Out with the Old?

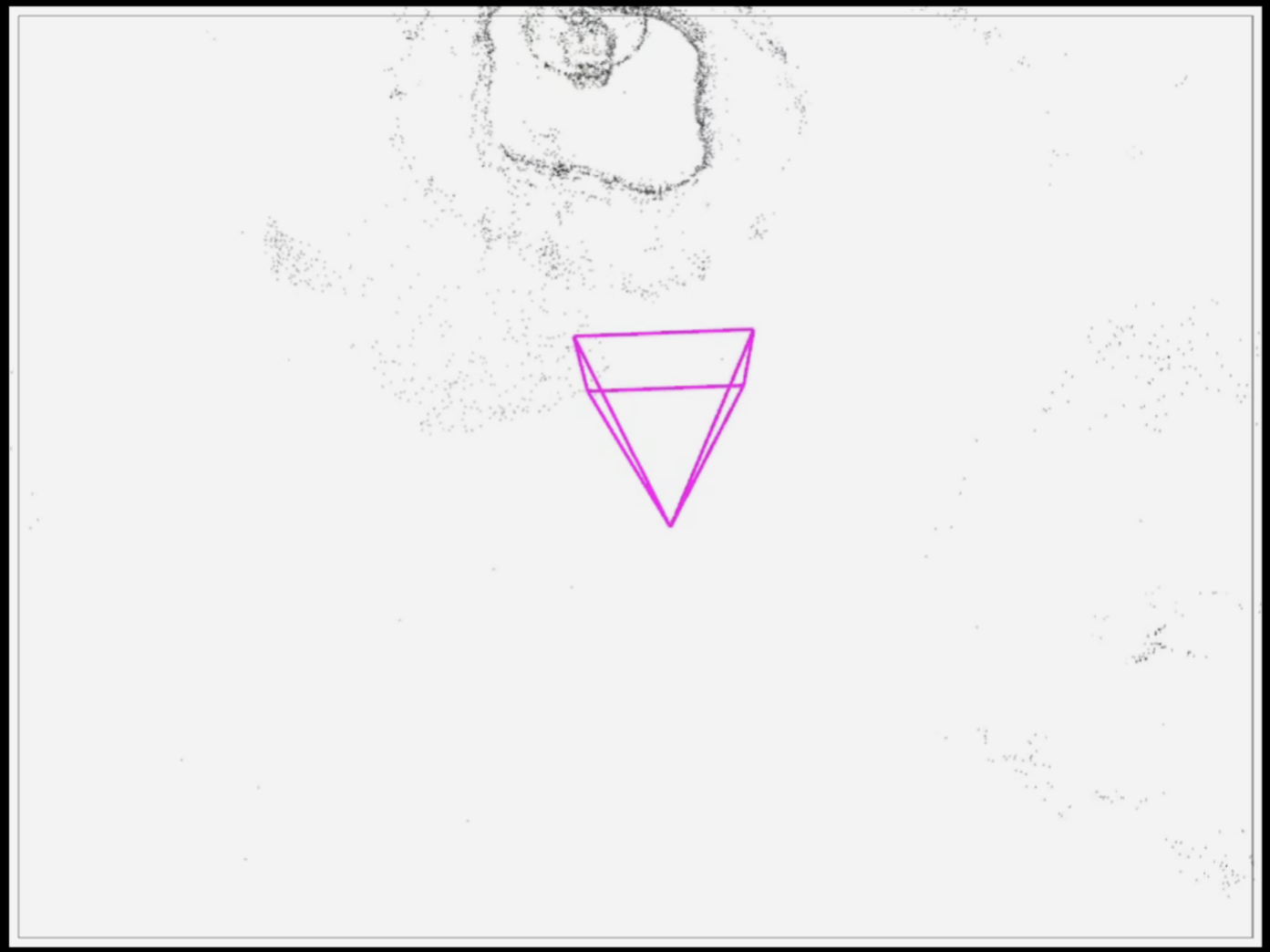
CNN vs. SIFT-based visual localization

Prof. Dr. Laura Leal-Taixé

Technische Universität München

Visual Localization

Compute exact position and orientation of query image (relative to 3D scene model)



Classic Localization Pipeline

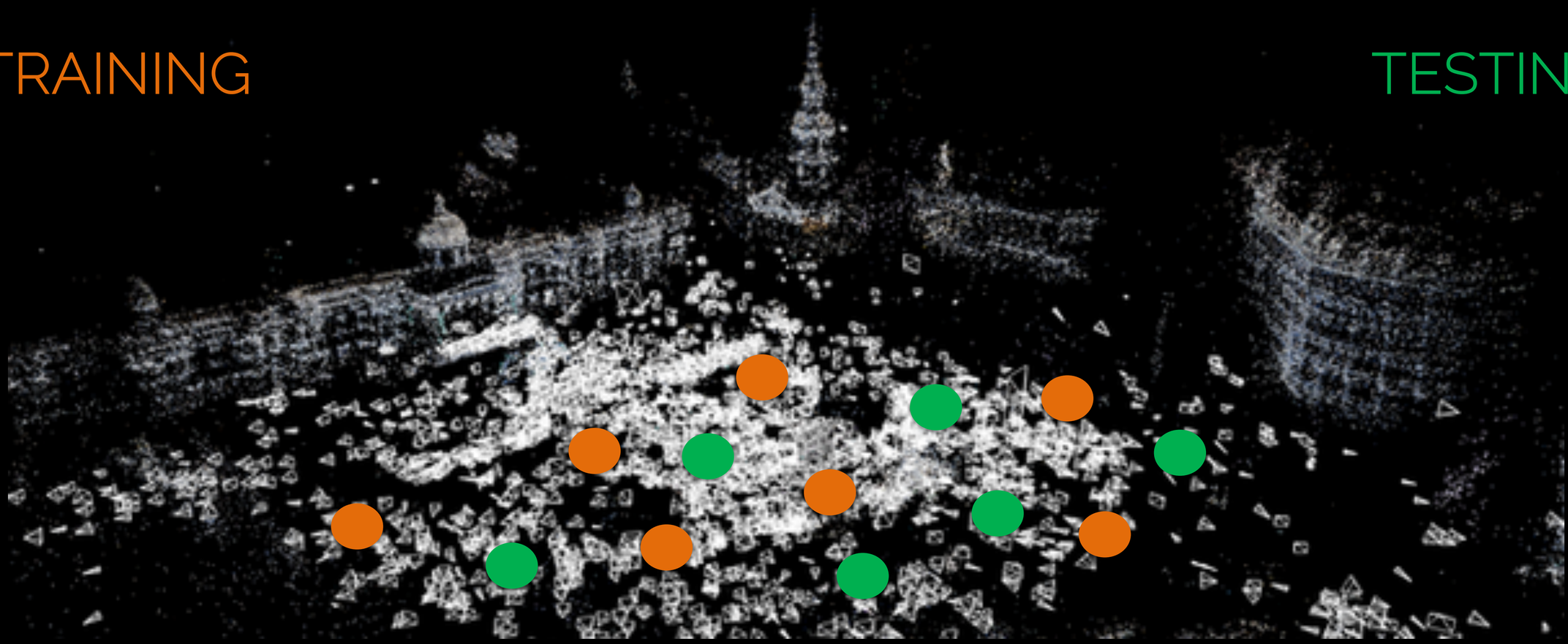
Extract Lo

Establish 2D-

Estimate Camera Pose
(RANSAC)

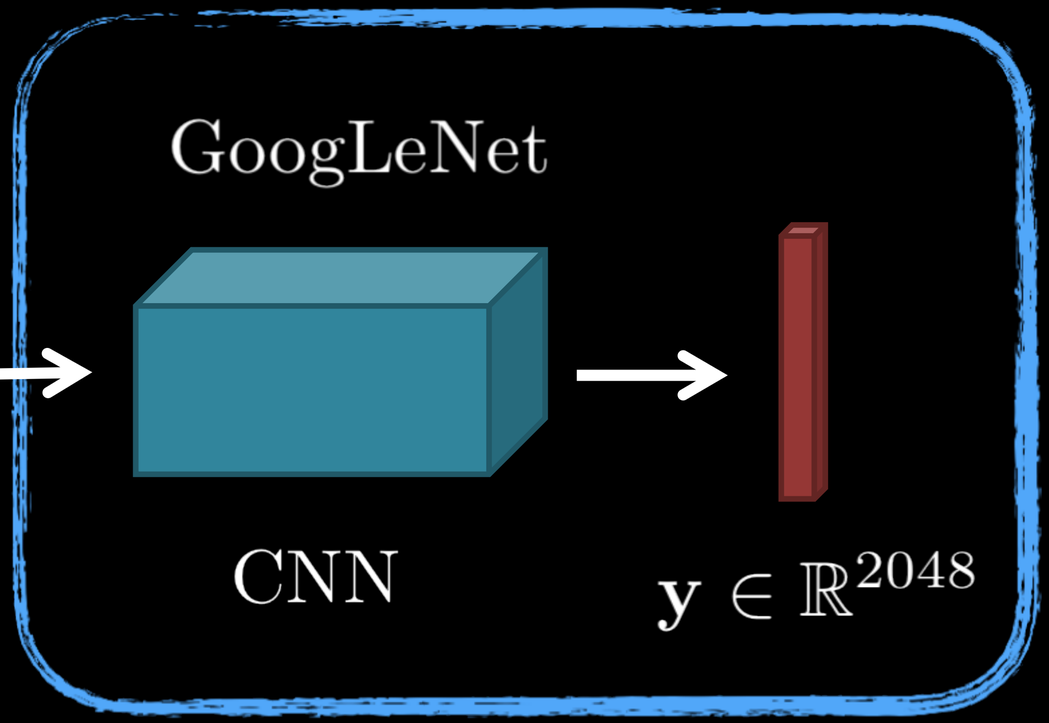
Learning visual localization?

Where do we get training data?

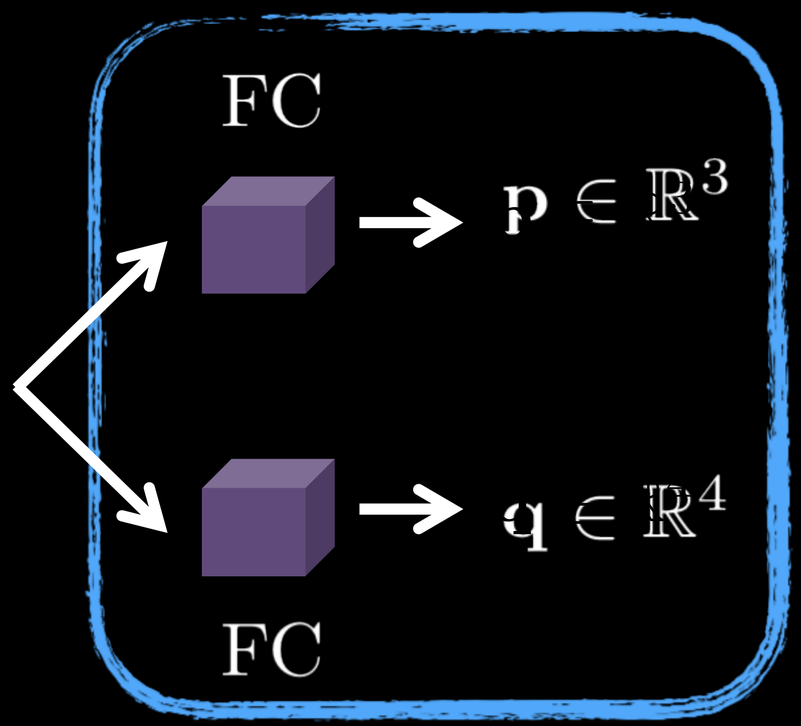


Structure-from-Motion gives us plenty of data

Related work: PoseNet



Non-Linear Embedding



Linear Regression

Kendall et al. PoseNet:. ICCV 2015

Outdoor localization

- Experiments on Cambridge Landmarks

Scene	Area or Volume	CNN
		PoseNet [22]
King's College	5600 m^2	1.92 m, 5.40°
Old Hospital	2000 m^2	2.31 m, 5.38°
Shop Façade	875 m^2	1.46 m, 8.08°
St Mary's Church	4800 m^2	2.65 m, 8.48°
Average All	–	2.08 m, 6.83°
Average by [41]	–	–

Walch, Hazirbas, Leal-Taixé, Sattler, Hilsenbeck, Cremers, Image-based localization using LSTMs for structured feature correlation. ICCV, 2017

Indoor localization: SIFT suffers

- Experiments on 7Scenes

		SIFT	CNN
Scene	Area or Volume	Active Search (w/o) [41]	PoseNet [22]
Chess	6 m^3	0.04 m, 1.96° (0)	0.32 m, 8.12°
Fire	2.5 m^3	0.03 m, 1.93° (1)	0.47 m, 14.4°
Heads	1 m^3	0.03 m, 2.59° (7)	0.29 m, 12.0°
Office	7.5 m^3	0.09 m, 3.61° (34)	0.48 m, 7.68°
Pumpkin	5 m^3	0.08 m, 3.10° (71)	0.47 m, 8.42°
Red Kitchen	18 m^3	0.07 m, 3.37° (0)	0.59 m, 8.64°
Stairs	7.5 m^3	not available	0.47 m, 13.8°
Average All	–	–	0.44 m, 10.4°
Average by [41]	–	0.06 m, 2.76°	–

Number of images that it failed to localize

Our new dataset TUM-LSI: SIFT dies

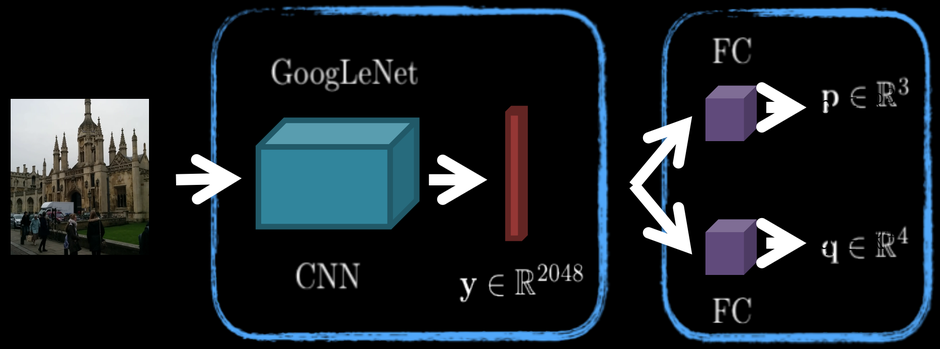
Scene	Area	# train/test	PoseNet [22]
Matriculation Hall	793 m^2	940/235	1.15 m, 3.93°
Hallway	2677 m^2	845/215	2.16 m, 6.99°
Floor	5575 m^2	875/220	1.87 m, 6.14°
Average			1.72 m, 5.68°

SIFT-based methods do not work at all!

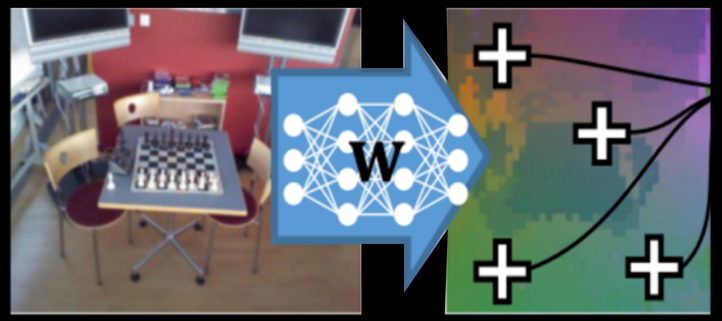
https://github.com/NavVisResearch/NavVis-Indoor-Dataset

Limitations of current methods

- One separate model is needed per scene



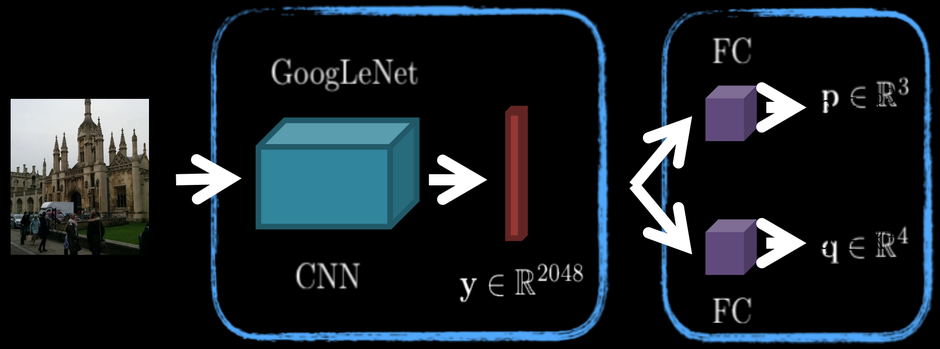
Kendall et al. PoseNet:. ICCV 2015

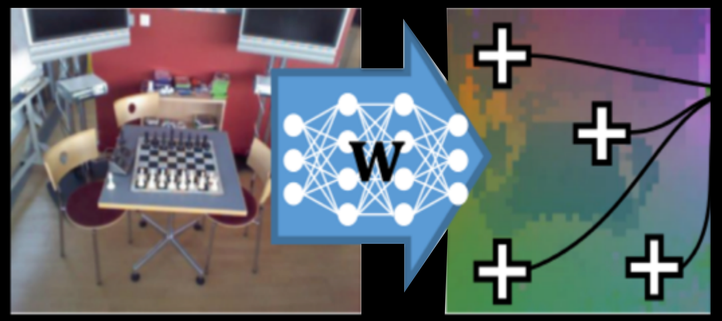


Brachmann et al. DSAC:. CVPR 2017

Limitations of current methods

- One separate model is needed per scene





- The most used loss function requires the setting of a hyperparameter which is again scene dependent

DOES NOT SCALE

PoseNet loss functions

- L2 loss

$$\mathcal{L}(y^*, y) = \| c - c^* \|_2 + \beta \| q - q^* \|_2$$

pose label

pose prediction

camera center position

scaling factor to balance the weighting between the rotation error and the position error

camera rotation in quaternion

✗ No geometry information

✗ Need to find its value per scene → can range from 200 to 2500

Kendall et al. PoseNet:. ICCV 2015

PoseNet loss functions

- Geometric loss function: Minimize re-projection error of 3D points visible in image

✓ More accurate results

✗ Network needs to be pre-trained on L2 loss

✗ One network per scene

Kendall et al. Geometric Loss Functions for Camera Pose Regression with Deep Learning. CVPR 2017

Relative Pose Estimation

- Use a neural network to predict relative poses

- Loss function: PoseNet L2 loss

$$\mathcal{L}(y^*, y) = \| c - c^* \|_2 + \boxed{\beta} \| q - q^* \|_2$$

✓ One network for all scenes

✗ Loss still depends on a scene-dependent hyperparameter

Laskar et al. Camera relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Network. ICCVW 2017

Limitations of current methods

- One separate model is needed per scene

- The most used loss function requires the setting of a hyperparameter which is again scene dependent

Proposed method

- One separate model is needed per scene

Relative Pose Estimation with Geometric Matching

✓ One network for all scenes

- The most used loss function requires the setting of a hyperparameter which is again scene dependent

Essential Matrix Regression

✓ No hyperparameter on the loss

Q. Zhou, T. Sattler, L. Leal-Taixé. Submitted to ECCV 2018

Proposed method

- One separate model is needed per scene

Relative Pose Estimation with Geometric Matching

✓ One network for all scenes

- The most used loss function requires the setting of a hyperparameter which is again scene dependent

Essential Matrix Regression

✓ No hyperparameter on the loss

Q. Zhou, T. Sattler, L. Leal-Taixé. Submitted to ECCV 2018

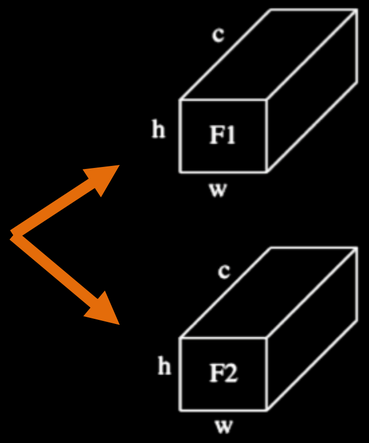
Regressing relative poses

- Siamese architecture
  - Shared weights on the input image pair
  - Feature concatenation

Geometric matching layer

- Mimic the geometrical matching process

Features from the siamese net



Fixed operation, no learnable parameters

I. Rocco et al. "Convolutional neural network architecture for geometric matching. CVPR 2017.

Proposed method

- One separate model is needed per scene

Relative Pose Estimation
with
Geometric Matching

✓ One network for all scenes

- The most used loss function requires the setting of a hyperparameter which is again scene dependent

Essential Matrix
Regression

✓ No hyperparameter on the loss

Q. Zhou, T. Sattler, L. Leal-Taixé. Submitted to ECCV 2018

Essential Matrix loss

- Essential matrix between query and training image $\mathbf{E}^* = [\mathbf{t}^*]_\times \mathbf{R}^*$

- L2 loss function on the entries of the essential matrix

$$\mathcal{L}_{ess}(\mathbf{E}^*, \mathbf{E}) = \| \mathbf{e} - \mathbf{e}^* \|_2$$

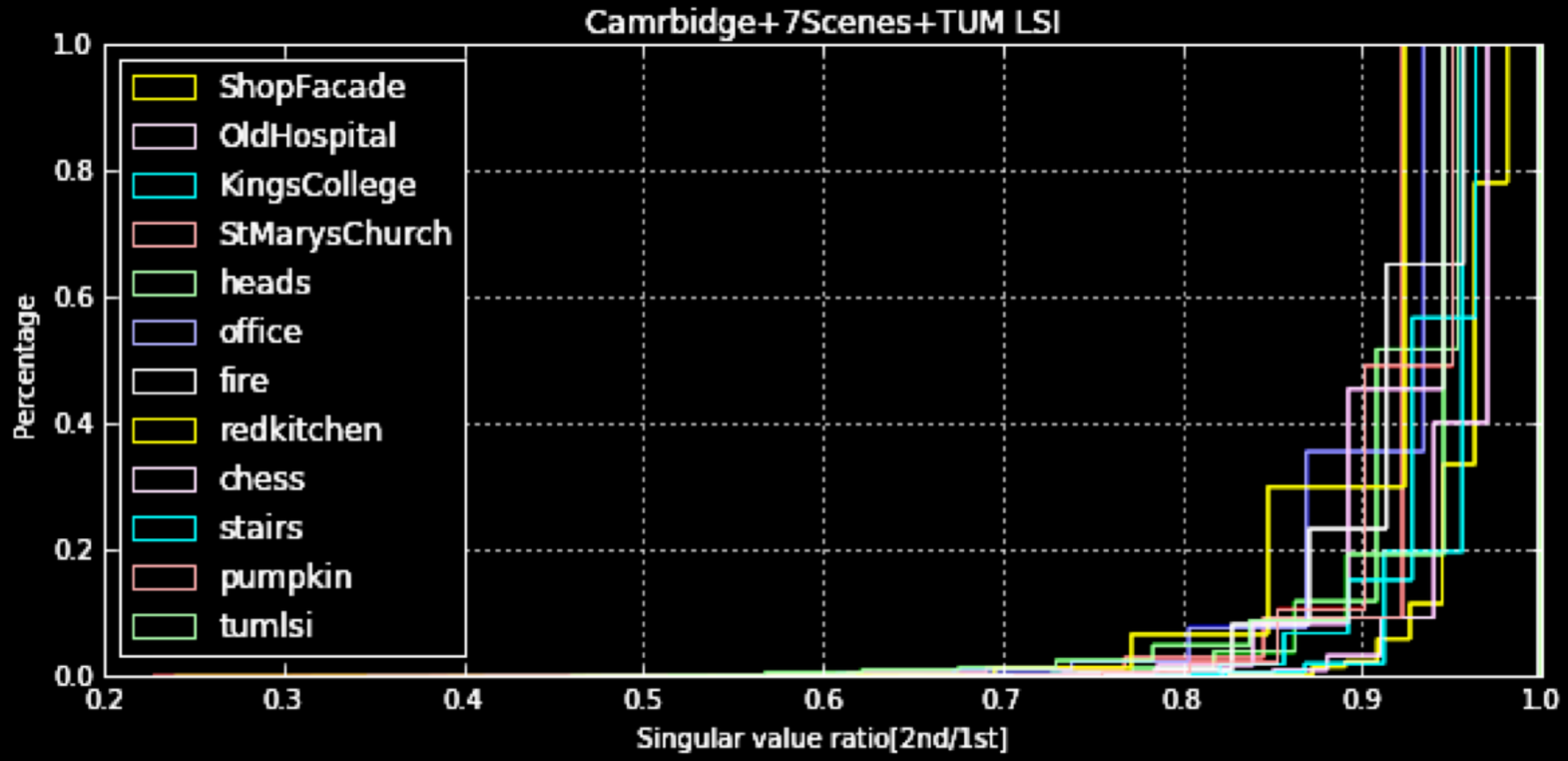
$$\| t \| = 1$$

Fixing the scale of translations to get consistent predictions

The loss is hyperparameter free!

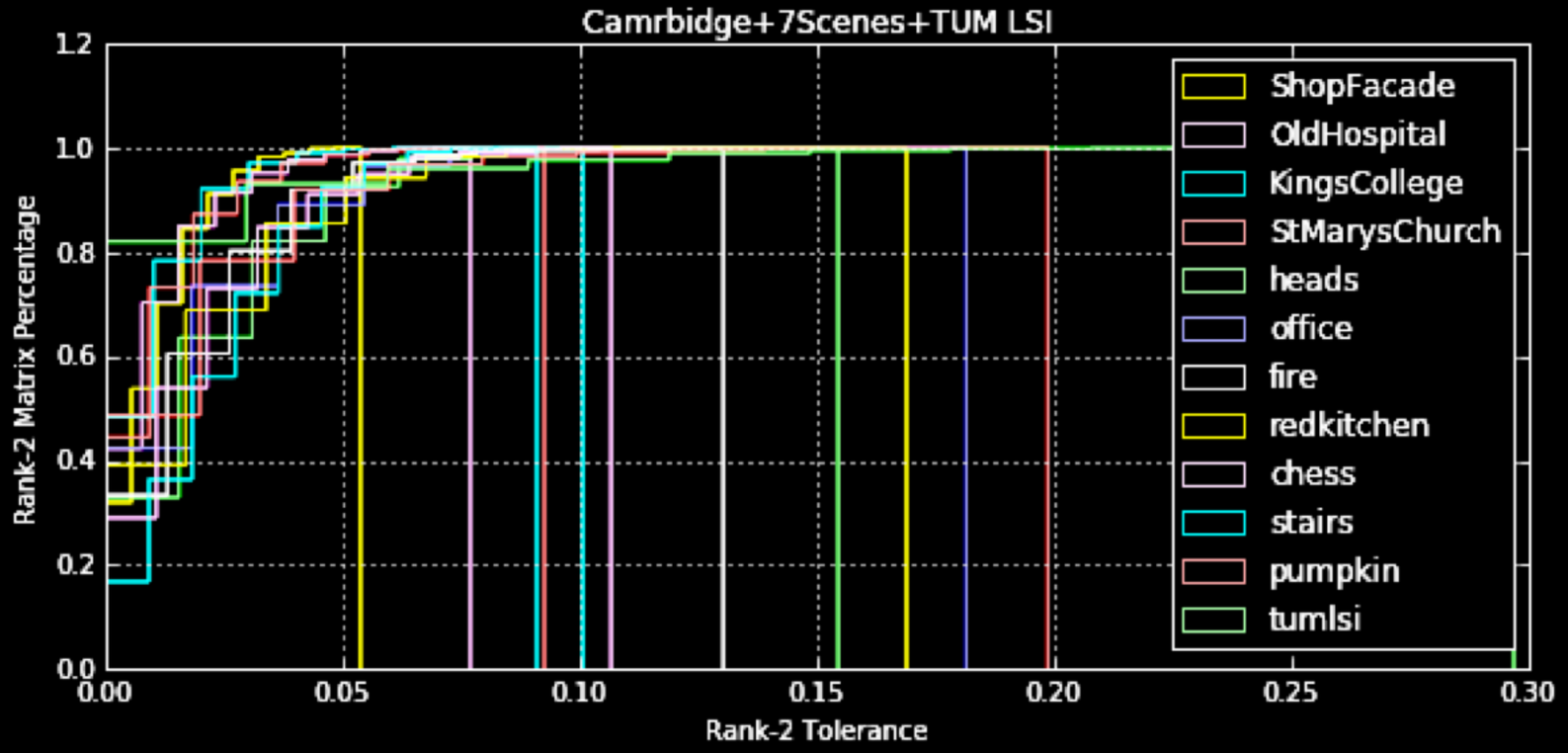
Are we really learning an essential matrix?

- Ratio between first two eigenvalues is close to 1? ✔



Are we really learning an essential matrix?

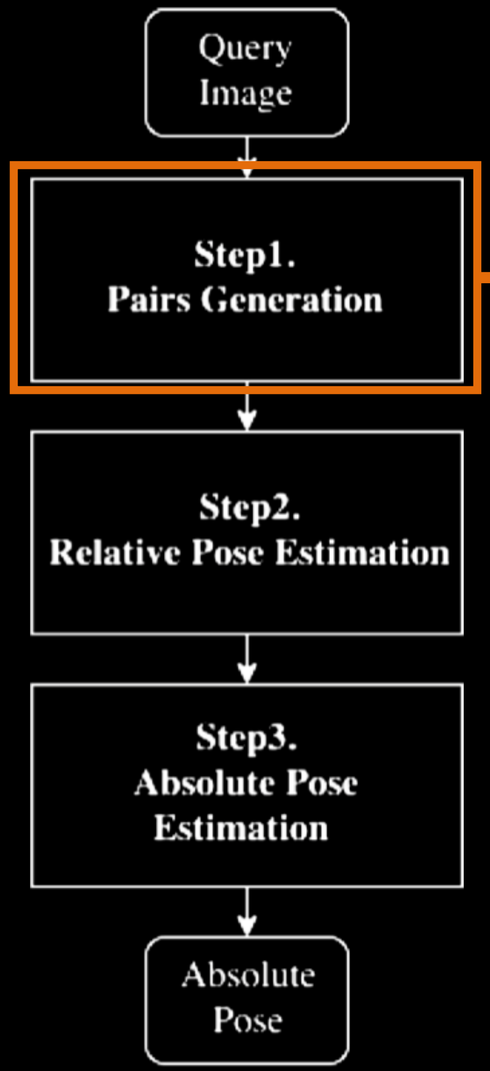
- Ratio between first two eigenvalues is close to 1? ✓
- Third eigenvalue close to 0? ✓



Are we really learning an essential matrix?

- Ratio between first two eigenvalues is close to 1? ✔
- Third eigenvalue close to 0? ✔
- We learn a good approximation of a true essential matrix
- Feature-based methods also project the resulting matrix to a rank 2 matrix (due to noise)
- Forcing this projection **does not have any effect** on the results on **any of the tested datasets**

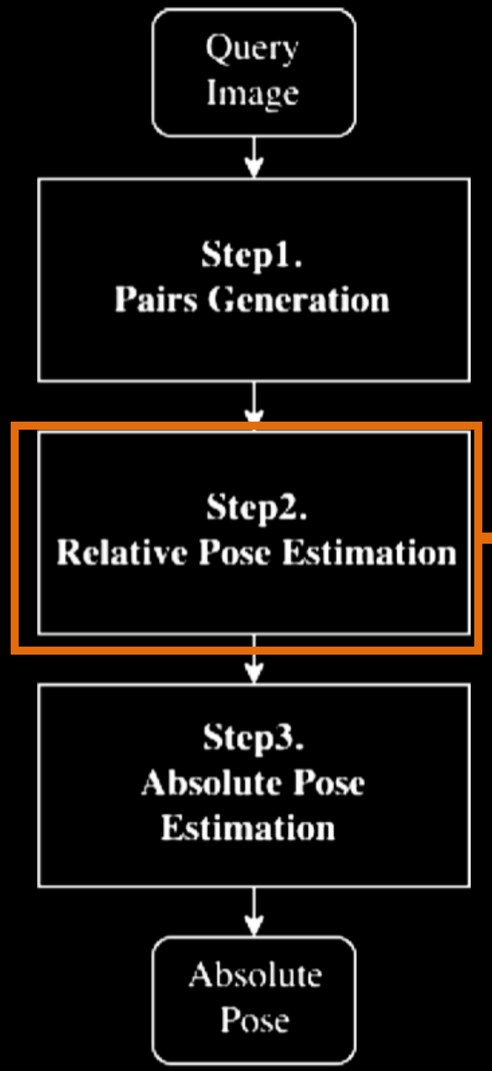
Full localization pipeline



A CNN model for Image retrieval to identify 30 nearest neighbors for the query image

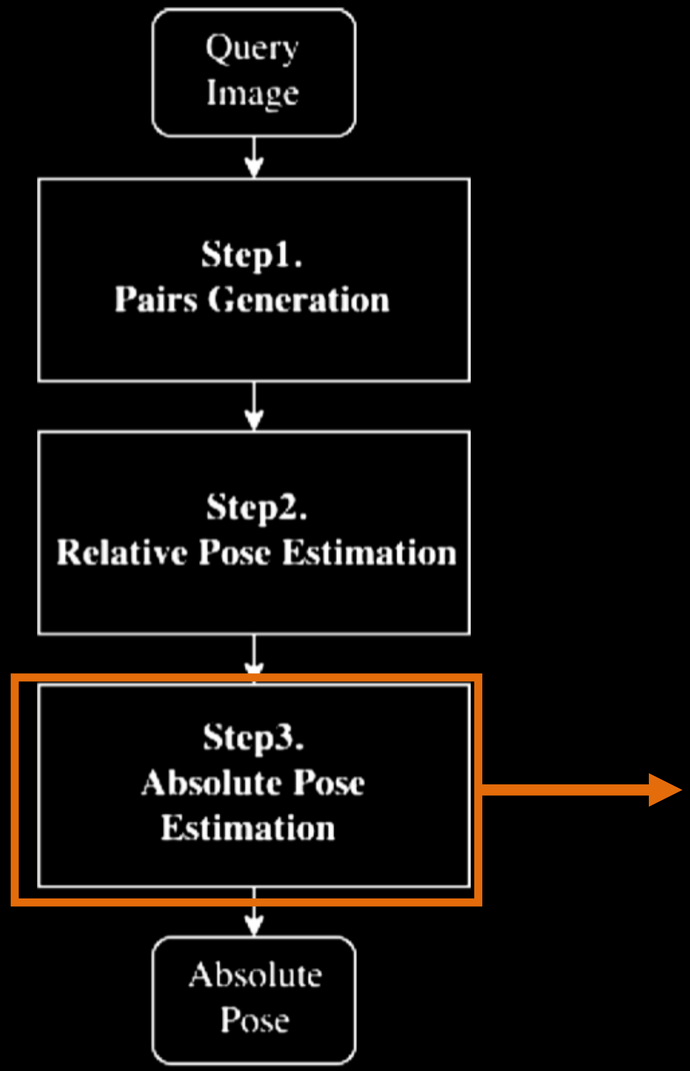
F. Radenovic´et al.. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples.. ECCV 2016

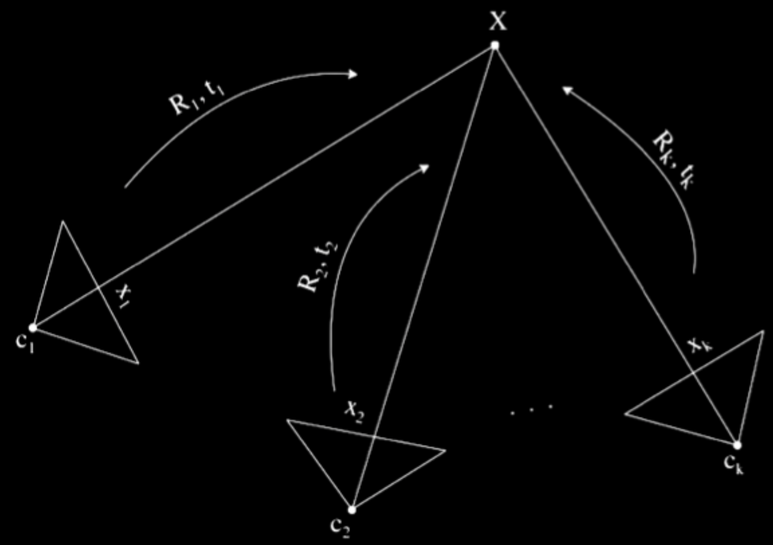
Full localization pipeline



EssNet to regress essential matrices and estimate relative poses

Full localization pipeline





Triangulation of the position and averaging of the rotation within a RANSAC loop.

Comparison to SOA

- Experiments on outdoor scenes, Cambridge Landmarks

Comparison to SOA

- Experiments on outdoor scenes, Cambridge Landmarks

Dataset	PoseNet [25]	Bayesian PoseNet [26]	LSTM PoseNet [28]	Geometric PoseNet [27]	VGGRegNet [32]	EssNet
St Mary's Church	$2.65m, 8.48°$	$2.11m, 8.38°$	$1.52m, 6.68°$	$1.57m, 3.32°$	$2.11m, 8.11°$	$1.39m, 5.61°$
ShopFacade	$1.46m, 8.08°$	$1.25m, 7.54°$	$1.18m, 7.44°$	$0.88m, 3.78°$	$0.63m, 5.73°$	$0.87m, 4.58°$
OldHospital	$2.31m, 5.38°$	$2.57m, 5.14°$	$1.51m, 4.29°$	$3.20m, 3.29°$	$1.50m, 4.03°$	$2.14m, 5.84°$
King's College	$1.92m, 5.40°$	$1.74m, 4.06°$	$0.99m, 3.65°$	$0.88m, 1.04°$	$1.06m, 2.81°$	$0.97m, 2.80°$
Average	$2.09m, 6.84°$	$1.92m, 6.28°$	$1.30m, 5.52°$	$1.63m, 2.86°$	$1.33m, 5.17°$	$1.34m, 4.71°$

- SOA positional error with better rotational error

Comparison to SOA

- Experiments on outdoor scenes, Cambridge Landmarks

Dataset	PoseNet [25]	Bayesian PoseNet [26]	LSTM PoseNet [28]	Geometric PoseNet [27]	VGGRegNet [32]	EssNet
St Mary's Church	$2.65m, 8.48°$	$2.11m, 8.38°$	$1.52m, 6.68°$	$1.57m, 3.32°$	$2.11m, 8.11°$	$1.39m, 5.61°$
ShopFacade	$1.46m, 8.08°$	$1.25m, 7.54°$	$1.18m, 7.44°$	$0.88m, 3.78°$	$0.63m, 5.73°$	$0.87m, 4.58°$
OldHospital	$2.31m, 5.38°$	$2.57m, 5.14°$	$1.51m, 4.29°$	$3.20m, 3.29°$	$1.50m, 4.03°$	$2.14m, 5.84°$
King's College	$1.92m, 5.40°$	$1.74m, 4.06°$	$0.99m, 3.65°$	$0.88m, 1.04°$	$1.06m, 2.81°$	$0.97m, 2.80°$
Average	$2.09m, 6.84°$	$1.92m, 6.28°$	$1.30m, 5.52°$	$1.63m, 2.86°$	$1.33m, 5.17°$	$1.34m, 4.71°$

- Only geometric loss is more accurate in rotation but 21% less accurate in translation

Comparison to SOA

- Experiments on outdoor scenes, Cambridge Landmarks

Dataset	PoseNet [25]	Bayesian PoseNet [26]	LSTM PoseNet [28]	Geometric PoseNet [27]	VGGRegNet [32]	EssNet
St Mary's Church	$2.65m, 8.48°$	$2.11m, 8.38°$	$1.52m, 6.68°$	$1.57m, 3.32°$	$2.11m, 8.11°$	$1.39m, 5.61°$
ShopFacade	$1.46m, 8.08°$	$1.25m, 7.54°$	$1.18m, 7.44°$	$0.88m, 3.78°$	$0.63m, 5.73°$	$0.87m, 4.58°$
OldHospital	$2.31m, 5.38°$	$2.57m, 5.14°$	$1.51m, 4.29°$	$3.20m, 3.29°$	$1.50m, 4.03°$	$2.14m, 5.84°$
King's College	$1.92m, 5.40°$	$1.74m, 4.06°$	$0.99m, 3.65°$	$0.88m, 1.04°$	$1.06m, 2.81°$	$0.97m, 2.80°$
Average	$2.09m, 6.84°$	$1.92m, 6.28°$	$1.30m, 5.52°$	$1.63m, 2.86°$	$1.33m, 5.17°$	$1.34m, 4.71°$

One network trained per scene

One single network

What can we do with more data?

- Training relative poses allows us to leverage more data

Dataset	1DSfM	Paris	1DSfM+Paris	
St Mary's Church	$1.75m, 6.88°$	$1.26m, 6.28°$	$1.19m, 5.13°$	$1.39m, 5.61°$
ShopFacade	$1.32m, 6.28°$	$1.01m, 7.07°$	$0.82m, 5.61°$	$0.87m, 4.58°$
OldHospital	$1.92m, 5.84°$	$2.06m, 5.13°$	$1.71m, 4.58°$	$2.14m, 5.84°$
King's College	$0.99m, 3.62°$	$0.84m, 3.24°$	$0.83m, 2.81°$	$0.97m, 2.80°$
Average	$1.50m, 5.66°$	$1.29m, 5.43°$	$1.14m, 4.53°$	$1.34m, 4.71°$

Pre-trained on a new dataset and fine-tuned on Cambridge Landmarks

Trained on Cambridge Landmarks

Out with the Old?

- SIFT-based methods still outperform CNN-based methods but there is potential in indoor scenes

- EssNet: relative pose estimation
  - mimics the visual localization pipeline → takes advantage of multiple view geometry
  - One network to perform localization in any scene

- Generalization still has a long way to go: if no images of the test scene are shown to the network, performance drops

Thank you!

Prof. Laura Leal-Taixé

https://dvl.in.tum.de

Technische Universität München