CVPR
JUNE 17-21, 2024
SEATTLE, WA

Learning to Produce Semi-dense Correspondences for Visual Localization

Khang Truong Giang[1], Soohwan Song[2], Sungho Jo[1]
[1]KAIST, [2]Dongguk University

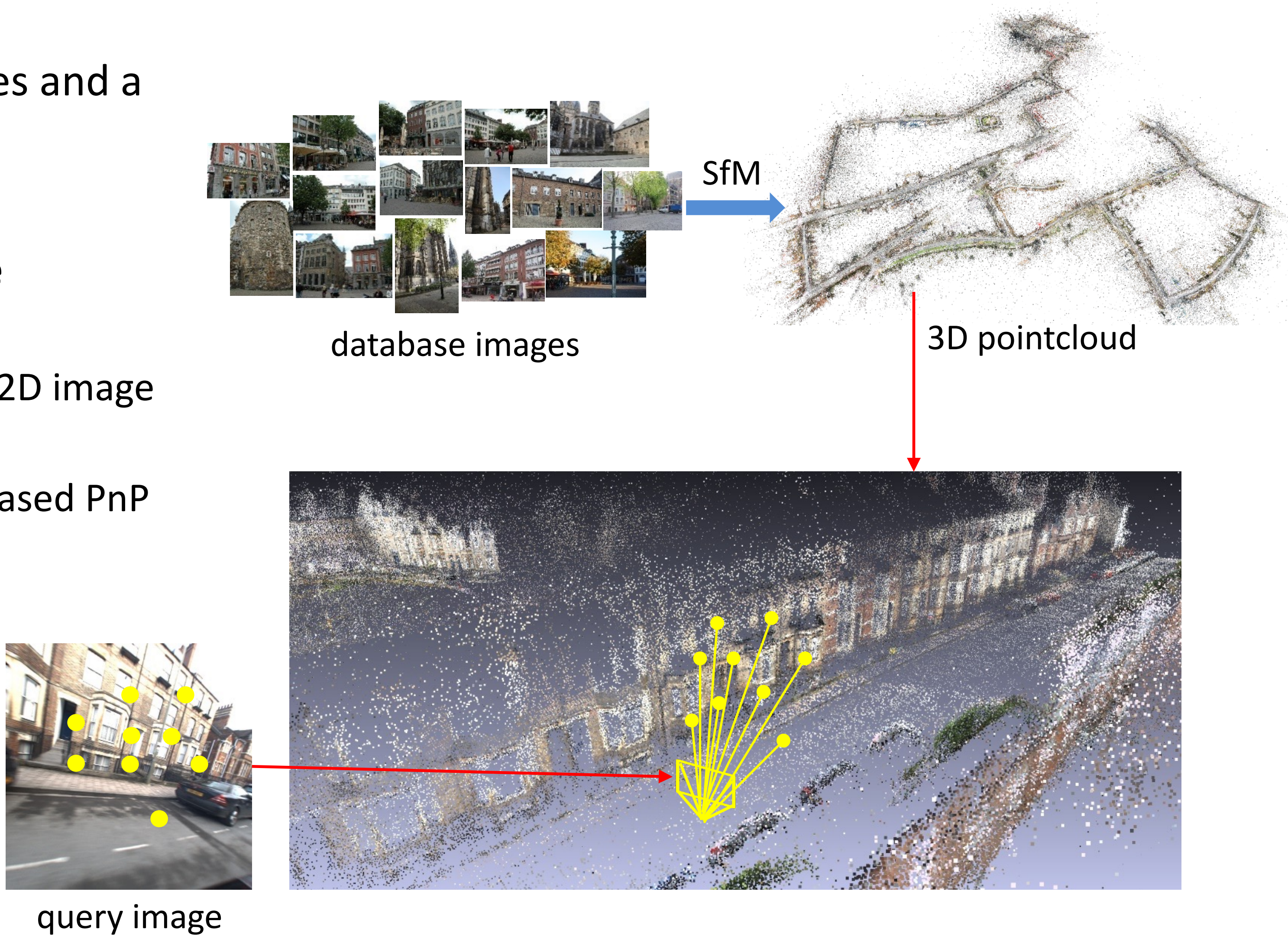
Background

Visual Localization

Estimate camera poses from input images and a known 3D representation of a scene

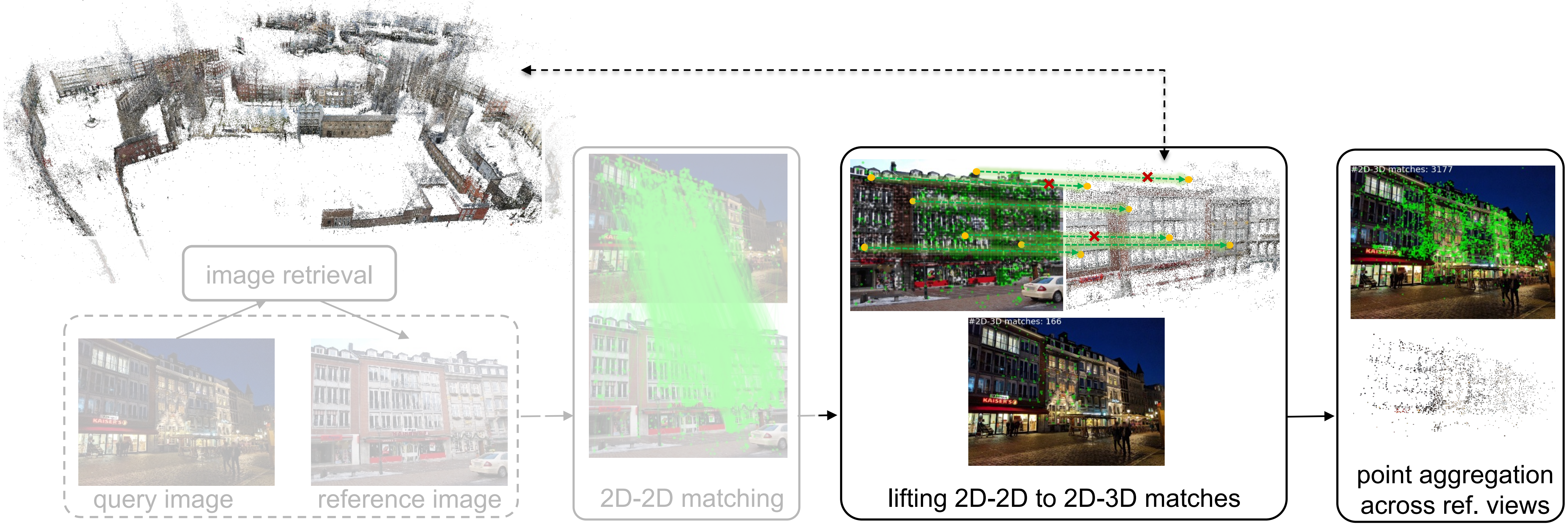
Most common approach: **2D-3D Feature Matching (FM)**

- establish correspondences between 2D image pixels and 3D scene coordinates
- find camera poses using a RANSAC-based PnP algorithm



Existing Challenges of FM-based methods

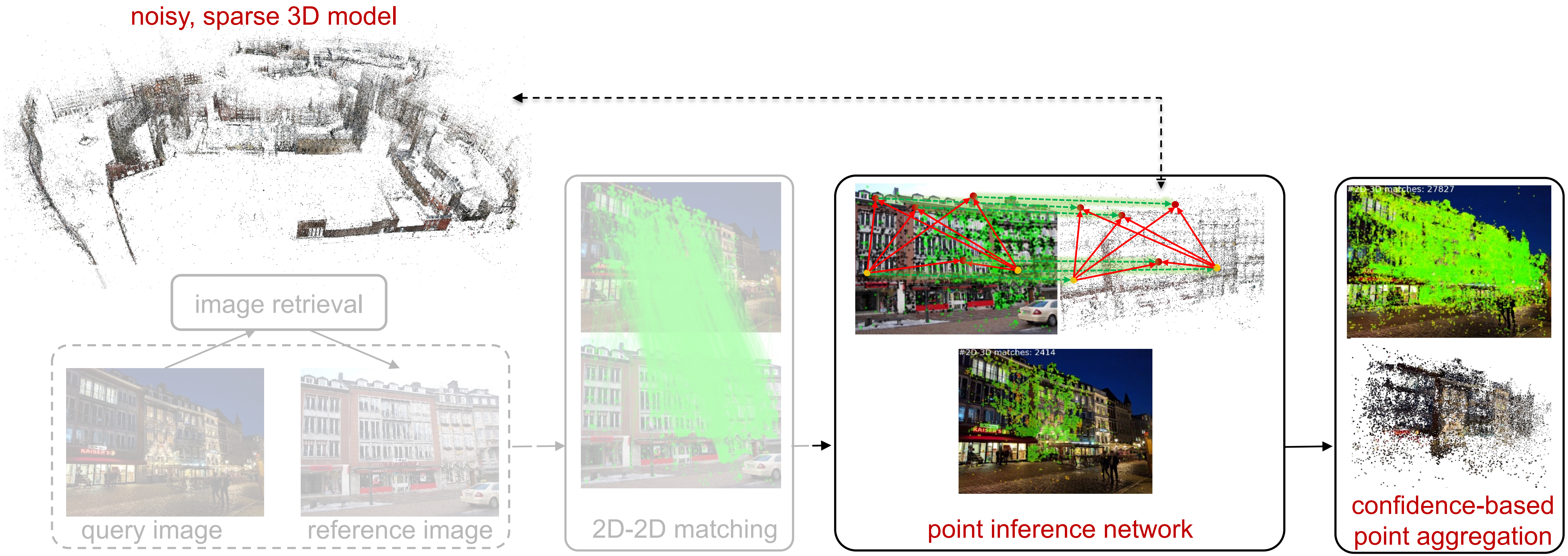
a robust FM-based 3D model built from SfM



Existing SOTA pipeline, HLoc

Motivation

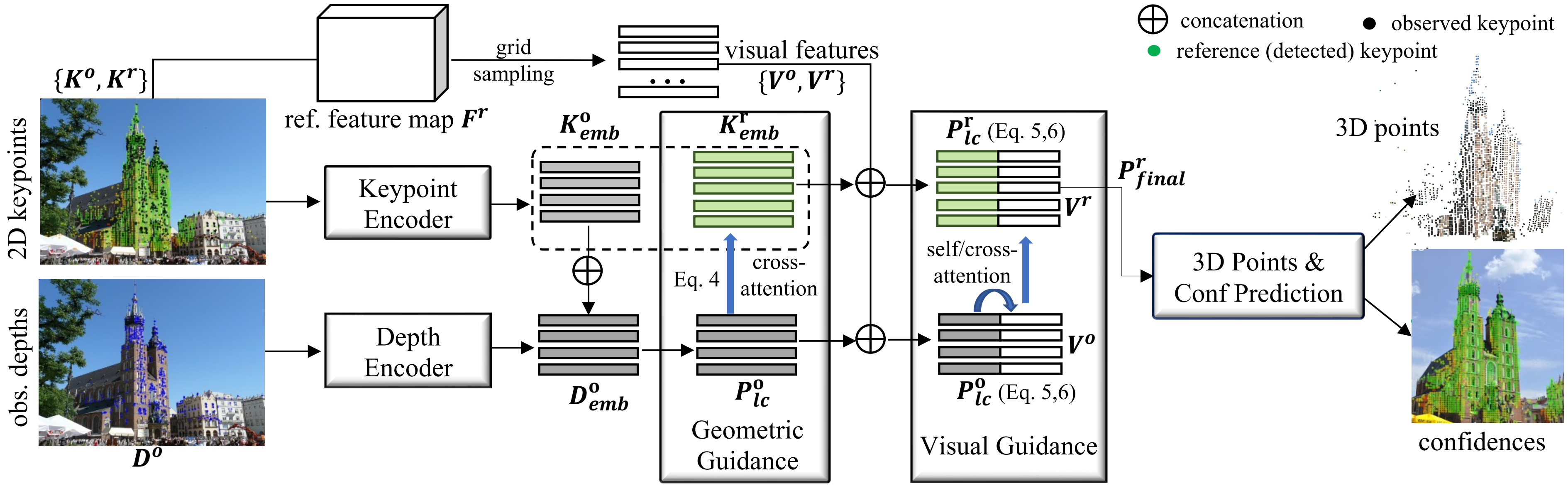
Our approach



DeViLoc

Proposed Method

Point Inference Network (PIN)



Confidence-based Point Aggregation

$$p_j^{agg} = \frac{\sum_{i,n} \boldsymbol{I}\left(Q_s(k_i^n) = k_j\right) c_i^n p_i^n}{\sum_{i,n} \boldsymbol{I}\left(Q_s(k_i^n) = k_j\right) c_i^n}$$

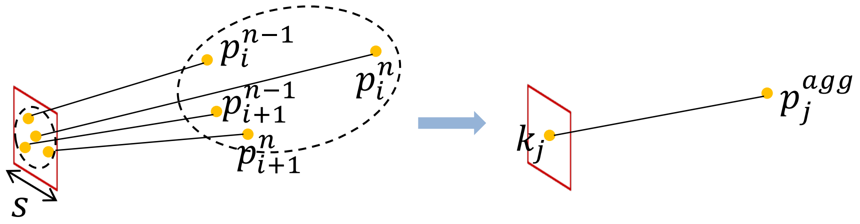
p_j^{agg}: merged 3D point for keypoint k_j

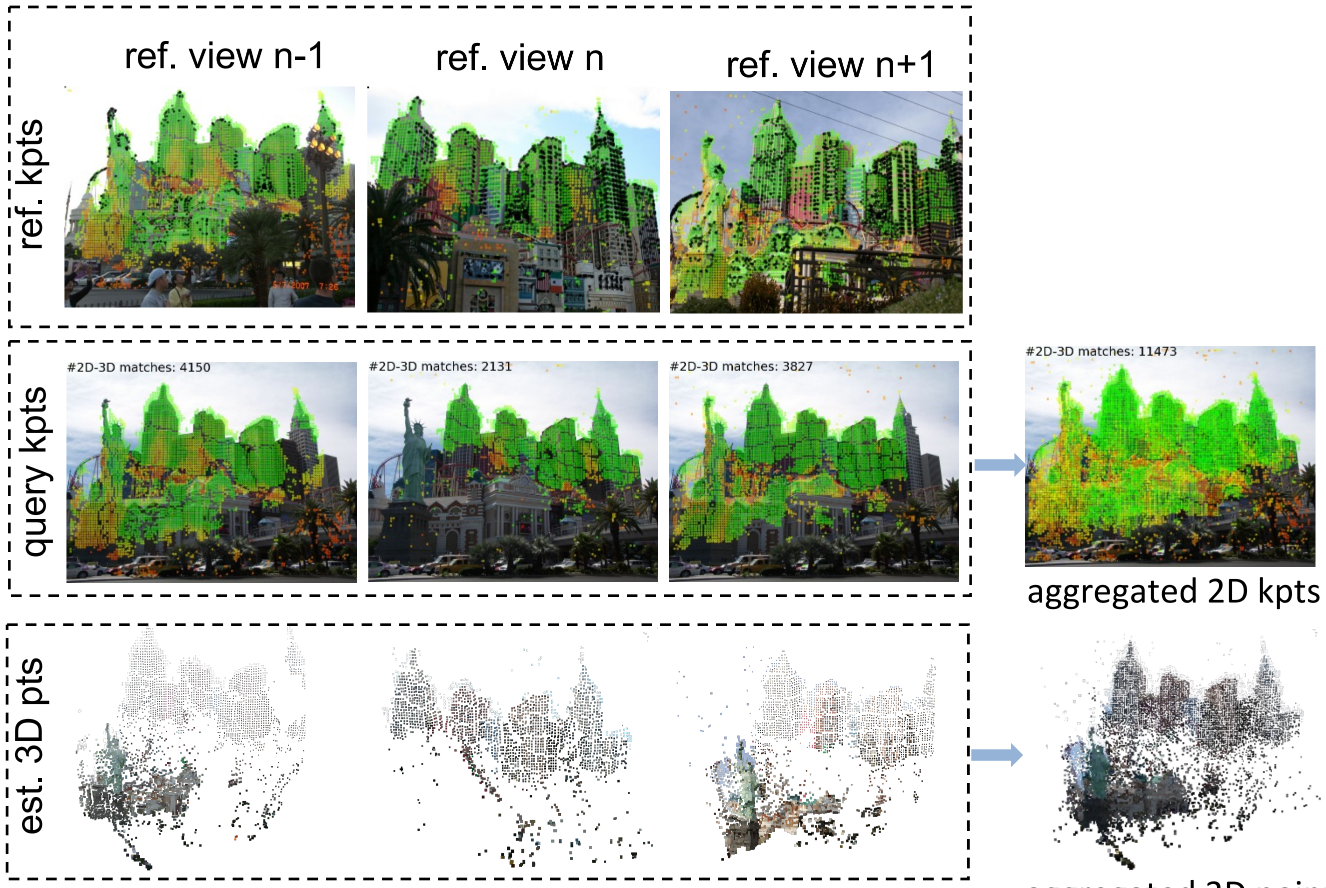
$Q_s(k_i^n) = round\left(\frac{k_i^n}{s}\right) * s$: quantization function

k_i^n: keypoint i associated with reference image n

p_i^n: 3D point i corresponding to keypoint k_i^r

c_i^n: confidence value for 3D point p_i^n





aggregated 2D kpts

aggregated 3D points

Experiments

Challenging benchmarks

	Methods	Aachen Day-Night		RobotCar-Seasons		Extended CMU-Seasons		
		Day	Night	Day-all	Night-all	Urban	Suburban	Park
D	ESAC [7]	42.6 / 59.6 / 75.5	6.1 / 10.2 / 18.4	-	-	-	-	-
S	AS [48]	85.3 / 92.2 / 97.9	39.8 / 49.0 / 64.3	50.9 / 80.2 / 96.6	6.9 / 15.6 / 31.7	81.0 / 87.3 / 92.4	62.6 / 70.9 / 81.0	45.5 / 51.6 / 62.0
S	D2Net [20]	84.8 / 92.6 / 97.5	84.7 / 90.8 / 96.9	54.5 / 80.0 / 95.3	20.4 / 40.1 / 55.0	94.0 / 97.7 / 99.1	93.0 / 95.7 / 98.3	89.2 / 93.2 / 95.0
S	S2DNet [23]	84.5 / 90.3 / 95.3	74.5 / 82.7 / 94.9	53.9 / 80.6 / 95.8	14.5 / 40.2 / 69.7	-	-	-
S	HLoc[SP] [18, 42]	80.5 / 87.4 / 94.2	68.4 / 77.6 / 88.8	53.1 / 79.1 / 95.5	7.2 / 17.4 / 34.4	89.5 / 94.2 / 97.9	76.5 / 82.7 / 92.7	57.4 / 64.4 / 80.4
S	PixLoc [44]	64.3 / 69.3 / 77.4	51.0 / 55.1 / 67.3	52.7 / 77.5 / 93.9	12.0 / 20.7 / 45.4	88.3 / 90.4 / 93.7	79.6 / 81.1 / 85.2	61.0 / 62.5 / 69.4
S	HLoc[SP+SG]	**89.6** / 95.4 / **98.8**	86.7 / **93.9** / **100.**	**56.9** / 81.7 / **98.1**	33.3 / 65.9 / 88.8	95.5 / 98.6 / **99.3**	90.9 / 94.2 / 97.1	85.7 / 89.0 / 91.6
S	LBR [65]	88.3 / **95.6** / **98.8**	84.7 / **93.9** / **100.**	56.7 / 81.7 / **98.2**	24.9 / 62.3 / 86.1	-	-	-
S	HLoc+PixLoc	84.7 / 94.2 / **98.8**	81.6 / **93.9** / **100.**	**56.9** / **82.0** / **98.1**	**34.9** / 67.7 / 89.5	**96.9** / **98.9** / **99.3**	93.3 / 95.4 / 97.1	87.0 / 89.5 / 91.6
S	HLoc[TopicFM][24]	88.8 / 94.7 / 97.9	86.7 / 92.9 / **100.**	-	-	-	-	-
SD	NeuMap [58]	80.8 / 90.9 / 95.6	48.0 / 67.3 / 87.8	-	-	-	-	-
SD	DeViLoc (Ours)	87.4 / 94.8 / 98.2	**87.8** / **93.9** / **100.**	**56.9** / 81.8 / **98.0**	31.3 / **68.9** / **92.4**	95.7 / 98.4 / **99.2**	**97.1** / **98.3** / **99.4**	**92.1** / **95.1** / **96.3**

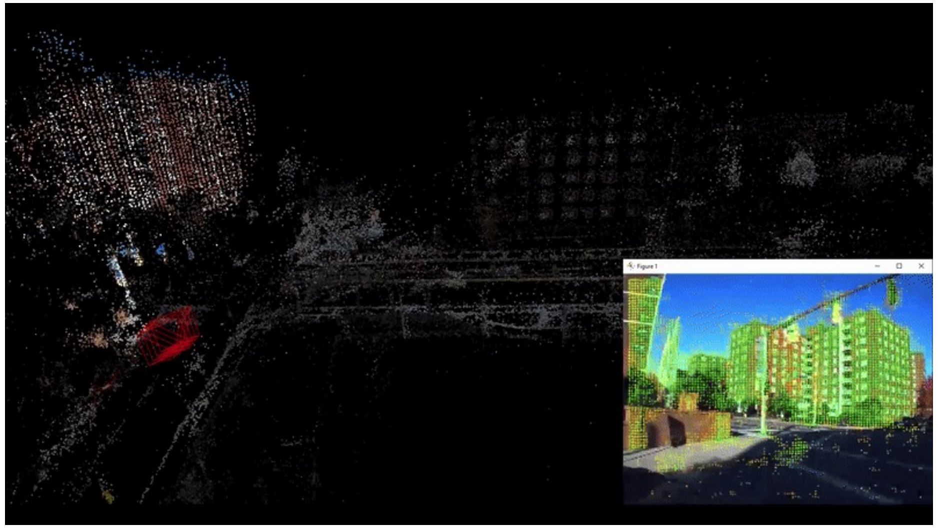
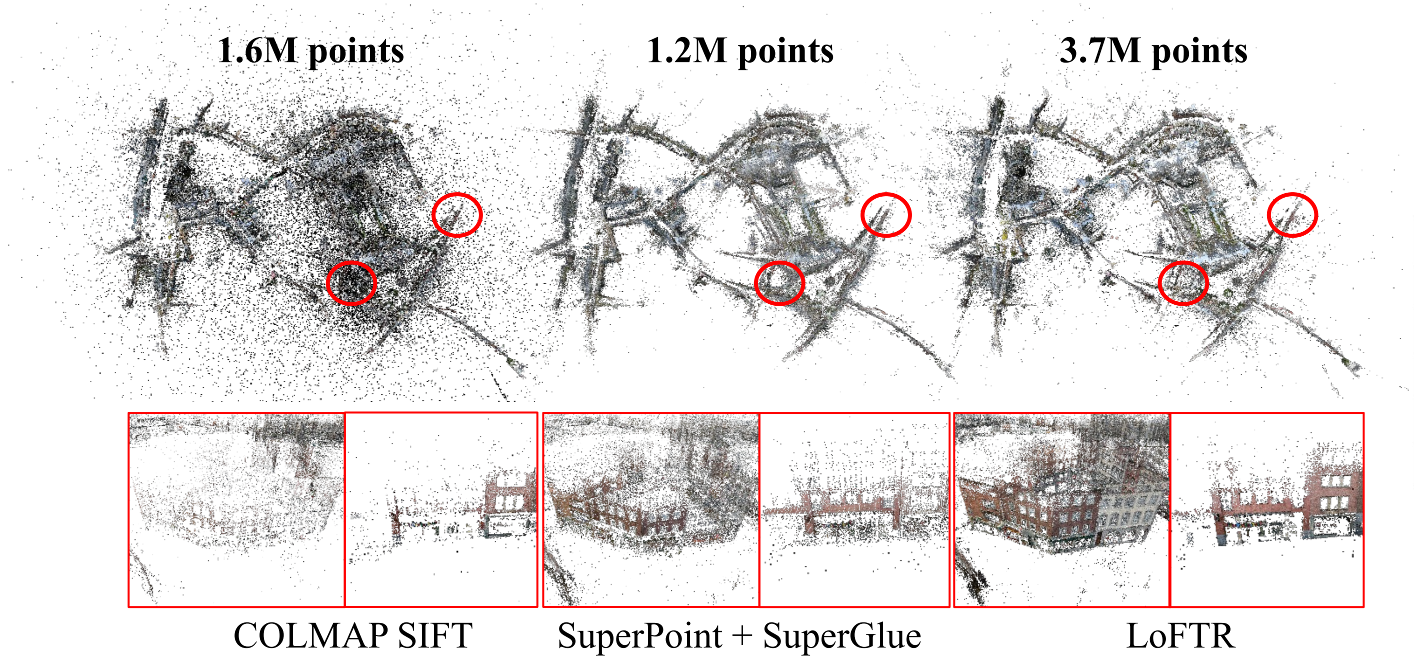


Illustration of pose estimation on CMU-Seasons

Experiments

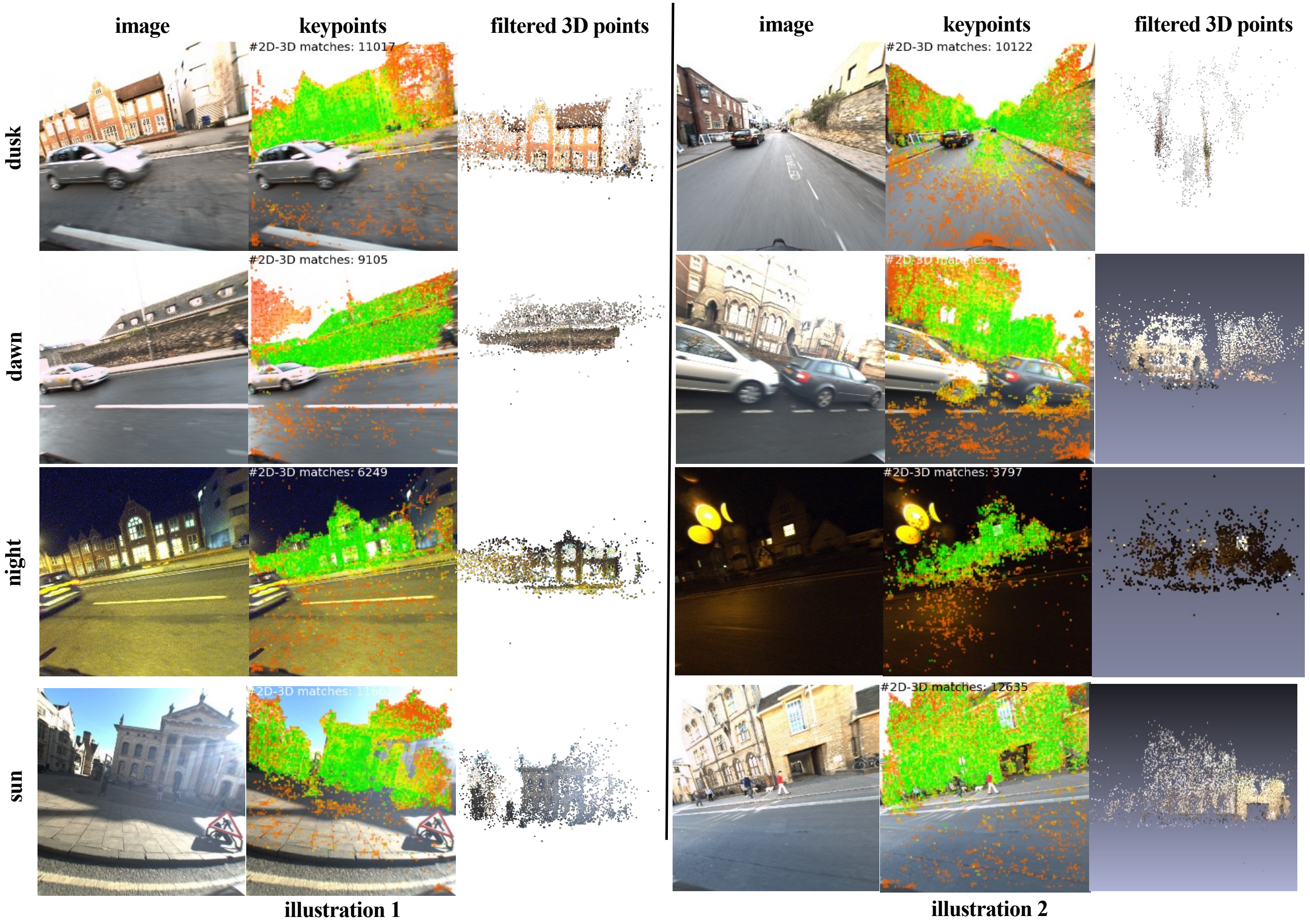
Robustness with noisy/sparse 3D pointcloud inputs



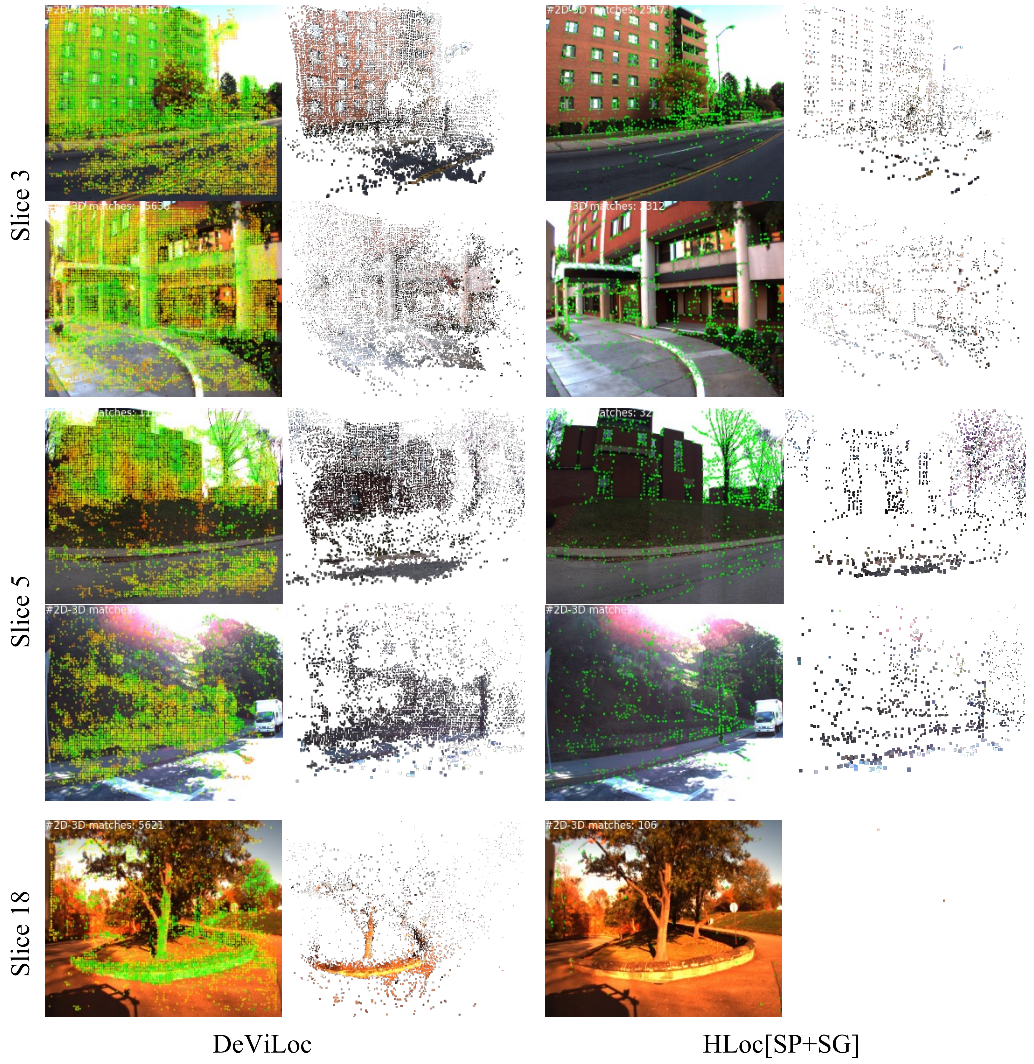
Models	Day	Night
	(0.25m,2°) / (0.5m,5°) / (5.0m,10°)	
DeViLoc[half-SIFT]	87.5 / 94.1 / 97.9	86.7 / 92.9 / **100.**
DeViLoc[SIFT]	87.4 / 94.8 / 98.2	87.8 / **93.9** / **100.**
DeViLoc[SP+SG]	87.3 / **95.3** / **98.3**	**88.8** / 92.9 / **100.**
DeViLoc[LoFTR]	**87.9** / 94.7 / 98.2	**88.8** / 92.9 / **100.**

Experiments

Qualitative Results



RobotCar dataset



CMU dataset

Thanks for listening!