Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions

Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, Cédric Demonceaux

Noname manuscript No.
(will be inserted by the editor)

Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions

Nathan Piasco · Désiré Sidibé · Valérie Gouet-Brunet · Cédric Demonceaux

Received: date / Accepted: date

Abstract Image indexing for lifelong localization is a key component for a large panel of applications, including robot navigation, autonomous driving or cultural heritage valorization. The principal difficulty in long-term localization arises from the dynamic changes that affect outdoor environments. In this work, we propose a new approach for outdoor large scale image-based localization that can deal with challenging scenarios like cross-season, cross-weather and day/night 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.

We show through extensive evaluation that our method can improve localization performances, especially in challenging scenarios when the visual appearance of the scene has changed. Our method is able to leverage both visual and geometric clues from monocular images to create discriminative descriptors for cross-season localization and effective matching of images acquired at different time periods. Our method can also use weakly annotated data to localize night images across a reference dataset of daytime images. Finally we extended our method to reflectance modality and we compare multi-modal descriptors respectively based on geometry, material reflectance and a combination of both.

Keywords Localization · Image retrieval · Side modality learning · Depth from Monocular · Global Image Descriptor

N. Piasco
VIBOT ERL CNRS 6000, ImViA Université Bourgogne Franche-Comté
Univ. Paris-Est, LaSTIG, IGN, ENSG, F-94160 Saint-Mandé, France

D. Sidibé
Université Paris-Saclay - Univ Evry
IBISC, 91020, Evry, France

V. Gouet-Brunet
Univ. Paris-Est, LaSTIG, IGN, ENSG, F-94160 Saint-Mandé, France

C. Demonceaux
VIBOT ERL CNRS 6000, ImViA, Université Bourgogne Franche-Comté

1 Introduction

Visual-Based Localization (VBL) is a central topic in a large range of domains, from robotics to digital humanities, involving advanced computer vision techniques [57]. It consists in retrieving the location of a visual input according to a known absolute reference. VBL is exploited in various applications such as autonomous driving [48], robot navigation or SLAM loop closing [45], augmented reality [75], navigation in cultural heritage collections [6, 66, 11], etc. In this paper, we address VBL as a content-based image retrieval (CBIR) problem where an input image is compared to a referenced pool of localized images. This image-retrieval-like problem is two-step: descriptor computation for both the query (online) and the reference images (offline) and similarity association across the descriptors. Since the reference images are associated to a location, by ranking images according to their similarity scores, we can deduce an approximate location for the query. Numerous works have introduced image descriptors well suited for image retrieval for localization [3, 39, 28, 65, 42]; we present in figure 1 our learned descriptor and the entire image localization pipeline. The final localization obtained by such a system can be used as it or as initializa-

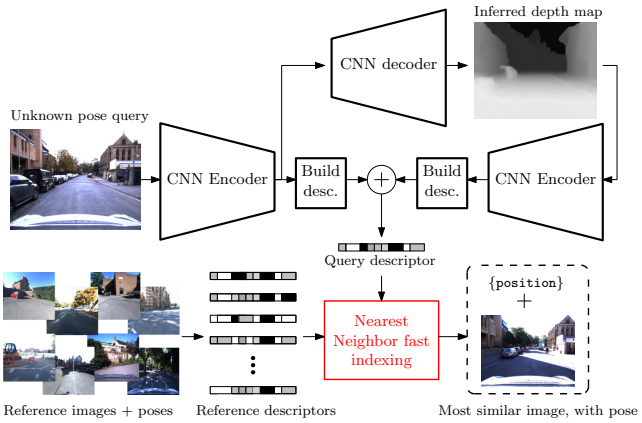

Fig. 1 Our method at test time: we rely *only on radiometric data* (= monocular images) to build low dimensional global image features used for localization. The reconstructed depth map used for image description makes our method more robust to visual changes that may occur in long-term localization scenarios.

tion for pose refinement method [70, 67, 58, 25]. In this paper, we focus on building a discriminative image descriptor for initial pose localization by image retrieval, so we do not investigate pose refinement methods.

One of the main challenges of outdoor image-based localization remains the mapping of images acquired under changing conditions: cross-season images matching [53], comparison of recent images with reference data collected a long time ago [80], day to night place recognition [81], etc. Recent approaches use complementary information in order to address these visually challenging localization scenarios: geometric information through point cloud [71, 72] or depth maps [16] and semantic information [5, 16, 53]. However geometric or semantic information are not always available or can be costly to obtain, especially in robotics or mobile applications when the sensor or the computational load on the system is limited, or in digital humanities when the data belong to ancient collections.

In this paper, we propose an image descriptor capable of reproducing the underlying scene geometry from a monocular image, in order to deal with challenging outdoor large-scale image-based localization scenarios. We introduce dense geometric information as side training objective to make our new descriptor robust to visual changes that occur between images taken at different times. Once trained, our system can be used on monocular images only to construct an expressive descriptor for image retrieval. This kind of system design is also known as side information learning [32], as it uses geometric and radiometric information only during the training step and pure radiometric data for the image localization.

The paper is organized as follows. In section 2, we first revisit recent works related to our method, including: state of the art image descriptors for large scale outdoor localization, method for localization in changing environment and side information learning approaches. In section 3, we describe in detail our new image descriptor trained with side depth information. In section 4 we give insight on our implementation and the dataset we used and we illustrate the effectiveness of the proposed method on six challenging scenarios in section 5. We discuss in section 6 about the complicated night to day localization scenario, and in section 7 we present a variation of our method using dense object reflectance map instead of depth maps. Section 8 finally concludes the paper.

Here, we extend our method presented in [60] with four original contributions: we report results on three new challenging scenarios from a dataset different from the one used to train the system, we investigate the impact of fine tuning the model for night to day localization, we compare our proposal with a domain adaptation method and show that these two approaches can be successfully combined together. We also extend the proposed method to another side modality, object reflectance, instead of depth map.

2 Related Work

In this section, we briefly discuss the state of the art for image-based retrieval applied to localization, before introducing some works that focus on the challenging outdoor localization scenario and we conclude by an overview of side-information learning (in our case, the side information is geometric) methods with deep learning.

2.1 CBIR for outdoor visual localization

Image descriptor. We tackle the task of localization as a problem of Content-Based Image Retrieval (CBIR). Standard image descriptors for image retrieval in the context of image localization are usually built by combining sparse features with an aggregation method, such as BoW, VLAD or DenseVLAD [81]. Before pooling the local features together, we can balance the contribution of some local clues in the global representation, *e.g.* to reduce visual burst [35, 82, 49]. Specific re-weighting scheme dedicated to image localization have been introduced in [4]. Global hand-crafted descriptors, like GIST [54], have also been used to perform image retrieval for localization [66, 7, 30].

Learned image descriptor. With the recent progress of image representation through deep neural network, many new types of image descriptors and their quantized counterpart have emerged [40, 13, 14]. Arandjelović et al. [3] introduce NetVLAD, a convolutional neural network that is trained to learn a well-suited image descriptor for image localization. Their proposal is trained using a triplet loss on multi-temporal data extracted from the Google street view time machine and they introduce a soft and differentiable VLAD layer for end-to-end optimization. Numerous other CNN image descriptors have been proposed in the literature: Gordo et al. [28] use a region proposal network to extract salient regions in the images and Radenović et al. [65] show that simple generalized-mean (GeM) pooling with proper training can produce state-of-the-art image retrieval results. NetVLAD layers have been improved by adding a self spatial attention mechanism [39] and advanced training losses [42]. Learned image descriptors have become a key component for numerous visual localization methods [78, 70, 68, 67, 58, 59, 25, 55], therefore we decide to build on these recent advances and use this learned image descriptor specially designed to solve the localization task.

Re-ranking candidates. Once the reference images have been efficiently ranked [51, 38] according to their similarity with the image query, re-ranking methods can be used. In addition to classical re-ranking methods like query expansion [18, 17] or spatial verification [56], Sattler et al. [69] introduce a re-ranking routine to improve the localization performances on large-scale outdoor area. They adapt the concept of geometric burstiness in images [35] to burstiness in places, by taking advantages of the known location of reference images. In [87, 88] authors use graph reasoning based on the location of the retrieved candidates in order to infer the most likely query position. These different refinement methods can be easily include in our visual localization framework as post-processing step.

2.2 Localization in challenging condition

As mentioned in the introduction, the main challenge in image-based localization is induced by visual changes due to time. Naseer et al. [52] show that using a combination of handcrafted and learned descriptors make the final image descriptor more robust to visual changes in images taken at different times. [22] introduces temporal consistency by considering a sequence of images. In our proposal, we suppose that we have access to only one image during the localization process.

Domain adaptation. An efficient way to handle visual changes in VBL is to use domain adaptation methods [43, 83]. Germain et al. [24] explicitly introduce the acquisition condition in the training pipeline of their CNN image descriptor using branching architecture. With this prior, they are able to create an image representation almost domain invariant. In [62, 2], authors synthesize new images to match the appearance of reference images, for instance to narrow the gap between daytime and nighttime images. This domain adaptation method allows to compare data with drastic visual changes, hence as in [24], we need priors about the nature of the changes that will occur. Using a similar approach, authors of [61] train a neural network to remove image noise induced by rain drops over camera lens.

Semantic information. An efficient method to improve robustness of VBL method in challenging condition is to rely on additional modalities. Numerous works [77, 80, 53, 73] enhance their visual descriptors by adding semantic information. For instance, in [53], authors use only the time-stable region of the scene (*e.g.* buildings) before computing their image descriptor. Semantic region consistency check is used in [23, 80] to reject wrong correspondences between matched images. In [73], authors design a multi-modal attention model for long-term image localization. Their system takes as input the image and its corresponding pixel segmentation and output a robust image descriptor.

Geometric information. Although semantic representation is useful for long term localization, it may be costly to obtain. Therefore, several methods rely on geometric information like point clouds [71, 72], or 3D structures [81]. Some methods rely only on geometric information, like in [84], where authors fuse PointNet [63] neural architecture with a NetVLAD [3] layer in order to perform point cloud retrieval. In [72], the presented method is based on 3D auto-encoder to capture a discriminative data representation for localization in challenging condition. Geometric information has the advantage of remaining more stable across time, compared to visual information but, similarly as semantic information, is not always available. That is why we decide to use depth information as side information in combination with radiometric data to learn a powerful image descriptor.

2.3 Depth as side information

As mentioned previously, complementary modalities, like geometry or semantic, may not be always available

at test time. This could be due to limitation on the computational resources of an embedded system or to the source of the input (different sensor, old data) during the localization process. For this reason, we make available the geometric information used in this work only during the offline training step and we rely on side information learning to benefit from this auxiliary modality at test time.

Learning using privileged information. The problem we tackle here is a well studied problem in computer vision called learning using privileged information [85, 74, 15, 50] or side information learning. Recent work from [41] casts the side information learning problem as a domain adaptation problem, where source domain includes multiple modalities and the target domain is composed of a single modality. Another successful method has been introduced in [32]: authors train a deep neural network to hallucinate features from a depth map only presented during the training process to improve objects detection in images. The closest work to ours, presented in [86], uses recreated thermal images to improve pedestrian detection on standard images only. Our system, inspired by [86], learns how to produce depth maps from images to enhance the description of these images.

Cross-modality retrieval. Even though they share similarities, the problems of side-information and cross-modality are not directly related. Cross-modality retrieval is the task of computing a common description of the same scene, measured from different modalities [36, 19]. For instance, recovering the closest image regarding the corresponding depth map (with only this geometric information as input) is a cross-modality problem. In our application, we are not interested in cross-modality retrieval as we determine that the radiometric information (*e.g.* the image) is our main modality, and is always available. We are trying to use multiple modalities in a constrained setting (the auxiliary modality is not always available, only during training) to build an image descriptor more efficient, because less restrictive, than cross-modality descriptor [32].

Depth from monocular image for localization. Modern neural networks architectures can provide reliable estimation of the depth associated to monocular image in a simple and fast manner [20, 26, 47]. This ability of neural networks has been used in [79] to recover the absolute scale in a SLAM mapping system. In [58], authors introduce a method that use the generated depth map to recover the exact pose of an image query by ICP refinement. An incremental improvement of this work is presented in [59], where the authors consider a more generic PnP formulation to refine the camera pose. Loo et al. [44] use the depth estimation produced by a CNN to improve a visual odometry algorithm by reducing the incertitude related to the projected 3D points. In this work, we use the depth information obtained by a neural network as stable features across season changes.

3 Method

We use the descriptor architecture introduced in our previous work [60]. The method, presented in figure 2, is composed of:

- a CNN image encoder E_I linked to a feature aggregation layer d_I that produces a compact image descriptor,
- a CNN image decoder D_G used to reconstruct the corresponding depth map according to the monocular image,
- a CNN depth map encoder E_D linked to a feature aggregation layer d_D that produces a compact depth map descriptor,
- a fusion module that concatenates the image and depth map descriptor.

3.1 Training routine

Trainable parameters are θ_I the weights of image encoder and descriptor $\{E_I, d_I\}$, θ_D the weights of the depth encoder and descriptor $\{E_D, d_D\}$ and θ_G the weights of the decoder used for depth map generation.

For training our system, we follow standard procedure of descriptor learning based on triplet margin losses [3]. A triplet $\{x, x^+, x^-\}$ is composed of an anchor image x, a positive example x^+ representing the same scene as the anchor and an unrelated negative example x^-. The first triplet loss acting on $\{E_I, d_I\}$ is:

$$L_{f_{\theta_I}}(x, x^+, x^-) = max\left(\lambda + \left\|f_{\theta_I}(x) - f_{\theta_I}(x^+)\right\|_2 - \left\|f_{\theta_I}(x) - f_{\theta_I}(x^-)\right\|_2, 0\right), \quad (1)$$

where $f_{\theta_I}(x)$ is the global descriptor of image x and λ an hyper-parameter controlling the margin between positive and negative examples. f_{θ_I} can be written as:

$$f_{\theta_I}(x) = d_I(E_I(x)), \quad (2)$$

where $E_I(x)$ represents the deep feature maps extracted by the encoder and d_I the function used to build the final descriptor from the feature.

We train the depth map encoder and descriptor $\{E_D, d_D\}$ in a same manner, with the triplet loss of equation (1), $L_{f_{\theta_D}}(\hat{z}, \hat{z}^+, \hat{z}^-)$, where $f_{\theta_D}(z)$ is the global

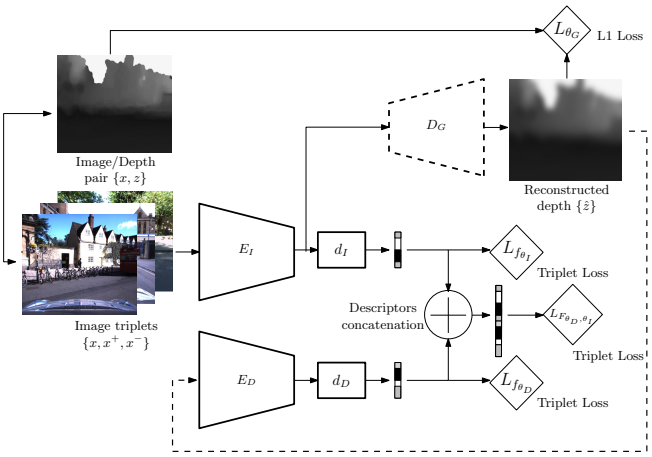

Fig. 2 Image descriptors training with auxiliary depth data (our work):
two encoders are used for extracting deep features map from the main image modality and the auxiliary reconstructed depth map (inferred from our deep decoder). These features are used to create intermediate descriptors that are finally concatenated in one final image descriptor.

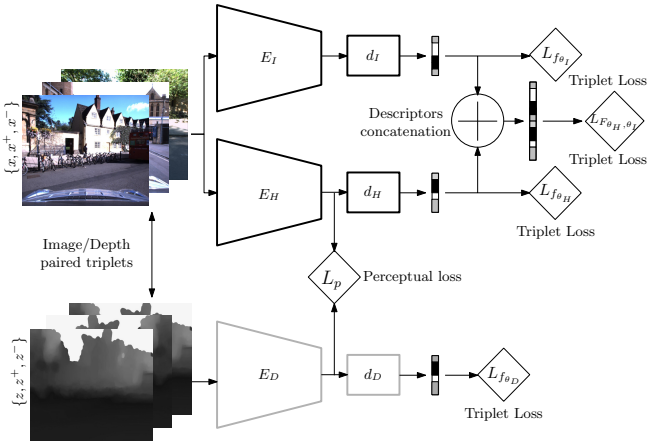

Fig. 3 Hallucination network for image descriptors learning:
we train an hallucination network, inspired from [32], for the task of global image description. Unlike the proposed method (see figure 2), hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the depth map itself.

descriptor of depth map z and $\hat{z}$ is the reconstructed depth map of image x by the decoder D_G:

$$\hat{z} = D_G(E_I(x)). \tag{3}$$

Decoder D_G uses the deep representation of image x computed by encoder E_I in order to reconstruct the scene geometry. Notice that even if the encoder E_I is not especially trained for depth map reconstruction, its intern representation is rich enough to be used by the decoder D_G for the task of depth map inference. We choose to use the features already computed by the first encoder E_I instead of introducing another encoder for saving computational resources.

The final image descriptor is trained with the triplet loss $L_{F_{\theta_I,\theta_D}}(x, x^+, x^-)$, where $F_{\theta_I,\theta_D}(x)$ denotes the fusion of image descriptor and depth map descriptor:

$$F_{\theta_I,\theta_D}(x) = fuse\left(f_{\theta_I}(x), f_{\theta_D}(\hat{z})\right). \tag{4}$$

In order to train the depth map generator, we use a L_1 loss function:

$$L_{\theta_G} = \|z - \hat{z}\|_1. \tag{5}$$

We apply L_1 penalty to minimize the blur effect on our reconstructed depth maps [34]. As we use a Sigmoid activation function at the end of our decoder, limiting the range of our depth value to 1, we do not need to use more sophisticated function like hybrid L_1/L_2 Huber loss [89] for faster convergence.

The whole system is trained according to the following constraints:

$$(\theta_I, \theta_D) := arg\min_{\theta_I,\theta_D}\left[L_{f_{\theta_I}} + L_{f_{\theta_D}} + L_{F_{\theta_I,\theta_D}}\right], \tag{6}$$

$$(\theta_G) := arg\min_{\theta_G}\left[L_{\theta_G}\right]. \tag{7}$$

We use two different optimizers: one updating θ_I and θ_D weights regarding constraint (6) and the other updating θ_G weights regarding constraint (7). Because decoder D_G relies on feature maps computed by encoder E_I (see equation (3)), at each optimization step on θ_I we need to update decoder weights θ_G to take in account possible changes in the image features. We finally train our entire system, by alternating between the optimization of weights $\{\theta_I, \theta_D\}$ and $\{\theta_G\}$ until convergence.

3.2 Hard mining and swapping in triplet ranking loss

Hard negative minning policy. Hard mining is a crucial step in metric learning [3, 65, 27, 33]. We construct our triplets like in [3], using the GPS-tag information provided with the data. We gather N triplets $\left\{x, \{x_i^+\}_{i\in[1,M_p]}, \{x_i^-\}_{i\in[1,M_n]}\right\}$ composed of one anchor, M_p positive examples and M_n negative examples. Negative examples are easy to collect as we only have to consider all the data located further than a given distance threshold (according to the GPS information), resulting in a large number of negative examples ($M_n \approx 2000$ in our experiment).

Because M_n is too large, exact hard mining examples is not tractable. In [3], authors store a fixed representation of the negatives examples that is used for negative mining. They update the representation of all negative examples as soon as the new representation computed by their model differs to much from the stored one. We adopt a different approach with a small overhead in terms of computation but taking into account model updates directly. At each iteration, we randomly select a subset of M_n^{sub} negative examples from the entire pool, and compute the true hard negative example from this subset. This strategy also acts as regularization during training as the negative training examples are different at each epoch.

Our mining is performed according to the final feature representation F_{θ_I,θ_D} and the image triplet are the same for the three triplet losses $L_{f_{\theta_I}}, L_{f_{\theta_D}}$ and $L_{f_{\theta_I,\theta_D}}$.

Anchor and positive swapping. We also adopt the swapping technique introduced in [8]. It simply consists in choosing the most confusing pair between {anchor, negative} and {positive, negative} examples:

$$L_{swap}(x,x^+,x^-) = max\left(\lambda + \left\|f(x)-f(x^+)\right\|_2 - min\left(\left\|f(x)-f(x^-)\right\|_2, \left\|f(x^+)-f(x^-)\right\|_2\right), 0\right). \quad (8)$$

Multiple examples. Finally, we use all the positive examples and the mined M_n^{hard} hard negative examples from the initial pool of negative examples, to compute a normalized triplet ranking loss:

$$L_{final}\left(x, \left\{x_i^+\right\}_{i\in[1,M_p]}, \left\{x_i^-\right\}_{i\in[1,M_n^{sub}]}\right) = \frac{1}{M_p M_n^{hard}} \sum_{i=1}^{M_p}\sum_{j=1}^{M_n^{hard}} L_{swap}\left(x, x_i^+, x_j^-\right). \quad (9)$$

3.3 Descriptors fusion and dimension reduction

We test several functions for the fusion of the descriptors, the one introduced in equation 4, in order to benefit as much as possible from the complementarity of the main and the auxiliary modalities. We compare: simple descriptors concatenation, hand-tuned descriptors scalar weighting, trained scalar weighting [76], trained modal attention mechanism at the level of descriptors and trained spatial and modal attention mechanism at the level of the deep features [73]. We found that all the fusion policies perform similarly, so we use the simple concatenation operator to fuse the descriptors. Indeed, the modalities fusion are learned by our system through the triplet loss $L_{F_{\theta_I,\theta_D}}$, making the system aware of what is important and complementary in the radiometric and geometric domain, without the need of a complex fusion method.

We can reduce the dimension of the final descriptor by applying PCA + whitening [3, 64, 65, 28]. After the convergence of the whole system we reuse the images from the training dataset to compute the PCA parameters.

3.4 Side information learning with hallucination

We compare our method of side information learning with a state-of-the-art approach system, named hallucination network [32]. The hallucination network is originally designed for object detection and classification in images. We adapt the work of [32] to create an image descriptor system that benefits from depth map side modality during training. The system is presented in figure 3.

Hallucination descriptor. The key component of Hoffman et al. [32] proposal is the hallucination network. The task of the hallucination branch is, with images as input, to reproduce feature maps that would have been obtained by a network trained with depth map rather than the depth map itself. The hallucination network shares the same architecture for the principal and the auxiliary branches. The hallucination descriptor is composed of an encoder E_H and a descriptor d_H with trainable weights θ_h. It is trained with triplet ranking loss $L_{f_{\theta_H}}$ under the constraint of a perceptual loss [37]:

$$L_p(x,z) = \left\|E_H(x) - E_D(z)\right\|_2. \quad (10)$$

This constraint can be interpreted as knowledge distillation [31]. Final image descriptor $F_{\theta_I,\theta_H}(x)$ is obtained by concatenating $f_{\theta_I}(x)$ and $f_{\theta_H}(x)$.

Overall training. Training routine presented in [32] is two-step: we first optimize weights θ_D of the auxiliary descriptor with loss $L_{f_{\theta_D}}(z,z^+,z^-)$ and, secondly, we initialize hallucination weights θ_H with pre-trained weights θ_D and solve the following optimization problem:

$$(\theta_I,\theta_H) := arg\min_{\theta_I,\theta_H} \alpha\left[L_{F_{\theta_{I,H}}}(x,x^+,x^-) + L_{f_{\theta_I}}(x,x^+,x^-) + L_{f_{\theta_H}}(x,x^+,x^-)\right] + \gamma\left[L_p(x,z) + L_p(x^+,z^+) + L_p(x^-,z^-)\right], \quad (11)$$

where α and γ are weighting constants, set after hyperparameters finetuning to $\alpha =$ and $\gamma =$.

In the original paper [32], during this final training step, all the networks weights were optimized jointly. However, we have found that freezing the weights θ_D of the auxiliary descriptor $\{E_D, d_D\}$ in this final training step leads to better results. Metric learning is a more complicated optimization problem than fully supervised object detection training (the task targeted in the original contribution of Hoffman et al. [32]), we deduce that reducing the number of parameters to optimize for this problem leads to more stable convergence. We apply the same triplet mining that the one used in our method during the two steps of training.

Like our proposal, this method requires triplets of RGB-D data to be trained and, at test time, the principal and hallucination descriptors are used on images only and the auxiliary descriptor $\{E_D, d_D\}$ is dropped.

Advantages and drawbacks. One advantage of the hallucination network over our proposal is that it does not require a decoder network, resulting in a architecture lighter than ours. However, it needs a pre-training step, where image encoder and depth map encoder are trained separately from each other before a final optimization step with the hallucination part of the system. Our system does not need such initialization. Training the hallucination network requires more complex data than the proposed method. Indeed, it needs to gather triplets of images, and depth map pairs, which require to know the absolute position of the data [3, 42], or to use costly algorithms like Structure from Motion (SfM) [26, 65, 39].

One advantage of our method over the hallucination approach is that we have two unrelated objectives during training: learning an efficient image representation for localization and learning how to reconstruct scene geometry from an image. It means that we can train several parts of our system separately, with different source of data. Especially, we can improve the scene geometry reconstruction task with non localized {*image, depth map*} pairs. These weakly annotated data are easier to gather than triplet, as we only need calibrated system capable of sensing radiometric and geometric modalities at the same time. We will show in practice how this can be exploited to fine tune the decoder part to deal with complex localization scenarios in section 5.2.

4 Implementation details

This section presents the datasets used for training and testing our method as well as details about our implementation and a short presentation of the competitors compared to our proposal.

4.1 Dataset

We have tested our proposal on the *Oxford Robotcar* public dataset [46] and on the *CMU Visual localization* dataset [9] from the city of Pittsburg. These are common datasets used for image-based localization [71] and loop closure algorithms involving neural networks training [62] under challenging conditions.

4.1.1 Training data

We exploit the temporal redundancy present in Oxford Robotcar dataset to build the images triplets needed to train our CNN. We build 400 triplets using three runs acquired at dates: `15-05-19, 15-08-28` and `15-11-10`, and we select an area of the city different from the one used for training our networks for validation. Our triplets creation process is explained in section 3.2 and we fix $M_n^{sub} = 20$ and $M_p = 4$ in such a way that one triplet example is composed of 25 images.

Depth modality is extracted from the lidar point cloud. When re-projected in the image frame coordinate, it produces a sparse depth map. Since deep convolutional neural networks require dense data as input, we pre-process these sparse modality maps with the inpainting algorithm from [10] in order to densify them. We drop depth values larger than 100 meters in order to produce depth maps with value in $[0, 1]$, consistent with the sigmoid decoder output.

4.1.2 Testing data

We propose six testing scenarios, 3 on each datasets. For the Oxford Robotcar dataset, the reference dataset is composed of 1688 images taken every 5 meters along a path of 2 km, when the weather was overcast. Query sets are composed of approximately 1000 images. The three query sets are:

- Oxford – Long-term (LT): queries have been acquired 7 months after the reference images under similar weather conditions,
- Oxford – Snow: queries have been acquired during a snowy day,
- Oxford – Night: queries have been acquired at night, resulting in radical visual changes compared to the reference images.

For the CMU Visual localization dataset, the reference dataset is composed of 1944 images with a sunny weather and the three query sets are (approximately 1000 images by set):

- CMU – Long-term (LT): queries have been acquired 10 months after the reference images under similar weather conditions,

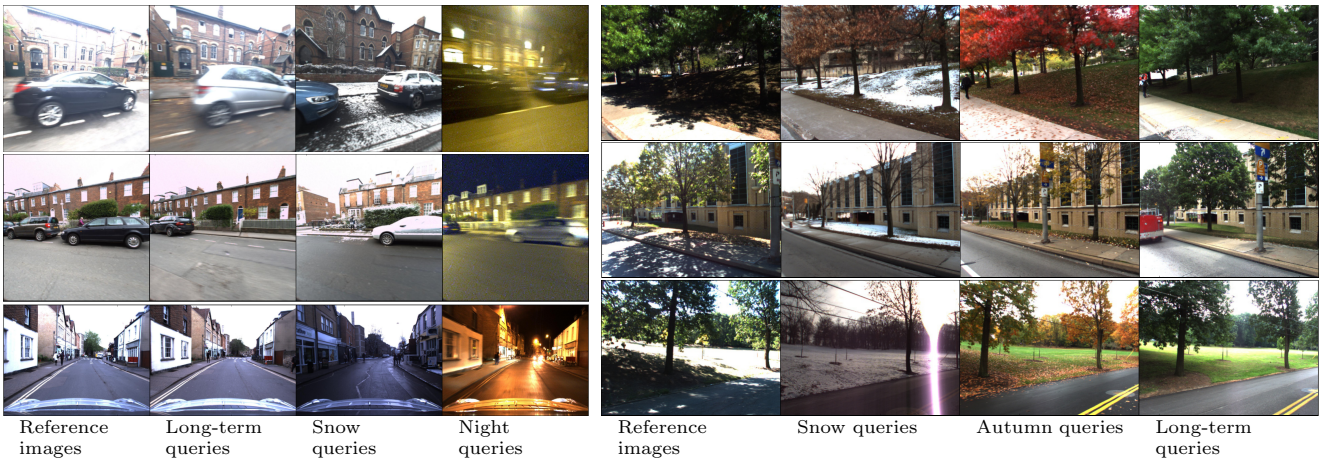

Fig. 4 Examples of test images : we evaluate our proposal on 6 challenging localization sequences. Query image samples and the closest reference images in the database are presented from Oxford Robotcar [46] (left) and CMU season dataset [9] (right).

– CMU – Snow: queries have been acquired during a snowy day,
– CMU – Autumn: queries have been acquired during Autumn, featuring warm-coloured foliage and low sunlight compare to the reference data.

Query examples are presented in figure 4.

4.1.3 Evaluation metric

For a given query, the retrieved reference images are ranked according to the cosine similarity score computed over their descriptors. To evaluate the localization performances, we consider two evaluation metrics:

– **Recall @N:** we plot the percentage of well localized queries regarding the number N of returned candidates. A query is considered well localized if one of the top N retrieved images lies within $25m$ radius from the ground truth query position.
– **Top-1 recall @D:** we compute the distance between the top ranked returned database image position and the query ground truth position, and report the percentage of queries located under a threshold D (from 15 to 50 meters), like in [88]. This metric qualifies the accuracy of the localization system.

4.2 Implementation

Our proposal is implemented using Pytorch as deep learning framework, ADAM stochastic gradient descent algorithm for the CNN training with learning rate set to 1e-4, weight decay to 1e-3 and λ in the triplet loss of equation 1 equal to 0.1. We use batch size between 10 and 25 triplets depending of the size of the system to train, convergence occurs rapidly and takes around 30 to 50 epochs. We perform both positive and negative hard mining, as in [65]. Images and depth maps are re-sized to 224×224 pixels before training and testing.

4.2.1 Encoder architectures

We test the fully convolutional part of Alexnet and Resnet18 architectures for features extraction. As shown in [58], we use the truncated version of Resnet18 to increase the spatial resolution of the final features block. Weights are initialized with the pre-trained weights on ImageNet. We always use Alexnet encoder to extract features from raw depth map, reconstructed depth map, or hallucinated depth map. Indeed the quality of our depth map is usually very low, and we have found that using deeper network does not significantly improve localization results. We transform the 1-channel depth map into 3-channels jet colorization depth map in order to benefit from the convolutional filters learned on ImageNet. We do not use the 3-channels HHA depth map representation introduced in [29] as it have been shown to perform equivalently to jet colorization [21].

4.2.2 Descriptor architectures

We test the two state-of-the-art image descriptors MAC and NetVLAD. MAC [64] is a simple global pooling method that takes the maximum of each feature map from the encoder output. NetVLAD [3] is a trainable pooling layer that mimics VLAD aggregation method. For all the experiments, we set the number of NetVLAD clusters to 64. Finally, both MAC and NetVLAD descriptors are L_2 normalized.

By combining Alexnet or Resnet encoder with MAC or NetVLAD descriptor pooling, we obtain 4 global image descriptor variants.

4.2.3 Decoder architecture

The decoder used in our proposal is based on Unet architecture and inspired by network generator from [34]. Dimension up-sampling is performed through inverse-convolutions layers. Decoder weights are initialized randomly.

4.3 Competitors

We compare the four following global image descriptors:

1. *RGB only (**RGB**):* simple networks composed of encoder + descriptor trained with images only, without side depth maps information. Networks are trained on Robotcar dataset following the standard procedure of image descriptor training with triplet loss [3, 65].
2. *Our proposal (**RGB(D)**):* network that uses pairs of aligned image and depth map during training step and images only at test time. We follow training procedure as explained in 3.1.
3. *Hallucination network (**RGB(H)**):* we compare our version of hallucination network, trained on aligned triplets of images and depth maps. We follow training procedure described in the previous section 3.
4. *Oracle descriptor **(RGBD)**:* we compare our proposed descriptor with an oracle version of our work, that has access to the ground truth depth map of images during testing. We train this descriptor independently with image and depth map pairs (*i.e* we do not reuse our RGB(D) network and substitute the generated depth map by the ground truth depth maps).

For fair comparison, as **RGB(D)**, **RGB(H)** and **RGBD** image descriptors are obtained by concatenating two full-size descriptors (see section 3.3), we perform PCA to reduce the size of the final descriptor of all four methods to 2048.

5 Experiments

As a first step, we conduct preliminary experiments to justify design choices for our method. Then, in the second part of this section, we compare the localization performances of the proposed image descriptors.

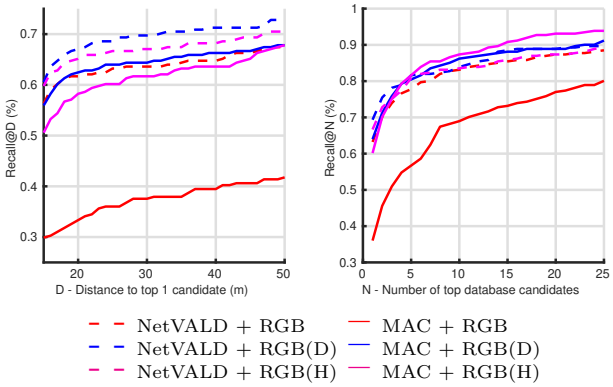

Fig. 5 Comparison of descriptors pooling layer: NetVLAD [3] pooling layer performs better than MAC [65] in our preliminary experiment, whatever the tested method.

Table 1 Contribution of the depth side information during training.

Network		Top-1 recall@D			Recall@N	
Name	#Param.	@15	@30	@50	@1	@5
RGB + MAC	2.5M	46.7	56.7	60.9	56.3	76.6
RGB^+ + MAC	7.9M	51.0	61.0	66.7	60.1	79.3
RGB(D) + MAC	7.9M	**55.9**	**64.4**	**67.8**	**64.0**	**80.5**

5.1 Preliminary results

5.1.1 Contribution of the depth information

In this paragraph, we investigate the impact on localization performances provided by the side geometric information on our method. To ensure a fair comparison in terms of number of trainable parameters, we introduce RGB^+ network that has the same architecture as our proposed method. We train RGB^+ with images only to compare the localization results against our method that uses side depth information. For training RGB^+, we simply remove the loss introduced in equation (3), and make the weights of the decoder trainable when optimizing triplets losses constraints. Results on the validation dataset with encoder architecture Alexnet are presented in table 1.

Increasing the size of the system results in a better localization (RGB^+ + MAC versus RGB + MAC). However, our RGB(D) + MAC system always produces higher localization results facing RGB^+ + MAC, which shows that the side depth information provided during training is wisely used to describe the image location.

5.1.2 Descriptor comparison

In figure 5, we present the localization scores of the three different methods on the validation set with Alexnet as base encoder. It clearly demonstrates the superi-

ority of the NetVLAD pooling layer compared to the MAC descriptor. Thus, we only use NetVLAD as pooling layer for the rest of the experiments, in combination with Alexnet or Resnet encoder architecture. Still, this preliminary experiment has shown that the proposed method can be used in combination with various descriptor pooling layers.

5.2 Localization results

Localization results on the six query sets are presented in figure 6. We also show, in figure 7, some examples of top-1 returned candidate by the different descriptors. Both methods trained with auxiliary depth information (hallucination RGB(H) and our RGB(D)) perform on average better than the RGB baseline. This shows that the geometric clues given during the training process can be efficiently used for the task of image-only retrieval for localization. This observation is confirmed with the results obtained on the oracle descriptor: RGBD outperforms, by a large margin, all the other methods. This result also shows that our method could achieve better performances with a more realistic depth reconstruction. Compared to hallucination network, our method shows better results, both in terms of recall and precision. We report results for the hallucination network only with encoder Alexnet as we were not able to obtain stable training when using a deeper architecture. This may due to the limited amount of data we have for finetuning the method (see section 4.1).

We obtain convincing localization results for the CMU query sets (figure 6 d-f). It means that our method is able to generalize well on unseen architectural structures for the depth map creation and the extraction of discriminative clue for localization. The RGBD oracle descriptor cannot be tested on CMU dataset because there are no depth maps for this dataset.

Our method shows the best localization improvement on the Oxford - Snow query sets (figure 6-b) and CMU – Snow (for encoder Alexnet, see figure 6-e). Standard image descriptors are confused by local changes caused by the snow on the scene whereas our descriptor remains confident by reconstructing the geometric structure of the scene (see figure 7, CMU-Snow 1^{st} row). Similar results should be intended regarding Oxford – Night query set (figure 6-c), however our proposal is not able to improve localization accuracy for this particular scenario. We investigate the night to day localization problem specifically in the following section.

6 Challenging localization scenarios

As mentioned previously, at first glance our method is not well designed to perform the challenging task of night to day matching. In this section, we conduct experiments in order to explain the results previously obtained and we propose an enhanced version of our descriptor performing much better on this challenging scenario.

6.1 Fine tuned descriptor

Night to day localization. Night to day localization is an extremely challenging problem: our best RGB baseline achieves a performance less than 13% recall@1. This can be explained by the huge difference in visual appearance between night and daytime images, as illustrated in figure 4. Our system should be able to improve the RGB baseline relying on the learned scene geometry, which remains the same during day and night. Unfortunately, we use training data exclusively composed of daytime images, thus making the decoder unable to reconstruct a depth map from an image taken at night. The last line of figure 8 shows the poor quality of decoder output after initial training. In order to improve the decoder's performances, we propose to use weakly annotated data to fine tune the decoder part of our system. We collect 1000 pairs of image and depth map acquired at night and retrain only decoder weights θ_G using the loss of equation (5). Figure 8 presents the qualitative improvement of the inferred depth map after the fine tuning. Notice that domain adaptation methods [83], potentially better than our fine tuning routine, could have been used to improve the quality of our depth map generated at night. However, for this experiment, we focus on the potential gain for localization permitted by the design of our system, rather than on finding the most efficient manner to adapt our method to night domain.

With the level of data annotation (*i.e* without absolute pose information) used to fine-tune our method, such post-processing trick cannot be used to improve RGB and RGB(H) image descriptors. Indeed, for the standard image descriptor and the hallucination network training we need to know the location of the night data to build images triplets with aligned pairs of anchors and positive images from the night and day domain. We believe that this type of annotated data are more complicated to gather than the one we use to finetune our system: we only need a calibrated system with synchronized acquisition of radiometric and geometric modalities, *e.g.* a stereovision system. For instance, for finetuning our model, we use a night run

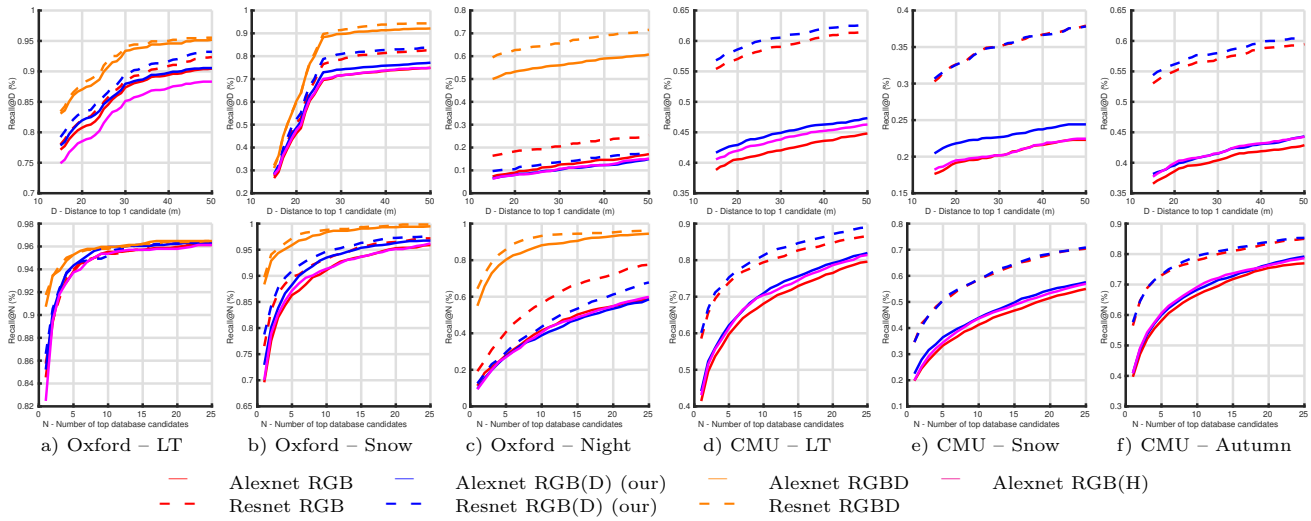

Fig. 6 Comparison of our method RGB(D) versus hallucination network RGB(H), networks trained with only images RGB and oracle descriptor RGBD using both images and depth maps at test time: we report results for backbone network encoder Resnet (- -) and Alexnet (–). Our method (in blue) is superior in every scenario facing hallucination network (in magenta). It also beats, with a significant margin, networks trained with only images (in red). All the methods failed on the very challenging night to day scenario (b). Curves best viewed in color.

from the Robotcar dataset with a low quality GPS signal, which makes impossible the automatic creation of triplets.

We show in figure 9-c that we are able to nearly double the localization performances by only fine tuning a small part of our system. Our best network achieves 23% recall@1 against 13% recall@1 for the best RGB baseline. We present some daylight images returned after the nearest neighbor search in figure 10. Even with blurry images, our method is able to extract useful geometric information to improve the matching (see figure 10, 3^{rd} row).

Impact of fine tuning on other environments. In this section, we measure the impact of the fine tuning process on other localization scenarios. Performances could decrease if our system *"forgets"* how to produce depth map from daylight images. To prevent that, we integrate half of daylight images with the night images in the training data used for fine tuning.

We show results of the fine-tuned network on figure 9. Localization accuracy remains stable after the fine tuning. We even observe slight increase in the localization performances for some scenarios (figure 9-b): thank to the fine tuning with night images, the decoder has improved the depth map generation of dark images acquired during daytime. The fact that fine tuning our system, to deal with hard localization scenarios, do not negatively impact the performances on other environment makes our new method well suited for real applications when we cannot predict what will be the outdoor conditions.

6.2 Comparison with domain adaptation

Domain adaptation method. Domain adaptation has been successfully applied to day to night image matching. We propose in this section to compare our method with the method proposed in [2]. This method consists in projecting night query images in the same domain as the reference data: daytime images. Afterwards, authors perform local image matching to estimate a precise pose of the night queries. They use ComboGAN [1], a generative adversarial network (GAN), to compute the transformation from night domain to day domain. An interesting property of this GAN model is that it does not need aligned images of the target and the source domain to be trained (data that could be costly to collect, as seen previously).

We decide to setup two experiments to compare our proposal with this domain translation approach:

- one using the night to day domain transformer GAN trained on Oxford Robotcar dataset provided by the authors,
- one with a winter to summer domain transformer GAN trained using the authors' code on CMU dataset.

For evaluation, we first transform the *challenging* domain image query into the source domain, that is the same as the reference images. Then we compute the global image descriptors and perform the similarity comparison. We setup the second experiment, winter to summer domain adaption, first in order to see how this method performs on a easier scenario than the day night localization and second to evaluate the general-

Fig. 7 Visual inspection of selected examples: we show top-1 retrieved candidate after the nearest neighbor search for the different descriptor. Red box indicates a wrong match and green box a proper one (*i.e.* retrieved image lies in 25m radius from the query ground truth position). RGB([D/R/DR]) are our descritpor trained with, respectively, depth, reflectance, depth and reflectance auxiliary modality.

ization capability of the GAN. Indeed, with this second experiment, we can use both the CMU snow and Oxford snow query sets for testing, while the GAN being trained only on CMU data. We present in figure 11 examples of images translated from source to target domain.

One drawback of domain adaptation methods is that we have to know in advance the source and target domains to apply the right transformation to the data. This constraint does not exist with our method as the depth maps is invariant to the image domain. Additionally, we have shown in section 6.1 that even if the depth maps cannot be properly recovered because of image domain shift, finetuning can be applied to improve the method without impacting other scenario (cf. figure 9).

Results. In figure 12, we show results of domain adaptation experiments. For the very challenging night to day localization scenario, the method using domain adaptation combined with RGB descriptors achieves very good results. It performs better than our finetuned method applied directly on night images, emphasizing the major role of the radiometric modality for image description. Furthermore, combining the domain adaptation

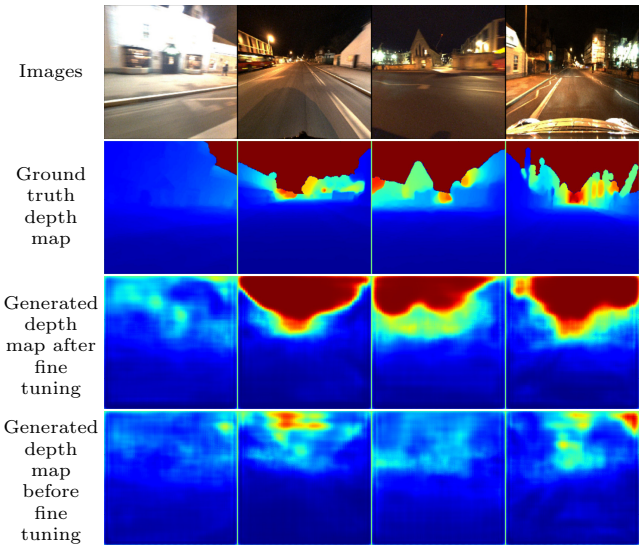

Fig. 8 Effect of fine tuning with night images on decoder output:. Decoder trained with daylight images is unable to reconstruct the scene geometry (bottom line). Fine tuning the network with less than 1000 pairs {image, depth map} acquired by night highly improves appearance of the generated depth maps. Maps best viewed in color.

pre-processing with our RGB(D) descriptor leads to the best localization results on this query set.

For the easier winter-to-summer localization scenario, despite the visually correct results obtained with the image translation model (5 last rows of figure 9), localization results are slightly worst than the one obtained without using domain adaptation. In particular, the domain adaptation pre-processing has a strong negative impact on the Snow Robotcar localization scenario. This can be explained by the poor cross dataset (train on CMU and test on Oxford) capability adaptation of the tested method.

From these experiments, we can draw the conclusion that domain adaption is very effective in *extremely* challenging localization scenario but does not handle more subtle visual changes from season changes. Furthermore, such method seems to be very sensitive to the data used for training and shows poor generalization capabilities. As a comparison, our method permits consistent improvement for various localization scenarios on different datasets and can be also combined with a domain adaptation approach.

7 Laser reflectance as side information

In this section we investigate the use of another modality replacing the depth map in order to evaluate the generalization capabilities of the proposed framework. We use lidar reflectance values as auxiliary modality for these experiments.

7.1 Laser reflectance

Lidar reflectance is defined by the proportion of the signal returned to the laser sensor after hitting an object in the scene. Reflectance characterizes the material property of an object. We use the reflectance information provided in the Robotcar dataset [46]. Reflectance values range from 0 to 1 indicating if the object has reflected from 0 to 100% of the original laser beam. We proceed the sparse reflectance data in the same manner as the depth map using inpainting algorithm from [10] to produce dense reflectance maps, and use exactly the same decoder architecture for the reflectance map and the depth map. Examples of dense reflectance map are presented in figure 13.

7.2 Reflectance versus Depth

We report in figure 14 results using reflectance map during the descriptor training (**RGB(R)**, in gray). We also illustrate in figure 7 the localization accuracy of the different methods by comparing the top-1 retrieved candidate after descriptors comparison. Localization accuracy is slightly worst when using the reflectance map than the results obtained while using the depth map. Still, reflectance information is beneficial as it increases the results over the RGB only descriptor. We can draw the conclusion that scene geometry is more informative for long term localization than reflectance property of observed objects.

We find that the reflectance side information signal enhances the image descriptor by leveraging visual clues of material with particular property: low reflectance capability (like windows, see figure 7, 2^{nd} row) or inversely very high light reflecting property (*e.g.* traffic signs, see figure 7, last row). In a different way, depth map training supervision provides interesting building shapes understanding (see the recognized tower building on figure 7, CMU - LT 2*nd* row).

7.3 Multi-modal complementarity of Reflectance and Depth

In this final experiment, we compare the performances of a single side modality training descriptor and a multiple side modalities training descriptor. We slightly modify our original system to benefit from both depth and reflectance information. The modified network is presented in figure 15. We report localization results of the three methods, depth map as side information (**RGB(D)**, in blue), reflectance map as side informa-

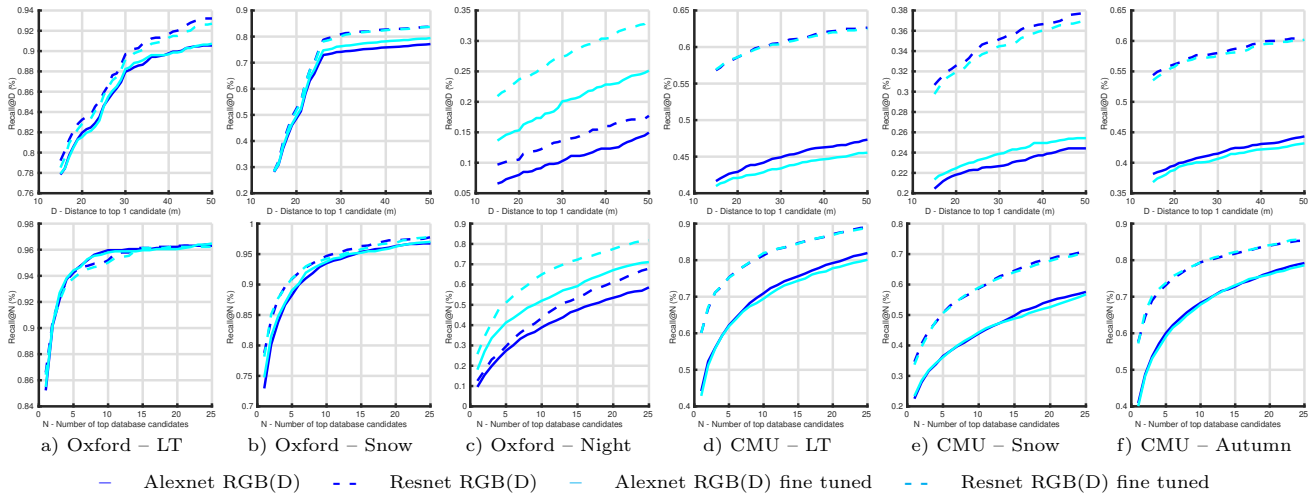

Fig. 9 Results after fine tuning: we are able to drastically improve localization performance for the Oxford – Night challenging scenario (c) by only fine tuning the decoder part of our network with weakly annotated data. Curves best viewed in color.

Fig. 10 Night to day image matching: we show top-1 retrieved candidate after the challenging night to day localizations scenario. Red box indicates a wrong match and green box a proper one (*i.e.* retrieved image lies in 25m radius from the query ground truth position). -A denotes Alexnet and -R truncated Resnet18 backbone used with NetVLAD.

tion (**RGB(R)**, in gray) and depth and reflectance map as side information (**RGB(DR)**, in green), in figure 14.

We do not observe systematic improvement when using both modalities. Nevertheless we obtain best localization results for 4 out of 5 query sets (figure 14 b, c & e). We observe that modality combination is beneficial only if each modal information performs equivalently when used alone. In other words, if one modality is a lot more informative than the other on a specific dataset (for instance depth over reflectance for the query set CMU - Snow, figure 14-d), the combination of the both will cancel potential benefit given by the most informative modality. On figure 7, we can observe successful image localization on very challenging examples: CMU - LT 1^{st} row, where the closest reference image is highly overexposed and on Oxford - Snow 1^{st} row with this very confounding image query.

Fig. 11 Examples of domain translated images: first line are the original images and bottom line are the domain-translated images. Night-to-day transfer is performed using authors' model [2] and we train our own model for winter-to-summer domain translation using images from the CMU dataset. Artifact are present in the last two rows summer images (*e.g.* vegetation on buildings), showing the poor adaptation capability of the presented model to the Oxford Robotcar dataset images.

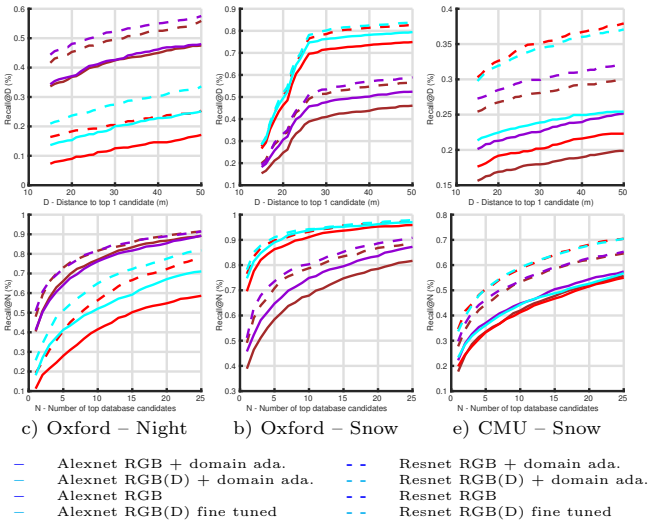

Fig. 12 Comparison with domain-adaptation method: we use two different GANs in these experiments: one for night to day domain transfer and another one for snow to summer transfer. Domain adaptation performs well for the very challenging night to day scenario and the use of our descriptor, on top of the domain adaptation pre-processing, further improves the localization performances.

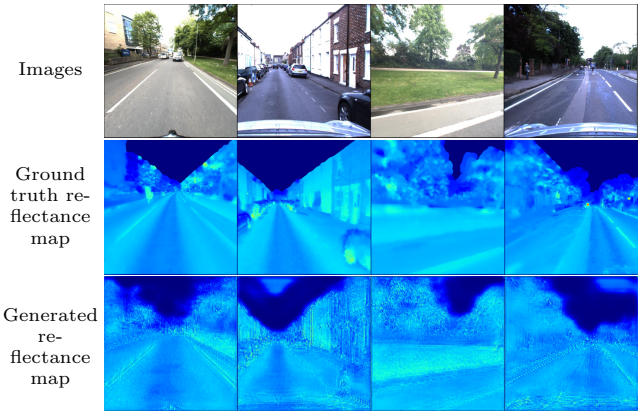

Fig. 13 Examples of dense reflectance map: the lighter the color, the higher the reflection of the material. Reflectance map highlights reflective areas, like road marking, road sign, vegetation and cars. Figure best viewed in colors.

These preliminary results concerning the use of multiple modalities during the training process of the descriptor are encouraging. Still, additional experiments have to be performed. In particular the behavior of the proposal according to the joint use of these modalities indicate that we have to focus the final descriptor fusion; modality-aware aggregation descriptor or more complex attention mechanism may be considered [73].

8 Conclusion

We have introduced a new competitive global image descriptor designed for image-based localization under challenging conditions. Our descriptor handle visual changes between images by learning the geometry of the scene. Strength of our method remains in the fact that it needs geometric information only during the learning procedure. Our trained descriptor is then used on images only. Experiments show that our proposal is much more efficient than state-of-the-art localization methods [3, 65], including methods based on side information learning [32]. Our descriptor performs especially well for challenging cross-season localization scenario, therefore it can be used to solve long-term place recognition problem. We additionally obtain encouraging results for night to day image retrieval. We also compare our approach with domain adaptation methods and we demonstrate that these two methods can be used jointly for efficient localization in a very challenging scenario. Finally, we show that our method can generalize to over auxiliary modality supervision during training. We use lidar reflectance to illustrate this generalization capability.

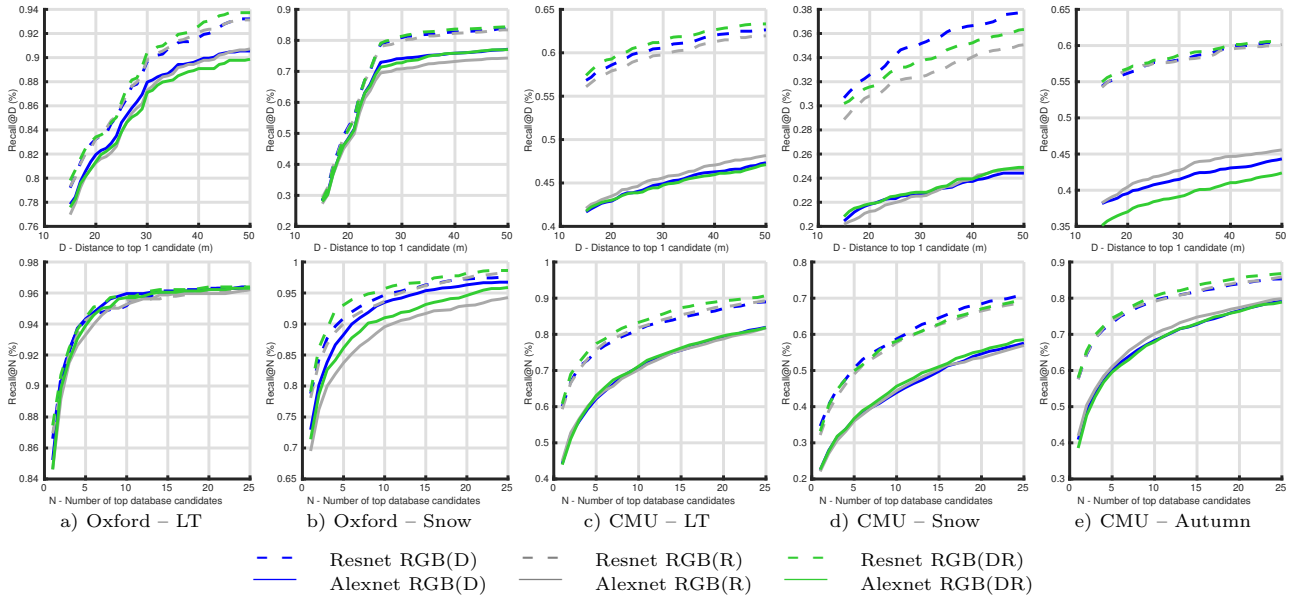

Fig. 14 Comparison of depth map and reflectance map as side information. The geometric information (in blue) remains more informative than the reflectance map (in gray) for the task of image description for localization. However, when combined (in green), depth map and reflectance map can benefit from each other and produce the most discriminative image descriptors for scenarios a, b, c & e. Curves best viewed in color.

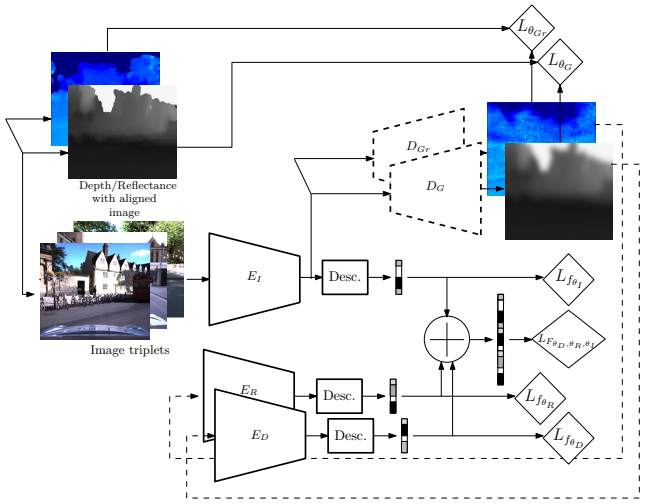

Fig. 15 Multi-modal training: we modify the training policy presented in figure 1 to handle multi-modality. Each generative modal branch (D_G and D_{Gr}) can be trained separately. Modality descriptors are trained jointly through the final triplet loss $L_{F_{\theta_D,\theta_R,\theta_I}}$.

In a future work, we will go deeper on the use of other modalities as side information sources, like semantic [73], and focusing on multi-modal fusion. We also want to study the generalization capability of our system, by considering a different image-based localization task like direct pose regression [12].

Acknowledgements We would like to acknowledge the French ANR project pLaTINUM (ANR-15-CE23-0010) for its financial support and Marco Bevilacqua for kindly sharing the code of his inpainting algorithm used in this research. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

1. Anoosheh, A., Agustsson, E., Timofte, R., and Van Gool, L. (2018). Combogan: Unrestrained scalability for image domain translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 783–790.
2. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., and Van Gool, L. (2019). Night-to-day image translation for retrieval-based localization. In *International Conference on Robotics and Automation (ICRA)*, pages 5958–5964. IEEE.
3. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2017). NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 5297–5307.
4. Arandjelović, R. and Zisserman, A. (2014). DisLocation : Scalable descriptor. In *Asian Conference on Computer Vision (ACCV)*.
5. Ardeshir, S., Zamir, A. R., Torroella, A., and Shah, M. (2014). GIS-assisted object detection and geospatial localization. In *European Conference on Com-*

puter Vision (ECCV), volume 8694 LNCS, pages 602–617.

6. Aubry, M., Russell, B. C., and Sivic, J. (2014). Painting-to-3D model alignment via discriminative visual elements. *ACM Transactions on Graphics (ToG)*, 33(2):1–14.
7. Azzi, C., Asmar, D., Fakih, A., and Zelek, J. (2016). Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization. In *British Machine Vision Conference (BMVC)*, number 2, pages 1–12.
8. Balntas, V., Riba, E., Ponsa, D., and Mikolajczyk, K. (2016). Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3.
9. Bansal, A., Badino, H., and Huber, D. (2014). Understanding how camera configuration and environmental conditions affect appearance-based localization. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 800–807.
10. Bevilacqua, M., Aujol, J. F., Biasutti, P., Brédif, M., and Bugeau, A. (2017). Joint inpainting of depth and reflectance with visibility estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 125:16–32.
11. Bhowmik, N., Weng, L., Gouet-Brunet, V., and Soheilian, B. (2017). Cross-domain Image Localization by Adaptive Feature Fusion. In *Joint Urban Remote Sensing Event (JURSE)*.
12. Brachmann, E. and Rother, C. (2018). Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
13. Cao, Y., Long, M., Wang, J., Zhu, H., and Wen, Q. (2016). Deep quantization network for efficient image retrieval. In *AAAI Conference on Artificial Intelligence*.
14. Cao, Z., Long, M., Wang, J., and Yu, P. S. (2017). Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE international conference on computer vision*, pages 5608–5617.
15. Chevalier, M., Thome, N., Hénaff, G., and Cord, M. (2018). Classifying low-resolution images by integrating privileged information in deep cnns. *Pattern Recognition Letters*, 116:29–35.
16. Christie, G., Warnell, G., and Kochersberger, K. (2016). Semantics for UGV Registration in GPS-denied Environments. *arXiv preprint*.
17. Chum, O., Mikul, A., Perdoch, M., and Matas, J. (2011). Total Recall II : Query Expansion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
18. Chum, O., Philbin, J., Sivic, J., Isard, M., and Zisserman, A. (2007). Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE International Conference on Computer Vision (ICCV)*.
19. Deng, C., Chen, Z., Liu, X., Gao, X., and Tao, D. (2018). Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903.
20. Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1–9.
21. Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust RGB-D object recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, volume 2015-Decem, pages 681–687.
22. Garg, S., Suenderhauf, N., and Milford, M. (2018a). Don't Look Back: Robustifying Place Categorization for Viewpoint- and Condition-Invariant Place Recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*.
23. Garg, S., Suenderhauf, N., and Milford, M. (2018b). LoST? Appearance-Invariant Place Recognition for Opposite Viewpoints using Visual Semantics. *Robotics Science and Systems (RSS)*.
24. Germain, H., Bourmaud, G., and Lepetit, V. (2018). Efficient Condition-based Representations for Long-Term Visual Localization. *arXiv preprint*.
25. Germain, H., Bourmaud, G., and Lepetit, V. (2019). Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. *arXiv preprint*.
26. Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
27. Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2016). Deep Image Retrieval: Learning Global Representations for Image Search. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 241–257.
28. Gordo, A., Almazán, J., Revaud, J., and Larlus, D. (2017). End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision (IJCV)*, 124(2):237–254.
29. Gupta, S., Girshick, R., Arbeláez, P., and Malik, J. (2014). Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision (ECCV)*, volume 8695 LNCS, pages 345–360.
30. Hays, J. and Efros, A. A. (2008). IM2GPS: Estimating Geographic Information From a Single Image.

In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 05.

31. Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
32. Hoffman, J., Gupta, S., and Darrell, T. (2016). Learning with Side Information through Modality Hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834.
33. Iscen, A., Tolias, G., Avrithis, Y., and Chum, O. (2018). Mining on Manifolds: Metric Learning without Labels.
34. Isola, P., Zhu, J.-Y. Y., Zhou, T., and Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134.
35. Jégou, H., Douze, M., and Schmid, C. (2009). On the burstiness of visual elements. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1169–1176.
36. Jiang, Q.-Y. and Li, W.-J. (2017). Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240.
37. Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.
38. Johnson, J., Douze, M., and Jégou, H. (2017). Billion-scale similarity search with GPUs. *arXiv preprint*.
39. Kim, H. J., Dunn, E., and Frahm, J.-M. (2017). Learned Contextual Feature Reweighting for Image Geo-Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
40. Lai, H., Pan, Y., Liu, Y., and Yan, S. (2015). Simultaneous feature learning and hash coding with deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3270–3278.
41. Li, W., Chen, L., Xu, D., and Van Gool, L. (2018). Visual Recognition in RGB Images and Videos by Learning from RGB-D Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(8):2030 2036.
42. Liu, L., Li, H., and Dai, Y. (2018). Deep Stochastic Attraction and Repulsion Embedding for Image Based Localization. *arXiv preprint*.
43. Long, M., Cao, Y., Cao, Z., Wang, J., and Jordan, M. I. (2018). Transferable representation learning with deep adaptation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3071–3085.
44. Loo, S. Y., Amiri, A. J., Mashohor, S., Tang, S. H., and Zhang, H. (2019). CNN-SVO: Improving the Mapping in Semi-Direct Visual Odometry Using Single-Image Depth Prediction. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 1.
45. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P., and Milford, M. J. (2016). Visual Place Recognition: A Survey. *IEEE Transactions on Robotics (TRO)*, 32(1):1–19.
46. Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016). 1 year, 1000 km: The Oxford RobotCar dataset. *The International Journal of Robotics Research (IJRR)*.
47. Mahjourian, R., Wicke, M., and Angelova, A. (2018). Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
48. Milford, M. J. and Wyeth, G. F. (2012). SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1643–1649.
49. Morago, B., Bui, G., and Duan, Y. (2016). 2D Matching Using Repetitive and Salient Features in Architectural Images. *IEEE Transactions on Image Processing (ToIP)*, 7149(c):1–12.
50. Mordan, T., Thome, N., Henaff, G., and Cord, M. (2018). Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In *Advances in Neural Information Processing Systems*, pages 1310–1322.
51. Muja, M. and Lowe, D. G. (2009). Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 1–10.
52. Naseer, T., Burgard, W., and Stachniss, C. (2018). Robust Visual Localization Across Seasons. *IEEE Transactions on Robotics (TRO)*, 34(2):289–302.
53. Naseer, T., Oliveira, G. L., Brox, T., and Burgard, W. (2017). Semantics-aware Visual Localization under Challenging Perceptual Conditions. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2614–2620.
54. Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3):145–175.

55. Paulin, M., Mairal, J., Douze, M., Harchaoui, Z., Perronnin, F., and Schmid, C. (2017). Convolutional Patch Representations for Image Retrieval: An Unsupervised Approach. *International Journal of Computer Vision (IJCV)*, 121(1):149–168.
56. Philbin, J., Chum, O., Isard, M., Sivic, J., and Zisserman, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
57. Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2018). A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109.
58. Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2019a). Geometric Camera Pose Refinement With Learned Depth Maps. In *IEEE International Conference on Image Processing (ICIP)*.
59. Piasco, N., Sidibé, D., Demonceaux, C., and Gouet-Brunet, V. (2019b). Perspective-n-Learned-Point: Pose Estimation from Relative Depth. In *British Machine Vision Conference (BMVC)*.
60. Piasco, N., Sidibé, D., Gouet-Brunet, V., and Demonceaux, C. (2019c). Learning Scene Geometry for Visual Localization in Challenging Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*.
61. Porav, H., Bruls, T., and Newman, P. (2019). I Can See Clearly Now : Image Restoration via De-Raining. In *IEEE International Conference on Robotics and Automation (ICRA)*.
62. Porav, H., Maddern, W., and Newman, P. (2018). Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer. *IEEE International Conference on Robotics and Automation (ICRA)*.
63. Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2016). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
64. Radenović, F., Tolias, G., and Chum, O. (2016). CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 3–20.
65. Radenović, F., Tolias, G., and Chum, O. (2017). Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
66. Russell, B. C., Sivic, J., Ponce, J., and Dessales, H. (2011). Automatic alignment of paintings and photographs depicting a 3D scene. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*.
67. Sarlin, P.-E., Cadena, C., Siegwart, R., and Dymczyk, M. (2019). From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
68. Sarlin, P.-E., Debraine, F., Dymczyk, M., Siegwart, R., and Cadena, C. (2018). Leveraging Deep Visual Descriptors for Hierarchical Efficient Localization. In *Conference on Robot Learning (CoRL)*, pages 1–10.
69. Sattler, T., Havlena, M., Schindler, K., and Pollefeys, M. (2016). Large-Scale Location Recognition and the Geometric Burstiness Problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
70. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., and Pajdla, T. (2018a). Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8601–8610.
71. Sattler, T., Maddern, W., Torii, A., Sivic, J., Pajdla, T., Pollefeys, M., and Okutomi, M. (2018b). Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
72. Schönberger, J. L., Pollefeys, M., Geiger, A., and Sattler, T. (2018). Semantic Visual Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
73. Seymour, Z., Sikka, K., Chiu, H.-p., Samarasekera, S., and Kumar, R. (2018). Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization. *arXiv preprint*.
74. Sharmanska, V., Quadrianto, N., and Lampert, C. H. (2013). Learning to rank using privileged information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 825–832.
75. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., and Fitzgibbon, A. (2013). Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937.
76. Sizikova, E., Singh, V. K., Georgescu, B., Halber, M., Ma, K., and Chen, T. (2016). Enhancing Place Recognition using Joint Intensity - Depth Analysis and Synthetic Data. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 1–8.
77. Stenborg, E., Toft, C., and Hammarstrand, L. (2018). Long-term Visual Localization using Semantically Segmented Images. *arXiv preprint*.

78. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. J. (2015). Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free. In *Robotics Science and Systems (RSS)*.
79. Tateno, K., Tombari, F., Laina, I., and Navab, N. (2017). CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
80. Toft, C., Stenborg, E., Hammarstrand, L., Brynte, L., Pollefeys, M., Sattler, T., and Kahl, F. (2018). Semantic Match Consistency for Long-Term Visual Localization. In *European Conference on Computer Vision (ECCV)*.
81. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., and Pajdla, T. (2015). 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
82. Torii, A., Sivic, J., Okutomi, M., and Pajdla, T. (2013). Visual Place Recognition with Repetitive Structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 37, pages 2346–2359.
83. Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176.
84. Uy, M. A. and Lee, G. H. (2018). PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
85. Vapnik, V. and Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557.
86. Xu, D., Ouyang, W., Ricci, E., Wang, X., and Sebe, N. (2017). Detection, Learning cross-modal deep representations for robust pedestrian. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5363–5371.
87. Zamir, A. R. and Shah, M. (2010). Accurate image localization based on google maps street view. In *European Conference on Computer Vision (ECCV)*, volume 6314 LNCS, pages 255–268.
88. Zamir, A. R. and Shah, M. (2014). Image geolocalization based on multiplenearest neighbor feature matching using generalized graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1546–1558.
89. Zwald, L. and Lambert-Lacroix, S. (2012). The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*.