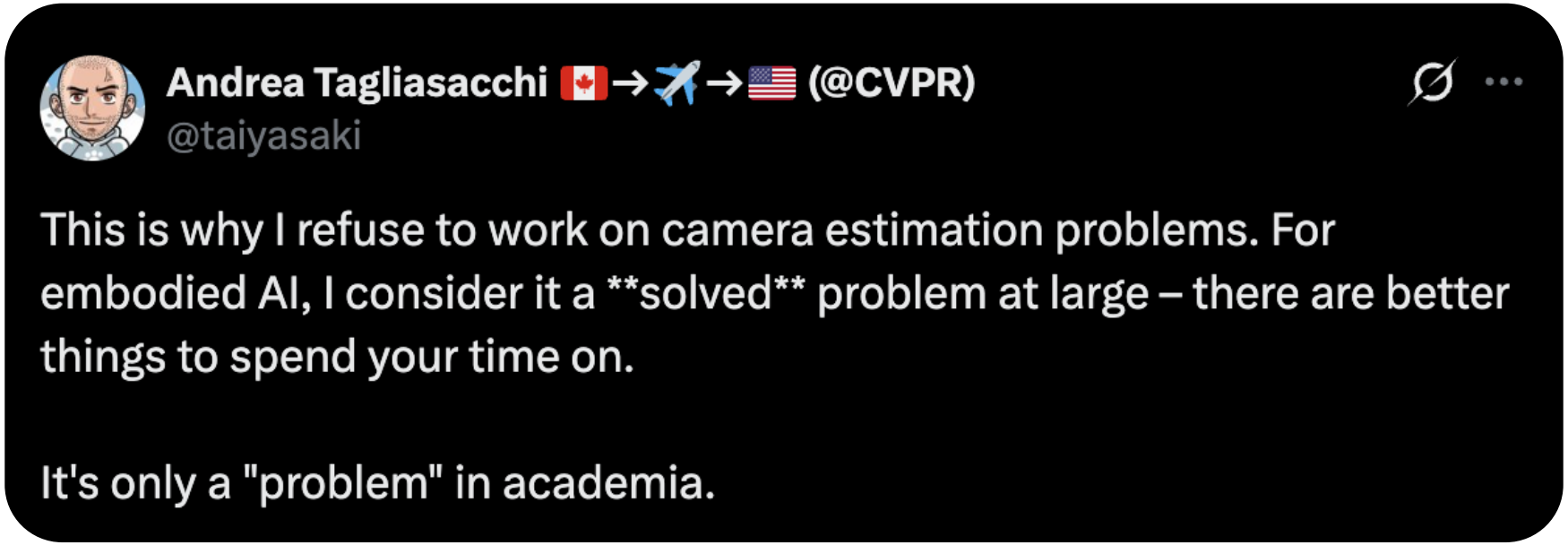
Do E2E Reconstruction Models Need More Geometry?

Paul-Edouard Sarlin

Workshop on End-to-End 3D Learning, CVPR 2026
2026-06-03

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

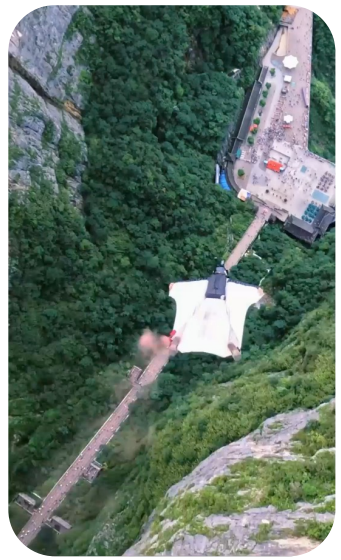
This covers much of:

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

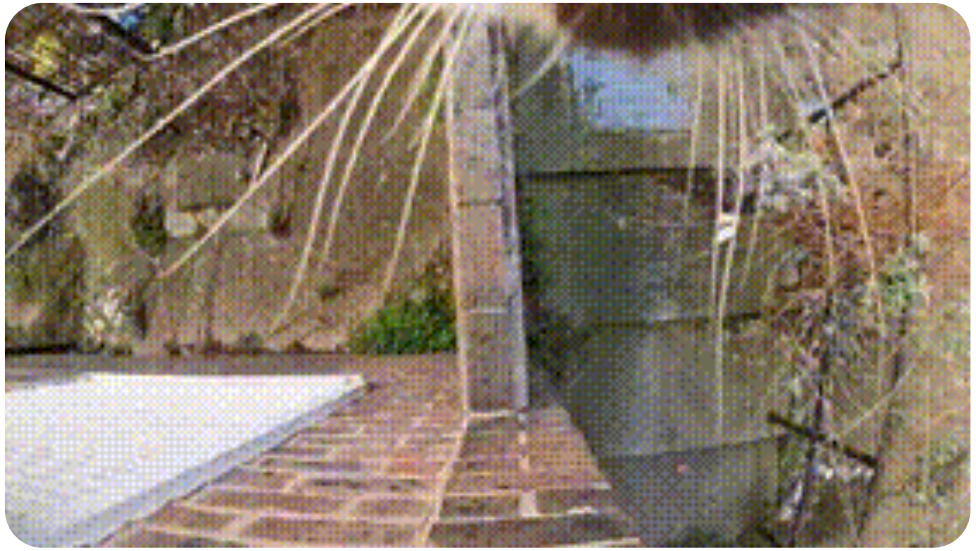
is camera pose estimation already solved?

Not for monocular videos:

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



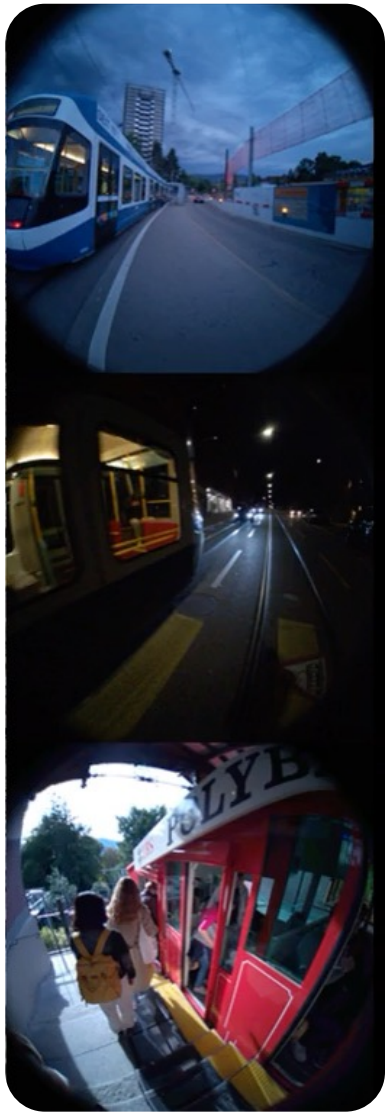
Not if we refer to mapping:
3D for every single pixel



Example: egocentric perception

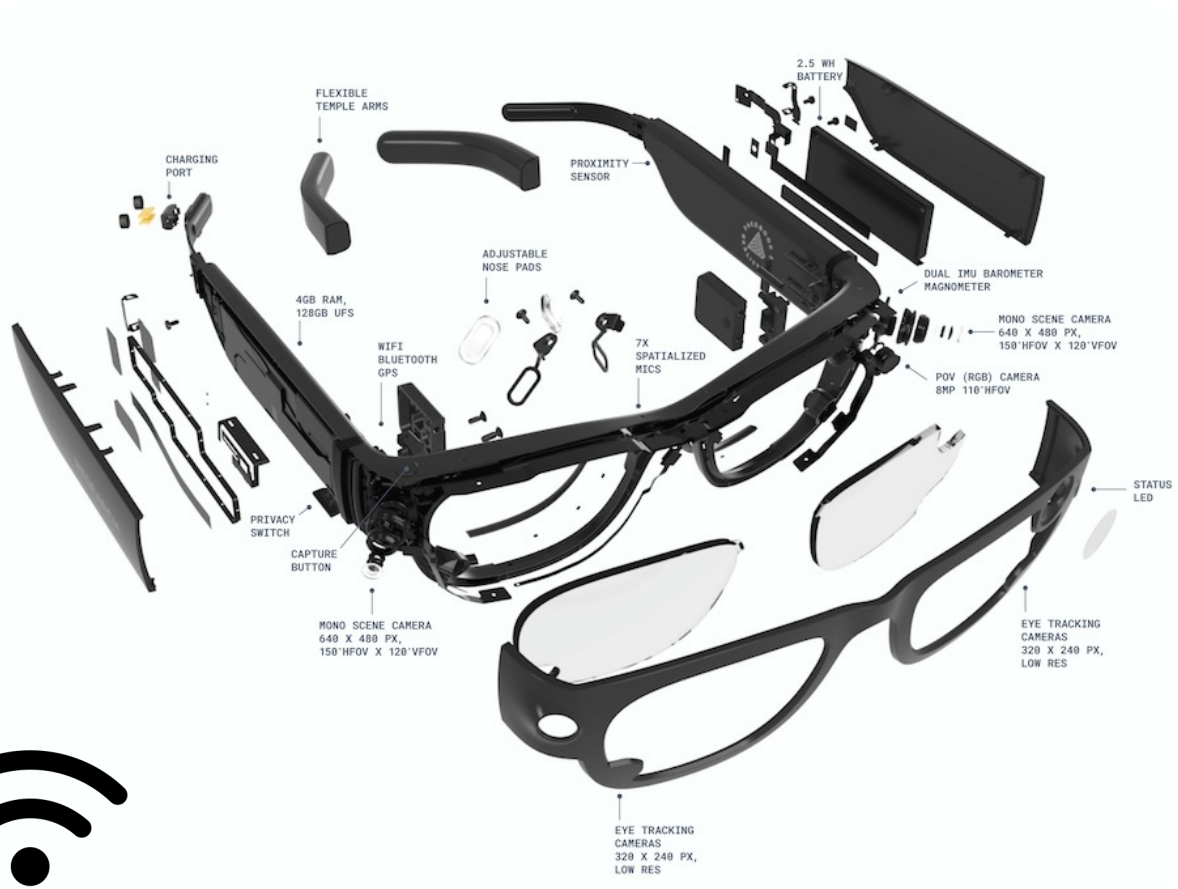
Opportunistic capture, byproduct of the user's motion

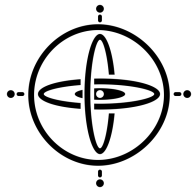
Challenges: motion, constraints, calibration



This isn’t a monocular problem!

- Many cameras
- Non-image sensors
- High-frequency (>60 FPS)
- Dense coverage (crowd-sourced)





inertial units

GPS

BT/WiFi

microphone array

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

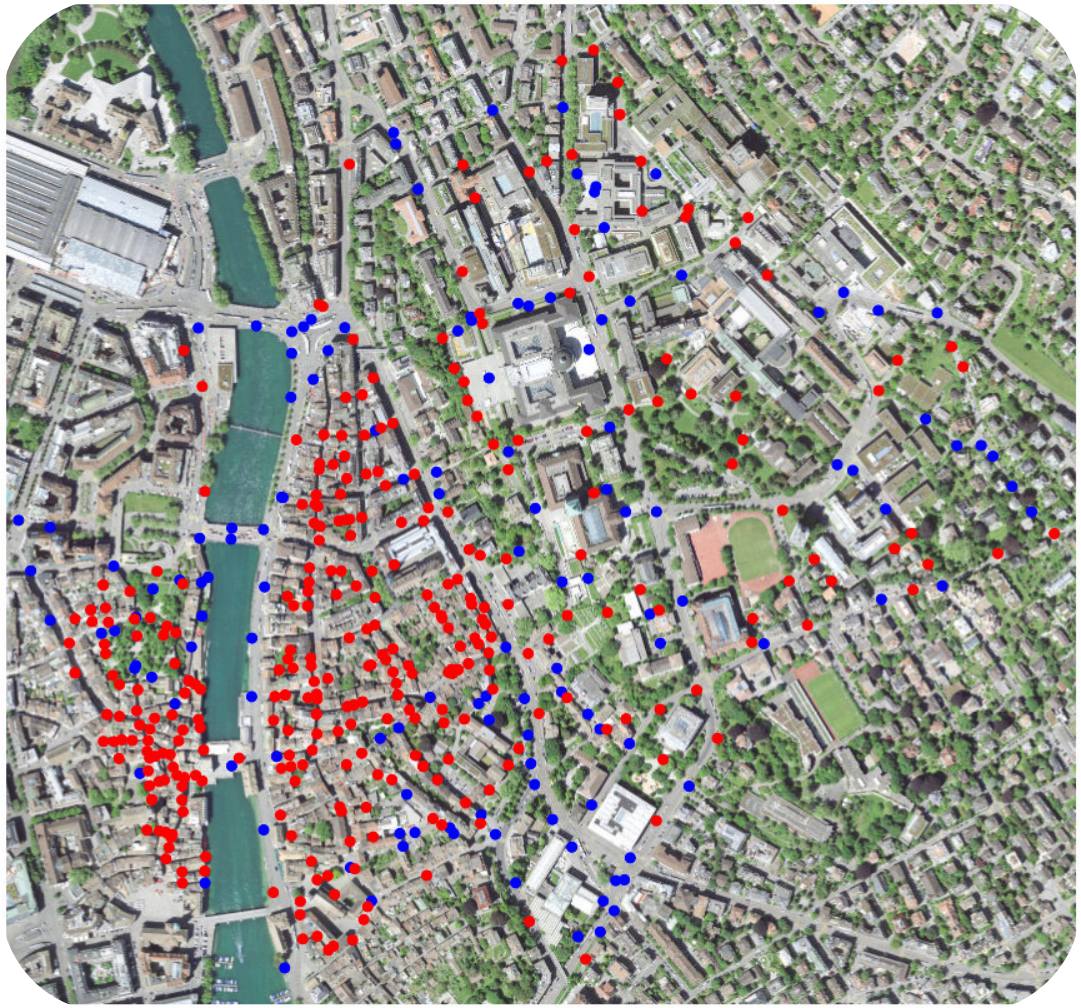
lamaria.ethz.ch

The LaMAria dataset

70km
22 hours
10-45min each

The LaMAria dataset

Control points with cm accuracy
Measure 0.1% metric scale drift



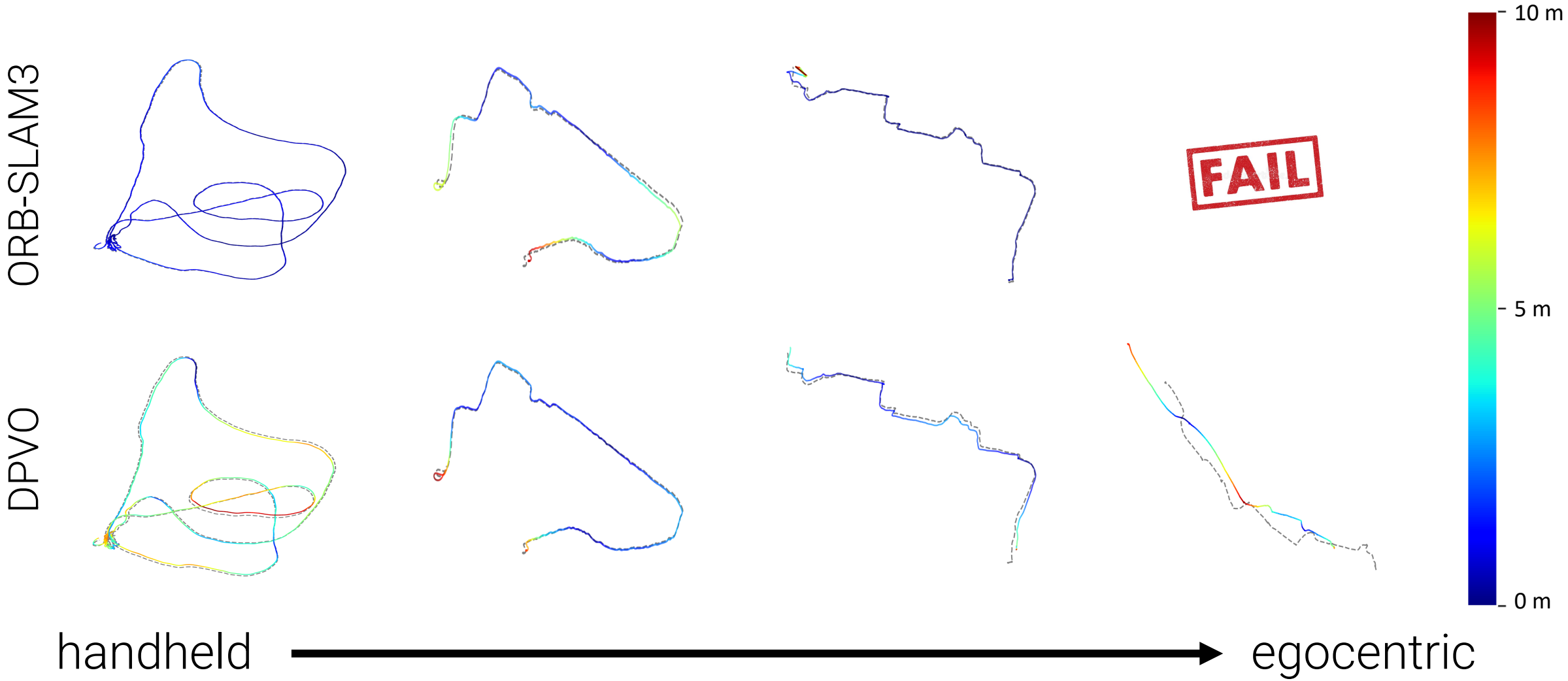
Limitations of existing benchmarks

dataset	data			sensors			ground-truth		challenges		
	motion	environment	multi-seq	multi-cam	IMU	others	source	accuracy	duration	dynamics	lighting
EuRoC [10]	drone	small	no	yes	yes	no	mocap	∼cm	<3 min	no	partial
TartanAir [67]	random	large	no	yes	no	depth,LiDAR	synthetic	perfect	→20 min	yes	yes
4Seasons [68]	car	large	yes	yes	yes	GNSS	VI+GNSS	>dm	350 km	moderate	yes
VBR [8]	car,handheld	large	partial	yes	yes	LiDAR, GNSS	LiDAR+IMU +RTK-GNSS	∼cm	→50 min	moderate	partial
TUM-RGBD [61]	handheld	small	no	no	no	depth	mocap	<cm	<3 min	no	no
TUM-VI [60]	handheld	medium	no	yes	yes	no	mocap	<cm	→25 min	no	no
ADVIO [53]	handheld	medium	yes	no	yes	no	VI-SLAM	∼dm	∼3 min	moderate	no
ETH3D-SLAM [59]	handheld	small	no	yes	yes	depth	mocap	<cm	<4 min	moderate	yes
NewerCollege [52, 70]	handheld	medium	yes	yes	yes	LiDAR	LiDAR-SLAM	∼cm	→26 min	no	no
Hilti-Oxford [71]	handheld	medium	yes	yes	yes	LiDAR	surveying	<cm	→17 min	no	partial
Hilti-UZH [45]	robot,handheld	medium	yes	yes	yes	LiDAR	surveying	<cm	→12 min	no	partial
LaMAR [56]	head-mounted handheld	medium	yes	yes	uncalibrated	GNSS, WiFi, BT	V-SLAM +LiDAR	∼dm	∼5 min	moderate	yes
LaMAria (ours)	head-mounted handheld	large	yes	yes	yes (×2)	GNSS, WiFi, BT	surveying	∼cm	→45 min	people,tram funicular	yes

Next level of scale and accuracy of GT poses

Monocular SLAM works only on easier data

Monocular SLAM works only on easier data



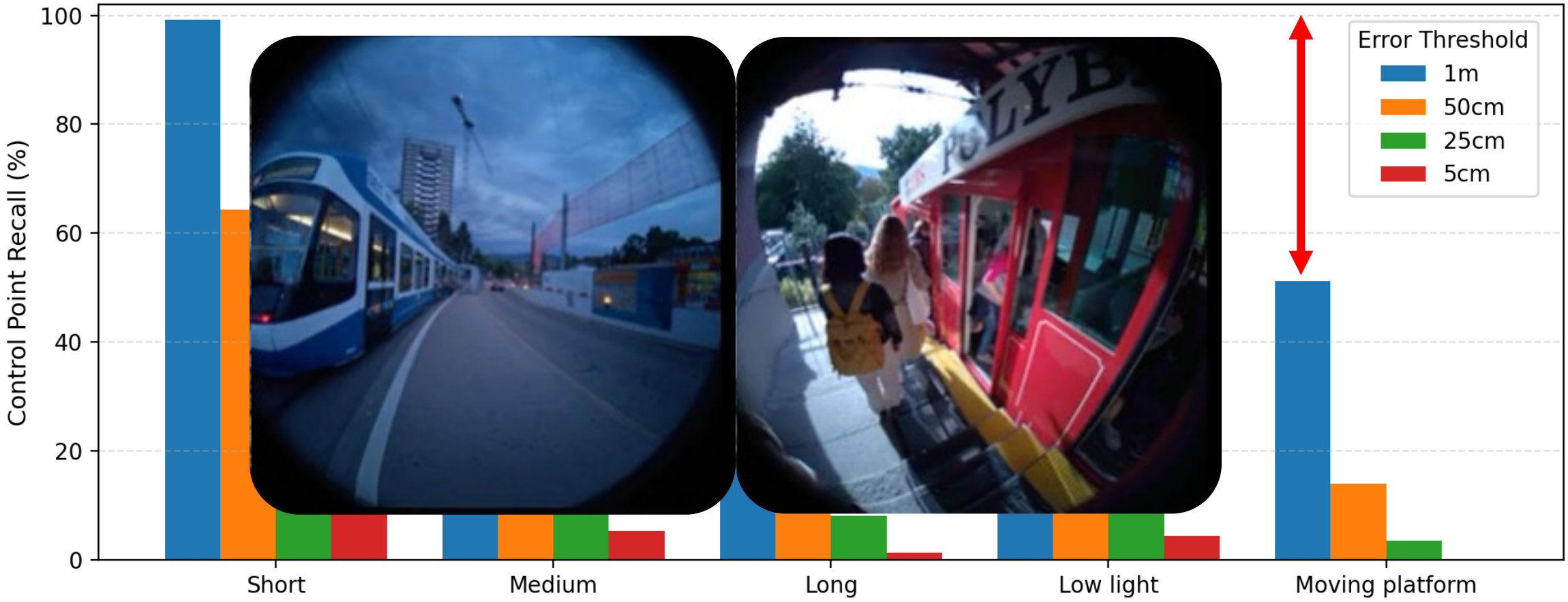
Open-source systems are far behind industry

🟠 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

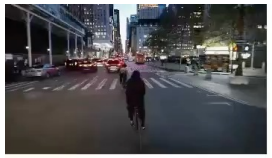
E2E models for mono SLAM

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-VIO

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

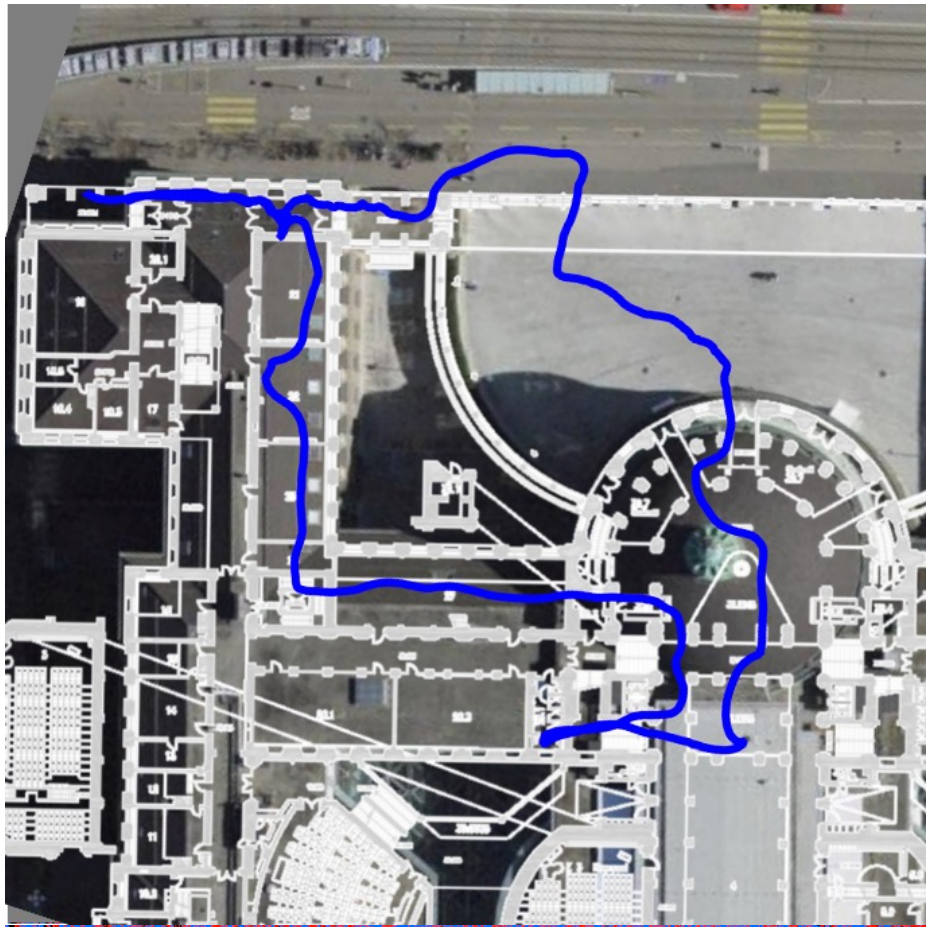
Are we comparing against the right classical baselines?

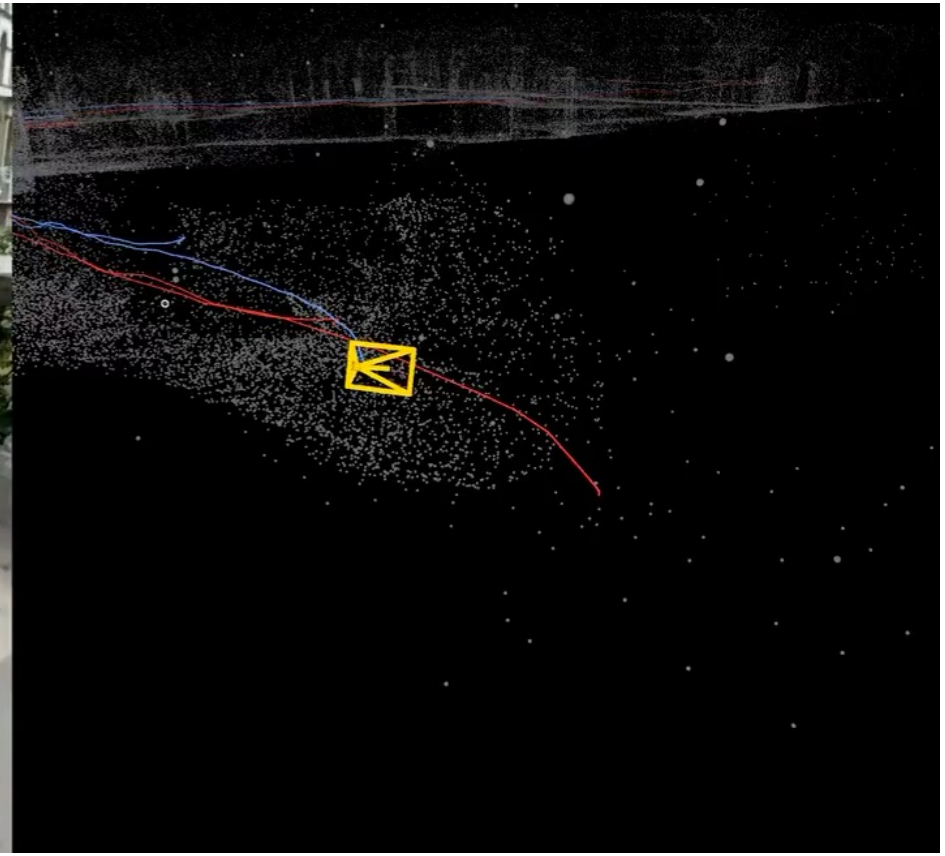
SLAM
for sequences
designed for real-time
Causal, limited robustness
Fast and efficient
Limited self-calibration

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





work done by Zador Pataki

How it works

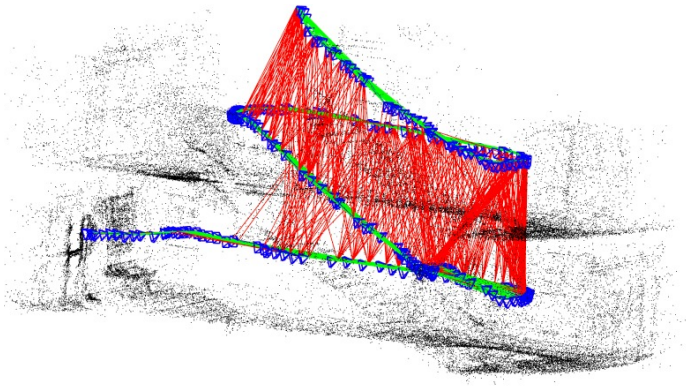
	SLAM	SfM	VidMap
Global, acausal	❌	✅ GLOMAP	✅
Adaptive keyframing	✅	❌	✅
Temporal constraints	✅	❌	✅
Offline, maximize robustness	❌	✅	✅

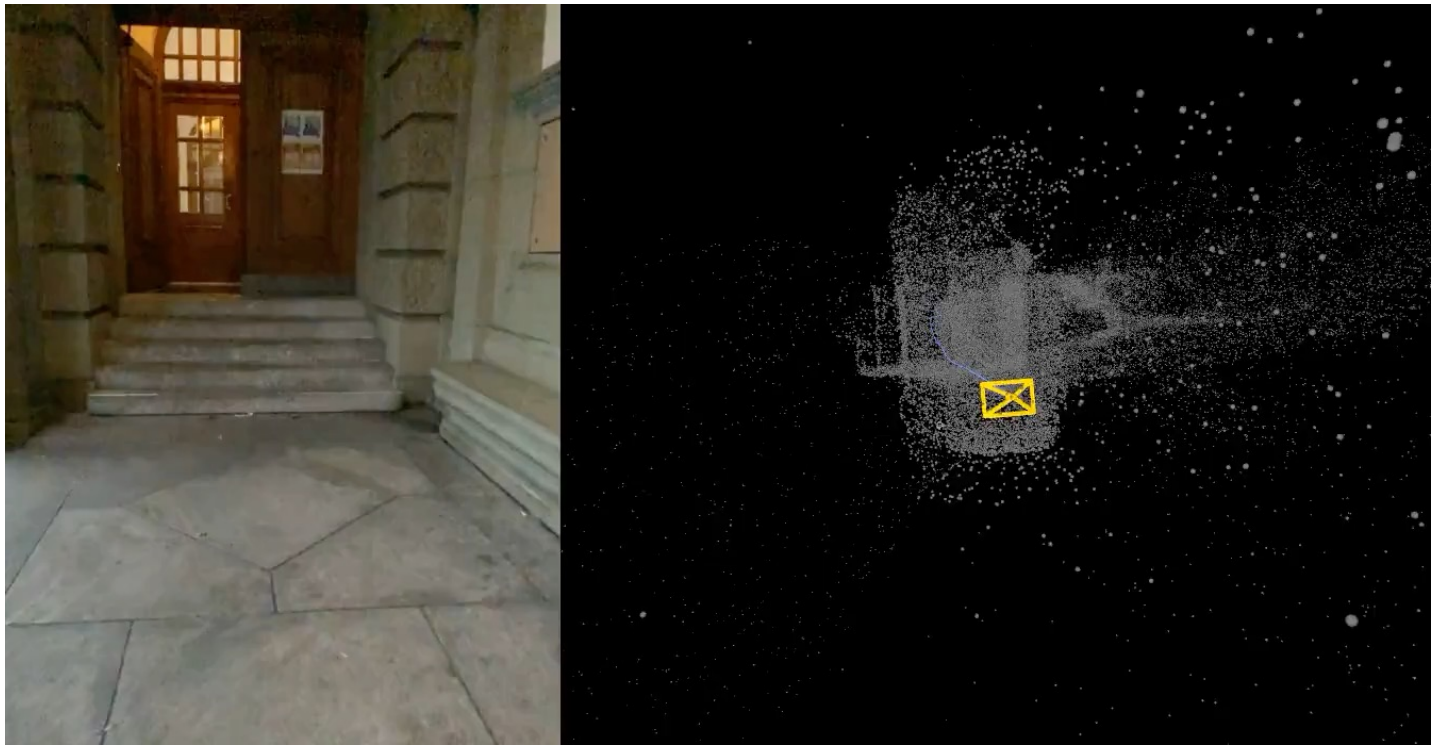
Take the best of SfM and SLAM by augmenting global SfM with:

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

Addressing failures of classical approaches

Pure rotation, lack of texture, scale drift, symmetries…

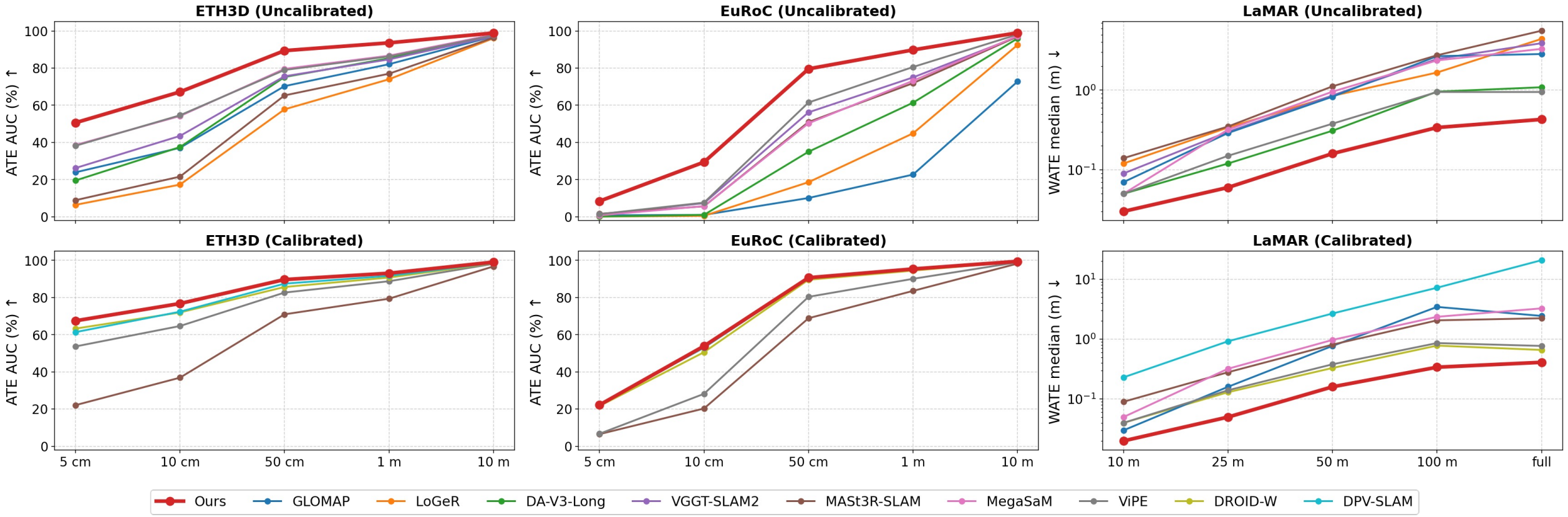




Results on hard benchmarks

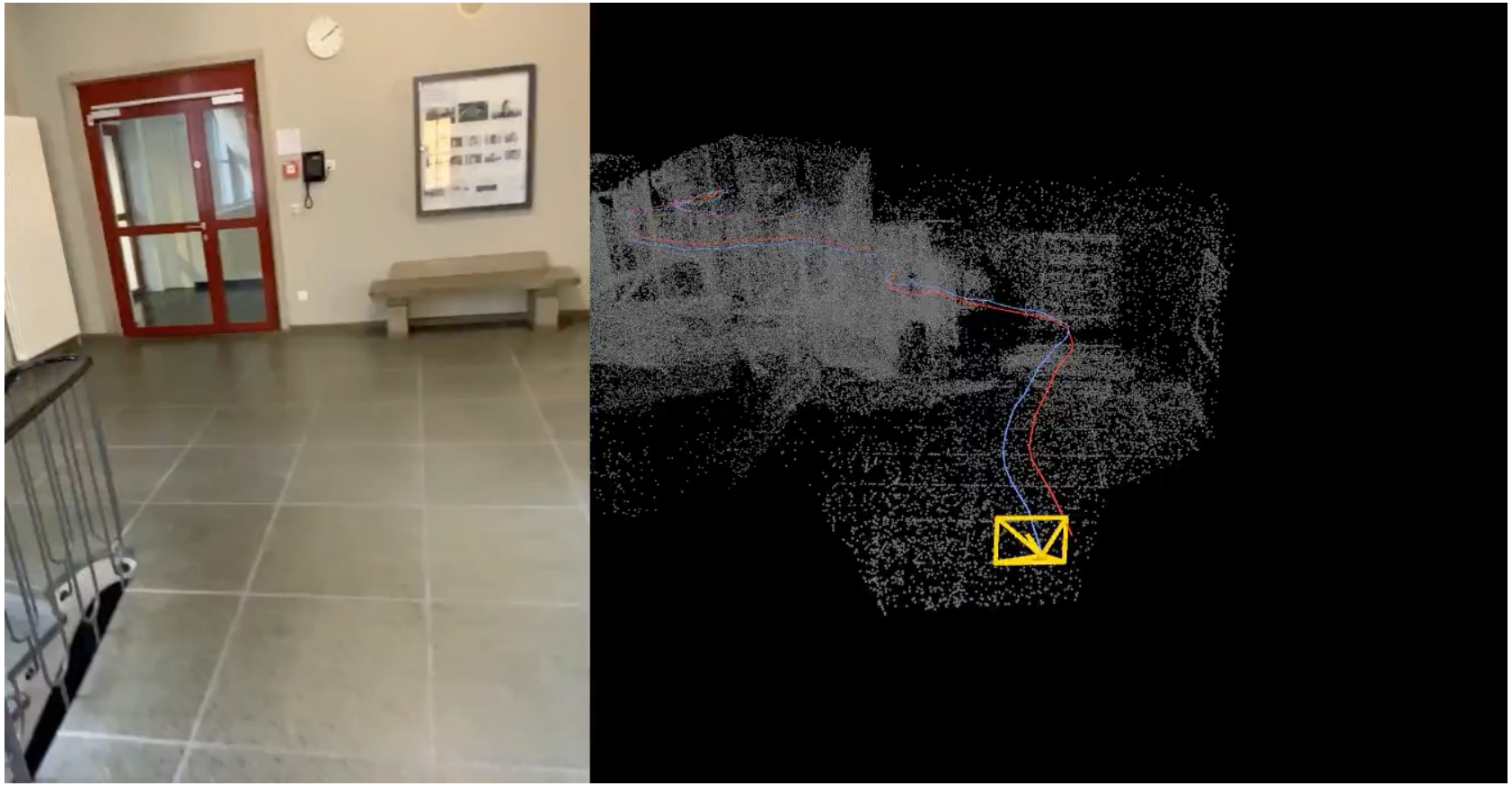
AUC (higher=better) on ETH3D & EuRoC

ATE (lower=better) on phone videos



Outperforms E2E models in robustness, accuracy, scalability

To learn more,
join the Image Matching Workshop



Tomorrow 13:10-13:55
image-matching-workshop.github.io

SfM/SLAM
in 2019

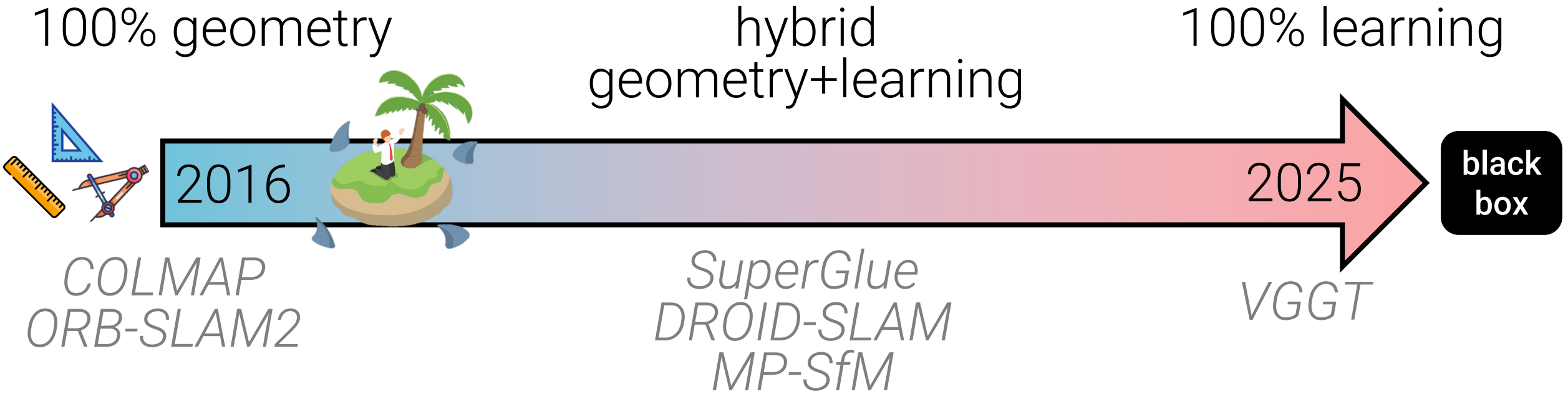
Deep
Learning

Deep
Learning

Deep
Learning

More Deep
Learning

Full swing towards E2E models



What did we gain?

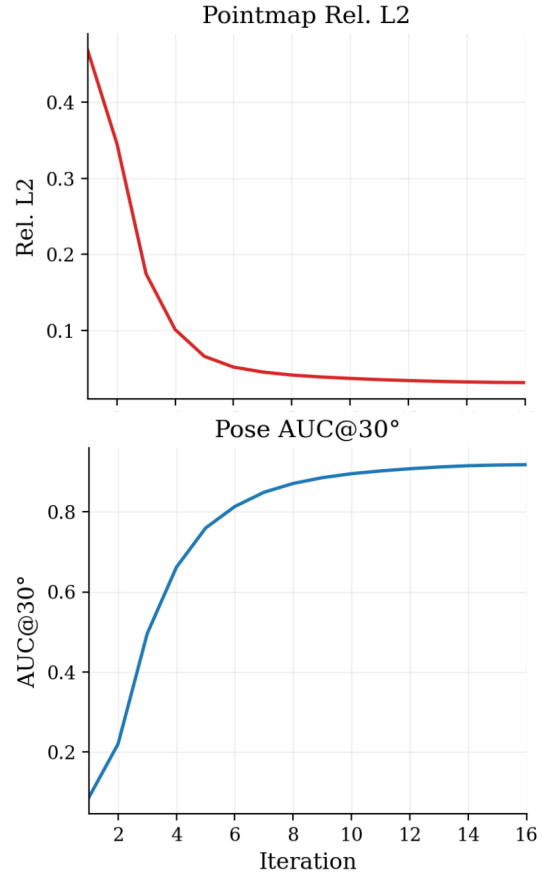
- Robustness – more powerful priors
- Simplicity – reduced engineering, “only” a Transformer
- Domain adaptation – no tuning, “only” data

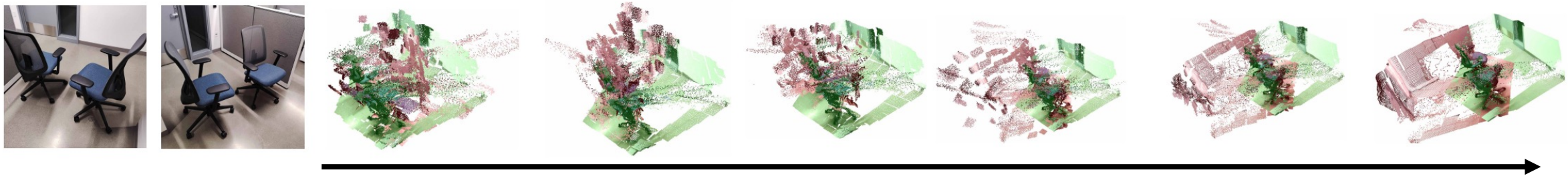
What happened to optimization?

Differentiable optimization did not work as expected…

But e2e models are solving a global optimization:

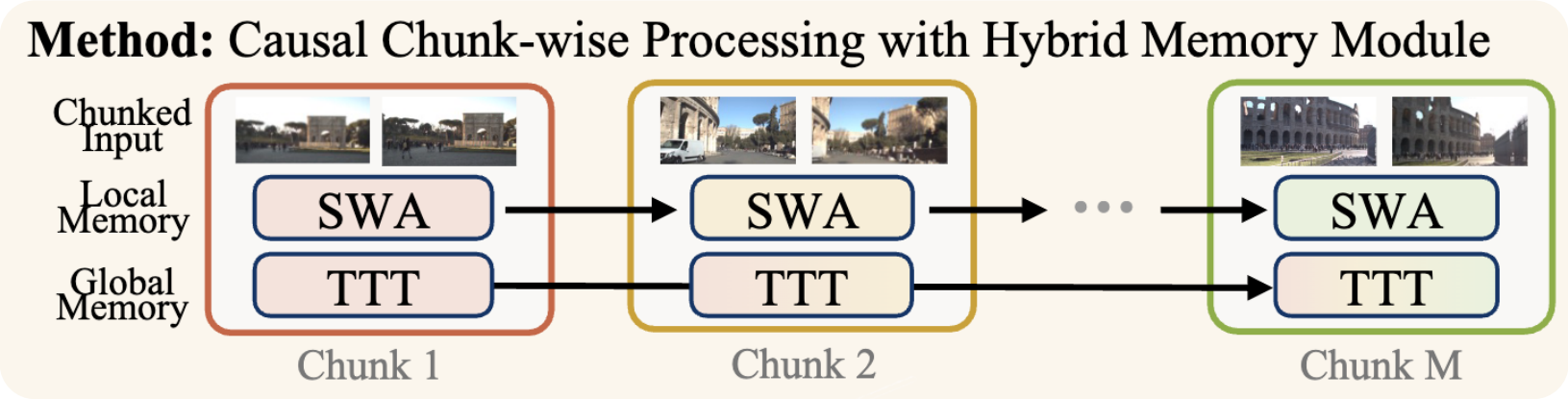
- Convergence behavior [Starý et al]
- Weight sharing [DejaView]
- Conditioning [Pow3r, MapAnything]



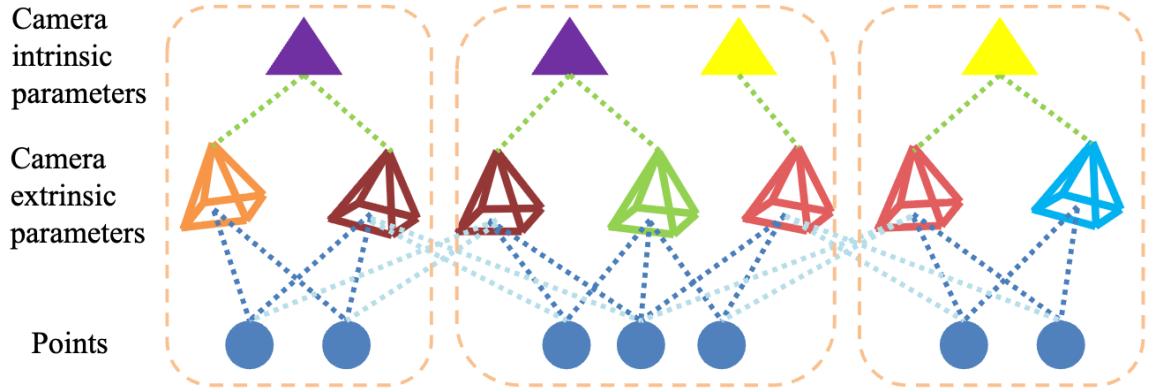


What happened to optimization?

Test-Time Training (TTT) of subsequences [LoGeR, TTT3R, ZipMap]



looks like distributed optimization with ADMM [Zhang et al].



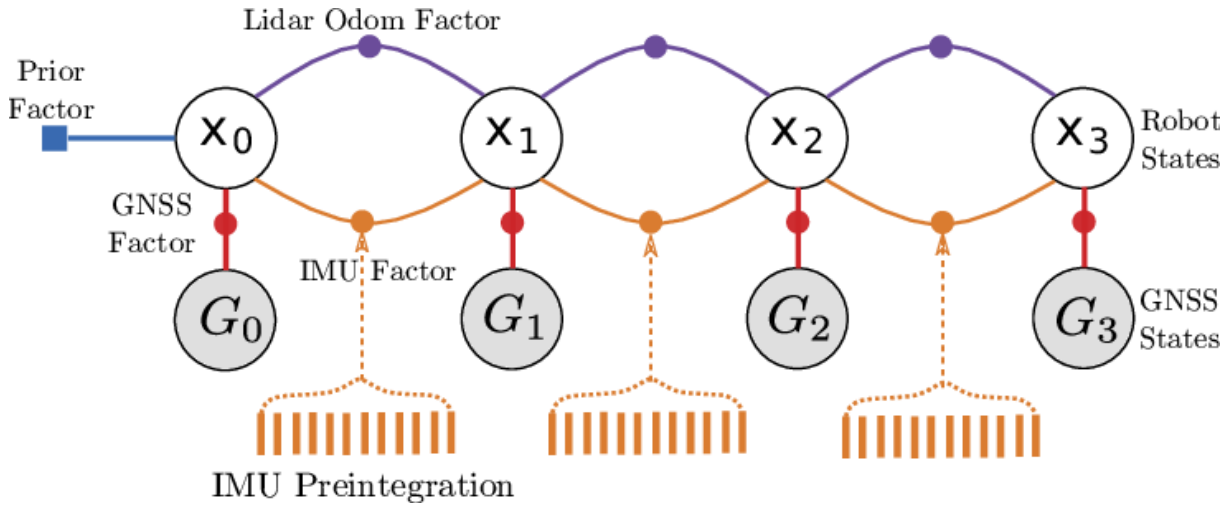
But what did we lose?

Missing benefits of optimization

- Precision: 3DGS requires refinement by classical BA

Method	Test-time Opt.	AUC@3°	AUC@5°	AUC@10°	Runtime
VGGT (ours)	✗	39.23	52.74	71.26	**0.2s**
VGGT + BA (ours)	✓	**66.37**	**75.16**	**84.91**	1.8s

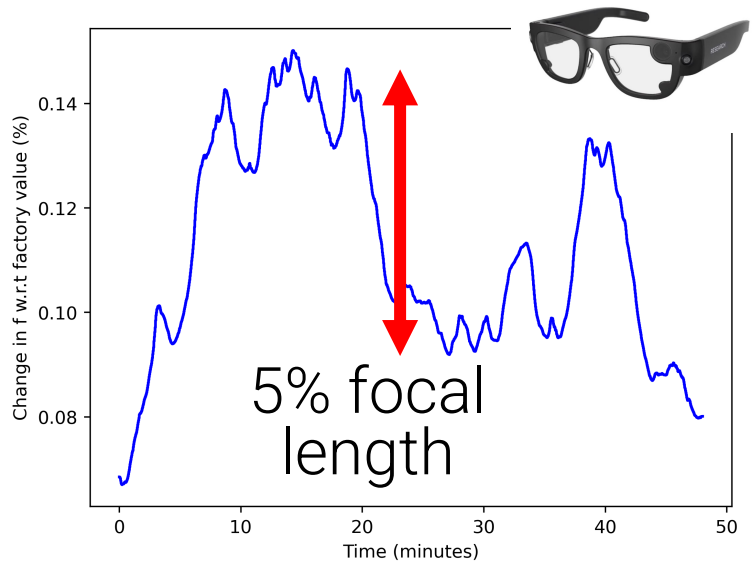
- Multi-sensor fusion (aka factor graphs): IMUs, GNSS, etc.



Missing benefits of optimization

Flexible modeling:

- Lens distortion (non-pinhole)
- Sharing intrinsics across views
- Rig constraints
- Physics (temperature)
- Pose priors in different coordinate frames



Missing benefits of optimization

Uncertainties! Which points/poses are (un)correlated along each axis?

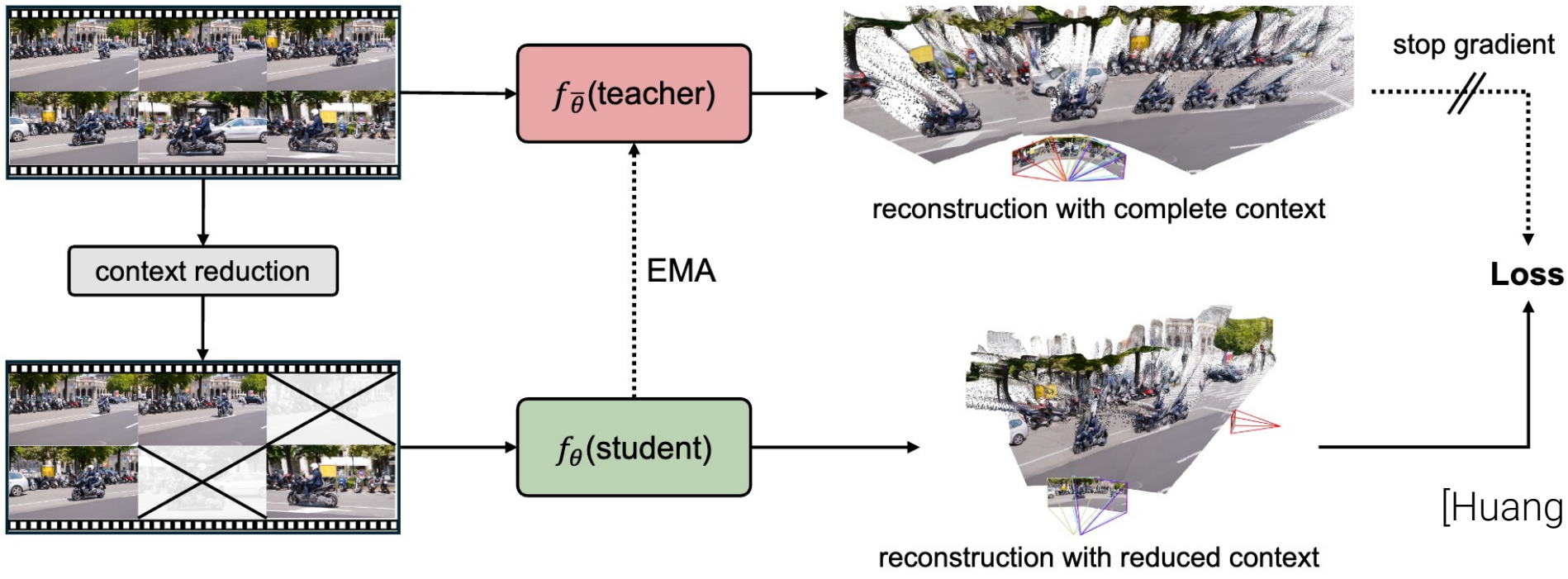
- Sensor fusion: gauge fixing and unobservability
- Outlier rejection: loop closure, unstructured collections (IMC 2025)

We need structured outputs!

Missing benefits of optimization

Uncertainties! Which points/poses are (un)correlated along each axis?

- SSL via bootstrapping: self-distillation rather than COLMAP-distillation

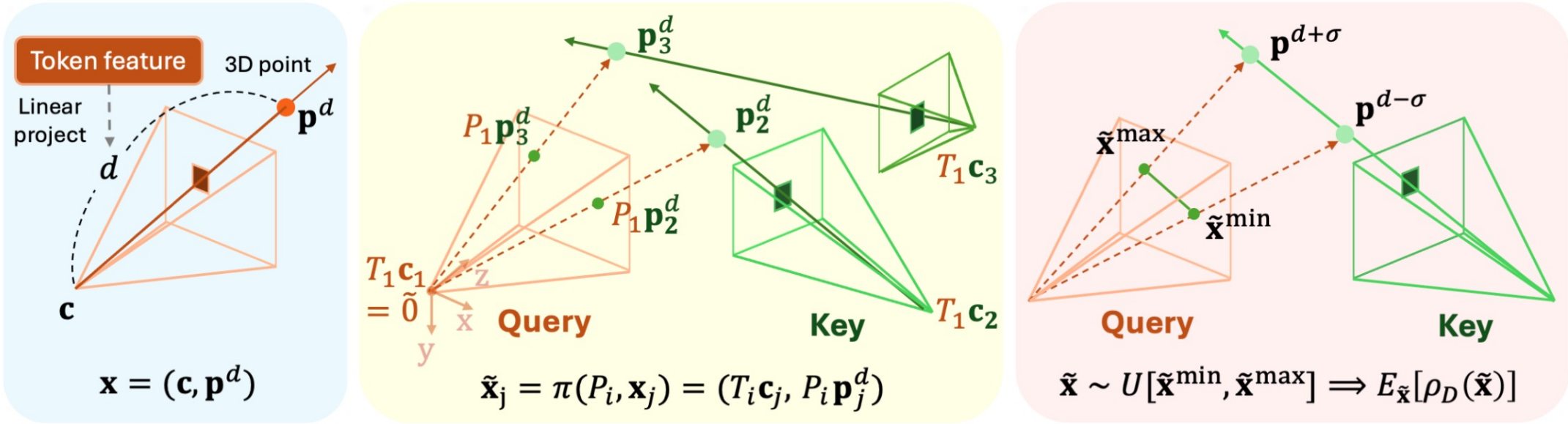


[Huang et al, 2026]

Future paths for e2e geometry models?

Structured internal representations:

- RayRoPE has some great ideas, but what if poses are unknown?
- Coupling more geometric quantities: point maps, normals, calibration

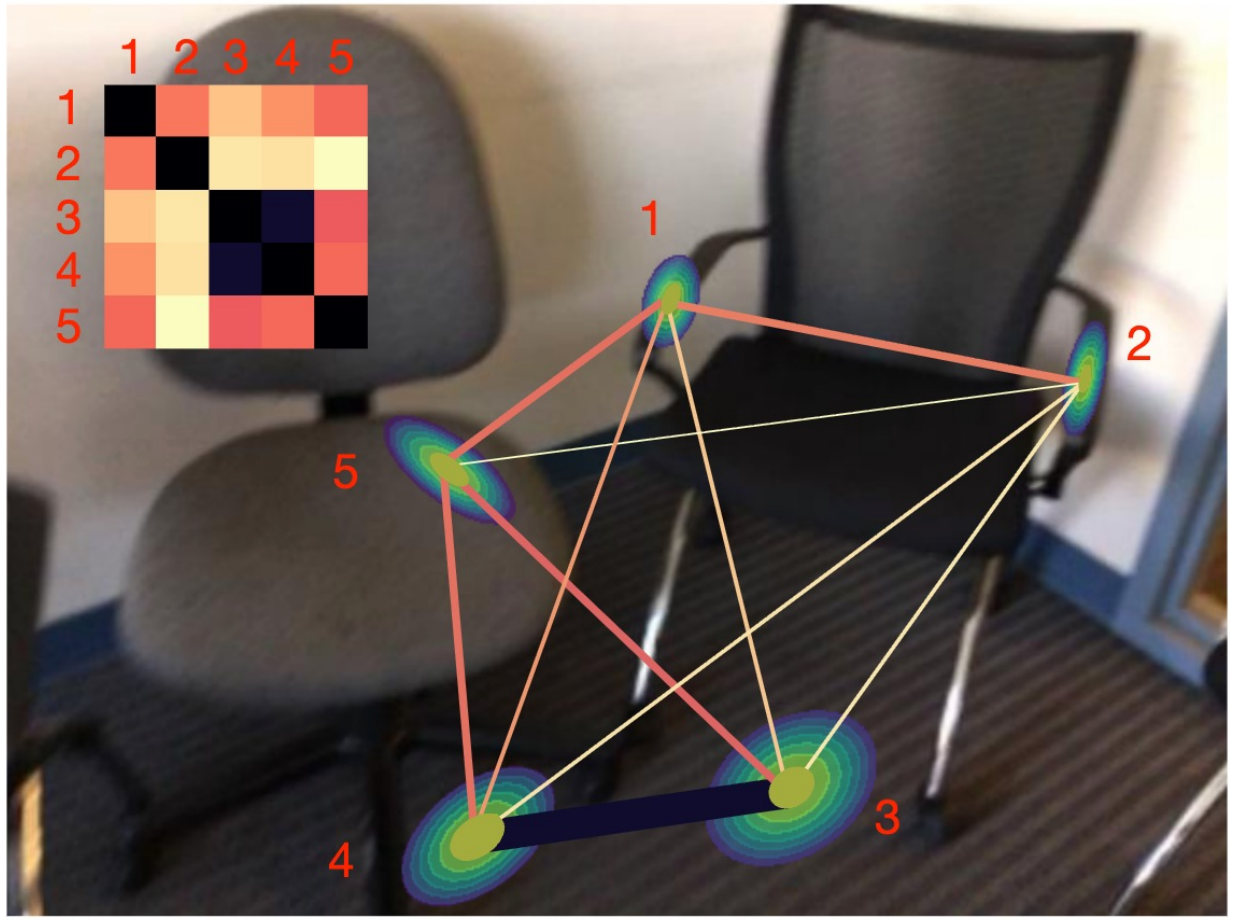


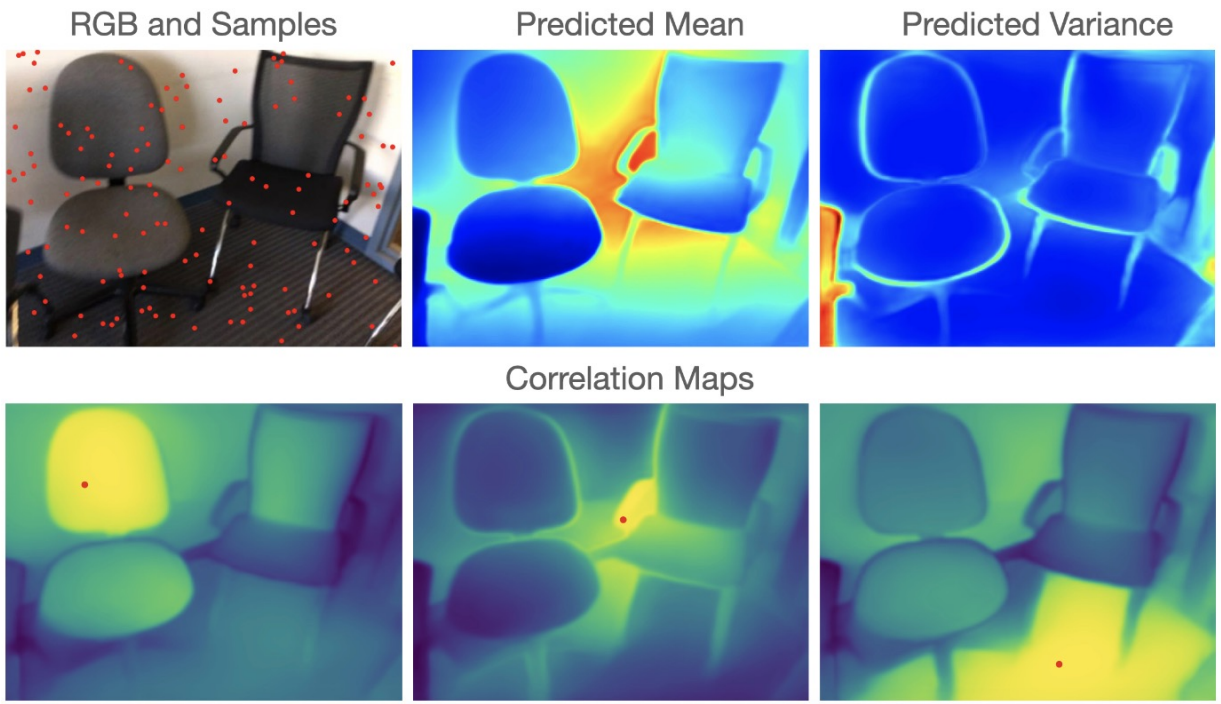
(a) Ray segment positions **(b) Encodings on projected rays** **(c) Expected RoPE**

[Wu et al, 2025]

Future paths for e2e geometry models?

Decoding joint covariances for depth and poses





[Dexheimer & Davison, CVPR 2023]

Is scaling & data curation all we need?

How do we learn long-range reasoning & structures?

- Training on long sequences is wasteful
- Required for doppelgangers, higher-order symmetries, etc.
- Curriculum learning: from short to long sequences?
- Mining related subsequences?

Constraints:

- don't make the training overly expensive
- don't bottleneck information or impair convergence
- don't introduce failures when geometric models don't fit the data

Conclusions

- Yes, e2e models are **impressive**! But they don’t yet dominate
- **Hybrid approaches** still have their own merit, e2e can be a **liability**
- How to retain the benefits of geometry
  while learning even higher-level priors e2e?
- Will this be solved by pure model+data scaling? with SSL?

Thank you!

psarlin.com