Learning Scene Geometry for Visual Localization in Challenging Conditions

Nathan Piasco[1,2] Désiré Sidibé [1] Valérie Gouet-Brunet [2] and Cédric Demonceaux [1]

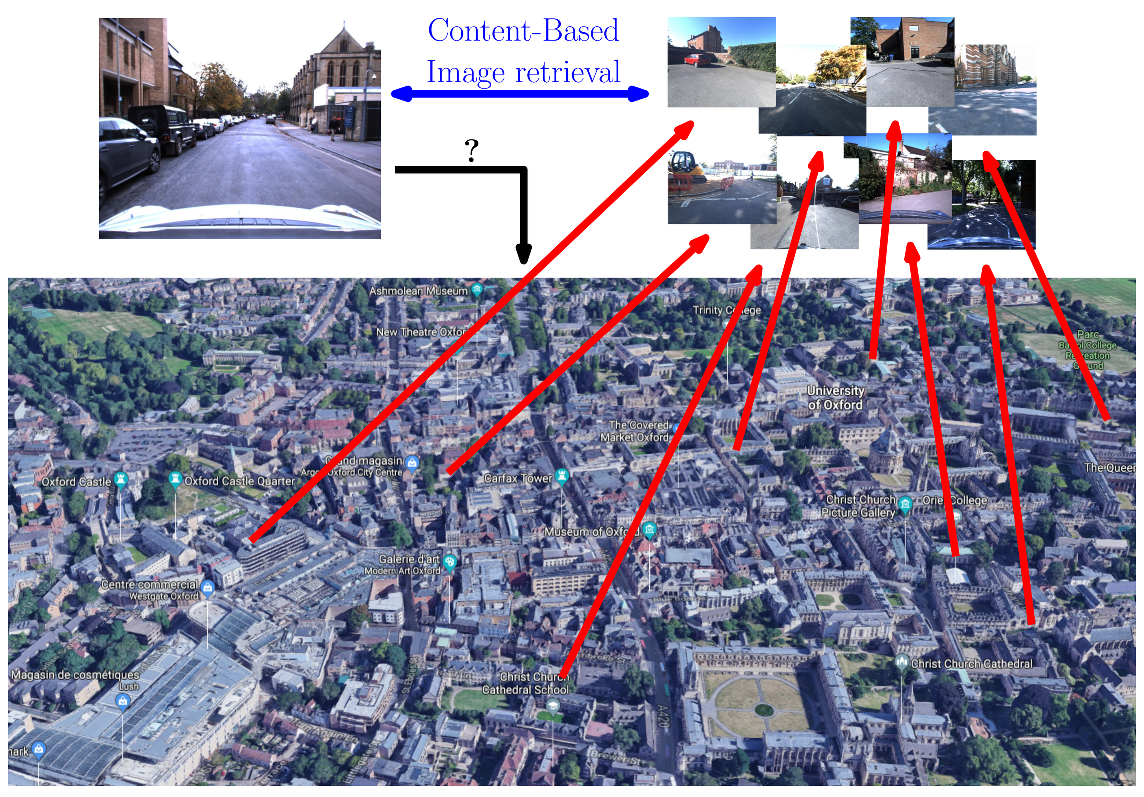
[1]ViBot ERL CNRS 6000, ImViA, Université Bourgogne Franche-Comté [2]LaSTIG, IGN, ENSG, Université Paris-Est, F-94160 Saint-Mandé, France

Abstract

We propose a new approach for outdoor large scale image based localization that can deal with challenging scenarios like cross-season, cross-weather, day/night and long-term localization. The key component of our method is a new learned global image descriptor, that can effectively benefit from scene geometry information during training. At test time, our system is capable of inferring the depth map related to the query image and use it to increase localization accuracy.

Problem statement

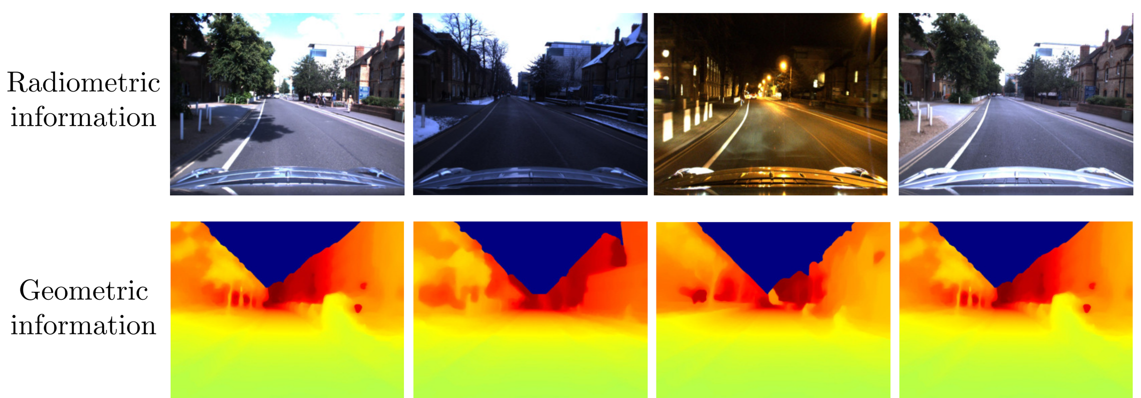
We want to find the position of an image query according to a known reference.



1. Collect geolocalized images on the area of interest.
2. Cast the image localization problem as an **image-retrieval problem**.
3. Transfer the pose of the closest retrieved candidate to the query.

Challenge in visual based localization

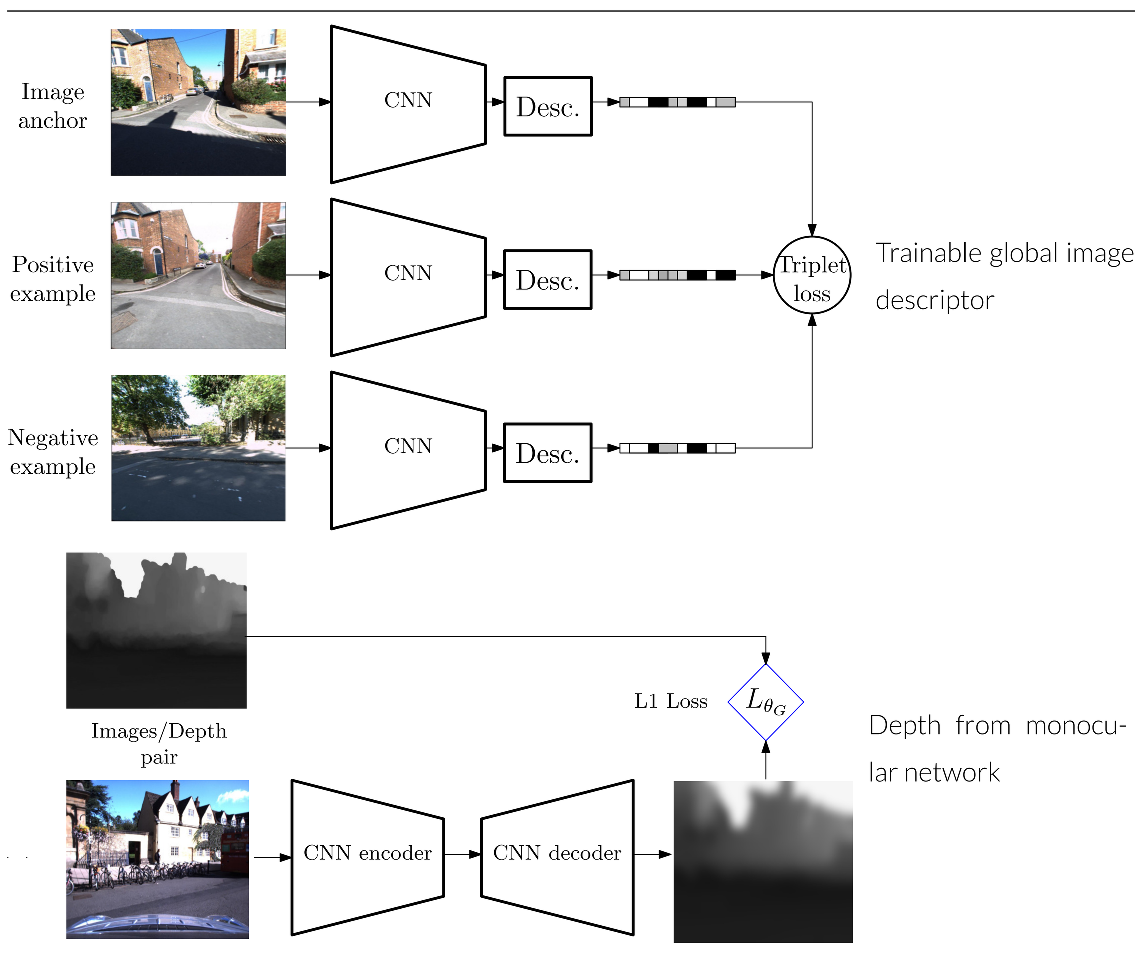
Drastic **visual changes** occur due to season/day-night cycles.



However, geometric information still remains the same. Unfortunately, geometric information is not always available.

How to use partial geometric information to improve image descriptor for localization?

System components



Global image descriptor

Triplet loss penalizes difference between anchor & positive example and similarity between anchor & negative example:

$$L = \max\left(\lambda + \|f(q_{im}) - f(q_{im}^{+})\|_2 - \|f(q_{im}) - f(q_{im}^{-})\|_2, 0\right),$$

where $\{q_{im}, q_{im}^{+}, q_{im}^{-}\}$ is an image triplet, $f(x_{im})$ the global descriptor of image x_{im} and λ an hyper-parameter controlling the margin between positive and negative examples.

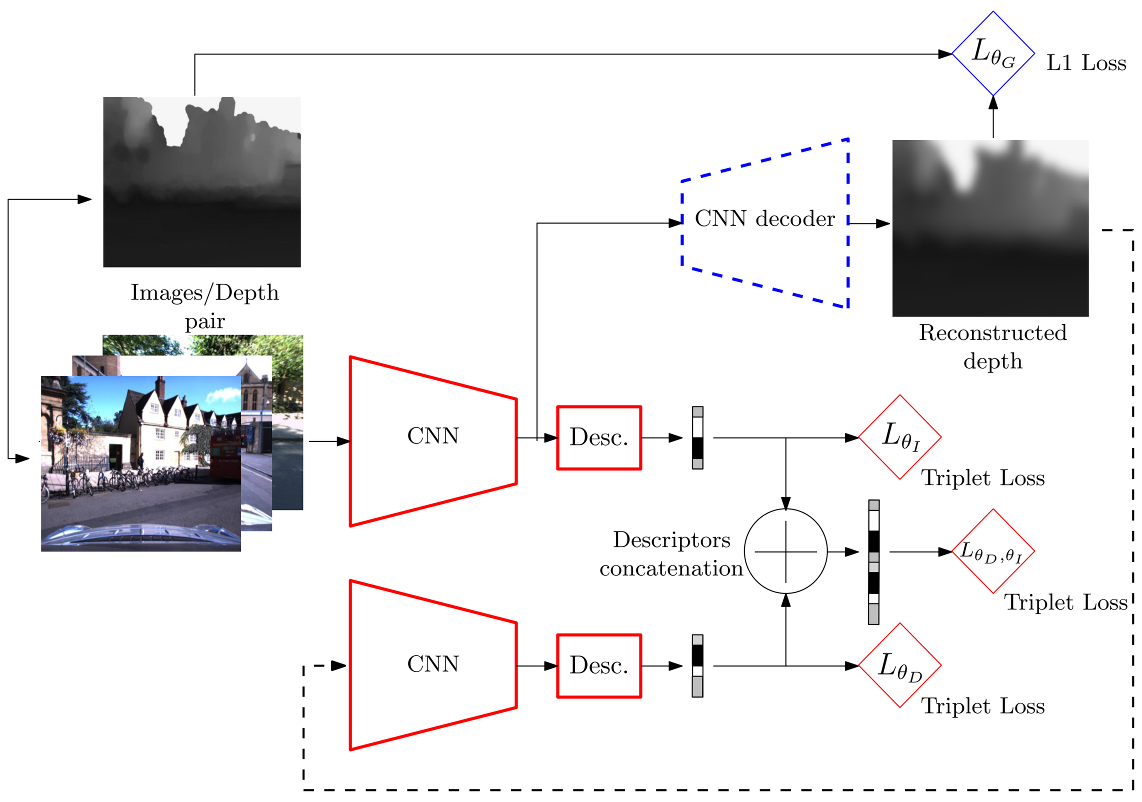
Depth from monocular

We use an encoder/decoder architecture to generate depth map from monocular images. Training is done in a supervised manner by minimising L_1 loss function:

$$L = \|G(I) - D_I\|_1,$$

where $G(I)$ is the generated depth map from image I and D_I the ground truth depth map associated to image I.

Learning through missing modality

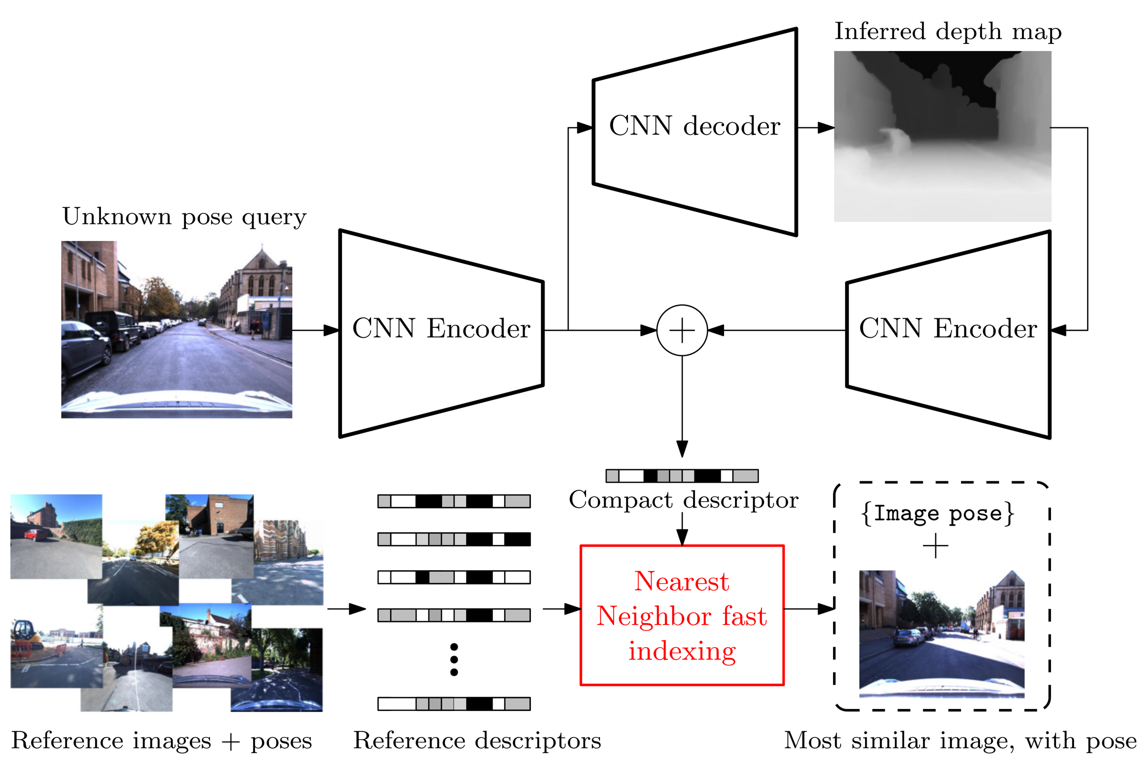


1. We use triplet loss to produce a strong image descriptor.
2. Latent image representation is given to a CNN decoder to reproduce the scene geometry.
3. We use a second CNN to produce a strong depth map descriptor.
4. Final descriptor is obtained by concatenating image and depth map descriptors.

Our proposal is trained with two different types of data:

- Image triplet
- Pair of image and associated depth map

System deployment



The depth information is only needed during the training step!

Dataset & Implementation

Training parameters:

- `pytorch` framework
- adam optimizer with lr=0.0001 and wd=0.001
- batches of 10 or 25 triplets, trained up to 50 epochs

Encoder		Descriptor	
Alexnet (A)	Resnet18 (Rt)	MAC [4]	NetVLAD [1]

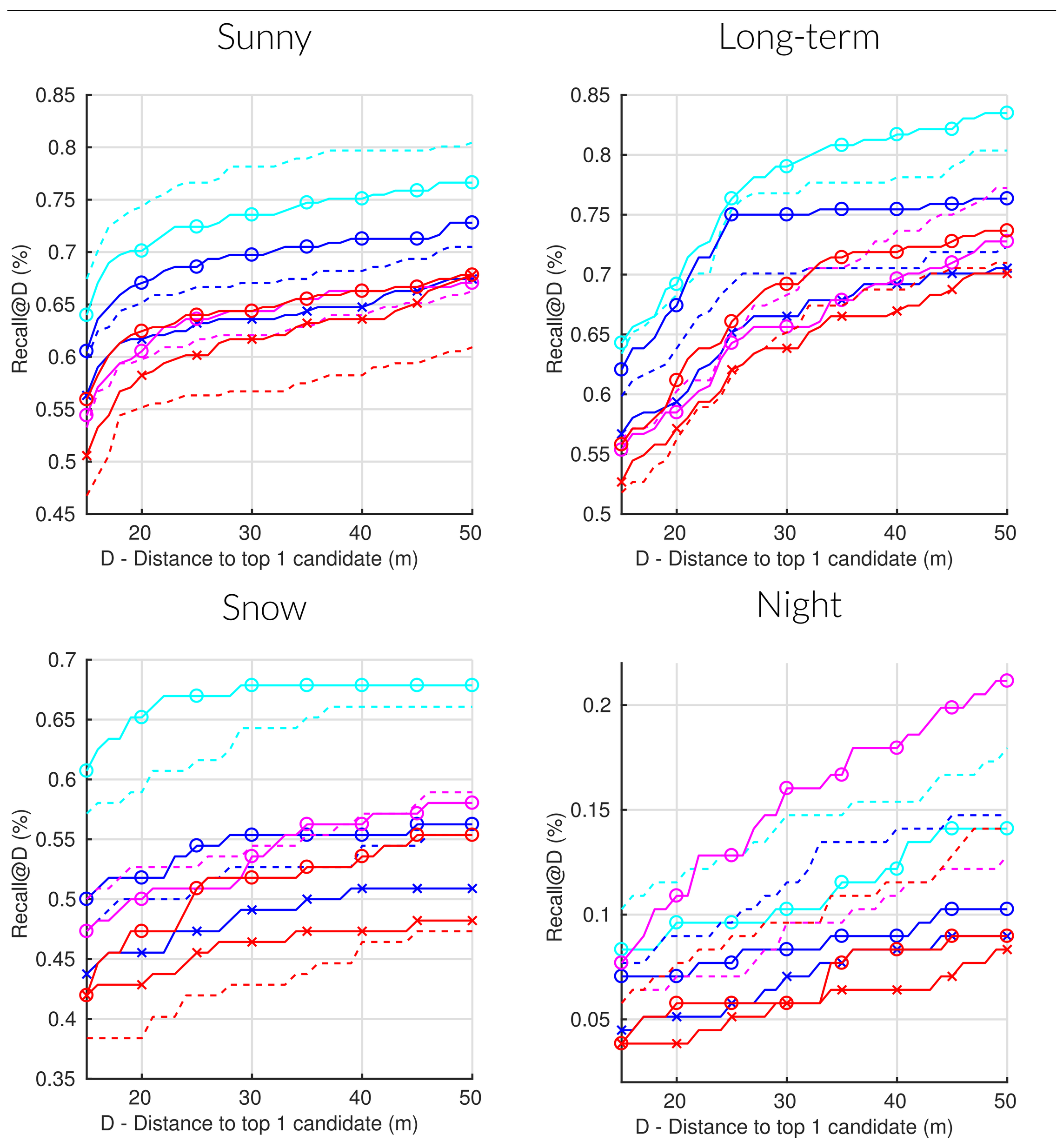
Table 1. Four possible combination of encoder/descriptor

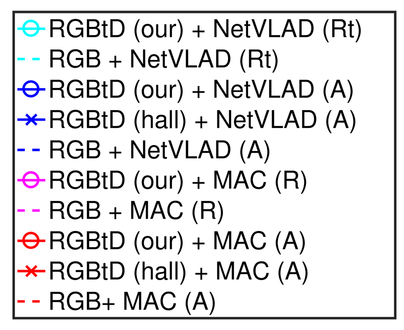
Competitors:

- Only RGB (dotted line)
- Hallucination network [2] (plain line with cross)

We train and test our proposal on RobotCar dataset with 4 different localization scenarios [3].

Results

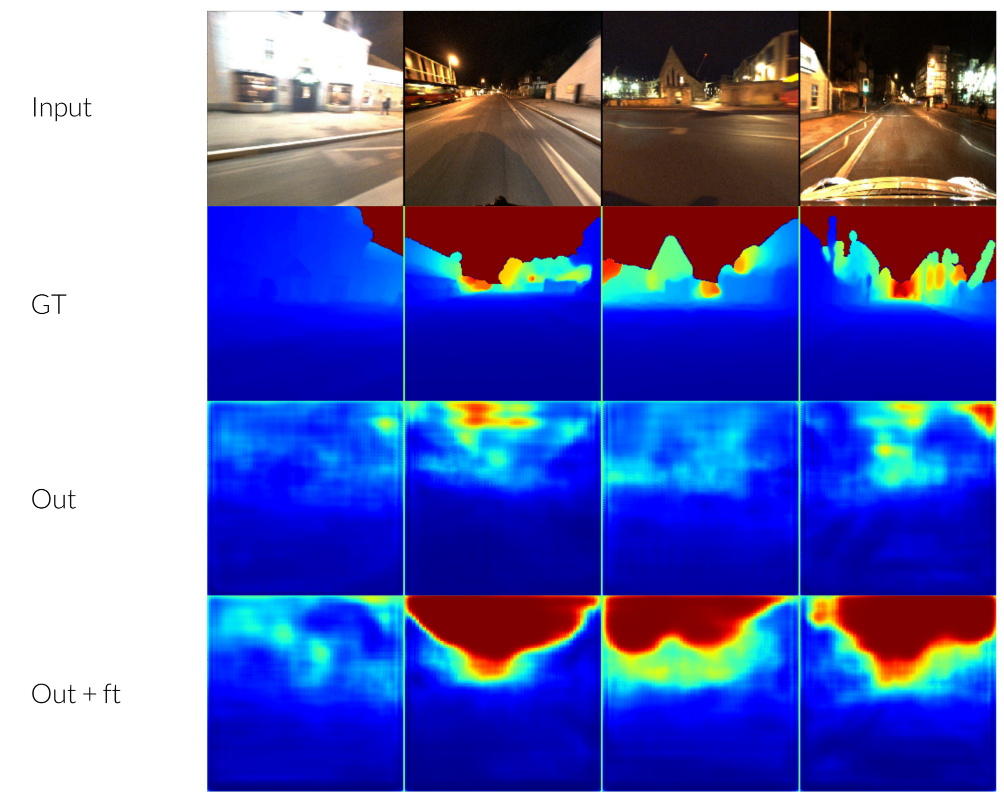




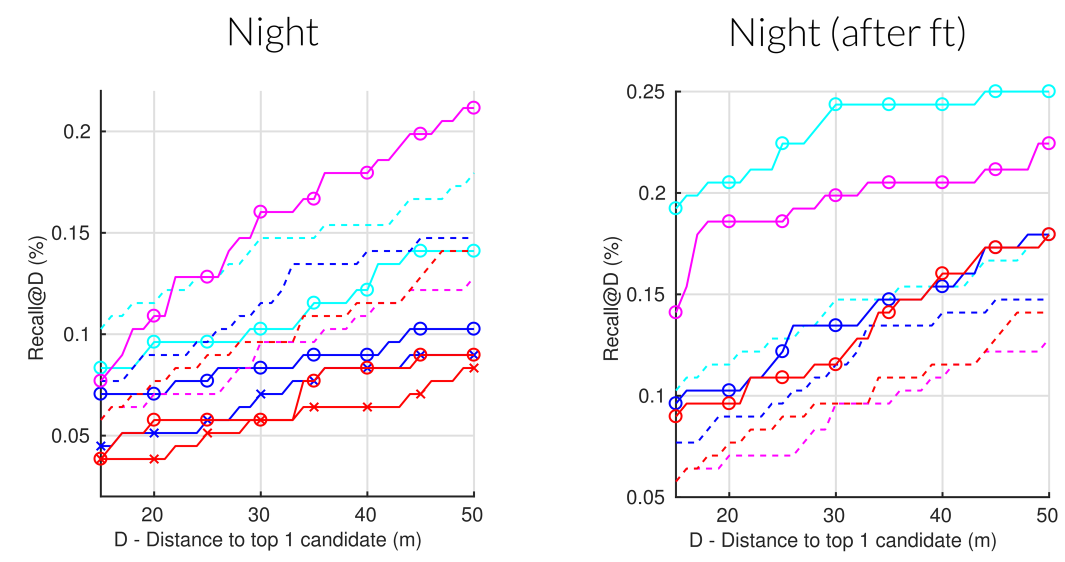
Metric: we compute the distance between **the top ranked** returned database image position and the query ground truth position and report the percentage of queries well located under a threshold D.

Improving night to day localization

Our network is *not able* to generate proper depth maps from night images.



Thanks to the design of our method, we can improve generation performances of the decoder without impacting the descriptors networks.



Perspectives

- Test our proposal on other modalities.
- Implement our method for other visual localisation tasks (e.g. direct pose regression).

Acknowledgments

We acknowledge the French ANR project pLaTINUM (ANR-15-CE23-0010) for its financial support. We also acknowledge NVIDIA Corporation for the donation of the Titan Xp GPU.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 5297–5307, 2017.

[2] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with Side Information through Modality Hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016.

[3] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research (IJRR)*, 2016.

[4] Filip Radenović, Giorgos Tolias, and Ondej Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.