Department of Computer Science

Deep Learning Methods for Semantic Matching, Image Retrieval and Camera Relocalization

Zakaria Laskar

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 191/2020

Deep Learning Methods for Semantic Matching, Image Retrieval and Camera Relocalization

Zakaria Laskar

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held via remote technology on 9 December 2020 at 13.

Aalto University
School of Science
Department of Computer Science

Supervising professor
Prof. Juho Kannala, Aalto University, Finland

Preliminary examiners
Prof. Giorgos Tolias, Czech Technical University, Czech Republic
Dr. Relja Arandjelović, Deepmind, United Kingdom

Opponent
Prof. Giorgos Tolias, Czech Technical University, Czech Republic

Aalto University publication series
DOCTORAL DISSERTATIONS 191/2020

ISBN 978-952-64-0145-4 (printed)
ISBN 978-952-64-0146-1 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0146-1

Unigrafia Oy
Helsinki 2020

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Zakaria Laskar

Name of the doctoral dissertation
Deep Learning Methods for Semantic Matching, Image Retrieval and Camera Relocalization

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 191/2020

Field of research Computer Science

Manuscript submitted 20 August 2020 **Date of the defence** 9 December 2020

Permission for public defence granted (date) 9 November 2020 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Image matching is a central component in many computer vision applications. The field has progressed significantly with the advancement of deep learning models such as convolutional neural networks. The thesis makes several contributions in advancing the performance of existing CNN based approaches in closely related problem areas of image matching, namely semantic matching, image retrieval and image based localization.

In this thesis, the problem of data and ground-truth labelling efficiency for training CNN models is studied in the context of semantic matching. A weakly supervised method is presented to address the problem of learning using small training datasets. The method first generates additional training samples using existing data and proposes a novel loss function based on cyclic consistency to regularize the training process. Results show that the proposed method can learn from weakly labelled data without pixel level correspondence information.

In the next part of the thesis, we study the application of both global and local image matching to the problem of image retrieval. In the problem of particular landmark retrieval the thesis studies the role of contextual information in global query image representation which is generally ignored by existing approaches to remove noisy background information. An attention model is proposed that uses bottom-up saliency to modulate contextual information in intermediary CNN representations in a top-down manner. On the other hand, to address the challenges due to local variations in city-scale retrieval, the thesis proposes a geometric verification method using CNN based image matching. In addition, it proposes method for improving the accuracy and efficiency of the image matching method.

Lastly, the thesis demonstrates methods utilizing the key concepts from image matching and image retrieval to address problems in the field of image based localization. In contrast to existing approaches the proposed method can be applied to novel scenes not seen during training and scales favourably with the size of the environment. In addition, a challenging indoor localization dataset is made publicly available to address limitation of existing datasets.

Keywords computer vision, machine learning, deep learning, camera relocalization, image retrieval, image matching

ISBN (printed) 978-952-64-0145-4 **ISBN (pdf)** 978-952-64-0146-1

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2020

Pages 132 **urn** http://urn.fi/URN:ISBN:978-952-64-0146-1

To my Namma,
Anowara Begum

Preface

Dissertation is a symbolic conclusion of the doctoral study but it feels like the first step of a long and exciting journey into the field of research. The beginning of this journey has been wonderful for which I would like to acknowledge several people who have directly or indirectly helped in the fruitful completion of the thesis.

Firstly, I would like to acknowledge my supervisor Prof. Juho Kannala for introducing me to the critical aspects of research. His patience and humility is unshakeable and is the role model of an advisor I want to be. Thank you for putting up with my mistakes and giving freedom to pursue independent research. I would also like to show my gratitude to Dr. Daniel Herrera, Prof. Esa Rahtu, Prof. Fazal Talukdar, Prof Janne Heikkilä, Dr. Sami Huttunen for assisting during the early stages of pursuing research that eventually transformed into a stepping stone for my doctoral thesis.

I am grateful to the reviewers of the thesis, Prof. Giorgos Tolias from Czech Technical University, Prague, and Dr. Relja Arandjelović from Deepmind, London for their valuable comments and constructive feedback.

The thesis would not have been possible without the ever-generous colleagues I had the privilege to work with. Special thanks to Dr. Iaroslav Melekhov whose discipline and work ethics is an inspiration. Further, I would like to acknowledge Dr. Hamed Tavakoli for generously filling in the role of the second supervisor despite his pressing commitments. I would also like to acknowledge both my current and former colleagues for wonderful conversations and exciting extra-curricular activities: Prof. Arno Solin, Dr. Juha Ylioinas, Dr. Markus Ylimäki, Rinu Boney, Santiago Cortes, Shuzhe Wang, Surya Kalia, Xiaotian Li, Yuxin Hou, Yi Zhao, Dr. Zinelabinde Boulkenafet.

The darkness of the Finnish winter was illuminated by the presence of wonderful people who lent their selfless support throughout the years. The brightest stars among them were Kunal, Faisal, and Vinod with whom I not only shared the happiness of friendship but also the struggle in navigating various aspects of life in a faraway land. Despite the long distance, the

following people never felt short of providing motivation and much needed guidance: Abrar, Akash, Aseem, Ashwini, Disha, Furqan, Jahidul, Karthik, Kushal, Madhurjya, Raheef, Rashad, Zeeshan. Thank you and I hope to reciprocate with the best of my abilities. I would also like to thank my lawn tennis buddies Quoc, Thanh, Kevin, Rohan, and Aalto Tennis for providing an escape from hectic academic life.

The tree is only as strong as its roots, and I am fortunate to be rooted in a family that greatly values education. A humble acknowledgment to my parents, Ayesha Begum and Misbahur Rahman for staying strong in the face of academic and economic uncertainty as I embarked on a journey to Finland. Finally, my heartfelt gratitude to Abdul Waris, Aktar, Asraf, Fazal, Faroque, Hifzur, Raihena, Jasmine, Miftahur, Mehebubur, Rousonara, Sahera, Saleha, Sharmin, Dr. Shahnowaj, Shahida, Shamsuddin, Sufia, for showing how to strive towards better purpose despite limited means.

Helsinki, November 6, 2020,

Zakaria Laskar

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Zakaria Laskar, and Juho Kannala. Semi-supervised Semantic Matching. *European Conference on Computer Vision. Geometry Meets Deep Learning Workshop (ECCVW)*, pp. 444–455, 2018.

II Zakaria Laskar, Hamed Rezazadegan Tavakoli, and Juho Kannala. Semantic Matching by Weakly Supervised 2D Point Set Registration. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 1061–1069, December 2019.

III Zakaria Laskar, and Juho Kannala. Context Aware Query Image Representation for Particular Object Retrieval. *Scandinivian Conference on Image Analysis (SCIA)*, pp. 88–99, January 2017.

IV Zakaria Laskar, Iaroslav Melekhov, Hamed Rezazadegan Tavakoli, Juha Ylioinas, and Juho Kannala. Geometric Image Correspondence Verification by Dense Pixel Matching. *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2510–2519, 2020.

V Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Networks. *International Conference on Computer Vision. Geometry Meets Deep Learning Workshop (ICCVW)*, pp. 929–938, 2017.

Author's Contribution

Publication I: "Semi-supervised Semantic Matching"

Laskar proposed the topic and also designed and conducted the experiments. He had the main responsibility in writing the article while Kannala reviewed and proposed suggestions to the manuscript.

Publication II: "Semantic Matching by Weakly Supervised 2D Point Set Registration"

Laskar had the main responsibility in designing and conducting the experiments. Laskar and Tavakoli reviewed the results and wrote the article. Kannala proposed suggestions to improve the final manuscript.

Publication III: "Context Aware Query Image Representation for Particular Object Retrieval"

Laskar had the main responsibility in writing the article. He also proposed the topic and designed the experiments. Kannala reviewed the methods and results and also provided valuable feedback to the manuscript.

Publication IV: "Geometric Image Correspondence Verification by Dense Pixel Matching"

Laskar proposed the methods and designed the experiments. Laskar and Melekhov conducted the experiments and wrote the manuscript. Kannala also contributed in writing the article. The co-authors reviewed and proposed suggestions to the manuscript.

Publication V: "Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Networks"

Laskar proposed the method and designed the experiments. Laskar, Melekhov and Surya conducted the experiments and wrote the article. Kannala reviewed the method and also contributed to the final manuscript.

Abbreviations

AEPE Average End Point Error

AR Augmented Reality

CD Chamfer Distance

CNNs Convolutional Neural Networks

DGC-Net Dense Geometric Correspondence Network

DoF Degree of Freedom

DoG Difference of Gaussian

ICP Iterative Closest Point

LiDAR Light Detection and Ranging

mAP Mean Average Precision

NN Nearest Neighbours

PCA Principal Component Analysis

PCK Percentage of Correctly transferred Keypoints

P*n*P Perspective-n-Point

RANSAC Random Sample Consensus

R-MAC Regional Maximum Activation of Convolutions

RNNs Recurrent Neural Networks

ROI Region of Interest

RPN Region Proposal Network

SfM Structure from Motion

SIFT Scale Invariant Feature Transform

SLAM Simultaneous Localization and Mapping

SVM Support Vector Machines

TPS Thin plate Spline

VR Virtual Reality

1. Introduction

The later half of the 20^{th} century has seen rapid technological progress in hardware devices that can capture and process digital images. The devices have become particularly ubiquitous in recent times ranging from handheld devices such as mobile phones, tablets, medical imaging devices and surveillance cameras among many others. With development in the field of artificial intelligence, researchers aimed to mimic the human visual system using images [26]. This led to the emergence of computer vision as a discipline for processing and extracting meaningful information from raw pixels in an image.

Wide range of application areas for computer vision emerged ranging from industrial to medical, military and entertainment [127, 37, 13]. In industries manual processes such as fault detection could now be automated using computer vision [89]. Similarly, in healthcare, medical systems are now assisted by vision algorithms [63, 38]. Large amount of online images have led to the development of computer vision algorithms to process these images in an automated manner. Systems dependant on human computer interaction such as electronic toll collector can be done using plate recognition systems based on computer vision. The field of sports has greatly benefited using automated tracking systems that track the players and the ball (e.g. in football, tennis) to generate detailed post match summary and also in-match assistance such as Hawk-eye to assist the human decision process.

With the advent of smartphones and smart devices, location based services have opened a new frontier for users. To facilitate such services, the users need to be accurately localized w.r.t a given environment. Options such as GPS are not accurate indoors and also do not provide 6-DoF localization which has many applications. Some of them being Augmented (AR) and Virtual Reality (VR) applications that are at the forefront of current consumer entertainment business. At the core of these AR, and VR technologies are algorithms solving computer vision problems. These technologies allow users in market places or museums to actively interact with their environment. Augmented reality based mobile games such as

Pokemon Go by *Niantic Inc* have been tremendously popular. Apart from games, AR and VR technologies have also found applications in therapeutic treatments [140] of stress and anxiety disorders as well as Alzheimer's. It provides the affected patients the ability to engage in simulated reality which they would otherwise not experience in their handicapped state. In similar lines, these technologies also provide the opportunity to educate engineers using virtual reality [2] in situations where monetary and safety costs are high.

In addition to entertainment, computer vision is at the forefront in powering future autonomous agents to reduce human labour and interference in various industries such as automotive [39], phone assembly lines [89], agriculture [144] etc. Earlier, robots were limited to assembly lines, operating in fixed environments performing pre-defined set of actions. With the advancement of computer vision algorithms, different avenues of deploying robots have opened up. Self-driving cars, item delivery robots in hotels, or indoor offices represent some of the potential applications. Perception sensors are the key components that assist such robots to navigate the environment around them. While LiDAR provides accurate geometric information about the environment in the form of depth data, the cost of procurement and deployment restricts its usage in different settings. On the other hand cameras are low cost alternatives and a richer semantic information of the environment can be extracted from images such as distinguishing pedestrians on the the side-walks from the cars on the road [49]. Soon robots are expected to be a household assistant doing a variety of tasks by perceiving and interacting with the environment based on language, vision and sound sensors.

To make the above applications possible, years and years of research have been invested in the field of computer vision. Recognizing objects in an image and its location is a fundamental step towards mimicking human vision. This is addressed by several areas of research in computer vision such as image recognition [61], object detection [34, 98], semantic segmentation [9] among many others. David Marr described modelling 3D world from 2D images as fundamental blocks towards emulating the human brain [79]. This formed the paradigm of early computer vision approaches for scene understanding such as 3D scene reconstruction [48], depth prediction [111] that aim to directly model the 3D world from 2D images. On the other hand, optical flow [124] or camera/object tracking [24] aim to provide important cues about the 3D world without explicitly modelling the 3D world. For example, optical flow or object tracking can provide information to a car about the movement dynamics of other objects on the road e.g. pedestrians or other cars without directly computing the 3D model of those objects. Instance retrieval and image-based localization are also closely coupled problems associated with indirectly modelling the 3D environment [5, 130, 92]. The objective is to effectively and efficiently

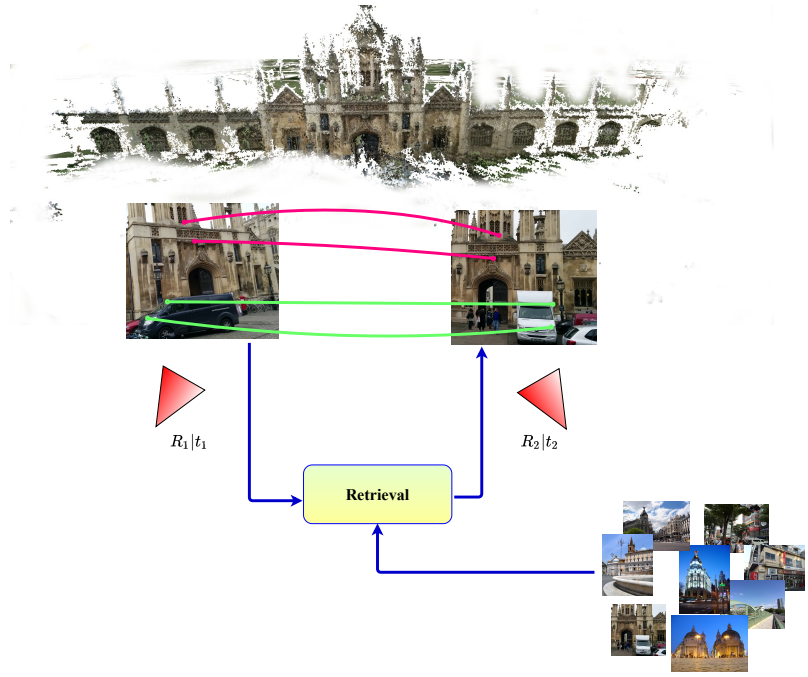

Figure 1.1. Some topics covered in the thesis: i) retrieving relevant database images w.r.t query image, ii) local descriptor matching (shown in red) between images observing the same 3D point, iii) semantic matching (shown in green) between different instances of the same object category such as truck, and iii) localizing an image w.r.t a 3D environment [57] by camera pose estimation.

compute representations for the 3D world. Then, as a robot or a user navigates the environment, it can localize itself w.r.t the environment by simply querying with a 2D image of its current location.

1.1 Scope of the thesis

One of the main challenges with computer vision problems has been finding methods for suitable and robust representation of images [87]. The seminal work of SIFT [78] laid the foundation for rapid progress in the field of computer vision. These coupled with Bag-of-Visual-Words [122] and Support Vector Machines [25] allowed learning complex decision boundaries for solving various computer vision problems. However, the accuracy lacked behind human performance on many frontiers such as image recognition [121]. As computer architecture got better, deep learning based solutions from 1970s [69] became a viable option as they allow to learn both the representation and the decision boundaries in an end-to-end manner. The key breakthrough can be attributed to the creation of ImageNet [27] dataset and the work by Krizhevsky [121]. The proposed CNN architectures were able to match human performance on the challeng-

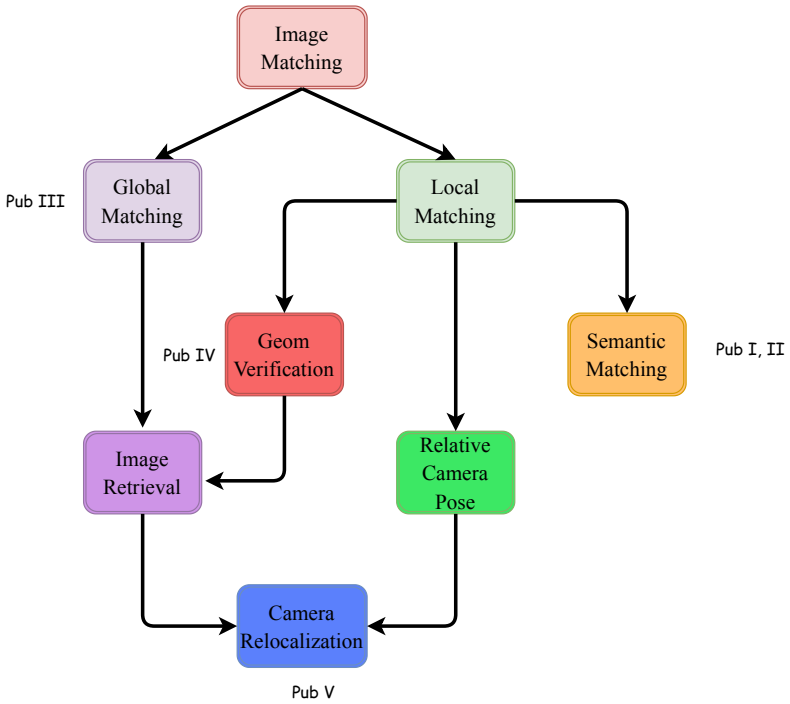

Figure 1.2. Relations of the publications. Publication I,II and IV address local descriptor matching: I and II introduce weakly supervised learning, IV improves standard geometric verification. Publication III, IV address the problem of image retrieval in the fast global and slow but accurate local matching settings. Publication V improves camera relocalization using image retrieval and local matching based relative camera pose estimation.

ing image recognition task. Since then various architectures [50, 126], datasets [40], optimization algorithms [59] have been proposed which allowed machines to surpass on many complex tasks such as playing complex Atari games [117]. Other fields such as object detection [98], semantic segmentation [51], speech recognition [3], and natural language processing [29] were quick to extend the success of CNN models to their respective domains.

In this thesis, the aim is to study several closely related problems in the area of computer vision, namely semantic matching, image retrieval and image-based localization as shown in Figure. 1.1. The problem overlaps with other domains such as optical flow [123, 35], image-based 3D reconstruction [104, 32], autonomous navigation [136], and scene understanding [52]. The applications range from self-driving cars, AR/VR based immersive technologies, to image and product search. In healthcare retrieving relevant imaging results of previous patients can accelerate prognosis. Although the topics studied in the thesis may seem diverse, but are closely connected as shown in Figure. 1.2. The central theme of the thesis is to learn robust image representations using CNN models in the context of

image matching and its applications to the above mentioned problems. In addition, the topics pertaining to CNN models involve learning from weakly supervised data, improving generalization and scalability, producing new datasets for the scientific community, and extending existing vision systems based on hand-crafted descriptors to CNN based representations.

Typically, CNN models require large amount of training data [27]. In addition, ground-truth labels are required to supervise the training process. To address the challenges of obtaining large amount of labelled data, semi-supervised or weakly supervised methods are proposed [12, 138, 96]. The thesis addresses similar problems in the context of semantic matching. Methods for generating more training data and deriving constraints to drive the learning process are studied. The central part of the thesis is the problem of image retrieval. Given a query image, the objective is to retrieve similar images from a large database. The role of contextual information outside the query region of interest is studied. Furthermore, the problem of geometric verification is revisited in the light of CNN representations.

An important part of the thesis is the approach to image-based localization. The work is built on CNN based image localization approaches [57, 56] which have shown that CNN models can be used to directly regress camera pose parameters from input RGB image. In this thesis the generalization and scalability of existing CNN based methods is studied.

1.2 Contributions

More concretely, the main contributions of the thesis are listed below:

- A process for generating weakly supervised data from small labelled datasets is presented for the problem of semantic matching.

- A cost function based on cyclic consistency is proposed to leverage the weakly supervised data for training end-to-end semantic matching CNN models.

- A top-down method for applying spatial attention based on a pre-defined attention map and bottom-up saliency measure is presented. The attention model is applied to incorporate contextual information for solving the particular object retrieval problem. In addition a weighted version of global aggregation of local CNN representations is proposed.

- A geometric verification system is proposed based on CNN representations to improve global image retrieval methods.

- Methods for improving the accuracy and efficiency of CNN based image

matching is presented.

- Methods for improving scalability, and generalization of existing end-to-end camera localization CNN models is proposed.

- A new indoor localization dataset is open-sourced to benchmark indoor localization methods.

1.3 Outline of the thesis

This thesis consists of an overview and an appendix, which includes the original articles. The rest of the overview has the following structure. Chapter 2 provides a review of the problem of semantic matching in the context of instance matching and introduces methods for training semantic matching CNNs with weakly supervised data. Chapter 3 concentrates on the problems of image retrieval and presents approaches for improving retrieval performance based on global representations and local spatial re-ranking strategies. Chapter 4 studies different CNN-based methods for the problem of image-based localization and their limitations. A system based on image retrieval and relative camera pose estimation using CNN representations is introduced to address the limitations such as scalability and generalization. A summary of the publications is provided in Chapter 5, and some concluding remarks and possible avenues for future research are presented in Chapter 6.

2. Semantic Matching

This Chapter focuses on the problem of finding pixel level correspondences between semantically related objects such as animals, cars etc in any given image pair. The task is challenging due to variations in shape, illumination, geometry, occlusion and scale. The problem can be formulated as finding robust representations of each (-sub)pixel and models for obtaining accurate matches between these representations. First an overview is provided of the standard techniques used to solve the problem of semantic matching, which is followed by a brief discussion of the challenges and motivations leading to the contributions of the thesis.

2.1 Background

In addition to semantic matching, finding correspondences is a key component in a wide variety of computer vision applications such as structure-from motion (SfM), depth estimation, image retrieval, simultaneous localization and mapping (SLAM), object and scene tracking, and optical flow [35]. Broadly, image matching can be categorized into works based on hand-crafted pixel representations and those based on learned representations such as CNNs. Furthermore, the matching function can also be hand-crafted such as methods based on (approximate) Nearest Neighbor (NN) similarity or learnt in a data-driven manner [35, 82, 100, 101] predicting 6 degree-of-freedom relative transformations or 2 dimensional optical flow. The next section goes into more details related to methods of instance matching as they also lay the foundation for the subsequent Chapters in this thesis. Later in the section its application and challenges in the domain of semantic matching is discussed.

Hand-crafted descriptors The seminal work by Lowe [78] introduced the local descriptor SIFT which laid the foundation for much of the developments in computer vision in general and image matching in particular. The algorithm outputs keypoint locations and corresponding descriptors for reliable matching. The descriptor is computed using histograms of

aggregated image gradients. In order to increase repeatibility and robustness of descriptors across similar images under various changes such as viewpoint due to geometric transformations, illumination, scale, several improvements were later proposed [6, 11, 31, 83, 129].

With the availability of large datasets, data-driven approaches [70, 120, 15] demonstrated significant increase in image matching performance compared to earlier hand-crafted models. These models learn a mapping from local image patch centered around each pixel to a new discriminative subspace. Random forest and SVM are some examples of such models. As machines got faster and datasets got bigger, CNN based deep parametric models started to dominate the scene [54, 90, 46, 119, 149]. The parameters of the CNN are optimized to learn a descriptor space, where descriptors of similar (positives) patches are closer than dissimilar (negative) ones. However, the process of selecting these positives and negatives was found to be important in learning robust representations [134, 119]. In particular, it was observed that random sampling lead to the selection of easy positives and negatives. The network stops learning due to the vanishing gradient problem resulting from low error signals. This was addressed in later works [85, 128] where hard positives and negatives are mined using the current state of the CNN parameters. This leads to increased robustness and accuracy in a variety of image matching problems. One of the limitations of patch based descriptor learning is the increased training and evaluation cost as the network needs to be evaluated separately for each patch per image. Recently, several approaches [32, 114, 148] address this issue by jointly optimizing descriptors from the whole image in an end-to-end manner.

2.2 Semantic matching

Semantic matching is a problem category that falls under the broader set of image matching problems mentioned above. As such it shares similar frameworks and models such as general image matching problems. For example, earliest works such as SIFTFlow [76, 146] are based on hand-crafted local descriptors such as SIFT and DAISY [129]. The key difference to standard image matching is a hierarchical optimization of matching cost to obtain a dense pixel correspondence output.

Concurrent to the success in general image matching problems, CNNs descriptors were also used in the semantic matching task [44]. However, the descriptors used for this task were extracted using off-the-shelf CNN network parameters trained for classification problems on the ImageNet [27] dataset. It was observed [44] that descriptors from such pre-trained CNNs did not generalize well to the domain of semantic matching. An alternative is to train the network directly on the task of semantic matching. How-

ever, training deep neural network models requires large training datasets with labelled ground-truth data. For the task of semantic matching one of the following forms of pixel correspondence supervision is required : *i)* geometric transformations like affine, thin-plate spline, homography or relative camera position with depth, and, *ii)* flow fields that contain pixel level correspondence. Popular datasets like Proposal-Flow [44] consists of about only 1400 image pairs. Each image is labelled with sparse keypoints and each image pair with corresponding key-point correspondences. Instance level image matching problems generate large datasets for training neural networks using SfM based techniques to generate 3D models of popular landmarks from images available on the internet [143]. In contrast, generating datasets with ground-truth transformations for semantic matching is a challenging task. Some of the challenges are non-rigid transformation, intra-class variations such as different breeds of dogs or cats, which cannot be modelled by a single 3D model of any object. Furthermore, training using ground-truths data on small datasets brings the risk of overfitting network parameters on the training set. Nevertheless, existing methods [45, 58, 100, 101] avoid directly using the sparse keypoint correspondence information to fine-tune CNN parameters using region level correspondence supervision or self-supervision.

Publications I and II propose methods to learn a dense pixel correspondence function parameterized by a CNN such as [100, 101] using weakly labelled data. The weak supervision is in the form of image level correspondence which in the context of semantic matching does not have a quadratic cost due to pairwise labelling. That is, given the object categories for each image in some dataset, pairwise combination of images in the same category results in potentially quadratic number of image pairs. We show how to generate such dataset from existing small scale datasets such as Proposal-Flow [44] and propose constraints to learn from such image pairs with limited or without explicit pixel-to-pixel correspondence. Data augmentation methods such as synthetically transforming images can generate large number of image pairs but fail to capture wider distribution of transformations observed in the real life. However, data augmentation methods can be considered complimentary to our proposed weakly-supervised methods.

2.2.1 Weakly Supervised Learning

In Publication I the semi-supervised setting is considered where small number of image pairs with sparse keypoint correspondence information are available along with large amount of unlabelled image pairs. For labelled image pairs we used the standard objective function whereby pixels in correspondence should have small Euclidean distance between their respective descriptors. If the correspondence function itself is parameter-

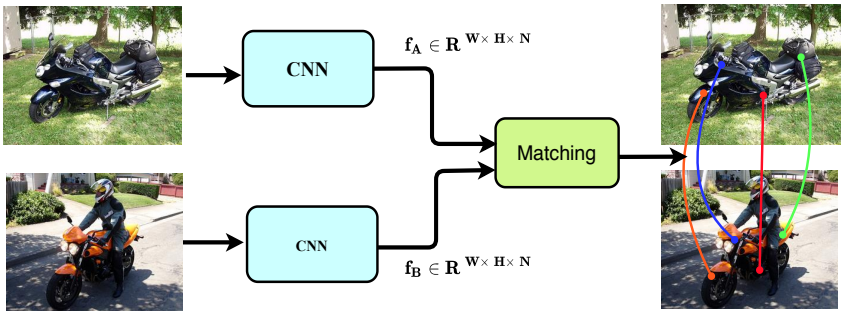

Figure 2.1. An overview of a general CNN architecture for correspondence estimation. Note that the details vary according to target tasks and design options. In general the network consists of a Siamese CNN encoder branch with shared parameters that outputs coarse representations for corresponding input image. The representations from each image are then passed through the matching layer consisting of a global [100] or local [123] correlation layer. The output of the matching layer is used to establish pixel correspondences between input images.

.

ized, for e.g. a CNN that directly estimates optical flow or transformation parameters such as affine or thin-plate spline (TPS), then the supervised loss is the standard pixel re-projection error between pixels, p in correspondence under the estimated correspondence mapping. A general design overview of such a correspondence estimation CNN network is presented in Figure. 2.1. In practise, the CNN network [100] is used in the thesis that predicts affine and TPS parameters, $\hat{\theta}_{AB}$ in an iterative and end-to-end manner from a given input image pair I_A, I_B. The estimated parameters can then be used to compute the dense correspondence field, $\mathbf{T}$ to compute the supervised loss function, L_s.

$$L_s = \frac{1}{M} \sum_{i=1}^{M} ||\mathbf{T}_{\hat{\theta}_{AB}}(p_a) - p_b||_2 \tag{2.1}$$

For unlabelled image pairs, the key observation made is that the descriptors of corresponding pixels should be mutual nearest neighbors in the descriptor space under Euclidean distance metric. Formally, for the considered CNN model, the mutual neighborship can be realized by using cyclic consistency constraint on the forward and backward transformation parameters. That is, the input image pair is passed twice through the CNN by reversing the input image order resulting in forward and backward estimates of the transformation. Under the cyclic consistency constraint the projection of source grid points, $g \in G$ on the target image are constrained to re-project back to the original source points under the backward transformation. In other words, the combined mapping should be close to identity. That is

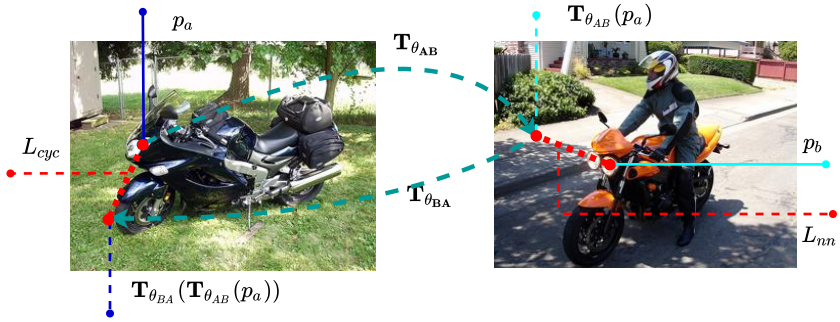

Figure 2.2. A pictorial illustration of the proposed weakly supervised training objective. Dotted lines represent estimations while solid lines are true observations. The error L_{nn} is the Euclidean distance between projected source points, $\mathbf{T}_{\theta_{AB}}(p_a)$ onto the target image and corresponding target points, p_b. The cyclic consistency loss, L_{cyc} measures the Euclidean distance between source points, p_a and its projection under the combined forward-backward mapping, $\mathbf{T}_{\theta_{BA}}(\mathbf{T}_{\theta_{AB}}(p_a))$

.

$$L_{us} = \frac{1}{|G|}\sum_{i=1}^{|G|} ||\mathbf{T}_{\hat{\theta}_{BA}}(\mathbf{T}_{\hat{\theta}_{AB}}(g)) - g||_2 \tag{2.2}$$

It is to be noted that cyclic consistency alone cannot converge to any form of optimal solution as the model can simply converge to identity mappings. However, when combined with the supervised loss function an identity mapping will produce high loss for the labelled pairs. Publication I optimizes the combined loss function, $L = L_s + L_{us}$ to learn semantic matching from semi-supervised data.

The labelling constraint is further relaxed in Publication II, where only sparse keypoint locations per image is required. In theory, such information can also be obtained from keypoint detectors, but the idea remains for future work. Due to the lack of direct keypoint correspondence a point-set matching loss is proposed based on nearest neighbor assignment. That is, given a source-target image pair, the estimated geometric transformation parameters is used to project the source keypoints onto the target image plane. Thereafter, each projected source point is assigned the nearest ground-truth target key-point based on Euclidean distance. The NN loss function is presented below

$$L_{nn} = \frac{1}{M}\sum_{i=1}^{M} \min_{p_b \in \mathbb{P}_B} ||\mathbf{T}_{\hat{\theta}_{AB}}(p_a) - p_b||_2 \tag{2.3}$$

In addition, another objective function, Chamfer distance (CD) based on similar principles of nearest neighbor assignment is used to learn semantic keypoint correspondence. CD additionally measures the distance between

each target point and its nearest-neighbor projected source point.

$$L_{cd} = \frac{1}{M}\sum_{i=1}^{M} \min_{p_b \in \mathbb{P}_B} ||\mathbf{T}_{\hat{\theta}_{AB}}(p_a) - p_b||_2 + \frac{1}{N}\sum_{i=1}^{N} \min_{p_a \in \mathbb{P}_A} ||\mathbf{T}_{\hat{\theta}_{AB}}(p_a) - p_b||_2 \tag{2.4}$$

The nearest neighbor constraint is simple and allows to match unordered sets of keypoints with different set cardinalities. However it fails to generalize to real world image pairs with significant viewpoint variation. If the initial nearest neighbor assignment fails, then the network converges to incorrect solutions. Thereby, similar to Publication I, a second loss function based on cyclic consistency is added to the nearest neighbor loss. The cyclic consistency is applied to the source keypoints, i.e. the pre-defined grid points in Equation 2.2 is replaced with, $p_a \in \mathbb{P}_A$. If the nearest-neighbor target point is assigned incorrectly to a source point, then the backward transformation will restrict the convergence of the points under the nearest-neighbor constraint. Thus the network has to search the space of transformation parameters such that a source point is projected close to a target point which re-projects back to the original source point under the backward transformation. A pictorial representation of the proposed weakly supervised loss function is shown in Figure. 2.2.

Due to good network initialization obtained by pre-training on synthetic augmentations [100], the proposed weakly supervised loss converges and provides improvement in semantic matching performance. It is to be noted that keypoint supervision is only used during training while at test time only a pair of RGB images is required. The proposed loss functions are computed in both the forward and backward direction. The final loss is the sum of the combined loss values.

2.2.2 Weakly Labelled Dataset

The weakly supervised methods in Publications I and II are trained and evaluated on Proposal-Flow dataset. The dataset has been a widely used standard benchmark for evaluating semantic matching [101, 45, 100]. The dataset consists of 1400 image pairs selected from PASCAL-VOC [44] dataset. The images are selected from 20 object categories and are annotated with corresponding key-point locations.

To leverage the power of unlabelled image dataset to model better variations in geometric transformation, additional training image pairs are generated from the given training set using multiple ways. First, the image pairs in the training set are randomly flipped, which resulted in a total of 2500 training pairs from the original 700. This dataset is labelled D_l. This dataset still does not model enough variations as the original training

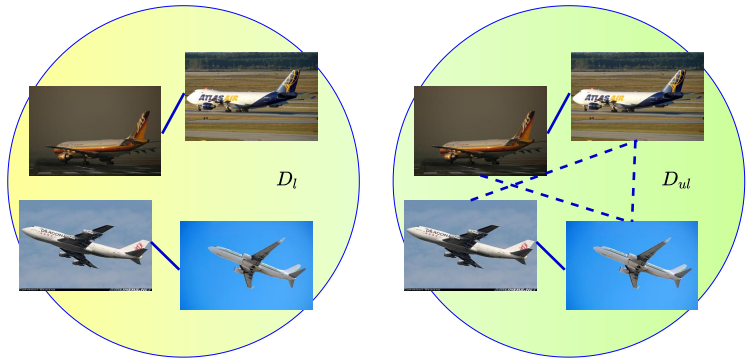

Figure 2.3. An overview of the weakly labelled semantic matching dataset. On the left is shown the original dataset, D_l [44] with an edge representing the corresponding training image pairs. On the right is the new dataset, D_{ul} generated by pairwise permutations resulting in additional edges or training pairs. Notice that the pairs in D_{ul} have larger viewpoint variation compared to D_l.

pairs themselves have very limited variations in geometric transformation as shown in Figure 2.3.

It is observed that different image pairs from the same object category had wider geometric variations between themselves as shown in Figure 2.3. Thereby, we list all the images from the original training set under their respective object category. Thereafter, new image pairs are generated by random category specific pairwise combination of images. As some object classes have more samples, the generation of image pairs is limited to 100 per object category. This balances the contributions from different object categories and prevents the learnt CNN model to be biased towards transformations from the dominant object classes. This results in a new set of image pairs labelled, D_{ul} with an additional 1800 pairs.

The combined set $D = D_l \cup D_{ul}$ forms our weakly-labeled training dataset. We observed that direct flips of 100 test image pairs were present in D. Although existing methods ignore this bias, we decided to remove these pairs from D.

2.3 Results and Discussion

In this Chapter, Section 2.1 provided a brief introduction to the field of local descriptor matching, and then, Section 2.2 briefly introduced the problem of semantic matching. In addition, methods and limitations of CNN based approaches were reviewed and an introduction to proposed solutions in Publications I and II was discussed.

In summary, methods for using small scale datasets with sparse annotations and cost efficiency in generating correspondence labels is addressed in Publications I and II. Firstly, a simple method is proposed to generate more data by augmenting current small scale semantic correspondence datasets.

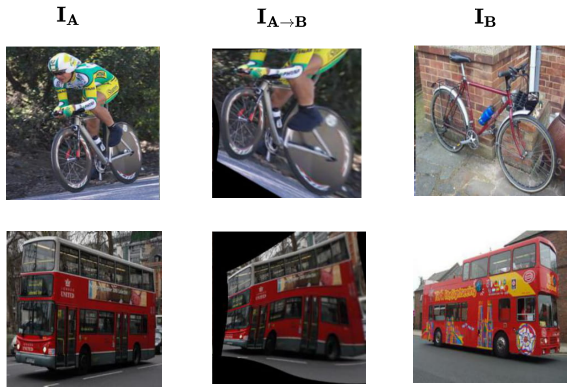

Figure 2.4. Some failure cases of Publication II. the figure shows the source, $\mathbf{I_A}$ and target images, $\mathbf{I_B}$ and the source image aligned onto the target image under the estimated transformation, $\mathbf{I_{A\to B}}$. The network fails to account for the small local variations in symmetrically shaped objects. However, smoothness of the global object symmetries suggests the network has learnt canonical object representations using the proposed method

.

Secondly, a novel objective function based on cyclic consistency is proposed to train CNN models on the new augmented dataset. In Publication I, the cyclic consistency loss is used along with a standard supervised loss based on pixel re-projection error. The supervision is required only for the small set of original image pairs while cyclic consistency loss is applied to pairs from the augmented dataset. Furthermore, the supervision constraint is further relaxed from requiring sparse keypoints per image and a sparse pixel correspondence mapping between images to only requiring keypoint locations per image in Publication II. The supervised loss is replaced with a point-set matching objective based on nearest neighbor match assignment.

The results presented in Publications I and II demonstrate superior semantic matching performance in terms of the standard metric of percentage of correctly transferred keypoints (PCK). Both methods in Publications I and II obtain an improvement of 12-15% over baseline method [100] in semantic PCK metric on Proposal-Flow dataset. One of the main limitations of the presented work is its inability to handle very large viewpoint transformations such as presented in Figure 2.4. This can partly be attributed to the absence of training image pairs with similar viewpoint variations. However, the incorrectly predicted transformation as presented indicates that the network maintains the global and local structure in the predicted transformation. This implies the network learns a canonical form of the semantic objects from just the keypoint locations and proposed matching objective function.

A number of later works have proposed weakly supervised methods for semantic matching [84, 77, 19]. In addition, [84] proposed a new dataset, *SPair-71k* to address the data scarcity problem. The dataset

consists of total 70,958 pairs of images collected from PASCAL 3D+ [33] and PASCAL VOC 2012 [145]. The dataset has richer annotations in the form of semantic keypoint locations, object segmentation masks, bounding boxes, viewpoint, scale etc. The problem of semantic matching also has an important requirement in computer graphics with image morphing between different object categories [1]. In addition the problem of camera relocalization [112, 75, 64] have shown improvements in performance by incorporating semantic information in the matching process. Due to changes in seasonal weather, large variation in appearance occurs in outdoor images. Incorporating semantic information provides meta information leading to robust image matching. This is similar to the theme of the Chapter which is modelling image matching across appearance variations arising from intra-class variations.

3. Image Retrieval

This Chapter addresses the problem of image retrieval in the context of recent developments in the field of deep neural networks. Similar to the problem of semantic matching, image retrieval also borrows models and frameworks from the problem of instance matching. Instance matching methods introduced in the previous Chapter perform matching using local descriptors. As such these methods though accurate fail to scale to large scale datasets. Image retrieval methods addresses this bottleneck using global aggregation of local descriptors to perform image matching. Algorithms addressing the scalability-accuracy trade-off are addressed in Publication III and IV. The proposed methods are built on existing advancements in the field which is summarize below.

The seminal work of Sivic *et al*. [122] first addressed the scalability of instance matching problems by extending the idea of bag-of-words model from text retrieval. The local descriptors in each image were encoded into the bag-of-words histogram representation and image matching score was computed using the term frequency-inverted document frequency (tf-idf) score. Later works introduced more robust global encoding methods such as Fisher vectors [91], VLAD [7] and accurate global matching method such as selective matching kernels [130].

Similarly, several works [97, 8, 41, 131] propose global encoding of local descriptors from ImageNet pre-trained CNNs. Instance level tasks such as object retrieval require more fine-grained decision boundaries to separate different instances of the same category. In contrast, category based problems such as object classification, semantic matching, aim to suppress intra-class variations to solve general computer vision tasks. As such, the performance of CNN descriptors lagged the standard hand-crafted descriptors. The ideas from local patch matching [54, 90, 46, 119, 149, 85, 128] were extended to full image matching for learning global image representations [8, 5, 94, 42]. The advancements were in two directions:generation of large scale landmark retrieval datasets with ground-truth without manual annotation, and algorithms to compute global representations from local CNN descriptor maps. While [5] addressed retrieval in the challenging

Figure 3.1. Example images from the Oxford [92] and Paris [93] datasets. Four query images from these datasets are shown with the corresponding region of interest (ROI) in red colored bounding box. Existing methods only encode the image region inside the ROI avoiding interference from the region exterior to ROI (context). Next to each query an irrelevant database image is shown with a similar pattern as the ROI. The aim of Publication III is to use global context in query images to lower the ranking of such irrelevant database images.

large city scale, others [8, 94, 42] targeted particular object retrieval such as popular landmarks. Therefore, the datasets used for fine-tuning the CNN parameters were also different. For instance, in [5], the authors collected geo-tagged images from Google Street View in urban cities such as Pittsburgh and Tokyo. On the other hand, [8, 94, 42] used Flickr and other image search engines to query images of landmarks that were popular tourist destinations. The images were labelled either using available geo-tags or using SfM to create 3D models. Subsequently, image similarity labels were computed using overlap of camera field-of-views. Concurrent to the weakly supervised data collection, advancements in global encoding methods were also achieved. Arandjelovic *et al*. [5] proposed NetVLAD, an end-to-end learnable VLAD layer, while [8, 94, 42] proposed aggregating local descriptors using pooling layers such as global sum or max pooling, and regional pooling (R-MAC) [131].

Global representations allows image retrieval pipelines to scale to billions of database images by means of faster approximate similarity matching. However, compression of large number of local descriptors introduces various forms of quantization noise. The presence of variations in illumination, occlusion, and viewpoint further affects the global representations.

3.1 Object Retrieval

Publication III studies methods for improving the robustness of global encoding methods in the context of particular object retrieval. The objective is to retrieve images similar to the object of interest in a given image. Each

query image is labelled with the region of interest (ROI) specifying the bounding box coordinates containing an object of interest as shown in Figure 3.1. Existing methods [94, 42] only encode the cropped ROI into a global representation. This suppresses interference from background patterns. On the other hand, information outside the ROI can also provide facilitatory contextual signals that improves the retrieval accuracy. As the facilitatory or inhibitory nature cannot be known apriori, the use of such contextual information remains a tightly coupled problem. This problem is addressed in Publication III where we extend the computational model of hippocampus spatial attention, introduced by Mozer *et al*.in 1998 [88], to jointly encode the ROI and its context in the final global representation.

Top down attention has received significant success in computer vision applications such as image classification [86, 16] and image captioning [4] among many other recent works. While these works attempt to find the attention map using parameterized functions such as a neural network, our method attempts to solve how to apply the attention map. The above attention mechanisms crop the input image at locations of high attention predicted by the attention network. The cropped input patches are feed-forwarded through the representation network to obtain final predictions. Cropping image patch loses structural information which is important as CNN parameters are trained to respond to both low and high level structures or patterns. In the context of image retrieval, Chum *et al*. [22] proposed a context expansion algorithm to incorporate descriptors beyond query ROI to improve retrieval performance. The method first obtains an initial ranked list of database images using the local descriptors inside the ROI. Then the model iteratively adds descriptors beyond ROI that are co-observed in the initial top ranked images and also passed a costly spatial geometric verification step. In contrast the proposed method jointly encodes the context without requiring any costly verification steps or querying the database multiple times.

The method requires the CNN representations for an input image from different layers, $\mathbf{X} = \{\mathbf{X}^l\}_{l=1}^{L}$, where $\mathbf{X}^l \in \mathbb{R}^{W^l \times H^l \times N^l}$. The main stages of the proposed method are as follows. Firstly, a bottom-up approach computes a 2D saliency map using the 3D representation maps from the final convolutional layer of the CNN. The saliency map is obtained by summing the representations along the channel dimension followed by max normalization, $M = \sum_{n=1}^{N^L} X_n^L \in [0,1]^{W^L \times H^L}$ where $X_n^L \in \mathbf{X}^L, n = 1...N^L$. Secondly, in a top-down fashion the saliency map is used to modulate representations from the an intermediate layer, $\mathbf{X}^l$ beyond the ROI. In practise we consider the projection of the ROI, R^l onto the intermediate layer,l that is considered for modulation. The saliency map is resized to match the resolution of the representation maps at that given layer. For each channel, X_n^l modulation is then performed by scaling activations outside ROI, $X_n^l(p)$, $p \notin R^l$ with the factor $g(M(p))$, where $M(p)$ is the corresponding saliency

value at 2D location, p. More concretely,

$$\tilde{X}_n^l(p) = \begin{cases} X_n^l(p), & \text{if } p \in R^l \\ g(M(p))X_n^l(p), & \text{if } p \notin R^l, \end{cases} \tag{3.1}$$

$g(.)$ is a monotonic function [88] :

$$g(a) = \lambda_1 + \lambda_2 a^{\phi} \tag{3.2}$$

The constants λ_1, $\lambda_2 \in (0,1)$ are so chosen such that the function $g(.)$ always maintains a value less than one i.e. $g(.) < 1$. ϕ non-linearly suppresses activations with weak saliency level. The modulated intermediate representations, $\tilde{X}_n^l \in \tilde{\mathbf{X}}^l$ are feed-forwarded through the remaining parts of the network to obtain context encoded representations, $\tilde{\mathbf{X}}^L$.

To encode the final 3D representation maps into global vector representations, R-MAC encoding technique is used. The standard R-MAC defines several overlapping pre-defined regions in the form of a multi-scale grid covering the spatial dimensions of the representation map. The local CNN representations in each region are max-pooled resulting in a single vector representation per region. The regional representations are then $l2$ normalized and sum-aggregated followed by a final $l2$ normalization. The aggregation step suffers from the drawback of assigning equal weights to each regional representation. As the regions are pre-defined and independent of the image content, responses from background clutter can negatively interfere with the final aggregated representation. To circumvent this issue, Gordo *et al*. [42] proposed to learn the parameters of a Region Proposal Network (RPN) that generates regions around potential landmarks in the image. However, such a method has the drawback of requiring to train the RPN separately for the given task. Instead, a weighted aggregation scheme is proposed whereby each regional representation is weighted with the corresponding saliency value derived from the bottom-up saliency map introduced earlier. The saliency values are scalar estimates of regional saliency and are computed by max-pooling the saliency map over a R-MAC defined grid.

The proposed methods are based on the CNN network fine-tuned for landmark retrieval [43]. Additional training is not required while the attention model can be seamlessly integrated into any network architecture. Three baseline methods are considered for evaluation:i)Full query, ii) Cropped ROI, and iii) Cropped Activation. Full query encodes the full query image (ROI + context) into a global representation. Cropped ROI refers to the encoding of only the ROI into a fixed length representation. Finally, instead of cropping the ROI in the image space, cropped activation crops the ROI projection on the final layer representations. Networks with growing receptive field size are able to encode context in the final layer activations. That is, cropped activation encodes additional information beyond the ROI

Figure 3.2. Qualitative results produced by Publication IV on Tokyo24/7. Each column corresponds to a query, and the top-1 NN database images by NetVLAD [5] and the verification method proposed in Publication IV. Similarity in global properties such as geometry, illumination and viewpoint results in the failure of NetVLAD. Publication IV utilizes local similarity and geometric verification to retrieve correct database images (bottom row).

defined by the receptive field size of the final layer activations. Thus, it can be seen as a special case of the proposed attention model. The methods are evaluated on the challenging Oxford [92] and Paris [93] datasets using the mean Average Precision (mAP) metric.

3.2 Geometric Verification

Particular landmark retrieval poses the problem of retrieving objects belonging to the same landmark category with geometry and design variations. The challenge becomes more prominent in city-scale retrieval problems where geometry is fairly ubiquitous across the city. The distinguishing factor has to additionally account for both local and global patterns. Under such circumstances global image representations cannot accurately retrieve relevant database images purely from global similarity. Figure 3.2 shows some failure cases for global image encoding methods such as NetVLAD [5]. In this context the problem of city-scale retrieval using dataset such as Tokyo 24/7 is studied in Publication IV. The dataset also features challenging day-night queries.

To overcome the limitations of global representations, a second re-ranking step consisting of geometric verification of local descriptor matches is performed. Geometric verification finds the local descriptor matches that are inliers with respect to some geometric models such as homography, affine or epipolar transformations. Traditionally, there is a vast literature [21] covering different aspects of geometric verification using local hand-crafted descriptors such as SIFT. In line with the success of CNN representations in image matching, we extend its application to the problem of geometric

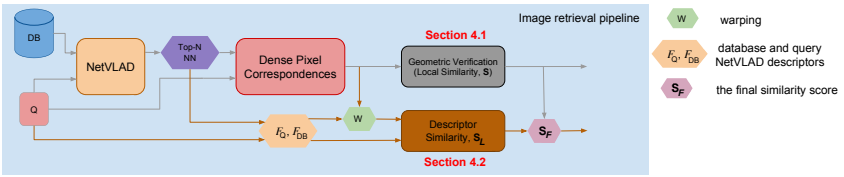

Figure 3.3. Overview of the proposed pipeline. Given a query image, we first rank the database images based on global similarity (*e.g.* using NetVLAD). In the next stage dense pixel correspondences are computed between the query and top N ranked database images. These correspondences are then verified by the proposed similarity functions utilizing geometry and CNN based image descriptors to re-rank database images according to the input query.

verification in Publication IV.

3.2.1 Overview

The image retrieval system considered in Publication IV consists of the following three stages: *1)* Given a query image, a shortlist of top ranked NN database images is obtained using global image representations; *2)* Dense pixel correspondence is obtained between query and each top ranked NN database image using a CNN based correspondence network; *3)* A subset of local matches obtained using RANSAC based verification and cyclic consistency are used to compute a set of similarity scores, which provides the final re-ranking of the shortlist.

The particular architecture of the aforementioned retrieval pipeline used in this work is illustrated in Fig. 3.3. That is, we use NetVLAD [5] for the first stage, our own modified version of DGC-Net [80] for the second stage, and the proposed approach with a novel similarity metric for the third stage. Here NetVLAD is used for retrieval, but also other global image level descriptors could be used instead.

Methods proposed in Publication IV makes contributions related to stages *2)* and *3)* above and are described in the following sections. Our modifications to the DGC-Net architecture are described in Section 3.2.2. The geometric verification method is presented in Section 3.2.3.

3.2.2 Image Matching

The DGC-Net architecture is fully convolutional and consists of a pre-trained encoder such as VGG16 and a series of stacked decoder layers. Given an image pair, the encoder output representation maps for each image are passed through a global correlation layer and the first decoder layer that outputs a coarse correspondence map. The remaining decoder layers refine the coarse correspondence map using finer details from the lower layers of the encoder. In this thesis, Publication IV proposed methods to the global correlation and the refinement decoder layers resulting in improved and efficient matching.

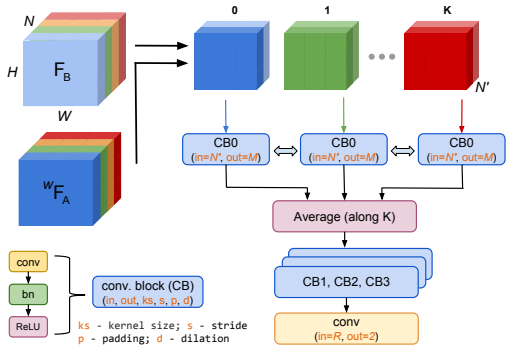

Figure 3.4. Overview of the unified correspondence map decoder (UCMD) D_c. The feature maps of the target F_B and the warped source ${}^{w}F_A$ images have been split into K tensor blocks of channel length $N\prime$ and then concatenated along the channel dimension. Further, each concatenated tensor is complemented by the correspondence map estimates $H \times W \times 2$ from the coarser layer and then fed into a convolutional block $CB0$ with $2N'$ inputs and shared weights. The output feature maps of $CB0$ are then averaged and processed by the remaining layers of the decoder to produce refined pixel correspondence estimates.

The global correlation layer performs cross-correlation of encoder representations from each image. This results in a 4D match tensor containing the match probability. The first decoder layer in DGC-Net takes the normalized cross-correlated match tensor as input and outputs a coarse correspondence field. In doing so it fails to take into account the spatial and cyclical coherence of the 4D matches. In other words, good matches are expected to be located in dense regions with higher similarity scores in its spatial neighborhood in the 4D match tensor. In addition, the matches should be mutually consistent in both forward and backward direction. To ensure local and cyclical consistency, we integrate the Neighborhood Consensus Network (NC-Net) [102] in between the global correlation and the first decoder layer.

The encoder-decoder pairing is fairly general and usually consists of concatenating a pair of encoder maps that are inputted to the corresponding decoder layer consisting of several convolutional layers. The concatenation operation conditions each decoder layer to account for the dimensionality of the corresponding encoder maps. In contrast, Publication IV removes this constraint using block-wise concatenation of the l^{th} layer encoder maps. That is, the input encoder map, $\mathbf{F}^l \in \mathbb{R}^{W^l \times H^l \times N^l}$ is divided along the channel dimension into K^l non-overlapping 3D tensor blocks, $\mathbf{F}^l = \{\mathbf{F}^l_k\}_{k=1}^{K^l}$, where $\mathbf{F}^l_k \in \mathbb{R}^{W^l \times H^l \times N'}$ and $K^l = N^l/N'$. Each of the warped source and target block representations, ${}^{w}\mathbf{F}^l_{k,A}, \mathbf{F}^l_{k,B}, k = 1...K^l$ are concatenated and feed-forwarded through the first layer of the decoder. The block outputs are averaged and passed through the remaining convolutional layers of the decoder. This allows using a universal decoder across multiple encoder layers. The parameter, N' is assigned a small value of 16.

3.2.3 Verification System

The improved DGC-Net is applied to the problem of geometric verification which is presented next. The pipeline consists of a series of geometric verification steps outlined below.

The first stage combines cyclic consistency with RANSAC based geometric model fitting. As DGC-Net was initially proposed to operate on similar images, its functionality does not naturally extend to dissimilar images. That is, for both similar and dissimilar image pairs, DGC-Net outputs a locally smooth correspondence map. As such, standard application of RANSAC based model fitting estimates large inlier counts, I for dissimilar image pairs as well. To circumvent this issue only those inliers are kept that are cyclically consistent, C. That is, those inliers for which the combined mapping under the forward and backward correspondence map is close to identity mapping are considered cyclically consistent. The number of such cyclically consistent inliers are used to compute a new similarity score, S that re-ranks a shortlist of top ranked images from the initial ranked list obtained by NetVLAD.

$$S = \frac{|C|}{|I|} \cdot exp\left(-\frac{\beta}{|C|}\right), \tag{3.3}$$

where β is a constant. As $|C|/|I|$ is a ratio, the exponential term is added to down-weight the similarity cost for image pairs which have fewer cyclically consistent correspondences in the inlier set. β is a constant set to 240x240, the fixed input image resolution.

Although the above mentioned method improves the re-ranking, it fails to address outliers resulting from large similarities in structure, texture, and layout between buildings in the particular setting of city-scale retrieval. This can be related to the fact that inlier count only reveals information about the geometry and fails to take into account both local and global similarities. Consequently, to compute these similarities we first extract both local and global representations using NetVLAD for each image. The NetVLAD architecture consists of a VGG16 encoder and a NetVLAD layer that performs soft visual word assignment. The visual word centers are learned in an end-to-end fashion during training. The local representations are extracted from the encoder output while the global representations are extracted from the final NetVLAD layer output. The local similarity, S_L between an image pair is then computed using cosine similarity between their respective local representations at the cyclically consistent match locations. On the other hand, global dissimilarity, G is computed using Euclidean distance between the respective global representations. The three similarity scores, S,S_L,G are fused to obtain the final similarity scores, S_F. In order to save computational time, we compute the above final similarity based 2nd re-ranking only for the top-ranked images in the

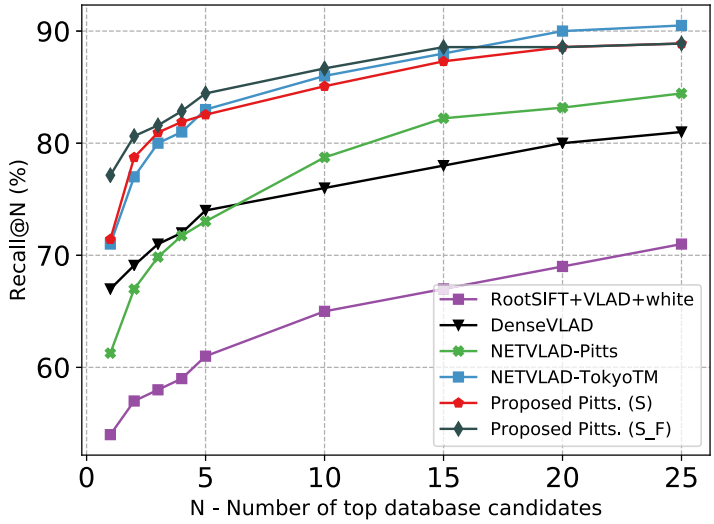

Figure 3.5. Comparison of the proposed methods versus state-of-the approaches for place recognition.

1st re-ranked list obtained using S.

$$S_F = \log_{10}(S_L \cdot S) \cdot 10^{-G} \tag{3.4}$$

The similarity fusion Equation 3.4 (see Publication IV for more details) is empirically derived with trial and error on validation dataset. The possibility of better models to fuse the respective similarity scores is tested using a neural network. The input to the neural network are C, I, S_L and G. In the arxiv version [68], the supplementary section of Publication IV is added containing a detailed analysis of the ablation studies. The proposed method is also compared against the standard SIFT based geometric verification. InLoc, a multi-stage geometric verification pipeline based on CNN [125] is used as a baseline. However, the last view synthesis stage in InLoc is ignored due to the requirement of depth maps. Retrieval performance is evaluated on the challenging Tokyo247 dataset [132] using Recall@N metric. Furthermore, the retrieval performance is measured by plugging the system in an image-based localization system. As will be detailed in the next Chapter, image retrieval is the first and usually the key component in localization pipelines. Therefore, improvement in retrieval stage should reflect on the final localization performance as well. The localization performance is measured on the Aachen [107] and CMU Seasons [107] datasets.

Methods	Condition, $5m, 10^\circ$				
	Aachen Day-Night		CMU-Seasons		
	day	night	urban	suburban	park
HF-Net [104]	94.2	76.5	97.9	92.7	80.4
D2-Net [32]	93.4	74.5	-	-	-
Active Search [108]	96.6	43.9	-	-	-
NetVLAD-Pitts	81.7	64.3	78.9	77.0	63.2
Publication IV (S_F)	*84.7*	*68.4*	*89.1*	*77.1*	*63.3*

Table 3.1. Localization performance on the Aachen and CMU-Seasons datasets (higher is better). The best performance among *image retrieval* based approaches (Row 4-5) is highlighted as *italic*.

3.3 Results and Discussion

The Chapter presented the problem of image retrieval using CNN based image representations. Both the problem of particular landmark retrieval and city-scale image retrieval are studied. A brief summary of the key findings are discussed below.

In summary Publication III proposed a model to incorporate contextual information in query image to improve particular object retrieval. A technical report on arxiv [66] extends the attention model to the database side as well. The proposed method brings consistent improvement over baseline methods in the final retrieval performance on Oxford5k and Paris6k datasets. Furthermore, cropped activation based setting has a comparable performance to the proposed method which can be attributed to the large receptive field size of ResNet101 architecture used in the paper.

In Publication IV, the problem of geometric verification is revisited. Unlike traditional verification systems based on SIFT, the proposed method is based on local CNN descriptor matching. Firstly, modifications to the end-to-end correspondence network, DGC-Net are proposed by improving the matching layer and compacting the decoder architecture. The modified DGC-Net attains higher accuracy on HPatches dataset in terms of average end point error (AEPE) metric. On the other hand, the universality of the proposed decoder allows addressing inference budget constraints by replacing the VGG16 encoder to MobileNetv2 [103] without requiring any re-training the full encoder-decoder network. Results show that the proposed decoder with a MobileNetv2 encoder achieves similar performance as VGG16 encoder on Tokyo247 while reducing the number of parameters from 8**M** to 1**M**.

The two-stage verification pipeline outperforms baseline methods on both Tokyo247 and localization benchmarks Aachen and CMU-Seasons datasets as shown in Figure 3.5 and Table.3.1 respectively. It is to be noted that the corresponding baseline in both benchmarks is NetVLAD-Pitts

which is the NetVLAD version trained on Pittsburgh dataset [133]. The localization pipeline is kept the same in Aachen and CMU-Seasons i.e. the local 2D-2D matching is done using the modified DGC-Net. Better accuracy in query camera pose estimation *given the same localization pipeline and image matching method* shows that our approach retrieves higher quality database images compared to NetVLAD-Pitts.

Geometric verification using local CNN descriptors are now well studied [118] and widely used in image-based localization tasks. The most prominent methods are SuperPoint [28], D2-Net [32] that extract keypoint locations from high level CNN representations followed by local matching based on mutual nearest neighbor search. However, unlike our proposed method, the local spatial consistency of the match is not taken into account prior to verification stage. A preliminary result [68] obtained by applying our verification method only to the SuperPoint locations results in much improved retrieval accuracy. A recent work [62] combines a similar network for image based localization in outdoor scenes.

4. Camera Relocalization

Image-based localization is a popular topic in the field of geometric computer vision. The task is to localize a query image in an environment represented by a set of database images with corresponding pose or some form of 3D model as shown in Figure 4.1. In this Chapter we provide a high level connection between instance matching for geometry estimation and image retrieval to solve the problem of image-based localization. Image retrieval with global image representations (such as those presented in Chapter 3) allows for efficient and scalable retrieval of database images that hypothesize candidate locations for the query image to be localized. On the other hand deep learning based geometry estimation between images (similar to Chapter 2) provides more accurate estimates of query image pose w.r.t the relevant database images. The application of image retrieval and geometry estimation is a well studied problem in image-based localization and is extended to deal with CNN based models in this Chapter.

First, a brief overview of existing approaches based on hand-crafted descriptors and CNN based methods is presented. Then, the limitations of existing CNN based methods and the proposed approach in Publication V is discussed.

4.1 Related Work

We briefly summarize prior works in the field of image-based localization. Existing works can be broadly divided into two categories: place recognition based and image-based methods.

Place recognition based methods approximate the location of the query using the pose of the nearest neighbor database images [5, 18, 105, 132, 142]. The relevant database images are obtained using image retrieval methods summarized in the previous Chapter. However, the accuracy of the query pose is limited by the spacing between the relevant database images and the query image. Torii *et al*. [132] utilize depth maps to generate novel

Figure 4.1. Image-based localization aims to recover the 6 DoF camera pose w.r.t to an environment [57] from an input image (e.g. captured by a mobile phone camera). The environment is represented as a 3D point cloud or a database of images with respective poses and local 2D keypoint representations.

view images around the NN database images. The synthetically generated images are added to the database improving the localization performance.

Image-based methods compute the query pose utilizing the 3D representation of a scene obtained from SfM. Thereafter, 3D scene points are matched with local 2D descriptors in the query image to obtain tentative correspondences. The candidate 2D-3D matches are verified using RANSAC [36] followed by query pose estimation using Perspective-n-Point algorithm [60]. To address large costs resulting from local descriptor matching in large scenes, several approaches [20, 74, 106] propose accelerating the correspondence search by terminating the search procedure as soon as enough inlier matches have been obtained. Other methods [17, 53, 109, 108] utilize an intermediate place recognition step to obtain candidate query locations represented by the top retrieved relevant database images. The query pose is estimated by restricting the matching only to the 3D points visible in the NN database images.

Recently, it has been shown that data-driven techniques provide efficient solutions to the pose estimation problem. Shotton *et al*. [116] train a regression forest to regress 3D scene locations for pixel in an input RGB-D image. Thus the method bypasses the costly local matching stage by directly regressing 2D-3D matches. Subsequently, the candidate corre-

spondences are verified using RANSAC and the query pose estimated. In contrast, Valentin *et al*. [137] skip the correspondence estimation step by exploiting the uncertainty of the predicted 3D pixel locations to estimate query pose. Brachmann *et al*. [14] propose an end-to-end trainable CNN architecture for camera pose estimation. The network estimates 2D-3D matches from monocular images followed by RANSAC based geometric verification in a full differentiable manner. Li *et al*. [71, 72, 73] proposed data-augmentation, angle-based reprojection loss and hierarchical 2D-3D match estimation to advance the performance of [14].

In contrast to the above structure based image localization methods, Kendall *et al*. [55, 56, 57] proposed a CNN architecture, PoseNet that directly regresses the camera pose from an input RGB image. Later works based on similar principles by Walch *et al*. [141] propose context aggregation using LSTM units to improve camera localization. Valada *et al*. [135, 95, 136] incorporate semantic information to robustly estimate camera pose estimation. In turn, Clark *et al*. [23] utilizes recurrent neural networks (RNN) to capture temporal geometric cues from video sequences outputting accurate global camera pose. However, end-to-end CNN camera pose estimation pipelines are limited due to lack of generalization [110]. More recently, several approaches have been proposed that combine CNN based local descriptors with traditional localization pipeline [108]. Shi *et al*. [115] proposed a method combining semantic information with structure-based methods that leads to improved localization performance. Sarlin *et al*. [104] propose a hierarchical localization approach based on coarse-to-fine localization. NetVLAD [5] global descriptors are used to retrieve NN database images to obtain a coarse estimate of query pose. Thereafter, using local descriptors from SuperPoint [28] an accurate 6-DoF query camera pose is estimated. Several recent works [32, 99] have advanced the method of structured localization using local CNN descriptors.

4.2 Localization Pipeline

The key limitation of end-to-end CNN based localization approaches is that the learning process is strongly coupled with the coordinate frame of the scene. For example, end-to-end camera pose estimation pipelines [55, 56, 57, 141] learn a mapping from image pixels to the corresponding camera pose. On the other hand, structured approaches [14] learn a similar mapping from pixels to 3D scene coordinates. In both cases, the final output space (pose/scene coordinates) are *dependent* on the coordinate frame of the scene specific training data. Due to the above coupling existing approaches [55, 56, 57, 141, 14] suffer from the following drawbacks. Firstly, the existing models are not scalable with increasing number of scenes. This follows from the fact that with increasing number of scenes the network

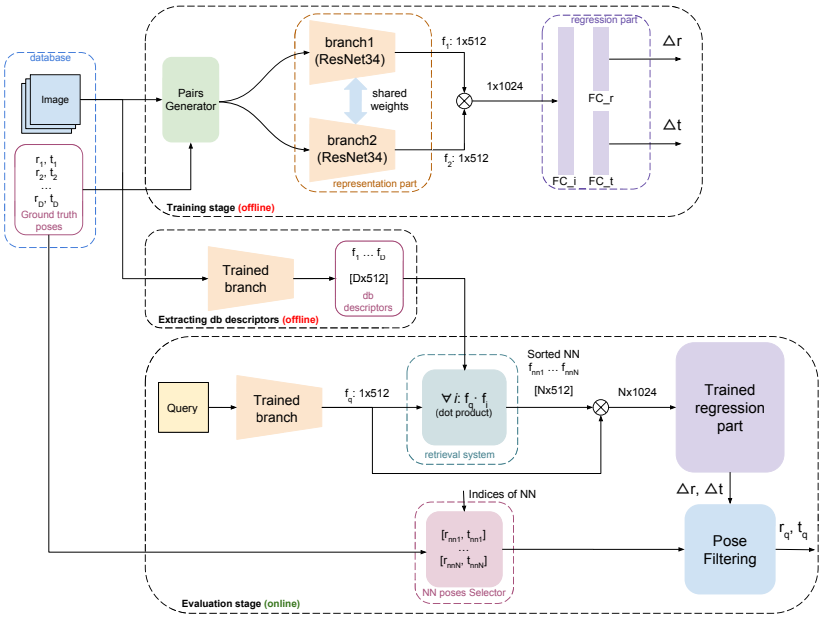

Figure 4.2. Pipeline of the proposed system. An ImageNet pre-trained Siamese based CNN network is trained to directly regress relative pose between a pair of cameras (top). At test time, the encoder of the trained model is used to compute representations of database and query images. Then, the dot product in the descriptor space is used to retrieve relevant database images. Consequently, query descriptor and its top N ranked database representations are concatenated and fed to the regression part of the network to predict pairwise relative pose. Finally, the proposed fusion algorithm naturally coalesces relative pose estimates and ground truth absolute poses to produce the full 6-DoF query location.

has to maintain a global representation of the joint space of images and corresponding camera pose. Secondly, the existing methods are inefficient due to the requirement of scene-specific training and storing of multiple models per scene. Thirdly, following the scene-specific training and evaluation constraint, the existing models do not generalize to novel scenes without re-training the models. Publication V addresses these limitations by proposing a localization framework that decouples the learning from the scene coordinate frame. The proposed pipeline is inspired by classical pipelines [106] that localize query image using sequential steps consisting of image retrieval, relative and absolute camera pose estimation. Publication V adopts the above framework while replacing the hand-crafted descriptors with learning based CNN models. The proposed approach is summarized next.

The first stage of the localization pipeline illustrated in Figure 4.2 consists of an offline stage of training a Siamese CNN network for estimating relative camera pose from an input image pair. The training image pairs are sampled independently from the joint space of training data from all the scenes. As a result the parameters of the CNN network can be trained

using image pairs of any scene thereby being able to improve towards generic relative pose estimator. Each Siamese branch first encodes the image into a global vector representation. A second shared branch then predicts the relative camera pose by concatenating global image representations from each image in an image pair. This conditioning of the relative camera pose estimator network on the global image representations allows for efficient storage and evaluation costs. That is large scale database images can be stored as compact global representations, while the costly image encoding stage can be avoided at evaluation time. This stage of training CNN is performed only once.

The second stage uses the trained CNN model for localization in multiple disjoint scenes. To localize a query image in a given scene, we assume representative database images from the scene and their respective camera poses to be available. First, a retrieval stage is used to identify nearest neighbor (NN) relevant database images. This is done by first extracting query and database global image representations using the trained model followed by a simple dot product to measure representation similarity. The absolute camera poses of the NN database images posits candidate hypothesis for the query image pose. However, as the query and its (true) NN database image can have significant viewpoint variation, we obtain a finer estimate of query camera pose using the trained relative camera pose estimator. That is, we concatenate and feed-forward global representations of the query and its NN database images in a pairwise manner through the relative pose network to obtain candidate query pose estimates.

As some of the retrieved NN database images can be outliers, the final stage of the pipeline performs a novel RANSAC based outlier estimation method to obtain the inlier query translation and rotation estimates. The translation estimates are simply averaged to obtain the final query location. On the other hand, query rotation estimate is obtained using a robust rotation averaging algorithm [47].

4.3 University Dataset

One of the contributions of Publication V is a new indoor localization dataset, *University* to fairly evaluate the performance of the proposed approach. In addition, the standard indoor localization benchmark dataset, *7Scenes* is also used to validate the localization performance. Both the datasets are summarized below.

7Scenes is a widely popular indoor localization dataset [23, 56, 57, 81] introduced by Microsoft. It consists of RGB-D images from 7 different scenes, namely: *Chess, Fire, Heads, Office, Pumpkin, Red Kitchen* and *Stairs*. Each scene exhibits significant variations in camera viewpoint and perceptual aliasing from similar objects with different spatial distributions. Existing

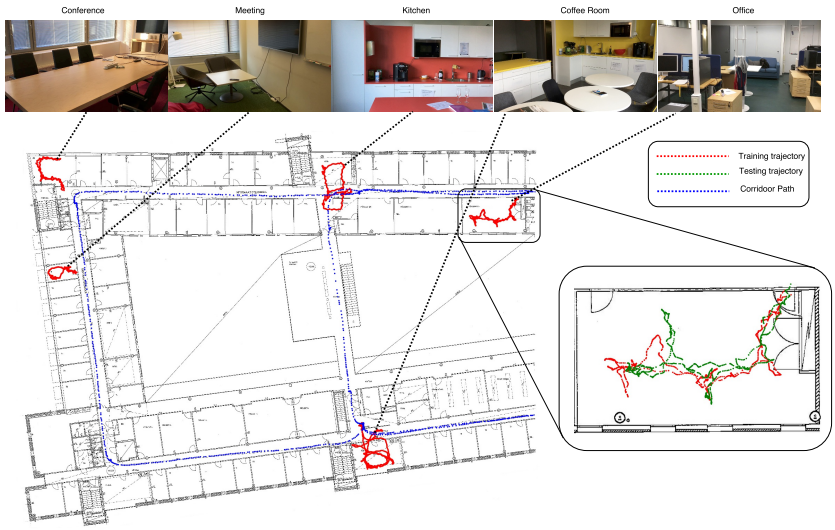

Figure 4.3. The University dataset. The proposed large-scale indoor localization dataset consists of 5 different scenes:*Conference, Meeting, Kitchen, Office*, and *Coffee Room* registered to a common global coordinate system. The detailed view shows the training (-red) and testing (-green) camera trajectories.

baseline approaches train and evaluate their CNN models independently for each scene. On the other hand, the proposed approach trains a single CNN model across the images from each scene. However, the final model is evaluated in a scene specific manner.

University dataset is introduced to address the limitations of the *7Scenes* dataset. As each scene in *7Scenes* dataset is registered to separate coordinate system, it limits the evaluation of baseline methods jointly across multiple scenes with a larger spatial distribution. This is fundamental to the application of developed methods in real life scenarios where large scale environments such as big shopping markets, museums etc consist of multiple sub-scenes.

The released dataset, *University* has a similar data structure as *7Scenes* allowing easy usage by existing models. The dataset is constructed by navigating a large indoor space in a University environment using a hand-held device (Google Project Tango and iPhone6s). The collected data consists of high resolution camera images from iPhone6s and corresponding raw camera pose trajectory from the Tango tablet. The raw camera odometry outputted by Tango tablet suffers from substantial drift [65]. The initial camera trajectory along with manually generated constraints using the classic checkerboard images are jointly optimized using the pose-graph optimization. The result is a drift corrected globally consistent camera odometry.

From the whole dataset we select camera images and poses from a set of

disjoint scenes, namely: *Office, Meeting, Kitchen, Conference*, and *Coffee Room* as shown in Fig. 4.3. Similar to *7Scenes* each scene contains multiple traversals through the scene. The dataset contains 9694 training and 5068 test images respectively.

4.4 Results and Discussion

Camera relocalization unifies the topics presented earlier in the thesis. More concretely, this Chapter presented the contributions of Publication V. Results and review of the current developments in the field of CNN based camera relocalization is discussed next.

In summary, Publication V advances the generalization and scalability of CNN based camera pose estimation. Unlike traditional CNN based approaches such as PoseNet, a coarse to fine localization approach is proposed based on image retrieval and relative camera pose estimation. Results in Publication V demonstrate that the proposed method performs comparably or better on the *7Scenes* dataset compared to baseline method based on PoseNet. The proposed method overcomes the limitations of baseline methods that require scene-specific training resulting in a single model per scene. For example on *7Scenes*, PoseNet requires 7 models and thus 7 times more parameters that are learnt compared to the proposed method that requires only 1 model trained jointly on all scenes. The scalability is further tested on the larger *University* dataset. Due to the design of the dataset, the baseline and the proposed methods were trained jointly on all the scenes. Results demonstrate that the proposed method clearly outperforms baseline *PoseNet* based pose regression method. Furthermore, the proposed method generalizes favourably to novel scenes not seen during training.

Several works such as *RelocNet* [10],CamNet [30], and SANet [147] extend the scene agnostic training of CNN based camera relocalization methods. While *RelocNet* and *CamNet* improve the loss function for training relative camera pose estimation network and the intermediate image retrieval step, *SANet* creates a hierarchical scene representation for iterative scene co-ordinate prediction and fusion. On the other hand, structured methods [28, 32, 99, 104] based on 2D-3D matching and local CNN descriptors have recently dominated the localization performance benchmarks[1]. Such methods require a 3D scene representation created using database images and SfM pipelines such as COLMAP [113]. An interesting direction of future work can be to unify the end-to-end relative camera pose estimation with local 2D-3D or 2D-2D matching. Possible application scenarios include low textured and illuminated scenes such as the interior of industry plants etc. An initial estimate from end-to-end network output

[1] https://www.visuallocalization.net/benchmark/

can be used to initialize the local matching process.

5. Summary of the Original Articles

We now briefly summarize the main findings of the thesis in this Chapter. The articles are reprinted and included in the appendix of the thesis.

Publication I addresses the problem of semantic matching in the context of learning from large amount of data with weak supervision. While existing methods are limited by supervisory data, the proposed method regularizes training signals from small amount of supervised data with large amount of unlabelled data. To regularize the learning process, an algorithm based on forward backward consistency of estimated transformation parameters is proposed. Publication II extends the idea of Publication I further to show that the proposed constraint can accurately learn geometric transformation without explicit correspondence information. In addition, method to generate large amount of weakly labelled data from existing small semantic matching datasets is presented. The results show the proposed method outperforms baseline methods on the challenging semantic correspondence task.

Publication III addresses the problem of retrieving objects of interest such as landmarks from a large collection of database images. Existing approaches only encode the region of interest containing the landmark in the query image using a CNN. The paper improves on this form of query encoding by additionally encoding the contextual information outside the region containing the landmark. The proposed approach first computes a bottom-up saliency measure using high level representations which is then used to attenuate contextual representations in intermediate CNN layers in a top-down fashion. The high level representations are re-computed in a second feed-forward pass from the modulated intermediate CNN representations. The final high level representations are then encoded into global vector representations using regional aggregation of local representations. Furthermore, a new method is presented that additionally weights the regional representations before aggregating. The combined method consistently outperforms baseline methods in the task of landmark retrieval.

Publication IV revisits the application of geometric verification in image

retrieval. The problem of image retrieval is studied in the context of large city-scale datasets. As large urban environments depict similar global imagery such as building structures, CNN based global encoding methods are inherently limited in uniquely encoding fine-grained local variations in textures, patterns etc. The paper addresses this issue by proposing a re-ranking strategy based on a series of sequential geometric verification steps. The first stage re-ranks a shortlist using a similarity function that computes the number of matches from a CNN based dense correspondence network that are inliers to a geometric model using RANSAC. Only those inliers are considered that are mutual nearest neighbors. In the second verification stage, a stricter and costlier step re-ranks the re-ranked shortlist from the previous step using local and global representation similarities. The representations are extracted using a CNN fine-tuned for retrieval in urban environments. The different similarities are combined to compute the final query-database similarity which is used for the second re-ranking of the shortlist. Results demonstrate significant improvement over both baseline global image matching results and a single stage verification step. Furthermore, the paper makes contributions to improve the accuracy and efficiency of the dense correspondence network. The accuracy is improved by constraining the matching correlation layer in the network to be cyclically and locally consistent. Multiple encoder specific decoder layers that perform the same function of refining correspondence are replaced with a single refinement decoder layer. This universal decoder can operate on any encoder layer representations. The universality is further tested by pairing the trained universal decoder with a new but compacter encoder at test time. Comparable image matching performance is obtained at a significant improvement in memory and computational costs.

Publication V studies the problem of camera relocalization. The paper proposes a novel CNN based localization system that performs coarse-to-fine query camera pose estimation using advances in image retrieval and relative camera pose estimation. More concretely, the proposed method utilizes a retrieval step to obtain coarse estimate of query camera location using pose hypothesis from the top-ranked database images. Thereafter, multiple candidate finer query pose estimates are obtained using a relative camera pose estimation network. The candidate pose hypotheses are passed through a novel and robust pose fusion algorithm to obtain the final query camera pose. Due to the intermediate image retrieval step, the proposed method is able to disentangle the representation learning step from the scene co-ordinate resulting in several advancements over existing CNN based direct pose regression methods. First improvement is in terms of scalability of localization systems across several scenes. Existing methods require a single CNN model for each scene whereby multiple models need to be stored in memory in addition to the increased training costs. Secondly the proposed method generalizes to new scenes

previously unseen during training. As another contribution, the paper introduces a new dataset for addressing large-scale indoor localization. The dataset consists of several scenes registered to the common co-ordinate frame and presents challenging conditions such as texture-less surfaces etc. Results across benchmark and proposed datasets demonstrate competitive performance compared to baseline methods while improving generalization and scalability.

6. Conclusion

This thesis has addressed a number of closely related computer vision problems, namely, semantic matching, image retrieval and image-based localization. The central theme of the thesis is the problem of image matching using CNN based representations and extending its application in prior standard computer vision pipelines such as geometric verification and image based localization. The thesis provides a general overview of the connections between targeted problems. For example, pairwise geometric image matching is used in both semantic matching and also to improve image retrieval using geometric verification. On the other hand, image retrieval and pairwise geometry estimation is also used to improve scalability and generalization of CNN based localization systems.

Efficiency is one of the recurring themes in the thesis. Obtaining large amount of labelled data has been a challenging aspect to training deep CNN models. Data and labelling efficiency can be obtained by generating more training signals from small labelled datasets. We proposed methods addressing data and label efficiency to improve CNN models for semantic matching. A recent work of ours also addresses this problem in the context of image retrieval [67]. Furthermore, the proposed universal decoder for end-to-end image matching allows for efficiency in memory and computational cost during inference. Inefficiency of existing end-to-end camera relocalization systems is addressed in terms of scalability and memory. The proposed localization framework provides unifies frameworks from classical localization pipelines with CNN based models.

Several avenues exist for extending the works presented in the thesis. The weakly supervised method for learning semantic matching depends on the supervision of keypoints. Detection of such keypoints can be learned jointly with the proposed semantic matching in an end-to-end manner [150]. Given the renewed interest of the computer vision community in attention models [139], it would be interesting to extend its application to the domain of image retrieval. In particular, bottom-up attention can be applied with our proposed spatial top-down model to enhance the representations of weakly represented regions. Weak response in the final representation

results from occlusion, smaller scale etc. Learning an end-to-end attention model to increase the response of potential landmark regions can address the above challenges and improve image retrieval and by extension image based localization systems. Extending the computer vision systems presented in the thesis to an end-to-end learning based system is a possible future direction. For instance, the localization system can be made end-to-end by making the final pose fusion step differentiable. Currently, learning is restricted to the relative camera pose estimation step that operates on image pairs. The next stage in the presented localization pipeline is the pose filtering and fusion step that takes into account representations and poses of relevant database images. By making this next stage in the pipeline differentiable the structure in data manifold can be exploited to learn better representations and pose estimates.

Bibliography

[1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *In ACM Transactions on Graphics (TOG)*, 2018.

[2] Abdul-Hadi Abulrub, Alex Attridge, and Mark Williams. Virtual reality in engineering education: The future of creative learning. 2011.

[3] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning (ICML)*, 2016.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European Conference on Computer Vision (ECCV)*, 2014.

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2017.

[10] Vassileios Balntas, Shuda Li, and Victor Prisacariu. Relocnet: Continuous metric learning relocalisation using neural nets. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding (CVIU)*, 2008.

[12] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] Christopher M Bishop. Pattern recognition and machine learning. In *springer*, 2006.

[14] 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.

[15] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: binary robust independent elementary features. In *IEEE European Conference on Computer Vision (ECCV)*, 2010.

[16] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[17] S. Cao and N. Snavely. Minimal Scene Descriptions from Structure from Motion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[18] 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)*, 2011.

[19] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[20] S. Choudhary and P. J. Narayanan. Visibility Probability Structure from SfM Datasets and Applications. In *European Conference on Computer Vision (ECCV)*, 2012.

[21] Ondrej Chum and Jiri Matas. Matching with prosac-progressive sample consensus. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[22] Ondřej Chum, Andrej Mikulik, Michal Perdoch, and Jiří Matas. Total recall ii: Query expansion revisited. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[23] R. Clark, S. Wang, A. Markham, N. Trigoni, and H.i Wen. VidLoc: A deep spatio-temporal model for 6-DoF video-clip relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE Conference on Computer Vision and Pattern Recognition. (CVPR)*, 2000.

[25] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 1995.

[26] Daniel Crevier. Ai: the tumultuous history of the search for artificial intelligence. In *Basic Books, Inc.*, 1993.

[27] 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)*, 2009.

[28] D. DeTone, T. Malisiewicz, and A. Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2018.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

[30] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera re-localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[31] J. Dong and S. Soatto. Domain-Size Pooling in Local Descriptors: DSP-SIFT. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[32] 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.

[33] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 2015.

[34] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2009.

[35] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick Van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[36] 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.

[37] David A Forsyth and Jean Ponce. Computer vision: a modern approach. In *Prentice Hall Professional Technical Reference*, 2002.

[38] Junfeng Gao, Yong Yang, Pan Lin, and Dong Sun Park. Computer vision in healthcare applications. *Hindawi*.

[39] Paul Gao, Hans-Werner Kaas, and Det Mohr. Automotive revolution–perspective towards 2030 how the convergence of disruptive technology-driven trends could transform the auto industry. 2016.

[40] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[41] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision (ECCV)*, 2014.

[42] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision (ECCV)*, 2016.

[43] Albert Gordo, Jon Almazan, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision (IJCV)*, 2017.

[44] Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[45] Kai Han, Rafael Rezende, Bumsub Ham, Kwan-Yee Wong, Minsu Cho, Cordelia Schmid, and Jean Ponce. Scnet: Learning semantic correspondence. In *IEEE Conference on Computer Vision (ICCV)*, 2017.

[46] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A.C. Berg. MatchNet: Unifying Feature and Metric Learning for Patch-Based Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[47] Richard Hartley, Khurrum Aftab, and Jochen Trumpf. L1 rotation averaging using the Weiszfeld algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

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

[49] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Pedestrian detection: The elephant in the room. *arXiv preprint arXiv:2003.08799*, 2020.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[51] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[52] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. In *European Conference on Computer Vision (ECCV)*, 2020.

[53] A. Irschara, C. Zach, J. 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.

[54] M. Jahrer, M. Grabner, and H. Bischof. Learned local descriptors for recognition and matching. In *Computer Vision Winter Workshop (CVWW)*, 2008.

[55] A. Kendall and R. Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[56] 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.

[57] A. Kendall, M. Grimes, and R. Cipolla. Convolutional networks for real-time 6-DOF camera relocalization. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2015.

[58] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[59] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[60] 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)*, 2011.

[61] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[62] Grzegorz Kurzejamski, Jacek Komorowski, Lukasz Dabala, Konrad Czarnota, Simon Lynen, and Tomasz Trzcinski. Superncn: Neighbourhood consensus network for robust outdoor scenes matching. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*.

[63] Anthony R Lanfranco, Andres E Castellanos, Jaydev P Desai, and William C Meyers. Robotic surgery: a current perspective. In *Lippincott, Williams, and Wilkins*, 2004.

[64] Mans Larsson, Erik Stenborg, Carl Toft, Lars Hammarstrand, Torsten Sattler, and Fredrik Kahl. Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[65] Zakaria Laskar, Sami Huttunen, Daniel Herrera, Esa Rahtu, and Juho Kannala. Robust loop closures for scene reconstruction by combining odometry and visual correspondences. In *IEEE International Conference on Image Processing (ICIP)*, 2016.

[66] Zakaria Laskar and Juho Kannala. Context aware query image representation for particular object retrieval. In *Scandinavian Conference on Image Analysis (SCIA)*, 2017.

[67] Zakaria Laskar and Juho Kannala. Data-efficient ranking distillation for image retrieval. In *Asian Conference on Computer Vision (ACCV)*, 2020.

[68] Zakaria Laskar, Iaroslav Melekhov, Hamed R Tavakoli, Juha Ylioinas, and Juho Kannala. Geometric image correspondence verification by dense pixel matching. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020.

[69] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

[70] V. Lepetit and P. Fua. Keypoint Recognition Using Randomized Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2006.

[71] X. Li, J. Ylioinas, and J. Kannala. Full-Frame Scene Coordinate Regression for Image-Based Localization. In *Proceedings of Robotics: Science and Systems*, 2018.

[72] 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 Workshop (ECCVW)*, 2018.

[73] Xiaotian Li, Shuzhe Wang, Yi Zhao, Jakob Verbeek, and Juho Kannala. Hierarchical scene coordinate classification and regression for visual localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[74] Y. Li, N. Snavely, and D. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *European Conference on Computer Vision (ECCV)*, 2010.

[75] Konstantinos-Nektarios Lianos, Johannes L Schonberger, Marc Pollefeys, and Torsten Sattler. Vso: Visual semantic odometry. In *European Conference on Computer Vision (ECCV)*, 2018.

[76] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T Freeman. Sift flow: Dense correspondence across different scenes. In *European Conference on Computer Vision (ECCV)*, 2008.

[77] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[78] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV*.

[79] David Marr. Vision: A computational investigation into the human representation and processing of visual information. 1982.

[80] Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. Dgc-net: Dense geometric correspondence network. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[81] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Image-based localization using hourglass networks. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[82] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS*, 2017.

[83] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2005.

[84] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[85] A. Mishchuk, D. Mishkin, F. Radenović, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2017.

[86] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[87] Hans P Moravec. Rover visual obstacle avoidance. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.

[88] Michael C Mozer and Mark Sitton. Computational modeling of spatial attention. *Attention*, 9:341–393.

[89] I.F. of Robotics. Executive summary world robotics 2018 industrial robots. In *Available online on http://www.ifr.org*, pages 13–22, 2018.

[90] C. Osendorfer, J. Bayer, S. Urban, and P. van der Smagt. Convolutional Neural Networks Learn Compact Local Image Descriptors. In *International Conference on Neural Information Processing*, 2013.

[91] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[92] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[93] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[94] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision (ECCV)*, 2016.

[95] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 2018.

[96] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[97] Ali Sharif Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *arXiv preprint arXiv:1412.6574*, 2014.

[98] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[99] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[100] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[101] I. Rocco, R. Arandjelović, and J. Sivic. End-to-end weakly-supervised semantic alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[102] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[103] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[104] 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.

[105] 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.

[106] 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)*, 2017.

[107] 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.

[108] 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.

[109] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *British Machine Vision Conference (BMVC)*, 2012.

[110] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé. Understanding the Limitations of CNN-based Absolute Camera Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[111] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Springer, 2002.

[112] Johannes L. Schonberger, Marc Pollefeys, Andreas Geiger, and Torsten Sattler. Semantic visual localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[113] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[114] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, and Z. He. RF-Net: An End-to-End Image Matching Network based on Receptive Field. *arXiv:1906.00604*, 2019.

[115] Tianxin Shi, Shuhan Shen, Xiang Gao, and Lingjie Zhu. Visual localization using sparse semantic 3d map. In *IEEE International Conference on Image Processing (ICIP)*, 2019.

[116] 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)*, 2013.

[117] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 2018.

[118] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[119] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative Learning of Deep Convolutional Feature Point Descriptors. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[120] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning Local Feature Descriptors Using Convex Optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014.

[121] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2016.

[122] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision (ICCV)*, 2003.

[123] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[124] Richard Szeliski. *Computer vision: algorithms and applications*. 2010.

[125] 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.

[126] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[127] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. Probabilistic robotics. In *The MIT Press*, 2005.

[128] Y. Tian, X. Yu, B. Fan, and V. Balntas F. Wu, H. Heijnen. SOSNet: Second Order Similarity Regularization for Local Descriptor Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[129] E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(5):815–830, 2010.

[130] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[131] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *International Conference on Learning Representations (ICLR)*, 2016.

[132] 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.

[133] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[134] D. Ponsa V. Balntas, E. Riba and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference (BMVC)*, 2016.

[135] Abhinav Valada, Noha Radwan, and Wolfram Burgard. Deep auxiliary learning for visual localization and odometry. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[136] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[137] Julien Valentin, Matthias Nießner, Jamie Shotton, Andrew Fitzgibbon, Shahram Izadi, and Philip H. S. Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[138] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 2020.

[139] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[140] Sara Ventura, Rosa M Baños, and Cristina Botella. Virtual and augmented reality: New frontiers for clinical psychology. In *State of the Art Virtual Reality and Augmented Reality Knowhow*, pages 99–118. InTech, 2018.

[141] 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.

[142] T. Weyand, I. Kostrikov, and J. Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In *European Conference on Computer Vision (ECCV)*, 2016.

[143] S. Winder and M. Brown. Learning Local Image Descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[144] Xiaolong Wu, Stéphanie Aravecchia, Philipp Lottes, Cyrill Stachniss, and Cédric Pradalier. Robotic weed control using automated weed and crop classification. Wiley Online Library, 2020.

[145] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[146] Hongsheng Yang, Wen-Yan Lin, and Jiangbo Lu. Daisy filter flow: A generalized discrete approach to dense correspondences. In *IEEE International Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2014.

[147] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. Sanet: Scene agnostic network for camera localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[148] K.M. Yi, E. Trulls, V. Lepetit, and P. Fua. LIFT: Learned Invariant Feature Transform. In *European Conference on Computer Vision (ECCV)*, 2016.

[149] S. Zagoruyko and N. Komodakis. Learning to Compare Image Patches via Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[150] Xingyi Zhou, Arjun Karpur, Chuang Gan, Linjie Luo, and Qixing Huang. Unsupervised domain adaptation for 3d keypoint estimation via view consistency. In *European Conference on Computer Vision (ECCV)*, 2018.

Aalto-DD 191/2020

ISBN 978-952-64-0145-4 (printed)
ISBN 978-952-64-0146-1 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS