Perspective-n-Learned-Point: Relative Pose Estimation at Absolute Scale

BMVC 2019 Submission # 981

Abstract

In this paper we present an online camera pose estimation method that combines Content-Based Image Retrieval (CBIR) and pose refinement based on a learned representation of the scene geometry extracted from monocular images. Our pose estimation method is two-step, we first retrieve an initial 6 Degrees of Freedom (DoF) location of an unknown-pose query by retrieving the most similar candidate in a pool of geo-referenced images. In a second time, we refine the query pose with a Perspective-n-Point (PnP) algorithm where the 3D points are obtained thanks to a generated depth map from the retrieved image candidate. We make our method fast and lightweight by using a common neural network architecture to generate the image descriptor for image indexing and the depth map used to create the 3D points required in the PnP pose refinement step. We demonstrate the effectiveness of our proposal through extensive experimentation on both indoor and outdoor scenes, as well as generalisation capability of our method to unknown environment. Finally, we show how to deploy our system even if geometric information is missing to train our monocular-to-depth neural networks.

1 Introduction

Image-based localisation (IBL) consists in retrieving the exact 6 Degrees of Freedom (DoF) of an image query according to a known reference [24]. IBL is involved in various computer vision and robotics tasks, such as camera relocalisation for augmented reality or SLAM mapping[22], autonomous driving [8], robot or pedestrian localisation [31], cultural heritage [2], etc.

IBL can be considered as a visual place recognition problem [20] and solved using Content Based Image Retrieval (CBIR) [1]. Indeed, as the reference scene is described by a pool of geo-localised images, a coarse pose can be obtained by retrieving the closest reference image to the query. So far, the most successful approaches for IBL are methods matching 2D image features to a 3D reference point cloud, before using a Perspective-n-Point (PnP) algorithm to estimate the 6-DoF pose of the image query [32, 35]. Following these methods, new IBL systems have increased the localisation performances by relying on more and more complete and heavy geometric representation of the environment [37, 39]. However, when the underlying geometry of the scene is not available, or the computational resources allocated to the localisation framework are limited, such methods cannot be deployed.

With the recent advance in machine learning, Kendall et al. [14] introduce Posenet, a new compact system that directly regresses the pose of a given query image. Although Posenet has the advantages of being lightweight and relies on only-images data, Sattler et al.

[36] show that performances of such methods are, by design, less precise than CBIR-based pose estimation [41]. Another disadvantage of Posenet-like methods rely on the fact that a different model has to be trained for each new scene.

Based on these observations, we propose a new pose estimation method built on CBIR augmented by a subsequent pose refinement step, like in [3]. We use dense correspondences from the retrieved image and the query to refine its 6-DoF pose with a PnP algorithm. In order to obtain a position at absolute scale (which is not the case with traditional multi-view methods [10]), we exploit learning to reconstruct the depth map associated to the reference images [40]. We take advantages of the recent progress in depth estimation from monocular images to train our model with or without the supervision of ground truth depth maps [7, 8, 48]. In order to perform online IBL, we use the same neural model to compute the global image descriptor used in CBIR, the dense matching between the query and the retrieved image and to estimate the depth map associated to a single image. Thanks to this multi-task design, our system is compact and lightweight as Posenet while not necessitating the costly scene-specific training as mentioned earlier. Unlike traditional IBL method, our proposal do not requires heavy representation of the scene geometry as we exploit the capability of recent neural networks to learn the underlying structure of a scene from the radiometric appearance.

The rest of our paper is presented as follows: the next section is dedicated to a brief review of the related work, then the details of our method are presented in section 3. The obtained results with our proposal are discussed in section 4, and we finally conclude the paper in section 5.

2 Related work

2.1 Image-based localisation

Sattler et al. [32] have designed a state-of-the-art camera localisation where 2D hand-crafted features from the query image are matched to a large 3D point clouds created by Structure from Motion. Another successful approach have been presented in [33, 34] with a coarse to fine localisation pipeline by initial image indexing followed by feature registration on a local 3D model. In [39], the hand-crafted features usually used for image matching are replaced by dense matching using features block for pre-trained CNN with successive geometric verification steps using a complete 3D model of an indoor building. In [37], authors use, in combination with 3D geometry, semantic labelling of the scene to perform outdoor localisation at large scale. In our proposal, we adopt the coarse to fine localisation strategy while limiting the data required during the pose request to images only.

Learning approaches for camera localisation have also been considered since early work from [38] that uses regression forest at pixel level for fast pose estimation. Cavallari et al. [6] extended this work by reusing the forest structure for fast adaptation to unknown scene. Pose regression CNN-based methods [12, 14, 15, 30, 44] and more recently coordinates regression method [4, 17] are also well studied topics and provide compact localisation system relying on images only. We do not design our system as a direct image to pose regression method, as this approach cannot be generalised and needs specific training and model for each new environment. Closest work to our is a method called Relocnet [26], where authors use a two-step localisation approach consisting of a first pose estimation by CBIR followed by a relative pose estimation between two images with a CNN. By learning relative information, Relocnet can be used in various environment without specific training for each scenes.

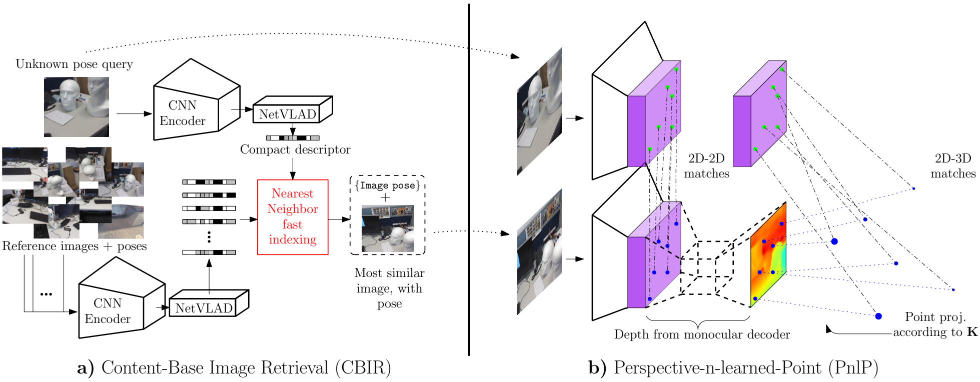

Figure 1: Pipeline of the proposed method. **a)** We retrieve initial pose of an image query using CBIR. **b)** We refine initial pose with a PnP algorithm where 2D to 3D matches are obtained through the reconstructed depth map of the reference image. Purple boxes are deep features blocs used for dense images matching.

2.2 Depth from monocular for localisation

Modern neural networks architectures can provide reliable estimation of the depth associated to monocular image in a simple and fast manner [7, 8, 21]. This ability of neural networks have been used in [40] to recover the absolute scale in a SLAM mapping system. Loo et al. [19] 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 [25], authors use the depth map generated from monocular images as stable features across season changes within a CBIR localisation framework. As in [40], we use the depth information obtained by a neural network to recover the absolute scale of the scene and for modelling the geometry of multiples environment within a common model.

3 Method

Workflow. Our method for fast image pose estimation is described in figure 1. The camera pose is estimated following this two-step algorithm:

a) We obtain the initial pose of the query image by Content-Based Image Retrieval (section 3.1).

b) Initial pose is refined by finding dense correspondences between the query image and the best retrieved image (section 3.2). Meanwhile, we use a neural network to create the depth map related to the retrieved image candidate (section 3.3). We use correspondences between the 2D points of the query image and the 3D points projected from the depth map to compute the real pose of the query using Perspective-n-Point (PnP) algorithm (section 3.4). We further denote our pose refinement method as Perspective-n-learned-Point (PnlP).

Notations. The aim of our method is to recover the camera pose $\mathbf{h_q} \in \mathbb{R}^{4\times4}$, represented by a pose matrix in homogeneous coordinates, corresponding to an input RGB image $\mathrm{I_q} \in$

$\mathbb{R}^{3\times H\times W}$. We know the matrix $\mathbf{K}\in\mathbb{R}^{3\times 3}$ of intrinsic parameters of the camera. We assume that we know the pose $\{\mathbf{h_r}^i\}_{i=1,\ \dots,\ N}$ of a pool of N references images $\{\mathrm{I_r}^i\}_{i=1,\ \dots,\ N}$ of the scene where we want to localise the query. These poses can be obtained by Structure from Motion (SfM) or by using external sensors. We denote as E, respectively D, a neural network encoder, respectively decoder.

3.1 Image retrieval

We cast the initial pose estimation task as a content-based image retrieval problem like in [3], since the reference data are augmented with 6 DoF pose information. In order to evaluate the similarity between the unknown pose query image $\mathrm{I_q}$ and the N reference images $\{\mathrm{I_r}^i\}_{i=1,\ \dots,\ N}$, we need to use a discriminative image representation. Recent works have shown that deep features extracted from convolutional neural network offer better global image representations compared to hand-crafted features [1, 9, 27, 28]. We use a state-of-the art global image descriptor for place recognition, NetVLAD [1], to describe the data by low-dimensional L_2 normalised vectors. The NetVLAD descriptor $\mathbf{f}$ is obtained by concatenating the dense feature from neural network encoder E: $\mathbf{f}=\mathrm{NetVLAD}(\mathrm{E}(\mathrm{I}))$.

We first compute reference descriptors $\{\mathbf{f_r}^i\}_{i=1,\ \dots,\ N}$ from the reference images. Then we compare the query descriptor $\mathbf{f_q}$ to the pre-computed descriptors by fast nearest neighbour indexing and retrieval:

$$\left\{\hat{\mathbf{f}}_\mathrm{r}{}^j\right\}_{j=1,\ \dots,\ K}=NN\left(\mathbf{f_q},\{\mathbf{f_r}^i\}_{i=1,\ \dots,\ N}\right), \tag{1}$$

where NN is the nearest neighbour matching function and $\hat{\mathbf{f}}_\mathrm{r}{}^j$, $j\in[1,K]$, the K closest reference descriptors to the query descriptor. We use cosine similarity to evaluate the similarity between two descriptors and K-D tree as indexing structure. We consider poses $\mathbf{h}_\mathrm{r}^j$, $j\in[1,K]$, as candidate poses of the image $\mathrm{I_q}$.

3.2 Dense correspondences

In order to refine the initial pose obtained by image retrieval, we compute correspondences between the query image and the closest retrieved image candidates. In [39], authors use the dense features extracted by a convolutional neural network in order to compute correspondences between images. We follow the same idea and use the latent representation already computed by the neural network encoder E to compute correspondences between the query image and the K retrieved candidates.

Local image descriptors are obtained from the latent image representation by concatenating the features at each position $(l,m)_{W_\mathrm{E},H_\mathrm{E}}$ (W_E and H_E are the spatial dimensions of the features map) along the depth of the features map [39, 45]. We subsequently L_2-normalise the extracted descriptors before matching. We consider only consistence matches by rejecting correspondences that do not respect the bidirectional test (nearest descriptors of image 1 in image 2 have to be the same as nearest descriptors of image 2 to image 1).

3.3 Depth from monocular

2D to 2D correspondences obtained by dense features matching (section 3.2) do not provide enough information to compute relative pose between images at absolute scale. Therefore, we propose to reconstruct the relative scene geometry from the camera to circumvent this limitation. Various recent deep learning generative models are able to properly reconstruct geometry associated to radiometric data, with full supervision training [7], weakly annotated data [8] or even in a self-supervised way [21].

We train an encoder/decoder jointly to predict the corresponding depth map M associated to an image: $\mathrm{M} = \mathrm{D}(\mathrm{E}(\mathrm{I}))$. With the generated depth map obtained by our neural network and the intrinsic parameters of the camera $\mathbf{K}$, we can project the 2D point $(l,m)^T$ to the corresponding 3D coordinate $\mathbf{p}$:

$$\mathbf{p} = \mathrm{M}^{l,m} \cdot \mathbf{K}^{-1}[l,m,1]^T. \tag{2}$$

3.4 Pose refinement

Thanks to the generated depth map (section 3.3) and the equation 2, we can project 2D points from retrieved images into 3D coordinates. 2D-2D correspondences obtained in section 3.2 can be interpreted as 2D-3D correspondences and we can use PnP algorithm to compute the relative transformation $\mathbf{h}_{\mathrm{r}\rightarrow\mathrm{q}}$ between the query image and the reference image. Final pose of query image I_{q} is obtained with the equation:

$$\mathbf{h}_{\mathrm{q}} = \mathbf{h}_{\mathrm{r}}\mathbf{h}_{\mathrm{r}\rightarrow\mathrm{q}}. \tag{3}$$

We embed the PnP algorithm within a RANSAC consensus where a sub-part of 2D-3D correspondences are evaluated at a time. As we have multiple reference candidates from image retrieval step (section 3.1), we select the pose with the largest proportion of inlier correspondences after the PnP optimisation. If the ratio of inlier is below a given threshold, we simply affect the pose of the retrieved image to the query.

3.5 System design and motivation

Multi-task model. In order to make our system fast and lightweight, we use a single encoder/decoder neural network for the three tasks needed in our pose estimation pipeline. That means with a single image forward, we obtain a compact global image description, dense local descriptors and a depth map corresponding to the observed scene.

Single task training policy. It exists in the literature methods for training deep neural network either for global image description [1, 9, 27], local features extraction and description [23, 29, 46] or depth from monocular estimation [7, 8, 21]. We decide to train our encoder/decoder network for the task of depth from monocular estimation because estimation of erroneous depth measurement will result in wrong estimation of the final pose. In the next section, we experimentally show that even if our network has not been trained especially for the task of image description or local feature matching, the latent features computed within the network embed enough high-level semantic to perform well on these tasks [39, 47].

Generalisation. Because we rely on a non-absolute representation of the scene geometry (depth is estimated *relatively* to the camera frame), our model is not limited to localisation on one specific scene like end-to-end pose estimation networks [4, 13]. The same trained network can be used to localise images in multiple indoor and outdoor scenes, and even on totally unknown environments.

4 Experiments

In this section, we present extensive experiments to evaluate our proposal. We consider two localisation scenarios: indoor static scenes (section 4.2) and outdoor dynamic scenes (section 4.3). We also divide our evaluation according to the data available to train our encoder/decoder architecture: fully supervised depth from monocular training (when ground-truth associated depth map are available during training), and unsupervised depth from monocular (when the only data available during training are video sequences with relative poses between images).

4.1 Implementation details

Datasets. We test our method on the following indoor localisation datasets: 7 scenes [38] and 12 scenes [42]. These datasets are composed of various indoor environments scanned with RGB-D sensors. We use the Cambridge Landmarks [14] dataset for outdoor evaluation. This dataset is composed of 6 scenes featuring dynamic changes (pedestrian and cars in movement during the acquisition) acquired by a cell-phone camera. 6-DoF image poses and camera calibration parameters are provided for these 3 datasets. For all the experiments, reference images used for the initial pose estimation with CBIR are taken from the training split and query images are taken from the testing split of the respective datasets.

As not ground truth depth maps are available for the Cambridge Landmarks scenes, we only perform outdoor experiments related to the unsupervised depth from monocular training.

Networks architecture and training. For both fully supervised and unsupervised depth from monocular experiments, we use a U-Net like convolutional encoder/decoder architecture [11] with multi-scale outputs [8]. For the unsupervised scenario, we also try to add some recurrent layers (LSTM) in the decoder to capture long term dependencies [18, 43]. We denote the fully convolutional architecture as **FC** and convolutional layers + recurrent layers architecture as **C+LSTM**. FC and C+LSTM encoders are identical, with 6.3M parameters, FC decoder has 16.7M parameters and C+LSTM decoder has 10.1M parameters.

During training and testing, images are resized to 224×224 pixels for indoor scenes, and 224×112 for outdoor images. The generated depth map is 4 times smaller than the RGB input. We use L_1 loss function for the fully supervised depth from monocular training. To learn depth from RGB in a unsupervised manner, we follow the training procedure of [48], using the ground truth relative pose between images and by adding SSIM loss function for radiometric comparison as in [21]. We train all the architecture with adam optimizer, learning rate of 10^{-4} divided by two every 50, respectively 5, epochs for the supervised, respectively unsupervised, training. Training takes approximately one day on our Nvidia Titan X GPU with a batch size is set to 24, respectively 12, for supervised, respectively unsupervised, training.

We train networks for indoor localisation on the 7 scenes dataset (using only sequences from the training split). The 12 scenes dataset is used to evaluate the generalisation capability of our method. For outdoor localisation, we train our two different architectures (FC and C+LSTM) on the Cambridge Landmarks dataset.

	Scene	Image retrieval FC-sup.	Image retrieval FC-unsup.	PnlP refinement FC-sup.	PnlP refinement FC-unsup.	Relocnet [3]	Posenet [13]
7-Scenes [38]	Chess	0.29/13.0	0.34/15.4	0.07/2.7	0.13/4.7	0.12/4.1	0.13/4.5
	Fire	0.40/15.5	0.48/19.3	0.07/3.2	0.22/8.2	0.26/10.4	0.27/11.3
	Heads	0.28/20.5	0.25/17.9	0.05/3.9	0.15/10.5	0.14/10.5	0.17/13.0
	Office	0.38/13.0	0.50/16.1	0.09/2.9	0.23/6.3	0.18/5.3	0.19/5.6
	Pumpkin	0.43/13.1	0.54/15.0	0.13/3.6	0.29/7.1	0.26/4.2	0.26/4.8
	Red Kitchen	0.23/9.5	0.26/10.5	0.05/2.0	0.12/3.3	0.23/5.1	0.23/5.4
	Stairs	0.46/14.9	0.49/15.5	0.40/9.2	0.48/12.2	0.28/7.5	0.35/12.4
12-Scenes [42]	Apt1-kitchen	0.12/7.7	0.14/9.2	0.09/4.1	0.14/5.0	-	-
	Apt1-living	0.12/6.8	0.13/6.7	0.08/2.9	0.10/3.3	-	-
	Apt2-kitchen	0.10/6.5	0.10/6.6	0.10/3.7	0.10/3.9	-	-
	Apt2-living	0.11/5.6	0.13/7.3	0.10/4.7	0.11/3.7	-	-
	Apt2-bed	0.13/7.0	0.12/7.1	0.12/5.7	0.15/5.0	-	-
	Apt2-luke	0.15/7.2	0.16/7.8	0.14/5.5	0.14/5.3	-	-
	Office 5a	0.12/5.3	0.13/6.3	0.09/3.6	0.14/4.6	-	-
	Office 5b	0.15/7.2	0.18/6.7	0.10/4.7	0.14/5.0	-	-

Table 1: Results on the **7 scenes** [38] and **12 scenes** [42] indoor datasets, we report median position/orientation error in meters/degree. We compare the first pose estimation (im. retrieval) and, the final image localisation (PnlP) of our method and two state-of-the-art approaches. Red numbers show failure cases when the pose refinement increases the initial pose error. Sup. and unsup. stand for supervised, respectively unsupervised, depth from monocular training.

Method parameters. We use NetVLAD layer with 64 clusters as global image descriptor for initial pose estimation. We concatenate features from the last convolutional layers of the encoder network, composed of 256 convolutional filters, resulting in a global descriptor of size 16384. Descriptor dimension can be further reduced with PCA projection [1]. We consider the 5-top retrieved candidates from the nearest neighbour search in the pose refinement process, resulting in a good trade-off between time consumption and pose estimation performances. For the final pose estimation, we use the fast `C++` PnP implementation from [16] and we set the inlier ratio threshold mentioned in section 3.4 to 10%.

4.2 Indoor localisation

Indoor localisation error on 7 scenes [38] dataset are presented in table 1. We compare our proposal with Relocnet [26] and Posenet [13] trained with a geometric-aware loss. At first glance, we find that the initial pose estimation with image retrieval produces decent results (first two columns), while the network used to produce the global image descriptor has not been trained to this particular task. After applying our PnlP pose refinement, the model trained in a fully supervised manner produces the most precise localisation among the presented methods.

For the unsupervised setting, we found that FC and C+LSTM architectures perform equivalently on the indoor dataset, thus we present only results of the FC architecture. We observe an average relative improvement of $\times 2.8/\times 3.5$, respectively $\times 1.8/\times 2.1$, for the supervised, respectively unsupervised, model in position/rotation from initial to PnlP refined pose. Compared to Posenet [13] our unsupervised model perform equivalently, while using the same trained network for all the 7 scenes, compared to one network by scene for Posenet. Our proposal clearly outperforms Relocnet [3] in a supervised setting, while producing comparable localisation for the model trained in an unsupervised manner. It is important to remind that Relocnet relies on two different networks: one trained especially to produce discriminative global image descriptors for CBIR and the second to estimate the relative pose between two images. Our method is lighter as it uses a single network and do

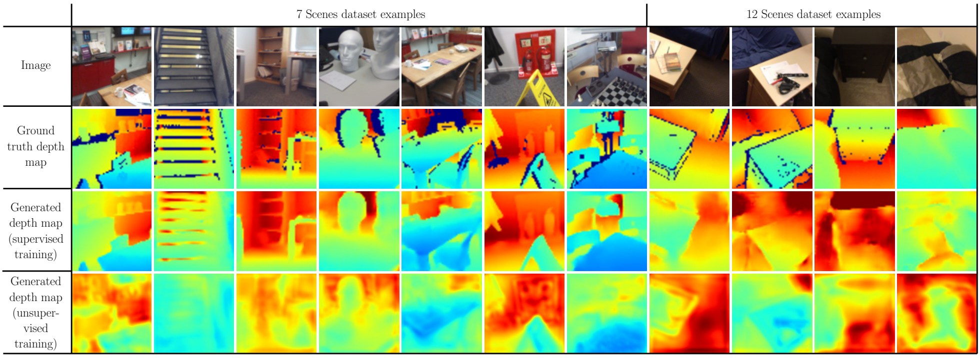

Figure 2: Visualisation of the depth map generated from RGB input with two networks trained with full supervision or without ground truth depth map in an unsupervised manner. In both configurations, networks are trained on the **7 scenes** dataset [38]. Examples from **12 scenes** [42] show networks generalisation capability.

		Great Court	Kings C.	Old Hosp.	Shop	St Mary's	Street
Im. retrieval	FC-unsup.	27.6/26.79	4.4/6.10	6.2/10.09	4.3/14.93	6.9/15.17	95.5/58.38
	C+LSTM-unsup	24.3/20.94	5.0/5.86	6.5/8.60	3.2/9.47	5.9/12.71	92.5/67.10
PnlP	FC-unsup.	25.5/22.64	2.9/2.98	4.9/6.37	1.8/5.78	3.5/6.99	76.2/51.91
	C+LSTM-unsup	13.2/10.07	2.7/3.10	3.5/5.55	1.1/3.38	2.6/5.85	69.5/52.07
	Posenet [13]	-	0.9/1.04	3.2/3.29	0.9/3.78	1.6/3.32	20.3/25.5

Table 2: Results on the Cambridge Landmarks [14] outdoor dataset, we report median position/orientation error in meters/degree. We compare our two network architectures, FC and C+LSTM, trained in an unsupervised manner.

not uses specific training for the task of global image description. We observe a failure case of our method for the scene stairs due to a poor initial pose estimation. This scene contains repetitive visual patterns that may confuse the CBIR localisation.

Generalisation. We also report on table 1 localisation error on 8 scenes of the 12 Scenes dataset [42]. For these experiments, we use the same network as mentioned earlier, trained on 7 Scenes dataset [38]. We observe an average relative improvement of $\times 1.2/\times 1.5$, respectively $\times 1.1/\times 1.6$, for the supervised, respectively unsupervised, model in position/rotation from initial to refined pose. Even though the pose refinement is not as effective as previously, it shows that our system can be used on completely new indoor environments. We also demonstrate, in figure 2, the generalisation capability of our method through the depth maps produced by our networks, from images taken on both known and unknown scenes. We notice that the poor localisation performance on the Apt2-bed scenes is closely related to the poor generated depth map on this scene (see figure 2, two last columns).

4.3 Outdoor localisation

As mentioned previously, we only test our unsupervised set-up for outdoor image pose estimation as the Cambridge Landmarks dataset [14] does not contain ground truth depth maps. Results are presented in table 2. PnlP performs well on outdoor scene, with a mean improvement of $\times 1.3/\times 1.4$ for FC architecture, and $\times 1.5/\times 1.6$ for C+LSTM, in position/rotation precision over initial pose given by CBIR. Superior performances of C+LSTM model can be

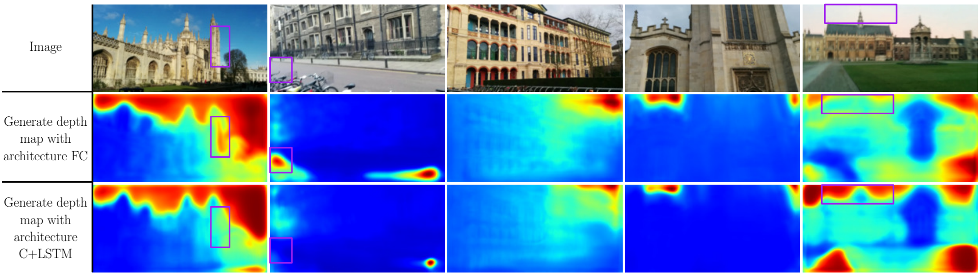

Figure 3: Visualisation of the depth map generated from RGB input by our two architectures, FC and C+LSTM, trained in an unsupervised manner on Cambridge Landmarks dataset [14]. Purple boxes show regions where C+LSTM network produces slightly better depth map reconstruction compared to FC.

explained by a better capability of the recurrent cells in the C+LSTM decoder for modelling the 3D structure of the scene, as shown in figure 3. Our method is not able to recover a proper pose for the scene Street. As same as for the indoor failure case, this is the result of a poor initial pose estimation at the CBIR preliminary step. Compared to Posenet [13], our method is marginally less precise but requires only one trained model compared to the 6 models needed by Posenet and can potentially be used on unknown scenes according to the previous indoor experiments. We do not compare our method to Relocnet [26] baseline because authors do not evaluate Relocnet on outdoor scenes.

4.4 Limitations

The final camera pose precision is highly dependent on the images returned by the CBIR inital step. Thus, our method performances are limited by the quality of the global image descriptor. Wrong initial pose estimation for stairs indoor scene and street outdoor environment cannot be recovered by PnlP pose refinement. It will be interesting to consider more discriminative image descriptors, and especially image descriptors that can benefit from the depth map related to the image [25].

The pose refinement is also very sensitive to the quality of the generated depth map. Artefacts present on depth map related to images of unknown scenes, see last 4 columns of figure 2, or wrong reconstruction, last column of figure 3, generate outliers for the PnlP optimisation.

5 Conclusion

We have introduced a new method for online IBL consisting of an initial pose estimation by CBIR followed by our new PnlP pose refinement. In order to achieve relative pose estimation at absolute scale, we learn to reconstruct the depth map associated to a monocular image to project into 3D densely matched 2D points between the query and the reference. The presented method is compact and fast as all the components needed by the localisation pipeline are computed thanks to the same neural network in a single forward pass. Because our network learns the depth relative to the camera frame, not the absolute geometric structure of the scene, it can be used in unknown environment without fine tuning or specific training.

In a future work, we will investigate multi-task learning in order to address all the computer vision problems involved in IBL jointly, namely global image description, dense correspondences between images and depth from monocular.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 5297–5307, 2017. ISSN 10636919. doi: 10.1109/CVPR.2016.572. URL http://arxiv.org/abs/1511.07247.

[2] Mathieu Aubry, Bryan C. Russell, and Josef Sivic. Painting-to-3D model alignment via discriminative visual elements. *ACM Transactions on Graphics (ToG)*, 33(2):1–14, 2014. ISSN 07300301. doi: 10.1145/2591009. URL https://hal.inria.fr/hal-00863615v3http://dl.acm.org/citation.cfm?id=2591009{%}5Cnhttp://pucrio.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwlV3dS8MwED9ERRTx225-QB98bV2bZG0f5-YYyMQHBfFFkjQRwW3qOv

[3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. RelocNet : Continous Metric Learning Relocalisation using Neural Nets. In *IEEE European Conference on Computer Vision (ECCV)*, 2018.

[4] Eric Brachmann and Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00489. URL http://arxiv.org/abs/1711.10228.

[5] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-Aware Learning of Maps for Camera Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00277. URL http://arxiv.org/abs/1712.03342.

[6] Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Victor A. Prisacariu, Luigi Di Stefano, and Philip H. S. Torr. Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade. *arXiv preprint*, pages 1–18, 2018. doi: arXiv:1810.12163v1. URL http://arxiv.org/abs/1810.12163.

[7] David Eigen, Christian Puhrsch, and Rob Fergus. 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, 2014. ISBN 10495258. doi: 10.1007/978-3-540-28650-9_5. URL http://arxiv.org/abs/1406.2283.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.699. URL http://arxiv.org/abs/1609.03677.

[9] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. End-to-End Learning of Deep Visual Representations for Image Retrieval. *International Journal of Computer Vision (IJCV)*, 124(2):237–254, 2017. ISSN 15731405. doi: 10.1007/s11263-017-1016-8.

[10] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[11] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.632.

[12] Alex Kendall and Roberto Cipolla. Modelling Uncertainty in Deep Learning for Camera Relocalization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[13] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL http://arxiv.org/abs/1704.00390.

[14] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.336.

[15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. 2017. URL http://arxiv.org/abs/1705.07115.

[16] Laurent Kneip and Paul Furgale. OpenGV: A unified and generalized approach to real-time calibrated geometric vision. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2014.

[17] Xiaotian Li, Juha Ylioinas, Jakob Verbeek, and Juho Kannala. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization. In *IEEE European Conference on Computer Vision Workshops (ECCVW)*, pages 1–17, 2018. URL http://arxiv.org/abs/1808.04999.

[18] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling. In *IEEE European Conference on Computer Vision (ECCV)*, volume 9906 LNCS, pages 541–557, 2016. ISBN 9783319464749. doi: 10.1007/978-3-319-46475-6_34.

[19] Shing Yan Loo, Ali Jahani Amiri, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. 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, 2019. doi: arXiv:1810.01011v1. URL http://arxiv.org/abs/1810.01011.

[20] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J. Leonard, David Cox, Peter Corke, and Michael J. Milford. Visual Place Recognition: A Survey. *IEEE Transactions on Robotics (TRO)*, 32(1):1–19, 2016. ISSN 15523098. doi: 10.1109/TRO.2015.2496823.

[21] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00594. URL http://arxiv.org/abs/1802.05522.

[22] Lili Meng, Jianhui Chen, Frederick Tung, James J Little, and Clarence W. de Silva. Exploiting Random RGB and Sparse Features for Camera Pose Estimation. In *British Machine Vision Conference (BMVC)*, pages 1–12, 2016.

[23] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning Local Features from Images. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2018. URL http://arxiv.org/abs/1805.09662.

[24] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on Visual-Based Localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, feb 2018. ISSN 00313203. doi: 10.1016/j.patcog.2017.09.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0031320317303448.

[25] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Learning Scene Geometry for Visual Localization in Challenging Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[26] Pulak Purkait, Cheng Zhao, and Christopher Zach. Synthetic View Generation for Absolute Pose Regression and Image Synthesis. In *British Machine Vision Conference (BMVC)*, 2018.

[27] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. URL http://arxiv.org/abs/1711.02512.

[28] Ali Sharif Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual Instance Retrieval with Deep Convolutional Networks. *arXiv preprint*, 4(3):251–258, 2014. ISSN 2186-7364. doi: 10.3169/mta.4.251. URL http://arxiv.org/abs/1412.6574.

[29] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood Consensus Networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, number Nips, 2018. ISBN 1810.10510v2. doi: arXiv:1810.10510v2. URL http://arxiv.org/abs/1810.10510.

[30] Soham Saha, Girish Varma, and C. V. Jawahar. Improved Visual Relocalization by Discovering Anchor Points. In *British Machine Vision Conference (BMVC)*, 2018. URL http://arxiv.org/abs/1811.04370.

[31] Torsten Sattler, Michal Havlena, Filip Radenović, Konrad Schindler, and Marc Pollefeys. Hyperpoints and fine vocabularies for large-scale location recognition. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 2102–2106, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.243.

[32] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, X(1), 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2611662.

[33] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, Fredrik Kahl, and Tomas Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. pages 8601–8610, 2017. doi: 10.1109/CVPR.2018.00897. URL http://arxiv.org/abs/1707.09092.

[34] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] Torsten Sattler, Will Maddern, Akihiko Torii, Josef Sivic, Tomas Pajdla, Marc Pollefeys, and Masatoshi Okutomi. Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL http://arxiv.org/abs/1707.09092.

[36] Torsten Sattler, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31–34, 2019. URL http://arxiv.org/abs/1903.07504.

[37] Johannes L. Schönberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic Visual Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00721. URL http://arxiv.org/abs/1712.05773.

[38] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, 2013. ISBN 978-0-7695-4989-7. doi: 10.1109/CVPR.2013.377.

[39] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor Visual Localization with Dense Matching and View Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00752. URL http://arxiv.org/abs/1803.10368.

[40] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL http:

//scielo.sld.cu/scielo.php?script=sci{_}arttext{&}pid=S0034-75232011000300003{&}lng=es{&}nrm=iso{&}tlng=es.

[41] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[42] Julien Valentin, Angela Dai, Matthias Niessner, Pushmeet Kohli, Philip Torr, Shahram Izadi, and Cem Keskin. Learning to navigate the energy landscape. In *International Conference on 3D Vision (3DV)*, pages 323–332, 2016. ISBN 9781509054077. doi: 10.1109/3DV.2016.41.

[43] Francesco Visin, Kyle Kastner, Kyunghyun Cho, Matteo Matteucci, Aaron Courville, and Yoshua Bengio. ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. *arXiv preprint*, pages 1–9, 2015. URL http://arxiv.org/abs/1505.00393.

[44] Florian Walch, Caner Hazirbas, Laura Leal-Taixé, Torsten Sattler, Sebastian Hilsenbeck, and Daniel Cremers. Image-based Localization with Spatial LSTMs. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. URL http://arxiv.org/abs/1611.07890.

[45] Aji Resindra Widya, Akihiko Torii, and Masatoshi Okutomi. Structure from motion using dense CNN features with keypoint relocalization. *IPSJ Transactions on Computer Vision and Applications*, pages 0–6, 2018.

[46] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. LIFT: Learned Invariant Feature Transform. In *IEEE European Conference on Computer Vision (ECCV)*, volume 9905, pages 467–483, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0. URL http://link.springer.com/10.1007/978-3-319-46448-0.

[47] Amir R Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722, 2018. URL http://openaccess.thecvf.com/content{_}cvpr{_}2018/papers/Zamir{_}Taskonomy{_}Disentangling{_}Task{_}CVPR{_}2018{_}paper.pdf.

[48] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised Learning of Depth and Ego-Motion from Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.700. URL http://arxiv.org/abs/1704.07813.