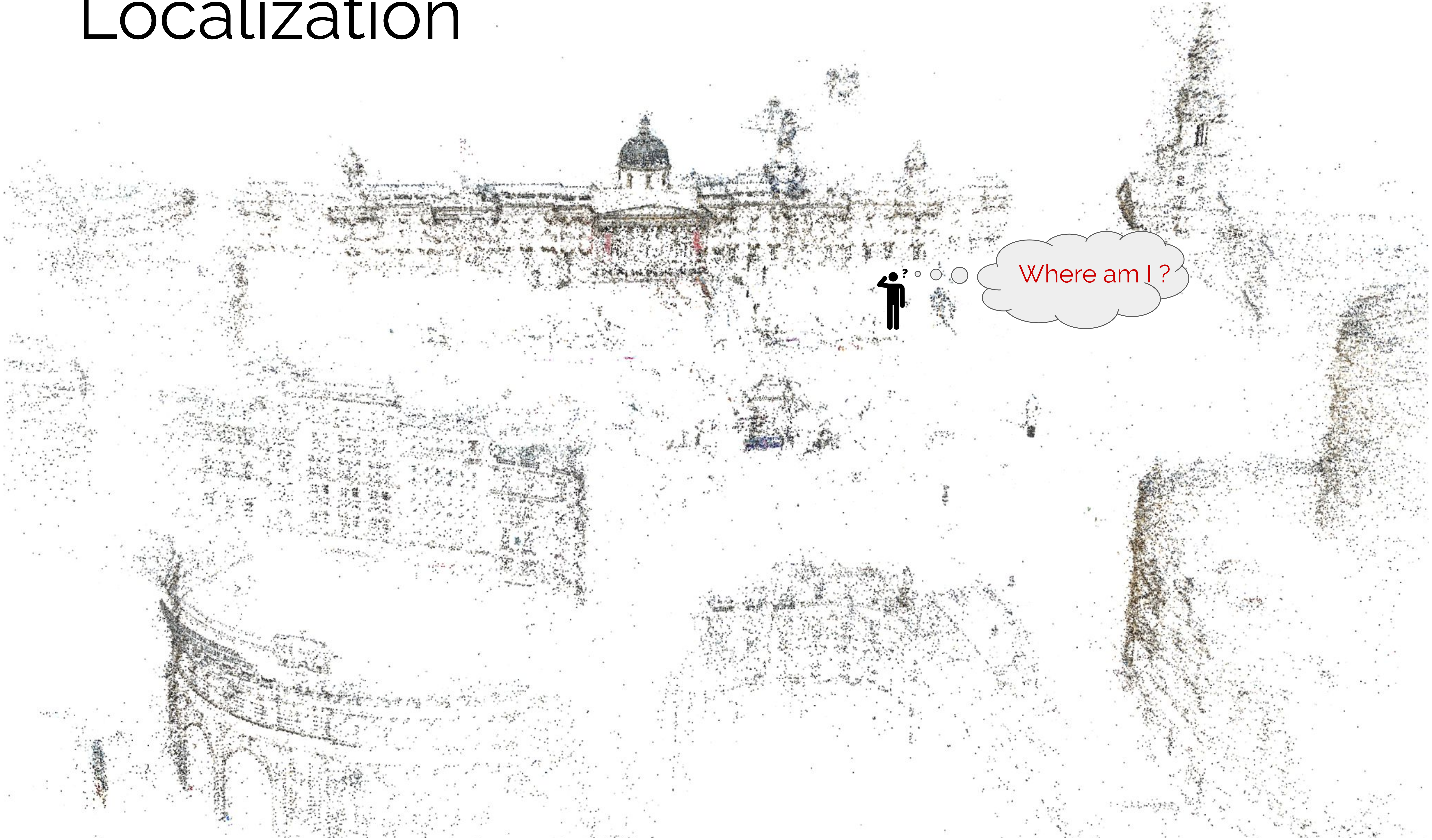
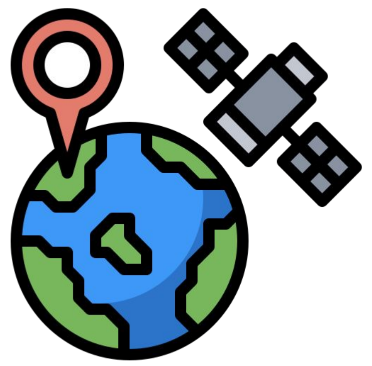
Exploring scene representations for visual localization

Laura Leal-Taixé | CVPR | June 2024

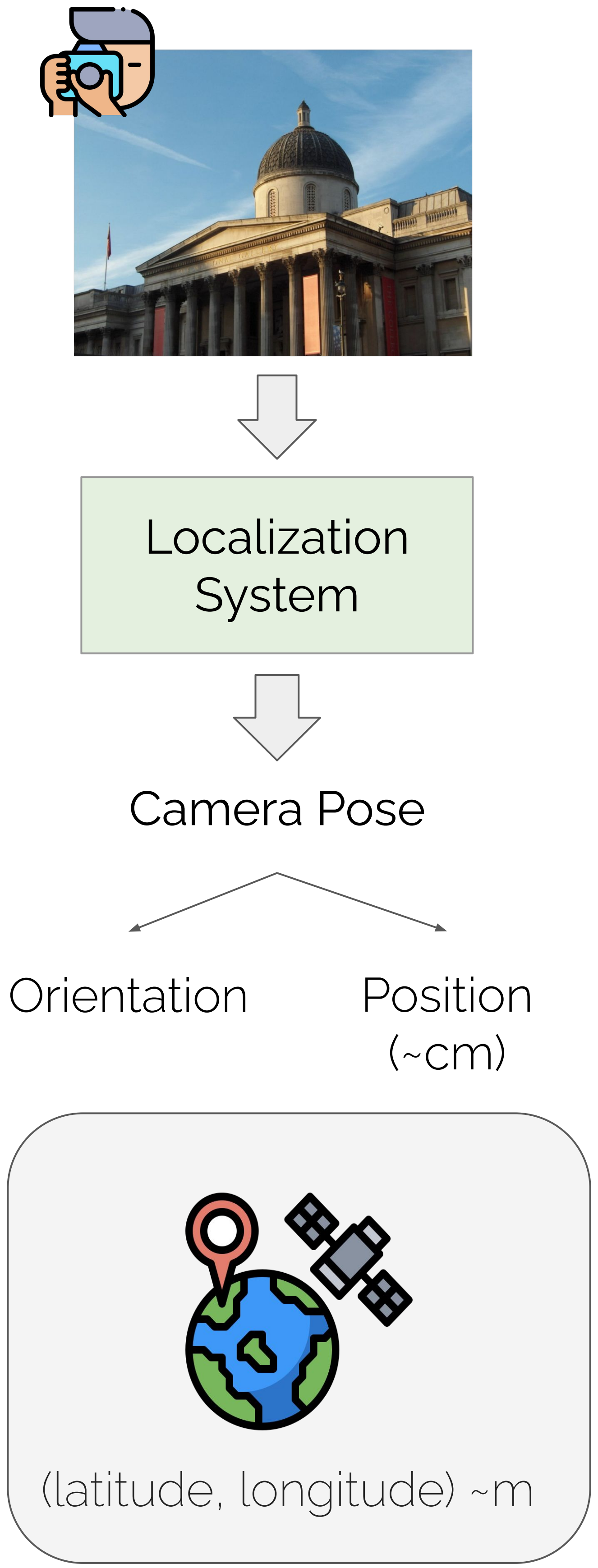
Localization



Global Positioning System (GPS)



Visual Localization

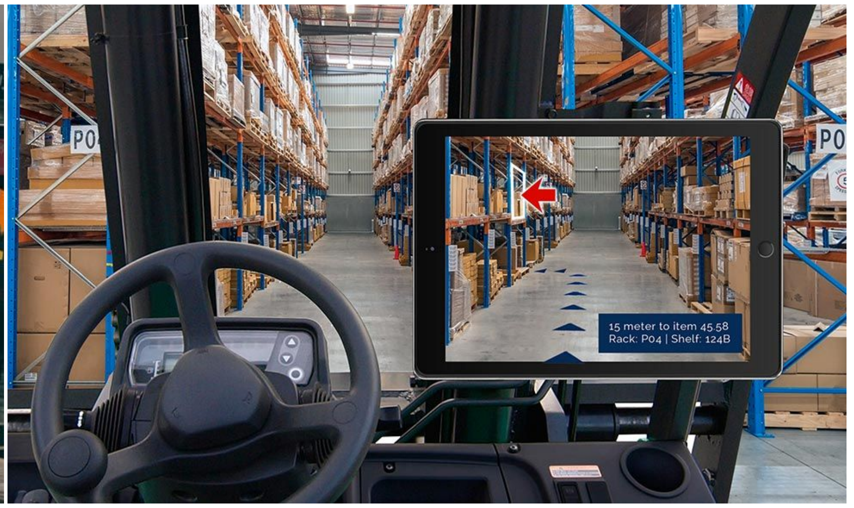


Applications

Indoor / Outdoor Navigation
(GPS-unavailable /unreliable)

src:https://techcrunch.com/2018/08/09/blippar-is-using-ar-to-help-customers-find-their-way-indoors/

src:https://xrlabs.co/how-ar-vr-experiences-can-enhance-tourism-experiences/



src:https://insidernavigation.com/ar-indoor-navigation/

src:https://advanced-robotics.ch/robot-for-events/

src:https://mashable.com/video/aeolus-robot-cleans-your-houses-serves-you-drinks-uses-vacuum

src:https://www.latimes.com/world-nation/story/2020-05-31/hello-and-welcome-robot-waiters-to-the-rescue-amid-virus

Autonomous Service Robots

AR / VR
(Require cm-mm accuracy)

src:https://blog.guidigo.com/blog/augmented-reality-is-coming-to-museums-this-fall-with-guidigo-ar-composer/

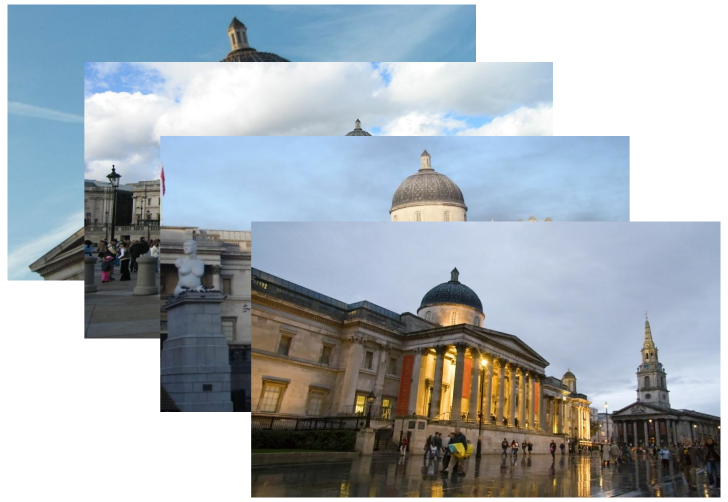
src: A mock-up of design app HoloStudio

src:Microsoft Hololens Project XRay Demo

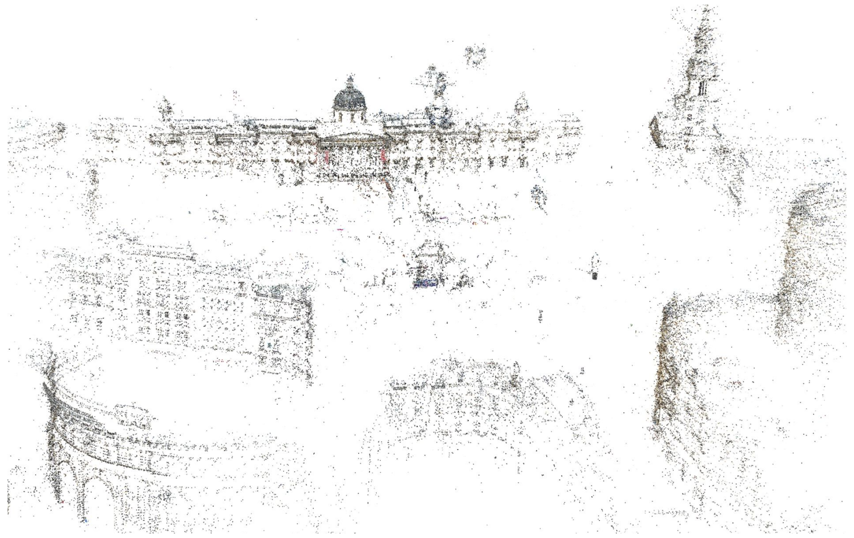
Localization System

Query	Scene Map	Method	Outputs

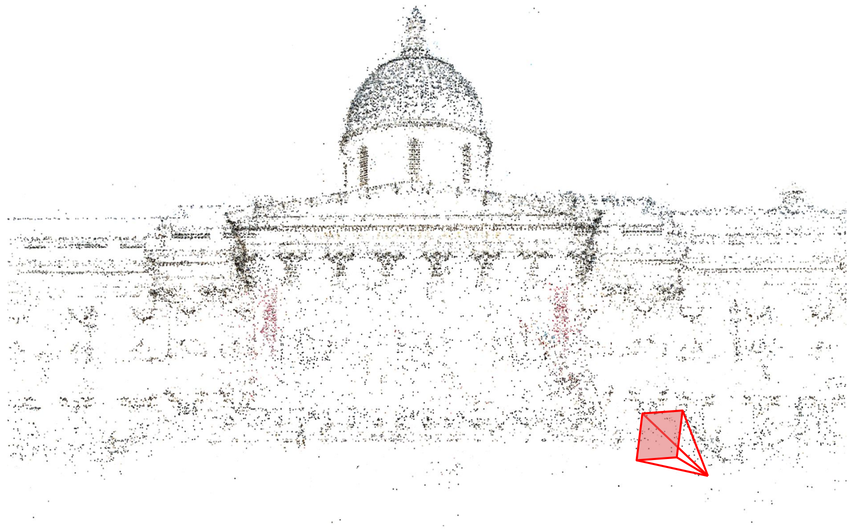
Query Image

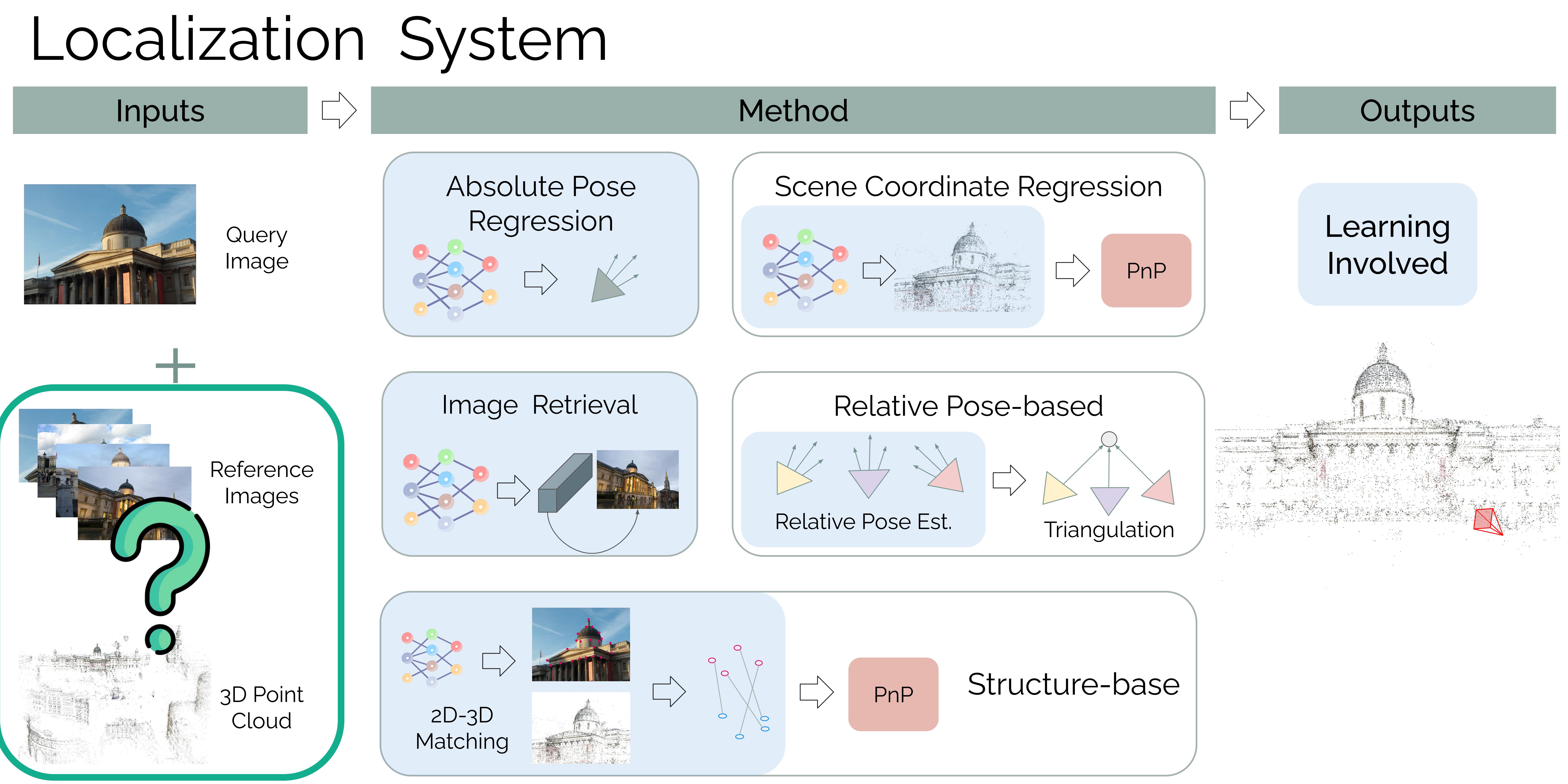


Reference Images



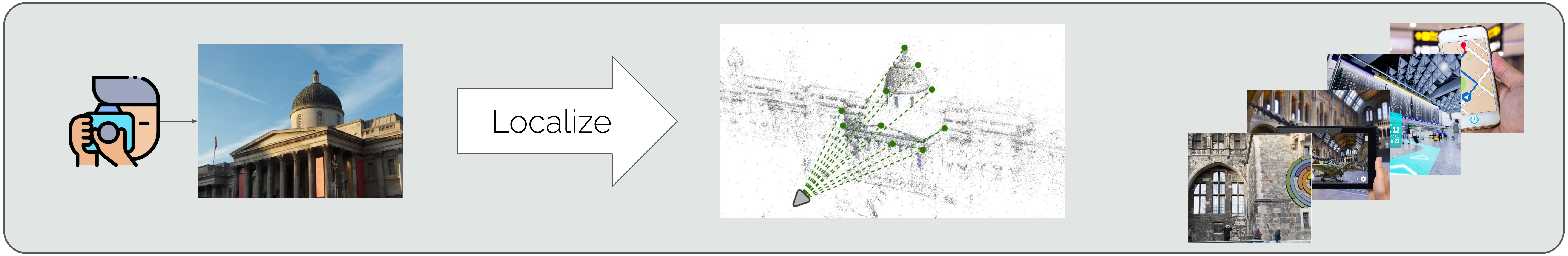
3D Point Cloud

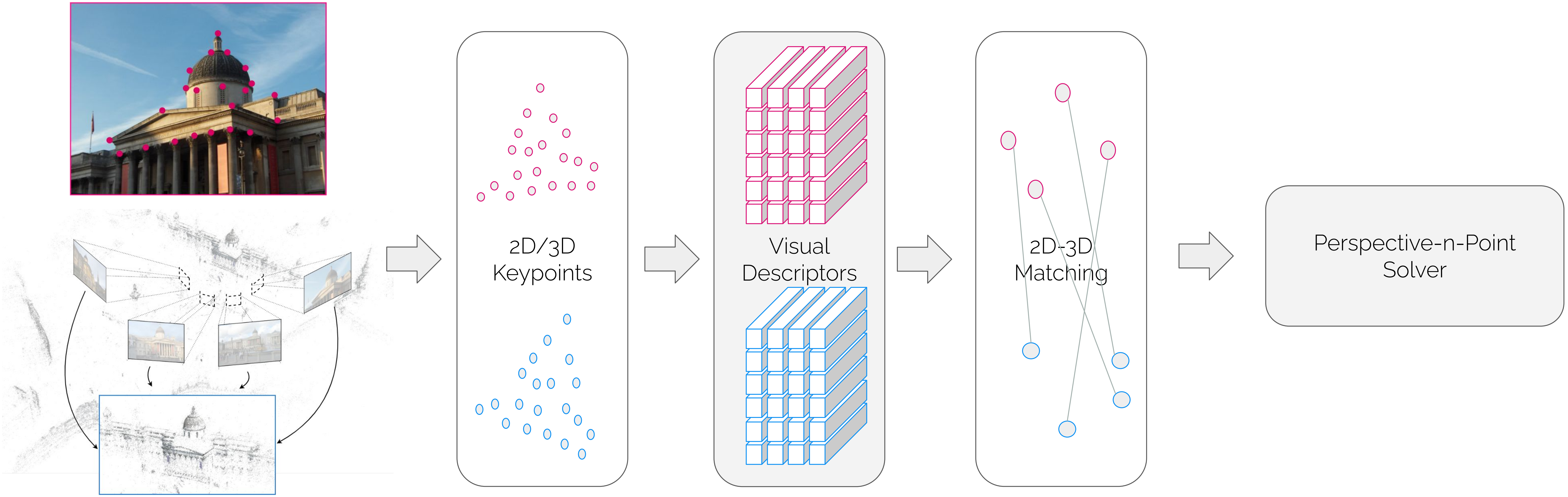




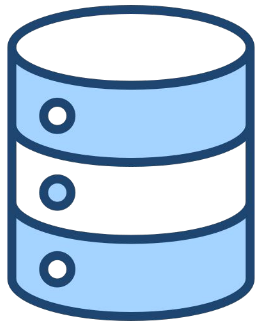
Localization System
Inputs
Method
Outputs
Query Image
Reference Images
3D Point Cloud
Absolute Pose Regression
Scene Coordinate Regression
PnP
Image Retrieval
Relative Pose-based
Relative Pose Est.
Triangulation
2D-3D Matching
PnP
Structure-base
Learning Involved

Structure-based Localization





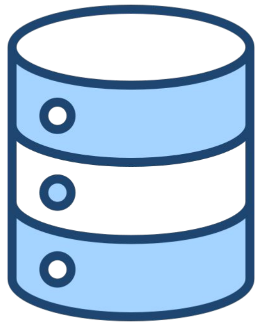
Practical Challenges



Storage Demand

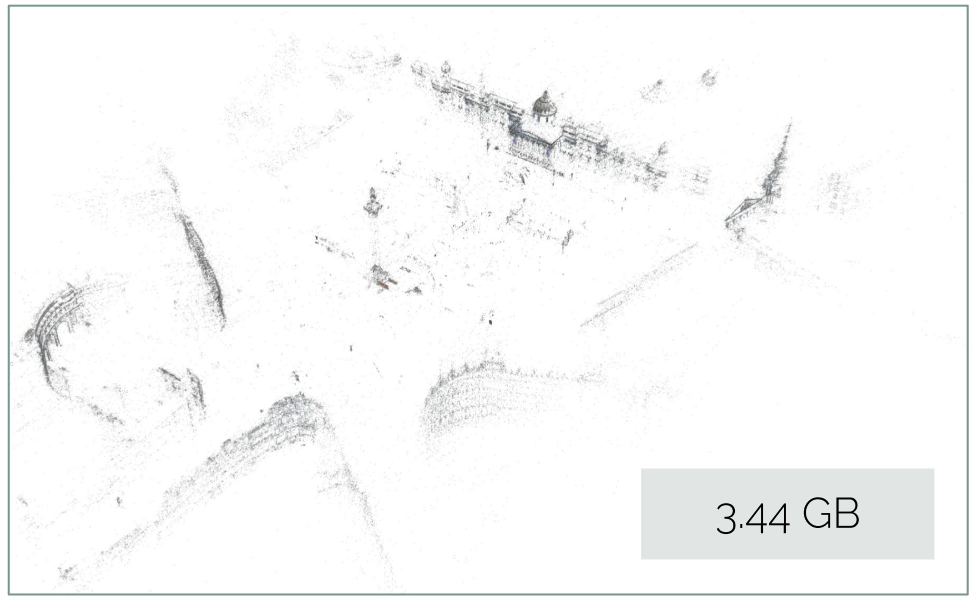
MegaDepth (192 scenes)	Camera	3D Points	Images	Point Descriptors		
				SIFT	CAPS	SuperPoint
Storage	15.73 MB	**3.44 GB**	157.84 GB	130.10 GB	520.38 GB	**1.041 TB**

Practical Challenges

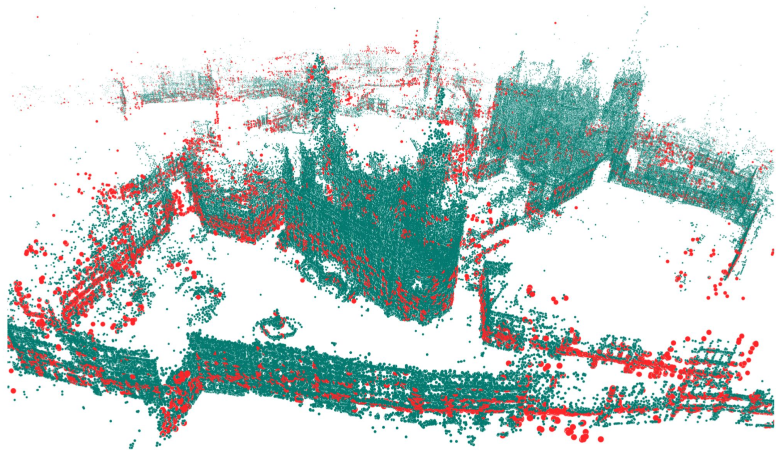


Storage Demand

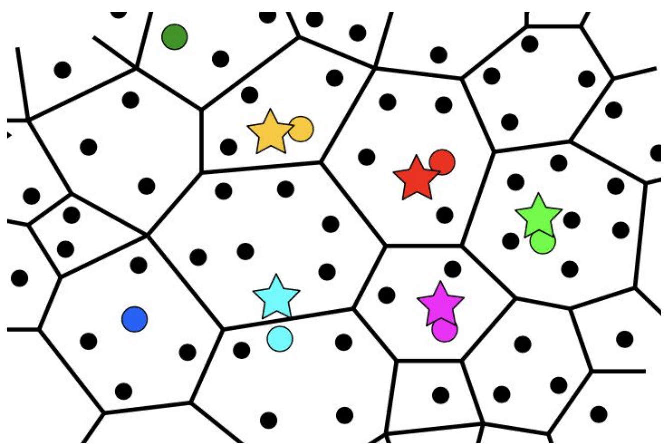
MegaDepth (192 scenes)	Camera	3D Points	Images	Point Descriptors		
				SIFT	CAPS	SuperPoint
Storage	15.73 MB	**3.44 GB**	157.84 GB	130.10 GB	520.38 GB	**1.041 TB**



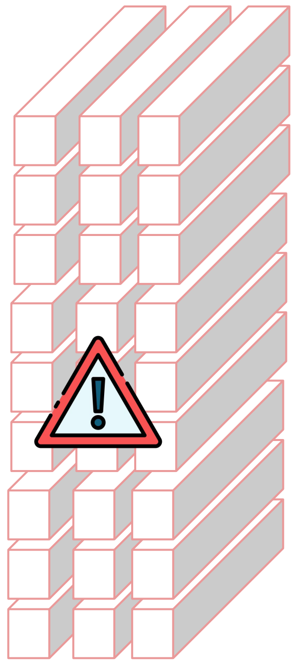
Scene Compression [1]



Desc Quantization [1, 2]



Scene Descriptor

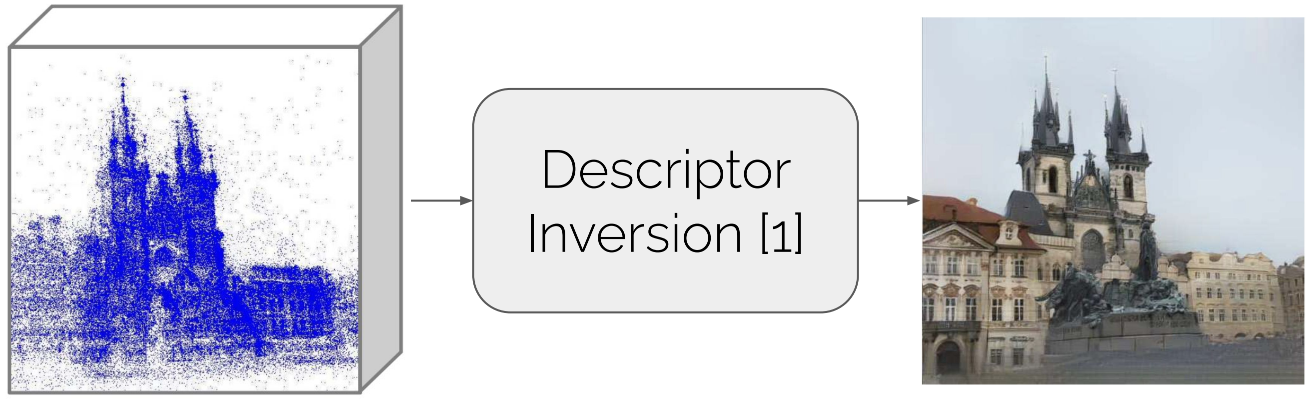


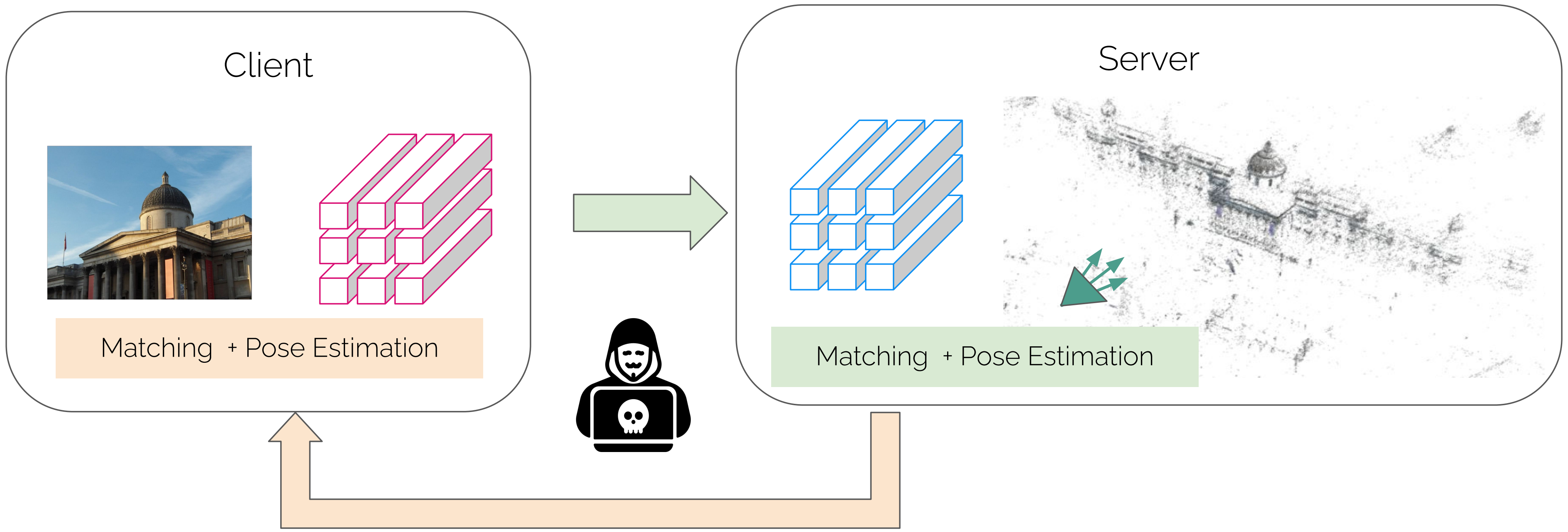
130 GB ~ 1.04 TB

[1] Camposeco, Federico, et al. "Hybrid scene compression for visual localization." CVPR19
[2] Sattler, Torsten, Bastian Leibe, and Leif Kobbelt. "Efficient & effective prioritized matching for large-scale image-based localization." PAMI16

Practical Challenges

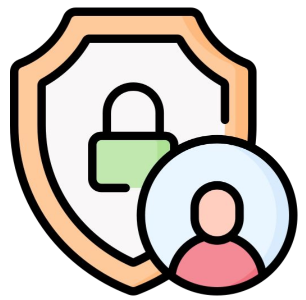
Privacy Risk



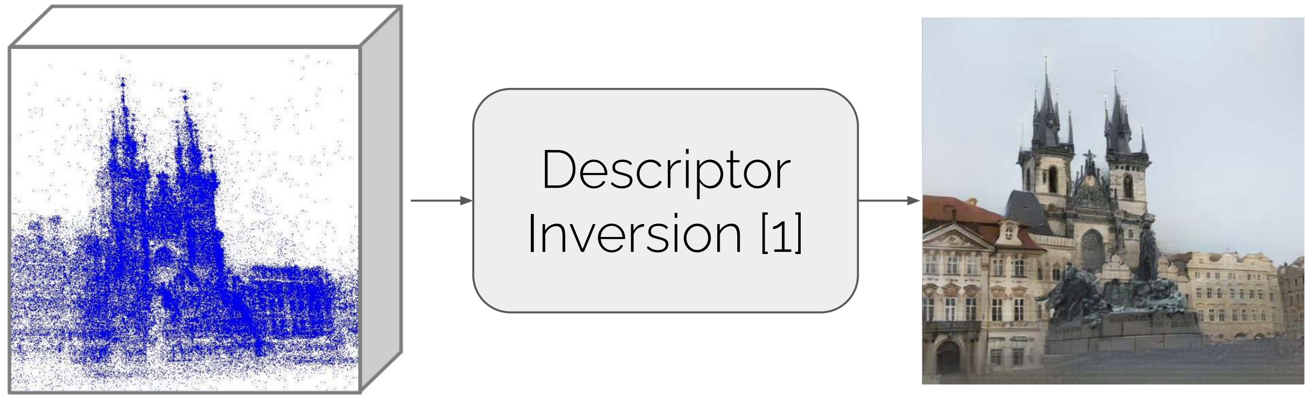


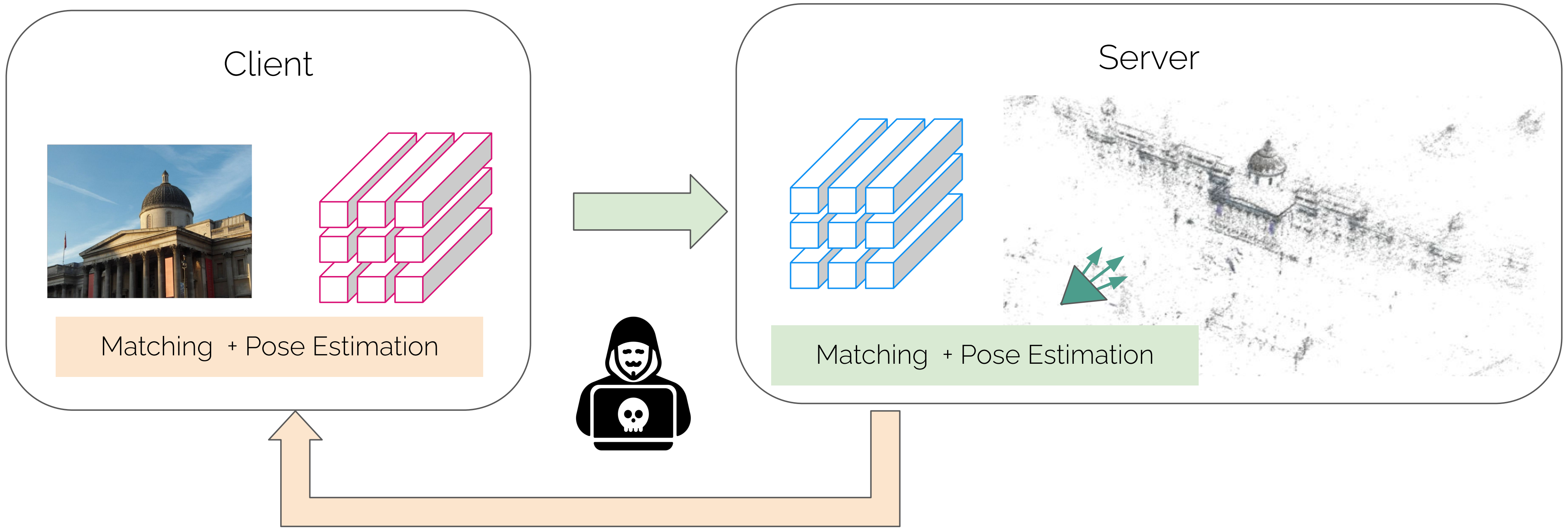
[1] Francesco, Pittaluga, et al Revealing Scenes by Inverting Structure From Motion Reconstructions. CVPR19

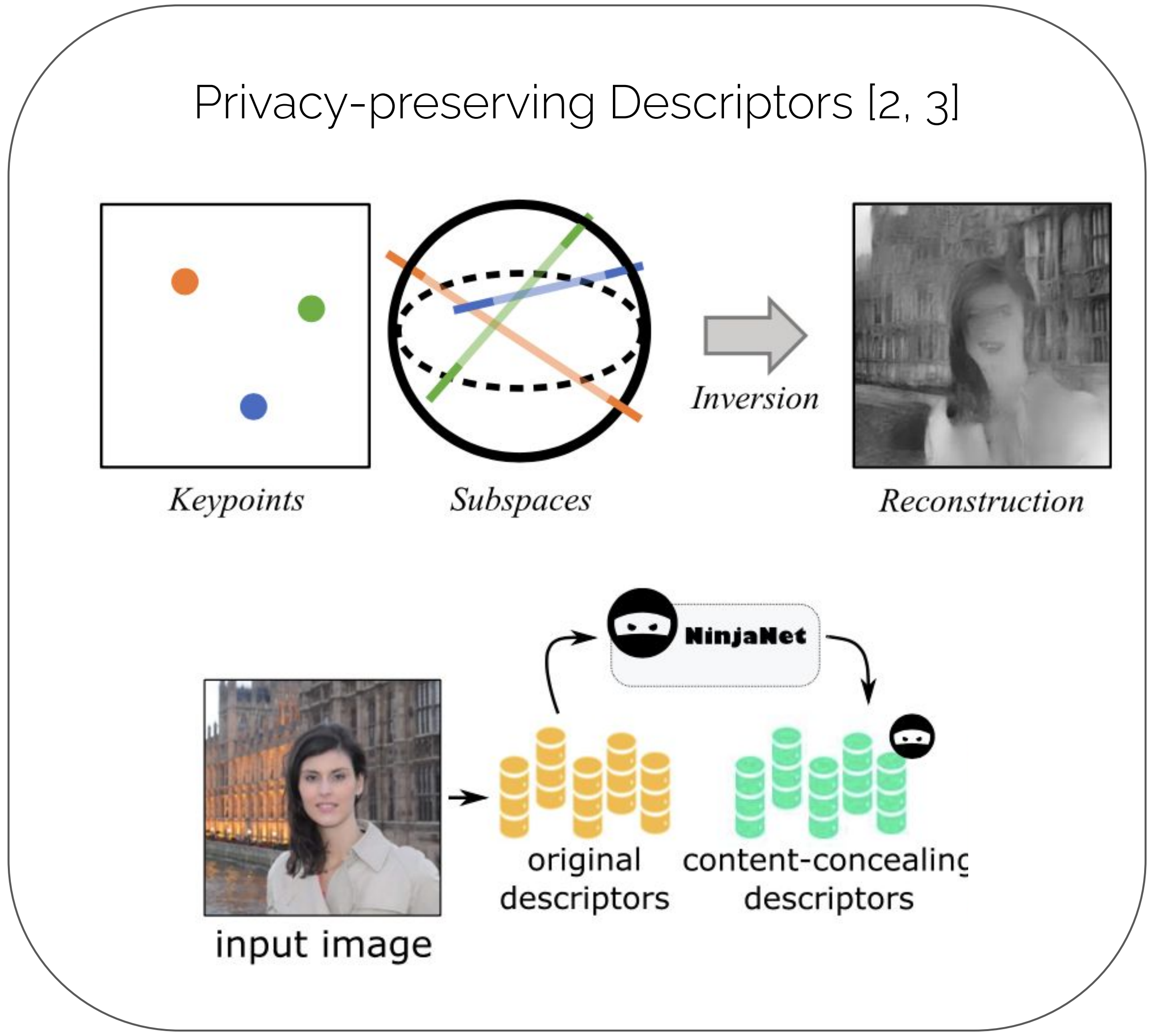
Practical Challenges



Privacy Risk



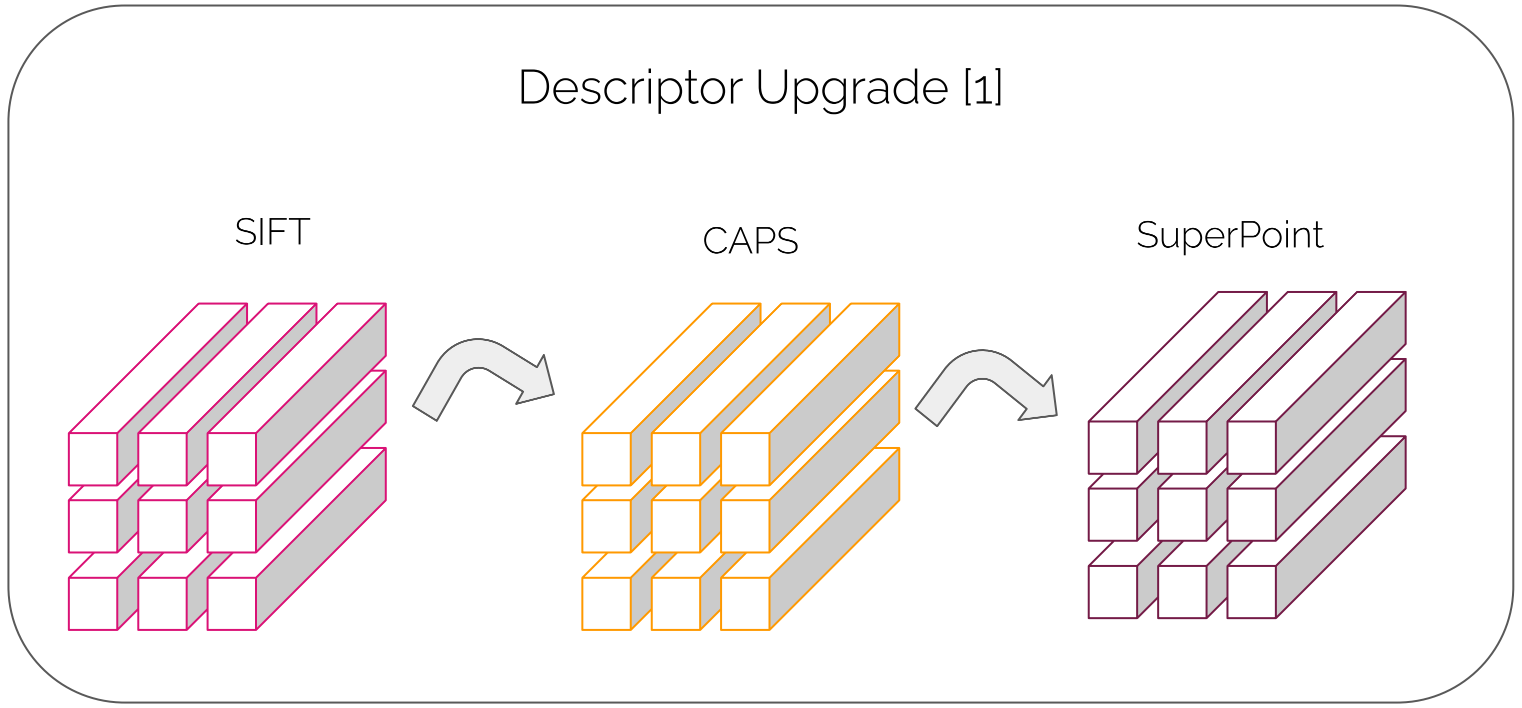


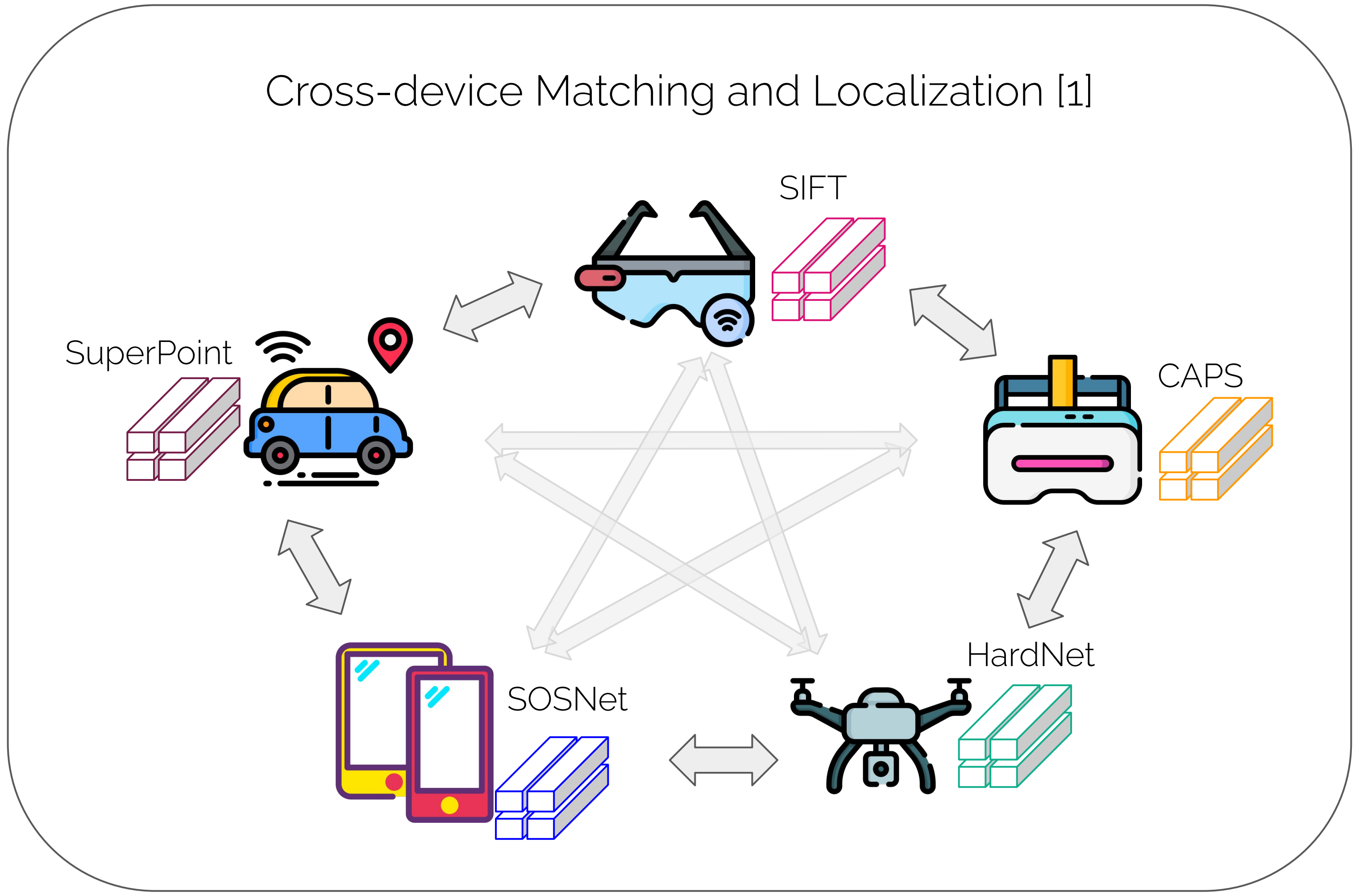


[1] Francesco, Pittaluga, et al Revealing Scenes by Inverting Structure From Motion Reconstructions. CVPR19
[2] Dusmanu, Mihai, et al. "Privacy-preserving image features via adversarial affine subspace embeddings." CVPR21.
[3] Ng, Tony, et al. "NinjaDesc: Content-Concealing Visual Descriptors via Adversarial Learning." CVPR22

Practical Challenges

Maintenance
Complexity

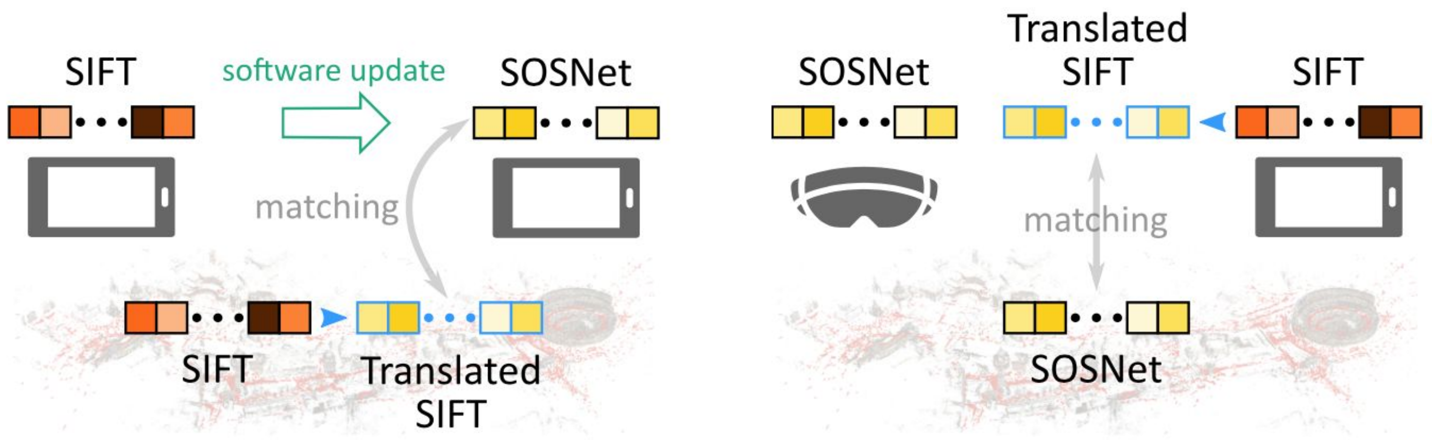


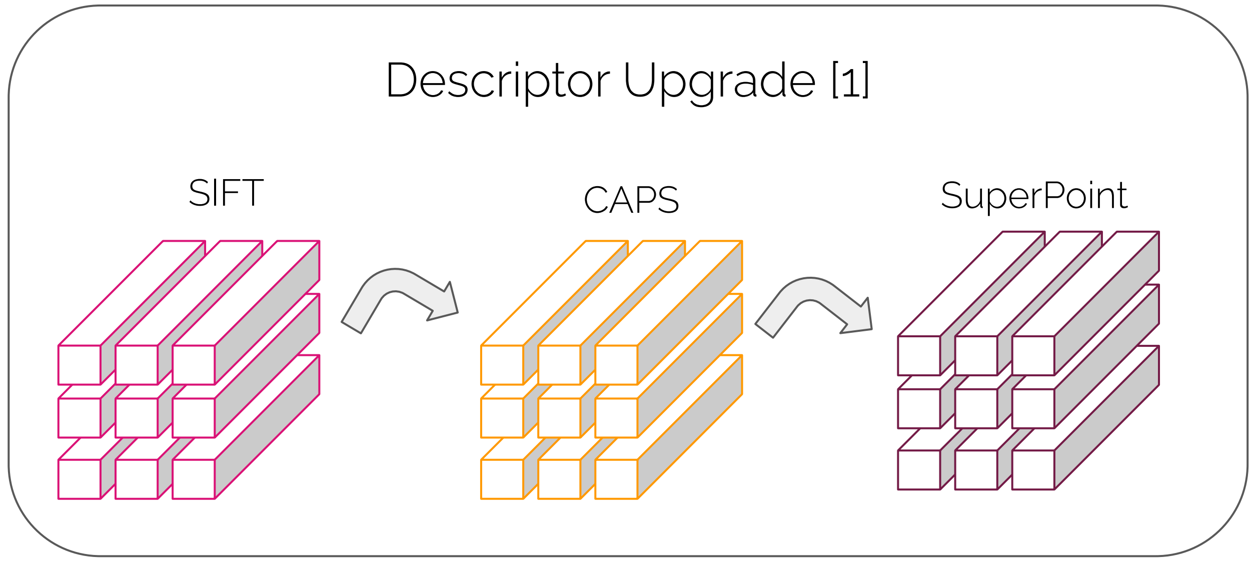


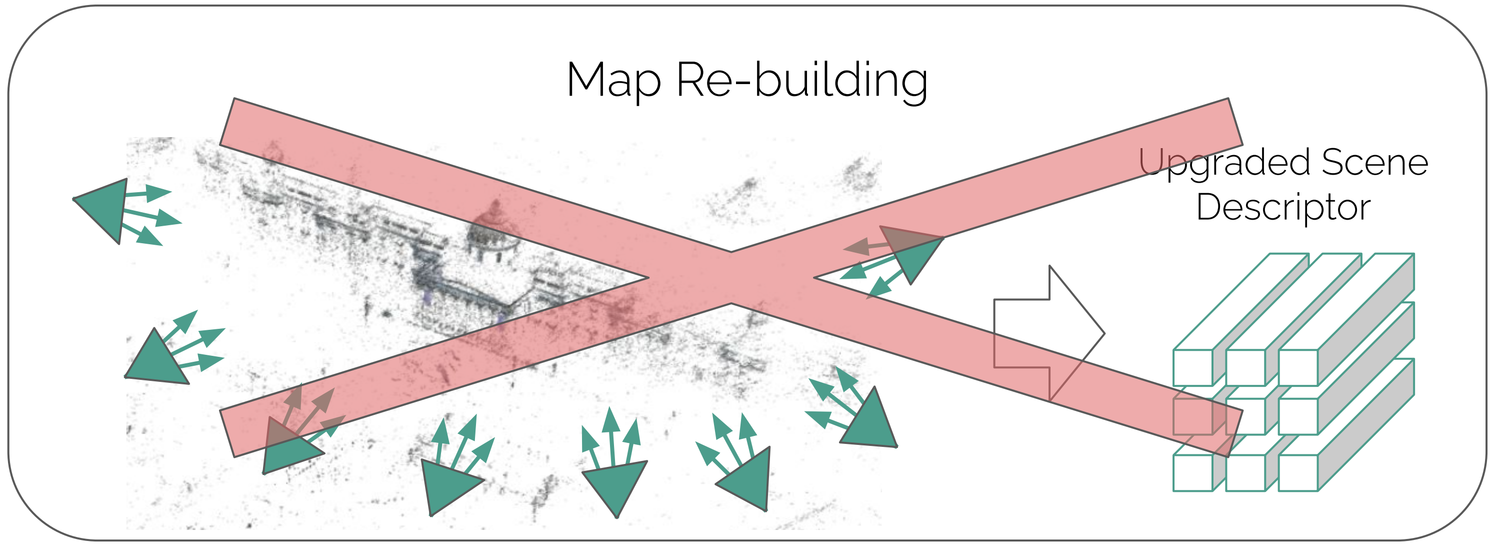
[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

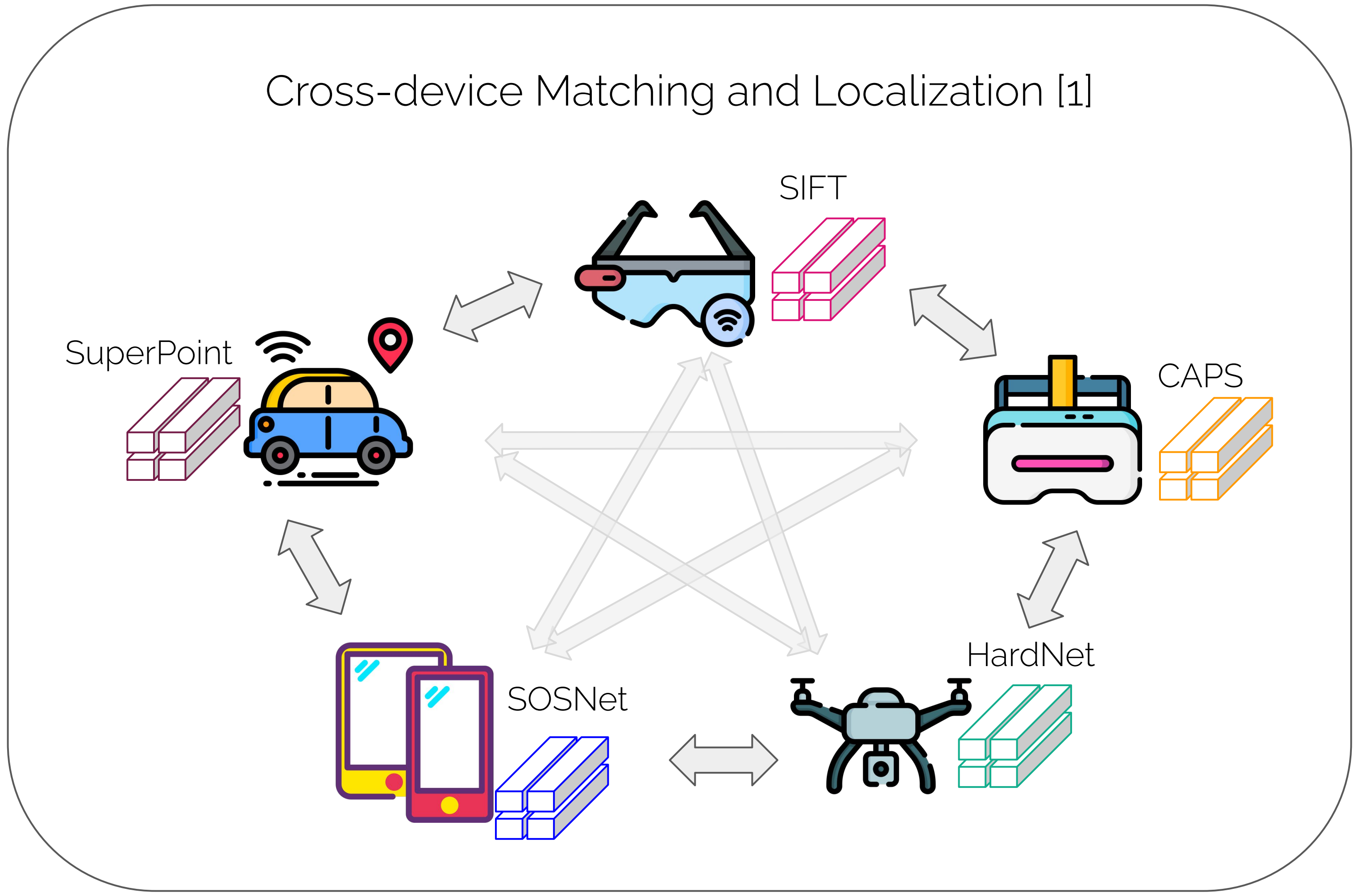
Practical Challenges

Maintenance Complexity



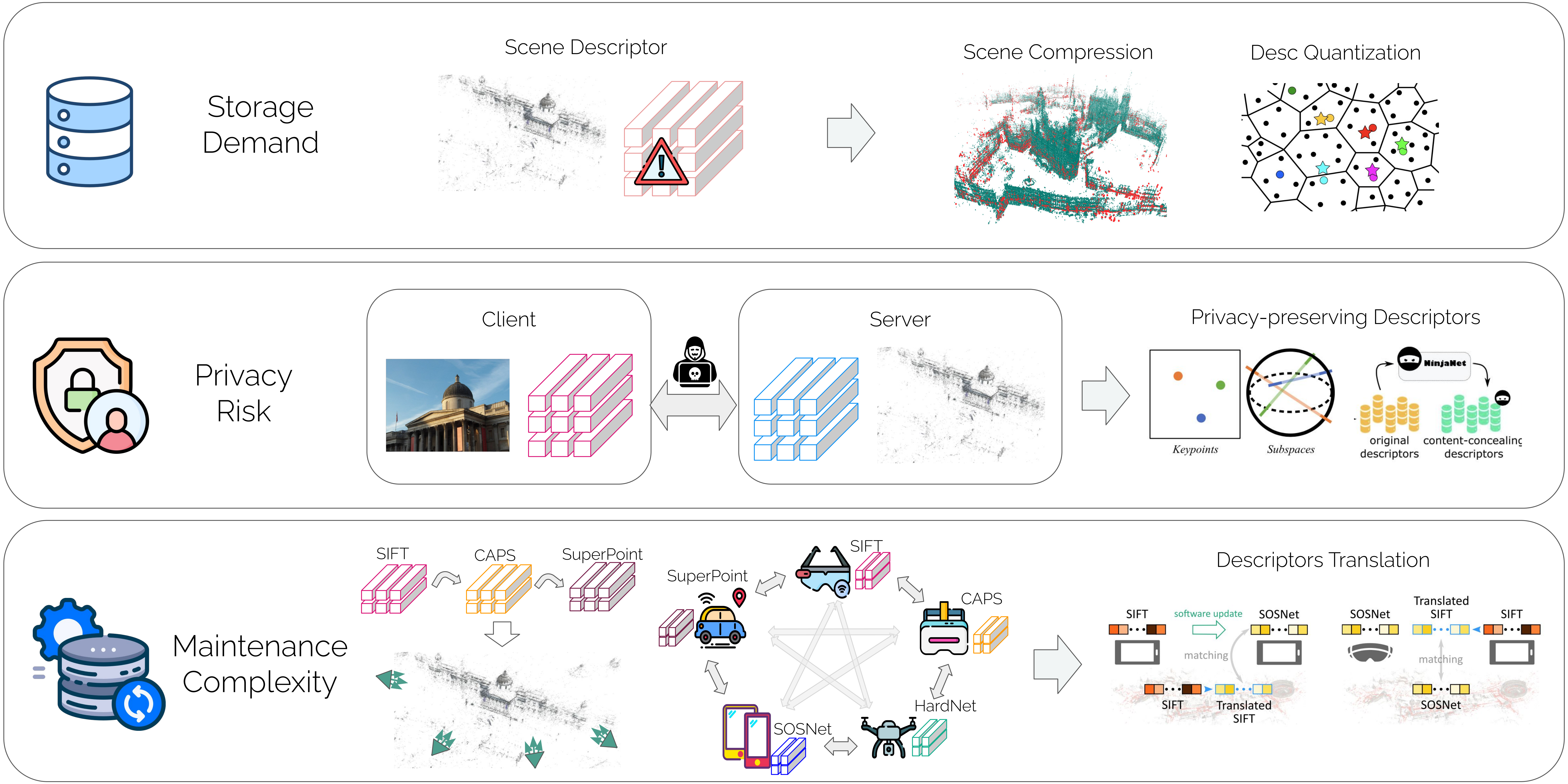




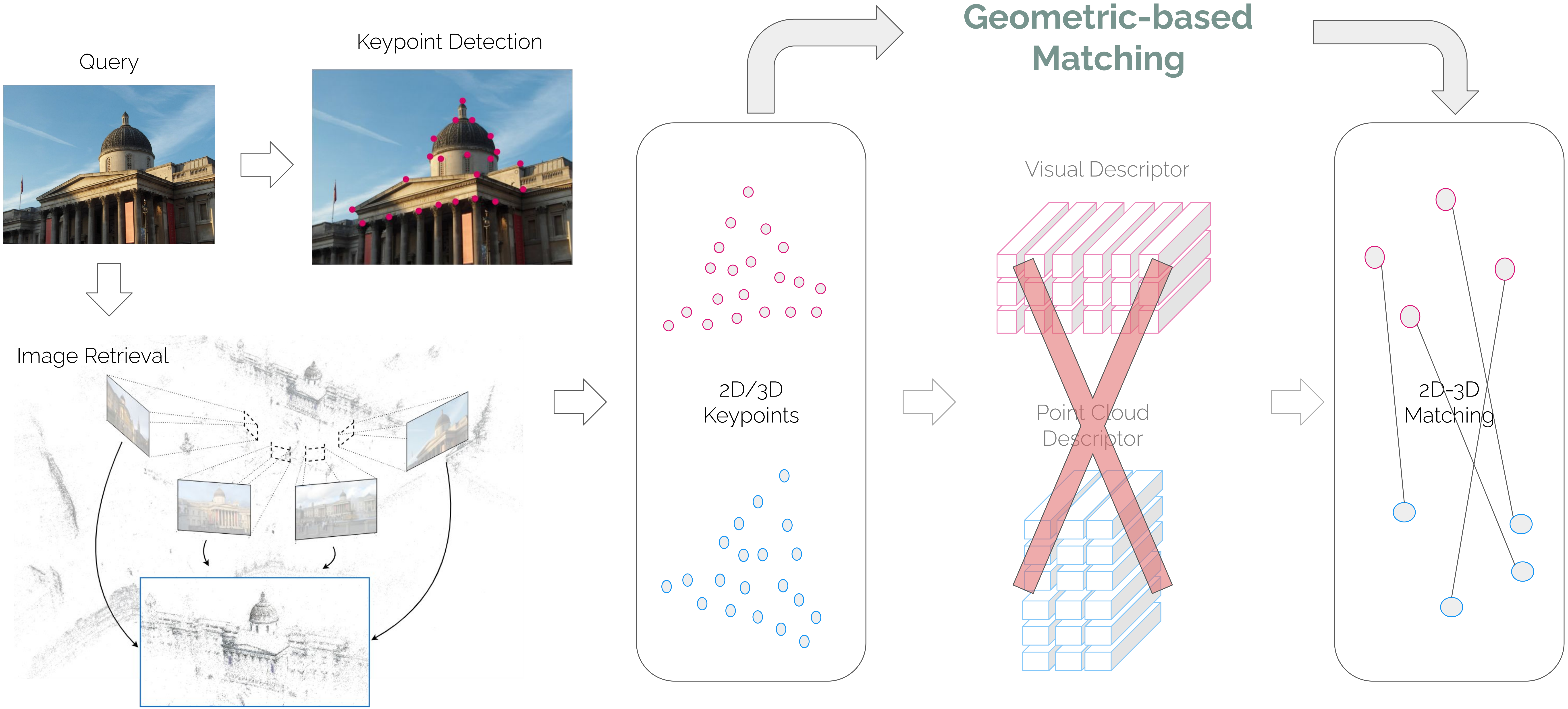


[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

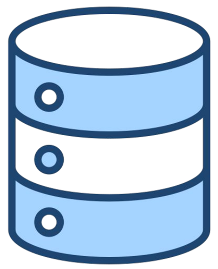
Practical Challenges



Geometric-based Matching

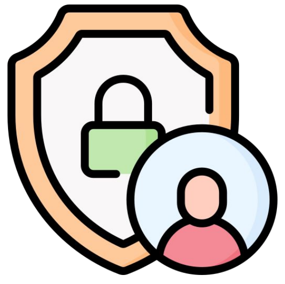


Geometric-based Matching

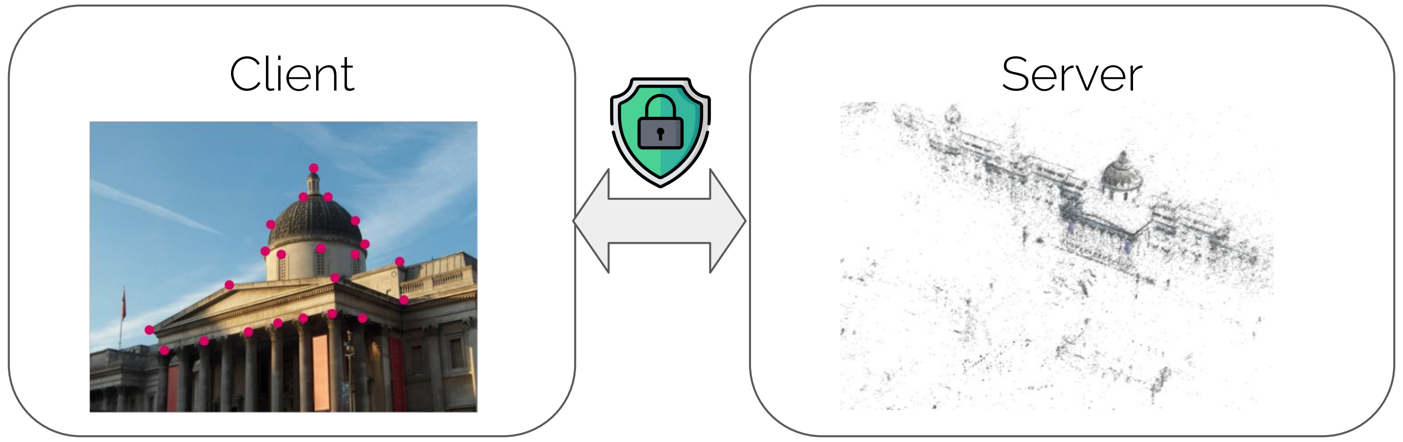


Storage
Demand

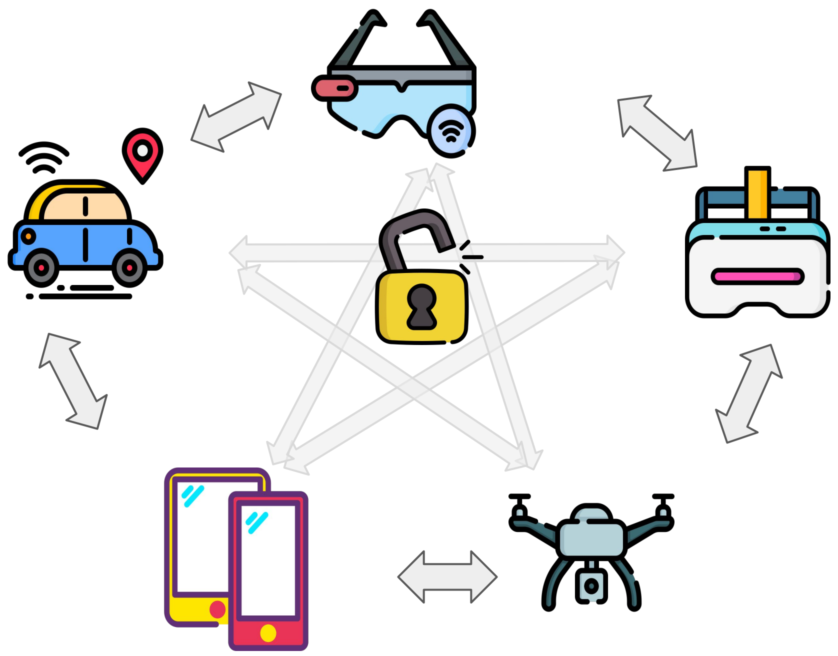
Minimal
Representation



Privacy
Risk



Maintenance
Complexity

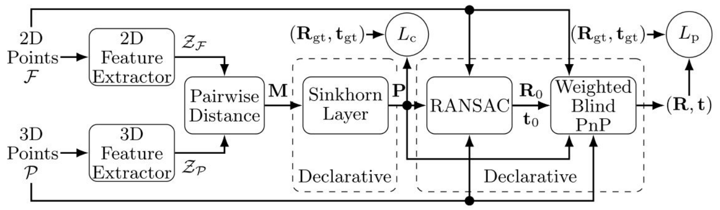


Scalable Large-scale Localization

- Low Storage
- Privacy Preserving
- No Descriptor Maintenance

Geometric-based matching and pose estimation

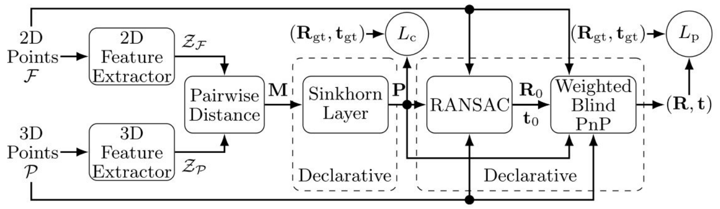
BPnPNet [4]
- Learning-based
- Declarative layers
- Degrades with outliers.

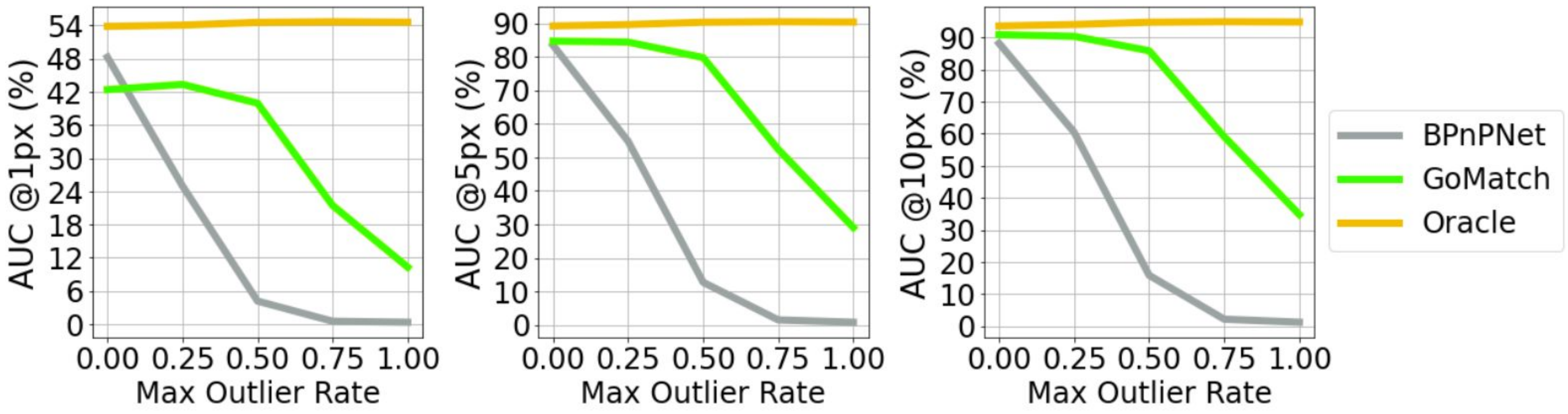


Geometric-based matching and pose estimation

Does not scale to real-world
localization settings!

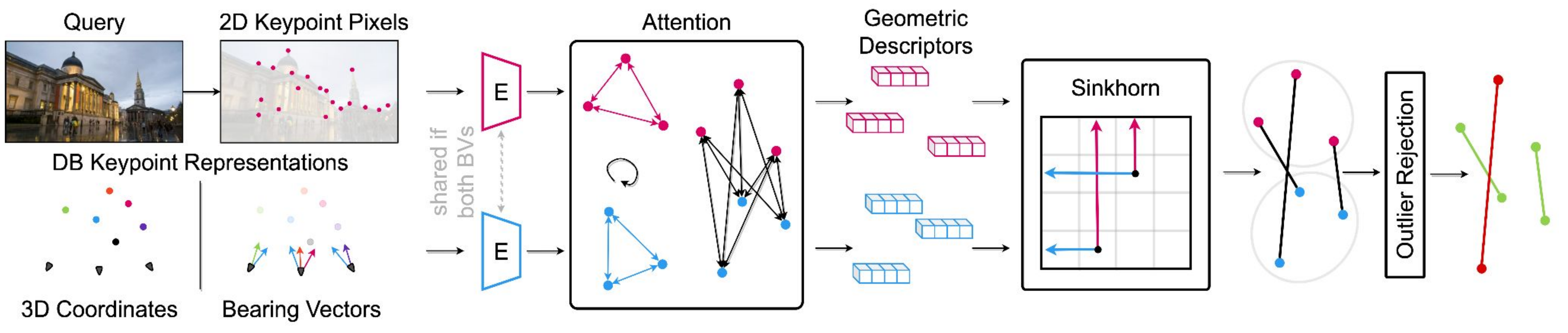
BPnPNet [4]
- Learning-based
- Declarative layers
- Degrades with outliers.





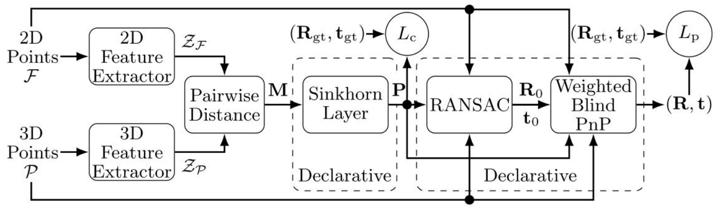
Geometric-based matching and pose estimation

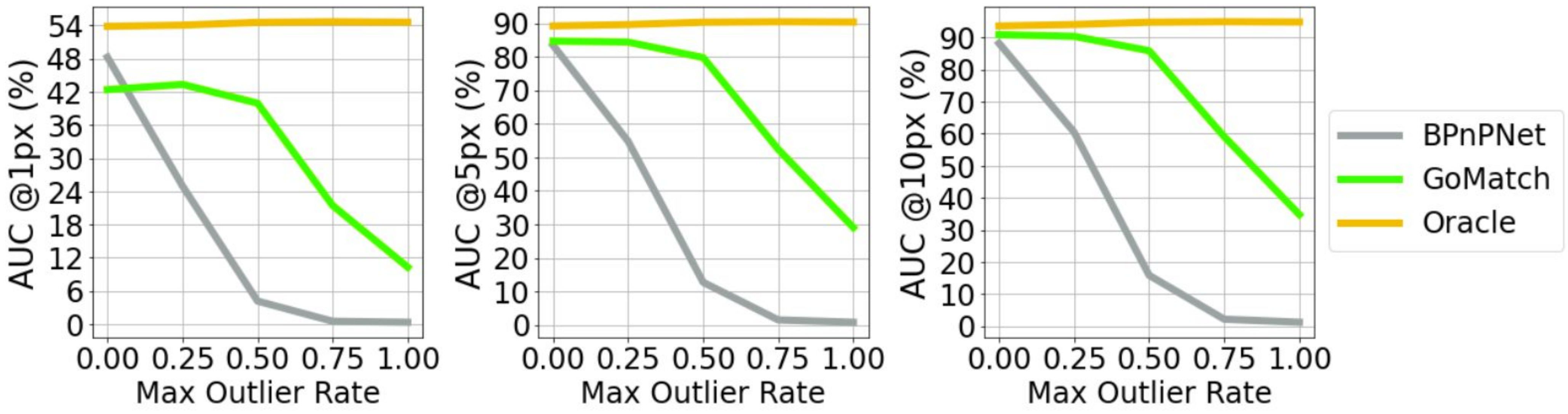
GoMatch



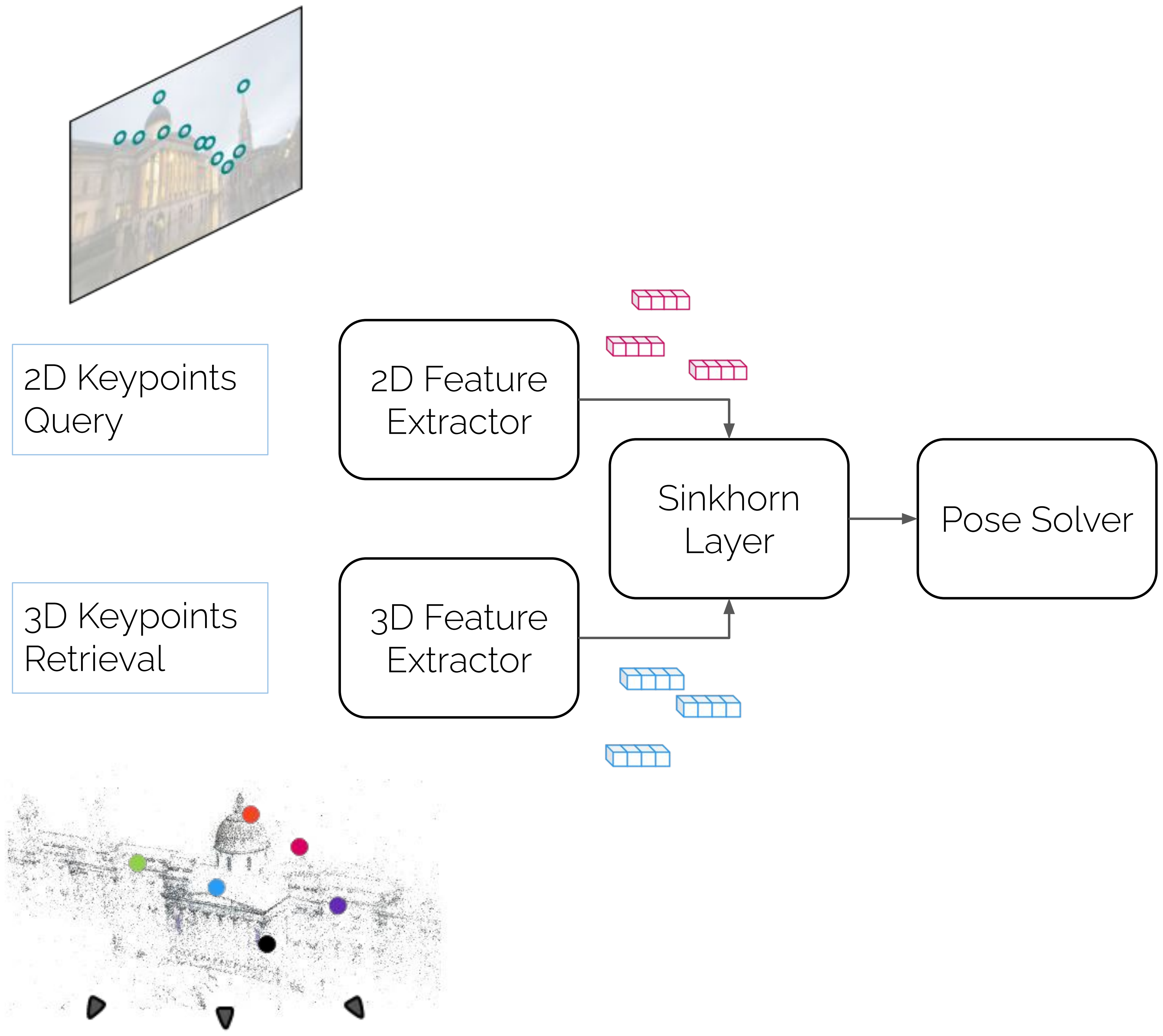
BPnPNet [4]

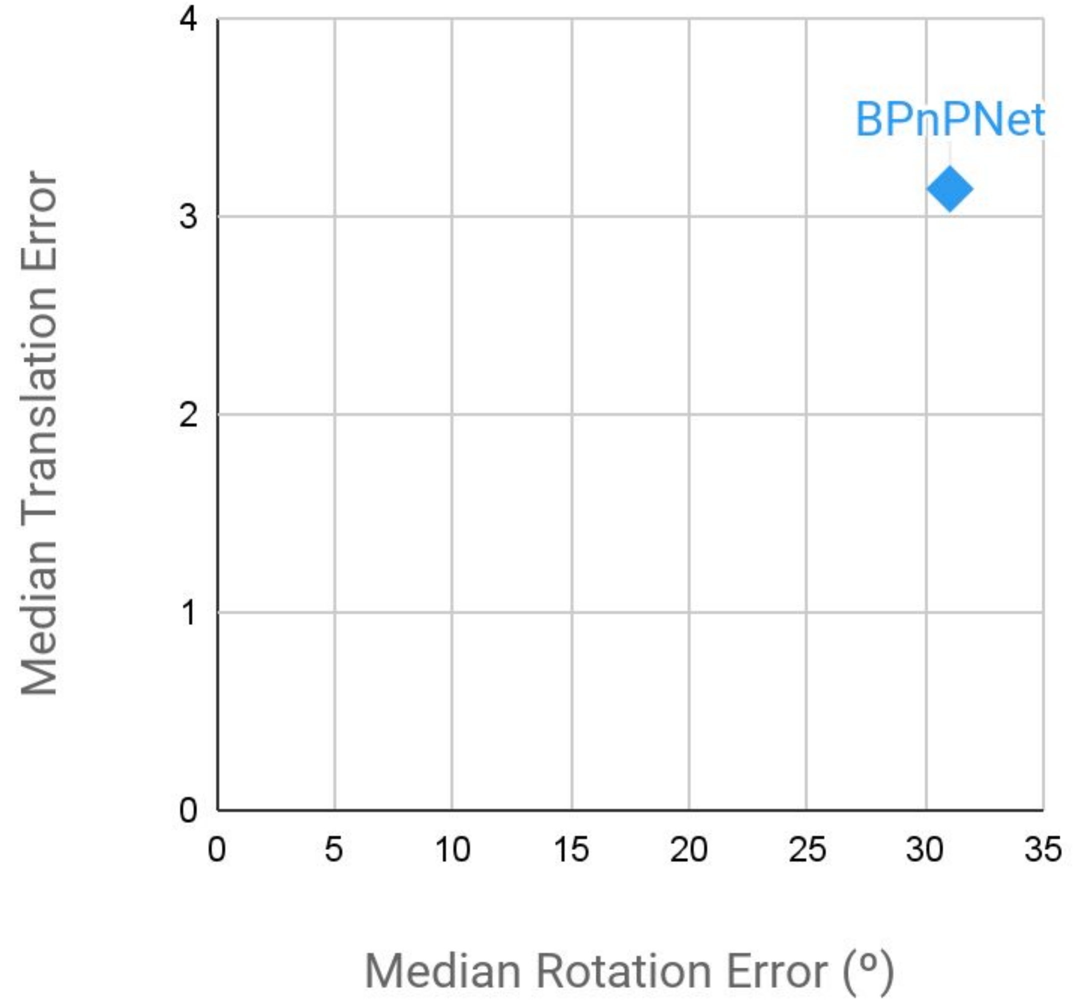
- Learning-based
- Declarative layers
- Degrades with outliers.



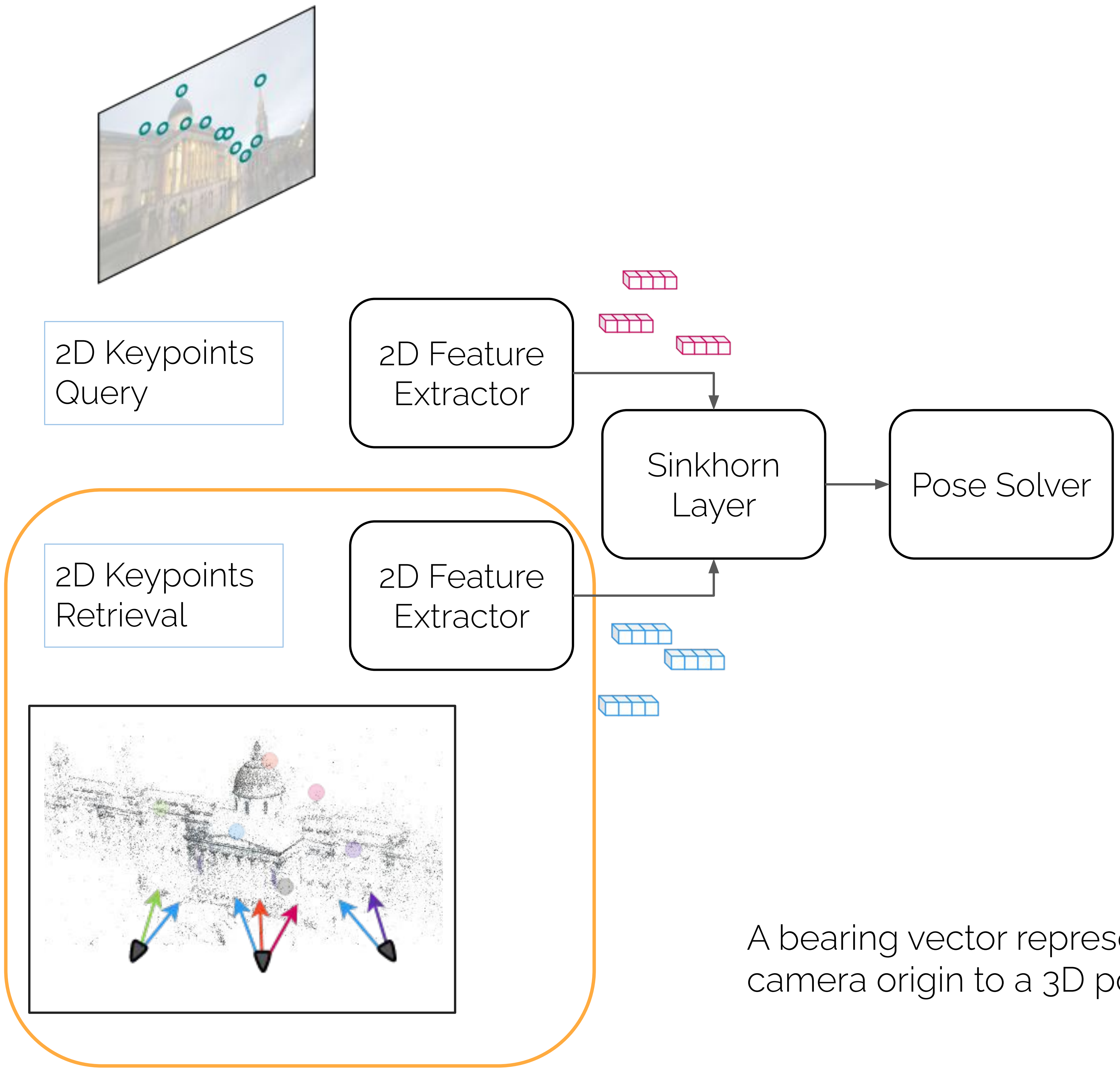


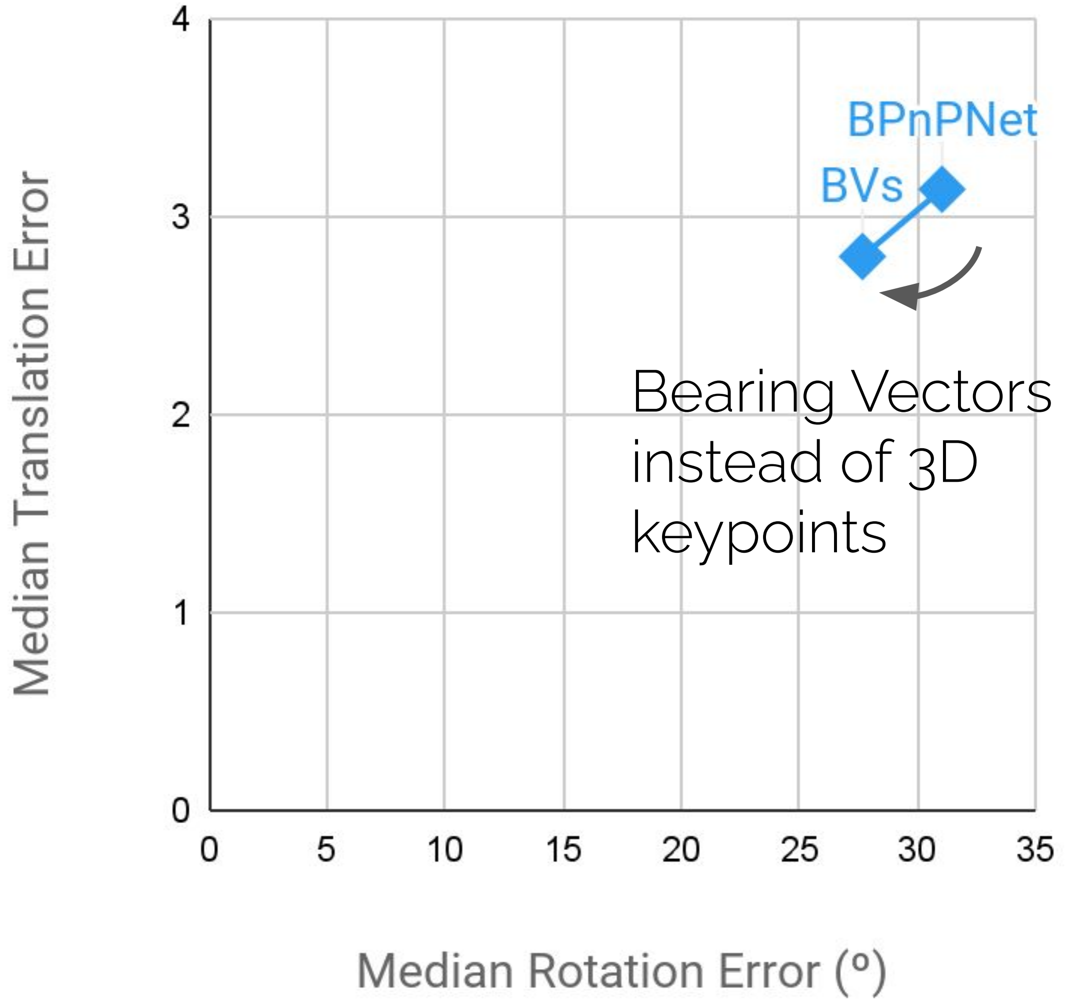
GoMatch Step-by-Step





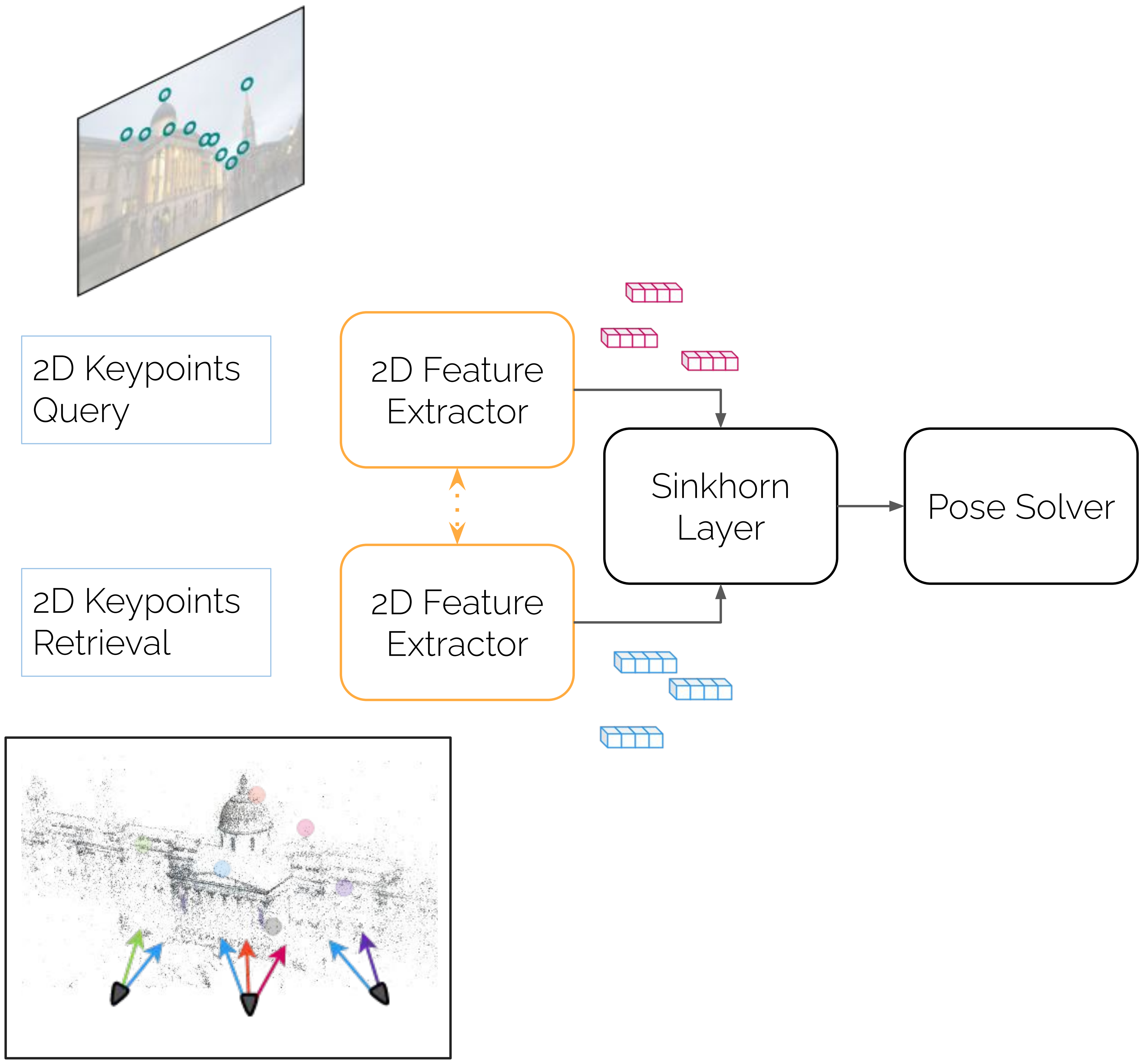
GoMatch Step-by-Step

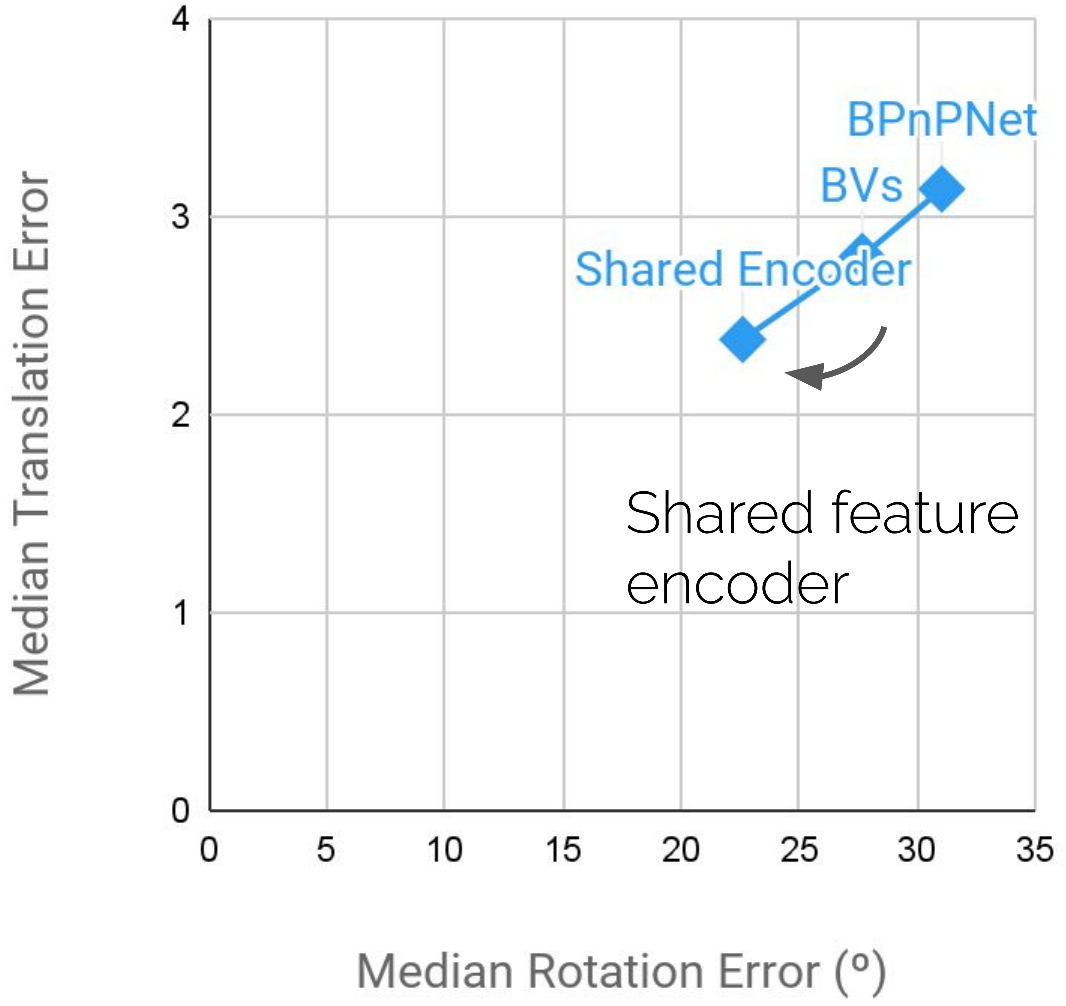




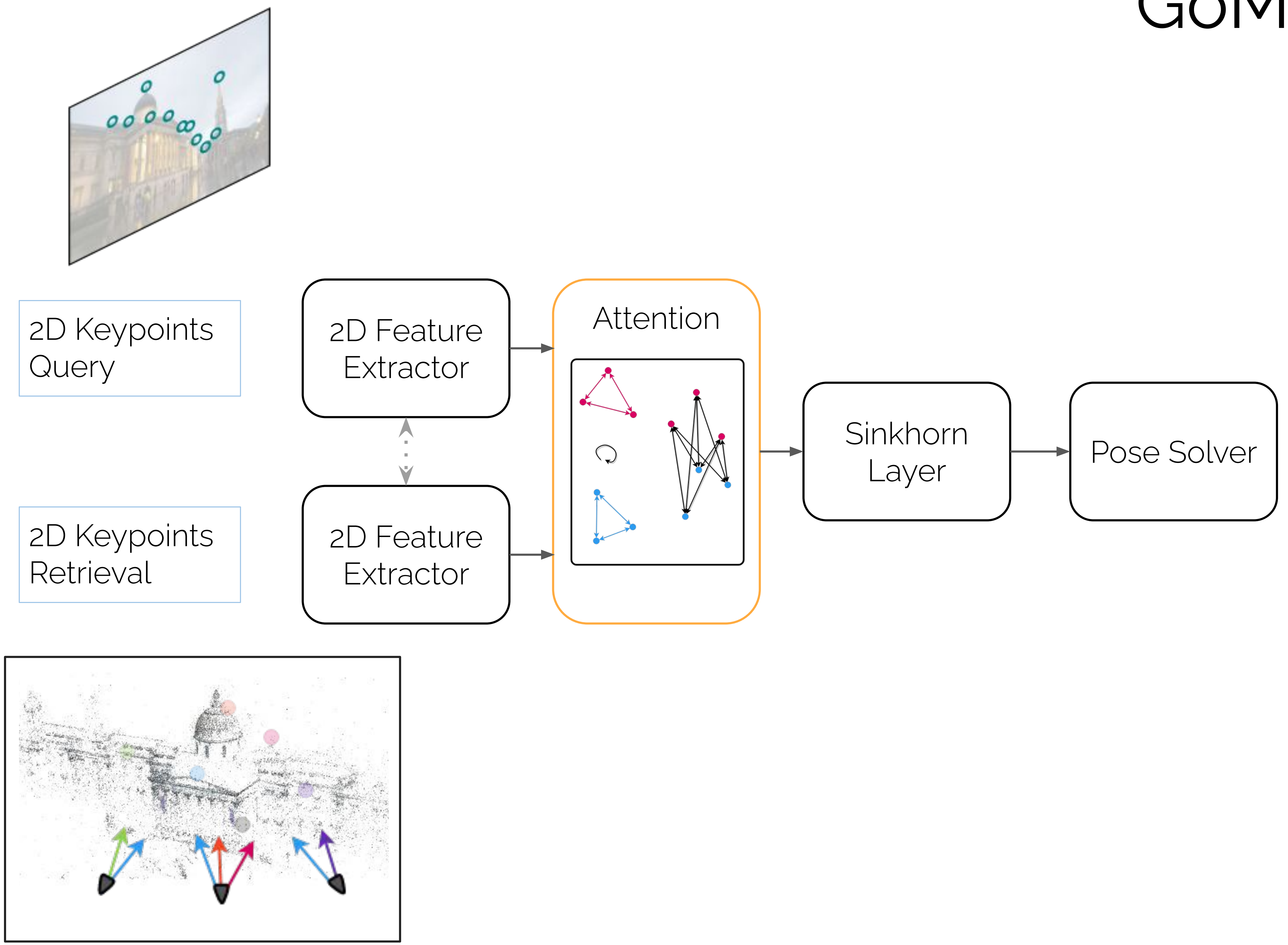
A bearing vector represents the direction from the reference camera origin to a 3D point in normalized coordinates.

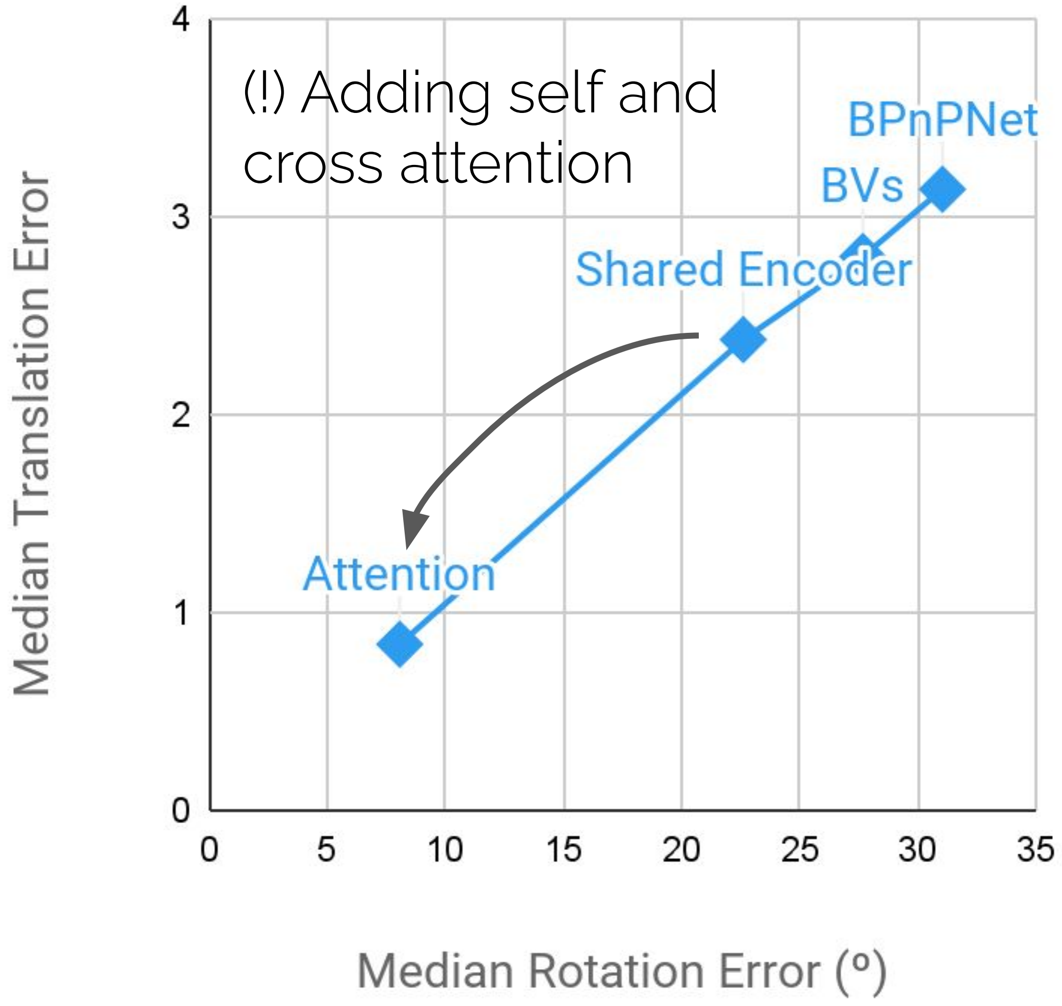
GoMatch Step-by-Step



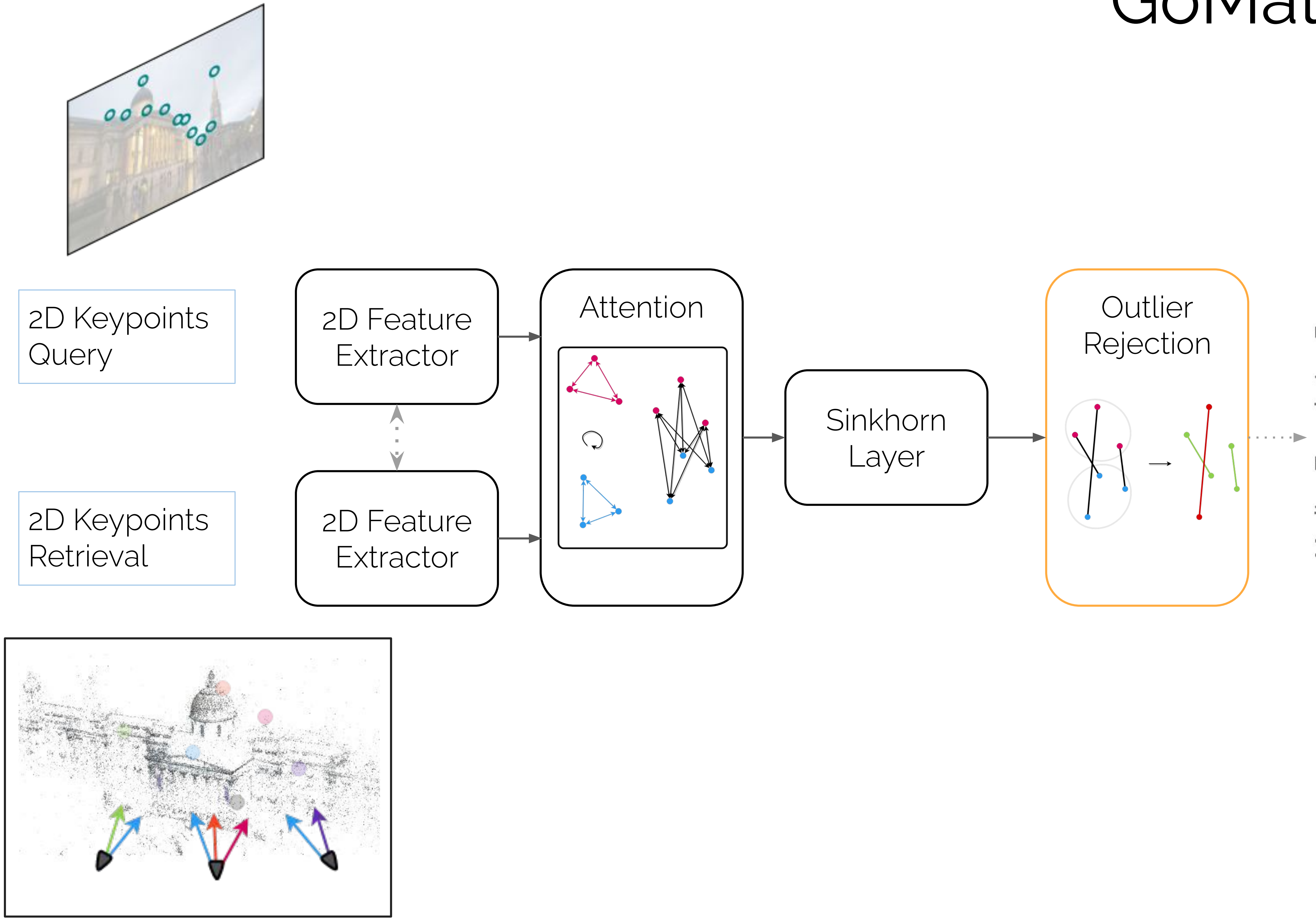


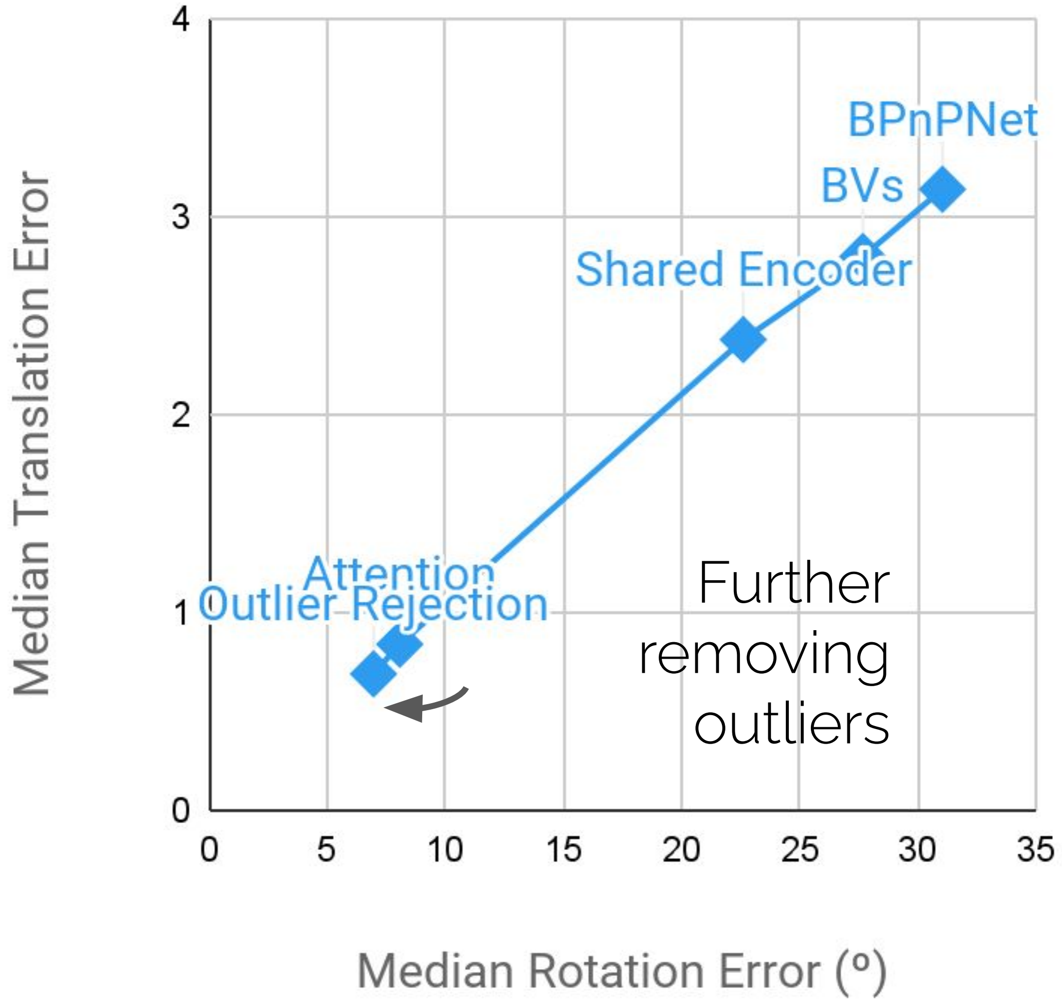
GoMatch Step-by-Step



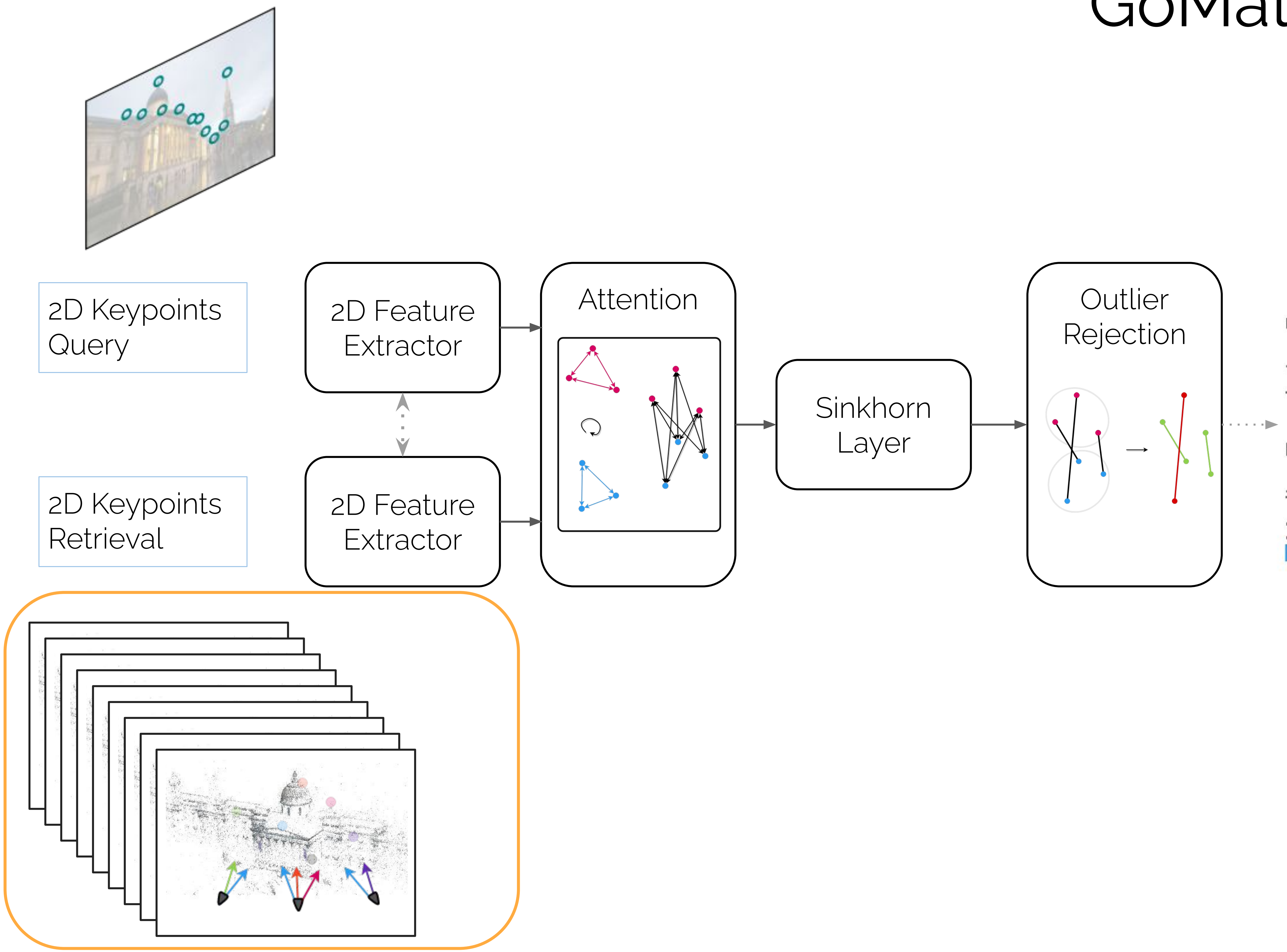


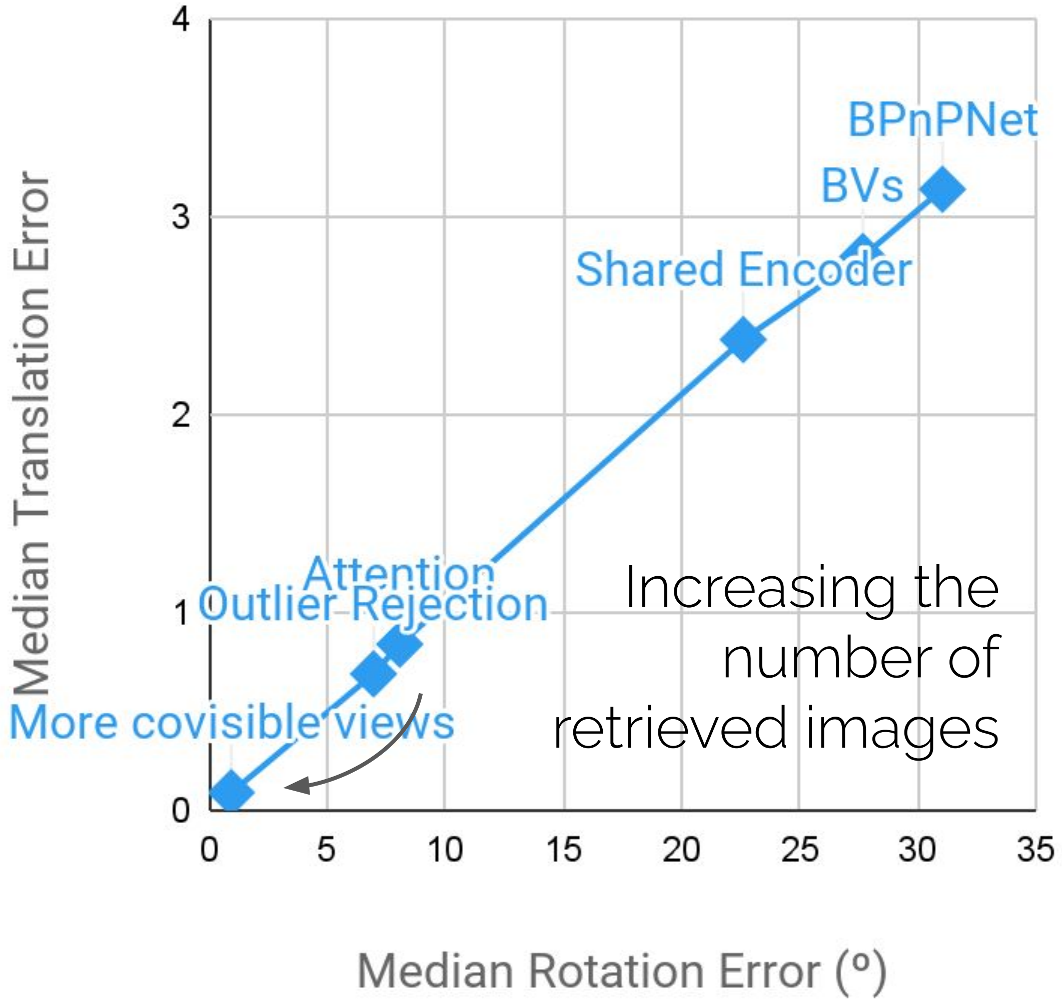
GoMatch Step-by-Step





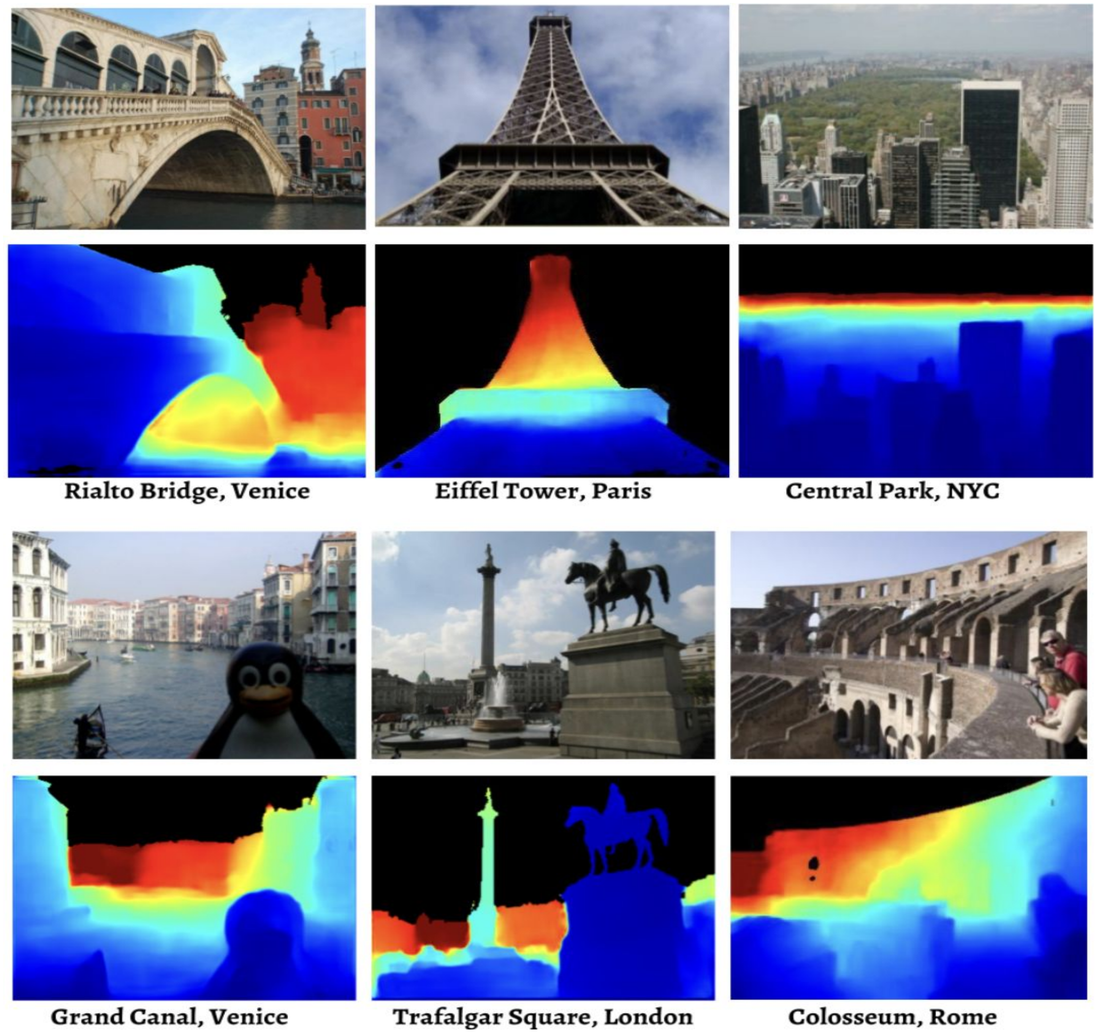
GoMatch Step-by-Step





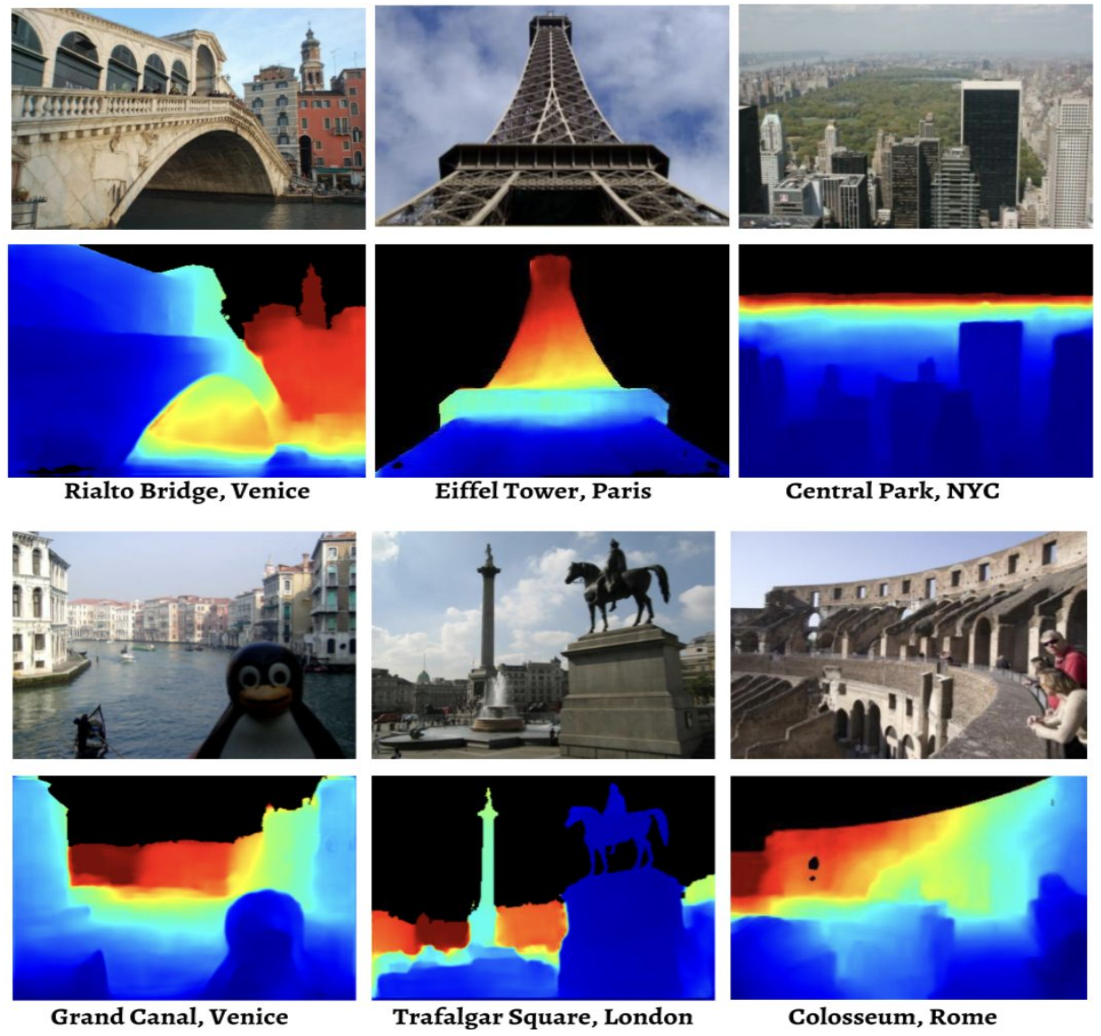
Generalization: outdoor/indoor and keypoints

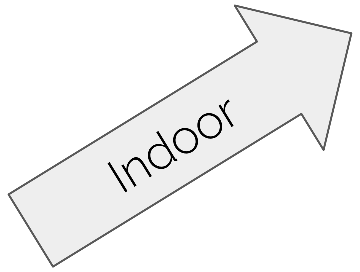
MegaDepth
(Outdoor w.SIFT)

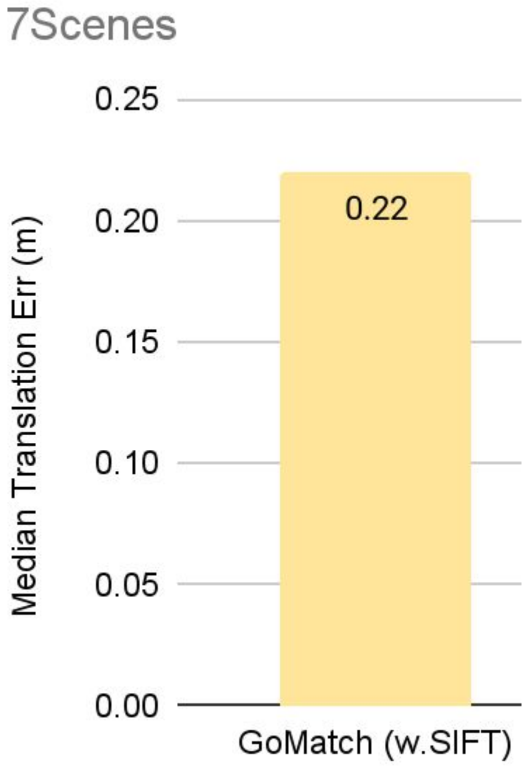


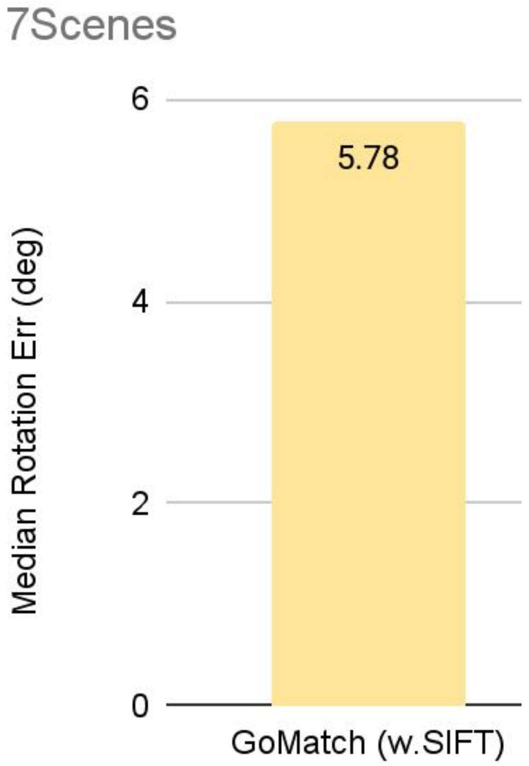
Generalization: outdoor/indoor and keypoints

MegaDepth
(Outdoor w.SIFT)



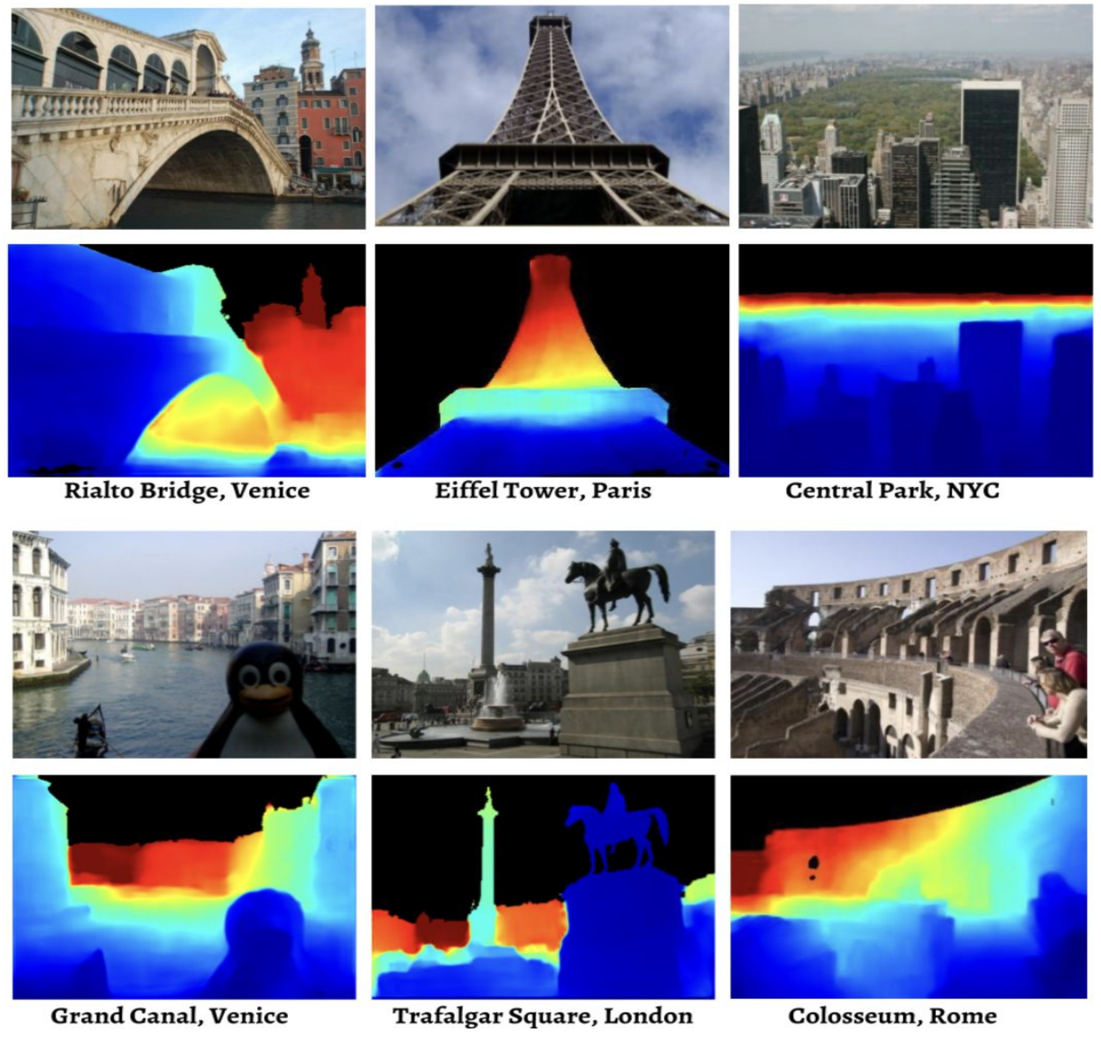




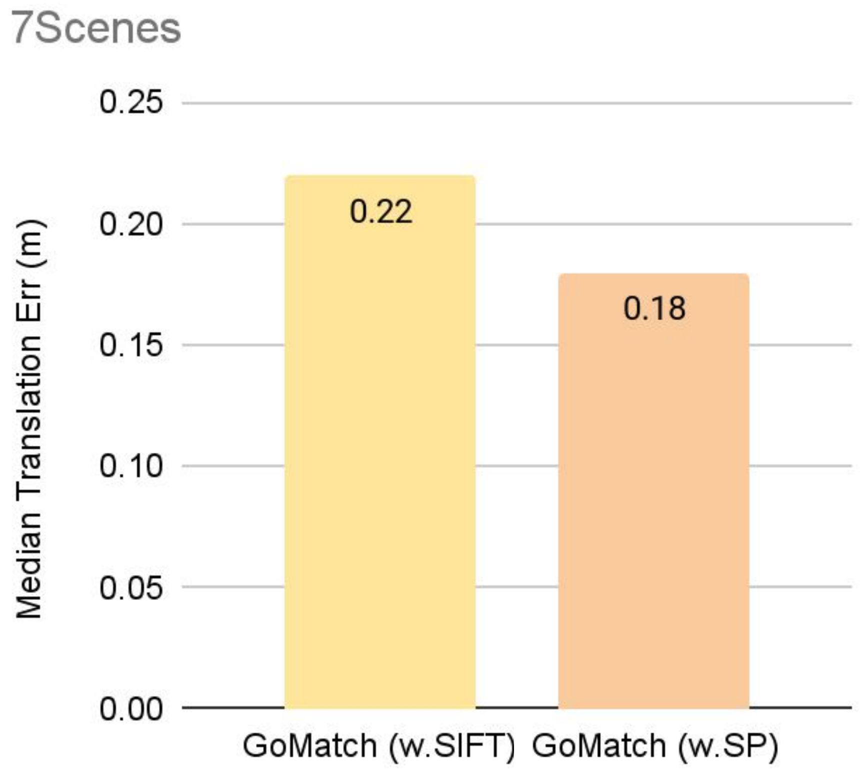


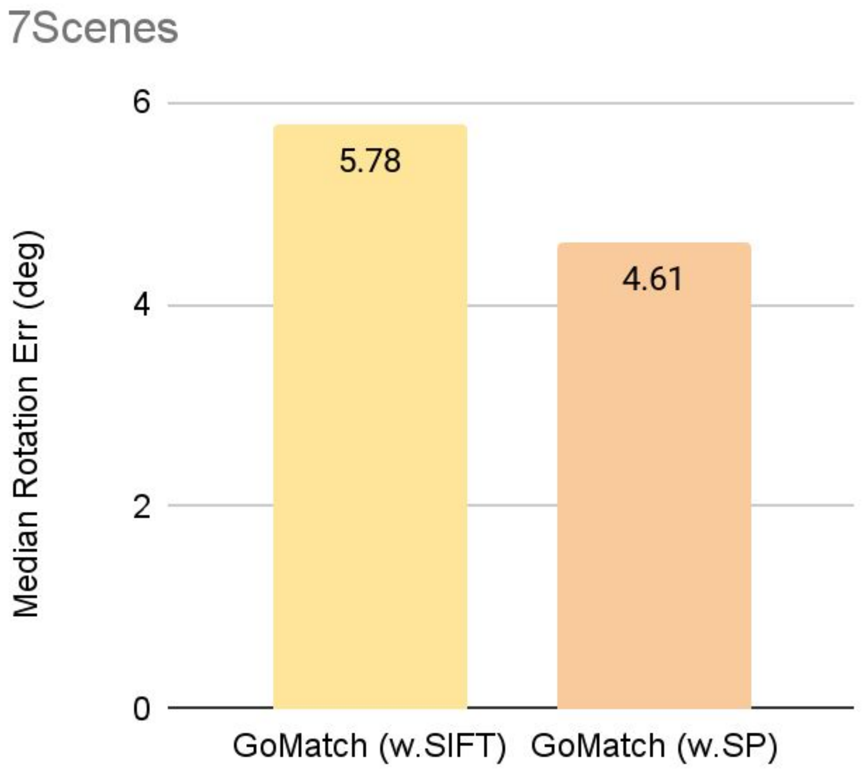
Generalization: outdoor/indoor and keypoints

MegaDepth
(Outdoor w.SIFT)

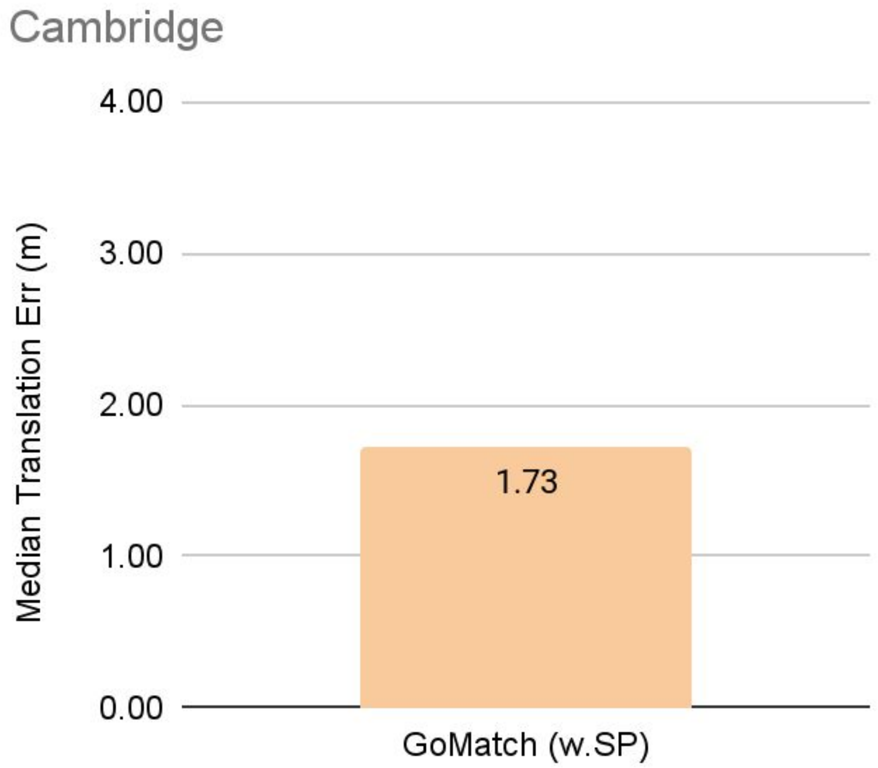


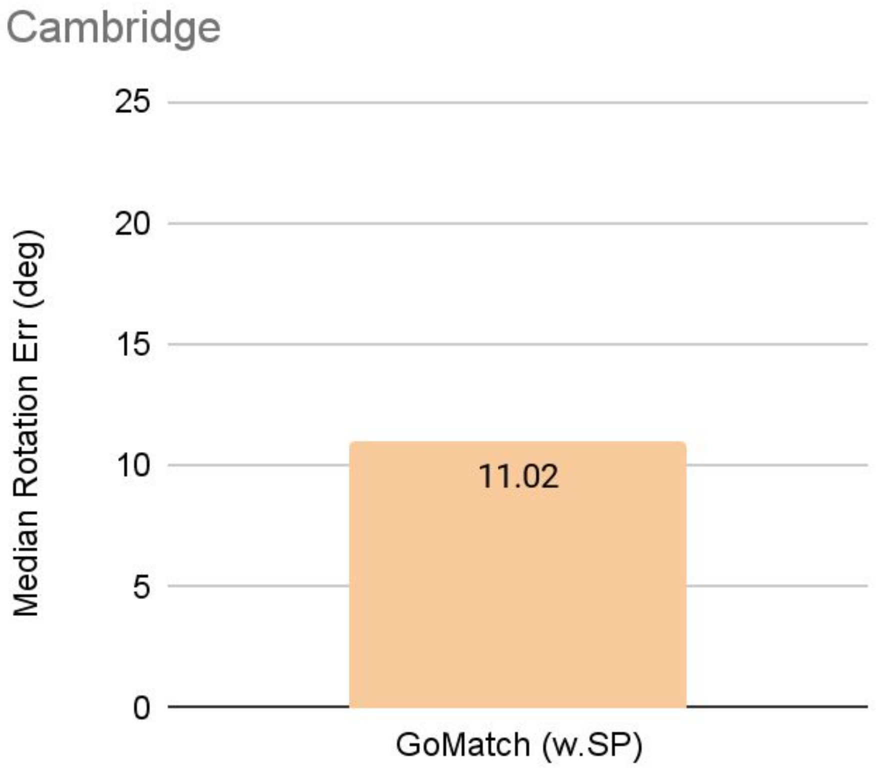
Indoor



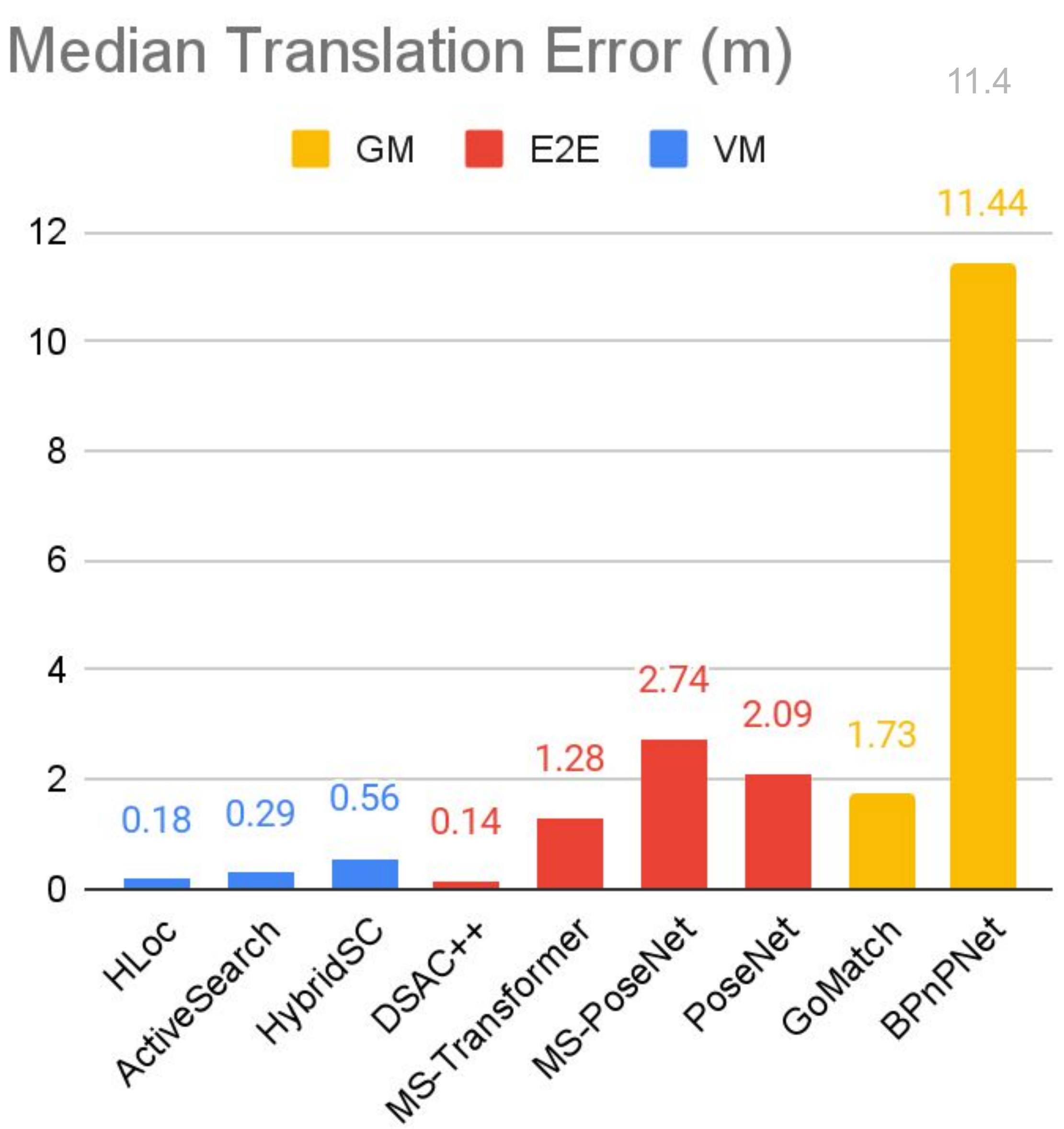


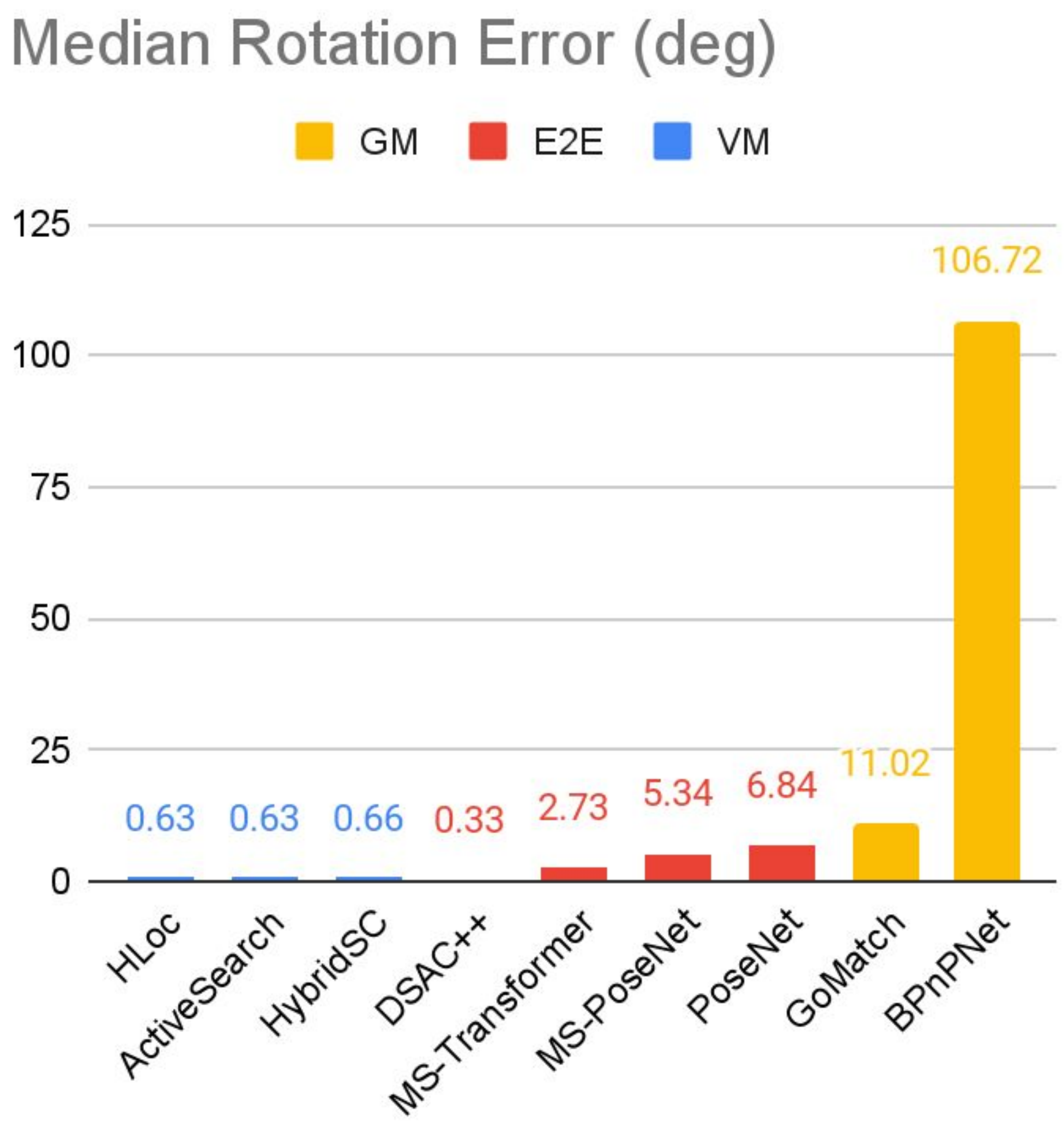
Outdoor



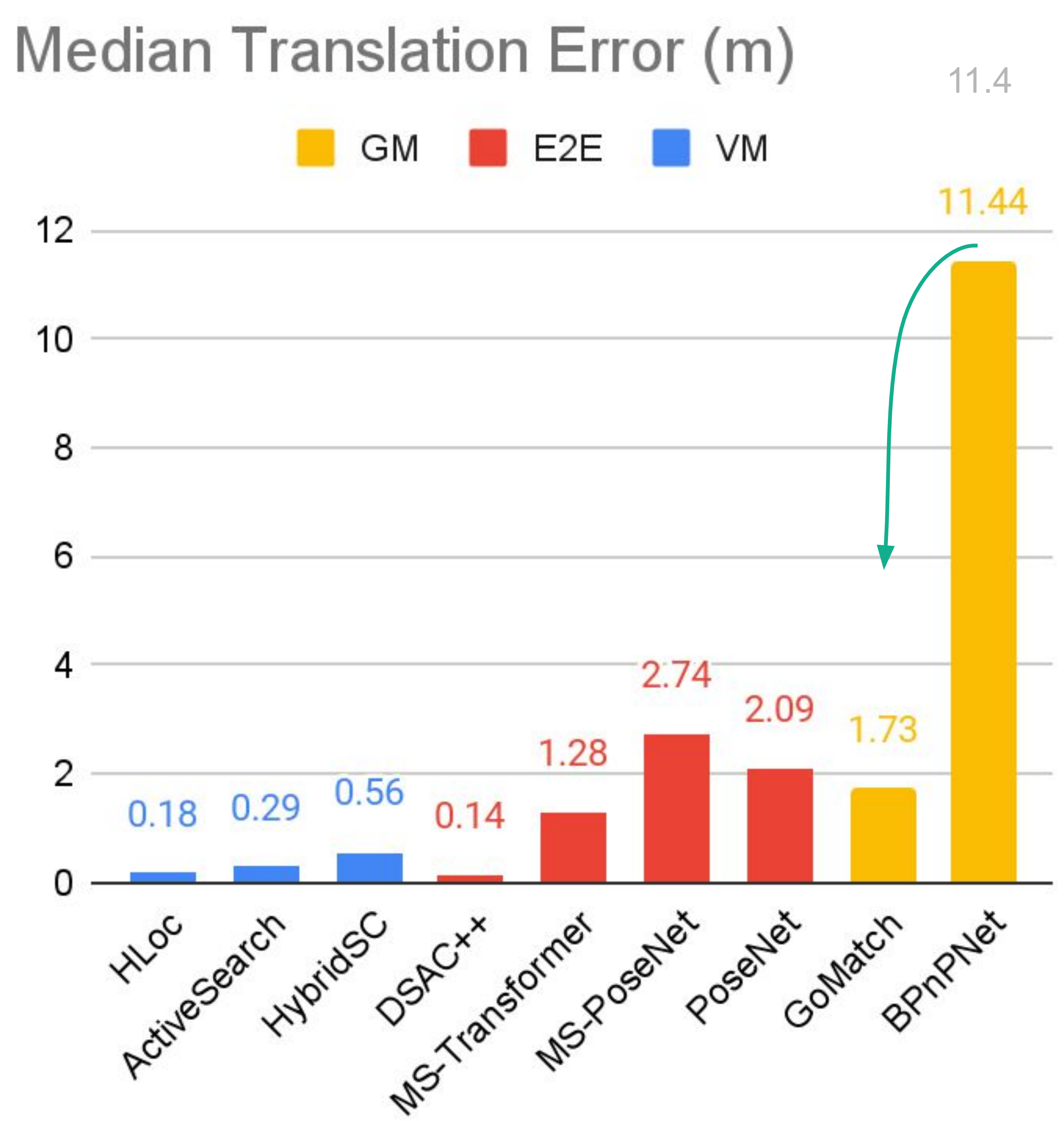


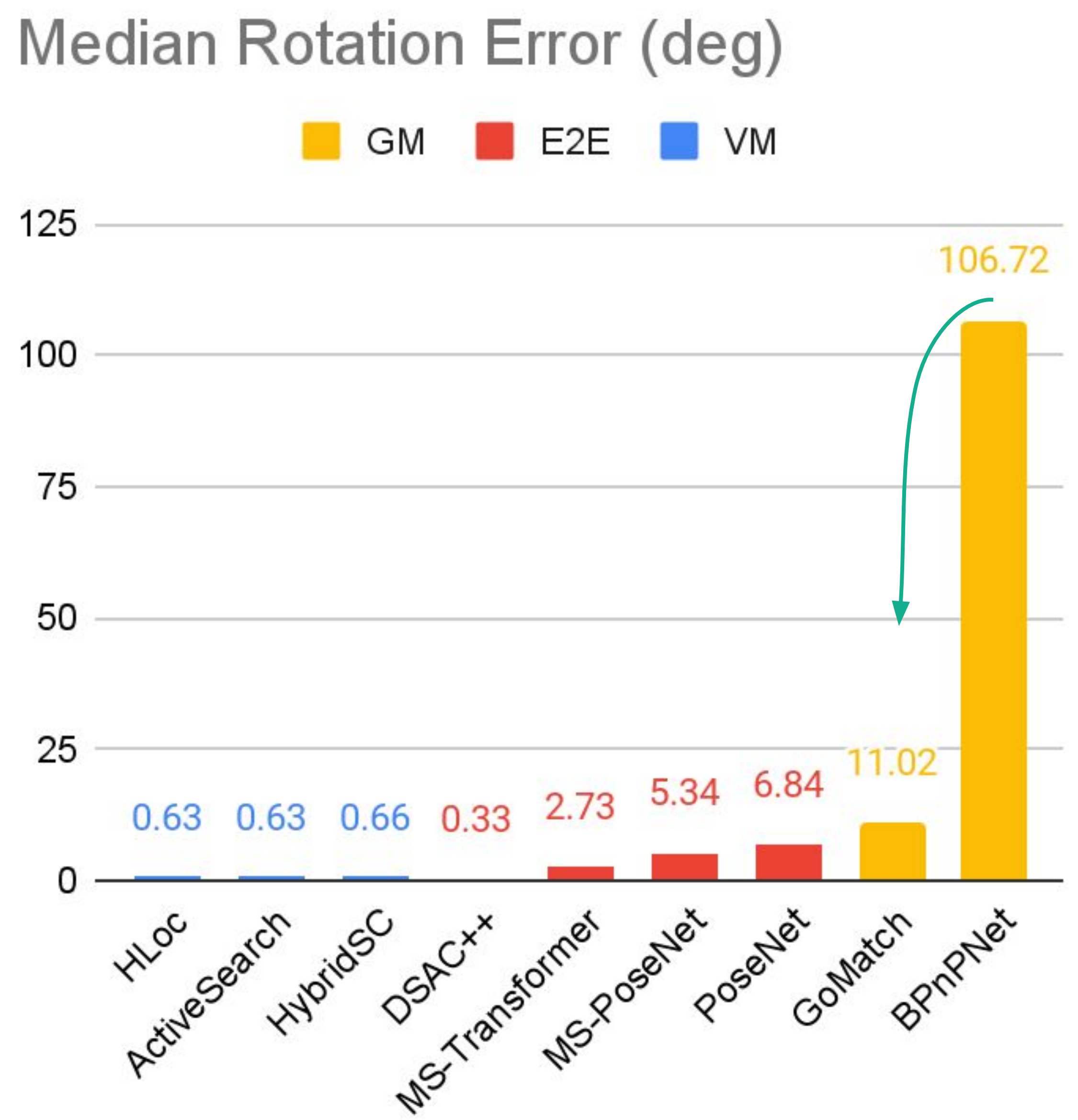
Comparison with SOTA – Cambridge Landmarks



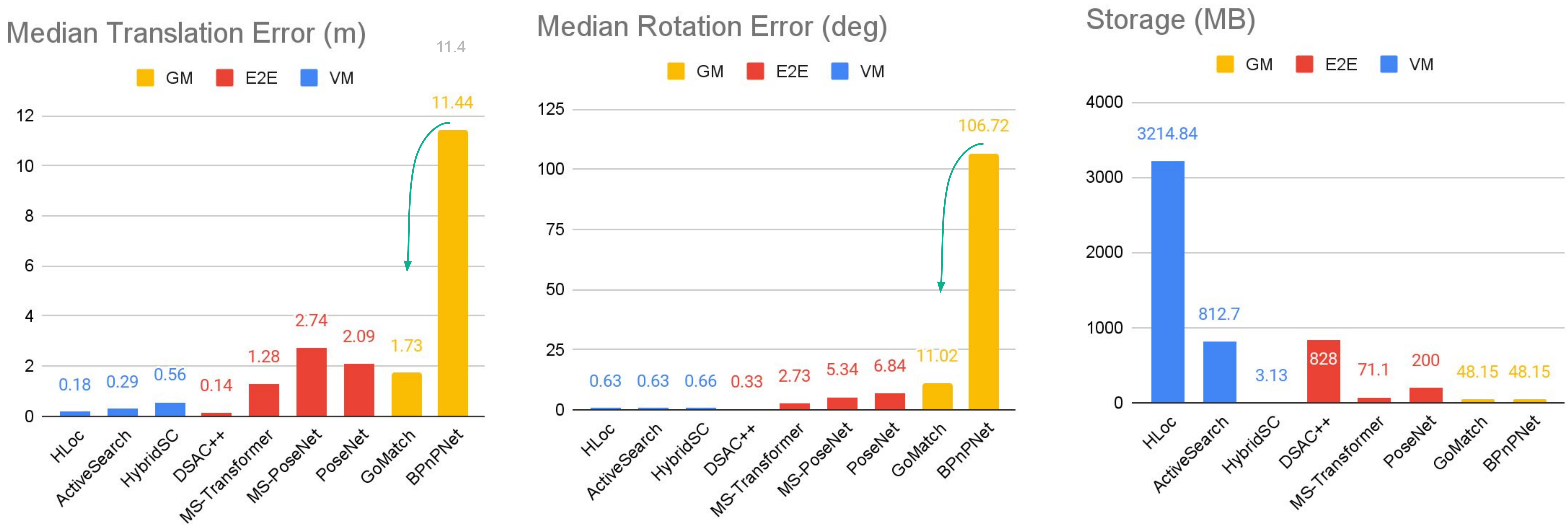


Comparison with SOTA – Cambridge Landmarks

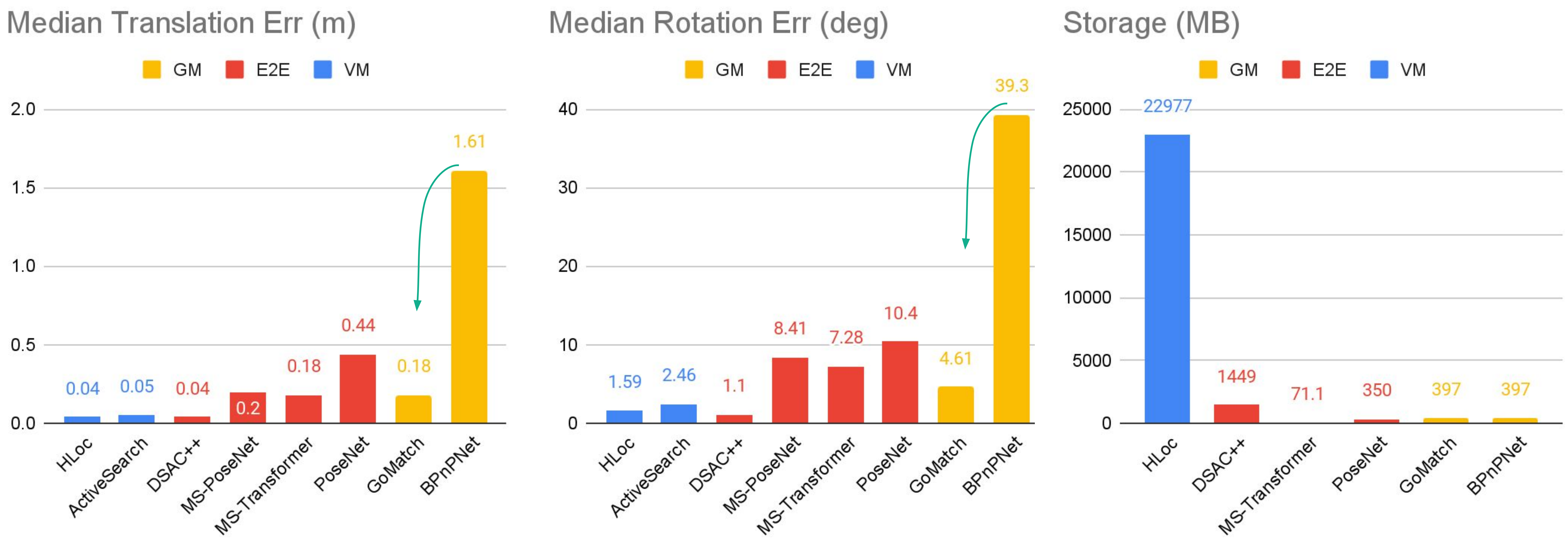




Comparison with SOTA – Cambridge Landmarks



Comparison with SOTA – 7 Scenes



Compare to VM – Cambridge Landmarks

	Method	Storage (MB)	No Desc. Maint.	Privacy	King's College	Old Hospital	Shop Facade	St. Mary's Church
					Median Pose Error (m, °) (↓)			
E2E	PoseNet [38]	200	✓	✓	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
	DSAC++ [6]	828	✓	✓	0.18/0.30	0.20/0.30	0.06/0.30	0.13/0.40
	MSPN [4]	-	✓	✓	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18
	MS-Transformer [65]	71.1	✓	✓	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99
VM	HybridSC [14]	3.13	✗	?	0.81/0.59	0.75/1.01	0.19/0.54	0.50/0.49
	Active Search [58]	812.7	✗	✗	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54
	HLoc [55](w.SP [22])	3214.84	✗	✗	0.16/0.38	0.33/1.04	0.07/0.54	0.16/0.54
	HLoc(w.SP+SG [56])	3214.84	✗	✗	**0.12/0.20**	**0.15/0.30**	**0.04/0.20**	**0.07/0.21**
GM	BPnPNet [11]	48.15	✓	✓	26.73/106.99	24.8/162.99	7.53/107.17	11.11/49.74
	GoMatch	48.15	✓	✓	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94

Compare to VM – Cambridge Landmarks

	Method	Storage (MB)	No Desc. Maint.	Privacy	King's College	Old Hospital	Shop Facade	St. Mary's Church
					Median Pose Error (m, °) (↓)			
E2E	PoseNet [38]	200	✓	✓	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
	DSAC++ [6]	828	✓	✓	0.18/0.30	0.20/0.30	0.06/0.30	0.13/0.40
	MSPN [4]	-	✓	✓	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18
	MS-Transformer [65]	71.1	✓	✓	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99
VM	HybridSC [14]	3.13	✗	?	0.81/0.59	0.75/1.01	0.19/0.54	0.50/0.49
	Active Search [58]	812.7	✗	✗	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54
	HLoc [55](w.SP [22])	3214.84	✗	✗	0.16/0.38	0.33/1.04	0.07/0.54	0.16/0.54
	HLoc(w.SP+SG [56])	3214.84	✗	✗	**0.12/0.20**	**0.15/0.30**	**0.04/0.20**	**0.07/0.21**
GM	BPnPNet [11]	48.15	✓	✓	26.73/106.99	24.8/162.99	7.53/107.17	11.11/49.74
	GoMatch	48.15	✓	✓	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94

Compare to VM – Cambridge Landmarks

	Method	Storage (MB)	No Desc. Maint.	Privacy	King's College	Old Hospital	Shop Facade	St. Mary's Church
					Median Pose Error (m, °) (↓)			
E2E	PoseNet [38]	200	✓	✓	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
	DSAC++ [6]	828	✓	✓	0.18/0.30	0.20/0.30	0.06/0.30	0.13/0.40
	MSPN [4]	-	✓	✓	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18
	MS-Transformer [65]	71.1	✓	✓	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99
VM	HybridSC [14]	3.13	✗	?	0.81/0.59	0.75/1.01	0.19/0.54	0.50/0.49
	Active Search [58]	812.7	✗	✗	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54
	HLoc [55](w.SP [22])	3214.84	✗	✗	0.16/0.38	0.33/1.04	0.07/0.54	0.16/0.54
	HLoc(w.SP+SG [56])	3214.84	✗	✗	**0.12/0.20**	**0.15/0.30**	**0.04/0.20**	**0.07/0.21**
GM	BPnPNet [11]	48.15	✓	✓	26.73/106.99	24.8/162.99	7.53/107.17	11.11/49.74
	GoMatch	48.15	✓	✓	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94

Compare to VM – Cambridge Landmarks

	Method	Storage (MB)	No Desc. Maint.	Privacy	King's College	Old Hospital	Shop Facade	St. Mary's Church
					Median Pose Error (m, °) (↓)			
E2E	PoseNet [38]	200	✓	✓	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
	DSAC++ [6]	828	✓	✓	0.18/0.30	0.20/0.30	0.06/0.30	0.13/0.40
	MSPN [4]	-	✓	✓	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18
	MS-Transformer [65]	71.1	✓	✓	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99
VM	HybridSC [14]	3.13	✗	?	0.81/0.59	0.75/1.01	0.19/0.54	0.50/0.49
	Active Search [58]	812.7	✗	✗	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54
	HLoc [55](w.SP [22])	3214.84	✗	✗	0.16/0.38	0.33/1.04	0.07/0.54	0.16/0.54
	HLoc(w.SP+SG [56])	3214.84	✗	✗	**0.12/0.20**	**0.15/0.30**	**0.04/0.20**	**0.07/0.21**
GM	BPnPNet [11]	48.15	✓	✓	26.73/106.99	24.8/162.99	7.53/107.17	11.11/49.74
	GoMatch	48.15	✓	✓	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94

Compare to VM – Cambridge Landmarks

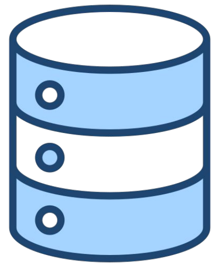
	Method	Storage (MB)	No Desc. Maint.	Privacy	King's College	Old Hospital	Shop Facade	St. Mary's Church
					Median Pose Error (m, °) (↓)			
E2E	PoseNet [38]	200	✓	✓	1.92/5.40	2.31/5.38	1.46/8.08	2.65/8.48
E2E	DSAC++ [6]	828	✓	✓	0.18/0.30	0.20/0.30	0.06/0.30	0.13/0.40
E2E	MSPN [4]	-	✓	✓	1.73/3.65	2.55/4.05	2.92/7.49	2.67/6.18
E2E	MS-Transformer [65]	71.1	✓	✓	0.83/1.47	1.81/2.39	0.86/3.07	1.62/3.99
VM	HybridSC [14]	3.13	✗	?	0.81/0.59	0.75/1.01	0.19/0.54	0.50/0.49
VM	Active Search [58]	812.7	✗	✗	0.42/0.55	0.44/1.01	0.12/0.40	0.19/0.54
VM	HLoc [55](w.SP [22])	3214.84	✗	✗	0.16/0.38	0.33/1.04	0.07/0.54	0.16/0.54
VM	HLoc(w.SP+SG [56])	3214.84	✗	✗	**0.12/0.20**	**0.15/0.30**	**0.04/0.20**	**0.07/0.21**
GM	BPnPNet [11]	48.15	✓	✓	26.73/106.99	24.8/162.99	7.53/107.17	11.11/49.74
GM	GoMatch	48.15	✓	✓	0.25/0.64	2.83/8.14	0.48/4.77	3.35/9.94

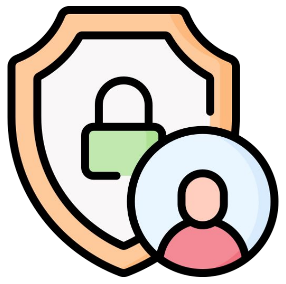
Conclusions

DGC-GNN: Leveraging Geometry and Color Cues for Visual Descriptor-Free 2D-3D Matching (CVPR24):
- adding sparse color information
- global-to-local GNN for matching

=> significantly boost performance

- Geometric localization is possible and (somewhat) SOTA

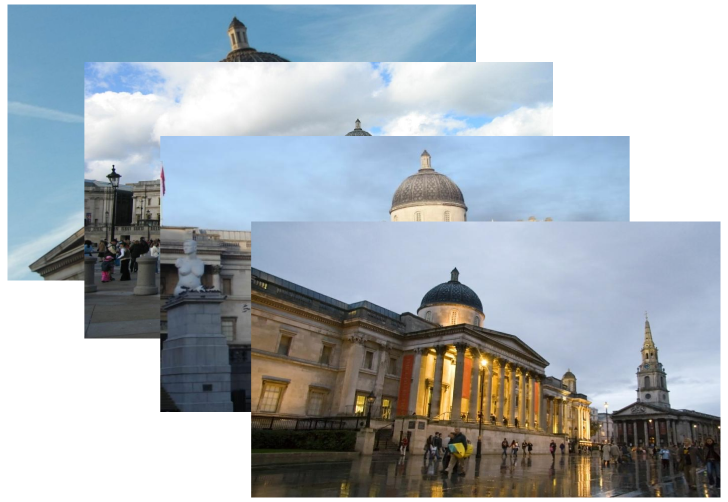




- Opens a new door for new work in privacy-aware, scalable localization

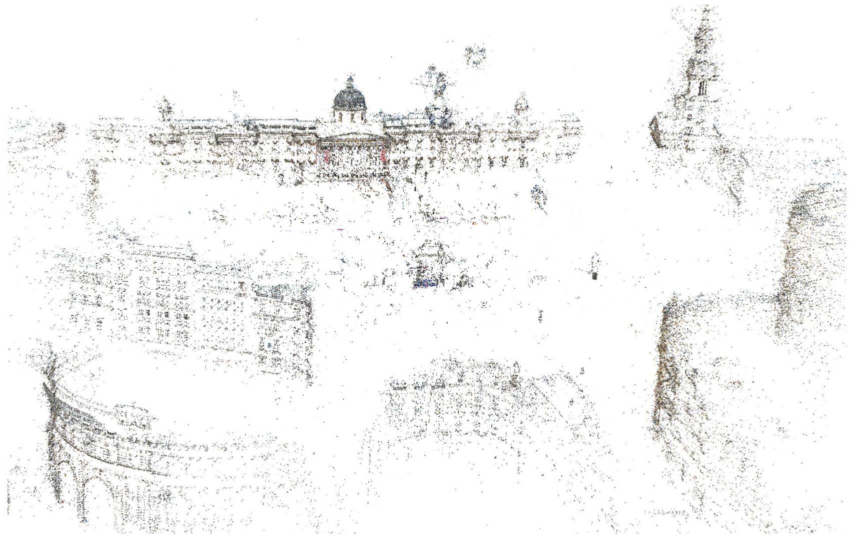
Localization System

Query | Scene Map | Method | Outputs



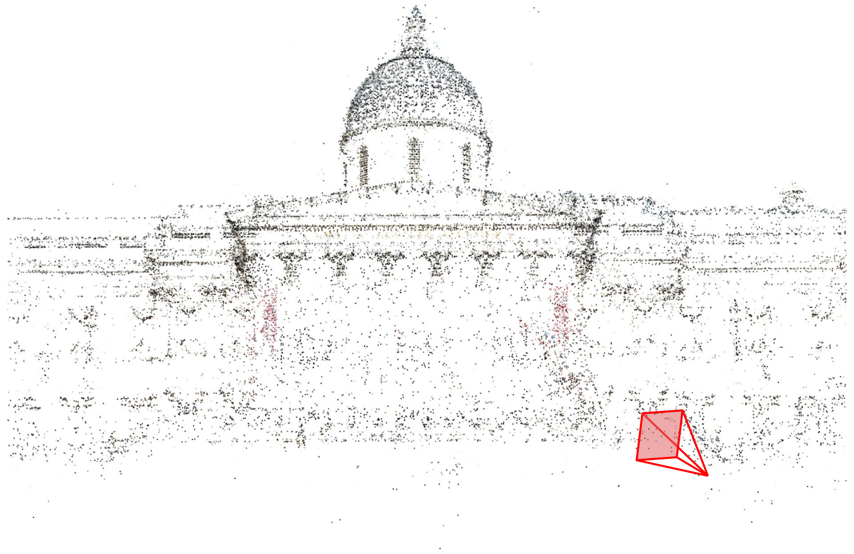
Reference Images

Query Image



Descriptor-free
Point Cloud

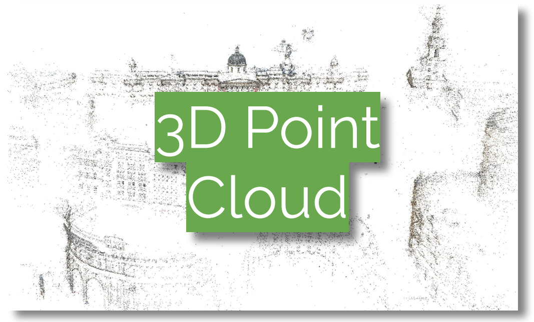
GoMatch



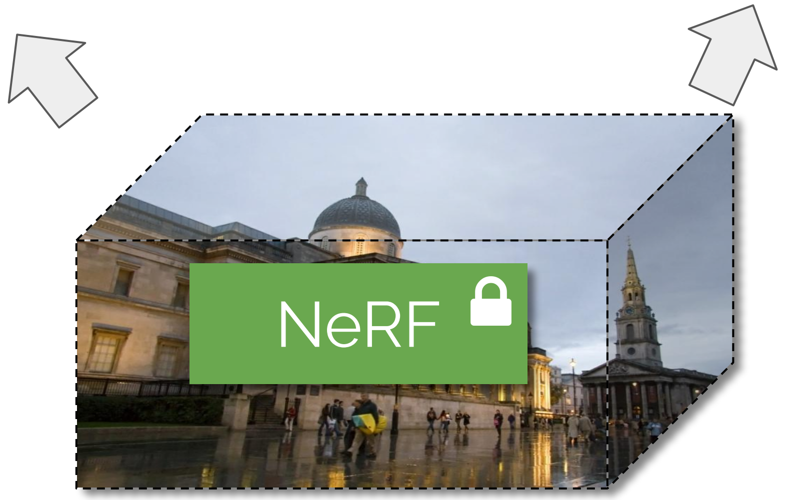
Even more compact
scene representation ?

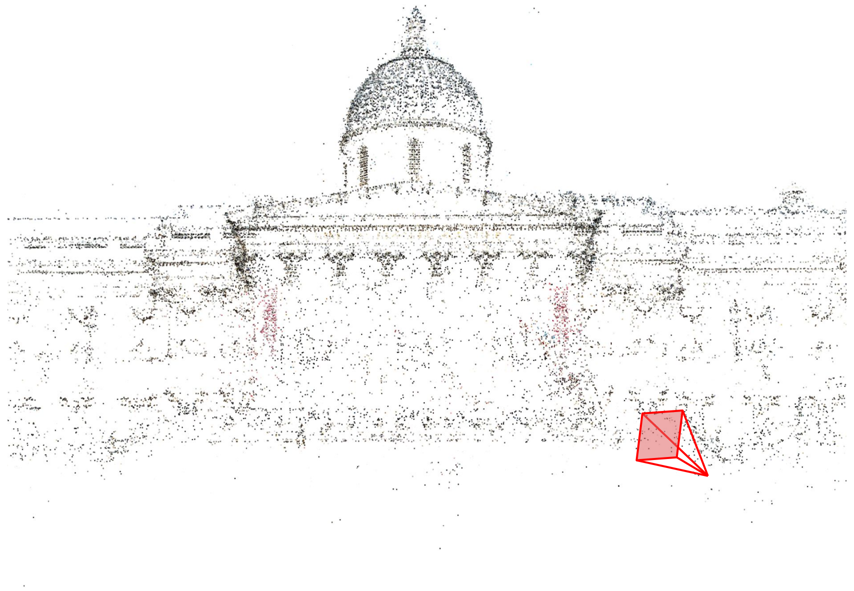
Localization System

Query + Scene Map → Method → Outputs



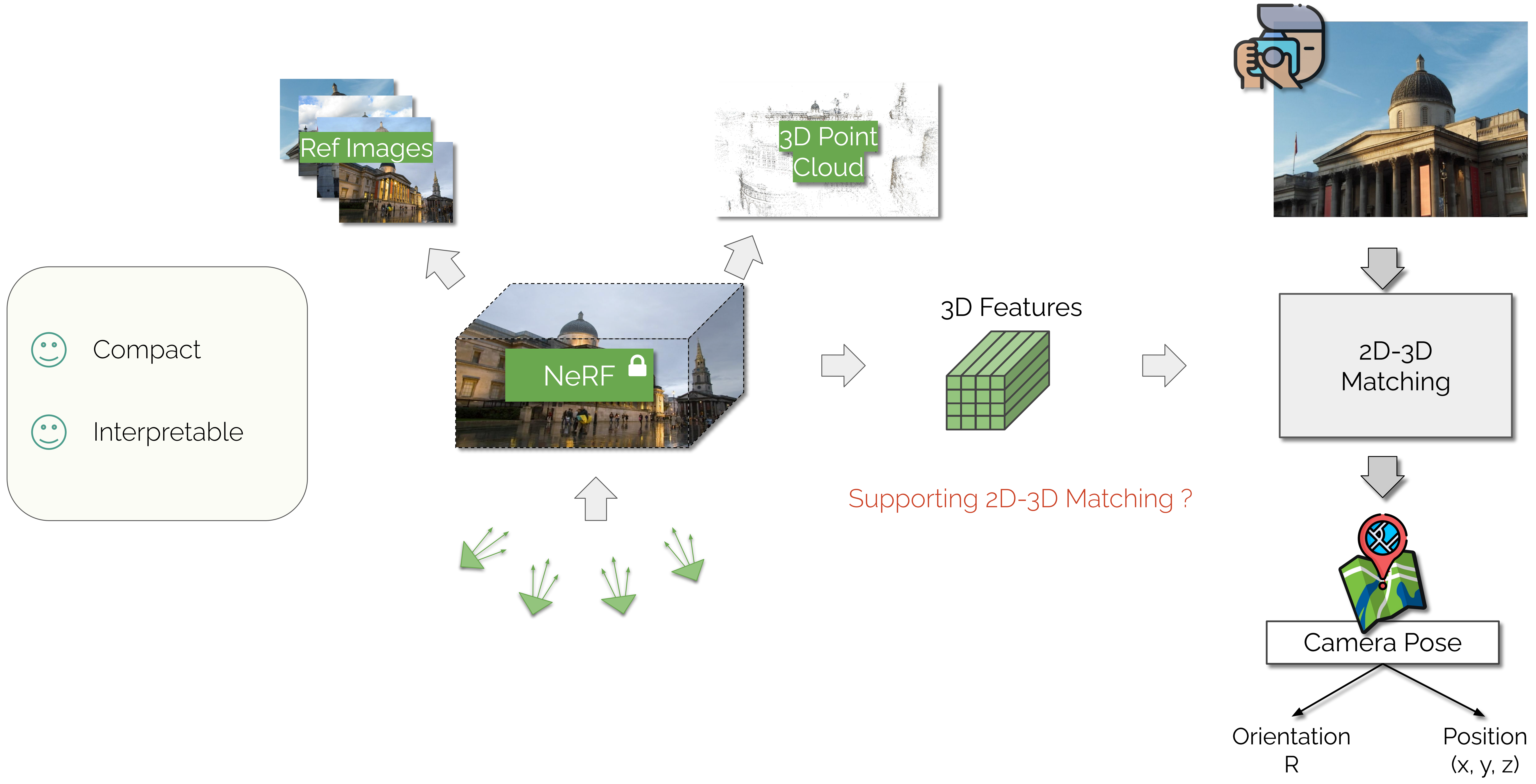

Query Image



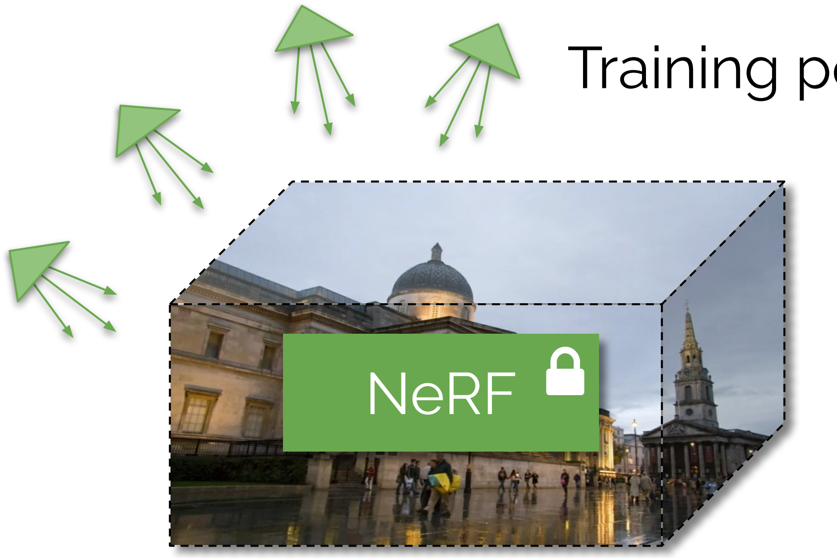


- Compact
- Interpretable

Introduction



Method



Training pose

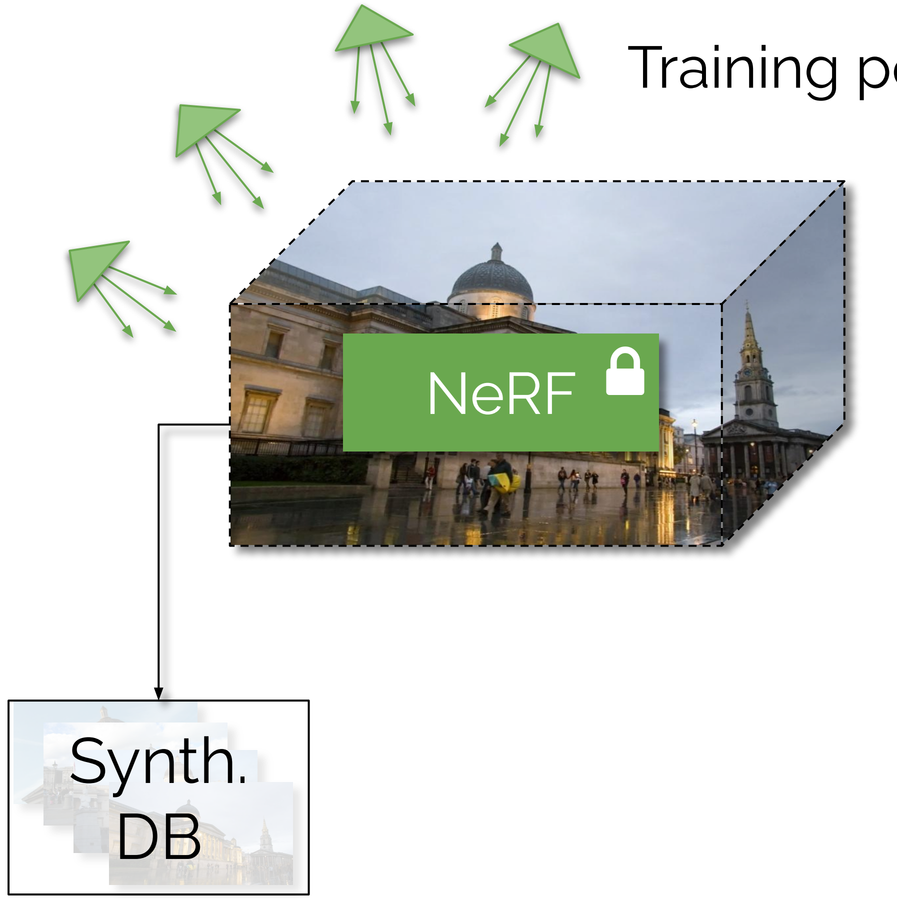
Query Image

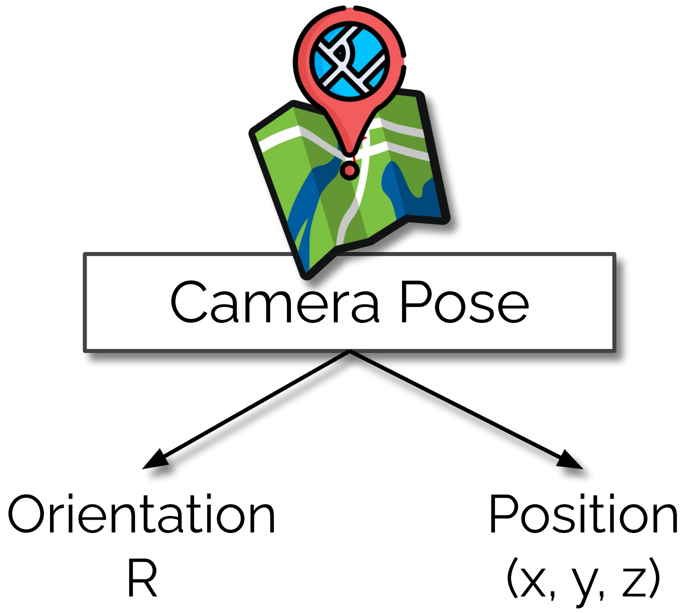
Camera Pose

Orientation
R

Position
(x, y, z)

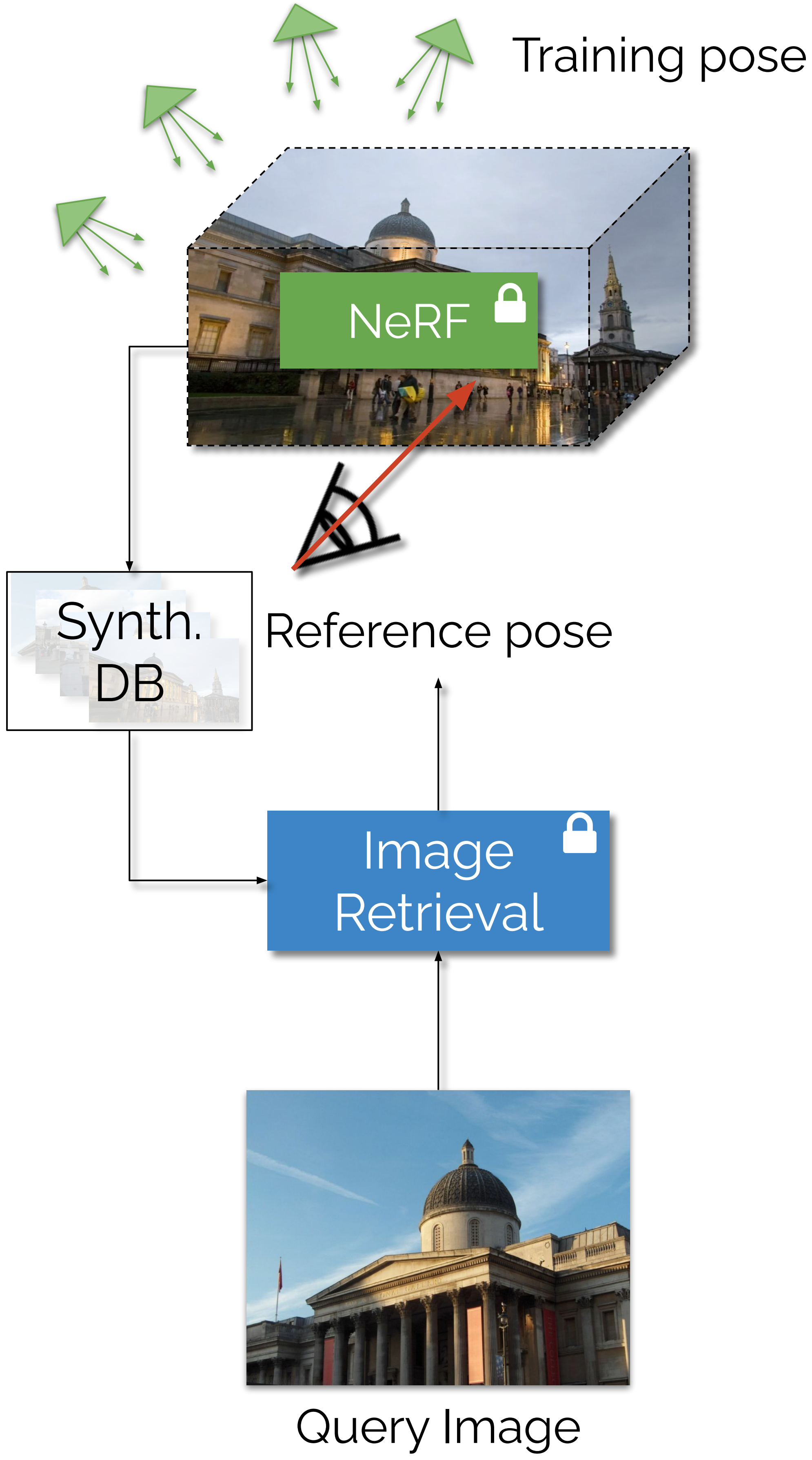
Method

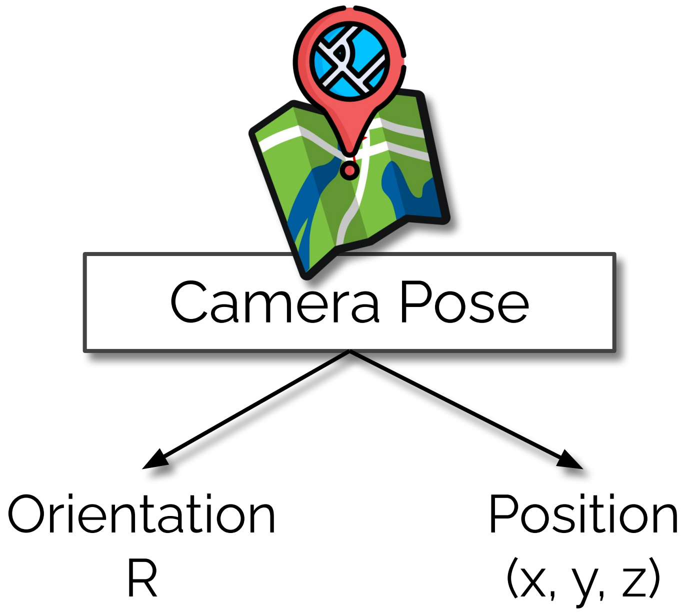




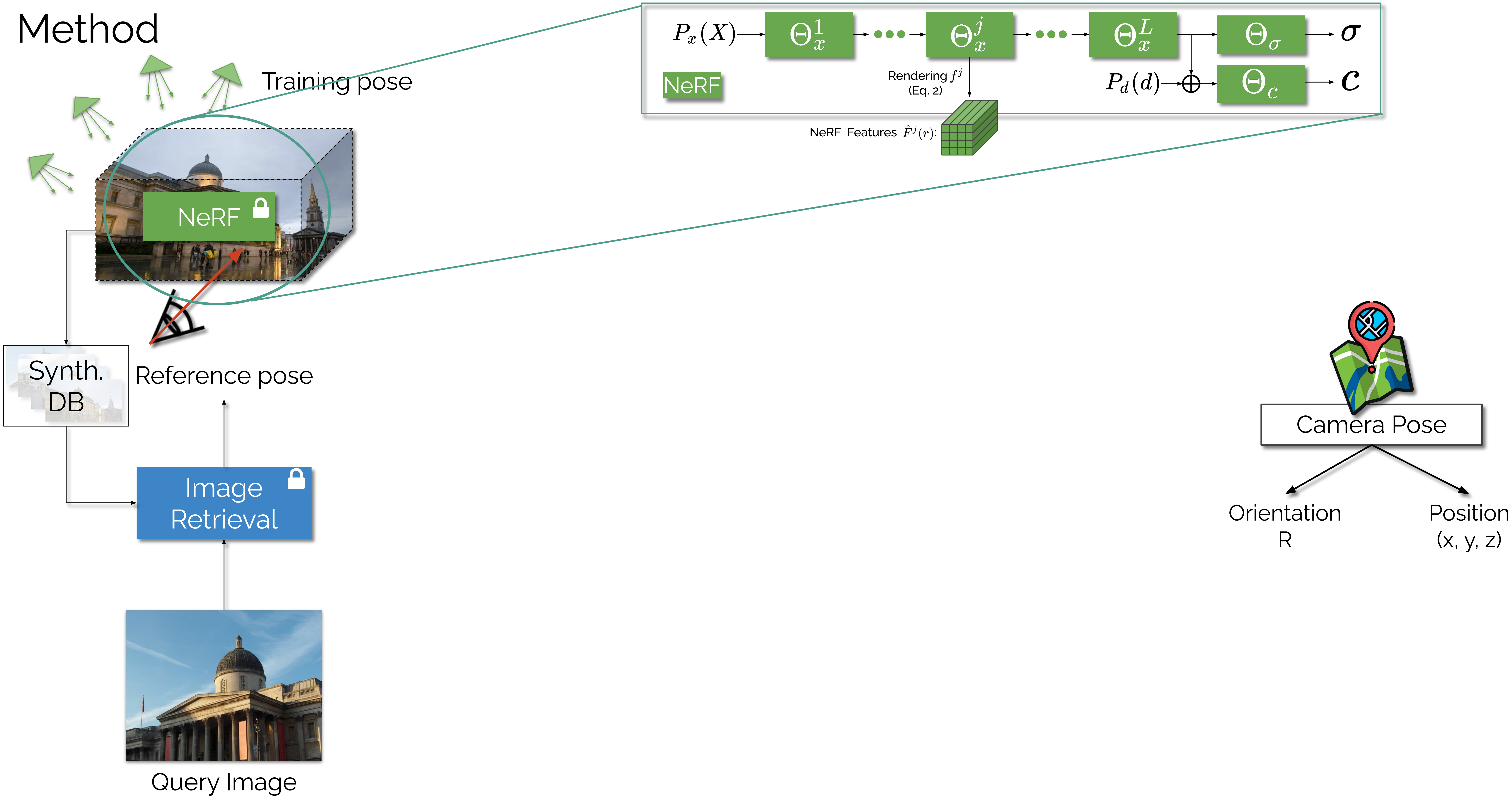

Query Image

Method


Training pose
NeRF
Synth. DB
Reference pose
Image Retrieval
Query Image

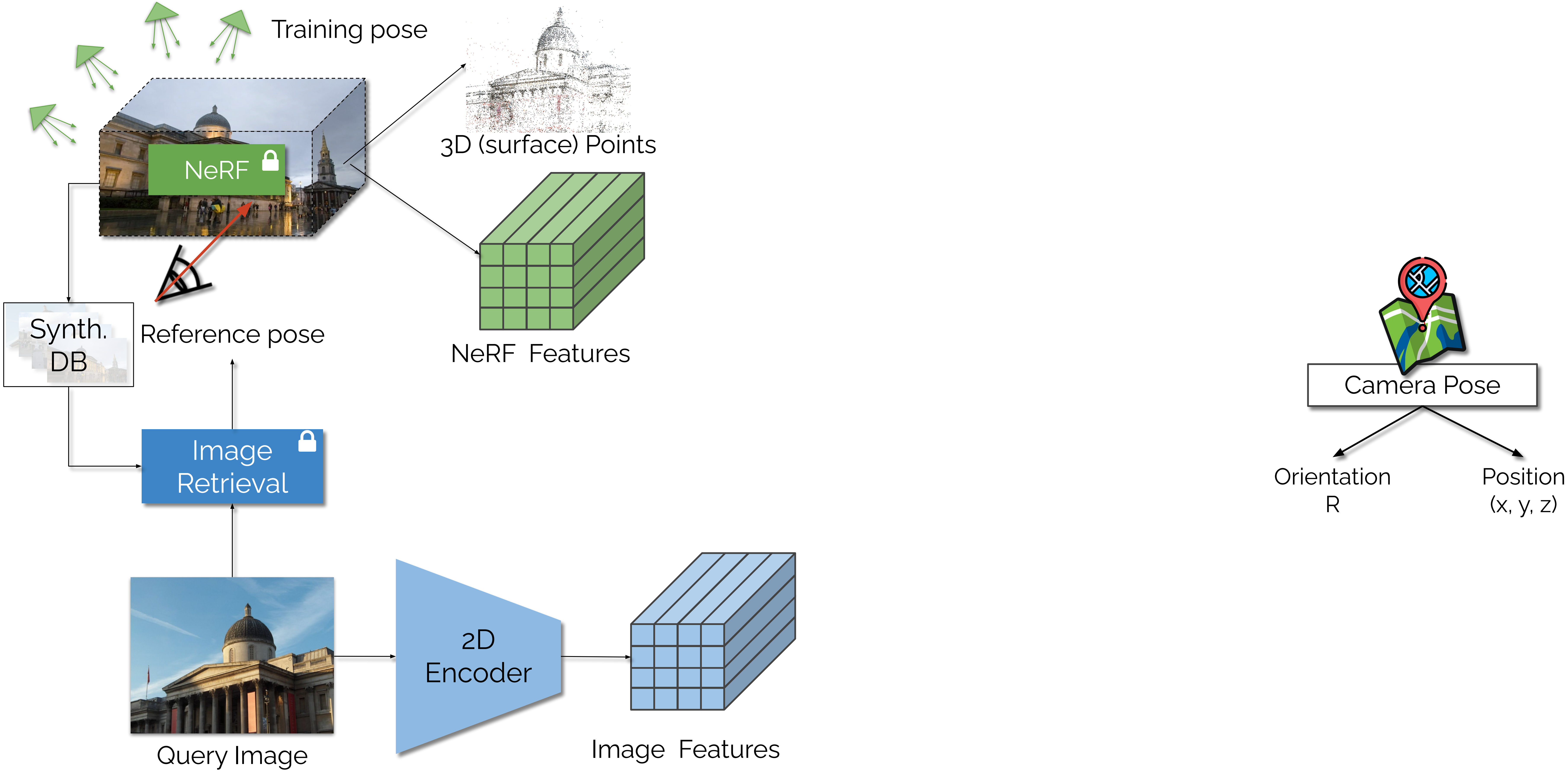

Camera Pose
Orientation R
Position (x, y, z)

Method



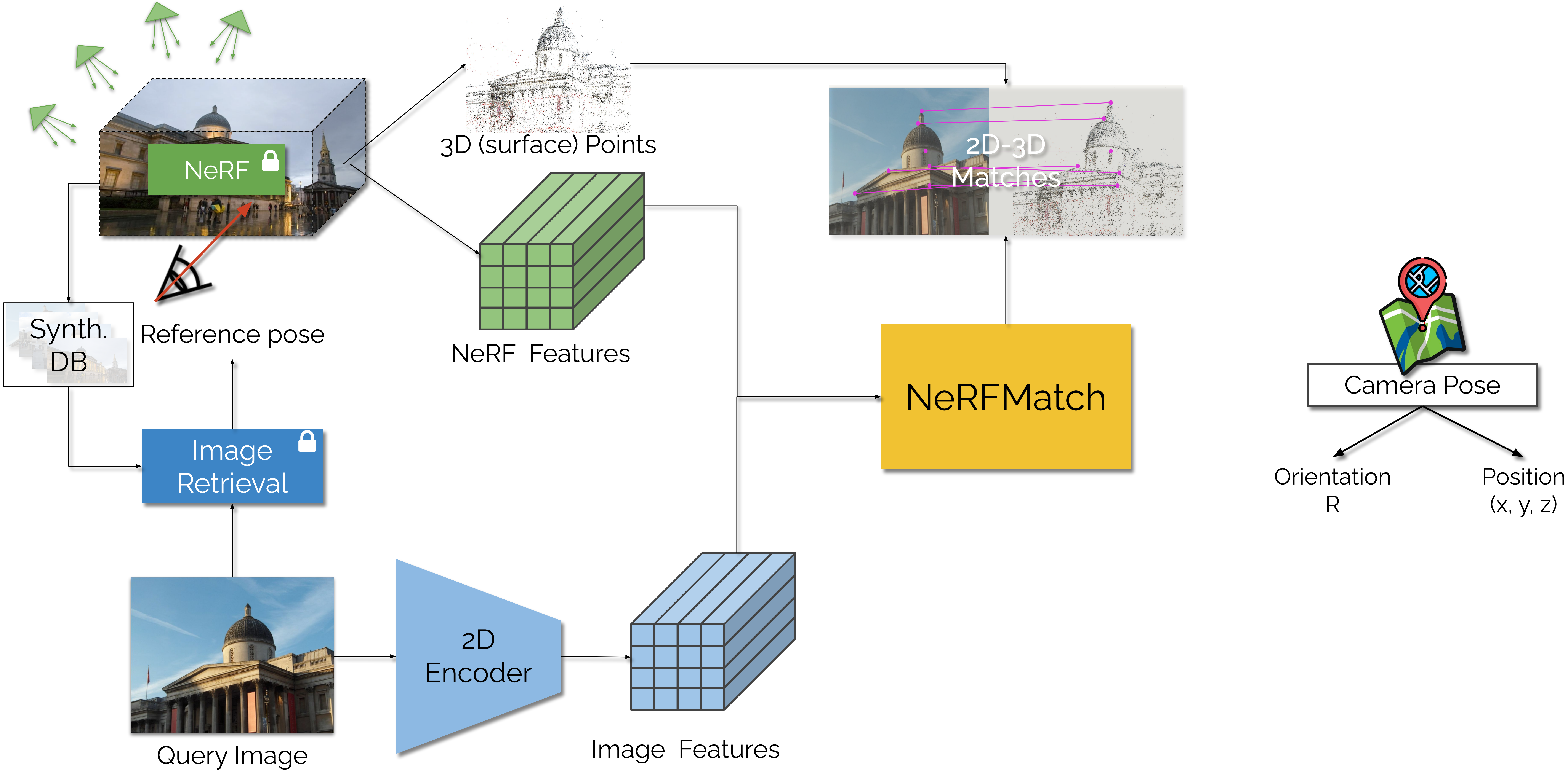
$P_x(X)$
Θ_x^1
Θ_x^j
Θ_x^L
Θ_σ
σ
NeRF
Rendering f^j (Eq. 2)
$P_d(d)$
Θ_c
c
NeRF Features $\hat{F}^j(r)$:
Training pose
NeRF
Synth. DB
Reference pose
Image Retrieval
Query Image
Camera Pose
Orientation R
Position (x, y, z)

Method

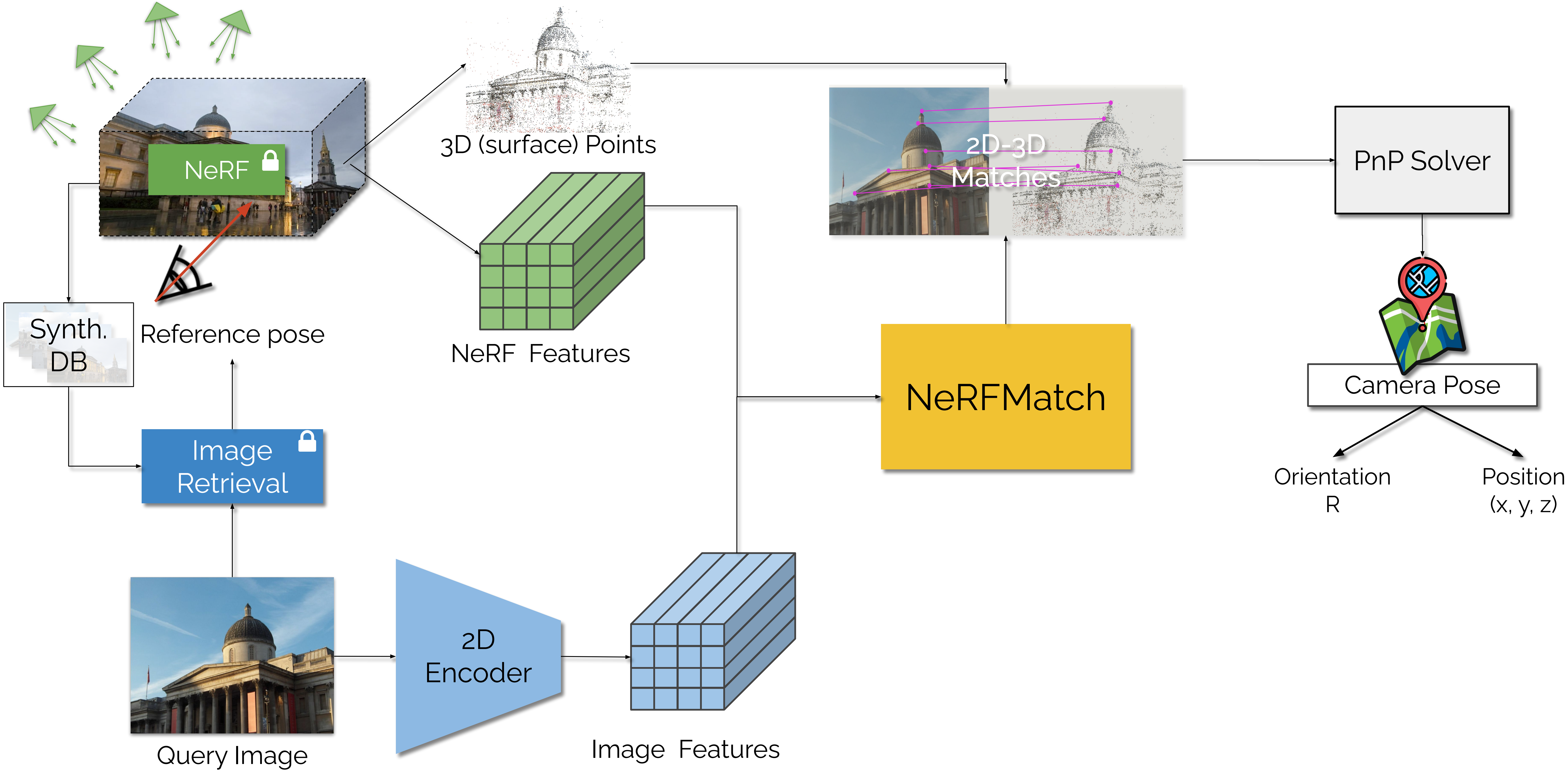


Training pose
3D (surface) Points
NeRF
NeRF Features
Synth.
DB
Reference pose
Image
Retrieval
Query Image
2D
Encoder
Image Features
Camera Pose
Orientation
R
Position
(x, y, z)

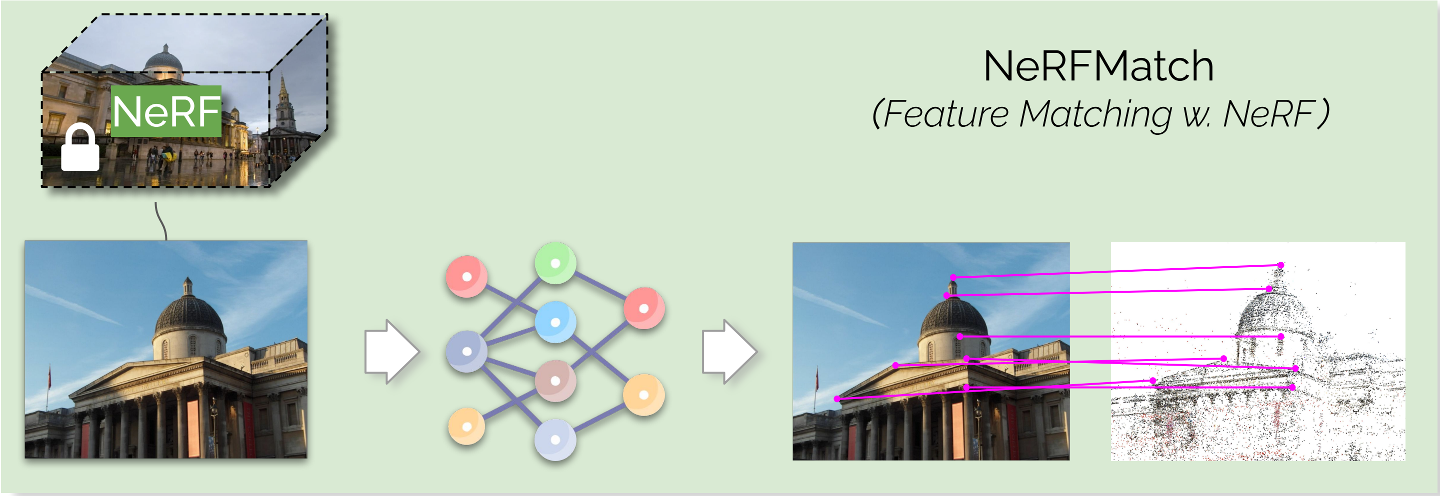
Method


NeRF
3D (surface) Points
2D-3D
Matches
NeRF Features
Synth.
DB
Reference pose
NeRFMatch
Camera Pose
Orientation
R
Position
(x, y, z)
Image
Retrieval
2D
Encoder
Query Image
Image Features

Method

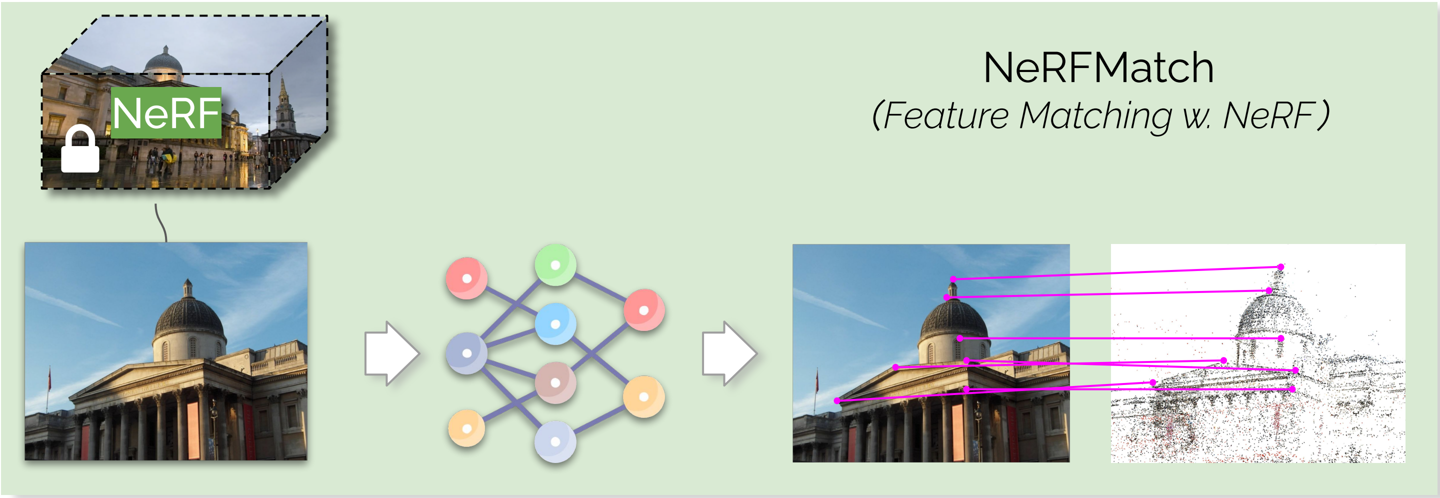


NeRF
3D (surface) Points
NeRF Features
2D-3D Matches
PnP Solver
Synth. DB
Reference pose
NeRFMatch
Camera Pose
Orientation R
Position (x, y, z)
Image Retrieval
2D Encoder
Query Image
Image Features



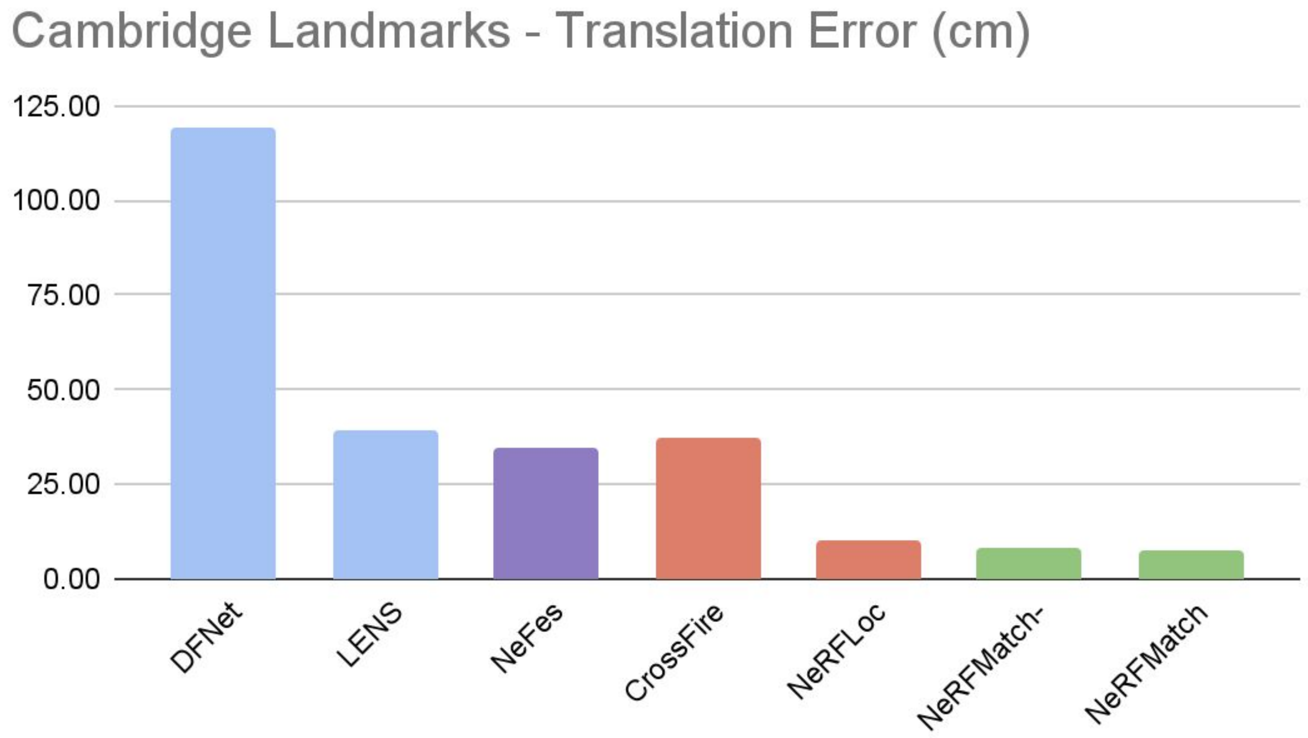
- NeRF not only provides 3D geometry but also comes with feature representation of 3D points that supports 2d-3D matching

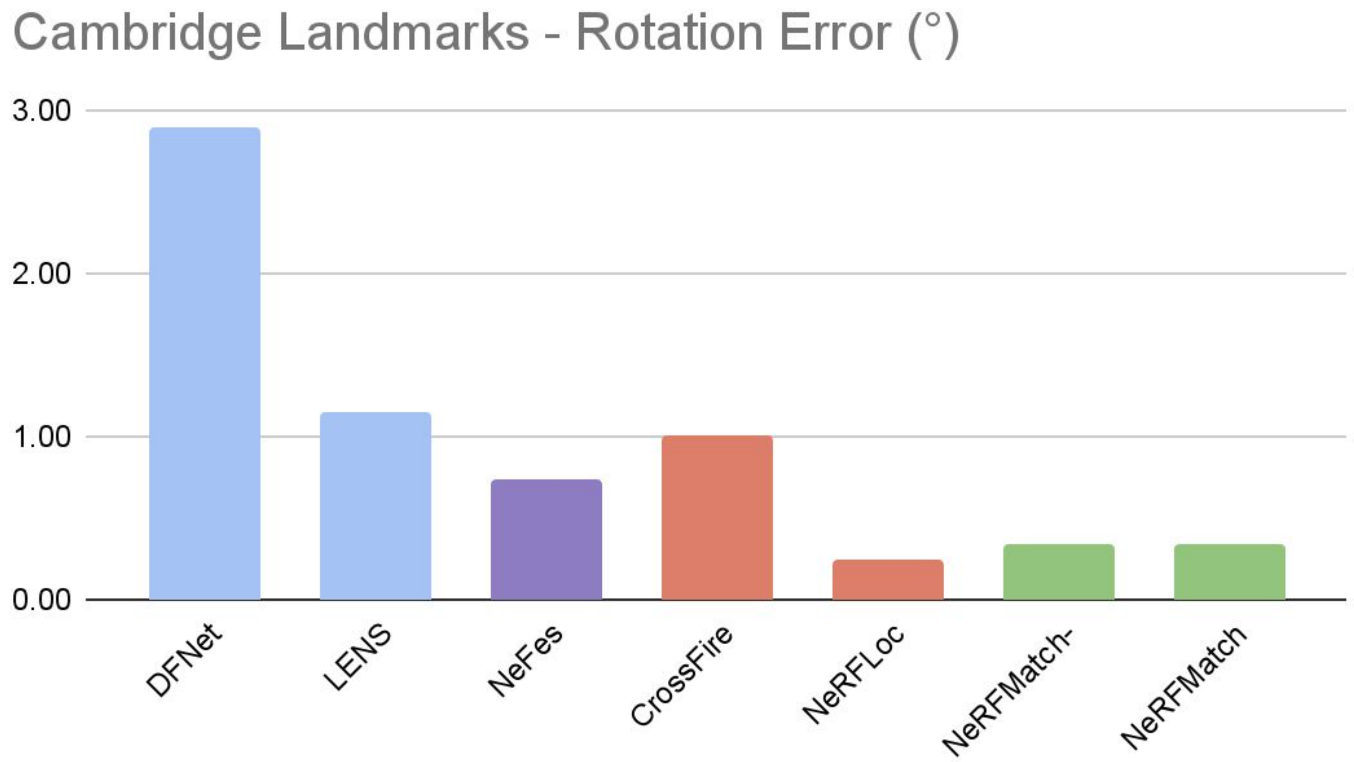
Metrics	Pt3D	Pe3D	f^1	f^2	f^3	f^4	f^5	f^6	f^7
Med. Translation (cm, $\downarrow$)	432.3	25.5	22.9	23.3	**21.8**	22.3	23.5	24.1	40.9
Med. Rotation ($^\circ$, $\downarrow$)	6.5	0.6	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	**0.5**	1.0
Localize Recall. (%, $\uparrow$)	2.1	60.1	64.8	64.2	**65.8**	64.1	63.3	62.1	43.2

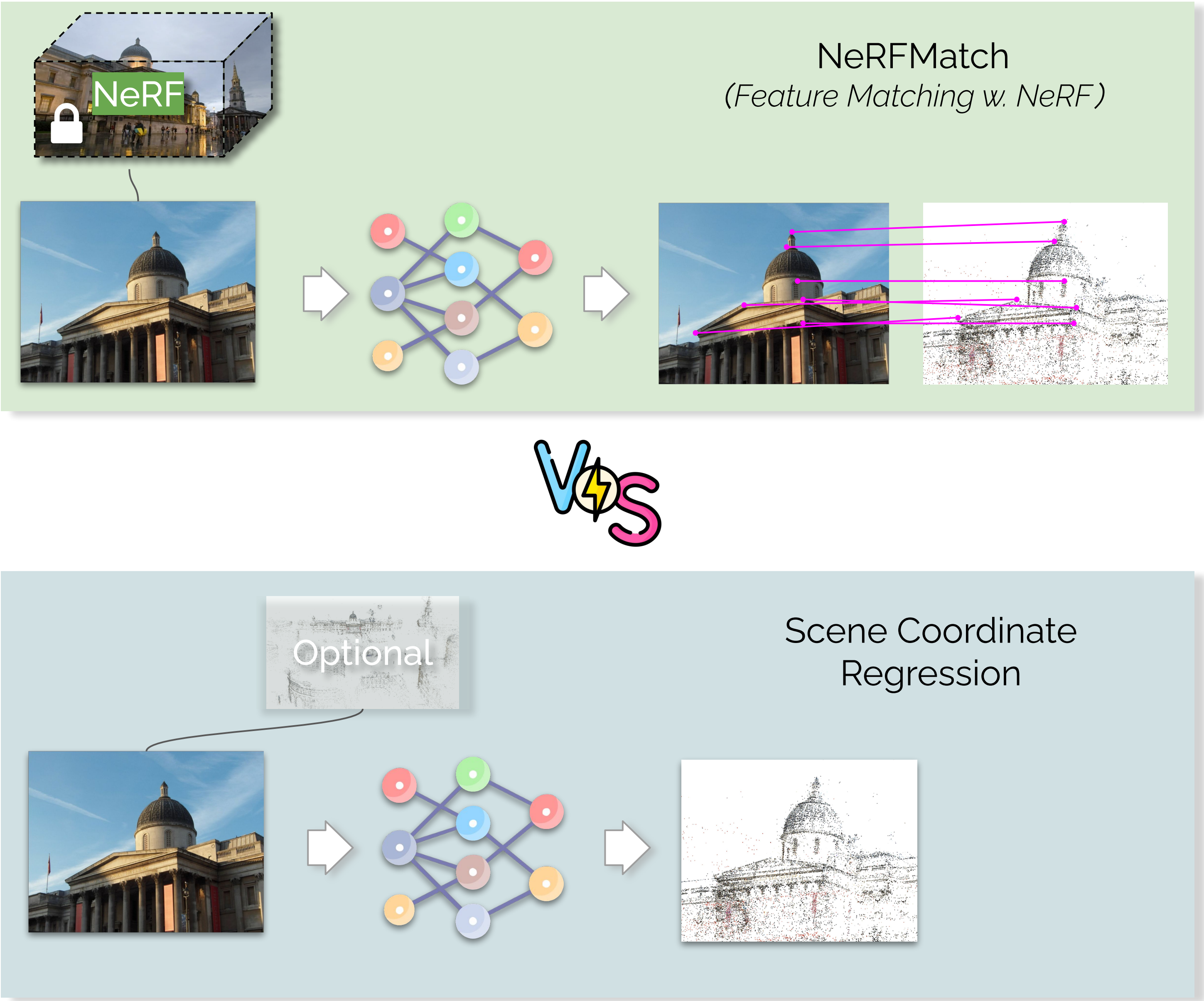


Compared to other ways of using NeRF:

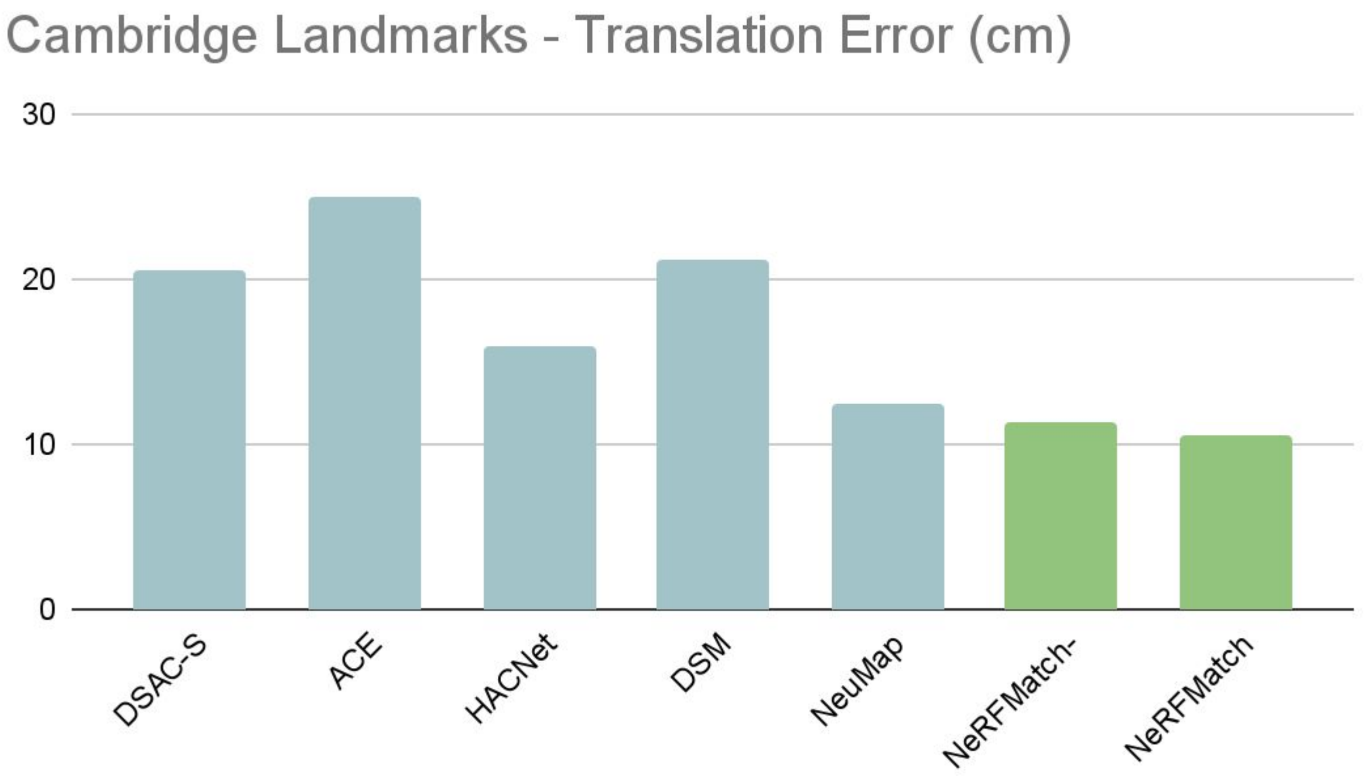
- Training Augmentation: DFNet, LENS
- Test-time refinement: NeFes
- Joint NeRF and matching training: CrossFire, NeRFLoc (requires depth image input)

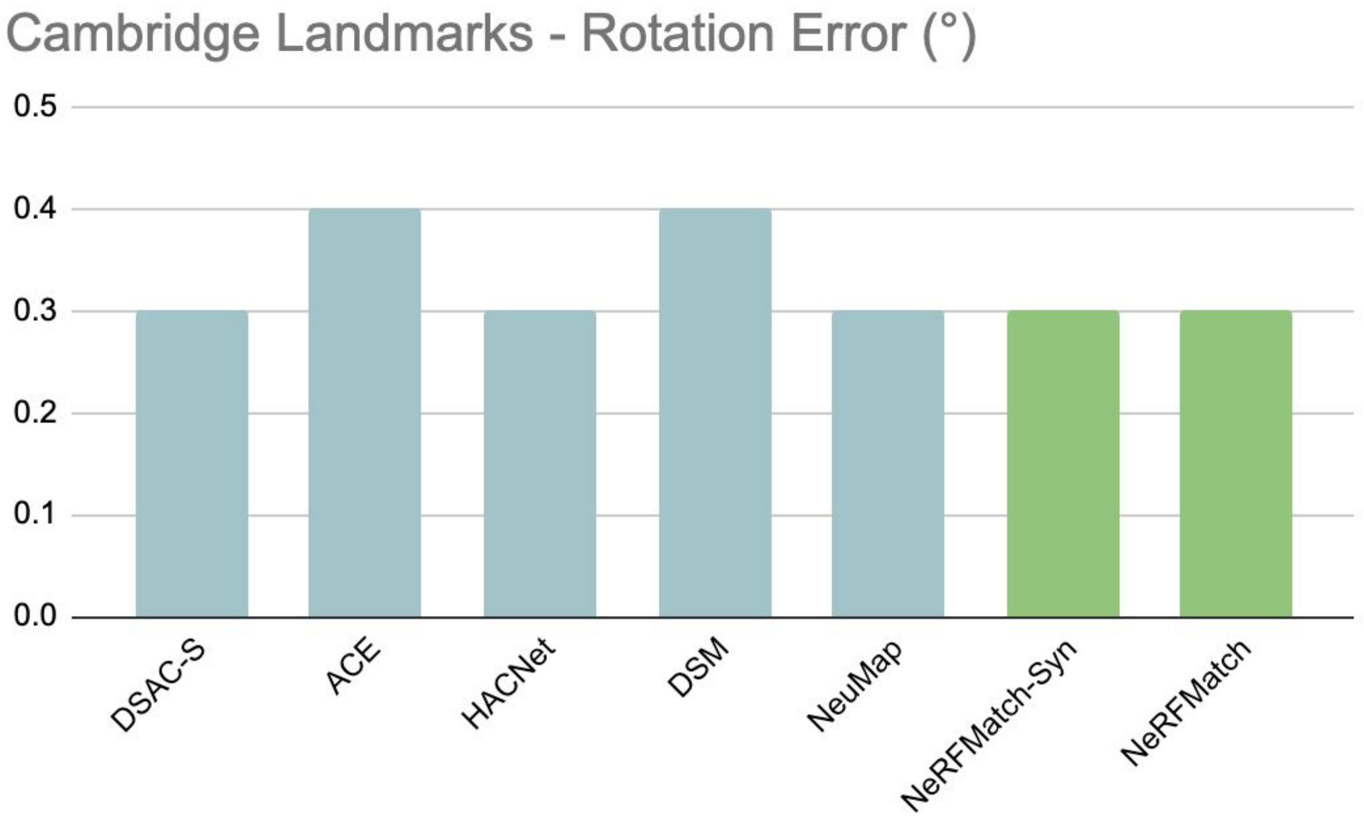


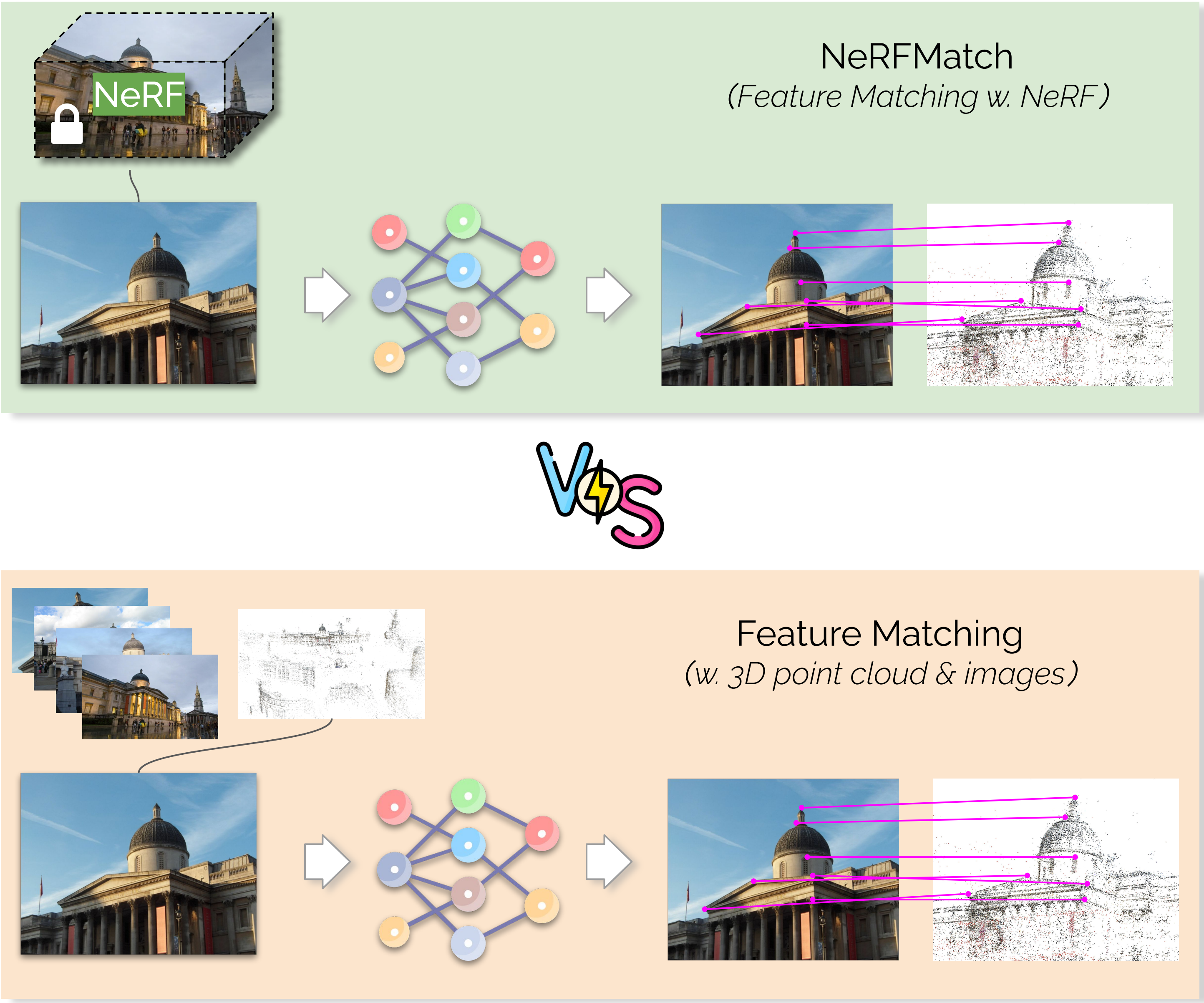




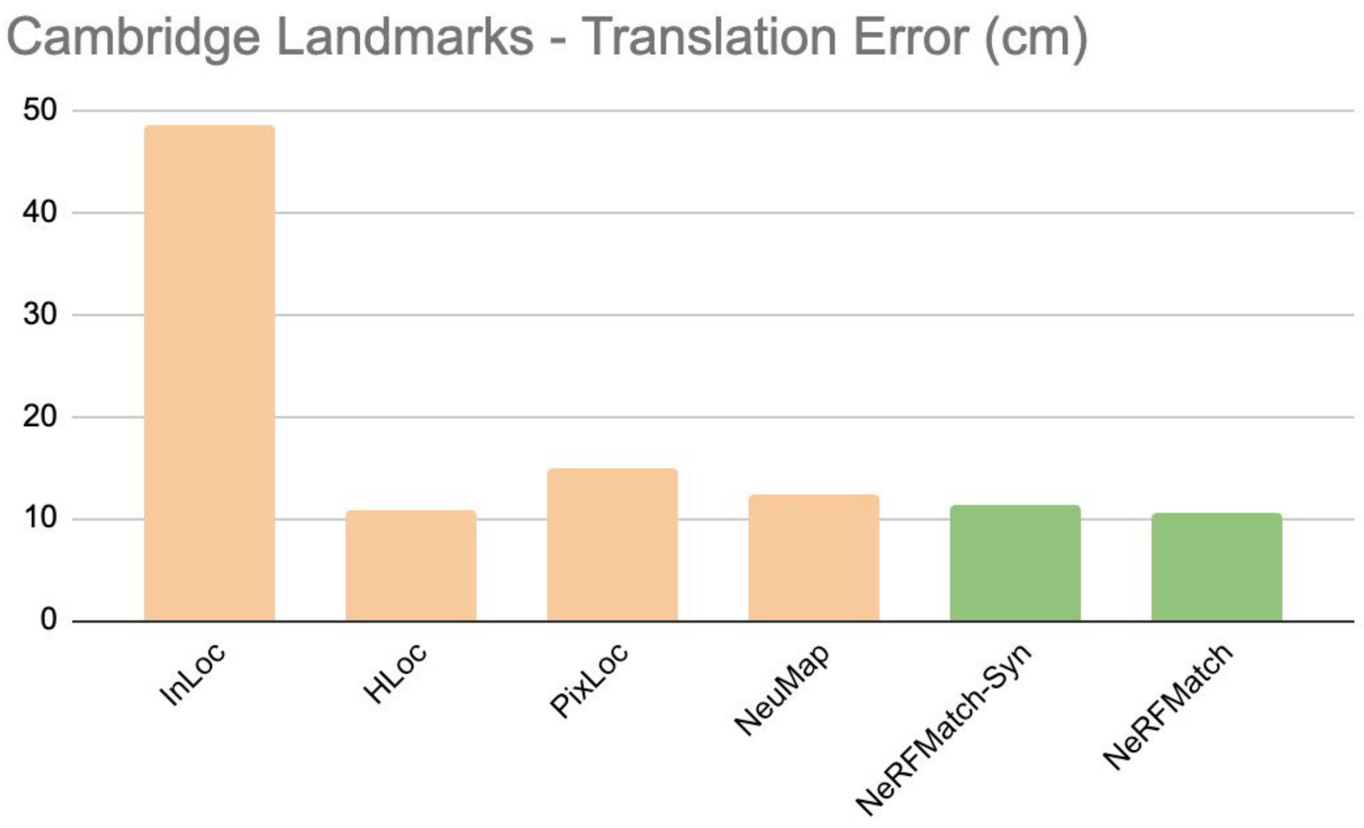
- Our method is similar to SCR methods where they map RGB to 3D points. Yet SCR represent 3D points with its coordiantes, while we use high-dimensional representation learned from NVS. And instead of regression, we learning a matching function to find a common ground between the (2d AND 3d) feature spaces. Our method outperform SCR on outdoor scenes.

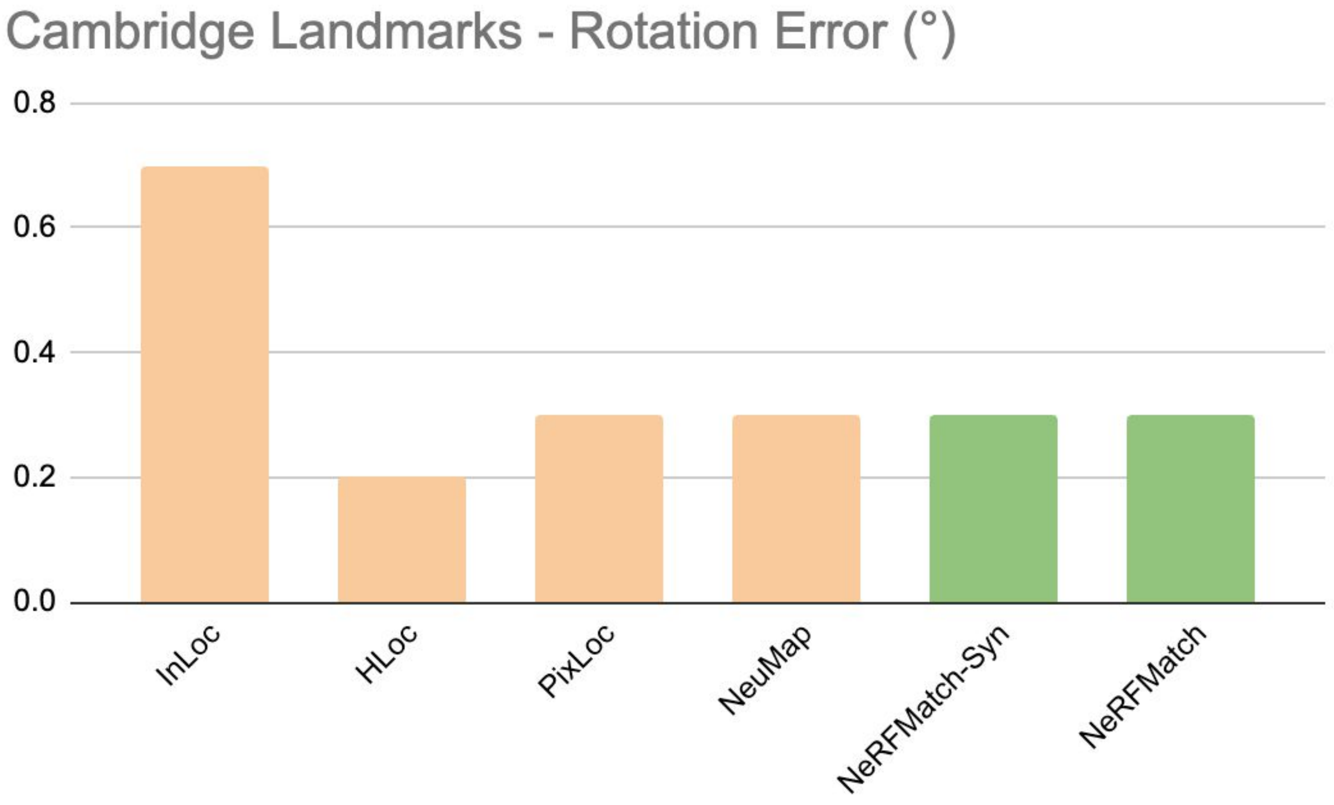


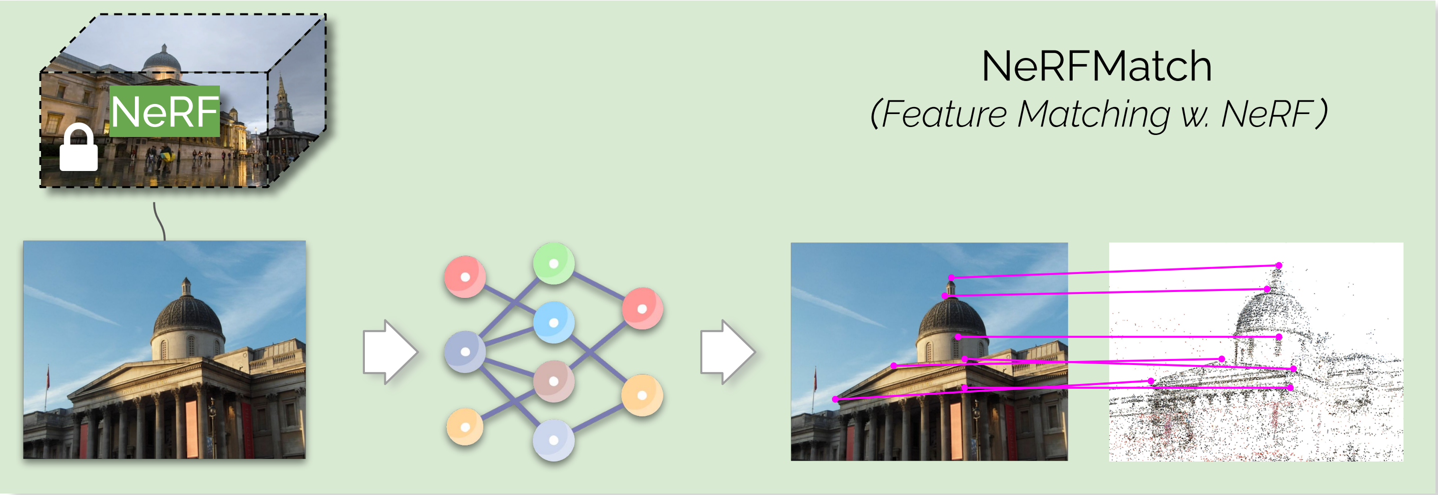




- We are on-par with the SOTA HLoc results on Cambridge Landmarks, which is quite challenging wild environment for NeRF training.







Indoor performance bottleneck vs SOTA

- **Depth inaccuracies**: NeRF predicted depth maps are used to compute pseudo ground-truth for matching supervision. Incorrect depth predictions can lead to misaligned feature correspondences. In contrast, image matching, SCR, and APR methods use more accurate labels like Colmap camera poses or 3D maps.

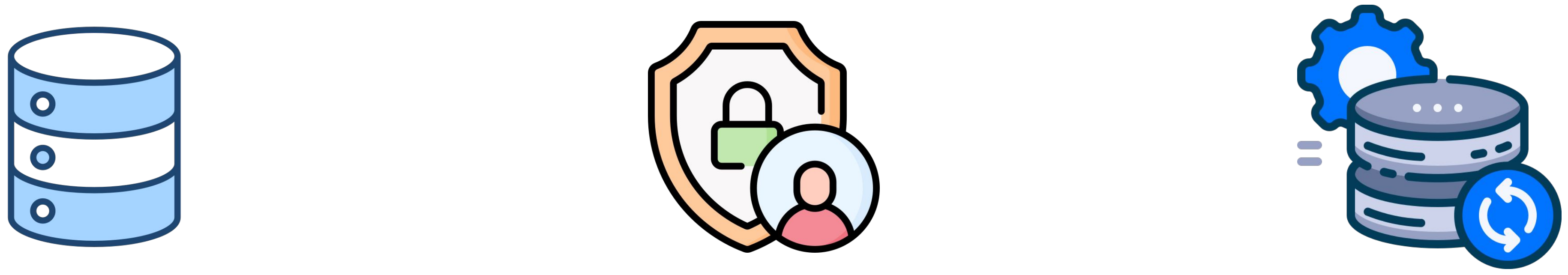
Indoor performance bottleneck vs SOTA

- Not good yet at **filtering inaccurate matches**, which has a large effect on small scenes.
- Better **scaling** to large-outdoor scene compared to regression-based methods.

Method	Scene	7-Scenes - SfM Poses - Indoor								
	Repres.	Chess	Fire	Heads	Office	Pump.	Kitchen	Stairs	Avg.Med↓	Avg.Recall↑.
MS-Trans. [53]	APR Net.	11/6.4	23/11.5	13/13	18/8.1	17/8.4	16/8.9	29/10.3	18.1/9.5	-
DFNet [17]	APR Net.	3/1.1	6/2.3	4/2.3	6/1.5	7/1.9	7/1.7	12/2.6	6.4/1.9	-
NeFeS [16]	APR+NeRF	2/0.8	2/0.8	2/1.4	2/0.6	2/0.6	2/0.6	5/1.3	2.4/0.9	-
DSAC* [10]	SCR Net.	0.5/0.2	0.8/0.3	0.5/0.3	1.2/0.3	1.2/0.3	0.7/0.2	2.7/0.8	1.1/0.3	**97.8**
ACE [6]	SCR Net.	0.7/0.5	0.6/0.9	0.5/ 0.5	1.2/0.5	1.1/0.2	0.9/0.5	2.8/1.0	1.1/0.6	97.1
DVLAD+R2D2 [45, 60]	3D+RGB	**0.4/0.1**	**0.5/0.2**	**0.4/0.2**	**0.7/0.2**	**0.6/0.1**	**0.4/0.1**	**2.4/0.7**	**0.8/0.2**	95.7
HLoc [48]	3D+RGB	0.8/**0.1**	0.9/**0.2**	0.6/0.3	1.2/**0.2**	1.4/0.2	1.1/**0.1**	2.9/0.8	1.3/0.3	95.7
NeRFMatch-Mini	NeRF+RGB	1.4/0.5	1.7/1.0	2.1/0.7	4.4/1.0	4.7/1.0	2.2/0.5	8.8/2.1	3.6/0.9	67.9
NeRFMatch	NeRF+RGB	0.9/0.3	1.3/0.4	1.6/1.0	3.2/0.7	3.3/0.7	1.3/0.3	7.5/1.3	2.7/0.7	75.3
NeRFMatch	NeRF	0.9/0.3	1.3/0.4	1.6/1.0	3.3/0.7	3.2/0.6	1.3/0.3	7.2/1.3	2.7/0.7	75.4

Conclusions

- Geometric localization is possible and (somewhat) SOTA

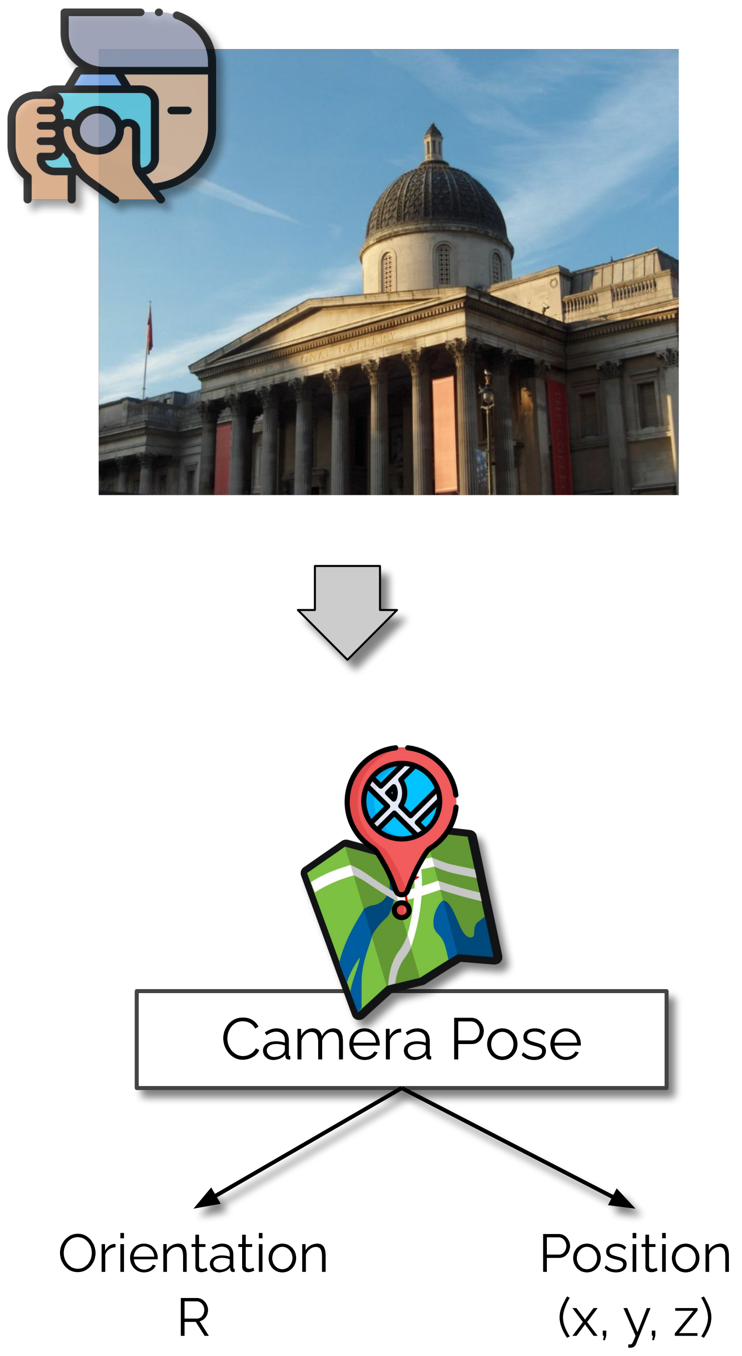


- Initial steps towards NeRF as the primary representation for visual localization

Compact

Interpretable

2D-3D matching

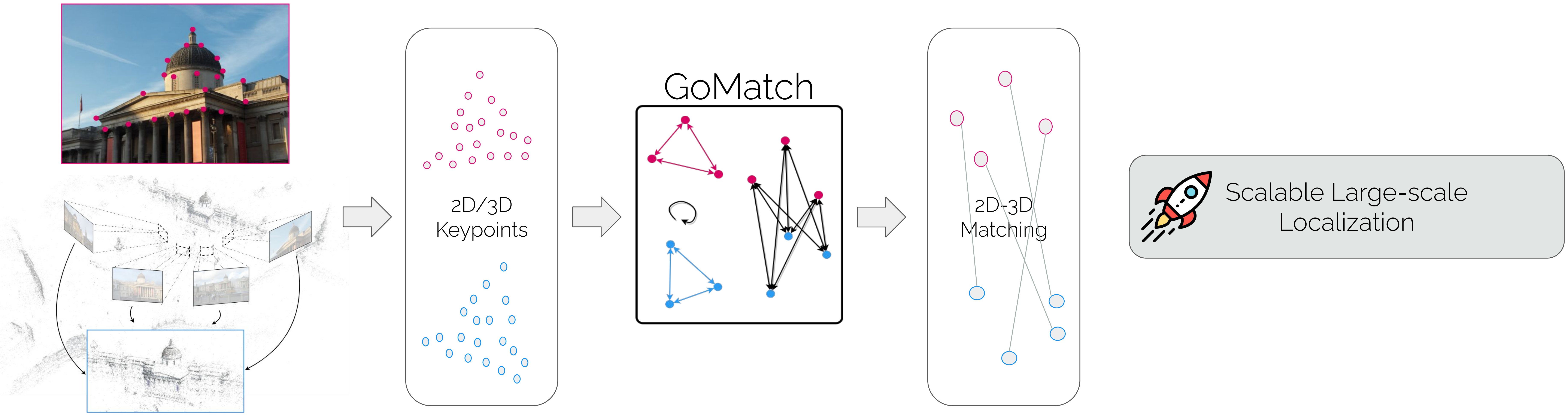


NVIDIA

Questions?

https://research.nvidia.com/labs/dvl/

Things that did not make the cut (yet?)



Qunjie Zhou

Sérgio Agostinho

Aljoša Ošep

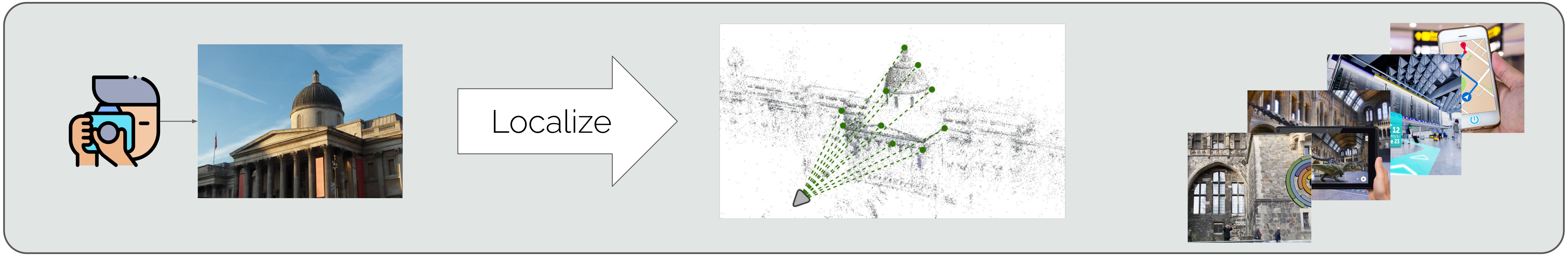
Laura Leal-Taixé

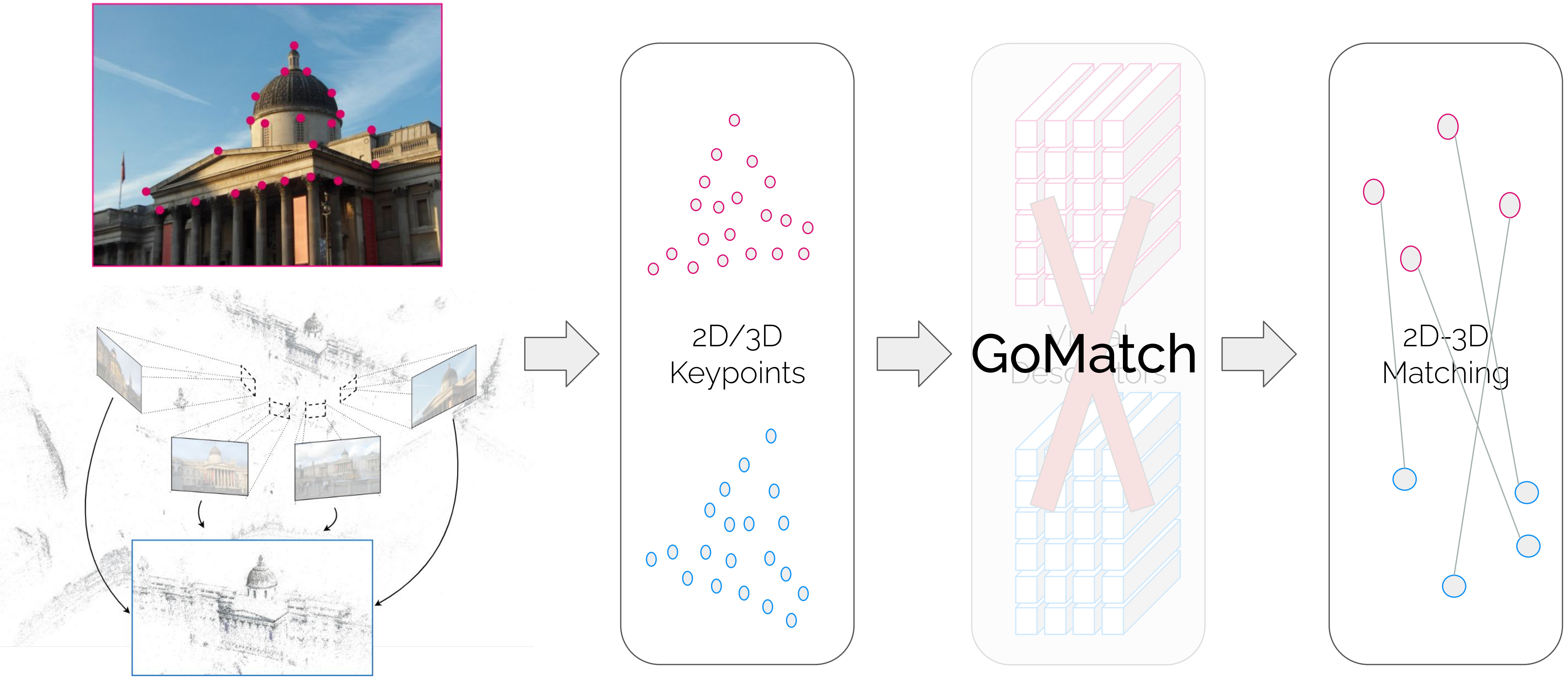
Visual Localization

Visual Localization

Visual Localization

Overview

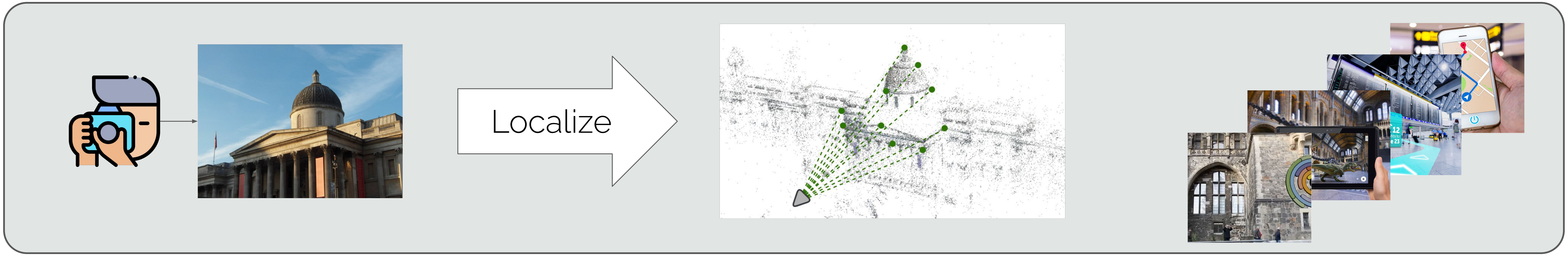


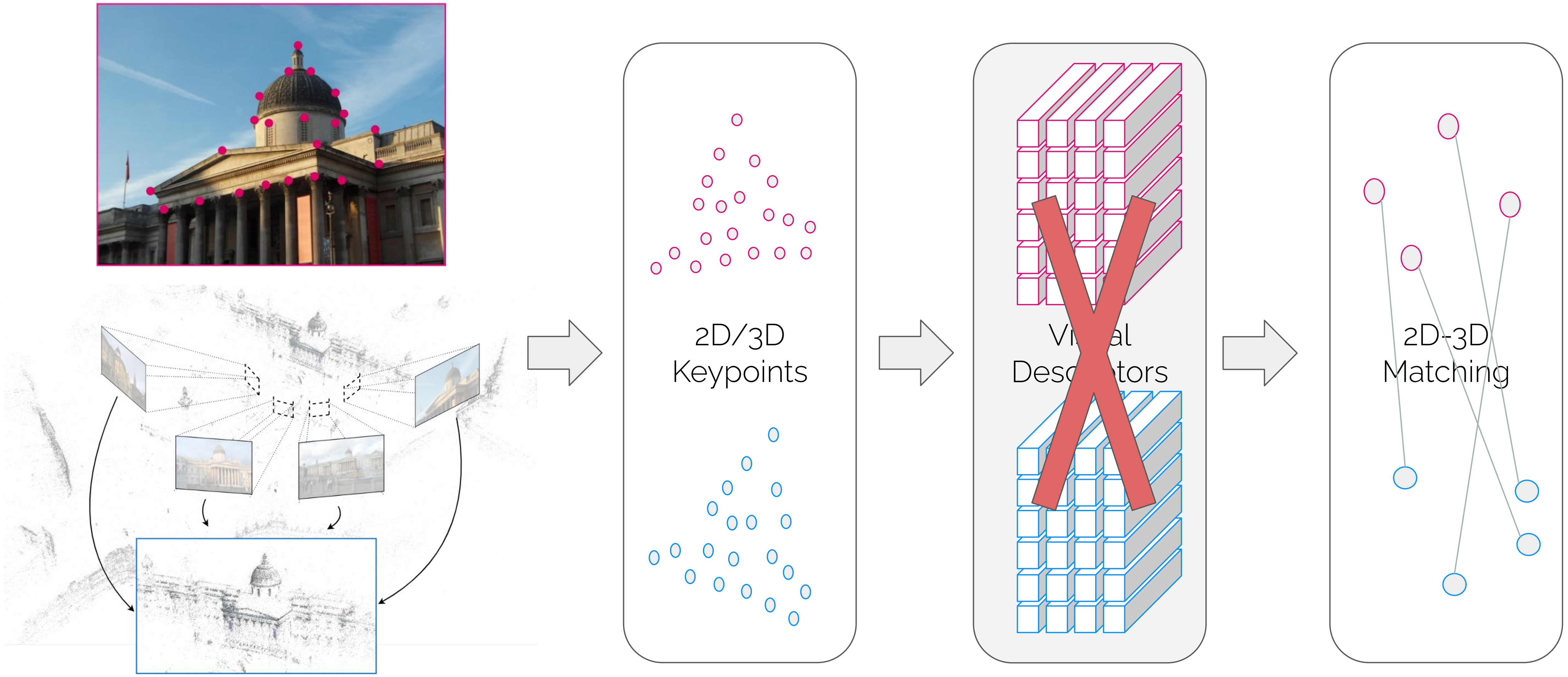


- Low Storage
- Privacy Preserving
- No Descriptor Maintenance

Scalable Large-scale Localization

Visual Localization

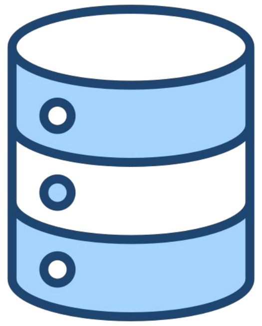




- Low Storage
- Privacy Preserving
- No Descriptor Maintenance

Scalable Large-scale Localization

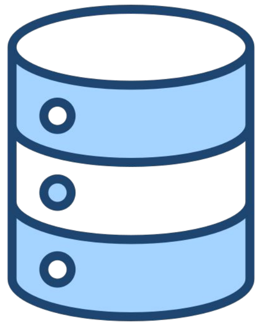
Practical Challenges



Storage Demand

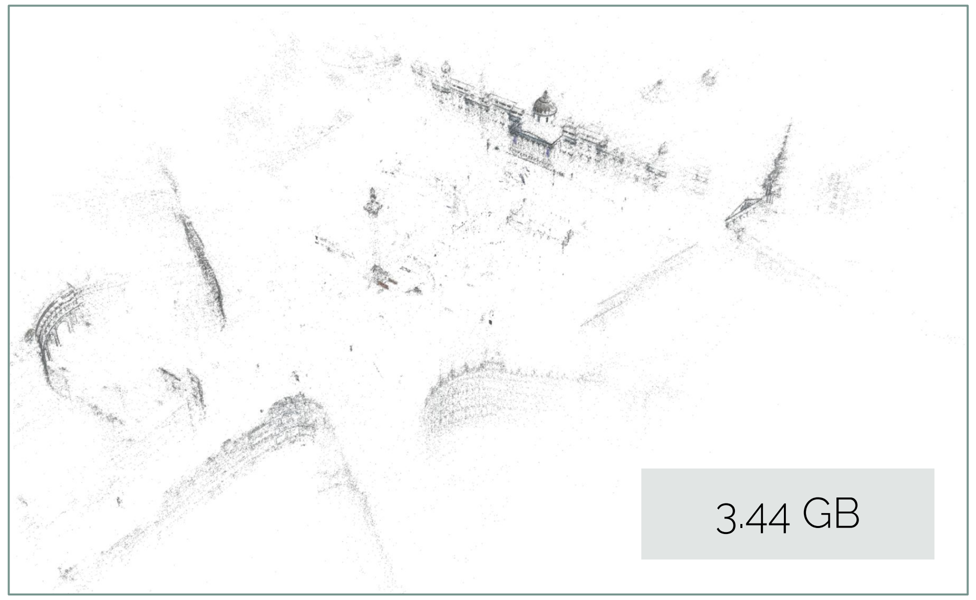
MegaDepth (192 scenes)	Storage	Desc Type	Data Type	Storage
Cameras	15.73 MB	SIFT	Uint8	133.33 GB
3D Points	**3.44 GB**	CAPS	FP32	523.83 GB
Images	157.84 GB	SuperPoint	FP32	**1.044 TB**

Practical Challenges

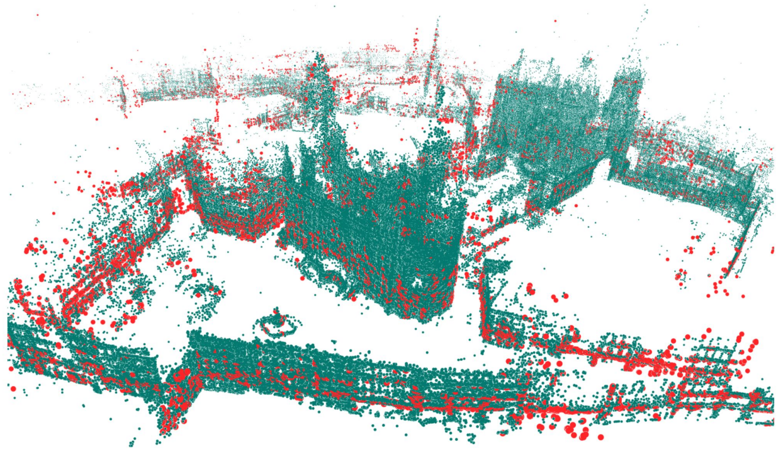


Storage Demand

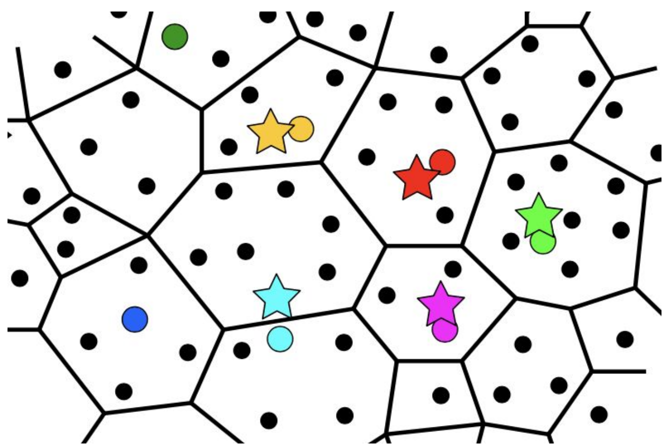
MegaDepth (192 scenes)	Storage	Desc Type	Data Type	Storage
Cameras	15.73 MB	SIFT	Uint8	133.33 GB
3D Points	**3.44 GB**	CAPS	FP32	523.83 GB
Images	157.84 GB	SuperPoint	FP32	**1.044 TB**



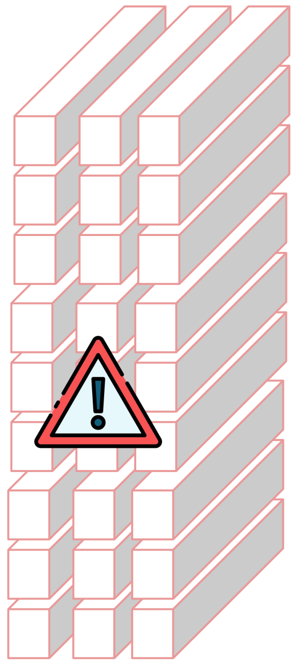
Scene Compression [1]



Desc Quantization [1, 2]



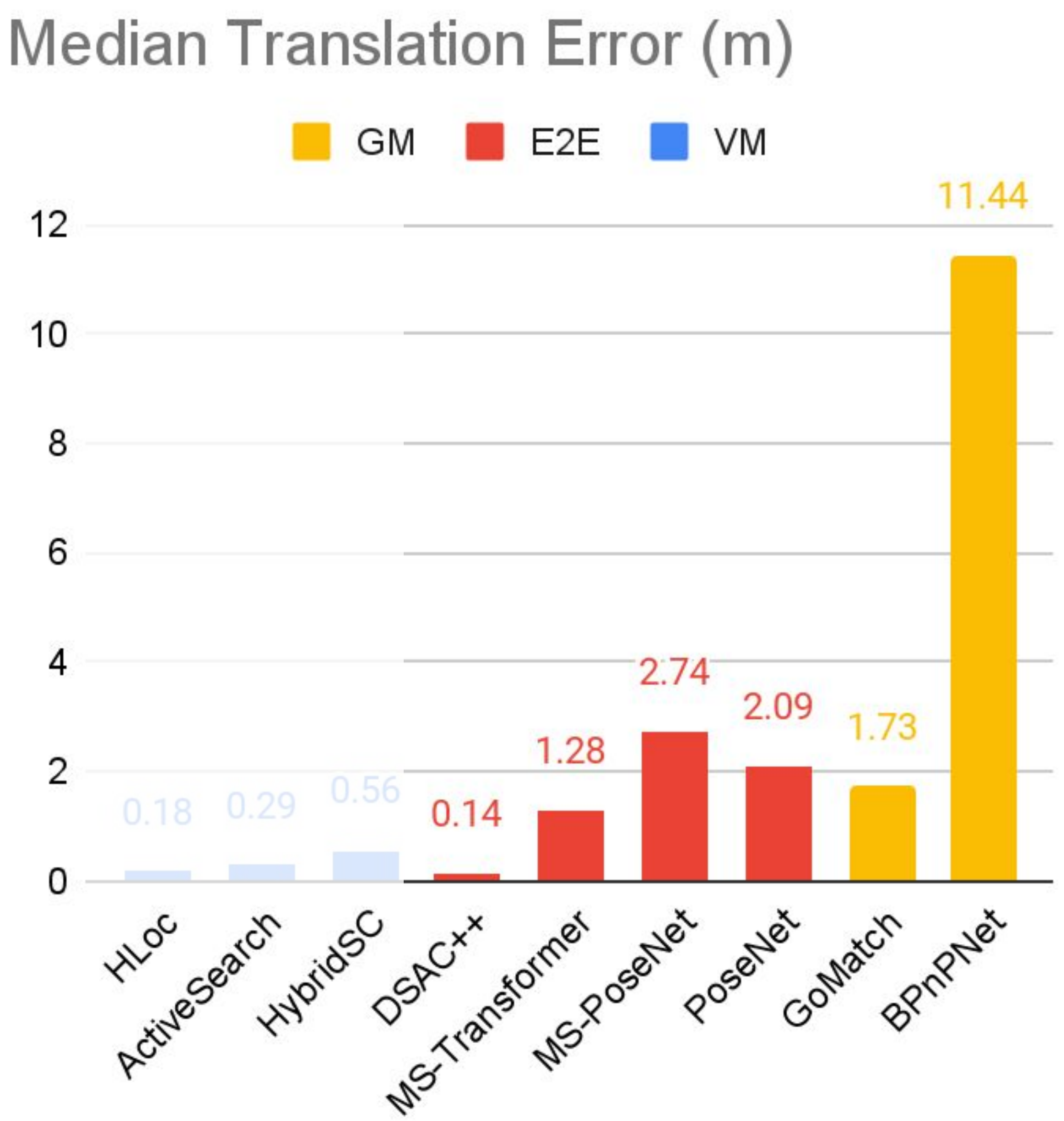
Scene Descriptor

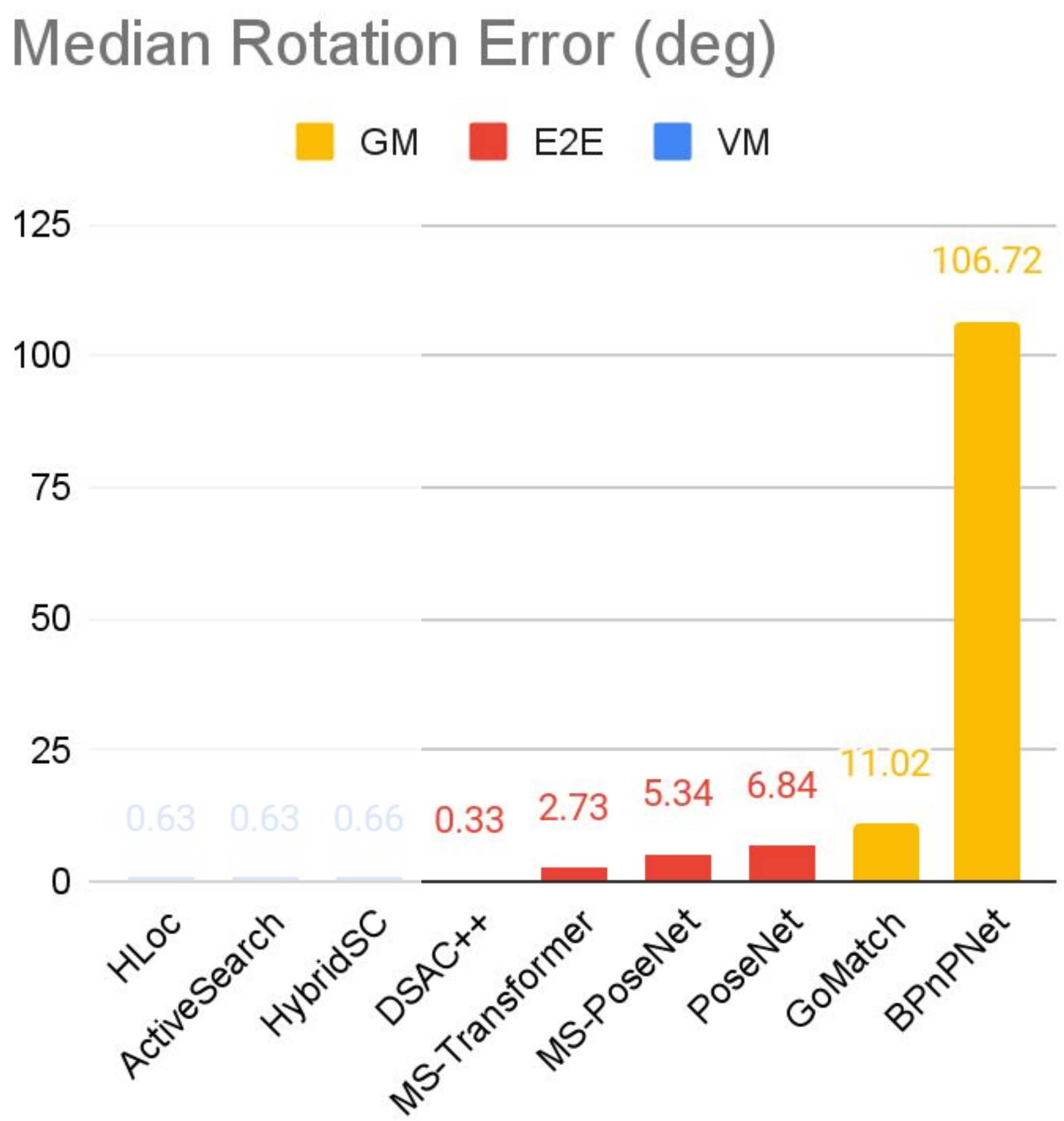


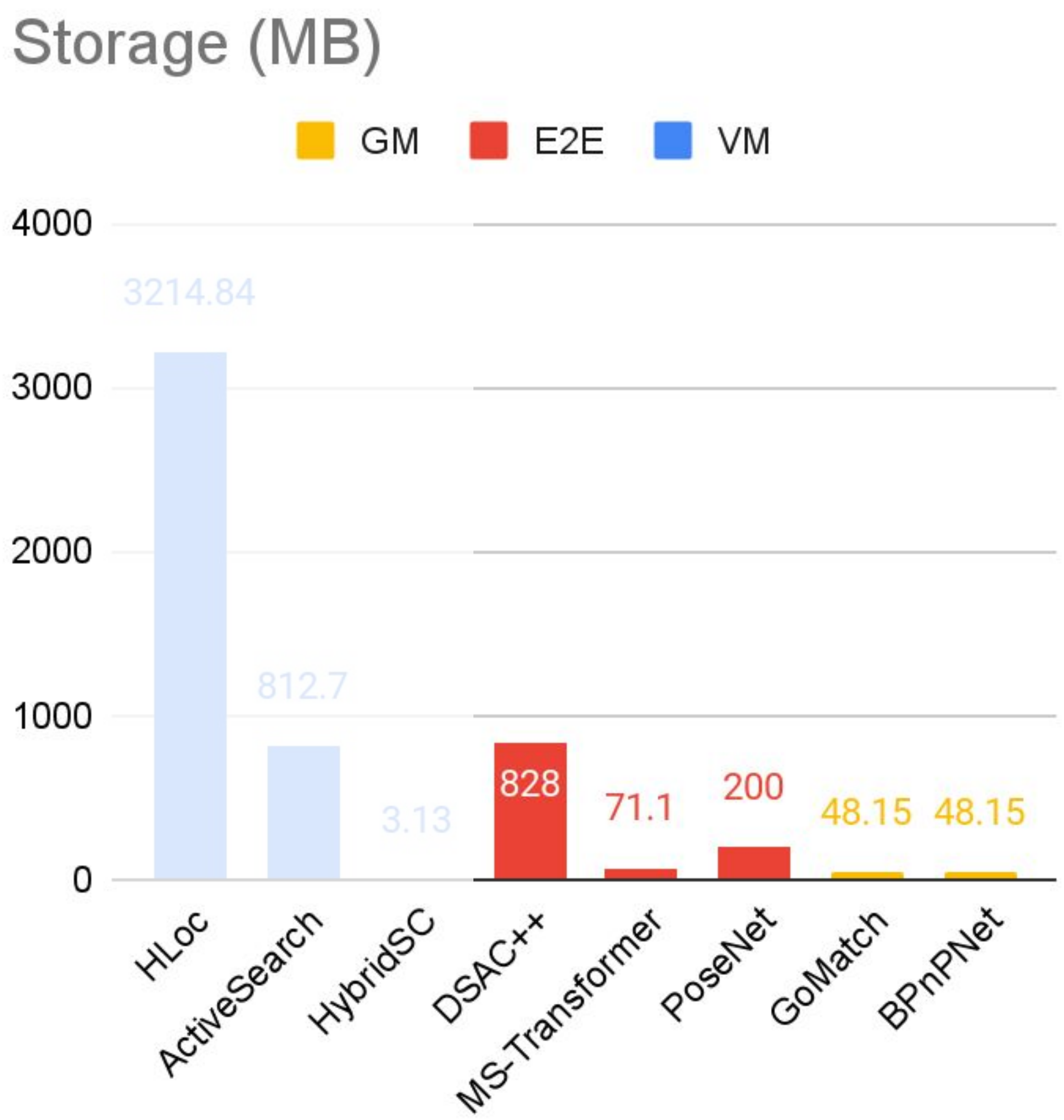
133 GB ~ 1.04 TB

[1] Camposeco, Federico, et al. "Hybrid scene compression for visual localization." CVPR19
[2] Sattler, Torsten, Bastian Leibe, and Leif Kobbelt. "Efficient & effective prioritized matching for large-scale image-based localization." PAMI16

Compare to E2E – Cambridge Landmarks

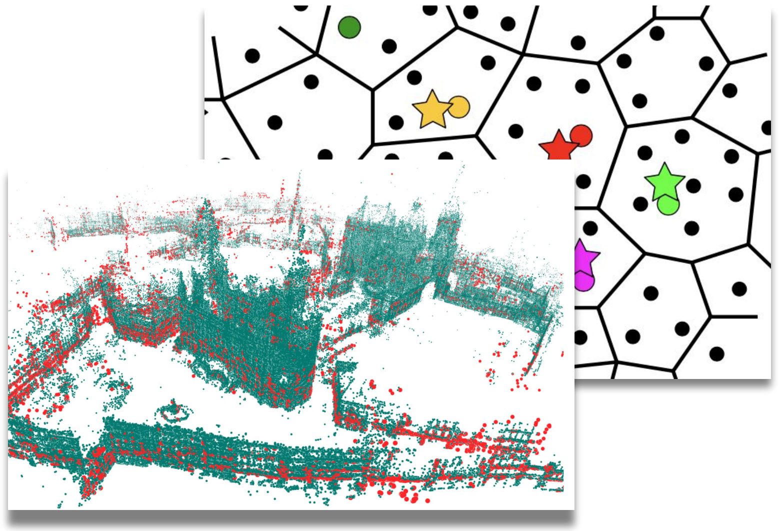




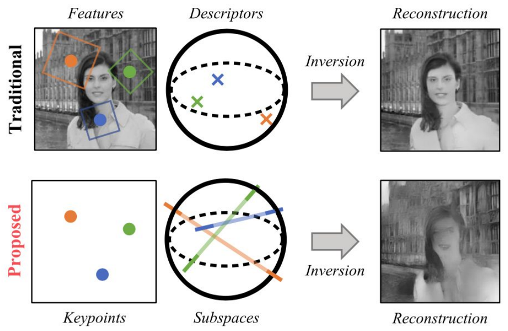


Existing Solutions

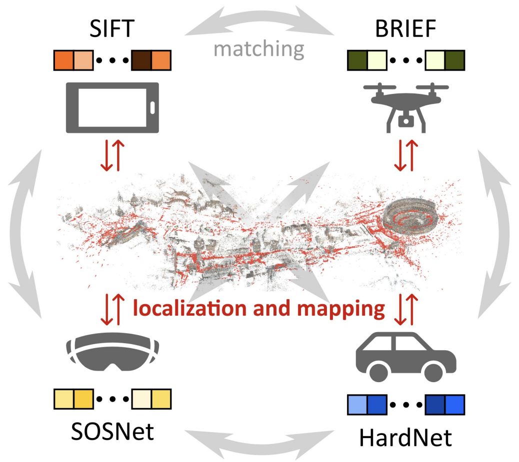
Storage / Memory Efficiency



Privacy Preserving



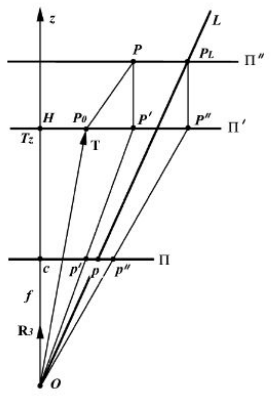
GoMatch



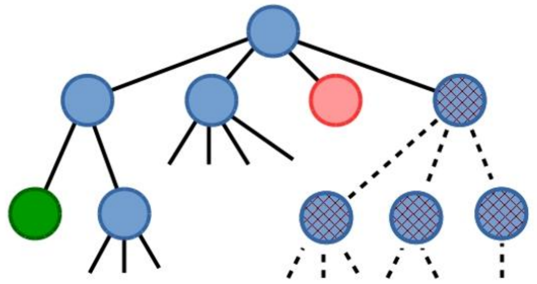
Descriptor Maintenance

Geometric-based matching and pose estimation

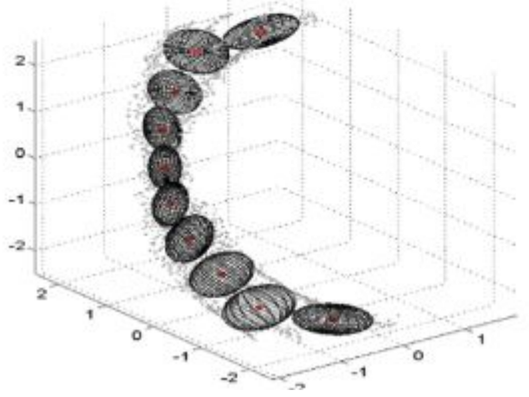
SoftPOSIT [1]
- Alternate step: softassign + POSIT
- Requires initialization
- Struggles with clutter, occlusions, repetitive patterns.
- Efficient



GOPAC [3]
- Globally optimal solution using Branch-and-Bound
- Prohibitive runtime requirements
- Cannot scale to large problems

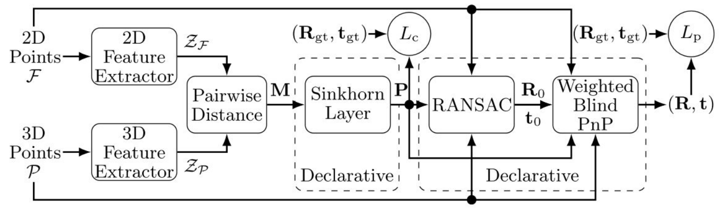


hegyhati.github.io



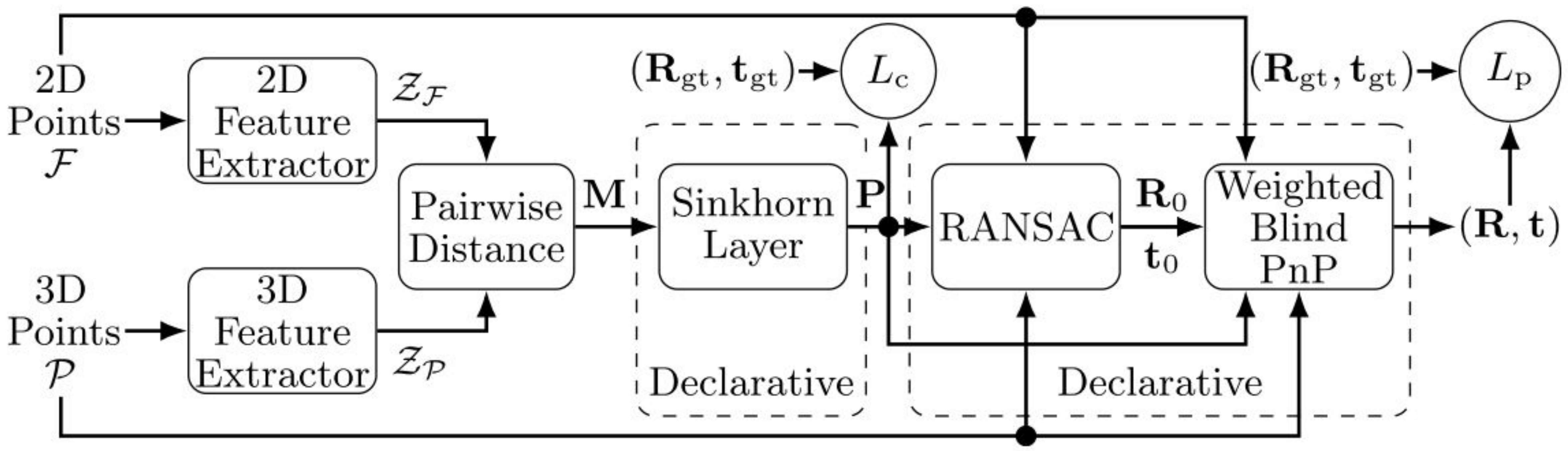
Bind PnP [2]
- Kalman-Filter to maintain correspondence hypotheses.
- Requires initialization of GMM pose priors
- Better handling of occlusion, clutter and repetitive patterns

BPnPNet [4]
- Learning-based geometric matching network
- Declarative layers to back propagate through Sinkhorn, RANSAC and the PnP solver.
- Performance substantially degraded with outliers.



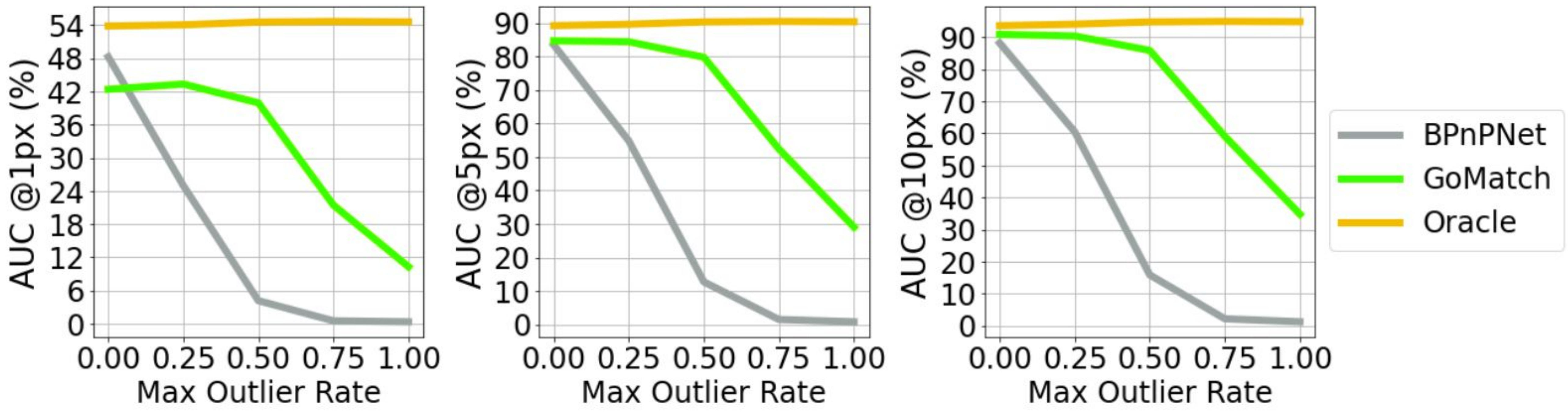
[1] David, Philip, et al. "SoftPOSIT: Simultaneous pose and correspondence determination." IJCV 2004
[2] Moreno-Noguer, Francesc et al. "Pose priors for simultaneously solving alignment and correspondence." ECCV 2008
[3] Campbell, Dylan, et al. "Globally-optimal inlier set maximisation for camera pose and correspondence estimation." PAMI 2018
[4] Campbell, Dylan, et al. "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization." ECCV 2020.

Geometric-Only Methods

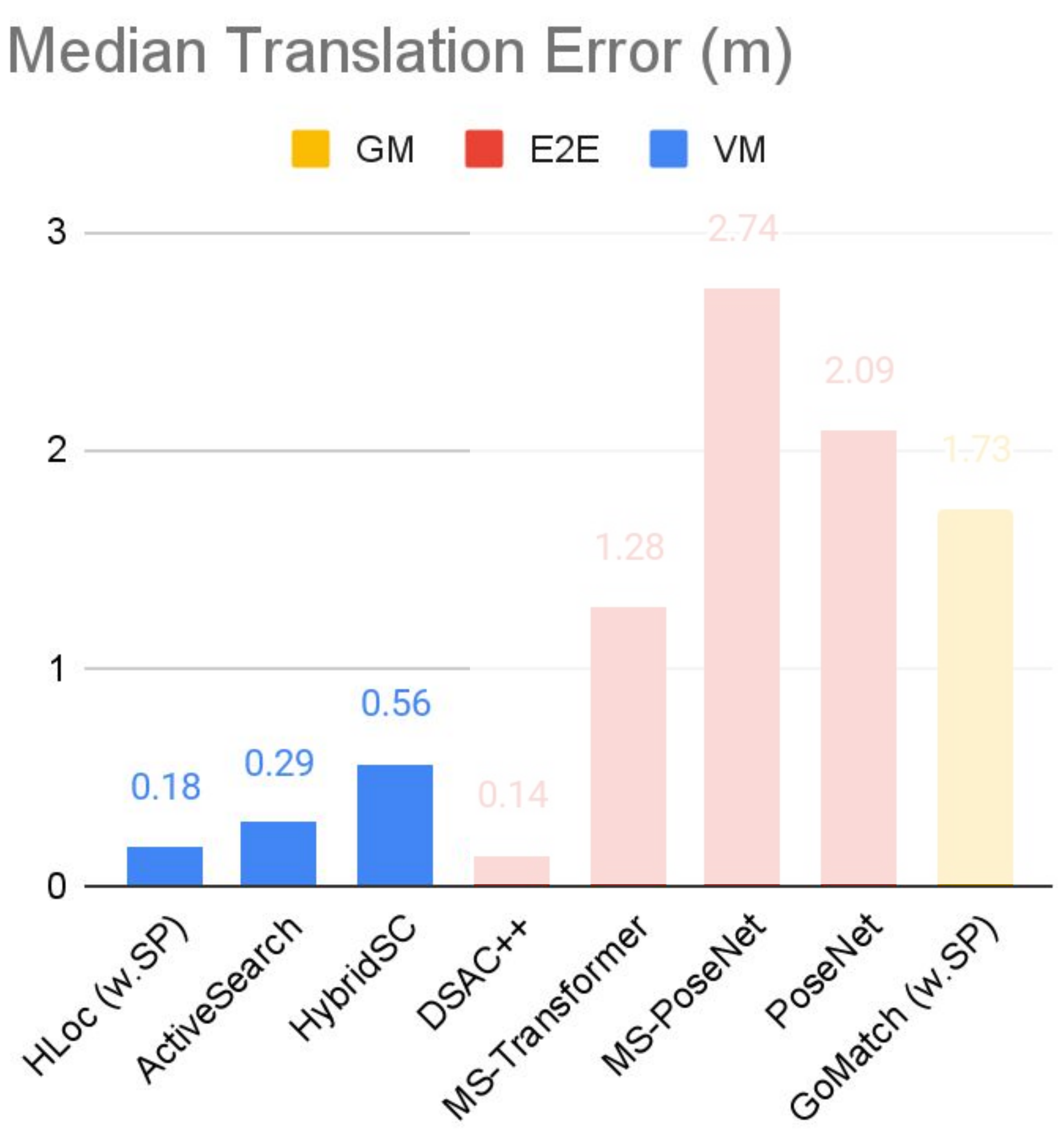


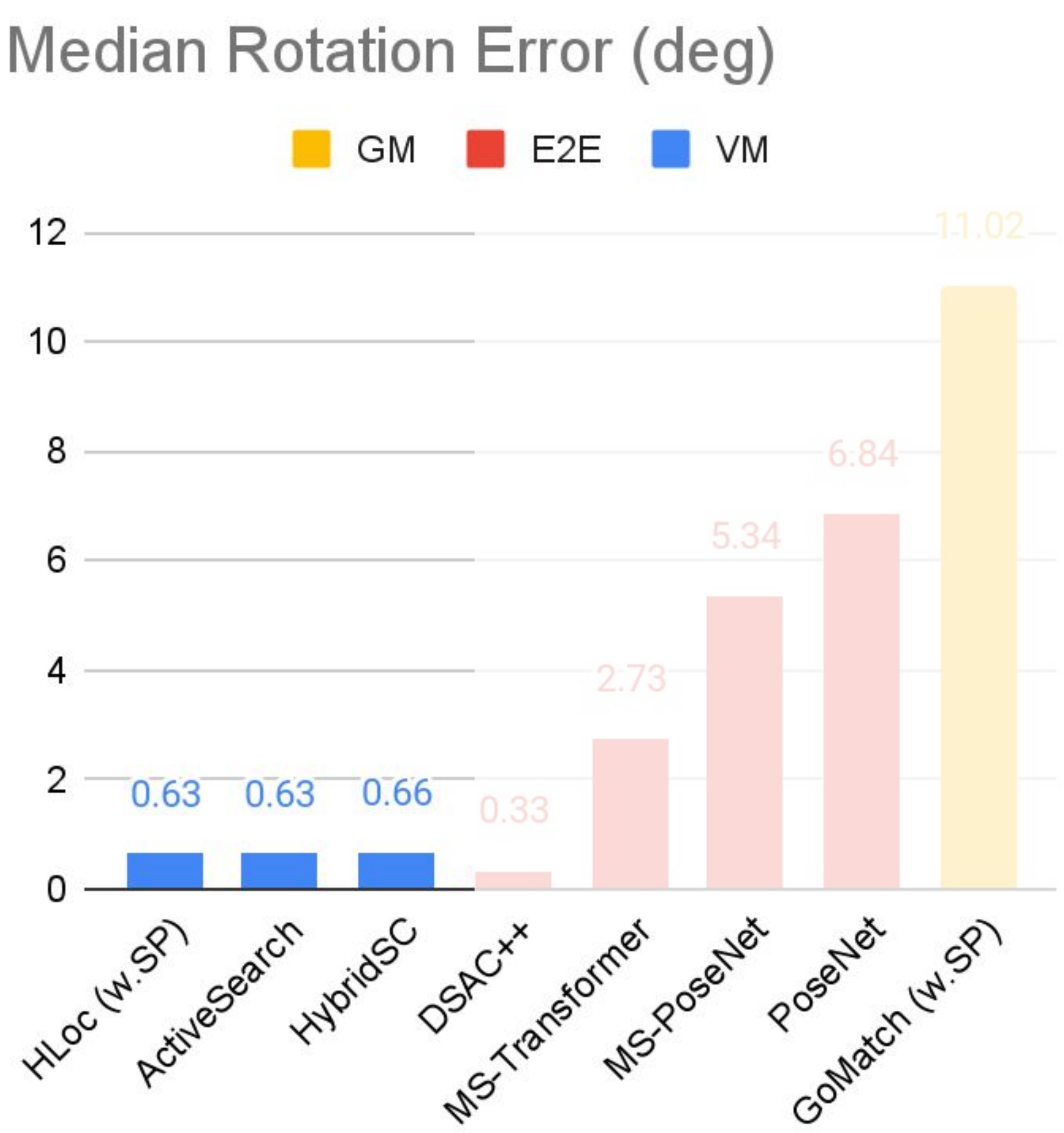
Liu Liu, et al. "Learning 2d-3d correspondences to solve the blind perspective-n-point problem." arXiv20

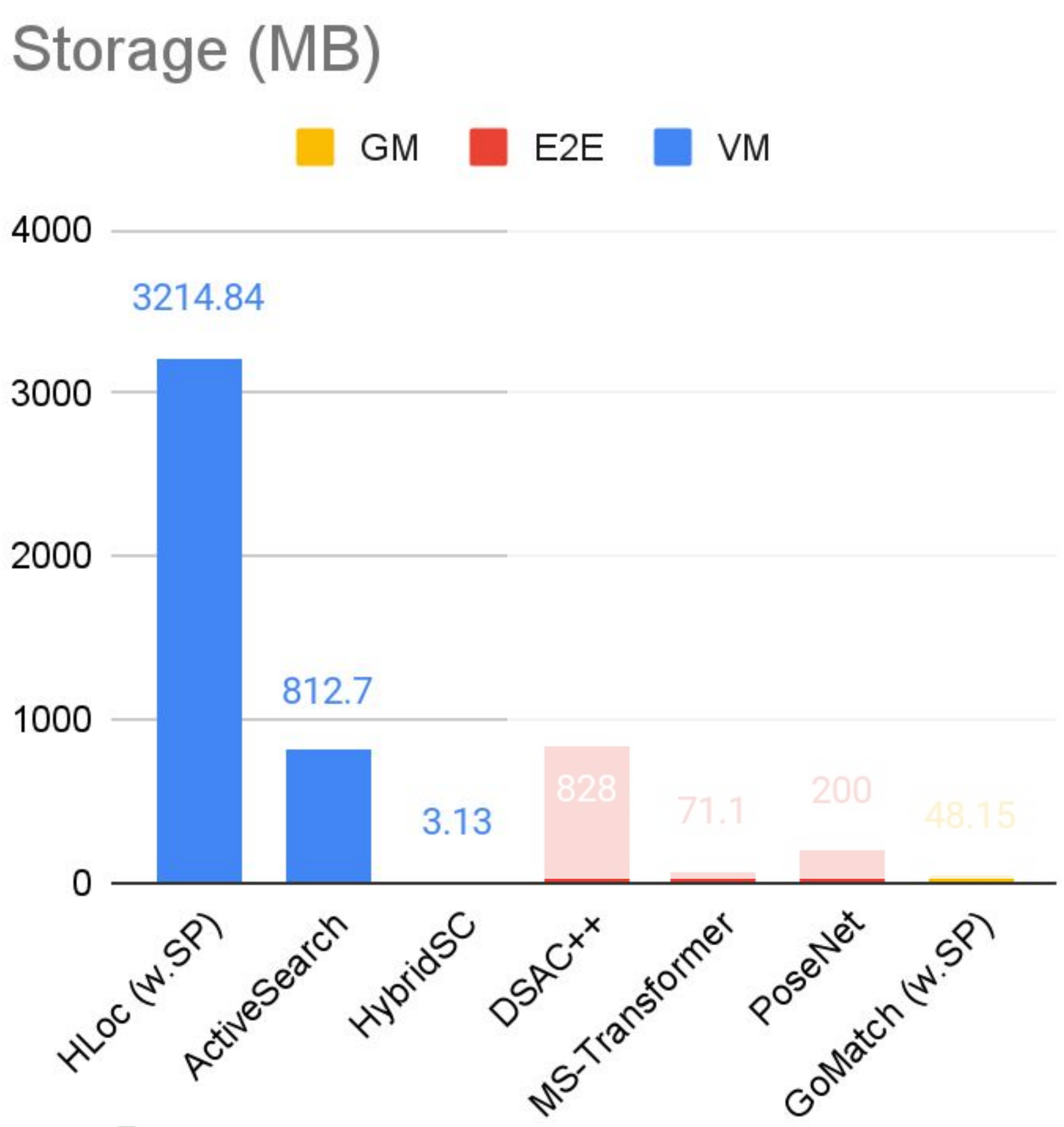
Dylan Campbell, Liu Liu, and Stephen Gould. "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization." ECCV20

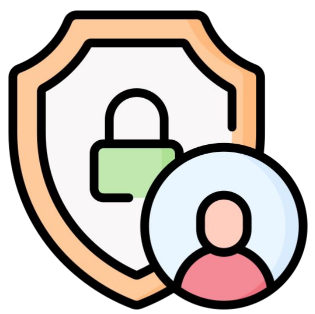


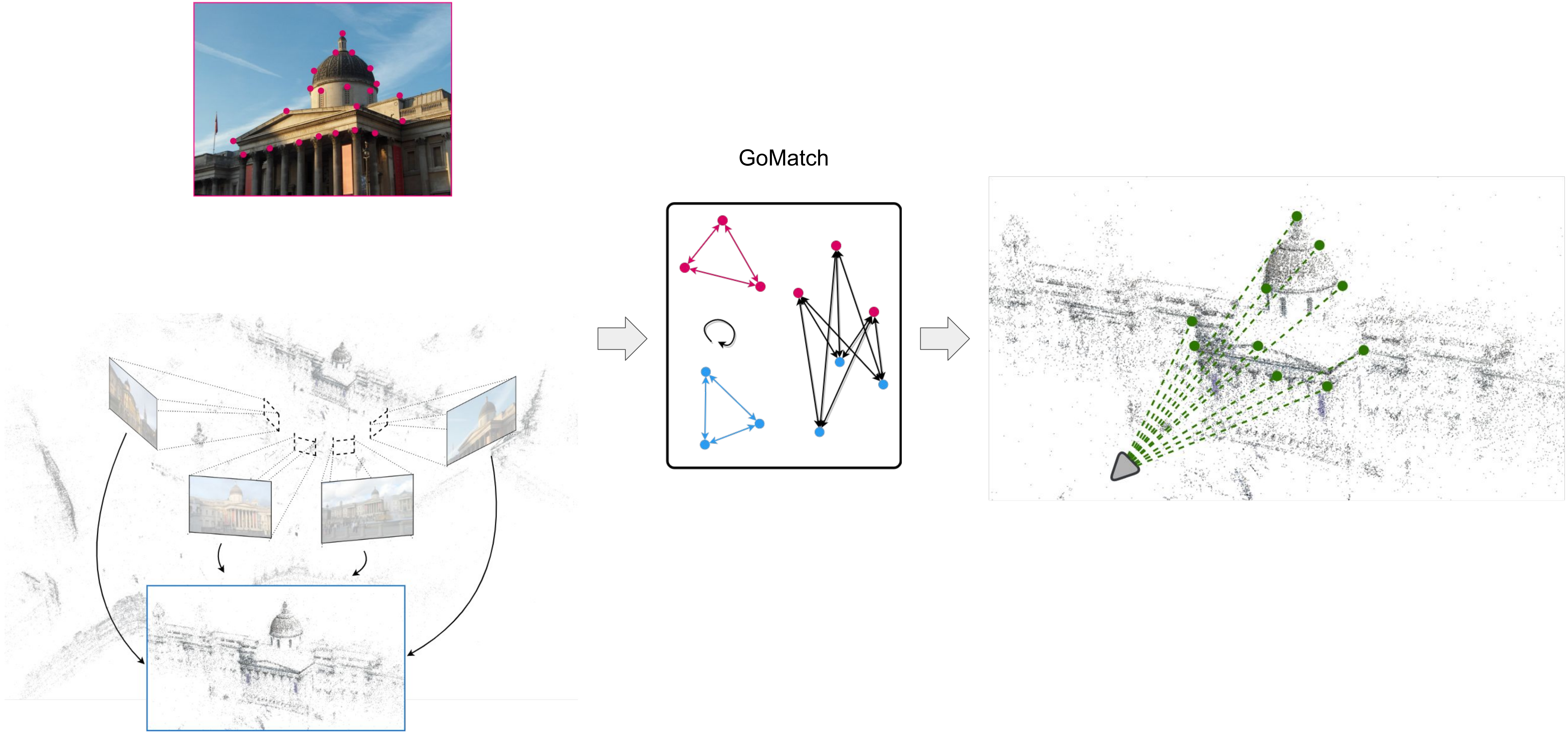
Outdoor Scene – Cambridge Landmarks



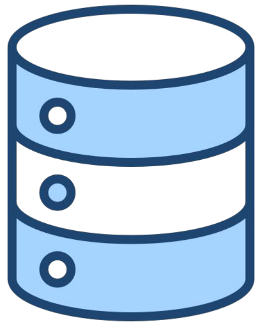






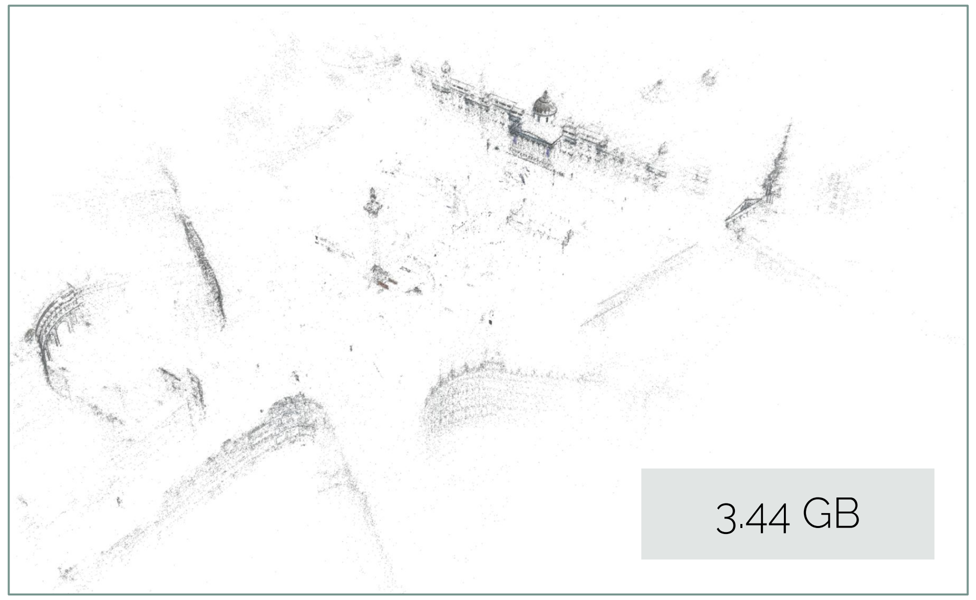

GoMatch

Practical Challenges

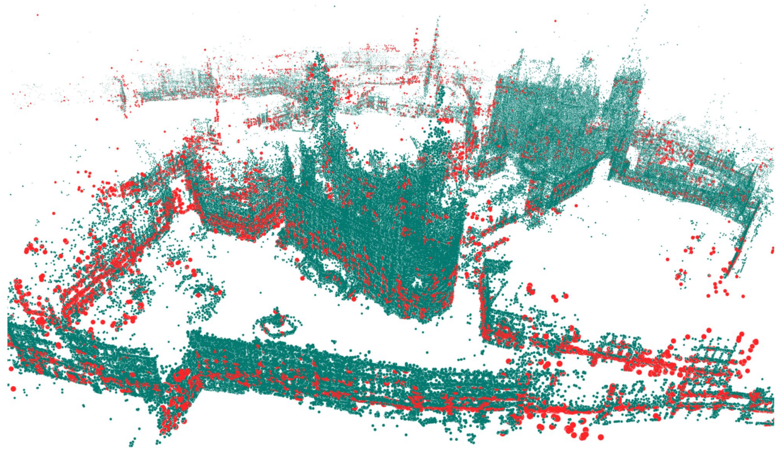


Storage Demand

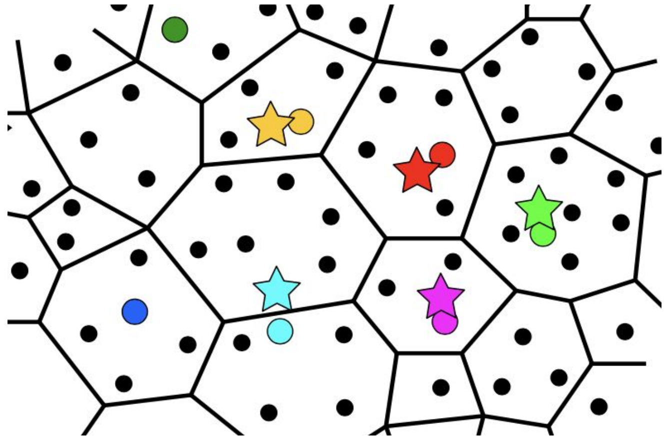
MegaDepth	Storage	Desc Type	Data Type	Storage
Cameras	15.73 MB	SIFT	Uint8	133.33 GB
3D Points *(192 scenes)*	3.44 GB	CAPS	FP32	523.83 GB
Images (*192* scenes)	157.84 GB	SuperPoint	FP32	1.044 TB



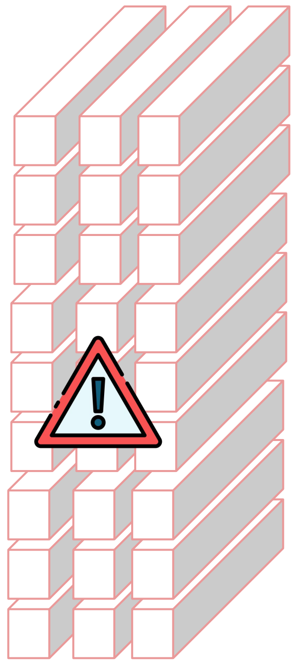
Scene Compression [1]



Desc Quantization [1, 2]

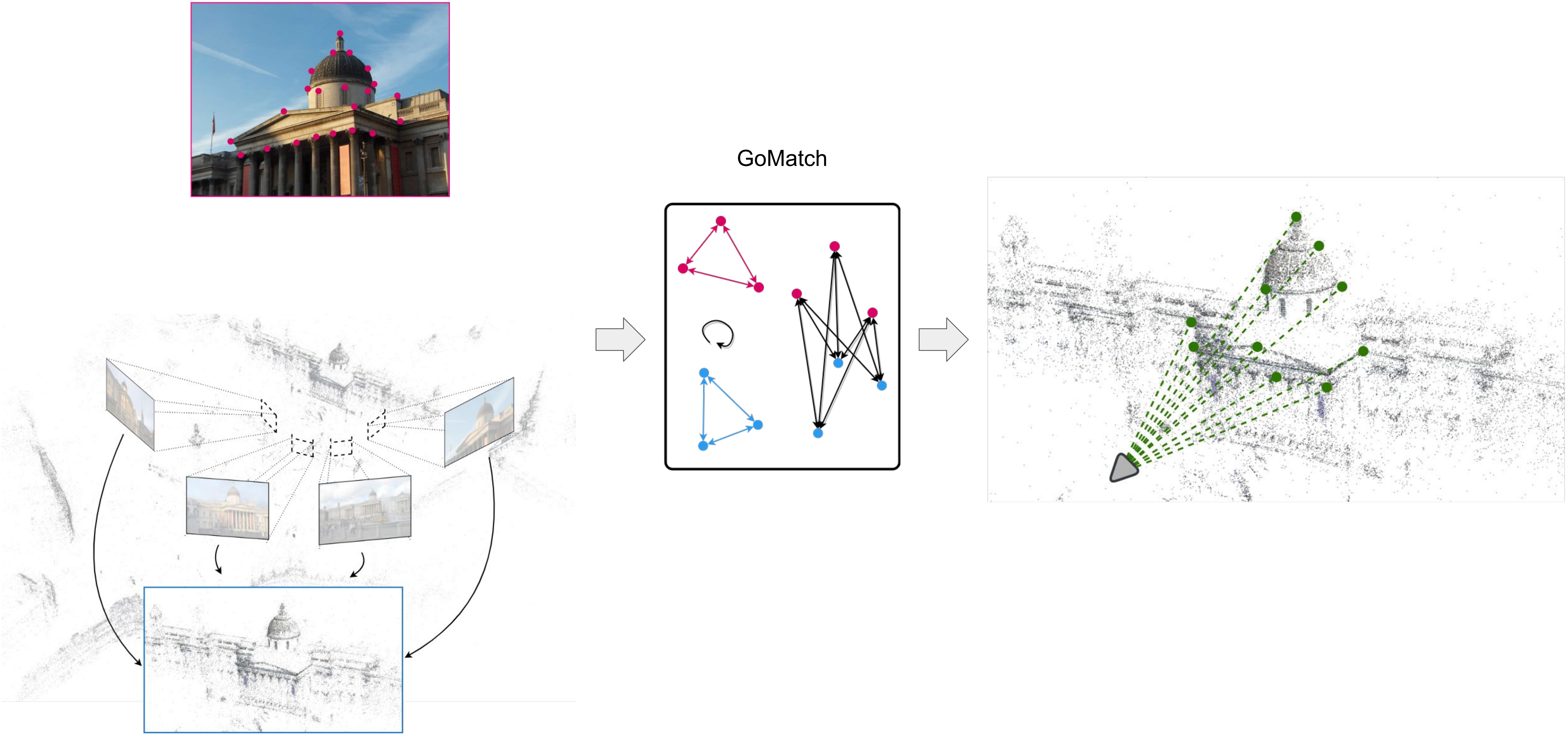


Scene Descriptor

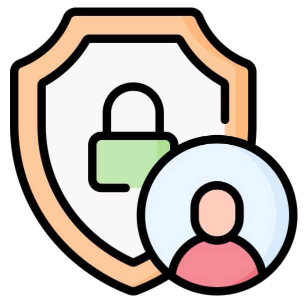


133 GB ~ 1.04 TB

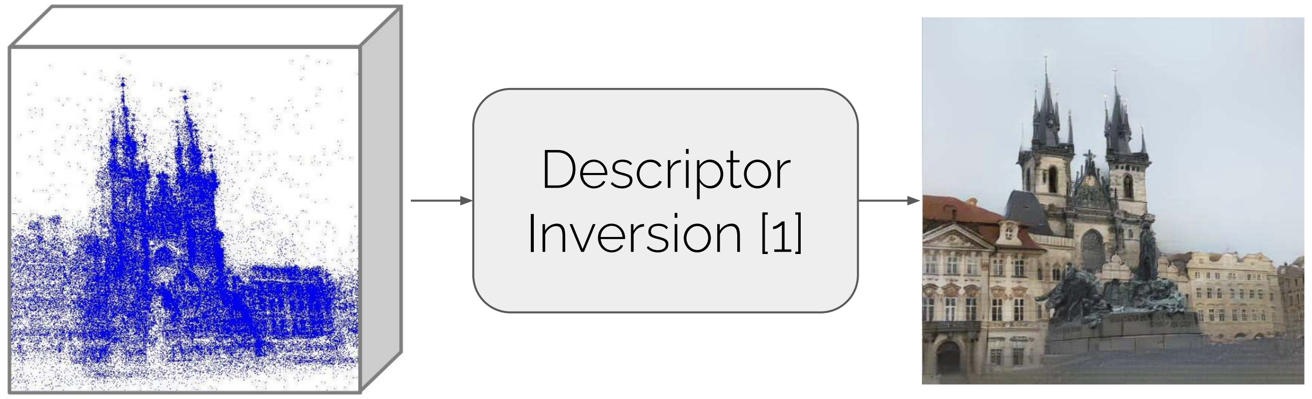
[1] Camposeco, Federico, et al. "Hybrid scene compression for visual localization." CVPR19
[2] Sattler, Torsten, Bastian Leibe, and Leif Kobbelt. "Efficient & effective prioritized matching for large-scale image-based localization." PAMI16

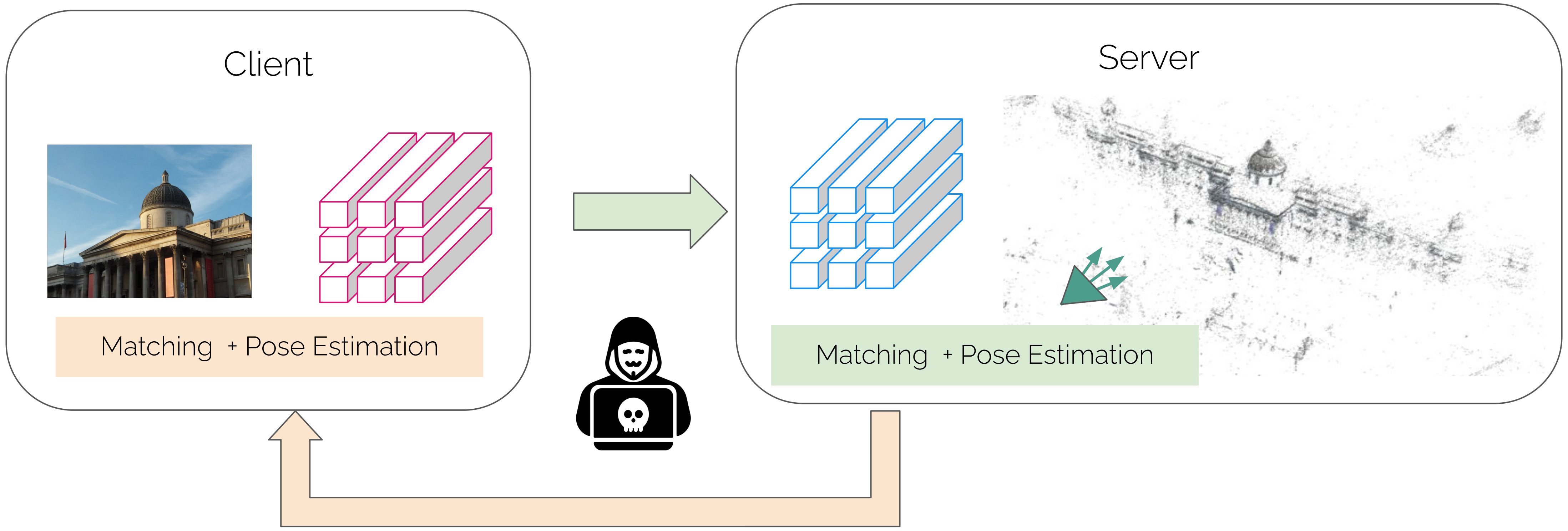

GoMatch

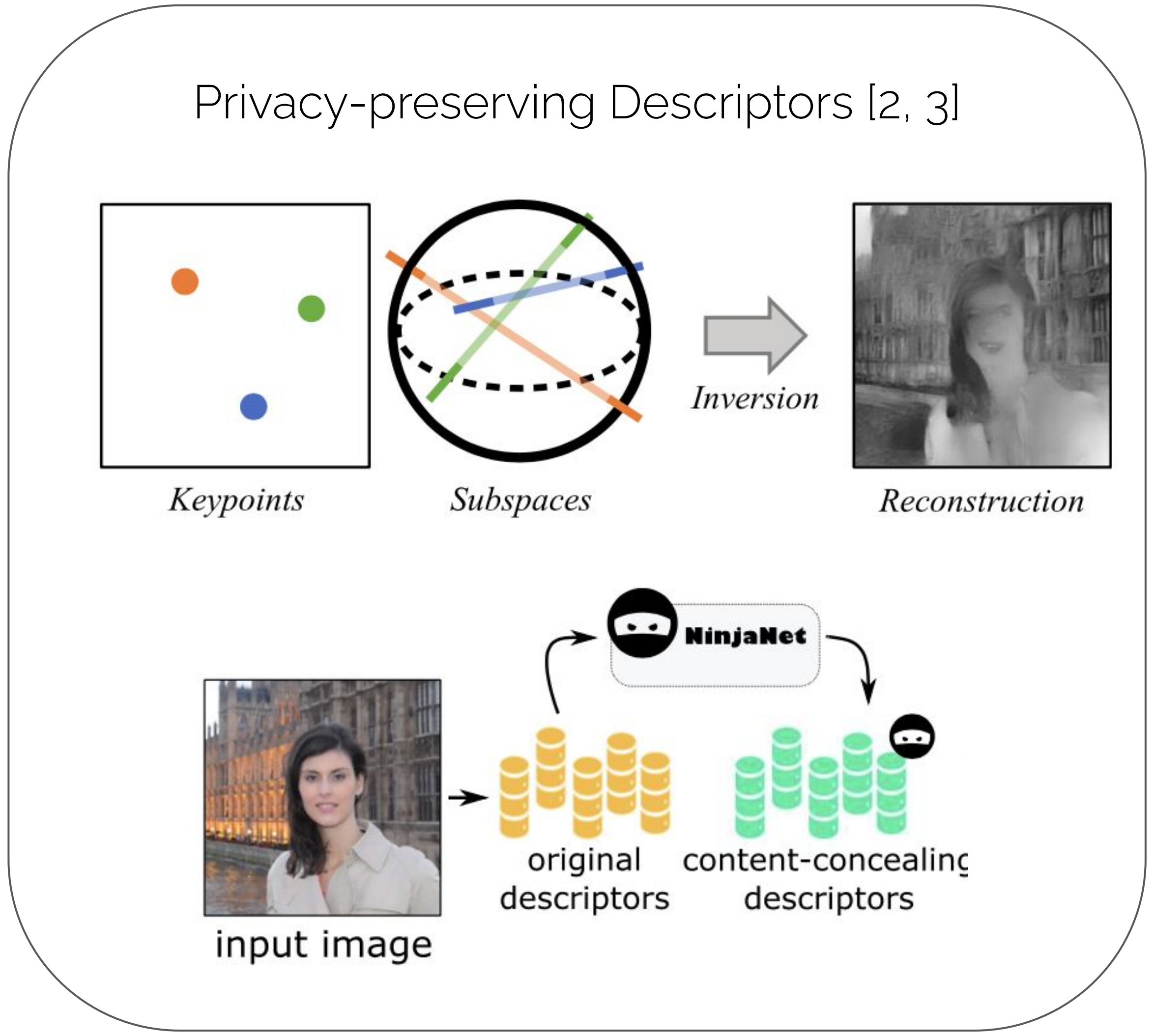
Practical Challenges



Privacy Risk



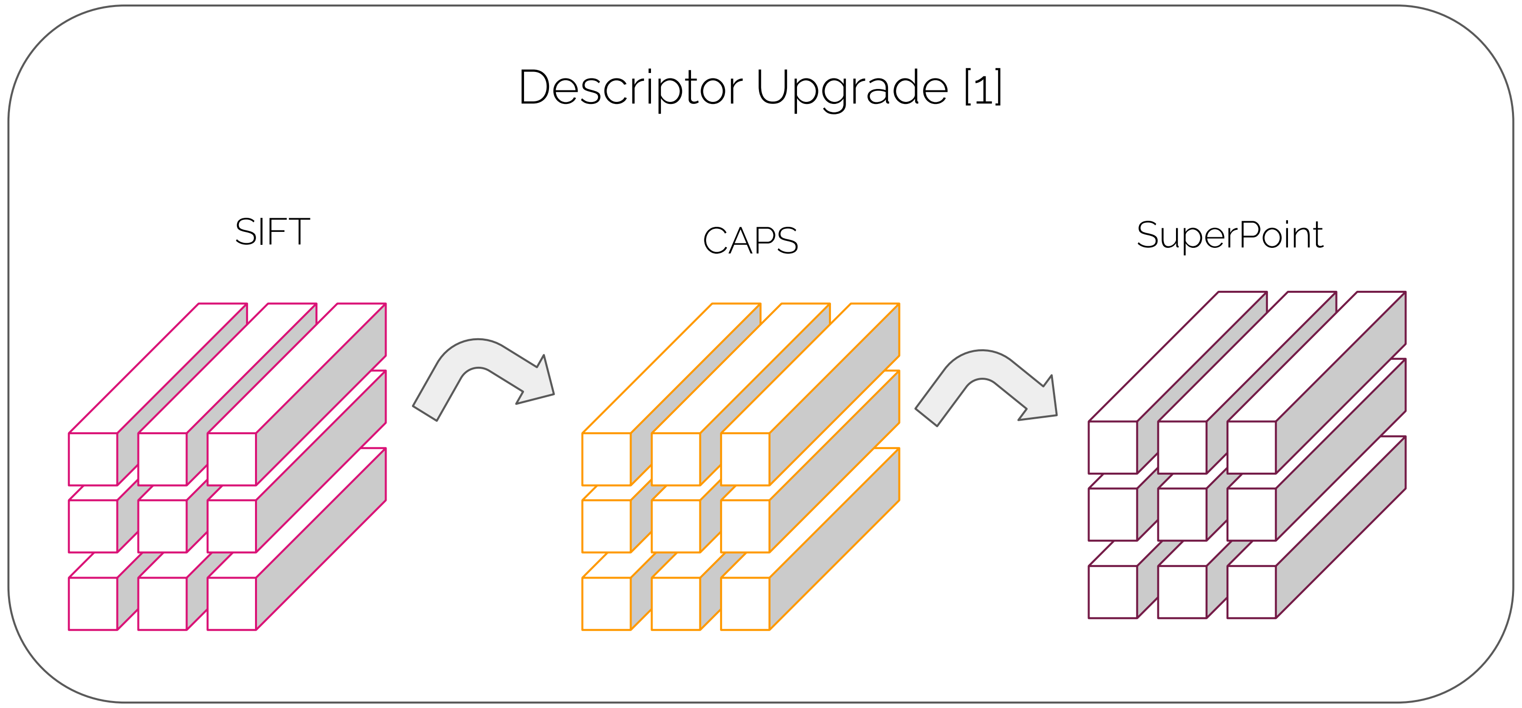


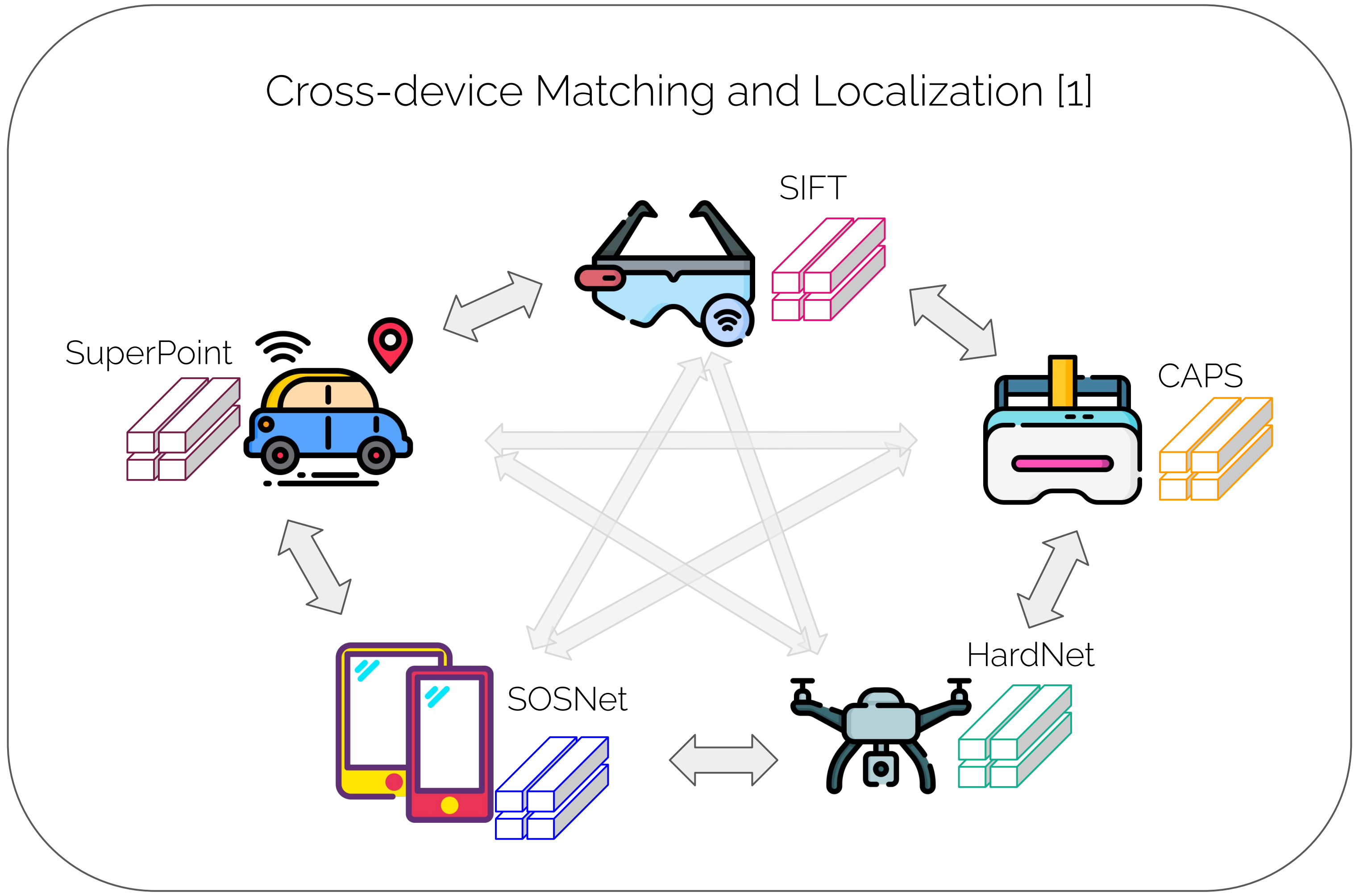


[1] Francesco, Pittaluga, et al Revealing Scenes by Inverting Structure From Motion Reconstructions. CVPR19
[2] Dusmanu, Mihai, et al. "Privacy-preserving image features via adversarial affine subspace embeddings." CVPR21.
[3] Ng, Tony, et al. "NinjaDesc: Content-Concealing Visual Descriptors via Adversarial Learning." CVPR22

Practical Challenges

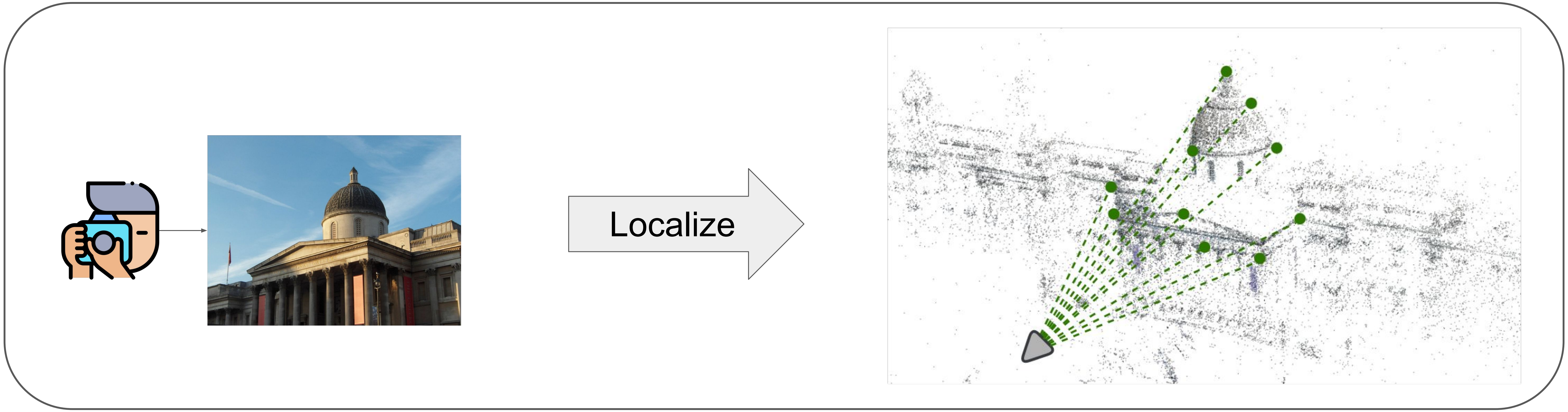
Maintenance
Complexity

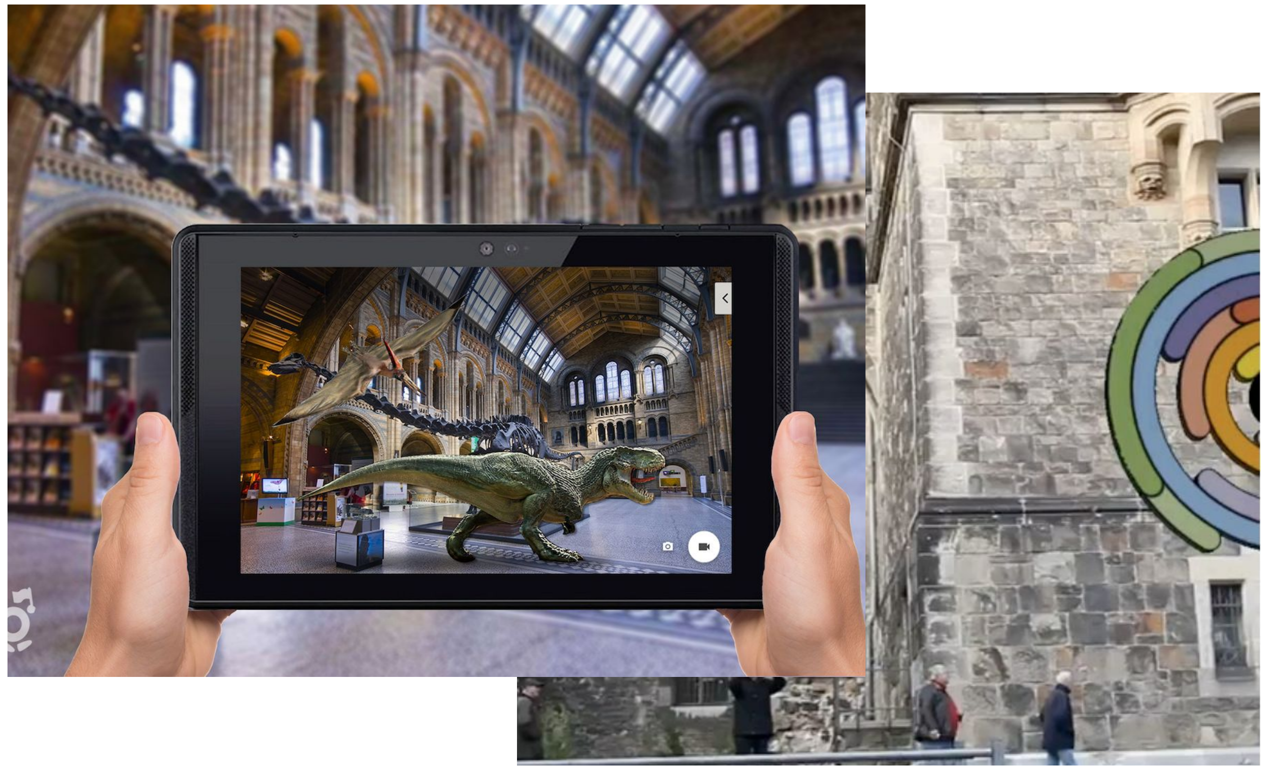




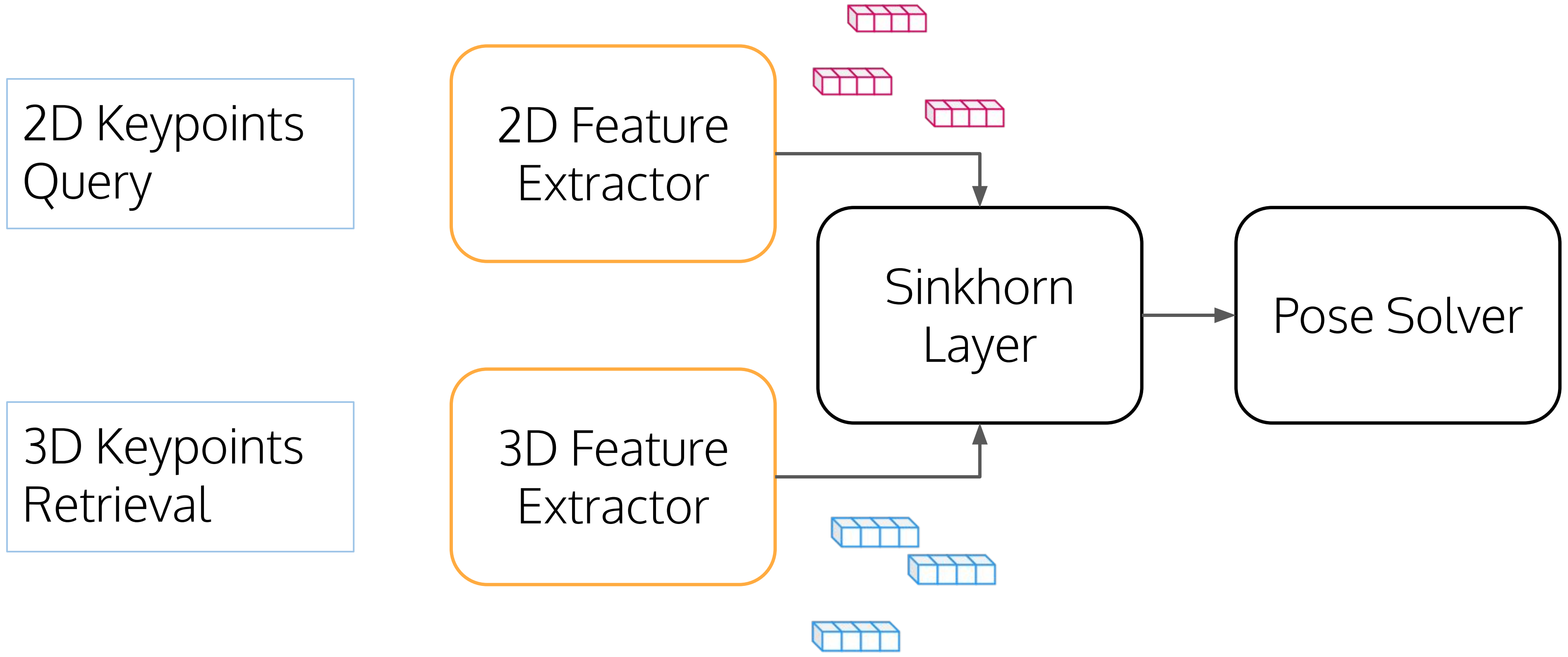
[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

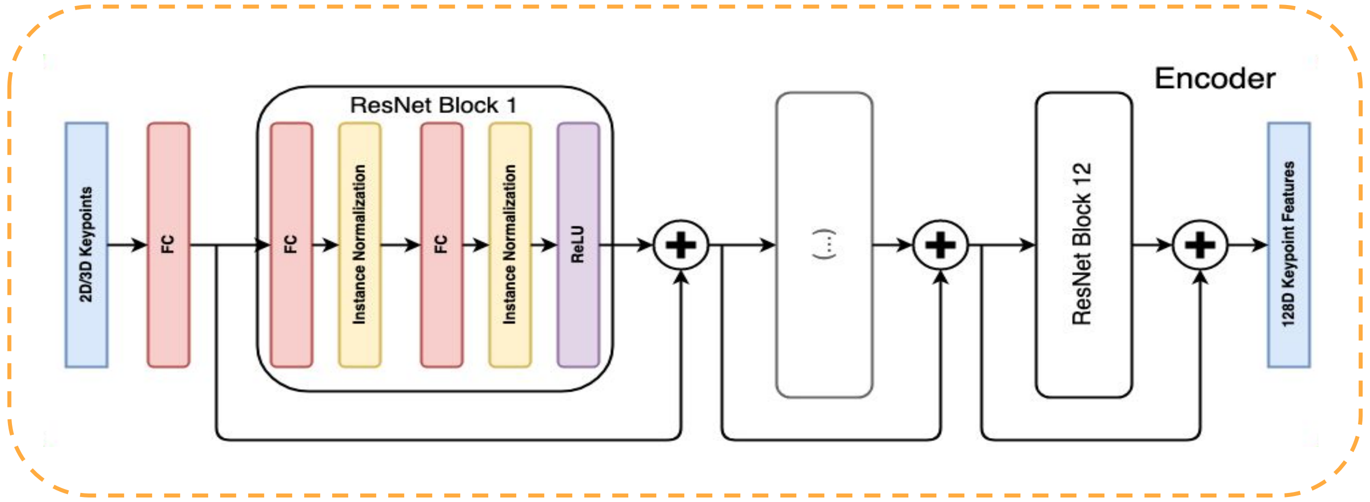
Visual Localization

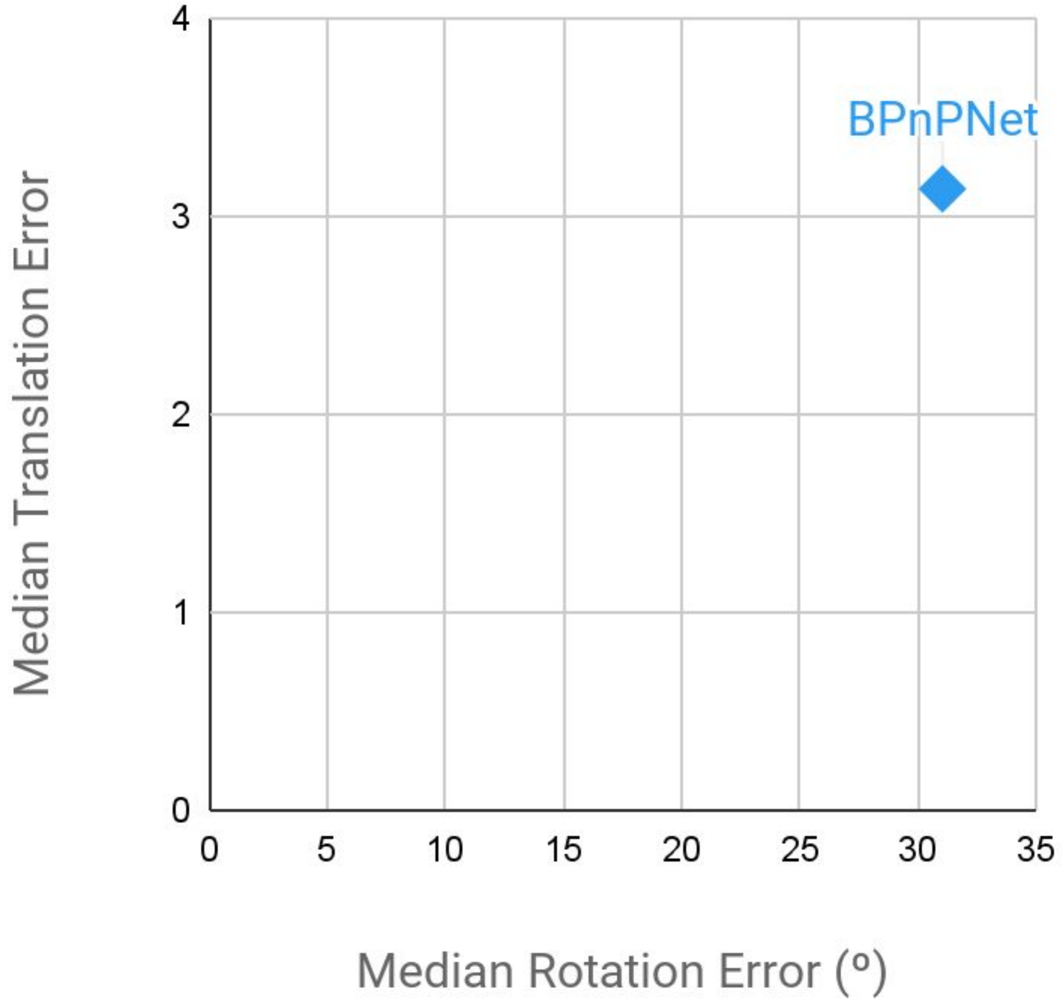




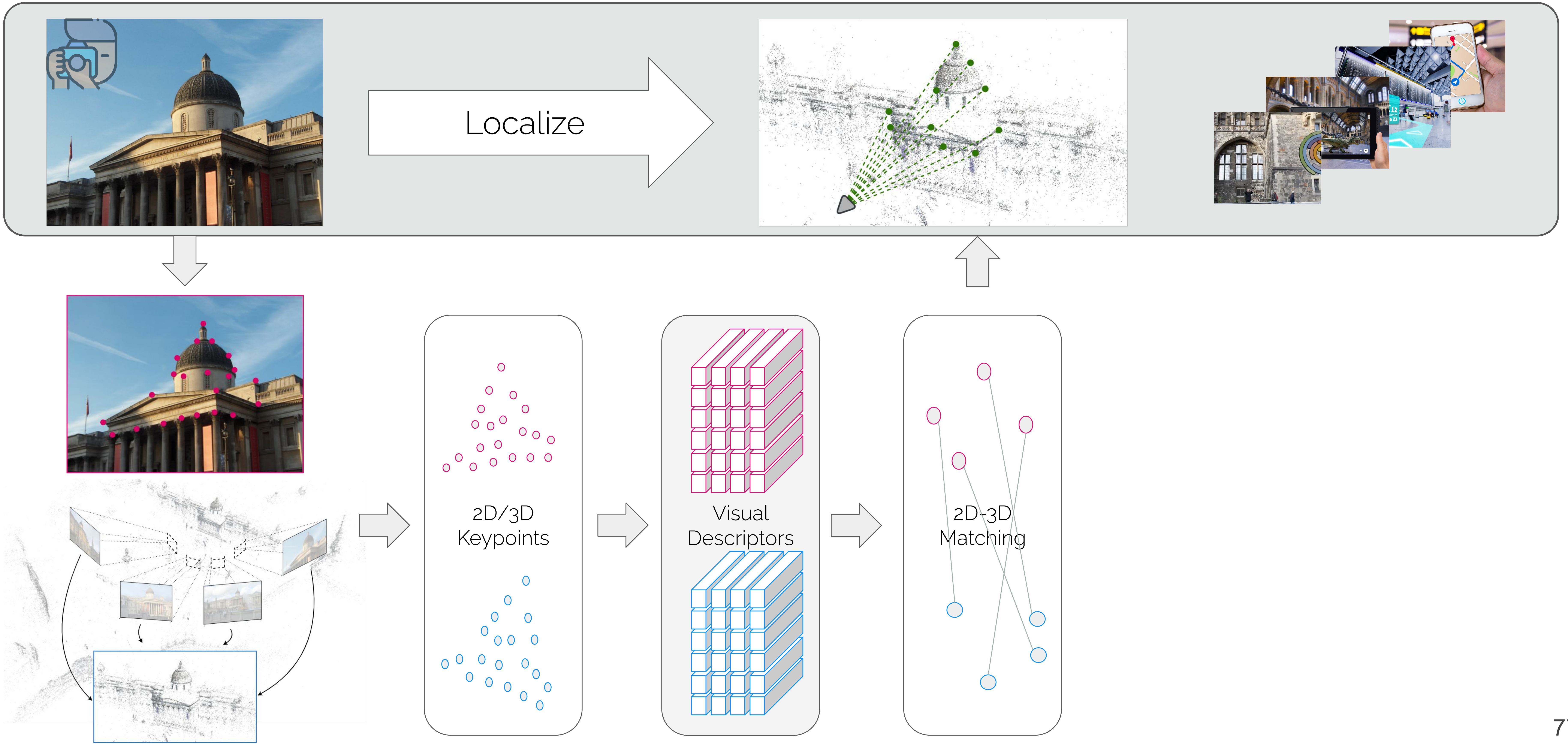
GoMatch Step-by-Step



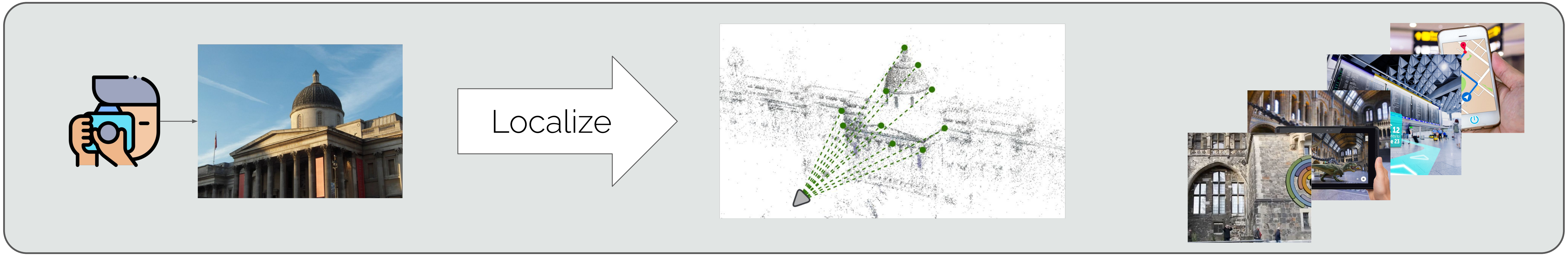


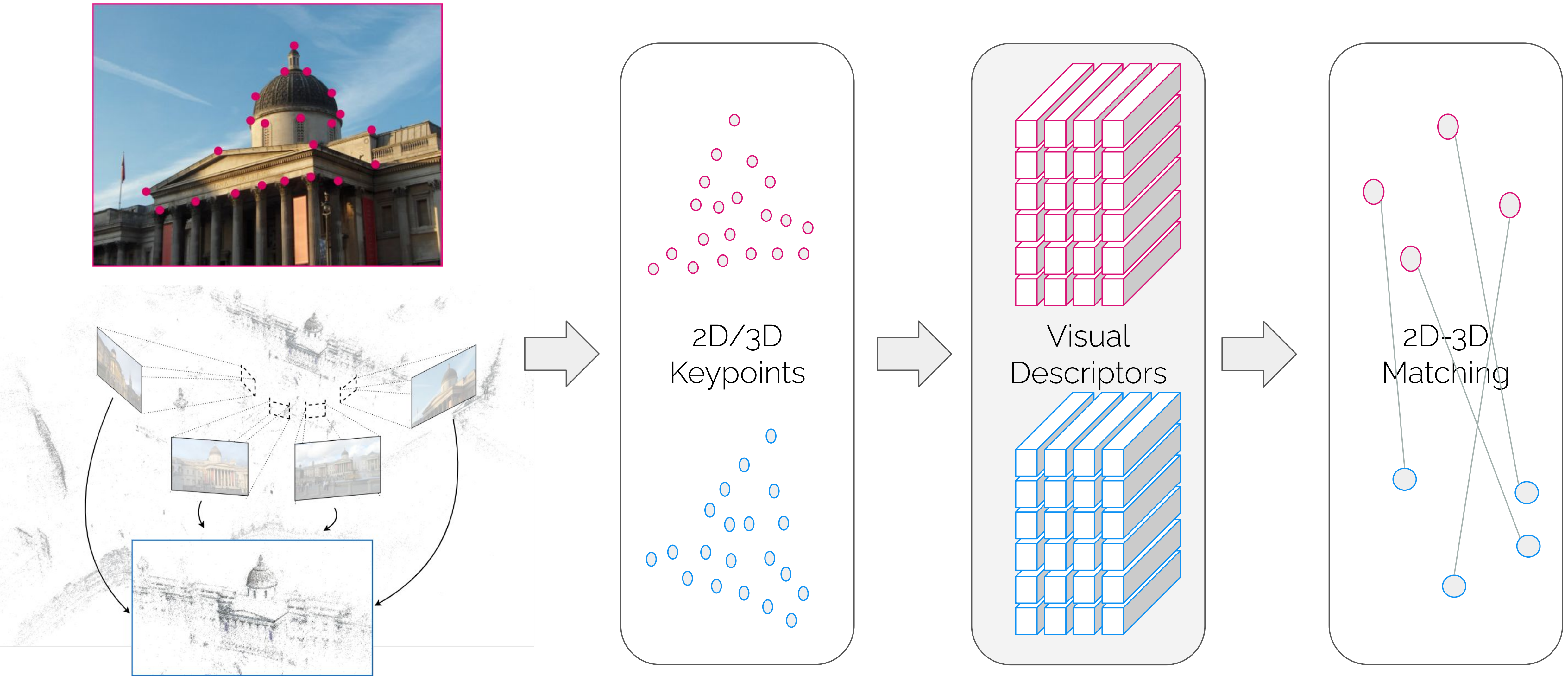


Introduction

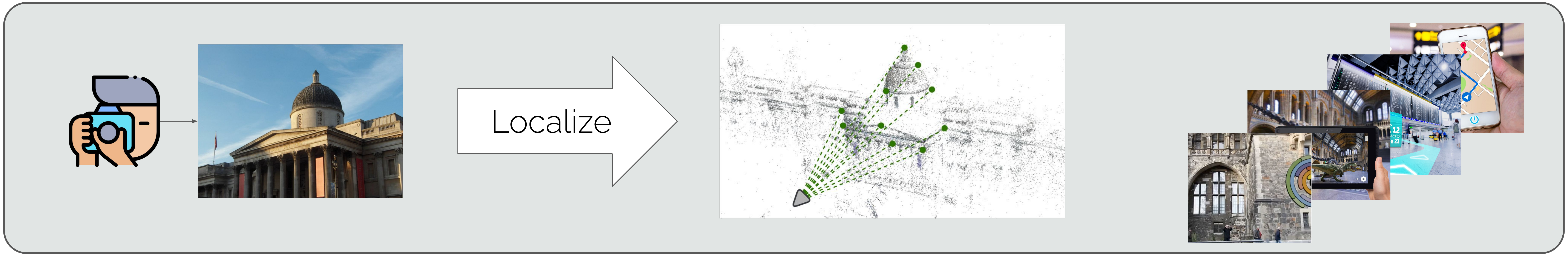


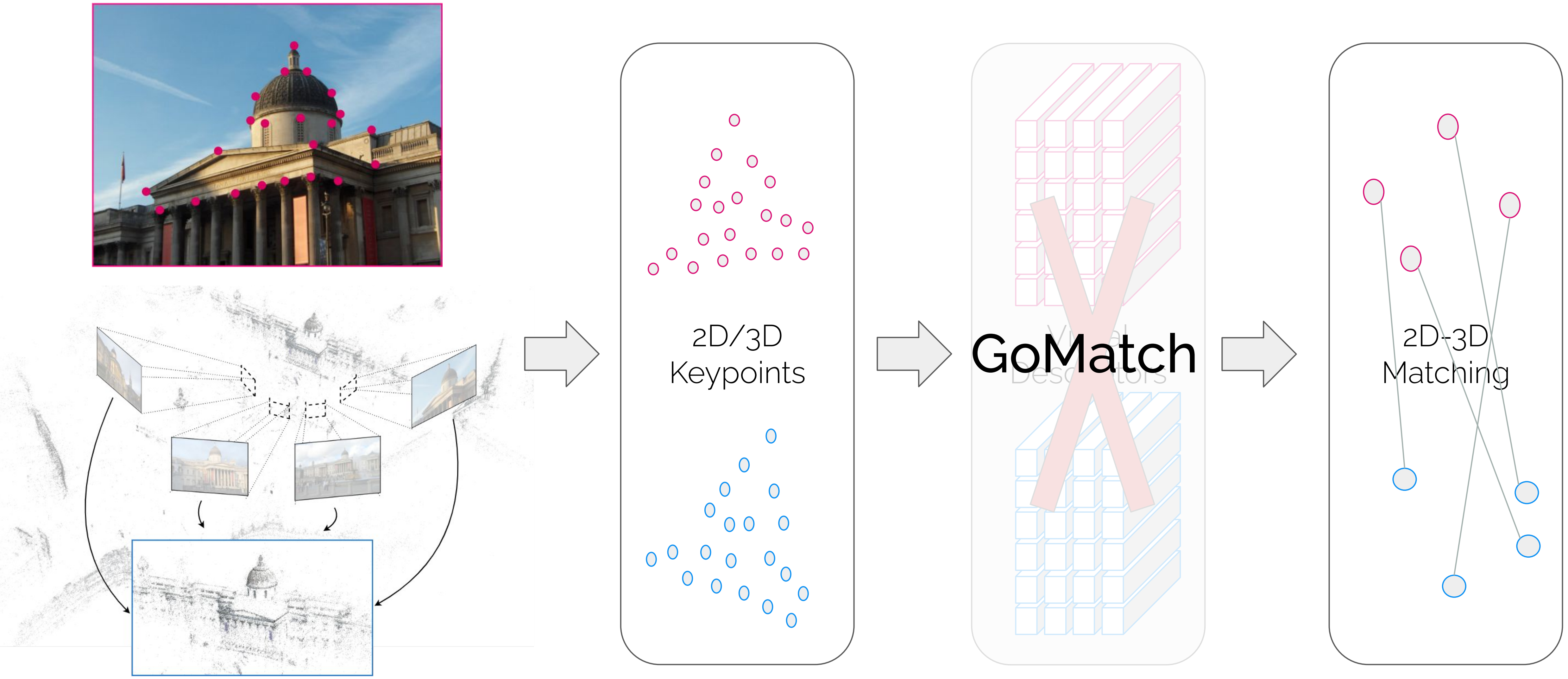
Overview


Localize


2D/3D
Keypoints
Visual
Descriptors
2D-3D
Matching

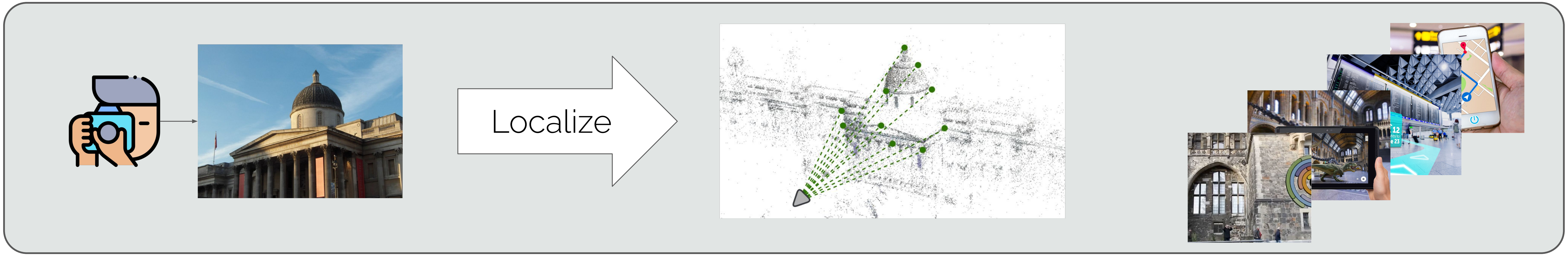
Overview

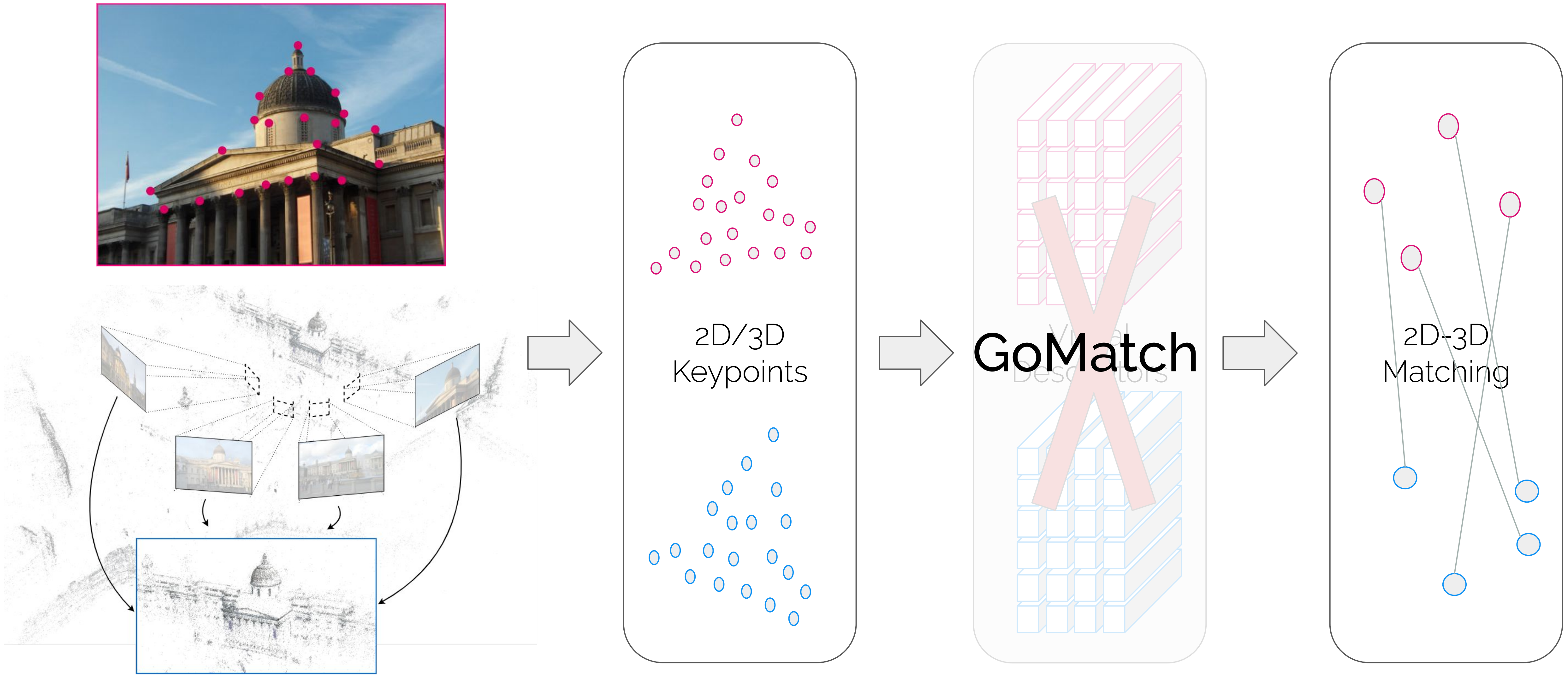




- Low Storage
- Privacy Preserving
- No Descriptor Maintenance

Overview



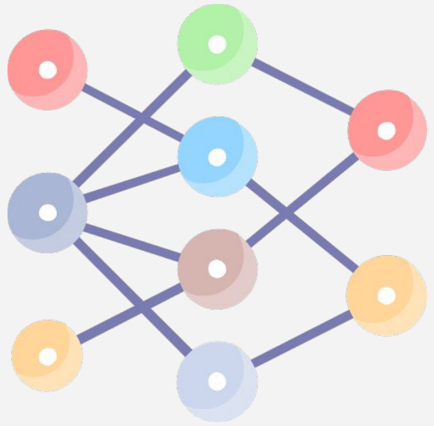


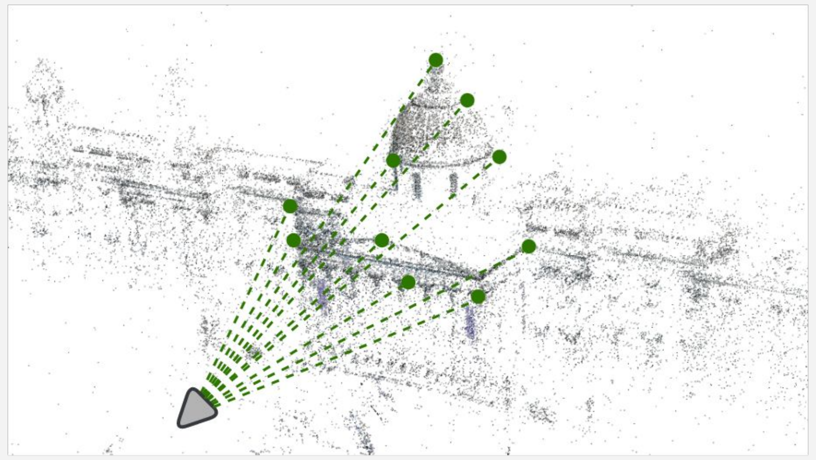
- Low Storage
- Privacy Preserving
- No Descriptor Maintenance

Scalable Large-scale Localization

Motivation

End-to-end

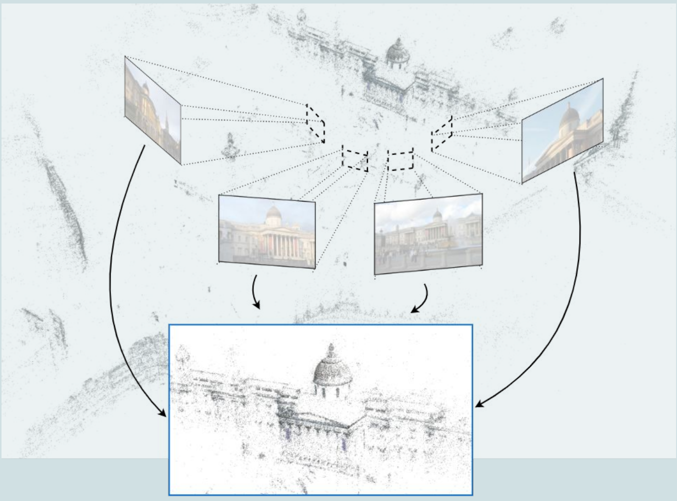


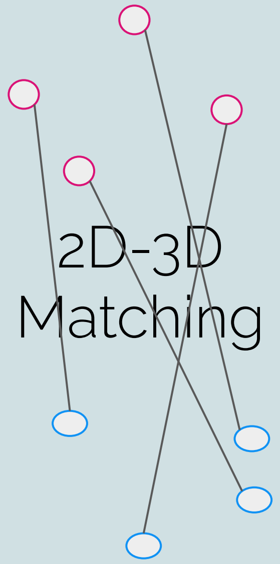


Camera Pose Regression
(PoseNet, MapNet, ...)

Scene Coordinate Regression
(DSAC, DSAC++, ...)

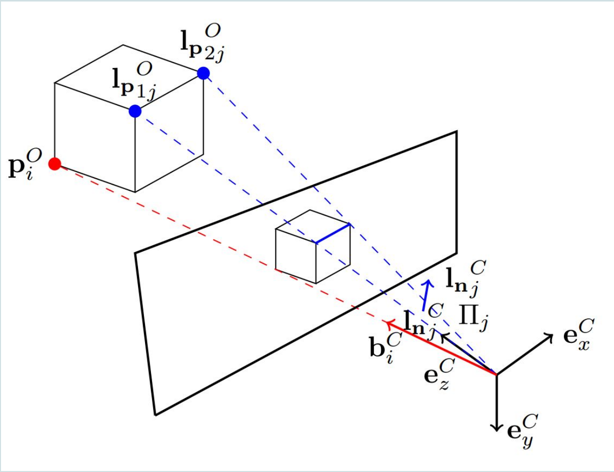
Structure-based





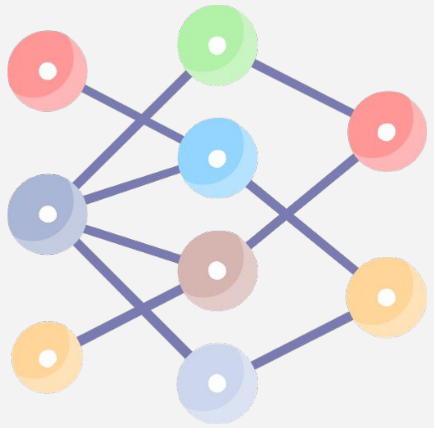
2D-3D
Matching

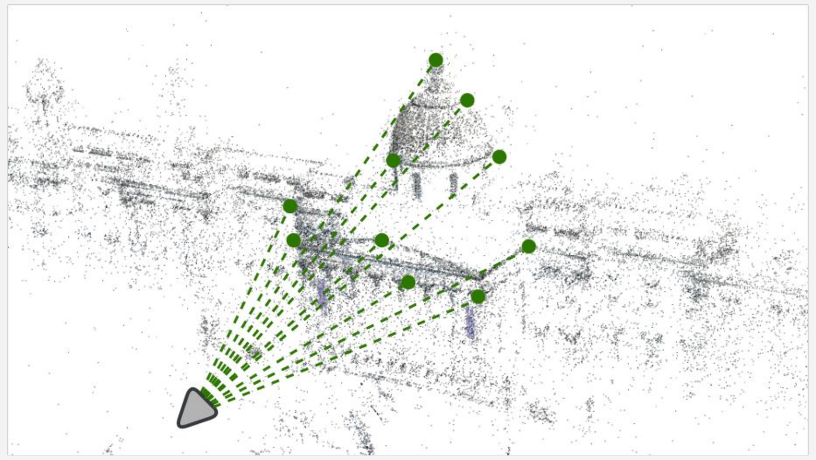
Perspective-n-Point
Solver



Structure-based
Localization
(HLoc, ActiveSearch, ...)

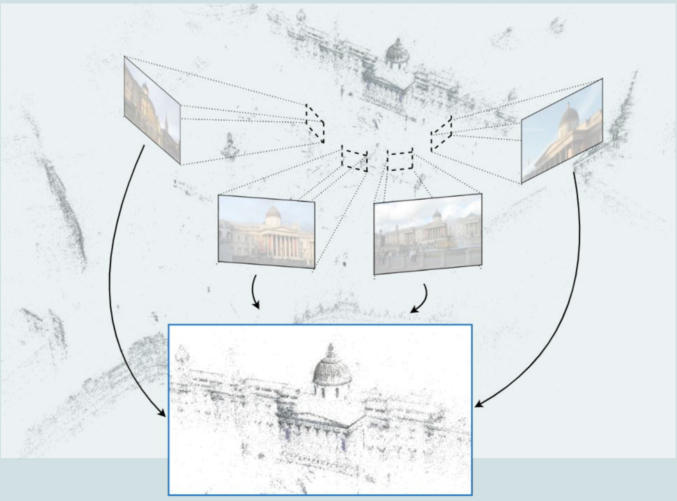
Visual Localization Approaches

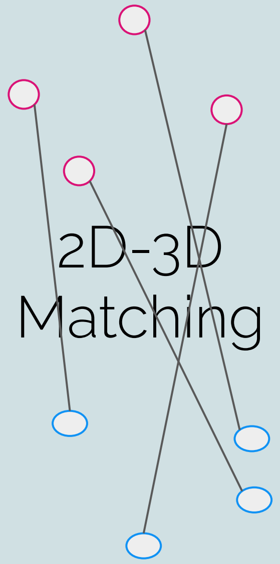




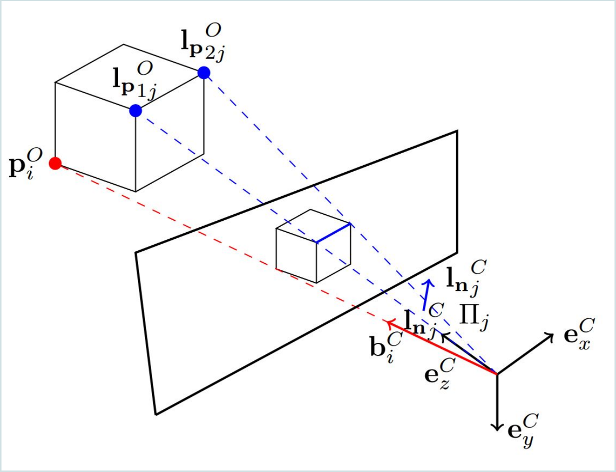
Camera Pose Regression
(PoseNet, MapNet, …)

Scene Coordinate Regression
(DSAC, DSAC++, …)



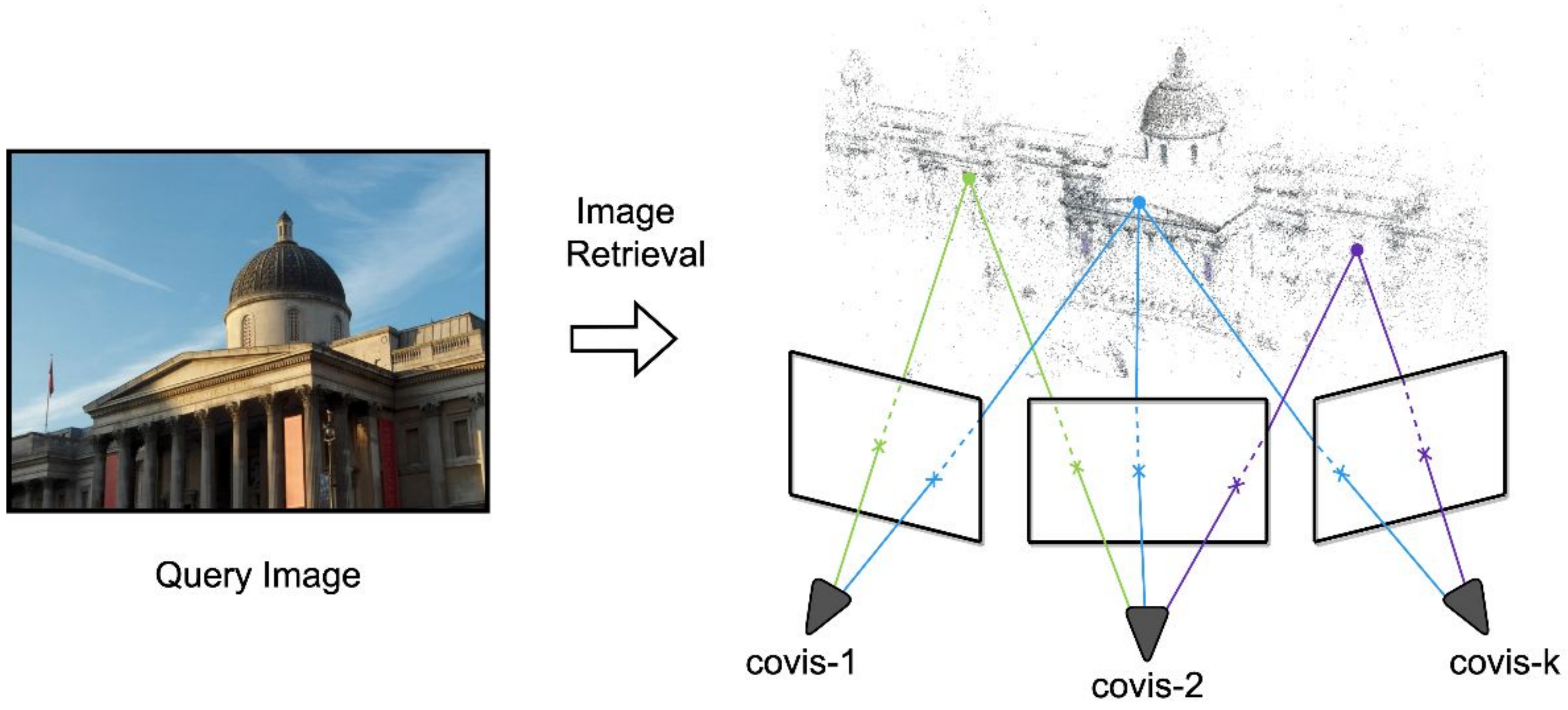


PnP Solver



Structure-based
Localization
(HLoc, ActiveSearch, …)

GoMatch



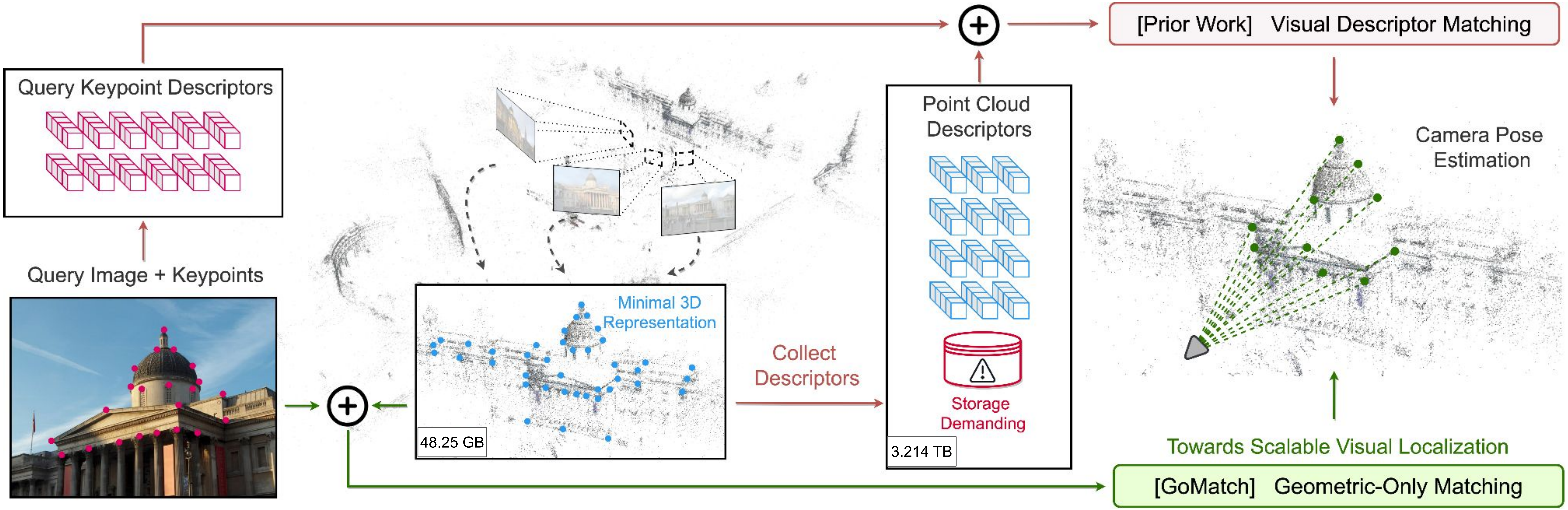
Localization Performance

Storage Requirements

Privacy

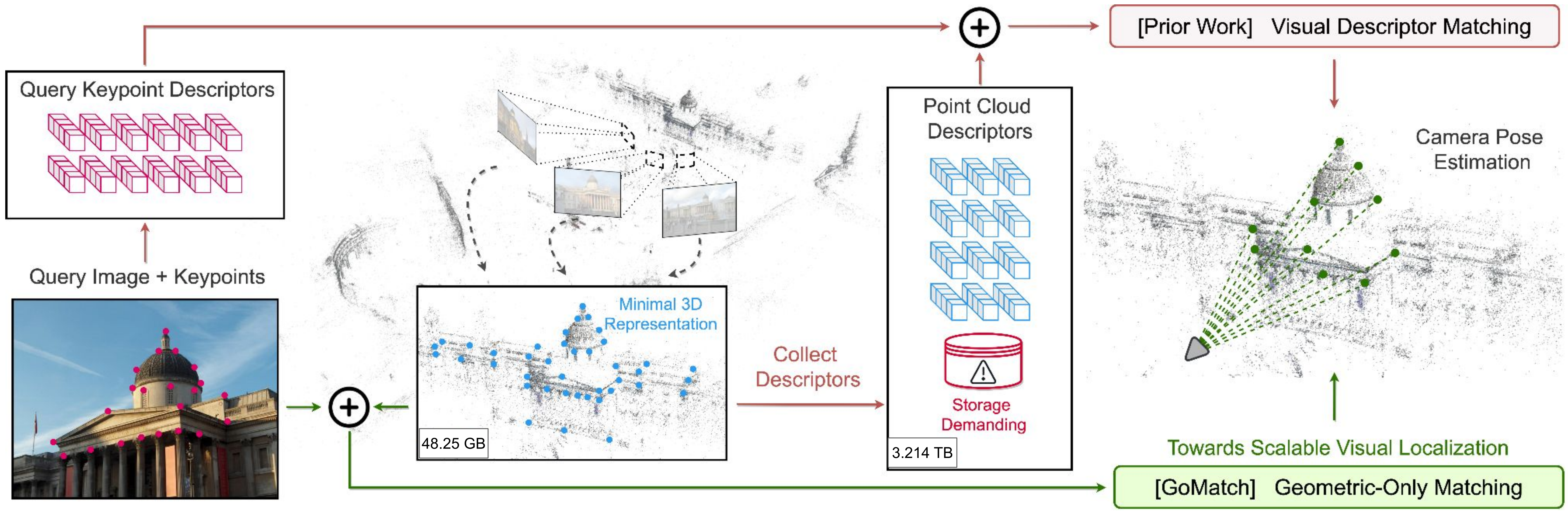
No Descriptor Maintenance

Significantly Lower Storage Requirements

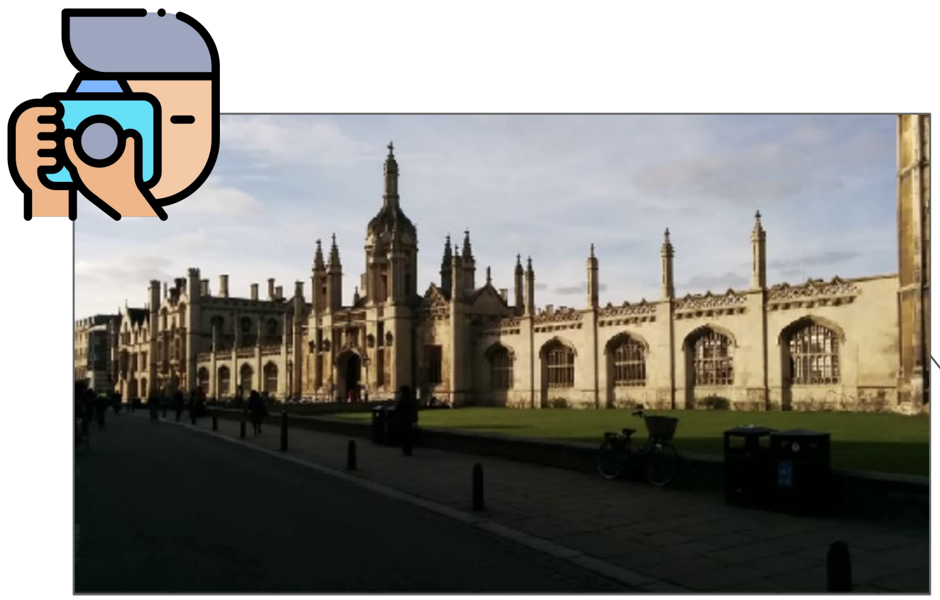


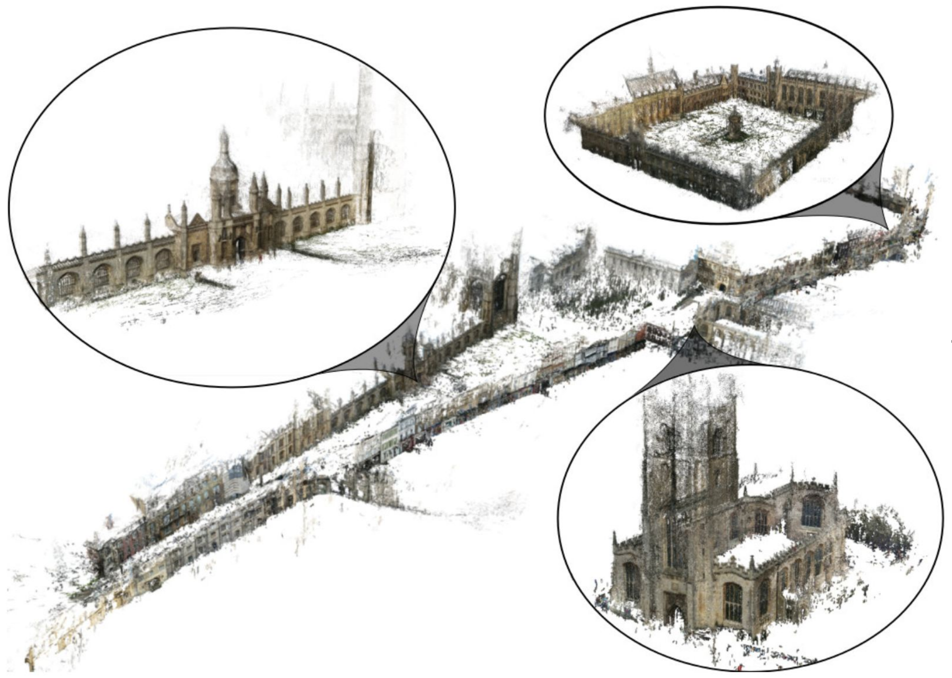
1.5% vs visual descriptors

Classical Structure-based Localization



Structure-based Approaches

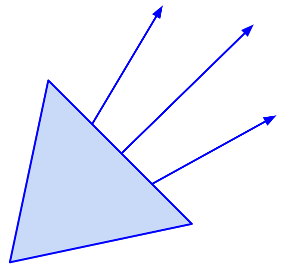


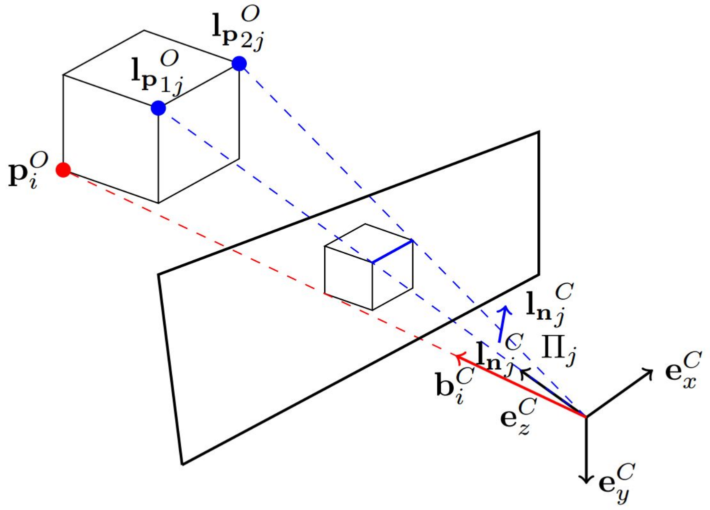


Scene representation

2D-3D Correspondences

PnP Solver

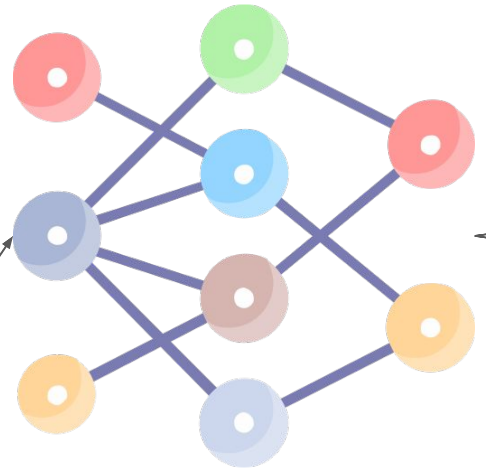




T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. PAMI2017

End-to-end Learned Localization

Sattler, Torsten, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixé. "Understanding the limitations of cnn-based absolute camera pose regression." CVPR19.



Scene representation

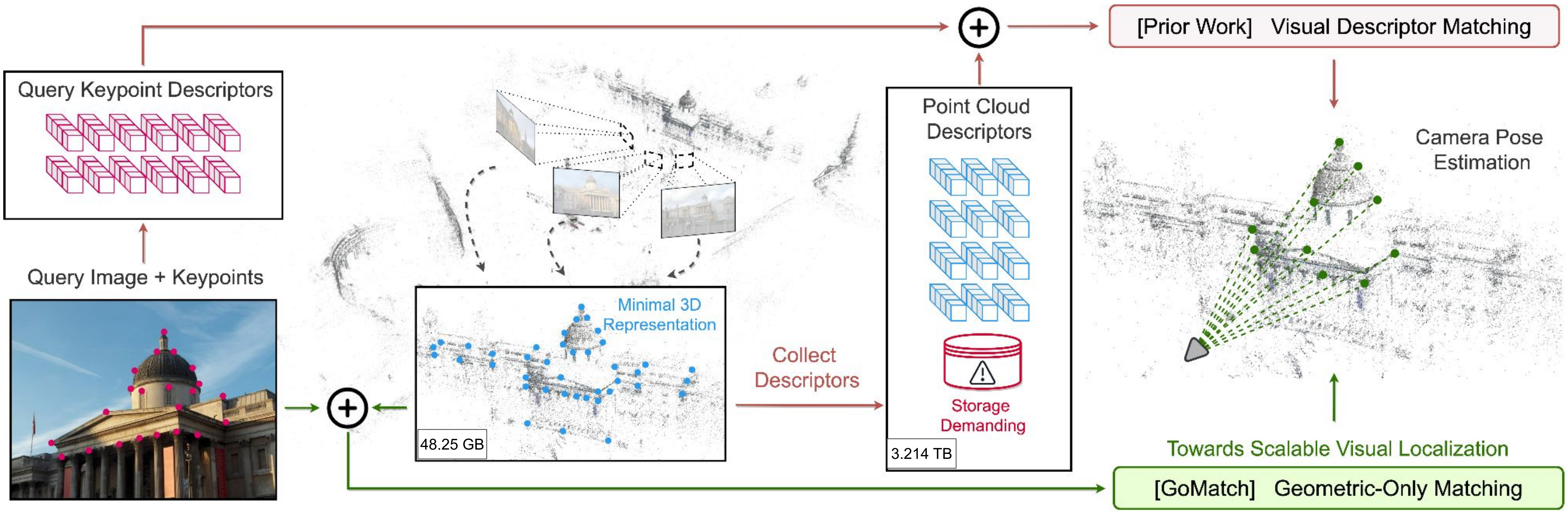
Scene Coordinate Regression (DSAC, DSAC++, …)

Camera Pose Regression (PoseNet, MapNet, …)

Image Retrieval (Netvlad, GeM, …)

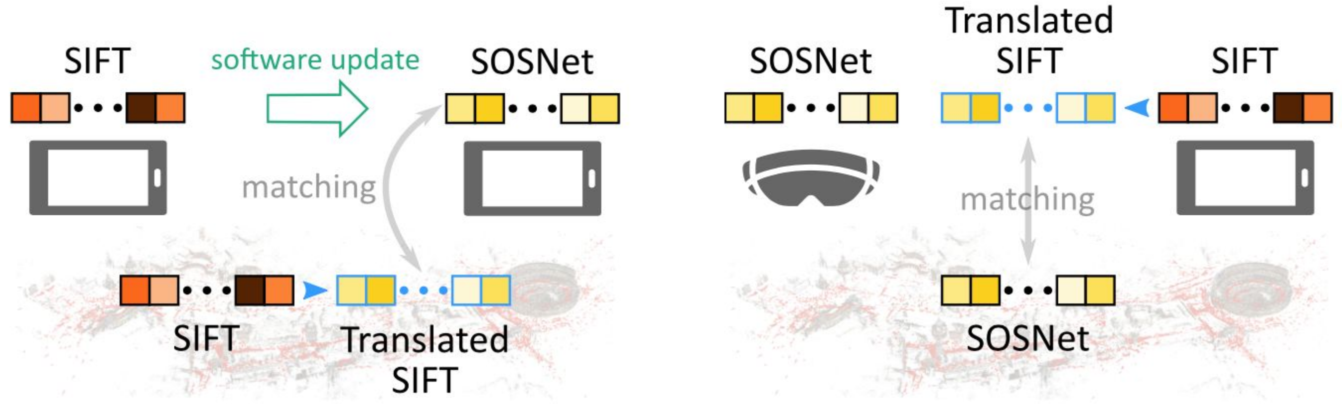
Relative Pose Estimation (EssNet, CamNet, …)

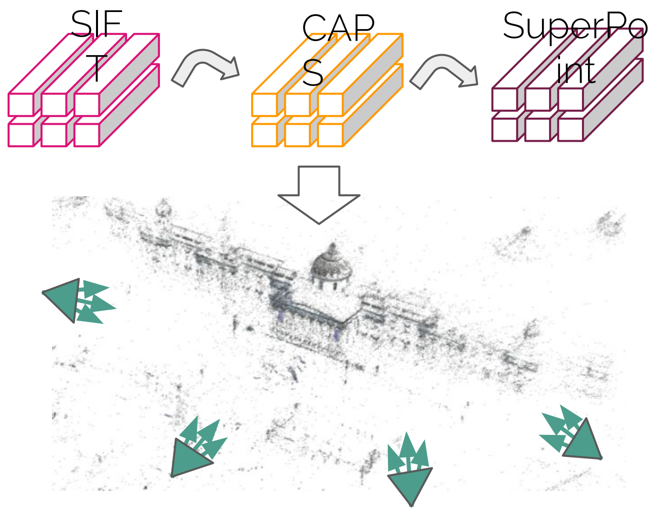
Storage Requirements

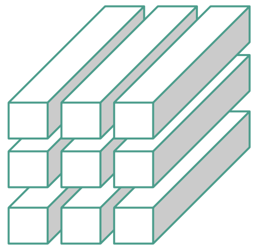


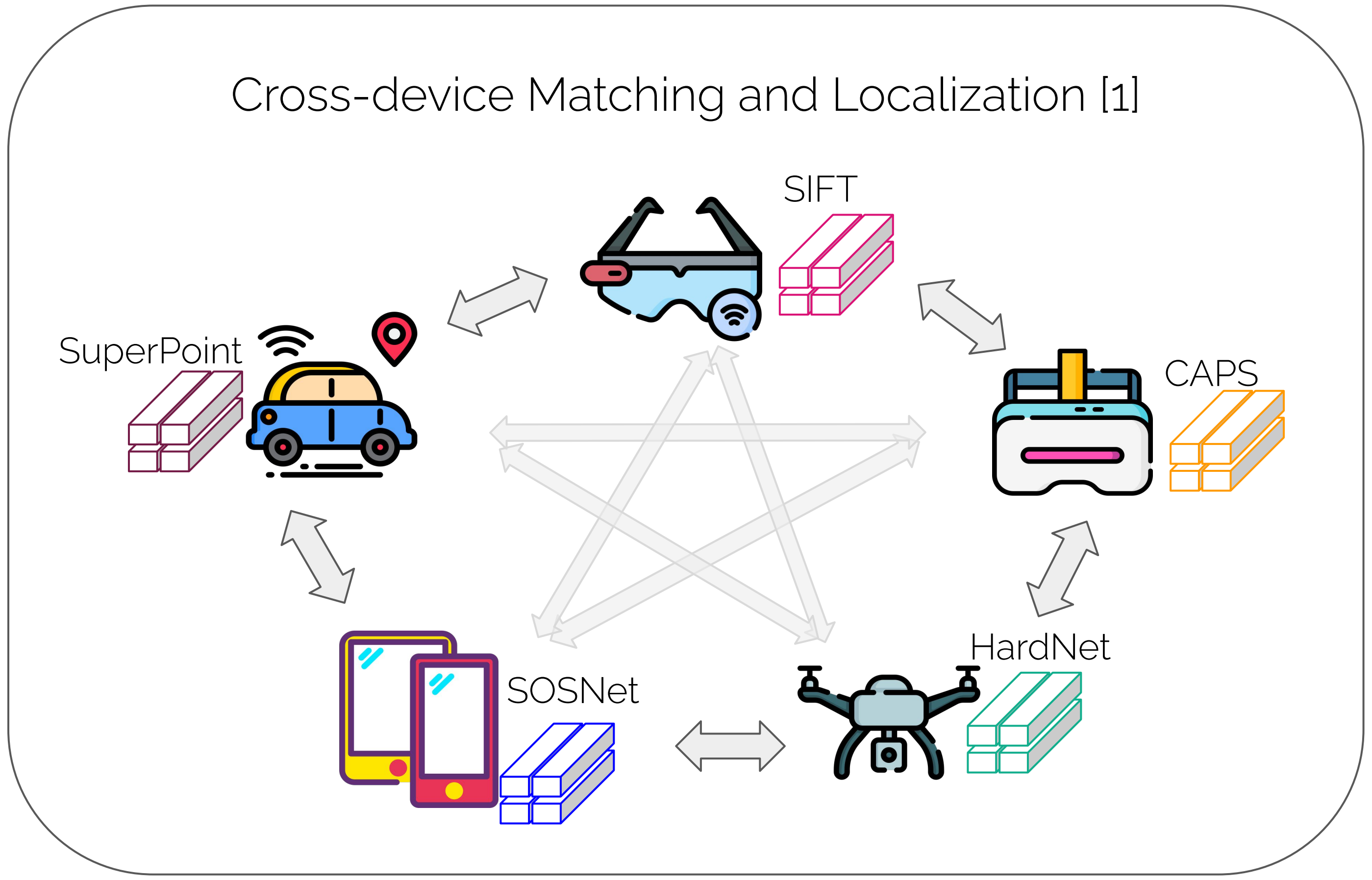
Practical Challenges

Maintenance
Complexity





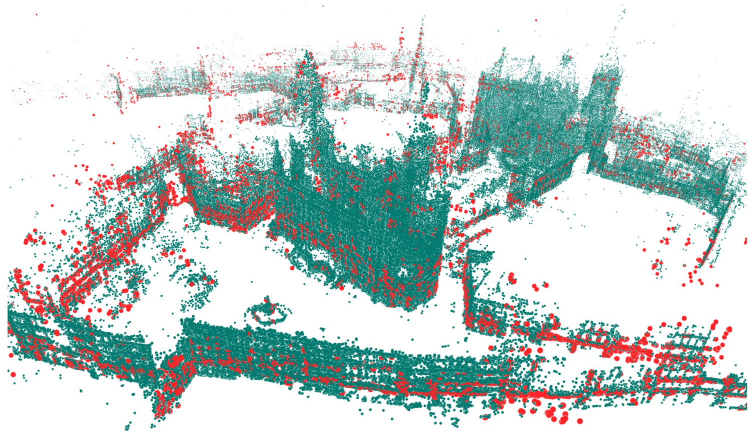


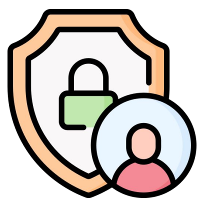


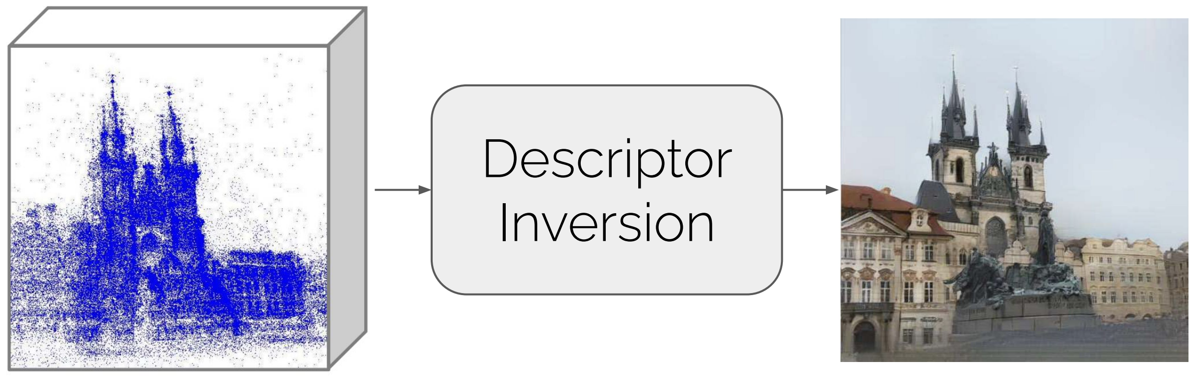
Cross-device Matching and Localization [1]

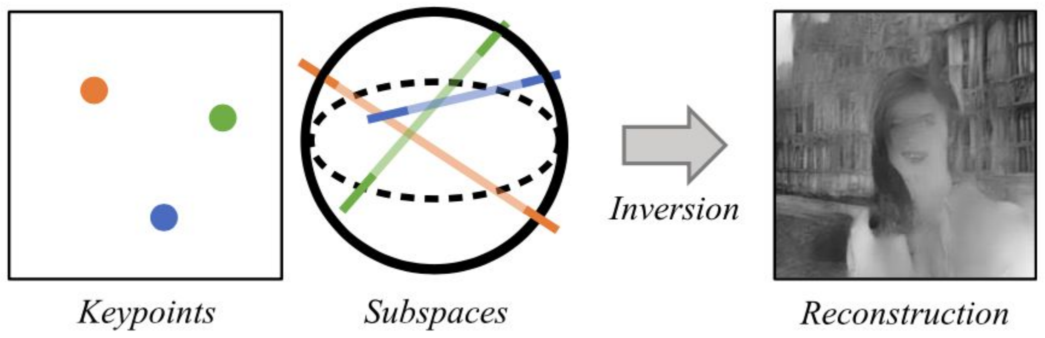
[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

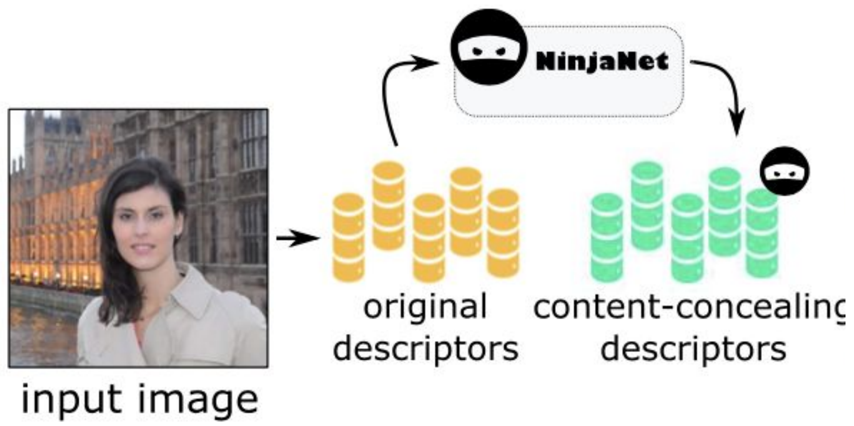
Practical Challenges

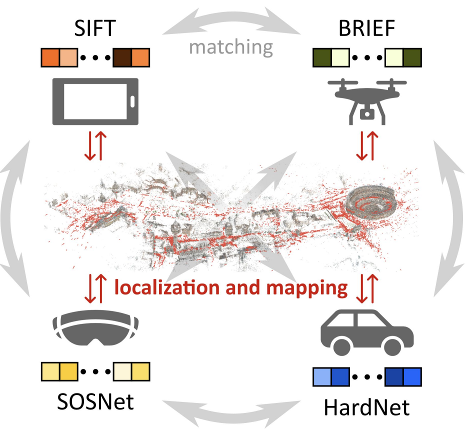


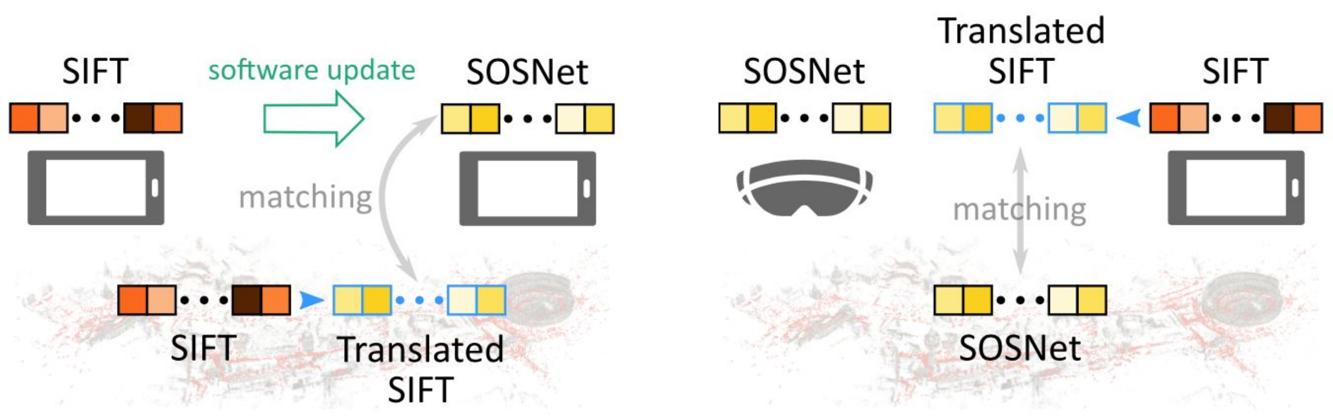




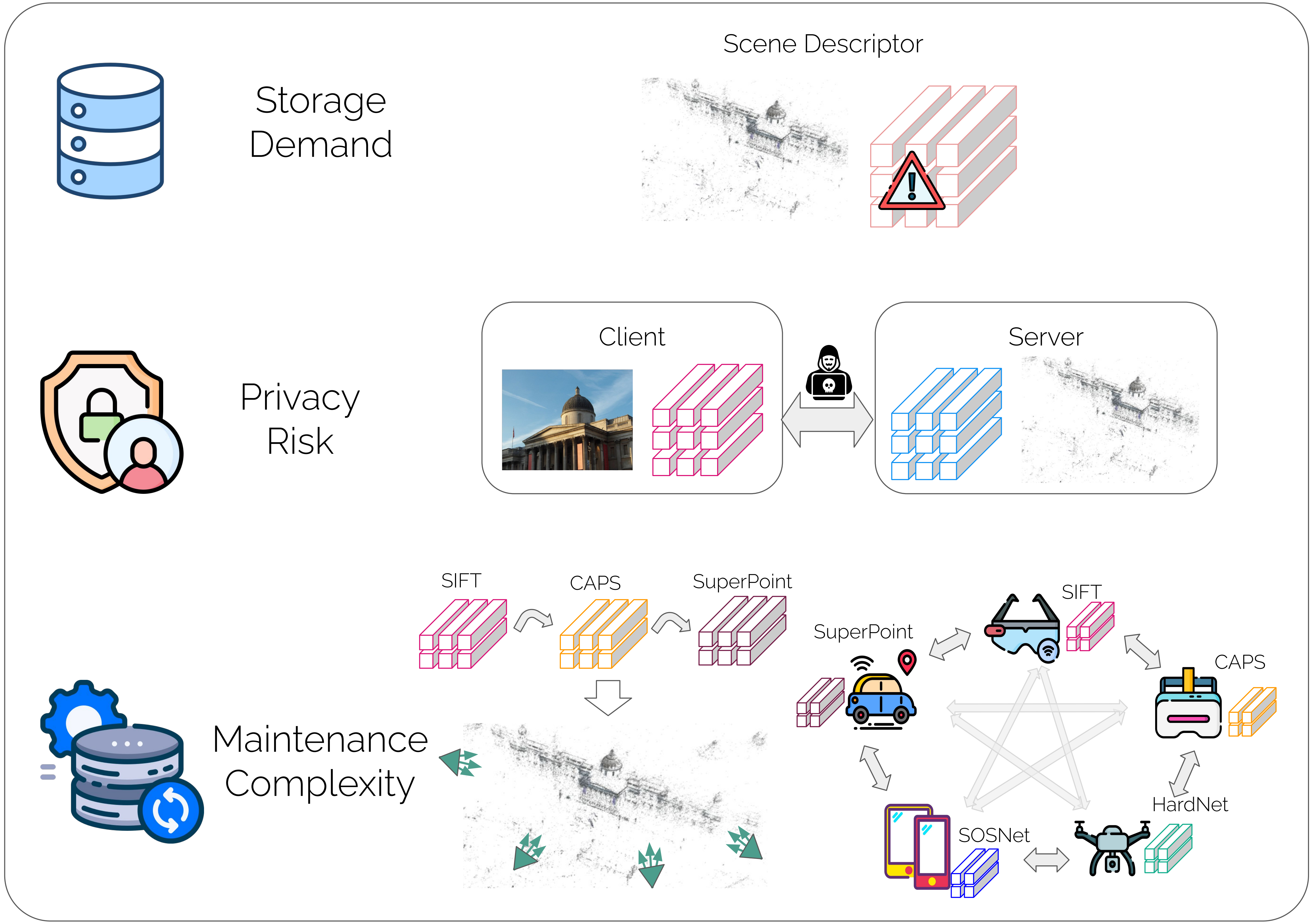




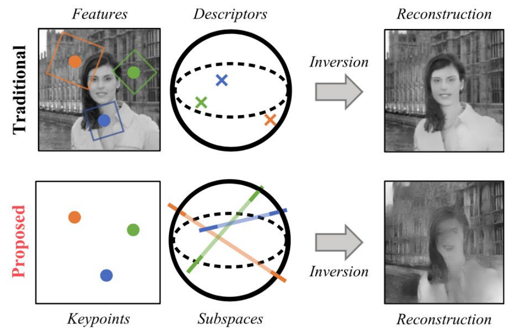




Practical Challenges



Existing Solutions

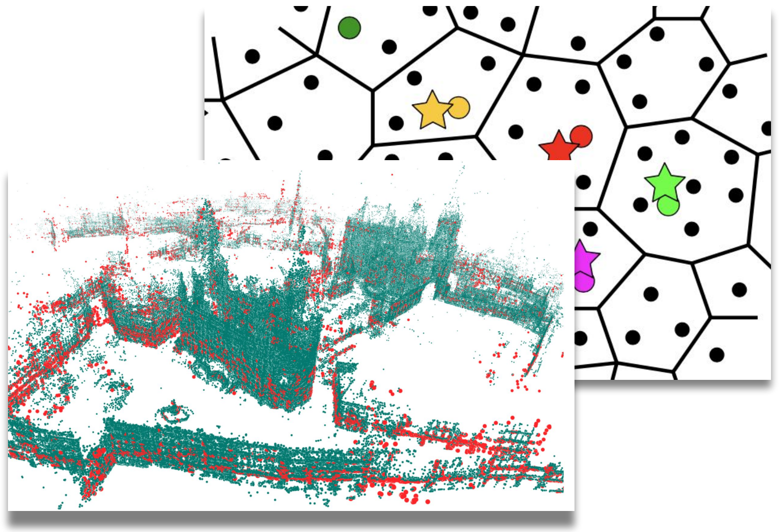


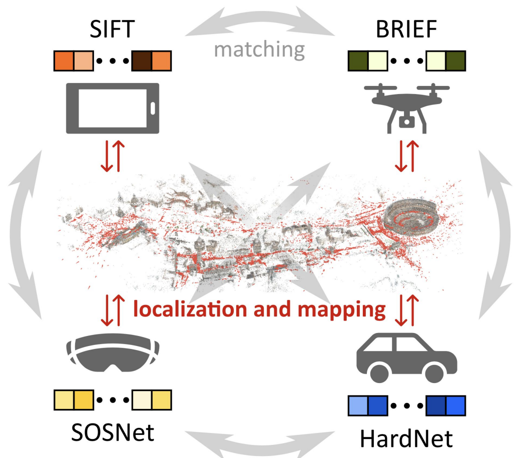
GoMatch

Storage / Memory Efficiency

Descriptor Maintenance

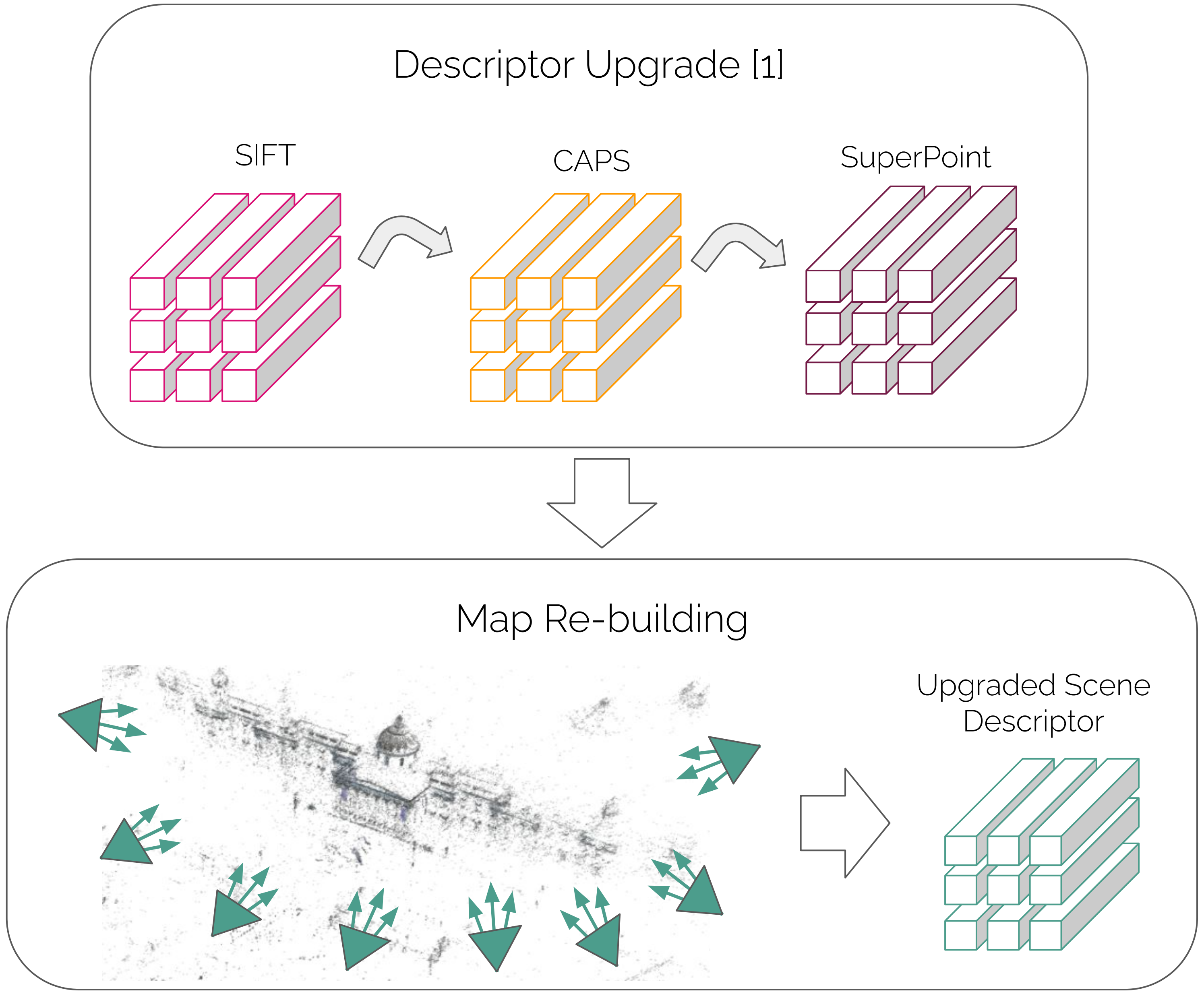
Privacy Preserving

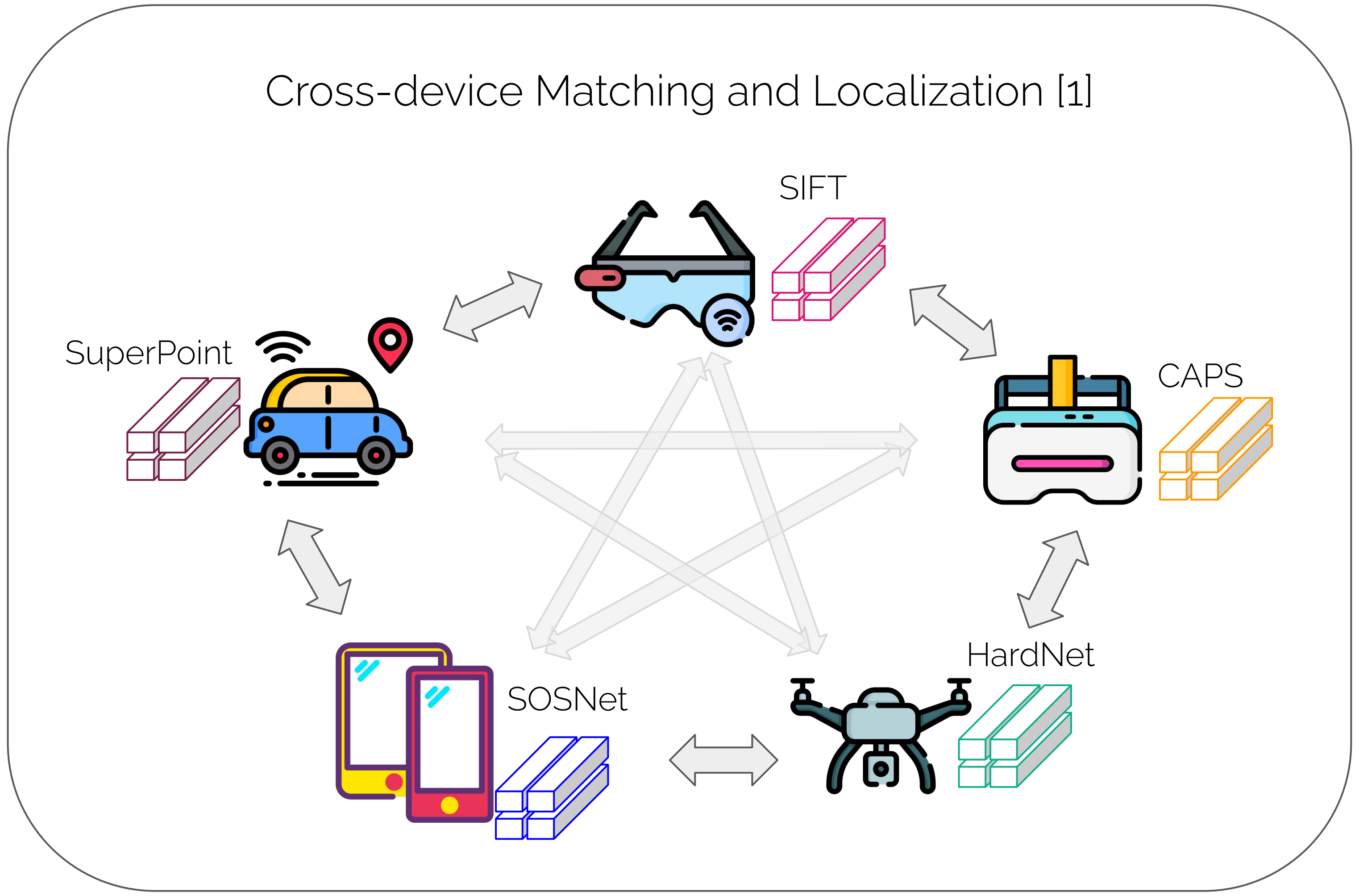




Practical Challenges

Maintenance
Complexity

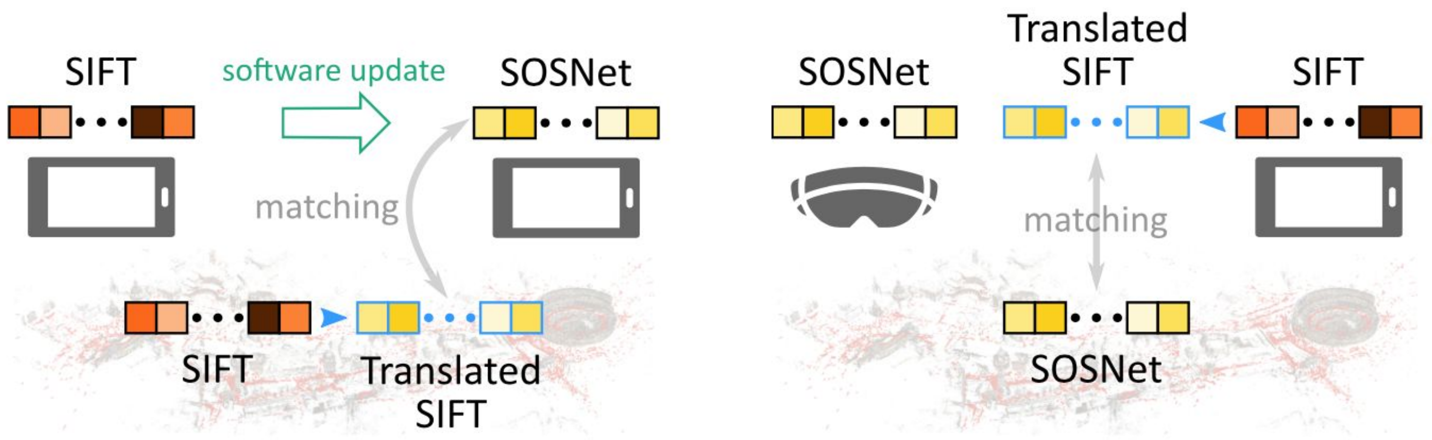




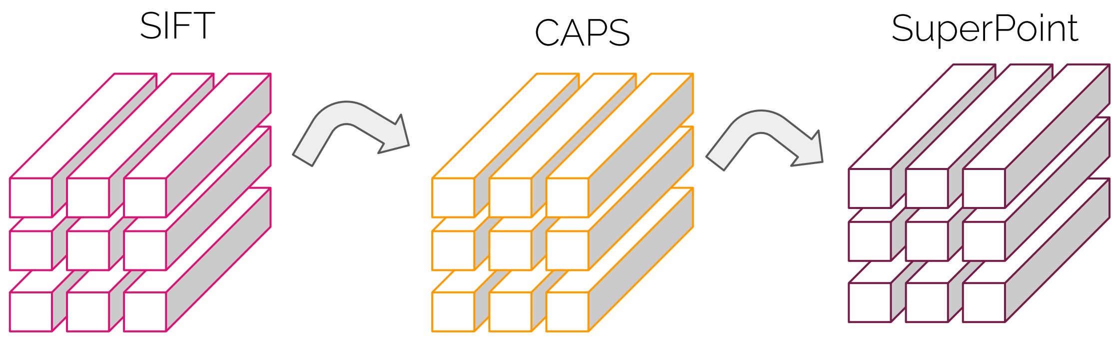
[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

Practical Challenges

Maintenance Complexity

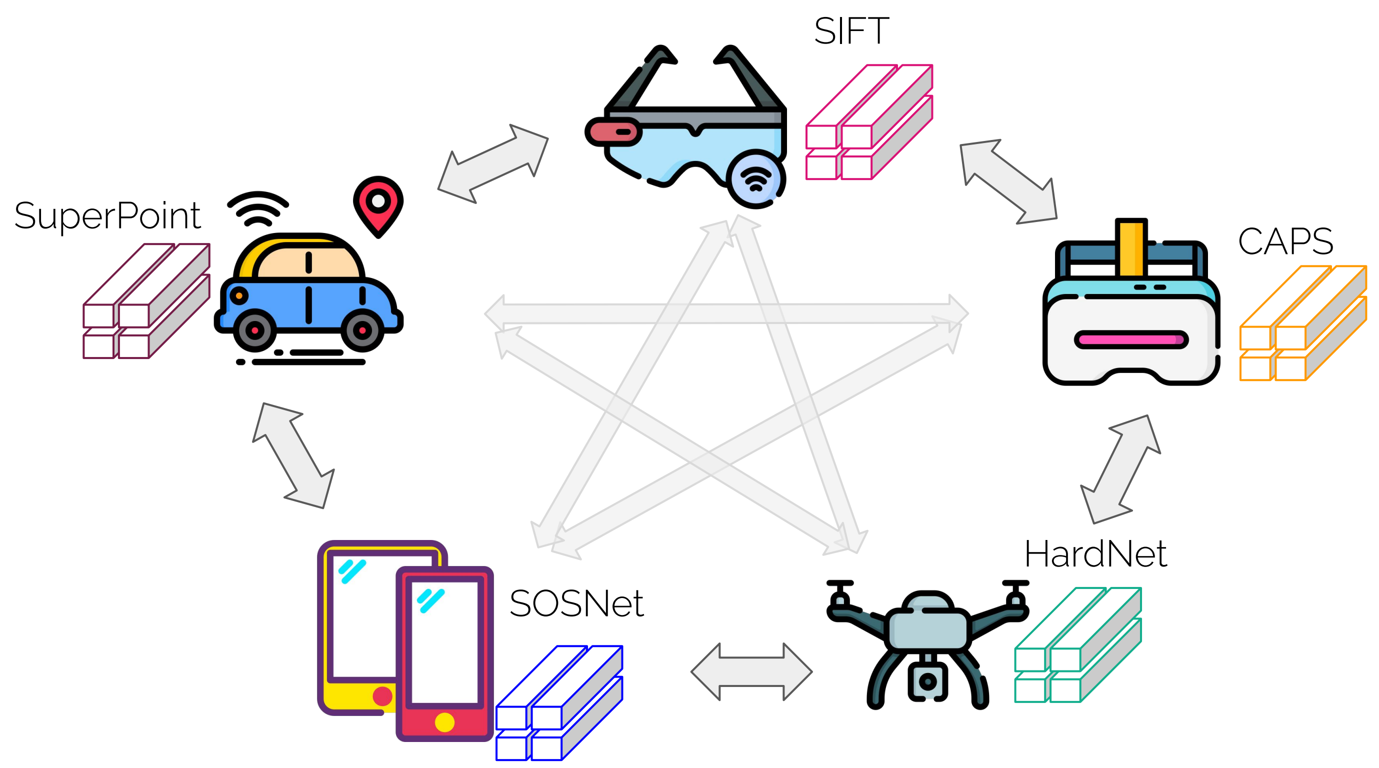


Descriptor Upgrade [1]



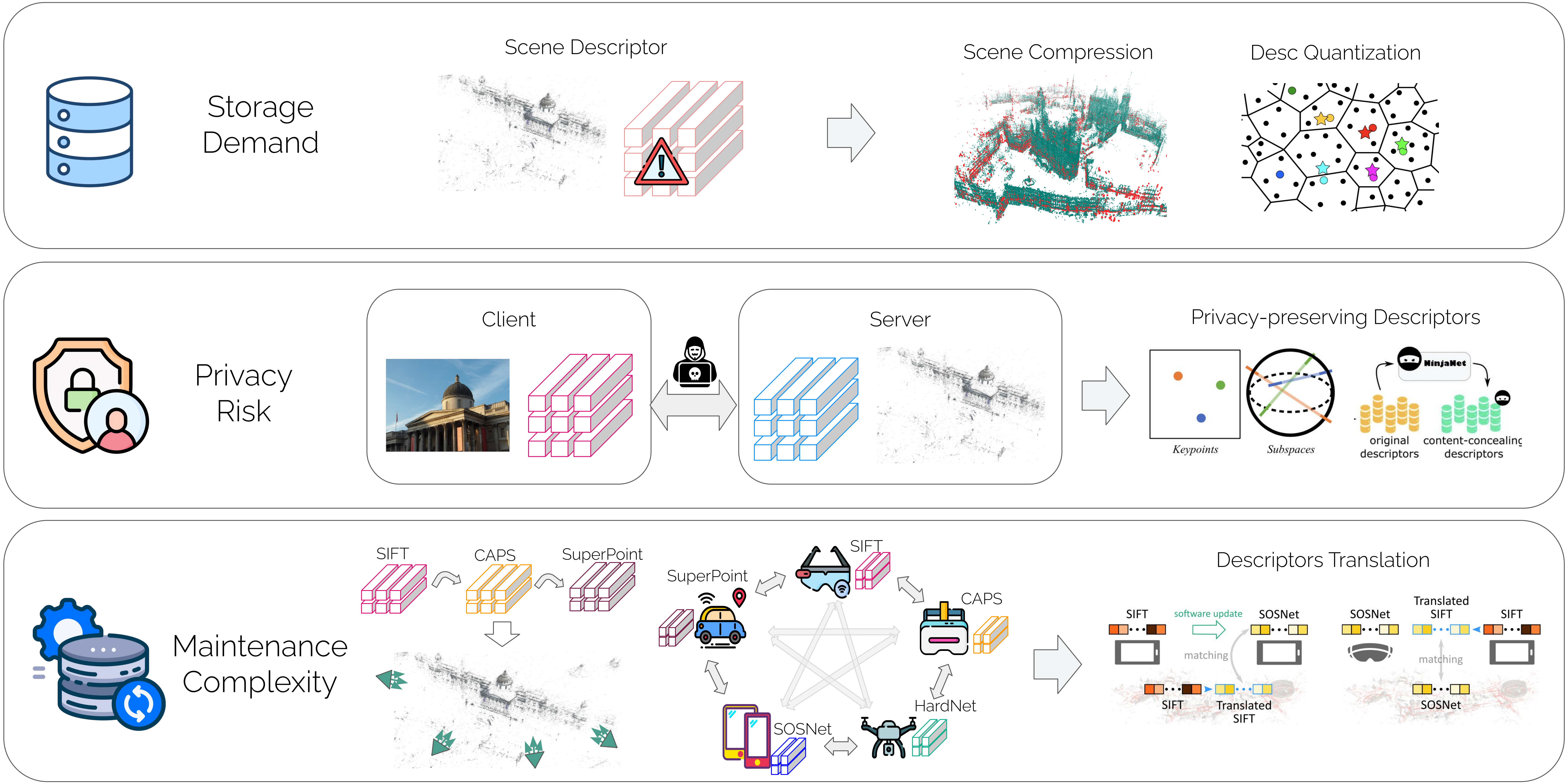
Map Re-building

Cross-device Matching and Localization [1]

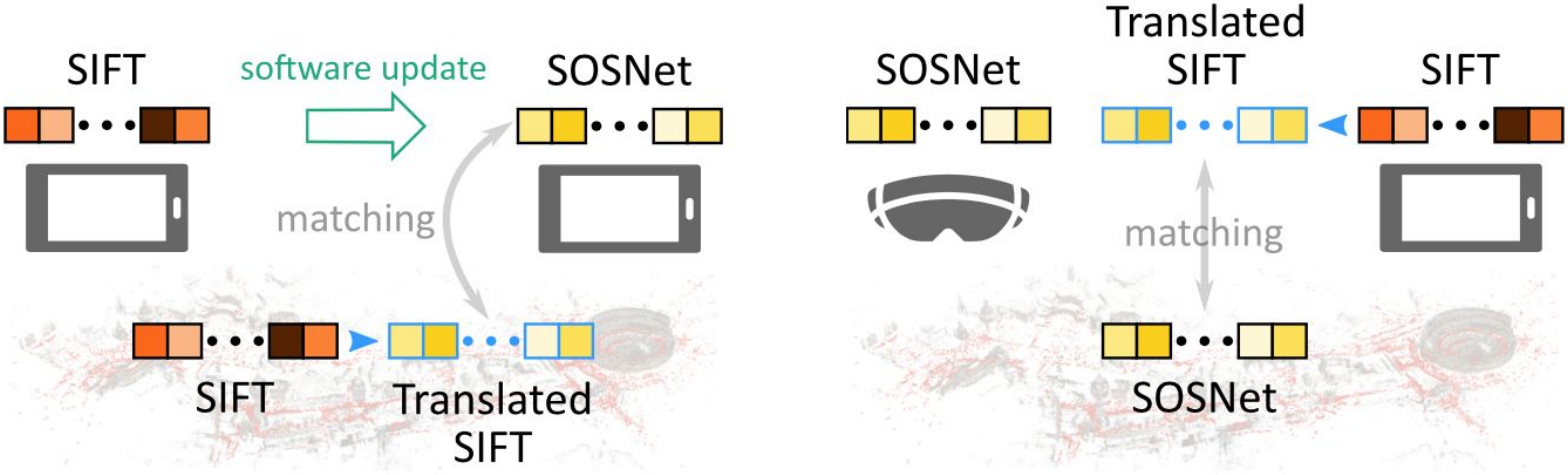


[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

Practical Challenges



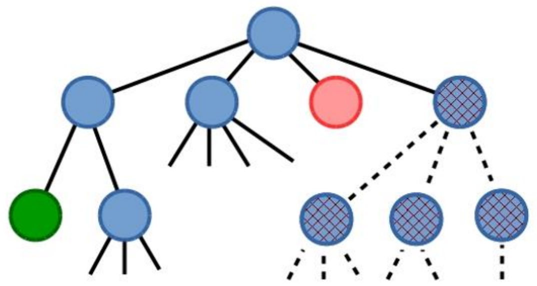
Maintenance Effort



[1] Dusmanu, Mihai, et al.. Cross-descriptor visual localization and mapping. ICCV21

Descriptor Maintenance

Geometric-based matching and pose estimation

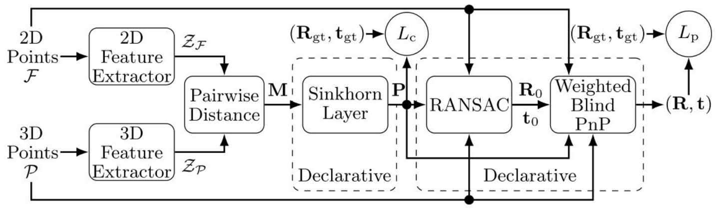


hegyhati.github.io

SoftPOSIT [1]
- Alternate step: softassign + POSIT
- Requires initialization
- Struggles with clutter, occlusions, repetitive patterns.
- Efficient

GOPAC [3]
- Globally optimal solution using Branch-and-Bound
- Prohibitive runtime requirements
- Cannot scale to large problems

Bind PnP [2]
- Kalman-Filter to maintain correspondence hypotheses.
- Requires initialization of GMM pose priors
- Better handling of occlusion, clutter and repetitive patterns



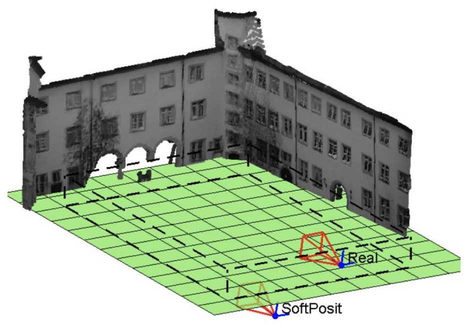
BPnPNet [4]
- Learning-based geometric matching network
- eclarative layers to back opagate through Sinkhorn, ANSAC and the PnP solver.
- erformance substantially egraded in the presence of outliers.

[1] David, Philip, et al. "SoftPOSIT: Simultaneous pose and correspondence determination." IJCV 2004
[2] Moreno-Noguer, Francesc et al. "Pose priors for simultaneously solving alignment and correspondence." ECCV 2008
[3] Campbell, Dylan, et al. "Globally-optimal inlier set maximisation for camera pose and correspondence estimation." PAMI 2018
[4] Campbell, Dylan, et al. "Solving the blind perspective-n-point problem end-to-end with robust differentiable geometric optimization." ECCV 2020.

Geometric-based matching and pose estimation

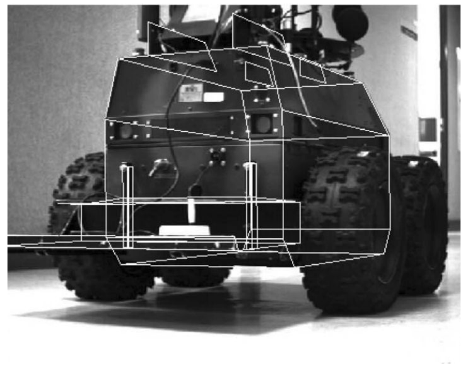
SoftPOSIT [1]
- Iterative softassign
- POSIT
- Requires initialization

Bind PnP
Moreno-Noguer et al.
2008

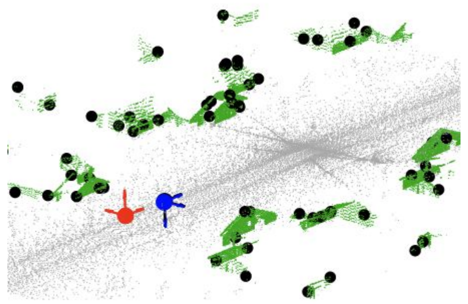


BPnPNet [4]
Campbell et al. 2018

SoftPOSIT
Dementhon et al.
2004



GOPAC
Campbell et al. 2018



[1] David, Philip, et al. "SoftPOSIT: Simultaneous pose and correspondence determination." IJCV
[

Geometric-based matching

ECCV'20 ONLINE
23-28 AUGUST 2020

Existing work

With 2D–3D correspondences:

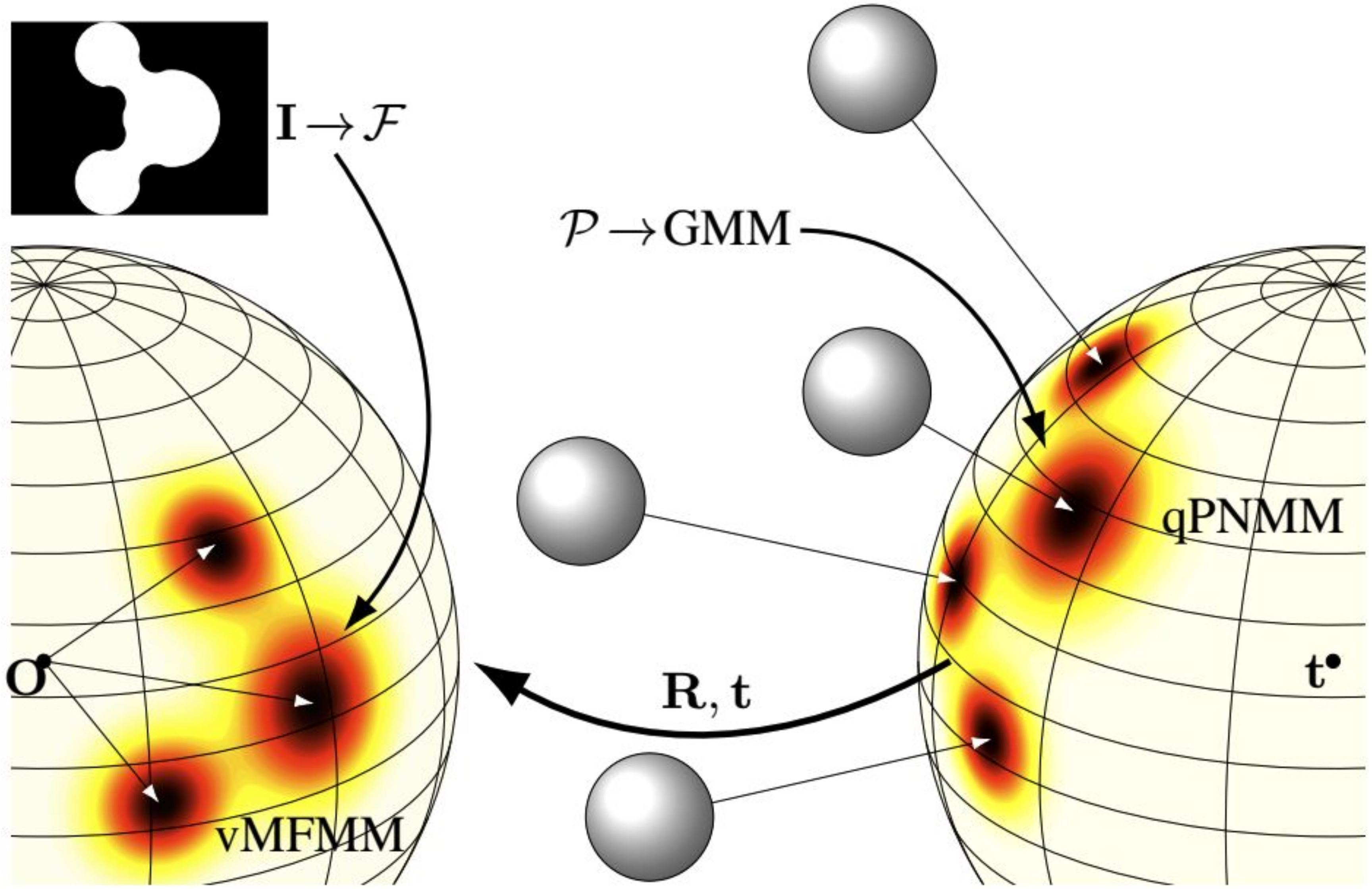
- Perspective-n-Point (PnP)
  - Gao *et al.* 2003; Lepetit *et al.* 2009
  - + RANSAC [Fischler & Bolles 1981]
  - + global optimisation [Li 2009]
  - + neural network [Dang *et al.* 2018]
- Sparse feature pipelines
  - Svärm *et al.* 2016; Sattler *et al.* 2017; Cavallari *et al.* 2017, 2019; Schönberger *et al.* 2018; Taira *et al.* 2018

Without 2D–3D correspondences:

- Learning-based camera pose
  - Kendall *et al.* 2015–2017; Cai *et al.* 2018; Brahmbhatt *et al.* 2018; Radwan *et al.* 2018; Walch *et al.* 2017; Brachmann *et al.* 2017, 2018, 2020 (DSAC)
- Optimization-based camera pose
  - **Local:** David et al. 2004 (SoftPOSIT); Moreno-Noguer et al. 2008 (BlindPnP)
  - **Global:** Grimson 1990; Jurie 1999; Brown *et al.* 2015; Campbell *et al.* 2019

Geometric-Only Methods

Campbell, Dylan, Lars Petersson, Laurent Kneip, Hongdong Li, and Stephen Gould. The alignment of the spheres: Globally-optimal spherical mixture alignment for camera pose estimation. CVPR 2019.



Visual Localization

Classical Structure-based Localization

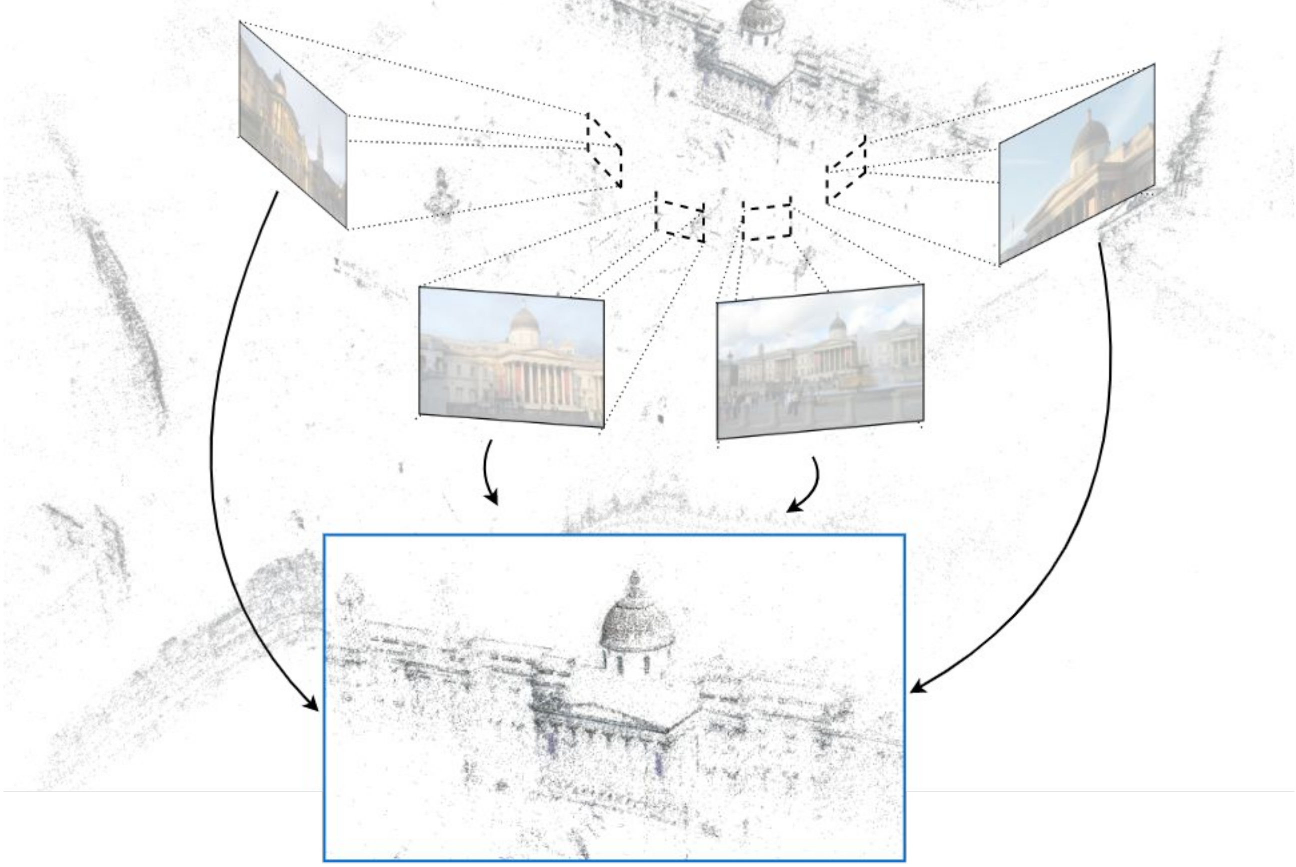
Query

Classical Structure-based Localization

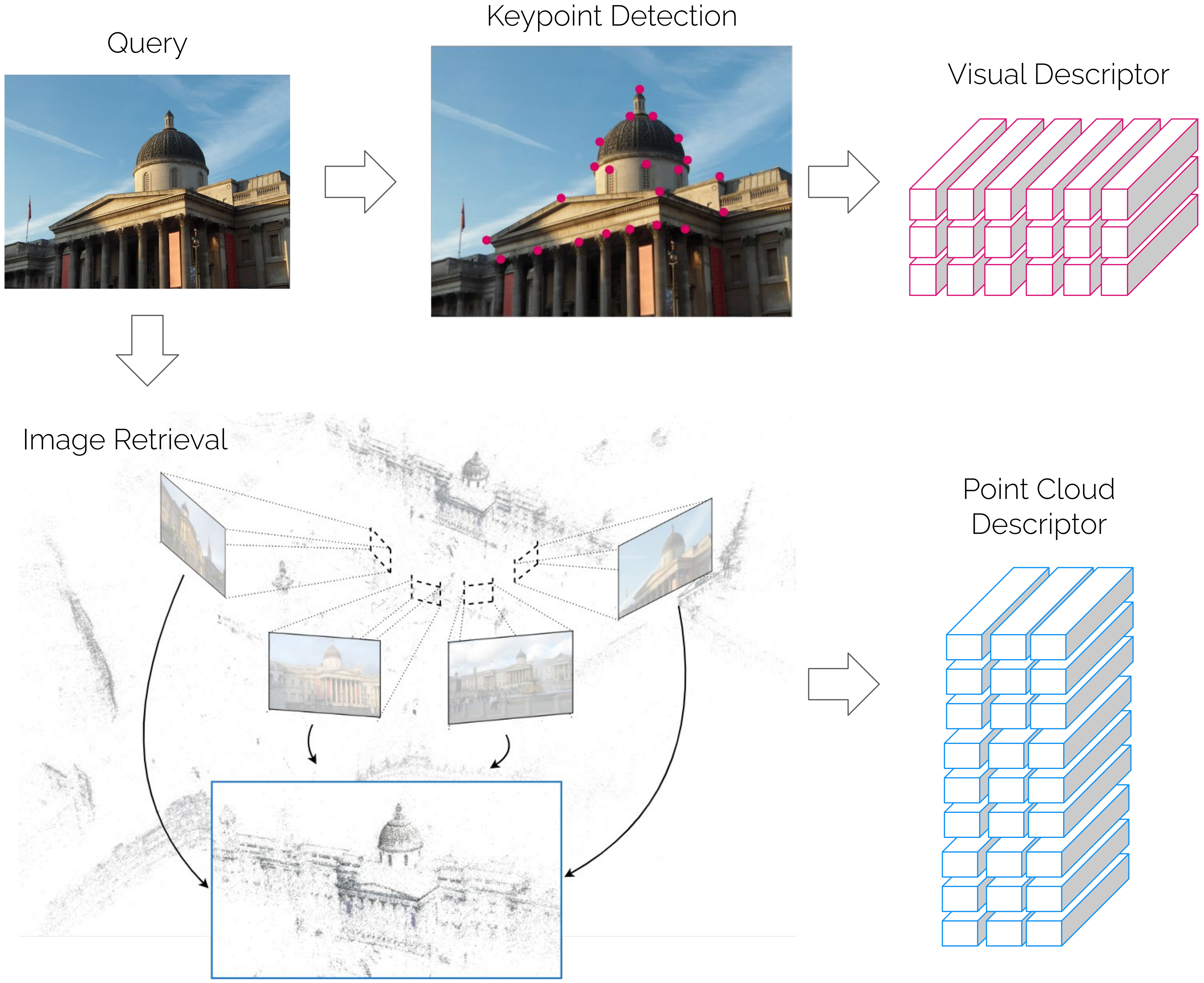
Query

Keypoint Detection

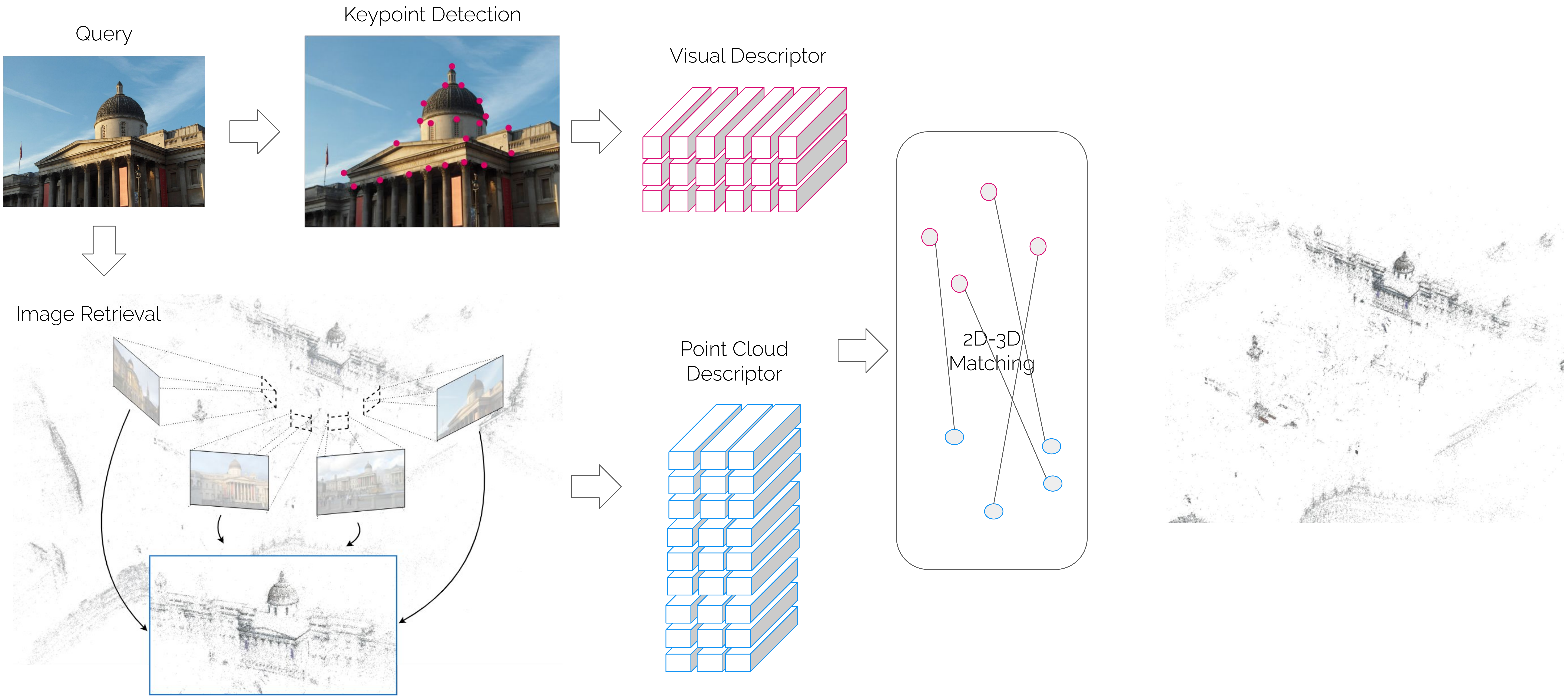
Image Retrieval



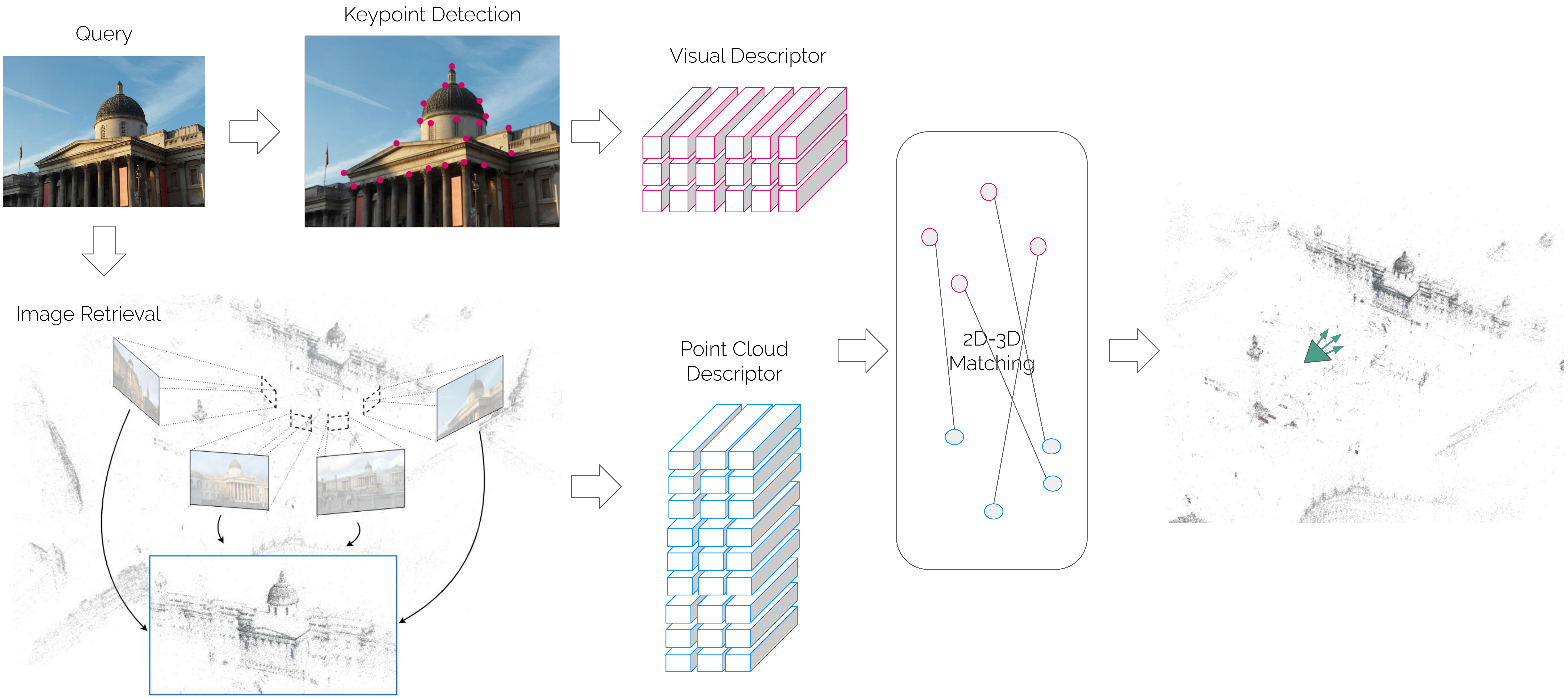
Classical Structure-based Localization



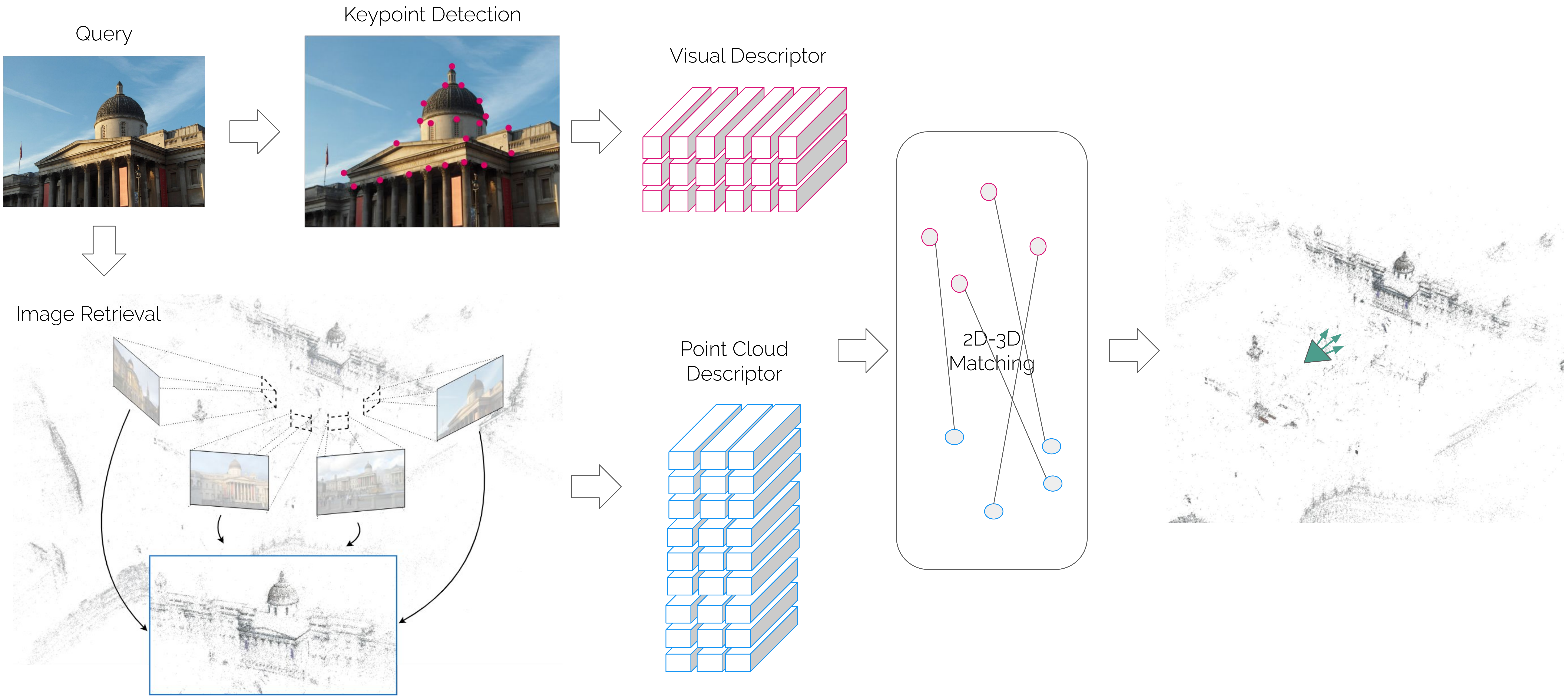
Classical Structure-based Localization



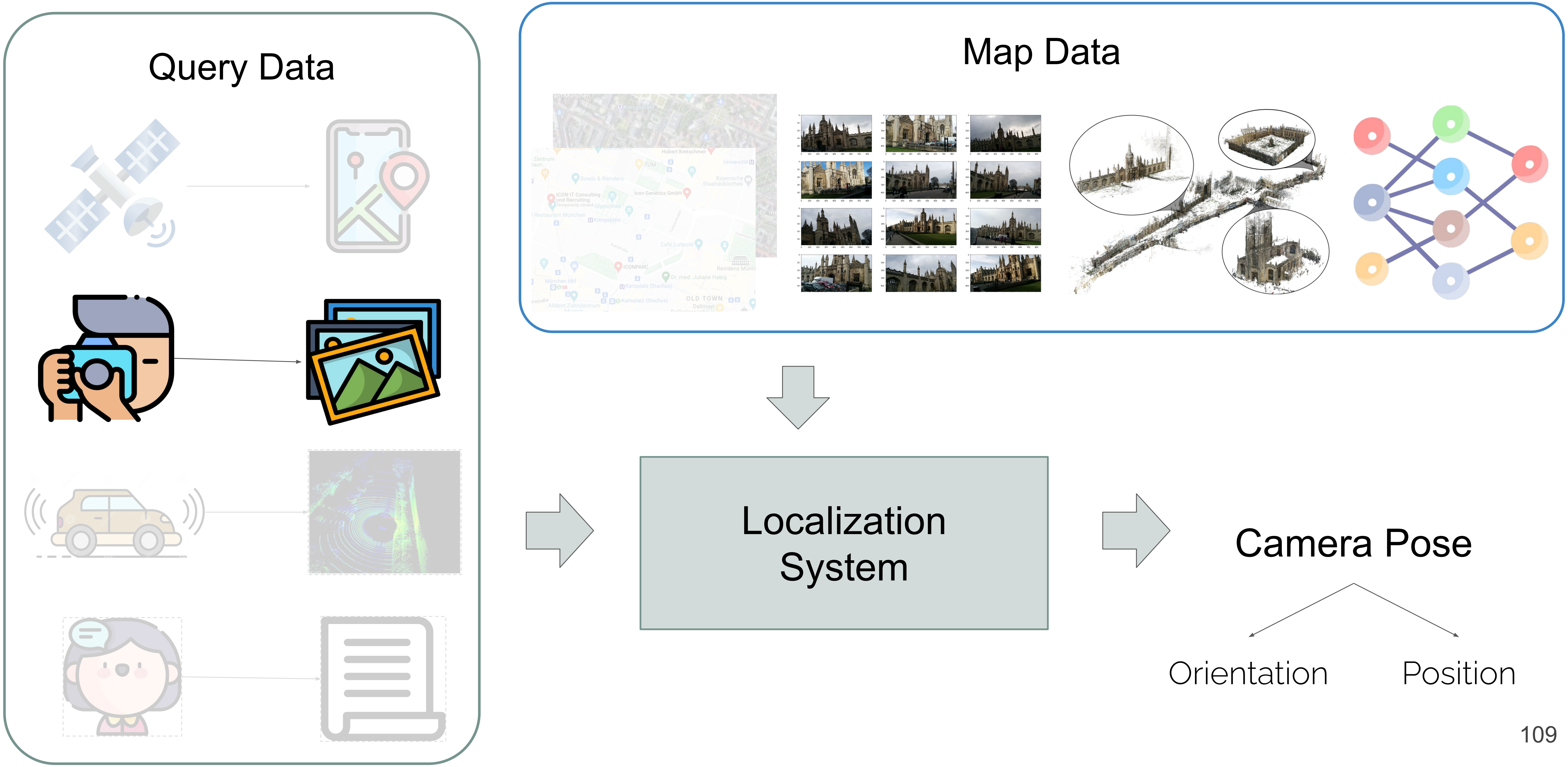
Classical Structure-based Localization



Classical Structure-based Localization



Visual Localization

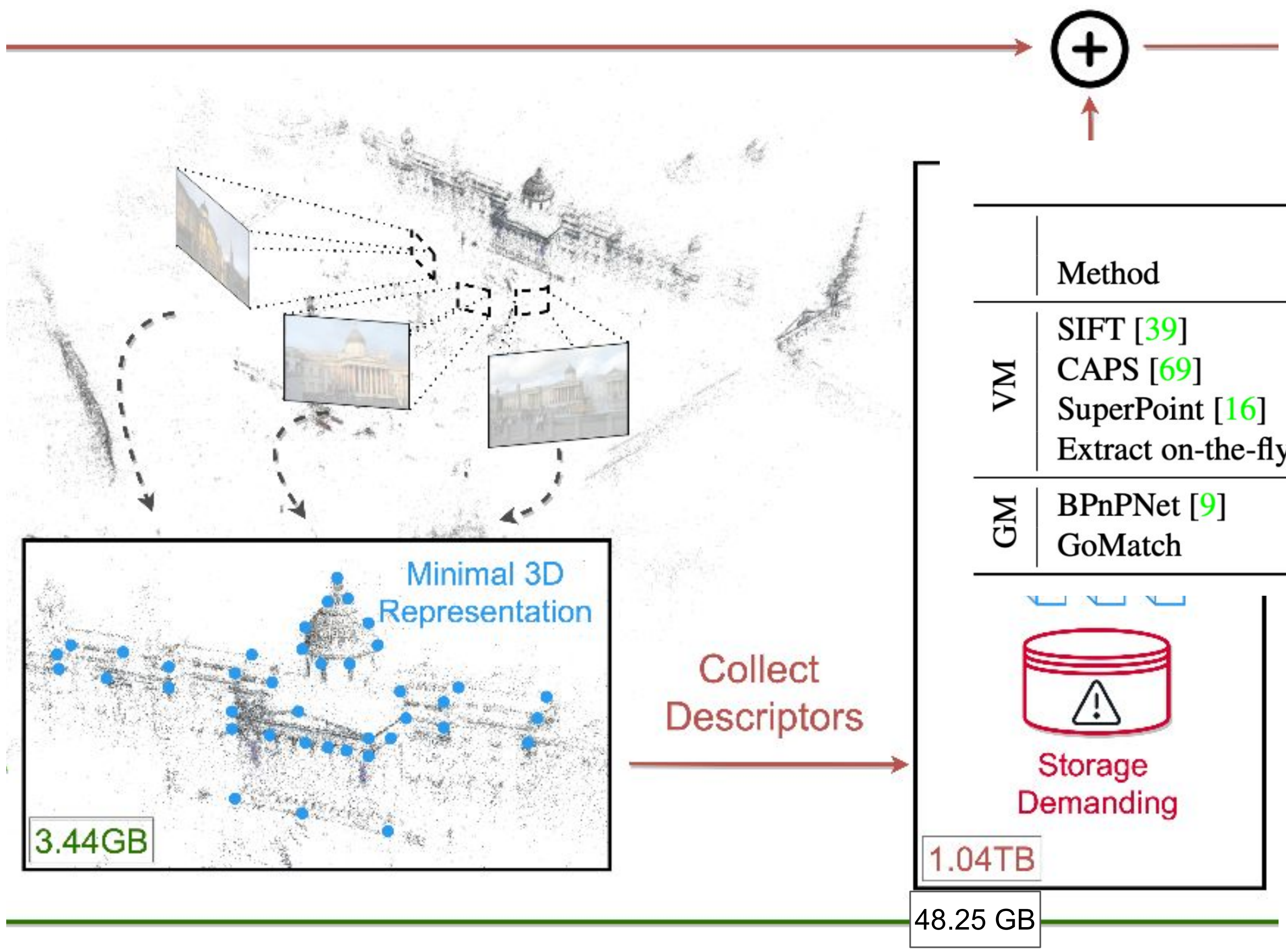


AR/VR

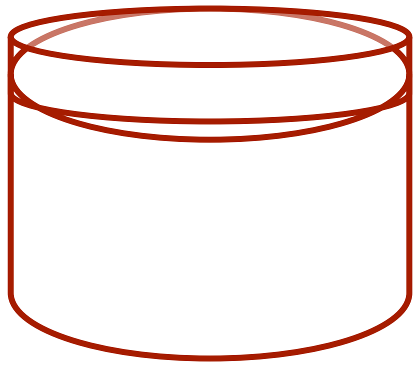
https://blog.helpdocs.io/guidigo/

[Middelberg, Sattler, Untzelmann, Kobbelt, Scalable 6-DOF Localization on Mobile Devices, ECCV 2014]

Storage Requirements

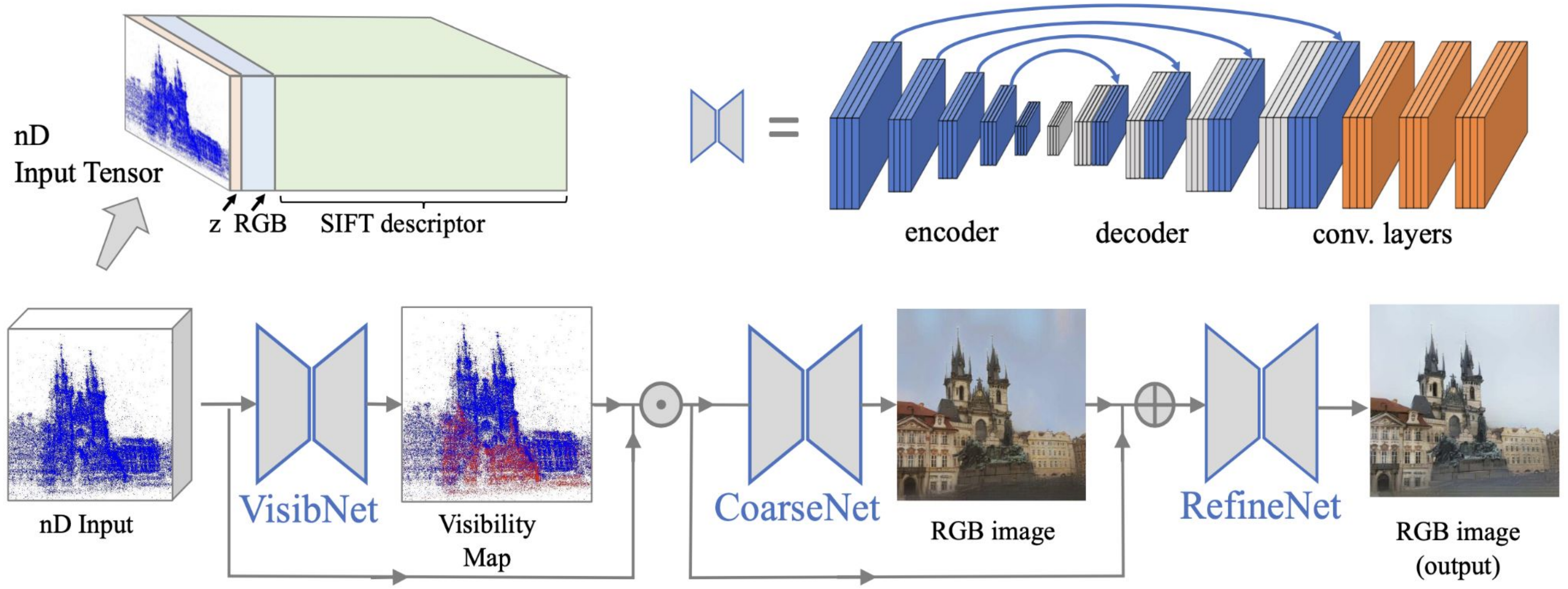


	Method	Easy Maintenance	Privacy	Database Storage (GB, ↓) Cameras (MB)	3D	Raw Imgs	Descs	Total
VM	SIFT [39]	✗	✗	15.73	3.44	✗	130.10 (uint8)	133.33
	CAPS [69]	✗	✗	15.73	3.44	✗	520.38 (fp32)	523.83
	SuperPoint [16]	✗	✗	15.73	3.44	✗	1040.76 (fp32)	1044.21
	Extract on-the-fly	✗	✗	15.73	3.44	157.84	✗	161.29
GM	BPnPNet [9]	✓	✓	15.73	3.44	✗	✗	**3.45**
	GoMatch	✓	✓					

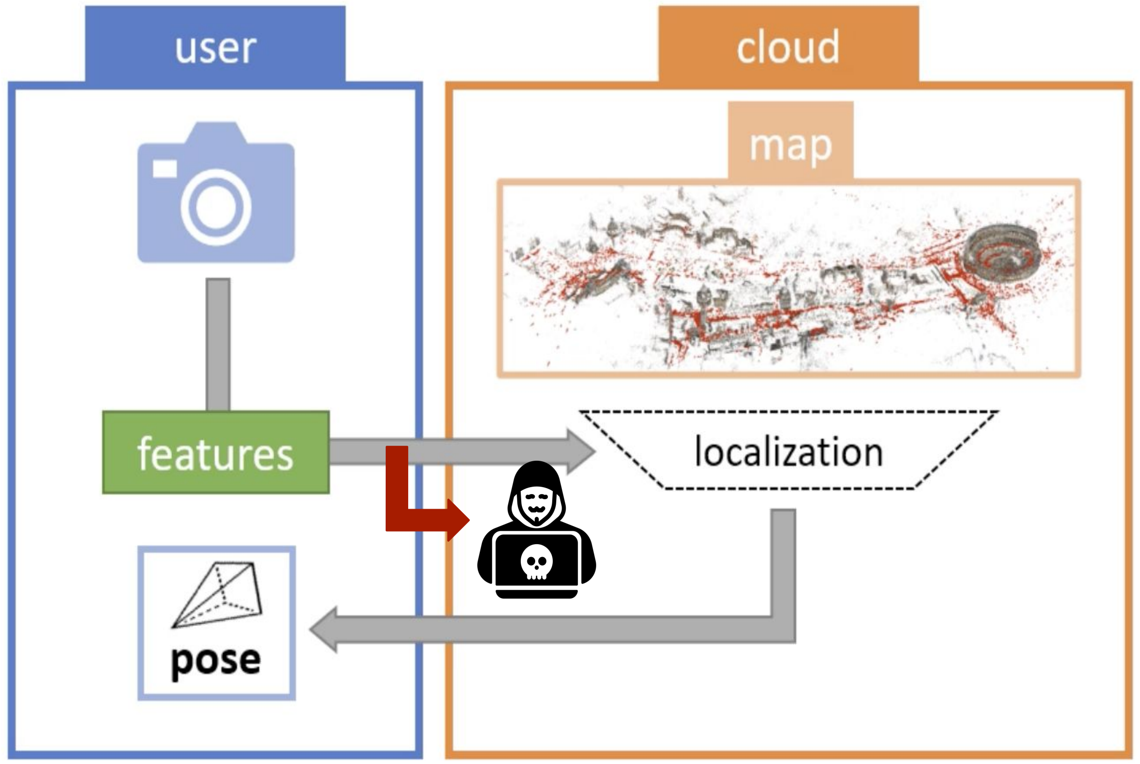


Privacy Challenge

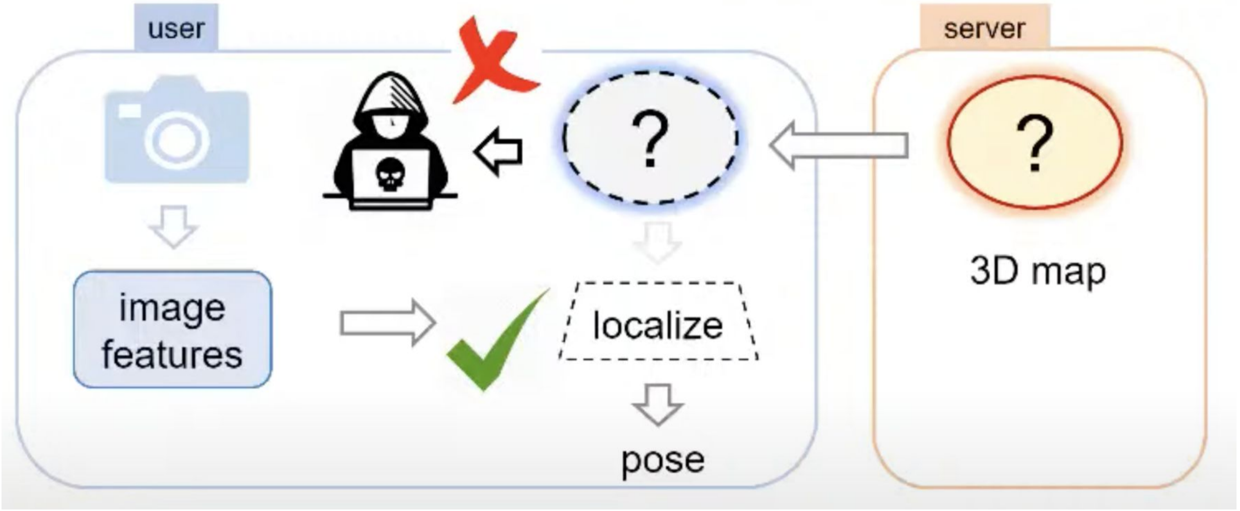
(a) SfM point cloud (top view) (b) Projected 3D points (c) Synthesized Image (d) Original Image



Francesco, Pittaluga, et al Revealing Scenes by Inverting Structure From Motion Reconstructions. CVPR19

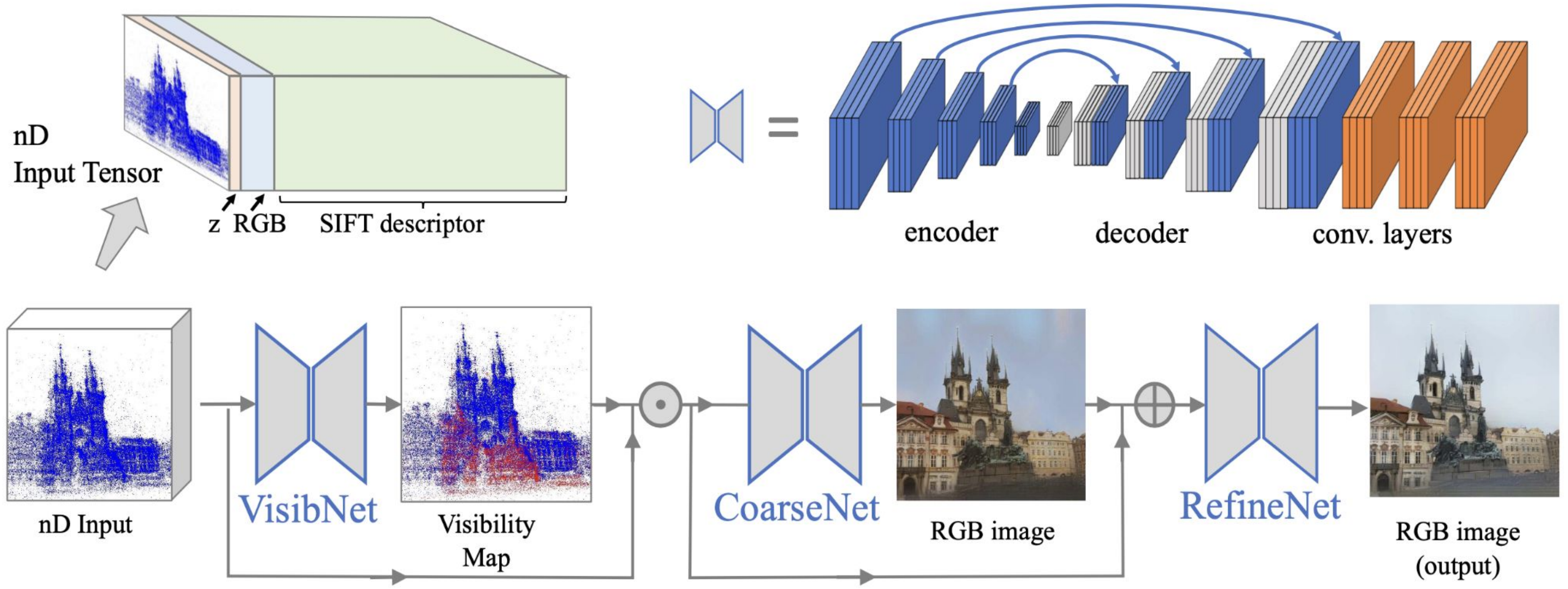


Man-in-the-middle Attack

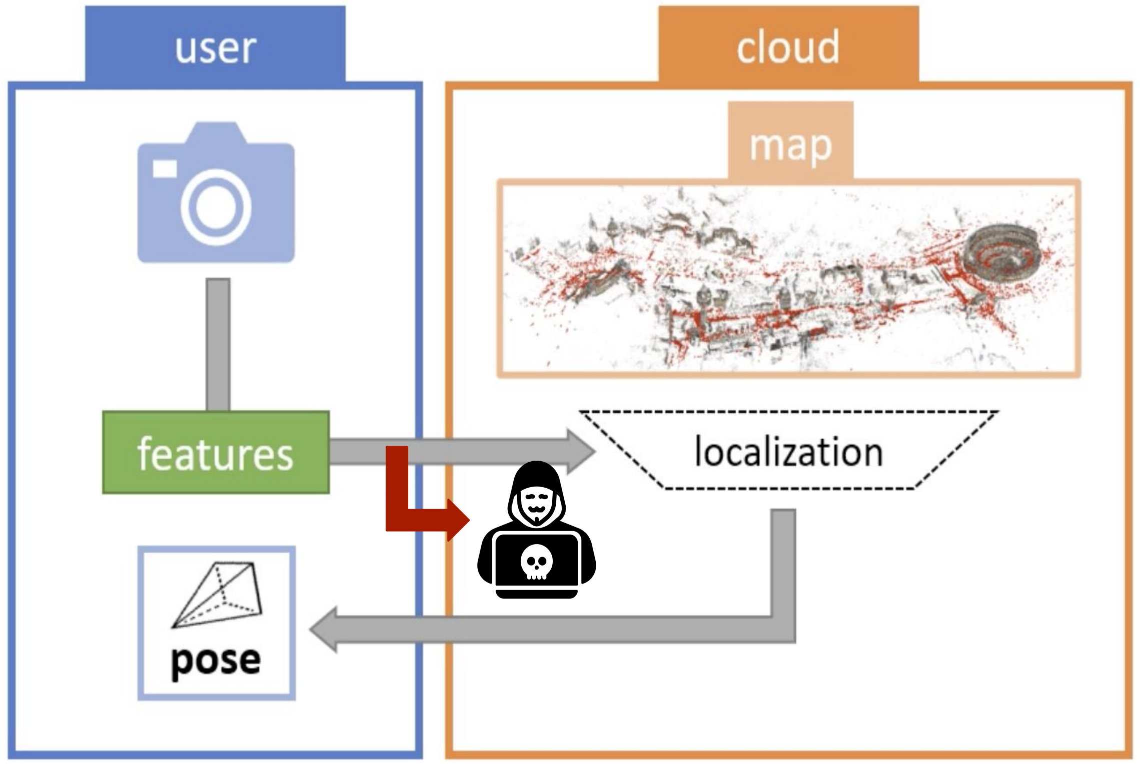


Privacy Challenge

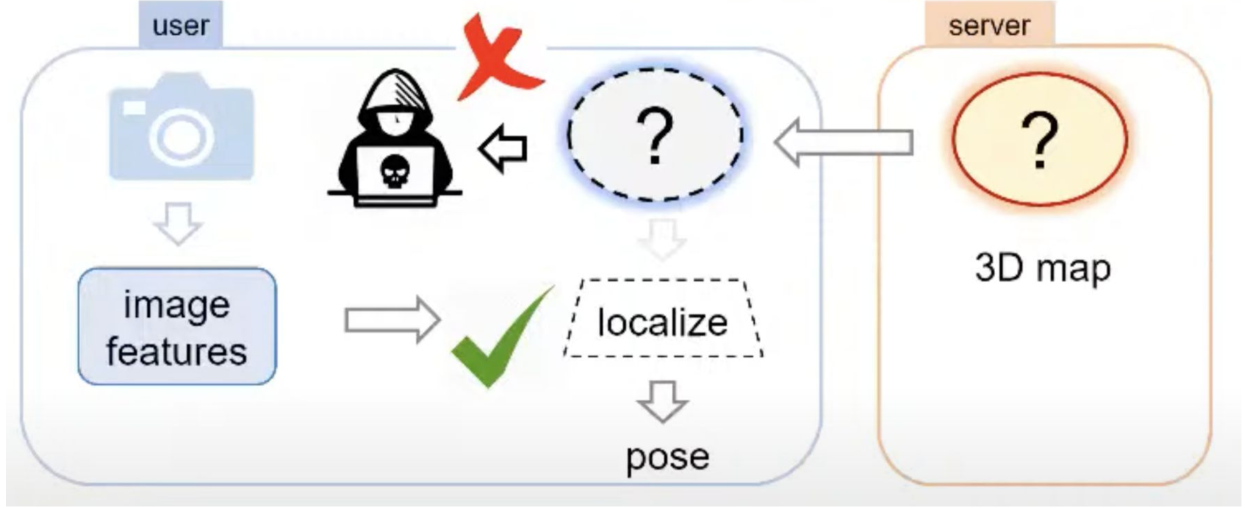
(a) SfM point cloud (top view) (b) Projected 3D points (c) Synthesized Image (d) Original Image



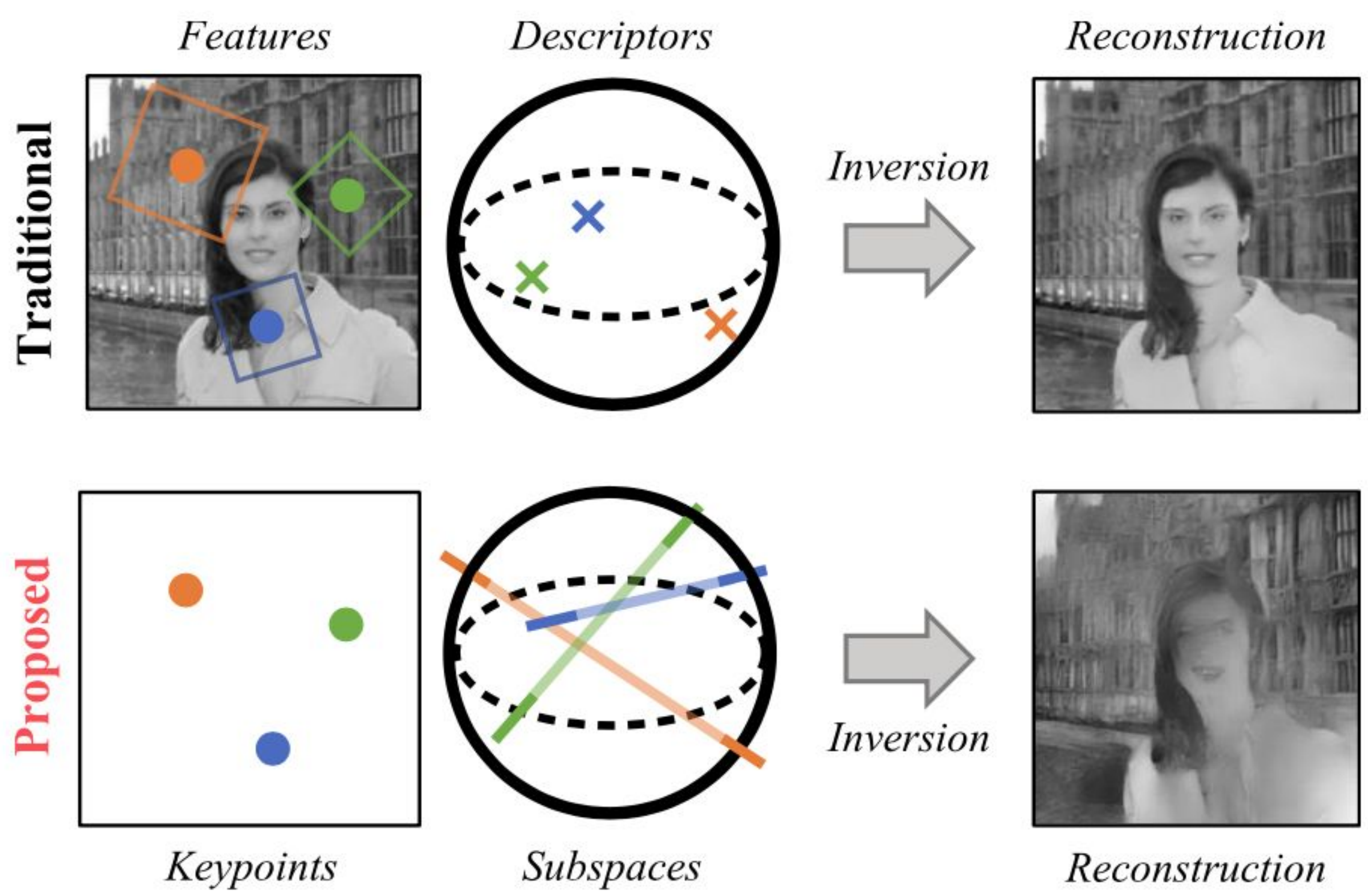
Francesco Pittaluga, Sanjeev J.Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing Scenes by Inverting Structure From Motion Reconstructions. CVPR19



Man-in-the-middle Attack



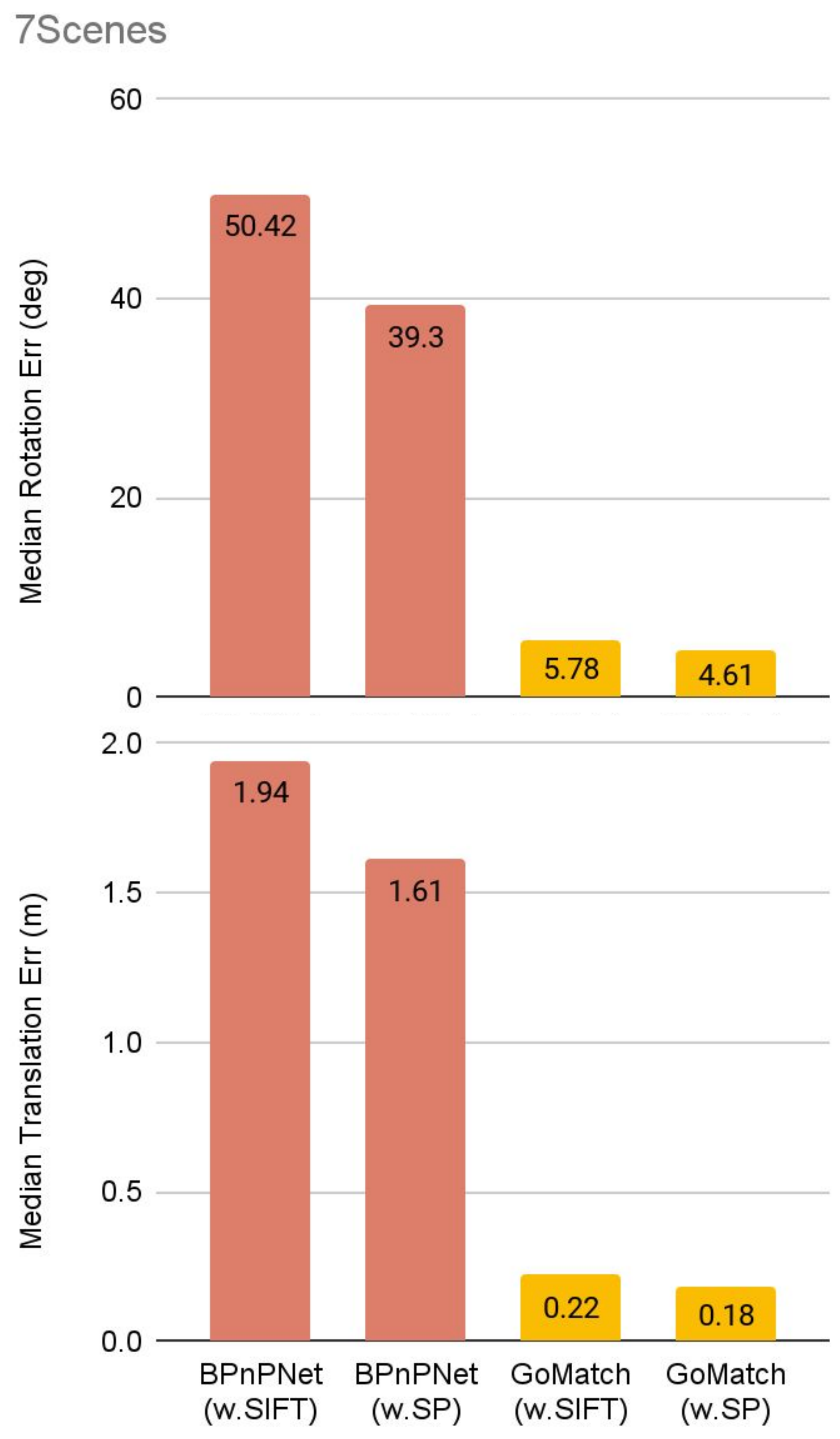
Privacy Challenge

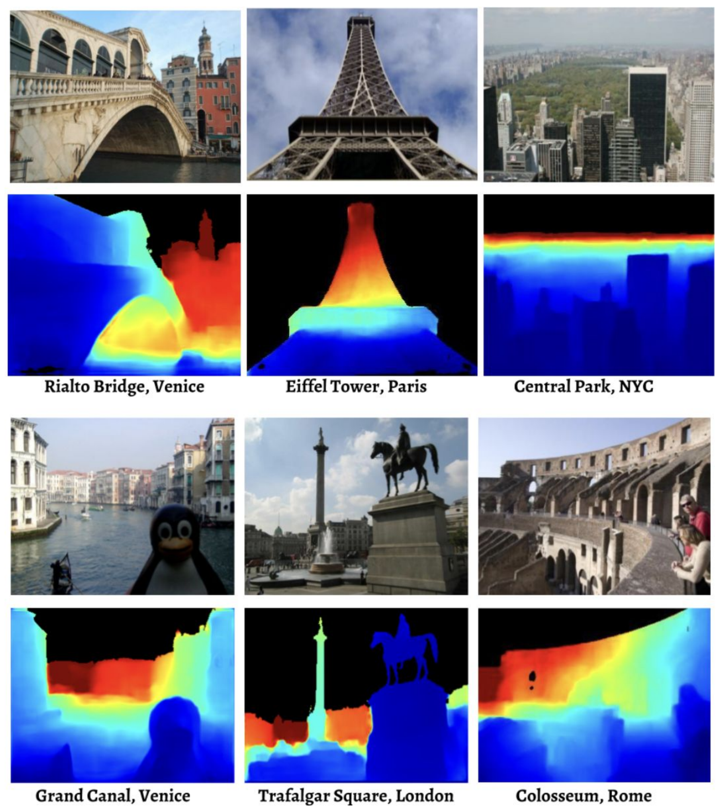


Dusmanu, Mihai, et al. "Privacy-preserving image features via adversarial affine subspace embeddings." CVPR21.

Ng, Tony, et al. "NinjaDesc: Content-Concealing Visual Descriptors via Adversarial Learning." CVPR22

Generalization





Evaluation

GoMatch

BPnPNet

MegaDepth
(Outdoor w.SIFT)