Department of Computer Science

# Deep Learning Methods for Image Matching and Camera Relocalization

Iaroslav Melekhov

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
**DOCTORAL DISSERTATIONS** 23/2020

# Deep Learning Methods for Image Matching and Camera Relocalization

**Iaroslav Melekhov**

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 at the lecture hall AS2 of the school on 21 February 2020 at 13.

**Aalto University**
**School of Science**
**Department of Computer Science**

**Supervising professor**

Prof. Juho Kannala, Aalto University, Finland and Prof. Esa Rahtu, Tampere University, Finland

**Preliminary examiners**

Prof. Atsuto Maki, KTH Royal Institute of Technology, Sweden
Dr. Vassileios Balntas, Scape Technologies, UK

**Opponent**

Prof. Victor Lempitsky, Skolkovo Institute of Science and Technology, Russia

Aalto University publication series
**DOCTORAL DISSERTATIONS** 23/2020

ISBN 978-952-60-8944-7 (printed)
ISBN 978-952-60-8945-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-8945-4

Unigrafia Oy
Helsinki 2020

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

**A!** **Aalto University**

## Abstract

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

**Author**
Iaroslav Melekhov

**Name of the doctoral dissertation**
Deep Learning Methods for Image Matching and Camera Relocalization

**Publisher** School of Science

**Unit** Department of Computer Science

**Series** Aalto University publication series DOCTORAL DISSERTATIONS 23/2020

**Field of research** Computer Science

**Manuscript submitted** 16 June 2019 **Date of the defence** 21 February 2020

**Permission for public defence granted (date)** 26 August 2019 **Language** English

□ **Monograph** ☒ **Article dissertation** □ **Essay dissertation**

**Abstract**

Deep learning and convolutional neural networks have revolutionized computer vision and become a dominant tool in many applications, such as image classification, semantic segmentation, object recognition, and image retrieval. Their strength lies in the ability to learn an efficient representation of images that makes a subsequent learning task easier. This thesis presents deep learning approaches for a number of fundamental computer vision problems that are closely related to each other; image matching, image-based localization, ego-motion estimation, and scene understanding.

In image matching, the thesis studies two methods utilizing a Siamese network architecture for learning both patch-level and image-level descriptors for measuring similarity using Euclidean distance. Next, it introduces a coarse-to-fine CNN-based approach for dense pixel correspondence estimation that can leverage the advantages of optical flow methods and extend them to the case of wide baseline between two images. The method demonstrates good generalization performance and it is applicable for image matching as well as for image alignment and relative camera pose estimation.

One of the contributions of the thesis is a novel approach for recovering the absolute camera pose from ego-motion. In contrast to the existing CNN-based localization algorithms, the proposed method can be directly applied to scenes which are not available at training stage and it does not require scene-specific training of the network, thus, improving the scalability. The thesis also shows that Siamese architecture can be successfully utilized in the problem of relative camera pose estimation achieving better performance in challenging scenarios compared to traditional image descriptors.

Lastly, the thesis demonstrates how the advances of visual geometry can help to efficiently learn depth, camera ego-motion, and optical flow for the task of scene understanding. More specifically, it introduces a method that can leverage temporally consistent geometric priors between frames of monocular video sequences and jointly estimate ego-motion and depth maps in a self-supervised manner.

**Keywords** computer vision, machine learning, deep learning, camera relocalization, image matching, scene understanding, ego-motion, image alignment

**ISBN (printed)** 978-952-60-8944-7 **ISBN (pdf)** 978-952-60-8945-4

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

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

**Pages** 145 **urn** http://urn.fi/URN:ISBN:978-952-60-8945-4

# Preface

The work presented in this thesis has been accomplished between the years 2016 and 2019 at the Department of Computer Science at Aalto University, and during memorable research internships at ETH Zürich (from November 2017 to March 2018) and Cambridge University tech spin-off Wayve (from October 2018 to March 2019).

Firstly, I would like to express my gratitude to my supervisors Prof. Juho Kannala and Prof. Esa Rahtu who have been great sources of brilliant ideas and advice through all these years. Without their support and encouragement I would have never finished such a long journey. Thank you for your guidance, patience and for the wonderful opportunities you have opened for me.

I am grateful to the reviewers of the thesis, Prof. Atsuto Maki from KTH and Dr. Vassileios Balntas from Scape Technologies, for their valuable comments and great feedback.

I have also had the pleasure to work with many talented researchers at the Department of Computer Science, that I have been part of. I would like to thank Zakaria Laskar with whom I worked particularly closely. I would like to extend my special thank you to a number of colleagues and friends. In alphabetical order: Rinu Boney, Santiago Cortes, Yuxin Hou, Xiaotian Li, Luiza Sayfullina, Arno Solin, Hamed R. Tavakoli, Juha Ylioinas. Thank you for various discussions during lunch, support, and for the great atmosphere in our research lab. I am also thankful to Hannakaisa Aikio from Oulu where my journey to computer vision began.

During my doctoral study I had an opportunity to collaborate with wonderful researchers from Zürich and Cambridge. I am grateful to Prof. Torsten Sattler for hospitality, insightful and illuminating discussions during my stay at ETH Zürich. I would also like to express my gratitude to Dr. Alex Kendall and all the folks from Wayve providing me a chance to be involved in the exciting and fascinating world of autonomous driving.

The generous financial support provided by Otto A. Malm Foundation, the Helsinki Institute for Information Technology (HIIT), the Nokia Foundation, and the Kaute Foundation is gratefully acknowledged.

I would like to thank my friends, Priscilla and Tobias for their positive attitude, jokes, good times, keeping me alive and making my life much colorful. I further wish to thank Marina and Aleksei for being by my side and helping and encouraging me.

Finally, I am mostly grateful to my parents Inna and Andrei for all the support, love, and motivation over the years.

Helsinki, January 20, 2020,

Iaroslav Melekhov

# Contents

**Preface** **1**

**Contents** **3**

**List of Publications** **5**

**Author's Contribution** **7**

**Abbreviations** **9**

**1. Introduction** **11**
1.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11
1.2 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . 12
1.3 Outline of the thesis . . . . . . . . . . . . . . . . . . . . . . 13

**2. Image Matching** **15**
2.1 Related work . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2.2 Matching based on learned descriptors . . . . . . . . . . . . 16
2.3 Dense pixel correspondence estimation . . . . . . . . . . . . 19
2.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21

**3. Camera Relocalization** **23**
3.1 Relative camera pose estimation . . . . . . . . . . . . . . . . 24
3.2 Image-based localization . . . . . . . . . . . . . . . . . . . . 25
3.3 An encoder-decoder network for camera relocalization . . 28
3.4 Recovering camera pose by image retrieval and ego-motion 28
3.5 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

**4. Geometric Scene Understanding** **31**
4.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
4.2 Visual scene understanding approaches . . . . . . . . . . . 32
4.3 On temporally consistent geometric prior . . . . . . . . . . 35
4.4 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36

**5. Summary of the Original Articles** **39**

**6. Conclusion** **43**

**Bibliography** **45**

**Publications** **55**

# 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** Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Image Patch Matching Using Convolutional Descriptors with Euclidean Distance. *Asian Conference on Computer Vision. Workshop on Interpretation and Visualization of Deep Neural Nets (ACCVW)*, pp. 638–653, 2016.

**II** Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese Network Features for Image Matching. *International Conference on Pattern Recognition (ICPR)*, pp. 378–383, December 2016.

**III** Iaroslav Melekhov, Aleksei Tiulpin, Torsten Sattler, Marc Pollefeys, Esa Rahtu, and Juho Kannala. DGC-Net: Dense Geometric Correspondence Network. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1034–1042, January 2019.

**IV** Iaroslav Melekhov, Juha Ylionas, Juho Kannala, and Esa Rahtu. Relative Camera Pose Estimation Using Convolutional Neural Networks. *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pp. 675–687, 2017.

**V** Iaroslav Melekhov, Juha Ylionas, Juho Kannala, and Esa Rahtu. Image-based Localization Using Hourglass Networks. *IEEE International Conference on Computer Vision. Geometry Meets Deep Learning Workshop (ICCVW)*, pp. 879–886, 2017.

**VI** Zakaria Laskar, Iaroslav Melekhov, Surya Kalia, and Juho Kannala. Camera Relocalization by Computing Pairwise Relative Poses Using Convolutional Neural Networks. *IEEE International Conference on Computer*

*Vision. Geometry Meets Deep Learning Workshop (ICCVW)*, pp. 929–938, 2017.

**VII** Iaroslav Melekhov, Esa Rahtu, Juho Kannala, Alex Kendall. TC-Net: Self-Supervised Monocular Video Scene Understanding Using Temporally Consistent Geometric Prior. *International Conference on Machine Learning. Self-Supervised Learning Workshop (ICMLW)*, 5 pages, April 2019.

# Author's Contribution

### Publication I: "Image Patch Matching Using Convolutional Descriptors with Euclidean Distance"

The topic was proposed by Kannala and Rahtu but Melekhov designed and conducted the experiments. He also had the main responsibility in writing the article while Kannala and Rahtu reviewed and proposed suggestions to the manuscript.

### Publication II: "Siamese Network Features for Image Matching"

Melekhov had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and evaluated the experiments.

### Publication III: "DGC-Net: Dense Geometric Correspondence Network"

Melekhov had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and conducted the experiments. The co-authors reviewed and proposed suggestions to the manuscript.

### Publication IV: "Relative Camera Pose Estimation Using Convolutional Neural Networks"

Melekhov had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and conducted the experiments. The co-authors gave valuable suggestions and comments during the discussions about the text, results and methods.

### Publication V: "Image-based Localization Using Hourglass Networks"

Melekhov had the main responsibility in writing the article. He also implemented all the models and methods used in the article as well as designed and conducted the experiments.

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

Laskar and Melekhov have equally contributed to the paper. Melekhov has the responsibility in writing the article, designed and implemented Deep Learning based models and conducted the experiments.

### Publication VII: "TC-Net: Self-Supervised Monocular Video Scene Understanding Using Temporally Consistent Geometric Prior"

The topic was proposed by Kendall and Melekhov. Melekhov had the main responsibility in writing the article, conducting the experiments and models implementation. Rahtu and Kannala gave valuable comments and suggestions to improve the article.

# Abbreviations

**ATE** Absolute Trajectory Error

**CNNs** Convolutional Neural Networks

**DGC-Net** Dense Geometric Correspondence Network

**DoG** Difference of Gaussian

**ICP** Iterative Closest Point

**LiDAR** Light Detection and Ranging

**NN** Nearest Neighbours

**PCA** Principal Component Analysis

**P*n*P** Perspective-$n$-Point

**RNNs** Recurrent Neural Networks

**SLAM** Simultaneous Localization and Mapping

**SfM** Structure from Motion

# 1. Introduction

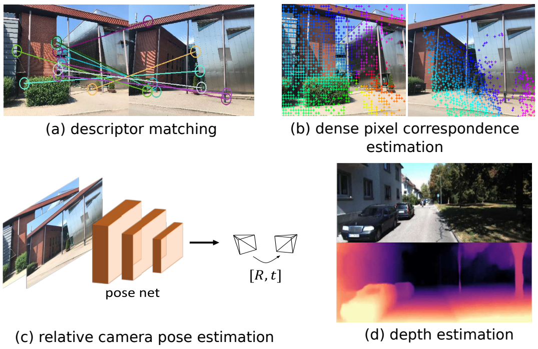
## 1.1 Motivation

The main goal of computer vision is to extract useful information from images. This has proved a surprisingly challenging task and it has occupied thousands of intelligent and creative minds all over the world. There has been remarkable recent progress in our understanding of computer vision, and the last decade has seen the first large scale deployments of consumer computer vision technology. Nowadays, computer vision systems are used in various applications in different fields, such as robot navigation, video game industry, human-computer interaction and medical imaging.

Computer vision is a very broad and multidisciplinary science and this thesis covers only a small part of it. It aims to contribute to the knowledge and to provide practical methods for closely related computer vision problems, such as image matching, estimation of absolute and relative camera pose, and scene understanding. Some of the topics presented in this thesis are illustrated in Figure 1.1. The proposed methods introduced in the thesis utilize only images taken by RGB cameras. High-quality cameras are affordable and probably most pervasive sensors available to us. For example, autonomous driving industry has been using very accurate, powerful and expensive hardware, such as LiDAR, for decades but it is still very far from full autonomy. According to recent news[1], the focus is shifting towards camera sensors providing the information about vast variety of situations that might be encountered in the wild. Thus, implementing algorithms which can process the data coming from the cameras in an efficient and reliable way is vital. This aim is very challenging, since there are quite many situations where the methods based on traditional techniques fail (see Figure 1.1).

In this thesis, the introduced algorithms rely on deep learning [34, 43] which has emerged as a powerful paradigm for understanding high di-

[1] https://techcrunch.com/2019/04/22/anyone-relying-on-lidar-is-doomed-elon-musk-says/



(a) descriptor matching

(b) dense pixel correspondence estimation

(c) relative camera pose estimation

(d) depth estimation

**Figure 1.1.** Some topics we cover in this thesis: 1) local descriptor matching; 2) dense pixel correspondence estimation; 3) relative camera pose estimation; 4) depth estimation (scene understanding). The results illustrated for descriptor matching and dense pixel correspondence estimation tasks have been obtained by using ORB descriptor [89] and DeepMatching [83]. As can be clearly seen, the accuracy is not perfect for this particular case, since local descriptors (pixel locations) with similar color should be matched in two images. We address these challenges in Chapter 2.

mensional data, such as visual imagery. Deep learning approaches have revolutionized many computer vision applications, such as pattern recognition, image retrieval [6, 95, 121], semantic segmentation [75, 76], object recognition, and image-based localization [53, 54, 114] achieving remarkable results. In addition to deep learning approaches, this thesis also demonstrates how geometric aspects can be successfully utilized in unsupervised learning providing additional form of supervision.

## 1.2 Contributions

All publications, software and project codes developed during this PhD are available in open access. The main contributions of the thesis are listed below:

- Two methods for image matching are introduced. The proposed end-to-end learned CNN-based image descriptors utilize a Siamese network structure trained in a supervised manner.

- A coarse-to-fine method for dense pixel correspondence estimation is presented (Publication III). Specifically, given a pair of images, the proposed approach, named DGC-Net, predicts dense and subpixel accurate estimates and can handle strong geometric transformations between two views. Along with image matching, DGC-Net is also applicable for image alignment, relative camera pose estimation, and for correspondence verification in image retrieval [60]. These topics are covered in the thesis.

The source code, pre-trained models and evaluation protocol of DGC-Net are available at `https://github.com/AaltoVision/DGC-Net`

- A CNN-based method for relative camera pose estimation is proposed (Publication IV). It leverages the idea of Siamese network structure, which has been described earlier, and can directly estimate relative camera pose from a pair of input images utilizing transfer learning from a large scale classification dataset. The source code, pre-trained models and evaluation protocol of the proposed approach are available at `https://github.com/AaltoVision/relativeCameraPose`

- Two methods related to image-based localization are presented. First, we propose a novel architecture (Publication V) consisting of a chain of convolution and upconvolution layers followed by a regression part which directly predicts absolute camera pose for a given image. The second camera relocalization approach (Publication VI) is based on the idea of relative camera pose estimation and image retrieval. Its source code, pre-trained models and evaluation protocol are available at `https://github.com/AaltoVision/camera-relocalisation`

- A CNN-based framework for scene understanding is introduced. The proposed approach, named TC-Net (Publication VII), can jointly learn depth maps and relative camera pose by exploiting temporal consistency between image frames throughout longer unlabeled video sequences. The source code, pre-trained models and evaluation protocol of TC-Net are available at `https://github.com/imelekhov/selfs_depth`

## 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 previous work on image matching and image descriptors and introduces end-to-end deep learning methods related to those tasks. Chapter 3 concentrates on image-based localization and relative camera pose estimation methods. Chapter 4 studies different CNN-based methods for the problem of scene understanding and introduces a self-supervised approach which can efficiently learn depth maps and ego-motion from monocular video sequences. 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. Image Matching

This chapter focuses on the three closely related computer vision problems, namely, matching of images and local image patches and pixel correspondence estimation. Section 2.1 provides a brief review of the prior art. Section 2.2 introduces formulations for end-to-end deep learning architectures for image patch matching presented in Publication I and Publication II. Finally, Section 2.3 summarizes the innovations of Publication III regarding the problem of dense pixel correspondence estimation. The main themes of the chapter are summarized in Section 2.4.

## 2.1 Related work

Finding correspondences between local image patches extracted from different views is a key component of many computer vision applications. For example, structure-from-motion (SfM), multi-view reconstruction, image retrieval, simultaneous localization and mapping (SLAM), object recognition and tracking require accurate computation of local image similarity. Due to importance of these problems various descriptors have been proposed for patch matching with the aim to represent distinctive image patches to be invariant under challenging viewing conditions. The local image descriptors can be broadly categorized into categories: **hand-crafted** descriptors which are designed using some prior knowledge and do not involve optimization procedure; and end-to-end **learned** descriptors which explicitly learn patch similarity from the data without manually designed features.

**Hand-crafted descriptors** The problem of deciding whether two image patches are similar or not is quite challenging especially in real world exhibiting various challenges, such as occlusions, illumination changes, changes in viewpoint, etc. The pivotal moment in patch matching was the introduction of SIFT [67] consisting of histograms of the aggregated image gradients characterizing the appearance of keypoints. However, it may not take into account all of the aforementioned factors in an optimal manner [50]. To address this limitation, real-valued [7, 12, 26, 70, 109] and binary [4, 5, 15, 62, 89] patch descriptors have been proposed.

**Learned descriptors** Descriptor learning can be formulated as finding a discriminative representation of a given image in a new subspace. Ke *et al.* [51] and Bursuc *et al.* [14] propose to utilize principal component analysis (PCA) to obtain the embeddings of SIFT and rootSIFT [7] features, respectively, while Lepetit *et al.* [61] embed the image patches by applying random forest. Simonyan *et al.* [103] demonstrate that descriptor learning can be formulated as a convex optimization problem achieving remarkable results in the local image patches benchmark. Rather than using Euclidean space, Calonder *et al.* [15] propose to utilize Hamming subspace which leads to a more computationally efficient method.

The pioneer works which utilized CNN-based representations for finding matching image patches were [47] and [78]. More recently, [39, 102, 123] propose CNN descriptors trained with two-branch architecture which significantly exceed the accuracy of manually engineered descriptors. However, in contrast to SIFT, in [39, 123] the feature representations of input patches are compared by a set of fully connected layers (match network) that learns a complex comparison metric. Nevertheless, Zagoruyko *et al.* [123] and Simo-Serra *et al.* [102] also conducted experiments in which the match network was replaced with Euclidean distance metric between the outputs of two branches and, hence, they can be directly compared to SIFT. In contrast to [102, 123], later work by Balntas *et al.* [112] applied a triplet distance objective loss function with random sampling of patch triplets. However, randomly sampled negative patch pairs can be easily separated from the positive ones [102], thus, it leads to inefficient training due to vanishing gradients. To address this problem, Mishchuk *et al.* [71] propose a novel loss for metric learning and carefully sample positive and negative patch pairs in the input mini-batch based on a distance matrix. Tian *et al.* [108] extended [112] and utilized second order similarity to learn local descriptors. Very recently, several deep learning methods have been proposed [28, 99, 121] to *jointly* learn the descriptor and detector and they demonstrate promising results in image matching and localization.

In the following sections, an overview of our methods for image matching proposed in Publication I, Publication II, and Publication III is provided. First, end-to-end CNN-based methods for image matching based on learned descriptors are presented. Then, the chapter introduces a deep learning architecture for task of dense pixel correspondence estimation.

## 2.2 Matching based on learned descriptors

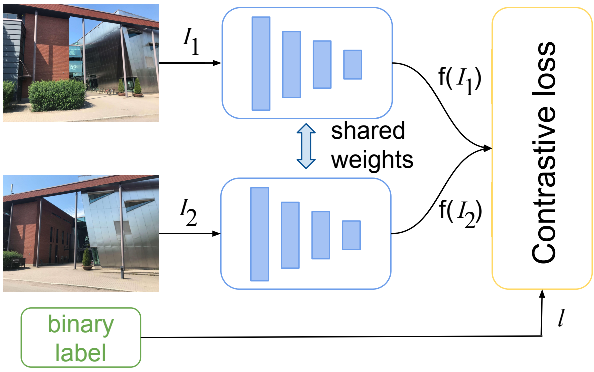


**Figure 2.1.** Schematic illustration of the learned descriptor based on a Siamese network structure. A pair of images $(I_1, I_2)$ is propagated through the network consisting of two identical branches and sharing the same set of weights. Feature representations of patches $(f(I_1), f(I_2))$ are extracted from the last layer of each branch separately and Euclidean distance is computed between them. Our objective is to learn a descriptor that minimizes the distance between similar pairs of images and maximizes it for dissimilar pairs. It is important to note that at test time (*i.e.* after learning) the feature descriptor $f$ can be computed independently for each individual patch since both branches are identical. See Section 2.1 for more details.

**Image patch matching** Similarly to [102, 123], in Publication I, we develop a supervised fully-convolutional CNN-based learned descriptor maps the raw input patch to a low dimensional feature vector so that the distance between representations is small for similar image patches and large otherwise. The model consists of two identical convolutional branches that share the same set of weights and parameters. Patches $P_1$ and $P_2$ are fed into branches and propagated through the model separately. The proposed method has been optimized by minimizing margin-based contrastive loss function [38, 102, 123] defined as follows (see Figure 2.1):

$$\mathscr{L}(P_1, P_2, l) = \frac{1}{2} l D^2 + \frac{1}{2}(1-l)\left\{max\left(0, m - D^2\right)\right\}, \tag{2.1}$$

where $l$ is a binary label which selects whether the input pair consisting of patch $P_1$ and $P_2$ is a positive ($l = 1$) or negative ($l = 0$), $m > 0$ is the margin for negative pairs and $D = \|f(P_1) - f(P_2)\|$ is the Euclidean distance between feature vectors $f(P_1)$ and $f(P_2)$ of input images $P_1$ and $P_2$. Since the distance metric is $L_2$ norm, the proposed approach can be considered as a direct replacement of SIFT descriptor.

In order to train and evaluate the proposed approach, the Photo Tourism patches dataset introduced by [119] has been used. The dataset consists of 1.5M grayscale patches with ground-truth positive and negative patch pairs extracted from images of the Statue of Liberty, Notredame and Yosemite by using Difference of Gaussian (DoG) interest point detector and matched by utilizing 3-D multi-view reconstruction. Pairs of patches corresponding to the same 3-D point are defined to be matching if they also originate from DoG interest points detected with sufficiently similar scale and orientation [119]. Pairs of patches sampled from different 3-D points are non-matching. Since many raw patches of the Photo Tourism dataset exhibit significantly different contrast, we apply histogram equalization to enhance the intensity and demonstrate (Publication I) that it helps to improve performance metric. In addition, based on the visualization of the false negative pairs of the dataset, we investigate whether the proposed descriptor could be made more robust to spatial misalignment by utilizing spatial transformer networks introduced by Jaderberg *et al.* [46]. Specifically, the spatial transformer is a differentiable module performing explicit spatial transformations of input feature maps. The spatial transformer explicitly estimates geometric transformations of the input image pair parameterized as affine transformation matrix. Since it has been placed at the beginning of the proposed descriptor architecture, it can directly transform the preprocessed input image patches. Since we aim to compensate errors caused by rotation, translation, and scaling, the number of parameters estimated by the spatial transformer equals four. As it is shown in Publication I, utilizing the spatial transformer module leads to better results compared to the original descriptor. We refer the interested reader to Publication I for more details about the descriptor architecture and evaluation procedure.

**Whole-image matching** In Publication II, the ideas proposed in Publication I have been extended to finding matching and non-matching pairs of images across large database of landmarks. Specifically, a Siamese neural architecture is used to explicitly learn the whole-image similarity measure. A structure-from-motion dataset proposed by Cao *et al.* [16] is utilized for the experiments. The dataset represents five crowd-sourced image collections, each corresponding to a popular landmark (London Eye, Tate Modern, San Marco, Times Square, and Trafalgar). We evaluated different types of weight initialization of the network branches and demonstrated that the proposed approach has promising results of generalization on unseen landmark datasets. Similarly to [9], the results could be improved further by fine-tuning the network on the data having similar distribution as the evaluation dataset.

Such learned image representations proposed in Publication II are well suited for the image retrieval task. Instance-level image retrieval is a computer vision problem that aims to retrieve all images that contain the same object instance as query from a potentially large database of images. There have been quite many works investigating the possibility of using CNN-based features for image retrieval. These include methods utilizing off-the-shelf CNN features [9, 82], finetuning CNN features from models trained for classification task on an external set of Landmark images [9, 35]; explicitly learn a ranking loss [6, 35] rather than the classification objective function [9].

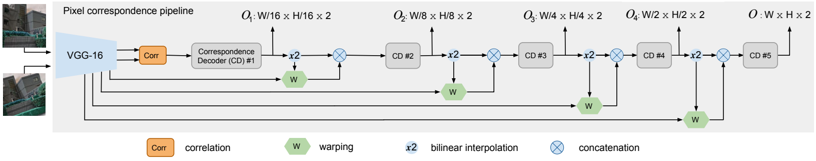


**Figure 2.2.** Overview of the proposed iterative architecture `DGC-Net` for dense pixel correspondence estimation consisting of four major components: 1) the **feature pyramid creator**. 2) the **correlation layer** estimates the pairwise similarity score of the source and target feature descriptors. 3) the fully convolutional **correspondence map decoders** predict the dense correspondence map between input image pair at each level of the feature pyramid. 4) the **warping layer** warps features of the source image using the upsampled transforming grid from a correspondence map decoder.

## 2.3 Dense pixel correspondence estimation

Finding similar and dissimilar image pairs can be accomplished based on the number of pixel-level correspondences between two views. In addition to the application of image matching, finding correspondences between images is a fundamental problem in many computer vision tasks, such as image retrieval, visual localization, image alignment, and relative camera pose estimation (Chapter 3). In general, one way to establish a pixel-wise correspondence field between images is based on applying feature descriptors to an image pair and utilizing nearest neighbour criterion to match keypoints globally. However, these approaches do not produce dense correspondences explicitly and apply interpolation or local affine transformations [66] to turn a sparse set into a pixel-wise correspondences.

The problem of dense pixel correspondence estimation is closely related to optical flow estimation task where CNN-based approaches have recently achieved remarkable results [27, 44, 48, 80, 106]. While optical flow methods produce very accurate results for the small pixel translation and limited appearance variation scenarios, they hardly deal with the strong geometric transformations that we consider in Publication III. Rocco *et al*. [84, 85] proposed a CNN-based approach for determining correspondences between two images and applying it to instance-level and category-level tasks. In contrast to optical flow methods, [84] comprises a matching layer calculating global correlation between target and reference feature maps without any spatial constraint. The method predicts either affine or thin plate splines (TPS) based geometric transformations. In later work [86], Rocco *et al*.propose locally and globally constrained matching network on top of the global correlation layer which leads to improvement in instance and semantic matching. In contrast to [84, 85], Publication III proposes a more general approach handling more diverse transformations, learning dense pixel correspondences between a pair of images and operating in an end-to-end fashion.

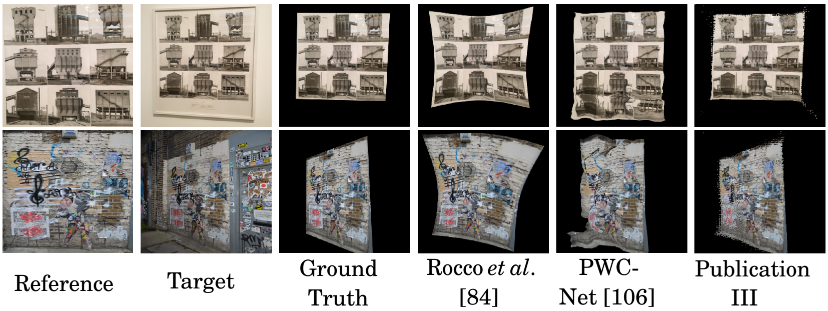


**Figure 2.3.** Qualitative image alignment results produced by different algorithms on the HPatches [11] dataset. The reference image is warped (aligned) using pixel correspondence estimates predicted by DGC-Net. The proposed model produces more accurate correspondence map leading to better image alignment.

The architecture of the proposed method, named DGC-Net, is schematically illustrated in Figure 2.2. A pair of input images is fed into a module (the feature pyramid creator) consisting of two pre-trained CNN branches which construct a feature pyramid. The correlation layer takes feature maps of the source and target images from the top (coarse) level of the pyramid and computes the pairwise similarity between them. Then, the correspondence map decoder (CD) takes the output of the correlation layer and directly predicts pixel correspondences for this particular layer of the pyramid. The estimates are then refined in an iterative manner by using a chain of correspondence map decoders. Such a coarse-to-fine architecture is influenced by optical flow methods [80, 106] where a hierarchical feature representation is utilized to refine the estimates of pixel displacements between two views. In contrast to [106], where the correlation volume is computed for the raw features in a restricted area around the center pixel, we compute global correlation and apply L2-normalization before and after the correlation layer to strongly down-weight ambiguous matches (see Figure 2.2).

DGC-Net is trained in a weakly-supervised manner using synthetic geometric transformations obtained by applying random homography [24] to the Pascal VOC 2011 [30] and Tokyo Time Machine [6] datasets. We have evaluated DGC-Net on the HPatches [11] dataset consisting of 59 outdoor image sequences and exhibiting varying photometric and geometric changes. Some examples of HPatches are illustrated in Figure 2.3. The proposed method achieves significantly better image matching performance compared to optical flow [44, 106] and dense pixel correspondence estimation [84] approaches. The corresponding values of Percentage of Correct Keypoints (PCK) metric are provided in Publication III. In addition to image matching, we demonstrate the use of DGC-Net to complicated computer vision tasks such as image alignment and relative camera pose estimation (Chapter 3). Figure 2.3 shows qualitative results for image alignment produced by DGC-Net and strong baseline methods. The proposed method can handle drastic changes between two views where other baseline methods fail.

Our recent work [60] has extended DGC-Net to the problem of image retrieval and demonstrated that the combination of global image descriptors, such as NetVLAD [6] and geometrically verified pixel correspondences obtained by DGC-Net lead to state-of-the-art results on several challenging datasets, *i.e.* Tokyo 24/7 [110], InLoc [107] and Aachen Day-Night [93].

## 2.4 Discussion

This chapter focused on two closely related tasks, namely image matching and dense pixel correspondence estimation. The main conclusions within the chapter are briefly summarized below.

**Image matching.** End-to-end learned CNN-based image descriptors suitable for image matching have been introduced. The proposed descriptors encompass two identical fully-convolutional branches and learn a similarity measure by using a contrastive loss function (*cf.* Equation 2.1). The chapter showed that the proposed patch matching descriptor achieves better matching performance on the Photo Tourism dataset compared to the strong baseline methods.

**Dense pixel correspondence estimation.** A coarse-to-fine CNN model for dense pixel correspondence estimation has been proposed. The method, named DGC-Net, is leveraging the advantages of optical flow methods and extends them to the case of large geometric transformations between two input images providing dense and subpixel accurate estimates. It has been trained on synthetic transformations and demonstrates very good generalization performance to unseen, realistic, data. In addition to image matching application, we demonstrate that the estimates produced by DGC-Net can be used for the problem of camera relocalization (Chapter 3) achieving better results compared to SIFT descriptor.

# 3. Camera Relocalization

This chapter addresses two closely related problems; estimating the absolute and relative camera pose (position and orientation) in three dimensional space. The ability to estimate the absolute (relative) camera pose from the visual scene representation is essential for many computer vision applications such as structure-from-motion, visual scene understanding, SLAM, navigation of autonomous vehicles, and augmented reality. The terms *image-based localization*, *camera relocalization*, and *absolute camera pose estimation* are used interchangeably through this chapter.

Many conventional camera relocalization methods proposed in the literature [94, 96, 97, 98] are based on hand-crafted local image features, such as ORB [89], SIFT [67], and SURF [12]. Specifically, these methods require a 3-D point cloud model where each 3-D point is associated with its 2D projection in the image. The list of tentative 2D-3D matches is then geometrically verified by RANSAC [31]. The verified correspondences are utilized in P*n*P algorithm to recover the absolute pose. However, the local image features obtained by hand-crafted feature detectors and descriptors perform poorly under varying environmental and weather conditions (see Figure 3.1).

This chapter studies proposed CNN-based approaches for estimating the absolute and relative camera pose. Section 3.1 focuses on a framework for relative camera pose estimation between two views which is inspired by the advances of deep learning methods for image-based localization [52, 54, 114]. Our system takes RGB images from both cameras and directly regresses the relative camera pose (relative rotation and translation). The method is simple in the fact that it consists of two identical convolutional neural networks with shared weights trained in an end-to-end manner. The proposed approach is scalable and can predict the relative camera pose more robustly, compared with traditional methods based on hand-crafted descriptors [12, 67, 89] across different challenging conditions (large viewpoint changes, textureless surfaces, repetitive structures).

Utilizing CNNs in the image-based localization problem has been pioneered by Kendall *et al.* [54]. Their method, named PoseNet, casts camera

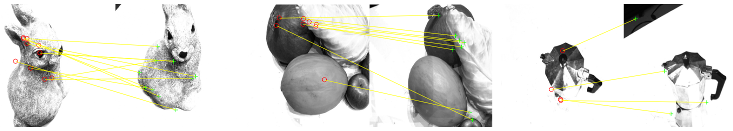


**Figure 3.1.** Scenarios where traditional approaches (SIFT) are not able to establish reliable 2D pixel correspondences. *Left*: very large viewpoint changes, thus most of inliers (correspondences) are not correct; *center*: the correct inliers concentrate on a small region; *right*: there is insufficient number of correspondences due to textureless scene (object with reflecting surface).

relocalization as a regression problem, where 6-DoF camera pose is directly predicted from a monocular image by leveraging transfer learning from very large scale classification datasets. Although PoseNet overcomes many limitations of the feature-based methods, its localization performance lacks behind traditional approaches in typical cases where local features perform well. Section 3.2 demonstrates how the advances of semantic segmentation methods can be utilized in the camera relocalization problem. The proposed novel architecture consisting of a symmetric "encoder-decoder" network structure allows to better collect information available in the input image and to add more context to the regression process which helps to improve localization performance.

Finally, Section 3.4 studies how to utilize relative camera pose estimates (Section 3.1) to recover the absolute pose. The main limitation of CNN-based methods directly regressing the absolute camera pose [52, 54, 114] is the fact that they have to be trained and evaluated scene-wise when the scenes are registered to different coordinate frames. This causes complications, especially if we need to localize the camera across several scenes simultaneously. The proposed approach can alleviate such issues and predict the camera pose for any scenes without re-training.

## 3.1 Relative camera pose estimation

This section describes the proposed CNN-based approach for relative camera pose estimation presented in Publication IV. Note, the terms *relative camera pose* and *ego-motion* are used interchangeably. Classical localization techniques that rely on different local image features [12, 15, 67, 89] have recently been accompanied by deep CNN based methods. Konda *et al*. [58] adopt a classification approach to the problem of relative camera pose estimation, where a shallow CNN architecture along with the softmax function are utilized to predict the relative transformation between two consecutive video frames using a pre-defined set of discretized velocities and orientation. Similarly, DeepVO proposed by Mohanty *et al*. [72] comprises two AlexNet-like CNN branches acting as inputs to a stack of fully connected layers coupled with a regression layer. Ummenhofer *et al*. [111] proposed a CNN architecture for depth and relative camera motion estimation. They utilized multiple tasks in the learning phase to provide additional supervision in order to get more accurate depth maps and camera motion estimates.

In Publication IV, we address the problem of relative camera pose estimation with deep learning. Compared with [58, 72], the proposed approach is applicable for general unrestricted camera motion and for wide baseline view pairs. We introduce a pipeline for training a Siamese-based CNN to regress a 7-dimensional relative camera pose vector $\Delta p$ containing the relative orientation vector $\Delta q$ (4-dimensional quaternion), and the relative position, *i.e.* translation vector $\Delta t$ (3-dimensional), so that $\Delta p = [\Delta q, \Delta t]$. Specifically, the proposed network takes a pair of RGB images optimizing the following objective function[1]:

$$\mathscr{L} = \left\|\Delta t - \Delta\hat{t}\right\| + \beta \left\|\Delta q - \frac{\Delta\hat{q}}{\|\Delta\hat{q}\|}\right\|, \tag{3.1}$$

where $\Delta\hat{q}$ and $\Delta\hat{t}$ are the ground-truth relative orientation and translation, respectively. In practice, the magnitude of $\Delta\hat{q}$ and $\Delta\hat{t}$ varies a lot making the optimization process intractable. Therefore, a constant weighting term $\beta$ is used to keep the estimated values to be nearly equal leading to an improved ego-motion estimate. Furthermore, the proposed model is complemented by a spatial pyramid pooling layer [42] achieving two objectives. First, that layer allows to robustly handle input images with different spatial resolution. Second, applying spatial pyramid pooling is the key to even more accurate relative pose estimations as it opens the door for larger images to be used during training which, as empirically demonstrated, improve the results without changing the network structure. The proposed approach is trained in a supervised manner. Structure-from-motion datasets [118] are utilized to automatically generate ground truth data. Unlike [111], the proposed approach does *not* require any additional supervisory signals such as depth maps for training which is beneficial in practice.

Later work has extended the proposed approach to jointly learn semantics and geometry (odometry and global pose) of the scene [3, 73, 113], a probabilistic sequence-to-sequence visual odometry framework [116], and topometric localization [77].

## 3.2 Image-based localization

Visual localization approaches can be broadly divided into two categories: appearance-based place recognition and image-based pose estimation meth-

[1] Note that the following loss function is valid for learning the absolute pose. In this case, we simply drop $\Delta$ from the equation.

ods.

**Appearance-based approaches** model localization as an image retrieval problem, where the unknown location of the query is estimated using the locations of the most visually similar database images [6, 18, 91, 110, 117]. This can be modelled by employing image retrieval techniques such as Bag-of-Words [104], or more compact representations such as VLAD or Fisher vectors [8, 49]. In order to handle large changes in appearance due to illumination (day/night) or change of seasons, Torii *et al.* [110] develop a view synthesis method that can create virtual views from novel viewpoints by using associated approximate depth maps to warp the original images. These synthetic images are then added to the database resulting in a higher localization performance. Deep learning methods have also shown to be very efficient to directly learn image descriptors suitable for image retrieval. Most of the deep learning methods focus on designing image representations by leveraging models pre-trained on large image classification datasets such as ImageNet [23]. A significant improvement could be achieved by finetuning pre-trained models for retrieval on localization datasets [6, 9] and using a ranking loss [6, 35].

Visual localization problem can be solved by casting it as a classification task. Specifically, the scene is divided into a number of individual places (landmarks) which are then used to train either a CNN-based or SVM classifier employing the bag-of-visual-words representation to identify which place is illustrated in a given image.

**Image-based pose estimation methods** are structure-based localization methods estimate the 6-DoF pose by using a 3-D scene model obtained from structure-from-motion. Specifically, they create a set of 2D-3D correspondences between 3-D points and local features extracted from a query image. Finally, the pose of the query image is established by applying RANSAC [31] loop in combination with a Perspective-n-Point algorithm [56]. The main limitation of this approach is the descriptor matching stage that turns out to be computationally demanding if the 3-D scene models are large such as those for large-scale urban environments. To address this issue, some methods [19, 65, 92] utilize prioritized search techniques which terminate the correspondence search procedure as soon as a certain amount of inliers has been achieved, while other approaches restrict matching only to the 3-D points visible in the top-retrieved database images [16, 45, 95].

Recently, it has been shown that machine learning based approaches have great potential to approach image-based localization problem. Kendall *et al.* [52, 53, 54] propose PoseNet – a CNN architecture which can directly regress the camera pose from an input RGB image. Later works have improved localization performance achieved by PoseNet. Walch *et al.* [114] applied LSTM units to enhance the context of features extracted from the

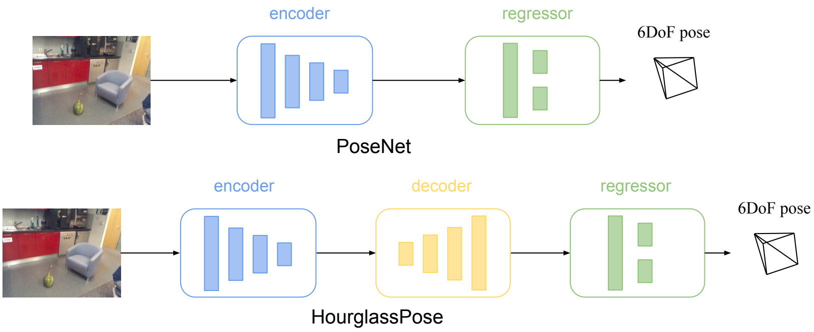


**Figure 3.2.** Overview of our proposed `Hourglass-Pose` architecture for image-based localization. Compared to PoseNet [54], we propose to use an additional decoder to recover fine-grained visual information from the input image to improve absolute camera pose estimates. Such a symmetric "encoder-decoder" network structure is also known as an hourglass architecture [76].

input image leading to significant improvement in localization. In turn, Clark *et al*. [20] utilized LSTMs to predict camera translation from short video sequences as input. The architecture encompasses a bidirectional recurrent neural network (RNN), which captures geometric dependencies between adjacent frames of the input sequence yielding refined accuracy of the global camera pose. Nevertheless, traditional absolute camera pose regression methods are less accurate than structure-based localization approaches due to lack of generalization from the training data [96]. However, recent work by Shi *et al*. [100] has demonstrated that combining structure-based method and image-based method with semantic information is beneficial and can outperform pure structure-based localization approaches such as Active Search [94]. Sarlin *et al*. [90] propose a hierarchical localization approach that simultaneously utilizes local image features using the SuperPoint [25] architecture and the global descriptor computed by a NetVLAD [6] layer for accurate 6-DoF localization.

Rather than directly regressing camera pose, Shotton *et al*. [101] apply a regression forest (SCoRF) to establish a set of 2D-3D correspondences inferred from an input RGB-D image and use RANSAC to estimate the pose of the image. Brachmann *et al*. [13] have extended the original SCoRF architecture to RGB-only images and proposed a new differentiable replacement of RANSAC leading to an end-to-end trainable CNN architecture and better localization performance. Li *et al*. [63, 64] have leveraged visual geometry cues and proposed angle-based reprojection loss function achieving improvement in localization compared to [13].

## 3.3 An encoder-decoder network for camera relocalization

In Publication V, similarly to [20, 52, 53, 114], we cast image-based localization as a regression problem and propose a CNN architecture for estimating camera pose from a single RGB-image. Seeking a potential way to improve localization performance, we adopt the ideas discovered in efforts solving the problems of image restoration [69], human pose estimation [75] and semantic segmentation [79, 87]. The proposed method, Hourglass-Pose (see Figure 3.2), which consists of a chain of convolution and upconvolution layers helps to add more context to the regression part process to better collect the overall information, from coarse structures to fine-grained object details, available in the input image. Specifically, the architecture has a bottom part (the encoder) that is used to encode the overall context and a latter part (the decoder) that recovers the fine-grained visual information by gradually increasing the size of the output feature map of the encoder towards the original resolution of the input image. The network is trained in an end-to-end manner by optimizing regression loss (see Equation 3.1) using Adam [55] solver. The proposed method has been evaluated on Microsoft 7-Scenes Dataset [101] containing RGB-D images covering different indoor locations, namely: Chess, Fire, Heads, Office, Pumpkin, Red Kitchen, and Stairs. The dataset was collected with a Kinect device and has been widely used for image-based localization [13, 20, 52, 53, 114] exhibiting significant variation in camera pose, motion blur and perceptual aliasing making the localization process based on SIFT-like features very challenging.

Table 3.1 shows the localization performance of Hourglass-Pose along with other CNN-based methods. The proposed approach outperforms PoseNet [54] on translation accuracy by a factor of 1.5 in all test scenes and substantially improves orientation accuracy. However, the main weakness of Hourglass-Pose is that, despite its good performance and robustness, it is a memory demanding architecture due to the regression part.

## 3.4 Recovering camera pose by image retrieval and ego-motion

Although learning-based approaches overcome many disadvantages of point-based methods, they still have certain limitations. For example, directly regressing the absolute camera pose constrains the current machine learning models to be trained and evaluated scene-wise when the scenes are registered to different coordinate frames. The reason for this is that the trained model learns a mapping from pixels to pose which is *dependent* on the coordinate frame of the training data belonging to a particular scene. To address this problem, Publication VI proposed a localization framework consisting of two modules: a Siamese CNN network for relative pose com-

**Table 3.1.** Camera localization performance of the proposed method and existing RGB-only CNN-based approaches for the *7-Scenes* dataset [101]. We follow original notation presented in [54] and provide median translation and orientation errors. In terms of localization error, our approach proposed in Publication VI is superior to other methods utilizing similar loss (3.1) such as PoseNet [54], LSTM-Pose [114], VidLoc [20] and Hourglass-Pose (Publication V) for the all scenes.

| Scene | Spatial Extent | PoseNet [54] | LSTM-Pose [114] | VidLoc [20] | Hourglass-Pose Publication V | PoseNet2 [53] | Publication VI (baseline) | Publication VI (proposed) |
|---|---|---|---|---|---|---|---|---|
| Chess | 3 × 2 × 1m | 0.32m, 8.12° | 0.24m, 5.77° | 0.18m, N/A | 0.15m, 6.53° | 0.13m, 4.48° | 0.12m, 6.69° | 0.13m, 6.46° |
| Fire | 2.5 × 1 × 1m | 0.47m, 14.4° | 0.34m, 11.9° | 0.26m, N/A | 0.27m, 10.84° | 0.27m, 11.3° | 0.31m, 13.36° | 0.26m, 12.72° |
| Heads | 2 × 0.5 × 1m | 0.29m, 12.0° | 0.21m, 13.7° | 0.14m, N/A | 0.19m, 11.63° | 0.17m, 13.0° | 0.16m, 13.78° | 0.14m, 12.34° |
| Office | 2.5 × 2 × 1.5m | 0.48m, 7.68° | 0.30m, 8.08° | 0.26m, N/A | 0.21m, 8.48° | 0.19m, 5.55° | 0.21m, 8.78° | 0.21m, 7.35° |
| Pumpkin | 2.5 × 2 × 1m | 0.47m, 8.42° | 0.33m, 7.00° | 0.36m, N/A | 0.25m, 7.01° | 0.26m, 4.75° | 0.25m, 7.89° | 0.24m, 6.35° |
| Red Kitchen | 4 × 3 × 1.5m | 0.59m, 8.64° | 0.37m, 8.83° | 0.31m, N/A | 0.27m, 10.15° | 0.23m, 5.35° | 0.22m, 9.35° | 0.24m, 8.03° |
| Stairs | 2.5 × 2 × 1.5m | 0.47m, 13.8° | 0.40m, 13.7° | 0.26m, N/A | 0.29m, 12.46° | 0.35m, 12.4° | 0.37m, 14.45° | 0.27m, 11.82° |
| Average | | 0.44m, 10.4° | 0.31m, 9.85° | 0.25m, N/A | 0.23m, 9.53° | 0.23m, 8.12° | 0.23m, 10.61° | 0.21m, 9.30° |

putation (Publication IV) and the localization pipeline. The input to the system is an RGB query image to be localized, and a database of images with their respective poses. At the first stage, we construct a set of training image pairs and use it to train a Siamese CNN to predict relative camera pose of each pair. It should be noted that the training image pairs can be independent of the scenes present in the localization database. Then, each trained branch of the network is considered a feature extractor and the extracted feature vectors can be utilized to identify the database images that are nearest neighbours (NN) to the query image in the feature space. Finally, relative pose estimates between the query and its neighbours are computed and then complemented with ground truth absolute location of the corresponding database images in a fusion algorithm producing the full 6-DoF camera pose. Specifically, from the shortlisted top ranked database images $d$, we select a pair $p^s = \{d^k, d^m\}$, where $p^s \subset d$ and $s = 1, 2, \ldots, \binom{N}{2}$. The translation direction predictions to the query $q$ from the images $p^s$ are triangulated to obtain the location/translation parameter of query camera, $t^s$. This gives us $\binom{N}{2}$ hypotheses for the query location which are then refined based on the angular distance between the query and the camera centers of the database images $p^r$, $p^r = d \setminus p^s$. Estimating rotation for the query camera can be estimated in a more straightforward way by using the following equation:

$$\Delta R^j = R_j^T R_q^j \tag{3.2}$$

where $R_j$ is the ground-truth orientation of the $j^{th}$ camera in $d$, $\Delta R^j$ is the relative orientation between the query and database image $j$ predicted by the trained network at the first stage of the proposed pipeline; $R_q^j$ is the $j^{th}$ hypothesis of the query camera orientation. For $N$ nearest neighbours it leads to $N$ hypotheses for query orientation. Instead of naïvely averaging the estimations, a consensus based filtering is used similarly to the process of estimating query translation. A robust rotation averaging algorithm [41] is then applied to obtain the final query orientation. According to the results presented in Table 3.1, the proposed approach can achieve competitive results compared with other CNN-based methods in camera relocalization. Although the improvement is not huge, it is worth to highlight that all the methods presented in Table 3.1 are trained in a scene-specific manner whereas the proposed approach is designed to overcome this fundamental limitation.

## 3.5 Discussion

This chapter focused on two closely related tasks; estimating absolute and relative camera pose. The main conclusions within the chapter are summarized below.

**Relative camera pose.** An end-to-end CNN-based architecture for egomotion estimation has been introduced. It predicts relative camera pose from an input pair of images and based on a Siamese network structure covered in Chapter 2. The proposed method is reliable providing egomotion estimates under challenging conditions where traditional methods based on hand-crafted image descriptors (Chapter 2) fail.

**Camera relocalization.** Two modifications to PoseNet [54]—a neural architecture for camera relocalization—have been proposed leading to significant improvement in localization. First, the original structure of PoseNet is modified by adding a CNN decoder to better collect the overall information, from coarse structures to fine-grained object details, available in the input image. Although the resulted structure, HourglassPose, outperforms PoseNet on indoor localization benchmarks, it has a very large memory footprint preventing using this model on mobile devices. Second, a novel image-based localization approach leveraging the advances of image retrieval (Chapter 2) and relative camera pose estimation is presented. The proposed approach generalizes well to previously unseen scenes and compares favourably to other CNN-based methods.

# 4. Geometric Scene Understanding

Chapter 3 of this overview has focused on applying deep learning to the camera relocalization problem and estimating camera ego-motion. This part explores the link between the knowledge we gained from the previous chapter and the scene understanding problem. Scene understanding requires knowledge of geometry and semantics and can be divided into many sub-tasks. This chapter focuses on geometric scene understanding, *i.e.* estimating the 3-D structure of a scene. The interest of this section is in the ideas how CNN-based methods can be employed to predict camera ego-motion and depth maps from monocular video sequences in a self-supervised setting.

## 4.1 Motivation

Understanding 3-D scene geometry from video is a long-standing and fundamental computer vision problem. The human visual system has a remarkable ability to make understanding of our 3-D world from its 2-D projection. Even in complex environments with multiple moving objects, people are able to maintain a feasible interpretation of the objects' geometry and depth ordering. The field of computer vision has long studied how to achieve similar capabilities by computationally reconstructing a scene's geometry from 2-D image data, but robust reconstruction remains difficult in many cases.

Scene understanding is a fundamental problem of computer vision which can be broadly classified into two categories; semantic and geometric scene understanding. Semantic scene understanding requires knowledge of semantic segmentation, instance segmentation and object detection/recognition [10, 57, 75]. This chapter studies geometric scene understanding which is the process of inferring the 3-D configuration of a scene without semantic labels. Specifically, it focuses on the problem of optical flow estimation, depth prediction, and ego-motion estimation. Each of these problems is highly ambiguous, thus learning a mapping from pixels to

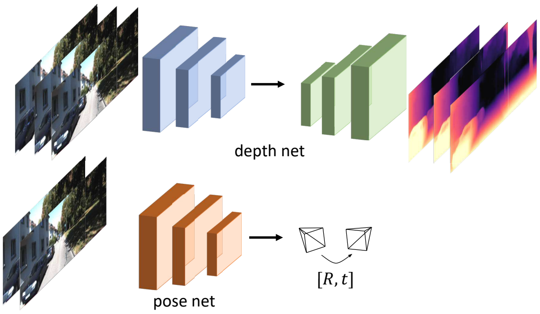


**Figure 4.1.** Schematic illustration of the method proposed in Publication VII to jointly learn scene depth and camera ego-motion from monocular video sequences. The network consists of two sub-nets (depth net and pose net, specifically) and can be trained end-to-end. We propose an additional term to the reprojection objective function efficiently utilizes temporal consistency between frames of the video sequence leading to more accurate depth and ego-motion estimates. See more details in Section 4.3.

depth, flow, and ego-motion becomes a challenging task without ground truth data. In Publication VII, we propose a novel CNN-based framework which can leverage temporally consistent geometric priors between frames of monocular video sequences and efficiently approach the problem of geometric scene understanding in a self-supervised manner.

## 4.2 Visual scene understanding approaches

Joint estimation of scene structure and camera motion can be performed by traditional structure-from-motion methods [74, 105]. Specifically, the pipeline of modern approaches consists of two stages, *i.e.* feature extraction and matching part which is followed by geometric verification procedure. In order to refine 3-D geometry scene estimates, bundle adjustment is iteratively applied during the reconstruction process. However, despite the success in robust feature descriptors [12, 89], traditional methods are still prone to outliers in some challenging cases such as texture-less objects and scenes with repetitive structures (Chapter 3).

In order to address these limitations, a wide range of deep learning based approaches has been recently proposed. These methods can be divided into two categories: supervised and self-supervised (unsupervised) approaches. Supervised deep learning methods have shown themselves capable to estimate depth maps from a single view by utilizing RGB-D datasets [111]. Although these methods have demonstrated very promising results, they require ground truth depth during training. However, collecting diverse and large datasets with accurate depth information is itself a very challenging task. Furthermore, constructing a dataset with optical flow ground truth labels is even more formidable compared to the RGB-D datasets. To alleviate these issues, existing optical flow estimation approaches [27, 44, 80, 106] utilize synthetic datasets, such as Flying Chairs [27] at training stage. While using synthetic training data is a promising idea, it is not trivial to generate sufficient amount of data exhibiting varied real-world appearance and motion. To address the requirement of expensive ground-truth data, several unsupervised approaches have been proposed by utilizing image reconstruction as a supervisory signal. Specifically, a set of images (either a monocular video sequence or a set of stereo image pairs) is provided to the model which minimizes the reprojection error between the target and the source view warped by the 2D optical flow. In the case of stereo pairs, by predicting the pixel disparities between the pair, a CNN model can be trained to perform monocular depth estimation at test time. Godard *et al.* [32] demonstrated that using geometric cues (left-right depth consistency) leads to superior performance in depth estimation compared to supervised methods.

Rather than using a stereo image pairs as model input, monocular video sequence looks an attractive alternative. However, in addition to estimating depth, the model needs to predict ego-motion (Chapter 3) between two frames during training which can be accomplished by an independent sub-network. The ego-motion estimates are then used to compute the optical flow representing the pixel displacement (movement) between the two input frames. It is important to note that movement of static parts (image background) in video sequences is induced by the camera motion and scene depth. Thus, the rigid optical flow between two nearby input frames $I_s$ and $I_t$ can be computed by using the depth of the reference view and relative camera pose between $I_t$ and $I_s$ as follows:

$$F_{t,s}^{rig} = KT_{t,s}D(x_t)K^{-1}x_t - x_t, \tag{4.1}$$

where $K$ is the camera intrinsics matrix; $D$ is the depth map for target frame; $T_{t,s}$ the relative camera pose transformation matrix between the target view and the source frames. The relative rigid flow $F_{t,s}^{rig}$ is then used to reconstruct the target frame from source image, i.e. $\tilde{I}_t = F_{t,s}^{rig}(I_s)$. The inconsistency between the reconstructed $\tilde{I}_t$ and original $I_t$ views is used to optimize the model. In addition to the background, there are usually dynamic objects in the input video which should be taken into account too.

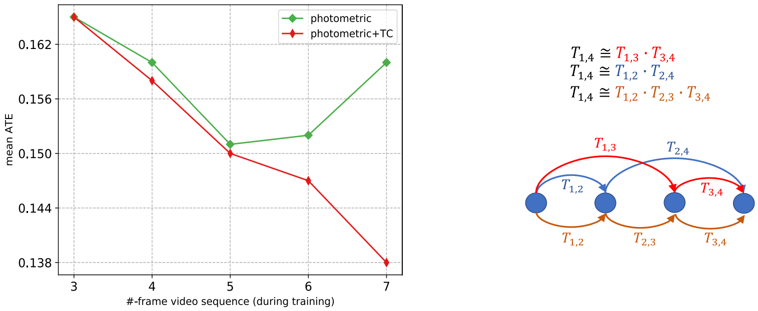
Joint estimation of scene depth (disparity) and relative camera pose from monocular video has been a long-standing problem. While initial methods have relied on ground truth information, such as ego-motion [21, 72, 111] and semantic masks [17], recent approaches have demonstrated a possibility to learn depth and ego-motion in a self-supervised manner. Zhou *et al.* [17] proposed a model consisting of two CNNs to predict relative camera pose and the depth in a coupled way. In order to cope with dynamic objects in video, [125] additionally predicts explainability masks to remove moving objects from the scene. In contrast to [125], [81, 122, 126] propose to explicitly learn the residual optical flow to deal with the non-rigid motion and apply forward-backward consistency to depth and optical flow estimates. Later work by Godard *et al*. [33] has extended [125] to propose a novel multi-scale sampling method and several modifications regarding reprojection loss allowing to minimize the gap between self-supervised and fully supervised methods. Wang *et al*. [115] propose a method that uses Recurrent Neural Networks (RNNs) and and can be optimized by utilizing multi-view image reprojection and forward-backward flow consistency losses for the task of scene understanding. Mahjourian *et al*.[68] present an unsupervised CNN-based approach for learning depth and ego-motion from monocular video by introducing a differentiable version of iterative closest point (ICP) method. Recent work by Guizilini *et al*. [37] proposed a novel loss function utilizing instantaneous velocity measurement to learn scale-aware depth. In addition to scene geometry, Gordon *et al*. [36] proposed a self-supervised approach to estimate camera intrinsics from monocular videos.

In Publication VII, we demonstrate how utilising the geometric consistency across a series of image frames can help to achieve an improvement upon existing self-supervised methods. The following sections provide more details about the proposed method.

**Datasets** In order to evaluate the proposed approach, two scene understanding datasets have been used. These datasets are summarized in this paragraph.

The KITTI driving dataset (KITTI raw) proposed by [1, 2] consists of a several outdoor scenes captured by driving vehicles with mounted cameras and depth sensors in different traffic conditions. The dataset has been widely used by other self-supervised approaches presented in this chapter. The depth ground truth data for this dataset is sampled at irregularly spaced points captured by LIDAR. Eigen *et al*. [29] provided the train and test splits avoiding duplicates and scenes where the car is stationary. In Publication VII, we use video clips from the train split to jointly learn the depth and camera ego-motion. In addition to the KITTI raw dataset, recent approaches [120, 122, 126] utilize the SYNTHIA [88] and CityScapes [22] datasets to pre-train the final model and improve the generalization.

For the task of ego-motion estimation the KITTI odometry dataset [1] is utilized. It contains 11 driving sequences with ground truth camera poses. Although most approaches [122, 125, 126] have utilized sequences 00-08 for fine-tuning the models, in Publication VII we use the dataset only for evaluation purposes. Traditionally, the absolute trajectory error (ATE) is used as a metric to compare estimated trajectories with ground-truth.



(a) Absolute trajectory error of the proposed method

(b) on the idea of temporally consistent ego-motion estimates

**Figure 4.2.** The proposed model improves when trained with longer video sequences, unlike baseline approaches; a) the plot shows absolute trajectory error (ATE) compared with training sequence length, computed for sequence 09 of the KITTI Odometry dataset [1]. The proposed temporal consistent term (TC) allows to achieve better localization results; b) schematic illustration of the idea of temporal consistency for the case of a 4-frame video sequence. Specifically, the relative camera pose transformation matrix between edge frames $T_{1,4}$ is consistent and considered to be a composite transformation of inner frames. See Publication VII for more detailed discussion.

## 4.3 On temporally consistent geometric prior

Utilizing reprojection error between pairs of images, a large family of monocular unsupervised scene-understanding methods (Section 4.2) has achieved promising results. In Publication VII, we have extended this idea and demonstrated how to leverage longer video sequences of images as a richer form of self-supervision. Rather than propose some modifications to the traditional network architecture [125], we rely on utilizing visual geometry cues. Specifically, given a sequence of consecutive frames, we constrain the network output to be geometrically consistent between each individual frame pair and the video sequence. Hence, we propose an additional term which is added to the reprojection loss function inferred by the rigid optical flow (Equation 4.1) and which enforces longer timescale ego-motion estimates to be consistent with the estimates between successive frames. As a result, this temporal consistency constraint leads to better localization performance efficiently handling large video sequences (*cf.* Figure 4.2).

The additional term is defined as follows. Complex multi stage rigid-body transformation can be represented as a composition of its independent stages. For example, if we have a 4-frame input video sequence (see Figure 4.2), relative camera pose transformation matrix between edge frames (1st and 4th) should be consistent with the transformation matrices of inner frames. Thus, temporal consistency term is defined as follows:

$$\mathscr{L}_{tc}^{ego} = \sum_{i=0}^{T-2} \sum_{j=2}^{T} \left\| f_p\left(T_{i,j}\right) - f_p\left(\prod_{k=i}^{j-1} T_{k,k+1}\right) \right\|. \tag{4.2}$$

**Table 4.1. Single-view depth estimation results** on *test split* of KITTI raw dataset [1]. The methods trained on KITTI raw dataset [1] are denoted by K. Models with additional training data from CityScapes [22] are denoted by CS+K. (D) denotes depth supervision, (B) denotes stereo input pairs, (M) denotes monocular video clips. The best performance for supervised methods is highlighted as italic and the most accurate results among self-supervised approaches are indicated as bold.

| | | Error metric ↓ | | | | Accuracy metric ↑ | | |
|---|---|---|---|---|---|---|---|---|
| Method | Dataset | Abs Rel | Sq Rel | RMSE | log RMSE | $\delta < 1.25$ | $\delta < 1.25^2$ | $\delta < 1.25^3$ |
| Eigen *et al.*. [29] | K (D) | 0.203 | 1.548 | 6.307 | 0.246 | 0.702 | 0.890 | 0.958 |
| Kuznietsov *et al.* [59] | K (B) / K (D) | *0.113* | *0.741* | *4.621* | *0.189* | *0.862* | *0.960* | *0.986* |
| Zhan *et al.* [124] | K (B) | 0.144 | 1.391 | 5.869 | 0.241 | 0.803 | 0.928 | 0.969 |
| Godard *et al.* [32] | K (B) | 0.133 | 1.140 | 5.527 | 0.229 | 0.830 | 0.936 | 0.970 |
| Godard *et al.*. [32] | CS+K (B) | 0.121 | 1.032 | 5.200 | 0.215 | 0.854 | 0.944 | 0.973 |
| Zhou *et al.* [125] no explainability mask | K (M) | 0.221 | 2.226 | 7.527 | 0.294 | 0.676 | 0.885 | 0.954 |
| Zhou *et al.* [125] | K (M) | 0.208 | 1.768 | 6.856 | 0.283 | 0.678 | 0.885 | 0.957 |
| Zhou *et al.* [125] updated | K (M) | **0.183** | 1.590 | 6.709 | 0.270 | **0.734** | 0.902 | 0.959 |
| Proposed (Publication VII) | K (M) | 0.188 | **1.529** | **6.431** | **0.265** | 0.726 | **0.905** | **0.962** |

Here, $T_{i,j}$ is the relative camera pose transformation matrix obtained from 6DoF ego-motion vector $p_{i,j}$. However, rather than minimize $L1$ distance between transformation matrices, we first decompose each matrix into orientation and translation component by applying $f_p$ function and optimize them independently. We use Euler angles to parameterize orientation since it can achieve better performance.

The results obtained by our best model have been presented in Table 4.1. Since depth and ego-motion are tightly coupled (see Equation 4.1) for rigid scenes, more accurate relative pose estimates lead to better performance in depth estimation. In addition to ego-motion, temporal consistency could be applied to optical flow and depth estimates for monocular sequences which will be an interesting starting point for future research. This example is a manifestation to the principles discussed throughout the whole dissertation; leveraging the multi-view geometry [40] advances can benefit a lot in various computer vision applications such as image matching (Chapter 2), image-based localization (Chapter 3), and scene-understanding.

## 4.4 Discussion

This chapter addressed two tasks, *i.e.* estimating depth maps and camera ego-motion (Chapter 3) from monocular video sequences. We briefly summarize the main conclusions within the chapter.

First, an end-to-end deep learning framework to jointly estimate depth maps and relative camera pose is introduced. The proposed method is trained in a self-supervised manner naturally utilizing geometric relationship between depth and camera motion of the scene. Second, the chapter showed that existing CNN-based approaches minimizing reprojection error between two frames can not fully utilize large video sequences. In fact, the absolute trajectory error is not improved if the length of the input video sequence increases. To address this problem, we proposed an additional term to the reprojection objective function which can exploit temporal consistency between frames throughout longer unlabeled video sequences has been proposed. This term provides a richer form of self-supervision leading to better performance compared to strong baseline methods.

# 5. Summary of the Original Articles

The thesis is based on seven articles which are briefly summarized and discussed in this chapter. The articles are reprinted and included in the appendix of the thesis.

Publication I and Publication II discuss the image patch and image matching problems, respectively. For image patch matching, we propose a learned image descriptor based on a Siamese network structure (*cf.* Section 2.1) which can be trained end-to-end so that the distance between feature representations is small for similar patches and large otherwise. The proposed descriptor significantly outperforms traditional methods. The paper also gives suggestions about using spatial transformer networks as a part of the descriptor aiming at better robustness to spatial misalignment. The ideas proposed in Publication I have been extended in Publication II to the problem of image matching and image retrieval, *i.e.* finding matching and non-matching pairs of images. The proposed learned image descriptor leverages the idea of transfer learning and demonstrates good generalization performance to unseen data.

Publication III studies the problem of dense pixel correspondence estimation by proposing a novel semi-supervised framework. The problem of predicting pixel correspondences is closely related to optical flow estimation. While optical flow methods produce very accurate results for the limited appearance variation scenarios, they typically do not deal with the strong geometric transformations caused by large viewpoint differences. The paper has addressed this issue and introduced an approach representing a hierarchical architecture which can iteratively refine pixel correspondence estimations between two views. Despite the fact that the proposed method has been trained on synthetic transformations, it can generalize well to real data exhibiting various illumination changes. In addition to image matching, we apply the method to the problem of relative camera pose estimation (Publication IV) and demonstrate that the model achieves favourable performance compared to traditional methods. The approach can be potentially useful also for other computer vision problems, such as image retrieval [60].

Publication IV extended PoseNet [54] and proposed a CNN-based method for relative camera pose estimation. We leveraged the knowledge about image matching gained from Publication I and Publication II and introduced a neural network architecture consisting of two CNN branches with shared weights taking a pair of RGB images as input and producing the relative camera rotation and translation as output. The proposed approach is robust to large viewpoint changes where traditional hand-crafted descriptors are not able to determine sufficient amount of correspondences to estimate relative pose accurately.

Publication V addresses the problem of image-based localization and introduces an end-to-end approach to predict absolute camera pose from a single RGB image. The proposed CNN architecture consists of a symmetric encoder-decoder pair followed by a regression part which directly estimates the pose. Such a hourglass structure helps to preserve the fine-grained information of the input image providing more context to the regression part of the architecture which leads to better localization performance compared to strong baseline methods [53, 54, 114]. However, this improvement has been achieved with a high memory footprint of the model making the proposed method intractable on mobile devices.

Similarly to Publication V, Publication VI studies the problem of camera relocalization. The paper proposes a novel deep architecture that leverages advances of image retrieval (Publication II) and relative camera pose estimation (Publication IV) to efficiently predict absolute camera pose. Specifically, the proposed approach localizes a given query image by using a CNN for first retrieving similar database images and then predicting the relative pose between the query and the database images with known poses. The absolute pose of the query is obtained by a new fusion algorithm based on triangulation from two relative translation estimates and geometric verification. An important take-home message of the paper is that existing deep learning methods for camera relocalization have to be trained and evaluated scene-wise when the scenes are registered to different coordinate systems. This causes complications, especially if one is interested in localization across several scenes simultaneously. The paper demonstrates that connections between Publication IV and image retrieval ideas benefit a lot leading to a scalable image-based localization approach. As another contribution, the paper introduces a challenging indoor localization dataset covering five different scenes registered to a common coordinate frame.

Finally, Publication VII explores the task of geometric scene understanding requiring the knowledge of camera ego-motion and scene depth maps. The paper introduces a CNN-based approach to *jointly* learn scene depth and ego-motion from monocular image sequences in a self-supervised manner. The proposed method leverages visual geometry and exploits temporal consistency between image frames throughout longer unlabeled video sequences. The main contribution of the paper is an additional term to the reprojection objective loss function which constrains the network output to be geometrically consistent between each individual frame pair and the longer sequence which leads to better performance in the monocular depth estimation problem.

# 6. Conclusion

This thesis has addressed a number of closely related computer vision problems—image matching, pixel correspondence estimation, absolute and relative camera pose estimation, and geometric scene understanding—by providing a set of end-to-end CNN-based methods optimizing the model with respect to the end goal. The connection between these topics is provided in the thesis. For example, dense pixel correspondences obtained by the method proposed in Publication III can be utilized in image matching and in the absolute and relative camera pose estimation pipelines. Further, the results of the thesis demonstrate, that achieving more accurate ego-motion estimates automatically improves depth estimates, since depth and ego-motion are tightly coupled. A particular emphasis has been placed on semi-supervised and self-supervised learning methods which leverage the advances of 3-D visual geometry in order to avoid the necessity of large annotated ground truth datasets which are very tedious and expensive to collect. For computer vision tasks considered in the thesis, we demonstrate that the proposed methods exhibit interesting properties and have achieved favourable results compared to traditional keypoint-based approaches in terms of both accuracy, and computational efficiency.

The themes presented in the thesis suggest some avenues for future research. Firstly, there are still many open questions related to applicability of CNN-based approaches in image-based localization (Publication VI). Although they provide promising results and learnable representations, the localization performance obtained by pure deep learning approaches is still behind the results produced by traditional methods, thus, additional steps, such as geometric verification could be required to improve the accuracy. It would be interesting to work on an end-to-end differentiable CNN-based approach which could outperform keypoint-based methods in localization benchmarks. Secondly, dense pixel correspondence estimation methods (Publication III) could be extended to the wide baseline case with background clutter. This potential research direction could bring benefits for 3-D reconstruction and camera relocalization. Lastly, in the context of geometric scene understanding from monocular videos, it would be useful if the proposed idea about temporal consistency of ego-motion could be extended to other modalities, such as depth and optical flow estimates to provide additional supervision and better performance. Such advances could have a great impact on the application of autonomous driving enabling an opportunity to rely more on camera sensors rather than expensive hardware, such as LiDAR.

# Bibliography

[1] C. Stiller A. Geiger, P. Lenz and R. Urtasun. Vision meets robotics: The KITTI dataset. In *The International Journal of Robotics Research (IJRR)*, volume 32, pages 1231–1237, 2013.

[2] P. Lenz A. Geiger and R. Urtasun. Are we ready for autonomous driving? The kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[3] N. Radwan A. Valada and W. Burgard. Deep Auxiliary Learning For Visual Localization And Odometry. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[4] P.F Alcantarilla, A. Bartoli, and A.J. Davison. KAZE Features. In *IEEE European Conference on Computer Vision (ECCV)*, 2012.

[5] P.F Alcantarilla, J. Nuevo, and A. Bartoli. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In *British Machine Vision Conference (BMVC)*, 2013.

[6] 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.

[7] 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.

[8] R. Arandjelović and A. Zisserman. All About VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[9] A. Babenko, A. Slesarev, A. Chigorin, and V.S. Lempitsky. Neural Codes for Image Retrieval. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(12):2481–2495, 2017.

[11] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008.

[13] 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.

[14] A. Bursuc, G. Tolias, and H. Jégou. Kernel Local Descriptors with Implicit Rotation Matching. In *ACM International Conference on Multimedia Retrieval*, 2015.

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

[17] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova. Depth Prediction Without the Sensors: Leveraging Structure for Unsupervised Learning from Monocular Videos. In *Association for the Advancement of Artificial Intelligence Conference (AAAI)*, 2019.

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

[20] 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.

[21] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni. VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2017.

[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] 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.

[24] D. DeTone, T. Malisiewicz, and A. Rabinovich. Deep Image Homography Estimation. In *Proc. in RSS Workshop on Limits and Potentials of Deep Learning in Robotics*, 2016.

[25] 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.

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

[27] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow with Convolutional Networks. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2015.

[28] 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.

[29] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2015.

[30] M. Everingham, L. Gool, C.K. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision (IJCV)*, 88(2):303–338, 2010.

[31] 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.

[32] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into Self-Supervised Monocular Depth Prediction. *arXiv:1806.01260v3*, 2019.

[34] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[35] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.

[36] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from Videos in the Wild: Unsupervised Monocular Depth Learning from Unknown Cameras. *CoRR*, abs/1904.04998, 2019.

[37] V. Guizilini, R. Ambrus, S. Pillai, and A. Gaidon. PackNet-SfM: 3D Packing for Self-Supervised Monocular Depth Estimation. *CoRR*, abs/1905.02693, 2019.

[38] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[39] 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.

[40] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2 edition, 2003.

[41] 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.

[42] K. He, X. Zhang, S. Ren, and J. Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

[43] G.E. Hinton, S. Osindero, and Y.-W. Teh. A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18(7):1527–1554, 2006.

[44] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] 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.

[46] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28*, pages 2017–2025. 2015.

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

[48] J. Janai, G. Fatma, R. Anurag, M. J. Black, and A. Geiger. Unsupervised Learning of Multi-Frame Optical Flow with Occlusions. In *IEEE European Conference on Computer Vision (ECCV)*, 2018.

[49] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating Local Image Descriptors into Compact Codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(9):1704–1716, 2012.

[50] J. Johansson, M. Solli, and A. Maki. An evaluation of local feature detectors and descriptors for infrared images. In *IEEE European Conference on Computer Vision (ECCV Workshop)*, 2016.

[51] Y. Ke and R. Sukthankar. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

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

[53] 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.

[54] 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.

[55] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

[56] 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.

[57] I. Kokkinos. UberNet: Training a 'Universal' Convolutional Neural Network for Low-, Mid-, and High-Level Vision using Diverse Datasets and Limited Memory. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[58] K. Konda and R. Memisevic. Learning visual odometry with a convolutional network. In *VISIGRAPP*, 2015.

[59] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-Supervised Deep Learning for Monocular Depth Map Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[60] Z. Laskar, I. Melekhov, H.R. Tavakoli, J. Ylioinas, and J. Kannala. Geometric Image Correspondence Verification by Dense Pixel Matching. *CoRR*, abs/1904.06882, 2019.

[61] V. Lepetit and P. Fua. Keypoint Recognition Using Randomized Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28:1465–79, 10 2006.

[62] S. Leutenegger, M. Chli, and R. Y. Siegwart. BRISK: Binary Robust invariant scalable keypoints. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2011.

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

[64] X. Li, J. Ylioinas, J. Verbeek, and J. Kannala. Scene Coordinate Regression with Angle-Based Reprojection Loss for Camera Relocalization. In *IEEE European Conference on Computer Vision Workshop (ECCVW)*, 2018.

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

[66] W.-Y. Lin, M.-M. Cheng, J. Lu, H. Yang, M. N. Do, and P. Torr. Bilateral Functions for Global Motion Modeling. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

[67] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 60:91–110, 2004.

[68] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised Learning of Depth and Ego-Motion from Monocular Video Using 3D Geometric Constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[69] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems 29*, pages 2802–2810. 2016.

[70] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(10):1615–1630, 2005.

[71] 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 30*, pages 4826–4837. 2017.

[72] V. Mohanty, S. Agrawal, S. Datta, A. Ghosh, V. D. Sharma, and D. Chakravarty. DeepVO: A deep learning approach for monocular visual odometry. *CoRR*, abs/1611.06069, 2016.

[73] A. Valada N. Radwan and W. Burgard. VLocNet++: Deep Multitask Learning For Semantic Visual Localization And Odometry. *IEEE Robotics And Automation Letters (RA-L)*, 3(4):4407–4414, 2018.

[74] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2011.

[75] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In *IEEE European Conference on Computer Vision (ECCV)*, 2016.

[76] A. Newell, K. Yang, and J. Deng. Stacked Hourglass Networks for Human Pose Estimation. In *IEEE European Conference on Computer Vision (ECCV)*, 2016.

[77] G.L. Oliveira, N. Radwan, W. Burgard, and T. Brox. Topometric Localization with Deep Learning. In *IEEE European Conference on Computer Vision Workshop (ECCVW)*, 2018.

[78] C. Osendorfer, J. Bayer, S. Urban, and P. van der Smagt. Convolutional Neural Networks Learn Compact Local Image Descriptors. In *Neural Information Processing*, pages 624–630, 2013.

[79] P.O. Pinheiro, T.-Y Lin, R. Collobert, and P. Dollár. Learning to Refine Object Segments. In *IEEE European Conference on Computer Vision (ECCV)*, 2016.

[80] A. Ranjan and M. J. Black. Optical Flow Estimation using a Spatial Pyramid Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[81] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, and M.J. Black. Adversarial Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[82] A.S. Razavian, J. Sullivan, S. Carlsson, and A. Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016.

[83] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. DeepMatching: Hierarchical Deformable Dense Matching. *International Journal of Computer Vision (IJCV)*, 120(3):300–323, 2016.

[84] 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.

[85] 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.

[86] I. Rocco, M. Cimpoi, R. Arandjelović, A. Torii, T. Pajdla, and J. Sivic. Neighbourhood consensus networks. In *Advances in Neural Information Processing Systems 31*, pages 1651–1662. 2018.

[87] O. Ronneberger, P.Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015.

[88] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[89] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: An Efficient Alternative to SIFT or SURF. In *IEEE Internation Conference on Computer Vision (ICCV)*, 2011.

[90] 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.

[91] 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.

[92] 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)*, 39(9):1744–1756, 2017.

[93] 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.

[94] 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.

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

[96] 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.

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

[98] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *IEEE European Conference on Computer Vision (ECCV)*, 2016.

[99] 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.

[100] T. Shi, S. Shen, X. Gao, and L. Zhu. Visual Localization Using Sparse Semantic 3D Map. *CoRR*, abs/1904.03803, 2019.

[101] 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.

[102] 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 Internation Conference on Computer Vision (ICCV)*, 2015.

[103] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning Local Feature Descriptors Using Convex Optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(8):1573–1585, 2014.

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

[105] N. Snavely, S.M. Seitz, and R. Szeliski. Modeling the World from Internet Photo Collections. *International Journal of Computer Vision (IJCV)*, 80(2):189–210, 2008.

[106] 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.

[107] 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.

[108] 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.

[109] 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.

[110] 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.

[111] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. DeMoN: Depth and motion network for learning monocular stereo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[112] 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.

[113] J. Vertens, A. Valada, and W. Burgard. SMSnet: Semantic Motion Segmentation using Deep Convolutional Neural Networks. In *International Conference on Intelligent Robots and Systems (IROS)*, 2017.

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

[115] R. Wang, S.M Pizer, and J.-M. Frahm. Recurrent Neural Network for (Un-)supervised Learning of Monocular Video Visual Odometry and Depth. In *CoRR*, 2019.

[116] S. Wang, R. Clark, H. Wen, and N. Trigoni. End-to-End, Sequence-to-Sequence Probabilistic Visual Odometry through Deep Neural Networks. *International Journal of Robotics Research (IJRR)*, 37:513–542, 2018.

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

[118] K. Wilson and N. Snavely. Robust Global Translations with 1DSfM. In *IEEE European Conference on Computer Vision (ECCV)*, 2014.

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

[120] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. LEGO: Learning Edge with Geometry all at Once by Watching Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

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

[122] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[123] 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.

[124] H. Zhan, R. Garg, C.S. Weerasekera, K. Li, H. Agarwal, and I. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[125] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[126] Y. Zou, Z. Luo, and J.-B. Huang. DF-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *IEEE European Conference on Computer Vision (ECCV)*, 2018.

Aalto-DD 23/2020

9 789526 089447

ISBN 978-952-60-8944-7 (printed)
ISBN 978-952-60-8945-4 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

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