ETH zürich

2021 ICCV OCTOBER 11-17 VIRTUAL

Microsoft

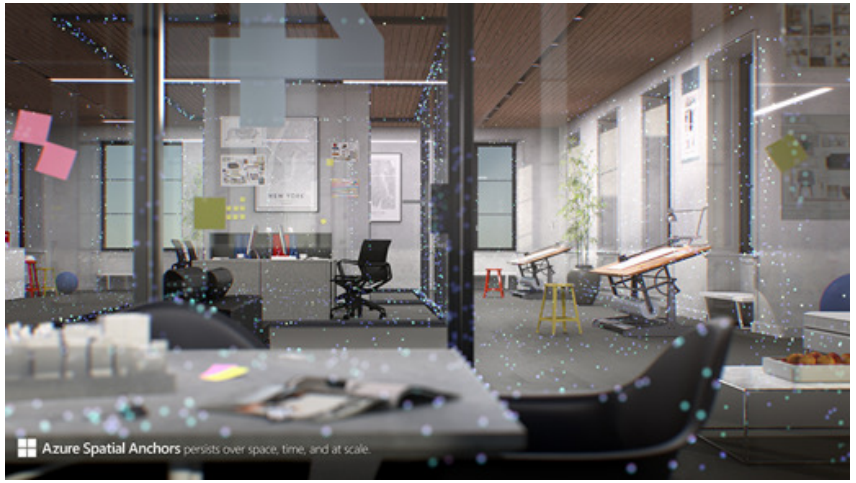
Cross-Descriptor
Visual Localization and Mapping

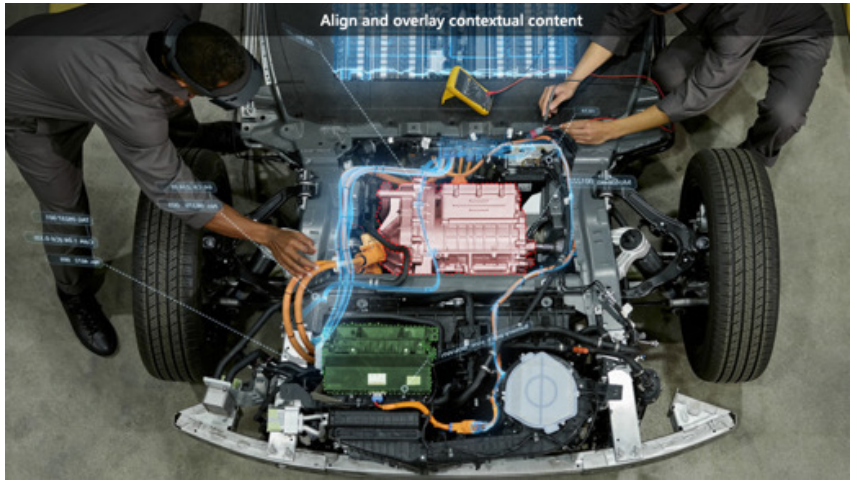


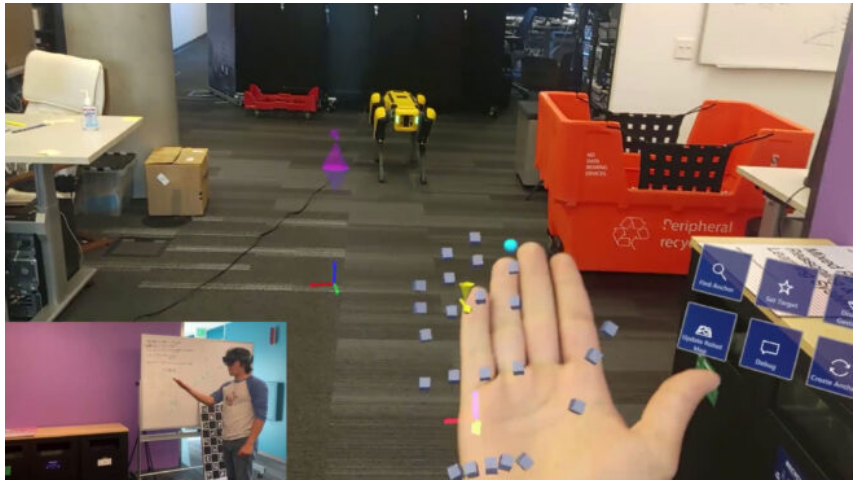

Mihai Dusmanu[1], Ondrej Miksik[2], Johannes L. Schönberger[2], Marc Pollefeys[1,2]

[1]ETH Zürich, [2]Microsoft

Motivation – Cloud-Based MR & Robotics







Sources

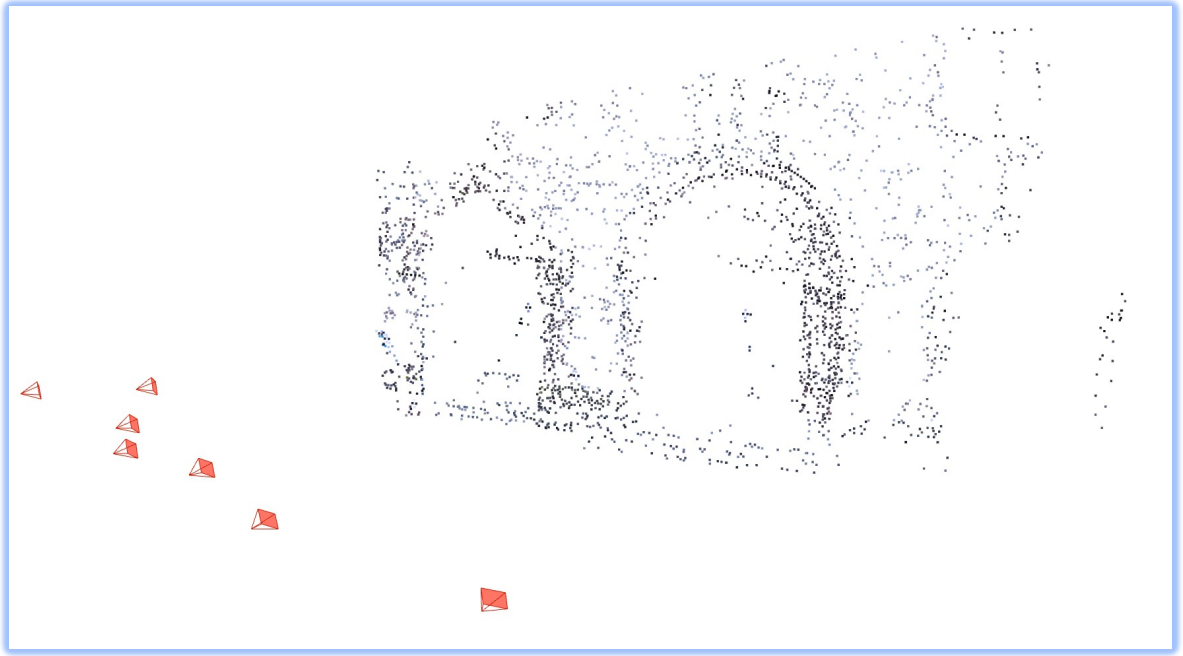
https://azure.microsoft.com/en-us/topic/mixed-reality/

https://www.microsoft.com/en-us/research/lab/mixed-reality-ai-zurich/

Visual Localization

Query Image

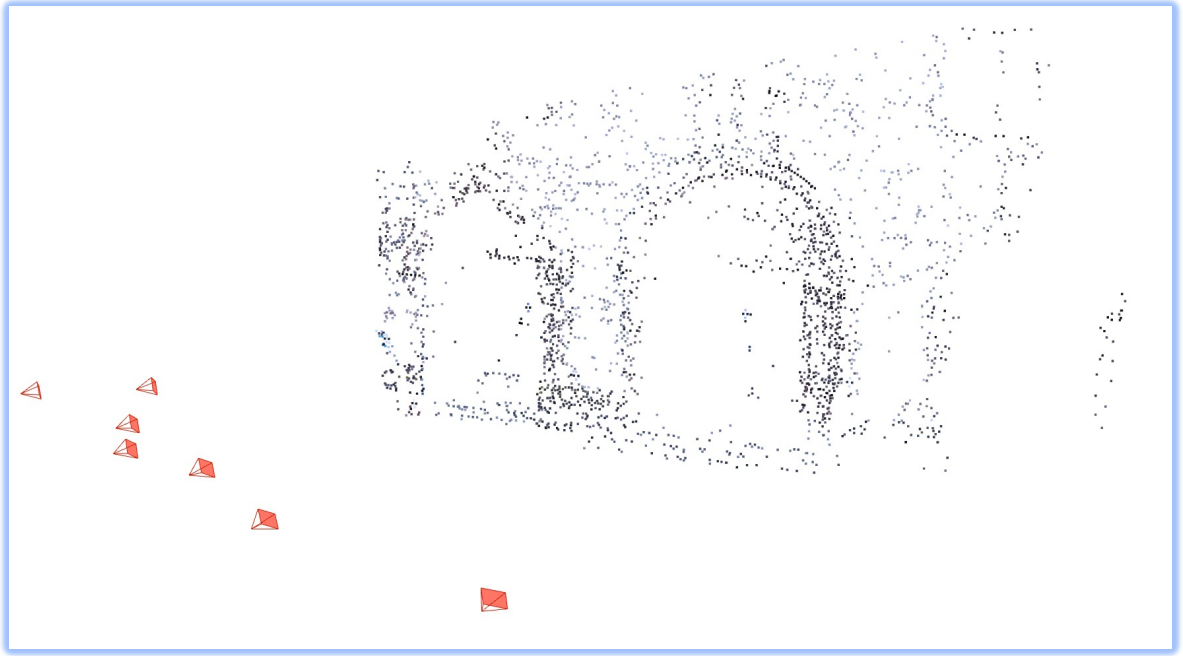
Reference 3D Model



Visual Localization

Query Image

Reference 3D Model



#1 (Local) Feature Extraction

Visual Localization

Query Image

Reference 3D Model

2D-3D correspondences

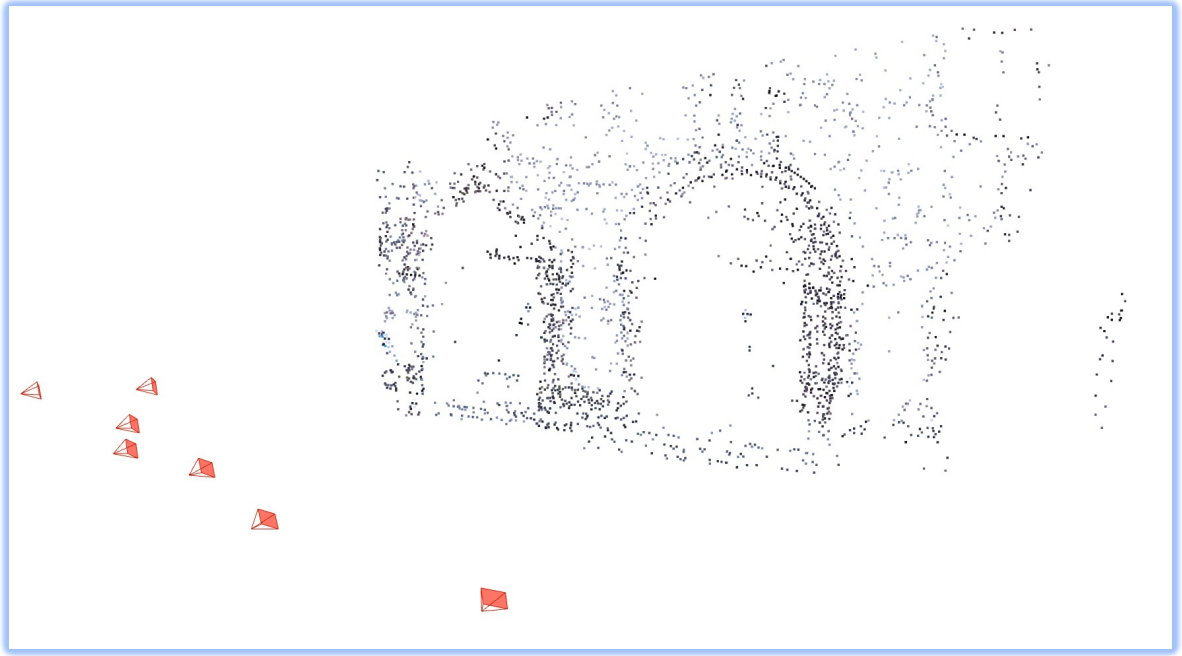
#2 Matching

Visual Localization

Query Image

Global Feature

Reference 3D Model



#2 Matching

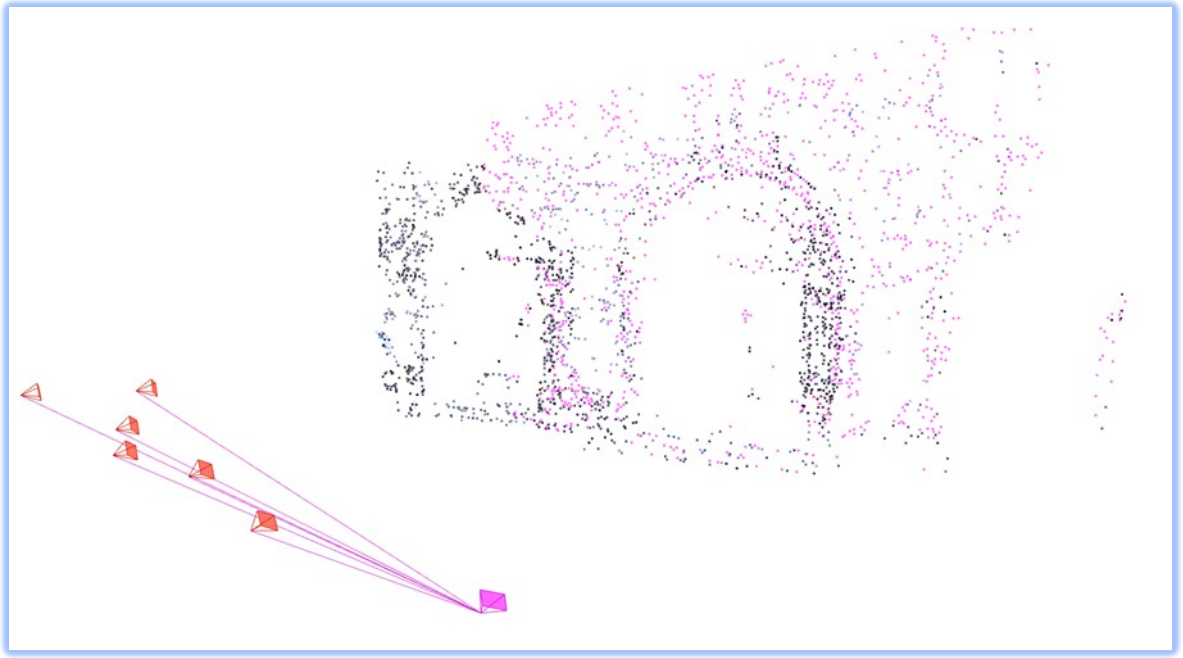
Visual Localization

Query Image

Global Feature

NN Search

Reference 3D Model



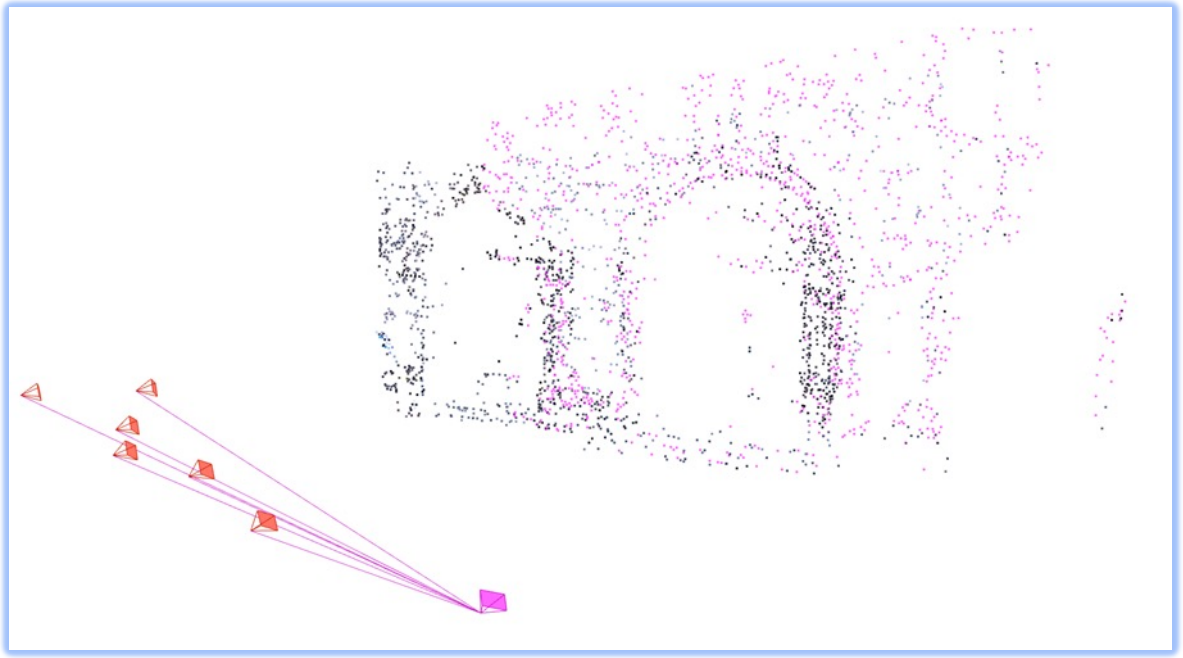
Step 1 – Image Retrieval

#2 Matching

Visual Localization

Query Image

Reference 3D Model



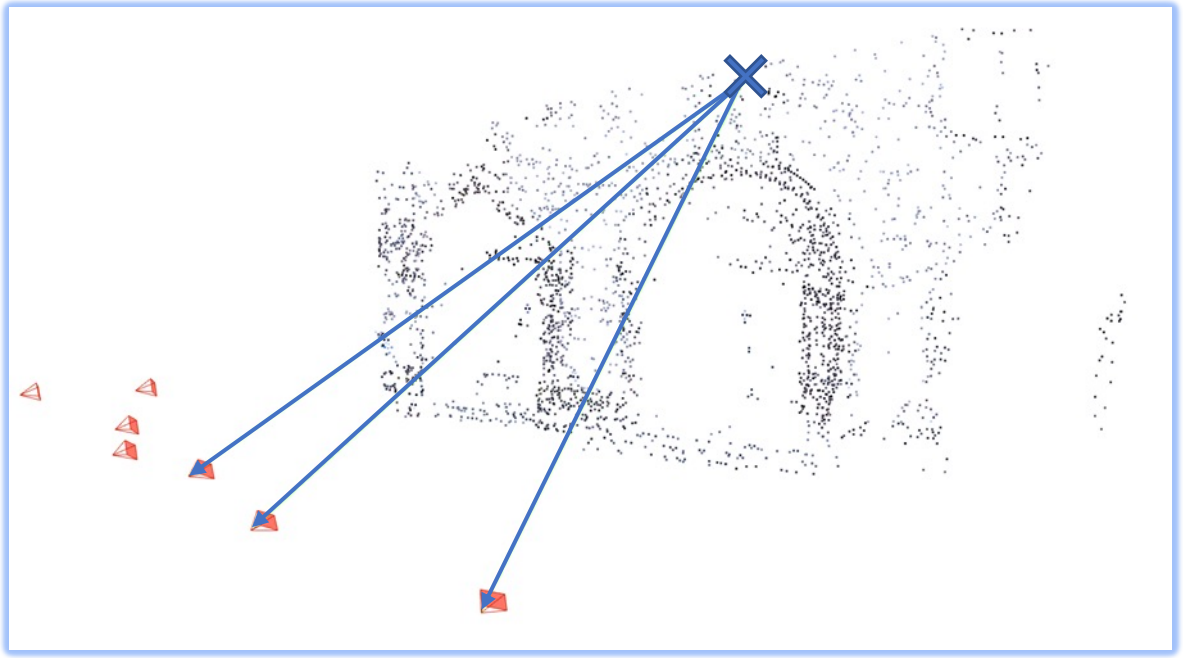
Step 2 – Image Matching

#2 Matching

Visual Localization

Query Image

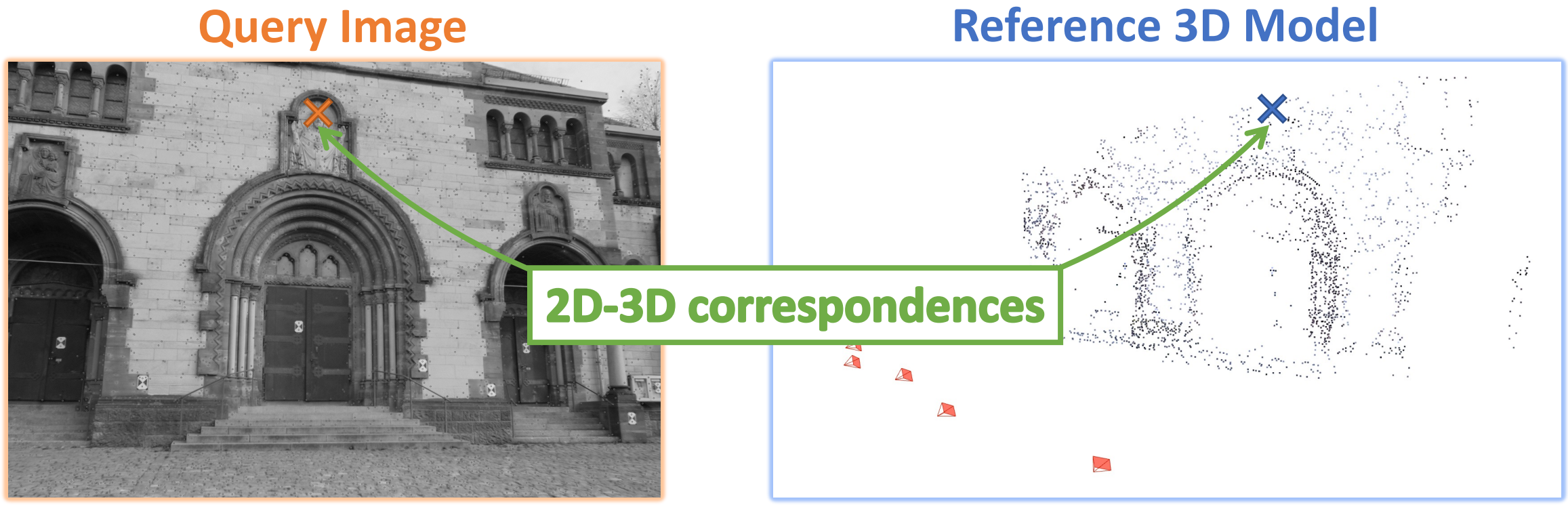
Reference 3D Model



Step 3 – Back-projection

#2 Matching

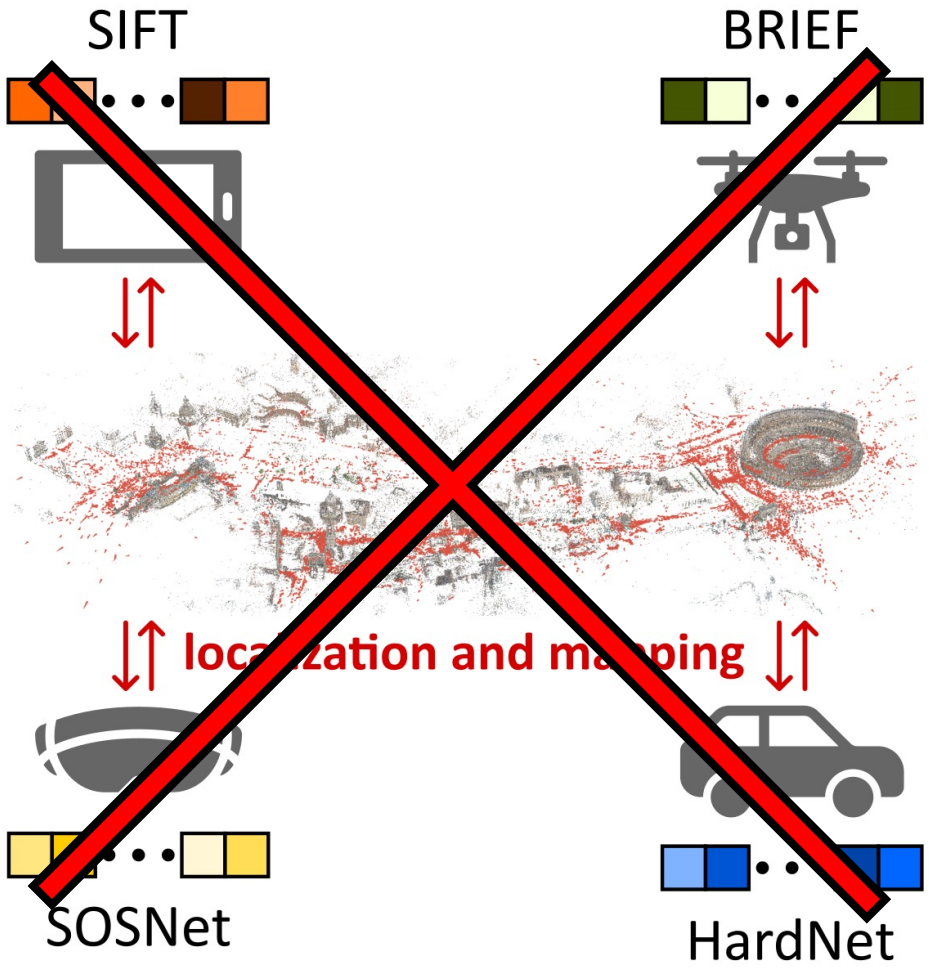
Visual Localization



#3 Pose Estimation (RANSAC)

Visual Localization

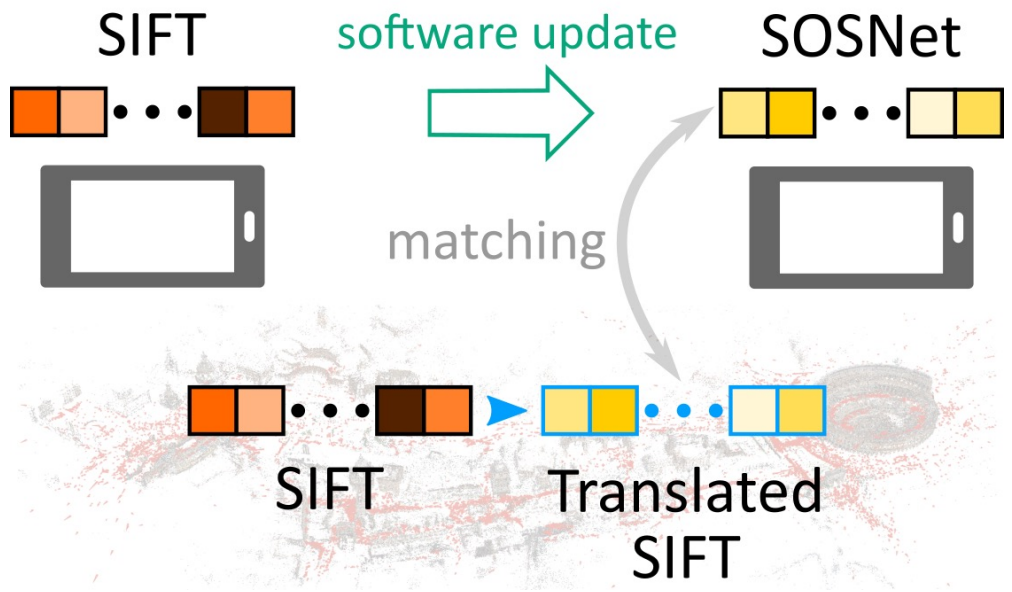
Motivation – Cloud-Based MR & Robotics



Scenarios

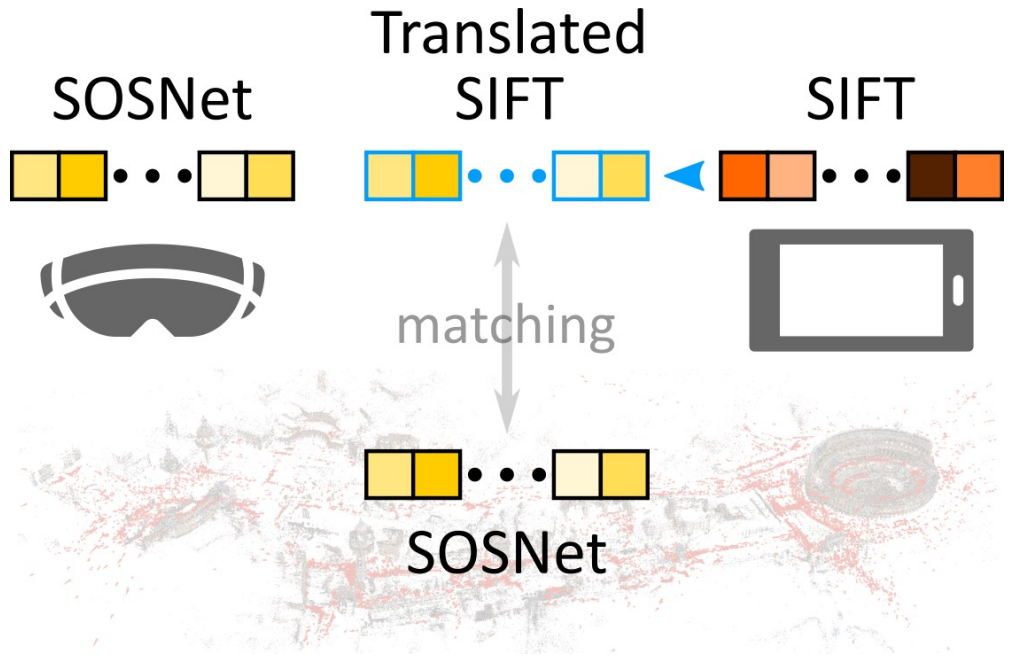
Scenarios – Continuous Deployment

- New feature representations
  - From handcrafted to learned
  - Change the dimensionality
  - Update the learned model
    - New loss / dataset / architecture
- Avoid remapping from scratch



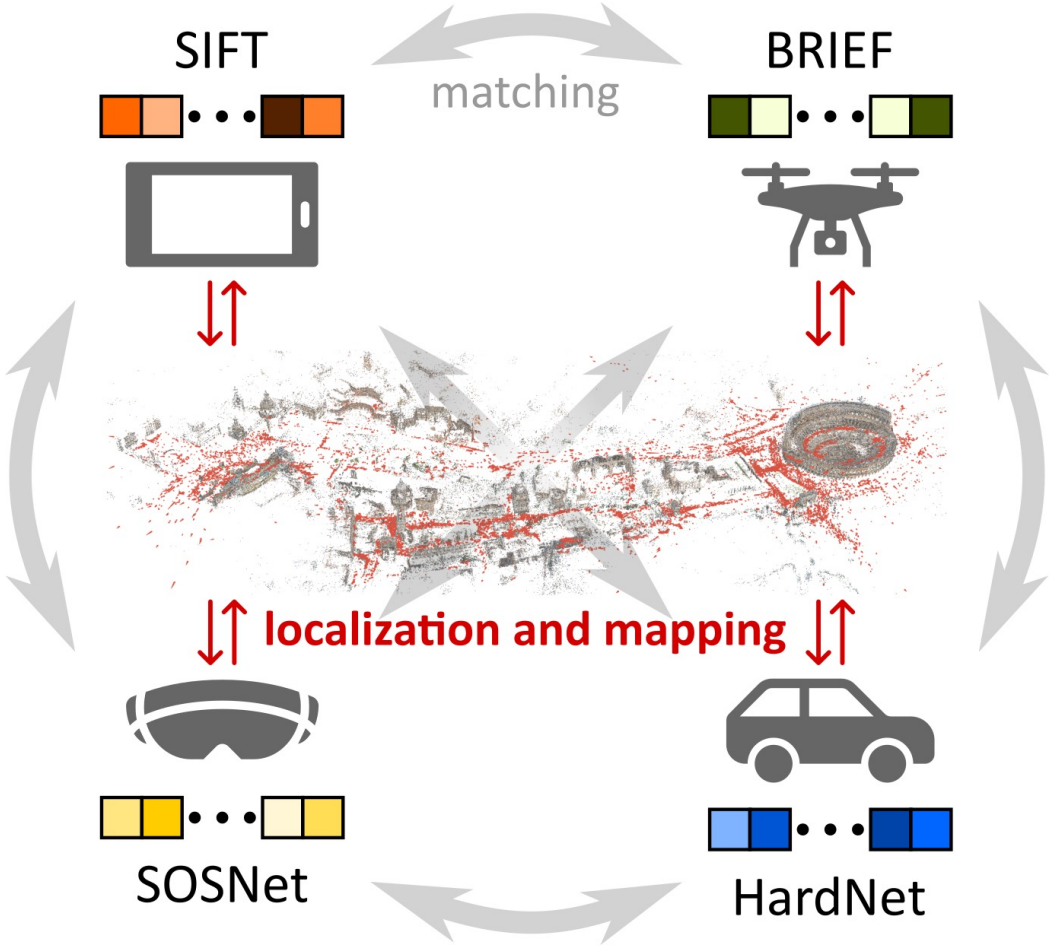
Scenarios – Cross-Device Localization

- Devices with different features
  - Legacy software
  - Hardware limitations
  - Different vendors
    - See discussion of limitations



Scenarios – Collaborative Mapping

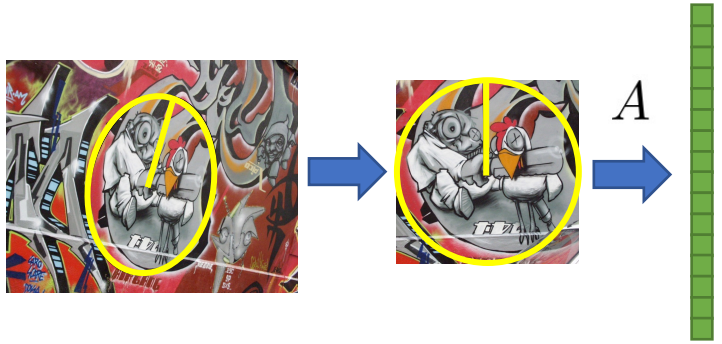
- Multiple devices simultaneously
- A common and consistent map



Descriptor Translation – Introduction

- Description algorithm

$$A : \mathcal{I} \rightarrow \mathbb{R}^n$$



- "Perfect" translation between two algorithms

$$\begin{aligned} A_1 &: \mathcal{I} \rightarrow \mathbb{R}^{n_1} \\ A_2 &: \mathcal{I} \rightarrow \mathbb{R}^{n_2} \end{aligned} \quad \Bigg| \quad \begin{aligned} t_{1\rightarrow 2} &: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2} \\ t_{1\rightarrow 2}(A_1(p)) &= A_2(p) \text{ for all patch } p \in \mathcal{I} \end{aligned}$$

- Idea: data-driven – approximate this function using MLPs

Descriptor Translation – Pair Network

- One MLP for each tuple
- Translation loss
  - Floating point descriptors

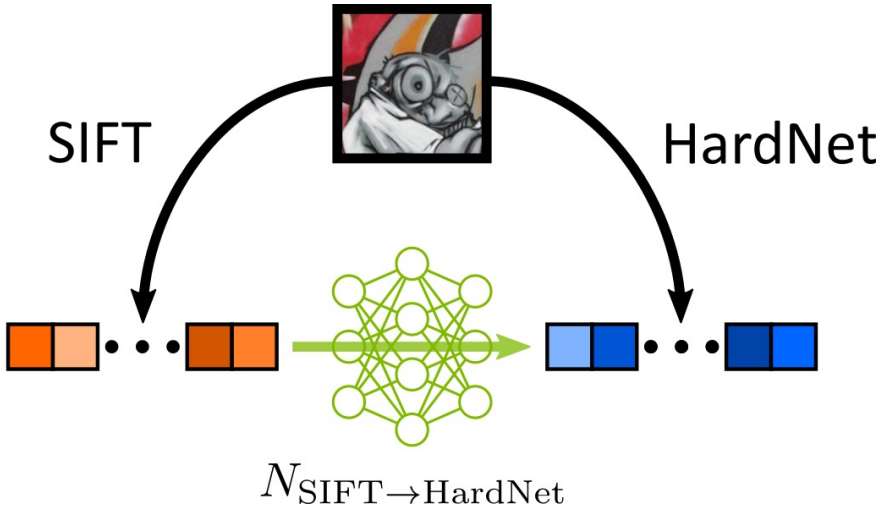
$$\mathcal{L}^T_{i\to j} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \|N_{i\to j}(A_i(p)) - A_j(p)\|$$

  - Binary descriptors

$$\mathcal{L}^T_{i\to j} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathrm{BCE}\left(N_{i\to j}(A_i(p)), A_j(p)\right)$$

- Descriptor distance

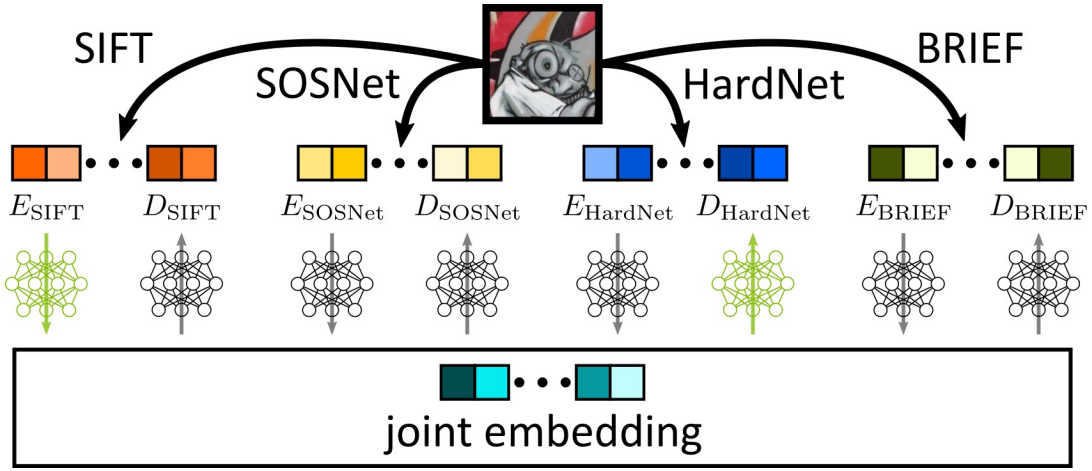
$$\|N_{i\to j}(d_i) - d_j\| \qquad \|d_i - N_{j\to i}(d_j)\|$$



Descriptor Translation – Joint Embedding

- One encoder and one decoder for each description algorithm
- Triplet margin loss for direct matching in the joint space

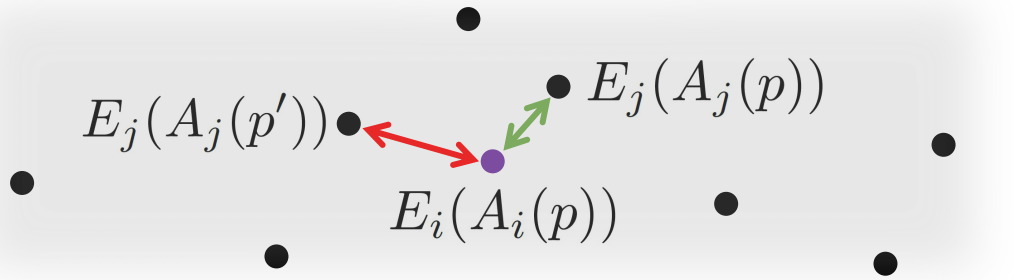
$$\mathcal{L}^{M}_{i\to j} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \max\left(m + \mathrm{pos}(p) - \mathrm{neg}(p), 0\right)$$



$$N_{i\to j} = D_j \circ E_i$$

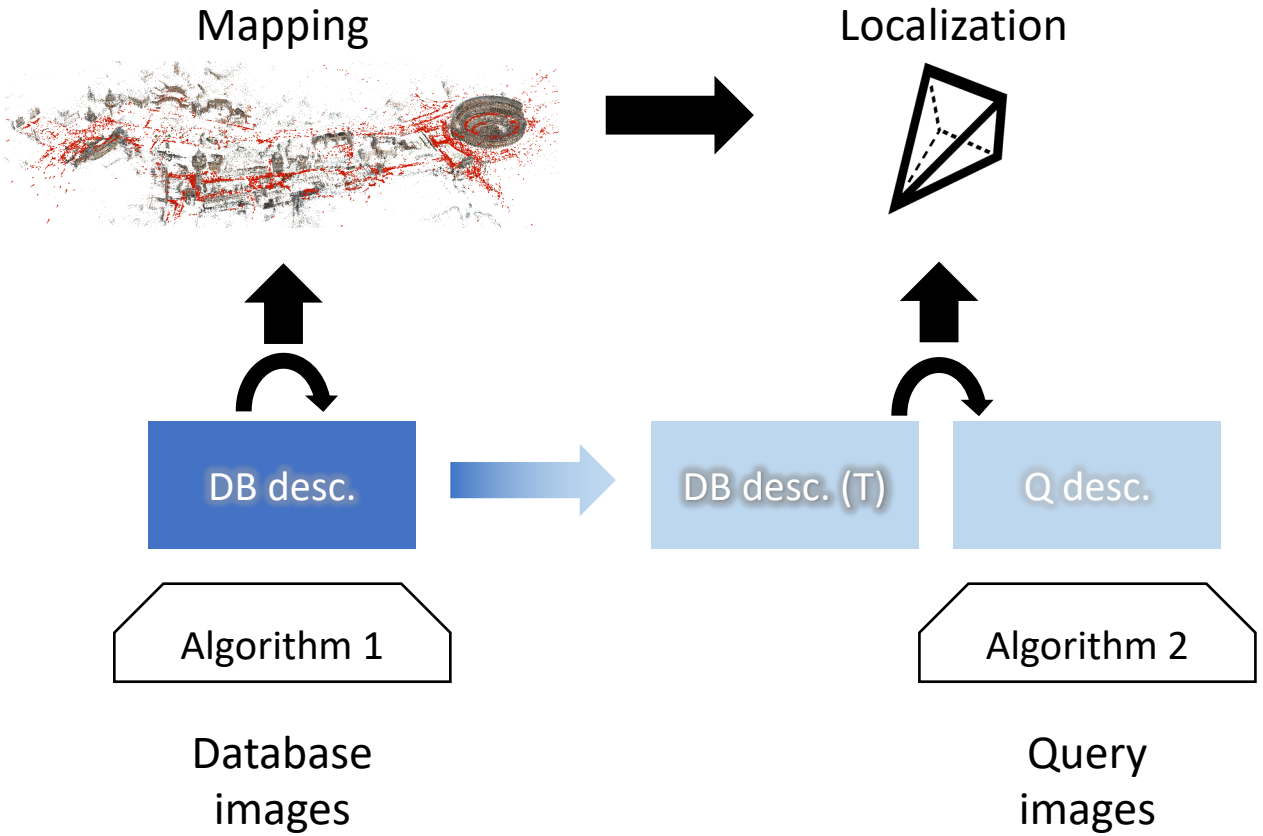
- Descriptor distance

$$\|E_i(d_i) - E_j(d_j)\|$$



Evaluation – Localization

Continuous Deployment



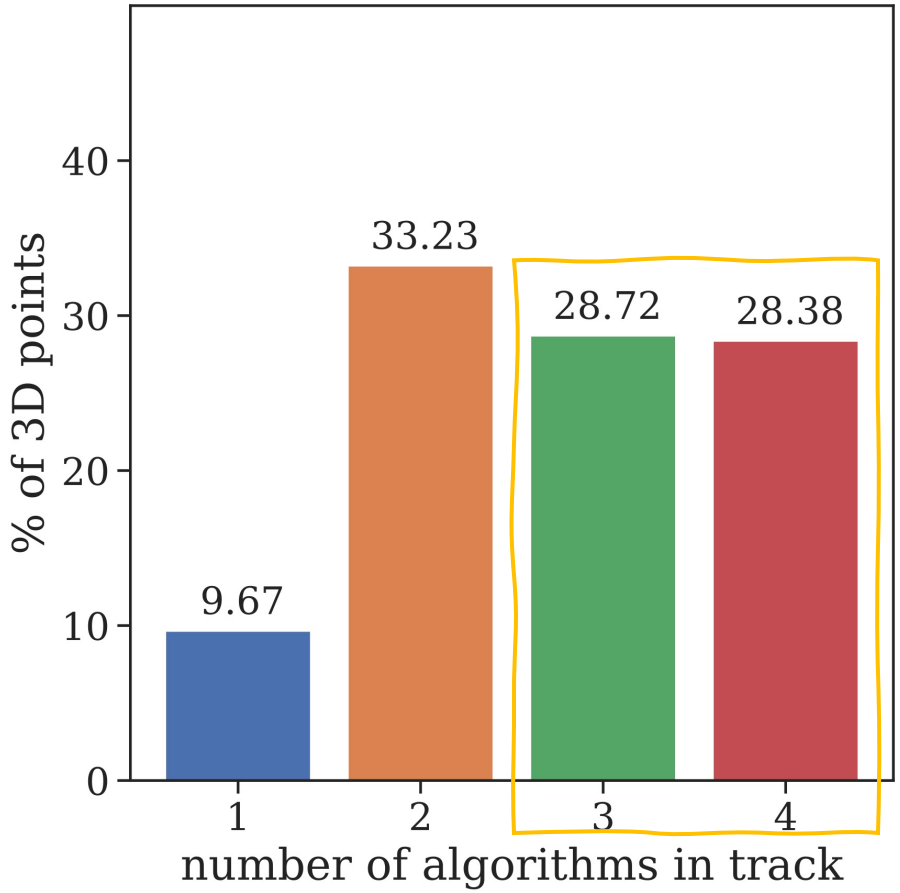
Scenario	Database descriptor	Query descriptor	% localized queries: Day (824 images) $0.25m, 2°$	Day (824 images) $0.5m, 5°$	Night (98 images) $0.25m, 2°$	Night (98 images) $0.5m, 5°$
Standard	BRIEF	BRIEF	76.1	81.4	32.7	36.7
	SIFT	SIFT	82.5	88.7	52.0	61.2
	HardNet	HardNet	86.2	92.2	64.3	72.4
	SOSNet	SOSNet	86.4	92.7	65.3	75.5
Continuous deployment	BRIEF →	SIFT	74.9 -1.2	80.5 -0.9	31.6 -1.1	36.7 0.0
		HardNet	81.4 +5.3	86.7 +5.3	44.9 +12.2	49.0 +12.3
		SOSNet	81.6 +5.5	86.9 +5.5	42.9 +10.2	46.9 +10.2
	SIFT →	BRIEF	66.6 -15.9	73.1 -15.6	19.4 -32.6	23.5 -37.7
		HardNet	83.4 +0.9	90.9 +2.2	59.2 +7.2	66.3 +5.1
		SOSNet	84.2 +1.7	91.4 +2.7	55.1 +3.1	62.2 +1.0
	HardNet →	BRIEF	70.5 -15.7	76.7 -15.5	22.4 -41.9	26.5 -45.9
		SIFT	81.2 -5.0	88.0 -4.2	41.8 -22.5	51.0 -21.4
		SOSNet	85.8 -0.4	92.4 +0.2	61.2 -3.1	68.4 -4.0
	SOSNet →	BRIEF	68.8 -17.6	74.8 -17.9	18.4 -46.9	20.4 -55.1
		SIFT	81.7 -4.7	87.5 -5.2	42.9 -22.4	49.0 -26.5
		HardNet	85.9 -0.5	92.4 -0.3	63.3 -2.0	69.4 -6.1

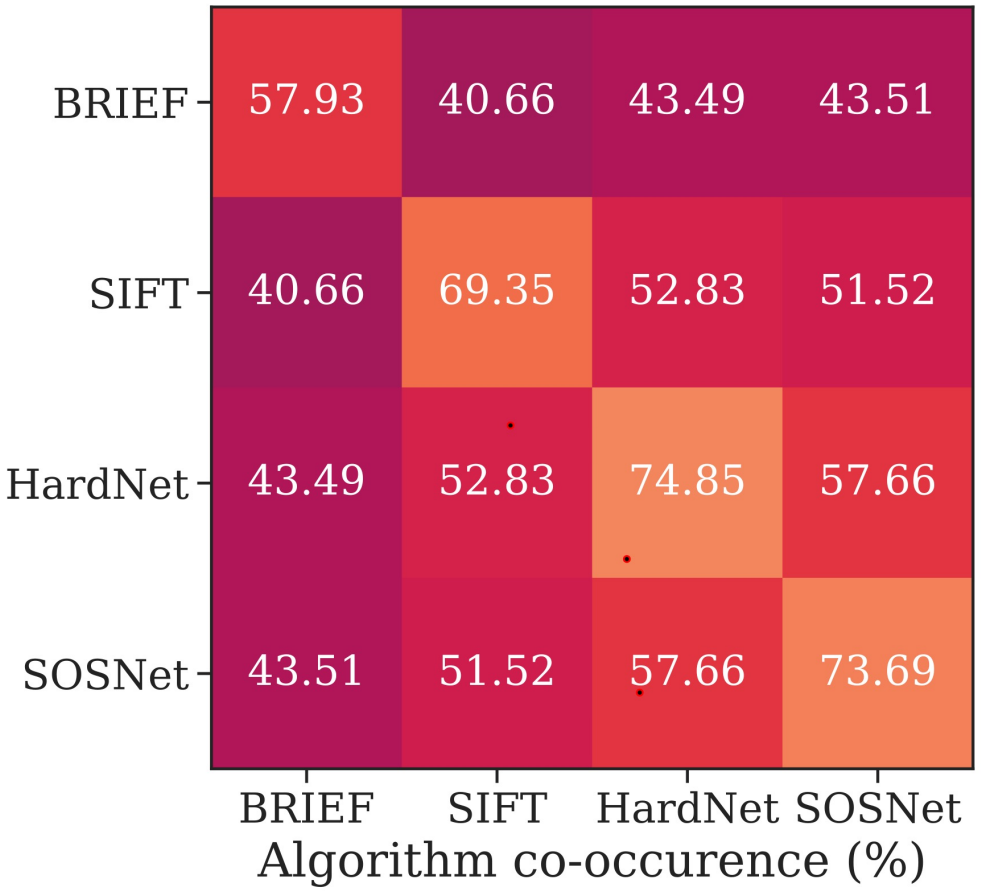
Evaluation – Mapping

- Split images in balanced subsets (one for each algorithm)
- Standard: each algorithm has **access to all images**
- Real-world: each algorithm has **access only to its split**
- Ours (cross-descriptor):
  - Embed: joint embedding
  - Progressive: decide online

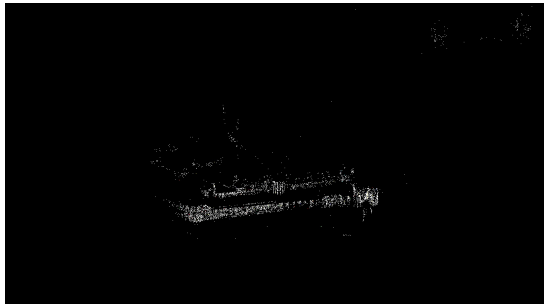
Dataset		*Tower of London* – 730 images					
	Method	% localized images $0.25m$ $2°$	$0.5m$ $5°$	∞	Num. 3D pts.	Track length	Reproj. error
Standard	BRIEF	64.9	68.5	74.2	48.1K	7.70	0.66
	SIFT	74.2	76.7	97.1	90.0K	7.14	0.81
	HardNet	83.0	87.7	100	104.5K	7.56	0.87
	SOSNet	85.2	89.2	100	101.3K	7.67	0.86
Real-world	BRIEF	10.4	11.5	11.8	5.8K	4.65	0.57
	SIFT	13.0	16.0	17.1	15.2K	4.55	0.74
	HardNet	16.0	17.8	18.6	22.1K	5.23	0.80
	SOSNet	17.3	18.8	23.0	22.9K	5.32	0.80
Ours	Embed	77.3	81.2	97.9	88.5K	7.57	0.81
	Progressive	79.2	83.2	96.6	76.0K	7.81	0.82

Evaluation – Mapping Details

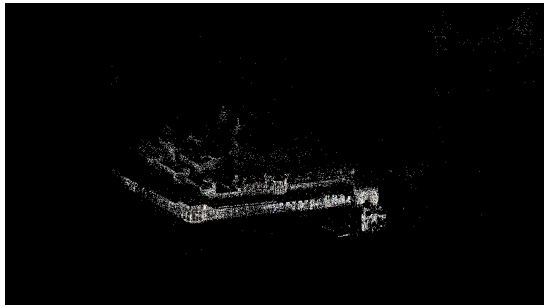




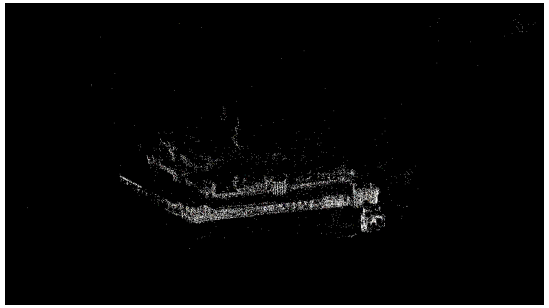
Evaluation – Mapping Models



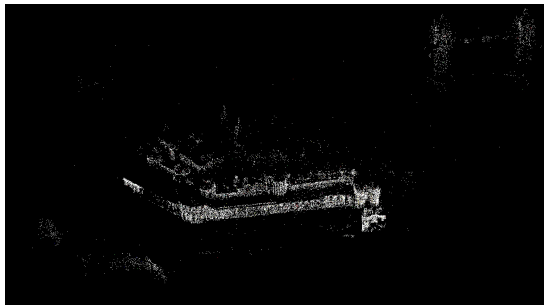
BRIEF



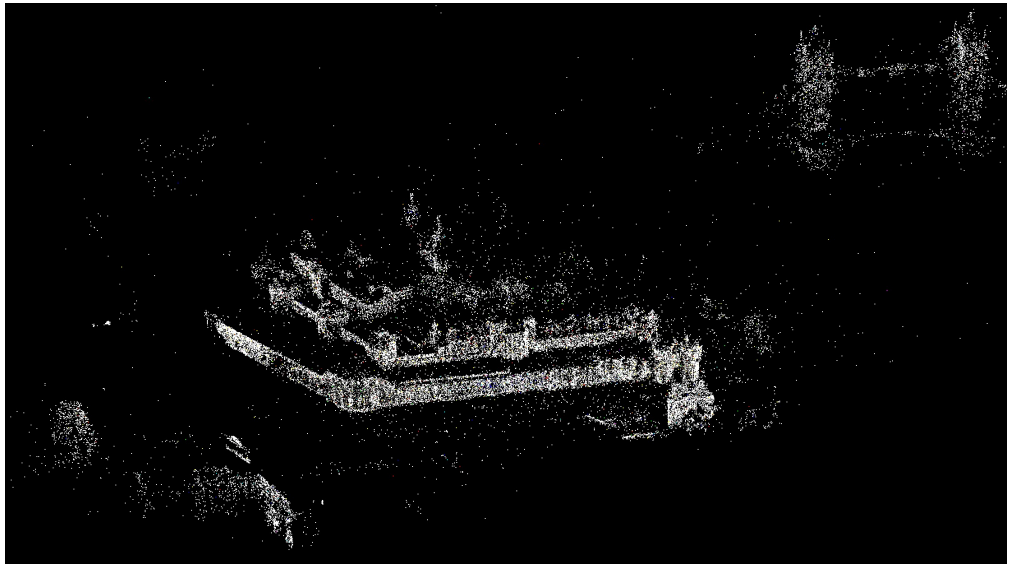
HardNet



SIFT



SOSNet

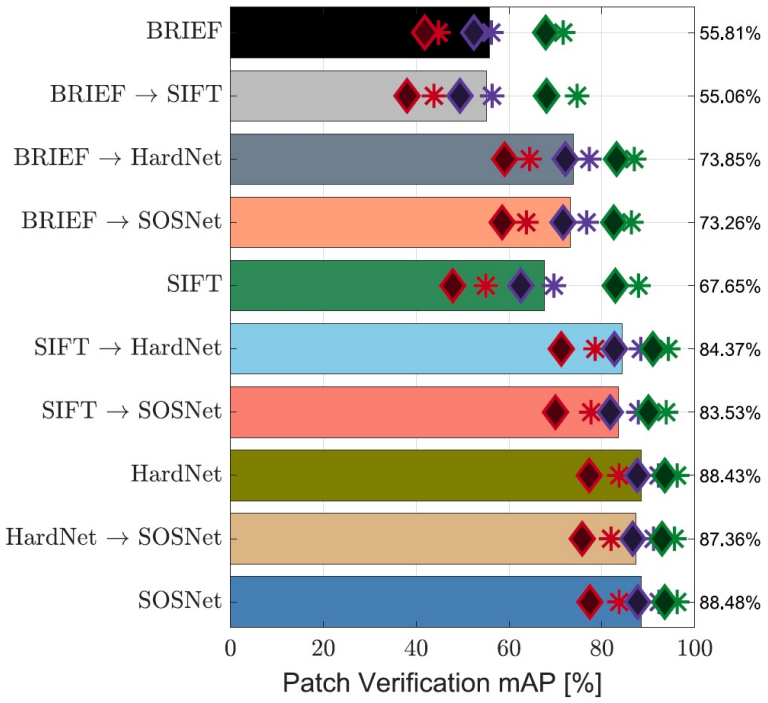


Embed

Evaluation – Other Results

Scenario	Database descriptor	Query descriptor	% localized queries DUC1 0.25m	DUC1 0.5m	DUC2 0.25m	DUC2 0.5m
Standard	BRIEF	BRIEF	25.3	36.4	22.9	41.2
	SIFT	SIFT	32.3	47.5	27.5	45.0
	HardNet	HardNet	36.4	52.5	30.5	54.2
	SOSNet	SOSNet	34.8	50.5	30.5	53.4
Cont. deployment	BRIEF →	SIFT	28.3	39.9	22.1	40.5
		HardNet	29.8	43.9	30.5	40.5
		SOSNet	31.8	43.4	23.7	40.5
	SIFT →	HardNet	36.4	50.0	31.3	50.4
		SOSNet	36.4	53.5	33.6	50.4
	HardNet →	SOSNet	33.3	48.5	30.5	55.7
Cross-device	SIFT ←	BRIEF	29.3	40.9	25.2	42.0
	HardNet ←	BRIEF	30.3	46.5	27.5	48.1
		SIFT	36.4	51.0	33.6	55.7
	SOSNet ←	BRIEF	29.8	44.9	29.0	45.0
		SIFT	34.8	51.0	33.6	53.4
		HardNet	37.4	50.5	29.0	49.6

InLoc – indoor localization



HPatches descriptor benchmark

	Descriptor	Stereo AUC (%) 5°	Stereo AUC (%) 10°	Multi-view AUC (%) 5°	Multi-view AUC (%) 10°	Real.
Standard	BRIEF	35.3	41.8	31.9	36.5	✓
	SIFT	41.4	49.2	41.4	48.7	✓
	HardNet	51.4	59.9	55.9	63.5	✓
	SOSNet	51.4	60.1	58.6	66.2	✓
Directional	BRIEF → SIFT	25.2	31.5	14.9	17.6	✗
	BRIEF → HardNet	35.3	42.7	36.5	40.8	✗
	BRIEF → SOSNet	39.8	47.5	40.3	46.9	✗
	SIFT → HardNet	42.7	51.1	48.7	55.4	✗
	SIFT → SOSNet	45.1	53.5	47.3	55.2	✗
	HardNet → SOSNet	49.4	57.8	56.9	64.3	✗
Embed	BRIEF, SIFT, 1/2	39.5	47.1	41.6	48.1	✓
	BRIEF, HardNet, 1/2	42.4	50.3	46.2	52.8	✓
	BRIEF, SOSNet, 1/2	41.3	48.9	45.2	51.9	✓
	SIFT, HardNet, 1/2	46.8	55.1	53.4	61.3	✓
	SIFT, SOSNet, 1/2	46.2	54.4	49.9	57.6	✓
	HardNet, SOSNet, 1/2	50.4	58.9	57.6	64.9	✓
	All, 1/4	42.3	50.1	46.7	53.5	✓

IMW Challenge

Limitations and Future Work

- Same keypoints assumption
  - Likely applicable when a single manufacturer is involved
  - **Cross-detector repeatability?**
  - **Feature inversion to reconstruct input image?**

- Translation function issues
  - Information loss
  - One-to-many associations
  - **Exploit local / global context to disambiguate?**

ETH zürich

2021 ICCV OCTOBER 11-17 VIRTUAL

Microsoft

Cross-Descriptor
Visual Localization and Mapping

Project Page

Mihai Dusmanu[1], Ondrej Miksik[2], Johannes L. Schönberger[2], Marc Pollefeys[1,2]

[1]ETH Zürich, [2]Microsoft