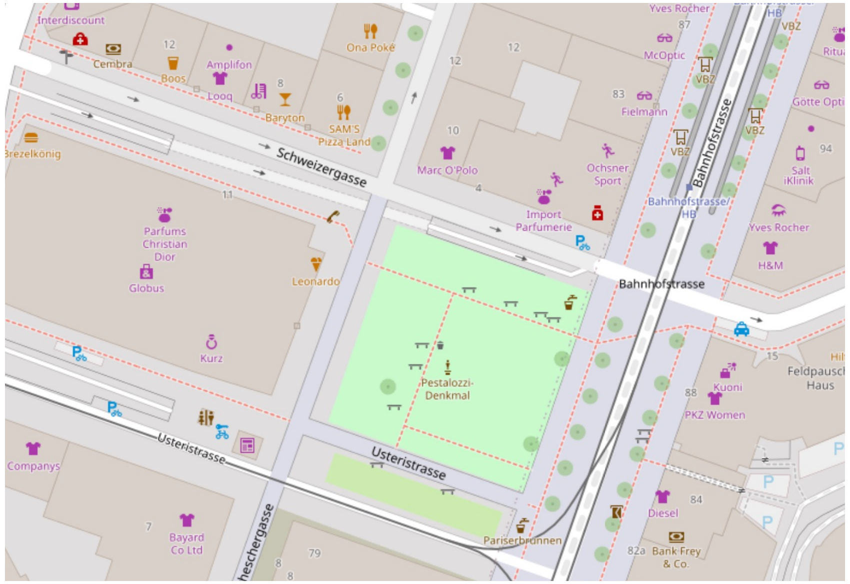
# Semantic 2D maps



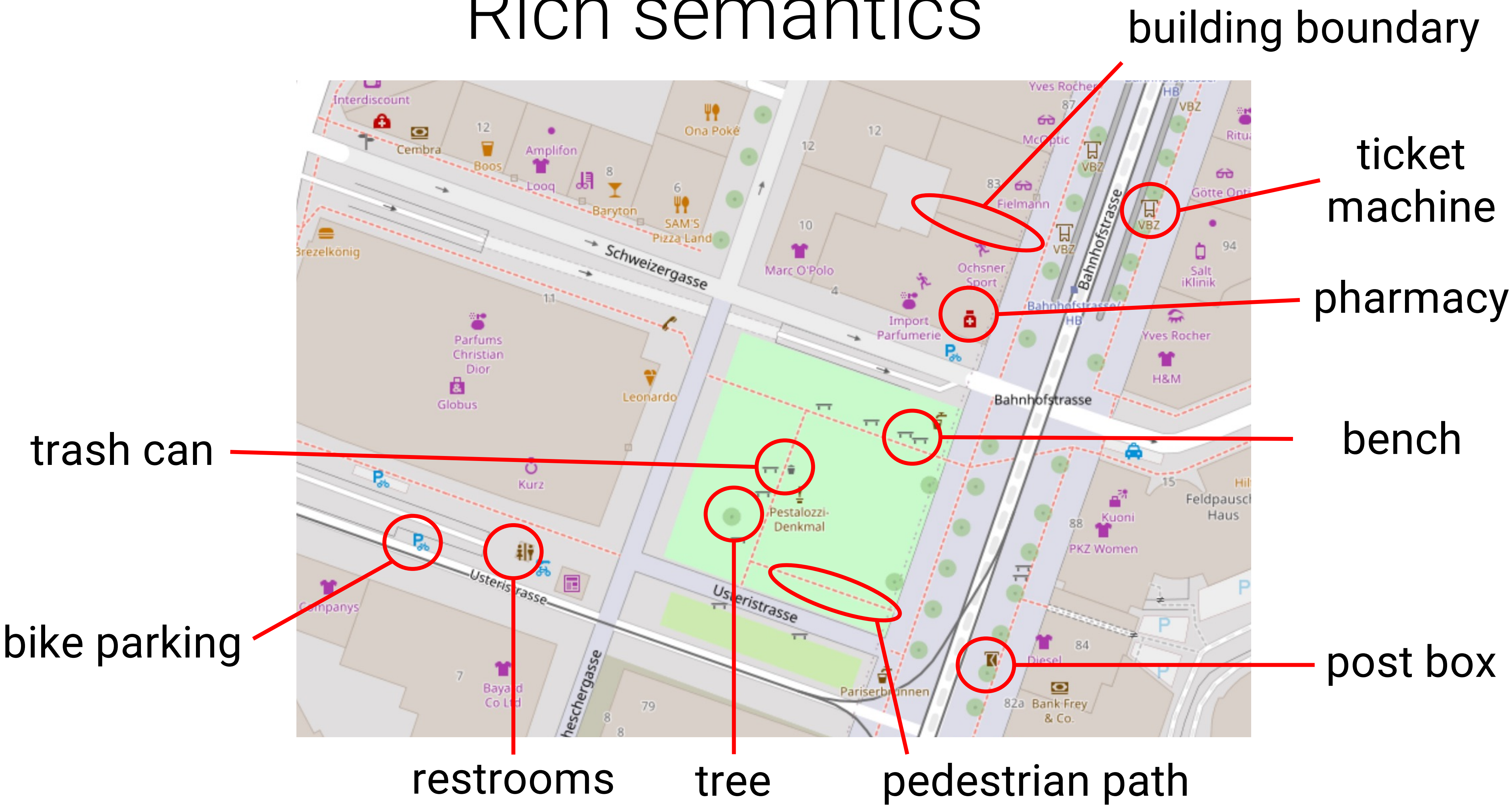
Planimetric



OpenStreetMap



# Rich semantics

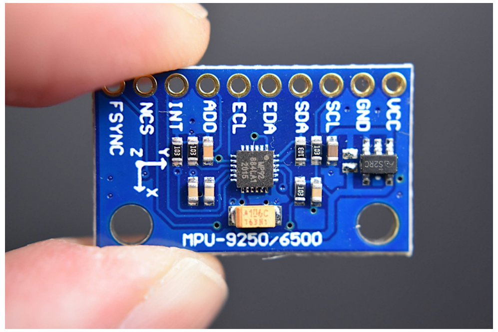


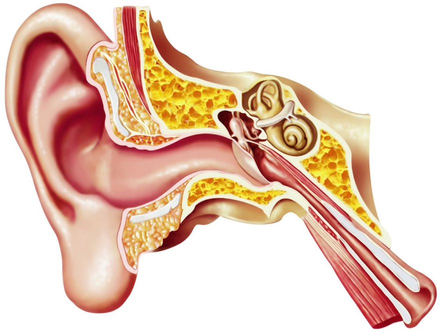
# Rich semantics

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

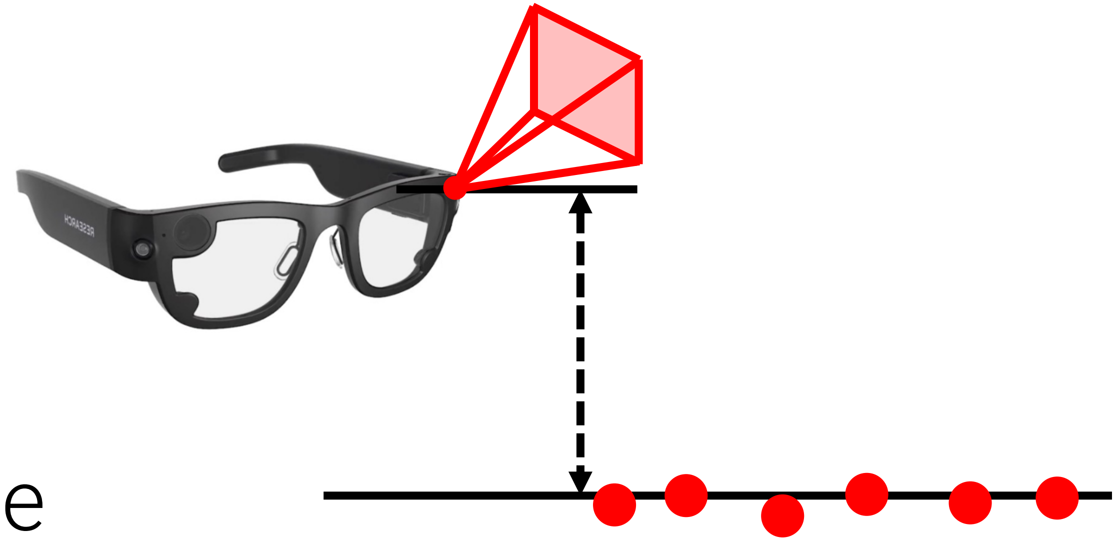
# Simplifying assumptions

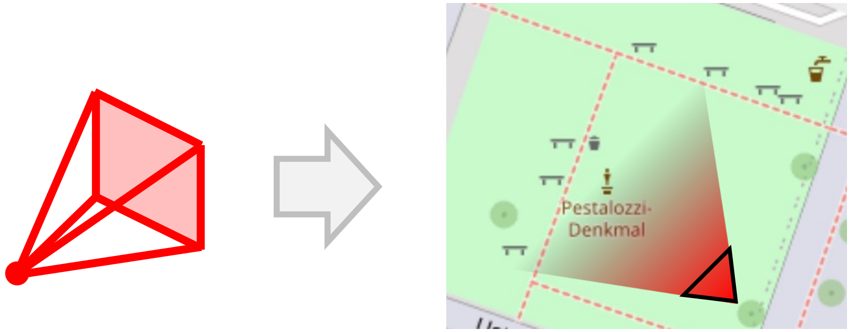
- Known gravity direction



- Unnecessary vertical position



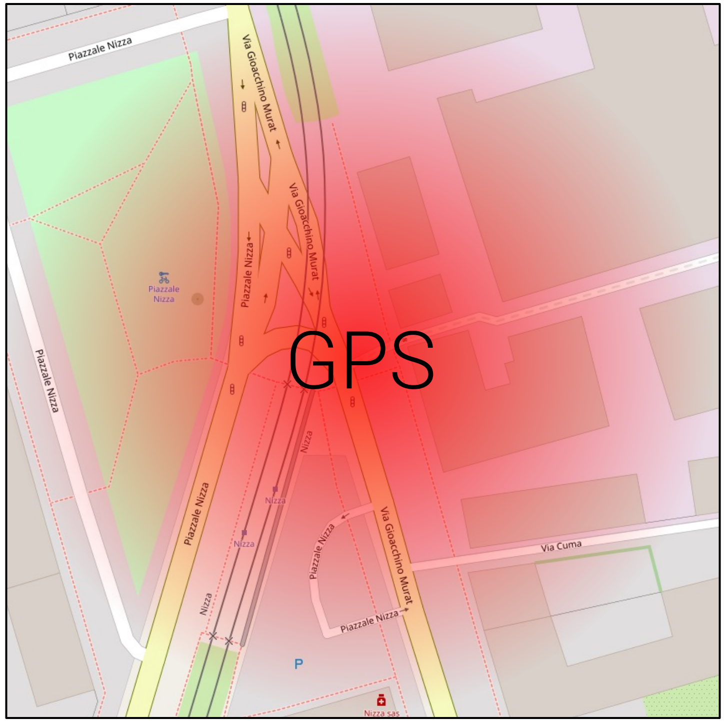



3-DoF pose
$(x, y, \theta)$

# Problem setup

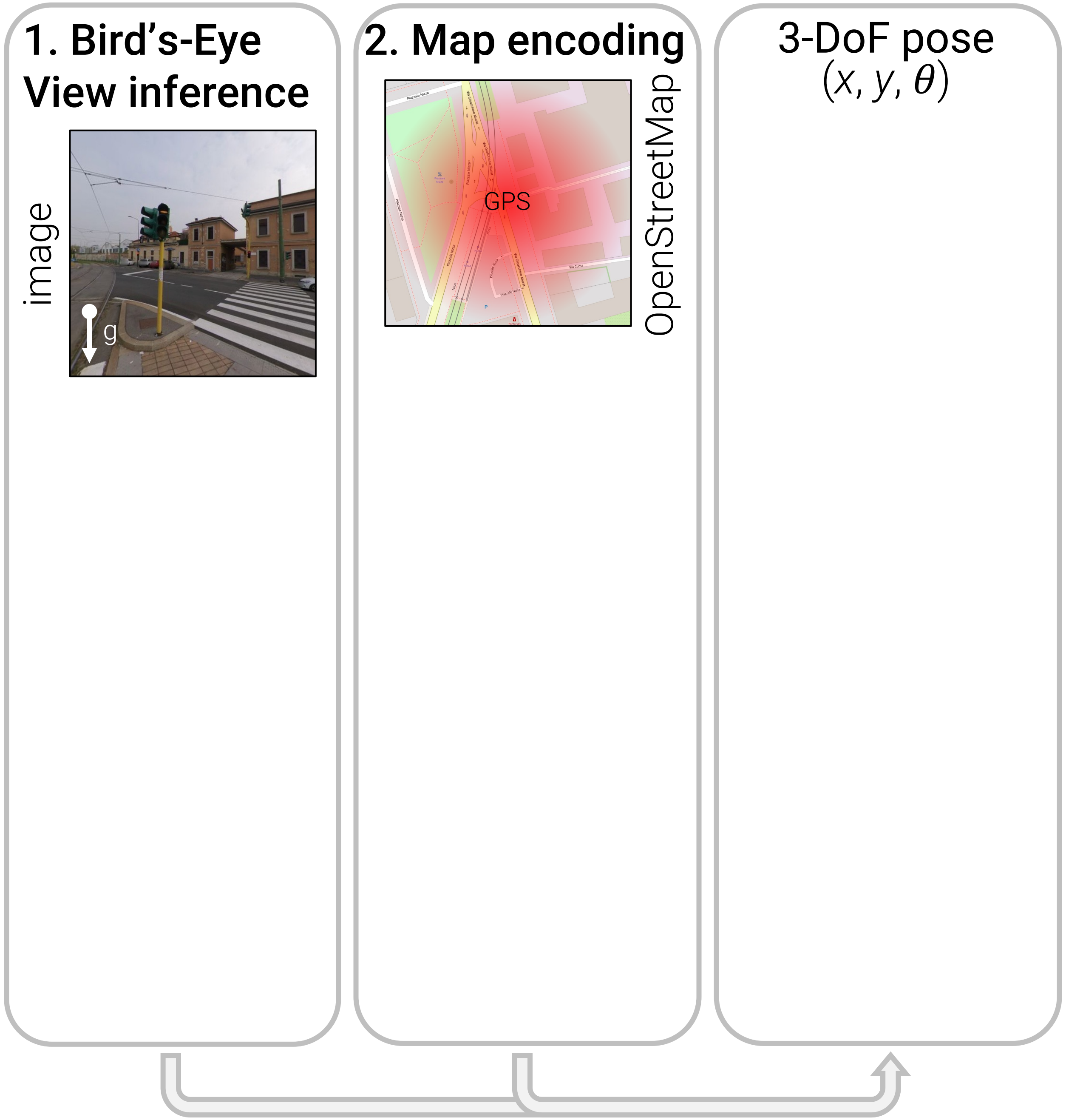

image
+ gravity

128m x 128m

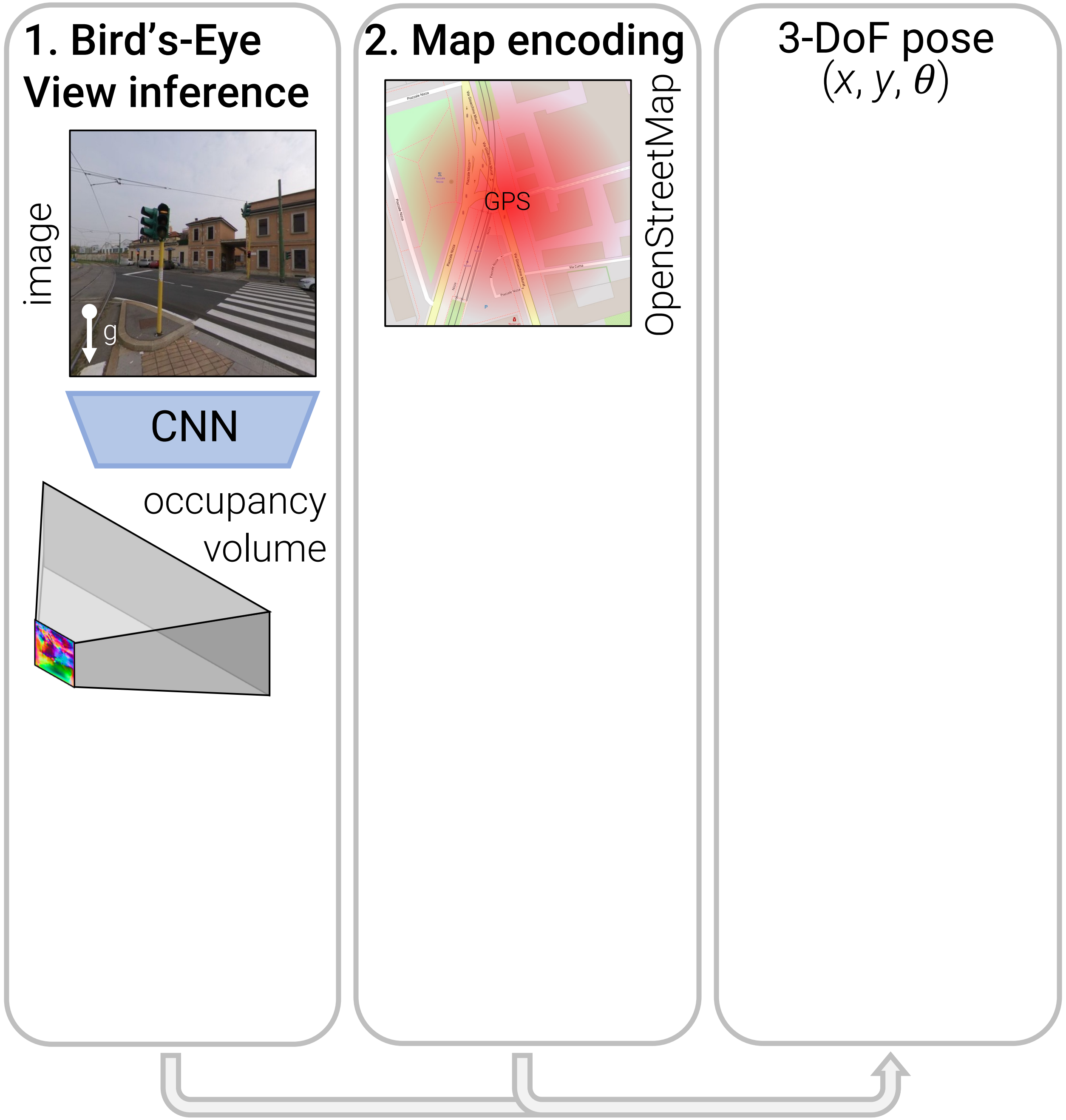



OpenStreetMap

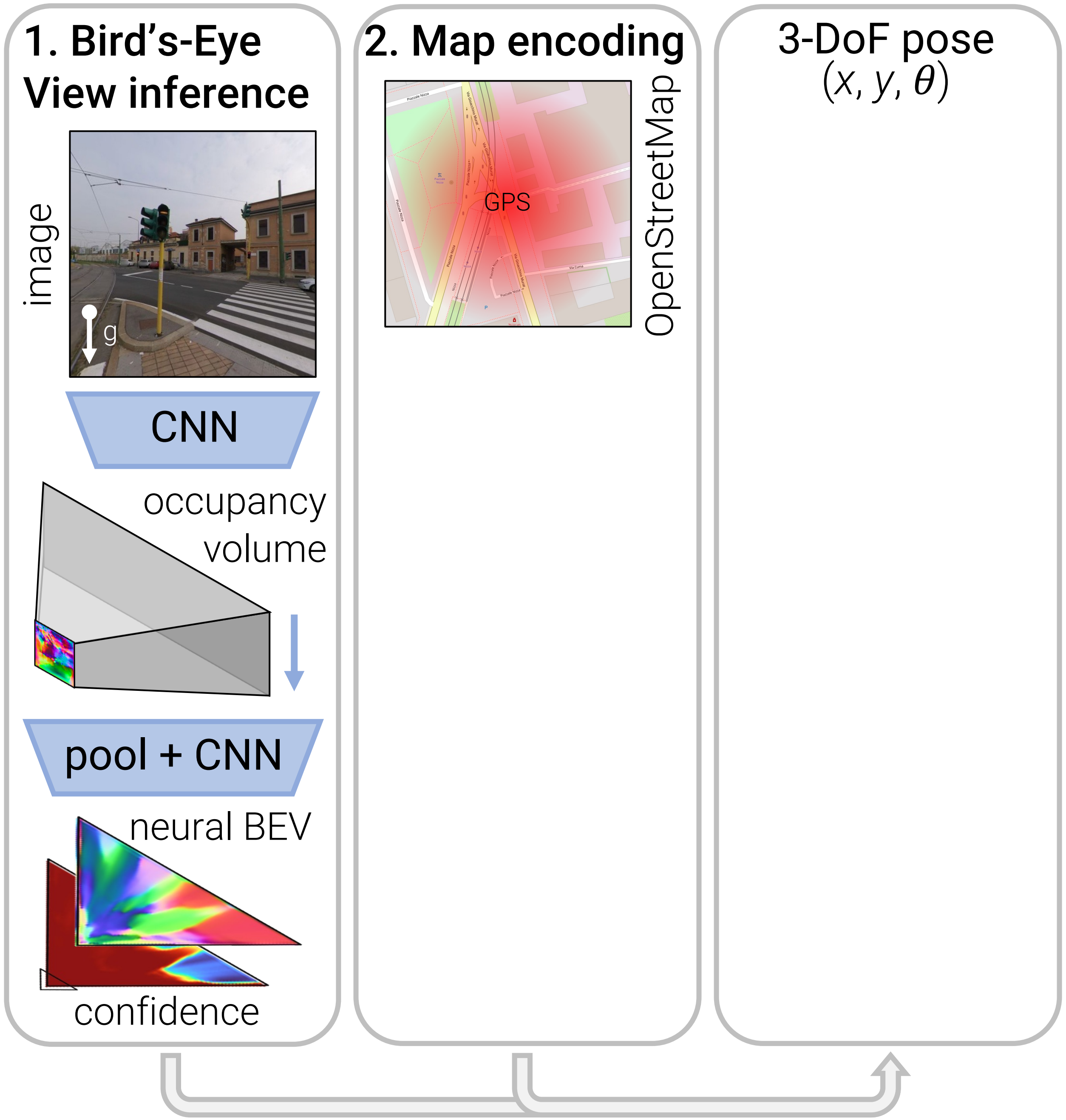
# The OrienterNet architecture



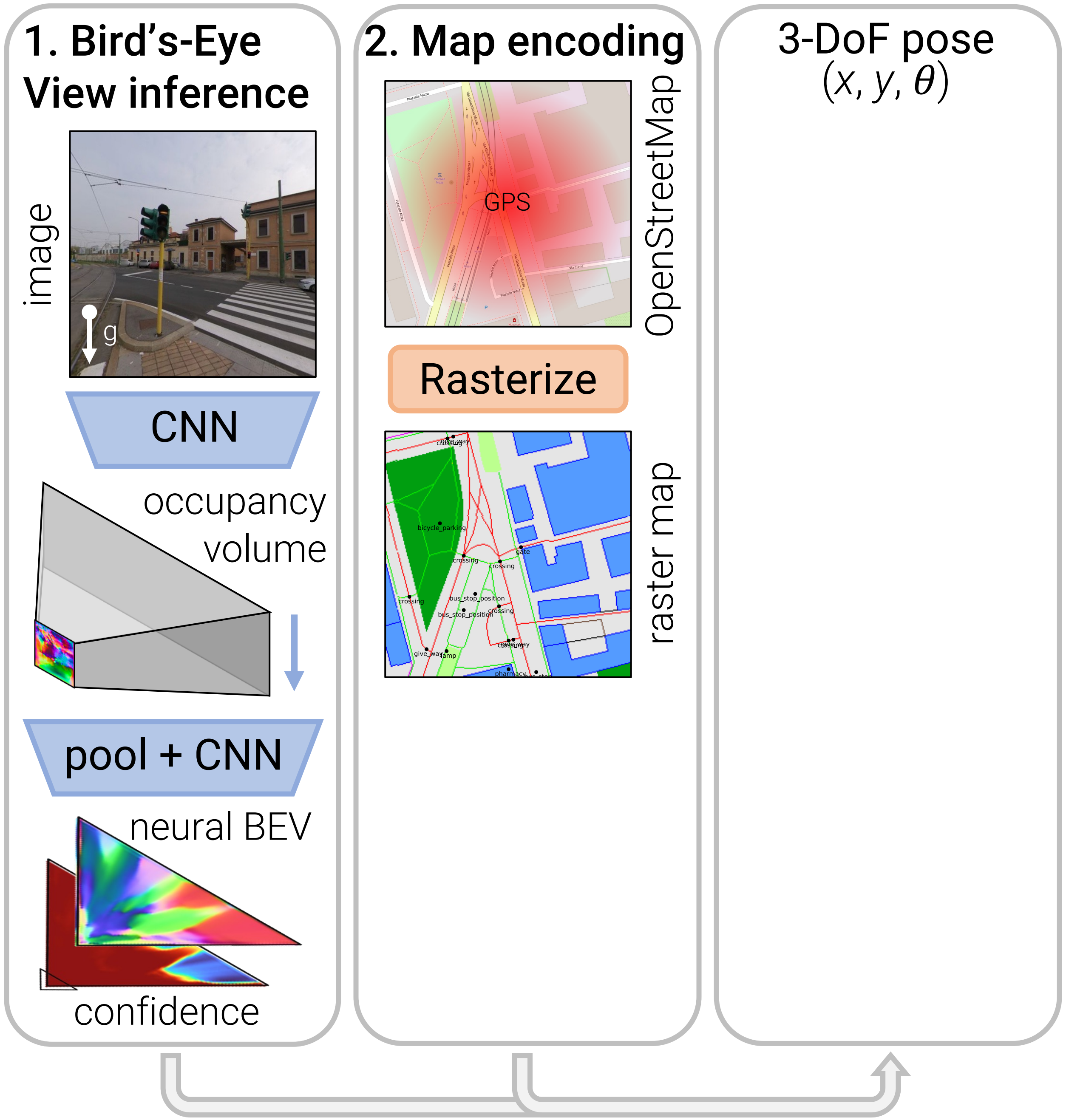
# The OrienterNet architecture



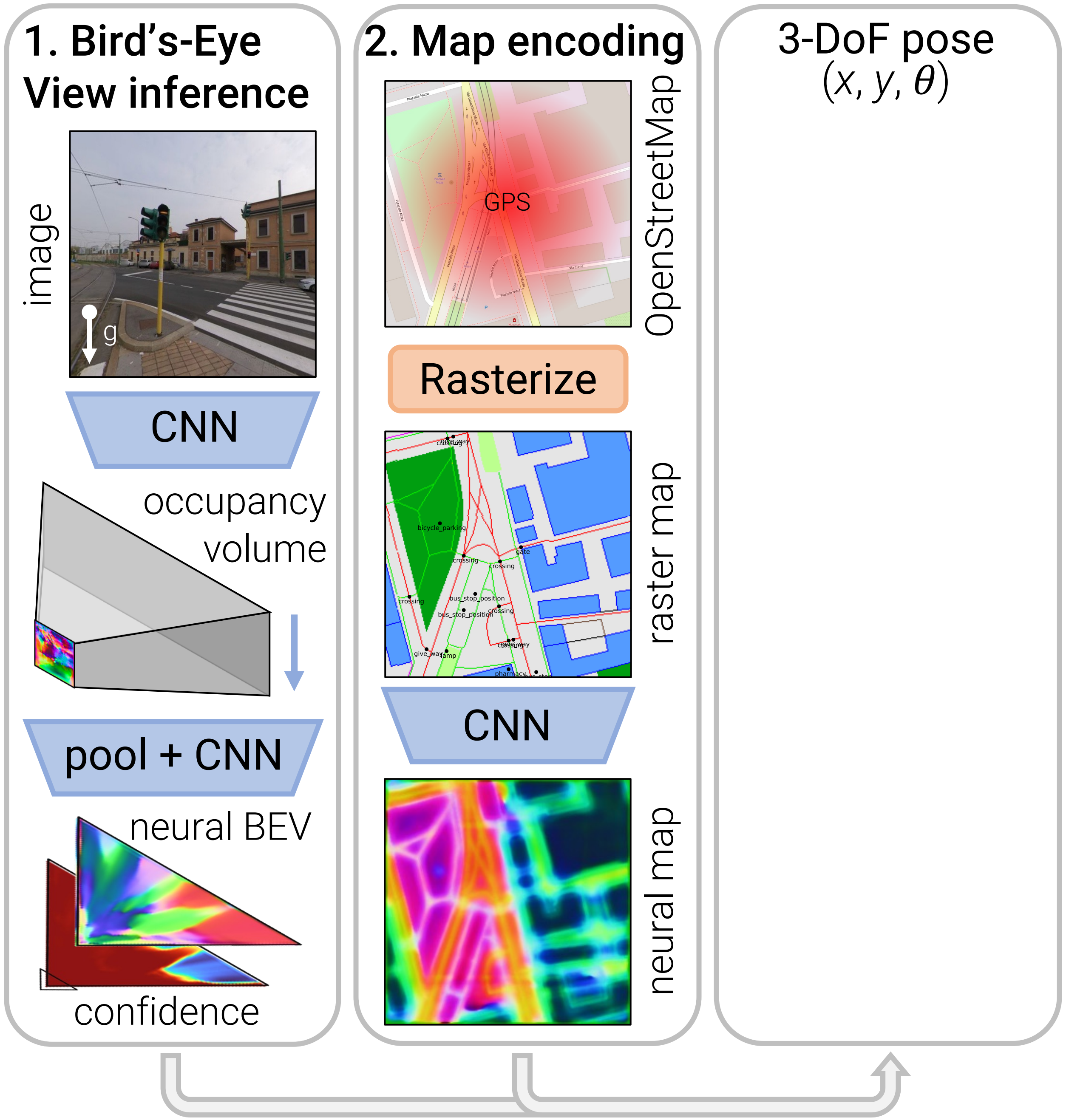
# The OrienterNet architecture



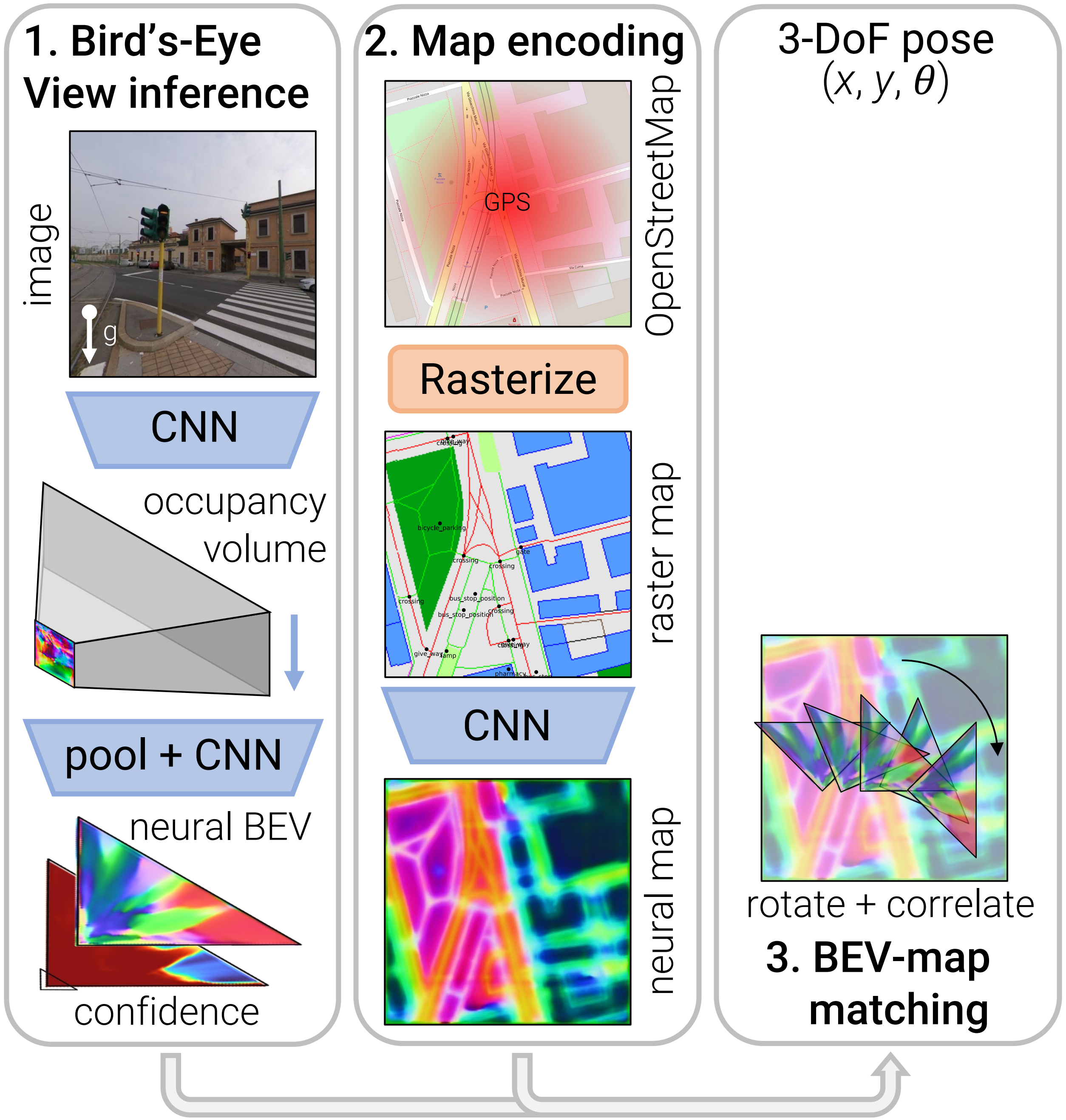
# The OrienterNet architecture



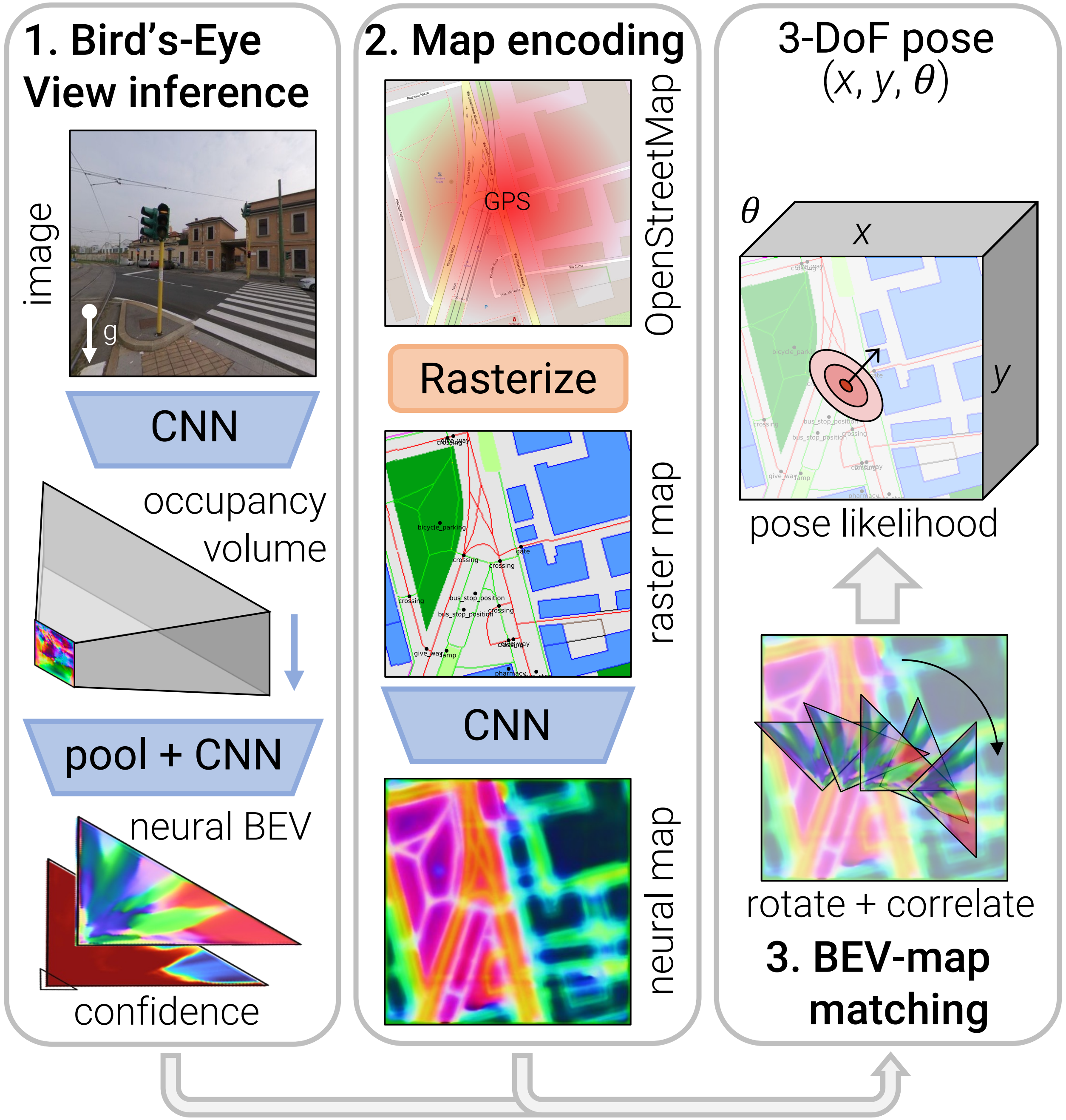
# The OrienterNet architecture



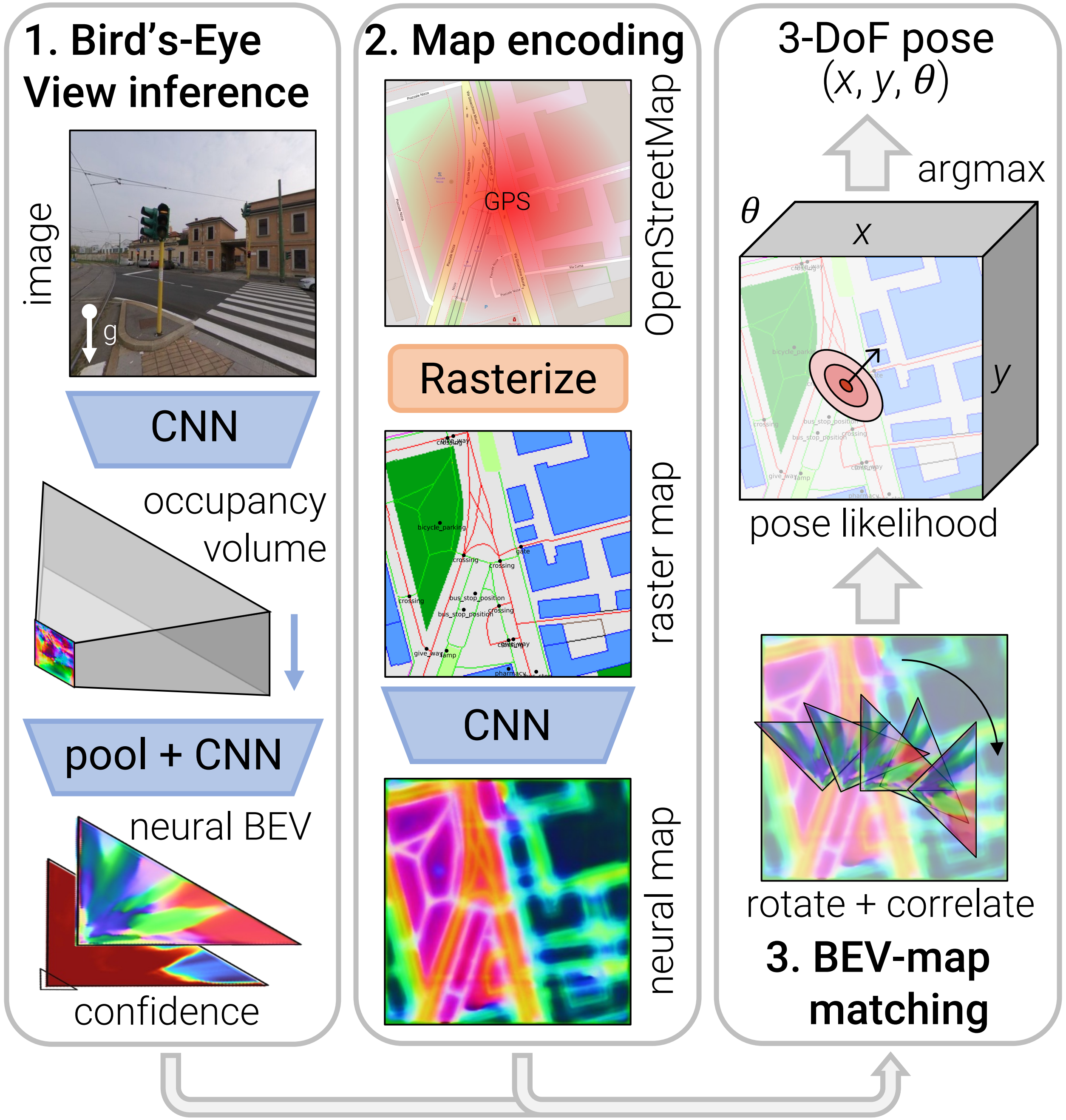
# The OrienterNet architecture



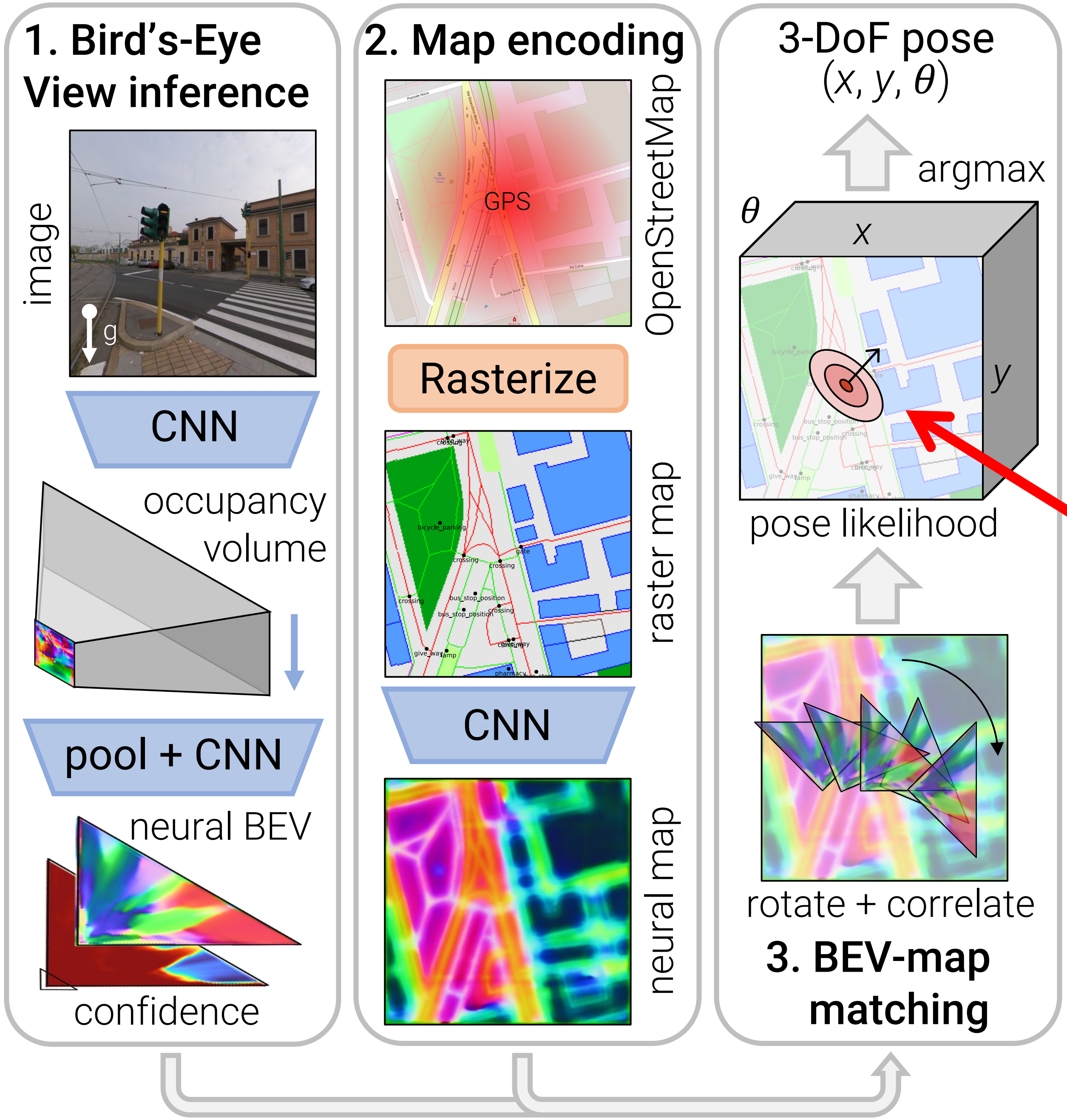
# The OrienterNet architecture



# The OrienterNet architecture



# The OrienterNet architecture



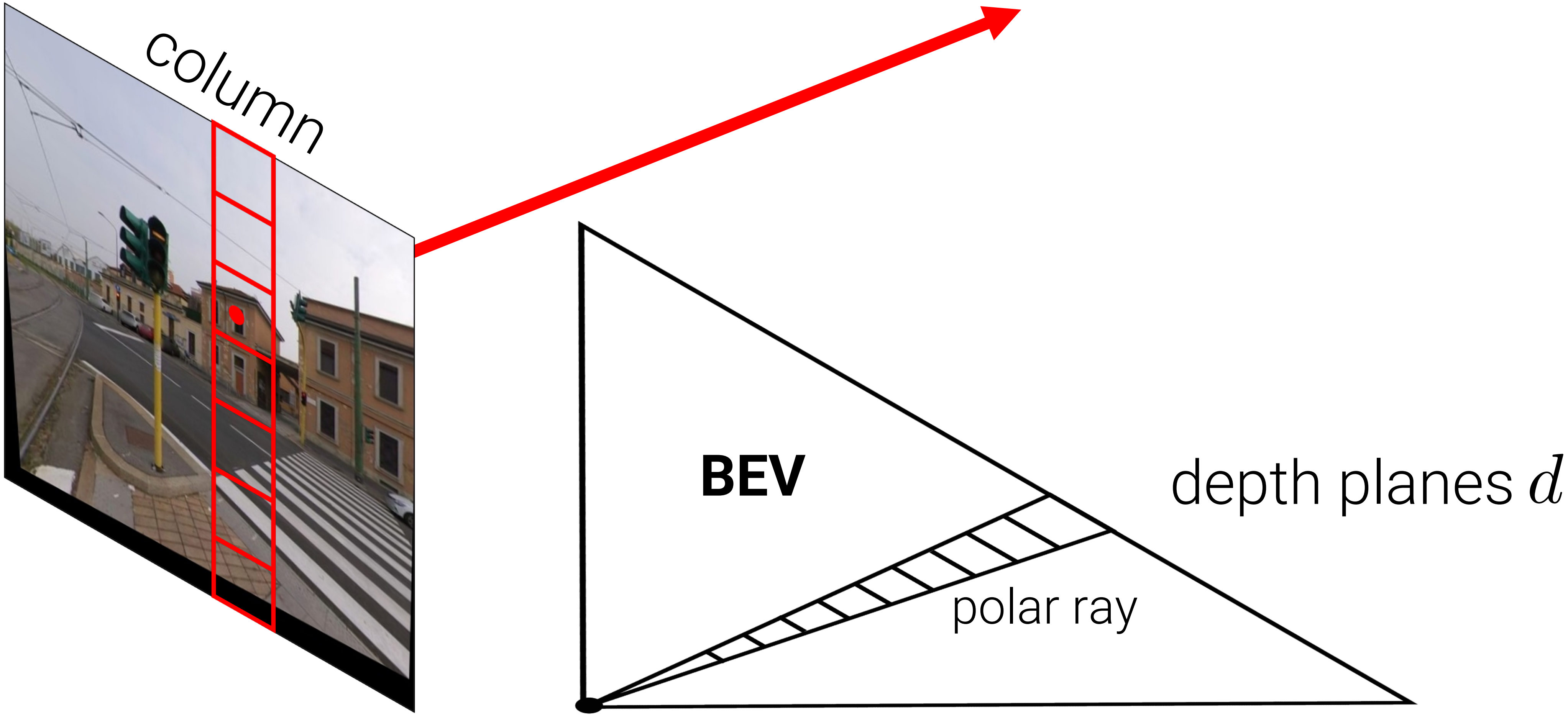

Maximize likelihood of GT pose

# 1. Bird's Eye View inference

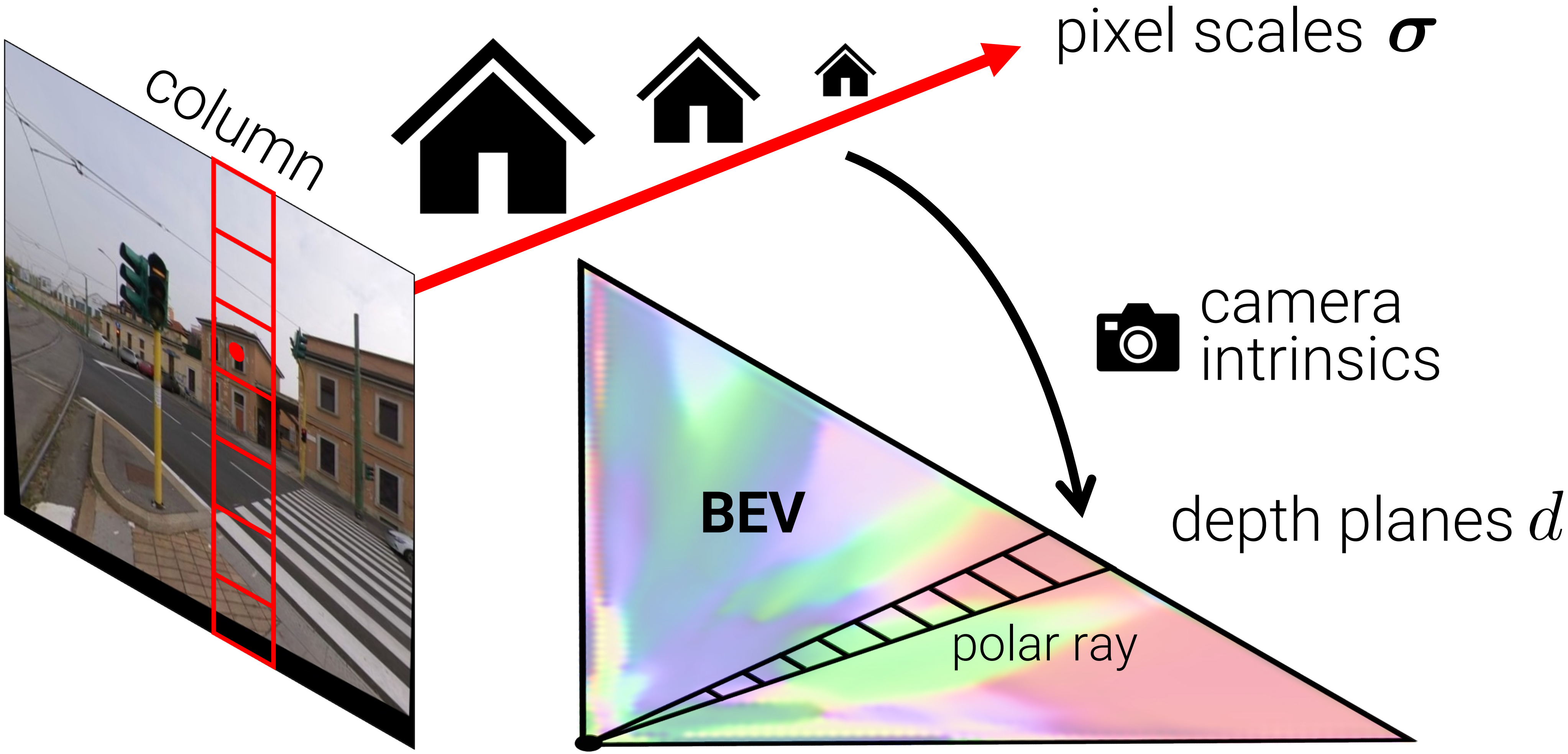
# 1. Bird's Eye View inference

gravity-aligned

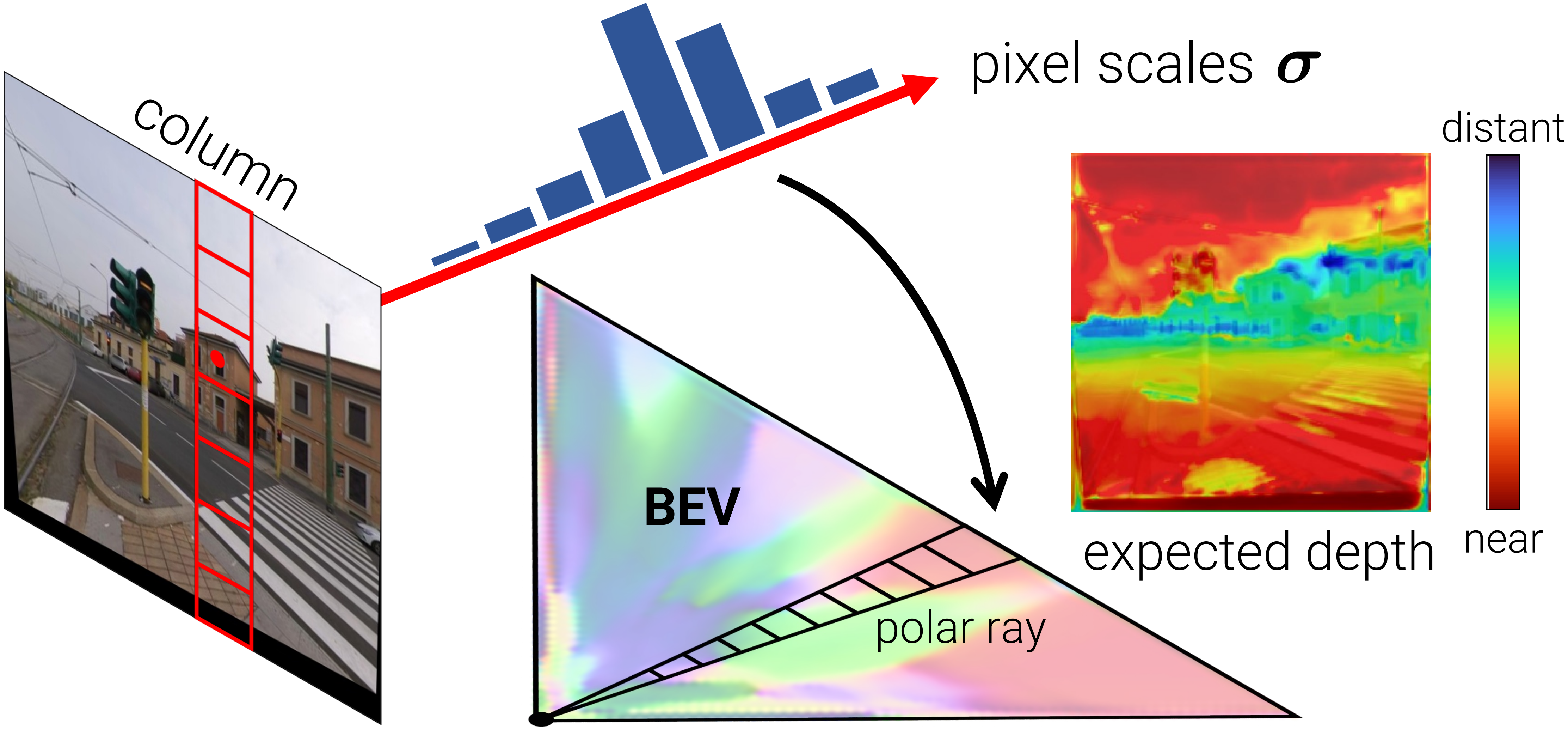
# 1. Bird’s Eye View inference



# 1. Bird's Eye View inference

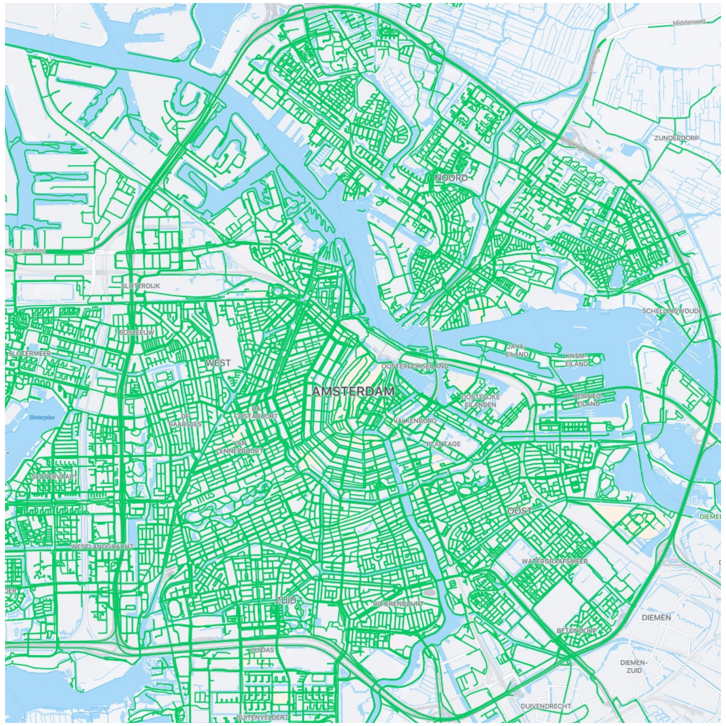


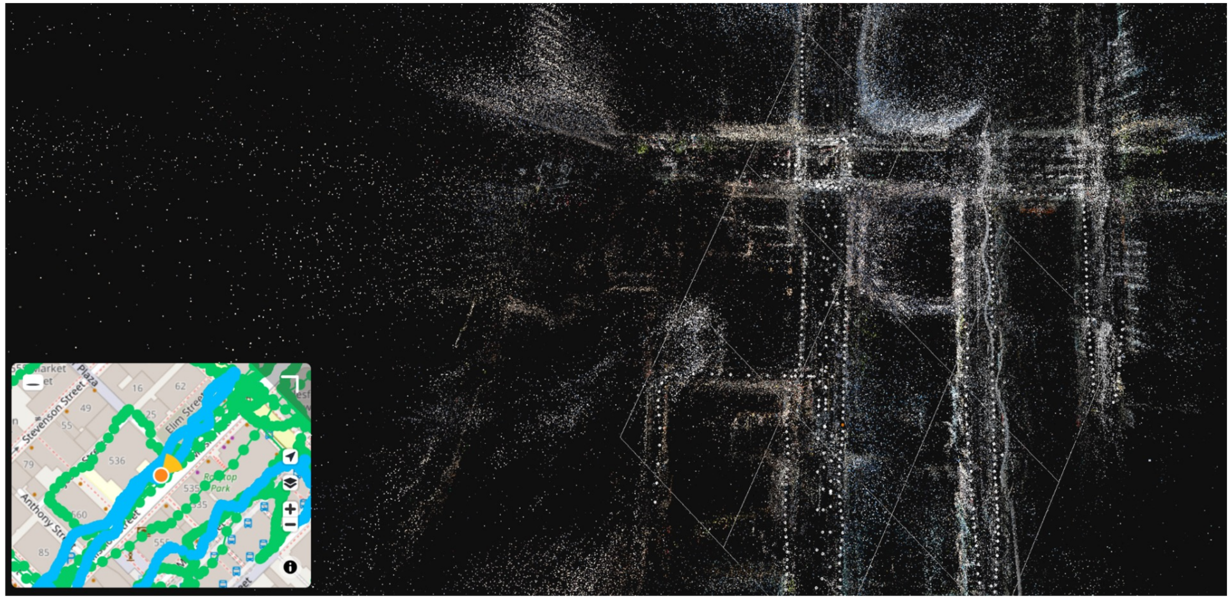
# 1. Bird's Eye View inference

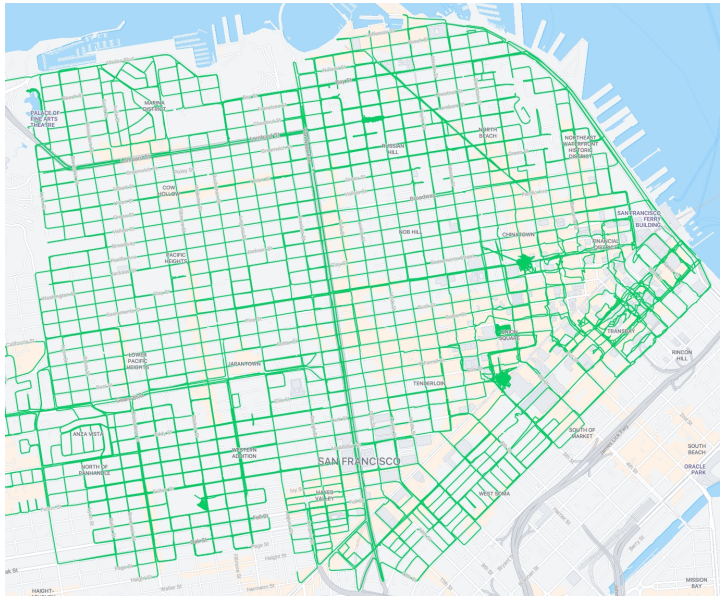


# Training a single strong model

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



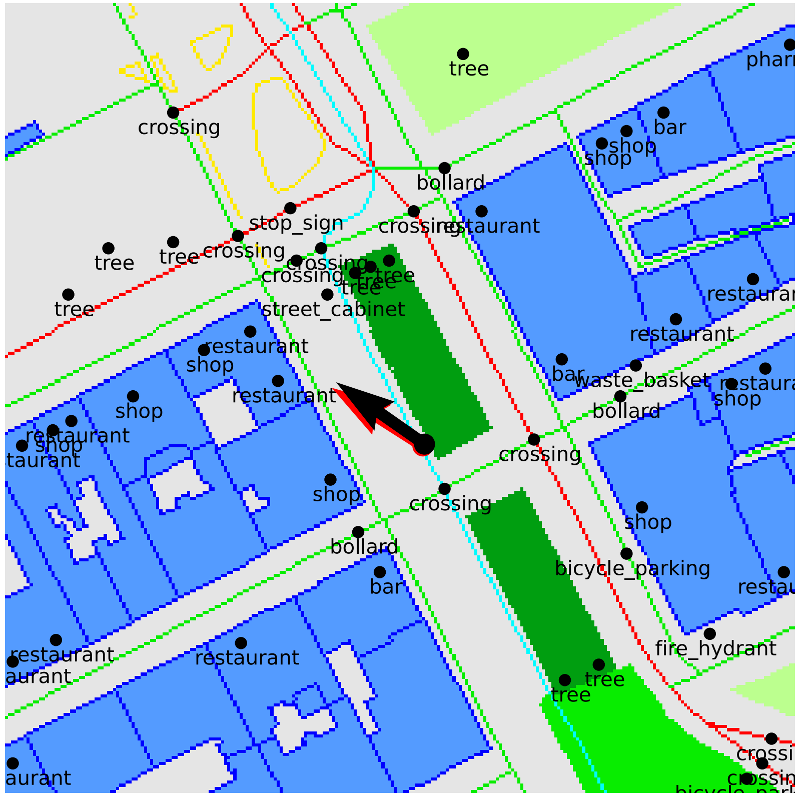



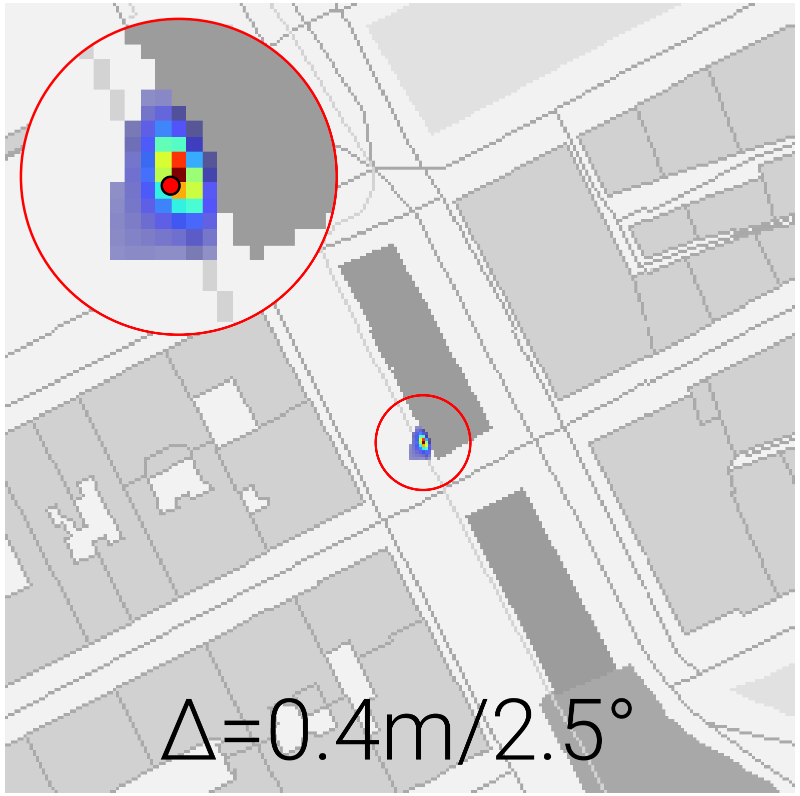


input image

raster map

likelihood



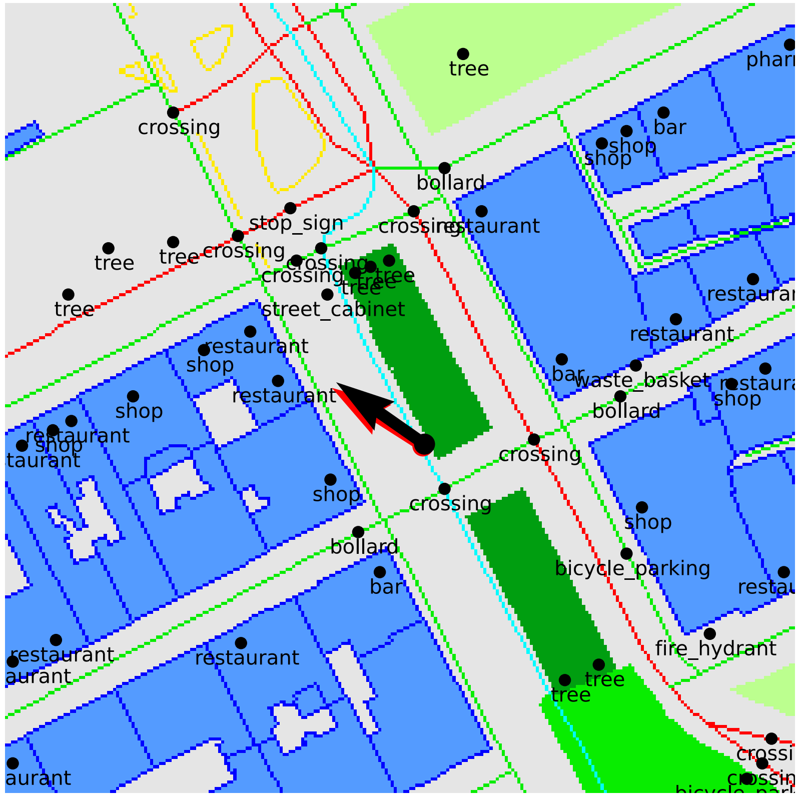




ground truth prediction

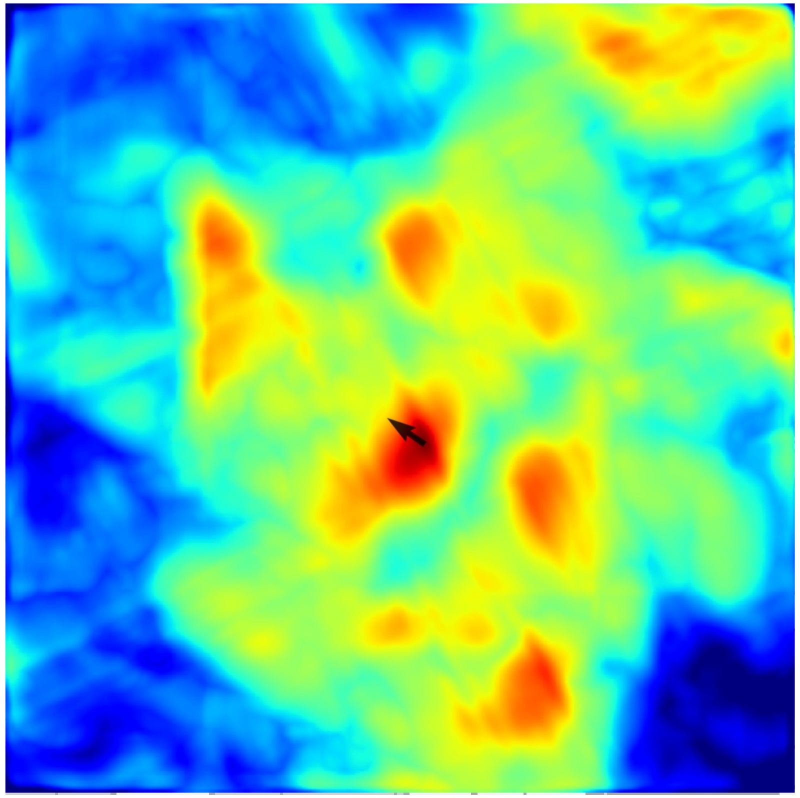
building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

input image

raster map

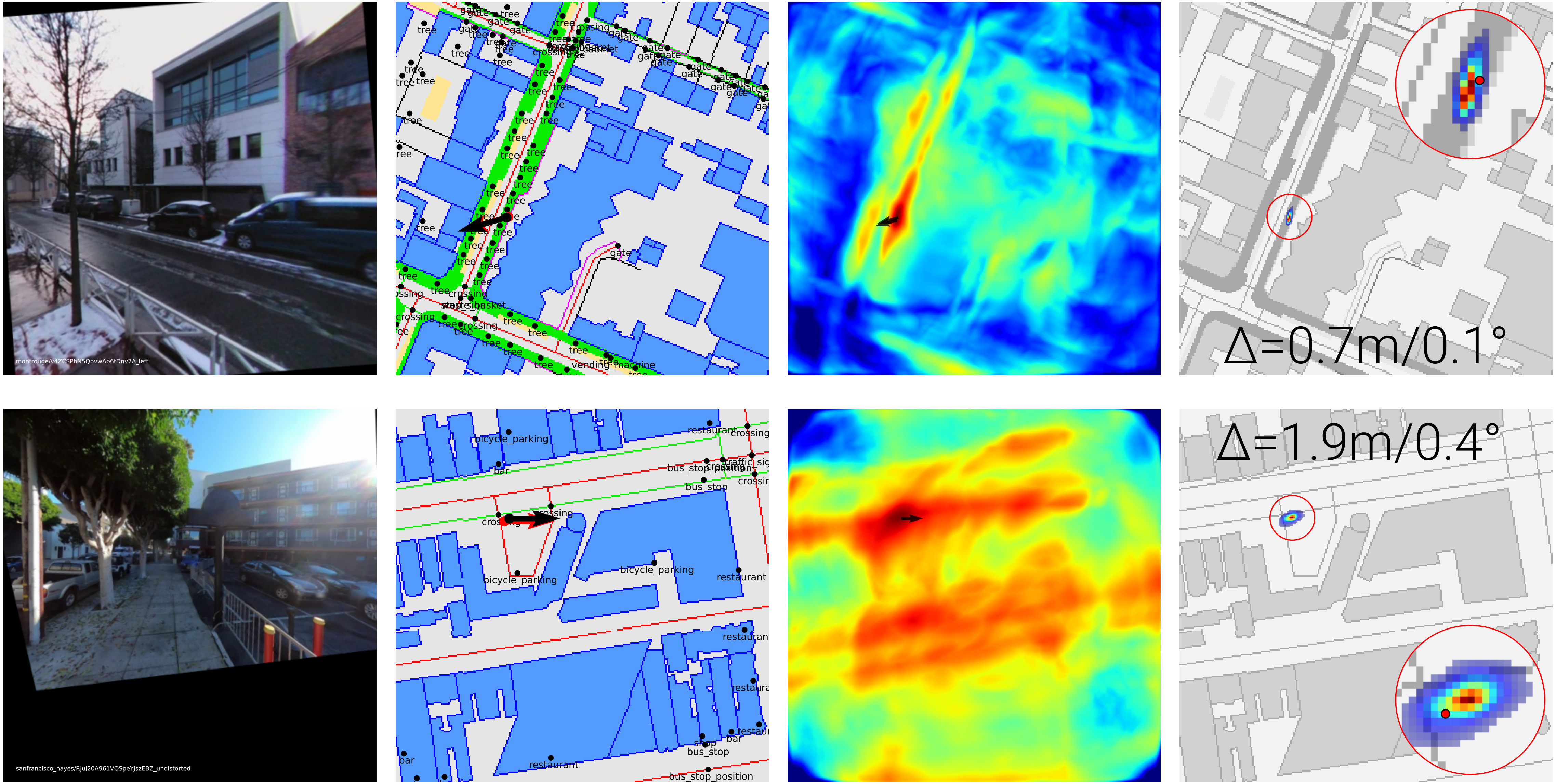



likelihood



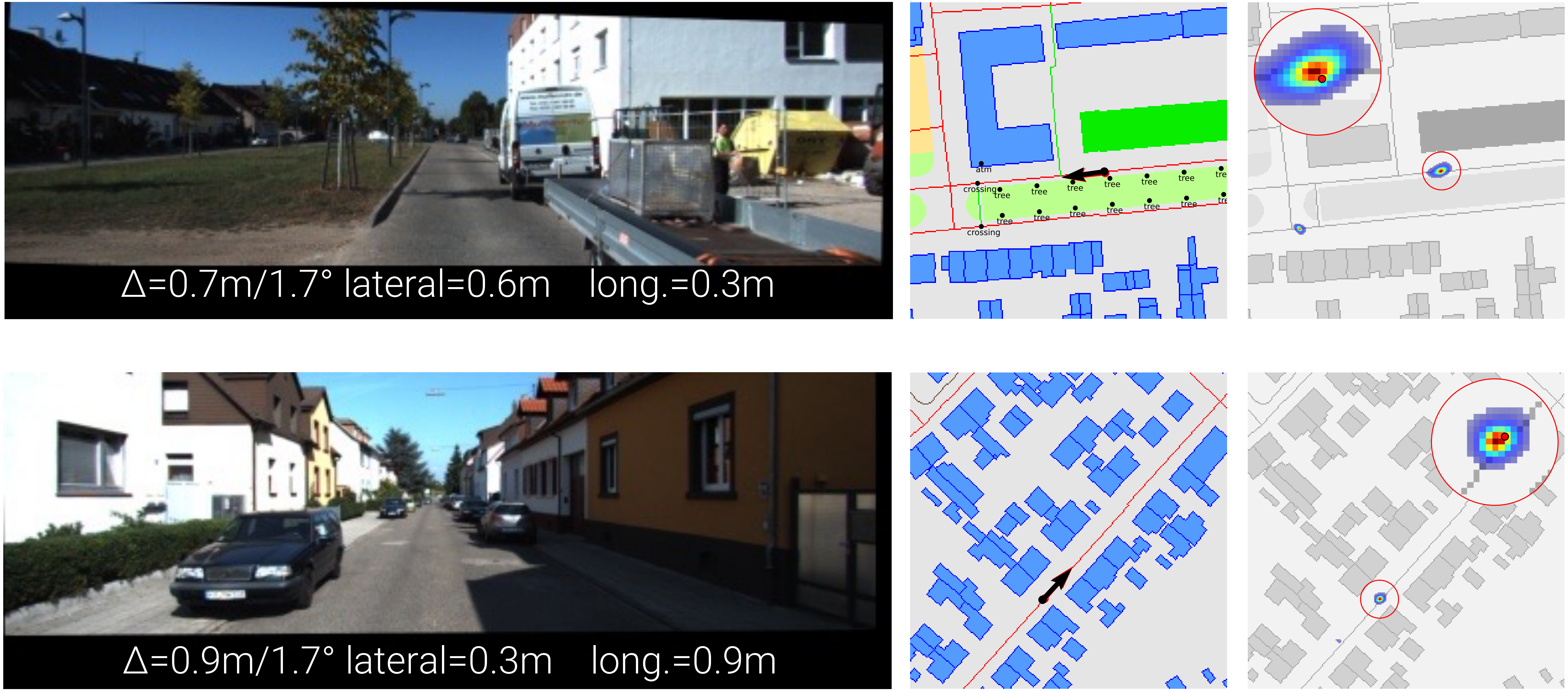
ground truth prediction

building area and outline, road, footway, cycleway, grass, park, playground, parking, fence input image raster map log-likelihood likelihood



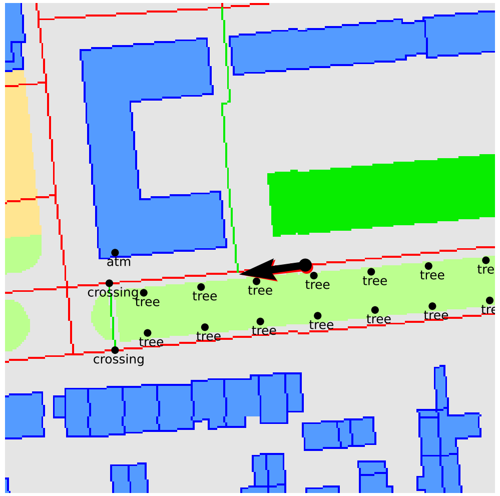

building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

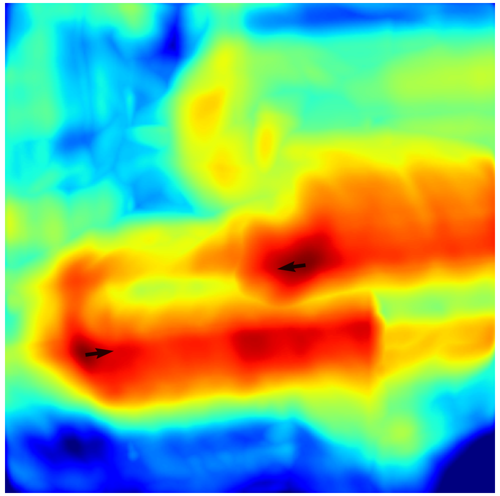
# Driving data – KITTI

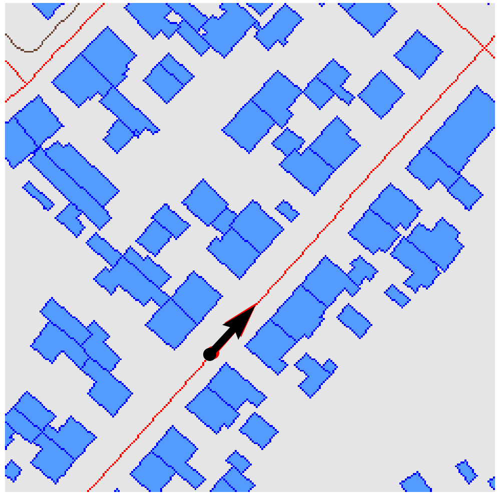


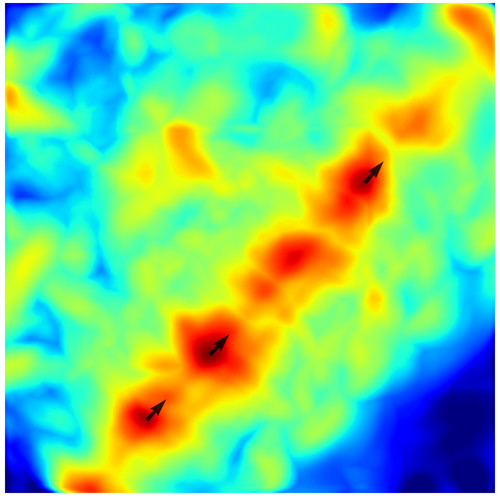

building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

# Driving data – KITTI

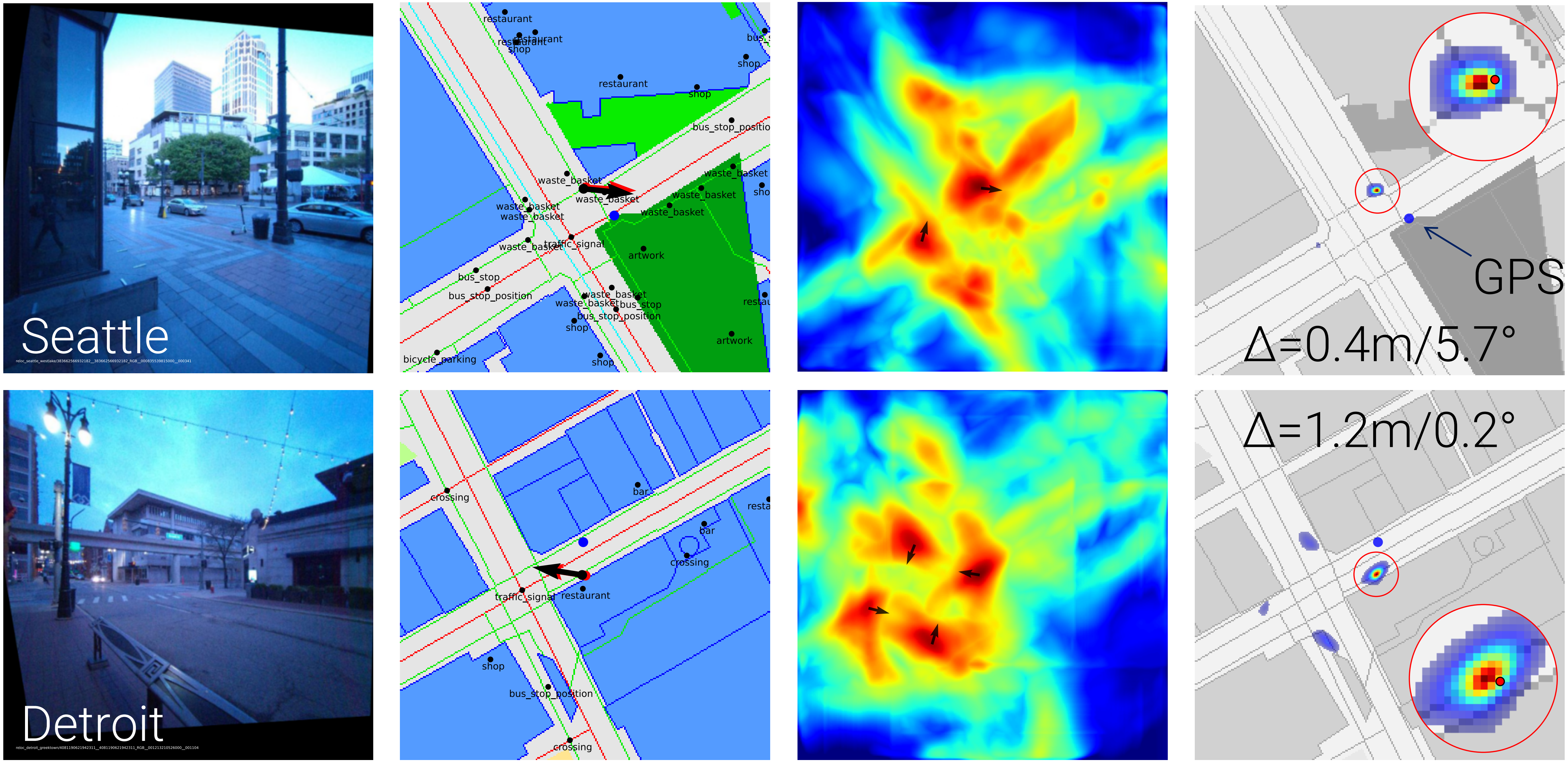





building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

# AR data – Aria glasses

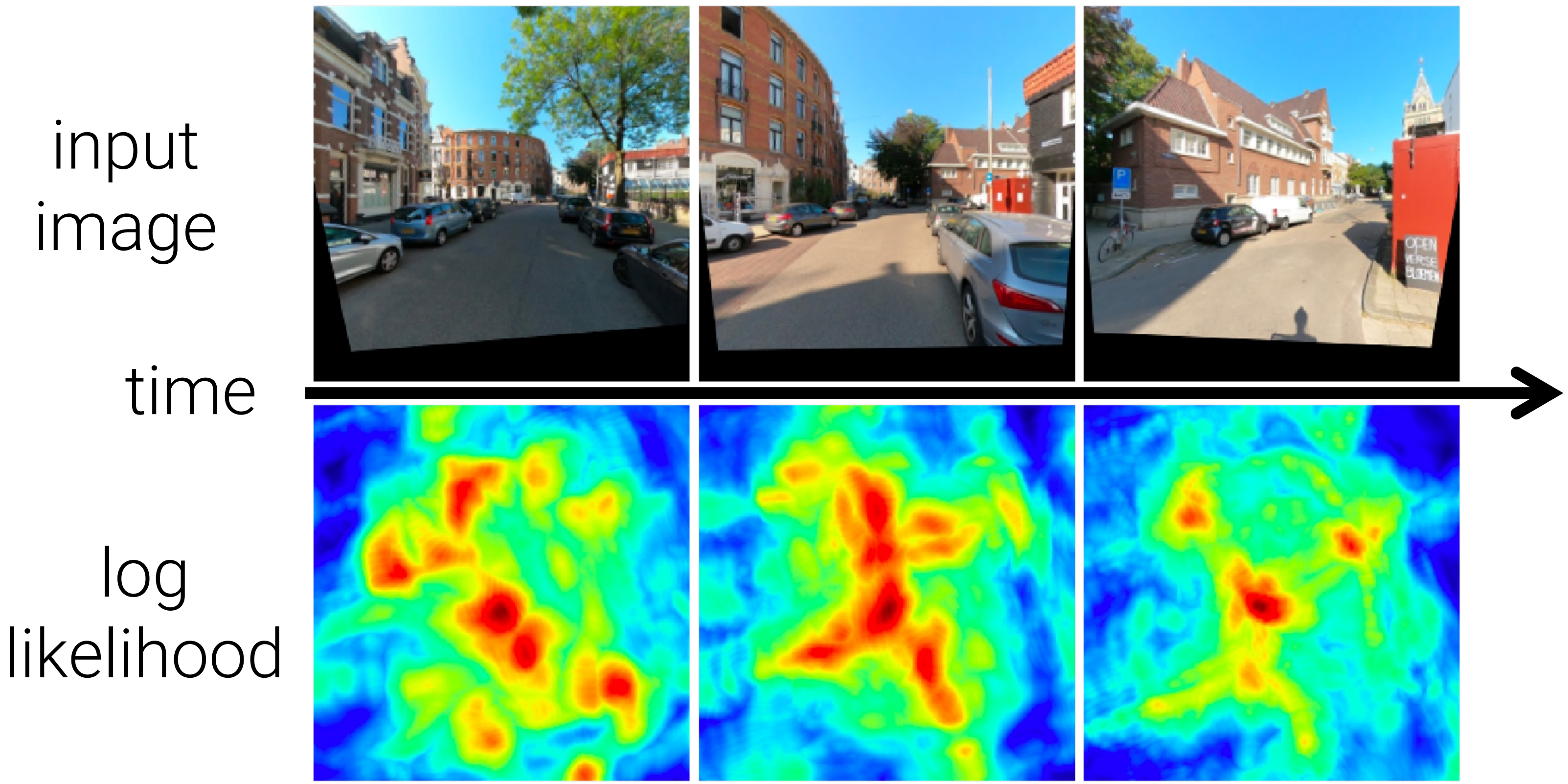



building ●area and ●outline, ●road, ●footway, ●cycleway, ●grass, ●park, ●playground, ●parking, ●fence

# Sequence localization

**Fuse successive predictions** assuming known relative poses

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



# Sequence localization

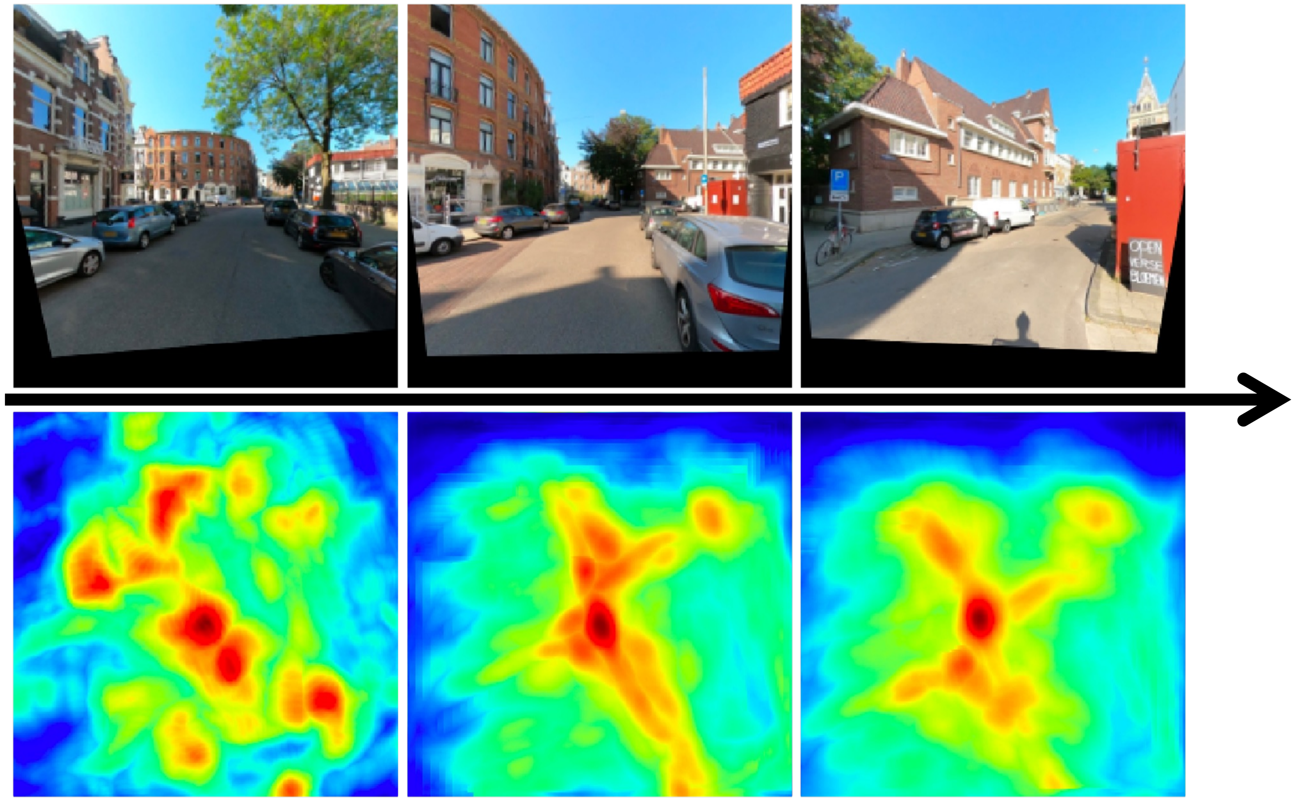
**Fuse successive predictions** assuming known relative poses

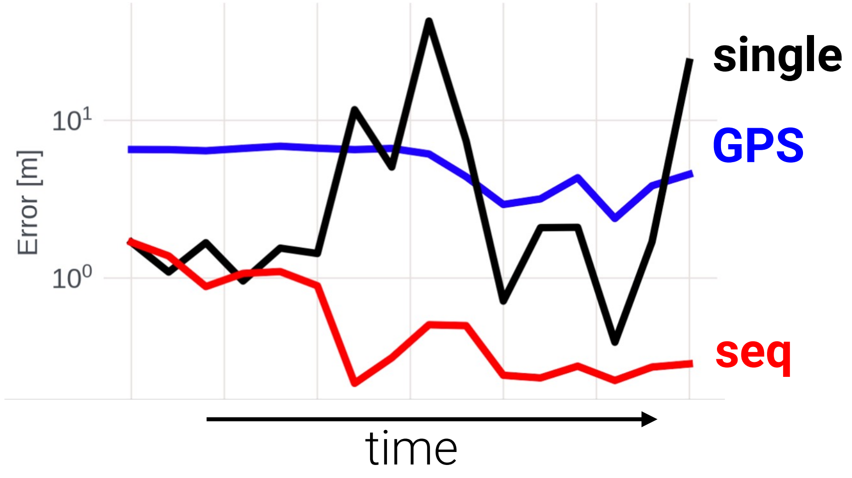
$$P(\boldsymbol{\xi}_i | \{\mathbf{I}_j\}, \mathrm{map}) = \prod_k P(\boldsymbol{\xi}_i \oplus \hat{\boldsymbol{\xi}}_{ij} | \mathbf{I}_j, \mathrm{map})$$

input image

time

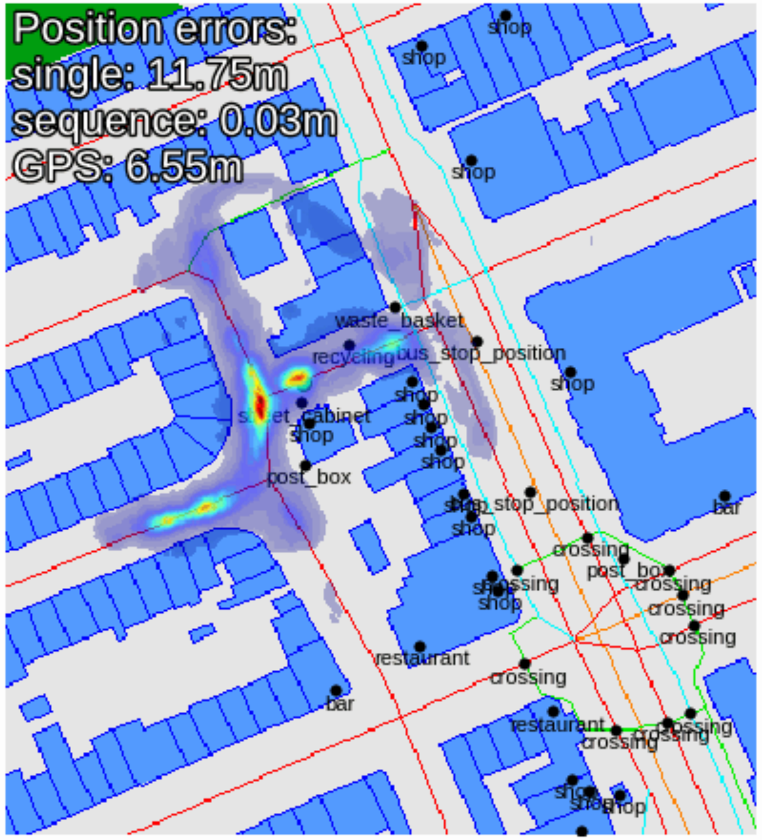
log likelihood



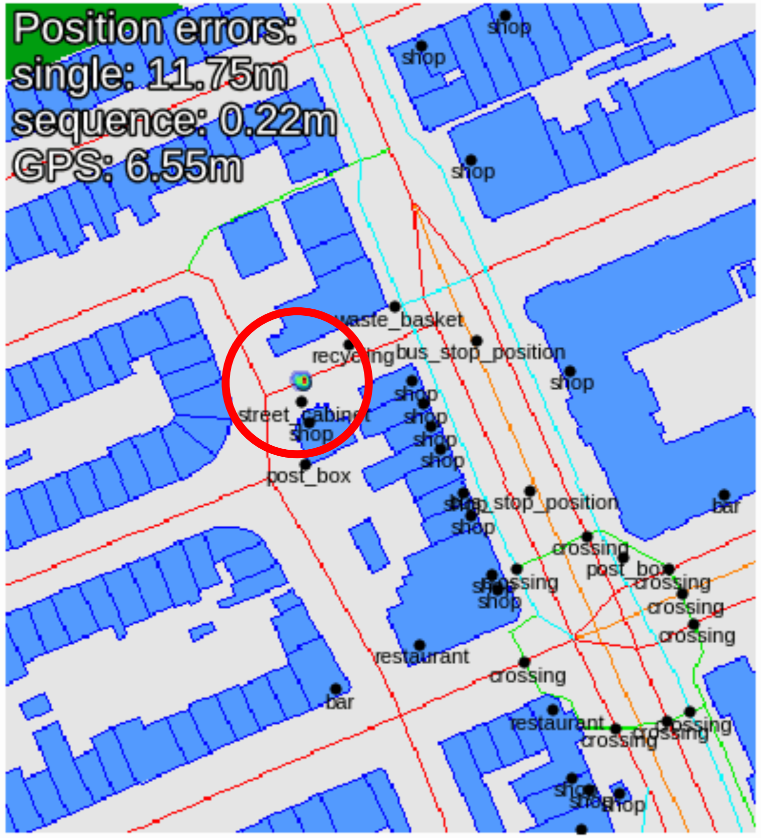


# Sequence localization

input image

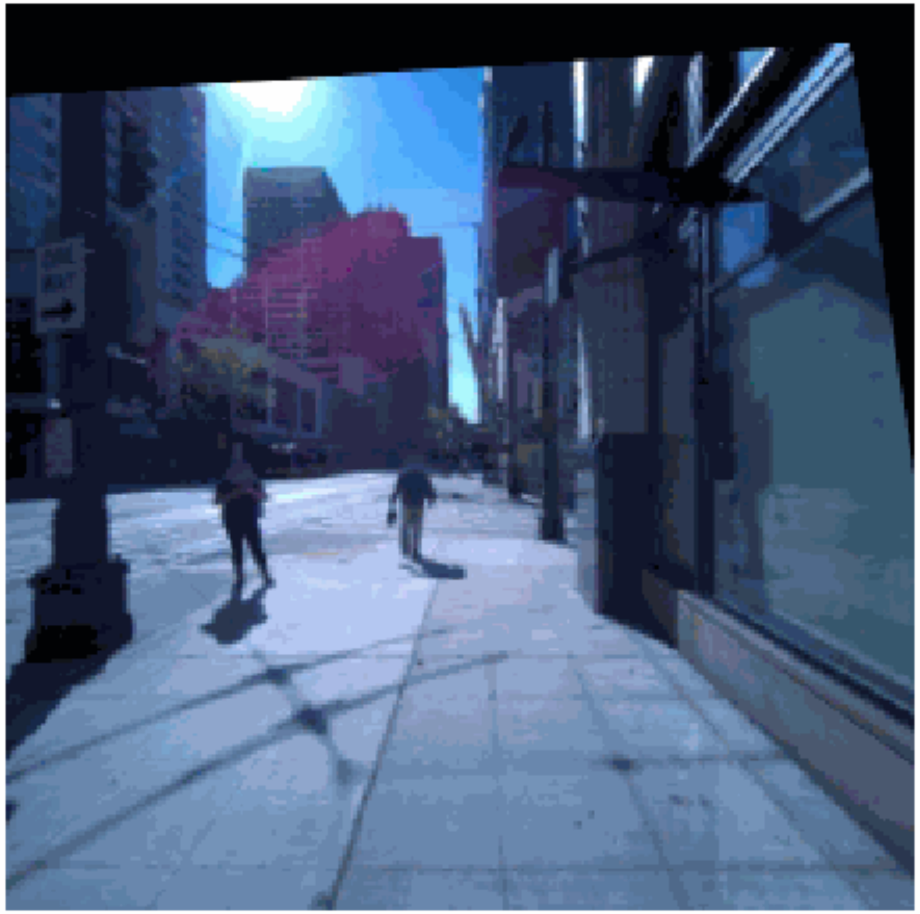



single-frame likelihood

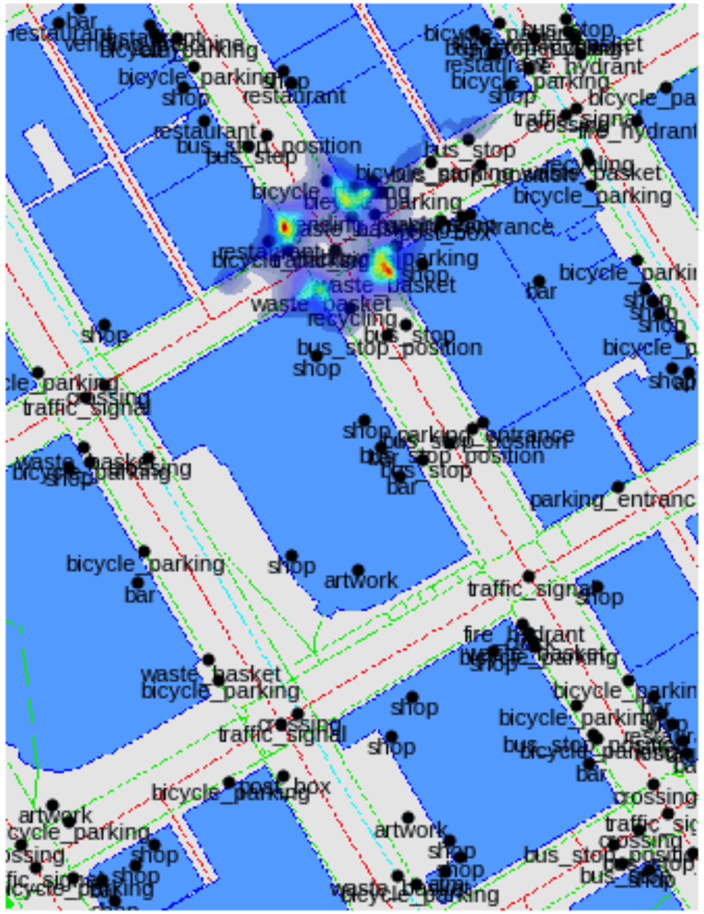



sequence likelihood

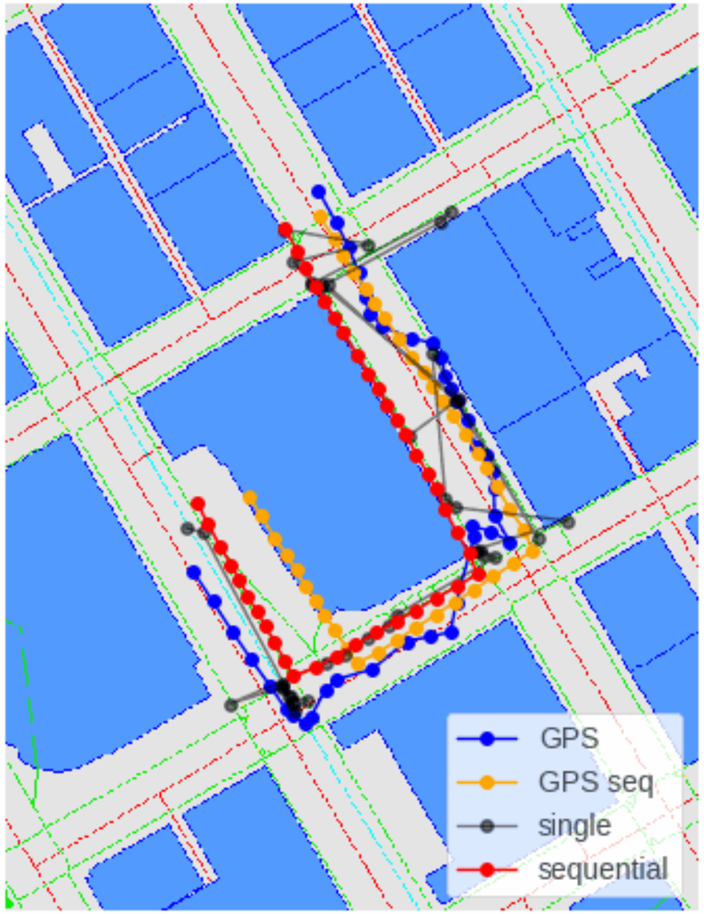
# Sequence localization – Aria



input image



single-frame likelihood




final trajectories

[1] ETH zürich

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]

CVPR 2023 psarlin.com/orienternet Poster THU-PM-098