lamar.ethz.ch

ECCV 2022 Tutorial on Localization and Mapping for AR

ETH zürich

ECCV TEL AVIV 2022

Microsoft

Tentative Schedule

1. Introduction and Motivations [1h]
2. Dataset and Ground-Truthing [1h]

Coffee break [15 mins]

3. Benchmarking Localization and Mapping [45 mins]
4. Practical guide and Conclusions [45 mins]

lamar.ethz.ch

The LaMAR Benchmark
Localization and Mapping for AR

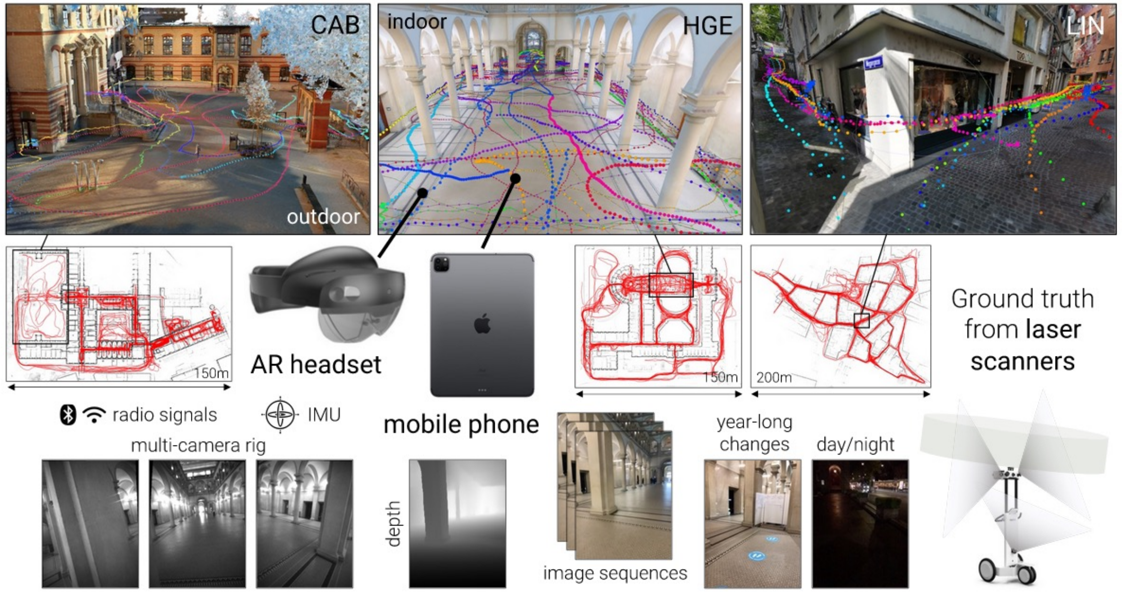
Home Dataset Leaderboard ECCV 2022 Tutorial

LaMAR: Benchmarking Localization and Mapping for AR

Paul-Edouard Sarlin[*1], Mihai Dusmanu[*1]
Johannes L. Schönberger[2], Pablo Speciale[2], Lukas Gruber[2], Viktor Larsson[2], Ondrej Miksik[2], Marc Pollefeys[1,2]

ETH Zurich[1], Microsoft[2]

European Conference on Computer Vision 2022



Come chat with us!
Poster 3.B
Hall B
Session 18
Poster 7
Thursday, 15:30-17:30

Organizers

Paul-Edouard Sarlin
ETH Zurich

Johannes L. Schönberger
Microsoft

Ondrej Miksik
Microsoft

Mihai Dusmanu
ETH Zurich

Viktor Larsson
Lund University

Marc Pollefeys
Microsoft & ETH Zurich

lamar.ethz.ch

LaMAR tutorial
1. Introduction & Motivations

Marc Pollefeys

ETH zürich

ECCV TEL AVIV 2022

Microsoft

Outline

a) Visual Localization and Mapping
b) Augmented Reality systems
c) Benchmarking & existing datasets

a) Visual Localization & Mapping

Applications

- Devices need to know **where** they are located in space
- Different accuracy requirements

Google Maps

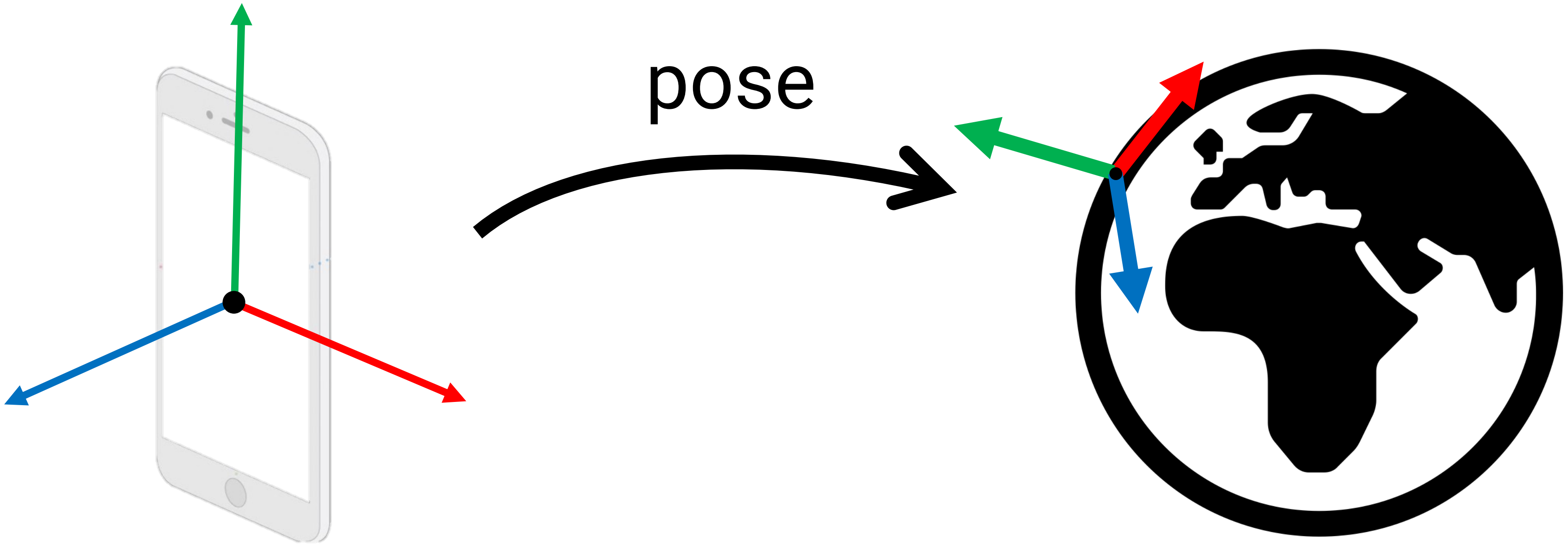
Microsoft Hololens

Sevensense Robotics

Positioning

Recover the pose of the device

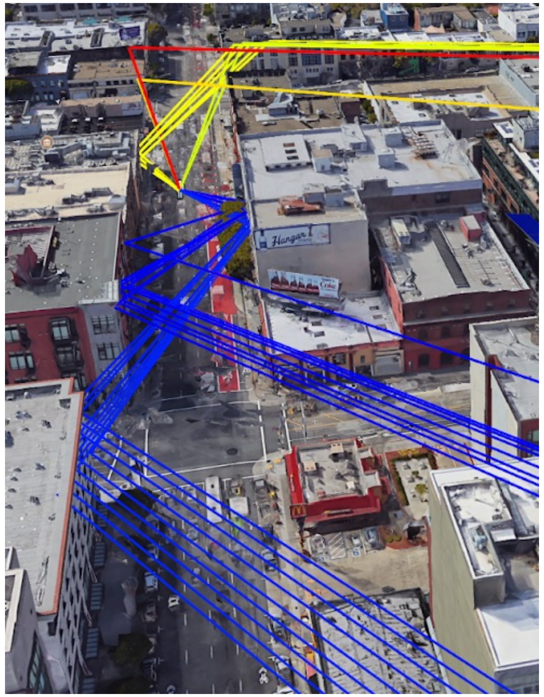
- 2D/3D translation? Rotation?
- w.r.t. a known reference frame



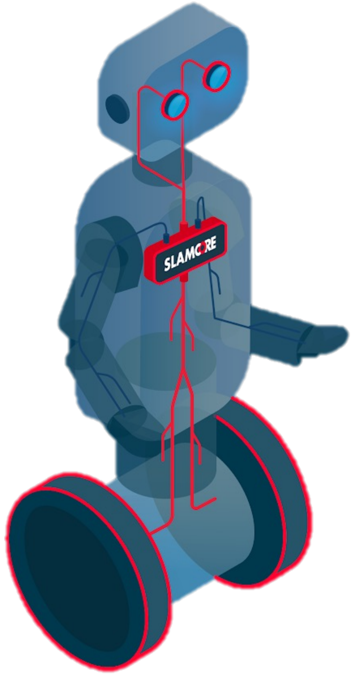
Apple ARKit

Positioning solutions

GPS
globally absolute
inaccurate



Wheel odometry
for robots
prone to drift



Vision
accurate
cameras are cheap

reference frame
= posed mapping images

Images from Google VPS / SLAMCORE / Sony

Challenges of long-term localization

Mapping and localization
at different times
→ the world changes

structure

appearance

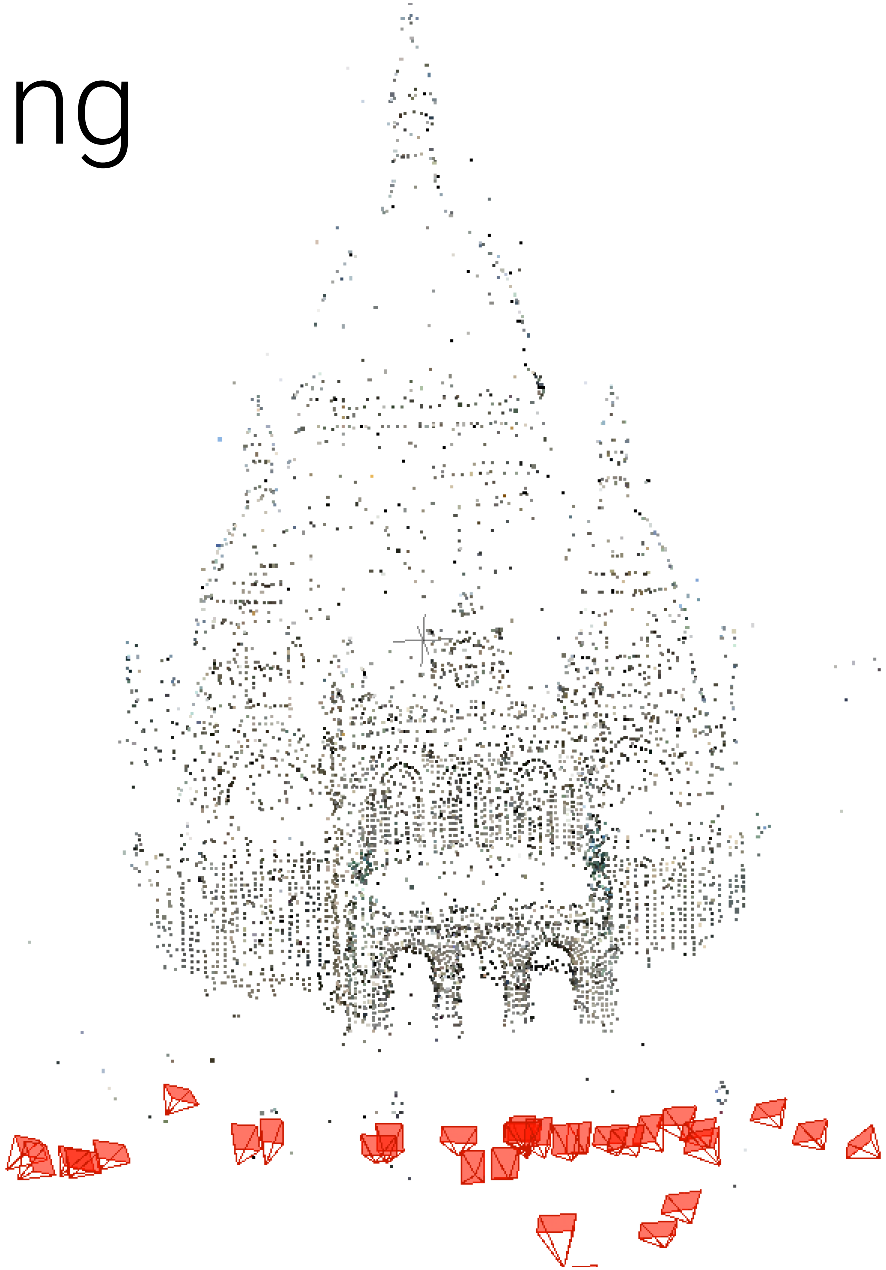
Visual mapping

A map: a representation
of **appearance/geometry** of the scene

- mapping images
  with poses + calibration

and/or

- a sparse **3D point cloud**
  with descriptors



Structure-from-Motion

SfM

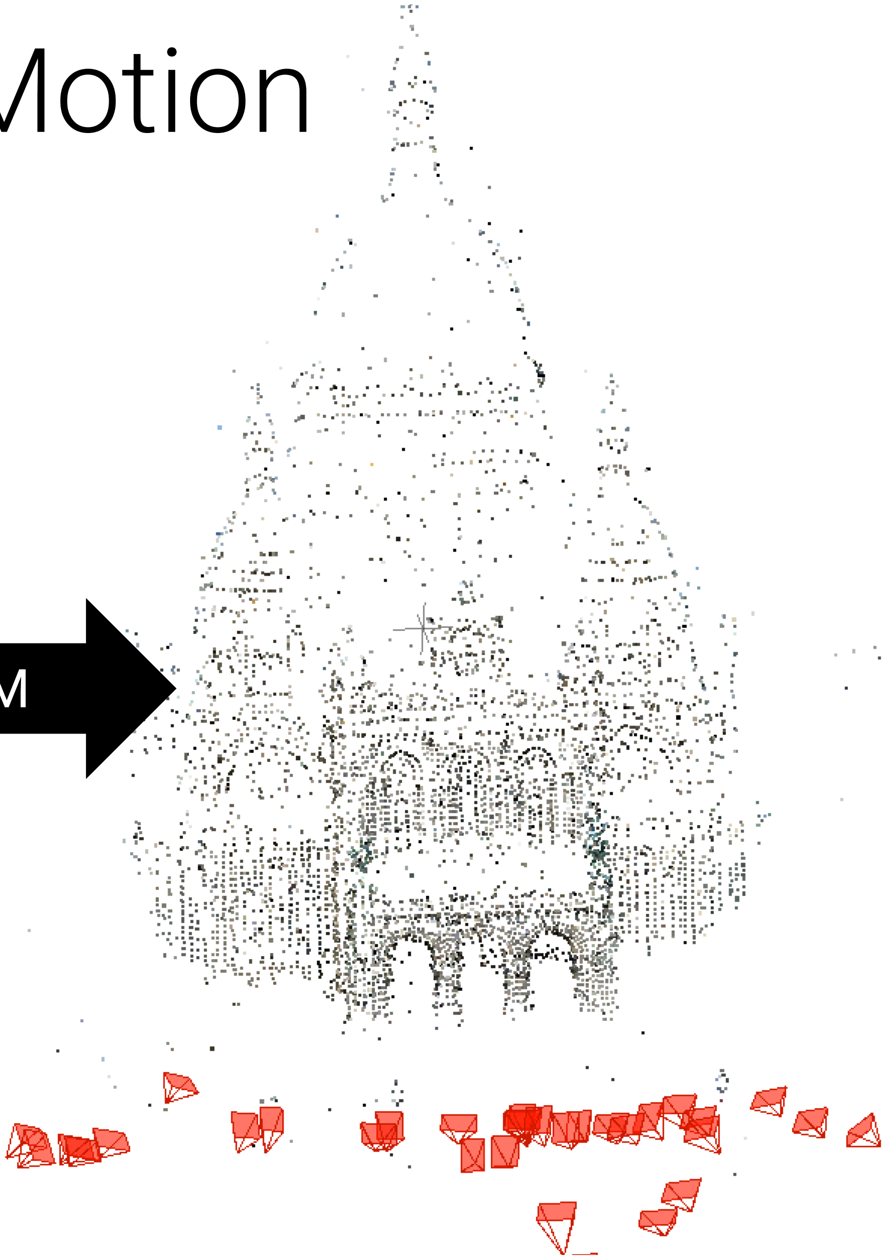


image matching + geometric optimization

1) Sparse image matching

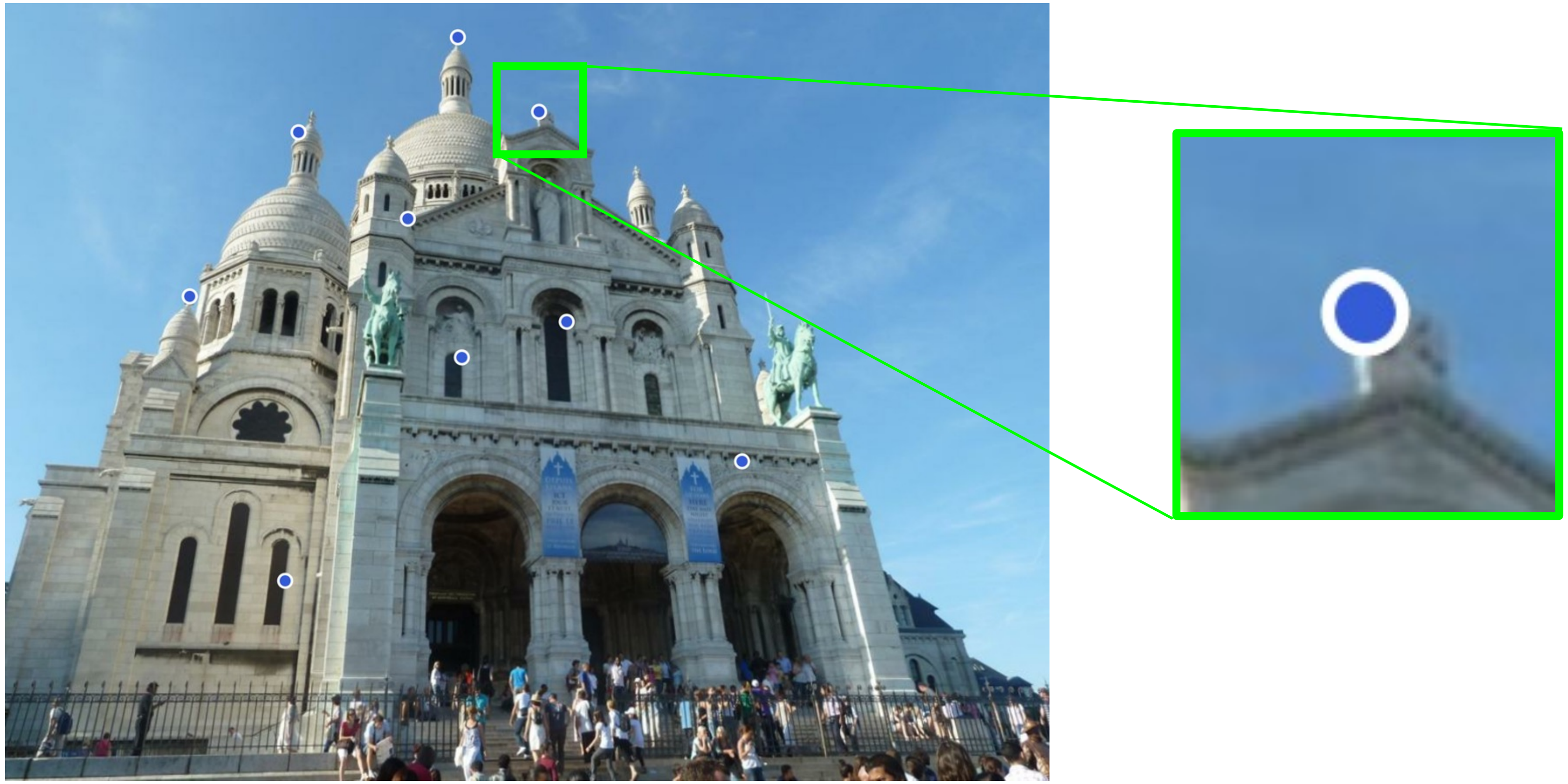
image → detect —keypoints→ describe → match

DoG
Harris
FAST
BRISK
SURF
AKAZE
LIFT
SuperPoint
D2-Net
Key.Net
R2D2
…

1) Sparse image matching

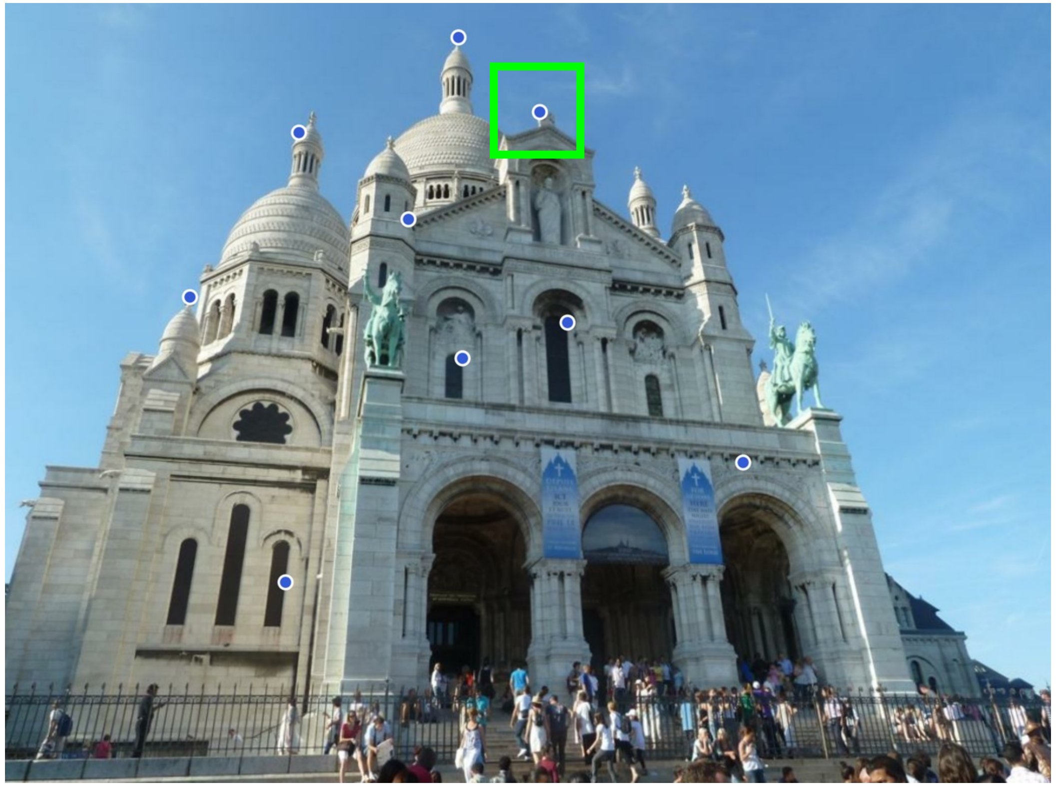
image → **detect** → keypoints → **describe** → descriptors → **match**

SIFT
SURF
BRIEF
ORB
L2-Net
HardNet
SOSNet
LIFT
SuperPoint
D2-Net
R2D2
…



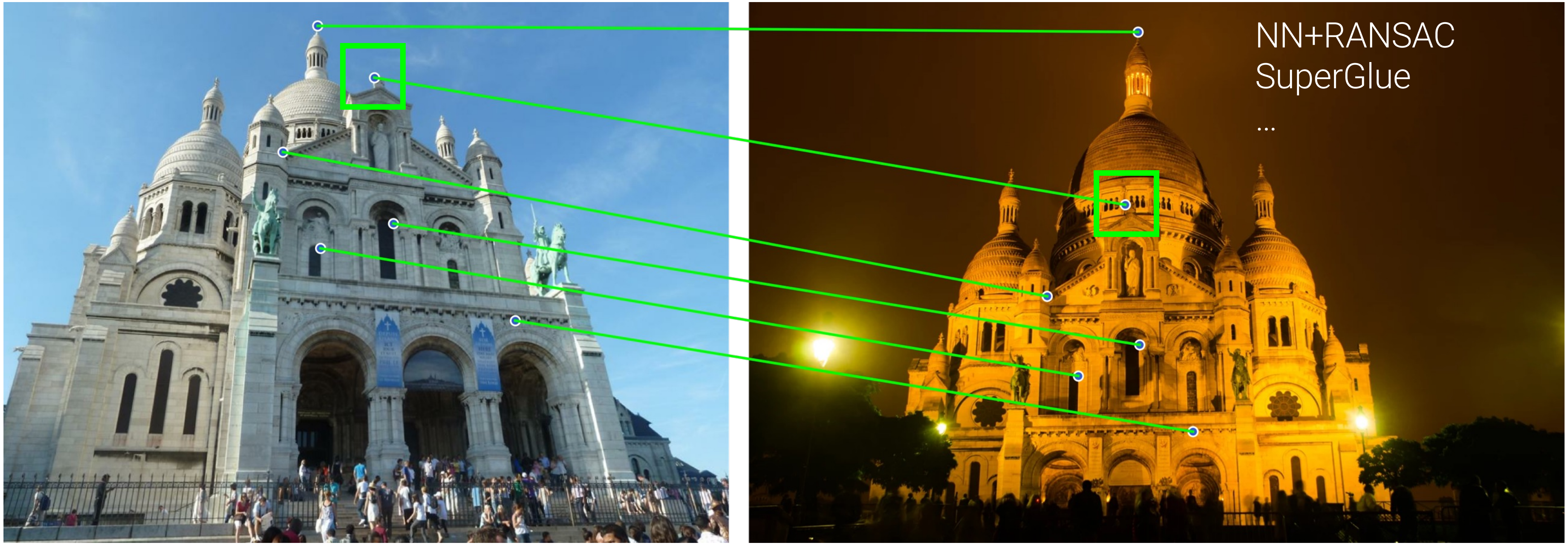
1) Sparse image matching

image → **detect** → keypoints → **describe** → descriptors → **match**

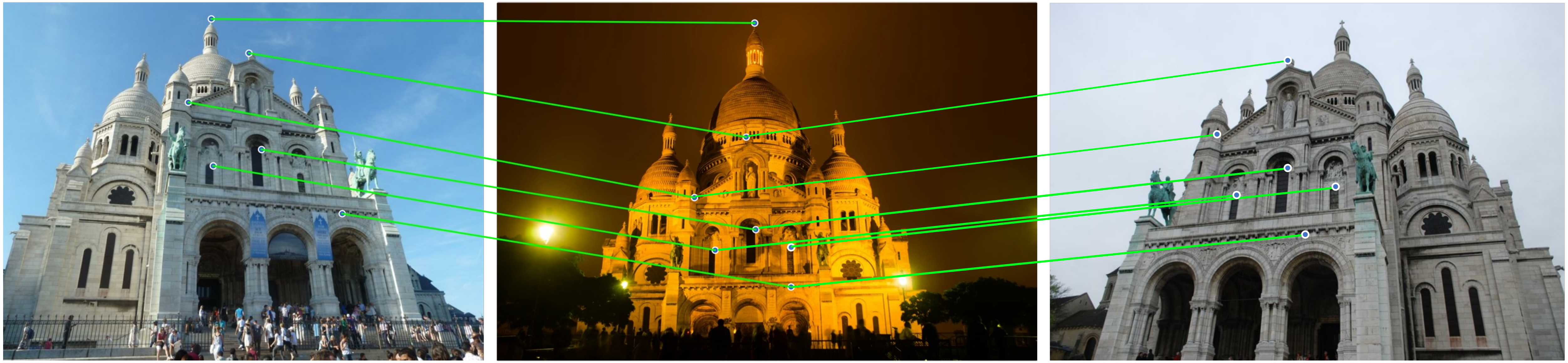


1) Sparse image matching

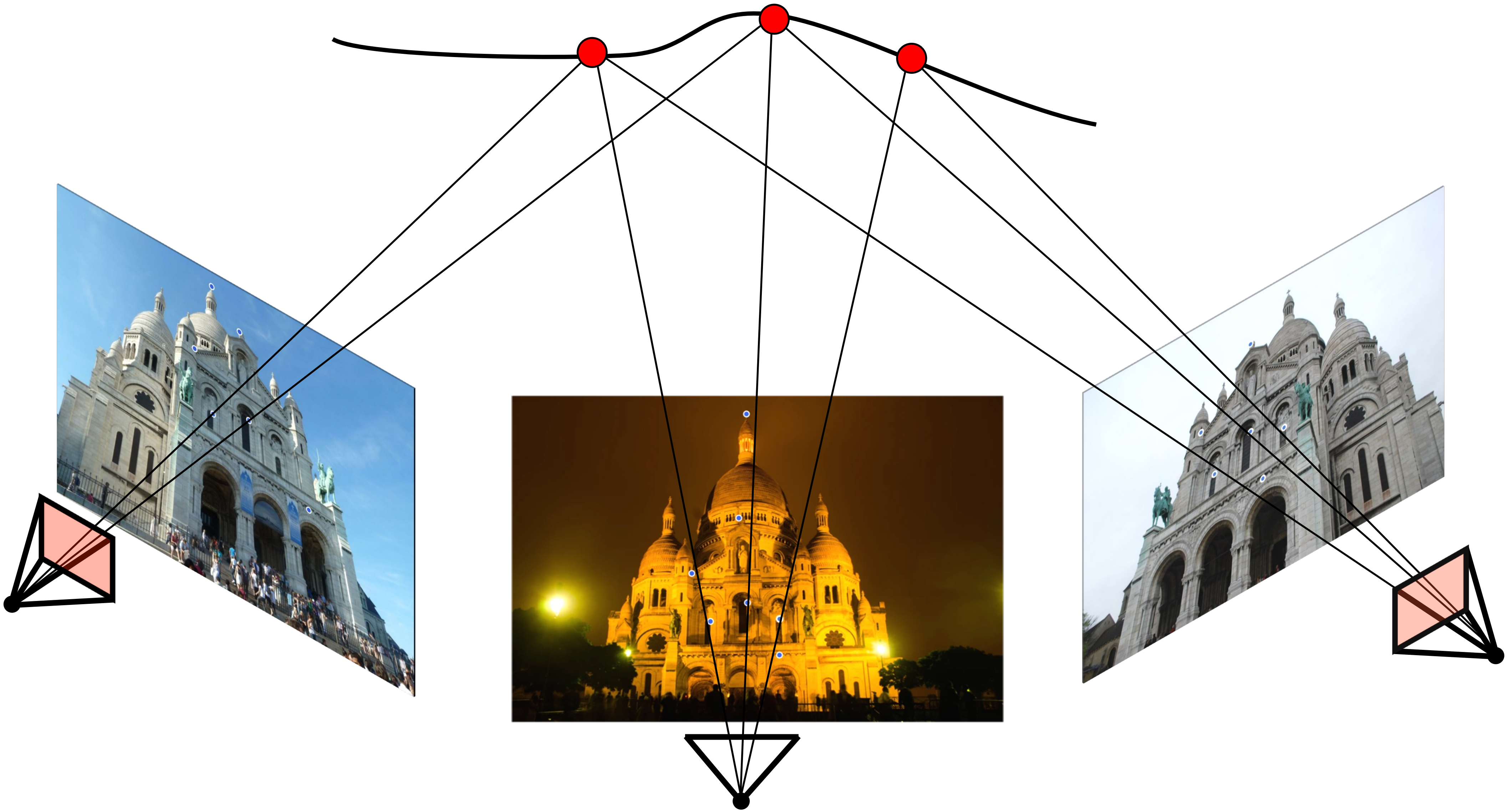
image → detect → keypoints → describe → descriptors → match



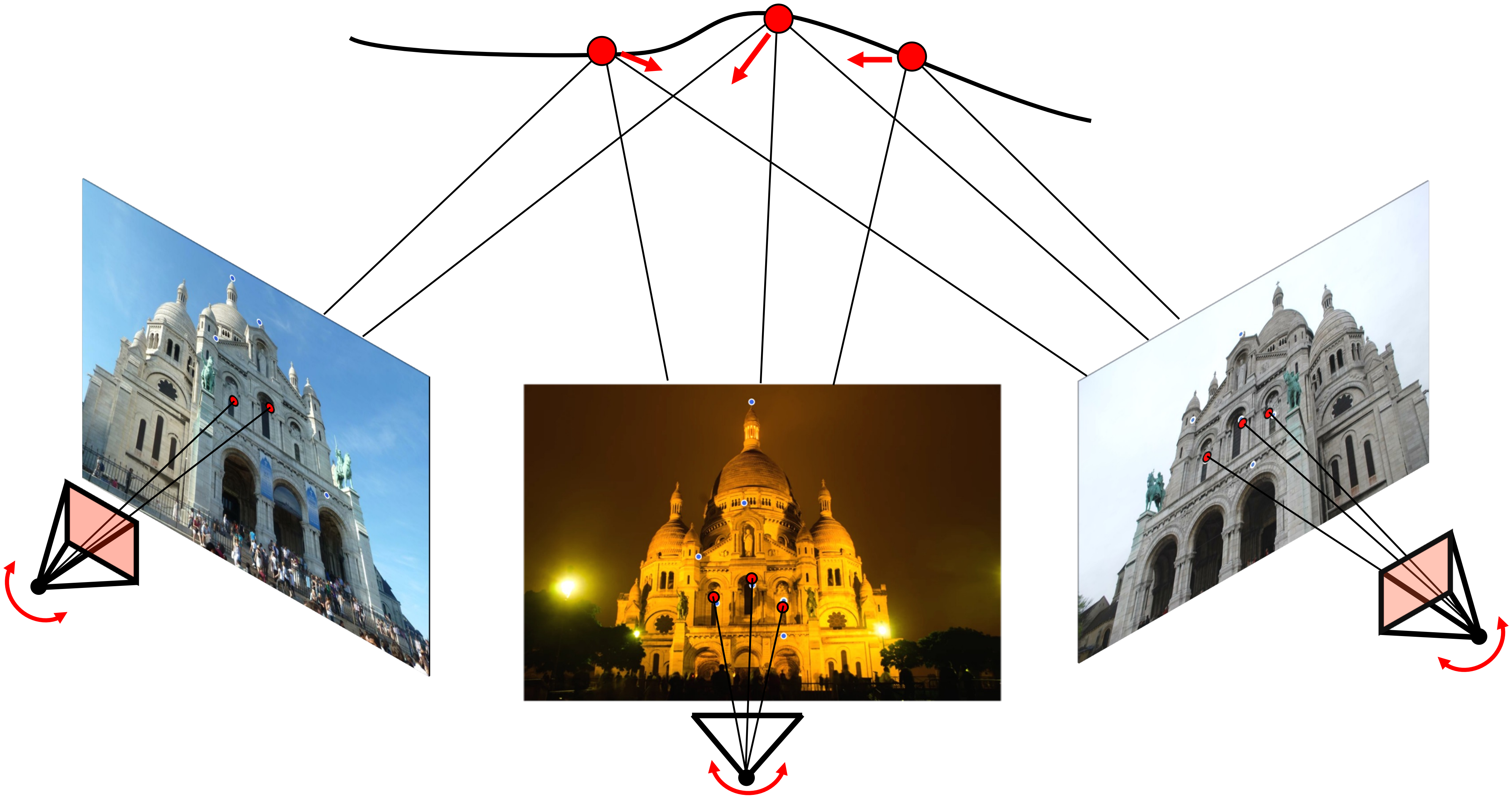
2) Triangulation



2) Triangulation



2) Triangulation



Mapping algorithms

Similar principles but different names

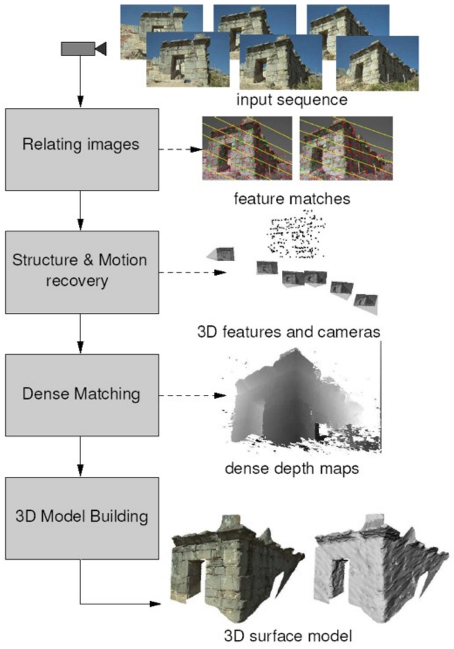
Structure-from-Motion

- once all images are captured
- offline
- typically for arbitrary internet image collections
- typically uses matching

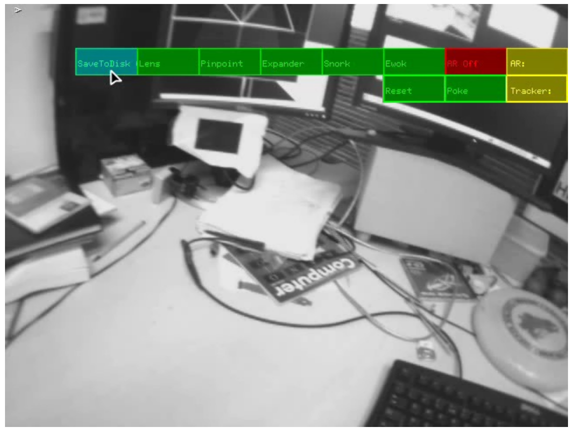
SLAM

- online estimation from videos
- minimal latency
- assumes continuous motion, typically leverages IMU
- typically uses tracking (except for loop closure)

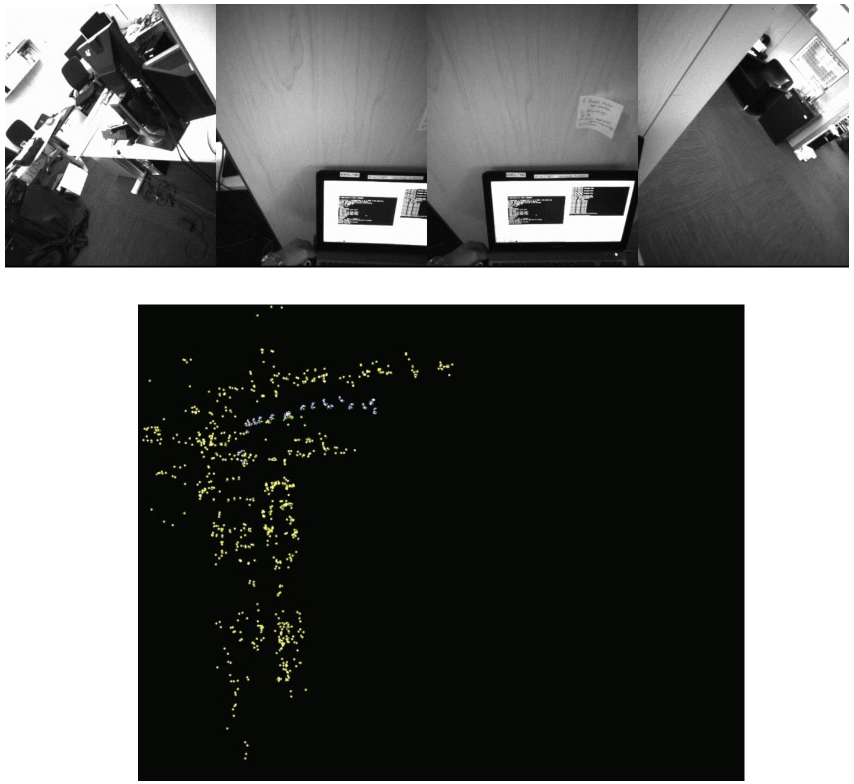
Mapping - history



Pollefeys et al ICCV98



Klein and Murray ISMAR07



HoloLens

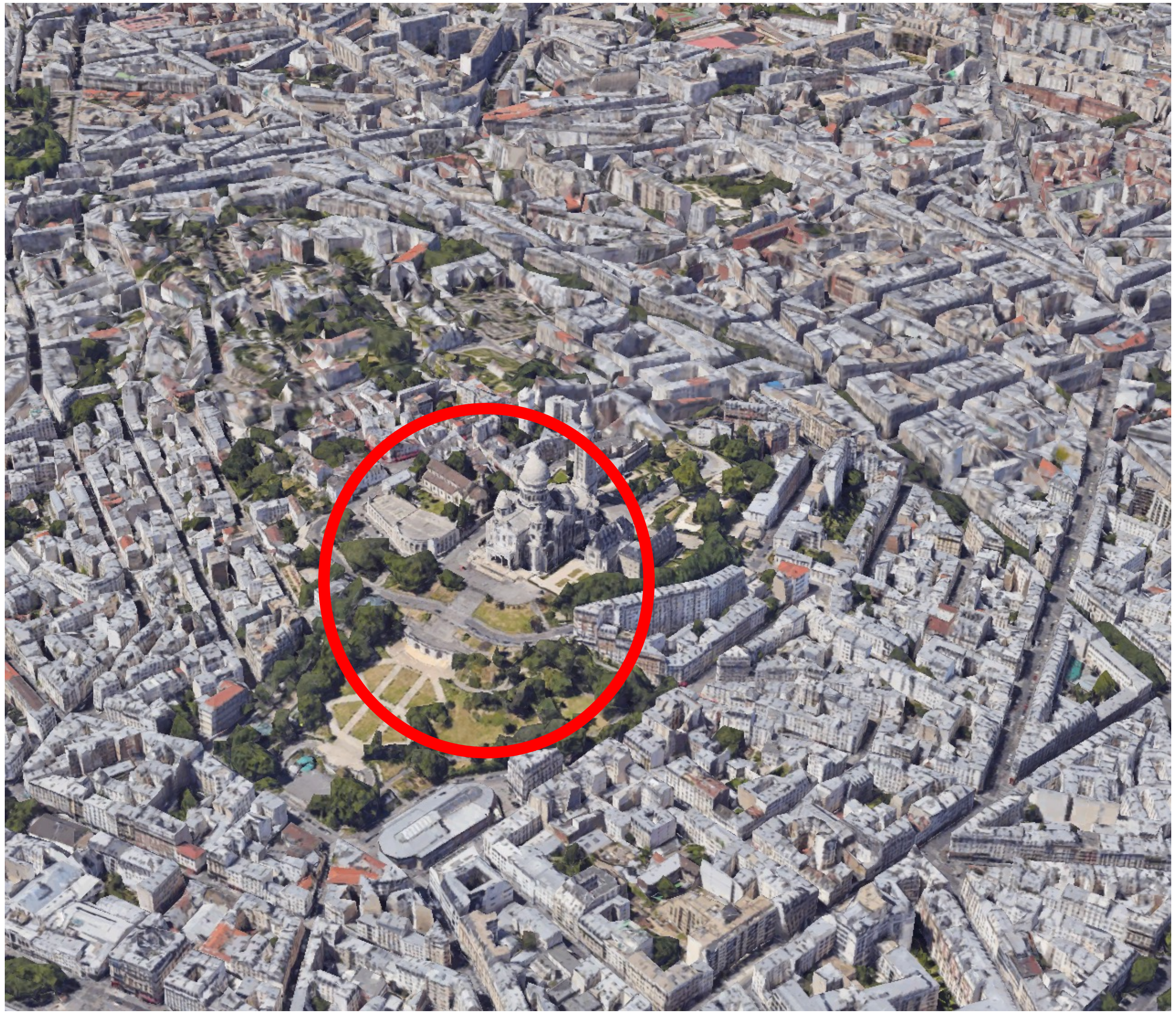
- extensive research over 20+ years
- good open-source tools: COLMAP, ORB-SLAM, etc.

Localization

Google Earth

Given a new image,
find its position w.r.t. the map

Localization

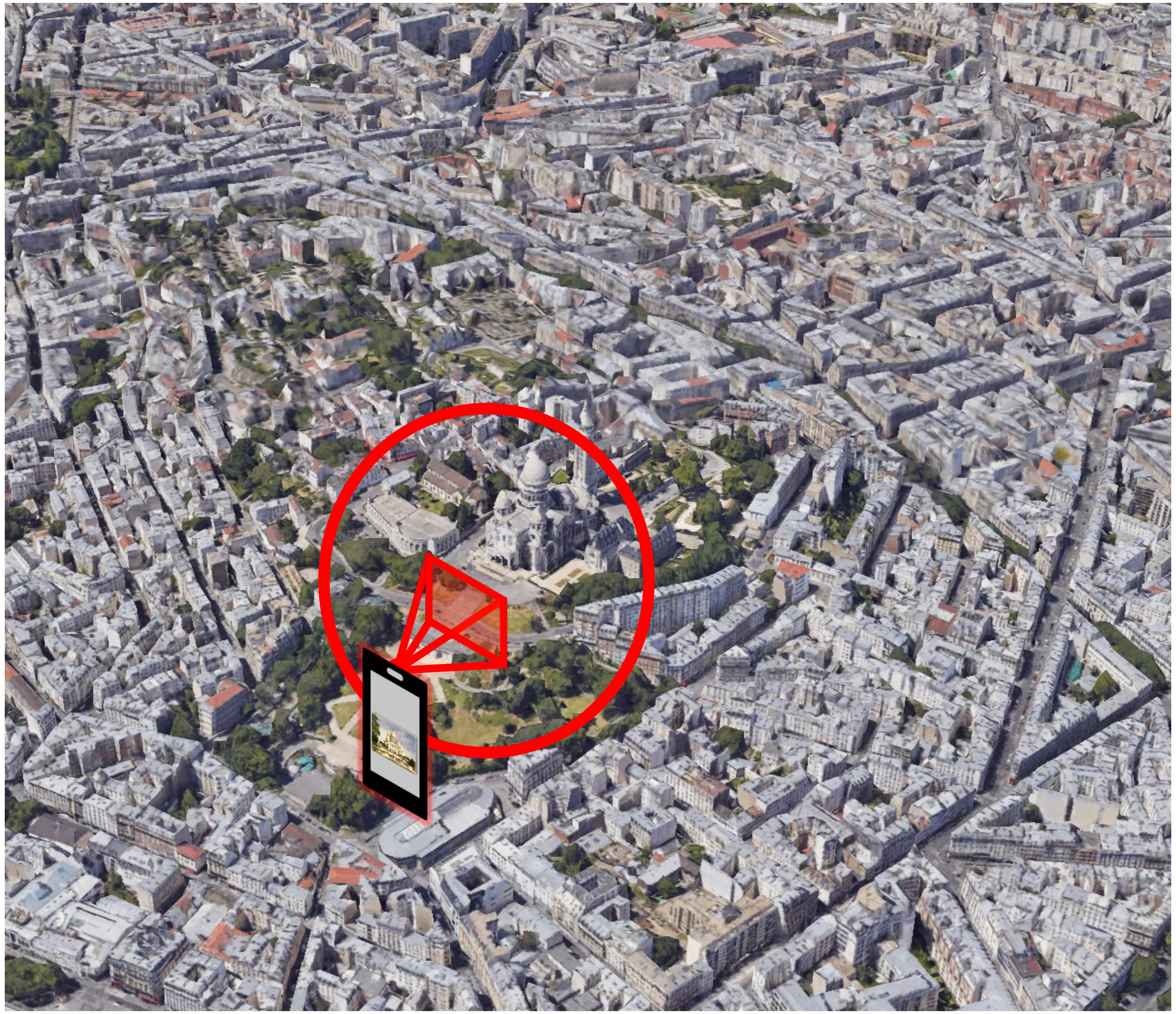


Google Earth

Given a new image,
find its position w.r.t. the map

- Place recognition: image similarity

Localization

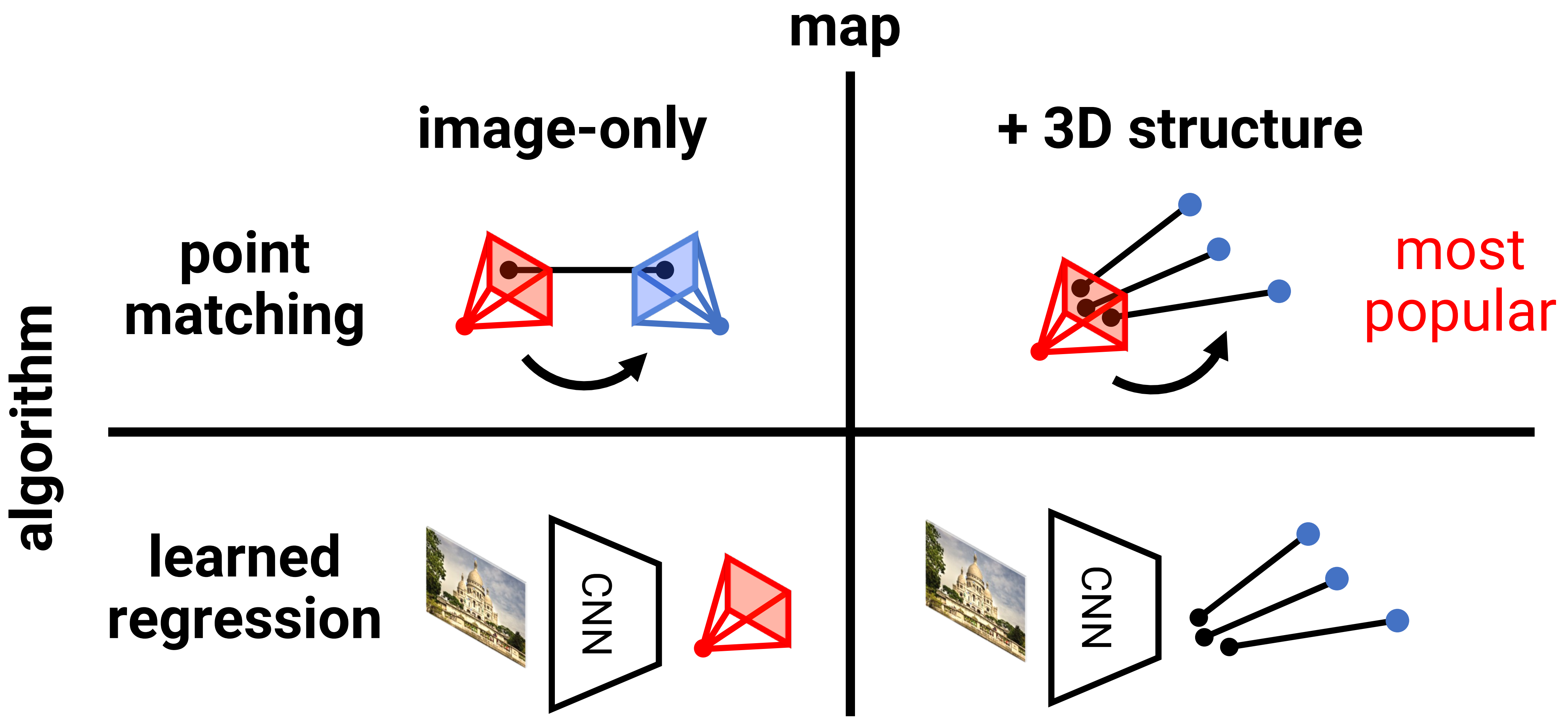


Google Earth

Given a new image,
find its position w.r.t. the map

- Place recognition: image similarity
- **6-DoF localization: R+t**

Localization – taxonomy

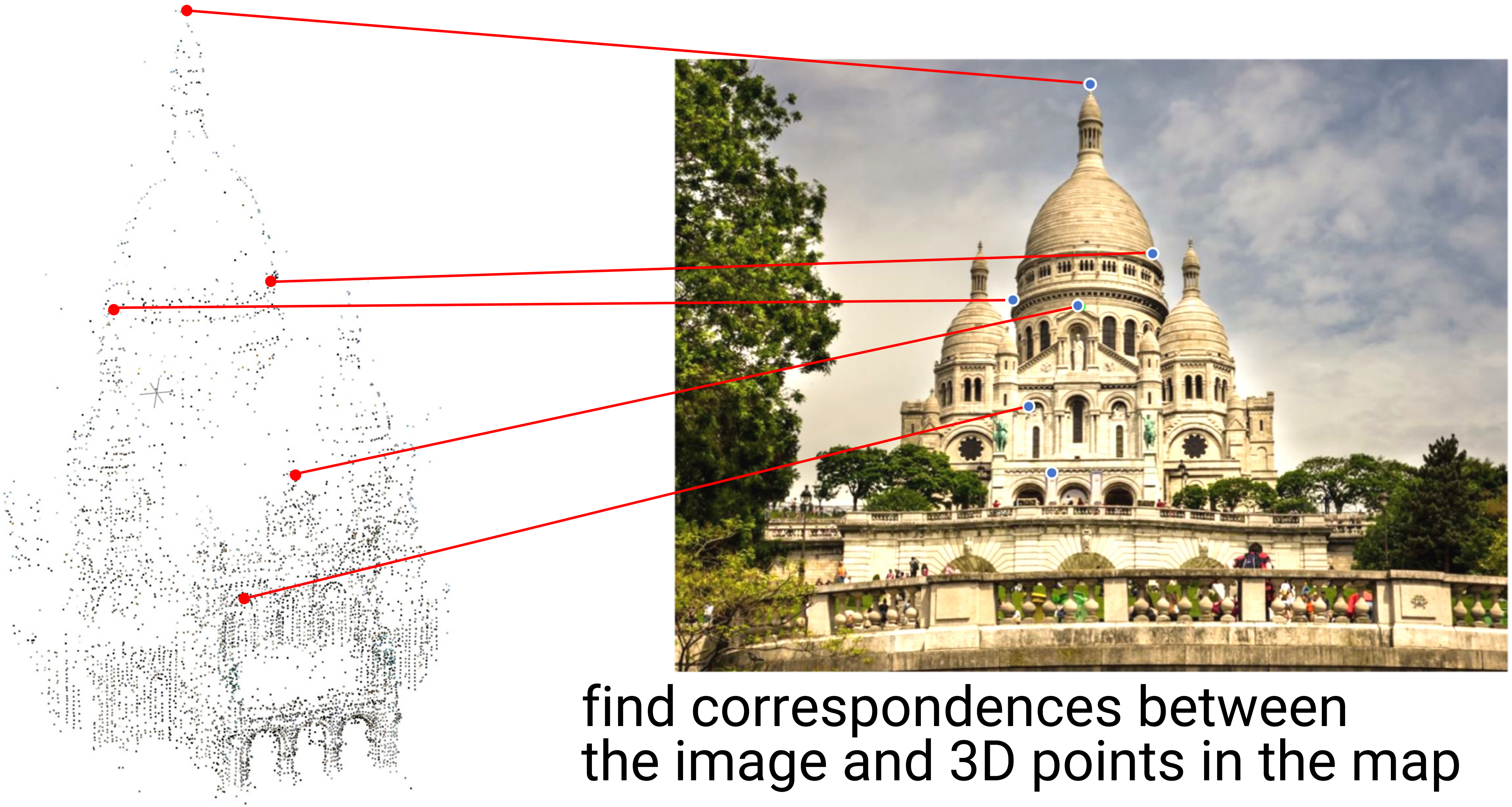


see ICCV 2021 Tutorial “Long-term visual localization”

Localization – structure-based

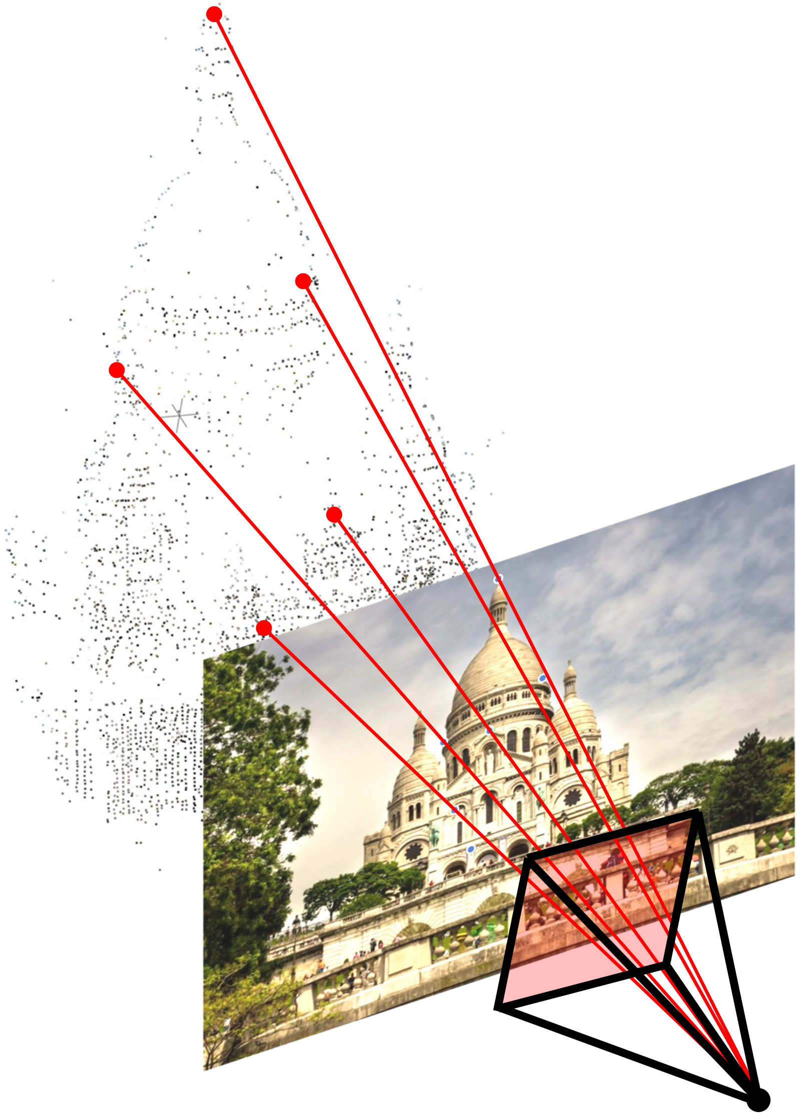
find correspondences between
the image and 3D points in the map

Localization – structure-based



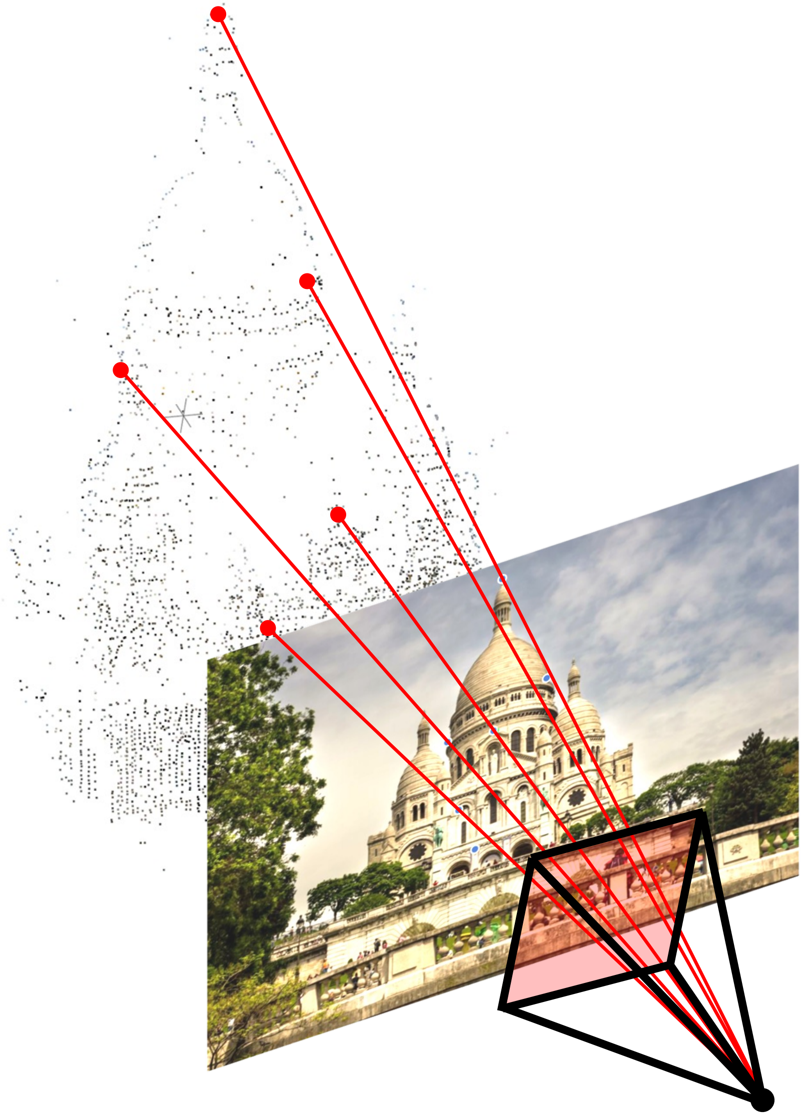
find correspondences between
the image and 3D points in the map

Localization – structure-based



minimize
reprojection errors
RANSAC + solver

Localization – structure-based

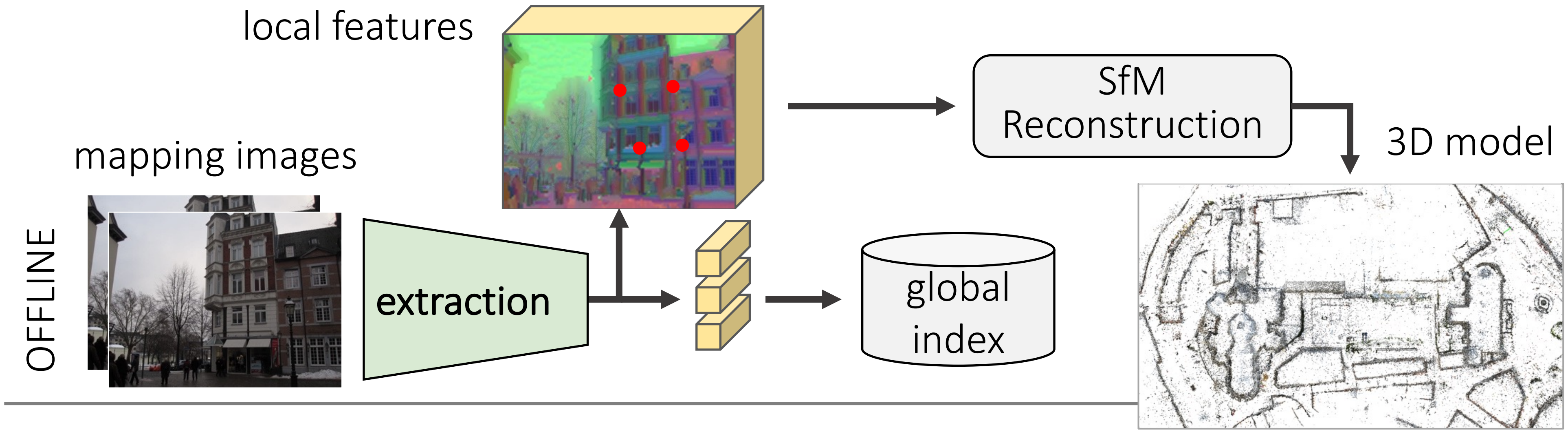


minimize
reprojection errors
RANSAC + solver

if map is large
→ efficient search
Active Search

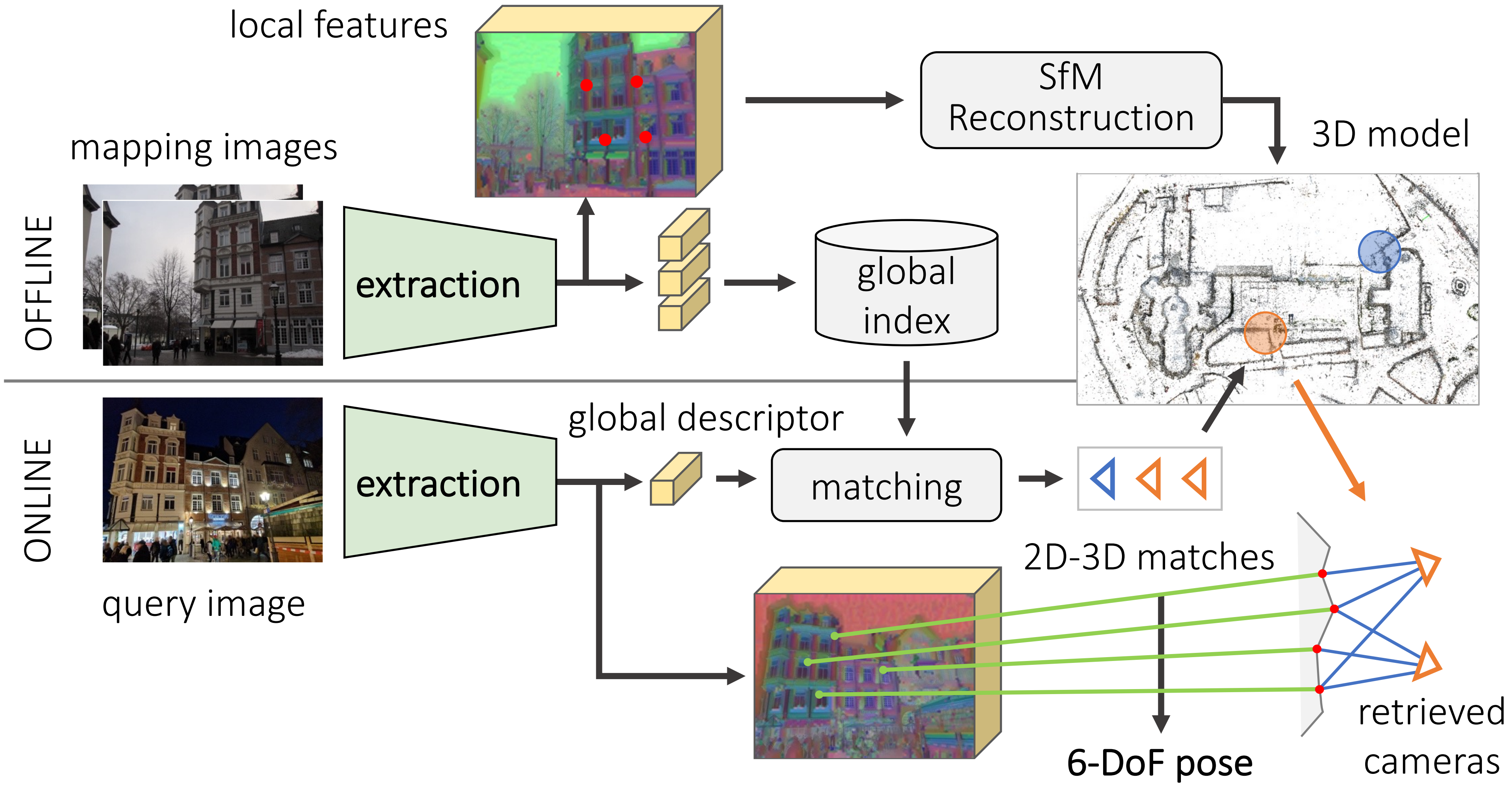
[Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization, Sattler et al, TPAMI 2017]

Hierarchical Localization



[From Coarse to Fine: Robust Hierarchical Localization at Large Scale, Sarlin et al., CVPR 2019]

Hierarchical Localization



[From Coarse to Fine: Robust Hierarchical Localization at Large Scale, Sarlin et al., CVPR 2019]

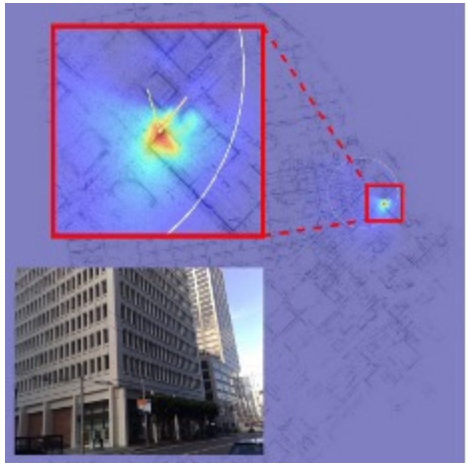
Hierarchical Localization

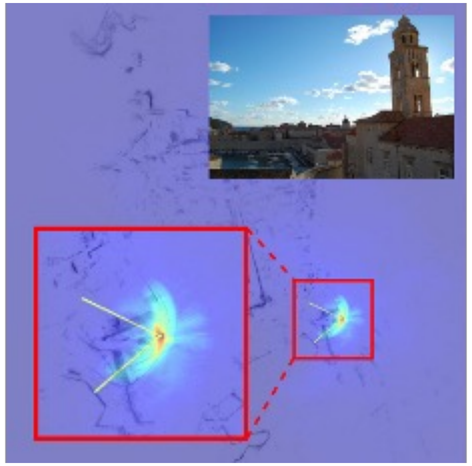
State of the art: many components use deep networks

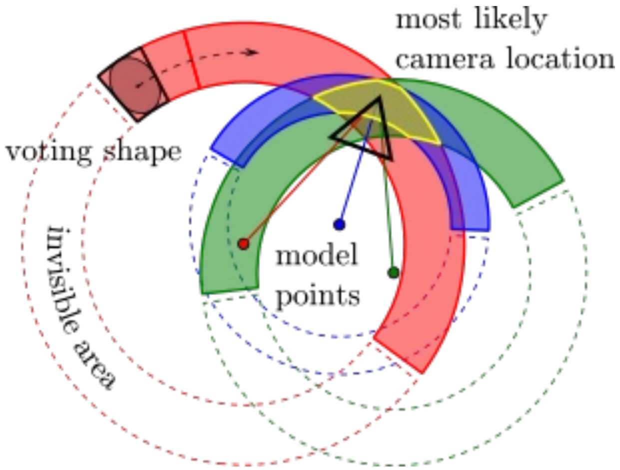
- Detection: DoG → SuperPoint, R2D2
- Description: SIFT → HardNet, SOSNet, SuperPoint, R2D2
- Matching: nearest neighbor search → SuperGlue; RANSAC
- Retrieval: BoW, VLAD → NetVLAD, AP-GeM

Open-source toolbox: hloc
github.com/cvg/Hierarchical-Localization

Camera Pose Voting for Large-Scale Image-Based Localization

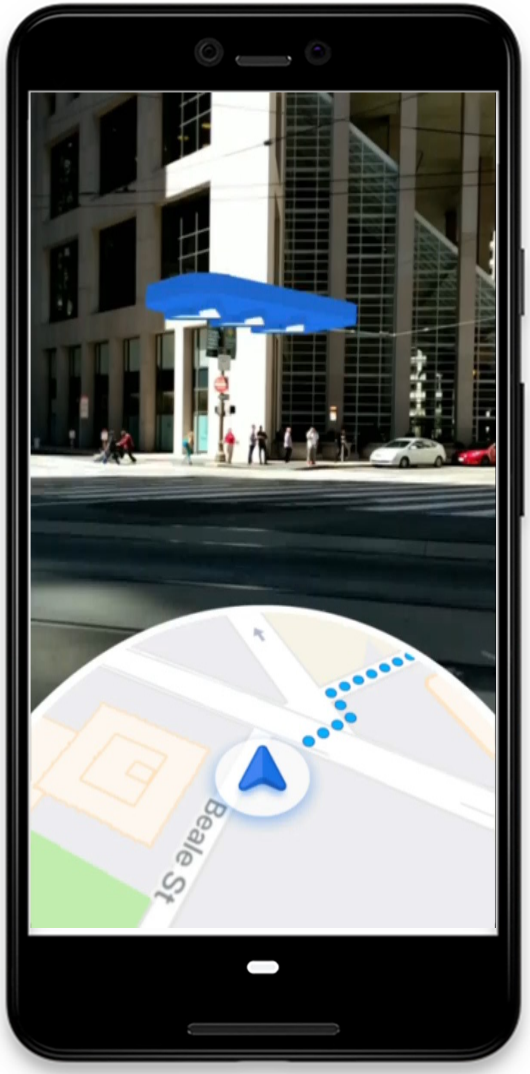






Zeisl, Sattler, Pollefeys ICCV'15

used inside
Google Maps
AR/Live View

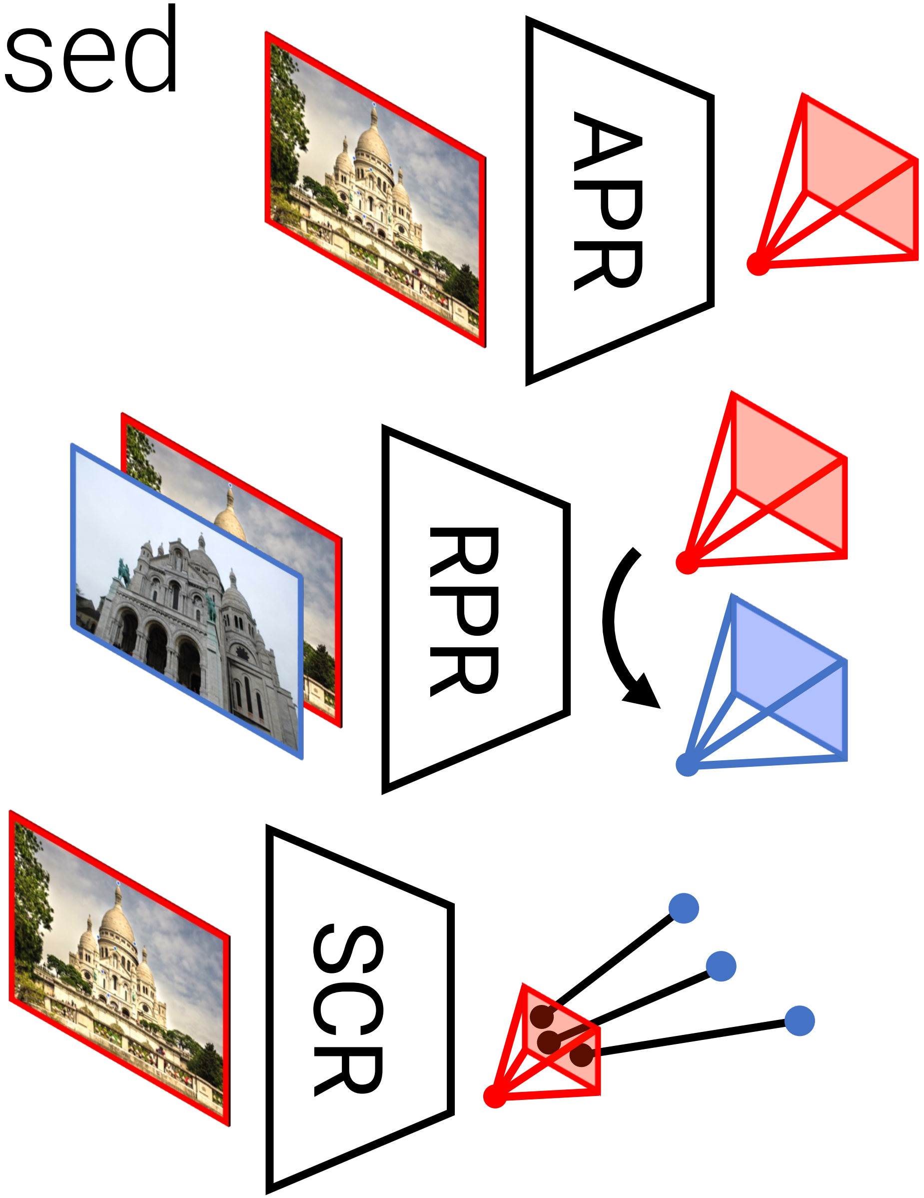


Localization – learning-based

- Regress geometry with deep networks
  - Absolute Pose
  - Relative Pose
  - Pixelwise 2D-3D matches
- Interesting but not practical yet

+ scene compression
+ learn data-dependent prior

- train for each new scene
- low generalization
- PR not as accurate as matching

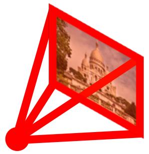


Localization & Mapping – summary

- Build a map:
  - Recover camera poses of mapping images
  - Optionally: triangulate a 3D point cloud with descriptors
- Localize each image individually:
  - feature matching + pose solver (+ image retrieval)
  - or regression via deep networks

Mapping
SfM/SLAM

Localization



Localization & Mapping – summary

- Assumptions:
  - Vision-only localization: single images
  - **Nice images curated for benchmarking**

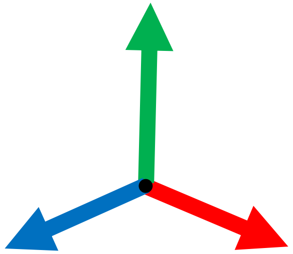
[Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions, Sattler et al., CVPR 2018]

b) Augmented Reality: constraints and opportunities

Why AR needs global localization

- Make virtual objects “stick” to the real word: local SLAM is enough
- Collaborative applications: *share* content between users
- Lifelong: *persist* content across time

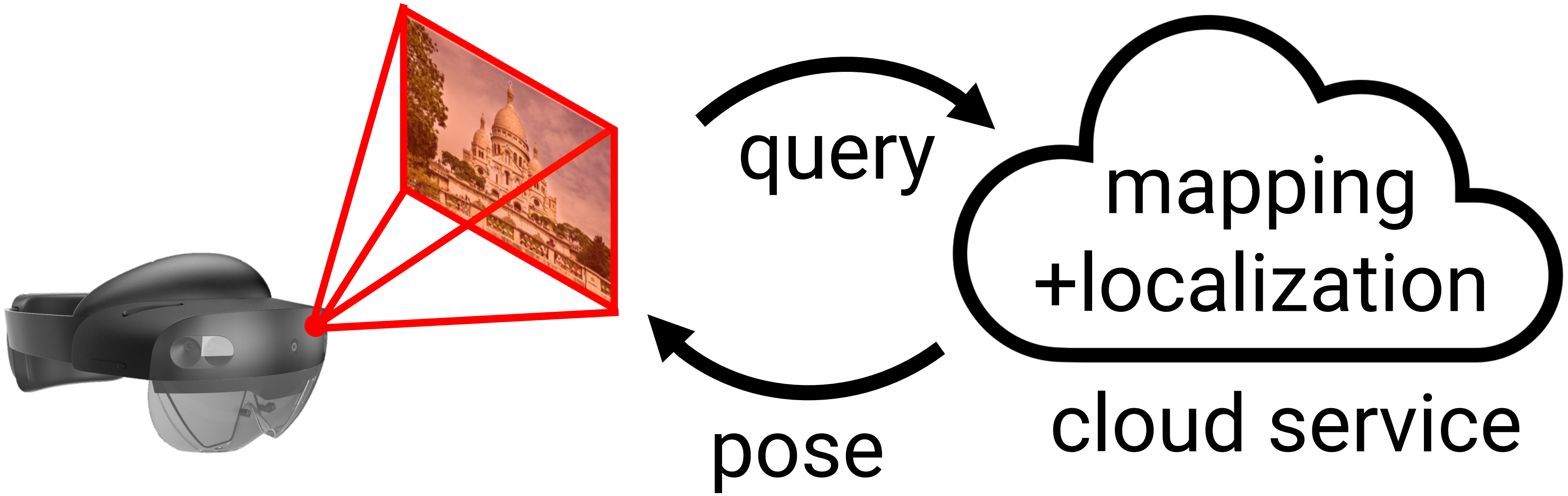
→ Global reference frame
common to all devices



Microsoft HoloLens 2 / Magic Leap 2 / Azure Spatial Anchors

High industry interest

- Commercial localization services
  - Microsoft ASA, Google VPS, Niantic Flagship
- Early demonstration by Middelberg et al.
- Map from AR devices or custom rigs



[Middelberg et al, Scalable 6-DOF Localization on Mobile Devices, ECCV 2014]

Challenges

- Localize regardless of user actions
  - Anywhere: unconstrained views, motions, scenes
  - Anytime: long after the map is built, at night, etc.
- Heterogenous devices:
  - Map and localize with different devices
  - Different cameras, sensors, etc.
- High accuracy requirement: pixel-aligned real-virtual

Apple iPad / Microsoft HoloLens 2 / Meta Project Aria / Magic Leap 2

Challenges

- Hardware limits:

minimize latency	bandwidth vs power	low on-device memory	maintain maps of the world

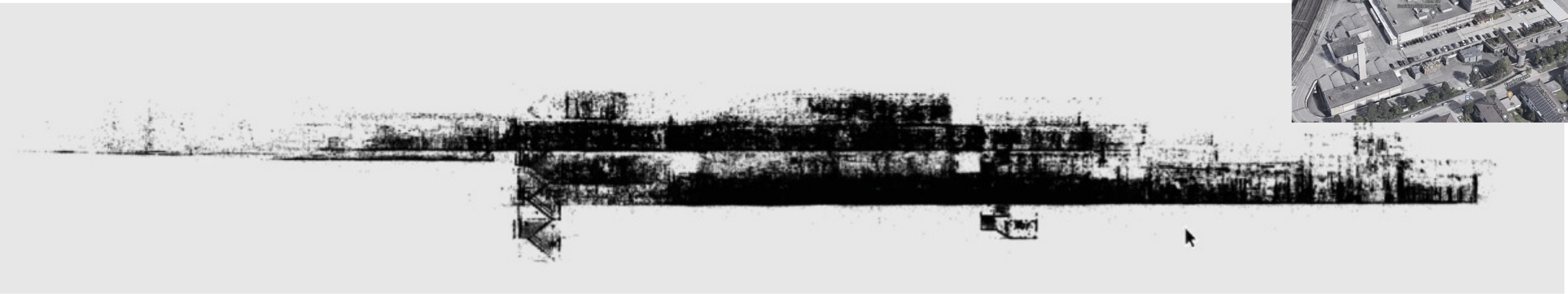
- Balance with cost of devices & data centers

Opportunities

- Multiple sensors: not just a single camera!
  - Multi-camera rig, factory calibrated
  - IMU
  - Radios: Bluetooth & WiFi
  - Depth
  - GPS
- Temporal streams: sensors are continuously in-use
  image sequences instead of single images
  + poses from on-device tracking

Opportunities

- Build map with crowd-sourced sensor data gathered by user
  - More scalable than specialized mapping teams
- Virtually unlimited data if large user base
  - Redundant data: select the best data for mapping



c) Existing datasets and benchmarks

Typical evaluation setup

1. Select a set of mapping images captured at time t
2. Select query images captured at time t+1
3. Build a map from mapping images given ground truth poses
4. Localize query images
5. Evaluate the localization with ground truth poses

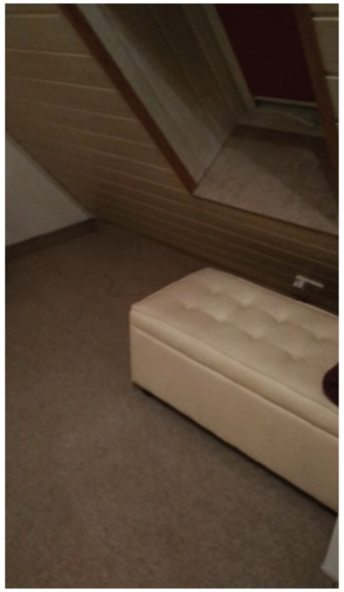
The need for benchmarks

- Incredible progress of the field in the past years
- All fueled by new benchmarks
  - Well-defined query vs mapping splits
  - Private ground truth poses
  - Public leaderboards
- visuallocalization.net, RIO-10, Image Matching Benchmark

Existing datasets – taxonomy

environment **type**

indoor ⟷ outdoor

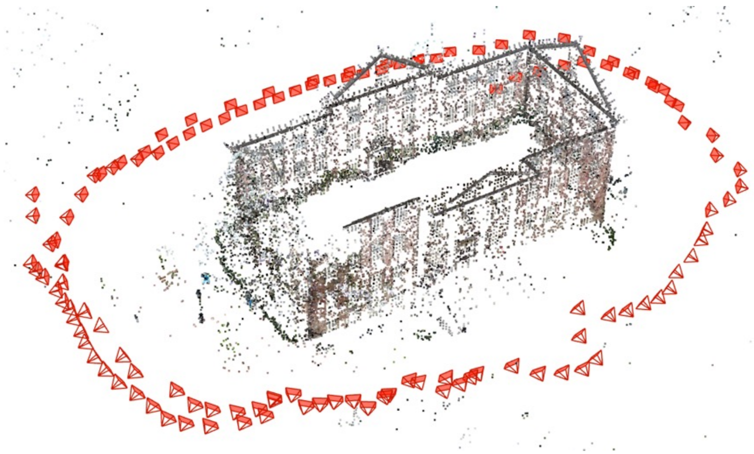


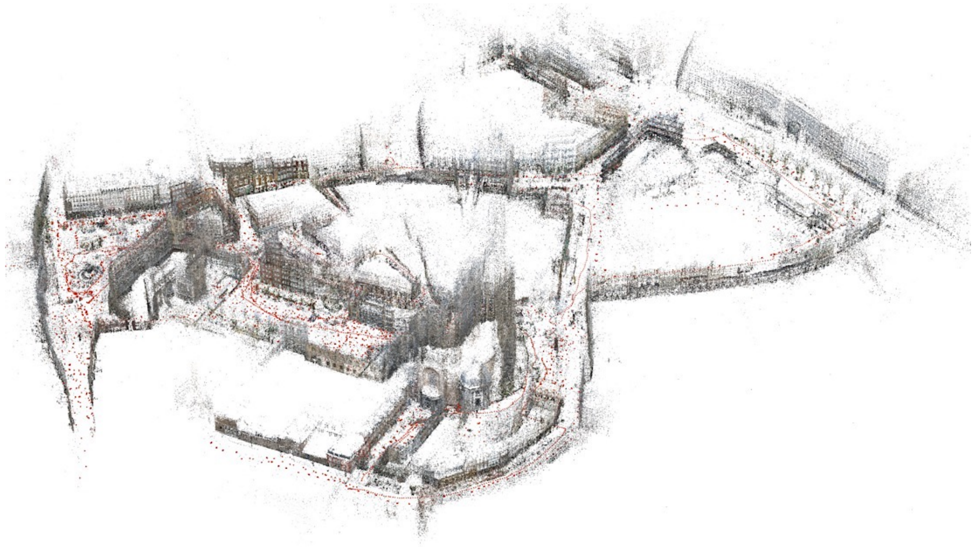
Existing datasets – taxonomy

environment **scale**

room ⟺ city

building





Existing datasets – taxonomy

environment **changes**

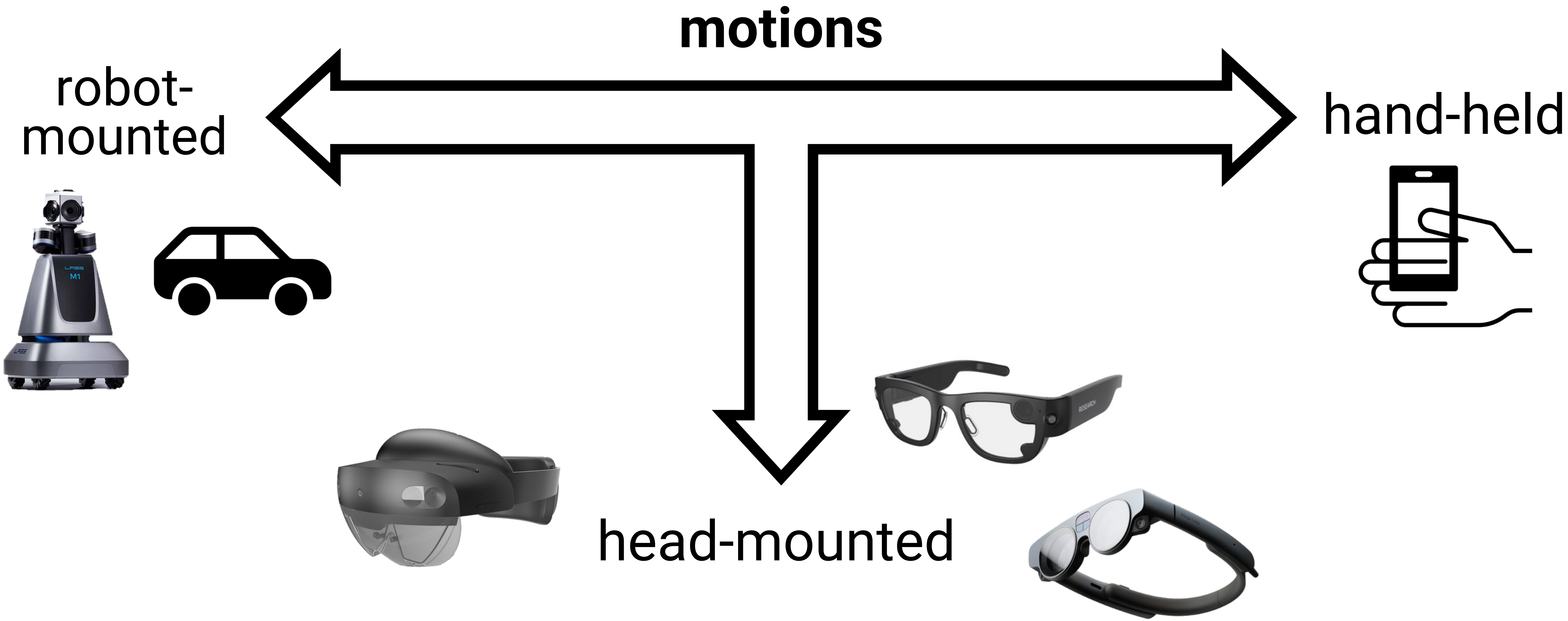
static ⟷ dynamic

Existing datasets – taxonomy

viewpoint diversity

constrained ⟺ arbitrary

Existing datasets – taxonomy



Microsoft HoloLens 2 / Meta Project Aria / Magic Leap 2 / Naver M1X

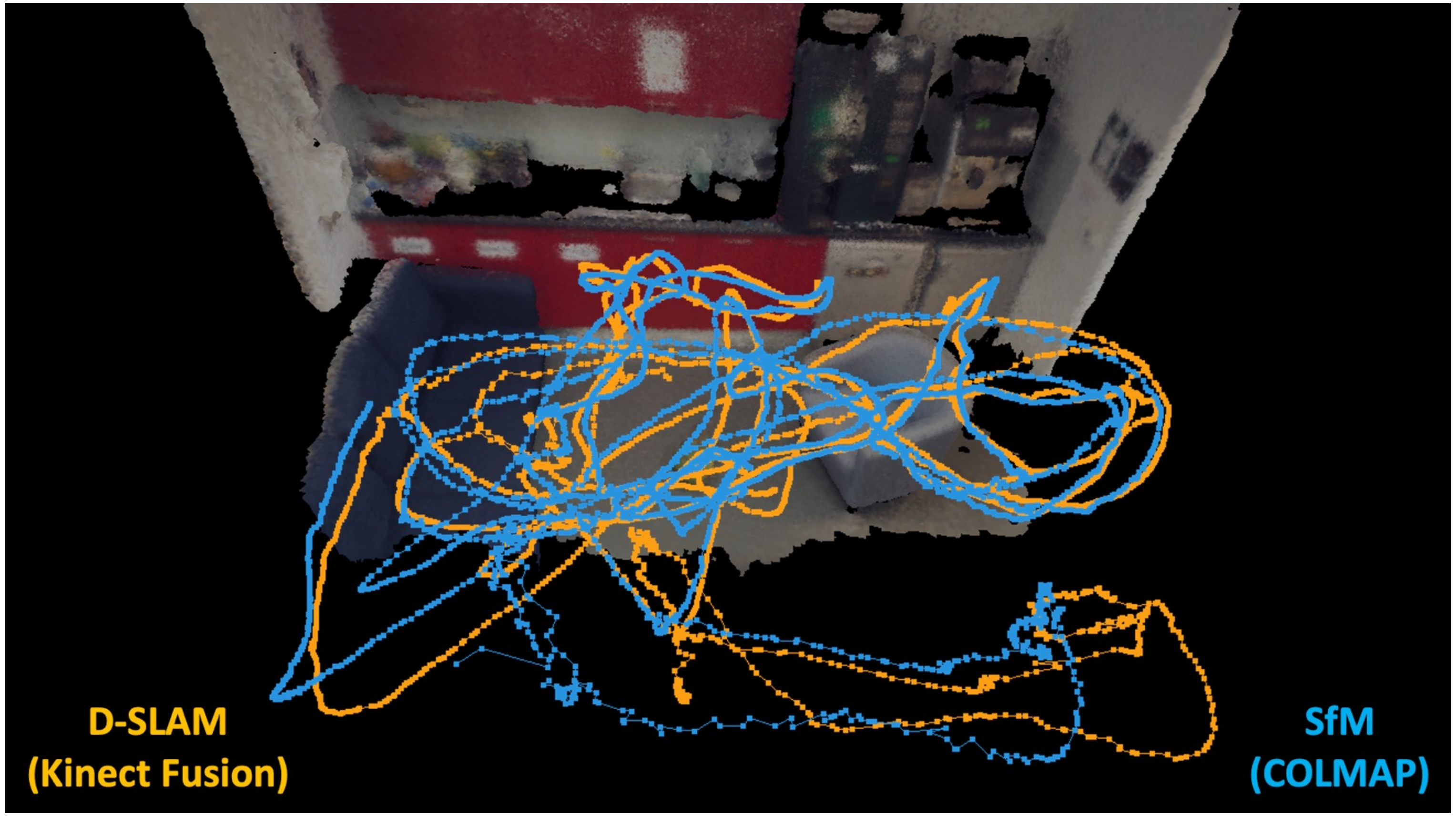
How datasets obtain their ground truth

- Manual labeling of correspondences
- Automatic SfM or SLAM
- Custom rigs with additional sensors like laser scanners
- Detailed 3D models of the scene

Naver M1X / Faro

How datasets obtain their ground truth

- Different GT approaches = different optima
- But there is a unique underlying GT – the geometry of the real world



[On the Limits of Pseudo Ground Truth in Visual Camera Re-Localization, Brachmann et al, ICCV 2021]

Popular datasets

Cambridge Landmarks, 7 scenes

+ pushed forward regression-based localization
+ representative of AR imagery (phones/kinect)
– mostly small-scale though
– (single) image only

[PoseNet: A Convolutional Network for Real-Time 6-DoF Camera Relocalization, Kendall et al, ICCV 2015]
[Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images, Shotton et al, CVPR 2013]

Popular datasets

Aachen Day-Night, PhotoTourism, San Francisco

+ pushed forward learned features & matching (esp night)
+ larger scale → push for scalability
+ long-term changes
+ crowd-sourced multi-device

– no guarantees in ground truth accuracy
– nice images from constrained viewpoints

[Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions, Sattler et al, CVPR 2018]
[Image Matching across Wide Baselines: From Paper to Practice, Jin et al, IJCV 2020]
[San Francisco Landmark Dataset for Mobile Landmark Recognition, Chen et al, 2011]

Popular datasets

Inloc, Baidu Mall, Naver Labs, RIO-10

+ pushed forward indoor localization
+ idea of moving beyond points
+ structural changes (furniture)
– InLoc: sparse mapping images
– Naver Labs: robot motion
– Most: only images

[InLoc: Indoor Visual Localization with Dense Matching and View Synthesis, Taira et al, CVPR 2018]
[A dataset for benchmarking image-based localization, Sun et al, CVPR 2017]
[Large-scale Localization Datasets in Crowded Indoor Spaces, Lee et al, CVPR 2021]
[Beyond Controlled Environments: 3D Camera Re-Localization in Changing Indoor Scenes, Wald et al, ECCV 2020]

Popular datasets

ETH3D

+ pushed dense reconstruction / MVS forward
+ millimeter accurate ground truth
– sensors / trajectories / views not representative of AR
– very sparse views
– no scene changes

[A Multi-view Stereo Benchmark with High-Resolution Images and Multi-camera Videos, Schöps et al, CVPR 2017]

Popular datasets

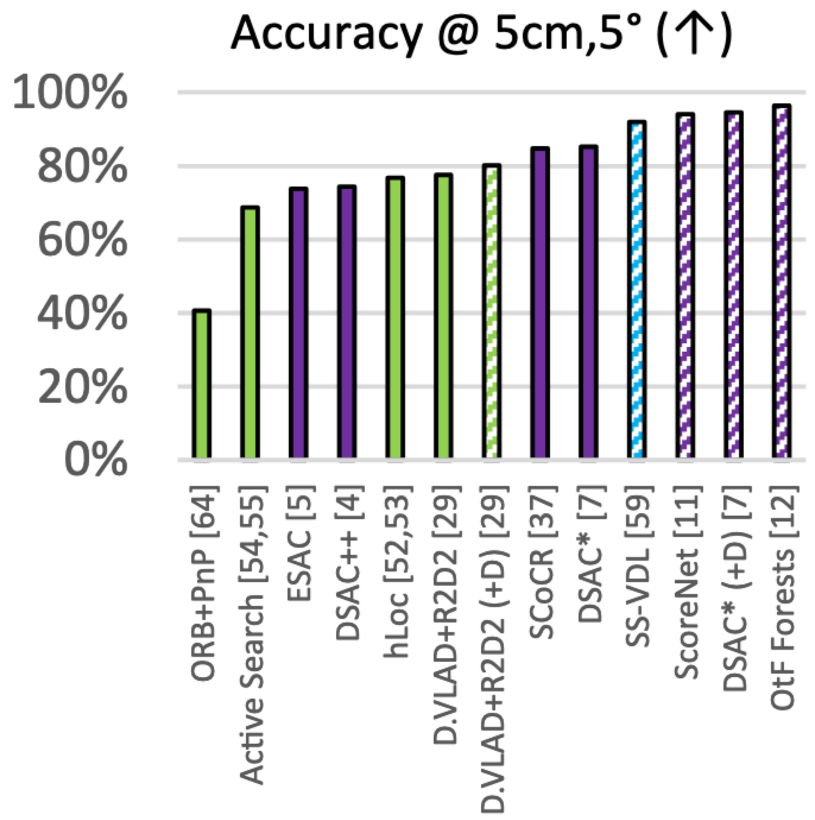
dataset	out/indoor	changes	scale	density	camera motion	imaging devices	additional sensors	ground truth	accuracy
Aachen [67,66]	✅❌	🌓🏗	★★⯪	★★☆	still images	DSLR	❌	SfM	>dm
Phototourism [34]	✅❌	🏃🏗	⯪☆☆	★★★	still images	DSLR, phone	❌	SfM	∼m
San Francisco [14]	✅❌	🏃🏗	★★★	★☆☆	still images	DSLR, phone	GNSS	SfM+GNSS	∼m
Cambridge [37]	✅❌	🏃🌨	⯪☆☆	★★☆	handheld	mobile	❌	SfM	>dm
7Scenes [73]	❌✅	❌	⯪☆☆	★★★	handheld	mobile	depth	RGB-D	∼cm
RIO10 [84]	❌✅	🪑	⯪☆☆	★★★	handheld	Tango tablet	depth	VIO	>dm
InLoc [77]	❌✅	🪑	★⯪☆	⯪☆☆	still images	panoramas, phone	lidar	manual+lidar	>dm
Baidu mall [76]	❌✅	🏃	★⯪☆	★★☆	still images	DSLR, phone	lidar	manual+lidar	∼dm
Naver Labs [40]	❌✅	🏃🪑	★★☆	★★☆	robot-mounted	fisheye, phone	lidar	lidar+SfM	∼dm
NCLT [12]	✅✅	🌨🪑	★★☆	★★☆	robot-mounted	wide-angle	lidar, IMU, GNSS	lidar+VIO	∼dm
ADVIO [57]	✅✅	🏃	★★☆	⯪☆☆	handheld	phone, Tango	IMU, depth, GNSS	manual+VIO	∼m
ETH3D [71]	✅✅	❌	⯪☆☆	★★☆	handheld	DSLR, wide-angle	lidar	manual+lidar	∼mm

Popular benchmarks are often saturated

Aachen Day-Night v1.1

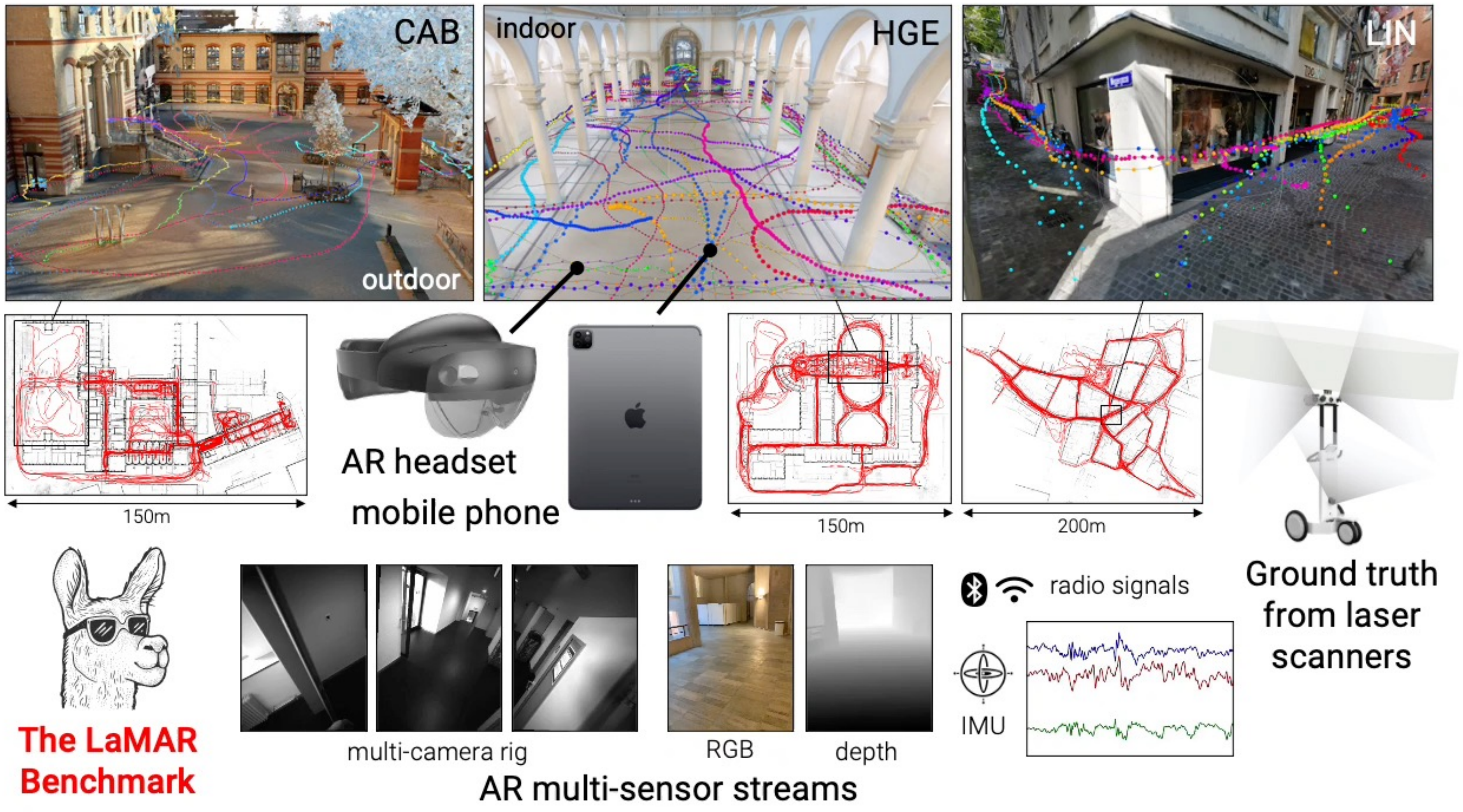
Method	day	night
MegLoc	90.5 / 97.3 / 99.8	77.5 / 92.7 / 100.0
OSpace	91.3 / 97.2 / 99.6	81.2 / 92.1 / 100.0
HHloc	90.5 / 97.1 / 99.8	77.0 / 92.1 / 100.0
4Fun	91.5 / 97.1 / 99.6	78.5 / 91.6 / 99.0
PtLine	90.0 / 96.7 / 99.5	80.6 / 92.1 / 100.0
KAPTURE-R2D2-FUSION-50-PYCOLMAP	91.3 / 97.0 / 99.5	78.5 / 91.6 / 100.0
KAPTURE-Fast-R2D2-FUSION-50-PYCOLMAP	91.0 / 96.6 / 99.6	78.5 / 91.6 / 100.0
KAPTURE-R2D2-FUSION	90.9 / 96.7 / 99.5	78.5 / 91.1 / 98.4
hloc-fusion	90.5 / 96.5 / 99.6	76.4 / 90.6 / 99.0
RLOCS_v3.0	89.8 / 96.7 / 99.5	74.9 / 90.6 / 100.0
DFM	90.3 / 96.5 / 99.5	74.3 / 91.6 / 99.5
KRNet	89.7 / 96.5 / 99.4	77.5 / 90.6 / 100.0
Hierarchical Localization - SuperPoint + SuperGlue	89.8 / 96.1 / 99.4	77.0 / 90.6 / 100.0

7 Scenes



We need a new benchmark

We need a new benchmark



Q&A