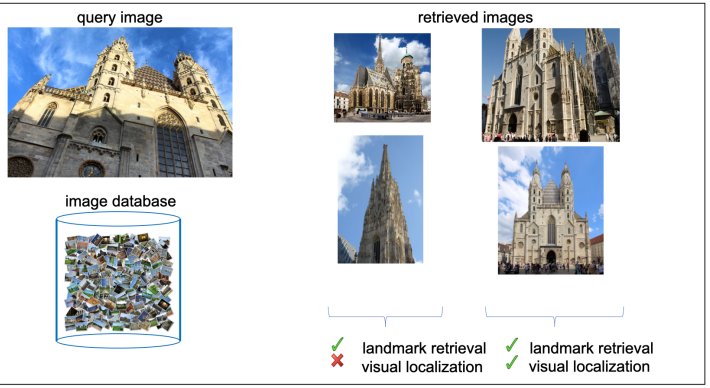
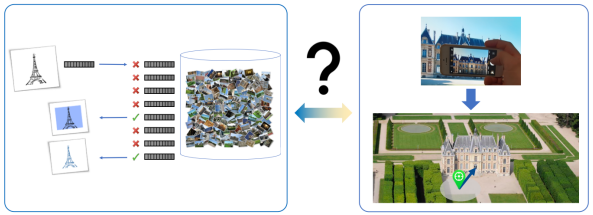
Benchmarking Image Retrieval for Visual Localization

[1]Noé Pion, [1]Martin Humenberger, [1]Gabriela Csurka, [1]Yohann Cabon, [2]Torsten Sattler
[1]NAVER LABS Europe, [2]Czech Technical University in Prague

This work received funding through the EU Horizon 2020 project RICAIP (grant agreement No 857306) and the European Regional Development Fund under IMPACT No. CZ.02.1.01/0.0/0.0/15 003/0000468.



How correlated is landmark retrieval / place recognition performance with visual localization performance?



https://github.com/naver/kapture-localization

Kapture

Benchmark

Aachen Day-Night-v1.1 [75]	Baidu-Mall [85]	RobotCar Seasons [56]
handheld scenario, augmented and mixed reality application		Autonomous driving
outdoor	indoor	outdoor
day-night, various viewpoints	close-up reflections and transparent surfaces, moving people	day-night, seasons, low image quality, dynamic traffic scenes

Handcrafted global features:
- **DenseVLAD** aggregates RootSIFT into an intra-normalized VLAD [90]

Learned global features:
- **NetVLAD** aggregates mid-level convolutional features extracted using VLAD [1]
- **AP-GeM** uses a generalized-mean pooling layer (GeM) to aggregate CNN-based descriptors [67]
- **DELG** is trained to extract both local and global features using one CNN [13]

Handcrafted local features:
- SIFT [53]

Learned local features:
- R2D2 [68]
- D2-Net [25]

Evaluation: https://visuallocalization.net
Metrics:
Visual localization accuracy:
- low (5m, 10°)
- medium (0.5m, 5°)
- high (0.25m, 2°)

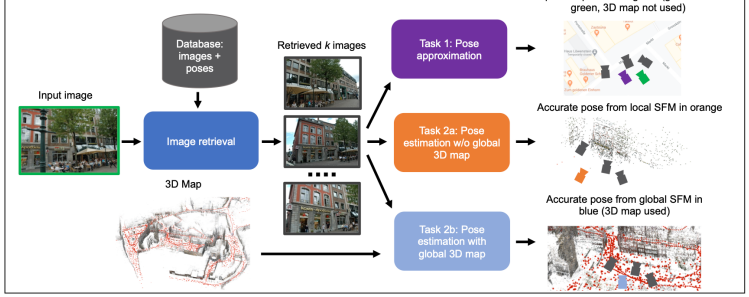
Landmark retrieval / Place recognition
- P@k: 1 if all retrieved imgs are relevant
- R@k: 1 if among the k imgs at least one is relevant

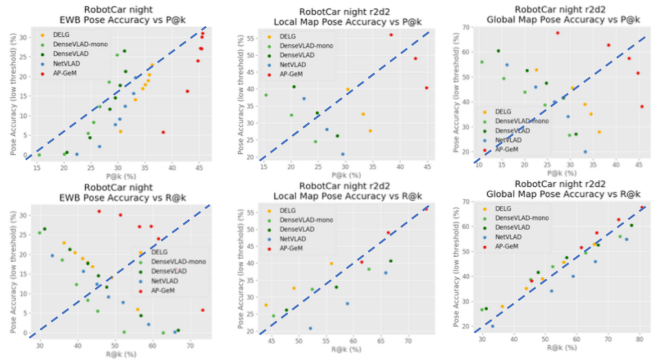
[1] Arandjelovic et al, NetVLAD: CNN Architecture forWeakly Supervised Place Recognition, CVPR'16
[13] Cao et al, Unifying Deep Local and Global Features for Efficient Image Search, arXiv'20
[25] Dusmanu et al, D2-Net: a Trainable CNN for Joint Description and Detection of Local Features, CVPR'19
[53] Lowe, Distinctive Image Features from Scale-Invariant Keypoints, IJCV'04
[67] Revaud et al, Learning with Average Precision: Training Image Retrieval with a Listwise Loss, ICCV'19
[68] Revaud et al, R2D2: Reliable and Repeatable Detectors and Descriptors, NeurIPS, 2019
[56] Maddern et al, The Oxford RobotCar Dataset, IJRR'17
[75] Sattler, et al, Benchmarking 6DoF Outdoor Visual Localization in Changing Conditions, CVPR'18
[85] Sun, A Dataset for Benchmarking Image-based Localization, CVPR'17
[90] Torii et al, 24/7 Place Recognition by View Synthesis CVPR'15

Tasks of Image Retrieval in Visual Localization

$$\mathbf{P}_q = \sum_{i=1}^{k} w_i \mathbf{P}_i$$

Interpolated pose in magenta (ground truth in green, 3D map not used)





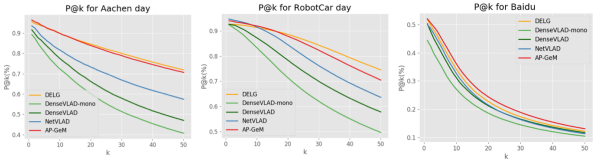
Task	Task 1: Pose approx.	Task 2a: Local SFM	Task 2b: Global SFM
P@k	weak	no	no
R@k	no	weak	yes

Correlation with R@k (Place Recognition) and P@k (Landmark Retrieval)

Landmark retrieval / Place recognition

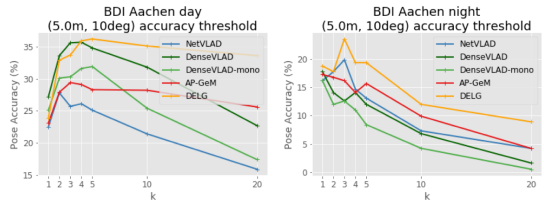
	$\mathcal{RO}$(m)	$\mathcal{RO}$ (h)	$\mathcal{RP}$(m)	$\mathcal{RP}$(h)
DenseVLAD	36.8	13.0	42.5	13.7
NetVLAD	37.1	13.8	59.8	35.0
AP-GeM	67.4	42.8	80.4	61.0
DELG	69.7	45.1	81.6	63.4

Mean Average Precision (mAP) on ROxford, RParis
https://github.com/filipradenovic/revisitop



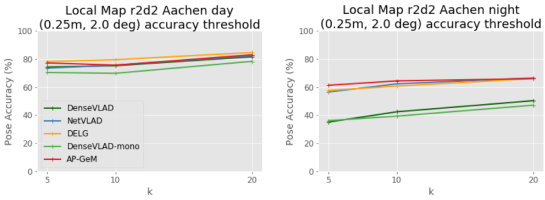
Learned descriptors outperform the handcrafted descriptor DenseVLAD

Pose Approximation

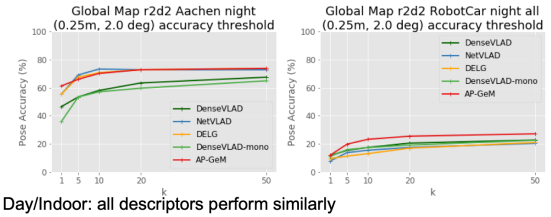


- DenseVLAD performs well on Aachen daytime images
- Night: learned > handcrafted

Accurate Pose Estimation with and without 3D Map



- Day: all descriptors perform similarly
- Night: learned descriptors significantly outperform DenseVLAD



- Day/Indoor: all descriptors perform similarly
- Night: learned descriptors > DenseVLAD (AP-GeM best