Improved Scene Landmark Detection for Camera Localization

Tien Do* Tesla

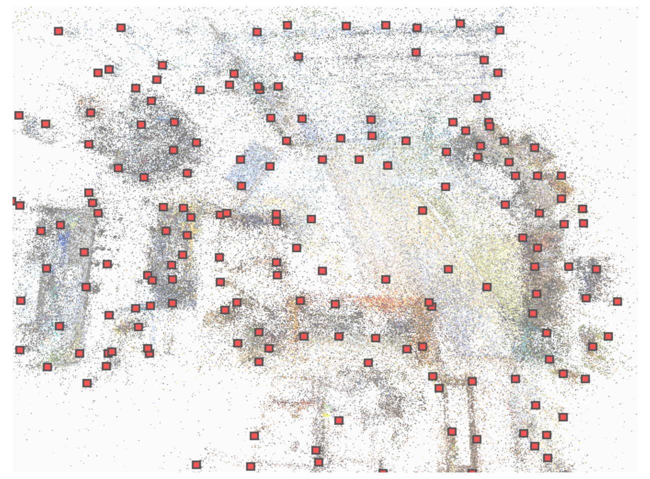
Sudipta N. Sinha Microsoft

* work done while author was at Microsoft.

3DV 2024

Recap: Scene Landmarks Detection (SLD)

Scene landmarks are salient scene-specific 3D points, that can be used for 6-DoF camera localization in pre-mapped scenes.



Approach:

- Specify scene landmark points in 3D scene coordinates.
- Train detector (CNN-based heatmap predictor) to predict visible scene landmarks (2D pixel locations) in RGB images.
- Compute 6-DoF camera pose from the 2D--3D scene landmark correspondences.

Our Contributions

- We show that the accuracy gap between **SLD** [B] and **hloc** [A] and **SLD**'s inability to handle many scene landmarks was due to insufficient model capacity in the SLD architecture.
- We propose to partition the landmark set and train an ensemble of networks, one per subset of landmarks.
- We propose a compact architecture, and a method to generate better training labels for training **SLD** and **SLD***.
- **SLD*** significantly outperforms **SLD** [B]. It is competitive with **hloc** [A] but 40X faster and 20X more storage efficient.

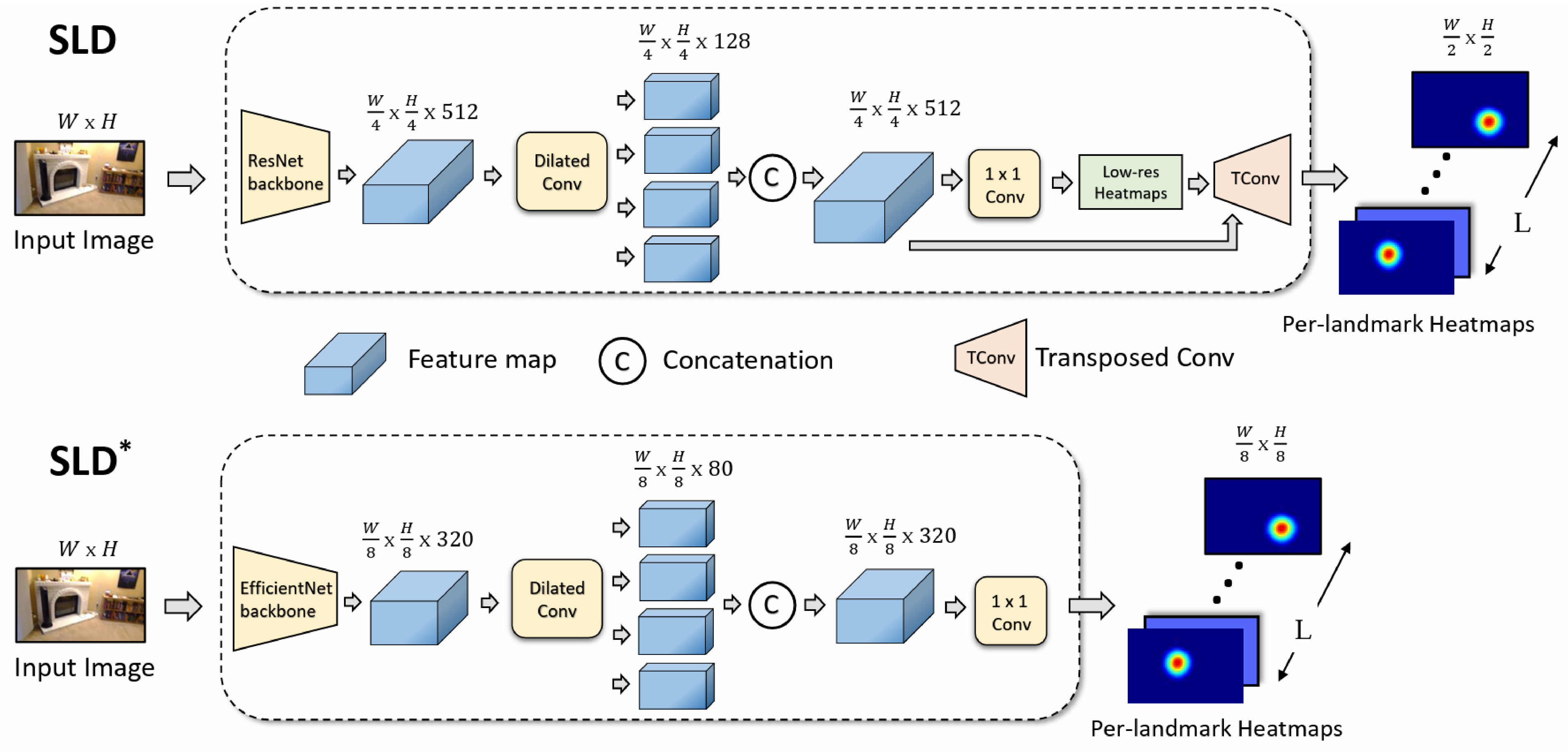
Related Work

A. [hloc] Sarlin et al., From Coarse to Fine: Robust Hierarchical Localization at Large Scale, CVPR 2019.

B. [SLD] Do et al., Learning to Detect Scene Landmarks for Camera Localization, CVPR 2022.

C. [DSAC*] Brachmann and Rother, Visual Camera Re-Localization From RGB and RGB-D Images using DSAC, T-PAMI 2022.

1. Compact Network Architecture (SLD*)



2. Partitioning the Landmark Set

- Partition the landmark set into mutually exclusive subsets. Train an ensemble of networks, one per subset.

subset size x #nets	200×1	300×1	100×3	100×4	125×6	125×8	125×12
R @ 5cm/5° ↑	46.0	50.8	61.1	63.0	66.6	**70.1**	69.1
Time (sec.) ↓	**0.05**	0.11	0.16	0.19	0.23	0.3	0.5
Size (MB) ↓	**15**	**15**	45	60	90	120	180

- Using more scene landmarks improves scene coverage and pose accuracy in larger scenes.

300 landmarks 1000 landmarks

Indoor-6 Dataset (*Do et. al. 2022*)

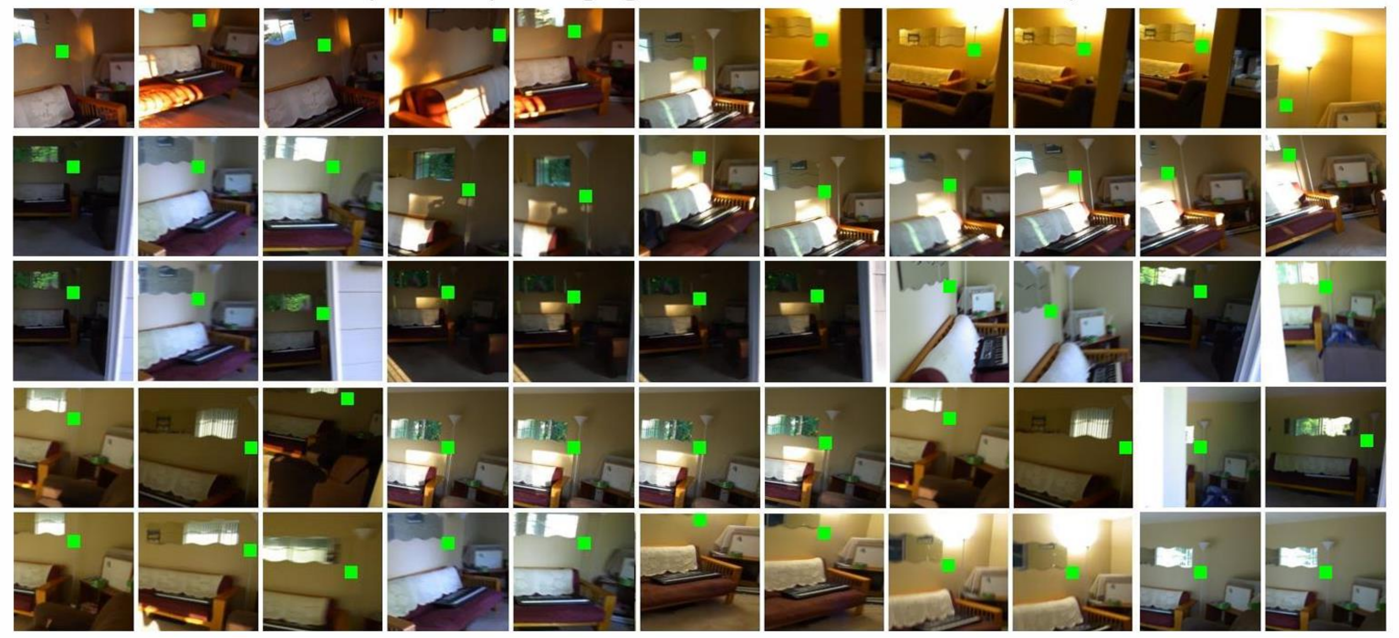
- Images span multiple days and times of day.
- Non-static geometry.
- Dramatic lighting changes.

3. Improved Visibility Estimation

- Reconstruct 3D mesh and use it to infer visibility of scene landmarks. Generates better training labels.

visible patches (leveraging SfM)

visible patches (leveraging SfM and dense reconstruction)

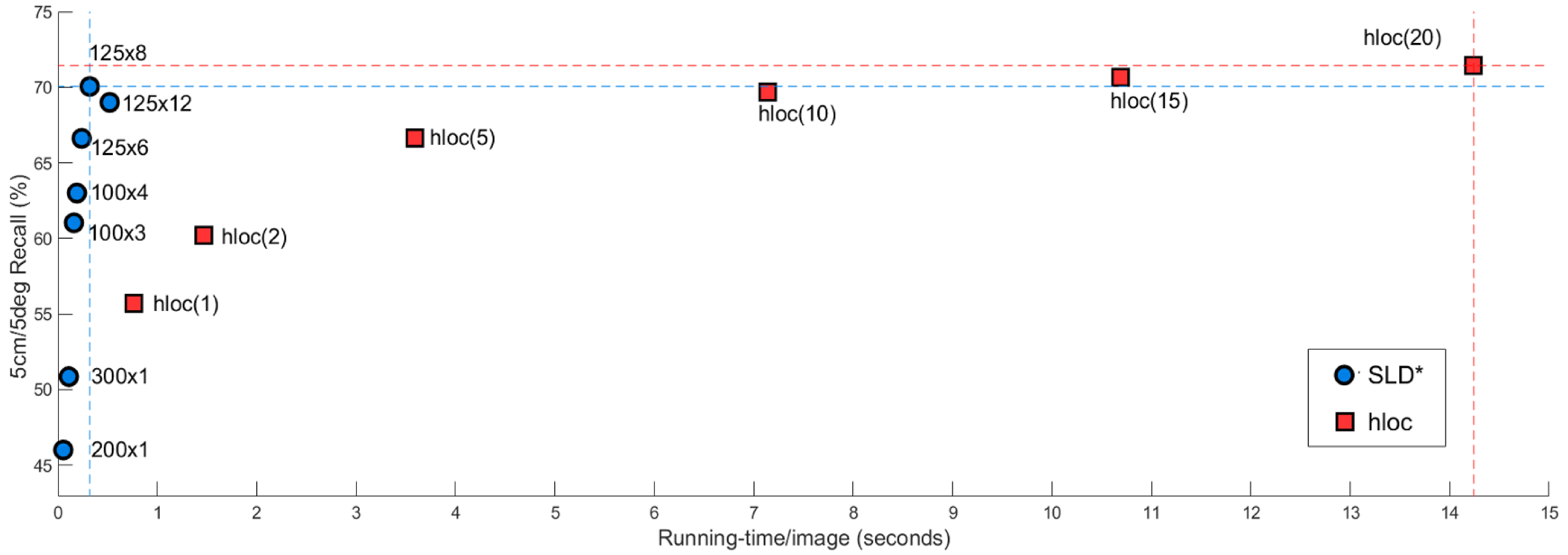


Results

	Scene	DSAC* [6]	NBE+SLD [11]	SLD [11]	SegLoc [21]	SLD* ours	hloc-l_{1000} [11]	hloc-l_{3000} [11]	hloc-A [11, 26]	hloc-B [26]	SLD* ours
#landmarks		n/a	300	300	n/a	300	1000	3000	n/a	n/a	1000
R@5cm/5° ↑	scene1	18.7	38.4	35.0	51.0	47.2	33.3	48.1	64.8	**70.5**	68.5
	scene2a	28.0	–	34.6	56.4	48.2	12.5	17.1	51.4	52.1	**62.6**
	scene3	19.7	53.0	50.8	41.8	56.2	48.3	61.9	81.0	**86.0**	76.2
	scene4a	60.8	–	56.3	33.8	67.7	34.8	39.2	69.0	75.3	**77.2**
	scene5	10.6	40.0	43.6	43.1	33.7	21.9	31.1	42.7	**58.0**	57.8
	scene6	44.3	50.5	48.9	34.5	52.0	47.4	59.1	79.9	**86.7**	78.0
R@5cm/5° ↑	avg.	30.4	45.5	44.9	43.4	50.8	33.0	42.8	64.8	**71.4**	70.1
Size (GB) ↓		0.027	0.135	0.020	0.161	0.015	0.17–0.21	0.2–0.5	0.7–2.4	0.7–2.4	0.120
Mem. (GB) ↓		0.85	1.35	1.2	–	0.99	1.3	1.3	1.3	1.3	0.99

- Recall @ 5cm/5° (in %), storage used (Size), and in-memory footprint (Mem.)
- **SLD*** is competitive with **hloc-B** (using latest code) but uses significantly less storage.

Accuracy/speed tradeoff of SLD* and hloc



- hloc's performance depends on the number of matched image pairs (1, 2, 5, … 20). 20 pairs has the best recall (71.4%) but a high running time of 14.2 seconds/image.
- Amongst seven **SLD*** ensembles, *125 x 8 = 1000 landmarks* has the best recall (70.1%) with running time of 0.3 sec./image. (40X faster than hloc).

Code & Data

https://github.com/microsoft/SceneLandmarkLocalization