Vision-based localization with discriminative features from heterogeneous visual data

Nathan Piasco

UBFC
UNIVERSITE
BOURGOGNE FRANCHE-COMTE

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PREPARÉE A UNIVERSITÉ DE DIJON

Ecole doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat en instrumentation et informatique de l'image

Par

M PIASCO NATHAN

Vision-based localization with discriminative features from heterogeneous visual data

Thèse présentée et soutenue au Creusot le 15/11/2019

Composition du Jury :

M Thome Nicolas	Professeur au Conservatoire National des Arts et Métiers	Rapporteur
M Lepetit Vincent	Professeur à l'Université de Bordeaux	Rapporteur
M Sivic Josef	Directeur de recherche à l'INRIA de Paris	Examinateur
M Torsten Sattler	Professeur associé à l'Université de Chalmers	Examinateur
Mme Gouet-Brunet Valérie	Directrice de recherche au LASTIG	Codirectrice de thèse
M Demonceaux Cédric	Professeur à l'Université de Bourgogne - Franche-Comté	Codirecteur de thèse
M Sidibé Désiré	Professeur à l'Université Evry Val d'Essonne	Encadrant de thèse

Reviewers:

Vincent Lepetit, Professor, University of Bordeaux

Nicolas Thome, Professor, Conservatoire national des arts et métiers

Day of the defense: 25/11/2019

Signature from head of PhD committee:

Publications

Peer-Review Journals Papers

1. **N. Piasco**, D. Sidibé, V. Gouet-Brunet and C. Demonceaux. Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions. (in submission)
2. **N. Piasco**, D. Sidibé, C. Demonceaux and V. Gouet-Brunet. A Survey on Visual-Based Localization: On the Benefit of Heterogeneous Data. Pattern Recognition, Volume 74, February 2018, pp.90-109.

Peer-Review International Conferences

1. **N. Piasco**, D. Sidibé, C. Demonceaux and V. Gouet-Brunet. Perspective-n-Learned-Point: Pose Estimation from Relative Depth. 2019 British Machine Vision Conference (BMVC), Cardiff, United Kingdom, September 2019.
2. **N. Piasco**, D. Sidibé, C. Demonceaux and V. Gouet-Brunet. Geometric Camera Pose Refinement with Learned Depth Maps. 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, September 2019.
3. **N. Piasco**, D. Sidibé, V. Gouet-Brunet and C. Demonceaux. Learning Scene Geometry for Visual Localization in Challenging Conditions. 2019 IEEE International Conference of Robotics and Automation (ICRA), Montreal, Canada, May 2019.

Peer-Review National Conferences

1. **N. Piasco**, D. Sidibé, V. Gouet-Brunet and C. Demonceaux. Apprentissage de modalités auxiliaires pour la localisation basée vision. Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP), Champs-sur-Marne, France, June 2018
2. **N. Piasco**, D. Sidibé, V. Gouet-Brunet and C. Demonceaux. Localisation Basée Vision : de l'hétérogénéité des approches et des données. ORASIS - Journées francophones des jeunes chercheurs en vision par ordinateur (ORASIS), Colleville-sur-Mer, France, June 2017.

Thesis

1. **N. Piasco**. Vision-based localization with discriminative features from heterogeneous visual data.

List of Abbreviations

ANR Agence Nationale de la Recherche

AR Augmented Reality

BoF Bag of Features

CAD Computer Aided Design

CBIR Content-based Image Retrieval

CNN Convolutional Neural Networks

DEM Digital Elevation Model

DL Deep Learning

DoF Degrees of Freedom

FT Fourier Transform

FV Fisher Vector

GAN Generative Adversarial Network

GeM Generalized-Mean

GIS Geographic Information System

GMM Gaussian Mixture Model

HPR Hidden Point Removal

ICLP Iterative Closest Learned Point

ICP Iterative Closest Point

LCD Loop Closure Detection

LDA Linear Discriminant Analysis

LSTM Long Short-term Memory

MAC Maximum Activation of Convolutions

mAP mean Average Precision

ML Machine Learning

MoE Mixture of Export

MTL Multi-task Learning

NN Neareast Neighbor

PCA Principal Components Analysis

pLaTINUM Cartographie Long Terme pour la Navigation Urbaine

PnLP Perspective-n-Learned-Point

RANSAC Random Sample Consensus

R-MAC Regional Maximum Activation of Convolutions

ROI region of interest

SfM Structure From Motion

SLAM Simultaneous Localization and Mapping

SPoC Sum-Pooled Convolutional Features

SVD Singular Value Decomposition

SVM Support Vector Machines

ToF Time of Flight

VBL Visual-based Localization

VLAD Vector of Locally Aggregated Descriptors

VO Visual Odometry

Contents

List of Figures

List of Tables

Acknowledgements

Je voudrais tout d'abord adresser mes remerciements à mes deux merveilleux directeurs de thèse. Merci Valérie de m'avoir considéré comme un véritable collègue et merci pour la pleine confiance que tu m'as accordée. Quant à Cédric, merci de m'avoir guidé et soutenu à tout moment pendant ces trois années, parfois difficiles. Je souhaiterais remercier Désiré qui a toujours répondu présent et qui est devenu un peu plus qu'un encadrant de thèse pour moi.

Tout au long de ma thèse, j'ai eu la chance de partager mon temps entre deux laboratoires de recherche. Je voudrais remercier toute l'équipe Parisienne du Lastig, permanents, doctorants et stagiaires, pour les moments de sciences, de sports et de rire que j'ai pu partager avec vous. Je remercie également toute la famille Creusotine du laboratoire ImViA qui m'a accueilli chaleureusement pour mes cours séjours hors métropole et avec qui j'ai pu partager raclettes, barbecues et très bon vin de Bourgogne. J'ai une pensé toute particulière pour mon logeur, David, qui m'a accueilli chez lui dès mon premier séjour et sans qui mes voyages au Creusot aurait été plus terne (et plus chaud par moment).

Enfin, je souhaiterais remercier celle qui partage ma vie depuis plus dix ans, qui m'a donné l'énergie de tout donner pendant ces trois année et sans qui je ne pourrais pas être aussi fier du travail que j'ai accompli.

Abstract

Visual-based Localization (VBL) consists in retrieving the location of a visual image within a known space. VBL is involved in several present-day practical applications, such as indoor and outdoor navigation, 3D reconstruction, etc. The main challenge in VBL comes from the fact that the visual input to localize could have been taken at a different time than the reference database. Visual changes may occur on the observed environment during this period of time, especially for outdoor localization. Recent approaches use complementary information in order to address these visually challenging localization scenarios, like geometric information or semantic information. However geometric or semantic information are not always available or can be costly to obtain. In order to get free of any extra modalities used to solve challenging localization scenarios, we propose to use a modality transfer model capable of reproducing the underlying scene geometry from a monocular image.

At first, we cast the localization problem as a Content-based Image Retrieval (CBIR) problem and we train a CNN image descriptor with radiometry to dense geometry transfer as side training objective. Once trained, our system can be used on monocular images only to construct an expressive descriptor for localization in challenging conditions. Secondly, we introduce a new relocalization pipeline to improve the localization given by our initial localization step. In a same manner as our global image descriptor, the relocalization is aided by the geometric information learned during an offline stage. The extra geometric information is used to constrain the final pose estimation of the query. Through comprehensive experiments, we demonstrate the effectiveness of our proposals for both indoor and outdoor localization.

Résumé

La localisation basée vision consiste à déterminer l'emplacement d'une requête visuelle par rapport à un espace de référence connu. Le principal défi de la localisation visuelle réside dans le fait que la requête peut avoir été acquise à un moment différent de celui de la base de données. On pourra alors observer des changements visuels entre l'environnement actuel et celui de la base de référence, en particulier lors d'application de localisation en extérieur. Les approches récentes utilisent des informations complémentaires afin de répondre à ces scénarios de localisation visuellement ambigus, comme la géométrie ou la sémantique. Cependant, ces modalités auxiliaires ne sont pas toujours disponibles ou peuvent être coûteuses à obtenir. Afin de s'affranchir de l'utilisation d'une modalité supplémentaire pour faire face à ces scénarios de localisation difficiles, nous proposons d'utiliser un modèle de transfert de modalité capable de reproduire la géométrie d'une scène à partir d'une image monoculaire.

Dans un premier temps, nous présentons le problème de localisation comme un problème d'indexation d'images et nous entrainons un réseau de neurones convolutif pour la description globale d'image, en introduisant le transfert de modalité radiométrie vers géométrie comme objectif secondaire. Une fois entrainé, notre modèle peut être appliqué à des images monoculaires pour construire un descripteur efficace pour la localisation en conditions difficiles. Dans un second temps, nous introduisons une nouvelle méthode de raffinement de pose pour améliorer la localisation obtenue à la première étape. Comme pour le descripteur d'image global, la relocalisation est facilitée par les informations géométriques apprises lors d'une étape préalable. L'information géométrique supplémentaire est utilisée pour contraindre l'estimation finale

de la pose de la requête. Grâce à des expériences approfondies, nous démontrons l'efficacité de nos propositions pour la localisation en intérieur et en extérieur.

Chapter 1

Introduction

This first chapter introduce the scientific environment of the thesis. We first present the global topic of long-term mapping followed by the introduction the overall project this thesis is part of. Then, we focus on the localization task, as it is the main topic of this research work.

1.1 Long-term mapping

Creating informative, accurate and detailed maps is a crucial step for many applications: pedestrian or vehicle navigation, data valuation, visualization, land or spatial planning, to name a few. Mobile mapping vehicles are able to collect and arrange large amount of data in order to create rich referential. However, these maps are fixed in time. Depending of the mapped area, the environment representation could be quickly outdated. Instead of recreating a completely new map, which is a costly operation, another solution is to locally update the map. This is what we called long-term mapping.

In order to include new data source to our initial representation, we need to located these information. In other words, we have to find the position of the up-to-date data according to the original map frame. Once we get a proper alignment between the two sources of data, we can update our main referential. In this thesis, we focus on the localization of visual data in the context of long-term mapping. Visual-based Localization (VBL) is not limited to the update of outdated referential and further applications are presented in the following. Thus, the core subject of this research is about localization of new visual data to a fixed, potentially outdated, map.

1.2 pLaTINUM project

This thesis is founded by the French Agence Nationale de la Recherche (ANR) and is part of the project named Cartographie Long Terme pour la Navigation Urbaine (pLaTINUM) (ANR-15-CE23-0010). pLaTINUM is a long-term mapping project composed of three parts: high quality multi-sources map creation, online visual-based urban navigation with user feedback and automatic map update for long-life usage. The first part is an offline mapping step from multi-modal data sources collected by a mobile mapping vehicle [31, 32, 206] that produces a high resolution textured mesh with radiometric, geometric and semantic information. Then, this map is used as a reference for an online visual navigation module. During the navigation, an agent sends visual feedback to the server in order to, in a third step, update the map if changes are detected between the reference map and the current observation. The city center of Rouen, in France, have been chosen to carry out the project experiments.

The subject of this thesis is the initial localization of the agent over the entire map before the start of the navigation. We detail in the next paragraphs the localization pipeline.

Summary of the map. The online localization task within the pLaTINUM project consists of finding the position a visual data from the agent over a summarized version of the global map. To summarize the initial textured mesh, we render a set of radiometric (RGB), depth (D) and semantically labeled (L) spheres at meaningful location for covering the entire area of possible navigation [252, 253, 254]. This representation contains all the modalities and most of the information from the original map while being lighter.

Initial agent query. In order to start the visually-guided navigation of the agent, we have to find its absolute position on the mapped area. We assume that the agent is not equipped with any global localization equipment, such as GPS, and carry only an embedded device to acquire visual information. This is a regular assumption in urban area where global positional system can suffer from buildings obstruction (*e.g.* the urban canyon effect that affects the GPS signal). In order to be globally located, the agent sends from his capture device a visual request to a server. By visual request, we regardless denote: monocular image, video sequence, pair of stereo images, semantically annotated image or combination of these.

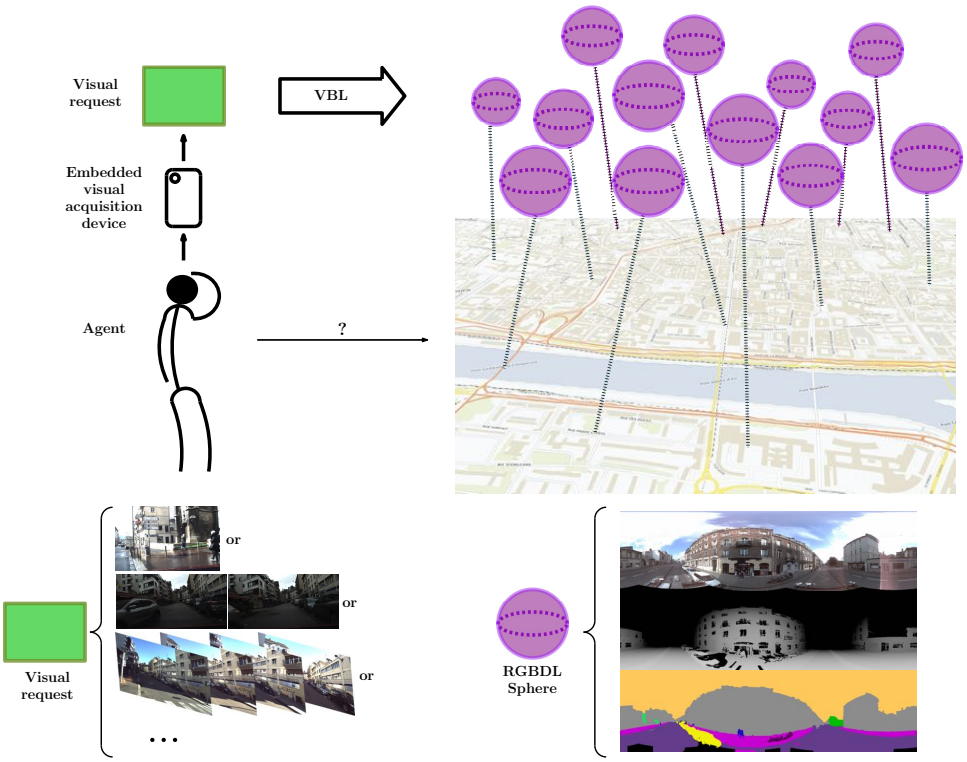


Figure 1.1: The localization task within the pLaTINUM project: we are interested in the localization of an agent on a large outdoor urban scene. The map is succinctly described by a set of geo-referenced RGBDL spheres. The agent sends a visual request from a mobile device and we have to apply VBL methods to find the closest RGBDL sphere to the agent.

Localization in a graph of spheres. Once the map has been summarized and the agent request received, the localization task can be compiled in this question: "which RGBDL spheres is located closest to the agent visual query?". In order to answer this question, we have to develop methods that can handle potentially multi-modal requests and compare them to augmented spherical images.

We present in figure 1.1 the localization task considered within the pLaTINUM project.

1.3 Visual-based Localization with heterogeneous data

As introduced in the previous section, with the work of this thesis, we aim to solve a VBL problem. In the following, we introduce the general formulation of VBL and the

particularity of our problem: the heterogeneity of the visual data.

1.3.1 Visual-based Localization

VBL consists of retrieving the location of a visual request within a known space representation [38, 333]. For instance, recovering the pose (position + orientation) of a camera that took a given photography according to a set of geo-localized images or a 3D model is a simple illustration of such a localization system [128, 263]. VBL has been an increasingly dynamic research subject in the last decade. This recent gain of interest is due to the provision of large geo-localized images database, the multiplication of embedded visual acquisition system (*e.g.* camera on smart-phone) and the limitation of usual localization system in urban environment (*e.g.* GPS signal failure in cluttered environment). Aforementioned localization problem is involved in several present-day practical applications, such as GPS-like localization system [10], indoor [50] or outdoor navigation [36], 3D reconstruction, models and databases update, cultural heritage [29], consumer photography [106, 322] and augmented reality [96]. VBL is also used in robotics Simultaneous Localization and Mapping (SLAM) for Loop Closure Detection (LCD) [91] or to solve the «kidnapped robot» scenario [66].

VBL is a very challenging problem. The main obstacle comes from the fact that the visual request we aim to localize have been taken at a different time than the database. Visual changes may occur on the observed environment during this period of time, especially if we target outdoor localization [168, 265]. For outdoor VBL, the appearance of the same scene observed from the query and the reference data can be different due to season changes [139], day-night cycle [227], weather conditions [228], mobile objects [295] (like cars or pedestrian) or urban evolution [251] (*e.g.* destruction or creation of buildings, change of street furniture). For the indoor case [290], visual changes can be generated by the modulation of the lightning [170], a rearrangement of a room, the people occupancy, etc. Differences in the request and the reference are also observable when the condition of the data acquisition differs. This can be due to the sensor architecture, *e.g.* database and reference images acquired by different cameras [174, 184] or to differences on the pose of the agent that acquire the data, resulting on important view point changes [158, 174, 293, 299, 312].

It is very challenging to address these corner case with image-only VBL. However, the use of over information, such as the scene geometry [304, 329] or a semantic understanding of the image [321], can circumvent the limitation of mono-modal VBL.

1.3.2 Heterogeneous data in VBL

VBL involves comparing a visual data for which we seek the location, the query, to a geo-located reference database. In conventional pipeline, the query and the reference data are from the same modalities: *e.g.* an image and a collection of geo-referenced images [4, 6] or a segment of a 3D model with semantic information to a the fully semantized 3D model [272]. It exists plenty of VBL methods relying on a single-modality such as radiometric information [162, 234], geometric information [304, 329] or semantic information [7] as well as methods based on multi-modal data: images and geometry [272], images and semantic [5], etc. The principal challenge comes when we observe an asymmetric representation of modalities between the request and the reference data. How to compare, or combine, data from different nature? This is a complex question, consequently, there is lack of method that benefit from heterogeneous data for the task of localization.

In this work, we are interested in VBL with heterogeneous data, *i.e.* the query and the reference data may not contain the same modalities. As mentioned in section 1.1, in the pLaTINUM project we are interested in localization of a heterogeneous visual query to a set of multi-modal geo-localized RGBDL spheres. That means we could encounter missing modalities or missing data within the queries comparing to the radiometric, depth and semantic modalities present in the reference data. We believe that heterogeneous and asymmetric data can be used wisely for the task of Visual-based Localization to overcome the limitation of single-modality systems. Therefore, within this thesis, we pursue research in this direction.

1.4 Thesis outlines

This thesis presents an original research work on the development of a new VBL method taking benefit from heterogeneous data.

An exhaustive review of existing VBL methods is presented in Chapter 2, with a special attention paid to data heterogeneity within these methods [220]. We first enu-

merate common processing among VBL methods and then we introduce the three classes of localization approaches. Open challenges as well as auxiliary data used to overcome limitation of visual-only systems are discussed thereafter.

Chap. 3 presents our learned global image descriptor augmented with auxiliary modalities and designed for the task of VBL in challenging conditions [223, 224]. We open the chapter with a brief review of descriptors for localization before introducing our global image feature extractor. The rest of the chapter is dedicated to exhaustive experiments on multiple challenging localization tasks.

In Chapter 4, we introduce a original image pose refinement step which uses missing geometric information to improve localization performances over our first method [221, 222]. Preliminary work and our final refinement pipeline are first presented. Then, we show the effectiveness of our method for localization in indoor environment before presenting furthers experiments on unsupervised auxiliary geometric learning for outdoor pose estimation.

Finally, Chap. 5 concludes the thesis and offers avenues for prospective work.

Chapter 2

Comprehensive study of Visual-Based Localization methods

In this chapter, we present the Visual-based Localization (VBL) problem by firstly focusing on the computer vision methods used in this area (sections 2.1-2.2) and, secondly, by regarding challenges in VBL (sections 2.3) and the nature of the data involved in the localization process (sections 2.4).

Topics addressed. We mainly focus on urban VBL as it represents the most studied end-user application in literature. This can be explained by the fact that most of the related applications take place in non-rural environment. As an illustration, VBL as GPS pedestrian localization system should be used when the presence of buildings disrupt the satellite signal. Most of the augmented reality applications are also designed for indoor or urbanized environment. Similar reasoning can be employed with robotic applications. Nowadays principal concerns about robots are related to human assistance or supervision and autonomous vehicles. Those services occur in indoor and outdoor man-made areas, therefore the robot localization should be studied for these sites. The other aspect that invites researchers to focus on urban environment is that large datasets are mainly describing cities or road networks. Indeed, they are the most reachable places. With the exception of airborne and satellite imageries, that are abundant all over the globe. Regarding aerial images, use cases differ as the data are usually already well organized and accurately geo-referenced. Finally, interest in urban VBL is motivated by the high number of remaining open challenges: handling the large number of different object

classes present in the scene, dealing with visual obstruction and dynamic changes induced by significant temporal variation, etc.

As well as VBL presents a heterogeneity about its end-user applications, methods and data involved in the process of localizing an image are various. These methods are divided into three categories: *Content-based Image Retrieval (CBIR) for image localization* [2, 162, 233], *6 Degrees of Freedom (DoF) image pose estimation* [33, 263] and *hierarchical localization* [257]. CBIR methods used in VBL slightly differ from classical vision object-retrieval algorithm [278] on two points: the images in the query and the database represent scenes rather than objects (*e.g.* street view panorama, buildings images, indoor scenes) and the performance of such system is evaluated according to the precision rate rather than the recall rate (i.e. a perfect VBL system should recover in its top ranked candidates documents that display the exact location of the query). On the other hand, pose estimation methods aim to recover instantly the 6 DoF pose of the query data. Where Structure From Motion (SfM) or Simultaneous Localization and Mapping (SLAM) techniques provide a *relative pose* of a sequence of data, VBL tackles the problem of retrieving the *absolute pose* of a query data according to a known representation. Nevertheless, this representation could have been built thanks to SfM or SLAM mapping module. Finally, hierarchical localization, also known as coarse to fine localization, relative pose regression or localization cascade, can be seen as a combination of the two aforementioned methods. These methods can be divided into two steps: as a first instance, the research scope is reduced to obtain a initial coarse query pose, then, this result is refined by computing the correct transformation between the query location and the first guess pose.

When designing a VBL system, the type of method is not the only parameter to consider. As pointed out in [168], robustness to environment appearance changes over time is a main concern. Data involved in the process of localization also define specifications of the system, like area covered by the VBL method or precision of the regressed pose. Data types are various in VBL: visual data, geometric information (provided by RGB-D camera, LIDAR, etc.) and semantic clues. Combination of different data in VBL aims to overcome limitation of images-only based method.

Related works. VBL is a well studied topic, and many contributions propose overviews of this domain. Brejcha and Čadík [38] present many works on VBL and classify them

depending on the environment for which the particular method was developed. Conversely, we focus our study on systems built for city-scale localization as it concerns the most VBL applications. Moreover, we propose a comprehensive description of the three types of methods used in VBL, and highlight the benefits of the use of heterogeneous data in the context of localization in challenging scenarios. Zamir et al. [333] gather recent articles to draw a large panorama of VBL, corroborating the growing importance of this domain in current research. This assumption is comforted regarding the many tutorials (CVPR2014, CVPR2015, CVPR2017, ECCV2018 and ICCV2019), workshops (CVPR2015 and CVPR2019) and challenges (Google Landmark Retrieval 2018, Google Landmark Retrieval 2019, Visual Localization Benchmark) about the Visual Localization problem in high impact international conferences.

Visual Place Recognition is a roboticist problem, defined in the general sense in [168] as the visual ability of a human, an animal or a robot to recognize an already visited place. It is a main concern for navigation, especially when we consider topological mapping [88]. Despite the fact that Visual Place Recognition shares huge similarity with VBL, the two problems differ on three major points. On the one hand, visual-based localization and visual place recognition purposes differ; where Place Recognition decides if a given place have already been seen, VBL produces a pose of the visual acquisition system. This explains the difference in their respective pipeline. Visual place recognition is composed of three main components; *the data processing module, the mapping module and the belief generation modules*, while visual-based localization does not consider the mapping module. On the other hand, the study presented here aims to consider VBL in a more general context. Communities and applications of the reviewed methods belong to the Computer Vision community [258], as well as the Robotics [88] community. Finally, we consider *heterogeneous* visual data without restriction, including: raw colour and grey-scale images, depth images, point cloud and 3D models, as well as semantic information extracted from aforementioned data.

However, we advise reader to refer to the recent surveys related to Visual Place Recognition [88, 136, 168] in order to capture a global panorama of existing approaches involved in localization process with visual data.

The rest of this chapter is organized as follows: in section 2.1, we introduce data representations used in VBL followed by a description of the three family of localization

methods in section 2.2. Section 2.3 analyses the problem of challenging association across data variability and in section 2.4 we present an overview of the different types of data used in VBL. Finally, in section 2.5, common datasets and trends in VBL are discussed.

2.1 Data Representation

What is the best manner for representing visual data? This central question, present in various Computer Vision, Robotic and Photogrammetric applications, leads up to numerous answers. The data representation, termed features, should incorporate as much as possible discriminant information from the initial visual document and be fast to compute and compare. We present in this section representations used in VBL. Table 2.1 summarizes the following presented features.

2.1.1 Local Features

Local features are widely used in VBL and more generally in Computer Vision. Their description occurs at pixel level among a local neighborhood of several points in the image. The description through local features is two-step: firstly detect salient region (the extraction phase) and then characterize them according to their neighborhood (the description phase).

Point features. Several criteria are taken into account for the selection of point features: scale, orientation and illumination invariance, as well as computational cost and descriptor vector dimension. A comprehensive list of local feature descriptors used in topological mapping in robotics can be found in Garcia-Fidalgo and Ortiz [88] survey. Krajník et al. [139] explore in-deep many combination of detector/descriptor for the specific task of images matching across seasons. The most used point feature in VBL remains the Hessian-affine detector [185] combined with SIFT [166] descriptor. Important contribution from Arandjelović and Zisserman [2] introduces RootSIFT which presents better results in matching step with minor overhead in computational load. SURF descriptors [26], light version of SIFT, are employed when real-time performance are required [64, 232, 285]. BRIEF descriptor rapidly computes a binary signature associated to a keypoint using pixel intensity comparisons. The advantage of binary descriptors is that they can be compared efficiently using hamming distance. ORB lo-

cal feature [248] counter the lake of angular invariability of BRIEF descriptor by using orientation-aware feature detector from FAST [246]. Binary descriptors are widely used in VBL [81, 102, 138, 139, 184, 193].

Learned local features is a well studied topic [45, 70, 71, 74, 204, 213, 330]. Features are detected and described through Convolutional Neural Networks (CNN) trained for the task of similar features association. Schönberger et al. [271] propose a recent comparison of hand-crafted and learned feature proposed and show, amongst others, that traditional local features perform the best in some scenario related to VBL. In [257], authors use SuperPoint [71] as local features for images localization. D2-Net [74] have also been successfully applied to VBL. Features block extracted from a CNN can also serve as densely sampled local keypoints [324] (each keypoint is extracted along the depth of the features block). This dense extraction of local features have been successfully used for VBL in [221, 290]. Attention mechanism can be added to select discriminative areas from the dense features block, as illustrated with the weakly supervised DELF [201] system trained for large scale image localization.

Geometric features. Visual data can be described by primitive geometric shapes. Despite the fact that geometric features are less compliant than point features, they include semantically meaningful information. For example, vertical lines are convenient descriptor in urban environment to represent buildings [12, 189, 238]. On the basis of this observation, Hays and Efros [106] introduce line extraction in combination with others descriptors to describe images. Works in [52] introduce a semantic line-based descriptor. The vertical lines are extracted using Canny filtering and coded into VCLH (Vertical Corner Line Hypothesis) for meaningful building corners representation. Contour extraction have also been employed by Russell et al. [250] to recover the pose of an image in a site of archaeological excavations. In the work of Baatz et al. [16], authors assume that skyline will be present in the data and use it as a geometric features to describe mountain panoramas. Dehaze segmentation is used to extract the skyline that is thereafter encoded in a curve bin descriptor.

Considering 3D data, several works use three-dimensional geometric features like normal vectors [149] or planar surfaces [83]. Bansal and Daniilidis [23] use PointRay (*i.e.* a 3D vectors aligning with an edge) extracted from a Digital Elevation Model (DEM) to represent building corners. With recent progress of Deep Learning (DL) on 3D

point cloud processing [230], local learned point feature descriptors for localization have emerged. 3DMatch [336], 3DFeat-Net [329] and PPFNet [68] are a 3D point descriptor trained for the task of 3D points to 3D points registration and used to localize a local laser scan within a reference point cloud.

Point features with geometric relations. The lack of geometric consistency across the whole image is a shortcoming associated with point features. Various contributions propose to overcome this limitation by adding local geometric information directly on the point descriptor [16, 117] or with the geometric association of numerous points [151, 163]. SIFT features contain scale and orientation information, that have been originally used in [117] through the Weak Geometry Consistency framework. Following the same idea, Baatz et al. [16] encode features relative pose in the image to perform geometric verification at matching time. Liu and Marlet [163] introduce a geometric descriptor called Virtual Line Descriptor (VLD) by connecting two local features with each other. The subsequent lines are used to reinforce the robustness of the matching process in VBL scenario [174]. Li et al. [151] propose a different pairwise geometric descriptor (PGM), showing great results on both urban and landscape scenes.

2.1.2 Global Features

Another description approach considers the image as a whole and produces one signature with high dimensionality (usually up to 4096 elements). Compared to local descriptors, global features are considered less robust in viewpoint changes, occlusion and local variations in the image. However, they are computationally less intensive to extract and capture a comprehensive description of the visual data. With the recent progress on Machine Learning (ML), a new class of very efficient global descriptor computed by CNN have emerged.

Hand-crafted features GIST descriptor introduced by Oliva and Torralba [202] is the most used hand-crafted global descriptor in VBL [15, 106, 250]. Azzi et al. [15] use features in a cascade scheme to first narrow the search scope with global feature GIST and then select the good candidates with local features SIFT. The raw image can serve as a descriptor, with systematic resizing in order to obtain thumbnail [63, 106] (potentially augmented with depth information [92]). Simple descriptor computed through an

histogram upon various criteria (color, texture [106] or depth [199]) also provides a fast global information. Taking the image as a whole in a different representation space that is more discriminant for similarity research can also be considered as global description. For instance, Fourier Transform (FT) is used by Wan et al. [316].

Learned features Democratization of CNN in computer vision domain leads to state-of-the-art techniques in image retrieval for urban scenes [6, 100, 132, 233]. Descriptors created through CNN are obtained using pooling mechanism on the computed features block [18]. We refer readers to section 3.1.1 for a detailed review on learned global image descriptor for the task of VBL.

Similar learned approaches have been recently used for 3D point cloud global description [272, 304]. Schönberger et al. [272] combine 3D convolution on point cloud (the 3D information is stocked in volumetric grid of voxels) and an self-supervised deep auto-encoder and use the low dimensional latent representation computed by the network as global point cloud descriptor. In [304] the PointNet [230] network is associated to a differentiable Vector of Locally Aggregated Descriptors (VLAD) module, called NetVLAD [6], to train in a supervised manner a discriminative global feature for fast localization.

2.1.3 Patch features

Patch features consider region of interest in the image, it can be interpreted as a compromise between local and global features. The patch could be manually extracted (with a fixed grid on an image, or a sliding window [67]) or automatically chosen in according to image saliency [178]. The discriminative HOG [67] descriptor has been used in VBL for capturing architectural cues of building and landmarks [14, 180, 189, 277]. In the work of [131, 200], MSER blob detector by Matas et al. [178] is used to extract visual information. Morago et al. [189] use a combination of local and patch features to describe repetitive shapes. Patch detector coupled with global descriptors are a common use in VBL, as illustrated in [100, 131, 287, 328]. Sünderhauf et al. [287] present promising works where the feature patches are automatically extracted with an edge boxes detector [342]. Another CNN approach is introduced to perform VBL in [100], authors use a custom region proposal network [242] to extract regions of interest (ROI) and compute a deep representation of the ROI.

Table 2.1: Features in VBL: Synthetic overview of features used in Visual-Based Localization.

Name	Feature type	Detector	Descriptor	Used in VBL
Pseudo Corners detector [188]	Point	✓	✗	[188, 189]
Hessian-affine [185]	Point	✓	✗	[2, 4, 118, 151, 262]
FALoG [320]	Point	✓	✗	[81]
SIFT [166]	Point	✓	✓	[157, 200, 269, 281, 328, 331, 332]
RootSIFT [3]	Point	✗	✓	[4, 184, 262, 300, 301]
SURF [26]	Point	✓	✓	[64, 151, 232, 281, 284, 309]
ORB [248]	Point	✓	✓	[102]
BRIEF [42]	Point	✗	✓	[138, 139]
BRISK [144]	Point	✓	✓	[81, 184, 193]
Learned descriptor	Point	✗	✓	[45, 139, 213, 221, 290]
Learned features	Point	✓	✓	[74, 201, 257]
Lines [310]	Geometric-2D	✓	✗	[12, 106, 189, 238]
Skyline	Geometric-2D	✓	✓	[16, 55, 303]
Contours [43]	Geometric-2D	✓	✗	[237, 250]
VLD [163]	Geometric-2D	✗	✓	[174]
VCLH [52]	Geometric-2D	✓	✓	[52]
PGM [151]	Geometric-2D	✗	✓	[151]
GIST [202]	Global	—	✓	[15, 106, 195, 250]
Tiny images	Global	—	✓	[63, 92, 106]
Histogram	Global	—	✓	[106, 199]
Fourier Transform	Global	—	✓	[316]
CNN	Global	—	✓	Refer to section 3.1.1
HOG [67]	Patch	✗	✓	[14, 180, 189, 277]
RPN [242]	Patch	✓	✗	[100]
Edge boxes [342]	Patch	✓	✗	[205, 287, 328]
MSER [178]	Blob	✓	✗	[131, 200]
Normal vector	Point-3D	✓	✗	[83, 149]
PointRay [23]	Point-3D	✓	✓	[23]
Learned 3D point descriptor	Point-3D	✗	✓	[68, 329, 336]
Planar surface	Patch-3D	✓	✗	[83]
Spherical function [267]	Global-3D	—	✓	[169]
Learned point cloud descriptor	Global-3D	—	✓	[272, 304]

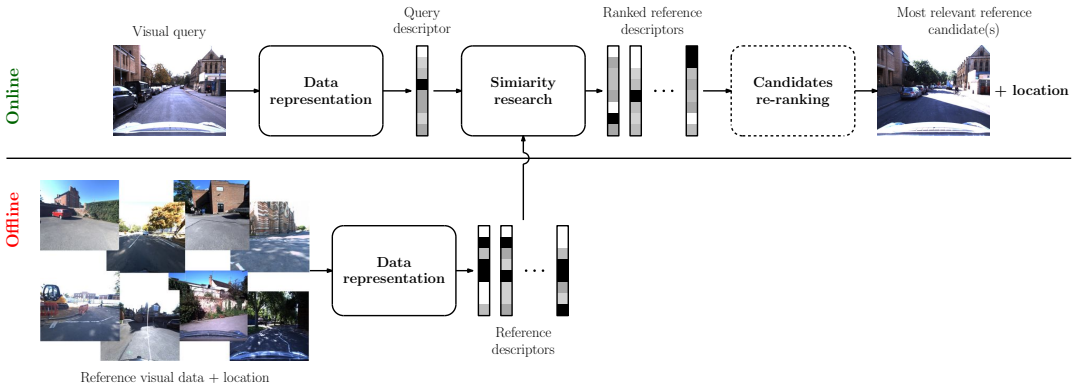


Figure 2.1: CBIR for localization: the location of a given request can be retrieved by comparing the query to a pool of geolocalized candidates. After the similarity research, the location associated to the top-ranked reference data is considered as the location of the query. Re-ranking of the reference data can be used in order to improve the relevance of the top-ranked candidates.

2.2 VBL methods

In the previous section, we have described data representation mostly used in VBL. The current section is dedicated to the method built on this representation to perform localization. As mentioned in the introduction, it exists three main family of methods:

- Content-based Image Retrieval (CBIR) for localization,
- 6 DoF camera pose estimation,
- coarse to fine localization.

2.2.1 CBIR for localization

The aim of CBIR methods is to retrieve a set of data presents in the database that are similar to an input query. This is a problem related to instance retrieval [337]. As the visual data used in VBL are augmented with geospatial information (*e.g.* a geotag associated to an image), retrieving documents comparable to the input provides an information on the possible location of the query. This localization method is three-step: description of the visual data, similarity association across the description vectors previously extracted and possible candidates re-ranking. CBIR for localization pipeline is illustrate in figure 2.1.

Efficient data representation for localization

As one of our main contribution in this thesis targets the design of a new learned global image representation for long-term CBIR for localization, we do not detail data description in this chapter. We refer reader to the first section of Chapter 3, Sect. 3.1, for a comprehensive review of image representation suited for localization by indexing.

For the subsequent tasks of the localization process, we assume that we are able to produce low dimensional vectors to describe the visual data.

Similarity Research

The similarity research step involve evaluating the sameness between the request descriptor (*i.e.* the description vector computed from the visual request we want to localize) and the reference descriptors. At the end of this step, we obtain a list of reference candidates, ranked accordingly to their similarity to the request.

Pre-processing. Dimension reduction of descriptor is often performed to reduce matching time and memory footprint. The most used technique remains the Principal Components Analysis (PCA). PCA is applied on high dimension vector, *e.g.* features block extracted from CNN layers ([6, 100]). PCA has also been used to reduce the size of local features aggregated vectors [131, 301] or global descriptors [199]. Gaussian Random Projection is applied in [205, 287] and in a different work, binary locality-sensitive hashing [286] is used instead. To reinforce data consistency, whitening could be applied to final features before the similarity search [6, 99, 100, 115, 233, 298].

Similarity metric. For most methods, comparison between descriptors is a trivial operation: it consists in a simple euclidean distance computation between vectors (with $L2$ norm or cosine similarity – *if the vectors are unitary* – as usually used function). Area correlation is another approach for computing data similarity. Simple forms of correlation like Sum of Squared Difference (SSD) or Sum of Absolute Difference (SAD) have been used in VBL to directly compare raw images [187, 225]. Wan et al. [316] use PC (Phase Correlation) on images described with FT (Fourier Transform) in order to be robust to shadow artifacts. In the work of [63], authors compare shadow invariant grey-scale images with Zero Mean Normalized Cross-Correlation (ZNCC).

Neareast Neighbor search. In some works, when the amount of data to compare remains acceptable, linear or brute-force retrieval procedure can be employed to retrieve the closest neighbors. Well-suited data structure (*e.g.* inverted-index or k-d-tree) can be used in order to speed-up the research process. Linear NN search is used in [6, 18, 100, 233, 286, 287, 331, 332], among others, where low-dimension global image descriptor (§2.1.2) are used to describe the data.

Exact nearest neighbors search becomes impracticable when the amount and/or dimensionality of the features are too large. Authors then turn to approximate nearest neighbor search to trade efficiency for rapidity, thus accepting some errors in the retrieved neighbors. Approximate matching involve hashing methods [95] and quantization frameworks [119, 200, 218]. Interested readers may see [317] for more details.

Several NN search algorithms are efficiently implemented in the `FLANN` library [194], and in the new Facebook `FAISS` library [124].

Machine Learning matching methods. Learning the distribution of the extracted features is an alternative to aforementioned NN search methods.

Support Vector Machines (SVM) classifiers are used in numerous works [14, 44, 160, 180, 277] to cast the similarity research as a classification task. Cao and Snavely [44] initially cluster the database according to the resemblance of the images. On top of this graph of similar images, they trained SVM for each cluster and at query time oppose the input image to all classifiers. By selecting the data associated to the SVM reaching the higher score of classification, this approach permits to quickly retrieve a pool of similar images. In [14, 180] authors train linear classifiers on HOG descriptors to robustly retrieve similar images that present extreme appearances changes. Aubry et al. [14] take the advantages of Linear Discriminant Analysis (LDA) data representation in order to avoid expensive SVM training (like hard negative mining used in [131, 277]). Similarly, Kim et al. [131] train SVM classifier to predict the robustness of extracted descriptors. This improves the matching process and reduces the number of features to compare against the database.

Lu et al. [169] introduce a Multi-task Learning (MTL) layout designed for features similarity association. Works from Torralba et al. [302] and Ni et al. [199] present VBL methods that are able to localize an input query among a set of predefined places. Authors embedded the recognition process into probabilistic framework, Gaussian Mixture

Model (GMM) in [302] and epitome in [199], trained upon images representing different areas. Such paradigms allow an easy integration of additional knowledge (such as depth information [199]).

Graph matching. Stumm et al. [284] introduce an innovative method based on graph matching. The visual vocabulary abstraction is employed and augmented with a graph of covisibility of the visual words in images. The graph is constructed as follows: nodes represent visual word detected in images and edges are created between two nodes if they are seen together in a same image. This formulation permits integrating geometric relations between the extracted features. Authors use a graph kernel for the similarity comparison among the query graph and the database [283, 285]. Notice that graph-based approaches are often employed when scenes are described by spatially organized semantic clues such as office furnitures [255] or street equipments [7].

Candidates re-ranking

Data can be processed after the similarity research to improve the final result. Post-processing methods are widely used to re-rank the candidate list, improving relevance of retrieved data.

Generic re-ranking. Query expansion is a post-process that re-query the database after a first retrieval step to increase the recall rate [59, 60, 297]. However, increasing the recall rate is not the main concern of VBL indirect method [259]. Indeed, as exposed in the introduction, a perfect VBL indirect system should retrieve at first position the closest visual document present in the database. However, more suitable top ranked candidates in the list of retrieved data could benefit to a subsequent pose estimation step [281]. The VBL system presented by Cao and Snavely [44] increase the diversity of retrieved images by introducing a probabilistic re-ranking on the assumption that the first ranked candidate is not a good one and by maximizing the probability that the second one is. On the other hand, geometric consistency check is often used to reject wrong matching. Relative pose between the query and the database candidates is computed by considering homography or multiple-view transformation, and candidates that produce the most consistent pose are ranked up. Philbin et al. [218] democratize the use of spatial verification by introducing prior on the pose of the photography by assuming a

top-oriented view. Authors perform spatial check hierarchically to get more flexibility between time computation and retrieval precision. The geometric transformation between the query and the candidate is usually computed with minimal solver embedded in a Random Sample Consensus (RANSAC) [85]. There exists multiple alternatives to the classical RANSAC algorithm. PROSAC by [58], used in [72], prioritize specific features during the random selection step. We can also enumerate LO-RANSAC used in [218] and AC-RANSAC in [231, 232]. Novel method F-SORT presented by Chan et al. [53] show outstanding result both in term on matching quality and computation efficiency. Notice that these algorithms, beside improving the relevance of the retrieved candidates, can give information about the relative pose of the query. That is why numerous 6 DoF pose estimation methods, presented thereafter in Sect. 2.2.2, rely also on these techniques.

Specific VBL re-ranking. Unlike conventional methods of object-retrieval, indirect VBL can benefit from geo-localization information associated to the documents present in the database. As discussed earlier, this information can be used to construct structured graph for the similarity search process [44, 299] or exploited to re-rank the candidates list [262, 331, 332]. Zamir and Shah [331] introduce this geographic re-ranking after a classical image-retrieval algorithm to quickly remove irrelevant candidates. Authors go one step further in [332] and embed the matching process within a Generalized Minimum Clique Graphs scheme to retrieve consistent candidates according to the GPS tag associated to the visual data. Sattler et al. [262] generalize the problem of visual burstiness introduced by [118] to a geographic level, introducing the concept of geometric burstiness. They improve the relevance of the ranked list of candidates using position and popularity meta-information of database images.

2.2.2 6 DoF pose estimation

At this point, we introduce camera pose estimation methods that instantly recover the exact 6 DoF pose of the query according to a known reference. Compared to CBIR approaches, 6 DoF pose estimation methods provide a more accurate query pose to the detriment of the area coverage. From this class of methods, we consider the two following approaches:

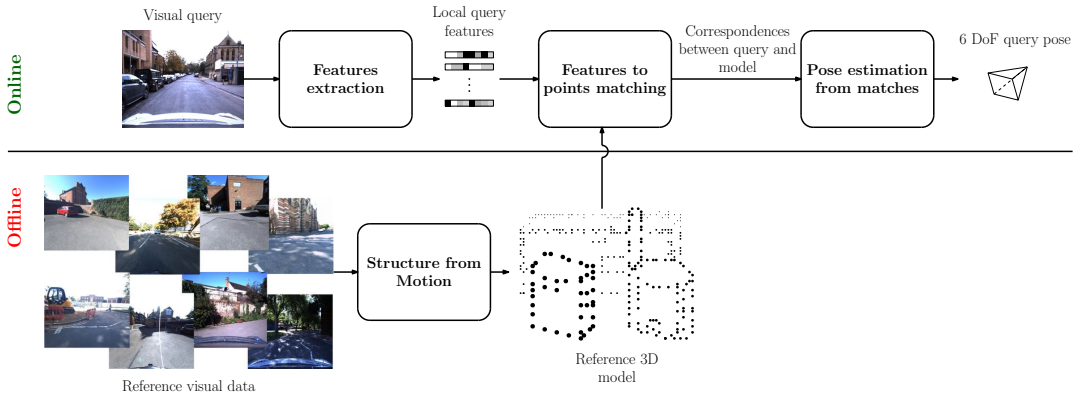


Figure 2.2: Structure-based VBL: reference data are used to construct a 3D model of the environment. During localization, the query is compared to this model to determine its 6-DoF pose.

Geometric methods: this class of methods, also known as structure-based methods, performs the global localization of the query by establishing correspondences between two-dimensional features extracted from a visual query and three-dimensional model of the environment (see figure 2.2).

Learned approaches: the second considered family of algorithms are methods that learns a model to directly regress from an input visual data to its corresponding pose. We distinguish between local [276] and global [128] learned methods.

Geometric methods

A widely represented family of VBL methods aims to regress the pose of a camera based on the analysis of a 3D point cloud reconstructed by SfM algorithms [191, 249, 270]. The principle of these methods is to establish 2D features to 3D points correspondences (F2P or 2D-3D in short terms). In a first step, three-dimensional representation of the environment is built thanks to many images. Triangulated points within this structure are associated to the local features (most of the time SIFT vectors [166]) extracted from all the images where the considered point is visible. At query time, local features from the query are matched against the set of pre-computed 3D points. Finally, 2D-3D correspondences permit a 6 DoF pose estimation of the acquisition system. Irschara et al. [111] introduce the first F2P method based on SfM environment representation. The overall pipeline of structured methods is presented in figure 2.2.

These methods have a lot in common with CBIR for localization approaches described in Section 2.2.1 as they share two major steps: feature extraction and data association. Yet, the use of a geometrically structured database introduces interesting elements not exploitable in a classical image-retrieval scheme [260].

Scalable localization. In [111], authors perform scalable VBL by registering the point cloud into synthetic visual documents covering the entire model. Sattler et al. [258] consider the original features to points correspondences scheme by [111] and introduce a Vocabulary-based Prioritized Search (VPS) inspired by BoF matching method. Heisterklaus et al. [108] introduce MPEG compression for visual document in order to speed-up the system. Works from [171, 184] tackle the problem of VBL embedded in a mobile device with limited memory storage and computational power. To achieve real-time performances, authors in [184] produce a very light 3D model to track the mobile camera in an urban environment. They send at regular interval key-frames to a server that is in charge of computing the global pose of the camera regarding a pre-produced point cloud. Aligning a light relative point cloud reconstructed with SfM to a bigger one have also been investigated in [169]. In the work described in [72], authors use the descriptor redundancy associated to 3D points to train random ferns on the top of each points. F2P matching time requirement is by the fact greatly reduced. Recent work by Feng et al. [81] reduce drastically the computational power requirement by considering fast point extractor and binary descriptors combined with an efficient similarity research. Authors show an order of magnitude in time reduction without any pose estimation performances deterioration.

Filtering wrong correspondences. Li et al. [153] reverse the conventional F2P process by searching from the point cloud correspondences in the image (P2F), instead of matching features from the image to points. This formulation causes an overhead in computation but is correctly handled by considering a compressed version of the SfM model and by implementing end-conditions and rejection cases in their algorithm. In work from [154], authors augment the P2F matching with hypothesis of co-occurrence of 3D points present in a close neighborhood. Based on similar observation, Sattler et al. [261] consider visibility graph to reject wrong matchings. Svarm et al. [288] consider the problem of VBL with F2P matching as a combinatorial optimization problem

and design a fast outliers rejection scheme. This promising work have been improved through [289, 335] contributions. In [141], authors extend the image registration problem to video registration. Temporal redundancy obtained from the visual flux facilitates the 2D-3D matching by adding smoothness constraints. Semantic filtering is also used to enforce matching consistency between features and points [295]. In [161], the proposed method achieve the best localization result on commonly used datasets without any prior assumptions on the query pose by introducing global co-visibility constraints with Markov network.

Pose estimation. F2P (as well as P2F) provides correspondences between 2D pixels and 3D points. Properly redefined by Hartley and Zisserman [105], perspective-n-point (PnP) formulation is the most common tool to recover the absolute camera pose according to a point cloud reconstructed by SfM.

Embedded in a random consensus scheme (see §2.2.1), six correspondences between the image and the 3D model are sufficient to retrieve the pose, if we have no information about the intrinsic parameters of the camera [72, 108, 153, 153]. This formulation is known as P6P and can be solved with Direct Linear Transformation (DLT [105]).

In particular cases, three correspondences between the image and the model are sufficient (P3P pose computation problem). Especially, the pose estimation problem can be reduced to a P3P formulation if the intrinsic parameters of the camera are known [111, 184], or if 3 or more DoF are fixed [232, 288, 289, 335]. In those particular cases, P3P solvers Kneip and Furgale [133] are mostly used to recover the pose.

In [282], authors are interested in privacy preserving personal information in augmented-reality consumer application. They argues that SfM point cloud can be targeted by inversion attack to recover the original state of the mapped scene. That is why they propose a 3D Line Cloud scene representation to replace the original point cloud and use 3D points to 3D lines association to recover the pose of a query image.

Learned methods

The last class of 6-DoF pose estimation methods cast VBL as a machine learning problem. A model is trained with geolocalized visual data in order to be able to predict the 6-DoF pose of an unknown visual request. Key components of learned methods are presented in figure 2.3. If the scene geometry is completely known, local method can

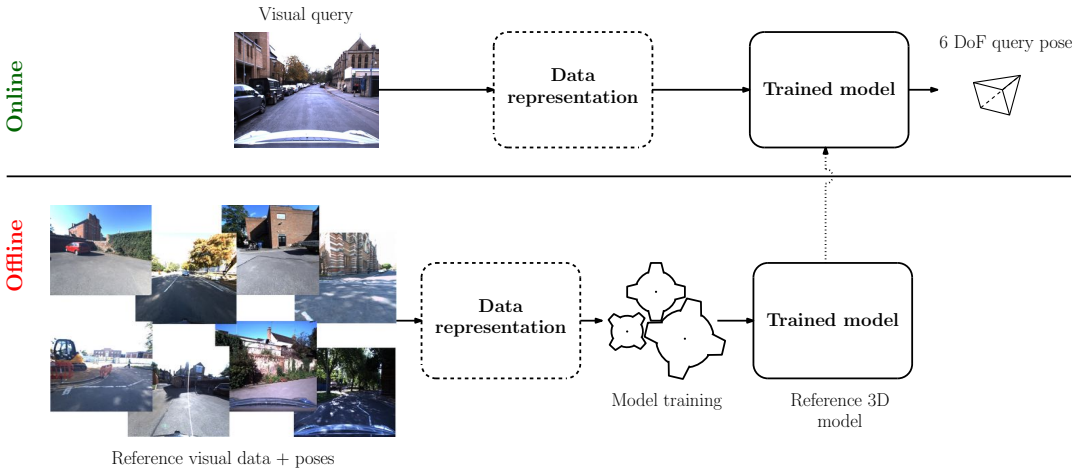


Figure 2.3: Learned method for VBL: a model is trained with geolocalized data in order to regress the right pose of a visual request. Once trained, the model is used to predict the pose of a unknown query. The data can be preprocessed (dotted blocks) before the evaluation by the model.

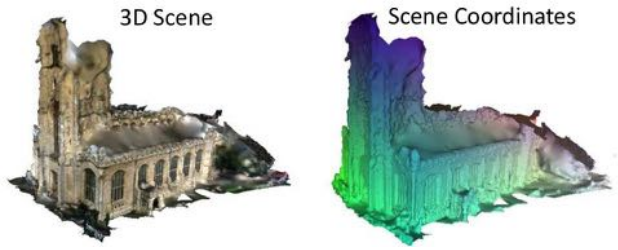


Figure 2.4: Scene coordinates repesentation: local learned-based VBL method rely on scene coordinates representation of the environment to retrieve the 6-DoF pose of a query. Model on the right is colorized according to the xyz position of the 3D points mapped in RGB color space. Figure from [33].

be applied [276] (*i.e.* the data involved in the learning processes are local feature or raw pixels). Otherwise, the method is global and entire images are used as input to the model [128].

Local methods. First introduced by Shotton et al. [276], local learned method for VBL are based on the scene coordinates representation of the environment. Scene coordinates representation associates to each pixel of an image its 3-dimension coordinate in a global scene frame. It means that we requires both image 6-DoF pose as well as scene geometry (*e.g.* depth map associated to the image) to create a scene coordinates representation. See figure 2.4 for a colorized example of scene coordinates representation. During training, we minimize the error between predicted 3D coordinates of each pixel

of a training set of images with the ground truth scene coordinates. Once the model is trained, we can use it to predict dense 3D coordinates at each pixel position of a new image. Finally, these 2D-3D correspondences (2D position of pixels' image and predicted 3D coordinates) are used to compute the 6-DoF of the query image. Similar algorithm as the ones used for structure-based methods (see section 2.2.2), embedded in a random consensus, are used in order to compute the query pose. This class of method is compact but not scalable: it requires to train one model by scene.

In the initial works by Shotton et al. [276], authors use a regression forest to learn the mapping from pixel to 3D scene coordinates. At query time, a handful of pixels from a depth camera frame are processed into the regression forest. This method is fast and precise and can be used on texture-less data. However, the depth information associated to each pixel is needed at test time. This initial method have been improved in [104], where authors take in consideration several candidates for the final pose regression obtained by multiple trained predictors. Valentin et al. [307] introduce mixture of Gaussian to represent the uncertainty associated with the regression forest prediction and significantly improve the 6 DoF estimation by embedding this information within the full camera pose regression step. The regression forest have been replaced by Neural Network (NN) in [177], bringing slightly better result at the cost of computational overhead. Meng et al. [182] consider only RGB images at query time. The loss in precision is compensated by a post pose refinement step based on nearest neighbors search with sparse extracted SIFT features (like in structured approaches, see section 2.2.2).

Duong et al. [73] introduce a patch-based approaches where area centered on keypoints extracted from standard image detector are used instead of raw pixels. During training, they use a CNN to predict 3D-coordinate from the extracted image patches. In major work of Brachmann and Rother [33], the complete image is used as input to a fully-CNN. The model predict a dense scene coordinates image and the authors perform end-to-end training regarding the final pose prediction thanks to a differentiable version of RANSAC (DSAC [35]). They obtain convening results using only images and they show that their method can be applied even if we do not know the full geometry of the scene. In [152], a new loss function is introduced to stabilize the training of fully-CNN for dense scene coordinates prediction.

To address the scene-dependent limitation of classical methods, Glocker et al. [97] design a system based on regression ferns to quickly associate an RGB-D image to a

binary feature. Ferns produce descriptor according to randomly initialized binary rules, and a look up table is maintained to directly associated image signature with 3D pose in the scene. Presented system is less precise that the one presented by Shotton et al. [276] but has the advantages of not relying on a heavy pre-processing step (*i.e.* the spawning of the regression forest). Along the same line, promising work of Cavallari et al. [49, 50] propose to quickly adapt a pre-trained regression forest to a new scene. This method permit to recover the pose of a RGB-D camera frame, more precisely than Glocker et al. [96, 97], without the costly step of complete regression forest training. Recent work in [51] applied a variant of the precedent method with fully-CNN architecture of [33] instead of regression forest. They fill coordinates reservoir at the end of a pretrained model with information about a new scene to quickly obtain a precise localization system. In a different manner in [80], authors propose to use multiple models distributed all over the area of interest. At query time, they use a Mixture of Export (MoE) for weighting predictions of each model, embedded in a differentiable RANSAC [34, 35] for end-to-end training.

Global methods. Introduced in 2015 by Kendall et al. [128], the first global learned method, PoseNet, consists of a fine-tuned CNN for the task of 6-DoF pose regression. The network is trained upon a set of paired image/pose and directly regress the 6-DoF pose of a camera from an image. The pose obtained through this method is not as accurate as the pose obtained with geometric or local method [33, 81] but provides great tolerance to changes in scale and appearance. Compared to scene coordinates-based methods [152, 276], CNN seems more appropriate to handle large environment and does not rely on depth information.

Recent improvement have been proposed by the original authors [126] to integrate an uncertainty estimation in the regression process. Liu et al. [164] consider this CNN architecture with only depth map information, acquirement with active depth sensor, for recovering the pose of a camera in complete obscurity. The work by Walch et al. [314] present a combination of a PoseNet [128] with a Long Short-term Memory (LSTM) units plugged at the output of the network in order to encode stronger spatial information from the image. This combination slightly improves the precision of the system. Authors of [122] propose a new method to gather supplementary image/pose pairs for the network training. They generate artificial images from a dense point cloud model obtained by SfM

thanks to a rendering software. Computer graphics shaders effects are added on some rendered views for simulating various illuminations. In [229], the view synthesis procedure is extended to real images thank to the known geometry of the scene. Contreras and Mayol-Cuevas [62] exploit this CNN architecture in order to create a fixed size map that can be improved by adding new trajectories. Authors were able to reduce the original size of the CNN by factor of three while maintaining similar localization performances on indoor scenes. Recent contribution [127] investigate new loss-functions for the training phase of the CNN, by adding constraints from multi-view geometry standards [105]. MapNet [37] introduces relative constraint between two consecutive images in order to train the regression network with unannotated data. For this training setup, the relative pose between two images is computed by a Visual Odometry (VO) algorithm [79, 86]. In a similar manner, VLocNet [306] benefits from relative pose information between consecutive frames by relying on two different networks with shared representation: one for absolute pose regression and the over one for relative pose estimation between two images. In a subsequent work [236, 305], authors show how multitask [129, 135, 334] learning can improve the pose estimation.

Although Posenet-like method has the advantages of being lightweight and relies on only-images, Sattler et al. [266] show that performances of such methods are less precise than CBIR-based pose estimation (section 2.2.1). They demonstrate that learned pose regression method are more likely to average the pose of the training examples [300] rather that computing a real pose based on geometric constraints. Another disadvantage of Posenet-like methods, and more generally of learned methods, rely on the fact that a different model has to be trained for each new scene.

Differently, recent work from Weyand et al. [322] consider the localization problem as a classification task. They perform a worldwide training on 126M images categorized into 26k places across the globe. According to a given image a CNN, named PlaNet, estimates the most likely location of the query over the split map. Localization of multiples photos taken from a common album can be performed by augmenting the original network with a LSTM layer. Vo et al. [313] push further the study of such a neural network and conclude that the features extracted from layers of PlaNet are more discriminative to determine the location of an input image that the CNN classifier itself. By extracting features instead of using a classification algorithm, their contribution is closer to the original CBIR world-wide localization method IM2GPS [106] (see section 2.2.1). Seo

et al. [273] introduce CPlaNet and use combinatorial partitioning in order to improve fine-grained localization. They use several models trained on different earth partitioning in such way that the final query location can be determine by considering the overlap of each model prediction.

2.2.3 Coarse to fine localization

The last family of VBL method we consider here are approaches called equivalently: coarse to fine, hierarchical or cascaded localization. In a word, the localization process is divided in two parts: initial coarse localization followed by a 6-DoF precise and relative pose estimation on the reduced search area.

Initial localization

Initial step of the localization; its provides a coarse pose or position information or multiple location hypothesis.

External sensors. The location prior can be obtained through extern sensors, like a GPS that provide 3-DoF position information [8, 9, 12, 54, 225]. In order to recover the orientation as well, magnetic compass can be used [288, 289, 335].

CBIR initial localization. Most of the hierarchical localization methods rely on CBIR to obtain an initial localization [74, 244, 250, 256, 257, 264, 264, 281, 290, 291, 299] (see section 2.2.1). In addition to obtain a coarse location information (from the geo-reference information attached to the retrieved candidates), this approach provides visually similar data that can be used afterwards for the refinement step [244, 290, 291]. The most used image descriptor for retrieval in hierarchical methods is the CNN learned global descriptor NetVLAD [6] (more information about NetVLAD can be found in next chapter).

Pose refinement

The second localization step used the prior location information to generate a finer pose estimation. This step often permits to obtain a full 6-DoF pose.

Image-based refinement. Innovative contribution from Torii et al. [299] refine the query location with a linear interpolation in the feature space domain of the closest database images. The database is arranged with a graph representation, where images represent the nodes and the edges encode spatial relation, *i.e.* images that are close to each other (according to their GPS-tag) are connected. The exact position of the query is guessed according to a linear combination of GPS information of the initially retrieved database images. Although promising, this method relies on complete panorama images, limiting its range of applications. Similar method have been used in [266] to demonstrate that Posenet-like methods (§ 2.2.2) can only average learned pose from data instead of reasoning on the image geometry to compute real 6-DoF pose.

Song et al. [281] introduce a purely image-based 6-DoF pose estimation after CBIR indexing. Authors first compute relative poses of the query image regarding two retrieved candidates by computing the essential matrix from 2D-2D correspondences [105]. These two relative poses are used together to find the exact 6-DoF query pose. Such approach has been extended in [339], where authors investigated the replacement of the conventional essential matrix estimation by a direct regression through a DL model.

In [74, 290, 291], authors establish a dense correspondences between the query image and the top-k retrieved candidates. Because the reference data are augmented with 3D depth map information (only indoor environment are considered), 6-DoF query pose can be retrieved from these correspondences (paragraph 2.2.2). This method, called InLoc, have been improved in [244] by replacing the dense correspondences operation by a CNN capable of establishing local matches between a pair of images by features block correlation.

In [21], the relative pose estimation between two images (the request and the top-retrieved candidates from initial CBIR) is directly computed by a two-streams CNN. As well as absolute pose regression described earlier (§ 2.2.2), relative pose regression with deep models is a well studied topic [78, 143, 146, 181, 251].

Model-based refinement. Arth et al. [11] present a system that recover the pose of a smart-phone camera by confronting an image to a subset of 3D points that should be visible in the query according to a prior pose information. This idea of local SfM model have been extended in [247, 264], where only a subset of the reference images (the ones retrieved in the initial step) is used to compute the 3D model. Authors of [256, 257]

use multiple position hypothesis obtained by clustering location information initially obtained by CBIR. The query image is registered on these different places and the most likely final pose is selected according to inlier count (§ 2.2.2). Shi et al. [275] use a similar approach, with the augmentation of the 3D point cloud with semantic information.

Russell et al. [250] investigate techniques to retrieve the pose of a painting or sketch according to a photo-realistic 3D model. Given a coarse pose prior, the query location is refined by establishing edges correspondences with the real model, followed by Iterative Closest Point (ICP) like algorithm. ICP pose refinement has also been used in [16], where authors register mountain images to a DEM. Similarly, Arth et al. [12] introduce a method to estimate a fine pose of a mobile camera to initialize Augmented Reality (AR) applications or SLAM systems. Authors refine the initial pose by matching geometric features to buildings outlines extracted from a 2.5D map. This method have been extended recently with advanced geometric features extraction models [8, 9, 10]. In [272], a local 3D point cloud augmented with semantic information is located in relation to a larger model. After a initial retrieval step, the query point cloud is precisely located using 3D-3D correspondences (with semantic agreement) followed by ICP.

Poglitsch et al. [225] introduce a particle filter to perform localization. The particles are randomly generated over a 3D model from a coarse position information. Widely spread in robotic community, particle filters have also been used to refine a coarse pose of a mobile robot in known ground 3D space [176] or an aerial map [41, 57].

2.3 Data with Dissimilar Appearances

As pointed out by Lowry et al. [168], permanent changes occurring in our environment is a huge concern in vision domain. In VBL, to the difference of SLAM based navigation methods [88, 168], the environment representation (i.e. the database) is most of the time acquired at a single date and query can be opposed to the system years after. To take into account local changes of the environment the database needs to be updated. Depending on the size of the covered area, database update can be a costly operation. Thus, an ideal VBL system should be able to handle minor visual changes from various sources: daily and season cycle, difference in viewpoint or modifications of the local geometry of the scene. In this section we review selected VBL papers that tackle the problem of visual changes in the environment. We dedicate the second part of the section to localization

(a) Appearance changes (b) Cross-view (c) Cross-domain

Figure 2.5: Illustration of appearance changes present in VBL system: 2.5(a) Visual dissimilarity between the query (left) and the closest image in the database (right). Cause of the change, from top to bottom: viewpoint differences [174], daytime to nighttime image matching [227] and shadow interferences from [63]. 2.5(b) Cross-view localization system [150]: left represent the ground-level query image and right the bird's eye view of the same scene. 2.5(c) Cross-domain VBL system [14]: on the left the query painting and on the right the corresponding pose according to a 3D model.

methods that consider extreme appearance changes between the query and the database, namely: cross-view and cross-domain VBL systems.

2.3.1 Appearance changes

Viewpoint changes. Common visual acquisition systems capture a part of the environment lying inside the frustum of the sensor. Indeed, perspective camera are oriented-device and due to the complex geometry of our surrounding environment, viewpoint changes in visual data can impact drastically the appearance of a scene. Visual disturbance induced by viewpoint changes is illustrated in figure 2.5(a), top images. To handle those changes, local descriptors described in §2.1.1 have been widely used. By describing partial areas of the whole scene, local features are naturally robust to a certain amount of changes introduced by difference in viewpoint and occlusion. Wan et al. [311] treat extreme viewpoints changing (when the camera are facing each other) in repetitive lunar environment. To achieve VBL in such conditions they match the ground part of the image (which is subject to large affine distortion) with a fully affine invariant feature [190].

In [91], authors use semantically weighted keypoints to match images taken with opposite viewpoints.

Image rectification [87] is also employed in VBL to minimize appearance changes introduced by different viewpoints. With strong assumption on the environment where the localization is performed (*e.g.* such as Manhattan world assumption [52, 195]), images rectification ensure that facing direction of all visual data will be barely the same. With the hypothesis of an urban scene, vanishing points can be extracted [87, 105, 145] and images rectified to display front facing buildings [12, 52, 54, 189, 243].

Other approaches consist of filling the database with additional data to cover all the possible viewpoints for a given environment. Milford et al. [187] generate translated view on a database road-circuit for preventing miss-matches if the car, carrying the acquisition system, is moving on a different traffic way than the one used to collect the database. Notice the use of a depth from monocular images DL model to produce synthetic shifted-views. Work from [14, 111, 301] increases the number of documents in the database by automatic data generation to ensure that whatever the viewpoint of an incoming query, a document displaying a similar view can be retrieved. Majdik et al. [174] perform air-ground matching of picture taken by a Micro Air Vehicle (MAV) against street view images. The main challenge outlined in this paper is the large difference in angle viewpoint. Authors generate artificial view from both the database and the query image to handle the affine transformation introduced by altitude differences (inspired by the work of [190]). Inloc [290, 291] introduces dense pose verification with view synthesis for indoor localization: artificial viewpoints are rendered from a 3D model at putative query poses.

Long-term localization. As exposed in the introduction of this section, VBL methods need to be robust to visual changes present in images taken at different times. These differences can be induced by: illumination changes across season, daily cycle, weather conditions, or dynamic changes in the scene (see figure 2.5(a) for illustration). Lowry et al. [168, Section VII] explore exhaustively Visual SLAM methods that perform strong illumination invariance place recognition (*e.g.* SeqSLAM [186, 215, 216] or FAB-MAP [64, 65, 212]). In [167], authors present an invariant-free image representation in order to overcome perturbation induced by long-term localization. Rosen et al. [245] propose a model to take in account the features persistence, decreasing the probability of

encountering a feature that have been met for the first time a long time ago. In the work of Porav et al. [227], authors use Generative Adversarial Network (GAN) to produce time-shifted images before the localization task. For instance, they convert night-time images into daytime images to improve local feature matching across the query and references data. This method can also be used for inter-season localization. Bescos et al. [27] handle explicitly non-sustainable visual elements, such as car or pedestrian, in order to create invariant image representation. They detect disruptive objects in images, remove it and finally inpaint the modified element thanks to a GAN trained on synthetic data. In [196], a adaptive learned mask is applied on the input images. This mask aims to remove non-persistent elements upstream of the image description.

As mentioned previously, methods based on local descriptors are prompt to handle local changes in images due to dynamic modifications of the environment (*e.g.* vegetation growing, buildings construction or annihilation, presence of pedestrians or vehicles, partial occlusions, etc.). Several investigations have been led for designing robust descriptors to local geometric changes. In [131, 160], authors train SVM classifiers to discriminate strong and weak local features for the VBL task. This method, and its continuation [130], shows promising results where features are more often selected when they are attached to persistent objects, such as facades, and dismissed when they represent ephemeral or changing elements, such as people or trees. In general terms, pretrained network for semantic segmentation offer strong local description for long-term localization, as illustrated in [91, 192, 272, 275, 295].

In a more robotic-oriented-scenario, Mühlfellner et al. [193] investigate map invariance representation when multiple instances of the same environment are available. On the other hand, learned descriptors show good performances if trained for the specific inter-season matching task [45]. Arandjelović et al. [6] train a CNN for global description upon images from the Google Street View Time Machine to get diverse representation of the same scene captured over a period of ten years. From this kind of representation of the environment, persistent clues can be efficiently extracted [198]. Similarly, Kumar et al. [142] proposes a CNN approach for place recognition across seasons. In Germain et al. [93], authors use a CNN as global descriptor for localization trained with explicit information about the outdoor condition of each example. Once trained, their descriptor can be used for cross-condition localization, as long as external conditions (*e.g.* daytime, rain, etc.) are known at query time.

In the following we report methods focusing on one specific problem induced by long-term localization.

Dealing with seasons & weather. Valgren and Lilienthal [309] have shown that usual local features, like SIFT or SURF, are not well suited for similarity association across season cycles. GRIEF local descriptor [139] (derivatives of BRIEF [42]) or ORB feature [102] show better results for this task. Works described in [138, 139] model seasonal-like cycle in a probabilistic framework in order to downgrade features that are not likely to appear during a given period of time. A de-raining CNN filter is proposed in [228] to improve localization with bad weather condition. Their model is trained with both synthetic and real data.

Dealing with nocturnal illumination. In some application, especially for vehicle localization, VBL has to be performed during a complete day, including overnight [180, 187] (see middle example of figure 2.5(a)). Dense descriptors' extraction used in [301] exhibit promising result for daytime to overnight images matching. At first glance, artificial lights ubiquitous in urban scene can be considered as sources of disruption. However, Nelson et al. [197] focus on this particular clues to perform localization across only night road images. Anoosheh et al. [1], with an approach similar to [227], use a night-to-day GAN to improve localization of night images.

Dealing with shadows. Some researches focus on the specific perturbation introduced by shadow casting over images. Wan et al. [316] outline that satellite and overhead images can change drastically in appearance depending on the relative position of the sun during the day. Authors show that Fourier transforms can be used to create shadow-invariant image representation. Corke et al. [63] implement the shadow suppression method presented in [84] to localize street images with important depth artefacts projected by trees or buildings. This method still remains very sensor-dependent.

2.3.2 Cross-appearance localization

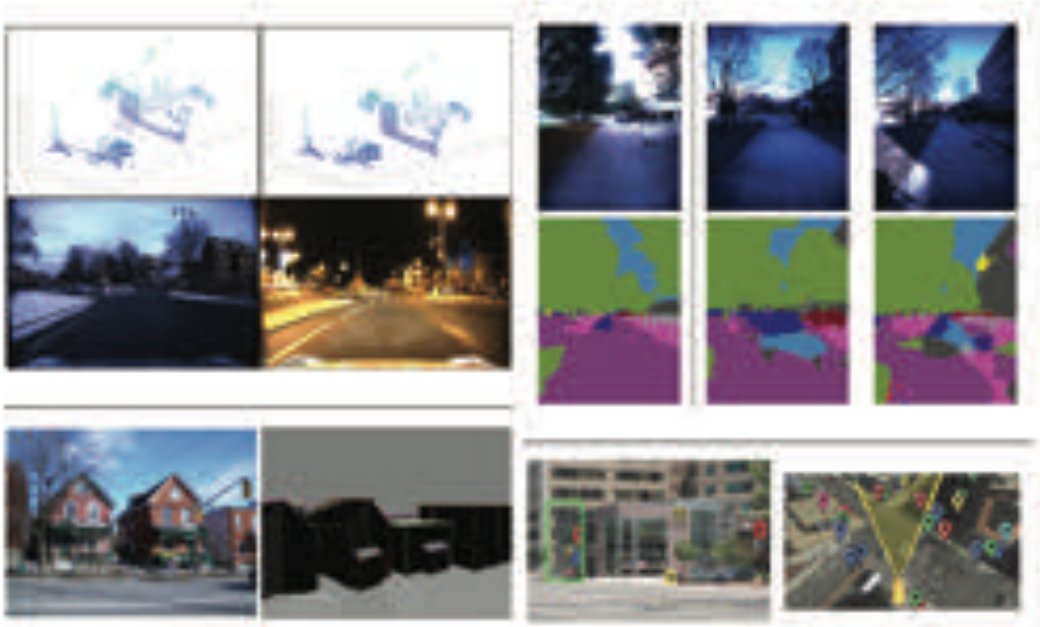
Subsequent part focus on methods that reach an extreme with change-invariance consideration by creating cross-appearance algorithms for VBL. We distinguish between two main categories of applications: cross-view VBL, where authors localize a ground-view

image against database of aerial images, and cross-domain VBL, where the purpose is to localize images of various nature (like ancient photographies, painting, sketching, etc.).

Cross-view Cross-view localization [48, 158, 293, 312, 325], also denoted as ultra-wide baseline matching [25], consider the problem of ground level localization from aerial-level set of photo shoots (see figure 2.5(b) for an illustration of data association targeted by cross-view systems). Cross-view VBL is motivated by the fact that satellite photographies are rich sources of information, available almost all over the globe. However, finding similarity between data acquired at a ground level and data captured with flying devices is a hard task due to the extreme change in viewpoint. In [312, 325], authors investigate the use of a CNN to automatically associate ground level images taken from street view service with fine-grained overhead images. Vo and Hays [312] compare several CNN architectures and conclude that triplet trained network provides the most suitable descriptors for cross-view matching. Rotation invariance between ground and overhead images is also studied through auxiliary loss and special training. Conditional GAN are explored in [241] to generate ground level image from aerial footage (or vice versa). Authors show that the latent representation learned by the CNN can be used as image feature for cross-view localization.

In [24, 25, 159], authors use bird's eye imagery to localize ground level snapshots. Bansal et al. [24] method relies on ground level images rectification, like methods focused on viewpoint changes (refer to §2.3.1).

Cross-domain Another field of research where the data association is very challenging is the cross-domain localization (an example of cross-domain VBL is presented in figure 2.5(c)). Russell et al. [250] work, followed by Aubry et al. [14] contribution, focus on the task of retrieving the pose of an old hand-drafted document (a sketch or a painting) according to a known realistic representation. In [14], hard training of HOG-based descriptors are used to capture the global shape of the architectural scene displayed in the documents, in the same manner as [277]. Results are impressive, but the used descriptor is not robust to viewpoint changes. Cross-domain techniques are also used to recover the pose of ancient photographies and to confront them with current data [19, 29]. Bhowmik et al. [29] study a new approach for pairing various local descriptors in order to increase



(a) Geometric information (b) Semantic information

Figure 2.6: Illustration of the data heterogeneity in VBL: 2.6(a) From top to bottom: PointNetVLAD [304] used to match point cloud for VBL and localization system built upon a DEM [179]. 2.6(b) From top to bottom: data segmentation to help queries (right) to database (left) visual association [272] and localization system with semantic information gathered from OpenStreetMap annotation [7].

retrieval results depending on the targeted dataset. This method have been successfully applied to match ancient photographies and nowadays images.

2.4 Data heterogeneity

Originally, pure-images were the dedicated data to VBL systems [243]. Still, conventional images for the task of localization have limitations, as mentioned in the previous section (see Section 2.3). The use of other type of data, such as geometric and semantic information, can circumvent these limitations.

2.4.1 Geometric information

The growing accessibility of geometric data promotes the development of systems based on depth information [206]. When available, depth information can directly improve the

result of VBL methods in term of robustness against visual changes and precision. We divide geometric information used in VBL on four main categories:

- **Weak Geometry:** basic primitives like plans or simplified geometric model,
- **Point Cloud:** unordered set of 3D points, triangulated from images or acquired from lidar,
- **Depth Camera:** locally dense geometry obtained by active sensors,
- **3D Model:** full geometric model covering the area of localization, hand-crafted or generated from other sources.

Weak geometry. In [54, 301], authors introduce weak geometric clues that describe principal 3D planes present in the scene. This information is then used to modify existing images in the database: for rectification purpose [54] or to generate more images in order to cover a larger area [301]. Cham et al. [52] use a 2D buildings outline map for VBL. From a given image, authors extract buildings corner and match them according to the map. Along the same line, VBL method from [8, 9, 10, 12] relies on a 2.5D city model (schematic buildings outlines boxes from OpenStreetMap[1]). 2D map is also used as geo-reference in the work from [40] (extended version in [41]) where authors produce, thanks to a stereo-camera, a path of a vehicle that is afterwards matched against the map. The matching process is embedded in a probabilistic framework to handle large environment. Baatz et al. [16] introduce the use of a DEM to perform localization in mountainous terrain [55, 237, 303]. Bansal and Daniilidis [23] extend this idea in urban localization to perform purely geometric VBL with images as query input and city DEM as reference. These purely geometric descriptions, also used in [57, 179, 237, 238] (see figure 2.6(a)), permit localization independently of the illumination conditions.

Point cloud. Previous section 2.2.2 emphasizes the growing importance of colourized point clouds obtained by SfM in VBL [81, 111, 153, 169, 171, 184, 258, 259, 261, 263, 288, 289, 335]. The addition of geometric relation by SfM improves retrieval performances [260] and permits precise pose estimation of the query, on the contrary of methods based on simple images collections.

[1] https://www.openstreetmap.org

In addition to the structured-based methods, some works focus on the localization of non-photogrammetric point cloud (*i.e.* not create from images with SfM algorithms), for instance acquired by laser scans [68, 77, 272, 304, 329, 336]. From these methods, we can distinguish between global [272, 304] and local [68, 77, 329, 336] point cloud description. Fluctuation in point density is handle trough 3D automatic completion with an autoencoder in [272] and by projection the 3D points in 2D in [77]. In order to be more robust to view point changes, methods in [68, 329] process points in an unordered manner. By considering only the geometric structure of the scene, such methods are prone to handle radiometric changes present in long-term localization scenario, as illustrate in [272, 304] (see top of figure 2.6(a)).

Depth camera. SfM reconstruction is a costly operation and laser sensor can be expensive and cumbersome for embedded VBL application. Depth cameras permit a direct and cheap perception of the 3D geometry of the scene (*e.g.* stereo camera, IR pattern projective camera, Time of Flight (ToF) camera, etc.). Raw data from those depth sensor are often used to add supplementary information channel to the VBL system. Works from [180, 199, 315] use disparity map from stereo camera. Several authors [97, 104, 276] used active depth camera that project infra-red pattern to train regression forest for localization. Similar technology is used in [147] to perform VBL in complete obscurity. In the work of [83], authors consider planar surfaces extracted from an RGB-D live stream as sustainable information for localization. The plane are organized within global graph where the pose of the camera can be retrieved quickly through sub-graph to global-graph matching. However, depth camera are not well suited for outdoor uses, reducing the range of applications to indoor scene.

3D model. Consistent 3D models (*i.e.* watertight 3D reconstruction) are also used to perform VBL task. For indoor localization, works from [207, 276, 290, 291] use textured model reconstructed from RGB-D sensor [276] or hand-crafted with dedicated software [207]. Salas-Moreno et al. [255] introduce Computer Aided Design (CAD) models to describe objects in an indoor environment and recover the pose of a depth camera when the pose tracking is lost. City-scale models are used by [14, 47, 208, 209, 225] to perform outdoor VBL.

2.4.2 Semantic information

Robustness and precision brought by geometric information has a significant cost in term on data acquisition, processing power and storage needs. Nevertheless, there is a good alternative and discriminant data representation: the semantic information. In addition to being generic regarding the original "raw" scene representation, semantic abstraction permits a discriminative and robust description of the scene. Semantic information used in VBL are classified between two classes: segmentation and categorization. Segmentation involves local methods that recognize within a data sub-parts with a semantic meaning (*e.g.* object detection in an image). On the other hand, categorization can be seen as global descriptors that associated semantic labels to a given data (*e.g.* scenes interpretation [69]).

Segmentation. In [7, 48, 57, 321], authors use object to object semantic correspondences to directly recover the pose of the query (illustration on figure 2.6(b)). Weinzaepfel et al. [321] create an object-of-interest database for localization. They demonstrate the robustness of their method to illumination changing on a synthetic museum dataset, where the object-of-interest are the paintings in a gallery. In [169], semantic segmentation is used to narrow the search scope by aggregating information about the detected objects in a room.

In a different manner, Arth et al. [12] present concrete application of semantic segmentation for localization by extracting building primitives to correct pose hypothesis (the image is segmented with a SVM classifier). In following works [8, 9, 10], the image localization is performed by segmenting principal semantic components (facades, normal to facades, corners, etc.) of a building and by optimizing the query pose based on this segmentation and a map. They use CNN trained in a weakly supervised manner to densely extract the architectural elements of buildings. On the other hand, several methods rely on annotated map [13, 318] or Geographic Information System (GIS) [7, 48, 231] to guide the localization.

Works described in [5, 192] consider the re-weighting of extracted local features in image according to the semantic class of the pixel obtained by image segmentation. Using this information, authors reduce the influence of local features that are not semantically

robust for VBL, like vegetation or cars. Thanks to progress in DL, specially in image semantic segmentation, new methods based on CNN have been successfully used to enforce coherence of local matching between images for localization [275, 291, 295].

Schönberger et al. [272] used latent representation computed by an autoencoder as global features for localization. They show that incorporating semantic segmentation information into the model representation improves significantly the localization performances of their descriptor. Similar conclusion are drawn in these following works [236, 274] In [90, 91], authors explicitly model the semantic classes of feature associated to a given region for localization of images with extreme view-point changes.

Categorization. Scene categorization [326] is a different manner to exploit semantic clues for VBL. High level semantic features have been popularized with the augmentation of labelled data and the accessibility of high computation power devices (GPU, Cloud Computing). ImageNet challenge introduced in 2009 by [69] permits the emergence of fast and robust classification methods, like the one described in [140]. Image classification produces a sub-sample of semantically identical images associated to a class. In VBL, classification can be used to decimate the database in order to proceed in a subsequent step to a more precise pose search. This method is successfully applied in [286], where the used CNN has a dual-purpose: narrowing the search scope by semantic labelling and producing a global descriptor by weight aggregation. In [199, 302], classical learning methods like Gaussian Mixture Models (GMM) or epitome are employed for associating images to a finite number of possible locations. Recent work from [89] use semantic categorization in order to establish transition probabilities from a given type of environment to another one. Authors embed this framework in the SeqSLAM algorithm, improving the global system accuracy. Finally, works presented in [273, 322] consider the problem of world-wide localization as a classification task. A CNN is trained to predict, from an raw image, its most probable location among a fixed number of Earth places.

2.4.3 Other modalities

Geometric and semantic are the principals modalities used in common with radiometric information to improve VBL. Yet, some works rely on thermal imaging [170], infrared sensor [30] or polarimetric cameras [239]. Thermal imaging makes possible the pose estimation of a camera in challenging low-light scenario [170]. On the other hand, sun light

waves measurement from polarimetric cameras enables attitude estimation in unknown outdoor environment [239].

2.4.4 Cross-data localization

We have presented so far three different kinds of information that can be used for VBL: radiometric, geometric and semantic. These types of data are commonly used together to improve localization. In this part, we consider the scenario where all types of data are not available at query time, for instance if the database uses more complete representation of the environment than the query input. It is a common scenario because some data are more difficult to acquire or required specific sensors (*e.g.* geometric information). In this case, methods have to deal with asymmetric representation of the environment in term of data type. We denote this problem cross-data VBL and classify the methods founded in the literature in two categories: methods using a common description regardless of the type of data and methods projecting one type of data within another data representation space.

Common description. Features to Points (F2P) VBL (see §2.2.2) oppose 2D images to 3D point cloud. In fact, all the features are exclusively extracted from images. On the other hand, semantic abstraction permit cross-data comparison by considering semantic object extracted from various types of data: images to 2D building outline map [52], images to map [7, 41, 48, 231], RGB-D data to DEM [57], etc. Referring to a similar physical entity, not necessary semantic, is also a manner to link information from various types of data. Images to DEM correspondences is performed in [23] based on a method relying on purely geometric clues extracted both in the image and the model. Recent work from [148, 279] use joint descriptors to merge RGB and depth data into a single feature.

Data projection. Another widely used method for combining data of different types consists of projecting one of the engaged data into the representation space of the other. For instance, lot of methods consider the challenging problem of registering photographies upon 3D models [12, 16, 128, 207, 208, 209]. Similarity comparison is performed thanks to synthetic images generated from the 3D models [14, 176, 225, 250, 290, 291]. Notice the synthesis of skyline profiles from DEM in the work of Baatz et al. [16]. In [14, 92, 111, 301],

special attention is paid to placement of the artificial cameras that generate virtual 2D views.

Generative models that create auxiliary modalities from images [334] are promising tools for cross-data localization. CNN that deduces normal direction (derivative of the depth map) and semantic segmentation from raw images have been successfully used in [291] for indoor localization.

2.5 Discussion

As VBL panorama is wide and varied, we first propose a review of recent datasets and evaluation metrics used for comparing different approaches. Afterwards, we highlight common usage and promising research avenues in VBL.

2.5.1 Datasets

Commonly used datasets in VBL are presented in table 2.2. Because of the important difference between 6-DoF pose estimation and CBIR for localization, there exists two kinds of datasets used in VBL: list of images (with basic position information or landmark/place tags) and strongly structured datasets (that can be composed of point cloud or fine geo-referenced images). Notice the growing number of publicly available datasets starring complete 3D scans of large cities [172, 183, 319]. As mentioned earlier, long-term localization in changing environment is an hot topic in robotic research. We observe therefore appearance of several datasets featuring multiple acquisitions of the same place over long period of time [46, 137, 138, 172].

For landmarks recognition in city (used method come mainly from CBIR for localization approaches), the most used dataset is the revisited version of Oxford and Paris landmarks [235]. Concerning precise 6-DoF pose estimation under changing condition, researchers prefer the recent benchmark from Sattler et al. [265] that compiles multiples datasets focused on long-term localization scenario.

Coverage and consistency. All the application relying on VBL do not cover the same area. For instance, a system design for car pose estimation should be able to localize a vehicle in a larger area than a pedestrian VBL system should do. Thus, there exists dataset with a coverage spreading from small indoor scenes to world-wide area.

Visual content within a dataset can be uniform, reference and queries are from the same sensor and captured at the same time, or inconsistent, queries have been taken under different condition or with different sensor than the references data. In order to reflect real-life conditions, outdoor datasets are often inconsistent, with visual dissimilarity between queries and references induced by acquisition sensors, dynamic changes in the scene (cars & pedestrians), weather conditions etc. Indoor datasets use more uniform data [276], thus targeting application like camera re-localization for robot navigation or augmented reality.

Evaluation metrics. Authors use various types of performances criteria in order to compare CBIR for localization methods. The recall @k, or recall @k%, is the most used metric. It represents the percentage of queries that present a good match within the k or k% top ranked images. A query is considered correctly localized if it lies inside a tolerance radius from its ground truth position (from 10 to 25m, depending on the dataset). Usually k is set to 10 or 1%. If we consider only the top 1 retrieved candidate for evaluation, distance from the ground truth is used as the main precision criteria. For places or landmarks recognition tasks [235], mean Average Precision (mAP) evaluate performances of the method.

Concerning 6-DoF pose estimation methods, authors simply compute the median (rarely the mean) of absolute position and orientation error relative to the ground truth. Another criterion can be extracted from the inlier count obtained against a robust geometric verification (for image based localization). A query is considered as successfully matched if enough inliers are found after RANSAC. However, such a metric does not ensure that the data is well localized according to the model [261]. Percentage of well localized images, *i.e.* under a given error threshold (*e.g.* 5cm & 5° for indoor scene), is also a current evaluation metric. Recently, Sattler et al. [265] introduce a more detailed metric by considering multiple level of precision (from fine to coarse localization) with different thresholds.

Table 2.2: Currently used datasets in VBL. Data type concerns the documents composing the database, not the one used as queries. RGB-D refer to data recorded with depth-cameras and RGB-S to information collected with standard cameras coupled with laser-scan. † 6 DoF available in [264].

Name	Application domain	Pose Information	Data Type
INRIA Holidays [dataset][117]	Landmark retrieval	Landmark tags	RGB
Oxford Buildings [dataset][218]	Landmark retrieval	Landmark tags	RGB
Paris [dataset][219]	Landmark retrieval	Landmark tags	RGB
Revisiting Oxford and Paris [dataset][235]	Landmark retrieval	Landmark tags	RGB
World Cities Dataset [296]	CBIR for localization	GPS	RGB
Pittsburgh 250k [300]	CBIR for localization	GPS	RGB
San Francisco Landmark [54]	Landmark retrieval	GPS† + Landmark tags	RGB
Pittsburgh Street View [332]	CBIR for localization	GPS + Compass	RGB
Tokyo 24-7 dataset [301]	CBIR for localization	GPS + Compass	RGB
Nordland train dataset	Inter-season matching	GPS	RGB
Stromovka dataset [137]	Inter-season matching	Inter-season image pairs	RGB
CH1/CH2 dataset [268]	Localization in mountain	GPS	RGB
GeoPose3K [39]	Localization in mountain	6 DoF Pose	RGB
Cambridge Dataset [128]	Camera localization	6 DoF Pose	SfM
Visual Localization Benchmark [265]	Long-term localization	6 DoF Pose	SfM
Rome16K [153]	Camera localization	6 DoF Pose	SfM
Dubrovnik6K [153]	Camera localization	6 DoF Pose	SfM
Aachen [260]	Camera localization	6 DoF Pose	SfM
Notre Dame dataset [280]	Camera localization	6 DoF Pose	SfM
7 scenes [276]	Multi-purpose (indoor)	6 DoF Pose	RGB-D
12 scenes [308]	Multi-purpose (indoor)	6 DoF Pose	RGB-D
Witham Wharf dataset [138]	Multi-purpose (indoor)	6 DoF Pose	RGB-D
North Campus dataset [46]	Multi-purpose (long-term loc.)	6 DoF Pose	RGB-S
KITTI dataset [183]	Multi-purpose	6 DoF Pose	RGB-S
CMU-Seasons [22]	Multi-purpose (long-term loc.)	6 DoF Pose	RGB-S
TorontoCity dataset [319]	Multi-purpose	6 DoF Pose	RGB-S
Oxford Robotcar [172]	Multi-purpose (long-term loc.)	6 DoF Pose	RGB-S

2.5.2 Trends in VBL

A quantitative comparison between all VBL systems is impossible due to the diversity in both methods and applications. Nevertheless, we refer reader to recent papers that quantitatively compare specific types of state-of-the-art methods. Concerning CBIR methods, following recent contributions [100, 233] show comprehensive comparisons. In [235], authors benchmark state-of-the-art methods on the Revisited Oxford and Paris dataset. F2P methods based on SfM are carefully compared on three papers [81, 263, 289]. A comprehensive comparison between 6-DoF pose estimation approaches, both learned and structured based, is presented in [266]. We refer readers to the online leader-board of the visual localization benchmark [265] for up-to-date bests methods for pose estimation under challenging conditions. In the following, we propose our qualitative analysis of VBL panorama.

Method development. As discussed earlier, there is a trade-off between the area covered and the precision reached by the VBL system; the survey of Brejcha and Čadík [38] provides a complete overview of this problem. CBIR for localization methods prioritize the space coverage, city scale [100] to word-wide [313], whereas full pose computation methods focus on precision and exact 6-DoF estimation [81] on reduced area.

Getting both wide area coverage and high precision of the query pose is the current challenge of VBL. As shown in [264], coarse to find localization systems (see section 2.2.3) are certainly a good alternative to achieve this objective. By firstly reducing the amount of data and in a second step recovering the exact pose of the query is a clever manner to target both pose precision and scalability. This cascaded localization pipeline is a well studied research area [15, 182, 247, 264, 281], and more and more recent works address the location problem in this way [94, 256, 257, 290, 291].

Benefit of heterogeneous data. All along this survey we emphasize the growing importance of multiple types of data (radiometric, geometric and semantic information) for the task of VBL. As discussed in section 2.4, using more sophisticate data aims to overcome shortcoming of radiometric based systems. Geometric data improve the final pose estimation [210, 211, 290] and are robust to radiometric changes that we encounter in long-term localization [304]. Semantic representation also offer promising results. Description at object-level is generic and compact in regard to the raw data [255] and

can be used to handle local changes in scene appearance and geometry [321]. Images, though, contain extremely explanatory clues [6] and are much more easier to collect compared to semantic and geometric data. That is why, when considered as complementary information, these three type of data offer the most effective scene representation for VBL [272, 291].

Machine learning in VBL. VBL benefits from the recent progress in machine learning. Recent global image descriptors for localization are CNN especially trained for this task [101, 201, 234] (see next chapter, section 3.1). End-to-end CNN for 6-DoF pose regression are also a growing research subject [37, 126, 127, 128, 236, 251, 314]. Regarding structured methods, classic gradient-based local features (*e.g.* SIFT) are progressively replaced by their learned counterpart [244, 257]. DL plays also an important role in the semantically-guided localization methods (section 2.4.2). High performances achieved by recent dense semantic segmentation models have permitted the emergence of novel VBL approaches [196, 272, 275, 294, 295]. VBL based on geometric information are also beneficing from DL progress. For instance, PointNet [230] has been successfully used to describe point cloud for large-scale localization [304]. New models capable of modality transfer, like depth from monocular images CNN [75], are well designed to handle cross-data localization scenario. Its can be used in VBL [291], as its have been used in autonomous navigation to improve SLAM systems [165, 292].

Runtime consideration Real-time performances and embedded architectures are constraints mainly present in the robotic community. In VBL, such criteria are not always taken into account. This can be explained by the fact that recovering the localization of an input query is a one-shoot action; *i.e.* it has to be performed only once compared to tracking systems [175] or SLAM algorithms [88]. Furthermore, as described in previous section 2.2, VBL methods are two-step: an offline and an online step. Computational time is mainly consumed during the offline step that can be computed in advance of the localization.

Yet, some authors manage to reduce the computational cost of their system [97, 171, 276]. For instance works from [56, 81] introduce a light version of F2P method, using binary local descriptors, and works from [62, 128, 322] embed their localization system in a compact CNN architecture loadable on a smart-phone. Middelberg et al. [184]

introduce a multi-scale scene representation for low-cost computation. The pose is firstly estimated according to a local representation of the scene before a global estimation on the full scene, computed on the cloud. Sarlin et al. [257] reduce the size of their CNN by training it through distillation.

2.6 Conclusion

In this chapter, we have been through the principals characteristic of a VBL system. We began by presenting common data description shared by localization method, then we introduced the three main classes of localization methods, CBIR for localization, 6-DoF pose estimation and coarse to fine localization, and we reviewed the principal VBL systems within these classes. In the second part of this chapter, we described the major challenges encountered in real-condition VBL scenarios and, in a second step, we have shown how auxiliary information about the scene, like the geometry or the semantic, can circumvent the limitation induced by the use of only radiometric data.

During the discussion, we have highlighted two major trends in modern localization systems. On the application level, the main challenge concerns the long-term localization scenario, where the queries and the reference data can be very different. From a methodological point of view, cascaded localization methods are providing the best trade-off between precision and coverage. That is why we decided to focus our research on the design of a 2-step VBL system well suited for long-term localization. In order to do so, our method will take advantages of learned geometric clues from a modality-transfer CNN. We have paid a particular attention at the implementation of our method. Thus, the proposed localization system is light and do not rely on a heavy scene representation: it is therefore suitable for embedded or robotic applications. The two following chapters, chapter 3 and chapter 4, are receptively dedicated to our new image descriptor for localization and to our pose refinement step based on learned geometry.

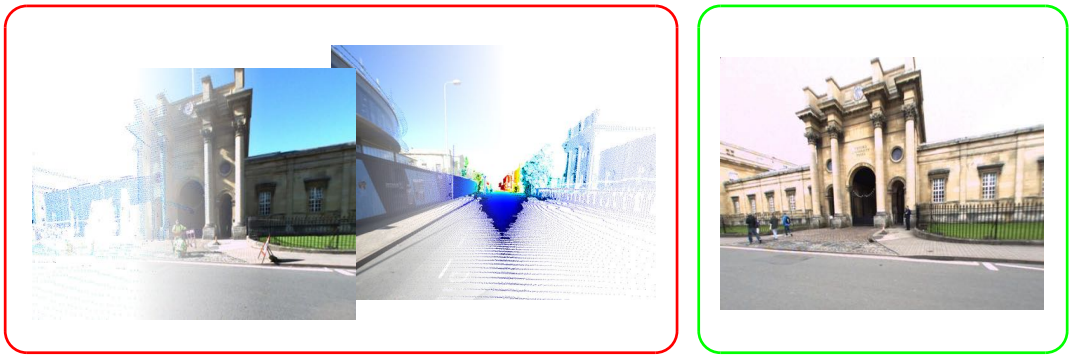
Chapter 3

Side modality learning for large-scale visual localization

In this chapter, we introduce the first step of our hierarchical localization method: a Content-based Image Retrieval (CBIR) method tuned for localization (see section 2.2.1). More precisely, we are interested in the design of an efficient global image descriptor for long-term localization.

As discussed in the previous chapter, one of the main challenges of Visual-based Localization (VBL) remains the mapping of images acquired under changing conditions: cross-season images matching [196], comparison of recent images with reference data collected a long time ago [295], day to night place recognition [301], etc. Recent approaches use complementary information in order to address these visually challenging localization scenarios: geometric information through point cloud [265, 272] or depth maps [57] and semantic information [7, 57, 196]. 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 cultural heritage when the data belong to ancient collections.

In this work, we are considering a scenario where we have an offline access to multimodal data but only-radiometric information during the online localization step (as illustrated in figure 3.1). This is a realistic scenario: a mobile mapping vehicle with multiple sensors is used once to gather initial information on the area of interest and then, agents are sending localization request with a low computational sensor like a smart-phone camera. Such setup is also in accordance with the specifications of the Cartographie Long



Multi-modal data available offline | Single-modality data available online

Figure 3.1: Data partitioning within a practical localization scenario: data available offline are richer than the one used during the localization task. We consider RGB (radiometric) + D (geometric) + R (material reflectance) as multi-modal data and only-RGB as single modality information.

Terme pour la Navigation Urbaine (pLaTINUM) project (see section 1.2), which this works is part of.

Based on these observations, 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 [110], as it uses geometric and radiometric information only during the training step and pure radiometric data for image localization.

The chapter is organized as follows. In section 3.1, we first revisit recent works related to image description for localization and side information learning approaches. In section 3.2, we describe steps that have led to the design of strong image descriptor trained with side depth information. In section 3.3 we give insight on our implementation and on the dataset we used and we illustrate the effectiveness of the proposed method on six localization scenarios in section 3.4. We discuss, in section 3.5, about the challenging night to day image matching problem and, in section 3.6, we present a variation of our method using dense object reflectance maps instead of depth maps. Section 3.7 finally concludes the chapter.

3.1 Related work

3.1.1 Image descriptor for localization.

As described in section 2.2.1 (figure 2.1), the first step of a CBIR as a localization method is the data description. This can be done using local descriptors, aggregating local features into a global vector or by computing a global signature from the raw image (section 2.1). In the following we review the two type of features mainly used in the literature:

- aggregated local descriptors,
- learned descriptors.

Aggregation of local descriptors

Local features (section 2.1.1) are prompt to produce a large number of descriptors for one single data, making the subsequent similarity search intractable. Hence, features aggregation is performed to reduce the dimensionality of the final descriptor vector. In VBL, the aggregation process emphasize specific features that are more beneficial for the localization task.

Quantization. Quantization methods have been widely adopted in image-retrieval domain since the pioneering contribution of Sivic and Zisserman [278]. They consider the problem of object retrieval in an image described through local features in the same manner as text document research. Words equivalent in image domain becomes local features and a dictionary is build upon a large set of features extracted from visual documents' database. These features are clustered to reduce the size of the dictionary; clusters' centroids are then called visual words. For each visual word in the dictionary, and inverted file is maintained to efficiently retrieve all the data that present this specific visual word. The Bag of Features (BoF) associates a vector of the dimension of the dictionary containing the visual word frequency of a specific visual document. With this representation, data similarity can be efficiently computed by a simple inner product of their respective visual word frequency vector.

BoF improvements. BoF original scheme [278] proceeds to a hard assignment from the extracted feature to the nearest visual word in the dictionary. However, depending on where the feature lies inside the Voronoi cell created in the clustering step, hard assignment can deteriorate the representation of the visual document. Soft assignment [219] methods have been considered by associating the features according to a linear combination of the k nearest visual words. Hamming embedding (HE), introduced by Jégou et al. [117], subdivides Voronoi cells and associates to each feature a binary signature to refine its position in the visual vocabulary. This method leads to excellent result in term of accuracy and rapidity and is still used in state-of-the-art VBL [4]. Inspired by Fisher Vector (FV) formulation [217], Jégou et al. [120] introduce Vector of Locally Aggregated Descriptors (VLAD) representation for image-based retrieval. The difference between a feature and its closest visual word is assigned to the final descriptor, instead of the visual word itself. The underlying idea behind VLAD representation have inspired various VBL methods [6, 131, 301, 328]. For instance, Kim et al. [131] introduce PBVLAD method to locally fuse SIFT features detected inside a MSER blob. Novel features aggregation method have been recently presented in [116].

Local features weighting. The weighting step is supposed to emphasize discriminative features regarding the similarity comparison. Original method by [278] used *tf-idf* weighting, relying on the occurrence frequency of the features in the database. Jégou et al. [118] handle the problem of intra and inter burstiness of visual words (*i.e.* the fact that a feature is more likely to appear in an image if it has already been detected once) by adapting the weight of the visual words before (inter-burstiness) and during (intra-burstiness) the query process. Torii et al. [300] tackle the problem of visual burstiness introduced by repetitive structures (abundant in urban environment) and introduce meta-features encompassing several similar descriptors (comparable both in their descriptor vector and their spatial location in the image). Such improvement permits a dense extraction of local features in images, bringing superior result in urban environment VBL [232, 301]. Another work from Morago et al. [189] also exploits the redundancy present in buildings facades. Recently, Arandjelović and Zisserman [4] improve *tf-idf* scheme by considering the descriptors' density in feature space. With their DisLoc weighing, 7% of the less discriminative visual words can be removed from the database without impacting the performances of the similarity computation. Mousavian et al. [192] introduce semantic

knowledge in the local feature weighting process, reducing the impact of features associated with non-relevant elements for localization (*i.e.* elements that are likely to change or disappear, such as trees and cars).

Learned descriptors

With the recent progress of image representation through deep neural network, learned descriptors have become a key component for numerous visual localization methods [94, 214, 256, 257, 265, 272, 287]. Therefore we decide to build on these recent advances and use a Convolutional Neural Networks (CNN) based image descriptor as base component of our system. We review in the following recent advances on learned representation for localization.

Off the shelf models. One of the first method based on learned model for the task of image retrieval have been introduced in [18]. Authors simply use pooling operation on the features block (= neural code) extracted from a CNN to compute the image descriptor (see figure 3.2). Although not trained for the specific task of image retrieval, such model benefits from initial training on images classification task on a huge amount of data (*e.g.* Imagenet dataset [69]). In [18, 286], authors show that the most discriminative descriptors for the task of image-retrieval, especially applied to place recognition [286], are extracted from mid-level convolutional layers instead of last fully-connected layers.

As shown in figure **??**, convolutional layers produce features block, composed of several activation maps stacked together. In order to capture a more discriminative representation from these features block, several activation map pooling methods can be applied. Maximum Activation of Convolutions (MAC) [240] reduce the features block by aggregating the maximum of each activation maps into a vector. Instead of maximum pooling, Sum-Pooled Convolutional Features (SPoC) [17] has shown superior results in image retrieval. More specific pooling method, carefully designed for localization, are presented in the next paragraph.

Learned descriptors can be combined with local or patch detectors, in order to obtain sparse representation of the data. In this case, features extracted from the image can be gathered into a single descriptor, like in the BoF framework. VLAD embedding is employed in [328]. In [205], patches are sorted according to their relative position in the image and aggregated in a Landmark Distribution Descriptor to improve the subsequent

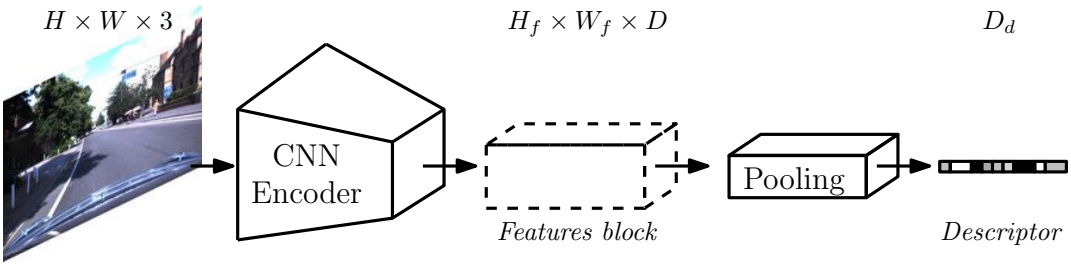


Figure 3.2: CNN image descriptor: modern learned descriptors are composed of two parts: a features extractor (encoder) and a descriptor pooling method. The features block has a lower spatial resolution that the input ($W \times H > W_f \times H_f$) with more channels ($D \gg 3$). The pooling method aims to reduce the size of the features block ($D_d \ll W_f \times H_f \times D$) while keeping discriminative clues.

similarity search. Zhi et al. [338] exploit the intensity response of each patches to discard descriptors with low intensity.

Specific architecture. Regional Maximum Activation of Convolutions (R-MAC) [298] is an improvement of the precedent MAC method, consisting of the computation of the MAC vector over regions of various sizes on the activation map. Gordo et al. [101] achieve state-of-the-art performances by combining MAC representation with a custom Region Proposal Network (RPN) that autonomously detects regions of interest on the activation map. They also add a differentiable Principal Components Analysis (PCA) layer (implemented with one fully connected layer) for dimension reduction. In [234], authors use Generalized-Mean (GeM) pooling, a trade of between mean and max pooling with a trainable parameter controlling the degree of spatial "focus" of the network. An entirely trainable aggregation layer, called NetVLAD, have been proposed in [6]. Authors design a differentiable architecture that aim to mimic VLAD aggregation scheme. In [112], authors create panorama features by aggregating multiple NetVLAD descriptor in a memory vector. Kim et al. [132] use the NetVLAD aggregation layer coupled with an Contextual Reweighing Network (CRN) to downgrade irrelevant features according to their local neighborhood, without the use of any manually annotated data. Along the same lines, Noh et al. [201] propose the DELF architecture for local features extraction. DELF relies on a self-spatial-attention mechanism to select discriminative local region on the features block.

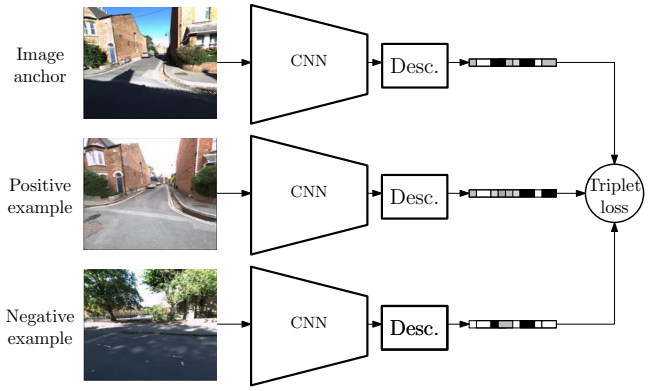


Figure 3.3: CNN descriptor training with images triplets: the triplet loss guided the training by penalizing difference of the anchor and the positive embeddings and the similarity of the anchor and the negative embeddings (see equation 3.2).

Training routine In recent works [6, 100, 233], authors tackle the problem of fine tuning a pre-trained network for the specific task of similar images association for localization. The shared idea is to construct images triplets composed of an anchor, a positive example (displaying the same scene as the anchor image with small view point or illumination variation) and a negative example (unrelated to the anchor image). An images triplet is presented in figure 3.3. Then, the training signal aims to enforce similarity between computed embedding of the anchor and the positive example and, conversely, to make the anchor and the negative descriptors far for each other in the embedding space. Multiple loss functions can be used to guide the descriptor training. Triplet ranking loss is used in [6, 101] while in [201, 233], authors choose a contrastive loss. A comprehensive evaluation of different objective functions is given by Liu et al. [162]. They also propose a Stochastic Attraction and Repulsion Embedding (SARE) loss function, that enforces both inter-place similarity and intra-place difference within the embedding space.

Arandjelović et al. [6] introduce a weakly supervised training approach, using the Google Time machine engine to automatically create large database of triplets. Work from [101, 201] make use of a large Landmark dataset for triplets creation. In [233, 234], Structure From Motion (SfM) is used to link multiple images over a large panel of places. The geometric verification provided by SfM allows to control the overlaps between anchors and positives examples. In metric learning, hard examples mining is a crucial step for creating meaningful embedding space. Hard negative mining is performed in [6, 100, 234] by selecting negative examples that are closer to the anchor in the embedding space. Iscen et al. [113] introduce a manifold distance to compare dataset examples. Using diffusion, they are able to mine more effective examples compare to standard mining methods.

3.1.2 Learning with side information

As mentioned previously, complementary modalities useful for localization, like geometry or semantic, may not be always available at test time. This could be due to limited resources (*e.g.* with embedded system), different sensors, localization of old data, etc. 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. Recent work from [150] 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 [110]: 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 [327], uses recreated thermal images to improve pedestrian detection on standard images only. Our system, inspired by [327], learns how to produce depth maps from images to enhance the description of these images.

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 [75, 98, 173]. This ability of neural networks has been used in [292] to recover the absolute scale in a Simultaneous Localization and Mapping (SLAM) mapping system. Loo et al. [165] 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. Taira et al. [291] rely on dense surface normal (derivative of the depth map) and semantic segmentation created by a convolutional encoder/decoder [334] for localization of indoor low-textured images.

3.2 Model architectures and training

Motivation. As illustrated in figure 3.4, outdoor conditions drastically impact visual appearance of a scene. It will be challenging for a descriptor relying only on the radiometric information to associate a similar embedding to the four images of figure 3.4. But, if we take a look at the underlying geometry in these images (*i.e.* the associated depth

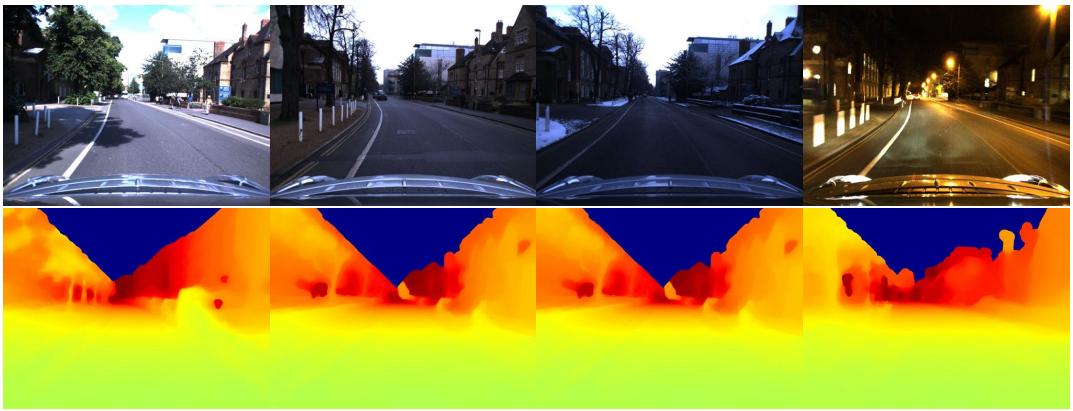


Figure 3.4: Visual changes between radiometric and geometric domain: due to outdoor conditions, visual aspect of images changes over time (top row), while, geometry of the scene (corresponding depth maps, bottom row) remains stable.

maps, bottom of figure 3.4), this information seems more stable over time. The central idea of our method is to use recent modality transfer network [75, 98, 173] (from images to depth maps) to provide invariant image representation to our CNN descriptor during training. At test time, the trained descriptor is used on images only.

3.2.1 Initial architecture

An overview of our first method can be seen in figure 3.5.

Principal descriptor. We build on recent advances in CNN image descriptor for designing our system. We use standard convolution features extractor linked to a pooling descriptor layer (figure 3.2). Formally, we denote f_p the principal features vector of image x computed by encoder E_p and descriptor P_p:

$$f_p(x) = P_p(E_p(x)). \tag{3.1}$$

We denote θ_p the weights of the image encoder and descriptor $\{E_p, P_p\}$. Notice that descriptor P do not necessary contains trainable parameters (if we consider MAC pooling method for instance).

Considering the images triplet $\{x, x^+, x^-\}$, as described in previous section (see figure 3.3), our CNN descriptor can be trained with the following triplet ranking loss [6]:

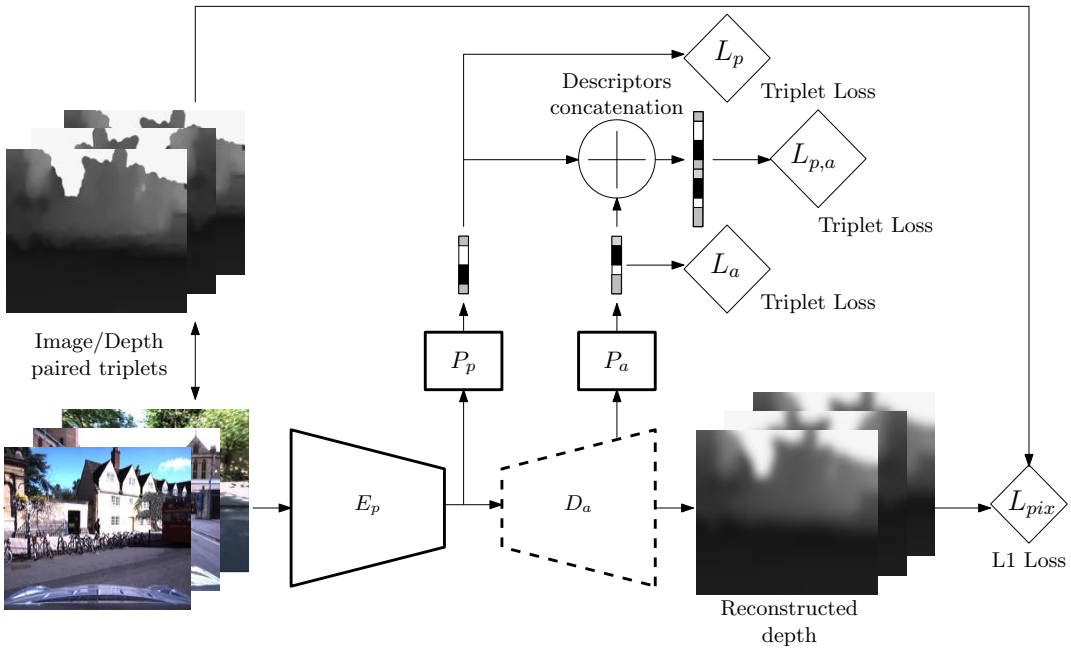


Figure 3.5: Training pipeline of our preliminary solution.

$$L_p(x, x^+, x^-) = max\left(\lambda + \left\|f_p(x) - f_p(x^+)\right\|_2 - \left\|f_p(x) - f_p(x^-)\right\|_2, 0\right), \tag{3.2}$$

where λ is an hyper-parameter controlling the margin between positive and negative examples.

Side geometry learning. In order to recover the geometric information from the radiometric signal, we use a fully convolutional decoder D_a [75]. Let $\{x, z\}$ be a pair of image and corresponding depth map, we can train our decoder to compute a depth map from an input image with the following loss function:

$$L_{pix}(x, z) = \left\|z - \hat{z}(x)\right\|_1, \tag{3.3}$$

where $\hat{z}(x) = D_a(E_p(x))$, is the output of the decoder. L_{pix} is a simple pixel loss penalizing absolute error between the output depth map $\hat{z}(x)$ and the target z.

Auxiliary descriptor. In order to take advantages of the learned depth representation in our final descriptor, we use intermediate deep features computed by D_a to create

another descriptor f_a:

$$f_a(x) = P_a(\bar{D}_a(E_p(x))), \tag{3.4}$$

where P_a is a descriptor and $\bar{D}_a$ designs an intermediate output extracted from decoder D_a. We do not use the reconstructed depth map $\hat{z}(x)$ (*i.e.* raw output of D_a) to produce vector $f_a(x)$ because it will be to sensitive to small viewpoint variations. Instead, we use an intermediate output from $\bar{D}_a$, that should be more meaningful compared to $\hat{z}(x)$ and less sensitive to viewpoint changes. Indeed, because the decoder upsample the feature maps, output of $\bar{D}_a$ has a smaller spatial resolution and is deeper in comparison to $\hat{z}(x)$. We apply a triplet ranking loss L_a (see equation 3.2) to train weights θ_a of decoder D_a and descriptor P_a.

Overall training. Finally, we combine the principal and auxiliary features, $f_p(x)$ and $f_a(x)$ in a common vector:

$$f_{p,a}(x) = [f_p(x), f_a(x)], \tag{3.5}$$

where $[\cdot]$ is the concatenation operation. Combined descriptor optimization is obtained through the last triplet ranking loss $L_{p,a}$ and the final optimization is defined by:

$$\begin{aligned}(\theta_p, \theta_a) := arg\min_{\theta_p,\theta_a} \Big[& L_p(x, x^+, x^-) + L_a(x, x^+, x^-) + L_{p,a}(x, x^+, x^-) \\ & + \frac{1}{3}\left(L_{pix}(x, z) + L_{pix}(x^+, z^+) + L_{pix}(x^-, z^-)\right)\Big]. \end{aligned} \tag{3.6}$$

Our initial method requires triplets of RGB-D data to be trained.

3.2.2 Hallucination network

We compare our method of side information learning with a state-of-the-art approach system, named hallucination network [110]. The hallucination network was originally designed for object detection and classification in images and has never been used for global image description. We adapt the work of Hoffman et al. [110] to create an image descriptor system that benefits from depth map side modality during training. Our adaptation of the hallucination network for image description is presented in figure 3.6.

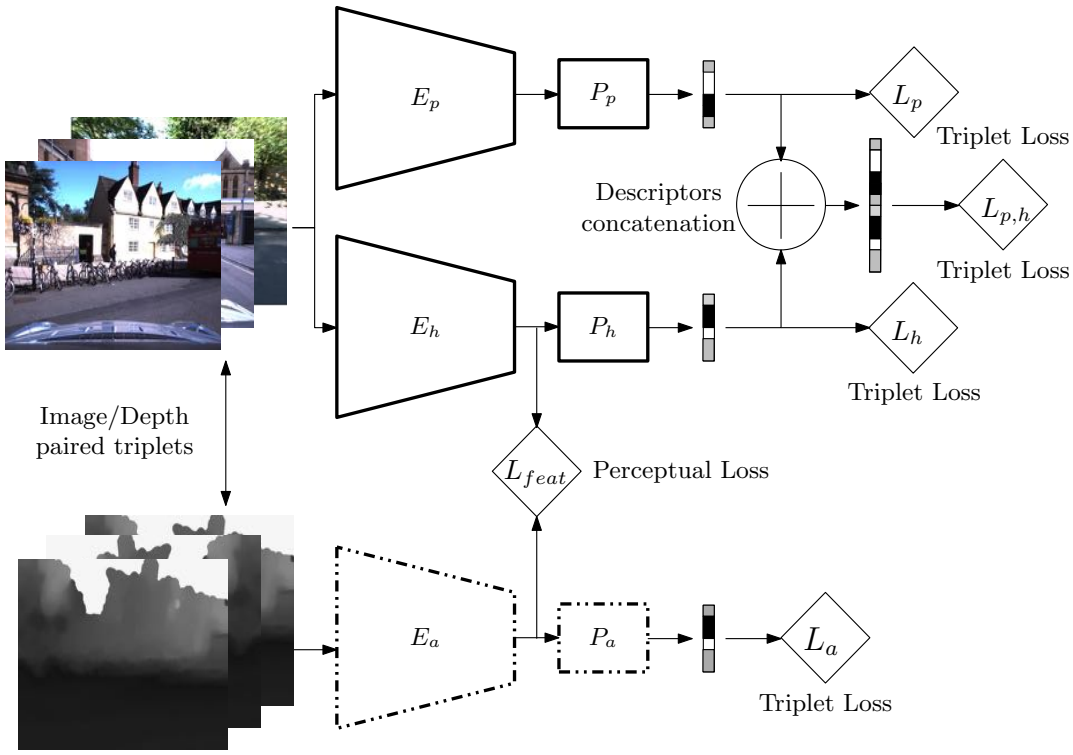


Figure 3.6: Hallucination network for image descriptors learning: we train an hallucination network, inspired from [110], for the task of global image description. Unlike the proposed method (see figure 3.7), hallucination network reproduces feature maps that would have been obtained by a network trained with depth map rather than the deep map itself.

Principal descriptor. Similar to our proposal, the system is composed of a principal image descriptor: encoder E_p + descriptor P_p, trained jointly through triplet ranking loss of equation 3.2.

Auxiliary descriptor. Hallucination architecture needs an auxiliary network for training purpose, that will be discarded at test time. This auxiliary branch focus on extracting significant information from the side modality (the depth map in our case). We design the auxiliary network similar to our principal branch: the depth map descriptor is composed of an encoder E_a linked to a descriptor P_a. The depth map descriptor is trained with a triplet ranking loss L_a, where the embeddings are directly computed from the truth depth maps:

$$f_a(z) = P_a(E_a(z)). \tag{3.7}$$

Hallucination descriptor. The key component of Hoffman et al. [110] 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 share the same architecture as the principal and the auxiliary branches. The hallucination descriptor is composed of an encoder E_h and a descriptor P_h with trainable weights θ_h. It is trained with triplet ranking loss L_h under the constraint of a perceptual loss [123]:

$$L_{feat}(x, z) = \|E_h(x) - E_a(z)\|_2 . \tag{3.8}$$

This constraint can be interpreted as knowledge distillation [109]. Final image descriptor is obtained by concatenating $f_p(x)$ and $f_h(x)$.

Overall training. Training routine presented in [110] is two-step: we first optimize weights θ_a of the auxiliary descriptor with loss $L_a(z, z^+, z^-)$ and, secondly, we initialize hallucination weights θ_h with pre-trained weights θ_a and solve the following optimization problem:

$$\begin{aligned}(\theta_p, \theta_h) := arg \min_{\theta_p, \theta_h} \big[\alpha \left[L_p(x, x^+, x^-) + L_h(x, x^+, x^-) + L_{p,h}(x, x^+, x^-)\right] \\ +\gamma \left[L_{feat}(x, z) + L_{feat}(x^+, z^+) + L_{feat}(x^-, z^-)\right]\big] ,\end{aligned} \tag{3.9}$$

where α and γ are weighting constants. During final optimization, weights θ_a are frozen.

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.

3.2.3 Discussion

Exploratory testing of our method has lead to unsuccessful results. During the training step, our network failed to produce at the same time a meaningful image representation for localization (losses L_p, L_a and $L_{p,a}$) and to reconstruct the scene geometry (loss L_{pix}). After analyzing our architecture, we came up with the following conclusion: the two target objectives are disrupting each other. This problem was due to the design of our initial method: weights modification computed by the triplet ranking losses were affecting both weights of encoder E_p and decoder D_a. At the same time, modifications induced by the loss L_{pix} (equation 3.3) were also correcting the same weights, making the system unable to converge.

We do not encounter the same problem with our implementation of hallucination network. The only loss functions that can interfere during the optimization are triplet ranking loss L_h and perceptual loss L_{feat}. Both losses lead to modification of hallucination encoder E_h weights. But targeted task of the two loss function are the same: L_h directly optimize the hallucination embedding for image retrieval and L_{feat} force the feature maps of encoder E_h to be close to the feature maps of encoder E_a, an encoder that have been trained for image retrieval task as well.

In the next section, we propose an improved version of our initial method that solves the aforementioned issue.

3.2.4 Final architecture

Our modified architecture, presented in figure 3.7, is composed of:

- a CNN image encoder E_p linked to a feature aggregation layer P_p that produces the principal image descriptor,
- a CNN image decoder D_a used to reconstruct the corresponding depth map according to the monocular image,

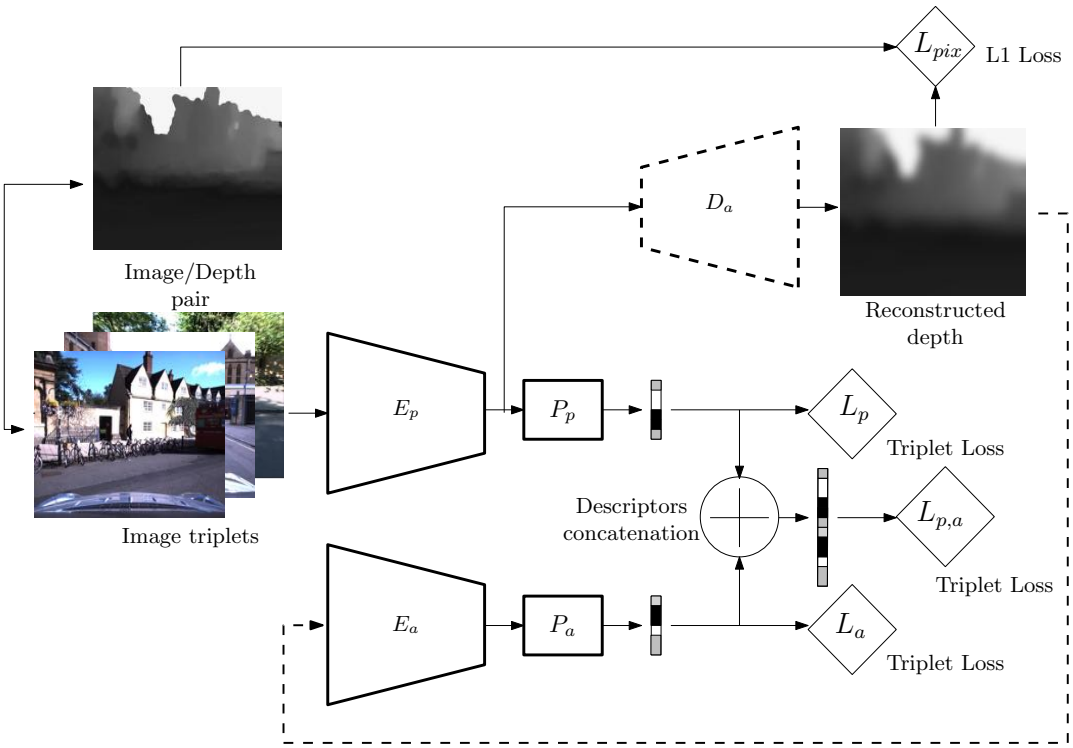


Figure 3.7: Image descriptors training with auxiliary depth data: two encoders are used for extracting deep features map from the main image modality and the auxiliary reconstructed depth map (inferred from our deep decoder). These features are used to create intermediate descriptors that are finally concatenated in one final image descriptor.

- a CNN depth map encoder E_a linked to a feature aggregation layer P_a that produces an auxiliary depth map descriptor,
- a fusion module that concatenates the image and depth map descriptor.

Training routine. Principal and auxiliary descriptor, $\{E_p, P_p\}$ and $\{E_a, P_a\}$, are trained separately with triplet ranking losses L_p and L_a. As before, loss $L_{p,a}$ is used for joint optimization. The auxiliary descriptor $\{E_a, P_a\}$ takes as input the reconstructed depth map $\hat{z}(x)$. Similar to our first architecture, decoder D_a take feature maps of encoder E_p to generate output $\hat{z}(x)$. We constrain $\hat{z}(x)$ to be close to ground truth depth map z with the pixel loss defined in equation 3.3.

We designate θ_p weights of encoder/descriptor pair $\{E_p, P_p\}$, θ_a weights of encoder/descriptor pair $\{E_a, P_a\}$ and θ_g weights of decoder D_a. The whole system is trained according to the following constraints:

$$(\theta_p, \theta_a) := arg \min_{\theta_p, \theta_a} \left[L_p(x, x^+, x^-) + L_a(x, x^+, x^-) + L_{p,a}(x, x^+, x^-) \right], \tag{3.10}$$

$$(\theta_g) := arg \min_{\theta_g} \left[L_{pix}(x, z) \right]. \tag{3.11}$$

We use two different optimizers: one updating θ_p and θ_a weights regarding constraint (3.10) and the other updating θ_g weights regarding constraint (3.11). Because decoder D_a relies on feature maps computed by encoder E_p, at each optimization step on θ_p 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 $\{\theta_p, \theta_a\}$ and $\{\theta_g\}$ until convergence. By removing E_p from the optimization of the depth from monocular objective (equation 3.11) and by adding an auxiliary encoder E_a, we get rid of the interfering tasks problem observed earlier (see § 3.2.3). Notice that even if the encoder E_p is not especially trained for depth map reconstruction, its intern representation is rich enough to be used by the decoder D_a for depth map reconstruction.

Advantages and drawbacks. One advantage of the hallucination network over our proposal is that it does not require a decoder network (D_a), resulting in a architecture lighter than ours. However, it needs a pre-training step, where image descriptor $\{E_p, P_p\}$ and depth map descriptor $\{E_a, P_a\}$ 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.

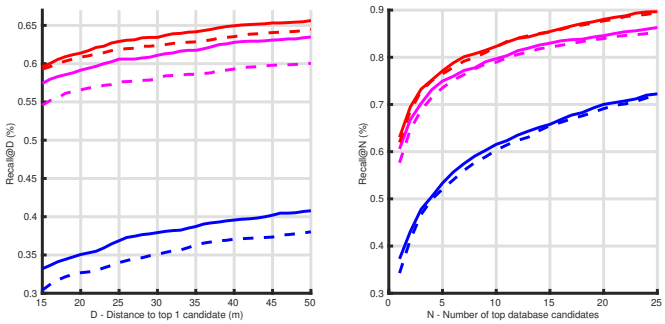
One advantage of our method over the hallucination approach is that we have two unrelated objectives during training: learning an efficient image representation for localization and learning how to reconstruct scene geometry from an image. It means that we can train several parts of our system separately, with different datasets. 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 triplets, 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 the following section 3.4.2.

3.2.5 Hard mining and swapping in triplet ranking loss

Hard negative minning policy. As mentioned in the previous section, hard mining is a crucial step in metric learning [6, 100, 113, 234]. We construct our triplets like in [6], using the GPS-tag information provided with the data. We gather N triplets $\left\{x, \left\{x_i^+\right\}_{i\in[1,M_p]}, \left\{x_i^-\right\}_{i\in[1,M_n]}\right\}$ 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 [6], 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 to the stored one. We adopt a different approach with a small overhead in term of computation but taking directly in account model updates. At each iteration, we randomly select a subset of M_n^{sub} negative examples from the entire pool, and compute the truth hard negative example from this subset. This strategy also act as regularization during training as the negative training examples are different at each epoch.

Anchor and positive swapping. We also adopt swapping technique introduced in [20]. It simply consists in choosing the most confusing pair between {anchor, negative} and



- - Individual optimization — Joint optimization
Test set: – CMU–LT – CMU–Autumn – CMU–Snow

Figure 3.8: Influence of the joint loss $L_{p,a}$: we report localization results of our method trained with or without the joint triplet ranking loss $L_{p,a}$. We detail the used metrics and datasets in section 3.3.

{positive, negative} examples:

$$L^{swap}(x, x^+, x^-) = max\left(\lambda + \left\|f(x) - f(x^+)\right\|_2 - min\left(\left\|f(x) - f(x^-)\right\|_2, \left\|f(x^+) - f(x^-)\right\|_2\right), 0\right), \quad (3.12)$$

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

$$L_{final}\left(x, \{x_i^+\}_{i\in[1,M_p]}, \{x_i^-\}_{i\in[1,M_n^{sub}]}\right) = \frac{1}{M_p M_n^{hard}} \sum_{i=1}^{M_p} \sum_{j=1}^{M_n^{hard}} L^{swap}\left(x, x_i^+, x_j^-\right). \quad (3.13)$$

3.2.6 Descriptors fusion and dimension reduction

Fusion policy. We try to replace our basic features fusion operator introduced in equation 3.5 (vectors concatenation) by more advanced functions, in order to benefit as much as possible from the complementarity of the principal and the auxiliary modalities. We investigate: hand-tuned descriptors scalar weighting, trained scalar weighting [279], trained modal attention mechanism at the level of descriptors and trained spatial and modal attention mechanism at the level of the deep features [274]. We found that all the fusion policies perform similarly. Indeed, as can be seen in figure 3.8, the modalities fusion are learned by our system through the triplet loss $L_{p,a}$, making the system aware of what is important and complementary in the radiometric and geometric domain, without

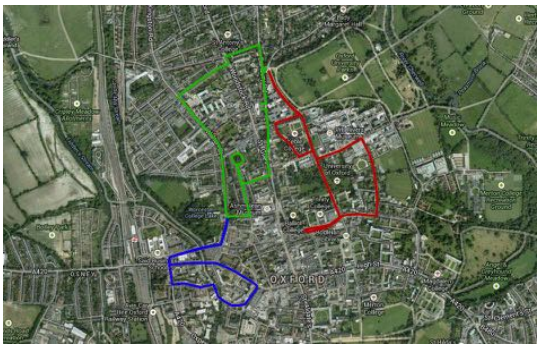


Figure 3.9: Train, validation and test zones: the green path delimits our training area, the blue trajectory the validation zone and the red path the test region. All the data are from the Oxford RobotCar dataset [172].

the need of a complex fusion method. Preliminary experiment presented in figure 3.8 also demonstrates the importance of using a triplet ranking loss on the combination of the two feature vectors f_p and f_a. In this preliminary experiment, we use Resnet18 as base encoder and NetVLAD [6] as pooling layer.

Post-treatment. After L_2 normalization, we reduce the dimension of the final descriptor by applying PCA + whitening [6, 101, 233, 234]. After the convergence of the whole system we reuse the images from the training dataset to compute the PCA parameters.

3.3 Implementation details

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

3.3.1 Datasets

We have tested our proposal on the *Oxford Robotcar* public dataset [172] and on the *CMU Visual localization* dataset [22] from the city of Pittsburg. These are common datasets used for image-based localization [265] and loop closure algorithm involving neural networks training [227] under challenging conditions.

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

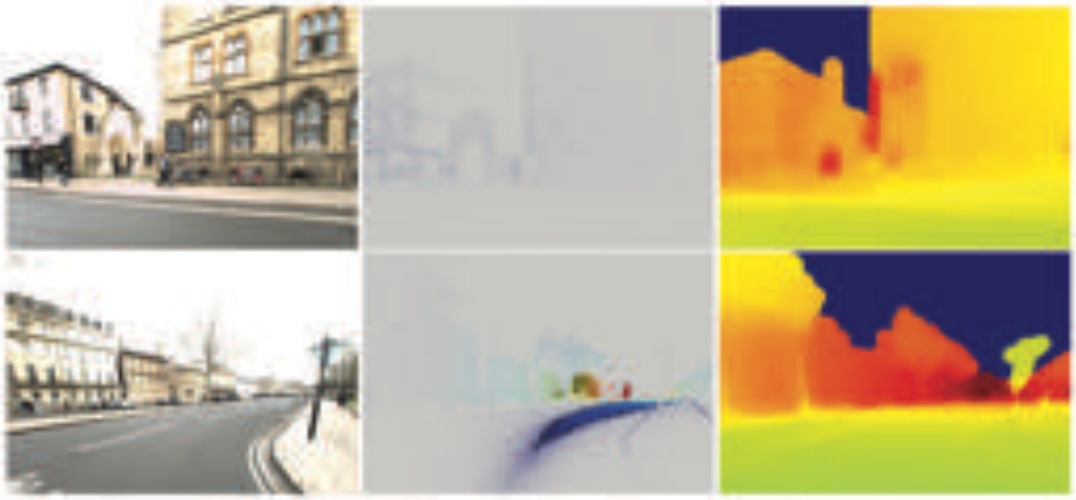


Figure 3.10: From point cloud to dense depth maps: from left to right: raw image, point cloud viewed from the camera frame, inpainted depth map by [28] (green is near, red is far).

an area of the city different from the one used for training our networks for validation, see figure 3.9.

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 [28] in order to make them dense. To avoid occlusion artifact (visible points that should not be visible because of occlusions) we use Hidden Point Removal (HPR) algorithm from Katz et al. [125] on the point cloud before further processing. We show in figure 3.10 example of depth maps generated through this pipeline. 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 (for insight about our model architecture, see annex ??).

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. The three query sets are:

- Oxford – Long-term (LT): queries have been acquired 7 months after the reference images under similar weather conditions.

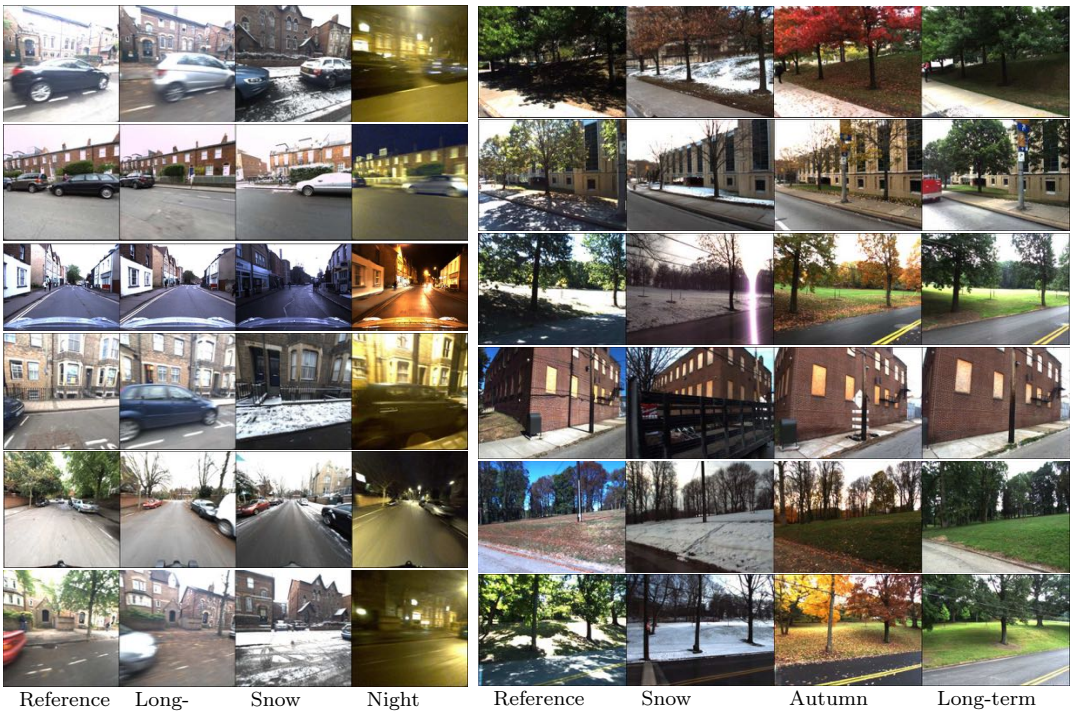


Figure 3.11: Examples of test images : we evaluate our proposal on 6 challenging localization sequences. Query image samples and the closest reference images in the database are presented from Oxford Robotcar [172] (left) and CMU season dataset [22] (right).

- 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:

- CMU – Long-term (LT): queries have been acquired 10 months after the reference images under similar weather conditions,
- 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 3.11.

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:

- **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 [332]. This metric qualifies the accuracy of the localization system.

3.3.2 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 3.2 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 hard negative mining

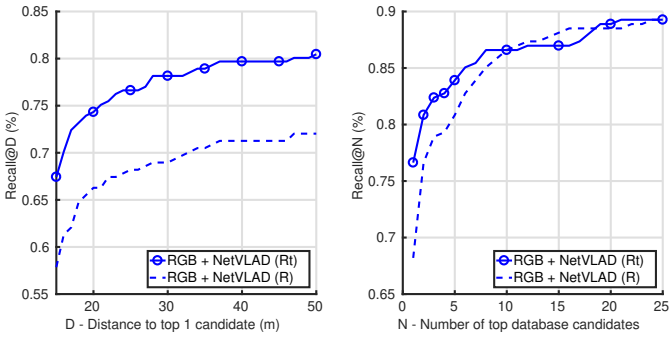


Figure 3.12: Full backbone resnet versus truncated version with NetVLAD : we show the importance of the spatial resolution of the feature maps used with NetVLAD layer. The truncated version of Resnet18 (Rt), two times lighter than the complete one (R), achieves much better localization results.

as explained in section 3.2.5 with: $M_p \in [1, 4]$, $M_n^{sub} = 25$ and $M_n^{hard} = 10$. Positive examples are chosen within a radius of 7 meters around the anchor image (according to GPS information) and we use the orientation given by position of the sensor on the Oxford mapping vehicle to ensure that cameras of positive examples are bearing in the same direction as the anchor camera (the vehicle is equipped with 4 cameras dispatched at the front, the back, the left side and the right side of the car). We set as negative examples data located further than 700 meters to the anchor. Images and depth maps are re-sized to 224×224 pixels before training and testing.

Encoder architectures. We use the fully convolutional part of Alexnet [140] and Resnet18 [107] (Resnet in short) architectures for features extraction. Weights are initialized with the pre-trained weights on ImageNet and firsts convolutional filter weights remain fixe during training. 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 [103] as it have been shown to perform equivalently to jet colorization [76].

Descriptor architectures. We test the two state-of-the-art image descriptors MAC [240] and NetVLAD [6] (see section 3.1). We experimented that NetVLAD descriptor combined with Resnet architecture 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 to use a truncated version of Resnet encoder, created by drooping the end of the network after the 13^{th} convolutional layer. Thus we obtain a feature block with greater spatial resolution: $256 \times 14 \times 14$ compared to $512 \times 7 \times 7$. Results on our validation set for both architectures are presented in figure 3.12. As the truncated version of Resnet encoder clearly dominates the full one, we use the truncated version for the following experiments.

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

Decoder architecture. The decoder used in our proposal is based on Unet architecture and inspired by network generator from [114]. Details about our decoder architecture can be found in appendix A.1. Decoder weights are initialized randomly.

3.3.3 Competitors

We compare the three 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 as explain in previous section, with triplet ranking loss of equation 3.2. We also train the RGB network on the aforementioned dataset, but only with the radiometric modality.

2. *Our proposal (**RGB(D)**):* introduced in the previous section (see figure 3.7) this architecture uses pairs of aligned image and depth map during training step and images only at test time.

3. *Hallucination network (**RGB(H)**):* our version of the hallucination network [110] (see figure 3.6), trained on aligned triplets of images and depth maps.

For fair comparison, as **RGB(D)** and **RGB(H)** image descriptors are obtain by concatenation two full-size descriptors (see section 3.2.6), we perform PCA to reduce the size of the final descriptor of all three methods to 2048.

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**

Table 3.1: Contribution of the depth side information during training.

3.4 Long-term localization

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.

3.4.1 Preliminary results

Contribution of the depth information

In this paragraph, we investigate the impact on localization performances provided by the side geometric information on our method. For a consistent 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 pixel loss introduced in equation (3.3), and make the weights of the decoder D_a trainable when optimizing triplets losses constraints. Results on the validation dataset with encoder architecture Alexnet and decoder MAC are presented in table 3.1.

Increasing the size of the system results in a better localization (RGB^+ + MAC > 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 create the final description.

Descriptor comparison

In figure 3.13, we present the localization scores of the three different methods on the validation set with Alexnet as backbone encoder. It clearly demonstrates the superiority of the NetVLAD pooling layer compared to the MAC descriptor in term of precision (recall@D). As we are more interested in precision than in recall, our concern is about localization, we only use NetVLAD as pooling layer for the rest of the experiments (in

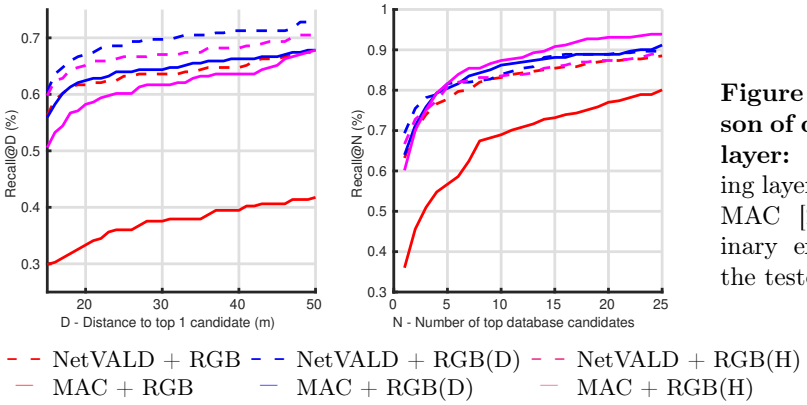


Figure 3.13: Comparison of descriptors pooling layer: NetVLAD [6] pooling layer perform better than MAC [240] in our preliminary experiment, whatever the tested method.

combination with Alexnet or Resnet encoder). Still, this preliminary experiment has shown that the proposed method can be used in combination with various descriptor pooling layers.

3.4.2 Localization results

Localization results on the six query sets are presented in figure 3.14. We also show, in figure 3.15 (3^{rd}, 5^{th} and 6^{th} columns), 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 confirm our intuition: geometric clues given during the training process can be efficiently used for CBIR for localization. In addition to that, 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.

We obtain convincing localization results for the CMU query sets (figure 3.14 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 features for localization.

Our method shows the best localization improvement on the Oxford - Snow query sets (figure 3.14-b) and CMU – Snow (for encoder Alexnet, see figure 3.14-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 3.15, CMU-Snow 1^{st} row). Similar results should be intended regarding Oxford –

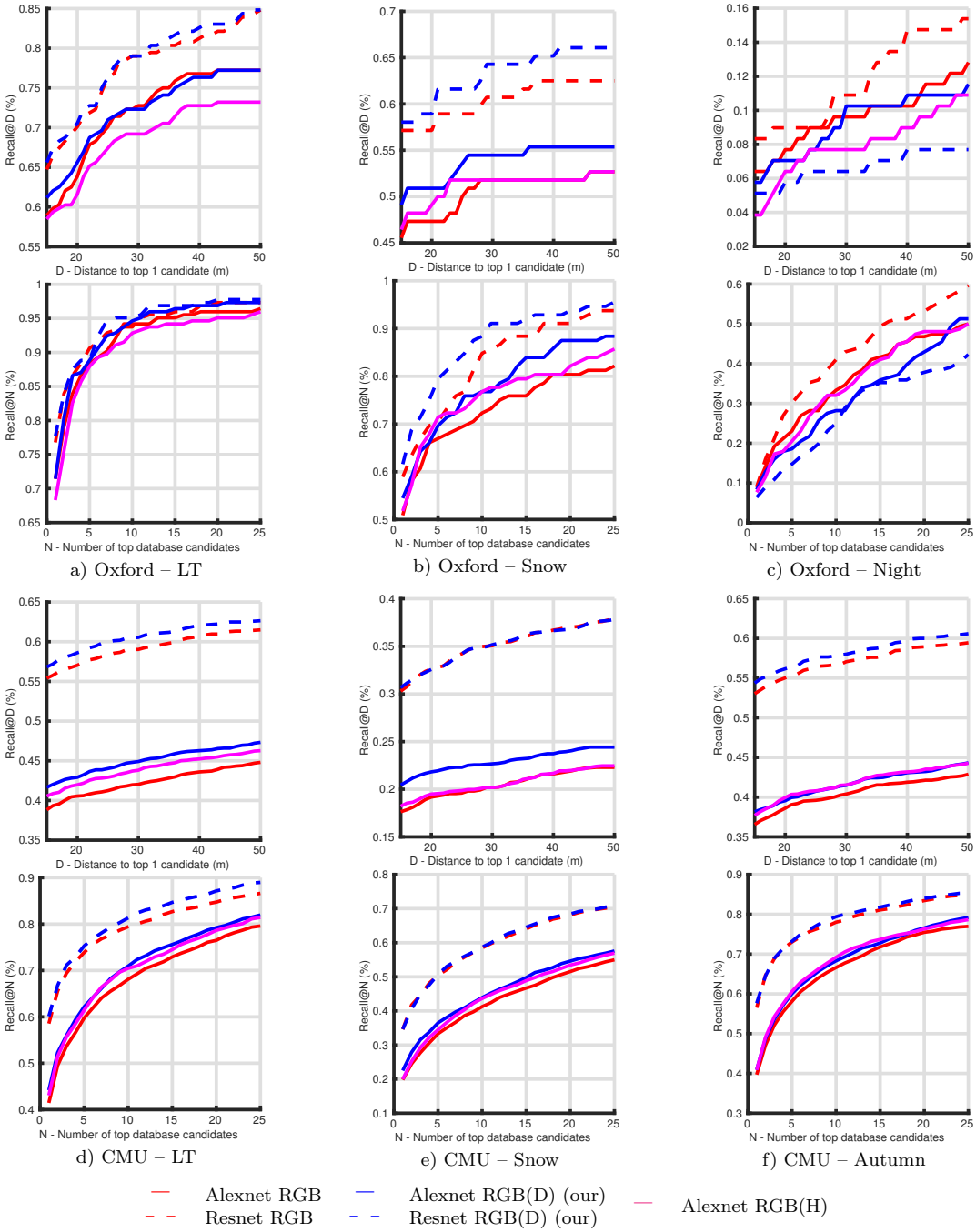


Figure 3.14: Comparison of our method RGB(D) versus hallucination network RGB(H) and networks trained with only images RGB: we report results for backbone network encoder Resnet (- -) and Alexnet (–). Our method (in blue) is superior in every scenario facing hallucination network (in magenta). It also 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 colors.

Figure 3.15: Comparison of top-1 retrieved images: 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). All the descriptor use Resnet18 as backbone, excepted RGB(H).

Night query set (figure 3.14-c), however, none of the tested methods are able to perform on this particular scenario. In the following section, we investigate why our method has failed on the night to day localization task.

3.5 Night to day localization scenarios

As mentioned previously, at first glance our method is not well designed to perform 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 the night to day image matching task.

3.5.1 Night to day localization

Night to day localization [1, 121] 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 and also by the poor quality of night images (motion blur), as illustrated in figure 3.11. 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 3.16 shows the poor quality of the estimated depth maps 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 (3.3). Figure 3.16 presents the qualitative improvement 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 3.17-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 3.18. Even with blurry images,

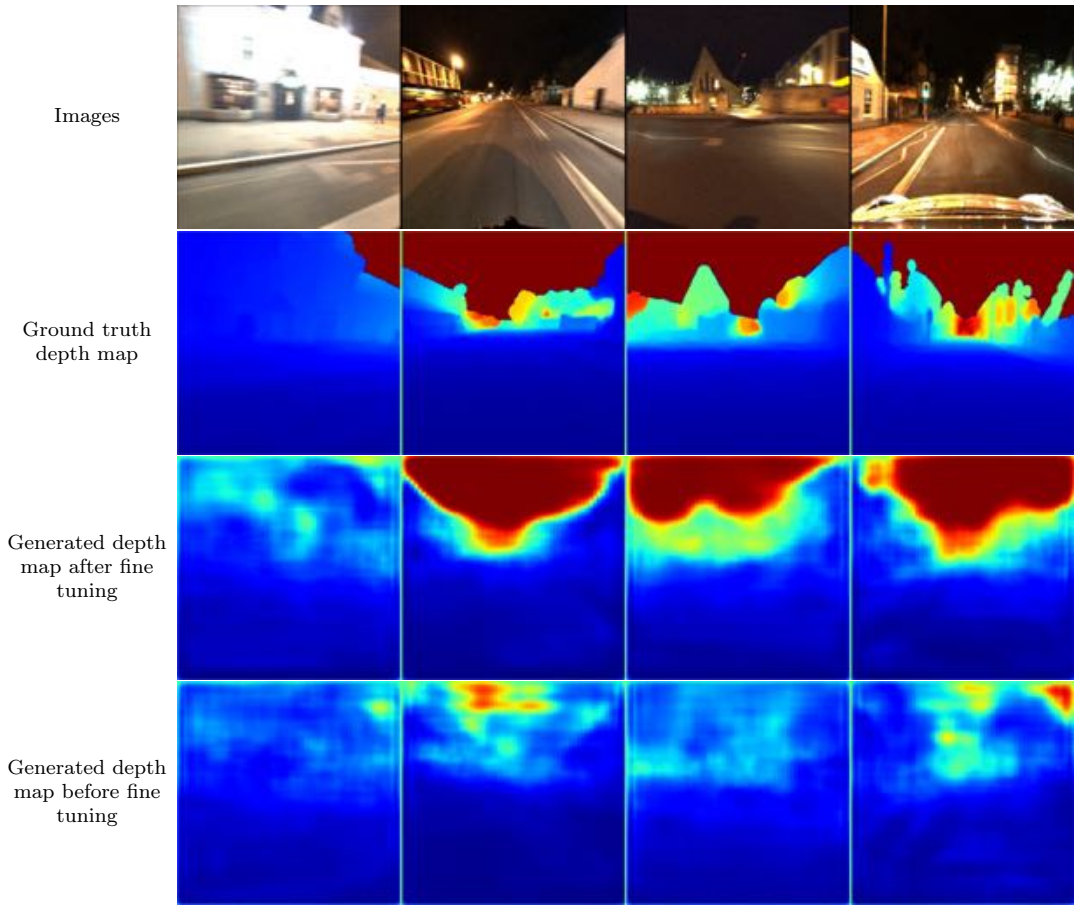


Figure 3.16: Effect of fine tuning with night images on decoder output:. Decoder trained with daylight images is unable to reconstruct the scene geometry (bottom line). Fine tuning the network with less than 1000 pairs {image, depth map} acquired by night highly improves appearance of the generated depth maps. Maps best viewed in color.

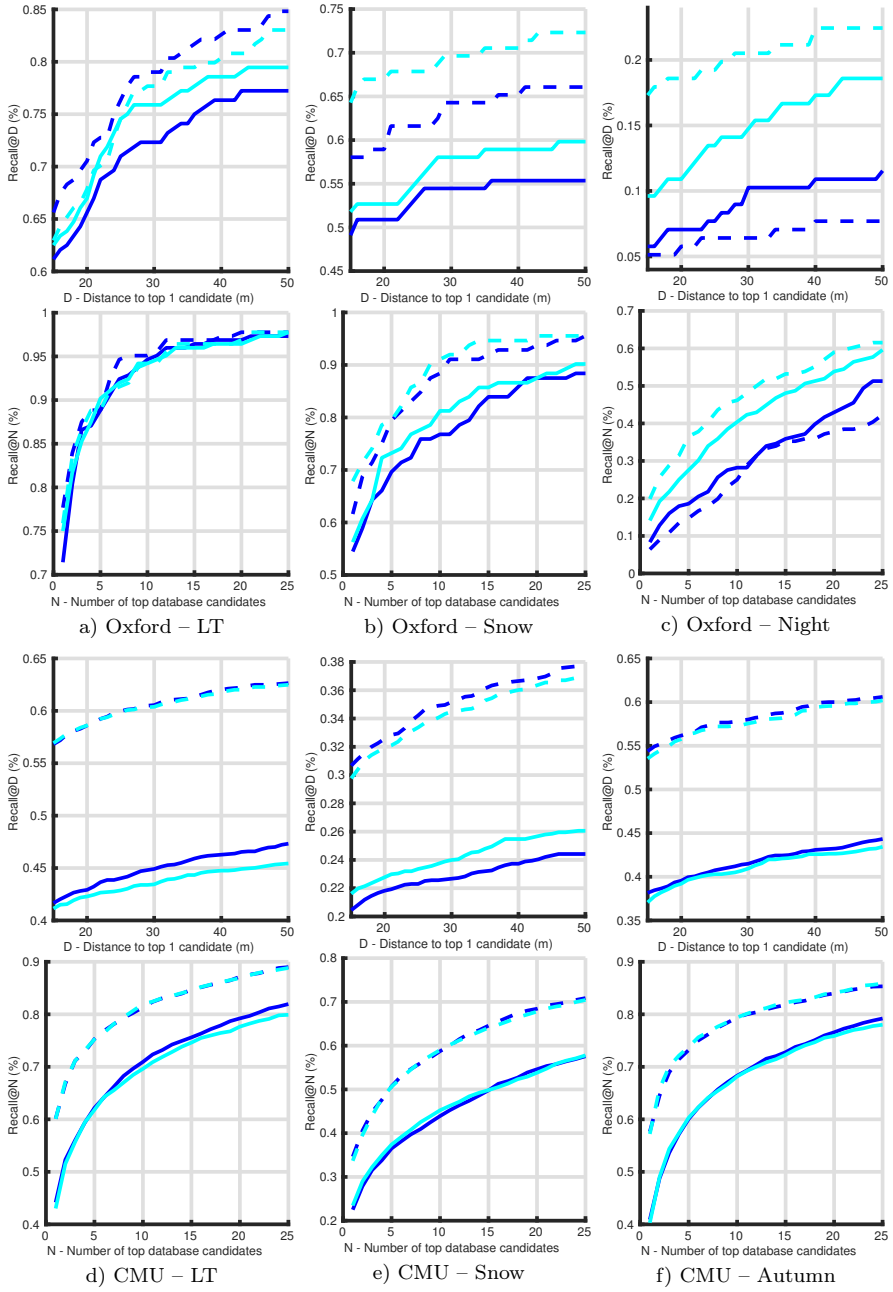


Figure 3.17: Results after fine tuning: we are able to drastically improve localization performance for the Oxford – Night challenging scenario (c) by only fine tuning the decoder part of our network with weakly annotated data. Curves best viewed in color.

our method is able to extract useful geometric information to improve the matching (see figure 3.18, 3^{rd} row).

3.5.2 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 3.17. Localization accuracy remains stable after the fine tuning. We even observe slight increase in the localization performances for some scenarios (figure 3.17-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.

3.6 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.

3.6.1 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 [172]. Reflectance values range from 0 to 1 indicating if the object has reflected from 0 to 100% of the original laser beam. We process the sparse reflectance data in the same manner as the depth map using inpainting algorithm from [28] to produce dense reflectance maps (§ 3.3.1), and use exactly the same decoder architecture for the reflectance map and the depth map. Examples of ground truth and reconstructed dense reflectance map are presented in figure 3.19.

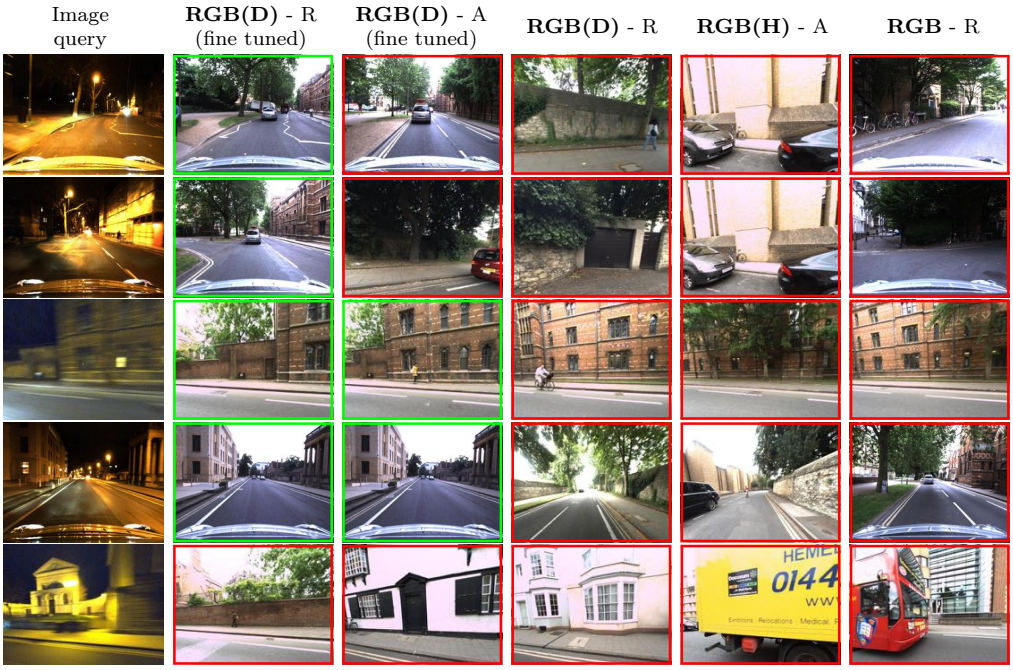


Figure 3.18: Comparison of top-1 retrieved images on night dataset: we show top-1 retrieved candidate after the challenging night to day localizations scenario. Red box indicates a wrong match and green box a proper one (*i.e.* retrieved image lies in 25m radius from the query ground truth position). -A denotes Alexnet and -R truncated Resnet18 backbone used with NetVLAD.

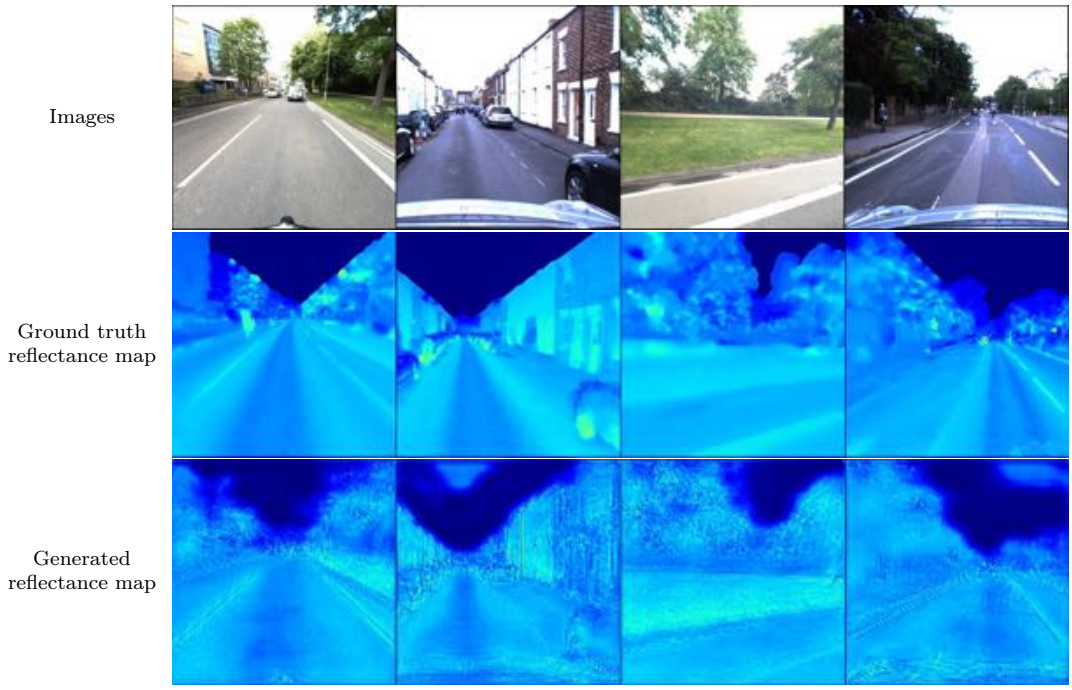


Figure 3.19: Examples of dense reflectance map: the lighter the color, the higher the reflection of the material. Reflectance map highlights reflective areas, like road marking, road sign, vegetation and cars. Figure best viewed in colors.

3.6.2 Reflectance versus Depth

We report in figure 3.20 results using reflectance map during the descriptor training (**RGB(R)**, in gray). We also illustrate in figure 3.15 localization performances of the different methods by comparing the top-1 retrieved candidate after similarity evaluation. 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 3.15, 2^{nd} row) or inversely very high light reflecting property (*e.g.* traffic signs, see figure 3.15, last row). In a different way, depth map training supervision provides interesting building shapes understanding (see the recognized tower building on figure 3.15, CMU - LT $2nd$ row).

The reflectance-augmented descriptor shows poor results on the snowy scenarios (figure 3.20 b-d). It is not surprising as the snow presents on the scene highly reflect the light, confusing our system based on material reflectance.

3.6.3 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, by adding an extra modal branch with refectance decoder D'_a and auxiliary encoder and descriptor $\{E'_a, P'_a\}$. The modified network is presented in figure 3.21. We report localization results of the three methods, depth map as side information (**RGB(D)**, in blue), reflectance map as side information (**RGB(R)**, in gray) and depth and reflectance map as side information (**RGB(DR)**, in green), in figure 3.20.

We do not observe systematic improvement when using both modalities. Nevertheless we obtain best localization results for 3 out of 5 query sets (figure 3.20 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

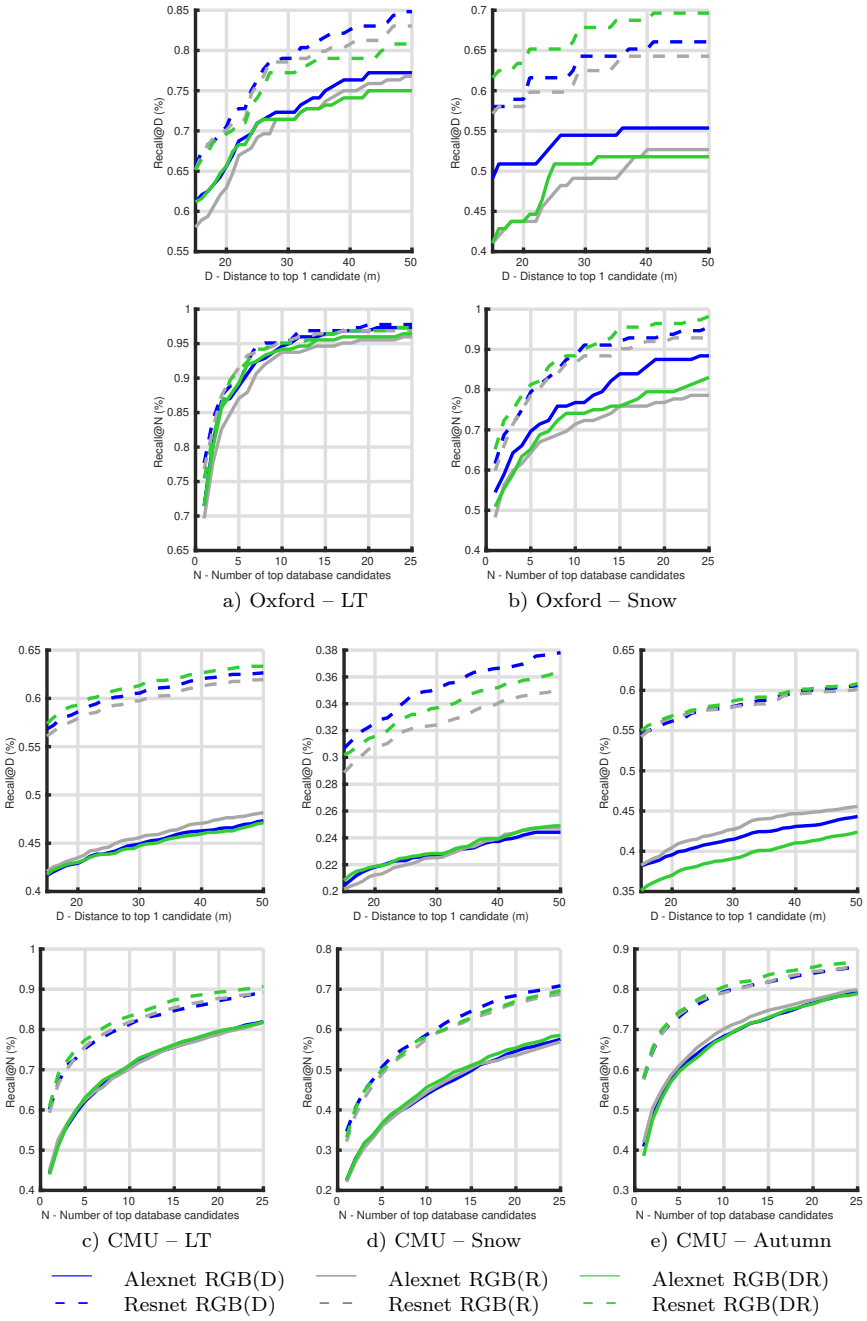


Figure 3.20: Comparison of depth map and reflectance map as side information. The geometric information (in blue) remains more informative than the reflectance map (in gray) for the task of image description for localization. However, when combined (in green), depth map and reflectance map can benefit from each other and produce the most discriminative image descriptors for scenarios b, c & e. Curves best viewed in colors.

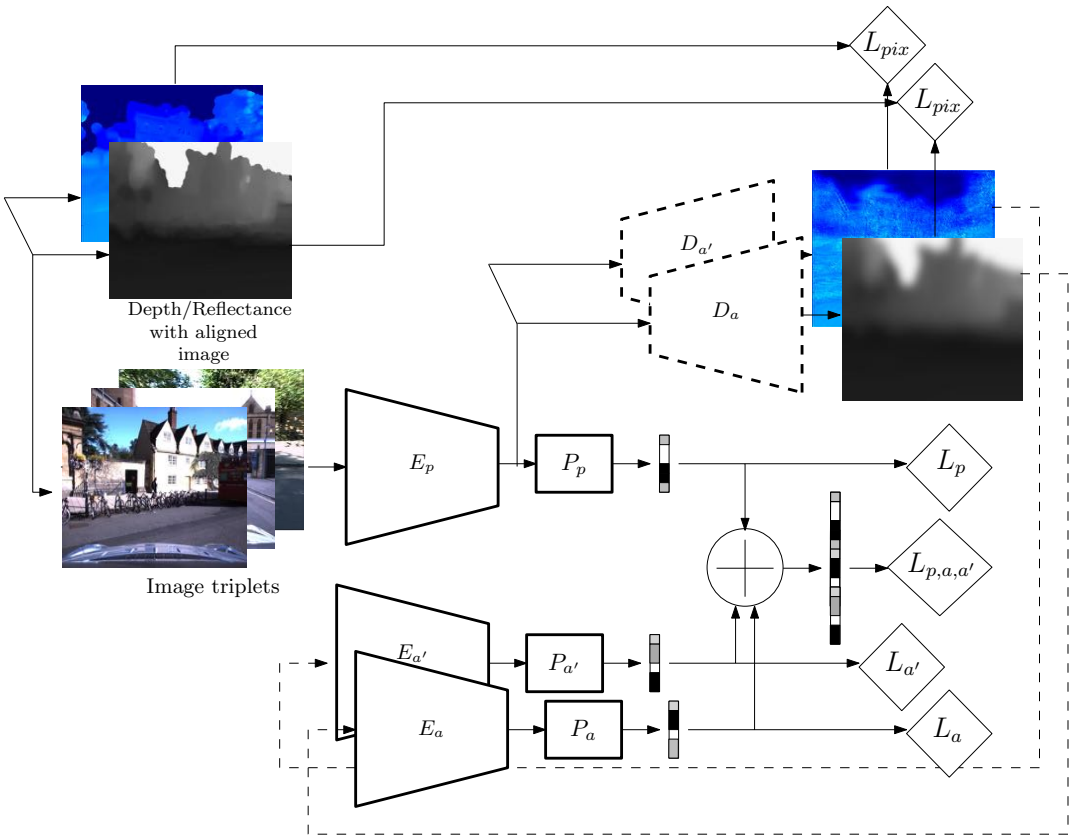


Figure 3.21: Multi-modal training: we modify the training policy presented in figure 3.7 to handle multi-modality. Each generative modal branch (D_a and $D_{a'}$) can be trained separately. Modality descriptors are trained jointly through the final triplet ranking loss $L_{p,a,a'}$.

than the other on a specific dataset (for instance depth over reflectance for the query set CMU - Snow, figure 3.20-d), the combination of the both will cancel potential benefit given by the most informative modality. As discussed before, the reflectance information can be source of disturbance if snow is present on the image, resulting on a worse scene description. On figure 3.15, 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.

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 more on the final descriptor fusion; modality-aware aggregation descriptor or more complex attention mechanism may be considered [274].

3.7 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 [6, 234], including methods based on side information learning [110]. 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. Finally we show that our method can generalize to over auxiliary modality supervision during training. We use lidar reflectance to illustrate this generalization capability.

In the next chapter, we will show how the learned depth can be used again in a pose refinement step.

Chapter 4

Refining visual localisation with learned geometric clues

Many applications in Computer Vision and Robotics require a precise initial pose estimation of the visual data acquisition system: augmented reality [12], visual odometry [209], SLAM [186] or visual servoing [175], to name a few. Coarse estimation provided from standard geo-localization system (*e.g.* GPS) or Content-based Image Retrieval (CBIR) for localization are not accurate enough for such applications, and other processing are required to initialize the system with a suitable pose. For instance, considering a system using pose initialization by CBIR, the reference database is never dense enough to ensure that there is a database example located at the same pose as the query. Thus, the exact 6-Degrees of Freedom (DoF) of the query cannot be recovered.

We introduced in the previous chapter a global image descriptor for localization under challenging condition. We are now considering the problem of pose refinement, in other words, the second step of our hierarchical Visual-based Localization (VBL) system. In section 2.2.3, we mentioned two main approaches for pose refinement: image-based and model based. As our first localization step rely on image indexing, an image-based refinement method is more logical to implement. Like in the global image descriptor introduced previously (see section 3.2), our objective is to take advantage of an auxiliary modality in our refinement step. We decide to use the learned depth maps to incorporate geometric constraints in our pose refinement process.

For computing the real 6-DoF pose of a query, we first compute dense correspondences between the query and the top retrieved images from our initial CBIR step. From these

correspondences, we can estimate relative pose information with geometric algorithms. In order to obtain a position at true scale (which is not the case with traditional multi-view methods [105]), we exploit the reconstructed depth map. 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 and can be used on various environment without specific retraining. Unlike model-based hierarchical VBL methods, our proposal does 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.

For a comprehensive review of hierarchical methods for VBL, please refer to section 2.2.3, chapter 2. The rest of this chapter is presented as follows: section 4.1 is dedicated to the workflow explanation of our method, then we introduce the two geometric algorithms used to compute the relative pose of the query 4.2. We present explanatory results for indoor localization on section 4.3 and we pursue more experiments on indoor localization on 4.4. Before the conclusion, we investigate unsupervised depth from monocular training as well as outdoor localization in section 4.5. We finally conclude the chapter, after a short discussion, in section 4.7.

4.1 Method

The camera pose is estimated following this four-step algorithm:

1. we obtain the initial pose of the query image by CBIR (section 4.1.1),
2. then we find dense correspondences between the query image and the best retrieved images (section 4.1.2),
3. meanwhile, we use a neural network to create the depth map related to the images (section 4.1.3),
4. finally, we use the dense correspondences as well as the reconstructed depth maps to compute relative poses from the retrieved candidates and the query image, using geometric reasoning (section 4.2).

The two first steps of our pose refinement method are illustrated in figure 4.1.

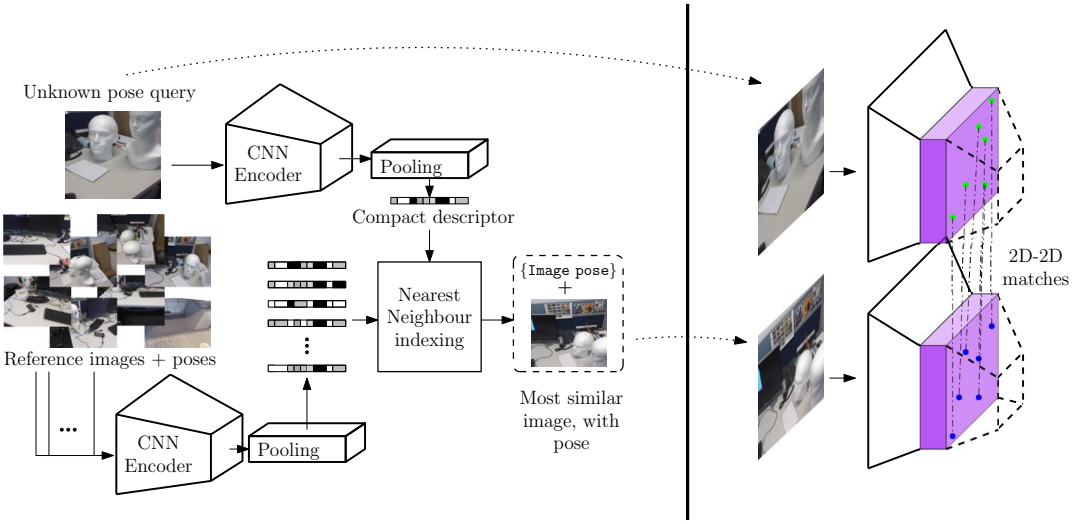


Figure 4.1: The two first steps of our relocalization pipeline:. **a)** We retrieve initial pose of an image query using CBIR. **b)** We make 2D dense correspondences between the retrieve image and the query. Purple boxes are deep features blocks used for dense images matching.

Notations. The aim of our method is to recover the exact 6 DoF camera pose $\mathbf{h}^q \in \mathbb{R}^{4\times4}$, represented by a pose matrix in homogeneous coordinates, corresponding to an input RGB image $x^q \in \mathbb{R}^{3\times H\times W}$. We know the matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ of intrinsic parameters of the camera. We assume that we know the pose $\{\mathbf{h}_i^r\}_{i\in[1,N]}$ of a pool of N reference images $\{x_i^r\}_{i\in[1,N]}$ of the scene where we want to localize the query. These poses can be obtained by Structure From Motion (SfM) algorithms or by using external sensors. We denote as E, respectively D, a neural network encoder, respectively decoder. P denotes the pooling layer used to create a global image descriptor for image indexing.

4.1.1 Image retrieval

CBIR for localization are deeply detailed in the previous chapter. We provide a short explanation in the following. We assume that the reference data are augmented with 6-DoF pose information and we cast the initial pose estimation task as a CBIR problem like in [21]. In order to evaluate the similarity between the unknown pose query image x^q and the N reference images $\{x_i^r\}_{i\in[1,N]}$, we need to use a discriminative image representation. We use deep global image descriptor for place recognition, to describe the data by low-dimensional L_2 normalized vectors. The image descriptor $f(x^q)$ is obtained

by concatenating the dense feature from neural network encoder E:

$$f(x^q) = P(E(x^q)). \tag{4.1}$$

In the following, we write $f(x^q)$ as f^q for readability.

We first compute reference descriptors $\{f_i^r\}_{i\in[1,N]}$ from the reference images. Then we compare the query descriptor f^q to the pre-computed descriptors by nearest neighbour indexing and retrieval:

$$\left\{f_j^{sim}\right\}_{j\in[1,K]} = sim\left(f^q, \{f_i^r\}_{i\in[1,N]}\right), \tag{4.2}$$

where sim is the nearest neighbour matching function and $f_j^{sim}, j \in [1, K]$, the K ranked 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 the retrieved poses $\mathbf{h}_{j\in[1,K]}^{sim}$ as initial candidate poses of the image x^q.

4.1.2 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 [201, 290, 324], 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. Since we only consider the K nearest neighbours to our query, dense features matching is tractable.

Local image descriptors $d_{l,m}$ 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 [290, 324]. We subsequently L_2-normalize the extracted descriptors before matching. We consider only consistence matches by rejecting correspondences that do not respect the bidirectional test (nearest descriptors from image 1 to image 2 have to be the same as nearest descriptors from image 2 to image 1).

4.1.3 Depth from monocular image

2D to 2D correspondences obtained by dense features matching (section 4.1.2) do not provide enough information to compute relative pose between images at absolute scale [105]. 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 [75], weakly annotated data [98] or even in an unsupervized fashion [173].

We train a Convolutional Neural Networks (CNN) encoder/decoder to predict the corresponding depth map $\hat{z}$ associated to an image:

$$\hat{z} = D(E(x)). \tag{4.3}$$

3D projection. With the generated depth map and the intrinsic parameters of the camera $\mathbf{K}$, we can project the 2D point in the image frame at coordinate $\{l, m\}$ to the corresponding 3D coordinate in the scene $\mathbf{p}_{l,m} \in \mathbb{R}^3$, relative to the camera frame:

$$\mathbf{p}_{l,m} = \hat{z}\,[l, m] \cdot \mathbf{K}^{-1}[l, m, 1]^T, \tag{4.4}$$

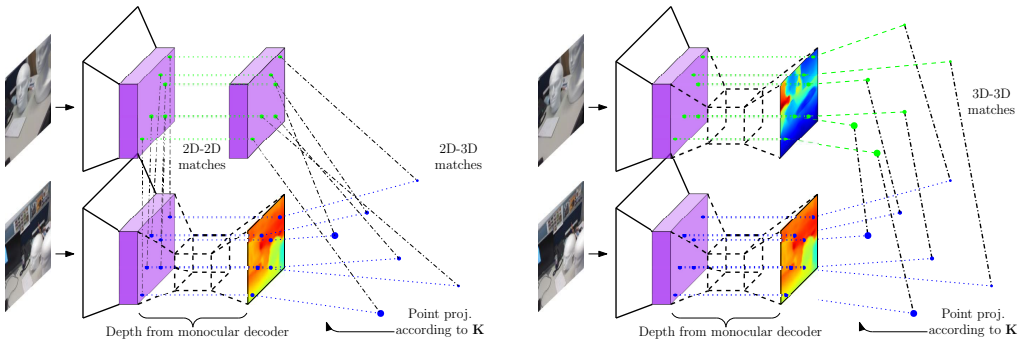
where $\hat{z}\,[l, m]$ is the metric depth value at position $\{l, m\}$ in the reconstructed depth map $\hat{z}$.

4.2 Relative pose estimation

We propose two alternatives to compute the relative pose between the query image x^q and the most similar retrieved images $x_j^{sim}, j \in [1, K]$:

1. an ICP-based method called Iterative Closest Learned Point (ICLP),
2. a PnP-based algorithm called Perspective-n-Learned-Point (PnLP).

The main differences between these two approaches are that ICLP is iterative and rely on the reconstructed depth map of the two densely matched images, while PnLP uses only one depth map. In figure 4.2, our two relative pose estimation are presented, side by side.



(a) PnlP (b) IClP

Figure 4.2: Our two method for relative pose computation: 4.2(a) we compute the relative 6-DoF from 2D-3D correspondances using a PnP-like method. 4.2(b) we manage to align two point clouds projected from learned depth maps to recover the realtive pose of the two images, using an ICP-like algorithm.

7 Scenes [276]	ICLP w/o deep features	ICLP w/ deep features
Chess	0.28/8.6	**0.23**/**5.4**
Fire	0.39/16.5	**0.30**/**14.1**
Heads	**0.18**/14.9	0.19/**14.1**
Office	0.41/13.4	**0.36**/**11.3**
Pumpkin	0.40/12.0	**0.35**/**7.4**
Kitchen	0.24/ 7.5	**0.19**/**4.9**
Stairs	0.57/12.2	**0.48**/**10.3**

Table 4.1: Point cloud registration: we report median error on position and orientation (m/°) for each scene. Its shows the importance of deep feature for point cloud matching in our method. Global image descriptors for image indexing are MAC descriptor, see section 4.3 for more details about the used datasets and our implementation.

4.2.1 Iterative Closest Learned Point

The general idea is to obtain the relative camera pose $\mathbf{h}^{\mathrm{r}\to\mathrm{q}}$ by registering the point cloud $\mathcal{P}^q$ obtained by projecting in 3D the depth map (equation 4.4) $\hat{z}^q$ of query x^q to the points cloud $\mathcal{P}_j^{sim}$ from the reference depth maps $\hat{z}_j^{sim}$ of images $x_j^{sim}, j \in [1, K]$. One reference point cloud is evaluate at a time and we chose the final registration that minimizes the point-to-point distance between the two point clouds.

Point matching. Refinement with ICP involves matching corresponding points between two point clouds in order to estimate a rigid transformation that minimizes the distances between the paired points. Standard approaches only consider the Euclidean distance between a single point and its nearest neighbours in the reference point cloud to establish matching, making the initial alignment between the two point clouds a crucial step to obtain correct results. We can rely on point descriptors to establish strongest

matches [226]. We use local descriptor $d_{l,m}$ introduced in section 4.1.2 and we associate to each projected point $\mathbf{p}_{l,m}$ the descriptor corresponding to the deep feature computed by the encoder E at the same spatial position $\{l, m\}$. The matching process remains the same as the one detailed in § 4.1.2, with the additional 3D point information added to the local features. We present in table 4.1 results of an exploratory experiment to estimate the benefit of adding the deep features for the point cloud alignment. We observe a clear improvement at almost no cost (the deep features are extracted from the already computed features block of the depth from monocular CNN).

Data: query point cloud to align $\mathcal{P}^q$ with augmented descriptors $\mathcal{D}^q$ and reference point cloud $\mathcal{P}^{sim}$ with associated descriptors $\mathcal{D}^{sim}$
Result: relative pose $\mathbf{h}^{\mathrm{r}\to\mathrm{q}}$, mean distance between matched points $\|\mathcal{M}\|_2$
$\mathbf{h}^{\mathrm{r}\to\mathrm{q}} \leftarrow \mathbf{I}_{4\times4}$;
$\mathbf{h}^{it} \leftarrow \mathbf{1}_{4\times4}$;
while $\left\|\mathbf{h}^{it} - \mathbf{I}_{4\times4}\right\|_{\mathrm{F}} \geq \epsilon$ **do**
  $\mathcal{P}^q \leftarrow \mathbf{h}^{\mathrm{r}\to\mathrm{q}}\mathcal{P}^q$;
  $\mathcal{M} \leftarrow$ `match_points` $\left(\left[\mathcal{P}^q, \mathcal{D}^q\right], \left[\mathcal{P}^{sim}, \mathcal{D}^{sim}\right]\right)$;
  $\mathbf{h}^{it} \leftarrow$ `relative_pose`$(\mathcal{M})$;
  $\mathbf{h}^{\mathrm{r}\to\mathrm{q}} \leftarrow \mathbf{h}^{it}\mathbf{h}^{\mathrm{r}\to\mathrm{q}}$;
end

Algorithm 1: Our ICLP algorithm, see section 4.1.2 for details about functions `match_points` and section 4.2.1 for function `relative_pose`. Expression $[\mathcal{P}^q, \mathcal{D}^q]$ denote the concatenation of the point coordinates with their corresponding deep feature d, as explained in section 4.2.1

Algorithm. Relative pose $\mathbf{h}^{\mathrm{r}\to\mathrm{q}}$ is obtain thank to the iterative algorithm detailed in 1. The `relative_pose` function computes the relative transformation between the matched points that minimizes the Euclidean difference between the two point clouds. We use classical relative pose estimation algorithm [226]:

- Rotation: we apply Singular Value Decomposition (SVD) on the matching matrix obtained by multiplication of the zeros-centered corresponding 3D points in each point cloud. Rotation matrix can be computed by multiplying the right-singular vectors matrix with the transposed of the left-singular vectors matrix.
- Translation: we obtain the translation component by aligning in the same frame the two point cloud centroids, using the rotation matrix, and evaluating the difference

between them.

We embed the pose computation within a Random Sample Consensus (RANSAC) algorithm, as the point cloud may contain erroneous data because it has been generated from image-only information by our encoder/decoder CNN. We run the algorithm for a fixed number of iterations, for each of the top K retrieved images by the first indexing step. We chose the final relative pose $\mathbf{h}_{best}^{\mathrm{r}\to\mathrm{q}}$ according to the minimal mean alignment error returned by our algorithm ($\|\mathcal{M}\|_2$).

4.2.2 Perspective-n-learned-Points

Thanks to the generated depth map (section 4.1.3) and the equation 4.4, we can project 2D points from retrieved images into 3D coordinates. 2D-2D correspondences obtained in section 4.1.2 can be interpreted as 2D-3D correspondences and we can use a PnP algorithm to compute the relative transformation $\mathbf{h}_{\mathrm{r}\to\mathrm{q}}$ between the query image and the reference image.

We embed our PnLP algorithm within a RANSAC consensus where a sub-part of 2D-3D correspondences is evaluated at a time. We use 3-points algorithm from [134], using the authors efficient implementation [133]. As we have K reference candidates from image retrieval step (section 4.1.1), we select the best pose $\mathbf{h}_{best}^{\mathrm{r}\to\mathrm{q}}$ as the one 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.

4.2.3 Final pose computation

We obtain final pose of query image x^{q} using the relation:

$$\mathbf{h}^{\mathrm{q}} = \mathbf{h}_{best}^{\mathrm{r}}\mathbf{h}_{best}^{\mathrm{r}\to\mathrm{q}}. \tag{4.5}$$

4.2.4 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. There are dedicated training pipeline for each of the computer vision tasks involved in our image pose estimation framework: methods for learning a global image descriptor [6, 101, 234], CNN designed to extract and describe local features [203, 244, 330] and systems that produce a depth map from a monocular image [75, 98, 173]. 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 [290, 334].

Generalization. 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 localization on one specific scene like end-to-end pose estimation networks [33, 127]. In other words, the same trained network can be used to localize images in multiple indoor and outdoor scenes, and even in totally unknown environments.

4.3 Preliminary results

In this section, we discuss about implementation details as well as training and testing datasets. Afterward, we present preliminary experiments to evaluate our two relative pose estimation methods.

4.3.1 Implementation

Dataset. We train and test our method on the 7 scenes indoor localization dataset [276]. This datasets are composed of various indoor environments scanned with RGB-D sensors. 6-DoF image poses and camera calibration parameters are provided. For all the experiments, reference images used for the initial pose estimation with VBL are taken from the training split and query images are taken from the testing split of the respective datasets.

	Initial localization (CBIR)		Pose refinement	
Scene	MAC (M)	NetVLAD (V)	V + ICLP	V + PnLP
Chess	0.31/14.9	0.29/13.0	0.12/4.5	**0.07/2.9**
Fire	0.49/16.7	0.40/15.5	0.25/8.9	**0.09/3.6**
Heads	0.28/20.5	0.20/16.0	0.18/9.9	**0.06/4.3**
Office	0.46/16.4	0.38/13.0	0.22/7.3	**0.06/4.3**
Pumpkin	0.50/15.0	0.43/13.1	0.21/6.2	**0.05/2.2**
Kitchen	0.30/11.2	0.23/9.5	0.15/4.5	**0.10/3.2**
Stairs	0.64/16.0	0.46/14.9	0.48/12.2	**0.44/10.3**

Table 4.2: Methods comparison: we report mean pose error (translation and orientation, m/°) of our two relative pose algorithms. Best results are shown in **bold**.

Network architecture and training. We use a U-Net like convolutional encoder/decoder architecture [114] with multi-scale outputs [98], see appendix A.2 for details. As the accuracy of our methods is highly correlated to the quality of the generated depth map, we use a more sophisticated network architecture than previously and we train it from scratch, without using pretrained weights. During training and testing, images are resized to 224×224 pixels. The generated depth map is 4 times smaller than the image input. We use L_1 pixel loss function for the fully supervised depth from monocular training. We train our architecture with Adam optimizer, learning rate of 10^{-4} divided by two every 50 epochs. Training our model using all the training sequences of the 7 scenes dataset takes approximately one day on our Nvidia Titan X GPU with a batch size sets to 24.

Method parameters. We compare MAC [240] and NetVLAD layer [6] with 64 clusters as global image descriptor for initial pose estimation. Deep local features are gathered from the second convolutional layer before the Relu activation of our architecture (appendix **??**), resulting in 56×56 feature vectors of dimension 64. We use this particular features block as it has the same spatial dimension as the generated depth map (*i.e.* 4 times smaller than the input). Concerning the ICLP algorithm: we set the maximum number of iteration to 100 and the threshold θ equal to 0.4. For the PnLP method, we set the inliers ratio threshold mentioned in section 4.2.2 to 10%.

4.3.2 Methods comparison

We report in table 4.2 initial result on the 7 scenes indoor dataset, using only the top-1 retrieved candidate (for computational time saving).

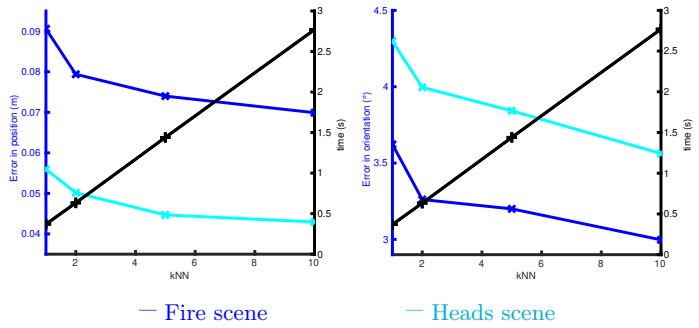


Figure 4.3: Trade of between precision/computation time: we compute for two scenes [276] the median absolute error in position and orientation for multiple number of k nearest neighbours used for pose refinement (see section 4.2.2).

Initial localization. We observe a comparable behavior than the one observed previously (section 3.4): NetVLAD pooling layer is more efficient than MAC for initial localization. In the following experiments, we only use the NetVLAD layer for global image descriptor.

ICLP vs PnLP. Our PnLP pose refinement clearly outperform the ICLP approaches. This can be explained by the fact that wrong estimations in the depth map values are penalizing twice the ICLP approach and only once the PnLP refinement. Thus, in the following, we use only the PnLP pose estimation for refinement.

Number of retrieved candidates. We use two scenes of the 7 scenes dataset to evaluate the impact of the number K of retrieved candidates on the final localization. Figure 4.3 shows improvement for 4 values of K, from 1 to 10. We found that $K = 5$ is a good trade-off between pose precision and time consumption. Localizing a query takes approximately one second using our non-optimized python code, including the image indexing step.

4.4 Indoor localization

In this section we present more detailed results on indoor localization, by comparing our method with state-of-the-art competitors and by evaluating our proposal on different environments than the one used for training.

Scene	*Image retrieval*	PnLP refinement	Relocnet [21]	Posenet [127]
Chess	*0.29/13.0*	**0.07/2.7**	0.12/4.1	0.13/4.5
Fire	*0.40/15.5*	**0.07/3.2**	0.26/10.4	0.27/11.3
Heads	*0.28/20.5*	**0.05/3.9**	0.14/10.5	0.17/13.0
Office	*0.38/13.0*	**0.09/2.9**	0.18/5.3	0.19/5.6
Pumpkin	*0.43/13.1*	**0.13/3.6**	0.26/4.2	0.26/4.8
Kitchen	*0.23/9.5*	**0.05/2.0**	0.23/5.1	0.23/5.4
Stairs	*0.46/14.9*	0.40/9.2	**0.28/7.5**	0.35/12.4

Table 4.3: Results on the 7 scenes [276] dataset: we report median position/orientation error in meters/degree. We compare the first pose estimation (im. retrieval, *in italics*) and, the final image localization (PnLP) of our method and two state-of-the-art approaches. Best localisation results are shown in **bold**.

4.4.1 Competitors

Indoor localisation error on 7 scenes [276] dataset are presented in table 4.3. We compare our proposal with Relocnet [229] and Posenet [127] trained with a geometric-aware loss function. Compared to Posenet [127] our model uses the same trained network for all the 7 scenes, compared to one network by scene for Posenet. 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 not uses specific training for the task of global image description.

4.4.2 Results

At first glance, we find that the initial pose estimation with image retrieval produces decent results (first column), while the network used to produce the global image descriptor has not been trained to this particular task. After applying our PnLP pose refinement, our method produces the most precise localization among the presented methods.

Figures 4.4-4.5 present estimated position at different steps of our method for 4 scenes. Our method is able to recover accurate positions even if there are not previous acquisitions close to the ground truth camera pose (see figure 4.4, top of the fire scene or figure 4.5, right of the heads scene).

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 localization.

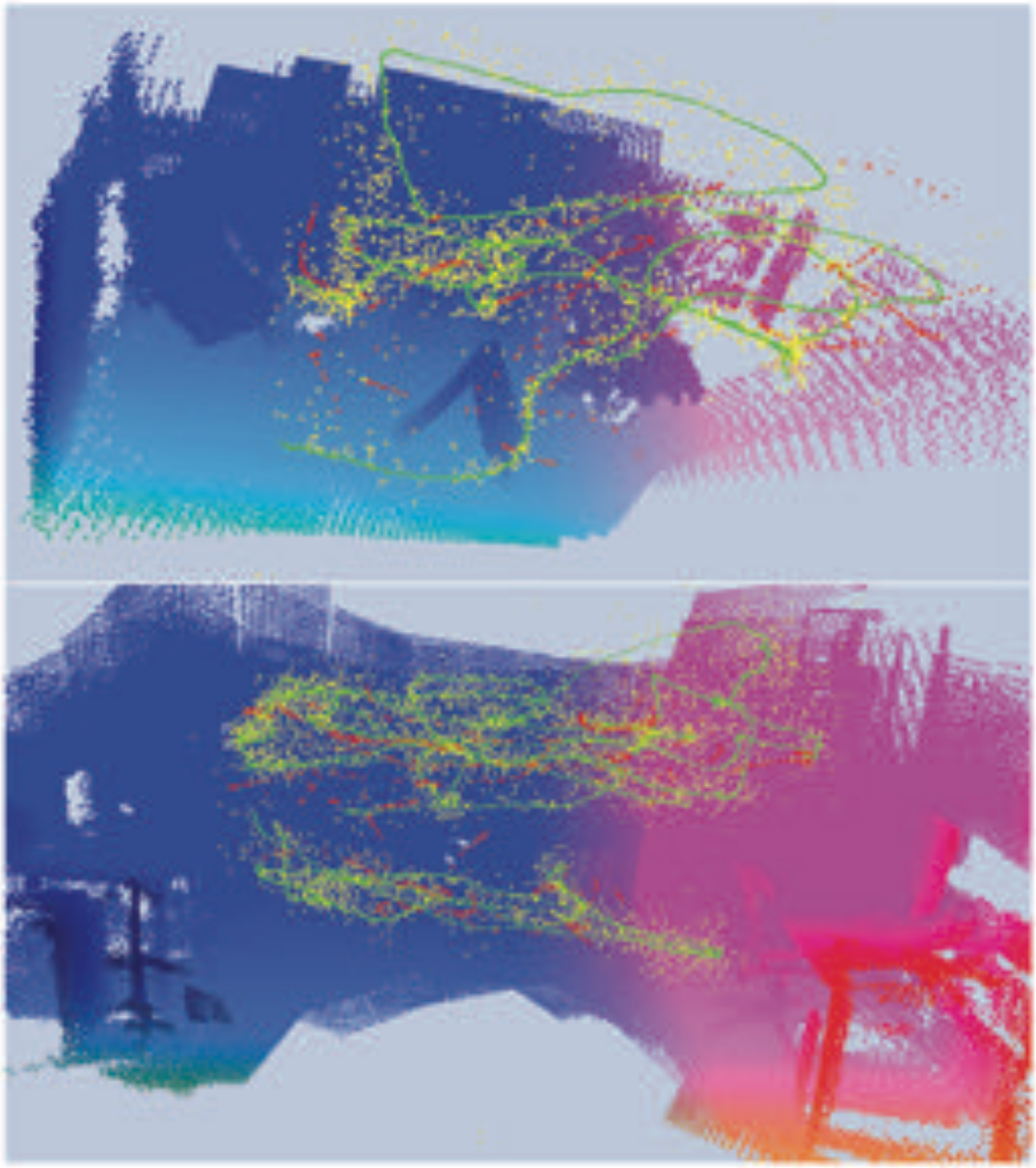


Figure 4.4: Visualization of the refined camera position: ground truth positions of test sequences are marked in green, initial position from image indexing in red and refined position after PnP in yellow. Environment from the 7 scenes dataset [276]: top fire and bottom office.

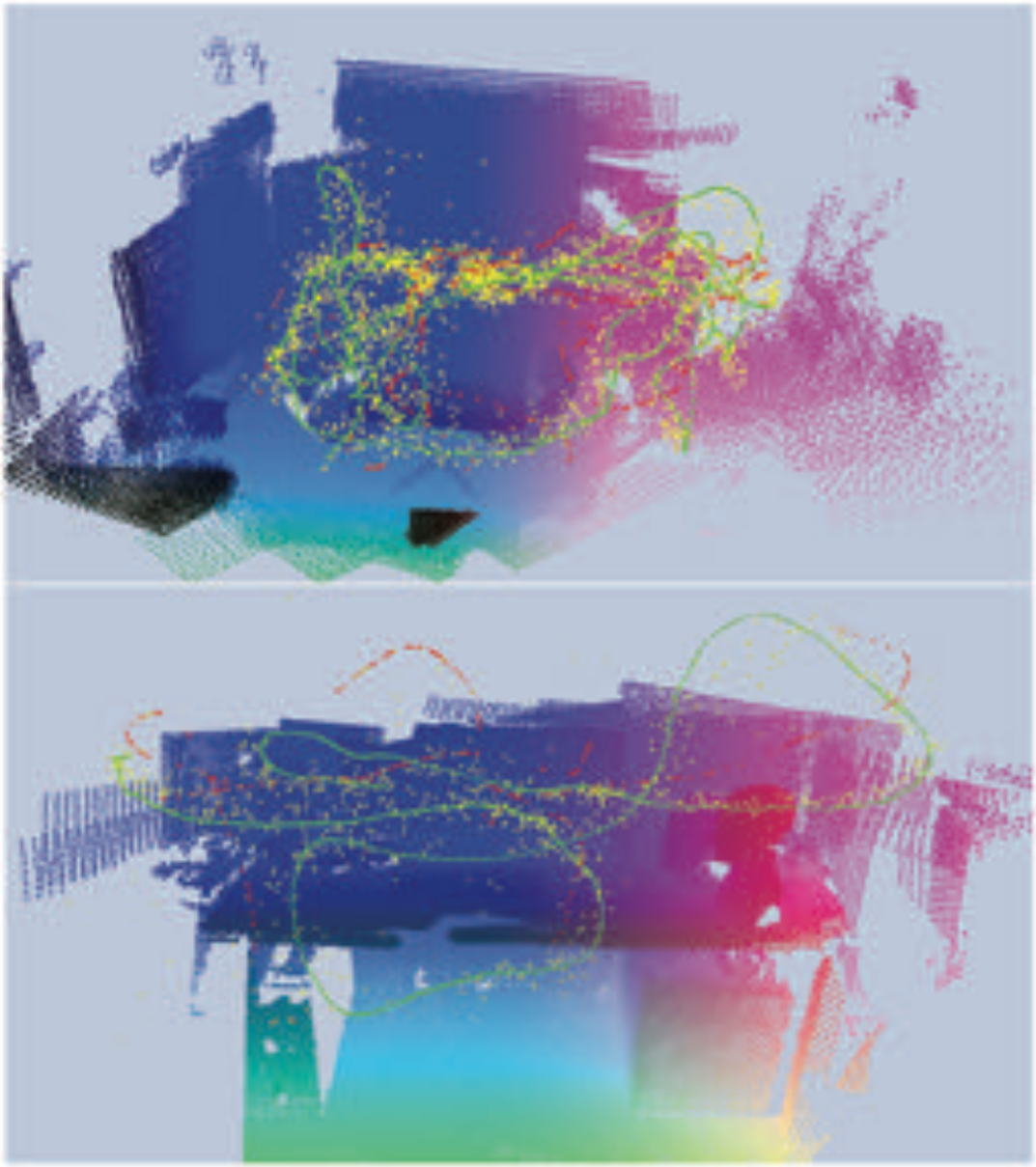


Figure 4.5: Visualization of the refined camera position: ground truth positions of test sequences are marked in green, initial position from image indexing in red and refined position after PnP in yellow. Environment from the 7 scenes dataset [226]: top chess and bottom heads.

Figure 4.6: Visualization of the refined camera position: ground truth positions of test sequences are marked in green, initial position from image indexing in red and refined position after PnP in blue. Environment from the 12 scenes dataset [308]: top apartment 1-kitchen and bottom manolis.

Scene	*Image retrieval*	PnLP
Apt1-kitchen	*0.12/7.7*	**0.09/4.1**
Apt1-living	*0.12/6.8*	**0.08/2.9**
Apt2-kitchen	***0.10***/*6.5*	**0.10/3.7**
Apt2-living	*0.11/5.6*	**0.10/4.7**
Apt2-bed	*0.13/7.0*	**0.12/5.7**
Apt2-luke	*0.15/7.2*	**0.14/5.5**
Office 5a	*0.12/5.3*	**0.09/3.6**
Office 5b	*0.15/7.2*	**0.10/4.7**
Lounge	*0.16/7.1*	**0.10/3.5**
Manolis	*0.13/6.3*	**0.09/3.7**
Gates362	*0.13/5.9*	**0.10/4.7**
Gates381	*0.15/7.7*	**0.11/4.4**

Table 4.4: Results on the 12 scenes [308] indoor dataset: we report median absolute position/orientation error in meters/degrees. Our model has been trained on a different dataset than the one used for testing. Best results are in **bold**.

4.4.3 Generalization

We report on table 4.4 localization errors the 12 Scenes dataset [308]. The 12 scenes dataset is used to evaluate the generalization capability of our method. For these experiments, we use the same network as mentioned earlier, trained on 7 Scenes dataset [276]. We observe an average relative improvement of $\times 1.2/\times 1.5$ in position/rotation from initial to refined pose, compare to $\times 2.8/\times 3.5$ on the 7 scenes dataset. 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 show on figure 4.6 positions recovered by our method on 2 of the 12 scenes of the dataset. Our method is able to compute new positions closer to the ground truth positions compare to the initial image retrieval step guess.

4.5 Unsupervised training and outdoor localization

In this section, we are interested in applied our method for outdoor localization. However, accurate dense depth map of outdoor environment for supervised depth from monocular CNN training are not easy to obtain. Thus, we decide to train our model in a unsupervised manner, *i.e.* without ground truth depth maps as training data. In the following, we first compare the impact of unsupervised training for indoor localization and then we test our proposal on outdoor scenes.

4.5.1 Unsupervised depth from monocular training

Training. To learn depth from RGB in a unsupervised manner, we follow the training procedure of [340], using the ground truth relative pose between images and by adding SSIM loss function for radiometric comparison as in [173]. We train the network with Adam optimizer, learning rate of 10^{-4} divided by two every 5 epochs. Training takes the same time as the supervised training, with a batch size reduced to 12.

Unsupervised depth from monocular at scale. It is not self-explanatory to claim that the depth maps produced from our unsupervised trained network [340] are at a real scale. In [340], authors use an auxiliary relative pose estimation network to make their method trainable with video sequences without any pre-processing. The counterpart is that the final CNN produces depth maps up to an unknown scale factor. Nevertheless, in our experiment they are at truth scale as we use the absolute 6-DoF camera pose (obtained by SfM or dense 3D fusion method [323]) to compute the relative position and orientation of the training images. Some learned depth maps can be found in figure 4.7, showing that unsupervised method leads to true scale depth values as long as it has been trained with true camera pose information.

	Scene	*Image retrieval* supervised	*Image retrieval* unsupervised	PnLP refinement supervised	PnLP refinement unsupervised
7-Scenes [276]	Chess	*0.29/13.0*	*0.34/15.4*	**0.07/2.7**	0.13/4.7
	Fire	*0.40/15.5*	*0.48/19.3*	**0.07/3.2**	0.22/8.2
	Heads	*0.28/20.5*	*0.25/17.9*	**0.05/3.9**	0.15/10.5
	Office	*0.38/13.0*	*0.50/16.1*	**0.09/2.9**	0.23/6.3
	Pumpkin	*0.43/13.1*	*0.54/15.0*	**0.13/3.6**	0.29/7.1
	Kitchen	*0.23/9.5*	*0.26/10.5*	**0.05/2.0**	0.12/3.3
	Stairs	*0.46/14.9*	*0.49/15.5*	**0.40/9.2**	0.48/12.2
12-Scenes [308]	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
	Lounge	*0.16/7.1*	*0.19/8.3*	**0.10/3.5**	0.13/4.7
	Manolis	*0.13/6.3*	*0.15/7.8*	**0.09/3.7**	0.12/4.5
	Gates362	*0.13/5.9*	*0.14/6.5*	**0.10**/4.7	0.11/**3.9**
	Gates381	*0.15/7.7*	*0.16/9.0*	**0.11/4.4**	0.13/5.1

Table 4.5: Results on the **7 scenes** [276] and **12 scenes** [308] indoor datasets, we report median position/orientation error in meters/degrees. Supervised (in purple) and unsupervised (in blue) refer to our model trained with, respectively without, truth depth maps as supervision signal. Best localization results are shown in **bold** and underlined numbers show failure cases when the pose refinement increases the initial pose error. Table best viewed in color.

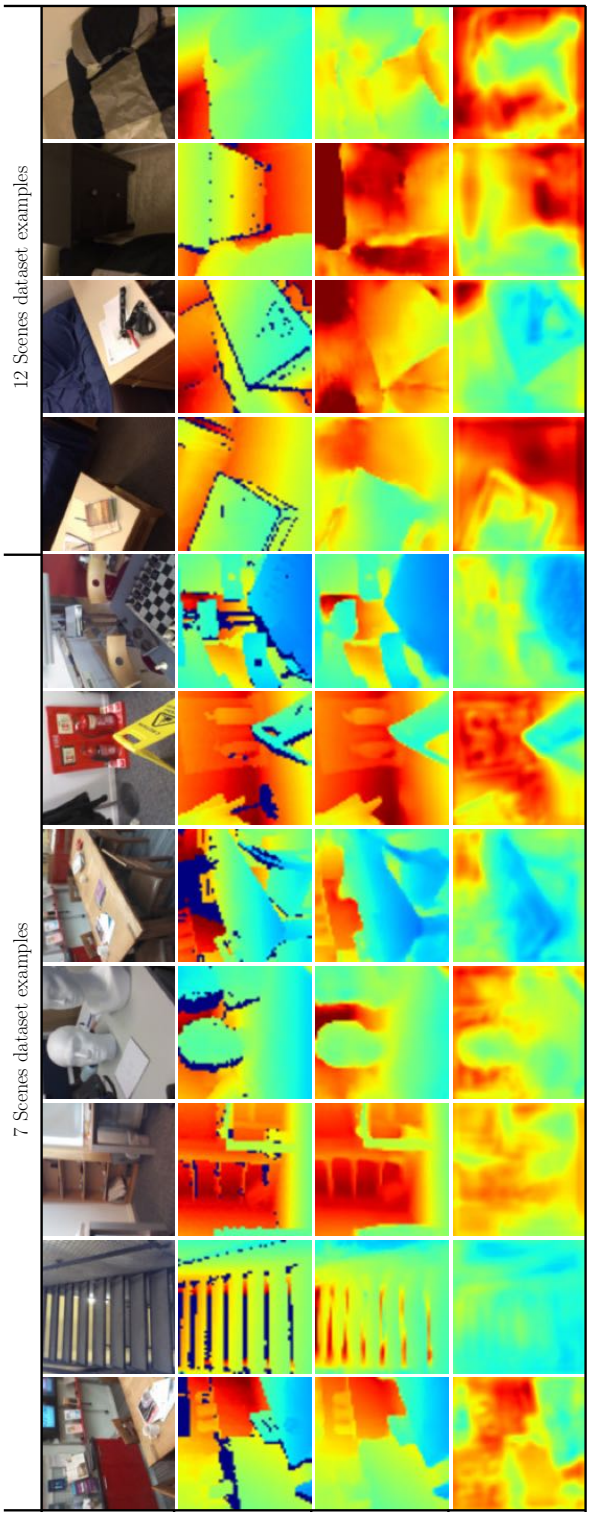


Figure 4.7: Visualization of the depth map generated from RGB input: from top to bottom: image, ground truth depth map, generated depth map (supervised training), generated depth map (unsupervised training). In both configurations (supervised and unsupervised), networks are trained on the **7 scenes** dataset [276]. Examples from **12 scenes** [308] show networks generalization capability.

4.5.2 Comparison with fully-supervised training

Localization results in indoor environments for the supervised and unsupervised training are shown in table 4.5. 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. Our unsupervised-trained proposal produces comparable localization as Relocnet and outperform Posenet baseline (see table 4.3). These encouraging results prove that, even without ground truth depth maps as supervision data, we can apply our refinement algorithm successfully.

For the novel scenes of the 12 scenes dataset, we observe an average relative improvement of $\times 1.2/\times 1.5$, compare to $\times 1.1/\times 1.6$, for the supervised, respectively unsupervised, model in position/rotation from initial to refined pose. We also show, in figure 4.7, the generalization capability of methods trained with or without ground truth depth maps, from images taken in both known and unknown scenes. We notice that the poor localization performance on the Apt2-bed scenes is closely related to the poor generated depth map on this scene (see figure 4.7, two last columns).

4.5.3 Outdoor localization

Dataset. We use the Cambridge Landmarks [128] 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. As no ground truth depth maps are available for the Cambridge Landmarks scenes, we only perform outdoor experiments related to the unsupervised depth from monocular training.

Network architecture and training. For the unsupervised scenario, we use a slightly different network architecture, composed of recurrent cells (Long Short-term Memory (LSTM)) in the decoder to capture long term dependencies [156, 311] (more details can be found in appendix A.2). During training and testing, images are rescaled to 224×112 pixels.

Results. Outdoor localization results are presented in table 4.6. PnLP performs well on outdoor scenes, with a mean improvement of $\times 1.5/\times 1.6$ in position/rotation precision over initial pose given by CBIR. 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

Scene	*Image retrieval*	PnLP	Posenet [127]
Great Court	*24.3/20.94*	13.2/10.07	-
Kings College	*5.0/5.86*	2.7/3.10	**0.9/1.04**
Old Hospital	*6.5/8.60*	3.5/5.55	**3.2/3.29**
Shop	*3.2/9.47*	1.1/**3.38**	**0.9**/3.78
St Mary's Church	*5.9/12.71*	2.6/5.85	**1.6/3.32**
Street	*92.5/67.10*	69.5/52.07	**20.3/25.5**

Table 4.6: Results on the Cambridge Landmarks [128] outdoor dataset: we report median position/orientation error in meters/degrees. We compare our network architectures trained in an unsupervised manner with Posenet [127]. Best localization results are shown in **bold**.

initial pose estimation at the CBIR preliminary step. Compared to Posenet [127], 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 [229] baseline because authors do not evaluate Relocnet on outdoor scenes and the source code is not yet available.

4.6 Discussion

The final camera pose accuracy is highly dependent on the images returned by the CBIR initial 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 like the one presented in previous chapter 3.

The pose refinement is also very sensitive to the quality of the generated depth map. Artifacts present in depth map related to images of unknown scenes or wrong reconstruction, as can be seen in the last 4 columns of figure 4.7, generate outliers for the PnLP optimization. This first work can be view as a proof of concept where we use very limited data for training our model. We can expect better results if we take advantages of bigger dataset, such as Megadepth dataset introduced in [155].

4.7 Conclusion

We have presented a new method for VBL consisting of an initial pose estimation by CBIR followed by a new relative pose estimation method. PnLP relies on densely matched 2D to 3D points between the query and the reference images, where the 3D points are project thank to the reconstructed depth map from a monocular image. The introduced method is compact and fast as all the components needed by the localization pipeline are computed using 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 the next chapter, we summarize our contributions on visual localization. We conclude the manuscript presenting potential improvement applicable to our system and discussing future works.

Chapter 5

Conclusion

In this final chapter, we summarize the main contributions of this thesis and we propose future research directions regarding the presented solutions.

5.1 Summary of the thesis

Throughout this thesis, we have focused on Visual-based Localization (VBL) in urban environment. We defined the boundaries of our research in the first chapter. In chapter 2, we reviewed exhaustively VBL related fields and methods, with a particular attention paid to challenges induced by long-term localization and to the data heterogeneity presents in the localization approaches. We came out with the conclusion that they are a lake of methods taking advantages of asymmetric data, *i.e.* different data modalities in the query side and in the database used as reference map. For instance, the query is often composed of a single modality, radiometric information, whereas the database is composed of multiple sensors information, like scene geometry, semantic and images.

In the chapter 3, we have proposed a new trainable global image descriptor for Content-based Image Retrieval (CBIR) for localization. Our method was built on the powerful NetVLAD [6] image descriptor, combined with a modality transfer model capable of generating a depth map from a single image. The particularity of our method remained in the fact that our descriptor could be trained using side modality that was not available during the task of localization. We showed that spreading out geometric clues within our pure radiometric descriptor improve the performances in challenging long-term localization scenarios. We were able to improve location accuracy for various realistic

scenarios, involving cross-season data and nocturnal images. We demonstrated that our method can take advantages of not-only geometric information but also reflectance given by laser scan sensor.

Chapter 4 was dedicated to our relocalization pipeline. Our method aimed to improve the localization given by our initial localization step. Using geometric reasoning, we refined 6-Degrees of Freedom (DoF) pose of the query to localize. In order to do so, we established dense 2D to 2D correspondences between the query and retrieved examples returned by our initial localization step. The matching was computed thanks to deep features extracted from a Convolutional Neural Networks (CNN). We defined our method to be lightweight and easily plugged after an existing CBIR for localization approach. In a same manner as our global image descriptor, the relocalization was aided by the geometric information learned during an offline stage. Indeed, the 2D to 2D matches were not sufficient to recover the truth 6-DoF pose of the query and the extra geometric information was used to constrain the final pose estimation. Through comprehensive experiments, we showed the effectiveness of our method in both indoor and outdoor scenes.

5.2 Scientific contributions

In the following, we summarize the major scientific contributions made during this thesis along with the corresponding publications:

Detailed review of VBL methods: we present a large panorama of image-based localization methods [220]. We propose a simple three-categories methods classification and we highlight common processing within the different VBL approaches. We also present in detail current challenges in localization as well as the different data types engaged in the localization process. We finally describe trends and common usages in the visual localization field as well as promising avenues for VBL.

Side modality trained global image descriptor: we introduce a new global image descriptor for VBL, trained with side geometric information [223]. By combining state-of-the-art global image descriptor along with modality transfer model, we are able to create discriminative image descriptors for localization in challenging conditions. To demonstrate the generalization capability of our proposal, we extend

this previous work with experiments on another dataset and we replace the geometric auxiliary modality by laser reflectance as side information during the training process [224].

Relative pose estimation from learned depth maps: we present a pose refinement method based on geometric alignment of learned depth map [221]. In a subsequent work [222], we improves the seminal method with a more efficient algorithm. We show the performances of the final method for pose refinement on indoor and outdoor environment.

5.3 Future Research

In this section, we enumerate possible improvements regarding the work presented in this thesis, as well as potential new research topics offered by new research results, including our contributions, and by ongoing proposal of various new datasets.

Towards an unified VBL pipeline. The two localization steps presented in this thesis have been developed in parallel during this thesis, resulting on two independent architectures. A straightforward improvement of this work would be the unification of these two frameworks, in order to get a complete two stage hierarchical localization method [257]. This unification would make possible the strict comparison of our localization results with the state of the art on VBL thanks to new challenging localization benchmarks, such as Sattler et al. [265].

Multi-task training. It will be interesting to investigate multi-task learning in order to address all the computer vision problems involved in VBL jointly. Our localization method involve global image description, establishments of dense correspondences between images and depth map generation from a monocular image. We only target one specific training task for our global image descriptor and our refinement method. Optimizing jointly the different tasks involved in our localization pipeline would certainly improve the overall precision of the system.

Heterogeneity in the geometry of acquisition. In this research work, we propose methods to deal with heterogeneity within the data modality. Indeed, we find a solution to benefit from extra modalities present in the reference dataset but not in the query side. Another interesting research oriented question would be: how to deal with visual data with different acquisition geometries? To be more specific, it would be interesting to tackle the problem of comparing perspective images with spherical ones [112, 237, 299, 331, 332]. Indeed, database coverage can be easily extended by using wide angle or omnidirectional cameras (*e.g.* google street view panorama). Furthermore, recent work have introduced a specific tool to exploit spherical geometry with deep learning: spherical CNN [61]. This new architecture has already been successfully used to solve a wide range of problems: room layout recovery from 360 images [82], depth estimation from spherical panorama [341], etc. We are convinced that similar approaches can be used to solve geometrically heterogeneous VBL problems.

Appendix A

Network architectures

A.1 Global image descriptor network

We use Alexnet or Resnet18 as backbone encoder for our image descriptor. NetVLAD layer is implemented as described in the original paper [6].

Our decoder is inspired by U-Net network architecture, *i.e.* we concatenate deep feature from encoder to recover fine details in the final depth maps. Up-sampling is performed through inverse convolutions and final activation is a sigmoid function, constraining the output depth (or reflectance) value to be in range $[0, 1]$. Non-linearity is assured by LeakyReLU and we use batch normalization and input data normalization.

A.2 Multitask pose refinement network

We build our own network, taking inspiration from the architecture Pix2Pix presented in [114]. Our encoder has 7 convolutional layers and our decoder has 5. We use skip connections between the encoder and the decoder. We use LeakyReLU non-linearity in our encoder and ReLu in our decoder and group normalization between layers. Final activation is a sigmoid function. Down-sampling in our encoder is performed through convolution operation only (using stride > 1). For up-sampling, we rely on bilinear interpolation followed by convolution for decoding the features maps to avoid artifact patterns induced by inverse convolution. Our final model as approximately $20M$ parameters (between Resnet18 and Resnet50).

A. NETWORK ARCHITECTURES

	Great Court	Kings C.	Old Hosp.	Shop	St Mary's	Street
FC	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	**13.2/10.07**	**2.7**/3.10	**3.5/5.55**	**1.1/3.38**	**2.6/5.85**	**69.5**/52.07

Table A.1: Results on the Cambridge Landmarks [128] outdoor dataset: we report median position/orientation error in meters/degrees. We compare our two network architectures trained in an unsupervised manner. Best localization results are shown in **bold**.

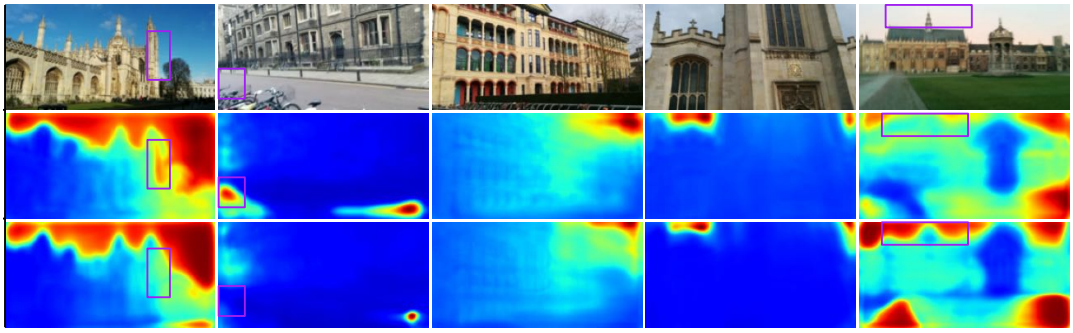


Figure A.1: Visualization of the depth map on outdoor scene: from top to bottom: image, generated depth map from architecture FC and generated depth map from architecture C+LSTM. Purple boxes show regions where C+LSTM network produces slightly better depth map reconstruction compared to FC.

For unsupervised depth from monocular training on outdoor scene, we add recurrent layers in our decoder to enforce spatial consistency. We replace two convolutional layers by bidirectional recurrent layer (4 Long Short-term Memory (LSTM) units: 1 for right to left, 1 for left to right, 1 for up to down and 1 for down to up). We design our spatial recurrent layers as in ReNet from [311]. We denote the fully convolutional architecture as **FC** and the recurrent variation as **C+LSTM**. We show in table A.1 the comparison of our two methods. Figure A.1 illustrates the better reconstruction capability of the **C+LSTM** architecture.

References

[1] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool. Night-to-Day Image Translation for Retrieval-based Localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number April, pages 2911–2918, 2012. ISBN 9781467312288. doi: 10.1109/CVPR.2012.6248018.

[3] R. Arandjelović and A. Zisserman. All About VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1585, 2013. ISBN 978-0-7695-4989-7. doi: 10.1109/CVPR.2013.207.

[4] R. Arandjelović and A. Zisserman. DisLocation : Scalable descriptor. In *Asian Conference on Computer Vision (ACCV)*, 2014.

[5] R. Arandjelović and A. Zisserman. Visual vocabulary with a semantic twist. In *Asian Conference on Computer Vision (ACCV)*, volume 9003, pages 178–195, 2014. ISBN 9783319168647. doi: 10.1007/978-3-319-16865-4_12.

[6] 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)*, pages 5297–5307, 2017. ISSN 10636919. doi: 10.1109/CVPR.2016.572.

[7] S. Ardeshir, A. R. Zamir, A. Torroella, and M. Shah. GIS-assisted object detection and geospatial localization. In *European Conference on Computer Vision (ECCV)*,

volume 8694 LNCS, pages 602–617, 2014. ISBN 9783319105987. doi: 10.1007/978-3-319-10599-4_39.

[8] A. Armagan, M. Hirzer, and V. Lepetit. Semantic Segmentation for 3D Localization in Urban Environments. In *Joint Urban Remote Sensing Event (JURSE)*, pages 3–6, 2017. ISBN 9781509058082.

[9] A. Armagan, M. Hirzer, P. M. Roth, and V. Lepetit. Learning to Align Semantic Segmentation and 2.5D Maps for Geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] A. Armagan, M. Hirzer, P. M. Roth, and V. Lepetit. Accurate Camera Registration in Urban Environments Using High-Level Feature Matching. In *British Machine Vision Conference (BMVC)*, pages 1–12, 2017.

[11] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide area localization on mobile phones. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 73–82, 2009. ISBN 978-1-4244-5390-0. doi: 10.1109/ISMAR.2009.5336494.

[12] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit. Instant Outdoor Localization and SLAM Initialization from 2.5D Maps. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 21(11):1309–1318, 2015. ISSN 10772626. doi: 10.1109/TVCG.2015.2459772.

[13] N. Atanasov, M. Zhu, K. Daniilidis, and G. J. Pappas. Localization from semantic observations via the matrix permanent. *The International Journal of Robotics Research (IJRR)*, 35(1-3):73–99, 2016. ISSN 0278-3649. doi: 10.1177/0278364915596589.

[14] M. Aubry, B. C. Russell, and J. 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.

[15] C. Azzi, D. Asmar, A. Fakih, and J. Zelek. Filtering 3D Keypoints Using GIST For Accurate Image-Based Localization. In *British Machine Vision Conference (BMVC)*, number 2, pages 1–12, 2016.

[16] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large Scale Visual Geo-Localization of Images in Mountainous Terrain. In *European Conference on Computer Vision (ECCV)*, volume 7573, pages 517–530, 2012. ISBN 978-3-642-33708-6.

[17] A. Babenko and V. Lempitsky. Aggregating local deep features for image retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 1269–1277, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.150.

[18] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural Codes for Image Retrieval. In *European Conference on Computer Vision (ECCV)*, pages 584–599, 2014. ISBN 978-3-319-10589-5. doi: 10.1007/978-3-319-10590-1_38.

[19] S. Bae, A. Agarwala, and F. Durand. Computational rephotography. *ACM Transactions on Graphics (ToG)*, 29(3):1–15, 2010. ISSN 07300301. doi: 10.1145/1805964.1805968.

[20] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference (BMVC)*, volume 2016-Septe, pages 119.1–119.11, 2016. doi: 10.5244/C.30.119.

[21] V. Balntas, S. Li, and V. Prisacariu. RelocNet : Continous Metric Learning Relocalisation using Neural Nets. In *European Conference on Computer Vision (ECCV)*, 2018.

[22] A. Bansal, H. Badino, and D. Huber. Understanding how camera configuration and environmental conditions affect appearance-based localization. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 800–807, 2014. ISBN 9781479936380. doi: 10.1109/IVS.2014.6856605.

[23] M. Bansal and K. Daniilidis. Geometric Urban Geo-Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis. Geo-localization of street views with aerial image databases. In *International Conference on Multimedia (MM)*, page 1125, 2011. ISBN 9781450306164. doi: 10.1145/2072298.2071954.

[25] M. Bansal, K. Daniilidis, and H. S. Sawhney. Ultra-wide baseline facade matching for geo-localization. In *European Conference on Computer Vision (ECCV)*, volume 7583 LNCS, pages 175–186, 2012. ISBN 9783642338625. doi: 10.1007/978-3-642-33863-2_18.

[26] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[27] B. Bescos, J. Neira, R. Siegwart, and C. Cadena. Empty Cities: Image Inpainting for a Dynamic-Object-Invariant Space. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[28] 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*, 125:16–32, 2017. ISSN 09242716. doi: 10.1016/j.isprsjprs.2017.01.005.

[29] N. Bhowmik, L. Weng, V. Gouet-Brunet, and B. Soheilian. Cross-domain Image Localization by Adaptive Feature Fusion. In *Joint Urban Remote Sensing Event (JURSE)*, 2017.

[30] F. Bonardi, S. Ainouz, R. Boutteau, Y. Dupuis, X. Savatier, and P. Vasseur. PHROG: A Multimodal Feature for Place Recognition. *Sensors*, 17(6):1167, 2017. ISSN 1424-8220. doi: 10.3390/s17051167.

[31] M. Boussaha, E. Fernandez-Moral, B. Vallet, and P. Rives. On the Production of Semantic and Textured 3D Meshes of Large scale Urban Environments from Mobile Mapping Images and LiDAR scans. In *Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP)*, 2018.

[32] M. Boussaha, B. Vallet, and P. Rives. Large Scale Textured Mesh Reconstruction from Mobile Mapping Images and LIDAR scans. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 4, pages 49–56, 2018. doi: 10.5194/isprs-annals-IV-2-49-2018.

[33] E. Brachmann and C. 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.

[34] E. Brachmann and C. Rother. Neural-Guided RANSAC: Learning Where to Sample Model Hypotheses. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[35] E. Brachmann, A. Krull, S. Nowozin, J. Shotton, F. Michel, S. Gumhold, and C. Rother. DSAC - Differentiable RANSAC for Camera Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.267.

[36] S. Brahmbhatt and J. Hays. DeepNav: Learning to navigate large cities. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2017-Janua, pages 3087–3096, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.329.

[37] S. Brahmbhatt, J. Gu, K. Kim, J. Hays, and J. 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.

[38] J. Brejcha and M. Čadík. State-of-the-art in visual geo-localization. *Pattern Analysis and Applications*, 2017. ISSN 1433-7541. doi: 10.1007/s10044-017-0611-1.

[39] J. Brejcha and M. Cadik. GeoPose3K: Mountain Landscape Dataset for Camera Pose Estimation in Outdoor Environments. *Image and Vision Computing*, 2017. ISSN 02628856. doi: 10.1016/j.imavis.2017.05.009.

[40] M. A. Brubaker, A. Geiger, and R. Urtasun. Lost! leveraging the crowd for probabilistic visual self-localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3057–3064, 2013. ISBN 1063-6919 VO -. doi: 10.1109/CVPR.2013.393.

[41] M. A. Brubaker, A. Geiger, and R. Urtasun. Map-based probabilistic visual self-localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(4):652–665, 2016. ISSN 01628828. doi: 10.1109/TPAMI.2015.2453975.

[42] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*

(ECCV), volume 6314 LNCS, pages 778–792, 2010. ISBN 364215560X. doi: 10.1007/978-3-642-15561-1_56.

[43] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (6):679–698, 1986.

[44] S. Cao and N. Snavely. Graph-Based Discriminative Learning for Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 112, pages 239–254, 2013. ISBN 1063-6919. doi: 10.1007/s11263-014-0774-9.

[45] N. Carlevaris-Bianco and R. M. Eustice. Learning visual feature descriptors for dynamic lighting conditions. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2769–2776, 2014. ISBN 9781479969340. doi: 10.1109/IROS.2014.6942941.

[46] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice. University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research (IJRR)*, 35(9):1023–1035, 2016. ISSN 0278-3649. doi: 10.1177/0278364915614638.

[47] T. Caselitz, B. Steder, M. Ruhnke, and W. Burgard. Matching Geometry for Long-term Monocular Camera Localization. In *IEEE International Conference on Robotics and Automation Workshop (ICRAW)*, 2016.

[48] F. Castaldo, A. R. Zamir, R. Angst, F. Palmieri, and S. Savarese. Semantic Cross-View Matching. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, volume 2016-Febru, pages 1044–1052, 2015. ISBN 9781467383905. doi: 10.1109/ICCVW.2015.137.

[49] T. Cavallari, S. Golodetz, N. A. Lord, J. Valentin, L. Di Stefano, and P. H. S. Torr. On-the-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[50] T. Cavallari, S. Golodetz, N. A. Lord, J. Valentin, V. A. Prisacariu, L. Di Stefano, and P. H. S. Torr. Real-Time RGB-D Camera Pose Estimation in Novel Scenes using a Relocalisation Cascade. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–18, 2018. doi: arXiv:1810.12163v1.

[51] T. Cavallari, L. Bertinetto, J. Mukhoti, P. Torr, and S. Golodetz. Let's Take This Online: Adapting Scene Coordinate Regression Network Predictions for Online RGB-D Camera Relocalisation. In *International Conference on 3D Vision (3DV)*, number ii, 2019.

[52] T. J. Cham, A. Ciptadi, W. C. Tan, M. T. Pham, and L. T. Chia. Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 366–373, 2010. ISBN 9781424469840. doi: 10.1109/CVPR.2010.5540191.

[53] J. Chan, J. A. Lee, and Q. Kemao. F-SORT : An Alternative for Faster Geometric Verification. In *Asian Conference on Computer Vision (ACCV)*, pages 1–15, 2016.

[54] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 737–744, 2011. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995610.

[55] Y. Chen, G. Qian, K. Gunda, H. Gupta, and K. Shafique. Camera geolocation from mountain images. In *International Conference on Information Fusion (Fusion)*, pages 1587–1596, 2015. ISBN 9780996452717.

[56] W. Cheng, W. Lin, K. Chen, and X. Zhang. Cascaded Parallel Filtering for Memory-Efficient Image-Based Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[57] G. Christie, G. Warnell, and K. Kochersberger. Semantics for UGV Registration in GPS-denied Environments. *arXiv preprint*, 2016.

[58] O. Chum and J. Matas. Matching with PROSAC-progressive sample consensus. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 220–226. IEEE, 2005.

[59] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *IEEE*

International Conference on Computer Vision (ICCV), 2007. ISBN 978-1-4244-1630-1. doi: 10.1109/ICCV.2007.4408891.

[60] O. Chum, A. Mikul, M. Perdoch, and J. Matas. Total Recall II : Query Expansion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[61] T. S. Cohen, M. Geiger, J. Koehler, and M. Welling. Spherical CNNs. In *International Conference on Machine Learning (ICML)*, number 3, pages 1–15, 2018.

[62] L. Contreras and W. Mayol-Cuevas. Towards CNN Map Compression for camera relocalisation. *arXiv preprint*, 2017.

[63] P. Corke, R. Paul, W. Churchill, and P. Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 2085–2092, 2013. ISBN 9781467363587. doi: 10.1109/IROS.2013.6696648.

[64] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research (IJRR)*, 27:647–665, 2008. ISSN 0278-3649. doi: 10.1177/0278364908090961.

[65] M. Cummins and P. Newman. Accelerating FAB-MAP with concentration inequalities. *IEEE Transactions on Robotics (TRO)*, 26(6):1042–1050, 2010. ISSN 15523098. doi: 10.1109/TRO.2010.2080390.

[66] R. Cupec, E. K. Nyarko, D. Filko, and L. Markasović. Recognition of Objects and Places in 3D Point Clouds for Robotic Applications. In *International Conference & Workshop on Mechatronics in Practice and Education (MECHEDU).*, 2015.

[67] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[68] H. Deng, T. Birdal, and S. Ilic. PPFNet: Global Context Aware Local Features for Robust 3D Point Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00028.

[69] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[70] D. DeTone, T. Malisiewicz, and A. Rabinovich. Toward Geometric Deep SLAM. *arXiv preprint*, 2017.

[71] D. Detone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 2018-June, pages 337–349, 2018. ISBN 9781538661000. doi: 10.1109/CVPRW.2018.00060.

[72] M. Donoser and D. Schmalstieg. Discriminative feature-to-point matching in image-based localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 516–523, 2014. ISBN 9781479951178. doi: 10.1109/CVPR.2014.73.

[73] N. D. Duong, A. Kacete, C. Sodalie, P. Y. Richard, and J. Royan. XyzNet: Towards Machine Learning Camera Relocalization by Using a Scene Coordinate Prediction Network. *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 258–263, 2018. doi: 10.1109/ISMAR-Adjunct.2018.00080.

[74] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[75] D. Eigen, C. Puhrsch, and R. 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.

[76] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard. 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, 2015. ISBN 9781479999941. doi: 10.1109/IROS.2015.7353446.

[77] G. Elbaz, T. Avraham, and G. Elbaz. 3D Point Cloud Registration for Localization using a Deep Neural Network Auto-Encoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number July, pages 4631–4640, 2017.

[78] S. En, A. Lechervy, and F. Jurie. RPNet: an End-to-End Network for Relative Camera Pose Estimation. In *European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[79] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(3):611–625, 2017.

[80] Eric Brachmann and Carsten Rother. Expert Sample Consensus Applied to Camera Re-Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[81] Y. Feng, L. Fan, and Y. Wu. Fast Localization in Large-Scale Environments Using Supervised Indexing of Binary Features. *IEEE Transactions on Image Processing (ToIP)*, 25(1):343–358, 2016. ISSN 1057-7149. doi: 10.1109/TIP.2015.2500030.

[82] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero. Corners for Layout: End-to-End Layout Recovery from 360 Images. *arXiv preprint*, pages 1–10, 2019.

[83] E. Fernandez-Moral, W. Mayol-Cuevas, V. Arevalo, and J. Gonzalez-Jimenez. Fast place recognition with plane-based maps. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2719–2724, 2013. ISBN 9781467356411. doi: 10.1109/ICRA.2013.6630951.

[84] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(1):59–68, 2006.

[85] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[86] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics (TRO)*, 33(2):249–265, 2016.

[87] W. Förstner and B. P. Wrobel. *Photogrammetric Computer Vision*. Springer, 2016.

[88] E. Garcia-Fidalgo and A. Ortiz. Vision-based topological mapping and localization methods: A survey. *Robotics and Autonomous Systems (RAS)*, 64:1–20, 2015. ISSN 09218890. doi: 10.1016/j.robot.2014.11.009.

[89] S. Garg, A. Jacobson, S. Kumar, and M. J. Milford. Improving Condition and Environment-Invariant Place Recognition with Semantic Place Categorization. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[90] S. Garg, N. Suenderhauf, and M. J. Milford. Don't Look Back: Robustifying Place Categorization for Viewpoint- and Condition-Invariant Place Recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018. ISBN 9781538630808.

[91] S. Garg, N. Suenderhauf, and M. J. Milford. LoST? Appearance-Invariant Place Recognition for Opposite Viewpoints using Visual Semantics. *Robotics Science and Systems (RSS)*, 2018.

[92] A. P. Gee and W. Mayol-Cuevas. 6D Relocalisation for RGBD Cameras Using Synthetic View Regression. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2012. ISBN 1-901725-46-4. doi: 10.5244/C.26.113.

[93] H. Germain, G. Bourmaud, and V. Lepetit. Efficient Condition-based Representations for Long-Term Visual Localization. *arXiv preprint*, 2018.

[94] H. Germain, G. Bourmaud, and V. Lepetit. Sparse-to-Dense Hypercolumn Matching for Long-Term Visual Localization. In *International Conference on 3D Vision (3DV)*, 2019.

[95] A. Gionis, P. Indyk, and R. Motwani. Similarity Search in High Dimensions via Hashing. In *Proceedings of the 25th VLDB Conference*, pages 518–529, 1999.

[96] B. Glocker, S. Izadi, J. Shotton, and A. Criminisi. Real-time RGB-D camera relocalization. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179, 2013. ISBN 9781479928699. doi: 10.1109/ISMAR.2013.6671777.

[97] B. Glocker, J. Shotton, A. Criminisi, and S. Izadi. Real-time RGB-D camera relocalization via randomized ferns for keyframe encoding. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 21(5):571–583, 2015. ISSN 10772626. doi: 10.1109/TVCG.2014.2360403.

[98] C. Godard, O. Mac Aodha, and G. 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.

[99] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In *European Conference on Computer Vision (ECCV)*, pages 1–17, 2014.

[100] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep Image Retrieval: Learning Global Representations for Image Search. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 241–257, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0.

[101] 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)*, 124(2):237–254, 2017. ISSN 15731405. doi: 10.1007/s11263-017-1016-8.

[102] S. Griffith and C. Pradalier. Survey Registration For Long–Term Natural Environment Monitoring. *Journal of Field Robotics*, 2017.

[103] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. 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, 2014. ISBN 9783319105833. doi: 10.1007/978-3-319-10584-0_23.

[104] A. Guzman-Rivera, K. Pushmeet, B. Glocker, J. Shotton, T. Sharp, A. Fitzgibbon, and S. Izadi. Multi-Output Learning for Camera Relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2014.

[105] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[106] J. Hays and A. A. Efros. IM2GPS: Estimating Geographic Information From a Single Image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 05, 2008. ISBN 9781424422432.

[107] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[108] I. Heisterklaus, N. Qian, and A. Miller. Image-based pose estimation using a compact 3D model. In *IEEE International Conference on Consumer Electronics Berlin (ICCE-Berlin)*, pages 327–330, 2014. ISBN 9781479961658.

[109] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *Annual Conference on Neural Information Processing Systems Workshop (NIPSW)*, 2015.

[110] J. Hoffman, S. Gupta, and T. Darrell. Learning with Side Information through Modality Hallucination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.96.

[111] A. Irschara, C. Zach, J.-m. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[112] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Panorama to panorama matching for location recognition. In *ACM International Conference on Multimedia Retrieval (ICMR)*, 2017. ISBN 9781450347013. doi: 10.1145/3078971.3079033.

[113] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Mining on Manifolds: Metric Learning without Labels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00797.

[114] P. Isola, J.-Y. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.632.

[115] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval : the benefit of PCA and whitening. In *European Conference on Computer Vision (ECCV)*, 2012.

[116] H. Jégou and A. Zisserman. Triangulation embedding and democratic aggregation for image search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[117] H. Jégou, M. Douze, and C. Schmid. Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search. In *European Conference on Computer Vision (ECCV)*, number October, pages 304–317, 2008. ISBN 9783540886815. doi: 10.1007/978-3-540-88682-2_24.

[118] H. Jégou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1169–1176, 2009. ISBN 9781424439935. doi: 10.1109/CVPRW.2009.5206609.

[119] H. Jégou, M. Douze, and C. Schmid. Product Quantization for Nearest Neighbor Search Herve. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 33(1):117–128, 2011. ISSN 1939-3539. doi: 10.1109/TPAMI.2010.57.

[120] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, C. Schmid, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1–12, 2012.

[121] T. Jenicek and O. Chum. No Fear of the Dark: Image Retrieval under Varying Illumination Conditions. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[122] D. Jia, Y. Su, and C. Li. Deep Convolutional Neural Network for 6-DOF Image Localization. *arXiv preprint*, (413113):1790–1798, 2016.

[123] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual Losses for Style Transfer and Super-Resolution. In *European Conference on Computer Vision (ECCV)*, pages 1–5, 2016. ISBN 978-3-319-46475-6. doi: 10.1007/978-3-319-46475-6_43.

[124] J. Johnson, M. Douze, and H. Jégou. Billion-scale similarity search with GPUs. *arXiv preprint*, 2017.

[125] S. Katz, A. Tal, and R. Basri. Direct visibility of point sets. In *ACM SIGGRAPH Conference on Computer Graphics*, 2007. doi: 10.1145/1275808.1276407.

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

[127] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[128] A. Kendall, M. Grimes, and R. 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.

[129] A. Kendall, Y. Gal, and R. Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[130] D.-K. Kim and M. R. Walter. Satellite Image-based Localization via Learned Embeddings. *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[131] H. J. Kim, E. Dunn, and J.-M. Frahm. Predicting good features for image geo-localization using per-bundle VLAD. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 1170–1178, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.139.

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

[133] L. Kneip and P. 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.

[134] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2976. IEEE, 2011.

[135] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6129–6138, 2017.

[136] I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems (RAS)*, 66:86–103, 2015. ISSN 09218890. doi: 10.1016/j.robot.2014.12.006.

[137] T. Krajník, J. Faigl, V. Vonásek, K. Košnar, M. Kulich, and L. Preucil. Simple yet stable bearing-only navigation. *Journal of Field Robotics*, 27(5):511–533, 2010. ISSN 15564959. doi: 10.1002/rob.20354.

[138] T. Krajník, J. P. Fentanes, O. M. Mozos, T. Duckett, J. Ekekrantz, and M. Hanheide. Long-Term Topological Localisation for Service Robots in Dynamic Environments using Spectral Maps. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, number Iros, pages 4537–4542, 2014. ISBN 9781479969333.

[139] T. Krajník, P. Cristoforis, K. Kusumam, P. Neubert, and T. Duckett. Image features for visual teach-and-repeat navigation in changing environments. *Robotics and Autonomous Systems (RAS)*, 88(November):127–141, 2017. ISSN 09218890. doi: 10.1016/j.robot.2016.11.011.

[140] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, pages 1097–1105, 2017.

[141] T. Kroeger and L. Van Gool. Video registration to SfM models. In *European Conference on Computer Vision (ECCV)*, volume 8693 LNCS, pages 1–16, 2014. ISBN 978-3-319-10601-4. doi: 10.1007/978-3-319-10602-1_1.

[142] D. Kumar, H. Neher, A. Das, D. A. Clausi, and S. L. Waslander. Condition and viewpoint invariant omni-directional place recognition using cnn. In *Conference on Computer and Robot Vision (CRV)*, pages 32–39. IEEE, 2017.

[143] Z. Laskar, I. Melekhov, S. Kalia, and J. Kannala. Camera Relocalization by Predicting Pairwise Relative Poses Using Convolutional Neural Network. *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 920–929, 2017. doi: 10.1109/ICCVW.2017.113.

[144] S. Leutenegger, M. Chli, and R. Siegwart. BRISK: Binary robust invariant scalable keypoints. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2548–2555. IEEE, 2011.

[145] J. Lezama, R. Von Gioi, G. Randall, and J.-M. Morel. Finding vanishing points via point alignments in image primal and dual domains. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 509–515, 2014.

[146] Q. Li, J. Zhu, R. Cao, K. Sun, J. M. Garibaldi, Q. Li, B. Liu, and G. Qiu. Relative Geometry-Aware Siamese Neural Network for 6DOF Camera Relocalization. *arXiv preprint*, pages 1–12, 2019.

[147] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu. Night-time indoor relocalization using depth image with Convolutional Neural Networks. In *IEEE International Conference on Automation and Computing (ICAC)*, pages 261–266, 2016. ISBN 9781862181311. doi: 10.1109/IConAC.2016.7604929.

[148] R. Li, Q. Liu, J. Gui, D. Gu, and H. Hu. Indoor Relocalization in Challenging Environments With Dual-Stream Convolutional Neural Networks. *IEEE Transactions on Automation Science and Engineering*, pages 1–12, 2017.

[149] S. Li and A. Calway. Absolute pose estimation using multiple forms of correspondences from RGB-D frames. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2016-June, pages 4756–4761, 2016. ISBN 9781467380263. doi: 10.1109/ICRA.2016.7487678.

[150] 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)*, 40(8):2030 2036, 2018. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2734890.

[151] X. Li, M. Larson, and A. Hanjalic. Pairwise geometric matching for large-scale object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 5153–5161, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7299151.

[152] X. Li, J. Ylioinas, J. Verbeek, and J. Kannala. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 1–17, 2018.

[153] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *European Conference on Computer Vision (ECCV)*, pages 791–804, 2010. ISBN 978-3-642-15551-2. doi: 10.1007/978-3-642-15552-9_57.

[154] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *European Conference on Computer Vision (ECCV)*, pages 15–29, 2012. ISBN 978-3-642-33717-8. doi: 10.1007/978-3-642-33718-5_2.

[155] Z. Li and N. Snavely. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00218.

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

[157] J. Z. Liang, N. Corso, E. Turner, and A. Zakhor. Image Based Localization in Indoor Environments. In *Computing for Geospatial Research and Application*, 2013.

[158] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 891–898, 2013. ISBN 978-0-7695-4989-7. doi: 10.1109/CVPR.2013.120.

[159] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning Deep Representations for Ground-to-Aerial Geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number JUNE, pages 5007–5015, 2015. ISBN 9781467369640.

[160] C. Linegar, W. Churchill, and P. Newman. Made to measure: Bespoke landmarks for 24-hour, all-weather localisation with a camera. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2016-June, pages 787–794, 2016. ISBN 9781467380263. doi: 10.1109/ICRA.2016.7487208.

[161] L. Liu, H. Li, and Y. Dai. Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[162] L. Liu, H. Li, and Y. Dai. Deep Stochastic Attraction and Repulsion Embedding for Image Based Localization. *arXiv preprint*, 2018.

[163] Z. Liu and R. Marlet. Virtual line descriptor and semi-local matching method for reliable feature correspondence. In *British Machine Vision Conference (BMVC)*, pages 11–16, 2012.

[164] Z. Liu, L.-Y. Duan, J. Chen, and T. Huang. Depth-Based Local Feature Selection for Mobile Visual Search. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

[165] S. Y. Loo, A. J. Amiri, S. Mashohor, S. H. Tang, and H. 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.

[166] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[167] S. Lowry and M. J. Milford. Supervised and Unsupervised Linear Learning Techniques for Visual Place Recognition in Changing Environments. *IEEE Transactions on Robotics (TRO)*, 32(3):600–613, 2016. ISSN 15523098. doi: 10.1109/TRO.2016.2545711.

[168] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. 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.

[169] G. Lu, Y. Yan, L. Ren, J. Song, N. Sebe, and C. Kambhamettu. Localize Me Anywhere, Anytime: A Multi-task Point-Retrieval Approach. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2434–2442, 2015. doi: 10.1109/ICCV.2015.280.

[170] G. Lu, Y. Yan, L. Ren, P. Saponaro, N. Sebe, and C. Kambhamettu. Where am I in the dark: Exploring active transfer learning on the use of indoor localization based on thermal imaging. *Neurocomputing*, 173:83–92, 2016. ISSN 18728286. doi: 10.1016/j.neucom.2015.07.106.

[171] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart. Get out of my lab: Large-scale, real-time visual-intertial localization. *Robotics Science and Systems (RSS)*, pages 37–46, 2015. doi: 10.15607/RSS.2015.XI.037.

[172] 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.

[173] R. Mahjourian, M. Wicke, and A. 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.

[174] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza. MAV urban localization from Google street view data. In *IEEE International Conference on Intelligent*

Robots and Systems (IROS), pages 3979–3986, 2013. ISBN 9781467363587. doi: 10.1109/IROS.2013.6696925.

[175] E. Marchand, H. Uchiyama, and F. Spindler. Pose Estimation for Augmented Reality : a Hands-On Survey. *IEEE Transactions on Visualization and Computer Graphics (ToVCG)*, 22(12):2633–2651, 2016. ISSN 1077-2626. doi: 10.1109/TVCG.2015.2513408.

[176] J. Mason, S. Ricco, and R. Parr. Textured Occupancy Grids for Monocular Localization Without Features. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5800–5806, 2011. ISBN 9781612843803. doi: 10.1109/ICRA.2011.5980506.

[177] D. Massiceti, A. Krull, E. Brachmann, C. Rother, and P. H. S. Torr. Random Forests versus Neural Networks - What's Best for Camera Relocalization? In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[178] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, 2004.

[179] B. C. Matei, N. Vander Valk, Z. Zhu, H. Cheng, and H. S. Sawhney. Image to LIDAR matching for geotagging in urban environments. In *IEEE Workshop on Applications of Computer Vision (WACV)*, pages 413–420, 2013. ISBN 9781467350532. doi: 10.1109/WACV.2013.6475048.

[180] C. McManus, B. Upcroft, and P. Newman. Scene Signatures : Localised and Pointless Features for Localisation. In *Robotics Science and Systems (RSS)*, 2014. ISBN 9780992374709.

[181] I. Melekhov, J. Kannala, and E. Rahtu. Relative Camera Pose Estimation Using Convolutional Neural Networks. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 1–12, 2017.

[182] L. Meng, J. Chen, F. Tung, J. J. Little, and C. W. de Silva. Exploiting Random RGB and Sparse Features for Camera Pose Estimation. In *British Machine Vision Conference (BMVC)*, pages 1–12, 2016.

[183] M. Menze and A. Geiger. Object Scene Flow for Autonomous Vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[184] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-DOF localization on mobile devices. In *European Conference on Computer Vision (ECCV)*, volume 8690 LNCS, pages 268–283, 2014. ISBN 9783319106045. doi: 10.1007/978-3-319-10605-2_18.

[185] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)*, 60(1):63–86, 2004.

[186] M. J. Milford and G. F. Wyeth. 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, 2012. ISBN 9781467314039. doi: 10.1109/ICRA.2012.6224623.

[187] M. J. Milford, S. Lowry, S. Shirazi, E. Pepperell, C. Shen, G. Lin, F. Liu, C. Cadena, and I. Reid. Sequence Searching with Deep-learnt Depth for Condition- and Viewpoint- invariant Route-based Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 18–25, 2015. ISBN 9781467367592.

[188] B. Morago, G. Bui, and Y. Duan. An Ensemble Approach to Image Matching Using Contextual Features. *IEEE Transactions on Image Processing (ToIP)*, 24 (11):4474–4487, 2015.

[189] B. Morago, G. Bui, and Y. Duan. 2D Matching Using Repetitive and Salient Features in Architectural Images. *IEEE Transactions on Image Processing (ToIP)*, 7149(c):1–12, 2016. ISSN 1057-7149. doi: 10.1109/TIP.2016.2598612.

[190] J.-M. Morel and G. Yu. ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, 2009.

[191] P. Moulon, P. Monasse, R. Perrot, and R. Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

[192] A. Mousavian, J. Kosecká, and J. M. Lien. Semantically guided location recognition for outdoors scenes. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2015-June, pages 4882–4889, 2015. ISBN 978-1-4799-6923-4. doi: 10.1109/ICRA.2015.7139877.

[193] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale. Summary Maps for Lifelong Visual Localization. *Journal of Field Robotics*, 23(0): 245–267, 2015. ISSN 14746670. doi: 10.1002/rob.

[194] M. Muja and D. G. Lowe. Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 1–10, 2009. ISBN 9789898111692. doi: 10.1.1.160.1721.

[195] A. C. Murillo, G. Singh, J. Kosecká, and J. J. Guerrero. Localization in urban environments using a panoramic gist descriptor. *IEEE Transactions on Robotics (TRO)*, 29(1):146–160, 2013. ISSN 15523098. doi: 10.1109/TRO.2012.2220211.

[196] T. Naseer, G. L. Oliveira, T. Brox, and W. Burgard. Semantics-aware Visual Localization under Challenging Perceptual Conditions. *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2614–2620, 2017.

[197] P. Nelson, W. Churchill, I. Posner, and P. Newman. From Dusk till Dawn: Localisation at Night using Artificial Light Sources. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5245–5252, 2015. ISBN 9781479969234. doi: 10.1109/ICRA.2015.7139930.

[198] P. Neubert, N. Sünderhauf, and P. Protzel. Superpixel-based appearance change prediction for long-term navigation across seasons. *Robotics and Autonomous Systems (RAS)*, 69(1):15–27, 2015. ISSN 09218890. doi: 10.1016/j.robot.2014.08.005.

[199] K. Ni, A. Kannan, A. Criminisi, and J. Winn. Epitomic location recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 31 (12):2158–2167, 2009. ISSN 01628828. doi: 10.1109/TPAMI.2009.165.

[200] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, 2006. ISBN 0769525970. doi: 10.1109/CVPR.2006.264.

[201] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-Scale Image Retrieval with Attentive Deep Local Features. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2017-Octob, pages 3476–3485, 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.374.

[202] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision (IJCV)*, 42(3): 145–175, 2001. ISSN 09205691. doi: 10.1023/A:1011139631724.

[203] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. LF-Net: Learning Local Features from Images. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2018.

[204] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. LF-Net: Learning Local Features from Images. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[205] P. Panphattarasap and A. Calway. Visual place recognition using landmark distribution descriptors. In *Asian Conference on Computer Vision (ACCV)*, 2016.

[206] 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*, 200(1):69–79, 2012.

[207] G. Pascoe, W. Maddern, and P. Newman. Robust Direct Visual Localisation using Normalised Information Distance. *British Machine Vision Conference (BMVC)*, pages 1–13, 2015.

[208] G. Pascoe, W. Maddern, and P. Newman. Direct Visual Localisation and Calibration for Road Vehicles in Changing City Environments. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2016-Febru, pages 98–105, 2015. ISBN 9781467383905. doi: 10.1109/ICCVW.2015.23.

[209] G. Pascoe, W. Maddern, A. D. Stewart, and P. Newman. FARLAP : Fast Robust Localisation using Appearance Priors. In *IEEE International Conference on*

Robotics and Automation (ICRA), pages 6366–6373, 2015. ISBN 9781479969227. doi: 10.1109/ICRA.2015.7140093.

[210] D. P. Paudel, A. Habed, C. Demonceaux, and P. Vasseur. LMI-based 2D-3D registration: From uncalibrated images to Euclidean scene. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4494–4502, 2015.

[211] D. P. Paudel, A. Habed, C. Demonceaux, and P. Vasseur. Robust and Optimal Sum-of-Squares-Based Point-to-Plane Registration of Image Sets and Structured Scenes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2048–2056, 2015.

[212] R. Paul and P. Newman. FAB-MAP 3D: Topological mapping with spatial and visual appearance. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2649–2656, 2010. ISBN 9781424450381. doi: 10.1109/ROBOT.2010.5509587.

[213] M. Paulin, M. Douze, Z. Harchaoui, J. Mairal, F. Perronnin, and C. Schmid. Local convolutional features with unsupervised training for image retrieval. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 91–99, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.19.

[214] M. Paulin, J. Mairal, M. Douze, Z. Harchaoui, F. Perronnin, and C. Schmid. Convolutional Patch Representations for Image Retrieval: An Unsupervised Approach. *International Journal of Computer Vision (IJCV)*, 121(1):149–168, 2017. ISSN 15731405. doi: 10.1007/s11263-016-0924-3.

[215] E. Pepperell, P. Corke, and M. J. Milford. All - Environment Visual Place Recognition with SMART. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014. ISBN 9781479936854.

[216] E. Pepperell, P. Corke, and M. J. Milford. Routed roads: Probabilistic vision-based place recognition for changing conditions, split streets and varied viewpoints. *The International Journal of Robotics Research (IJRR)*, 35(9):1057–1179, 2016. ISSN 0278-3649. doi: 10.1177/0278364915618766.

[217] F. Perronnin and Y. Liu. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. ISBN 9781424469833.

[218] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. ISBN 1424411807. doi: 10.1109/CVPR.2007.383172.

[219] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[220] N. Piasco, D. Sidibé, C. Demonceaux, and V. 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.

[221] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet. Geometric Camera Pose Refinement With Learned Depth Maps. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[222] N. Piasco, D. Sidibé, C. Demonceaux, and V. Gouet-Brunet. Perspective-n-Learned-Point: Pose Estimation from Relative Depth. In *British Machine Vision Conference (BMVC)*, 2019.

[223] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux. Learning Scene Geometry for Visual Localization in Challenging Conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[224] N. Piasco, D. Sidibé, V. Gouet-Brunet, and C. Demonceaux. Improving Image Description with Auxiliary Modality for Visual Localization in Challenging Conditions. *(in submission)*, 2019.

[225] C. Poglitsch, C. Arth, D. Schmalstieg, and J. Ventura. A particle filter approach to outdoor localization using image-based rendering. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 132–135, 2015. ISBN 9781467376600. doi: 10.1109/ISMAR.2015.39.

[226] F. Pomerleau, F. Colas, and R. Siegwart. A Review of Point Cloud Registration Algorithms for Mobile Robotics. *Foundations and Trends in Robotics*, 4(1):1–104, 2015. ISSN 1935-8253. doi: 10.1561/2300000035.

[227] H. Porav, W. Maddern, and P. Newman. Adversarial Training for Adverse Conditions: Robust Metric Localisation using Appearance Transfer. *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[228] H. Porav, T. Bruls, and P. Newman. I Can See Clearly Now : Image Restoration via De-Raining. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[229] P. Purkait, C. Zhao, and C. Zach. Synthetic View Generation for Absolute Pose Regression and Image Synthesis. In *British Machine Vision Conference (BMVC)*, 2018.

[230] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[231] X. Qu, B. Soheilian, and N. Paparoditis. Vehicle localization using mono-camera and geo-referenced traffic signs. *IEEE Intelligent Vehicles Symposium (IV)*, 2015-Augus:605–610, 2015. ISSN 1098-6596. doi: 10.1109/IVS.2015.7225751.

[232] X. Qu, B. Soheilian, E. Habets, and N. Paparoditis. Evaluation of SIFT and SURF for vision based localization. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41(July):685–692, 2016. ISSN 16821750. doi: 10.5194/isprsarchives-XLI-B3-685-2016.

[233] F. Radenović, G. Tolias, and O. Chum. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 3–20, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0.

[234] 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.

[235] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00598.

[236] N. Radwan, A. Valada, and W. Burgard. VLocNet++: Deep Multitask Learning for Semantic Visual Localization and Odometry. *IEEE Robotics and Automation Letters (RAL)*, 2018.

[237] S. Ramalingam, S. Bouaziz, P. Sturm, and M. Brand. SKYLINE2GPS: Localization in urban canyons using omni-skylines. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3816–3823, 2010. ISBN 9781424466757. doi: 10.1109/IROS.2010.5649105.

[238] S. Ramalingam, S. Bouaziz, and P. Sturm. Pose Estimation Using Both Points and Lines for Geolocation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2011.

[239] M. Rastgoo, C. Demonceaux, R. Seulin, and O. Morel. Attitude Estimation from Polarimetric Cameras. *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 8397–8403, 2018. ISSN 21530866. doi: 10.1109/IROS.2018.8593575.

[240] A. S. Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual Instance Retrieval with Deep Convolutional Networks. *arXiv preprint*, 4(3):251–258, 2014. ISSN 2186-7364. doi: 10.3169/mta.4.251.

[241] K. Regmi and A. Borji. Cross-View Image Synthesis using Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00369.

[242] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[243] D. Robertson and R. Cipolla. An Image-Based System for Urban Navigation. In *British Machine Vision Conference (BMVC)*, 2004. doi: 10.5244/C.18.84.

[244] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood Consensus Networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, number Nips, 2018. ISBN 1810.10510v2. doi: arXiv: 1810.10510v2.

[245] D. M. Rosen, J. Mason, and J. J. Leonard. Towards Lifelong Feature-Based Mapping in Semi-Static Environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2016.

[246] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision (ECCV)*, pages 1–14, 2006.

[247] A. Rubio, M. Villamizar, L. Ferraz, A. Penata-Sanchez, A. Ramisa, E. Simo-Serra, A. Sanfeliu, and F. Moreno-Noguer. Efficient Monocular Pose Estimation for Complex 3D Models. In *IEEE International Conference on Robotics and Automation (ICRA)*, volume 2, pages 1397–1402, 2015. ISBN 9781479969227.

[248] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[249] E. Rupnik, M. Daakir, and M. P. Deseilligny. MicMac, a free, open-source solution for photogrammetry. *Open Geospatial Data, Software and Standards*, 2(1):14, 2017.

[250] B. C. Russell, J. Sivic, J. Ponce, and H. Dessales. Automatic alignment of paintings and photographs depicting a 3D scene. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2011. ISBN 9781467300629. doi: 10.1109/ICCVW.2011.6130291.

[251] S. Saha, G. Varma, and C. V. Jawahar. Improved Visual Relocalization by Discovering Anchor Points. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2018.

[252] I. B. Salah, S. Kramm, C. Demonceaux, and P. Vasseur. Summarizing large scale 3D point cloud for navigation tasks. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2017. ISBN 9781538615256. doi: 10.1109/ITSC.2017.8317657.

[253] I. B. Salah, S. Kramm, C. Demonceaux, and P. Vasseur. Summarizing Large Scale 3D Mesh. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[254] I. B. Salah, S. Kramm, C. Demonceaux, and P. Vasseur. Navigability Graph Extraction from Large-Scale 3D Point Cloud. In *IEEE Conference on Intelligent Transportation Systems (ITSC)*, volume 2018-Novem, pages 3030–3035, 2018. ISBN 9781728103235. doi: 10.1109/ITSC.2018.8569447.

[255] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly, and A. J. Davison. SLAM++: Simultaneous localisation and mapping at the level of objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1352–1359, 2013. ISBN 978-0-7695-4989-7. doi: 10.1109/CVPR.2013.178.

[256] P.-E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, and C. Cadena. Leveraging Deep Visual Descriptors for Hierarchical Efficient Localization. In *Conference on Robot Learning (CoRL)*, pages 1–10, 2018.

[257] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[258] T. Sattler, B. Leibe, and L. Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *IEEE International Conference on Computer Vision (ICCV)*, pages 667–674, 2011. ISBN 9781457711015. doi: 10.1109/ICCV.2011.6126302.

[259] T. Sattler, B. Leibe, and L. Kobbelt. Improving image-based localization by active correspondence search. In *European Conference on Computer Vision (ECCV)*, volume 7572 LNCS, pages 752–765, 2012. ISBN 9783642337178. doi: 10.1007/978-3-642-33718-5_54.

[260] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *British Machine Vision Conference (BMVC)*, pages 76.1–76.12, 2012. ISBN 1-901725-46-4. doi: 10.5244/C.26.76.

[261] T. Sattler, M. Havlena, F. Radenović, K. Schindler, and M. 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.

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

[263] T. Sattler, B. Leibe, and L. 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.

[264] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[265] T. Sattler, W. Maddern, A. Torii, J. Sivic, T. Pajdla, M. Pollefeys, and M. Okutomi. Benchmarking 6DOF Urban Visual Localization in Changing Conditions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[266] T. Sattler, Q. Zhou, M. Pollefeys, and L. 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.

[267] D. Saupe and D. V. Vranić. 3D model retrieval with spherical harmonics and moments. In *Joint Pattern Recognition Symposium*, pages 392–397. Springer, 2001.

[268] O. Saurer, G. Baatz, K. Köser, L. Ladicky, and M. Pollefeys. Image Based Geolocalization in the Alps. *International Journal of Computer Vision (IJCV)*, 116 (3):213–225, 2016. ISSN 15731405. doi: 10.1007/s11263-015-0830-0.

[269] G. Schindler, M. Brown, and R. Szeliski. City-Scale Location Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. ISBN 1424411807.

[270] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[271] J. L. Schönberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number January, 2017.

[272] 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. doi: 10.1109/CVPR.2018.00721.

[273] P. H. Seo, T. Weyand, J. Sim, and B. Han. CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps. In *European Conference on Computer Vision (ECCV)*, 2018.

[274] Z. Seymour, K. Sikka, H.-p. Chiu, S. Samarasekera, and R. Kumar. Semantically-Aware Attentive Neural Embeddings for Image-based Visual Localization. *British Machine Vision Conference (BMVC)*, 2019.

[275] T. Shi, S. Shen, X. Gao, and L. Zhu. Visual Localization Using Sparse Semantic 3D Map. *arXiv preprint*, 2019.

[276] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. 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.

[277] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transactions on Graphics (ToG)*, 30(6):1, 2011. ISSN 07300301. doi: 10.1145/2070781.2024188.

[278] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003. ISBN 0769519504.

[279] E. Sizikova, V. K. Singh, B. Georgescu, M. Halber, K. Ma, and T. Chen. Enhancing Place Recognition using Joint Intensity - Depth Analysis and Synthetic Data. In *European Conference on Computer Vision Workshops (ECCVW)*, pages 1–8, 2016. ISBN 978-3-319-49408-1.

[280] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[281] Y. Song, X. Chen, X. Wang, Y. Zhang, and J. Li. 6-DOF Image Localization From Massive Geo-Tagged Reference Images. *IEEE Transactions on Multimedia (ToM)*, 18(8):1542–1554, 2016.

[282] P. Speciale, J. L. Schönberger, S. B. Kang, S. N. Sinha, and M. Pollefeys. Privacy Preserving Image-Based Localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5493–5503, 2019.

[283] E. Stumm. *Building Location Models for Visual Place Recognition*. PhD thesis, 2015.

[284] E. Stumm, C. Mei, and S. Lacroix. Location graphs for visual place recognition. In *IEEE International Conference on Robotics and Automation (ICRA)*, number May, 2015. doi: 10.1177/0278364915570140.

[285] E. Stumm, C. Mei, S. Lacroix, J. Nieto, M. Hutter, and R. Siegwart. Robust Visual Place Recognition with Graph Kernels. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4535–4544, 2016. doi: 10.1109/CVPR.2016.491.

[286] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. J. Milford. On the performance of ConvNet features for place recognition. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, volume 2015-Decem, pages 4297–4304, 2015. ISBN 9781479999941. doi: 10.1109/IROS.2015.7353986.

[287] 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. ISBN 9780992374716. doi: 10.15607/RSS.2015.XI.022.

[288] L. Svarm, O. Enqvist, F. Kahl, and M. Oskarsson. Accurate Localization and Pose Estimation for Large 3D Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. doi: 10.1109/CVPR.2014.75.

[289] L. Svarm, O. Enqvist, F. Kahl, and M. Oskarsson. City-Scale Localization for Cameras with Known Vertical Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 8828(c):1–1, 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2016.2598331.

[290] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. 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.

[291] H. Taira, I. Rocco, J. Sedlar, M. Okutomi, J. Sivic, T. Pajdla, T. Sattler, and A. Torii. Is This The Right Place? Geometric-Semantic Pose Verification for Indoor Visual Localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[292] K. Tateno, F. Tombari, I. Laina, and N. Navab. CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[293] Y. Tian, C. Chen, and M. Shah. Cross-View Image Matching for Geo-localization in Urban Environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[294] C. Toft, C. Olsson, and F. Kahl. Long-term 3D Localization and Pose from Semantic Labellings. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 650–659, 2017.

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

[296] G. Tolias and Y. Avrithis. Speeded-up, Relaxed Spatial Matching. In *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, 2011.

[297] G. Tolias and H. Jégou. Visual query expansion with or without geometry: Refining local descriptors by feature aggregation. *Pattern Recognition*, 47(10):3466–3476, 2014. ISSN 00313203. doi: 10.1016/j.patcog.2014.04.007.

[298] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *International Conference on Learning Representations (ICLR)*, pages 1–11, 2016.

[299] A. Torii, J. Sivic, and T. Pajdla. Visual localization by linear combination of image descriptors. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2011. ISBN 9781467300629. doi: 10.1109/ICCVW.2011.6130230.

[300] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla. Visual Place Recognition with Repetitive Structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 37, pages 2346–2359, 2013. ISBN 978-0-7695-4989-7. doi: 10.1109/TPAMI.2015.2409868.

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

[302] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *IEEE International Conference on Computer Vision (ICCV)*, volume 3, pages 273–280, 2003. ISBN 0-7695-1950-4. doi: 10.1109/ICCV.2003.1238354.

[303] E. Tzeng, A. Zhai, M. Clements, R. Townshend, and A. Zakhor. User-driven geolocation of untagged desert imagery using digital elevation models. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 237–244, 2013. ISBN 9780769549903. doi: 10.1109/CVPRW.2013.42.

[304] M. A. Uy and G. H. Lee. PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00470.

[305] A. Valada, N. Radwan, and W. Burgard. Incorporating Semantic and Geometric Priors in Deep Pose Regression. In *Robotics Science and Systems Workshop (RSSW)*, pages 1–4, 2018.

[306] A. Valada, N. Radwan, and W. Burgard. Deep Auxiliary Learning for Visual Localization and Odometry. In *IEEE International Conference on Robotics and*

Automation (ICRA), pages 6939–6946, 2018. ISBN 9781538630815. doi: 10.1109/ICRA.2018.8462979.

[307] J. Valentin, A. Fitzgibbon, M. Nießner, J. Shotton, and P. H. S. Torr. Exploiting Uncertainty in Regression Forests for Accurate Camera Relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4400–4408, 2015. ISBN 9781467369640.

[308] J. Valentin, A. Dai, M. Niessner, P. Kohli, P. H. S. Torr, S. Izadi, and C. 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.

[309] C. Valgren and A. J. Lilienthal. SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments. *Robotics and Autonomous Systems (RAS)*, 58(2):149–156, 2010. ISSN 09218890. doi: 10.1016/j.robot.2009.09.010.

[310] VC and P. Hough. Method and means for recognizing complex patterns, 1962.

[311] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio. ReNet: A Recurrent Neural Network Based Alternative to Convolutional Networks. *arXiv preprint*, pages 1–9, 2015.

[312] N. N. Vo and J. Hays. Localizing and Orienting Street Views Using Overhead Imagery. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 494–509, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0.

[313] N. N. Vo, N. Jacobs, and J. Hays. Revisiting IM2GPS in the Deep Learning Era. In *IEEE International Conference on Computer Vision (ICCV)*, number 1, 2017.

[314] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based Localization with Spatial LSTMs. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[315] W. Wan, Z. Liu, K. Di, B. Wang, and J. Zhou. A Cross-Site Visual Localization Method for Yutu Rover. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-4(May):279–284, 2014. ISSN 2194-9034. doi: 10.5194/isprsarchives-XL-4-279-2014.

[316] X. Wan, J. Liu, H. Yan, and G. L. K. Morgan. Illumination-invariant image matching for autonomous UAV localisation based on optical sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119:198–213, 2016. ISSN 09242716. doi: 10.1016/j.isprsjprs.2016.05.016.

[317] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen. A Survey on Learning to Hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13(9), 2017. ISSN 0162-8828. doi: 10.1109/TPAMI.2017.2699960.

[318] S. Wang, S. Fidler, and R. Urtasun. Lost shopping! Monocular localization in large indoor spaces. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 2695–2703, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.309.

[319] 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 *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[320] Z. Wang, B. Fan, and F. Wu. FRIF: Fast Robust Invariant Feature. In *British Machine Vision Conference (BMVC)*, 2013.

[321] P. Weinzaepfel, G. Csurka, Y. Cabon, and M. Humenberger. Visual Localization by Learning Objects-of-Interest Dense Match Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[322] T. Weyand, I. Kostrikov, and J. Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *European Conference on Computer Vision (ECCV)*, volume 9905, pages 37–55, 2016. ISBN 978-3-319-46447-3. doi: 10.1007/978-3-319-46448-0.

[323] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, and A. J. Davison. Elasticfusion: dense SLAM without a pose graph. In *Robotics and Autonomous Systems (RAS)*, 2015.

[324] A. R. Widya, A. Torii, and M. Okutomi. Structure from motion using dense CNN features with keypoint relocalization. *IPSJ Transactions on Computer Vision and Applications*, pages 0–6, 2018.

[325] S. Workman, R. Souvenir, and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, volume 11-18-Dece, pages 3961–3969, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.451.

[326] J. Wu, H. I. Christensen, and J. M. Rehg. Visual place categorization: Problem, dataset, and algorithm. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 4763–4770, 2009. ISBN 9781424438044. doi: 10.1109/IROS.2009.5354164.

[327] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe. Detection, Learning cross-modal deep representations for robust pedestrian. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5363–5371, 2017.

[328] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian. CNN vs. SIFT for Image Retrieval. In *International Conference on Multimedia (MM)*, pages 407–411, 2016. ISBN 9781450336031. doi: 10.1145/2964284.2967252.

[329] Z. J. Yew and G. H. Lee. 3DFeat-Net: Weakly Supervised Local 3D Features for Point Cloud Registration. *European Conference on Computer Vision (ECCV)*, 2018. ISSN 1083-3668. doi: 10.1117/1.3066900.

[330] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. In *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.

[331] A. R. Zamir and M. Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision (ECCV)*, volume 6314 LNCS, pages 255–268, 2010. ISBN 364215560X. doi: 10.1007/978-3-642-15561-1_19.

[332] 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)*, 36(8):1546–1558, 2014. ISSN 01628828. doi: 10.1109/TPAMI.2014.2299799.

[333] A. R. Zamir, A. Hakeem, L. Van Gool, M. Shah, and R. Szeliski. Large-scale visual geo-localization. *Advances in computer vision and pattern recognition*, 2016.

[334] A. R. Zamir, A. Sax, W. Shen, L. Guibas, J. Malik, and S. Savarese. Taskonomy: Disentangling Task Transfer Learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722, 2018. doi: 10.1109/CVPR.2018.00391.

[335] B. Zeisl, T. Sattler, and M. Pollefeys. Camera Pose Voting for Large-Scale Image-Based Localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2704–2712, 2015.

[336] A. Zeng, S. Song, M. Nießner, M. Fisher, and J. Xiao. 3DMatch: Learning the Matching of Local 3D Geometry in Range Scans. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 13, 2017.

[337] L. Zheng, Y. Yang, and Q. Tian. SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14(8), 2017.

[338] T. Zhi, L.-Y. Duan, Y. Wang, and T. Huang. Two-Stage Pooling of Deep Convolutional Features for Image Retrieval. In *IEEE International Conference on Image Processing (ICIP)*, 2016. doi: 10.1109/ICIP.2016.7532802.

[339] Q. Zhou, T. Sattler, M. Pollefeys, and L. Leal-Taixe. To Learn or Not to Learn: Visual Localization from Essential Matrices. *arXiv preprint*, 2019.

[340] T. Zhou, M. Brown, N. Snavely, and D. 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.

[341] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras. OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas. pages 1–18, 2018.

[342] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, pages 391–405. Springer, 2014.

SPIM

Titre : Localisation basée vision à partir de caractéristiques discriminantes issues de données visuelles hétérogènes

Mots clés : localisation, estimation de pose, indexation d'images

Résumé : La localisation basée vision consiste à déterminer l'emplacement d'une requête visuelle par rapport à un espace de référence connu. Le principal défi de la localisation visuelle réside dans le fait que la requête peut avoir été acquise à un moment différent de celui de la base de données. On pourra alors observer des changements visuels entre l'environnement actuel et celui de la base de référence, en particulier lors d'application de localisation en extérieur. Les approches récentes utilisent des informations complémentaires afin de répondre à ces scenarios de localisation visuellement ambigu, comme la géométrie ou la sémantique. Cependant, ces modalités auxiliaires ne sont pas toujours disponibles ou peuvent être coûteuse à obtenir. Afin de s'affranchir de l'utilisation modalité supplémentaire pour faire face à ces scenarios de localisation difficiles, nous proposons d'utiliser un modèle de transfert de modalité capable de reproduire la géométrie d'une scène à partir d'une image monoculaire.

Dans un premier temps, nous présentons le problème de localisation comme un problème d'indexation d'images et nous entrainons un réseau de neurones convolutif pour la description globale d'image en introduisant le transfert de modalité radiométrie vers géométrie comme objectif secondaire. Une fois entrainé, notre modèle peut être appliqué sur des images monoculaires pour construire un descripteur efficace pour la localisation en conditions difficiles. Dans un second temps, nous introduisons une nouvelle méthode de raffinement de pose pour améliorer la localisation donnée par notre première étape. De la même manière que notre descripteur d'image globale, la relocalisation est facilitée par les informations géométriques apprises lors d'une étape préalable. L'information géométrique supplémentaire est utilisée pour contraindre l'estimation finale de la pose de la requête. Grâce des expériences approfondies, nous démontrons l'efficacité de nos propositions pour la localisation en intérieur et en extérieur.

Title : Vision-based localization with discriminative features from heterogeneous visual data

Keywords : localization, pose estimation, image indexing.

Abstract : Visual-based Localization (VBL) consists in retrieving the location of a visual image within a known space. VBL is involved in several present-day practical applications, such as indoor and outdoor navigation, 3D reconstruction, etc. The main challenge in VBL comes from the fact that the visual input to localize could have been taken at a different time than the reference database. Visual changes may occur on the observed environment during this period of time, especially for outdoor localization. Recent approaches use complementary information in order to address these visually challenging localization scenarios, like geometric information or semantic information. However geometric
or semantic information are not always available or can be costly to obtain. In order to get free of any extra modalities used to solve challenging localization scenarios, we propose to use a modality transfer model capable of reproducing
the underlying scene geometry from a monocular image. At first, we cast the localization problem as a Content-based Image Retrieval (CBIR) problem and we train a CNN image descriptor with radiometry to dense geometry transfer as side training objective. Once trained, our system can be used on monocular images only to construct an expressive descriptor for localization in challenging conditions. Secondly, we introduce a new relocalization pipeline to improve the localization given by our initial localization step. In a same manner as our global image descriptor, the relocalization is aided by the geometric information learned during an offline stage. The extra geometric information is used to constrain the final pose estimation of the query. Through comprehensive experiments, we demonstrate the effectiveness of our proposals for both indoor and outdoor localization.

UBFC

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon