1 UNIVERSITY OF MINNESOTA

2 Microsoft

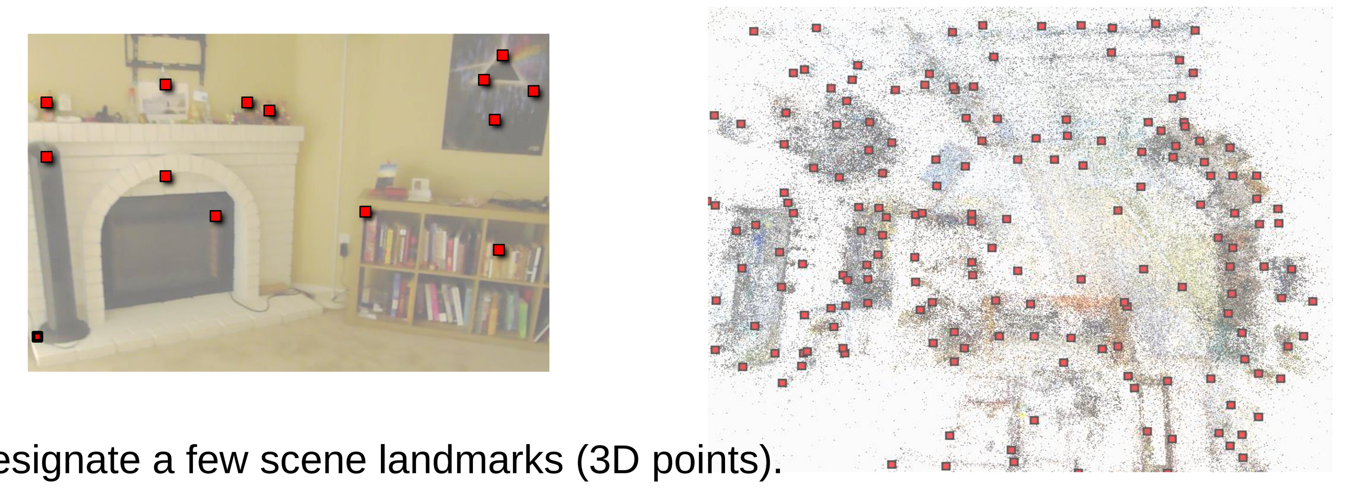
Learning to Detect Scene Landmarks for Camera Localization

Tien Do[1] Ondrej Miksik[2] Joseph DeGol[2] Hyun Soo Park[1] Sudipta N. Sinha[2]

Goal

We present a method to compute the exact 3D position and 3D orientation of the camera within a precomputed 3D map of the scene from a query image. We solve the task accurately and efficiently *i.e.*, without requiring extensive storage of visual features which helps to further address privacy concerns in existing localization techniques.

Main idea: Scene Landmarks Detection

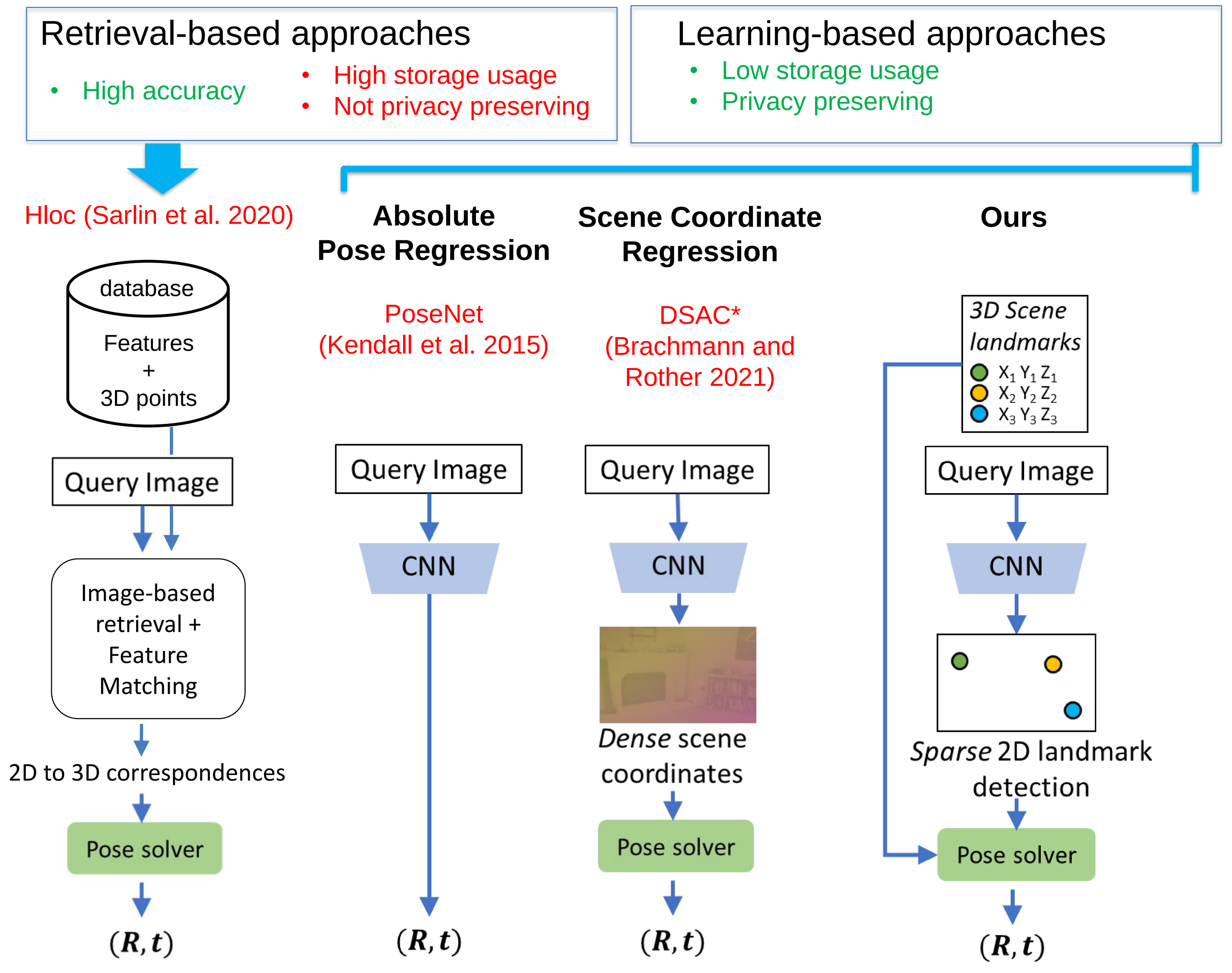


- Designate a few scene landmarks (3D points).
- Learn a detector to localize those scene landmarks in a query image.
- Estimate camera pose from the 2D-3D scene landmark correspondences.

Our contribution

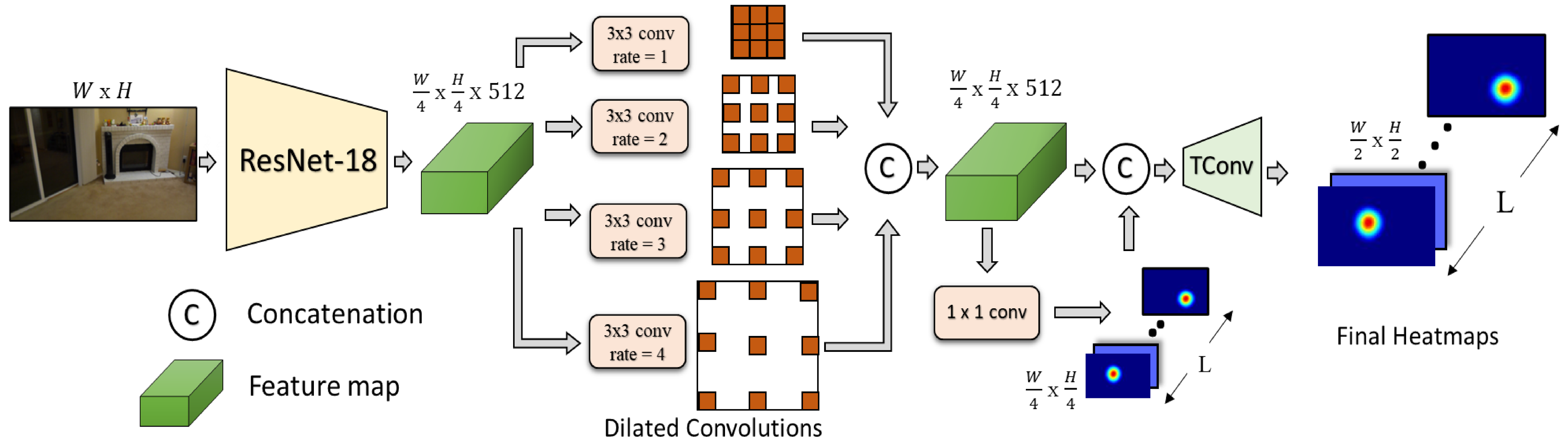
1. New formulation for *heatmap-based landmark localization* and *bearing angle estimation* for solving the camera localization problem.
2. New dataset to address challenging scenarios in indoor environments.
3. Superior results compared to existing learned localization methods

Comparison with related work



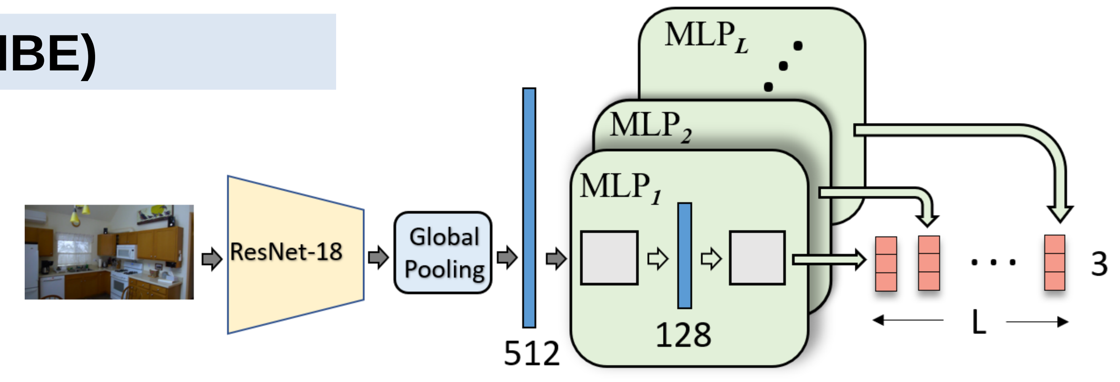
Scene Landmark Detector (SLD)

- Leverage mature CNN architecture for heatmap-based keypoint detection, commonly used in many detection and pose estimation tasks (face, body pose, hands, object, etc)



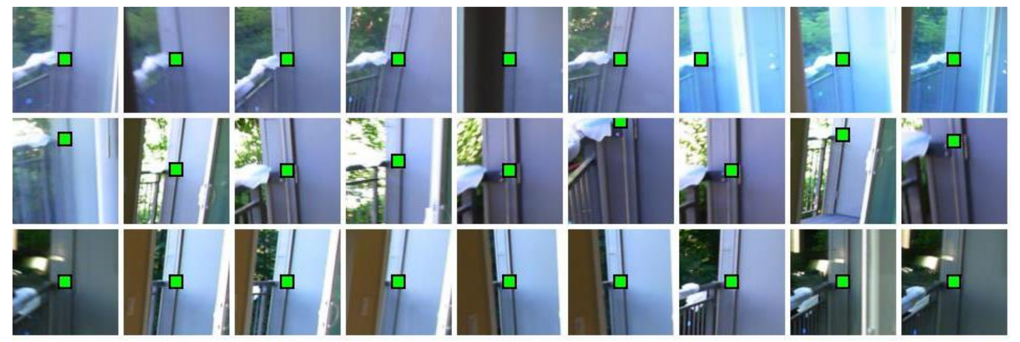
Neural Bearing Estimator (NBE)

- Directly predict landmark bearing vector (3D) from the image appearance.
- Can predict bearings for landmarks outside the camera's field-of-view.



Landmark Selection

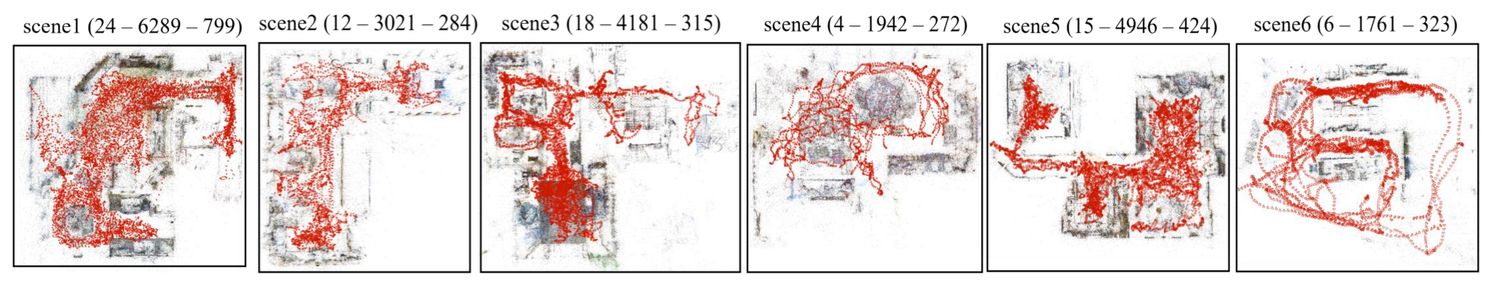
- Run structure-from-motion on training images.
- Select a subset of salient points (discriminative, repeatable, permanent) that maximizes scene coverage,
- Selected using an iterative greedy approach.



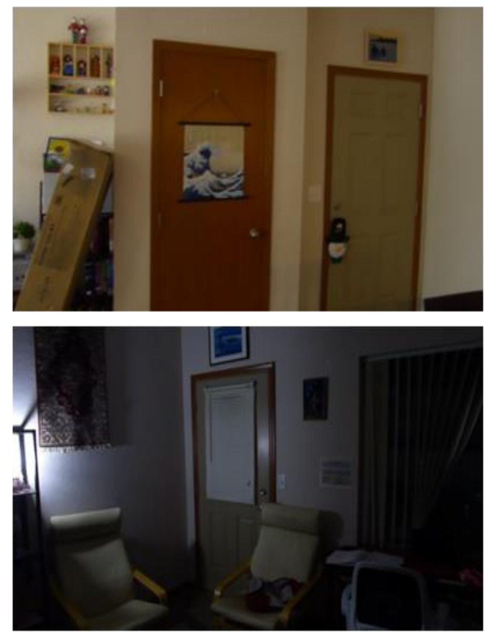
Example of training image patches depicting a scene landmark.

Indoor-6 dataset

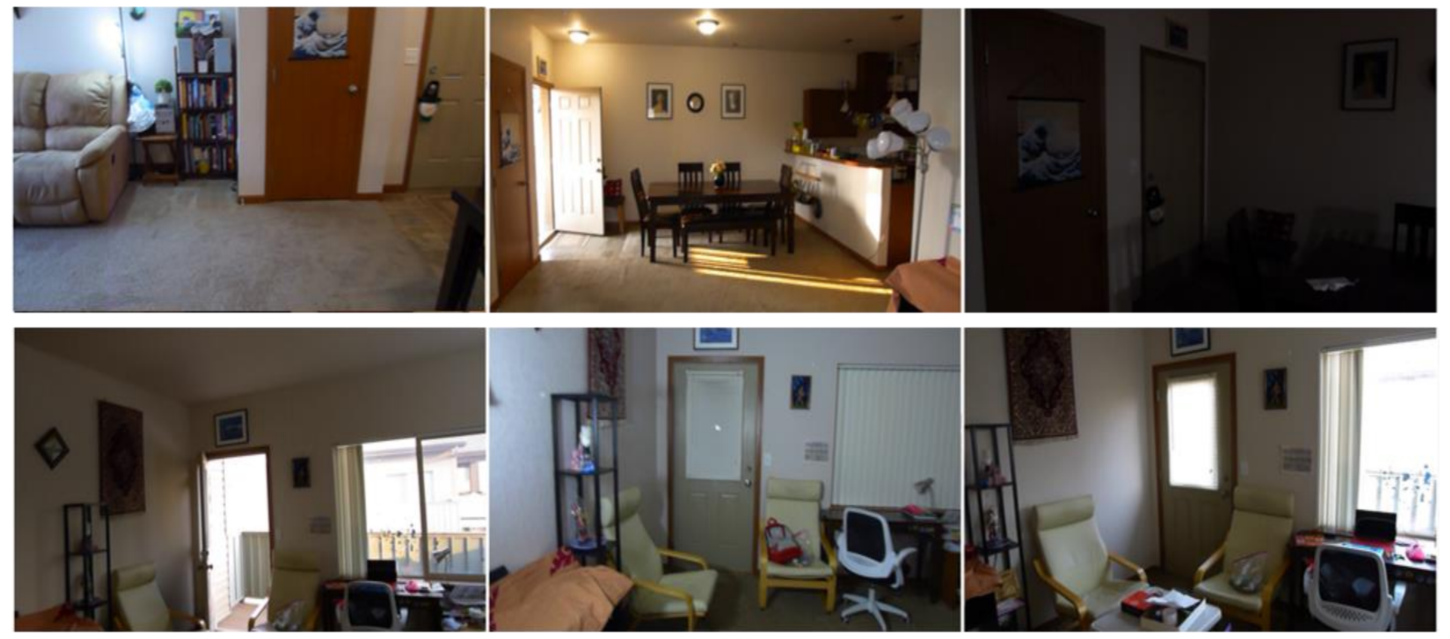
- Images span multiple days and times (incl. day/night images)
- Dramatic lighting variation
- Scene changes with time.



Test images



Train images



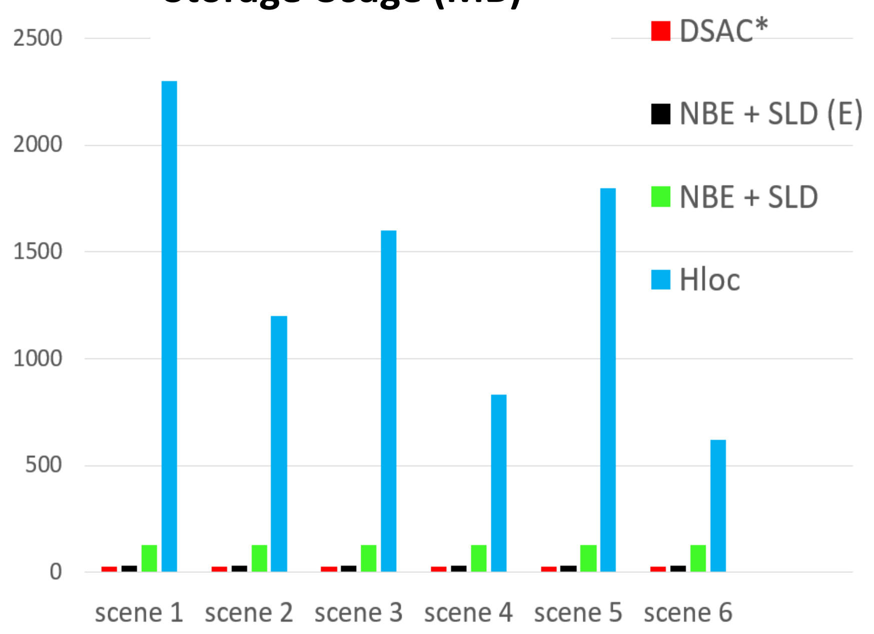
Qualitative Results

Results (Indoor-6)

Accuracy evaluation: **NBE+SLD**: uses a ResNet18 backbone, **NBE+SLD(E)**: uses an Efficient-Net (compact) backbone.

Method	INDOOR-6 scene1 (cm.)↓	(deg.)↓	(%)↑	scene2 (cm.)↓	(deg.)↓	(%)↑	scene3 (cm.)↓	(deg.)↓	(%)↑	scene4 (cm.)↓	(deg.)↓	(%)↑	scene5 (cm.)↓	(deg.)↓	(%)↑	scene6 (cm.)↓	(deg.)↓	(%)↑
PoseNet	159.0	7.46	0.0	193.0	8.42	0.0	141.0	9.26	0.0	109.4	7.84	0.0	179.3	9.37	0.0	118.2	9.26	0.0
NBE	22.3	4.03	2.0	29.9	4.88	2.1	24.7	4.85	2.9	39.9	5.35	1.5	37.8	5.28	0.0	30.8	6.60	0.3
DSAC*	12.3	2.06	18.7	17.5	3.4	12.3	13.1	2.34	19.7	5.5	0.84	44.9	40.7	6.72	10.6	6.0	1.40	44.3
NBE+SLD(E)	7.5	1.15	28.4	11.8	2.30	26.1	6.2	1.28	43.5	5.1	0.75	48.9	6.3	0.96	37.5	5.8	1.30	44.6
NBE+SLD	**6.5**	**0.9**	**38.4**	**7.4**	**1.6**	**37.0**	**4.4**	**0.91**	**53.0**	**4.0**	**0.63**	**62.5**	**6.0**	**0.91**	**40.0**	**5.0**	**0.99**	**50.5**
HLoc-L_{300}	-	-	12.9	-	-	7.0	-	-	27.3	-	-	44.5	-	-	9.7	-	-	28.4
HLoc-L_{1000}	8.7	1.20	33.3	-	-	25.4	5.5	1.02	48.3	4.3	0.64	56.6	-	-	21.9	5.6	1.10	47.4
HLoc-L_{3000}	5.3	0.73	48.1	-	-	31.3	3.4	0.65	61.9	3.6	0.54	69.5	-	-	31.1	3.7	0.71	59.1
HLoc	3.2	0.47	64.8	3.9	0.76	60.6	2.1	0.37	**81.0**	3.3	0.47	70.6	6.1	0.86	42.7	**2.1**	**0.42**	79.9
HLoc+SLD	**2.9**	**0.43**	**68.7**	**3.4**	**0.63**	**62.7**	**1.9**	**0.32**	**81.0**	**2.8**	**0.45**	**73.9**	**5.4**	**0.78**	**45.3**	**2.1**	**0.42**	**82.0**

Storage Usage (MB)



Ablation study:

Method	Patches	Res.	Aug.	L	INDOOR-6 (recall (5cm, 5°)) scene1	scene2	scene3	scene4	scene5	scene6	Average ↑
					# of visible points ≥ 8 ↑						
DSAC*	-	-	-	-	-	-	-	-	-	-	25.1
SLD	×	1/4	×	200	24.9	20.4	42.2	77.6	40.1	39.6	18.2
SLD	✓	1/4	×	200	77.2	38.0	53.0	94.1	72.2	66.3	28.2
SLD	✓	1/2	×	200	61.1	38.4	44.4	91.5	58.3	59.4	36.9
SLD	✓	1/2	✓	200	66.0	34.9	52.4	90.4	62.7	57.6	38.4
SLD	**✓**	**1/2**	**✓**	**300**	74.6	48.0	68.6	94.9	88.9	66.3	**42.7**
SLD	✓	1/2	✓	400	73.8	45.1	80.3	96.3	93.2	74.3	42.4

Main Insights:

- Our method **NBE+SLD** performs the best amongst learned localization methods.
- **Hloc** (with unlimited storage) outperforms all learned methods;
  - but its accuracy decreases as storage budget constraints are imposed.
- **Hloc+SLD** (the combination of both methods) works best; outperforms **Hloc**!

Results (7-scenes)

Method	7-SCENES chess (cm.)↓	(deg.)↓	(%)↑	fire (cm.)↓	(deg.)↓	(%)↑	heads (cm.)↓	(deg.)↓	(%)↑	office (cm.)↓	(deg.)↓	(%)↑	pumpkin (cm.)↓	(deg.)↓	(%)↑	redkitchen (cm.)↓	(deg.)↓	(%)↑	stairs (cm.)↓	(deg.)↓	(%)↑	recall (%)↑
MS-Transformer+	11	4.66	–	24	9.6	–	14	12.19	–	17	5.66	–	18	4.44	–	17	5.94	–	26	8.45	–	–
HLoc+	2.4	0.77	94.2	1.8	0.75	93.7	0.9	0.59	99.7	2.6	0.77	83.2	4.4	1.15	55.1	4.0	1.38	61.9	5.1	1.46	49.4	76.7
DSAC*+	1.8	0.59	97.8	1.7	0.77	94.5	1.0	0.66	98.8	2.7	0.79	83.9	3.9	1.05	62.0	3.9	1.24	65.5	3.5	0.93	78.0	82.9
NBE+SLD+	2.2	0.75	93.7	1.8	0.73	94.1	0.9	0.68	96.6	3.2	0.91	74.8	5.6	1.55	44.6	5.3	1.52	45.7	5.5	1.41	44.6	70.6
HLoc	0.8	**0.11**	**100**	0.9	**0.24**	99.4	0.6	**0.25**	**100**	**1.2**	**0.20**	**100**	**1.4**	**0.15**	**100**	1.1	**0.14**	**98.6**	2.9	0.80	72.0	95.7
DSAC*	**0.5**	0.17	99.9	0.8	0.28	98.9	**0.5**	0.34	99.8	1.2	0.34	98.1	1.2	0.28	99.0	**0.7**	0.21	97.0	2.7	0.78	**92**	**97.8**
NBE+SLD	0.6	0.18	**100**	**0.7**	0.26	**99.6**	0.6	0.35	98.4	1.3	0.33	95.8	1.5	0.33	94.4	0.8	0.19	96.6	**2.6**	**0.72**	85.2	95.7

Conclusion

Code and data available: https://github.com/microsoft/SceneLandmarkLocalization

We propose a new learned localization approach, where we designate scene-specific salient points as scene landmarks, leverage mature CNN architectures to detect them, and compute camera pose using a PnP solver from the 2D—3D scene landmark correspondences. Our method outperforms learned methods (DSAC*, etc.) but not yet as accurate as retrieval-based methods.