HAL
open science

Learning Scene Geometry for Visual Localization in Challenging Conditions

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

To cite this version:

Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, Cedric Demonceaux. Learning Scene Geometry for Visual Localization in Challenging Conditions. International Conference on Robotics and Automation, ICRA 2019, May 2019, Montréal, Canada. pp.9094-9100, 10.1109/ICRA.2019.8794221 . hal-02057378

HAL Id: hal-02057378

https://hal.science/hal-02057378v1

Submitted on 5 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning Scene Geometry for Visual Localization in Challenging Conditions

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

***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.**

We are able to increase recall@1 performances by 2.15% on cross-weather and long-term localization scenario and by 4.24% points on a challenging winter/summer localization sequence versus state-of-the-art methods. Our method can also use weakly annotated data to localize night images across a reference dataset of daytime images.

I. Introduction

Visual-Based Localization (VBL) is a central topic in robotics and computer vision applications [1]. It consists in retrieving the location of a visual query according to a known absolute reference. VBL is used in many applications such as autonomous driving, augmented reality, robot navigation or SLAM loop closing. In this paper, we address VBL as an image retrieval problem where an input image is compared to a reference pool of localized images. This image-retrieval-like problem is two-step: descriptor computation for both the query and the reference images and similarity association across the descriptors. Since the reference images are associated to a location, by ranking images according to their similarity scores we obtain an approximate location for the query. Numerous works have introduced image descriptors well suited for image retrieval for localization [2], [3], [4], [5], [6].

One of the main challenges of image-based localization remains the mapping of images acquired under changing conditions: cross-season images matching [7], long-term localization [8], day to night place recognition [9], etc. Recent approaches use complementary information in order to address these visually challenging localization scenarios (geometric information through point cloud [10], [11] or depth maps [12], semantic information [13], [12], [7]). However geometric or semantic information are not always available, especially in robotic applications when the sensors or the computational load on the robot are limited.

In this paper, we propose a image descriptor that learns, from an image, the corresponding scene geometry, in order to deal with challenging outdoor large-scale image-based localization scenarios. We introduce geometric information during the training step to make our new descriptor robust to visual changes that occur between images taken at different times. Once trained, our system is only used on images to construct a expressive descriptor for image retrieval. This kind of system design is also known as side information learning [14], as it uses geometric and radiometric information only during the training step and just radiometric data for the image localization. Our method is especially well-suited for robotic long-term localization when the perceptive sensor on the robot is limited to a camera [15], while having access to the full scene geometry off-line [16], [17], [18].

The paper is organized as follows. In section II, 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 III, we describe in detail our new image descriptor trained with side depth information. We illustrate the effectiveness of the proposed method on four challenging scenarios in section IV. Section V finally concludes the paper.

II. Related Work

Image descriptor for outdoor visual localization. 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 or VLAD. Specific features re-weighting scheme dedicated to image localization have been introduced in [19]. Authors of [20] introduce a re-ranking routine to improve the localization performances on large-scale outdoor area. More recently, [2] introduces NetVLAD, a convolutional neural network that is trained to learn a well-suited image representation for image localization. Numerous other CNN image descriptors have been proposed in the literature [3], [4], [5], [21], [6] and achieve state of the art results in image retrieval for localization. Therefore we use CNN image descriptors as base component in our system.

Localization in challenging condition. In order to deal with visual changes in images taken at different times, [22] uses a combination of handcrafted and learned descriptors. [23] introduces temporal consistency by using a sequence of images, while in our proposal we use only one image as input for our descriptor. In [24], authors synthesize new images to match the appearance of reference images, for instance they synthesized daytime images from night images. Numerous works [25], [8], [7] enhance their visual descriptors by

[1] ImViA-VIBOT, ERL CNRS 6000, Université Bourgogne Franche-Comté

[2] Univ. Paris-Est, LaSTIG MATIS, IGN, ENSG, F-94160 Saint-Mandé, France

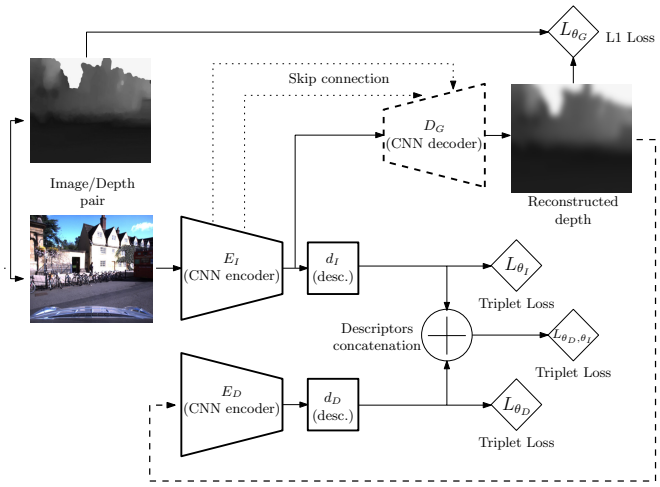

Fig. 1. **Image descriptors training with auxiliary depth data (our work):** two encoders are used for extracting deep features map from the main image modality and the auxiliary reconstructed depth map (inferred from our deep decoder). These features are used to create intermediate descriptors that are finally concatenated in one final image descriptor.

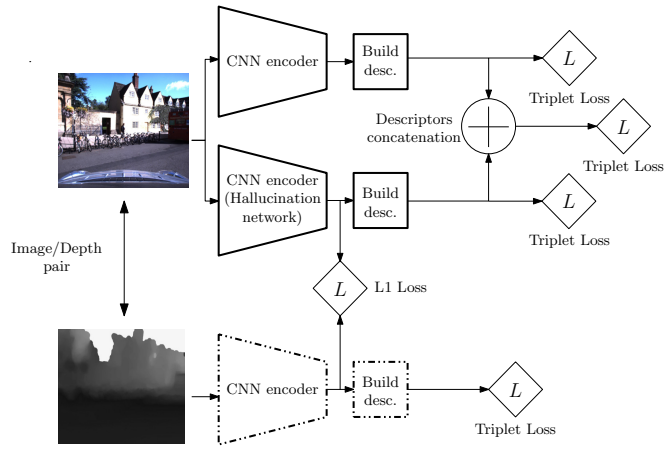

Fig. 2. **Hallucination network for image descriptors learning:** we train an hallucination network, inspired from [14], for the task of global image description. Unlike the proposed method (see figure 1), hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the depth map itself.

adding semantic information. Although semantic representation is robust for long term localization, it may be costly to obtain. Other methods rely on geometric information like point clouds [10], [11], or 3D structures [9]. Geometric information has the advantage of remaining more stable across time comparing to visual information but 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.

Learning with side information. Recent work from [26] casts the side information learning problem as a domain adaptation problem, where source domain includes multiples modalities and the target domain is composed of a single modality. Another successful method have been introduced in [14]: 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 [27], uses recreated thermal images to improve pedestrian detection on standard images only. Our system, inspired by [27], learns how to produce depth maps from images to enhance the description of these images.

III. Method

A. Overview

We design a new global image description for the task of image-based localization. We first extract dense feature maps from an input image with a convolutional neural network encoder (E_I). These feature maps are subsequently used to build a compact representation of the scene (d_I). State-of-the-art features aggregation methods can be used to construct the image descriptor, such as MAC [5] or NetVLAD [2]. We enhance this standard image descriptor with side depth map information that is only available during the training process. To do so, a deep fully convolutional neural network decoder (D_G) is used to reconstruct the corresponding depth map according to the input image. The reconstructed depth is then used to extract a global depth map descriptor. We follow the same procedure used before: we extract deep feature maps with an encoder (E_D) before building the descriptor (d_D). Finally, the image descriptor and the depth map descriptor are L_2 normalized to be concatenated into a single global descriptor. Figure 1 summarizes the whole process of our method. Once trained with geometric and radiometric information, the proposed method is used on images only, to create a descriptor tuned for image localization.

B. Training routine

Trainable parameters are θ_I the weights of encoder and descriptor $\{E_I, d_I\}$, θ_D the weights of the 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 [2]. A triplet $\{q_{im}, q_{im}^+, q_{im}^-\}$ is composed of an anchor image q_{im}, a positive example q_{im}^+ representing the same scene as the anchor and an unrelated negative example q_{im}^-. The first triplet loss acting on $\{E_I, d_I\}$ is:

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

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

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

where $E_I(x_{im})$ represents the deep feature maps extracted by the decoder 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), $L_{f_{\theta_D}}(\hat{q}_{depth}, \hat{q}_{depth}^+, \hat{q}_{depth}^-)$, where $f_{\theta_D}(x_{depth})$ is the global

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

$$\hat{x}_{depth} = D_G(E_I(x_{im})). \tag{3}$$

Decoder D_G uses the deep representation of image x_{im} 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 $L_{F_{\theta_I,\theta_D}}(q_{im}, q^+_{im}, q^-_{im})$, where $F_{\theta_I,\theta_D}(x_{im})$ denotes the concatenation of image descriptor and depth map descriptor: $F_{\theta_I,\theta_D}(x_{im}) = [f_{\theta_I}(x_{im}), f_{\theta_D}(\hat{x}_{depth})]$.

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

$$L_{\theta_G} = \|x_{depth} - \hat{x}_{depth}\|_1 . \tag{4}$$

The whole system is trained according to the following constraints:

$$(\theta_I, \theta_D) := arg \min_{\theta_I,\theta_D} \left[L_{f_{\theta_I}} + L_{f_{\theta_D}} + L_{F_{\theta_I,\theta_D}} \right], \tag{5}$$

$$(\theta_G) := arg \min_{\theta_G} \left[L_{\theta_G} \right]. \tag{6}$$

We use two different optimizers: one updating θ_I and θ_D weights regarding constraint (5) and the other updating θ_G weights regarding constraint (6). 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.

C. Hallucination network for image description

We compare our method of side information learning with a state-of-the-art approach system, named hallucination network [14]. The hallucination network is originally designed for object detection and classification in images. We adapt the work of [14] to create an image descriptor system that benefits from depth map side modality during training. Like our proposal, the trained hallucination network is used on images only and produce a global descriptor for image localization. The system is presented in figure 2. The main difference with our proposal is that the hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the deep map itself. We refer readers to [14] for more information about the hallucination network.

D. Advantages and drawbacks

One advantage of the hallucination network over our proposal is that it does not require a decoder network, resulting on 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 do not need such initialization. Training the hallucination network requires more complex data than the proposed method. Indeed, it needs to gather triplets of image, and depth map pairs, which require to know the absolute position of the data [2], [6], or to use costly algorithms like Structure from Motion (SfM) [28], [5], [3].

One advantage of our method over the hallucination approach is that we have two unrelated objectives during training: learning a efficient image representation for localization and learning how to reconstruct scene geometry from an image. It means 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 part IV-C.

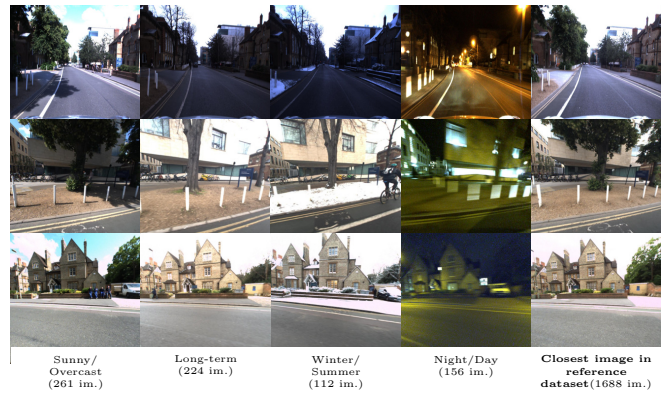

Fig. 3. **Examples of test images :** we evaluate our proposal on four challenging localization sequences. The number under the query set name indicates the amount of query images to compare against the 1688 reference images.

IV. Experiments

A. Dataset

We have tested our new method on the *Oxford Robotcar* public dataset [17]. This is a common dataset used for image-based localization [10] and loop closure algorithm involving neural networks training [24].

Training data. We use the temporal redundancy present in the dataset to build the images triplets to train our CNN. We build 400 triplets using three runs acquired at dates: `2015-05-19`, `2015-08-28` and `2015-11-10`. We selected an area of the city different from the one used for training our networks for validation. Depth modality is extracted from the lidar point cloud dataset of *Oxford Robotcar*. 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 inpainting algorithm from [29] in order to make them dense.

Testing data. We propose four testing scenarios on the same spatial area (different from the area used for training and validation). The reference dataset is composed of 1688 images taken every 5 meters along a path of 2 km, when the weather was overcast. The four query sets are:

- Sunny/Overcast: queries have been acquired during a sunny day.
- Long-term: queries have been acquired 7 months after the reference images under similar weather conditions.
- Winter/Summer: queries have been acquired during a snowy day.
- Night/Day: queries have been acquired at night, resulting in radical visual changes compared to the reference images.

Query examples are presented in figure 3.

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

a) 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 inside the $25m$ radius of the ground truth query position.

b) 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 [30]. This metric qualifies the accuracy of the localization system.

B. Implementation details

Our proposal is implemented by 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 triplet loss 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 [5]. Images and depth maps are re-sized to 224×224 pixels before training and testing.

Encoder architectures. We test the fully convolutional part of Alexnet and Resnet18 architectures for features extraction. The size of the final features block is $256 \times 13 \times 13$ for Alexnet and $512 \times 7 \times 7$ for Resnet. Initial weights are the ones obtained by training the whole network on ImageNet dataset. 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, we have found that using deeper network does not significantly improve localization results.

Descriptor architectures. We test the two state-of-the-art image descriptors MAC [5] and NetVLAD [2]. MAC is a simple global pooling method that takes the maximum of each feature map from the encoder output. NetVLAD 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.

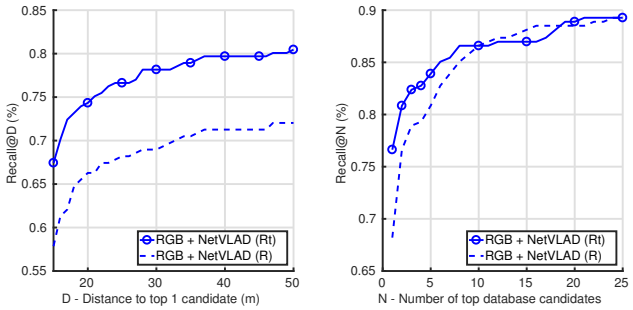

Fig. 4. **Resnet18 (R) versus truncated Resnet18 (Rt) in combination with NetVLAD pooling:** we show the importance of the spatial resolution of the deep feature maps of the encoder used with NetVLAD layer. The truncated version of Resnet18, more than two times lighter than the complete one, achieves much better localization results.

Decoder architecture. The decoder used in our proposal is based on Unet architecture and inspired by network generator from [31]. Dimension up-sampling is performed through inverse-convolutions layers. Decoder weights are initialized randomly.

C. Results

Baselines. We compare our method with two state-of-the-art baselines:

*a) RGB only (**RGB**):* simple networks composed of encoder + descriptor trained with only images, without side depth maps information. We evaluate 4 variants of networks, by combining Alexnet (A) or Resnet (R) encoder with MAC or NetVLAD descriptor pooling.

*b) RGB with Depth side information (**RGBtD**):* networks that use pairs of aligned image and depth map during training step and images only at test time. We compare our proposal with our version of hallucination network [14] (hall). We follow training procedure of [14] to train the hallucination network, whereas our proposal is trained as explained in III-B.

Truncated Resnet. We experimented that NetVLAD descriptor combined with Resnet architecture, RGB + NetVLAD (R), does not perform well. NetVLAD can be view as a pooling method that acts on local deep features densely extracted from the input image. We argue that the spatial resolution of the features block obtained with Resnet encoder is too low compared to the other architecture (for instance 13×13 for Alexnet compared to 7×7 for Resnet for an 224×224 input image). We propose a truncated version of Resnet encoder (Rt), created by drooping the end of the network after the 13th convolutional layer. Thus we obtain a feature block with greater spatial resolution: $256 \times 14 \times 14$ compared to $512 \times 7 \times 7$. Recall results on the *Sunny/Overcast* query set for both architectures are presented in figure 4. As

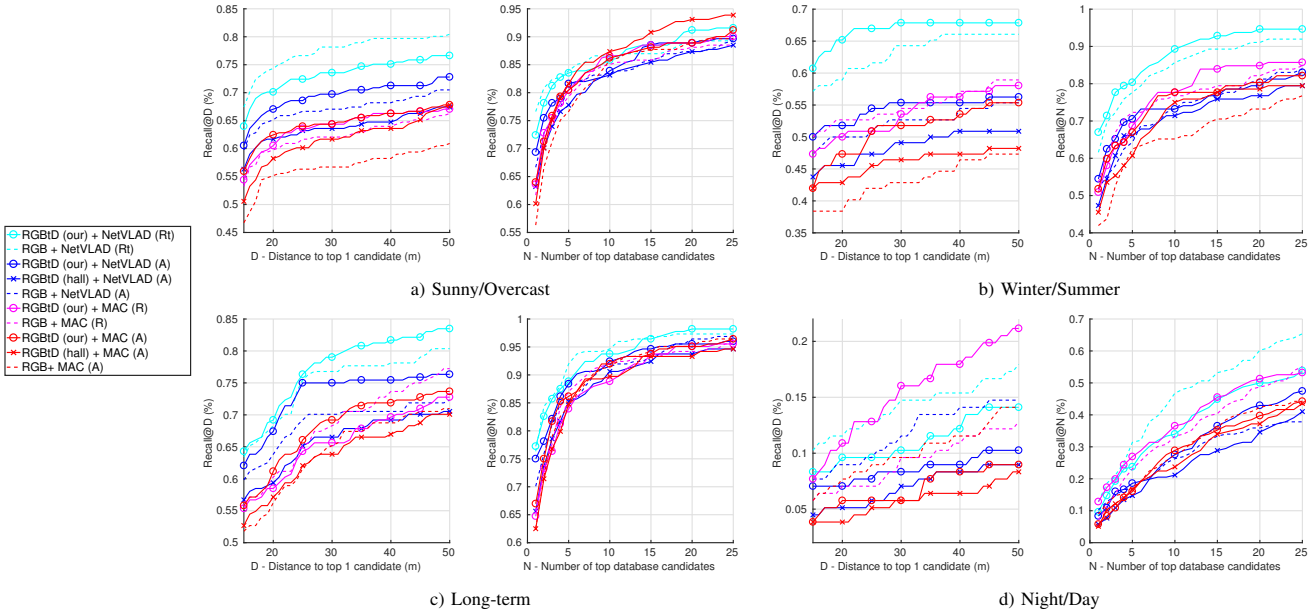

Fig. 5. **Comparison of our method versus hallucination network and networks trained with only images**: our method (-o-) is superior in almost every scenario facing hallucination network (-x-). It also beats, with a significant margin, networks trained with only images (−−). NetVLAD descriptors (blue and cyan curves) are superior to MAC (red and magenta curves), specially in terms of accuracy (Recall@D curve). Night/day dataset remains the most challenging one. Curves best viewed in colors.

TABLE I

RESULTS TOP-1 RECALL @D: MAC (A) & NETVLAD (RT).

Methods		Sunny			Long-term			Winter			All		
		@15	@25	@50	@15	@25	@50	@15	@25	@50	@15	@25	@50
RGB	MAC	46.7	56.3	60.9	51.8	62.5	71.0	38.4	42.0	47.3	45.6	53.6	59.7
	NetVLAD	67.4	76.6	80.5	63.4	76.3	80.4	57.1	61.6	66.1	62.6	71.5	75.6
Mean (all RGB)		56.8	65.3	69.4	57.8	68.7	75.2	48.2	51.8	56.9	**54.3**	**61.9**	**67.2**
Our	MAC	55.9	64.0	67.8	55.8	67.0	73.7	42.0	51.8	55.4	51.2	60.9	65.6
	NetVLAD	64.0	72.4	76.6	64.3	77.2	83.5	60.7	67.0	67.9	63.0	72.2	76,0
Mean (all our)		58.7	67.3	71.1	59.4	71.0	76.6	50.0	56.0	59.4	**56.0**	**64.8**	**69.0**

the truncated version of Resnet encoder clearly dominates the full one, we use the truncated version for the following experiments.

Discussion. Localization results on the four query sets are presented in figure 5. Both methods trained with auxiliary depth information (hall and our) 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. Compare to hallucination network, our method shows better results, both in term 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.

We also report on table I localization performances for the 3 daytime datasets (sunny, long-term and winter). We obtain our best localization results by combining truncated Resnet encoder with NetVLAD descriptor. However, for all combination of encode/decoder, our method increases the localization precision compare to the RGB baseline. This demonstrate the generalization capability of our method: we can either use lightweight architecture for online embedded localization or rely on greedier models to increase the overall localization precision. Our method only decreases the localization performances compare to the baseline when using Resnet+NetVLAD on the Sunny/Overcast query set. This is certainly because the training data are visually similar to the queries present in this scenario. It will be interesting to introduce attention mechanism to balance the relative importance of image and depth modality to overcome this limitation.

Our method shows the best localization improvement on the Winter/Summer query set. 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. Similar results should be intended regarding Night/Day query set (figure 5-d), however our proposal is not able to improve localization accuracy for this particular scenario. We investigate the night to day localization problem in the following.

Night to day localization. Night to day localization is an extremely challenging problem: the best RGB baseline achieves 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 3. 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 6 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

TABLE II

CONTRIBUTION OF THE DEPTH INFORMATION DURING TRAINING.

Query set	Network		Top-1 recall@D			Recall@N	
	Name	#Param.	@15	@30	@50	@1	@5
Sunny/ Overcast	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
	RGBtD + MAC	7.9M	**55.9**	**64.4**	**67.8**	**64.0**	**80.5**
Long-term	RGB + MAC	2.5M	51.8	65.2	71.0	62.5	84.4
	RGB$^+$ + MAC	7.9M	54.5	68.3	72.3	**67.0**	82.6
	RGBtD + MAC	7.9M	**55.8**	**69.2**	**73.7**	**67.0**	**86.2**
Winter/ Summer	RGB + MAC	2.5M	38.4	43.0	47.3	42.0	62.5
	RGB$^+$ + MAC	7.9M	36.6	42.0	43.0	41.1	56.3
	RGBtD + MAC	7.9M	**42.0**	**51.8**	**55.4**	**51.8**	**67.0**

1000 pairs of image and depth map acquired at night and retrain only decoder weights θ_G using loss of equation (4). Figure 6 presents the qualitative amelioration on the inferred depth map after the fine tuning. Such post-processing trick cannot be used to improve standard RGB image descriptors, because we need to know the location of the night data. For instance, we use a night run from the Robotcar dataset with a low quality GPS signal, that makes impossible the automatic creation of triplets that are essential for training a deep image descriptor. We show in figure 7 that we are able to nearly multiply by two 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.

Contribution of the depth information. In this paragraph, we investigate the impact on localization performances provided by the side geometry 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 of this experiment with encoder architecture Alexnet are presented in table II.

Increasing the size of the system results in a better localization on the two easiest query sets. Surprisingly RGB$^+$ system decreases localization performances on the winter queries compared to RGB. The system has probably overfitted on the training data that are visually close to queries of "Sunny" set and "Long-term" set, but quiet different from the queries of "Winter" set (see figure 3). Our RGBtD + 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.

V. 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 [2], [5], including methods based on side information learning [14]. 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.

In a future work we will investigate the use of other modalities as side information sources, like the reflectance factor provided by lidars. We also want to study the generalization capability of our system, by considering a different image-based localization task like direct pose regression [32].

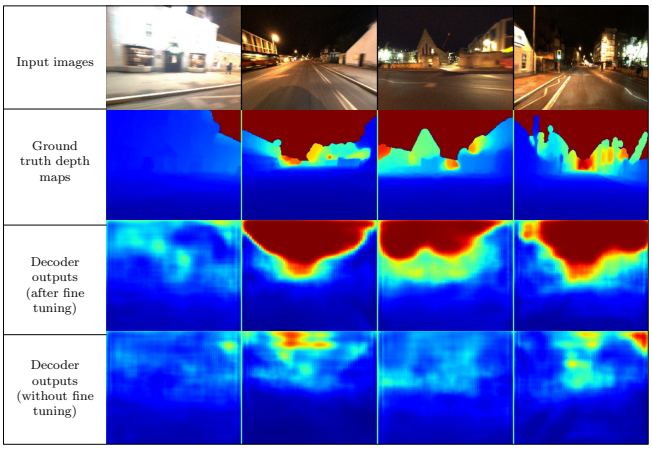

Fig. 6. **Effect of fine tuning with night images on decoder output:**. Decoder trained with daylight images is unable to reconstruct the scene geometry (bottom line). Fine tuning the network with less than 1000 pairs {image, depth map} acquired by night highly improves appearance of the generated depth maps. Maps best viewed in color.

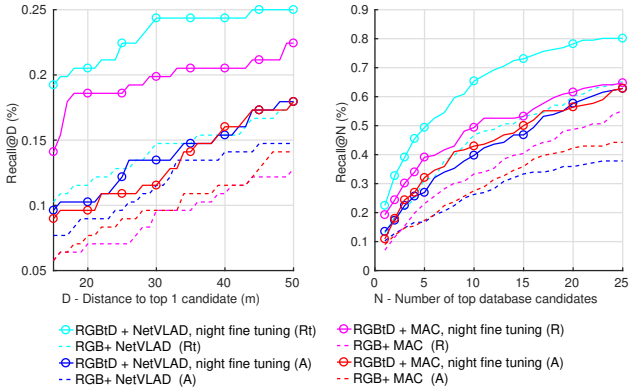

Fig. 7. **Results on Night/Day query set after fine tuning:** we are able to drastically improve localization performance for the Night/Day challenging scenario by only fine tuning the decoder part of our network with weakly annotated data. Curves best viewed in color.

ACKNOWLEDGMENTS

We would like to acknowledge the French ANR project pLaTINUM (ANR-15-CE23-0010) for its financial support. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet, "A survey on Visual-Based Localization: On the benefit of heterogeneous data," *Pattern Recognition*, vol. 74, pp. 90–109, feb 2018. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0031320317303448

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 5297–5307, 2017. [Online]. Available: http://arxiv.org/abs/1511.07247

[3] H. J. Kim, E. Dunn, and J.-M. Frahm, "Learned Contextual Feature Reweighting for Image Geo-Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "End-to-End Learning of Deep Visual Representations for Image Retrieval," *International Journal of Computer Vision (IJCV)*, vol. 124, no. 2, pp. 237–254, 2017.

[5] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN Image Retrieval with No Human Annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. [Online]. Available: http://arxiv.org/abs/1711.02512

[6] L. Liu, H. Li, and Y. Dai, "Deep Stochastic Attraction and Repulsion Embedding for Image Based Localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018. [Online]. Available: https://arxiv.org/pdf/1808.08779.pdf

[7] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard, "Semantics-aware Visual Localization under Challenging Perceptual Conditions," *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, pp. 2614–2620, 2017.

[8] C. Toft, E. Stenborg, L. Hammarstrand, L. Brynte, M. Pollefeys, T. Sattler, and F. Kahl, "Semantic Match Consistency for Long-Term Visual Localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 2018.

[9] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi, "Benchmarking 6DOF Urban Visual Localization in Changing Conditions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1707.09092

[11] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic Visual Localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1712.05773

[12] G. Christie, G. Warnell, and K. Kochersberger, "Semantics for UGV Registration in GPS-denied Environments," *arXiv preprint*, 2016. [Online]. Available: http://arxiv.org/abs/1609.04794

[13] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah, "GIS-assisted object detection and geospatial localization," in *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, vol. 8694 LNCS, no. PART 6, 2014, pp. 602–617.

[14] J. Hoffman, S. Gupta, and T. Darrell, "Learning with Side Information through Modality Hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 826–834. [Online]. Available: http://ieeexplore.ieee.org/document/7780465/

[15] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-DOF localization on mobile devices," *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, vol. 8690 LNCS, no. PART 2, pp. 268–283, 2014.

[16] N. Paparoditis, J.-P. Papelard, B. Cannelle, A. Devaux, B. Soheilian, N. David, and E. Houzay, "Stereopolis II: A multi-purpose and multi-sensor 3D mobile mapping system for street visualisation and 3D metrology," *Revue française de photogrammétrie et de télédétection*, vol. 200, no. 1, pp. 69–79, 2012.

[17] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research (IJRR)*, 2016.

[18] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "TorontoCity: Seeing the World with a Million Eyes," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. [Online]. Available: http://arxiv.org/abs/1612.00423

[19] R. Arandjelović and A. Zisserman, "DisLocation : Scalable descriptor," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2014.

[20] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-Scale Location Recognition and the Geometric Burstiness Problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] N. Sünderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. J. Milford, "Place Recognition with ConvNet Landmarks: Viewpoint-Robust, Condition-Robust, Training-Free," in *Robotics Science and Systems (RSS)*, 2015.

[22] T. Naseer, W. Burgard, and C. Stachniss, "Robust Visual Localization Across Seasons," *IEEE Transactions on Robotics (TRO)*, vol. 34, no. 2, pp. 289–302, 2018.

[23] S. Garg, N. Suenderhauf, and M. Milford, "Don't Look Back: Robustifying Place Categorization for Viewpoint- and Condition-Invariant Place Recognition," in *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, 2018. [Online]. Available: http://arxiv.org/abs/1801.05078

[24] H. Porav, W. Maddern, and P. Newman, "Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer," in *Proceedings of the IEEE International Conference of Robotics and Automation (ICRA)*, 2018. [Online]. Available: http://arxiv.org/abs/1803.03341

[25] E. Stenborg, C. Toft, and L. Hammarstrand, "Long-term Visual Localization using Semantically Segmented Images," *arXiv*, 2018. [Online]. Available: http://arxiv.org/abs/1801.05269

[26] W. Li, L. Chen, D. Xu, and L. Van Gool, "Visual Recognition in RGB Images and Videos by Learning from RGB-D Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 8, p. 2030 2036, 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8000401/

[27] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://arxiv.org/abs/1609.03677

[29] M. Bevilacqua, J. F. Aujol, P. Biasutti, M. Brédif, and A. Bugeau, "Joint inpainting of depth and reflectance with visibility estimation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 125, pp. 16–32, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.isprsjprs.2017.01.005

[30] A. R. Zamir and M. Shah, "Image geo-localization based on multiplenearest neighbor feature matching using generalized graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 36, no. 8, pp. 1546–1558, 2014.

[31] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[32] E. Brachmann and C. Rother, "Learning Less is More - 6D Camera Localization via 3D Surface Regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [Online]. Available: http://arxiv.org/abs/1711.10228