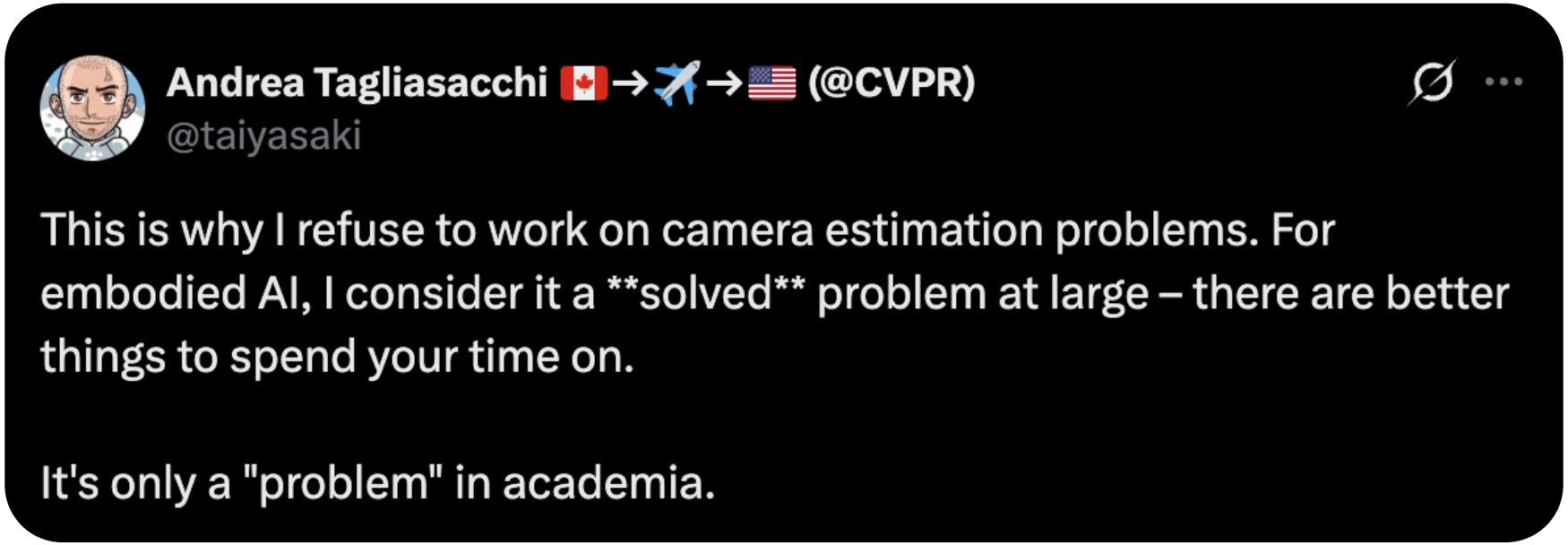
From Local Reconstruction to Global Geolocalization

Paul-Edouard Sarlin

Image Matching Workshop, CVPR 2026
2026-06-04

Spicy takes 🌶️

Last month on Twitter:



solved = achievable with high certainty given sufficient resources

is camera pose estimation already solved?

Yes when:

- We control the hardware
- We afford IMUs and/or multiple cameras
- We build a bespoke software stack

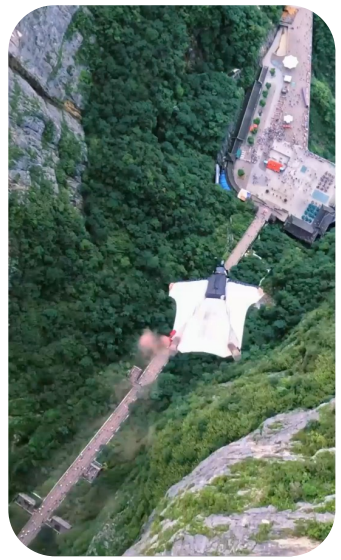
This covers much of:

- AR/VR – see Meta's Aria
- Mapping – see Google StreetView
- Robotics

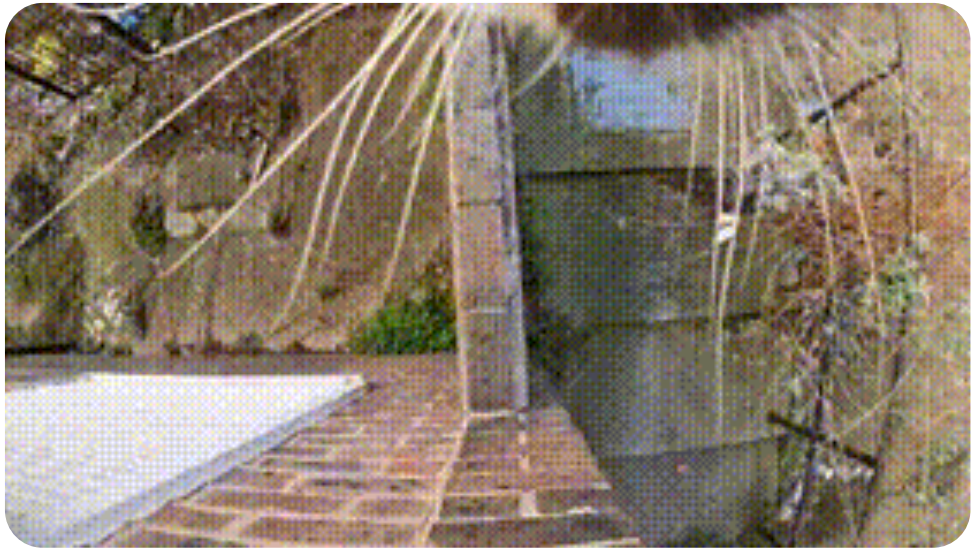
is camera pose estimation already solved?

Not for monocular videos:

- Internet videos
- Historical archives
- Consumer devices (phones)



Not if we refer to mapping:
3D for every single pixel



Benchmarking Egocentric Visual-Inertial SLAM at City Scale

Anusha Krishnan[1*] Shaohui Liu[1*] Paul-Edouard Sarlin[2*] Oscar Gentilhomme[1]
David Caruso[3] Maurizio Monge[3] Richard Newcombe[3] Jakob Engel[3] Marc Pollefeys[1,4]
[1]ETH Zurich [2]Google [3]Meta Reality Labs Research [4]Microsoft Spatial AI Lab

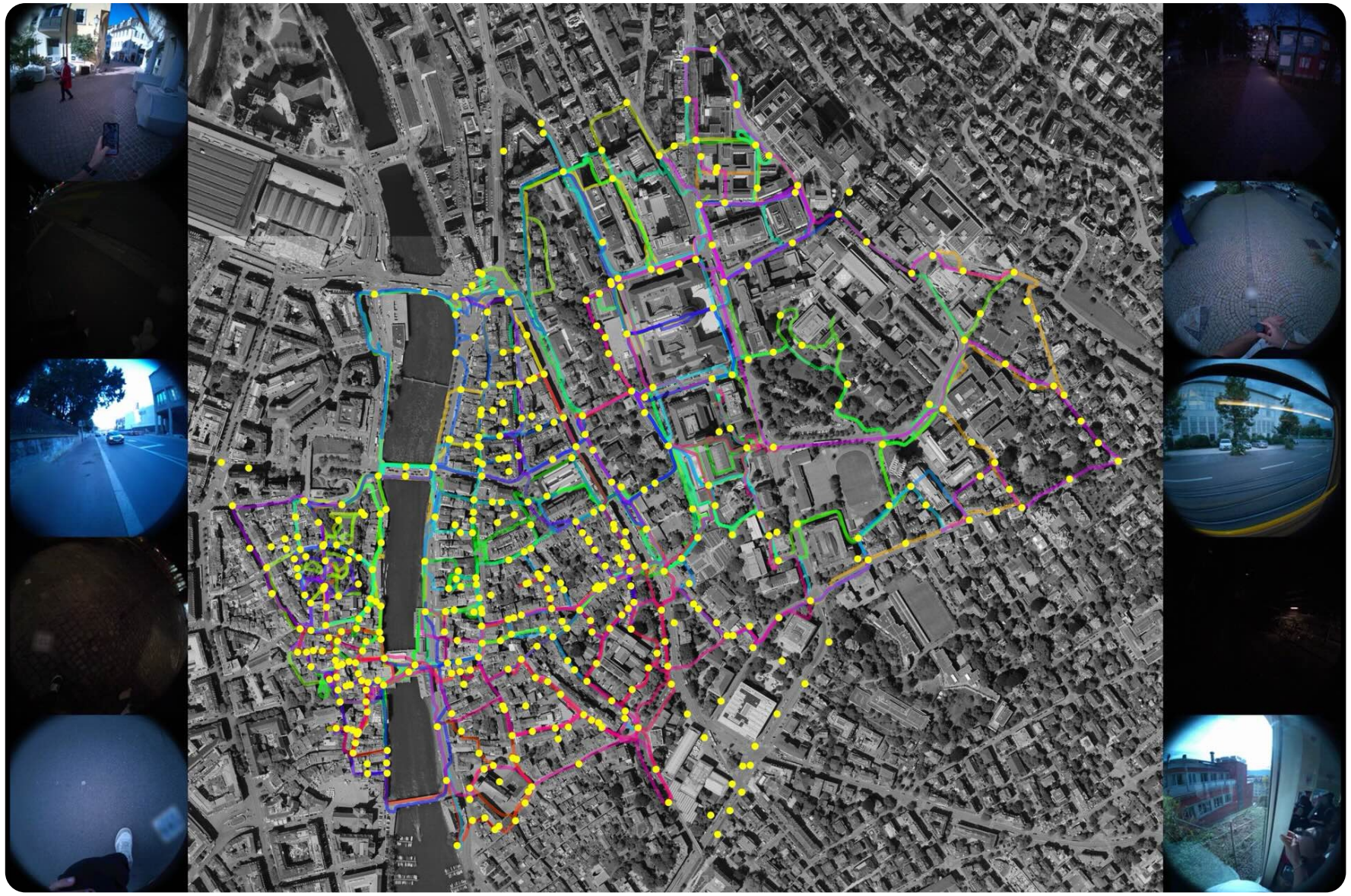
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Computer Vision and Geometry Lab ETH zürich

Meta Microsoft Google

lamaria.ethz.ch

The LaMAria dataset



70km
22 hours
10-45min each

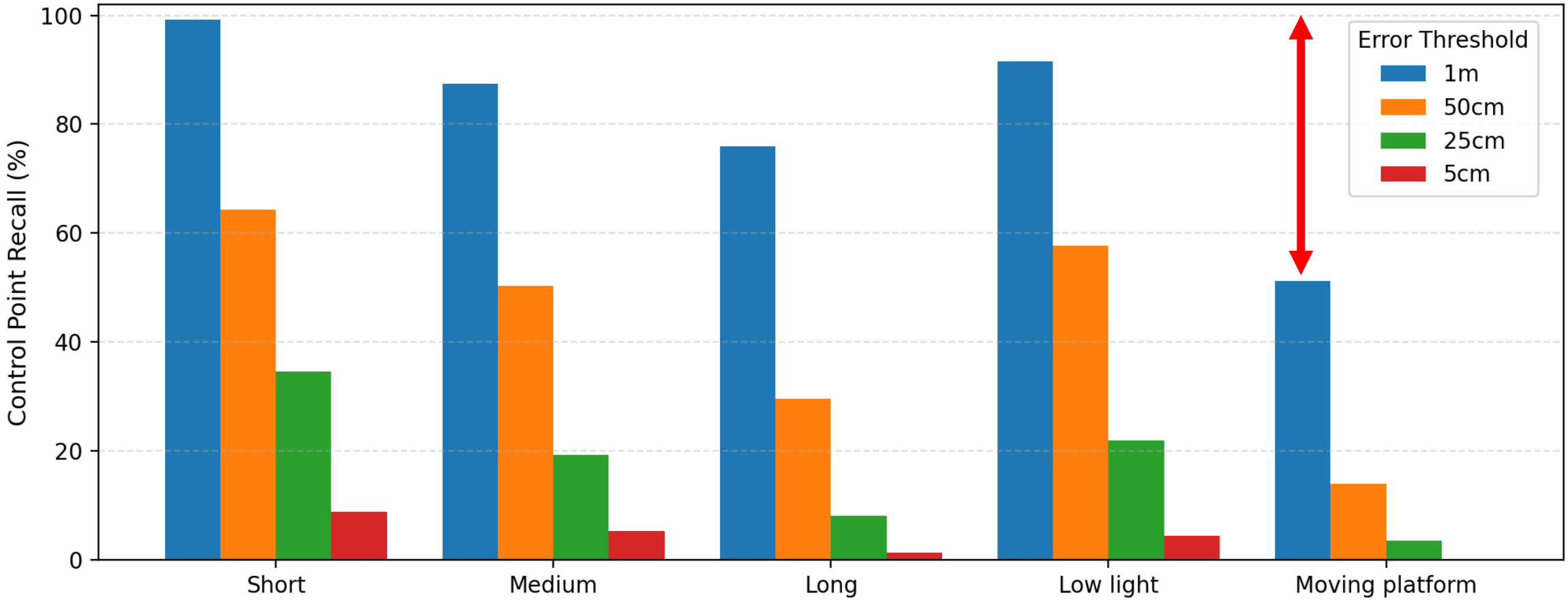
Open-source systems are far behind industry

🟠 monocular 🟢 monocular inertial 🔵 multi-camera inertial

method	causal	short			medium			long			challenge – low-light		
		score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑	score ↑	CP@1m ↑	R@5m ↑
DPVO	✓	9.4	1.7	21.3	5.2	1.0	10.8	1.2	0.0	1.9	3.4	0.2	7.5
DPV-SLAM	x	7.5	1.5	14.8	5.2	1.4	10.1	0.4	0.0	0.7	1.9	0.4	3.5
Kimera VIO	✓	6.3	2.9	12.6	6.6	1.7	15.1	6.3	1.7	14.3	4.2	2.7	6.4
ORB-SLAM3	x	28.3	13.4	67.1	20.3	4.4	57.0	14.2	2.3	40.6	6.2	0.6	12.5
OpenVINS	✓	18.1	4.4	45.7	10.9	2.3	27.9	4.7	0.5	12.3	7.9	2.4	17.6
OpenVINS + Maplab	x	22.9	8.1	50.8	13.1	4.1	29.0	5.8	1.3	13.3	9.6	2.9	19.3
OpenVINS	✓	22.2	6.2	57.9	17.8	5.7	46.1	10.6	1.7	25.8	16.9	6.2	38.2
OpenVINS + Maplab	x	26.0	9.5	61.1	21.3	7.3	50.6	12.6	1.9	30.3	16.5	4.6	37.9
OKVIS2	x	24.2	12.0	54.7	13.6	6.8	28.2	3.6	2.7	7.2	15.4	5.4	38.6
Aria's SLAM	x	90.7	99.2	–	78.5	87.4	–	70.8	75.9	–	84.2	91.6	–

Aria's MPS SLAM is far ahead of all current academic solutions.

Multi-camera inertial SLAM is not *fully* solved



Recall @ X of Aria's SLAM

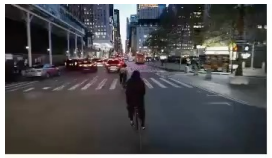
E2E models for mono SLAM

Great test bed for long-range architectures

MUCH recent interest: LoGeR , Lingbotmap, ZipMap, TT3R, …

This looks solved, right?

[LoGeR, 2026]
[Lingbotmap, 2026]



NYC, Outdoor, Dynamics, Night

Evaluated on easier datasets: KITTI, VBR, TUM-VO

LaMAria is way too hard for e2e models: fisheye, blur, long drift

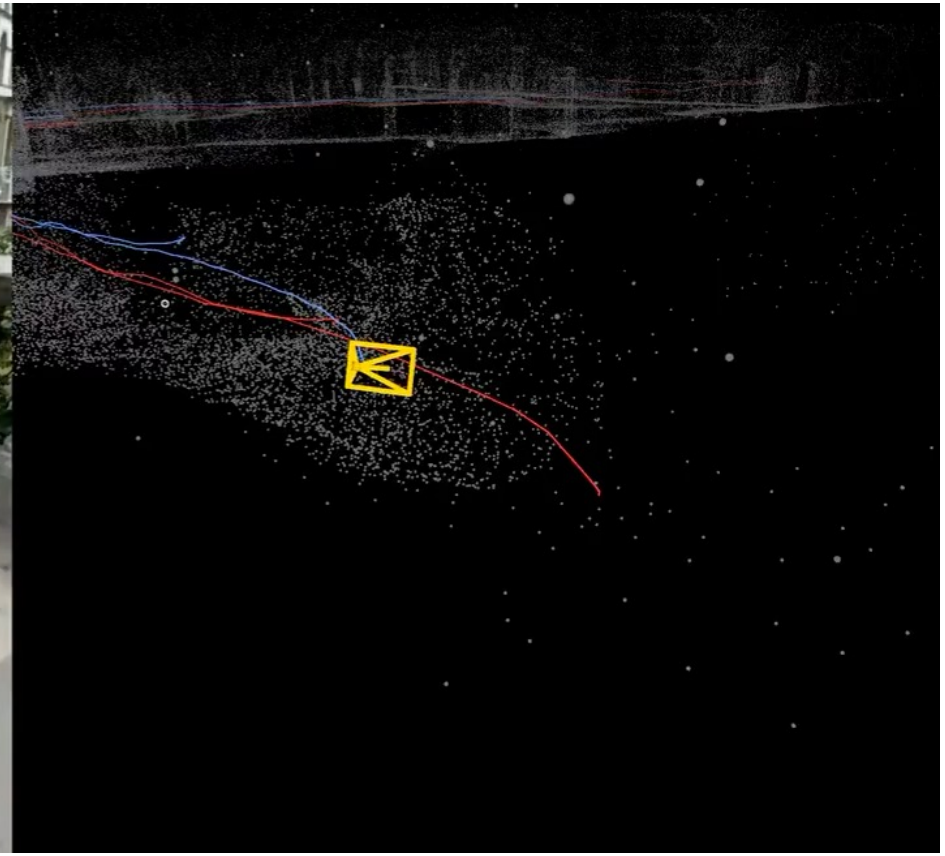
Are we comparing against the right classical baselines?

SLAM
for sequences
designed for real-time
Causal, limited robustness
Fast and efficient
Limited self-calibration

Structure-from-Motion
for unordered collections
designed for robust offline processing
Treats videos like bags of images
Slow for large inputs
Can process uncalibrated inputs

Can't we have the best of both?

Introducing VidMap: Geometry Strikes Back



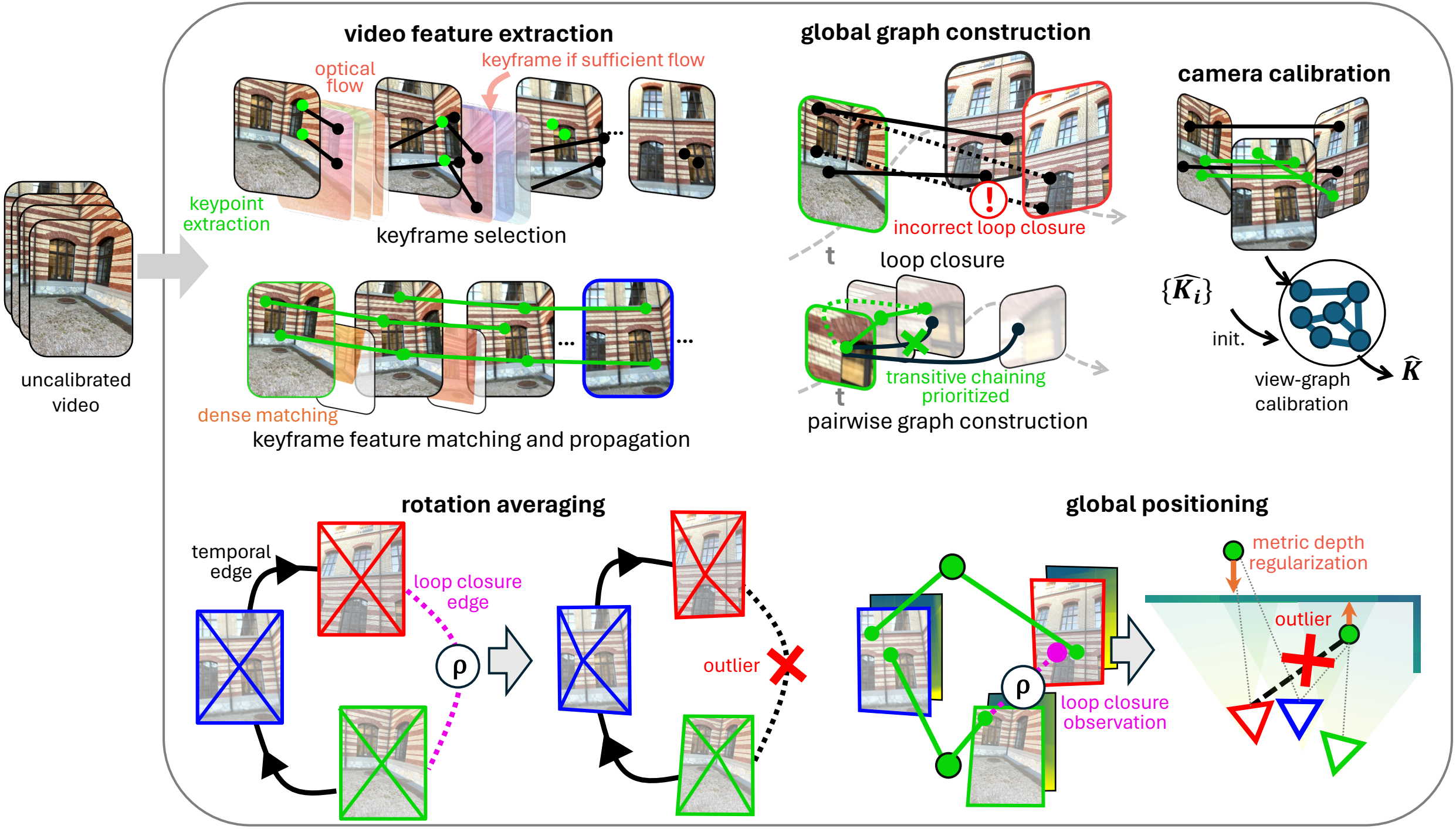
work done with Zador Pataki, coming soon to GitHub!

How it works

	SLAM	SfM	VidMap
Global, acausal	❌	✅ GLOMAP	✅
Adaptive keyframing	✅	❌	✅
Temporal constraints	✅	❌	✅
Offline, maximize robustness	❌	✅	✅

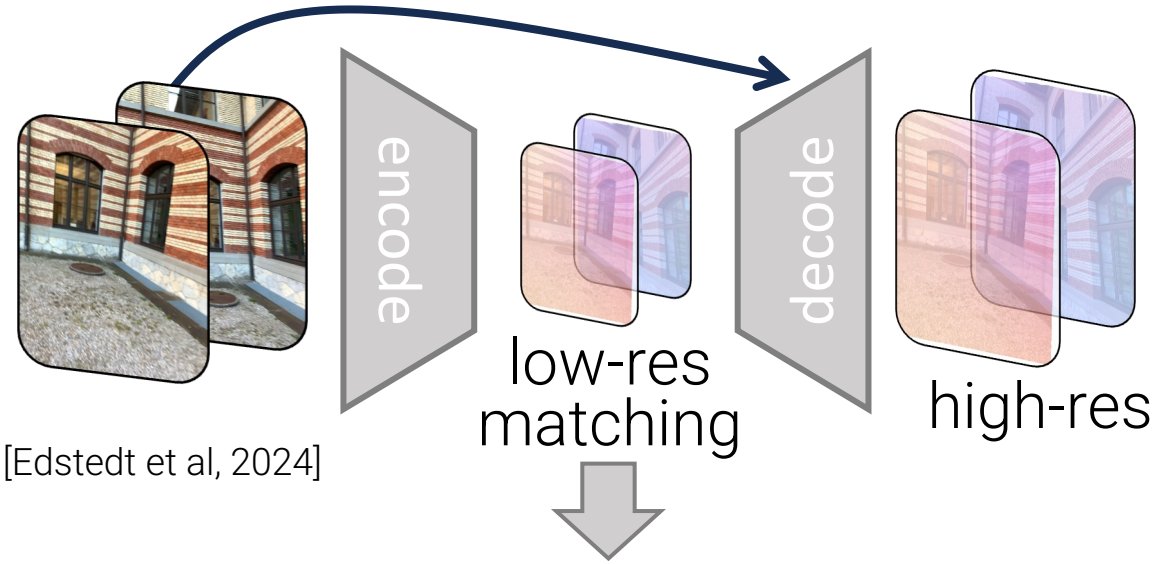
Take the best of SfM and SLAM by augmenting global SfM with:

- **Modern learned priors**: dense matching, depth, calibration, gravity
- **Adaptive keyframing** (fast!)
- **Temporal constraints**: tracking vs loop closure


uncalibrated video
video feature extraction
optical flow
keyframe if sufficient flow
keypoint extraction
keyframe selection
dense matching
keyframe feature matching and propagation
global graph construction
incorrect loop closure
t
loop closure
transitive chaining prioritized
t
pairwise graph construction
camera calibration
{K̂ᵢ}
init.
K̂
view-graph calibration
rotation averaging
temporal edge
loop closure edge
ρ
outlier
global positioning
ρ
loop closure observation
metric depth regularization
outlier

Tracking sparse points with dense matching

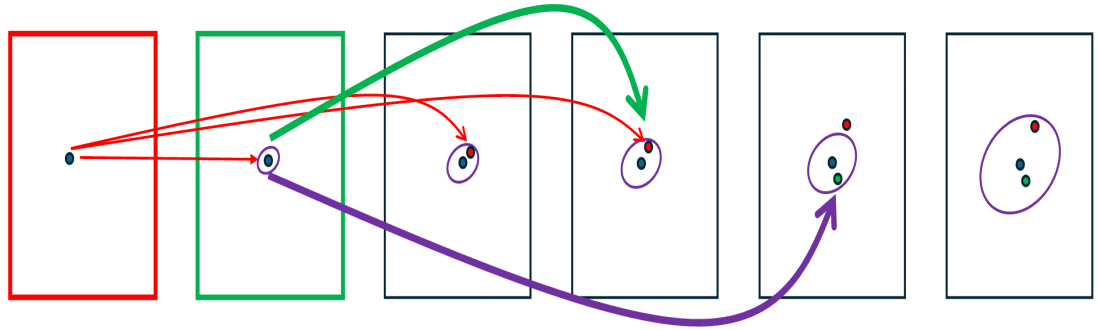
low-latency matching with RoMa



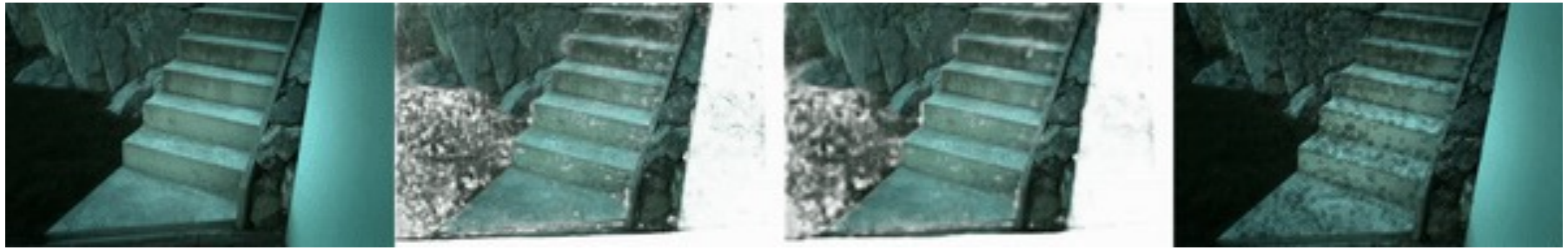
[Edstedt et al, 2024]

keyframe selection, early exit

transitive flow propagation



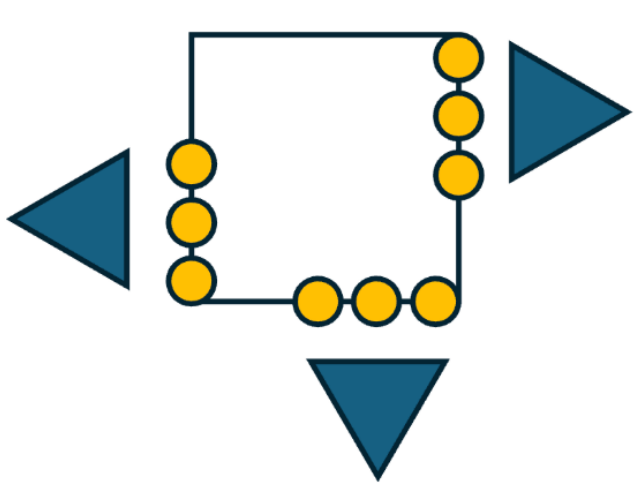
inspired by MFTIQ [Serych et al, 2024]



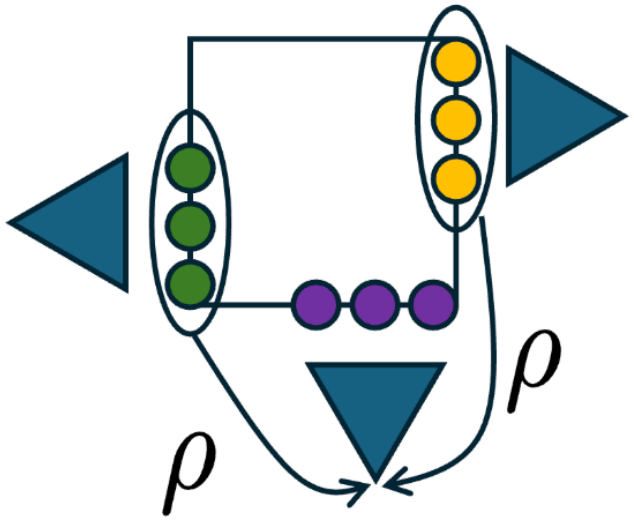
Handling symmetries

- Track duplication across tracking windows
- Prioritize temporal edges in verification and optimization

[IMW 2025]



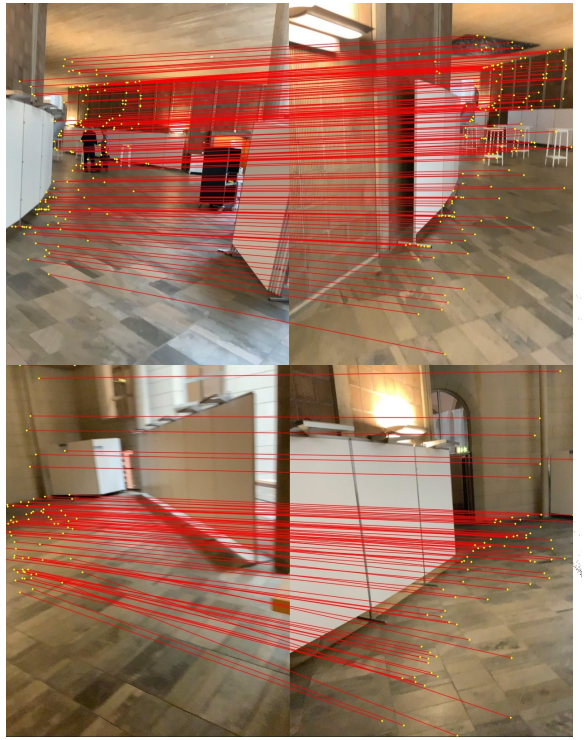
global mapping
GLOMAP

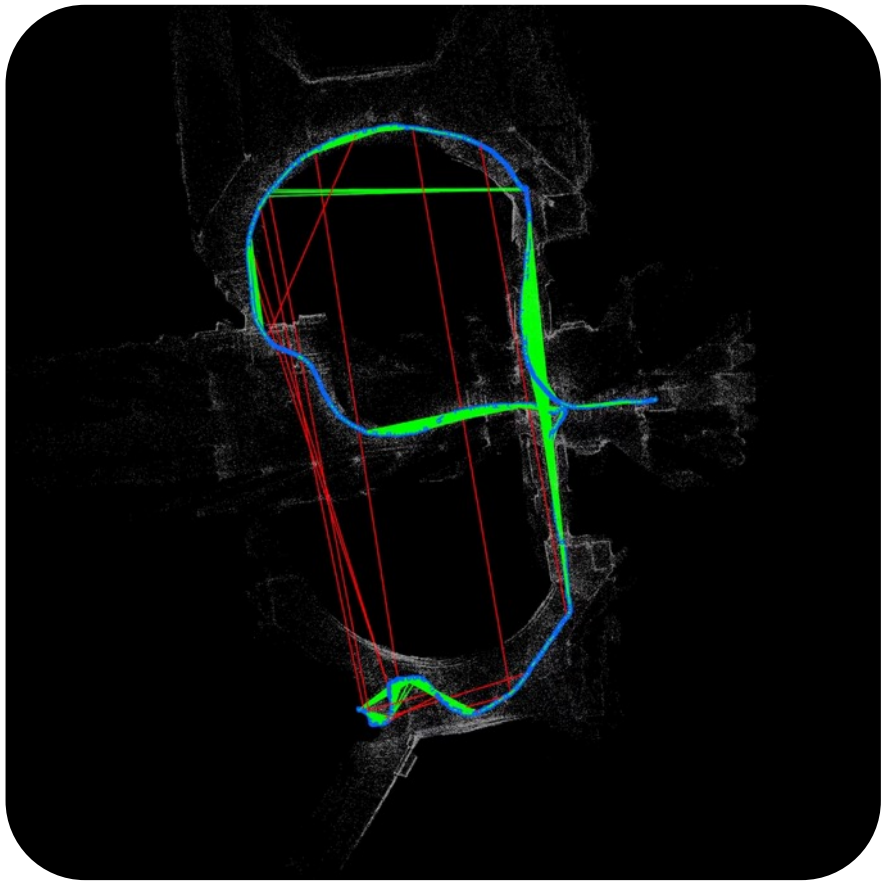


ours

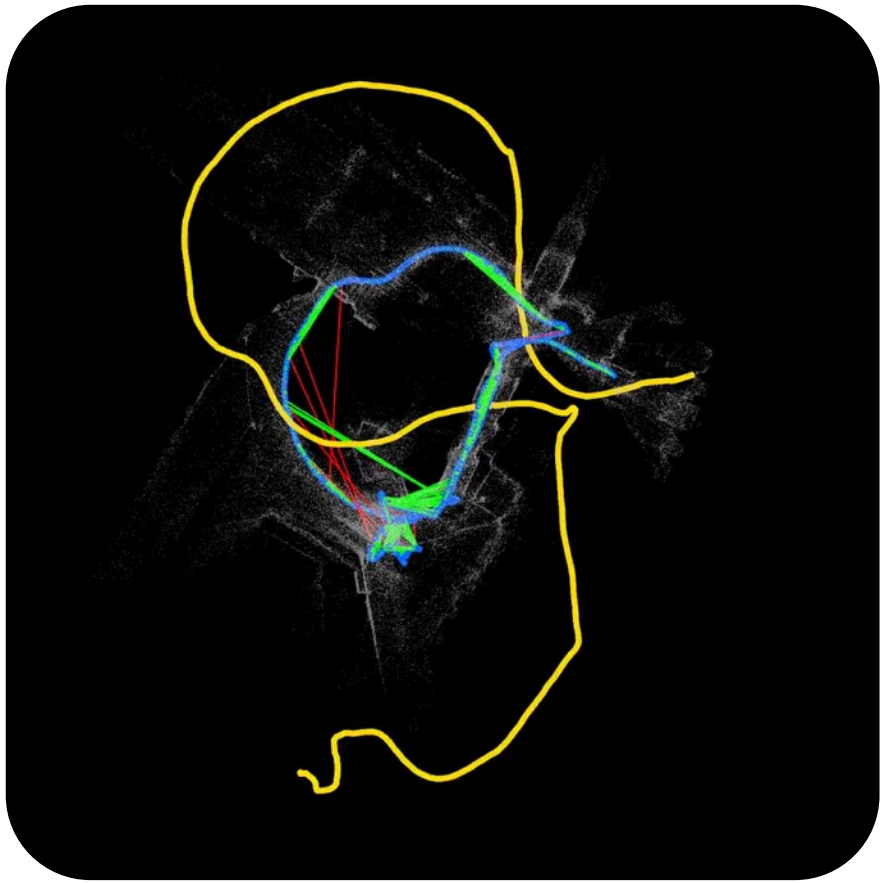
Addressing failures of classical approaches

Symmetries





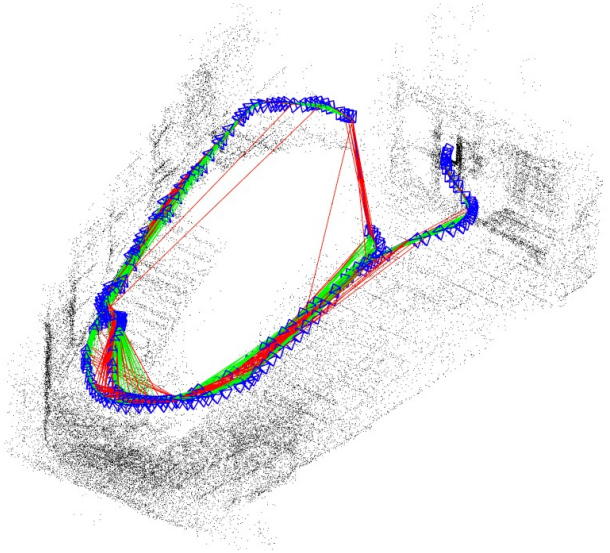
VidMap

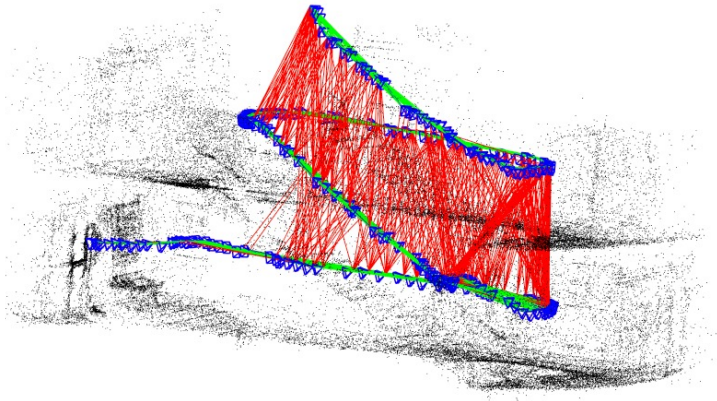


w/o handling symmetries

Addressing failures of classical approaches

Symmetries

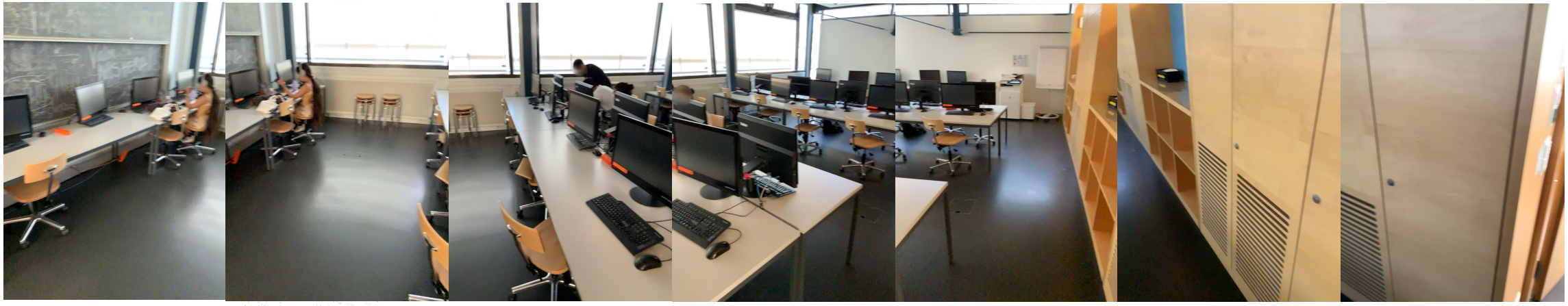


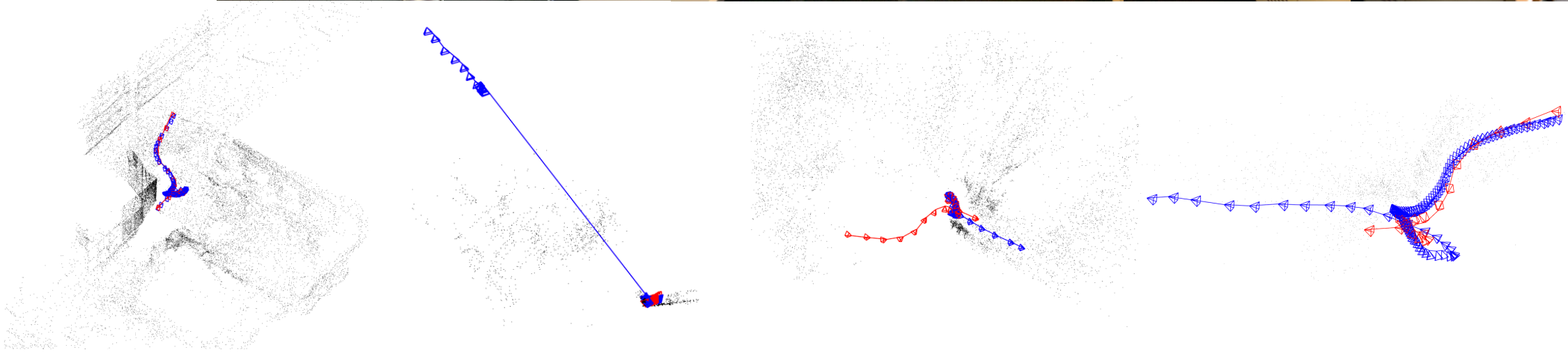


Visual cues aren’t sufficient!

Addressing failures of classical approaches

Pure rotation





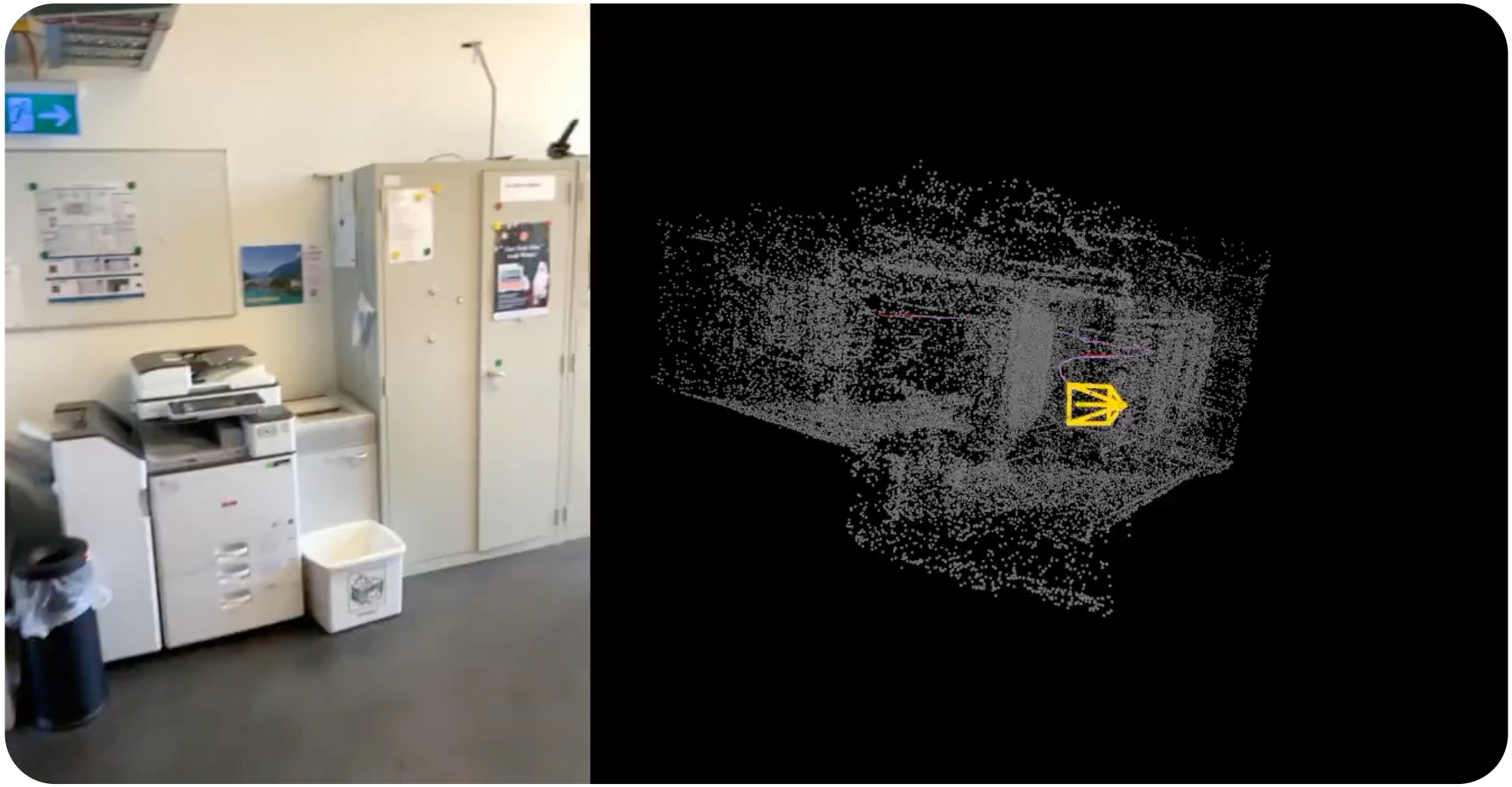
Ours | GLOMAP | COLMAP | DPV-SLAM

Addressing failures of classical approaches

Pure rotation

Addressing failures of classical approaches

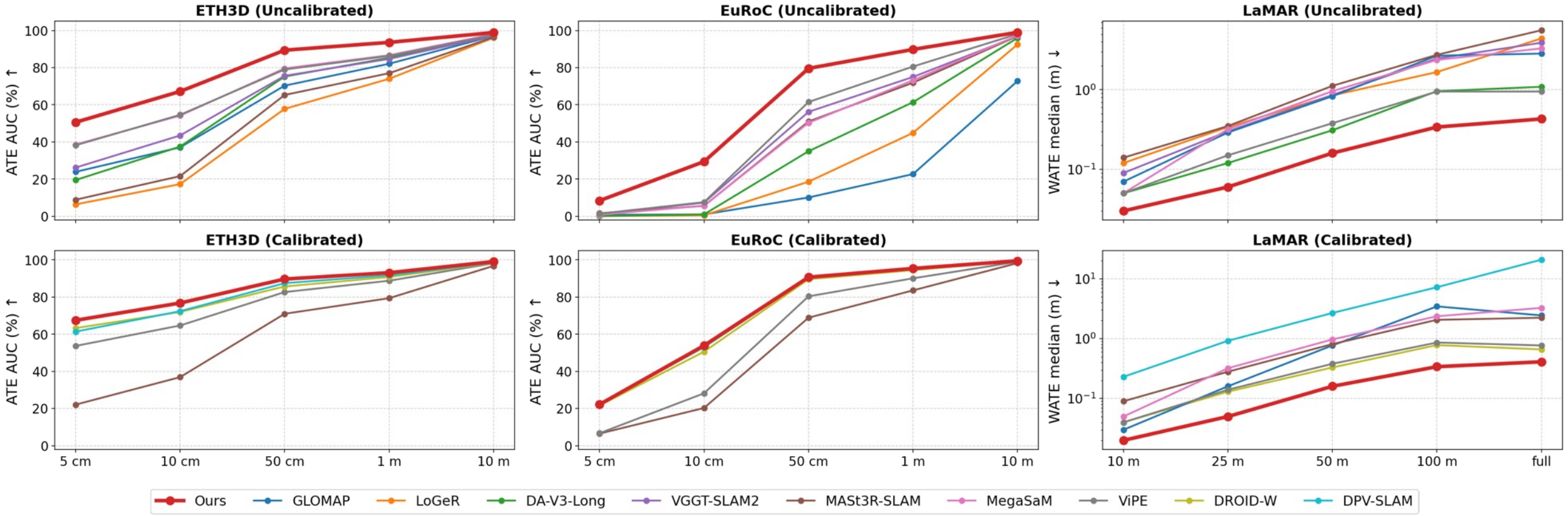
Motion blur



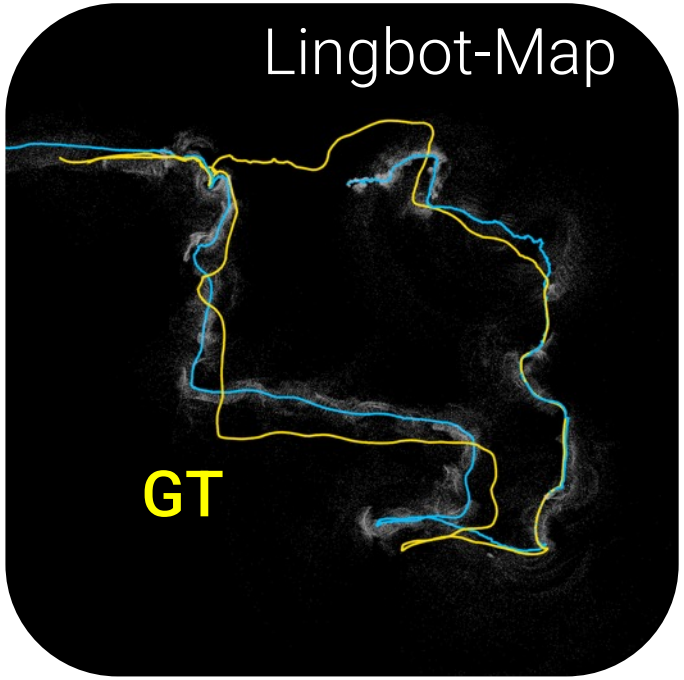
Results on hard benchmarks

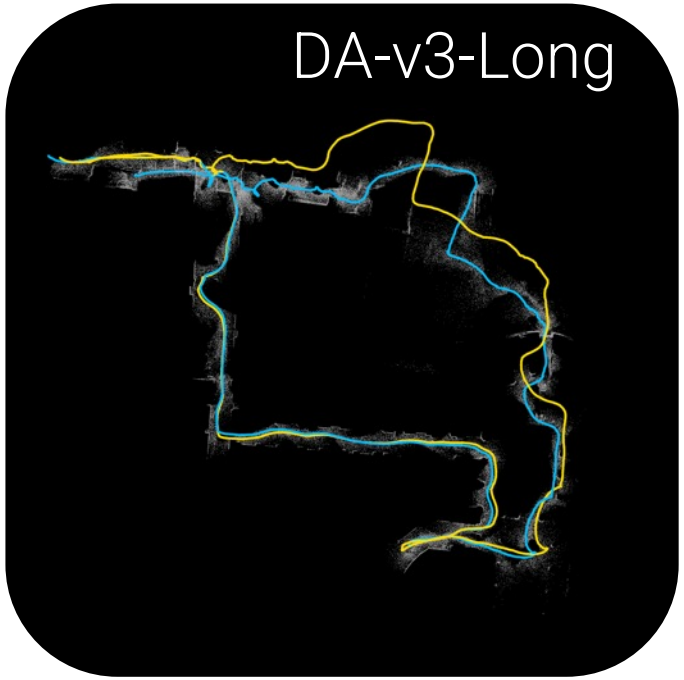
AUC (higher=better) on ETH3D & EuRoC

ATE (lower=better) on phone videos

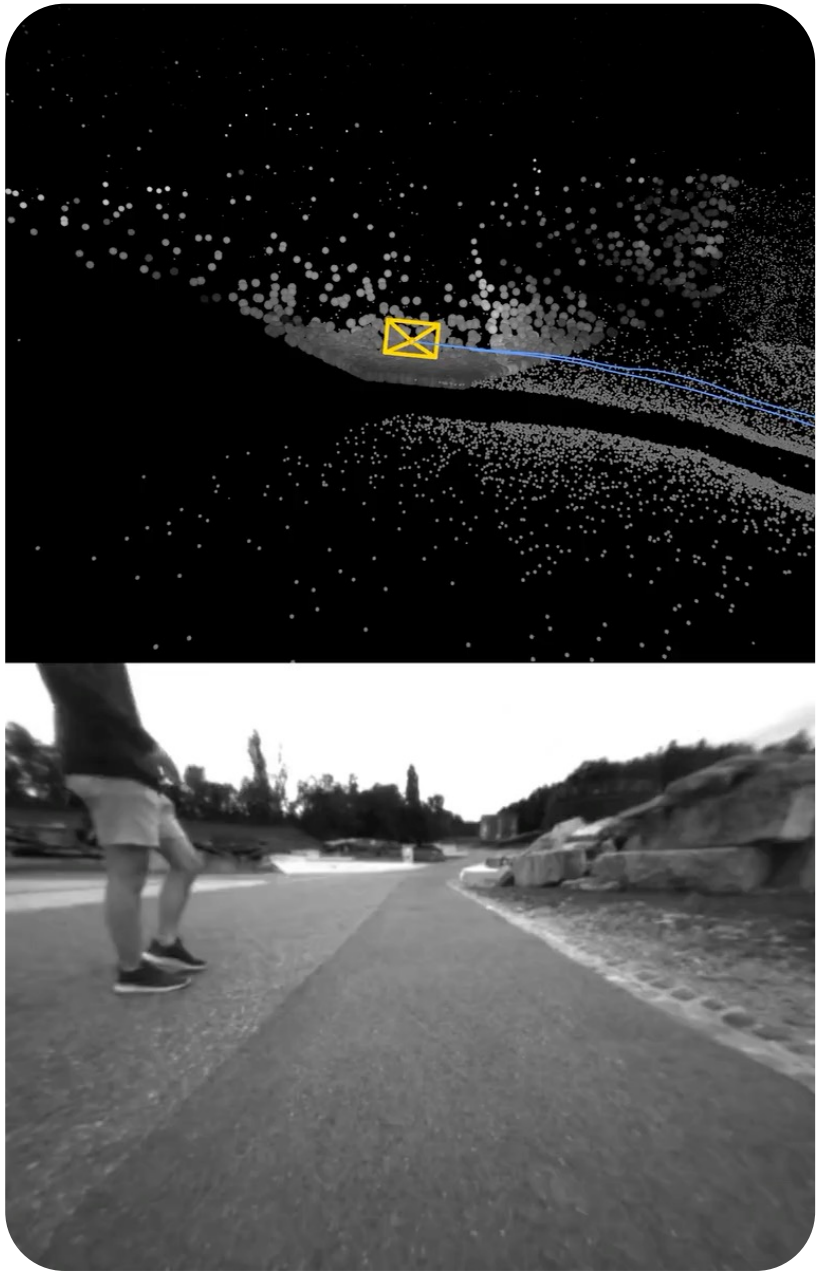


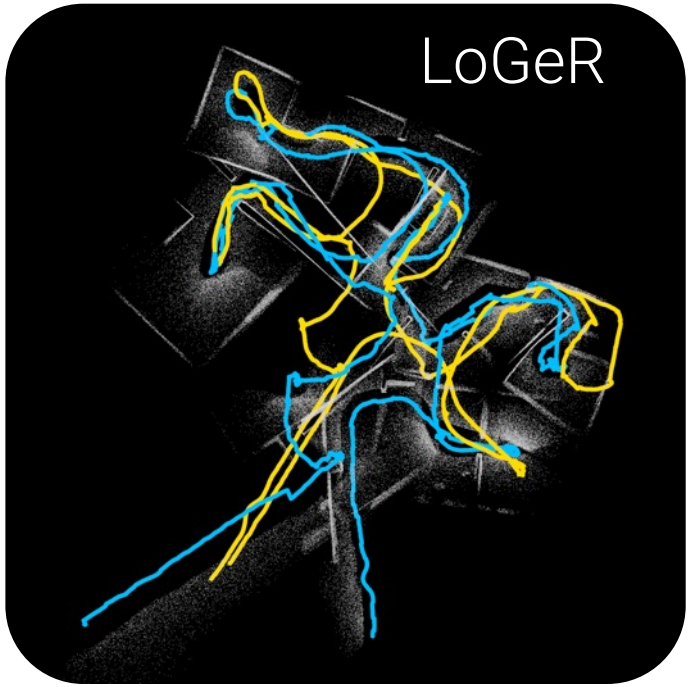
Outperforms e2e models in robustness and accuracy

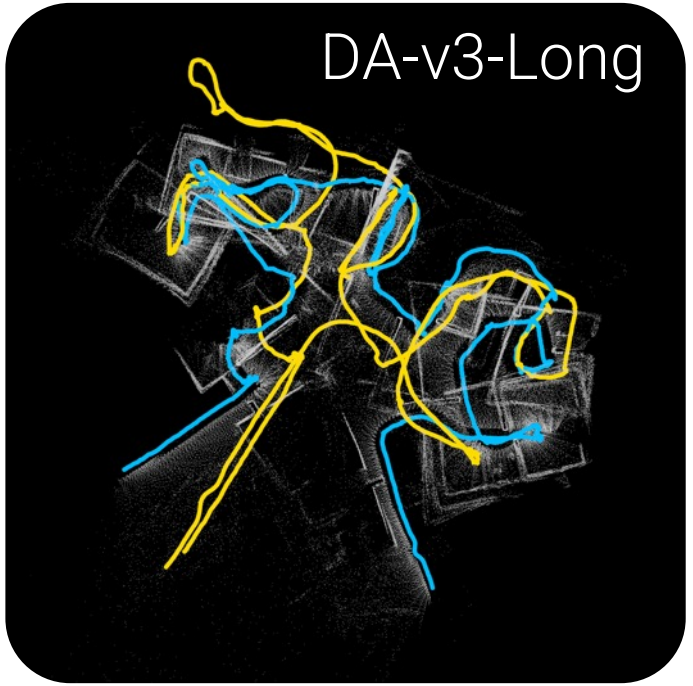

Lingbot-Map
GT

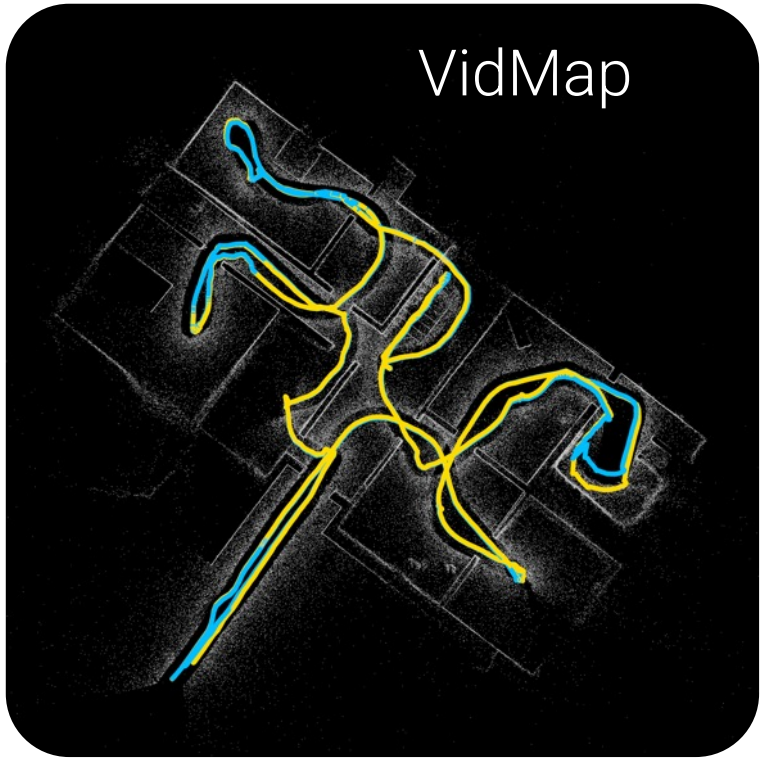

DA-v3-Long


VidMap



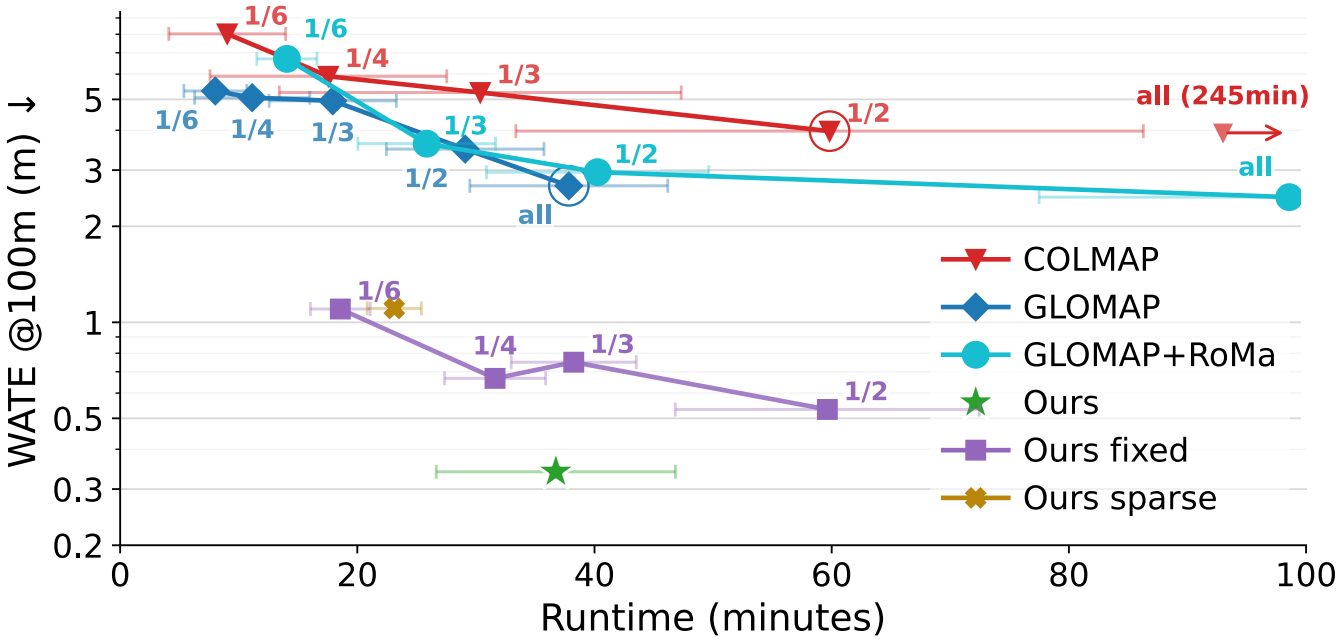


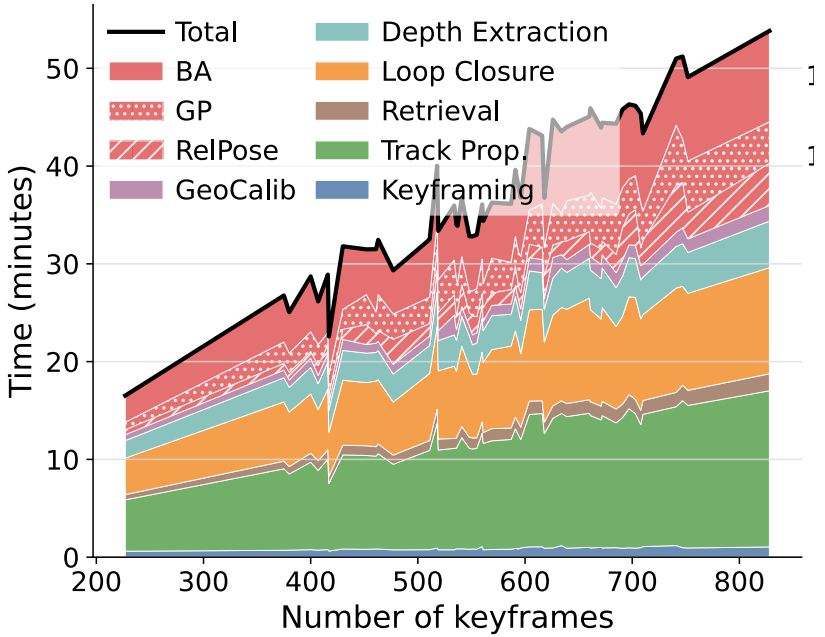


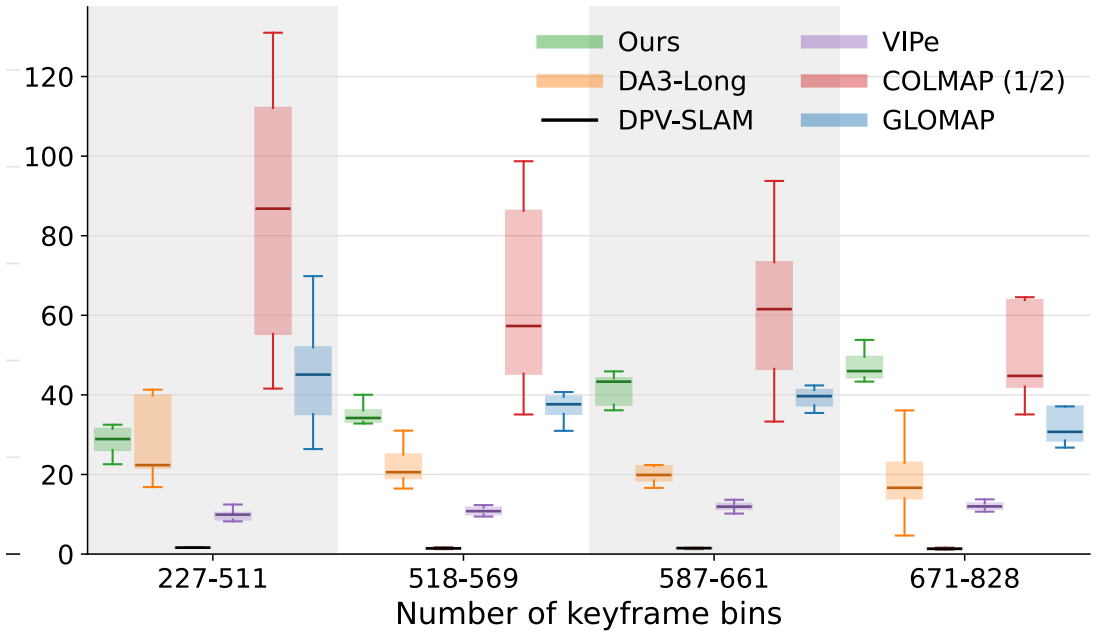


Generalizes well to OOD environments

Efficiency







- Adaptive keyframing helps
- Slower than SLAM,
  but embarrassingly parallel

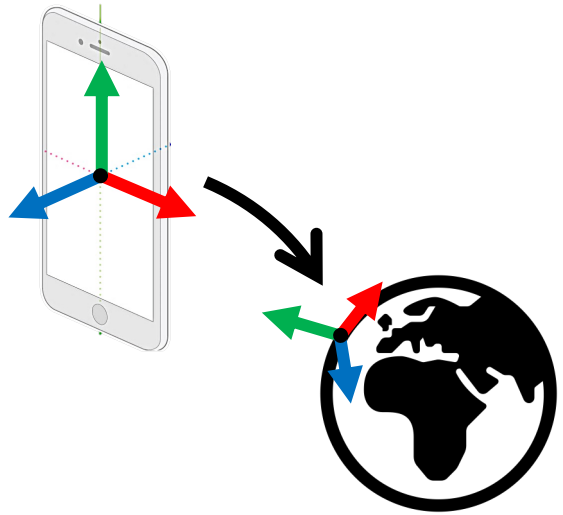
From videos to geolocalization

- Videos are posed in a local coordinate frames
- Next we want to anchor them in a global geo-coordinate system
  - To extract spatial information – crowd-sourced mapping
  - To co-optimize videos captured in the same location

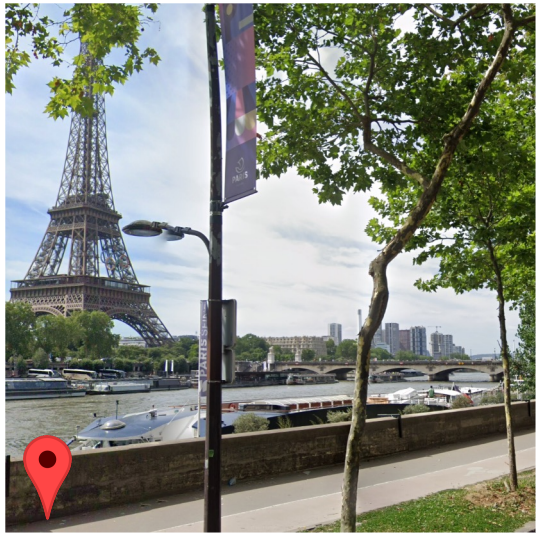
(39.743, -104.995)

(47.376, 8.547)

(52.371, 4.533)



Approach #1: retrieval of ground images *(place recognition)*



search by similarity

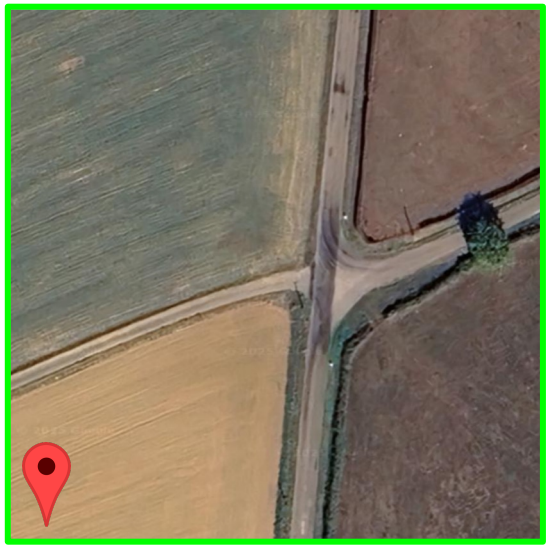
✅ Robust

✅ Simple to train

❌ Limited by coverage

❌ Expensive

Approach #2: retrieval of overhead images



search by similarity

✅ Dense coverage

✅ Scalable

❌ Large domain gap

❌ Requires negative mining

Approach #3: direct memorization

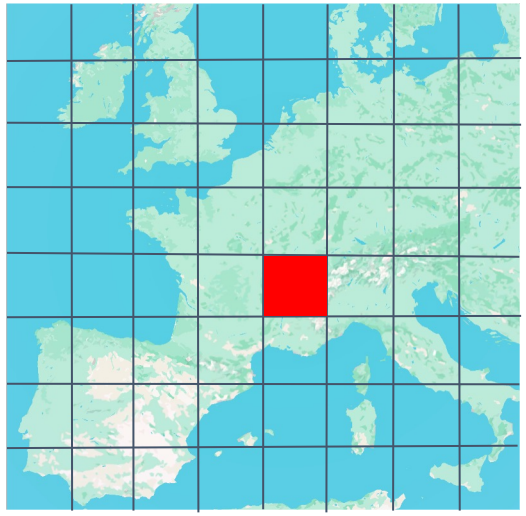
Regression

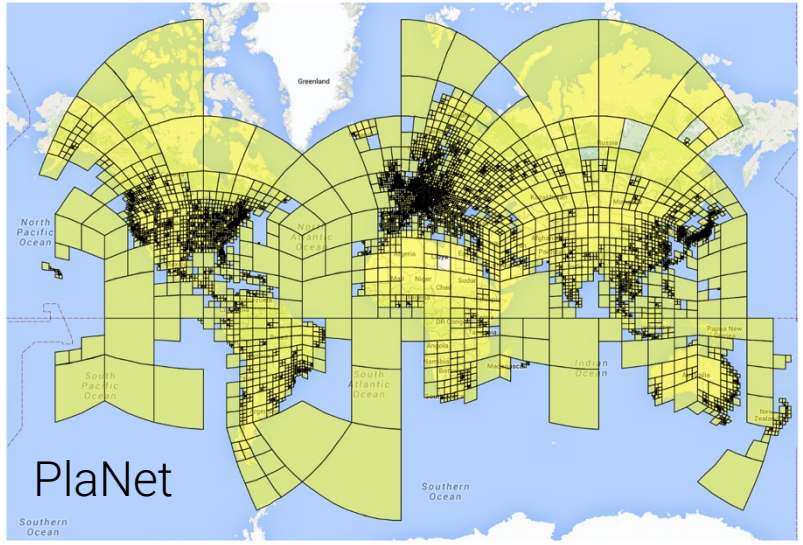
DNN

latitude
+ longitude

Classification

DNN





✅ Very scalable

❌ Very coarse

❌ Depends on data density

Summary

Ground retrieval

Overhead retrieval

Direct memorization

✅ Most accurate

❌ Not scalable

✅ Dense coverage

❌ Domain gap

✅ Simple to train

❌ Coarse

Scaling Image Geo-Localization to Continent Level


Philipp Lindenberger


Paul-Edouard Sarlin


Jan Hosang


Matteo Balice


Marc Pollefeys

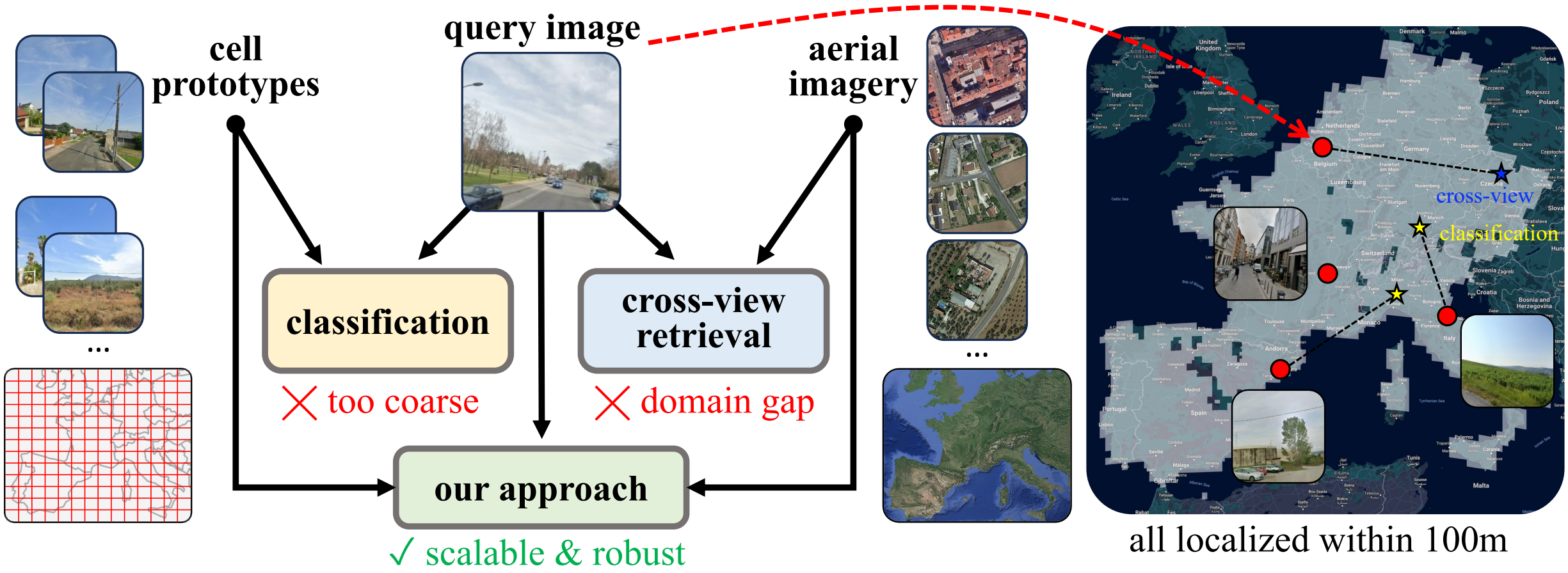

Simon Lynen


Eduard Trulls

NeurIPS 2025

scaling-geoloc.github.io

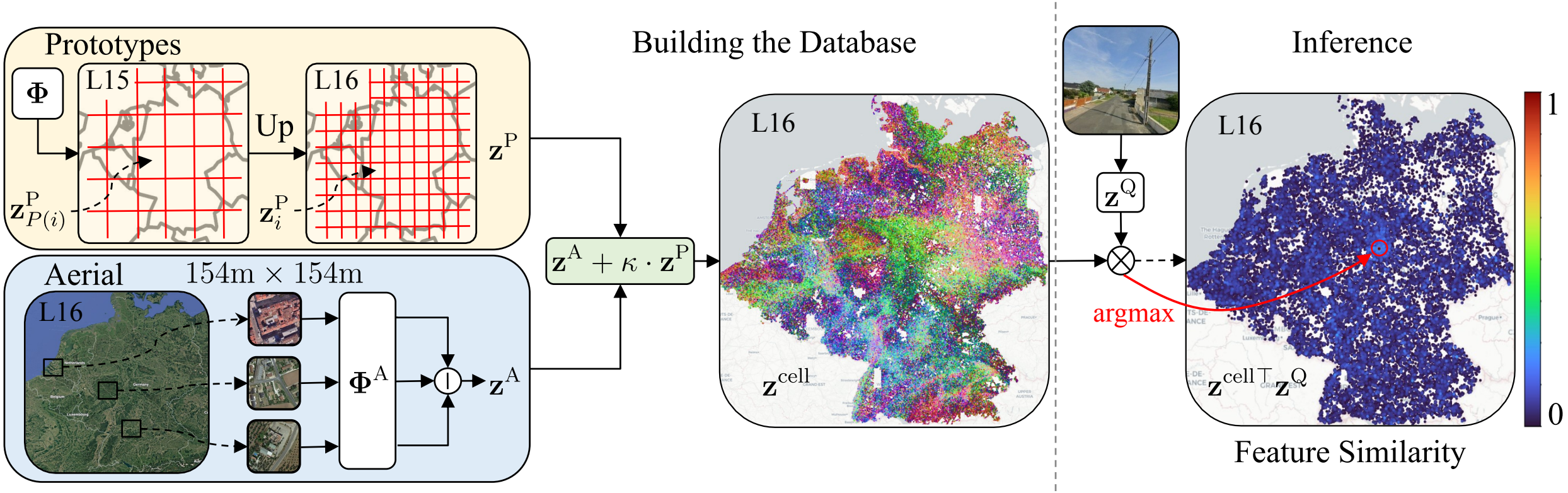
A hybrid approach



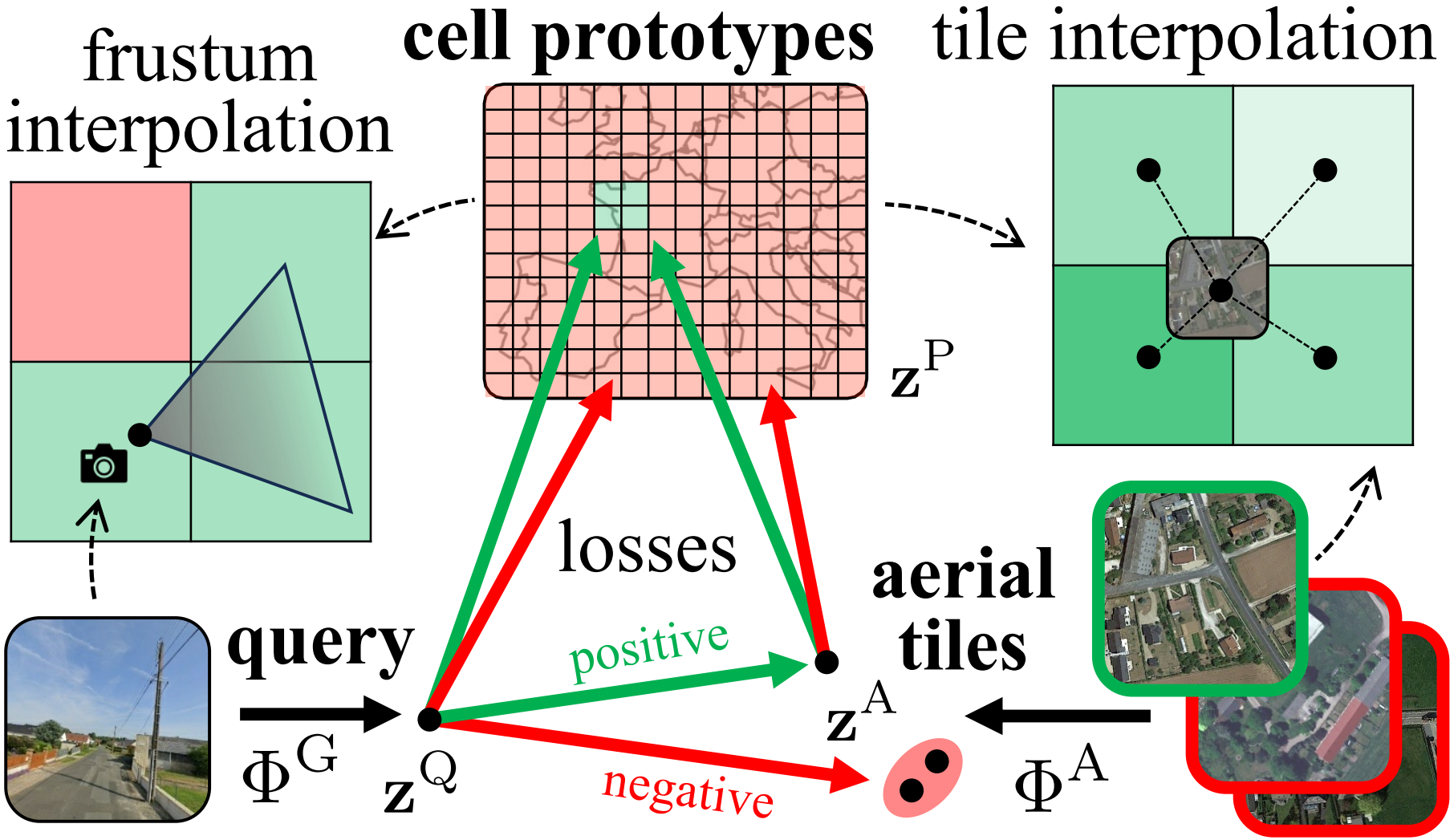
68% recall at 200m on Western Europe – 7M cells (150x150m each)

Qualitative results

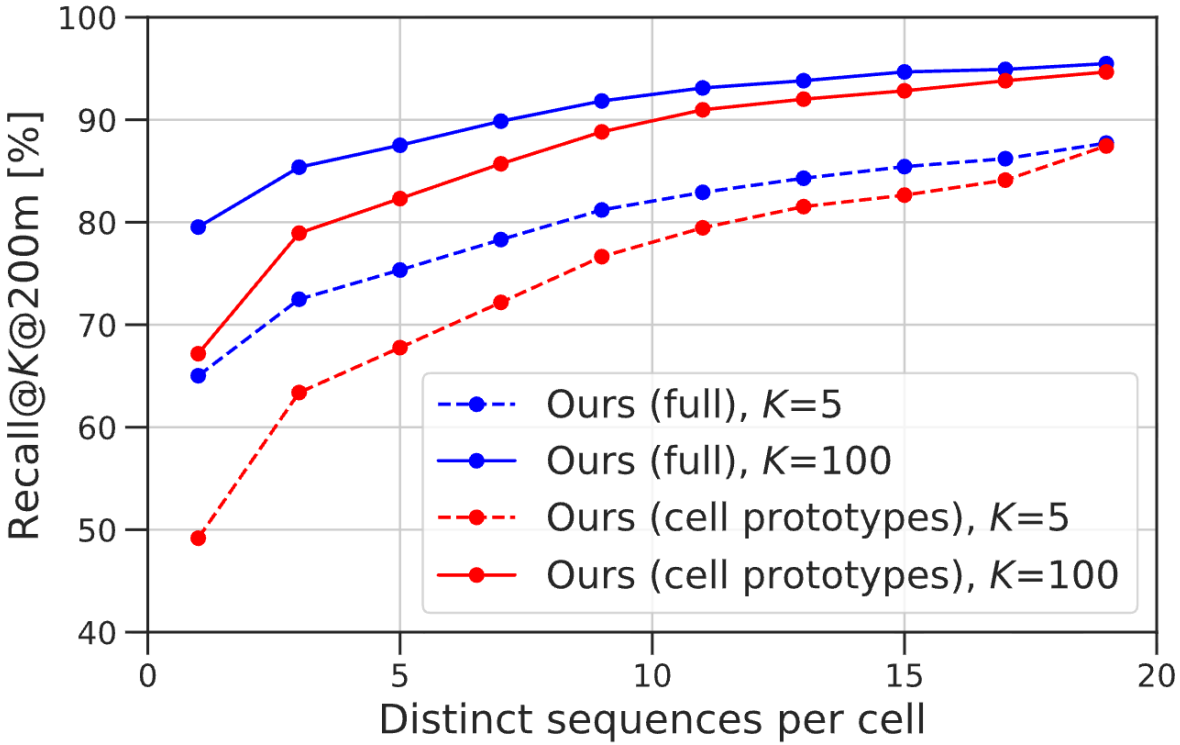
Inference algorithm

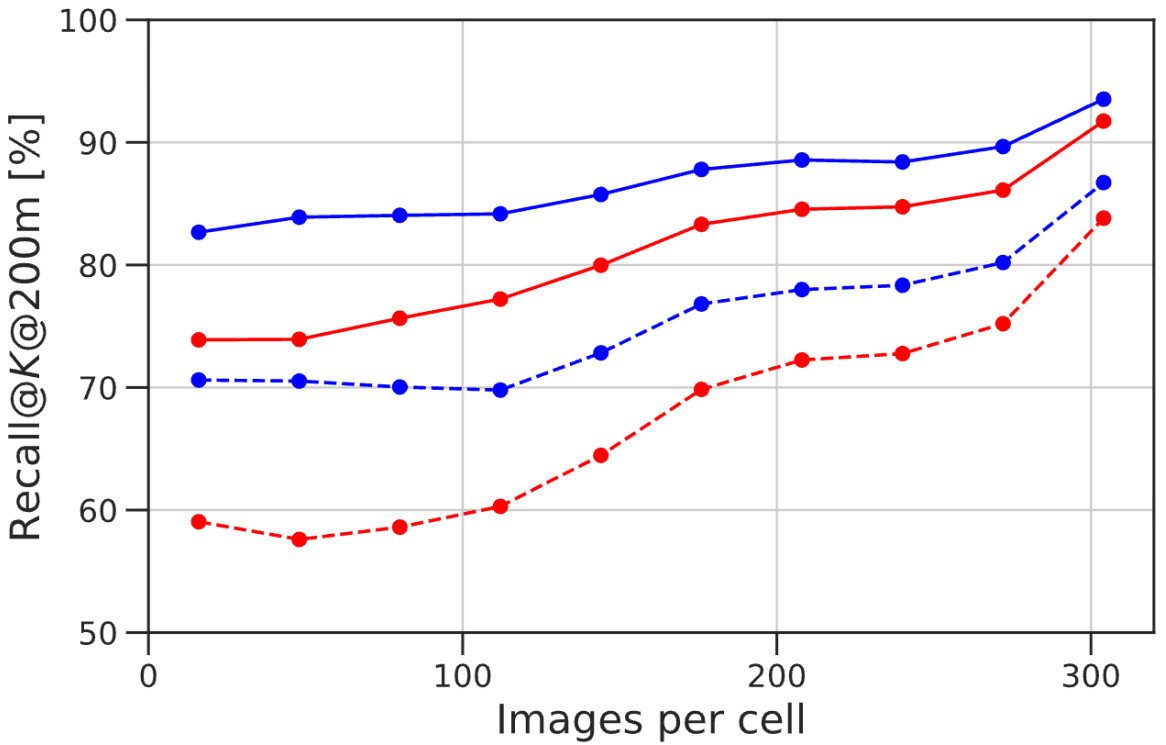


Supervision

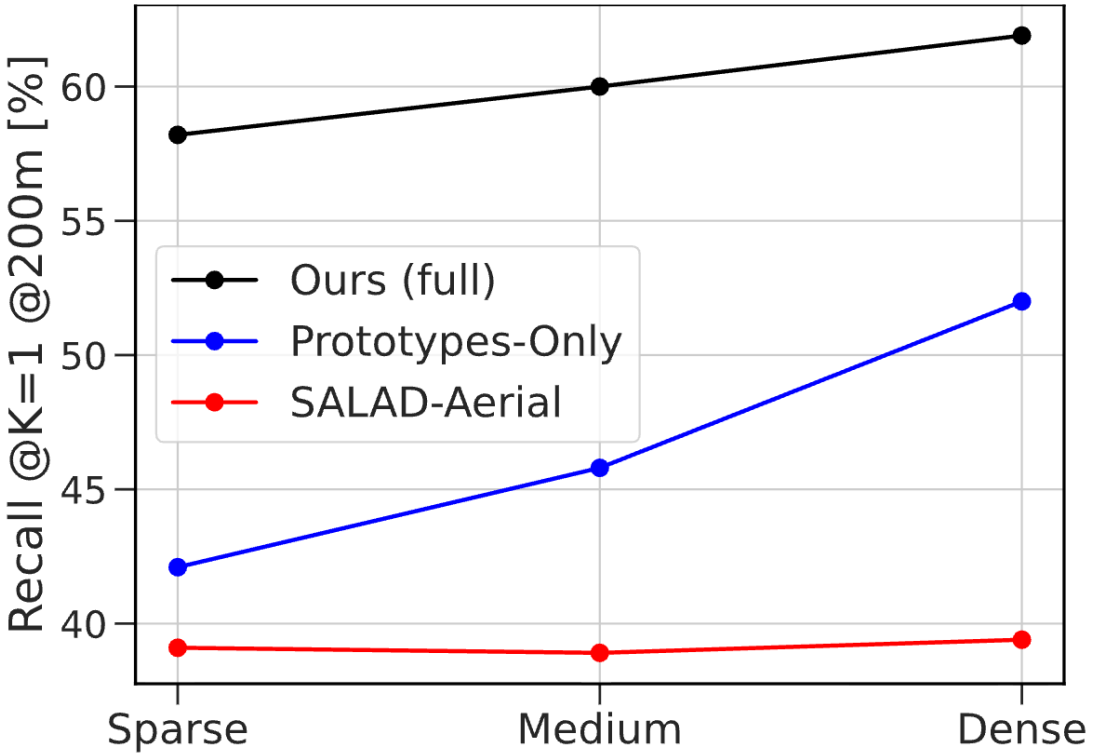


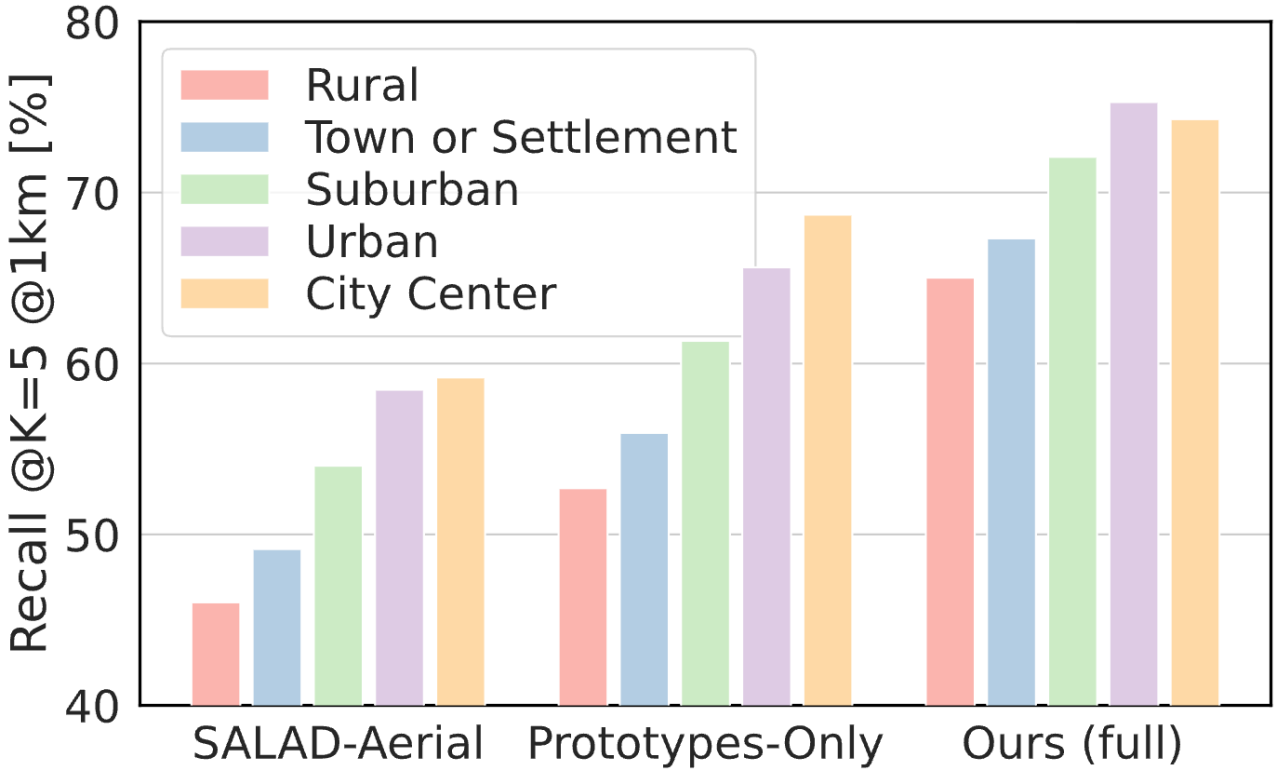
Overhead retrieval improves robustness to low data density



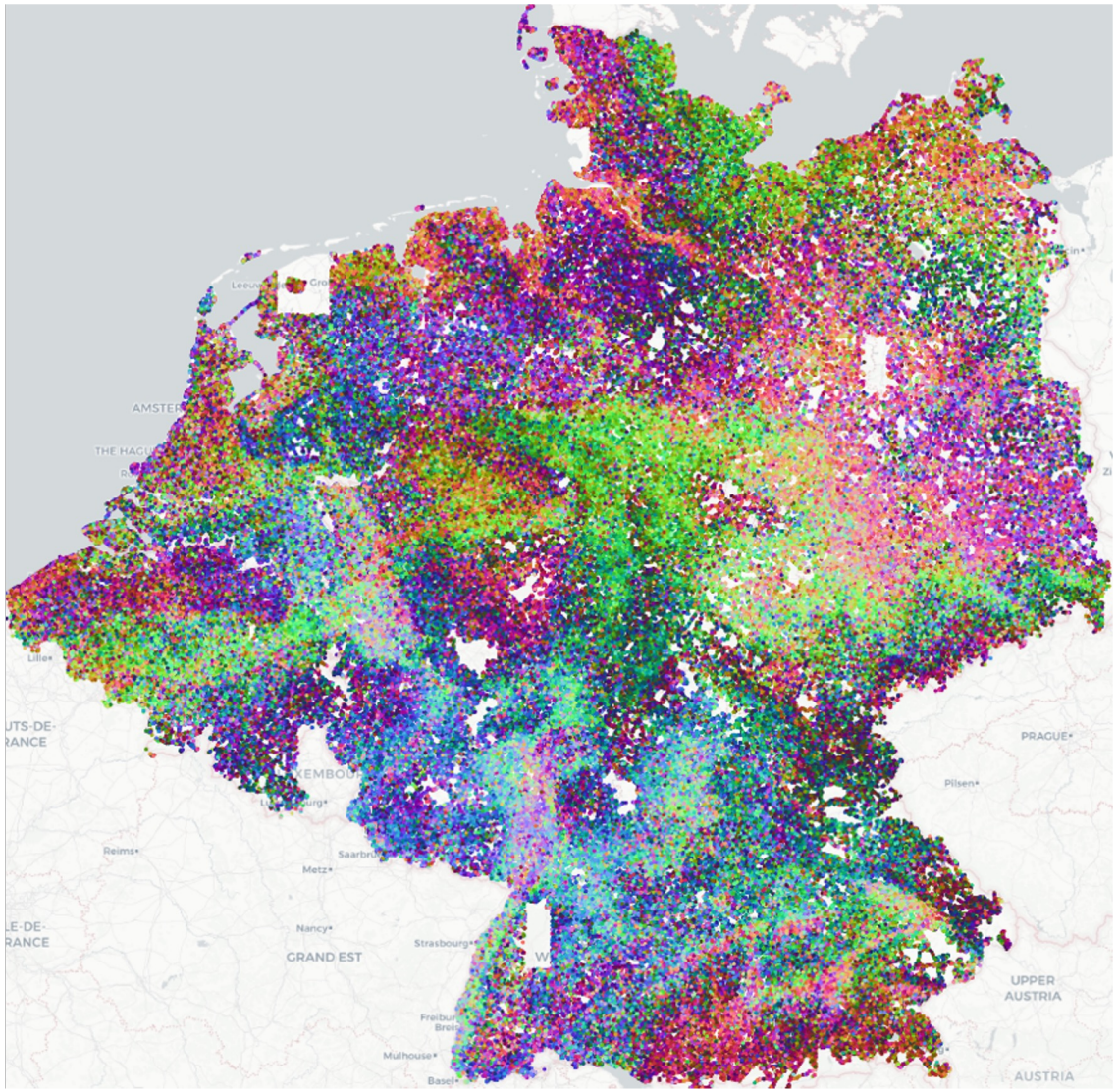


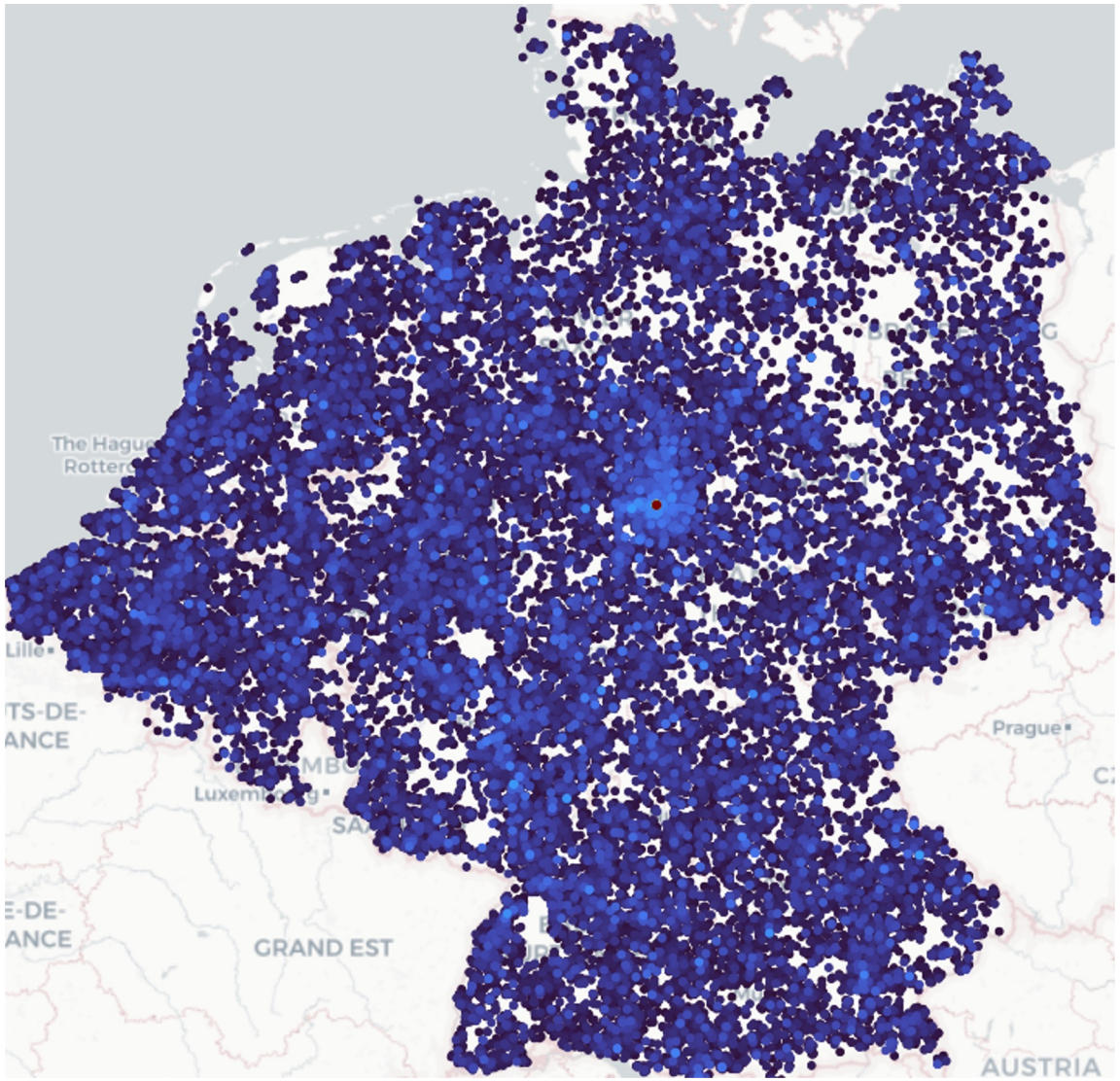
Increased accuracy in rural areas



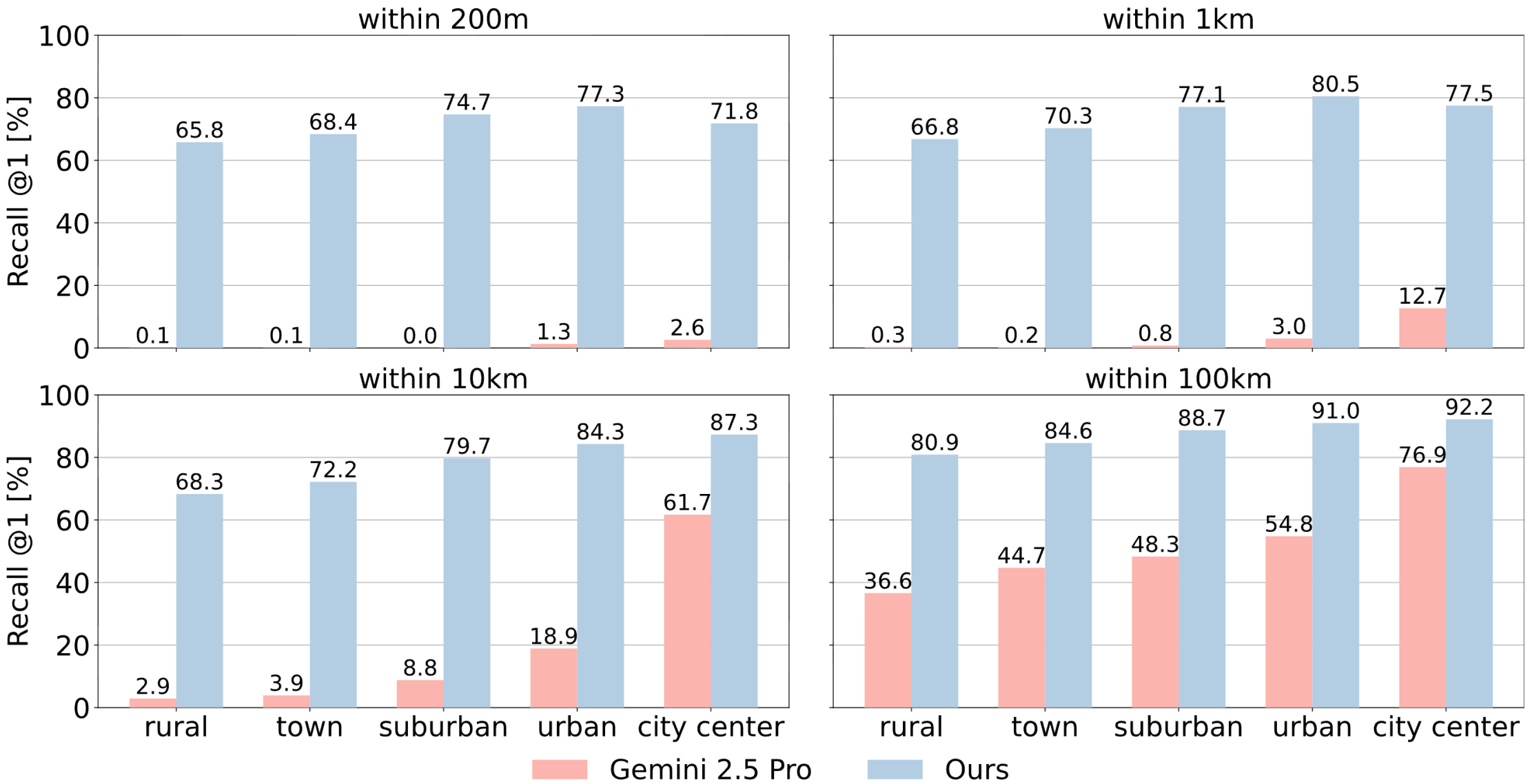


Semantics emerge from spatial supervision

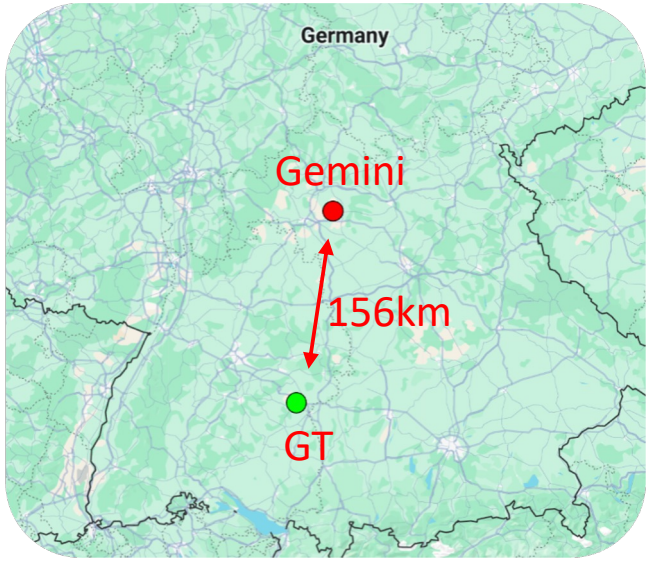




Comparison to VLMs



Example



Conclusions

Video reconstruction

- Yes, e2e models are **impressive**! But they don't yet dominate
- Classical approaches with learned priors are still competitive
- How do we make it easier for e2e models to use additional information like IMUs and calibrated multi-cameras?

Geolocalization

- How do we efficiently scale to the world?
- What other modalities can we leverage?

Thank you!

psarlin.com