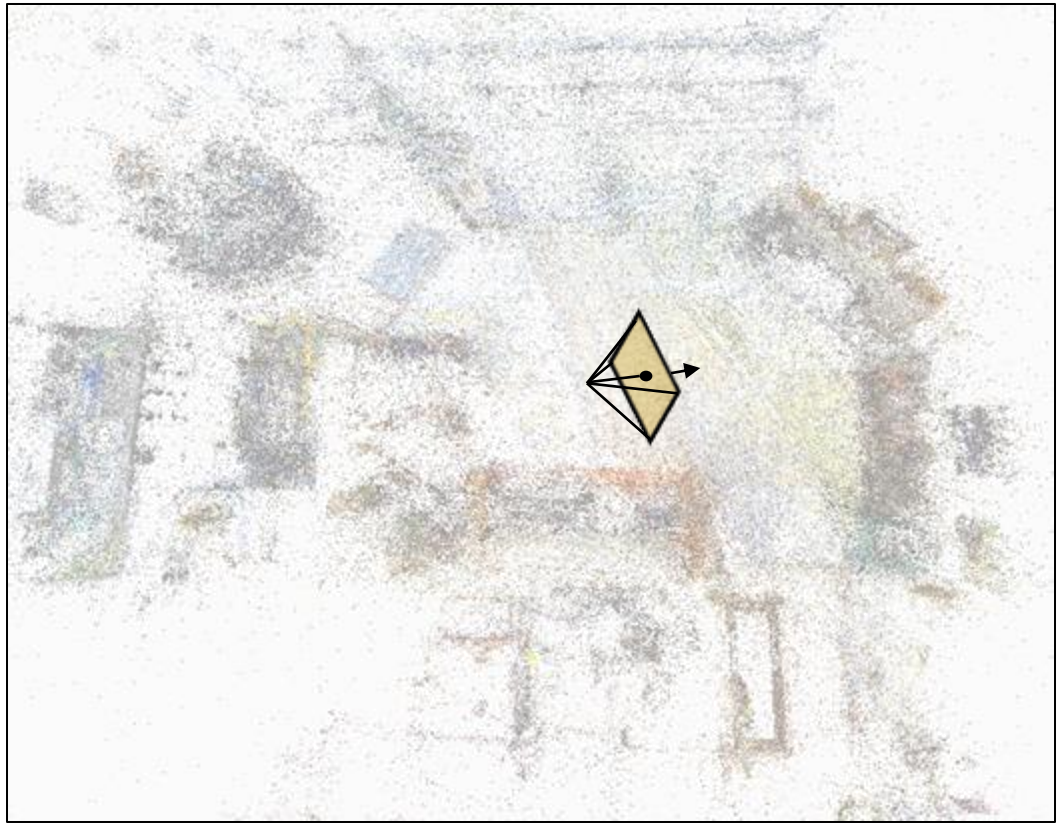
Learning to Detect Scene Landmarks for Camera Localization

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

[1]University of Minnesota [2]Microsoft

CVPR 2022

Camera localization problem



Given a query image, compute the 3D position and 3D orientation of the camera within a precomputed 3D map of the scene.

Related work

- Vast literature
- Retrieval-based methods
  - Hierarchical Localization (Hloc)

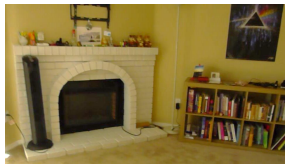
    Learning Feature Matching with Graph Neural Networks [Sarlin et al. 2020]
- Learned methods
  - Absolute pose regression (APR)

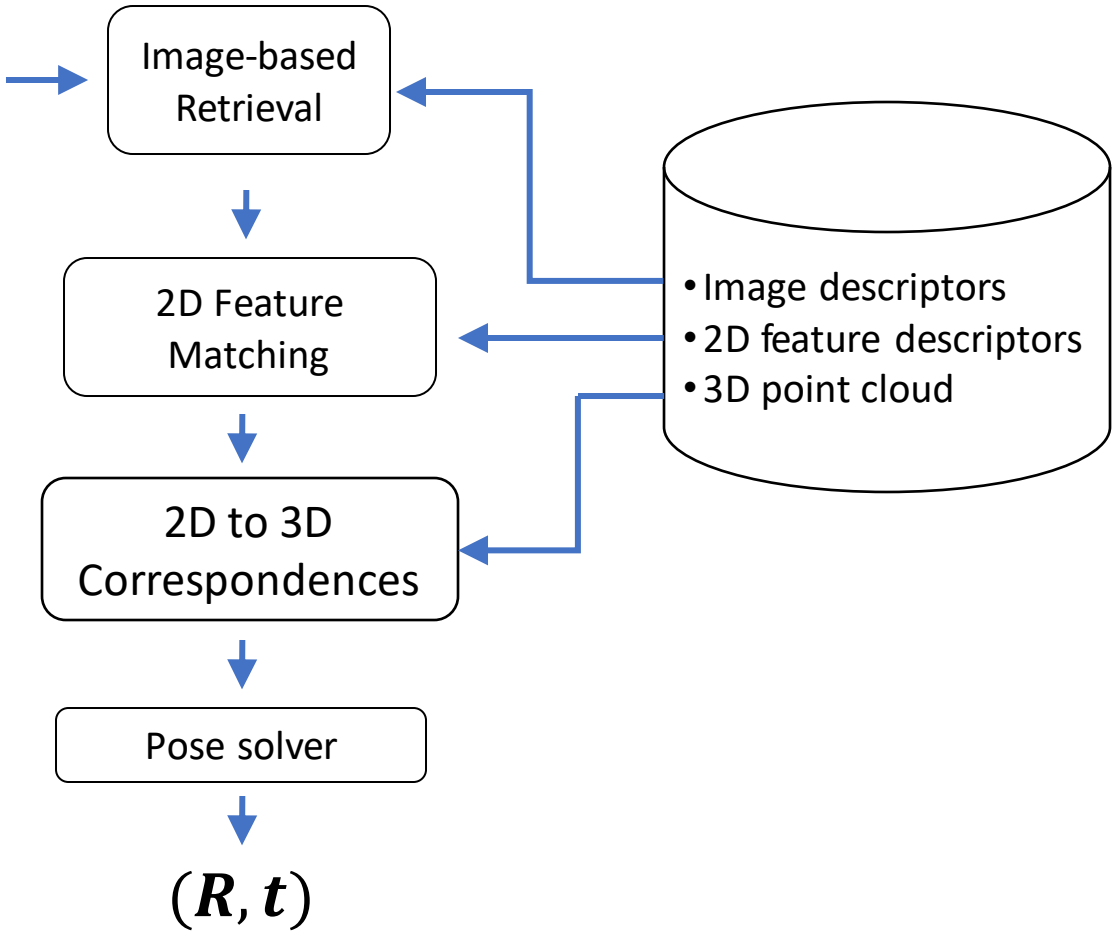
    PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization [Kendall et al. 2015]
  - Dense scene coordinate regression (SCR)

    Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images [Shotton et al. 2013]

    Visual Camera Re-Localization from RGB and RGB-D Images Using DSAC [Brachmann and Rother 2021]

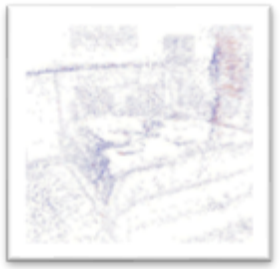
Retrieval-based methods





- Accurate
- High storage requirements
- Not privacy preserving
  - Image can be reconstructed from stored feature descriptors

Revealing scenes by inverting structure from motion reconstructions. [Pittaluga et al. 2019]



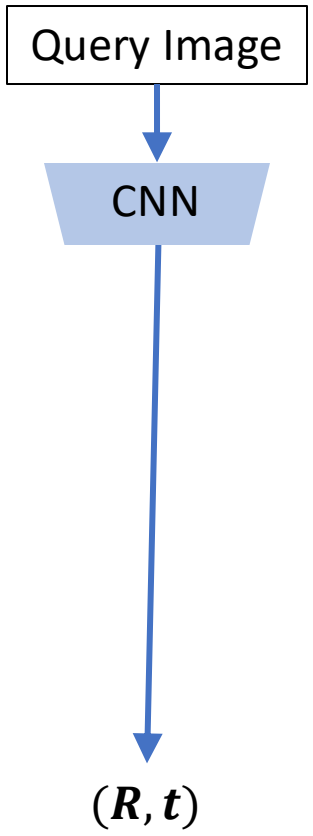
stored points + features

reconstructed image

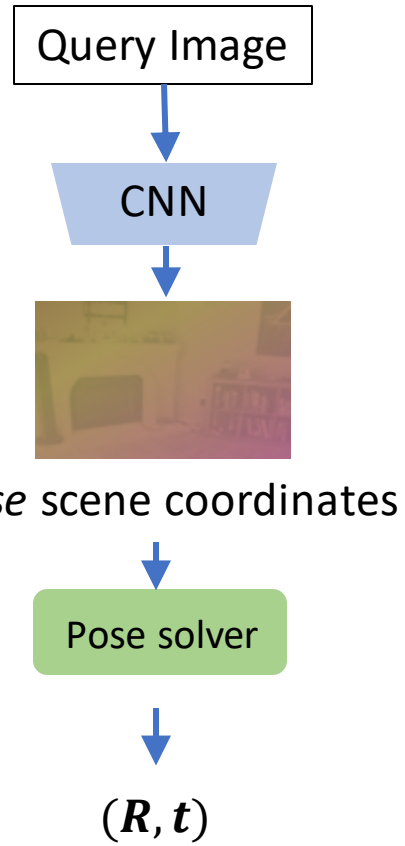
original image

Learned methods (low storage)

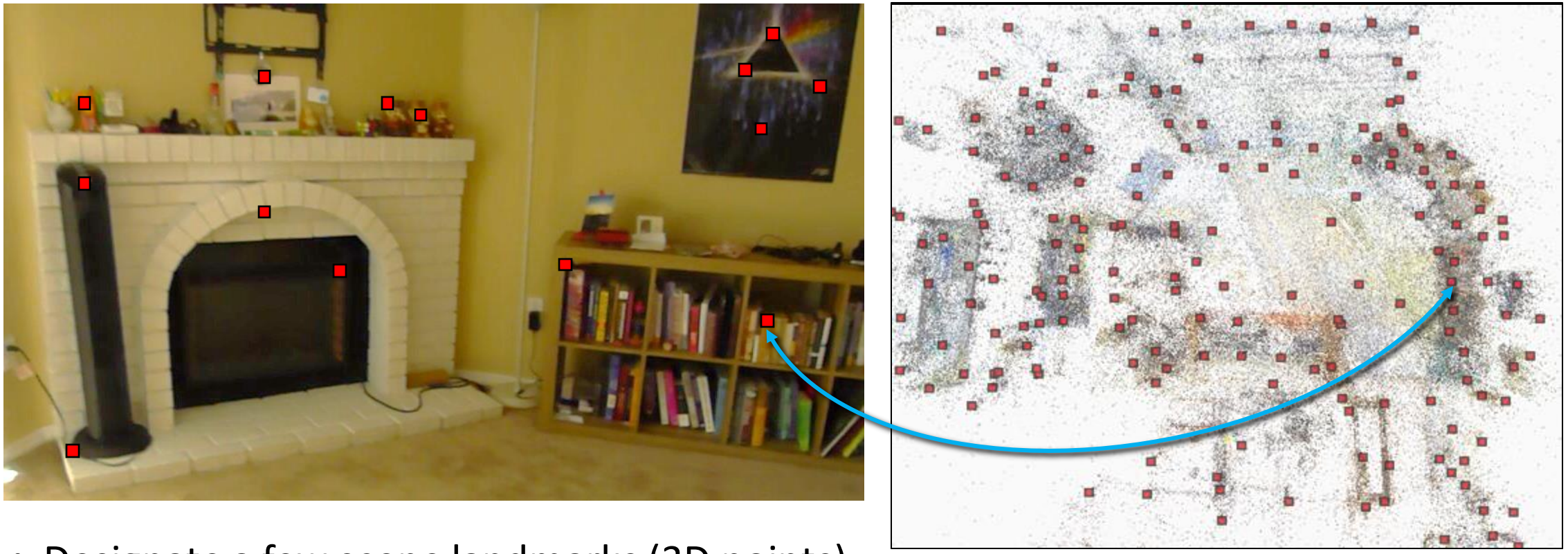
Abs. Pose Regression
(PoseNet)



Scene Coordinate Regression
(DSAC*)



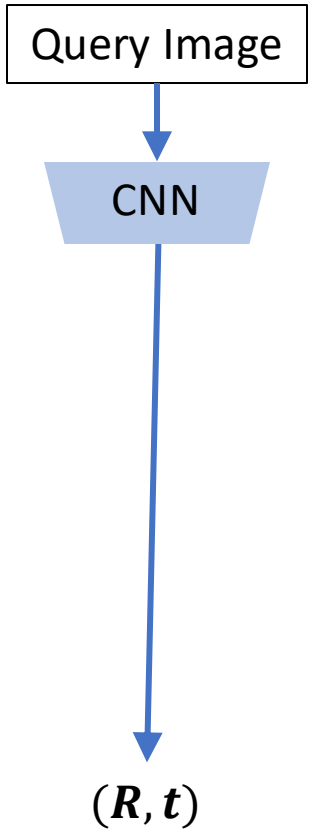
Main Idea



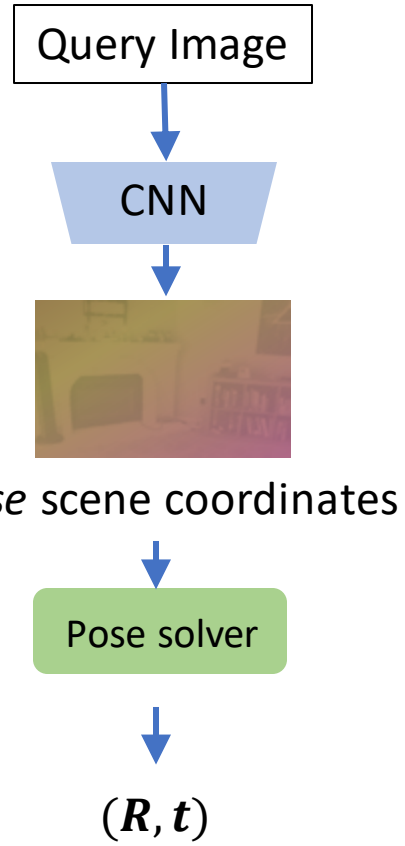
- Designate a few scene landmarks (3D points).
- Learn a detector to localize those scene landmarks in a query image.
- Estimate camera pose from the 2D-3D scene landmark correspondences.

Learned methods (low storage)

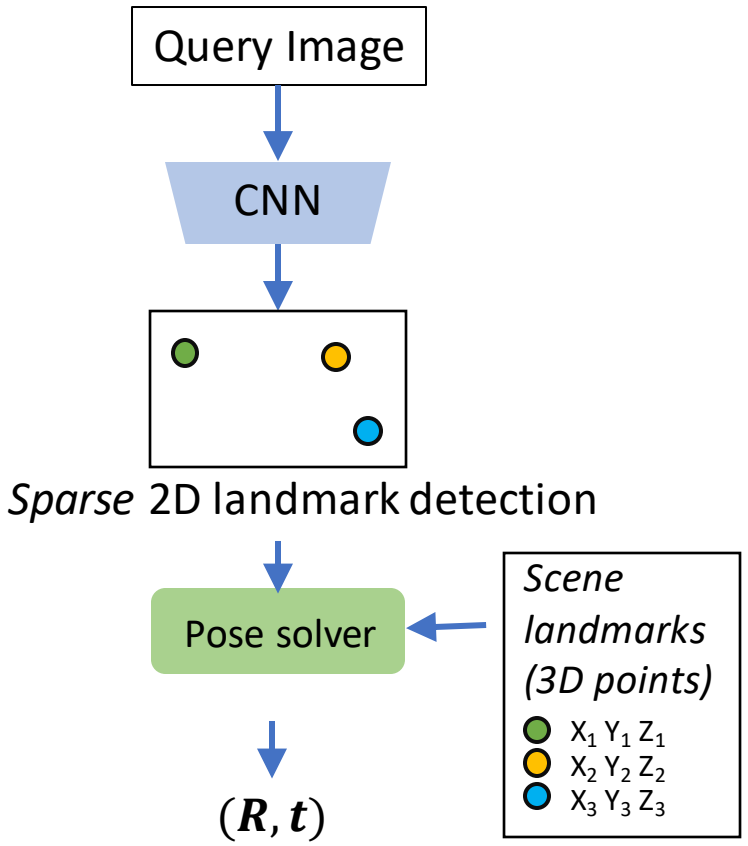
Abs. Pose Regression (PoseNet)



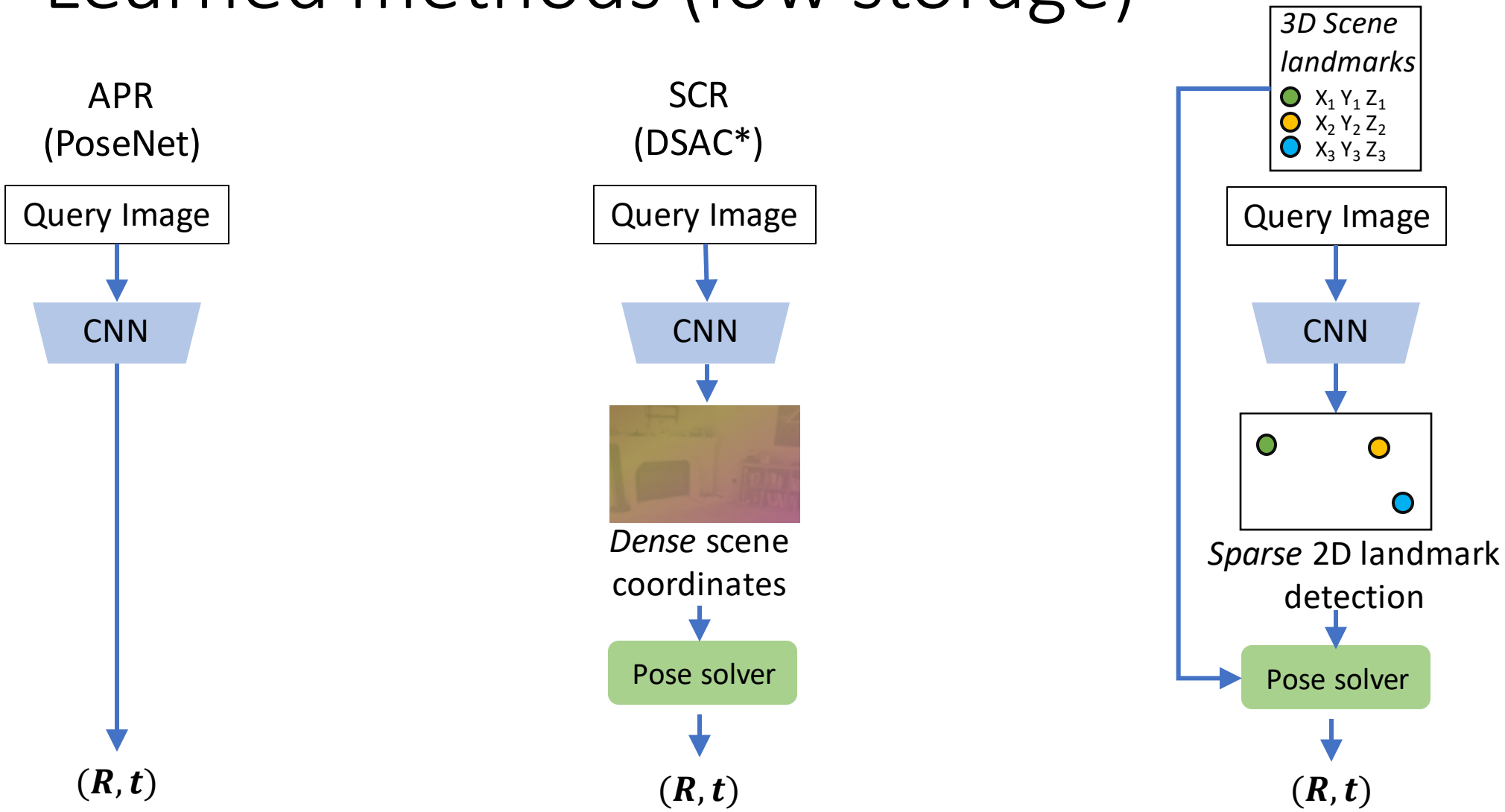
Scene Coordinate Regression (DSAC*, …)



Scene Landmark Detection (ours)

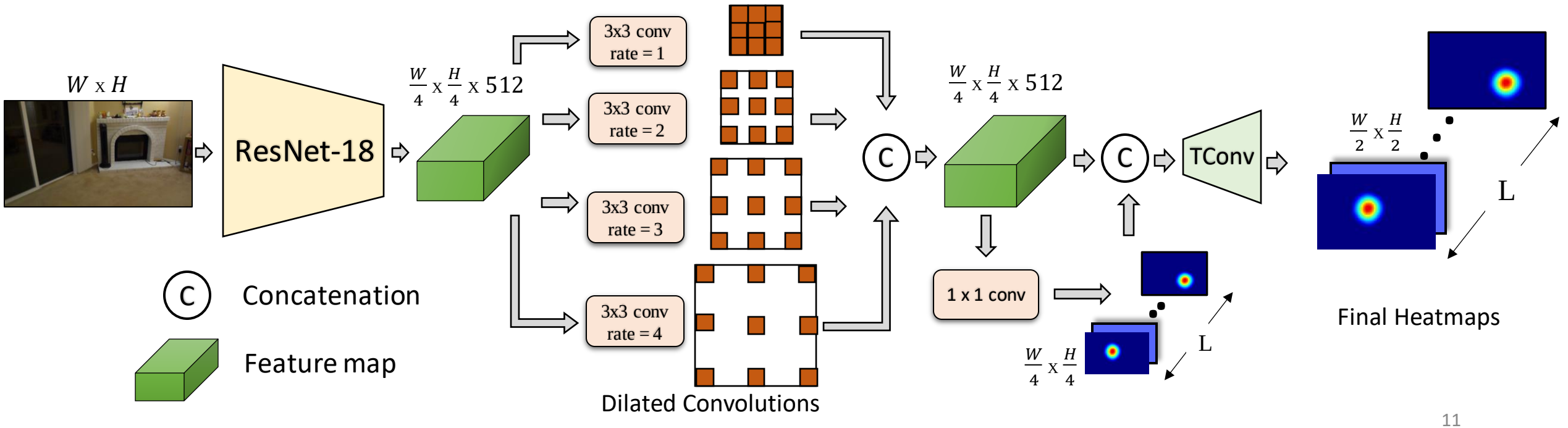


Learned methods (low storage)

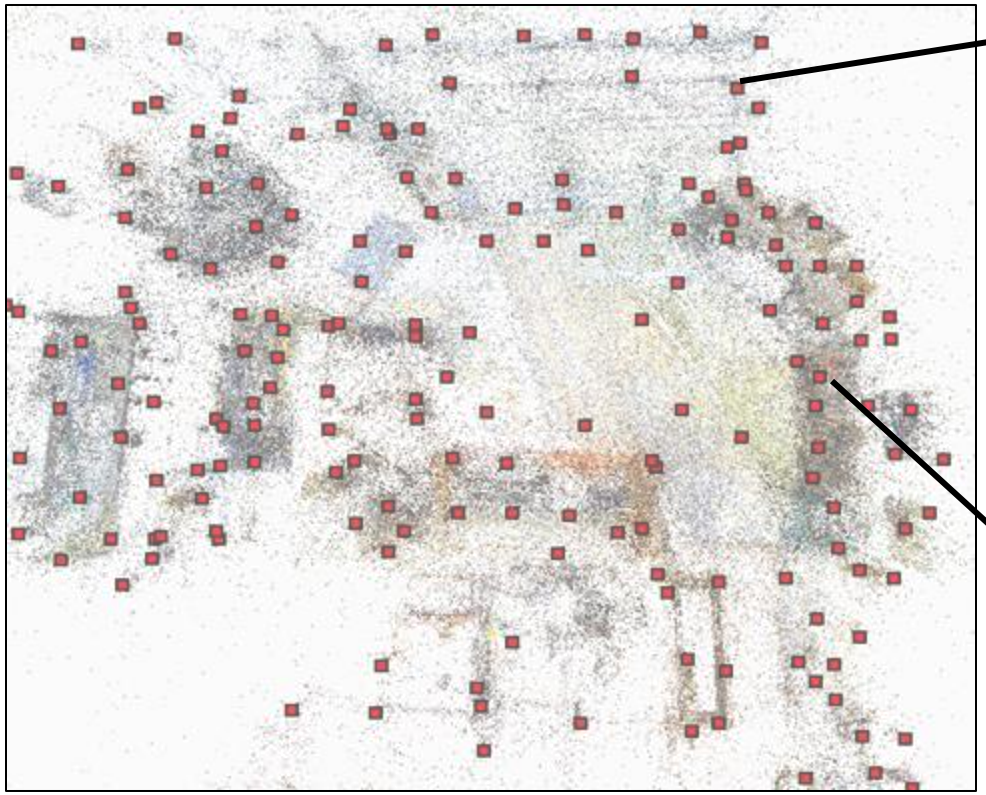


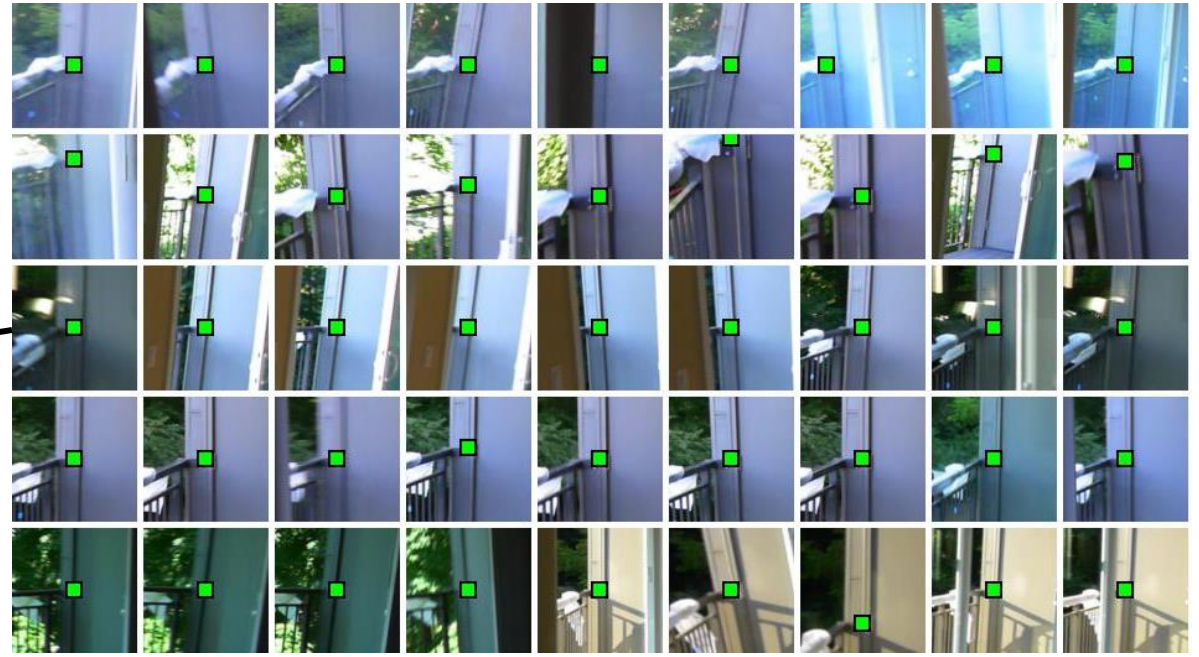
Scene Landmark Detector (SLD) Model

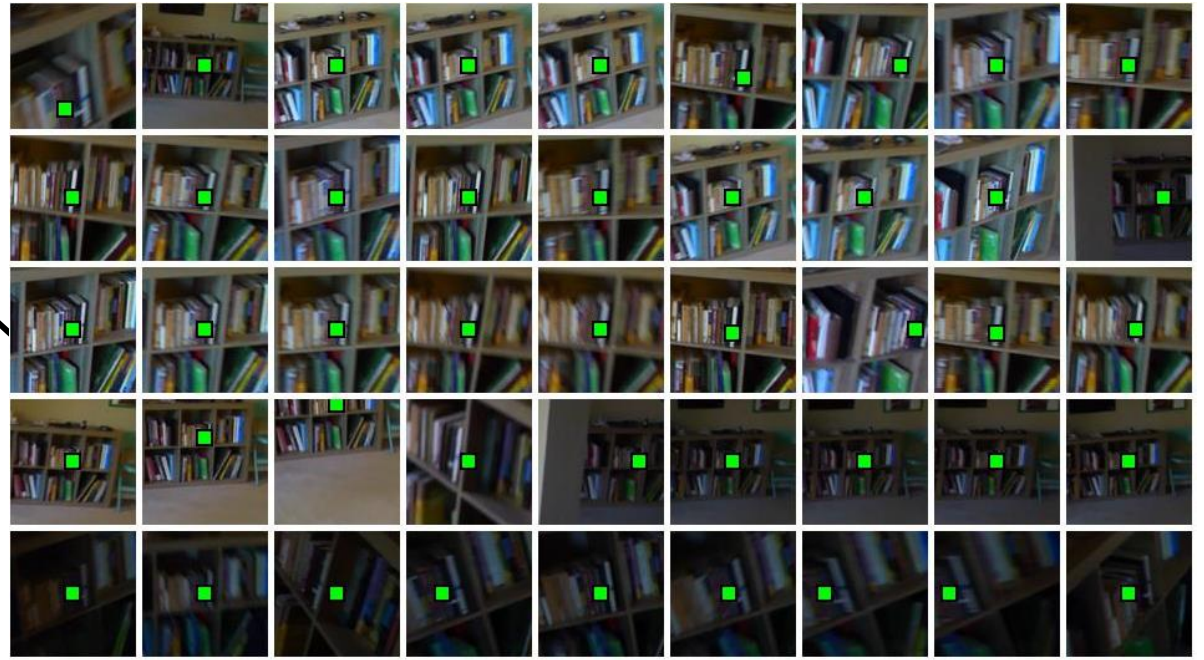
- Output heatmap for each landmark
- Dilated convolution architecture
- Mean Sq. Error (MSE) pixel-wise loss
- Homography and intensity data augmentation



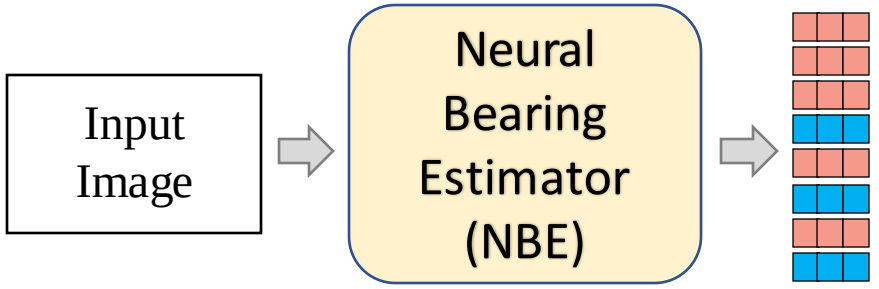
Example: training data

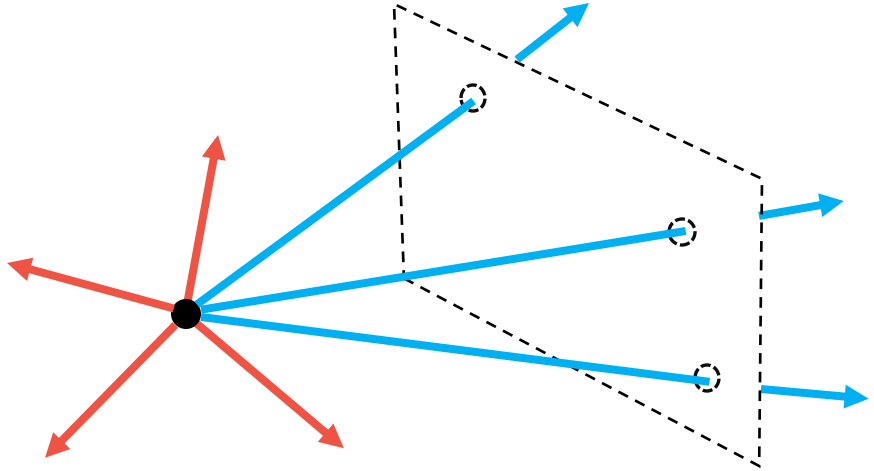






Neural Bearing Estimator (NBE)



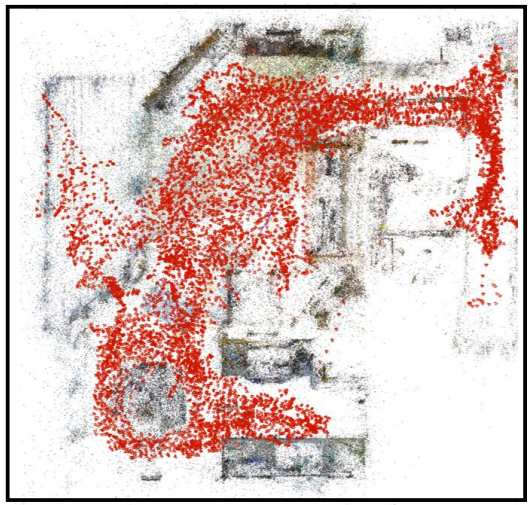


- From image appearance, directly predict landmark bearing vector (3D)
- Can do it for visible as well as invisible landmarks

Indoor-6 Dataset

- Multiple captures (different day and time) of the same scene
- SfM reconstructions

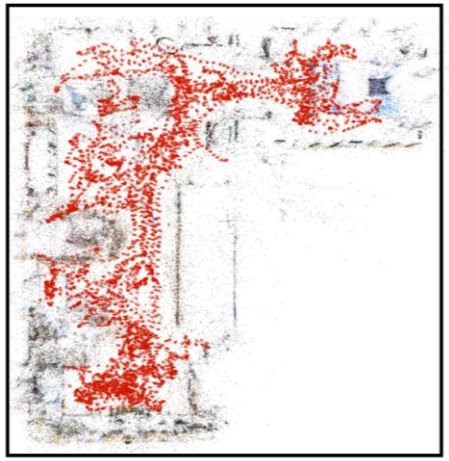
scene1 (24 – 6289 – 799)



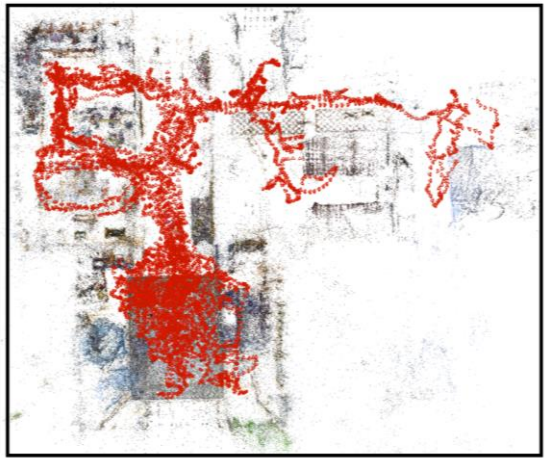
Test Images

Train Images

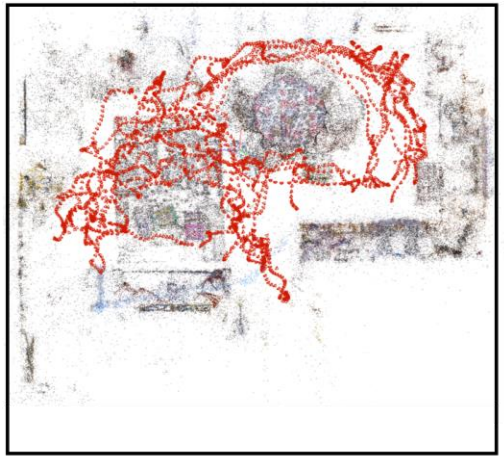
scene2 (12 – 3021 – 284)



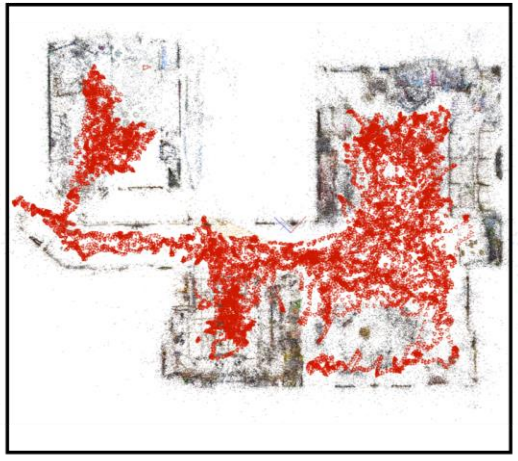
scene3 (18 – 4181 – 315)



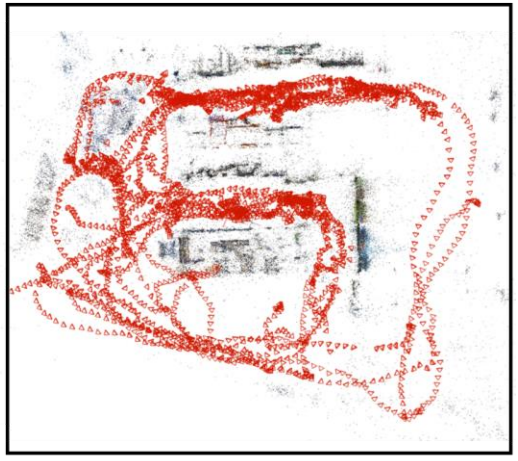
scene4 (4 – 1942 – 272)



scene5 (15 – 4946 – 424)



scene6 (6 – 1761 – 323)



Results

- NBE+SLD (ours) achieves the best performance among learned (low storage) methods
- NBE+SLD(E) outperforms SOTA DSAC* using similar network capacity
- NBE+SLD (ours) outperforms $Hloc_{1000}$ that uses 3x more landmarks

Recall (%) @ (5cm ,5°)

	Storage (MB)	scene1	scene2	scene3	scene4	scene5	scene6
PoseNet	12	0.0	0.0	0.0	0.0	0.0	0.0
DSAC*	27	18.7	12.3	19.7	44.9	10.6	44.3
$NBE+SLD(E)_{300}$	29	28.4	26.1	43.5	48.9	37.5	44.6
$NBE+SLD_{300}$	132	**38.4**	**37.0**	**53.0**	**62.5**	**40.0**	**50.5**

Conclusion

- New learned camera localization method that predicts pre-determined scene landmarks in images.
- Leverages mature heatmap-based keypoint detection architectures.
- Low storage, privacy preserving, and high accuracy
- Code & Dataset: github.com/microsoft/SceneLandmarkLocalization

UNIVERSITY OF MINNESOTA

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]

[1]University of Minnesota [2]Microsoft

CVPR 2022