

ON LEARNING PSEUDO-SENSORS TO IMPROVE EGOMOTION ESTIMATION FOR  
MOBILE AUTONOMY

by

Valentin Peretroukhin

A thesis submitted in conformity with the requirements  
for the degree of Doctor of Philosophy  
Graduate Department of Institute for Aerospace Studies  
University of Toronto

© Copyright 2019 by Valentin Peretroukhin

# Abstract

On Learning Pseudo-sensors to Improve Egomotion Estimation for Mobile Autonomy

Valentin Peretroukhin

Doctor of Philosophy

Graduate Department of Institute for Aerospace Studies

University of Toronto

2019

The ability to estimate *egomotion*, that is, to track one's own pose through an ~~unknown~~ environment, is at the heart of safe and reliable mobile autonomy. By inferring pose changes from sequential sensor measurements, egomotion estimation forms the basis of mapping and navigation pipelines, and permits mobile robots to self-localize within environments where external localization **sources** are intermittent or unavailable. Visual and inertial egomotion estimation, in particular, have become ubiquitous in mobile robotics due to the availability of high-quality, compact, and inexpensive sensors that capture **rich representations** of the world. To remain computationally tractable, ‘classical’ visual-inertial pipelines (like **visual odometry** and **visual SLAM**) make simplifying assumptions that, while permitting reliable operation in ideal conditions, often lead to systematic error. In this thesis, we present several data-driven learned *pseudo-sensors* that serve to augment conventional pipelines by inferring latent information from the same sensor data. Our approach retains much of the benefits of traditional pipelines, while leveraging high-capacity hyper-parametric models to extract complementary information that can be used to improve uncertainty quantification, correct for systematic bias, and improve robustness to difficult-to-model deleterious effects. We validate our pseudo-sensors on several kilometres of sensor data collected in sundry settings such as urban roads, indoor labs, and planetary analogue sites in the Canadian High Arctic.

# Epigraph

A little learning is a dangerous thing;  
drink deep, or taste not the Pierian  
spring: there shallow draughts  
intoxicate the brain, and drinking  
largely sobers us again.

---

ALEXANDER POPE

The universe is no narrow thing and the order within it is not constrained by any latitude in its conception to repeat what exists in one part in any other part. Even in this world more things exist without our knowledge than with it and the order in creation which you see is that which you have put there, like a string in a maze, so that you shall not lose your way. For existence has its own order and that no man's mind can compass, that mind itself being but a fact among others.

---

CORMAC McCARTHY

Elephants don't play chess.

---

RODNEY BROOKS

To all those who encouraged (or, at least, *never discouraged*) my intellectual wanderlust.

## Acknowledgements

This document would not have been possible without the generous support and guidance of my supervisor<sup>1</sup>, the perennial love of my family and friends<sup>2</sup>, and the limitless patience of my lab mates<sup>3</sup>. Thank you all.

---

<sup>1</sup>as well as all of my collaborators and academic mentors (special thanks to Lee)

<sup>2</sup>especially the support and encouragement of Elyse

<sup>3</sup>in humouring my insatiable need for debate and banter (special thanks to Lee)

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Autonomy and Humanity . . . . .	2
1.2	Mobile Autonomy and State Estimation . . . . .	3
1.3	The <i>State</i> of State Estimation . . . . .	6
1.4	The Learned Pseudo-Sensor . . . . .	7
1.5	Original Contributions . . . . .	8
<b>2</b>	<b>Mathematical Foundations</b>	<b>12</b>
2.1	Coordinate Frames . . . . .	12
2.2	Rotations . . . . .	13
2.2.1	Unit Quaternions . . . . .	14
2.3	Spatial Transforms . . . . .	15
2.3.1	Applying Transforms . . . . .	16
2.4	Perturbations . . . . .	16
2.5	Uncertainty . . . . .	18
<b>3</b>	<b>Classical Visual Odometry</b>	<b>19</b>
3.1	A Taxonomy of VO . . . . .	20
3.2	A Classical VO Pipeline . . . . .	20
3.2.1	Preprocessing . . . . .	20
3.2.2	Data Association . . . . .	21
3.2.3	Maximum Likelihood Motion Solution . . . . .	23
3.3	Robust Estimation . . . . .	25
3.4	Outstanding Issues . . . . .	27
<b>4</b>	<b>Predictive Robust Estimation</b>	<b>28</b>
4.1	Introduction . . . . .	28
4.2	Motivation . . . . .	29

4.3	Related Work . . . . .	30
4.4	Predictive Robust Estimation for VO . . . . .	30
4.4.1	Bayesian Noise Model for Visual Odometry . . . . .	30
4.4.2	Generalized Kernels . . . . .	32
4.4.3	Generalized Kernels for Visual Odometry . . . . .	33
4.4.4	Inference without ground truth . . . . .	36
4.5	Prediction Space . . . . .	38
4.5.1	Angular velocity and linear acceleration . . . . .	39
4.5.2	Local image entropy . . . . .	39
4.5.3	Blur . . . . .	39
4.5.4	Optical flow variance . . . . .	41
4.5.5	Image frequency composition . . . . .	42
4.6	Experiments . . . . .	42
4.6.1	Simulation . . . . .	43
4.6.2	KITTI . . . . .	44
4.6.3	UTIAS . . . . .	47
4.7	Summary . . . . .	50
<b>5</b>	<b>Learned Probabilistic Sun Sensor</b>	<b>52</b>
5.1	Introduction . . . . .	53
5.2	Motivation . . . . .	53
5.3	Related Work . . . . .	55
5.4	Sun-Aided Stereo Visual Odometry . . . . .	57
5.4.1	Observation Model . . . . .	57
5.4.2	Sliding Window Bundle Adjustment . . . . .	58
5.5	Orientation Correction . . . . .	59
5.6	Indirect Sun Detection using a Bayesian Convolutional Neural Network . . . . .	60
5.6.1	Cost Function . . . . .	61
5.6.2	Uncertainty Estimation . . . . .	61
5.6.3	Implementation and Training . . . . .	62
5.7	Simulation Experiments . . . . .	63
5.8	Urban Driving Experiments: The KITTI Odometry Benchmark . . . . .	71
5.8.1	Sun-BCNN Test Results . . . . .	74
5.8.2	Visual Odometry Experiments . . . . .	75
5.9	Planetary Analogue Experiments: The Devon Island Rover Navigation Dataset . . . . .	76
5.9.1	Sun-BCNN Test Results . . . . .	79

5.9.2	Visual Odometry Experiments . . . . .	80
5.10	Sensitivity Analysis . . . . .	82
5.10.1	Cloud Cover . . . . .	82
5.10.2	Model Generalization . . . . .	85
5.10.3	Mean and Covariance Computation . . . . .	88
5.11	Summary . . . . .	90
<b>6</b>	<b>Learned Pose Corrections</b>	<b>91</b>
6.1	Introduction . . . . .	91
6.2	Motivation . . . . .	92
6.3	Related Work . . . . .	93
6.4	System Overview: Deep Pose Correction . . . . .	94
6.4.1	Loss Function: Correcting SE(3) Estimates . . . . .	96
6.4.2	Loss Function: SE(3) Covariance . . . . .	96
6.4.3	Loss Function: SE(3) Jacobians . . . . .	97
6.4.4	Loss Function: Correcting SO(3) Estimates . . . . .	99
6.4.5	Pose Graph Relaxation . . . . .	99
6.5	Experiments . . . . .	100
6.5.1	Training & Testing . . . . .	100
6.5.2	Estimators . . . . .	102
6.5.3	Evaluation Metrics . . . . .	103
6.6	Results & Discussion . . . . .	109
6.6.1	Correcting Sparse Visual Odometry . . . . .	109
6.6.2	Distorted Images . . . . .	110
6.7	Summary . . . . .	110
<b>7</b>	<b>Learned Probabilistic Rotations</b>	<b>111</b>
7.1	Introduction . . . . .	111
7.2	Motivation . . . . .	112
7.3	Related work . . . . .	113
7.4	Approach . . . . .	114
7.4.1	Why Rotations? . . . . .	114
7.4.2	Probabilistic Regression . . . . .	115
7.4.3	Deep Probabilistic SO(3) Regression . . . . .	117
7.4.4	Loss Function . . . . .	119
7.5	Experiments . . . . .	121
7.5.1	Uncertainty Evaluation: Synthetic Data . . . . .	121

7.5.2	Absolute Orientation: 7-Scenes . . . . .	123
7.5.3	Relative Rotation: KITTI Visual Odometry . . . . .	123
7.6	Summary . . . . .	129
<b>8</b>	<b>Conclusion</b>	<b>130</b>
8.1	Summary of Contributions . . . . .	131
8.1.1	Predictive Robust Estimation . . . . .	131
8.1.2	Sun BCNN . . . . .	131
8.1.3	Deep Pose Corrections . . . . .	132
8.1.4	Deep Probabilistic Inference of $\text{SO}(3)$ with HydraNet . . . . .	132
8.2	Future Work . . . . .	132
8.3	Final Remarks . . . . .	133
8.4	Coda: In Search of Elegance . . . . .	134
<b>Appendices</b>		<b>137</b>
<b>A</b>	<b>PROBE: Isotropic Covariance Models through K-NN</b>	<b>138</b>
A.1	Introduction . . . . .	138
A.1.1	Theory . . . . .	138
A.1.2	Training . . . . .	139
A.1.3	Testing . . . . .	140
A.2	Experiments . . . . .	141
<b>B</b>	<b>Visual Odometry Implementation Details</b>	<b>143</b>
B.1	Overview . . . . .	143
B.2	Solution with Robust Loss . . . . .	144
B.3	Deriving the Necessary Jacobians . . . . .	144
<b>Bibliography</b>		<b>146</b>



# Notation

- $a$  : Symbols in this font are real scalars.
- $\mathbf{a}$  : Symbols in this font are real column vectors.
- $\mathbf{A}$  : Symbols in this font are real matrices.
- $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$  : Normally distributed with mean  $\boldsymbol{\mu}$  and covariance  $\mathbf{R}$ .
- $E[\cdot]$  : The expectation operator.
- $\underline{\mathcal{F}}_a$  : A reference frame in three dimensions.
- $(\cdot)^\wedge$  : An operator associated with the Lie algebra for rotations and poses. It produces a matrix from a column vector.
- $(\cdot)^\vee$  : The inverse operation of  $(\cdot)^\wedge$
- $\mathbf{1}$  : The identity matrix.
- $\mathbf{0}$  : The zero matrix.
- $\mathbf{p}_a^{c,b}$  : A vector from point  $b$  to point  $c$  (denoted by the superscript) and expressed in  $\underline{\mathcal{F}}_a$  (denoted by the subscript).
- $\mathbf{C}_{a,b}$  : The  $3 \times 3$  rotation matrix that transforms vectors from  $\underline{\mathcal{F}}_b$  to  $\underline{\mathcal{F}}_a$ :  $\mathbf{p}_a^{c,b} = \mathbf{C}_{a,b}\mathbf{p}_b^{c,b}$ .
- $\mathbf{T}_{a,b}$  : The  $4 \times 4$  transformation matrix that transforms homogeneous points from  $\underline{\mathcal{F}}_b$  to  $\underline{\mathcal{F}}_a$ :  $\mathbf{p}_a^{c,a} = \mathbf{T}_{a,b}\mathbf{p}_b^{c,b}$ .

# Chapter 1

## Introduction

To be sure, a writer cannot begin with a thesis; he must rather use his writerly sensitivity to intuit what is going on, even if he cannot understand its implications.



---

GARY MORSON, *How the great truth dawned*

### 1.1 Autonomy and Humanity

Autonomous systems, in some form, have been imagined and realized for the bulk of recorded history. In ancient Greek mythology, the god Hephaestus, the ‘patron of invention and technology’ (Mayor, 2019) was said to create talking mechanical hand-maidens, while early Hindu and Buddhist texts tell of *yantakara* that lived in Greece and created machines that helped in trade and farming. The secret methods of the *yantakara* (the early ‘roboticists’) were closely guarded, and mechanical assassins were said to pursue and kill any person who revealed their techniques<sup>1</sup> (Mayor, 2019).



Figure 1.1: A ‘robot’ rebellion from Karel Čapek’s 1920 play, *Rossum’s Universal Robots*.

---

<sup>1</sup>Please be careful distributing this thesis.

Since the industrial revolution, the idea of an autonomous machine—one that requires no, or very minimal, human intervention or oversight to operate—has been imagined in different ways. Depending on one’s perspective, autonomous machines have perennially promised to either usher in a utopia of freedom, or threatened to bring about an age of job loss and social upheaval that worsens socioeconomic divisions. Much like the Luddites of the 19th century, the social critics of the 21st century have continued the dialectic to understand the social ramifications of modern *yantakara* and their newly-created autonomous hand-maidens.

These controversial origins are embedded even within the modern name of for the academic field, *robotics*. The word *robot* comes from the title of a science fiction play, R.U.R.: Rossum’s Universal Robots, written by the Czech playwright Karel Čapek in 1920 (see Figure 1.2). In naming the play, the word *robot* was derived from the Slavic term for slave, *rab*, and its Czech derivative for serf labour, *rabota*, while the name *Rossum* was inspired by the Czech word for reason, or intellect. Indeed, the concept of enslaved or embodied intelligence is at the heart of modern definitions of the discipline of robotics (Redfield, 2019). Much of the popular culture surrounding robots (e.g., Shelley’s *Frankenstein*, Asimov’s *I, Robot*, Kubrick’s and Clarke’s *2001: A Space Odyssey*) also paints a complicated picture of humanity’s relationship with such enslaved machines. In this dissertation, we focus on improving a specific part of a modern mobile *autonomy* pipeline, while minimizing the use of term *robot* to avoid the maelstrom of philosophical and ethical problems that it connotes. We hope this work aids the march of technological progress towards a future which finds some Hegelian synthesis of autonomy and humanity—a future in which human-in-the-loop autonomous systems augment and improve the lot of many people while still negotiating and constantly considering the social costs that come with technological innovation.

## 1.2 Mobile Autonomy and State Estimation

While the looms and railroads of the industrial revolution were spurred by the discovery of steam engines and electricity, modern *mobile* autonomy was largely born out of the technological arms race of the cold war and the constraints and challenges associated with long-distance flight and extraterrestrial travel (see Grewal and Andrews (2010) for a history of one of the seminal algorithms in mobile autonomy, the Kalman filter). Indeed, much of the work on modern perception algorithms has its origins in the automated compilation of cold-war-era reconnaissance imagery and the design of extraterrestrial rovers like the Mars Exploration Rovers, *Spirit* and *Opportunity* (Scaramuzza and Fraundorfer, 2011). Similarly, much of the planning and control algorithms originate in American and Soviet defence-funded research (Nilsson, 1984; Thrun et al., 2006).

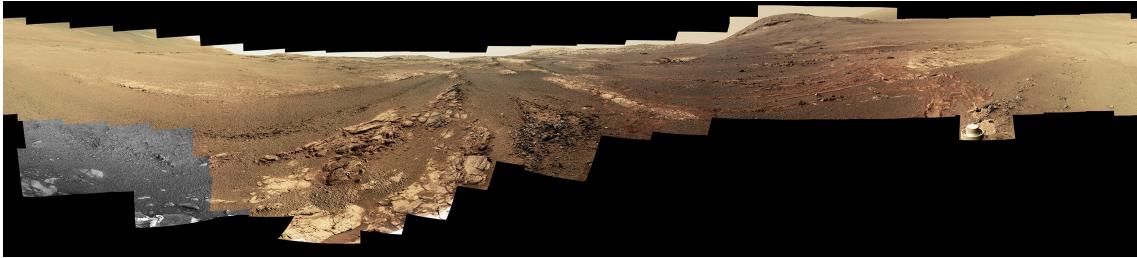


Figure 1.2: The last 360 degree panorama taken by the PanCam apparatus of the Mars Exploration Rover, *Opportunity*, at its final resting place on Mars, the western rim of the Endeavour Crater. Contact with *Opportunity* was lost shortly after this was captured, due to a severe dust storm (credit: NASA/JPL-Caltech/Cornell/ASU).

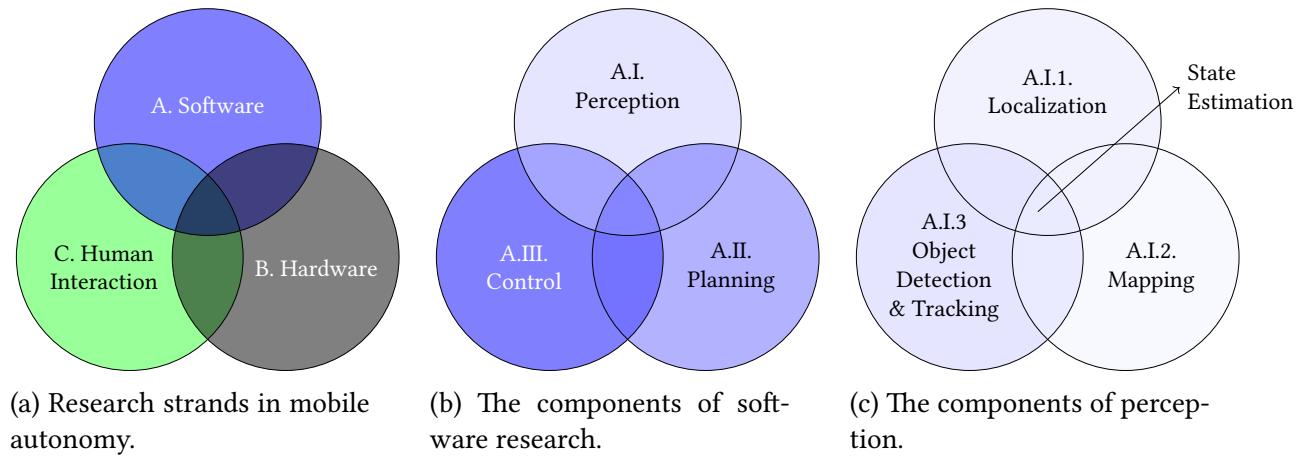


Figure 1.3: Venn diagrams of modern mobile autonomy.

Once confined to carefully-controlled factory floors, autonomous mobile platforms have now begun to show great promise in improving the safety of human transport, reducing the burden of repetitive, arduous jobs, and more efficiently leveraging limited resources for environmental monitoring. This newly-realized potential can be attributed to several factors: improvements in the cost and efficiency of computing devices (in terms energy efficiency, processing power, and overall size), the availability of relatively cheap, compact, high-quality sensors and rapid prototyping tools, and the development of open-source hardware, software platforms and datasets (e.g., the Robot Operating System, the KITTI Self-Driving Car dataset ([Geiger et al., 2013](#))).

Despite decades of research, mobile autonomy as a field still has nebulous demarcations between subfields. We have attempted to provide a general overview of the field through a series of Venn diagrams in Figure 1.3. At the highest level, the field can be roughly divided into those researchers who study and develop software, those who study and develop hardware, and those who study and analyze the interaction between autonomous systems (composed of both software and hardware) and humans (Figure 1.3a). There is, of course, a plethora of

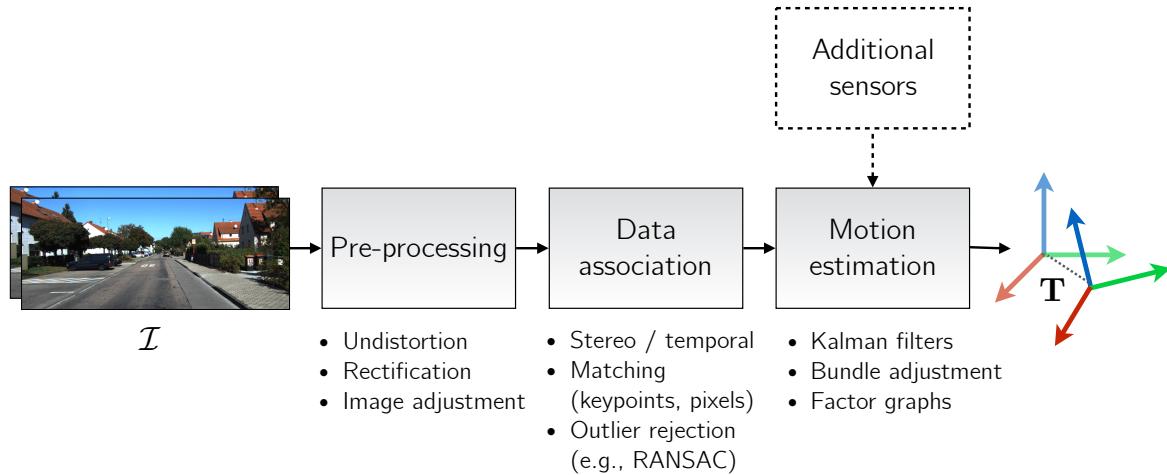


Figure 1.4: A ‘classical’ visual odometry pipeline consists of several distinct components that have interpretable inputs and outputs.

overlap between all three of these rough categories. Within the software realm, there has historically been a distinction between (1) those who study algorithms that deal with the perception of the interoceptive and exteroceptive data, (2) those who study how to use that data to plan action, and (3) those who study how to use those plans to control a system to execute that action (Figure 1.3b).

 Within perception, which is the focus of this dissertation, there are three general directions of research: localization, mapping, and object detection and tracking (Figure 1.3c). Localization and mapping can refer to self-localization, or egomotion estimation, which deals with the problem of estimating the pose of a moving platform through an unknown world, SLAM, simultaneous localization and mapping, which deals with the former problem and mapping simultaneously or finally, it can refer to localization within a known map given some hitherto unseen observation of that environment. The field of object detection and tracking (whether that be static objects like stop signs, or moving objects like humans, animals or vehicles) uses much of the same underlying mathematics as the former two, but has historically been a separate strand of research. Broadly, the overlap of all three of these pursuits within the field of autonomy and robotics is referred to as *state estimation* (Barfoot, 2017). To provide context for the central concept of this thesis, a *pseudo-sensor*, we will first outline a traditional approach to state estimation.

## 1.3 The State of State Estimation

Central to *classical* state estimation algorithms (which, in this context, refers to the bulk of state estimation research published during what Cadena et al. (2016) call the *classical* and *algorithmic-analysis* ages of SLAM research between 1986 and 2015) is the idea of a pipeline. A pipeline consists several distinguishable blocks that have interpretable inputs and outputs. By carefully processing information contained within sensor data, pipelines facilitate the construction of complex state estimation architectures that can fuse observations from sensors of varied modality to create rich models of the external world and infer the state of a mobile platform within it. This dissertation focuses on egomotion estimation: the problem of accurately and consistently estimating the *relative* pose of a moving platform. For this task, a variety of different sensors may be useful (e.g., lidar, stereo cameras, or inertial measurement units), and each may allow for various components of a state estimation pipeline. For cameras, egomotion estimation is typically referred to as *visual odometry* or VO for short. In this thesis, we will largely deal with the improvement of a ‘classical’ VO pipeline— we illustrate its major components in Figure 1.4.

Modern VO pipelines (e.g., Leutenegger et al. (2015); Cvišić and Petrović (2015); Tsotsos et al. (2015)) have achieved impressive localization accuracy on trajectories spanning several kilometres by carefully extracting and tracking sparse visual features (using *hand-crafted* algorithms) across consecutive images. Simultaneously, significant effort has gone to developing localization pipelines that eschew sparse features in favour of *dense* visual data (Alcantarilla and Woodford, 2016; Forster et al., 2014), typically relying on *loss functions* that use direct pixel intensities.

In the last five years, a significant part of the visual state estimation literature has also focused on the idea of replacing classical pipelines with parametric modelling through deep convolutional neural networks (CNNs) and data-driven training. Although initially developed for image classification (LeCun et al., 2015), CNN-based measurement models have been applied to numerous problems in geometric state estimation (e.g., homography estimation (DeTone et al., 2016), single image depth reconstruction (Garg et al., 2016), camera re-localization (Kendall and Cipolla, 2016), place recognition (Sünderhauf et al., 2015)). A number of recent CNN-based approaches have also tackled the problem of egomotion estimation, often purporting to obviate the need for classical visual localization pipelines by learning pose changes *end-to-end*, directly from image data (e.g., Melekhov et al. (2017), Handa et al. (2016), Oliveira et al. (2017)).

Despite this surge of excitement, significant debate has emerged within the robotics and computer vision communities regarding the extent to which deep models should replace ex-

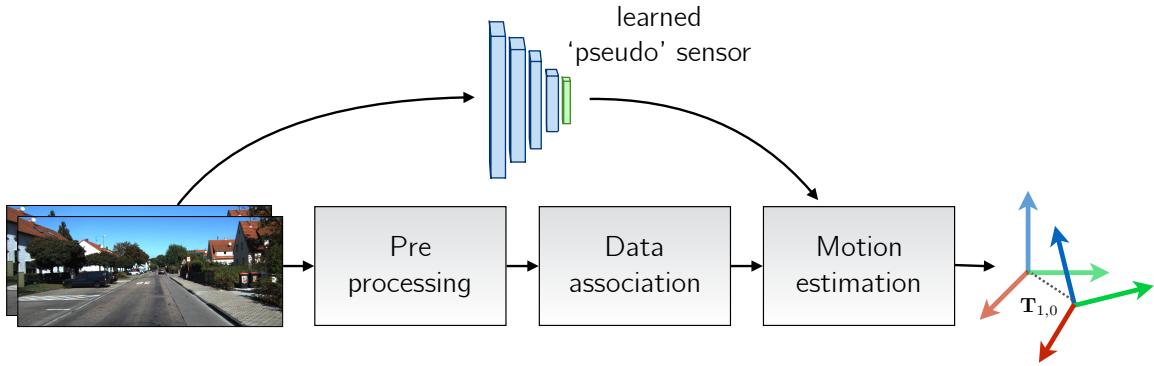


Figure 1.5: A learned *pseudo-sensor* extracts latent information from the same data stream.

isting geometric state estimation algorithms. Owing to their representational power, deep models may move the onerous task of selecting ‘good’ (i.e., robust to environmental vagaries and sensor motion) visual features from the roboticist to the learned model. By design, deep models also provide a straight-forward formulation for using *dense* data while being flexible in their loss function, and taking full advantage of modern computing architecture to minimize run-time. Despite these potential benefits, current deep regression techniques for state estimation often generalize poorly to new environments, come with few analytical guarantees, and provide only point estimates of latent parameters (see Table 1.1 for more details). Indeed, the most accurate visual egomotion pipelines at the time of writing<sup>2</sup> remain largely those based on carefully selected sparse features. Furthermore, there is recent empirical evidence (Zhou et al., 2019) that suggests that designing a pipeline with interpretable components (e.g., optical flow, scene segmentation) is crucial to generalization on various visual tasks.

## 1.4 The Learned Pseudo-Sensor

As state estimation enters the robust-perception age (Cadena et al., 2016), algorithms that work in limited contexts will need to be adapted and augmented to ensure they can operate over longer time-periods, and through sundry environments. Towards this end, we introduce the paradigm of the *learned pseudo-sensor*. Learned pseudo-sensors allow one to retain the benefits of classical state estimation pipelines while leveraging the representational power of modern data-drive learning techniques. Instead of completely replacing the classical pipeline, herein we present four ways in which machine learning can be used to train

<sup>2</sup>Based on the KITTI Odometry benchmark leaderboard at [http://www.cvlibs.net/datasets/kitti/eval\\_odometry.php](http://www.cvlibs.net/datasets/kitti/eval_odometry.php).

Table 1.1: A comparison of pipelines and end-to-end deep models for visual egomotion estimation.

	<b>Classical Pipelines</b>	<b>Deep Models</b>
<i>Maturity</i>	Decades of literature & domain knowledge	Nascent with few uses in mobile autonomy
<i>Interpretability</i>	Good, each component has interpretable input and output	Poor, often with no interpretable intermediate outputs
<i>Uncertainty</i>	Foundational to <i>probabilistic robotics</i>	Few nascent methods (Monte-carlo Dropout ( <a href="#">Gal and Ghahramani, 2016b</a> ), Bootstrap ( <a href="#">Osband et al., 2016</a> ))
<i>Robustness</i>	Empirically generalizable ( <a href="#">Zhou et al., 2019</a> )	Highly dependant on training data
<i>Flexibility</i>	Limited by ingenuity of designer	Limited by training data

a hyper-parametric model that extracts latent information from an existing visual and inertial data stream. By fusing the output of these sensors with the output of the pipeline, we can make the final egomotion estimates more accurate and more robust to difficult-to-model effects (Figure 1.5). To accomplish this fusion, we rely on two approaches. The first, which is used in Predictive Robust Estimation (PROBE, Chapter 4), treats the pseudo-sensor as a heteroscedastic noise model that can be incorporated into a maximum-likelihood loss. The uncertainty quantification provided by this pseudo-sensor is used to re-weight a predictively robust loss function which can then be minimized through traditional non-linear optimization routines. The second approach (used by Sun-BCNN, DPC-Net, and HydraNet, Chapters 5 to 7 respectively) produces geometric quantities (probabilistic estimates of an illumination source, SE(3) corrections to existing egomotion estimates, and independent probabilistic rotation estimates, respectively), that can be fused with the original pipeline through a factor graph optimization routine.

## 1.5 Original Contributions

This dissertation consists of several published contributions under the umbrella of *learned pseudo-sensors* that improve a canonical visual egomotion pipeline. Before detailing each pseudo-sensor, we present some mathematical foundations (Chapter 2) and a common base-

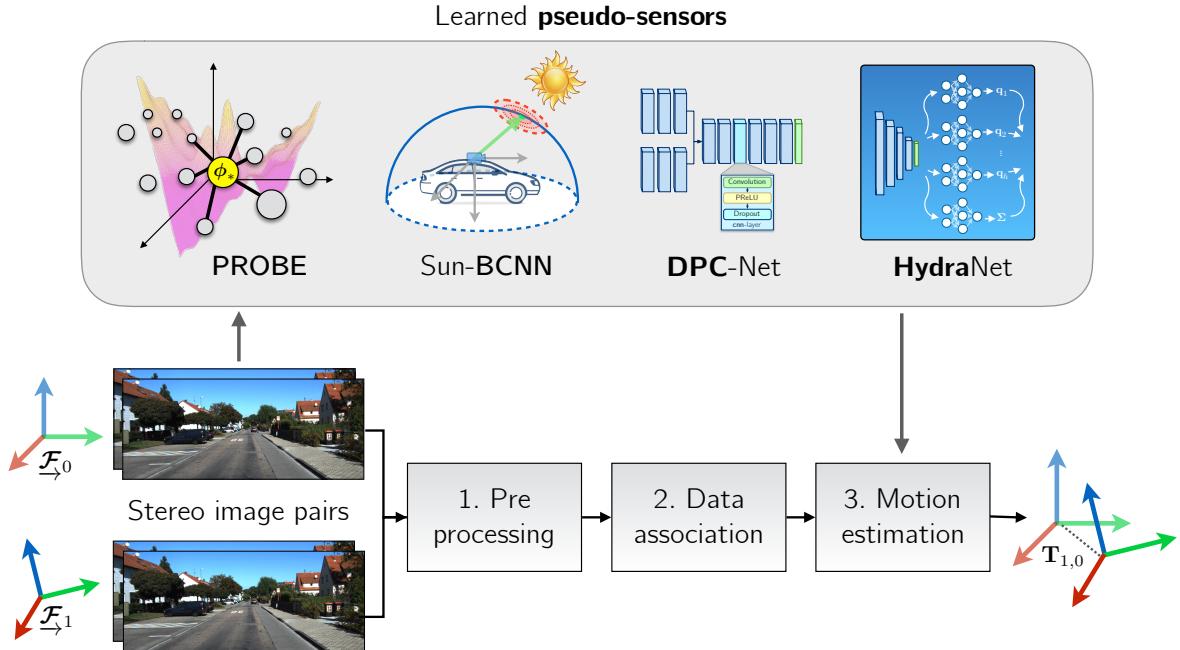


Figure 1.6: This dissertation details four examples of *pseudo-sensors* that improve 'classical' egomotion estimation through data-driven learning.

line for an indirect stereo visual odometry pipeline (Chapter 3) which all four methods build upon. In total, there are two journal papers and five conference papers associated with our work. Below, we briefly summarize each of the pseudo-sensors and list the publications that are associated with each.

### 1. Predictive Robust Estimation (PROBE),

Predictive Robust Estimation (Chapter 4, Appendix A) uses the method of Generalized Kernel estimation (Vega-Brown et al., 2014) to derive an efficient Bayesian model for stereo tracking uncertainty based a set of training reprojection errors. By setting a prior on covariance, we derive a robust loss that can be predictively scaled to improve the accuracy and consistency of an indirect stereo visual odometry pipeline. PROBE is associated with three publications:

- Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3668–3675, Hamburg, Germany
- Peretroukhin, V., Clement, L., and Kelly, J. (2015b). Get to the point: Active covariance scaling for feature tracking through motion blur. In *Proceedings of the*

*IEEE International Conference on Robotics and Automation Workshop on Scaling Up Active Perception*, Seattle, Washington, USA

- Peretroukhin, V., Vega-Brown, W., Roy, N., and Kelly, J. (2016). PROBE-GK: Predictive robust estimation using generalized kernels. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 817–824

## 2. Sun-BCNN: Learned sun sensor

Sun-BCNN (Chapter 5) is a technique to infer a probabilistic estimate of the direction of the sun from a single RGB image using a Bayesian Convolutional Neural Networks (BCNN). The method works much like dedicated sun sensors (Lambert et al., 2012), but requires no additional hardware, and can provide mean and covariance estimates that can be readily incorporated into existing visual odometry frameworks. It is associated with three publications listed below. Initial exploratory work was published at ISER 2016, and the BCNN improvement was presented at ICRA 2017. An additional journal paper summarizing the work of the prior two papers, adding data from the Canadian High Arctic and Oxford, and investigating the effect of cloud cover and transfer learning was published in the International Journal of Robotics’ Research, Special Issue on Experimental Robotics at the end of 2017. 

- Clement, L., Peretroukhin, V., and Kelly, J. (2017). Improving the accuracy of stereo visual odometry using visual illumination estimation. In Kulic, D., Nakamura, Y., Khatib, O., and Venture, G., editors, *2016 International Symposium on Experimental Robotics*, volume 1 of *Springer Proceedings in Advanced Robotics*, pages 409–419. Springer International Publishing, Berlin Heidelberg. Invited to Journal Special Issue
- Peretroukhin, V., Clement, L., and Kelly, J. (2017). Reducing drift in visual odometry by inferring sun direction using a bayesian convolutional neural network. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA’17)*, pages 2035–2042, Singapore
- Peretroukhin, V., Clement, L., and Kelly, J. (2018). Inferring sun direction to improve visual odometry: A deep learning approach. *International Journal of Robotics Research*, 37(9):996–1016

## 3. DPC-Net: Learned pose corrections

Deep Pose Correction (Chapter 6) is an approach to improving egomotion estimates through SE(3) pose corrections learned through deep regression with a supervised loss based on Lie theory. DPC takes as its starting point an efficient, classical localization

algorithm that computes high-rate pose estimates. To it, it adds a Deep Pose Correction Network (DPC-Net) that learns low-rate, ‘small’ *corrections* from training data that are then fused with the original estimates. DPC-Net does not require any modification to an existing localization pipeline, and can learn to correct multi-faceted errors from estimator bias, sensor mis-calibration or environmental effects. It is associated with a journal publication:

- Peretroukhin, V. and Kelly, J. (2018). DPC-Net: Deep pose correction for visual localization. *IEEE Robotics and Automation Letters*, 3(3):2424–2431.

#### 4. HydraNet: Learned probabilistic rotation estimation

Finally, HydraNet (Chapter 7) is a multi-headed network structure that can regress probabilistic estimates of rotation (elements of the matrix Lie group,  $\text{SO}(3)$ ) that account for both aleatoric and epistemic uncertainty. This uncertainty can then be used to fuse the output of HydraNet with the output of classical egomotion pipelines in a probabilistic factor graph formulation. It is associated with one publication:

- Peretroukhin, V., Wagstaff, B., and Kelly, J. (2019). Deep probabilistic regression of elements of  $\text{SO}(3)$  using quaternion averaging and uncertainty injection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’19) Workshop on Uncertainty and Robustness in Deep Visual Learning*, pages 83–86, Long Beach, California, USA.

# Chapter 2

## Mathematical Foundations

By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and, in effect, increases the mental power of the race.

---

ALFRED NORTH WHITEHEAD

### 2.1 Coordinate Frames

Before we can present the main contributions of this thesis, it will be useful to first outline the notation and mathematical foundations that underly the work. Throughout this thesis, we largely follow the notation of [Barfoot \(2017\)](#) when dealing with three-dimensional rigid-body kinematics.

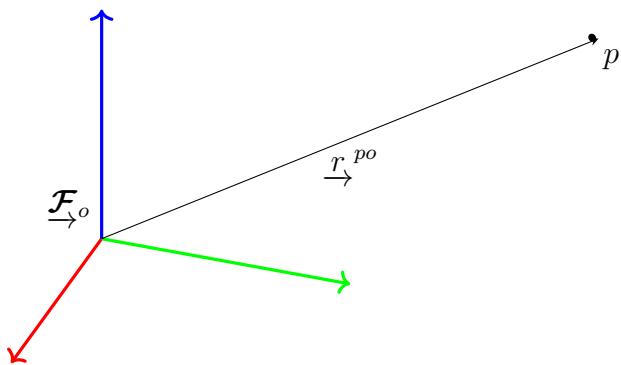


Figure 2.1: A position vector expressed in a coordinate frame.

We refer to a three-dimensional position vector,  $\underline{r}^{po}$ , as one that originates at the origin of a coordinate reference frame,  $\underline{\mathcal{F}}_o$ , and terminates at the point  $p$ . This geometric quantity has

the numerical coordinates  $\mathbf{r}_o^{po}$  when expressed in  $\underline{\mathcal{F}}_o$ . Often, we will refer to two reference frames such as a world or *inertial* frame,  $\underline{\mathcal{F}}_i$ , and a vehicle frame,  $\underline{\mathcal{F}}_v$ . Rotation matrices or rigid-body transformations that convert coordinates from  $\underline{\mathcal{F}}_i$  to  $\underline{\mathcal{F}}_v$  will be represented as  $\mathbf{T}_{vi}$ , and  $\mathbf{C}_{vi}$ <sup>1</sup>, respectively.

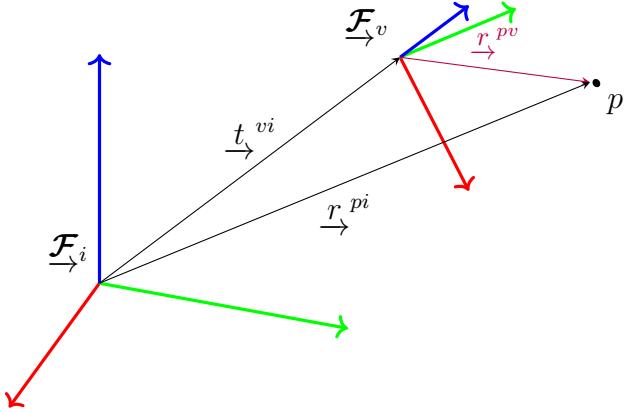


Figure 2.2: Two common references frames used throughout this thesis.

## 2.2 Rotations

The rotation matrix  $\mathbf{C}$  is a member of the matrix Lie group  $\text{SO}(3)$  (the Special Orthogonal group) and can be defined as a matrix as follows:

$$\text{SO}(3) = \{\mathbf{C} \in \mathbb{R}^{3 \times 3} \mid \mathbf{C}^T \mathbf{C} = \mathbf{1}, \det \mathbf{C} = 1\}. \quad (2.1)$$

### Active vs. Passive

An active (or *alibi*) rotation changes the coordinates of a position directly while implicitly assuming that the reference frame is fixed. A passive (or *alias*) rotation rotates the reference frame. Following Barfoot (2017), all rotation matrices in this dissertation are passive unless otherwise noted.

### Exponential and Logarithmic Maps

Since rotations form a matrix Lie group (we refer the reader to Solà et al. (2018) and Barfoot (2017) for a thorough treatment of Lie groups for state estimation), we can define a surjective

---

<sup>1</sup>We use  $\mathbf{C}$  and not  $\mathbf{R}$  for rotation matrices to avoid confusion with common notation for measurement model covariance.

exponential map<sup>2</sup> from three axis-angle parameters,  $\phi = \phi \mathbf{a}$ ,  $\phi \in \mathbb{R}$ ,  $\mathbf{a} \in S^2$ , to a rotation matrix,  $\mathbf{C}$ :

$$\mathbf{C} = \text{Exp}(\phi) = \exp(\phi^\wedge) = \sum_{n=0}^{\infty} \frac{1}{n!} (\phi^\wedge)^n \quad (2.2)$$

$$= \cos \phi \mathbf{1} + (1 - \cos \phi) \mathbf{a} \mathbf{a}^T + \sin \phi \mathbf{a}^\wedge, \quad (2.3)$$

where the wedge operator  $(\cdot)^\wedge$ <sup>3</sup> is defined as

$$\mathbf{a}^\wedge = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}^\wedge = \begin{bmatrix} 0 & -a_2 & a_1 \\ a_2 & 0 & -a_0 \\ -a_1 & a_0 & 0 \end{bmatrix}. \quad (2.4)$$

Equation (2.3) is known as the Euler-Rodriguez formula and it can also be derived geometrically, starting from Euler's theorem that any rotation can be expressed as an axis of rotation and an angle of rotation about that axis. Although the map in Equation (2.2) is surjective, we can define an inverse map if we restrict its domain to  $0 \leq \phi < \pi$ :

$$\phi = \text{Log}(\mathbf{C}) = \log(\mathbf{C})^\vee = \frac{\phi(\mathbf{C} - \mathbf{C}^T)^\vee}{2 \sin \phi}, \quad (2.5)$$

where  $\phi = \arccos \frac{\text{tr}(\mathbf{C}) - 1}{2}$  and the *vee* operator,  $(\cdot)^\vee : \mathbb{R}^{3 \times 3} \rightarrow \mathbb{R}^3$ , is defined as the unique inverse of the wedge operator  $(\cdot)^\wedge$ . Note Equation (2.5) is undefined at both  $\phi = 0$  and at  $\phi = \pi$ . In the former case, we can use a small-angle approximation and define

$$\text{Log}(\mathbf{C}) \approx (\mathbf{C} - \mathbf{1})^\vee \text{ when } \phi \approx 0. \quad (2.6)$$

The latter case (when  $\phi = \pi$ ) defines the *cut locus* of the space where  $\text{Exp}(\cdot)$  is not a covering map and both  $+\phi$  and  $-\phi$  map to the same  $\mathbf{C}$ . This *cut locus* is related to the idea that any three parameterization of  $\text{SO}(3)$  will have singularities associated with it.

## 2.2.1 Unit Quaternions

Another way (and historically, the original way) to represent a general rotation is to use a unit quaternion,  $\mathbf{q}$ . A unit quaternion has four parameters, a scalar value  $q_\omega$ , and a three-dimensional vector component,  $\mathbf{q}_v$ :

---

<sup>2</sup>We follow Solà et al. (2018) and also define *capitalized* map for notational clarity.

<sup>3</sup>This operator is sometimes also expressed as  $(\cdot)^\times$  or  $[\cdot]_\times$  and is known as the skew-symmetric operator.

$$\mathbf{q} = \begin{bmatrix} q_\omega \\ \mathbf{q}_v \end{bmatrix} \in S^3, \quad (\|\mathbf{q}\| = 1). \quad (2.7)$$

Unit quaternions also form a Lie group (Solà et al., 2018) and lie on a three-dimensional unit sphere within  $\mathbb{R}^4$ . This manifold represents a double cover of  $\text{SO}(3)$  (since both  $\mathbf{q}$  and  $-\mathbf{q}$  represent the same rotation)<sup>4</sup>. As with rotation matrices, we can define a surjective map from three parameters to the group itself,

$$\mathbf{q} = \text{Exp}(\boldsymbol{\phi}) = \begin{bmatrix} \cos \phi/2 \\ \mathbf{a} \sin \phi/2 \end{bmatrix}. \quad (2.8)$$

Similarly, we can also define a logarithmic map as

$$\boldsymbol{\phi} = \text{Log}(\mathbf{q}) = 2\mathbf{q}_v \frac{\arctan(\|\mathbf{q}_v\|, q_\omega)}{\|\mathbf{q}_v\|}. \quad (2.9)$$

To avoid issues with the double cover, we replace  $\mathbf{q}$  with  $-\mathbf{q}$  if  $q_\omega$  is negative before evaluating Equation (2.9). Also note again that Equation (2.9) is undefined when  $\phi = 0$ , but, importantly, we do not face any issues when  $\phi = \pi$  due to the half angle. As with rotation matrices, we can use small angle approximations to define:

$$\text{Log}(\mathbf{q}) \approx \frac{\mathbf{q}_v}{q_\omega} \left( 1 - \frac{\|\mathbf{q}_v\|^2}{3q_\omega^2} \right) \quad \text{when } \phi \approx 0. \quad (2.10)$$

A fantastic summary of the history of rotation parameterizations, unit quaternions and the story of Hamilton and Rodriguez can be found in Altmann (1989).

## 2.3 Spatial Transforms

The rigid body transform  $\mathbf{T}$  is also a member of the matrix Lie group, the Special Euclidian group  $\text{SE}(3)$  and can be defined as a  $4 \times 4$  matrix as follows:

$$\text{SE}(3) = \{ \mathbf{T} = \begin{bmatrix} \mathbf{C} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \in \mathbb{R}^{4 \times 4} \mid \mathbf{C} \in \text{SO}(3), \mathbf{t} \in \mathbb{R}^3 \}. \quad (2.11)$$

As a member of a matrix Lie group, it also admits a surjective exponential map,

---

<sup>4</sup> $\text{SO}(3)$  is diffeomorphic to  $\text{RP}(3)$  (real projective space) which can be made by 'identifying' antipodal points in  $S^3$  (unit sphere in  $\mathbb{R}^4$ ). Since unit quaternions belong to  $S^3$ , they represent a double cover of  $\text{SO}(3)$  since we have not identified  $\mathbf{q}$  and  $-\mathbf{q}$ .

$$\mathbf{T} = \text{Exp}(\boldsymbol{\xi}) = \exp(\boldsymbol{\xi}^\wedge) = \sum_{n=0}^{\infty} \frac{1}{n!} (\boldsymbol{\xi}^\wedge)^n \quad (2.12)$$

where  $\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\rho} \\ \phi \end{bmatrix} \in \mathbb{R}^6$  and the wedge operator is overloaded (following Barfoot (2017)) as follows:

$$\boldsymbol{\xi}^\wedge \triangleq \begin{bmatrix} \boldsymbol{\rho} \\ \phi \end{bmatrix}^\wedge = \begin{bmatrix} \boldsymbol{\phi}^\wedge & \boldsymbol{\rho} \\ \mathbf{0}^T & 0 \end{bmatrix}. \quad (2.13)$$

In practice, we can evaluate the exponential map through the Euler-Rodriguez formula (Equation (2.3)) and by computing the left-Jacobian of  $\text{SO}(3)$ ,  $\mathbf{J}$ ,

$$\mathbf{T} = \text{Exp}\left(\begin{bmatrix} \boldsymbol{\rho} \\ \phi \end{bmatrix}\right) = \begin{bmatrix} \mathbf{C}(\phi) & \mathbf{J}(\phi)\boldsymbol{\rho} \\ \mathbf{0}^T & 1 \end{bmatrix}, \quad (2.14)$$

where

$$\mathbf{J}(\phi) = \frac{\sin \phi}{\phi} \mathbf{1} + (1 - \frac{\sin \phi}{\phi}) \mathbf{a} \mathbf{a}^T + \frac{1 - \cos \phi}{\phi} \mathbf{a}^\wedge. \quad (2.15)$$

### 2.3.1 Applying Transforms

Applying our notation for coordinate frames (and referring back to Section 2.1), a transform,  $\mathbf{T}_{vi}$  can be expressed as

$$\mathbf{T}_{vi} = \begin{bmatrix} \mathbf{C}_{vi} & \mathbf{t}_v^{iv} \\ \mathbf{0}^T & 1 \end{bmatrix}. \quad (2.16)$$

This allows us to use the homogenous point representation for  $\mathbf{r}_i^{pi}$  and express the following relation:



$$\begin{bmatrix} \mathbf{r}_v^{pi} \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{C}_{vi} & \mathbf{t}_v^{iv} \\ \mathbf{0}^T & 1 \end{bmatrix}}_{\mathbf{T}_{vi}} \begin{bmatrix} \mathbf{r}_i^{pi} \\ 1 \end{bmatrix} \quad (2.17)$$

which is numerically equivalent to

$$\mathbf{r}_v^{pi} = \mathbf{C}_{vi} \mathbf{r}_i^{pi} + \mathbf{t}_v^{iv}. \quad (2.18)$$

## 2.4 Perturbations

When solving optimization problems that involve rotations or rigid-body transforms, it is often useful to consider a small *perturbation* about an operating point. By leveraging a core

property of Lie groups (that they are locally ‘Euclidian’), we can convert difficult non-linear problems into ones that have local linear approximations.

Using rotations as an example, we can perturb an operating point,  $\bar{\mathbf{C}} \triangleq \text{Exp}(\bar{\boldsymbol{\phi}})$ , in three different ways:

$$\mathbf{C} = \begin{cases} \text{Exp}(\delta\boldsymbol{\phi}^\ell) \bar{\mathbf{C}} & \text{left perturbation,} \\ \text{Exp}(\bar{\boldsymbol{\phi}} + \delta\boldsymbol{\phi}^m) & \text{middle perturbation,} \\ \bar{\mathbf{C}} \text{Exp}(\delta\boldsymbol{\phi}^r) & \text{right perturbation.} \end{cases} \quad (2.19)$$

We can relate all the left and middle perturbations through the left Jacobian of  $\text{SO}(3)$  with the following useful identity,

$$\text{Exp}((\boldsymbol{\phi} + \delta\boldsymbol{\phi}^m)) \approx \text{Exp}(\mathbf{J}(\boldsymbol{\phi})\delta\boldsymbol{\phi}^m) \text{Exp}(\boldsymbol{\phi}). \quad (2.20)$$

From this it follows that  $\delta\boldsymbol{\phi}^\ell \approx \mathbf{J}(\boldsymbol{\phi})\delta\boldsymbol{\phi}^m$  and elucidates why  $\mathbf{J}$  is called the *left* Jacobian.

In this thesis, we will use the left and middle perturbations when appropriate. Using small angle approximations, the Euler-Rodriguez formula (Equation (2.3)) yields  $\text{Exp}(\delta\boldsymbol{\phi}) \approx \mathbf{1} + (\delta\boldsymbol{\phi})^\wedge$ , which allows us to write the useful formula for the left perturbation:

$$\mathbf{C} = (\mathbf{1} + (\delta\boldsymbol{\phi}^\ell)^\wedge)\bar{\mathbf{C}}. \quad (2.21)$$

Similarly, we can write analogous expressions for a rigid body transform,  $\mathbf{T} \in \text{SE}(3)$ , as composed of an operating point  $\bar{\mathbf{T}} \triangleq \text{Exp}(\bar{\boldsymbol{\xi}})$ , and a small perturbation about that operating point:

$$\mathbf{T} = \begin{cases} \text{Exp}(\delta\boldsymbol{\xi}^\ell) \bar{\mathbf{T}} & \text{left perturbation,} \\ \text{Exp}(\bar{\boldsymbol{\xi}} + \delta\boldsymbol{\xi}^m) & \text{middle perturbation,} \\ \bar{\mathbf{T}} \text{Exp}(\delta\boldsymbol{\xi}^r) & \text{right perturbation.} \end{cases} \quad (2.22)$$

Now, we can also note a similar identity for  $\text{SE}(3)$ ,

$$\text{Exp}((\boldsymbol{\xi} + \delta\boldsymbol{\xi}^m)) \approx \text{Exp}((\mathcal{J}(\boldsymbol{\xi})\delta\boldsymbol{\xi}^m)) \text{Exp}(\boldsymbol{\xi}), \quad (2.23)$$

where  $\mathcal{J}$  is the left Jacobian of  $\text{SE}(3)$  and defined as

$$\mathcal{J}(\boldsymbol{\xi}) \triangleq \begin{bmatrix} \mathbf{J}(\boldsymbol{\phi}) & \mathbf{Q}(\boldsymbol{\xi}) \\ \mathbf{0} & \mathbf{J}(\boldsymbol{\phi}) \end{bmatrix}, \quad (2.24)$$

where  $\mathbf{Q}(\boldsymbol{\xi})$  can be evaluated analytically (see Barfoot (2017)). This again allows us to write  $\delta\boldsymbol{\xi}^\ell \approx \mathcal{J}(\boldsymbol{\xi})\delta\boldsymbol{\xi}^m$  and form a similar expression,

$$\mathbf{T} = (\mathbf{1} + (\delta\boldsymbol{\xi}^\ell)^\wedge)\bar{\mathbf{T}}. \quad (2.25)$$

To derive locally linear systems from sets of rigid-body transforms, or ‘poses’, we can apply Equation (2.25). To update an operating point, we solve for  $\delta\boldsymbol{\xi}^\ell$  and then use the constraint-sensitive update  $\mathbf{T} \leftarrow \text{Exp}(\delta\boldsymbol{\xi}^\ell)\bar{\mathbf{T}}$ .

## 2.5 Uncertainty

We can also use perturbation theory to implicitly define uncertainty on constrained manifolds (see [Barfoot and Fur-gale \(2014\)](#) for a thorough discussion).

Given a concentrated normal density,  $\delta\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \Sigma_{6 \times 6})$ , we can *inject* this unconstrained density onto the Lie group through left perturbations about some mean:

$$\mathbf{T} = \text{Exp}(\delta\boldsymbol{\xi})\bar{\mathbf{T}} \quad (2.26)$$

This allows us to keep track of a random variable,  $\mathbf{T}$ , by keeping its mean in group form,  $\bar{\mathbf{T}}$ , while its second statistical moment is stored as a standard  $6 \times 6$  covariance matrix,  $\Sigma$ .

We can define an analogous density for rotation matrices given normal densities over rotation perturbations  $\delta\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \Sigma_{3 \times 3})$ ,

$$\mathbf{C} = \text{Exp}(\delta\boldsymbol{\phi})\bar{\mathbf{C}}, \quad (2.27)$$

and also, for unit quaternions,

$$\mathbf{q} = \text{Exp}(\delta\boldsymbol{\phi}) \otimes \bar{\mathbf{q}} \quad (2.28)$$

where  $\otimes$  refers to the standard quaternion product operator [Sola \(2017\)](#).

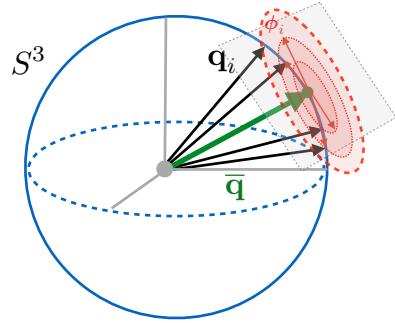


Figure 2.3: We can define uncertainty in the left tangent space of a mean element of a Lie group (here illustrated for unit quaternions).

# Chapter 3

## Classical Visual Odometry

Eventually, my eyes were opened, and I really understood nature.

— CLAUDE MONET

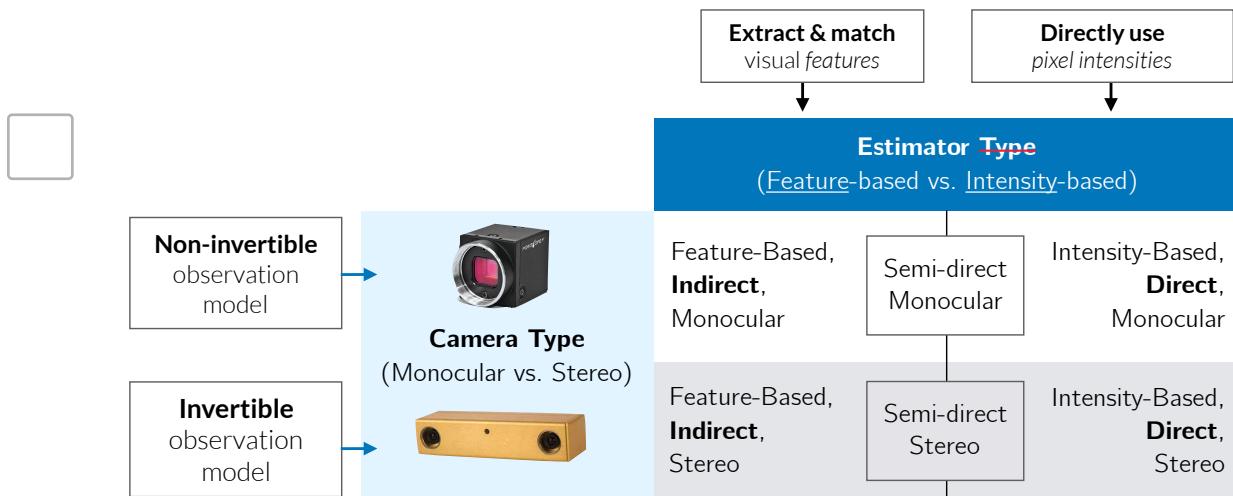


Figure 3.1: A taxonomy of different types of visual odometry.

Visual odometry (VO) has a rich history in mobile robotics and computer vision. As this dissertation largely deals with the improvement of a baseline visual odometry pipeline, we first outline the components of what we have chosen to be a canonical VO system. For a seminal tutorial on visual odometry and its more general cousin, visual SLAM, we refer the reader to two seminal papers: [Scaramuzza and Fraundorfer \(2011\)](#) and [Cadena et al. \(2016\)](#).

## 3.1 A Taxonomy of VO

VO can be largely divided along two dimensions (Figure 3.1): (1) the type of camera used to capture images (monocular vs. stereo) and (2) the type of data association used to compute motion estimates (indirect, or feature-based vs. direct, or pixel intensity-based).

**Monocular vs. Stereo:** Monocular VO methods use a single camera to infer motion and can use a single compact, low-power vision sensor. They do not require any extrinsic calibration but must rely on known visual cues or external information (e.g., wheel odometry, inertial measurements) to provide metric egomotion estimates. Conversely, stereo VO methods use a stereo camera to triangulate objects with metric scale. This allows stereo VO to provide metrically-accurate egomotion estimates. However, stereo methods rely on accurate extrinsic calibration, and their ability to resolve depth is limited by the baseline distance between the stereo pair and by the quality of stereo matches (which can be degraded by self-similar textures, occlusions, and foreshortening effects).

**Direct vs. Indirect:** The second distinction is based on the type of data association used to match sequential images and infer motion. Direct methods make the assumption of brightness constancy, and attempt to *directly* maximize the similarity of pixel intensities. Indirect methods, however, rely on image features detectors to extract a set of salient landmarks, and then match these landmarks across images (typically through some sort of invariant descriptor).

## 3.2 A Classical VO Pipeline

In this thesis, we apply our learned pseudo-sensors to a baseline stereo, indirect visual odometry pipeline (Figure 3.2) largely based on the work of [Furgale \(2011\)](#). We choose this baseline system for its computational efficiency and robustness. We briefly summarize the main components of the pipeline here.

### 3.2.1 Preprocessing

During preprocessing, we use a lens model (assumed to be known apriori) to undistort each stereo image. Further, using the camera extrinsic parameters (also assumed to be known), we *rectify* the stereo pair such that the images can be assumed to come from two cameras whose principal axes are parallel (Figure 3.3). Finally, we also assume that the stereo camera intrinsics are known a priori or compute them through a calibration process ([Furgale et al., 2013](#)).

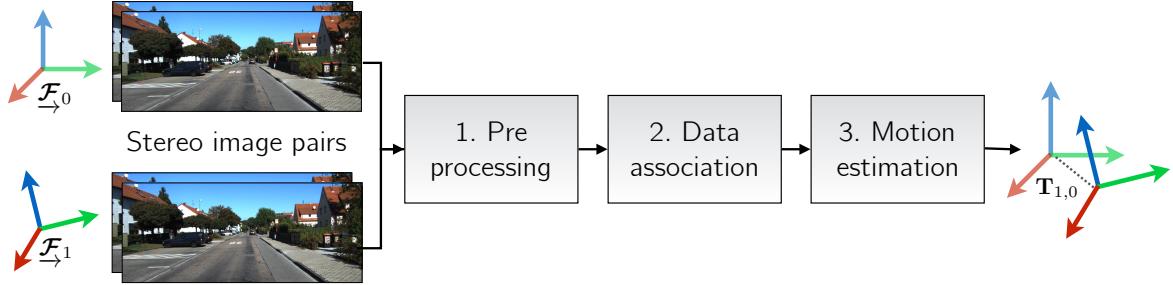
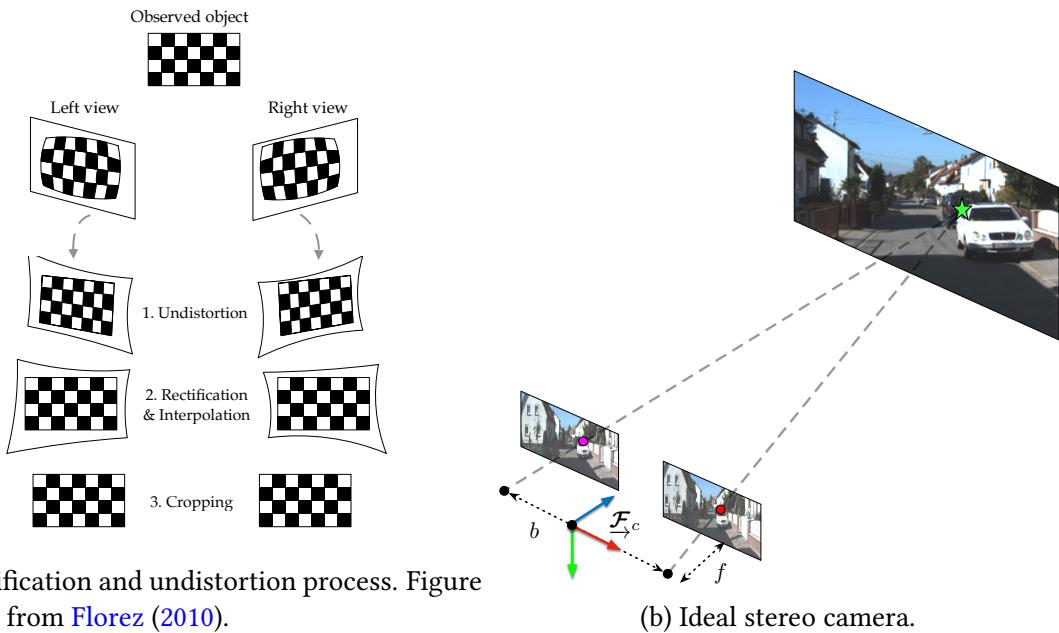


Figure 3.2: A ‘classical’ stereo visual odometry pipeline consists of several distinct components that have interpretable inputs and outputs.



(a) Rectification and undistortion process. Figure adapted from Florez (2010).

(b) Ideal stereo camera.

Figure 3.3: We pre-process stereo images (left) to simulate an ideal stereo camera (right).

### 3.2.2 Data Association

#### Feature Extraction and Matching

In this thesis, we focus on indirect stereo visual odometry for its computational efficiency. Although a number of different types of indirect feature extraction and matching methods can be used towards this end, we choose to use the `viso2` (Geiger et al., 2011) image feature extraction and matching algorithm as it is especially designed for sequential feature matching. In `viso2`, features are extracted using blob and corner masks with non-minimum and non-maximum suppression. Unlike other features detectors that do not assume a particular camera motion, `viso2` assumes a smooth camera trajectory that permits fast matching through a

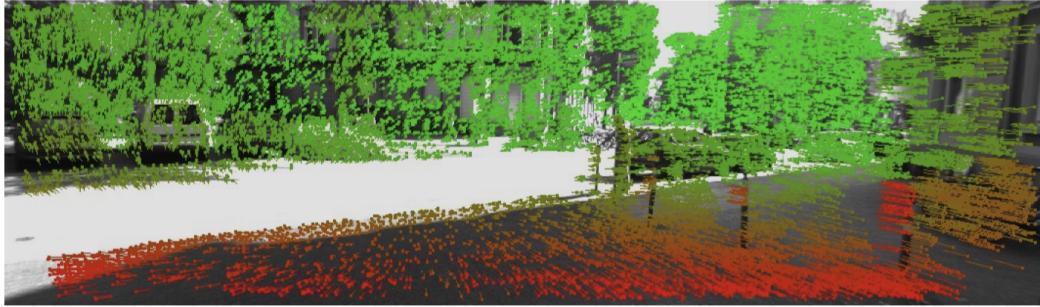


Figure 3.4: Feature tracking using libviso2, taken from [Geiger et al. \(2011\)](#). Colours correspond to depth.

simple sum-of-absolute-difference error metric based on Sobel filter responses. Features are matched across a stereo-pair and forward in time to ensure that a single feature exists across two consecutive stereo camera poses.

Each extracted feature corresponds to a point in space, expressed in homogeneous coordinates in the camera frame as  $\mathbf{p}_{i,c} := [p_1 \ p_2 \ p_3 \ p_4]^T \in \mathbb{P}^3$ . Given our intrinsics and extrinsic calibration parameters, our idealized stereo-camera model,  $\mathbf{f}$ , projects a landmark expressed in homogeneous coordinates into image space, so that  $\mathbf{y}_{i,c}$ , the stereo pixel coordinates of landmark  $i$  in the camera frame, are given by

$$\mathbf{y}_{i,c} = \begin{bmatrix} u_l \\ v_l \\ d \end{bmatrix} = \mathbf{f}(\mathbf{p}_{i,c}) = \mathbf{f}\left(\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}\right) = \mathbf{M} \frac{1}{z} \mathbf{p}_{i,c}, \quad (3.1)$$

where

$$\mathbf{M} = \begin{bmatrix} f & 0 & c_u & 0 \\ 0 & f & c_v & 0 \\ 0 & 0 & 0 & fb \end{bmatrix}. \quad (3.2)$$

Here,  $\{c_u, c_v\}$ ,  $f$ , and  $b$  are the principal points, focal length and baseline of the stereo camera respectively (computed through intrinsic calibration) and  $d \triangleq u_l - u_r$  is the *disparity* of the feature. Note that in this formulation, the stereo camera frame is in the left optical centre.

## Outlier Rejection

To filter out any ~~residual~~ outlier matches, we use a three-point random sample consensus algorithm (RANSAC, [Fischler and Bolles \(1981\)](#)) based on an analytic solution to the six degree-

of-freedom motion ([Umeyama, 1991](#)). 

### 3.2.3 Maximum Likelihood Motion Solution

Finally, we compute the rigid-body transform between two stereo camera frames using maximum likelihood estimation. We define the rigid-body transform,  $\mathbf{T}_t \in \text{SE}(3)$ , to be the rigid-body transform between two subsequent stereo camera poses,  $\underline{\mathcal{F}}_{c_0}$  and  $\underline{\mathcal{F}}_{c_1}$ . In other words,

$$\mathbf{T}_t = \mathbf{T}_{c_1 w} \mathbf{T}_{c_0 w}^{-1}, \quad (3.3)$$

where  $\underline{\mathcal{F}}_w$  is a privileged world frame. After data association, we assume we have a set of  $N_t$  matches,  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}_{i=1}^{N_t}$ , between visual landmarks in the subsequent camera frames. For each match, we define an error function,  $\mathbf{e}_i(\mathbf{T}_t)$ , that relates the rigid transform to these stereo feature matches. Throughout this dissertation, we assume that these errors are corrupted by zero-mean independent Gaussian noise with the (potentially heteroscedastic) covariance,  $\Sigma_{i,t}$ ;

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}(\mathbf{0}, \Sigma_{i,t}). \quad (3.4)$$

Under this noise model, the maximum likelihood transform,  $\mathbf{T}_t^*$ , is given by

$$\mathbf{T}_t^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmax}} \prod_{i=1}^{N_t} p(\mathbf{e}_i(\mathbf{T}_t)) = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_i(\mathbf{T}_t)^T \Sigma_{i,t}^{-1} \mathbf{e}_i(\mathbf{T}_t). \quad (3.5)$$

We will define the error function in two different ways.

#### Point Cloud Error

First, we can follow classical approach ([Maimone et al., 2007](#)) and define  $\mathbf{e}_i(\mathbf{T}_t)$  based on a three-dimensional point cloud error. To do this, we invert our stereo camera model to triangulate pairs of points in each frame,  $\mathbf{p}_{i,c_0} = \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})$  and  $\mathbf{p}_{i,c_1} = \mathbf{f}^{-1}(\mathbf{y}_{i,c_1})$ ,

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{D}(\mathbf{p}_{i,c_1} - \mathbf{T}_t \mathbf{p}_{i,c_0}), \quad (3.6)$$

where  $\mathbf{D} = \begin{bmatrix} \mathbf{1}_{3 \times 3} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{3 \times 4}$  converts homogenous coordinates into Euclidian coordinates. We can then follow [Maimone et al. \(2007\)](#) and assume each stereo projection is corrupted by additive Gaussian noise,

$$\mathbf{y}_{i,c} \sim \mathcal{N}(\bar{\mathbf{y}}_{i,c}, \mathbf{R}_{i,c}), \quad (3.7)$$

then we can compute a density on the error function itself through first order noise propagation

tion as

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}(\mathbf{0}, \Sigma_{i,t}), \quad (3.8)$$

where

$$\Sigma_{i,t} = \mathbf{D}\mathbf{G}_{i,c_1}\mathbf{R}_{i,c_1}\mathbf{G}_{i,c_1}^T\mathbf{D}^T + \mathbf{D}\mathbf{T}_t\mathbf{G}_{i,c_0}\mathbf{R}_{i,c_0}\mathbf{G}_{i,c_0}^T\mathbf{T}_t^T\mathbf{D}^T \quad (3.9)$$

with  $\mathbf{G}_{i,c} = \frac{\partial \mathbf{f}^{-1}}{\partial \mathbf{y}} \Big|_{\mathbf{y}_{i,c}}$ .

## Reprojection Error

Alternatively, we can represent reprojection errors in the second frame directly as

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})), \quad (3.10)$$

and assume the following simple noise model

$$\mathbf{e}_i(\mathbf{T}_t) \sim \mathcal{N}(\mathbf{0}, \Sigma_{i,t}) = \mathcal{N}(\mathbf{0}, \mathbf{R}_{i,t}), \quad (3.11)$$

where we abuse notation (slightly) and replace the index for the camera frames  $c_0$  or  $c_1$  with  $t$  to indicate that this covariance refers to the reprojection error that involves both sets of features.

Importantly, [Sibley et al. \(2007\)](#) show that using reprojection error (as compared to 3D point cloud error) results in less biased estimates for long-range stereo triangulation. Consequently, we favour this latter formulation in the large majority of our work (the one exception being the initial work on isotropic PROBE described in [Appendix A](#)).

## Solution via Gauss-Newton Optimization

In either case, we have now defined a weighted nonlinear least squares problem which can be solved iteratively using standard techniques. For our purposes, we opt to use Gauss-Newton optimization and follow [Barfoot \(2017\)](#) to optimize constrained poses.

Namely, at a given iteration  $n$ , we linearize the error function  $\mathbf{e}_i(\mathbf{T}_t)$ , about an operating point  $\mathbf{T}_t^{(n)} \in \text{SE}(3)$ , which results in a quadratic approximation to [Equation \(A.3\)](#). To linearize, we consider the left perturbations  $\delta\xi \in \mathbb{R}^6$  represented in exponential coordinates:

$$\mathbf{T}_t = \text{Exp}(\delta\xi) \mathbf{T}_t^{(n)} \approx (1 + \delta\xi^\wedge) \mathbf{T}_t^{(n)}. \quad (3.12)$$

This allows us to transform Equation (A.3) into a linear least squares objective in  $\delta\xi$ :

$$\mathcal{L}(\delta\xi) = \frac{1}{2} \sum_{i=1}^{N_t} (\mathbf{e}_i - \mathbf{J}_i \delta\xi)^T \Sigma_i^{-1} (\mathbf{e}_i - \mathbf{J}_i \delta\xi) \quad (3.13)$$

where  $\mathbf{J}_i = \left. \frac{\partial \mathbf{e}_i}{\partial \delta\xi} \right|_{\mathbf{T}_t^{(n)}}$ ,  $\mathbf{e}_i = \mathbf{e}_i(\mathbf{T}_t^{(n)})$ , and  $\Sigma_i = \Sigma_{i,t}(\mathbf{T}_t^{(n)})$ . The minimum to this objective can be solved for analytically by solving the normal equations. This results in the optimal parameters,

$$\delta\xi^* = \left( \sum_{i=1}^{N_t} \mathbf{J}_i^T \Sigma_i^{-1} \mathbf{J}_i \right)^{-1} \sum_{i=1}^{N_t} \mathbf{J}_i^T \Sigma_i^{-1} \mathbf{e}_i. \quad (3.14)$$

Given  $\delta\xi^*$ , we can update the operating point using the constraint-sensitive update

$$\mathbf{T}^{(n+1)} = \text{Exp}(\delta\xi^*) \mathbf{T}^{(n)}, \quad (3.15)$$

and iterate until convergence.

There are many reasonable choices for both the initial transform  $\mathbf{T}^{(0)}$  and for the conditions under which we terminate iteration. For most visual odometry applications, it suffices to initialize the estimated transform to identity, and iteratively perform the update given by eq. (3.15) until we see a relative change in the squared error of less than one percent after an update.

### 3.3 Robust Estimation

Since Equation (3.13) assigns cost values that grow quadratically with measurement error, it is very sensitive to outlier measurements. A common solution to this problem is to replace the quadratic loss function with one that is less sensitive to large measurement errors (MacTavish and Barfoot, 2015). These robust cost functions are collectively known as M-estimators<sup>1</sup>, and many variants exist. Each uses a re-weighting function,  $\rho(\cdot)$ , to define a new optimization problem,

$$\mathbf{T}_{\text{RLS}}^* = \underset{\mathbf{T} \in \text{SE}(3)}{\text{argmin}} \sum_{i=1}^N \rho \left( \sqrt{\mathbf{e}_i^T \Sigma_i^{-1} \mathbf{e}_i} \right) = \underset{\mathbf{T} \in \text{SE}(3)}{\text{argmin}} \sum_{i=1}^N \rho(\epsilon_i) = \underset{\mathbf{T} \in \text{SE}(3)}{\text{argmin}} \mathcal{L}_{\text{RLS}}(\mathbf{T}), \quad (3.16)$$

---

<sup>1</sup>M, for *maximum-likelihood-type* since they generalize the basic maximum likelihood solution (Barfoot, 2017).

where we have defined  $\epsilon_i \triangleq \sqrt{\mathbf{e}_i^T \Sigma_i^{-1} \mathbf{e}_i}$ . The basic idea with M-estimation is to use a  $\rho(\cdot)$  that reduces the influence of large  $\epsilon$  below that of the quadratic  $\rho(\epsilon) = \frac{1}{2}\epsilon^2$ . There are several examples of such functions, including,

$$\rho(\epsilon) = \begin{cases} \frac{c^2}{2} \log \left( 1 + \frac{\epsilon^2}{c^2} \right) & \text{Cauchy,} \\ \frac{1}{2} \frac{\epsilon^2}{c^2 + \epsilon^2} & \text{Geman-McClure (Geman et al., 1992),} \\ \begin{cases} \frac{\epsilon^2}{2} & \text{if } \epsilon < c \\ c\epsilon - \frac{c^2}{2} & \text{if } \epsilon \geq c \end{cases} & \text{Huber (Huber, 1964).} \end{cases} \quad (3.17)$$

where the constant  $c$  can be set with reference to asymptotic efficiency relative to a unit Gaussian (Holland and Welsch, 1977). To solve Equation (3.16), it is common in the literature to apply the technique of *iteratively reweighted least squares* (IRLS) (Holland and Welsch, 1977). To do this, we define a new non-linear least squares minimization problem,

$$\mathbf{T}_{\text{IRLS}}^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^N \mathbf{e}_i^T \mathbf{M}_i \mathbf{e}_i = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \mathcal{L}_{\text{IRLS}}(\mathbf{T}) \quad (3.18)$$

where we define these new weights,  $\mathbf{M}_i$ , based on an *influence function*,  $\psi(\cdot)$  as

$$\mathbf{M}_i = \frac{1}{\epsilon_i} \underbrace{\frac{\partial \rho}{\partial \epsilon} \Big|_{\epsilon_i}}_{\psi(\cdot)} \Sigma_i^{-1}, \quad (3.19)$$

and solve it using the Gauss-Newton approach presented in Section 3.2.3. We claim that upon convergence,  $\mathbf{T}_{\text{IRLS}}^*$  will also minimize Equation (3.16). To see why, consider that

$$\frac{\partial \mathcal{L}_{\text{RLS}}}{\partial \delta \xi} = \sum_i^N \frac{\partial \rho}{\partial \epsilon_i} \frac{\partial \epsilon_i}{\partial \mathbf{e}_i} \frac{\partial \mathbf{e}_i}{\partial \delta \xi} = \sum_i^N \frac{1}{\epsilon_i} \frac{\partial \rho}{\partial \epsilon_i} \mathbf{e}_i^T \Sigma_i^{-1} \frac{\partial \mathbf{e}_i}{\partial \delta \xi}, \quad (3.20)$$

where he have used the fact that  $\frac{\partial \epsilon_i}{\partial \mathbf{e}_i} = \frac{1}{\epsilon_i} \mathbf{e}_i^T \Sigma_i^{-1}$ . Now using our definition of  $\mathbf{M}_i$ , we can write,

$$\frac{\partial \mathcal{L}_{\text{RLS}}}{\partial \delta \xi} = \sum_i^N \mathbf{e}_i^T \underbrace{\frac{1}{\epsilon_i} \frac{\partial \rho}{\partial \epsilon_i} \Sigma_i^{-1} \frac{\partial \mathbf{e}_i}{\partial \delta \xi}}_{\mathbf{M}_i(\mathbf{T})} = \sum_i^N \mathbf{e}_i^T \mathbf{M}_i(\mathbf{T}) \frac{\partial \mathbf{e}_i}{\partial \delta \xi}. \quad (3.21)$$

where we have made the dependence on  $\mathbf{T}$  explicit. We could potentially proceed to set this gradient to 0 and attempt to solve for an optimal update  $\delta \xi$ . However, due to  $\mathbf{M}_i(\mathbf{T})$ , this may be difficult in general. Instead, we note that if we evaluate  $\mathbf{M}_i(\mathbf{T})$  at the current operating

point,  $\mathbf{T}^{(n)}$ , (i.e., we *re-weight* the loss) we are then left with the equivalent normal equations that solve  $\frac{\partial \mathcal{L}_{\text{IRLS}}}{\partial \delta \boldsymbol{\xi}} = \mathbf{0}$ .

Furthermore, upon convergence, our solution to the iteratively re-weighted problem  $\mathbf{T}^{(n)} = \mathbf{T}_{\text{IRLS}}^*$  will also minimize the robust objective Equation (3.16), since we must have that,

$$\left. \frac{\partial \mathcal{L}_{\text{IRLS}}}{\partial \delta \boldsymbol{\xi}} \right|_{\mathbf{T}_{\text{IRLS}}^*} = \left. \frac{\partial \mathcal{L}_{\text{RLS}}}{\partial \delta \boldsymbol{\xi}} \right|_{\mathbf{T}_{\text{IRLS}}^*} = \mathbf{0}. \quad (3.22)$$

## 3.4 Outstanding Issues

There are several outstanding limitations of classical visual odometry pipelines that we can address with learned pseudo-sensors.

Table 3.1: **Data efficiency vs. computational efficiency**

Synopsis	Addressed by
Classical VO pipelines face a difficult-to-optimize trade-off between using all of the information contained within image and while still remaining computationally tractable.	PROBE, DPC-Net, Sun-BCNN, HydraNet

Table 3.2: **Systematic bias**

Synopsis	Addressed by
Stereo visual odometry can incur systematic bias through poor extrinsic or intrinsic calibration, stereo triangulation errors, poor feature <i>spread</i> (i.e., concentration of features on one side of an image), and poor data association due self-similar textures.	DPC-Net

Table 3.3: **Homoscedastic uncertainty**

Synopsis	Addressed by
Stationary, homoscedastic noise in observation models can often reduce the consistency and accuracy of state estimates. This is especially true for complex, inferred measurement models.	PROBE, Sun-BCNN, HydraNet

# Chapter 4

## Predictive Robust Estimation

Information is the resolution of uncertainty.

---

Claude Shannon

### 4.1 Introduction

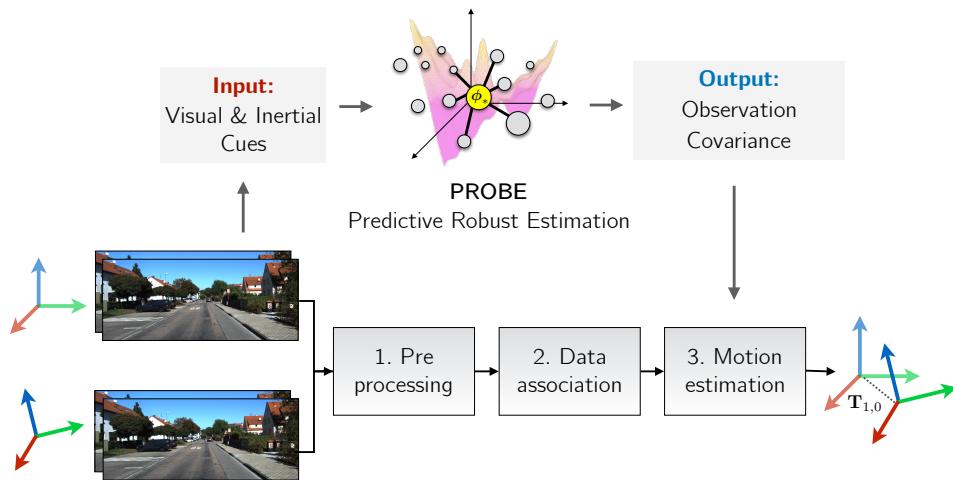


Figure 4.1: PROBE builds a predictive noise model for stereo visual odometry.

The first pseudo-sensor we present is a technique we call PRedictive ROBust Estimation, or PROBE. This approach uses non-parametric learning to build a model for anisotropic observation covariances for a stereo visual odometry pipeline. Namely, we apply the method of Generalized Kernels to a Bayesian treatment of covariance estimation. We show that by assuming a particular covariance prior over re-projection errors, we can then naturally derive

a robust least squares objective that resembles the widely-used Cauchy loss. The parameters of this robust loss are predicted (hence *predictive* robust estimation) for each error term as a function of a prediction space that we define.

PROBE was initially published as a simpler non-Bayesian technique that learned isotropic covariances through a k-nearest-neighbours approach (see Appendix A for more details). The following two publications summarize this initial technique:

1. Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3668–3675, Hamburg, Germany
2. Peretroukhin, V., Clement, L., and Kelly, J. (2015b). Get to the point: Active covariance scaling for feature tracking through motion blur. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Scaling Up Active Perception*, Seattle, Washington, USA

We significantly extended this technique to full anisotropic covariances and generalized kernels in the following publication:

- 
1. Peretroukhin, V., Vega-Brown, W., Roy, N., and Kelly, J. (2016). PROBE-GK: Predictive robust estimation using generalized kernels. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 817–824 .

We will present this latter technique in this chapter.

## 4.2 Motivation

Robot navigation relies on an accurate quantification of sensor noise or uncertainty in order to produce reliable state estimates. In practice, this uncertainty is often fixed for a given sensor and experiment, whether by automatic calibration or by manual tuning. Although a fixed measure of uncertainty may be reasonable in certain static environments, dynamic scenes frequently exhibit many effects that corrupt a portion of the available observations. For visual sensors, these effects include, for example, self-similar textures, variations in lighting, moving objects, and motion blur. Further, there may be useful information available in these observations that would normally be rejected by a fixed-threshold outlier rejection scheme. Ideally, we would like to retain some of these observations in our estimator, while still placing more trust in observations that do not suffer from such effects.

## 4.3 Related Work

There is a large and growing body of work on the problem of deriving accurate, consistent state estimates from visual data. Although our approach to noise modelling is applicable in other domains, for simplicity we focus our attention on the problem of inferring egomotion from features extracted from sequential pairs of stereo images.

Apart from simply rejecting outliers, a number of recent approaches attempt to select the optimal set of features to produce an accurate localization estimate from tracked visual features. For example, [Tsotsos et al. \(2015\)](#) amend Random Sample Consensus (RANSAC) with statistical hypothesis testing to ensure that tracked visual features have normally distributed residuals before including them in the estimator. Unlike our predictive approach, their technique relies on the availability of feature tracks, and [requires scene overlap to work continuously](#). In a different approach, [Zhang and Vela \(2015\)](#) choose an optimally observable feature subset for a monocular SLAM pipeline by selecting features with the highest *informativeness* - a measure calculated based on the observability of the SLAM subsystem. Observability, however, is governed by the 3D location of the features, and therefore cannot predict systematic feature degradation due to environmental or sensor-based effects.

## 4.4 Predictive Robust Estimation for VO

We present a principled, data-driven way to build a noise model for visual odometry. We leverage recent advances in covariance estimation ([Vega-Brown and Roy, 2013](#)) to formulate a predictive robust estimator for a stereo visual odometry pipeline. We frame the traditional non-linear least squares optimization problem as a problem of maximum likelihood estimation with a Gaussian noise model, and infer a distribution over the covariance matrix of the Gaussian noise from a predictive model learned from training data. This results in a Student's  $t$  distribution over the noise, and naturally yields a robust nonlinear least-squares optimization problem. In this way, we can predict, in a principled manner, how informative each visual feature is with respect to the final state estimate, which allows our approach to intelligently weight observations to produce more accurate odometry estimates.

### 4.4.1 Bayesian Noise Model for Visual Odometry

We adopt the motion solution for visual odometry based on reprojection errors presented in Section [3.2.3](#). In brief, this technique assumes independent Gaussian errors on stereo repro-

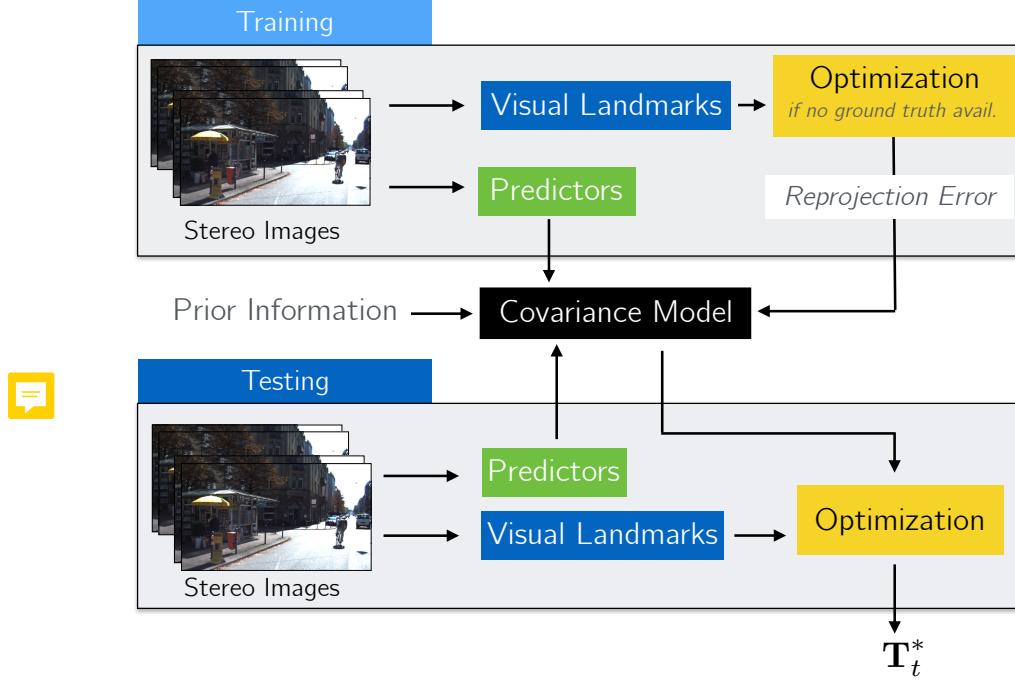


Figure 4.2: PROBE builds a predictive noise model for stereo visual odometry.

jections of a landmark from one frame,  $\underline{\mathcal{F}}_{c_0}$  into a subsequent frame,  $\underline{\mathcal{F}}_{c_1}$ :

$$\mathbf{e}_i(\mathbf{T}_t) = \mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_{i,t}). \quad (4.1)$$

Maximizing the likelihood of these errors is then equivalent to solving the following weighted non-linear least squares objective for  $\mathbf{T}_t \in \text{SE}(3)$  the rigid-body transform that transforms points in  $\underline{\mathcal{F}}_{c_0}$  to those in  $\underline{\mathcal{F}}_{c_1}$ :

$$\mathbf{T}_t^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmax}} \prod_{i=1}^{N_t} p(\mathbf{e}_{i,t}; \mathbf{T}_t, \mathbf{R}_{i,t}) \quad (4.2)$$

$$= \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_{i,t}^T \mathbf{R}_{i,t}^{-1} \mathbf{e}_{i,t}. \quad (4.3)$$

With PROBE, instead of treating  $\mathbf{R}_{i,t}$  as fixed, we build a model for it as a function of some useful *predictor*,  $\phi_{i,t}$ ,

$$\mathbf{R}_{i,t} = \mathbf{R}(\phi_{i,t}). \quad (4.4)$$

Each predictor can be computed based on the stereo track ( $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}$ ) and additional visual<sup>1</sup> and inertial cues, allowing us to model effects like motion blur and self-similar textures.

---

<sup>1</sup>Including potentially data from all four images in the pair of stereo images.

Further, instead of treating the covariance as a point function  $\mathbf{R}(\phi_{i,t})$ , we instead build a non-parametric model of covariance *density* based on a training dataset,  $\mathcal{D}$ ,

$$p(\mathbf{R}_{i,t}) = p(\mathbf{R}|\mathcal{D}, \phi_{i,t}). \quad (4.5)$$

We will seek the transform that then maximizes the posterior predictive distribution of the errors, given this posterior:

$$\mathbf{T}_t^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmax}} \prod_{i=1}^{N_t} \int p(\mathbf{e}_{i,t}; \mathbf{T}_t | \mathbf{R}_{i,t}) p(\mathbf{R}|\mathcal{D}, \phi_{i,t}) d\mathbf{R}. \quad (4.6)$$

Although at first this may seem unwieldy, we present an efficient method for computing the posterior and show that a particular formulation allows it to be marginalized out analytically to arrive at a simple posterior predictive distribution with a straight-forward objective for achieving a maximum likelihood egomotion transform.

#### 4.4.2 Generalized Kernels

To do this, we leverage Generalized Kernel (GK) estimation ([Vega-Brown et al., 2014](#)). GK estimation combines the benefits of kernel density estimation with Bayesian inference. The basic idea is as follows. Consider a **dataset** of inputs,  $\mathbf{x}$ , and outputs,  $\mathbf{y}$ , and a **dataset** of independent observations  $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$ . We are given a new ‘test’ input  $\mathbf{x}^*$ , and are asked to infer the likelihood of observing a given output at this input:

$$p(\mathbf{y}|\mathbf{x}^*, \mathcal{D}). \quad (4.7)$$

If we associate a set of latent parameters,  $\boldsymbol{\pi}$ , with each input  $\mathbf{x}$ , and assume a known likelihood function  $p(\mathbf{y}|\boldsymbol{\pi})$ , we can infer a distribution over  $\boldsymbol{\pi}$  and then marginalize it out to arrive at the desired likelihood

$$p(\mathbf{y}|\mathbf{x}^*, \mathcal{D}) = \int_{\boldsymbol{\pi}} \underbrace{p(\mathbf{y}|\boldsymbol{\pi}^*)}_{\text{Known likelihood function}} \underbrace{p(\boldsymbol{\pi}^*|\mathbf{x}^*, \mathcal{D})}_{\text{Parameter posterior}} d\boldsymbol{\pi}^*. \quad (4.8)$$

This is called the posterior predictive distribution. Using Bayes rule, we can write

$$p(\boldsymbol{\pi}^*|\mathbf{x}^*, \mathcal{D}) \propto \int \left( \prod_{i=1}^N p(\mathbf{y}_i|\boldsymbol{\pi}_i) d\boldsymbol{\pi}_i \right) p(\boldsymbol{\pi}_{1:N}, \boldsymbol{\pi}^*|\mathbf{x}^*, \mathbf{x}_{1:N}) \quad (4.9)$$

The technique of generalized kernels makes the assumption that the parameters  $\boldsymbol{\pi}_{1:N}$  are

conditionally independent given the target parameters,  $\boldsymbol{\pi}^*$ . This gives the distribution:

$$p(\boldsymbol{\pi}_{1:N}, \boldsymbol{\pi}^* | \mathbf{x}^*, \mathbf{x}_{1:N}) = \left( \prod_{i=1}^N p(\boldsymbol{\pi}_i | \boldsymbol{\pi}^* \mathbf{x}^*, \mathbf{x}_i) \right) p(\boldsymbol{\pi}^* | \mathbf{x}^*), \quad (4.10)$$

which combined with Equation (4.9) results in

$$p(\boldsymbol{\pi}^* | \mathbf{x}^*, \mathcal{D}) \propto \prod_{i=1}^N \underbrace{p(\mathbf{y}_i | \boldsymbol{\pi}^*, \mathbf{x}_i, \mathbf{x}^*)}_{\text{Extended likelihood}} \underbrace{p(\boldsymbol{\pi}^* | \mathbf{x}^*)}_{\text{Prior}}. \quad (4.11)$$

Now, the pièce de résistance of generalized kernels is that the *extended* likelihood can be written as function of the known likelihood  $p(\mathbf{y}_i | \boldsymbol{\pi}_i)$  if we assume it is the maximum entropy distribution whose information divergence from the likelihood is bounded by the metric  $\rho(\mathbf{x}^*, \mathbf{x}_i)$ . Specifically, in Vega-Brown et al. (2014), it is shown that under these assumptions, the extended likelihood must have the form:

$$p(\mathbf{y} | \boldsymbol{\pi}^*, \mathbf{x}, \mathbf{x}^*) \propto p(\mathbf{y} | \boldsymbol{\pi})^{k(\mathbf{x}^*, \mathbf{x})}, \quad (4.12)$$

where  $k(\cdot, \cdot)$  is a kernel function<sup>2</sup> that is uniquely defined by  $\rho$ . The intuition behind this is that we expect the extended likelihood to equal the known likelihood if  $\mathbf{x}^* = \mathbf{x}_i$  (and therefore  $\boldsymbol{\pi}^* = \boldsymbol{\pi}_i$ , resulting in  $p(\mathbf{y}_i | \boldsymbol{\pi}^*, \mathbf{x}_i, \mathbf{x}^*) = p(\mathbf{y}_i | \boldsymbol{\pi}_i)$ ) and diverge in some smooth way when  $\mathbf{x}^* \neq \mathbf{x}_i$ . Combining Equation (4.11) with Equation (4.12), we arrive at an expression for the posterior over parameters as

$$p(\boldsymbol{\pi} | \mathbf{x}, \mathcal{D}) \propto \prod_{i=1}^N p(\mathbf{y} | \boldsymbol{\pi})^{k(\mathbf{x}, \mathbf{x}_i)} p(\boldsymbol{\pi} | \mathbf{x}), \quad (4.13)$$

which can be evaluated in closed form for appropriate an appropriate likelihood and prior. Namely, for PROBE, we will assume Gaussian likelihoods for the reprojection errors (and therefore the observations  $\mathbf{y}_{i,c_1}$ ), and inverse Wishart priors for covariance matrices (this will result in inverse Wishart posteriors due to conjugacy). The input,  $\mathbf{x}$ , will be the vector of predictors  $\boldsymbol{\phi}$ .

#### 4.4.3 Generalized Kernels for Visual Odometry

In order to exploit conjugacy to a Gaussian noise model, we formulate our prior knowledge about this function using an inverse Wishart (IW) distribution over positive definite  $d \times d$

---

<sup>2</sup>i.e.,  $k(\mathbf{x}, \mathbf{x}) = 1 \forall \mathbf{x}$  and  $k(\mathbf{x}, \mathbf{x}') \in [0, 1] \forall \mathbf{x}, \mathbf{x}'$ .

matrices (the IW distribution has been used as a prior on covariance matrices in other robotics and computer vision contexts, see for example, ([Fitzgibbon et al., 2007](#))). This distribution is defined by a scale matrix  $\Psi \in \mathbb{R}^{d \times d} \succ 0$  and a scalar quantity called the degrees of freedom  $\nu \in \mathbb{R} > d - 1$ :

$$\begin{aligned} p(\mathbf{R}) &= \text{IW}(\mathbf{R}; \Psi, \nu) \\ &= \frac{|\Psi|^{\nu/2}}{2^{\frac{\nu d}{2}} \Gamma_d(\frac{\nu}{2})} |\mathbf{R}|^{-\frac{\nu+d+1}{2}} \exp\left(-\frac{1}{2} \text{tr}(\Psi \mathbf{R}^{-1})\right). \end{aligned} \quad (4.14)$$

We use the scale matrix to encode our prior estimate of the covariance, and the degrees of freedom to encode our confidence in that estimate. Specifically, if we estimate the covariance  $\mathbf{R}$  associated with predictor  $\phi$  to be  $\hat{\mathbf{R}}$  with a confidence equivalent to seeing  $n$  independent samples of the error from  $\mathcal{N}(\mathbf{0}, \hat{\mathbf{R}})$ , we would choose  $\nu(\phi) = n$  and  $\Psi(\phi) = n\hat{\mathbf{R}}$ . Given a sequence of observations and ground truth transformations,

$$\mathcal{D} = \{\mathcal{I}_t, \mathbf{T}_t\}, \quad t \in [1, N] \quad (4.15)$$

where

$$\mathcal{I}_t = \{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}, \phi_{i,t}\} \quad i \in [1, N_t], \quad (4.16)$$

we can use the procedure of generalized kernel estimation as described above to infer a posterior distribution over the covariance matrix  $\mathbf{R}_*$  associated with some query predictor vector  $\phi_*$ :

$$\begin{aligned} p(\mathbf{R}_* | \mathcal{D}, \phi_*) &\propto \prod_{i,t} \mathcal{N}(\mathbf{e}_{i,t} | \mathbf{0}, \mathbf{R}_*)^{k(\phi_*, \phi_{i,t})} \\ &\quad \times \text{IW}(\mathbf{R}_*; \Psi(\phi_*), \nu(\phi_*)) \end{aligned} \quad (4.17)$$

$$= \text{IW}(\mathbf{R}_*; \Psi_*, \nu_*). \quad (4.18)$$

Here,  $\mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$  as before. The function  $k : \mathbb{R}^M \times \mathbb{R}^M \rightarrow [0, 1]$  is a kernel function which measures the similarity of two points in predictor space. Note also that the posterior parameters  $\Psi_*$  and  $\nu_*$  can be computed in closed form (see [Vega-Brown et al. \(2014\)](#)) as

$$\Psi_* = \Psi(\phi_*) + \sum_{i,t} k(\phi_*, \phi_{i,t}) \mathbf{e}_{i,t} \mathbf{e}_{i,t}^T, \quad (4.19)$$

$$\nu_* = \nu(\phi_*) + \sum_{i,t} k(\phi_*, \phi_{i,t}). \quad (4.20)$$

If we marginalize over the covariance matrix, we find that the posterior predictive distribution is a multivariate Student's  $t$  distribution:

$$p(\mathbf{y}_{i,c_1} | \mathbf{T}_t, \mathbf{y}_{i,c_0}, \mathcal{D}, \phi_{i,t}) \quad (4.21)$$

$$= \int d\mathbf{R}_{i,t} \mathcal{N}(\mathbf{e}_{i,t}; \mathbf{0}, \mathbf{R}_{i,t}) \text{IW}(\mathbf{R}_{i,t}; \Psi_*, \nu_*) \quad (4.22)$$

$$= t_{\nu_* - d + 1} \left( \mathbf{e}_{i,t}; \mathbf{0}, \frac{1}{\nu_* - d + 1} \Psi_* \right) \quad (4.23)$$

$$= \frac{\Gamma(\frac{\nu_* + 1}{2})}{\Gamma(\frac{\nu_* - d + 1}{2})} |\Psi_*|^{-\frac{1}{2}} \pi^{-\frac{d}{2}} (1 + \mathbf{e}_{i,t}^T \Psi_*^{-1} \mathbf{e}_{i,t})^{-\frac{\nu_* + 1}{2}}. \quad (4.24)$$

Given a new landmark and predictor vector, we can infer a noise model by evaluating eqs. (4.19) and (4.20). In order to accelerate this computation, it is helpful to choose a kernel function with finite support: that is,  $k(\phi, \phi') = 0$  if  $\|\phi - \phi'\|_2 > \gamma$  for some  $\gamma$ . Then, by indexing our training data in a spatial index such as a  $k$ -d tree, we can identify the subset of samples relevant to evaluating the sums in eqs. (4.19) and (4.20) in  $\mathcal{O}(\log N + \log N_t)$  time. Algorithm 1 describes the procedure for building this model.

---

**Algorithm 1** Build the covariance model given a sequence of observations,  $\mathcal{D}$ .

---

```

function BUILDCOVARIANCEMODEL( $\mathcal{D}$ )
    Initialize an empty spatial index  $\mathcal{M}$ 
    for all  $\mathcal{I}_t, \mathbf{T}_t$  in  $\mathcal{D}$  do
        for all  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}, \phi_{i,c_0}\}$  in  $\mathcal{I}_t$  do
             $\mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ 
            Insert  $\phi_{i,t}$  into  $\mathcal{M}$  and store  $\mathbf{e}_{i,t}$  at its location
        end for
    end for
    return  $\mathcal{M}$ 
end function

```

---

Once we have inferred a noise model for each landmark in a new image pair, the maximum likelihood optimization problem is given by

$$\mathbf{T}_t^* = \underset{\mathbf{T}_t \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} (\nu_{i,t} + 1) \log (1 + \mathbf{e}_{i,t}^T \Psi_{i,t}^{-1} \mathbf{e}_{i,t}). \quad (4.25)$$

The final optimization problem thus emerges as a nonlinear least squares problem with a rescaled Cauchy-like loss function, with error term  $\mathbf{e}_{i,t}^T (\frac{1}{\nu_{i,t} + 1} \Psi_{i,t})^{-1} \mathbf{e}_{i,t}$  and outlier scale  $\nu_{i,t} + 1$ . This is a common robust loss function which is approximately quadratic in the reprojection error for  $\mathbf{e}_{i,t}^T \Psi_{i,t}^{-1} \mathbf{e}_{i,t} \ll \nu_{i,t} + 1$ , but grows only logarithmically for  $\mathbf{e}_{i,t}^T \Psi_{i,t}^{-1} \mathbf{e}_{i,t} \gg$

$\nu_{i,t} + 1$ . It follows that in the limit of large  $\nu_{i,t}$ —in regions of predictor space where there are many relevant samples—our optimization problem becomes the original least-squares optimization problem.

Solving nonlinear optimization problems with the form of Equation (4.25) is a well-studied and well-understood task, and software packages to perform this computation are readily available. Algorithm 2 describes the procedure for computing the transform between a new image pair, treating the optimization of Equation (4.25) as a subroutine.

---

**Algorithm 2** Compute the transform between two images, given a set,  $\mathcal{I}_t$ , of landmarks and predictors extracted from an image pair and a covariance model  $\mathcal{M}$ .

---

```

function COMPUTETRANSFORM( $\mathcal{I}_t, \mathcal{M}$ )
  for all  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}, \phi_{i,c_0}\}$  in  $\mathcal{I}_t$  do
     $\Psi, \nu \leftarrow \text{INFERNOISEMODEL}(\mathcal{M}, \phi_{i,t})$ 
     $g(\mathbf{T}) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}\mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ 
     $\mathcal{L} \leftarrow \mathcal{L} + (\nu + 1) \log \left( 1 + g(\mathbf{T})^T \Psi^{-1} g(\mathbf{T}) \right)$ 
  end for
  return  $\text{argmin}_{\mathbf{T} \in \text{SE}(3)} \mathcal{L}(\mathbf{T})$ 
end function

function INFERNOISEMODEL( $\mathcal{M}, \phi_*$ )
  NEIGHBORS  $\leftarrow \text{GETNEIGHBORS}(\mathcal{M}, \phi_*, \rho)$             $\triangleright \rho$  is the radius of support of kernel  $k$ 
   $\Psi_* \leftarrow \Psi(\phi_*)$ 
   $\nu_* \leftarrow \nu(\phi_*)$ 
  for  $(\phi_{i,t}, \mathbf{e}_{i,t})$  in NEIGHBORS do
     $\Psi_* \leftarrow \Psi_* + k(\phi_*, \phi_{i,t}) \mathbf{e}_{i,t} \mathbf{e}_{i,t}^T$ 
     $\nu_* \leftarrow \nu_* + k(\phi_*, \phi_{i,t})$ 
  end for
  return  $\Psi_*, \nu_*$ 
end function

```

---

We observe that Algorithm 2 is predictively robust, in the sense that it uses past experiences not just to predict the reliability of a given image landmark, but also to introspect and estimate its own knowledge of that reliability. Landmarks which are not known to be reliable are trusted less than landmarks which look like those which have been observed previously, where “looks like” is defined by our prediction space and choice of kernel.

#### 4.4.4 Inference without ground truth

Algorithm 1 requires access to the true transform between training image pairs. In practice, such ground truth data may be difficult to obtain. In these cases, we can instead formulate a likelihood model  $p(\mathcal{D}' | \mathbf{T}_1, \dots, \mathbf{T}_t)$ , where  $\mathcal{D}' = \{\mathcal{I}_t\}$  is a dataset consisting only of landmarks

and predictors for each training image pair. We can construct a model for future queries by inferring the most likely sequence of transforms for our training images. The likelihood has the following factorized form:

$$p(\mathcal{D}'|\mathbf{T}_{1:T}) \propto \int \prod_{i,t} d\mathbf{R}_{i,t} p(\mathbf{y}_{i,c_1}|\mathbf{y}_{i,c_0}, \mathbf{T}_t, \mathbf{R}_{i,t}) p(\mathbf{R}_{i,t}|\phi_{i,t}, \mathcal{D}, \mathbf{T}_{1:T}). \quad (4.26)$$

We cannot easily maximize this likelihood, since marginalizing over the noise covariances removes the independence of the transforms between each image pair. To render the optimization tractable, we follow previous work ([Vega-Brown and Roy, 2013](#)) and formulate an iterative expectation-maximization (EM) procedure. Given an estimate  $\mathbf{T}_{1:T}^{(n)}$  of the transforms, we can compute the expected log-likelihood conditioned on our current estimate:

$$Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}) = \int \left( \prod_{i,t} d\mathbf{R}_{i,t} p(\mathbf{R}_{i,t}|\mathcal{D}_{\setminus i,t}, \mathbf{T}_{1:T}^{(n)}) \right) \log \prod_{i,t} p(\mathbf{y}_{i,c_1}|\mathbf{y}_{i,c_0}, \mathbf{T}_t, \mathbf{R}_{i,t}). \quad (4.27)$$

This has the effect of rendering the likelihood of each transform to be estimated independently. Moreover, the expected log-likelihood can be evaluated in closed form:

$$Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}) \cong -\frac{1}{2} \sum_{t=1}^T \sum_{i=1}^{N_t} \mathbf{e}_{i,t}^T \left( \frac{1}{\nu_{i,t}^{(n)}} \boldsymbol{\Psi}_{i,t}^{(n)} \right)^{-1} \mathbf{e}_{i,t}. \quad (4.28)$$

The symbol  $\cong$  is used to indicate equality up to an additive constant. We can iteratively refine our estimate by maximizing the expected log-likelihood

$$\mathbf{T}_{1:T}^{(n+1)} = \underset{\mathbf{T}_{1:T} \in \text{SE}(3)^T}{\operatorname{argmax}} Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)}). \quad (4.29)$$

Due to the additive structure of  $Q(\mathbf{T}_{1:T}|\mathbf{T}_{1:T}^{(n)})$ , this takes the form of  $T$  separate nonlinear least-squares optimizations:

$$\mathbf{T}_t^{(n+1)} = \underset{\mathbf{T}_t \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_{i,t}^T \left( \frac{1}{\nu_{i,t}^{(n)}} \boldsymbol{\Psi}_{i,t}^{(n)} \right)^{-1} \mathbf{e}_{i,t}. \quad (4.30)$$

Algorithm 3 describes the process of training a model without ground truth. We refer to this process as PROBE-GK-EM, and distinguish it from PROBE-GK-GT (Ground Truth). We note that the sequence of estimated transforms,  $\mathbf{T}_{1:T}^{(n)}$ , is guaranteed to converge to a local maxima of the likelihood function ([Dempster et al., 1977](#)). It is also possible to use a robust loss function (Equation (4.25)) in place of Equation (4.30) during EM training. Although not

formally motivated by the derivation above, this approach often leads to lower test errors in practice. Characterizing when and why this robust learning process outperforms its non-robust alternative is outside the scope of this dissertation.

---

**Algorithm 3** Build the covariance model without ground truth given a sequence of observations,  $\mathcal{D}'$ , and an initial odometry estimate  $\mathbf{T}_{1:T}^{(0)}$ .

---

```

function BUILDCOVARIANCEMODEL( $\mathcal{D}'$ ,  $\mathbf{T}_{1:T}^{(0)}$ )
    Initialize an empty spatial index  $\mathcal{M}$ 
    for all  $\mathcal{I}_t$  in  $\mathcal{D}'$  do
        for all  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}, \phi_{i,t}\}$  in  $\mathcal{I}_t$  do
             $\mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t^{(0)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ 
            Insert  $\phi_{i,t}$  into  $\mathcal{M}$  and store  $\mathbf{e}_{i,t}$  at its location
        end for
    end for
    repeat
        for all  $\mathcal{I}_t$  in  $\mathcal{D}'$  do
            for all  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}, \phi_{i,t}\}$  in  $\mathcal{I}_t$  do
                 $\Psi, \nu \leftarrow \text{INFERNOISEMODEL}(\mathcal{M}, \phi_{i,t})$ 
                 $g(\mathbf{T}) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ 
                 $\mathcal{L} \leftarrow \mathcal{L} + g(\mathbf{T})^T (\frac{1}{\nu} \Psi)^{-1} g(\mathbf{T})$ 
            end for
             $\mathbf{T}_t \leftarrow \text{argmin}_{\mathbf{T} \in \text{SE}(3)} \mathcal{L}(\mathbf{T})$ 
             $\mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t^{(0)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ 
            Update the error stored at  $\phi_{i,t}$  in  $\mathcal{M}$  to  $\mathbf{e}_{i,t}$ 
        end for
    until converged
    return  $\mathcal{M}$ 
end function

```

---

## 4.5 Prediction Space

A crucial component of our technique is the choice of the vector of predictors  $\phi$ . In practice, feature tracking quality is often degraded by a variety of effects such as motion blur, moving objects, and textureless or self-similar image regions. The challenge is in determining predictors that account for such effects without requiring excessive computation. In our implementation, we use the following predictors, but stress that the choice of predictors can be tailored to suit particular applications and environments:

- Angular velocity and linear acceleration magnitudes

- Local image entropy
- Blur (quantified by the blur metric of [Crete et al. \(2007\)](#))
- Optical flow variance score
- Image frequency composition

We discuss each of these predictors in turn.

### 4.5.1 Angular velocity and linear acceleration

While most of the predictors in our system are computed directly from image data, the magnitudes of the angular velocities and linear accelerations reported by an IMU (if available) are in themselves good predictors of image degradation (e.g., image blur) and hence poor feature tracking. We do not explicitly correct for bias in linear accelerations because we expect real motion-induced acceleration to trump bias at the timescales of our test trials. As a result, there is virtually no computational cost involved in incorporating these quantities as predictors.

### 4.5.2 Local image entropy

Entropy is a statistical measure of randomness that can be used to characterize the texture in an image or patch. Since the quality of feature detection is strongly influenced by the strength of the texture in the vicinity of the feature point, we expect the entropy of a patch centered on the feature to be a good predictor of its quality. We evaluate the entropy  $S$  in an image patch by sorting pixel intensities into  $N$  bins and computing

$$S = - \sum_{i=1}^N c_i \log_2(c_i), \quad (4.31)$$

where  $c_i$  is the number of pixels counted in the  $i^{\text{th}}$  bin.

### 4.5.3 Blur

Blur can arise from a number of sources including motion, dirty lenses, and sensor defects. All of these have deleterious effects on feature tracking quality. To assess the effect of blur in detail, we performed a separate experiment. We recorded images of 32 interior corners of a standard checkerboard calibration target using a low frame-rate (20 FPS) Skybotix VI-Sensor stereo camera and a high frame-rate (125 FPS) Point Grey Flea3 monocular camera rigidly connected by a bar ([Figure 4.3](#)). Prior to the experiment, we determined the intrinsic and



Figure 4.3: The Skybotix VI-Sensor, Point Grey Flea3, and checkerboard target used in our motion blur experiments.

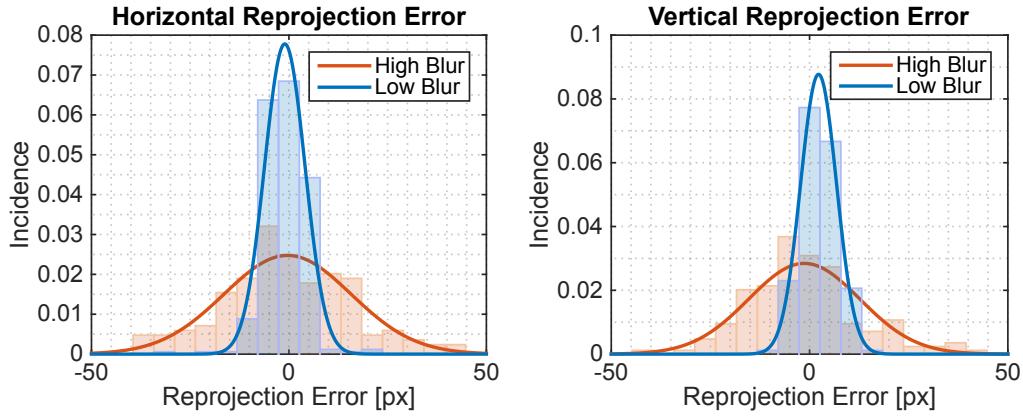


Figure 4.4: Reprojection error of checkerboard corners triangulated from the VI-Sensor and reprojected into the Flea3. We distinguish between high and low blur by thresholding the blur metric [Crete et al. \(2007\)](#).

extrinsic calibration parameters of our rig using the KALIBR<sup>3</sup> package [Furgale et al. \(2013\)](#). The apparatus underwent both slow and fast translational and rotational motion, which induced different levels of motion blur as quantified by the blur metric proposed by [Crete et al. \(2007\)](#).

We detected checkerboard corners in each camera at synchronized time steps, computed their 3D coordinates in the VI-Sensor frame, then reprojected these 3D coordinates into the Flea3 frame. We then computed the reprojection error as the distance between the reprojected image coordinates and the true image coordinates in the Flea3 frame. Since the Flea3 operated at a much higher frame rate than the VI-Sensor, it was less susceptible to motion blur and so we treated its observations as ground truth. We also computed a tracking error by comparing the image coordinates of checkerboard corners in the left camera of the VI-Sensor computed from both KLT tracking [Lucas and Kanade \(1981\)](#) and re-detection.

Figure 4.5 shows histograms and fitted normal distributions for both reprojection error and tracking error. From these distributions we can see that the errors remain approximately

<sup>3</sup><https://github.com/ethz-asl/kalibr>

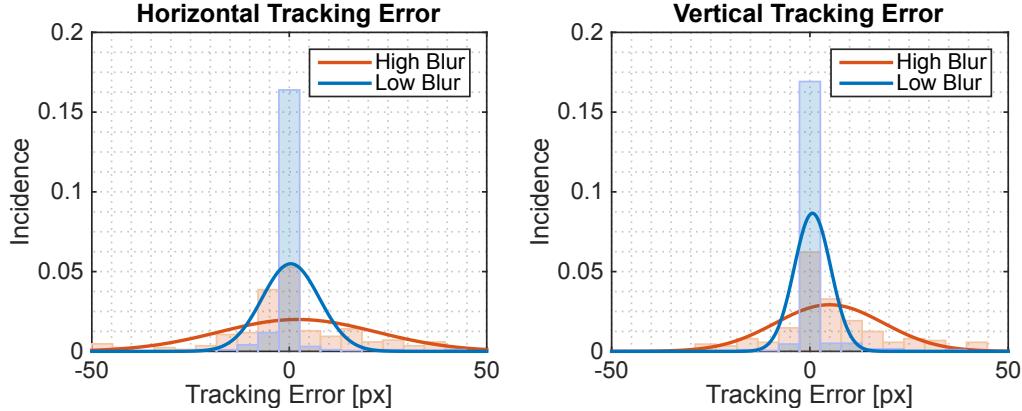


Figure 4.5: Effect of blur on reprojection and tracking error for the slow-then-fast checkerboard dataset. We distinguish between high and low blur by thresholding the blur metric [Crete et al. \(2007\)](#). The variance in both errors increases with blur.

zero-mean, but that their variance increases with blur. This result is compelling evidence that the effect of blur on feature tracking quality can be accounted for by scaling the feature covariance matrix by a function of the blur metric.

#### 4.5.4 Optical flow variance

To detect moving objects, we compute a score for each feature based on the ratio of the variance in optical flow vectors in a small region around the feature to the variance in flow vectors of a larger region. Intuitively, if the flow variance in the small region differs significantly from that in the larger region, we might expect the feature in question to belong to a moving object, and we would therefore like to trust the feature less. Since we consider only the variance in optical flow vectors, we expect this predictor to be reasonably invariant to scene geometry.

We compute this optical flow variance score according to

$$\log \left( \frac{\bar{\sigma}_s^2}{\bar{\sigma}_l^2} \right), \quad (4.32)$$

where  $\bar{\sigma}_s^2, \bar{\sigma}_l^2$  are the means of the variance of the vertical and horizontal optical flow vector components in the small and large regions respectively. Figure 4.6 shows sample results of this scoring procedure for two images in the KITTI dataset. Our optical flow variance score generally picks out moving objects such as vehicles and cyclists in diverse scenes.

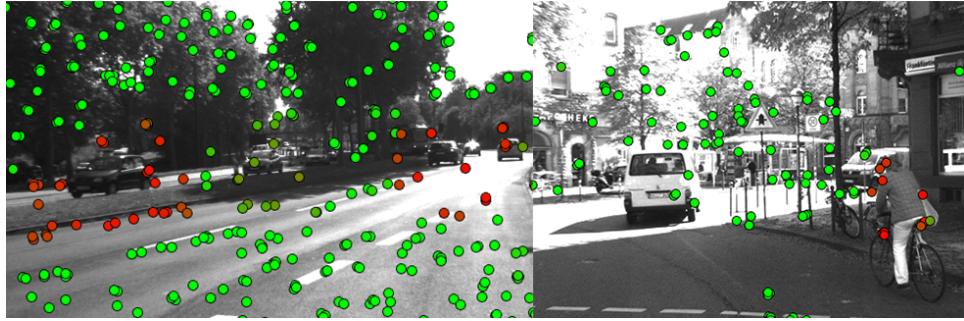


Figure 4.6: The optical flow variance predictor can help in detecting moving objects. Red circles correspond to higher values of the optical flow variance score (i.e., features more likely to belong to a moving object).

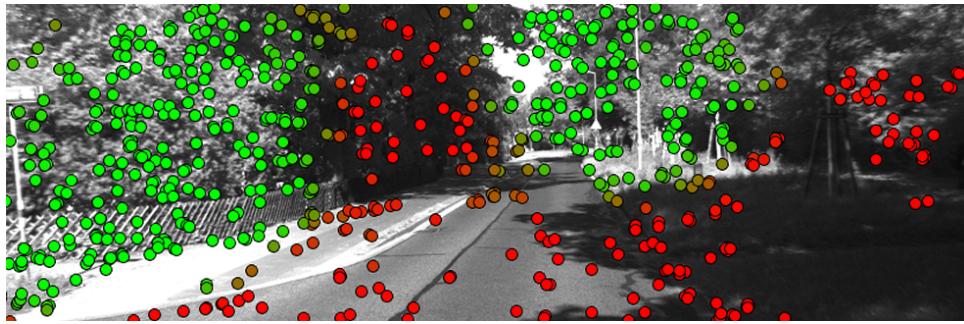


Figure 4.7: A high-frequency predictor can distinguish between regions of high and low texture such as foliage and shadows. Green indicates higher values.

#### 4.5.5 Image frequency composition

Reliable feature tracking is often difficult in textureless or self-similar environments due to low feature counts and false matches. We detect textureless and self-similar image regions by computing the Fast Fourier Transform (FFT) of each image and analyzing its frequency composition. For each feature, we compute a coefficient for the low- and high-frequency regimes of the FFT. Figure 4.7 shows the result of the high-frequency version of this predictor on a sample image from the KITTI dataset. Our high-frequency predictor effectively distinguishes between textureless regions (e.g., shadows and roads) and texture-rich regions (e.g., foliage).

## 4.6 Experiments

To validate PROBE-GK, we used three types of data: synthetic simulations, the KITTI dataset, and our own experimental data collected at the University of Toronto.

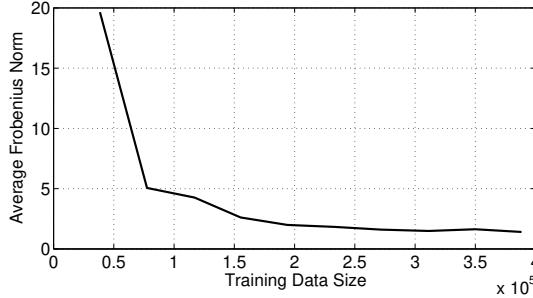


Figure 4.8: Mean Frobenius norm of the error between the estimated and true noise covariance as a function of training data size. The norm tends to zero as training data is added which indicates that PROBE-GK is learning the correct covariances.

### 4.6.1 Simulation

#### Monte-Carlo Verification

To begin, we verified that PROBE-GK can predict increasingly accurate estimates of the true error covariance as more training data is added. We developed a basic simulation environment consisting of a large amount of point landmarks being observed by a stereo camera. In our simulation, the camera traversed a single step in one direction, and recorded empirical reprojection errors based on ground truth poses. We simulated additive Gaussian noise on image coordinates, and used Monte Carlo simulations to estimate the true covariances. Figure 4.8 shows the mean Frobenius norm (as defined in [Barfoot and Furgale \(2014\)](#)) between the covariances estimated by PROBE-GK and the true covariances for a test trial. The mean norm tends to zero as more landmarks are added, indicating that PROBE-GK does learn the correct covariances.

#### Synthetic

Next, we formulated a synthetic dataset wherein a stereo camera traverses a circular path observing 2000 randomly distributed point features. We added Gaussian noise to each of the ideal projected pixel co-ordinates for visible landmarks at every step. We varied the noise variance as a function of the vertical pixel coordinate of the feature in image space. In addition, a small subset of the landmarks received an error term drawn from a uniform distribution to simulate the presence of outliers. The prediction space was composed of the vertical and horizontal pixel locations in each of the stereo cameras.

We simulated independent training and test traversals, where the camera moved for 30 and 60 seconds respectively (at a forward speed of 3 metres per second for final path lengths of 90 and 180 meters). Figure 4.10 and Table 4.1 document the qualitative and quantita-

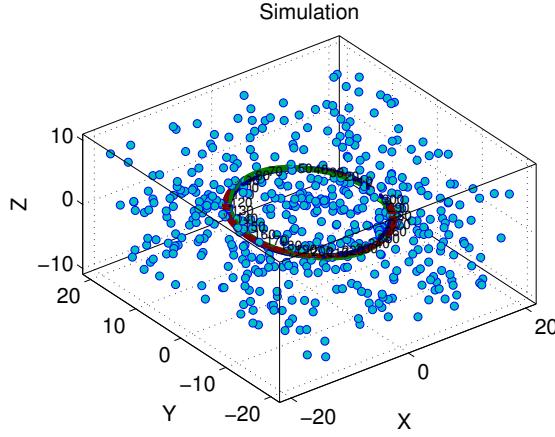


Figure 4.9: Our synthetic world. A stereo camera rig moves through a world with 2000 point features.

tive comparisons of PROBE-GK (trained with and without ground-truth) against two baseline stereo odometry frameworks. Both baseline estimators were implemented based on the reprojection-error-based VO pipeline described in Chapter 3. The first utilized fixed covariances for all reprojection errors, while the second used a modified robust cost (i.e. M-estimation) based on Student’s  $t$  weighting, with  $\nu = 5$  (as suggested in Kerl et al. (2013)). These benchmarks served as baseline estimators (with and without robust costs) that used fixed covariance matrices and did not include a predictive component.

Using PROBE-GK with ground truth data for training, we significantly reduced both the translation and rotational Average Root Mean Squared Error (ARMSE) by approximately 50%. In our synthetic data, the Expectation Maximization approach was able to achieve nearly identical results to the ground-truth-aided model within 5 iterations.

#### 4.6.2 KITTI

To evaluate PROBE-GK on real environments, we trained and tested several models on the KITTI Vision Benchmark suite (Geiger et al., 2013), a series of datasets collected by a car outfitted with a number of sensors driven around different parts of Karlsruhe, Germany. Within the dataset, ground truth pose information is provided by a high grade inertial navigation unit which also fuses measurements from differential GPS. Raw data is available for different types of environments through which the car was driving; for our work, we focused on the city, residential and road categories (Figure 4.11). From each category, we chose two separate trials for training and testing.

Figures 4.12 to 4.14 show typical results; Table 4.1 presents a quantitative comparison. PROBE GK-GT produced significant reductions in ARMSE, reducing translational ARMSE by

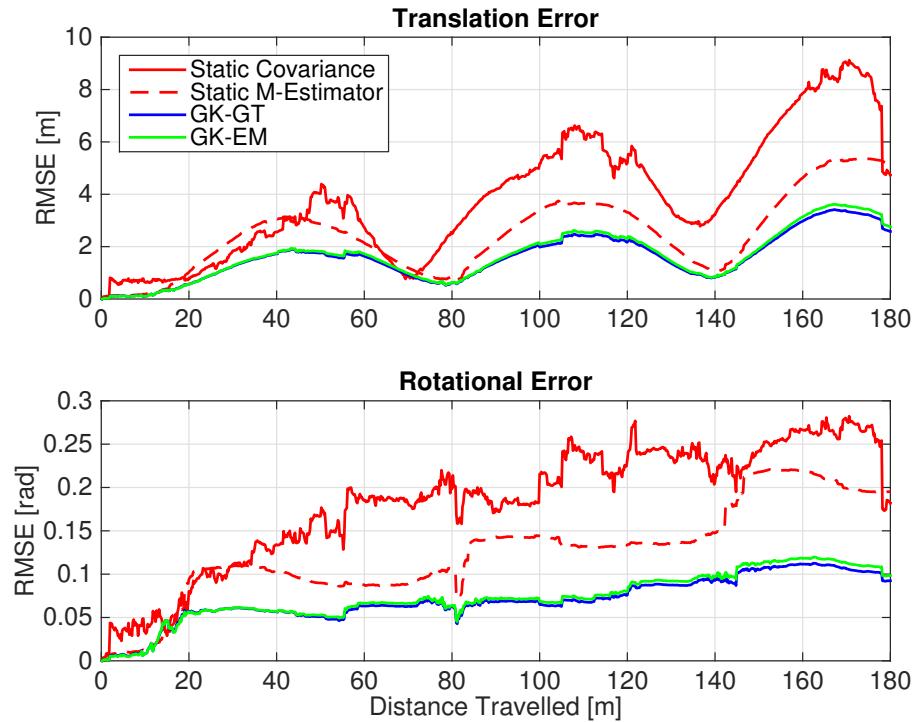


Figure 4.10: A comparison of translational and rotational Root Mean Square Error on simulated data (RMSE) for four different stereo-visual odometry pipelines: two baseline bundle adjustment procedures with and without a robust Student's  $t$  cost with a fixed and hand-tuned covariance and degrees of freedom (M-Estimation), a robust bundle adjustment with covariances learned from ground truth with algorithm 1 (GK-GT), and a robust bundle adjustment using covariances learned without ground truth using expectation maximization, with algorithm 3 (GK-EM). Note in this experiment, the RMSE curves for GK-GT and GK-EM very nearly overlap. The overall translational and rotational ARMSE values are shown in Table 4.1.



Figure 4.11: The KITTI dataset contains three different environments. We validate PROBE-GK by training on each type and testing against a baseline stereo visual odometry pipeline.

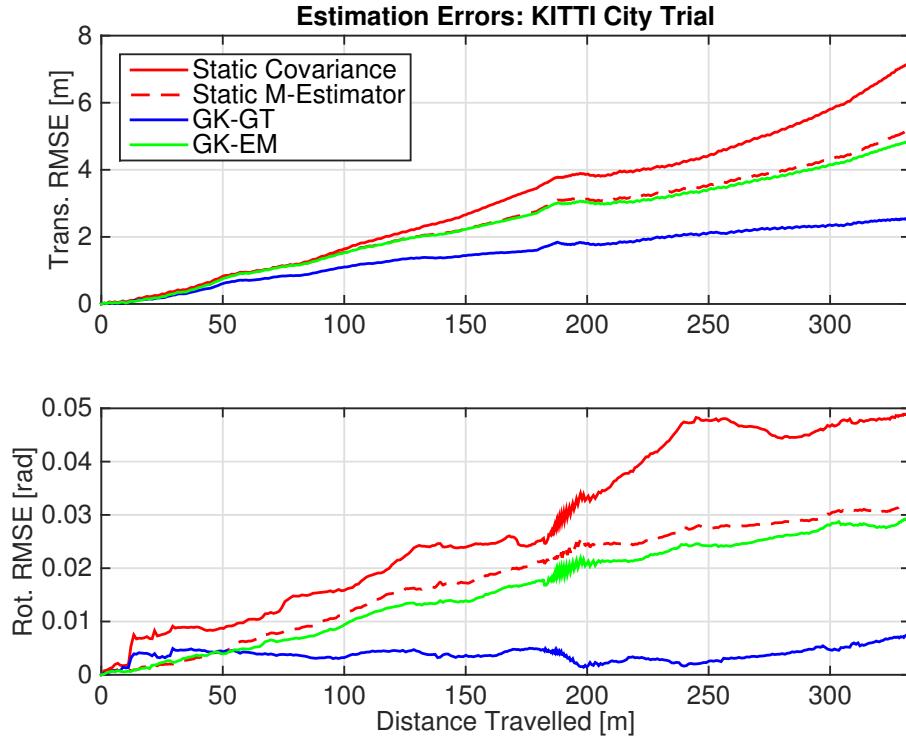


Figure 4.12: RMSE comparison of stereo odometry estimators evaluated on data from the city category in the KITTI dataset. See Table 4.1 for a quantitative summary.

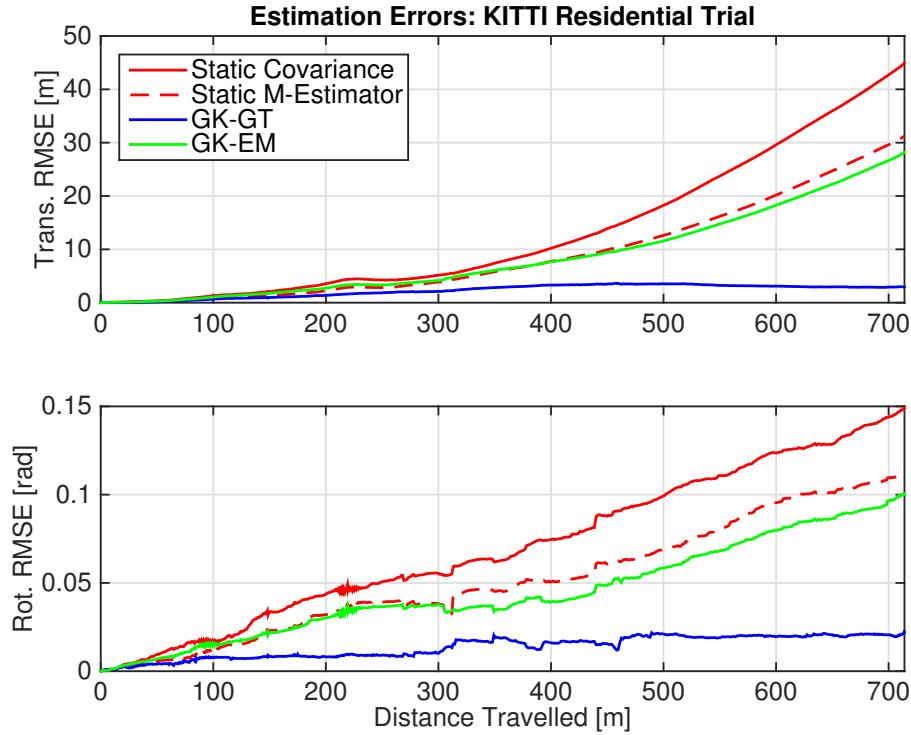


Figure 4.13: RMSE comparison of stereo odometry estimators evaluated on data from the residential category in the KITTI dataset.

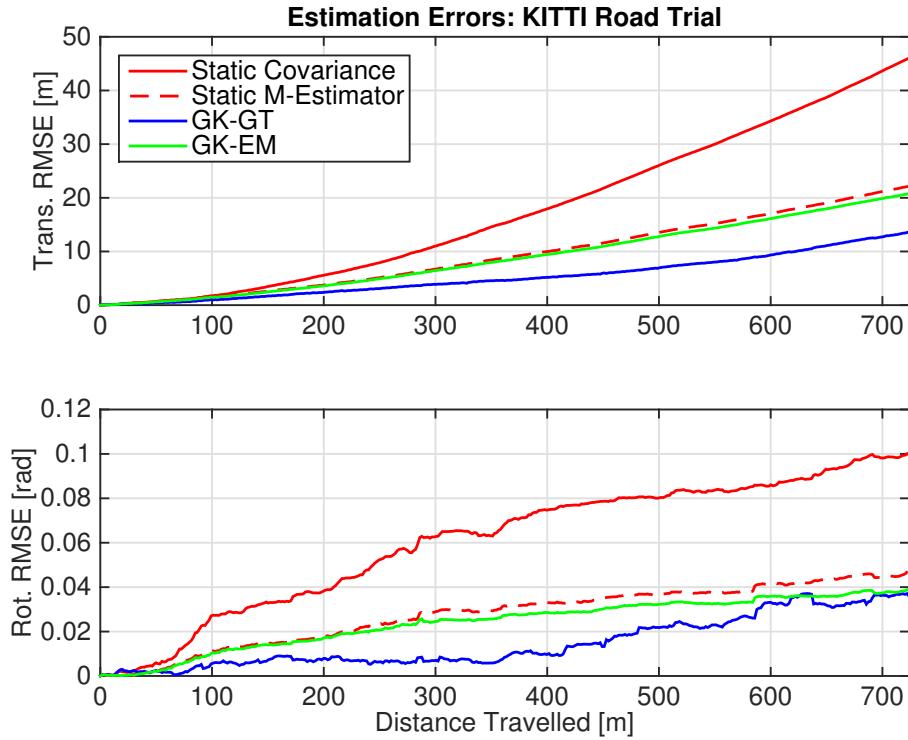


Figure 4.14: RMSE comparison of stereo odometry estimators evaluated on data from the road category in the KITTI dataset.

as much as 80%. In contrast, GK-EM showed more modest improvements; this is unlike our synthetic experiments, where both GK-EM and GK-GT achieved similar performance. We note that although our simulated data is drawn from a mixture of Gaussian distributions, the underlying noise distribution for real data may be far more complex. With no ground truth, EM has to jointly optimize the camera poses and sensor uncertainty. It is unclear whether this is feasible in the general case with no ground truth information.

Further, we observe that the performance of PROBE-GK depends on the similarity of the training data to the final test trials. A characteristic training dataset was important for consistent improvements on test trials.

### 4.6.3 UTIAS

To further investigate the capability of our EM approach, we evaluated PROBE-GK on experimental data collected at the University of Toronto Institute for Aerospace Studies (UTIAS). For this experiment, we drove a Clearpath Husky rover outfitted with an Ashtech DG14 Differential GPS, and a PointGrey XB3 stereo camera around the MarsDome (an indoor Mars analog testing environment) at UTIAS (Figure 4.15) for five trials of a similar path. Each trial was approximately 250 m in length and we made an effort to align the start and end points of

Table 4.1: Comparison of average root mean squared errors (ARMSE) for rotational and translational components. Each trial is trained and tested from a particular category of raw data from the synthetic and KITTI datasets.

	Length [m]	Trans. ARMSE [m]					Rot. ARMSE [rad]		
		Fixed Covar.	Static M-Estimator	GK-GT	GK-EM	Fixed Covar.	Static M-Estimator	GK-GT	GK-EM
Synthetic	180	3.87	2.49	1.59	1.66	0.18	0.13	0.070	0.073
City	332.9	3.84	2.99	1.69	2.87	0.032	0.021	0.0046	0.018
Residential	714.1	13.48	9.37	1.97	8.80	0.068	0.050	0.013	0.044
Road	723.8	17.69	9.38	5.24	8.87	0.060	0.027	0.015	0.024

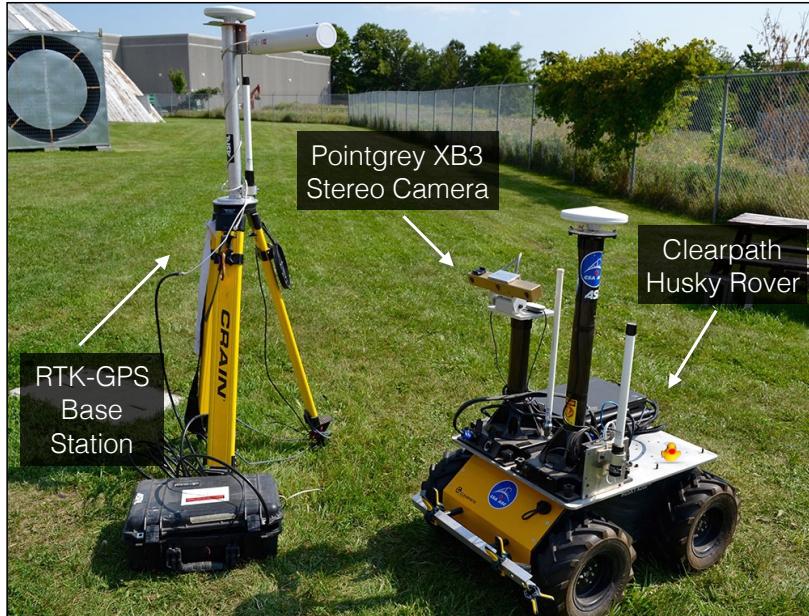


Figure 4.15: Our experimental apparatus: a Clearpath Husky rover outfitted with a PointGrey XB3 stereo camera and a differential GPS receiver and base station.

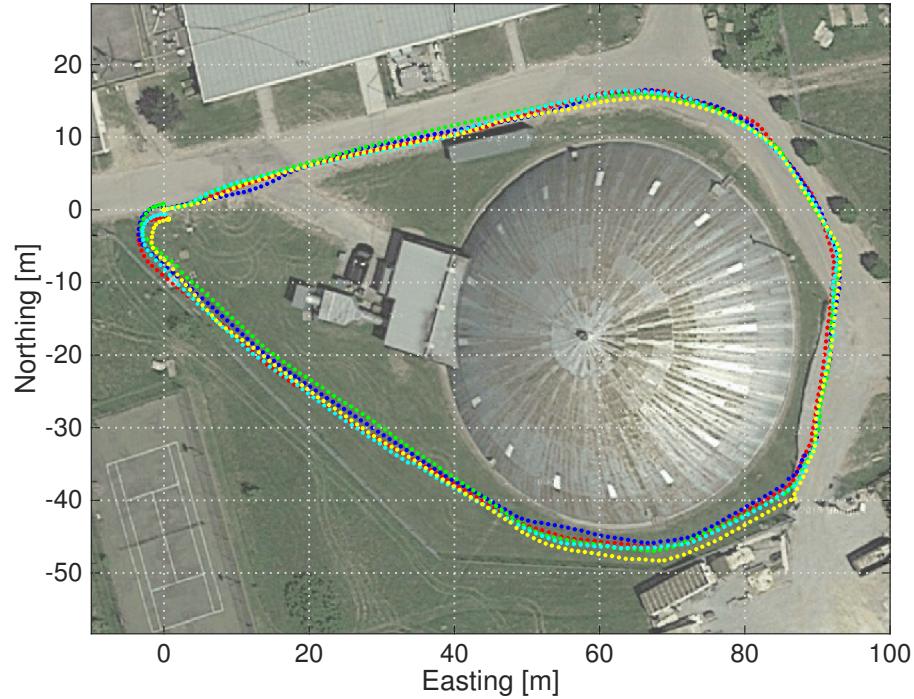


Figure 4.16: GPS ground truth for 5 experimental trials collected near the UTIAS Mars Dome. Each trial is approximately 250 m long.

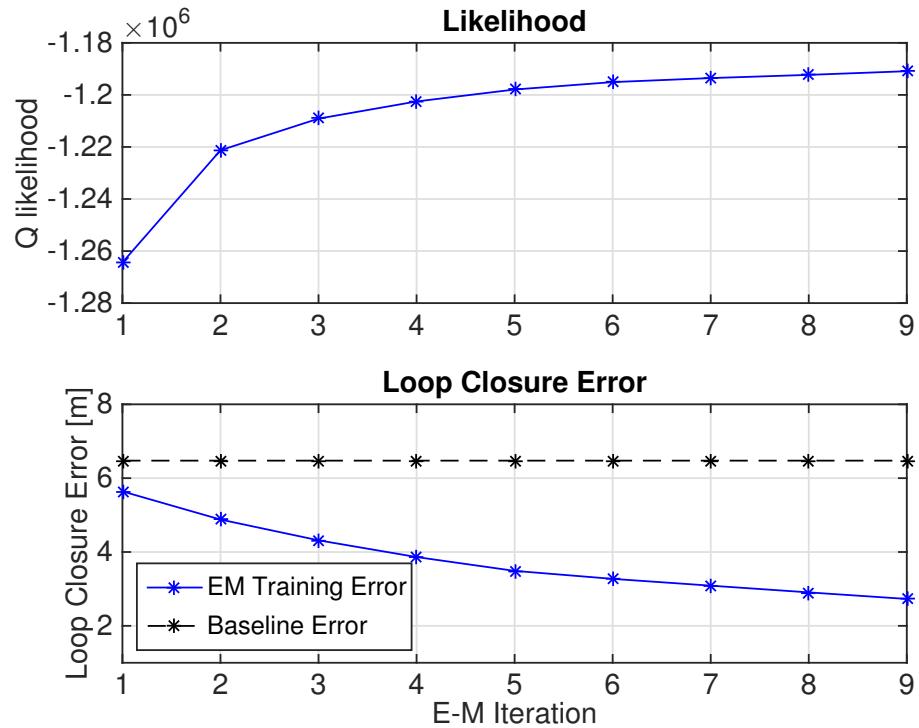


Figure 4.17: Training without ground truth using PROBE-GK-EM on a 250.2m path around the Mars Dome at UTIAS. The likelihood of the data increases with each iteration, and the loop closure error decreases, improving significantly from a baseline static M-estimator.

Table 4.2: Comparison of loop closure errors for 4 different experimental trials with and without a learned PROBE-GK-EM model.

Trial	Path Length [m]	Loop Closure Error [m]	
		PROBE-GK-EM	Static M-Estimator
2	250.3	3.88	8.07
3	250.5	3.07	6.64
4	205.4	2.81	7.57
5	249.9	2.34	7.75

each loop. We used the wide baseline (25 cm) of the XB3 stereo camera to record the stereo images. The approximate trajectory for all 5 trials, as recorded by GPS, is shown in Figure 4.16. Note that the GPS data was not used during training, and only recorded for reference.

For the prediction space in our experiments, we mimicked the KITTI experiments, omitting inertial magnitudes as no inertial data was available. We trained PROBE-GK without ground truth, using the Expectation Maximization approach. Figure 4.17 shows the likelihood and loop closure error as a function of EM iteration.

The EM approach indeed produced significant error reductions on the training dataset after just a few iterations. Although it was trained with no ground truth information, our PROBE-GK model was used to produce significant reductions in the loop closure errors of the remaining 4 test trials. This reinforced our earlier hypothesis: the EM method works well when the training trajectory more closely resembles the test trials (as was the case in this experiment). Table 4.2 lists the statistics for each test.

## 4.7 Summary

Predictive Robust Estimation (PROBE) applied Generalized Kernel estimation to improve on the uncorrelated and static Gaussian error models typically employed in stereo odometry. By building a non-parametric predictive model for the density of reprojection errors, we derived a robust least squares objective whose parameters were predicted based on training data. In summary, this chapter contributed

1. a probabilistic model for indirect stereo visual odometry, leading to a predictive robust algorithm for inference on that model,
2. an efficient approach to constructing the robust algorithm based on Generalized Kernel (GK) estimation,

3. a procedure for training our model using pairs of stereo images with known relative transforms, and
4. an iterative, expectation-maximization approach to train our GK model when the relative ground truth egomotion was unavailable.

# Chapter 5

## Learned Probabilistic Sun Sensor

He stepped down, avoiding any long look at her as one avoids long looks at the sun, but seeing her as one sees the sun, without looking.

---

Leo Tolstoy, *Anna Karenina*

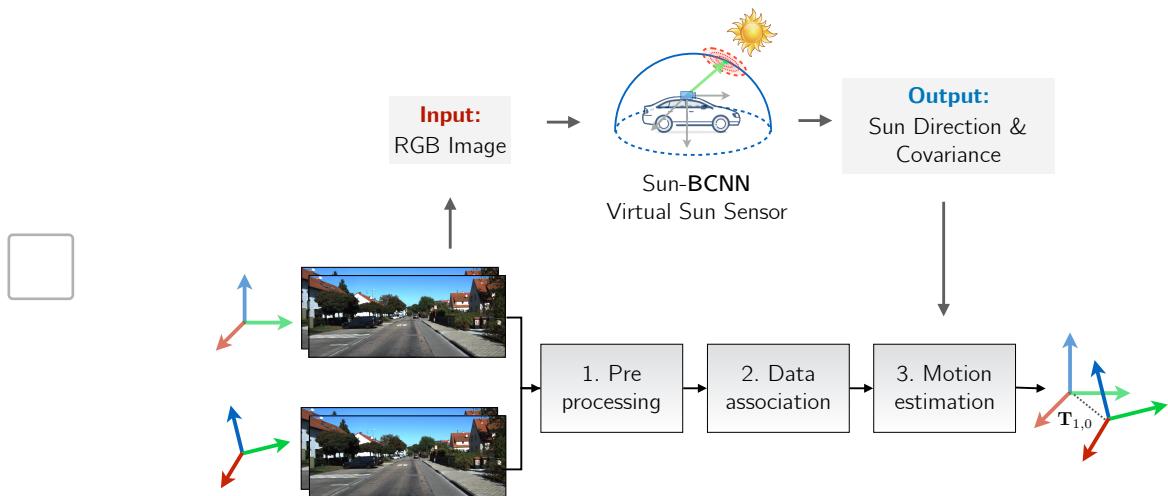


Figure 5.1: Sun-BCNN is a learned virtual sun sensor that outputs sun direction with an associated uncertainty based on a single RGB image. We use this as a source of orientation information within a privileged reference frame.

## 5.1 Introduction

Given that we can infer useful uncertainty information from images, is it possible to infer both uncertainty and some other geometric quantity that can aid in egomotion estimation? In this chapter, we present a pseudo-sensor that can do just that. Namely, we train a deep parametric model to act as a *virtual sun sensor* that mimics a hardware sun sensor (used in space-based applications to orient ~~rovers and~~ satellites) with no additional hardware. This project was a collaboration between myself and Lee Clement. While we were equally responsible for its conception and dissemination, Lee lead the integration of sun information into a visual odometry pipeline, while I designed and implemented the learning components.

This work was associated with three publications:

1. Clement, L., Peretroukhin, V., and Kelly, J. (2017). Improving the accuracy of stereo visual odometry using visual illumination estimation. In Kulic, D., Nakamura, Y., Khatib, O., and Venture, G., editors, *2016 International Symposium on Experimental Robotics*, volume 1 of *Springer Proceedings in Advanced Robotics*, pages 409–419. Springer International Publishing, Berlin Heidelberg. Invited to Journal Special Issue,
2. Peretroukhin, V., Clement, L., and Kelly, J. (2017). Reducing drift in visual odometry by inferring sun direction using a bayesian convolutional neural network. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'17)*, pages 2035–2042, Singapore,
3. Peretroukhin, V., Clement, L., and Kelly, J. (2018). Inferring sun direction to improve visual odometry: A deep learning approach. *International Journal of Robotics Research*, 37(9):996–1016.

This chapter is largely a reproduction of the latter journal publication which summarizes the approach.

## 5.2 Motivation

A crucial competency of any autonomous mobile robot is the ability to estimate its own motion through an operating environment. While there exists a rich body of literature on the topic of motion estimation using a variety of techniques such as lidar-based point cloud matching ([Zhang and Vela, 2015](#)) and visual-inertial odometry ([Leutenegger et al., 2015](#)), ego-motion estimation is fundamentally a process of dead-reckoning and will accumulate unbounded error over time. This accumulated error, or drift, can be limited by incorporating

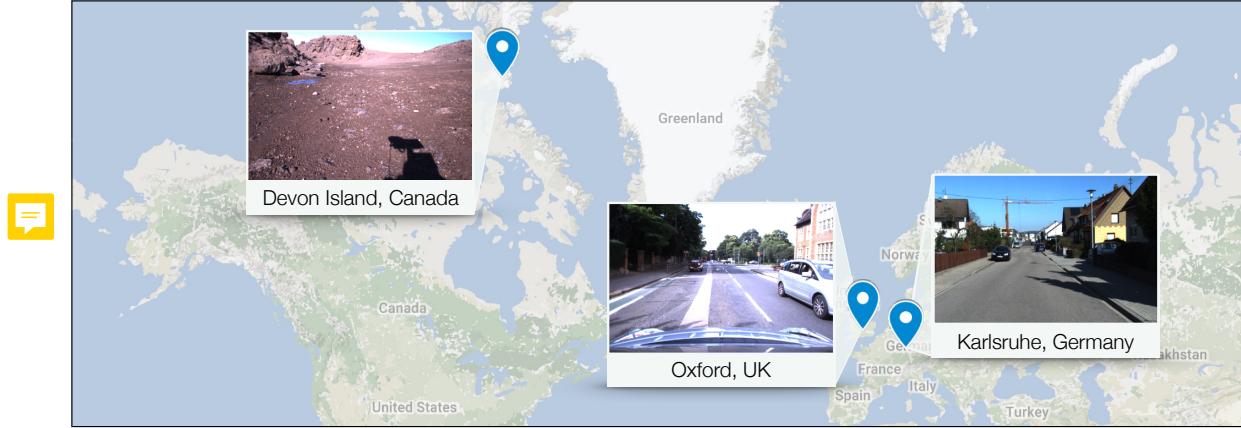


Figure 5.2: We train and test Sun-BCNN in a variety of environments ranging from urban driving in Europe to remote planetary analogue sites in the Canadian High Arctic. (Map data: Google, INEGI, ORION-ME.)

global information into the motion estimation problem. This frequently takes the form of a globally consistent map, loop closure detection, or reliance on additional sensors such as GPS to make corrections to the estimated trajectory. In many situations, however, a globally consistent map may be unavailable or prohibitively expensive to compute, loop closures may not occur, or GPS may be unavailable or inaccurate. In such cases, it can be advantageous to rely on environmental cues such as the sun, which can easily provide global orientation information since it is readily detectable and its apparent motion in the sky is well described by ephemeris models.

For visual odometry (VO) in particular, the addition of global orientation information can limit the growth of drift error to be linear rather than superlinear with distance traveled (Olson et al., 2003).

Sun-based orientation corrections have been successfully used in planetary analogue environments (Furgale et al., 2011; Lambert et al., 2012) as well as on board the Mars Exploration Rovers (MERs) (Eisenman et al., 2002; Maimone et al., 2007). In particular, Lambert et al. (2012) showed that incorporating sun sensor and inclinometer measurements directly into the motion estimation pipeline (as opposed to periodically updating

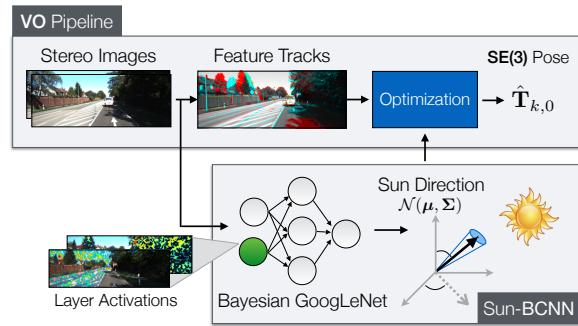


Figure 5.3: Our method uses a Bayesian Convolutional Neural Network (BCNN) to estimate the direction of the sun and to produce a principled uncertainty estimate for each prediction. We incorporate this *virtual sun sensor* into a stereo visual odometry pipeline to reduce estimation error.

the vehicle heading, as in earlier work) can significantly reduce VO drift over long trajectories.

In this work, we seek to answer the question of whether similar reductions in VO drift can be obtained solely from the image stream already being used to compute VO. The main idea here is that by reasoning over more than just the geometric information available from a standard RGB camera, we can improve existing VO techniques without needing to rely on a dedicated sun sensor or specially oriented camera. In particular, we leverage recent advances in Bayesian Convolutional Neural Networks (BCNNs) to demonstrate how we can build and train a deep model capable of inferring the direction of the sun from a single RGB image. Moreover, we show that our network, dubbed Sun-BCNN, can produce a covariance estimate for each observation that obviates the need for a hand-tuned or empirically computed static covariance typically used for data fusion in a motion estimation pipeline.

The remainder of this chapter begins with a discussion of related work, followed by an overview of the theory underlying BCNNs and a discussion of our model architecture, implementation, and training procedure. We then outline our chosen visual odometry pipeline, which is based is an adapted version of the pipeline describe in Chapter 3, and describe how observations of the sun can be incorporated directly into the motion estimation problem following the technique of [Lambert et al. \(2012\)](#). Finally, we present several sets of experiments designed to test and validate both Sun-BCNN and our sun-aided VO pipeline in variety of environments. These include experiments on 21.6 km of urban driving data from the KITTI odometry benchmark training set ([Geiger et al., 2013](#)), as well as a further 10 km traverse through a planetary analogue site taken from the Devon Island Rover Navigation Dataset collected in a planetary analogue site in the Canadian High Arctic ([Furgale et al., 2012](#)). We investigate the possibility of model generalization between different cameras and environments, and further explore the sensitivity of Sun-BCNN to cloud cover during training and testing, using data from the Oxford Robotcar Dataset ([Maddern et al., 2016](#)). We also examine the impact of different methods for computing the mean and covariance of a norm-constrained vector on the accuracy and consistency of the estimated sun directions.

### 5.3 Related Work

Visual odometry (VO), ~~a technique to estimate the motion of a moving platform equipped with one or more cameras~~, has a rich history of research including a notable implementation onboard the Mars Exploration Rovers (MERs) ([Scaramuzza and Fraundorfer, 2011](#)). Modern approaches to VO can achieve estimation errors below 1% of total distance traveled ([Geiger et al., 2013](#)). To achieve such accurate and robust estimates, modern techniques use careful visual feature pruning ([Cvišić and Petrović, 2015](#)), adaptive robust methods ([Alcantarilla and](#)

Woodford, 2016; Peretroukhin et al., 2016), or operate directly on pixel intensities (Engel et al., 2015).

Independent of the estimator, VO exhibits superlinear error growth, and is particularly sensitive to errors in orientation (Olson et al., 2003; Cvišić and Petrović, 2015). One way to reduce orientation error is to incorporate observations of a landmark whose position or direction in the navigation frame is known *a priori*. The sun is an example of such a known directional landmark. Accordingly, hardware sun sensors have been used to improve the accuracy of VO in planetary analogue environments (e.g., the Sinclair Interplanetary SS-411 sun sensor used by Furgale et al. (2011) and Lambert et al. (2012)), while the MERs articulated their Pancam apparatus to directly image the sun (Maimone et al., 2007; Eisenman et al., 2002). More recently, software-based alternatives have been developed that can estimate the direction of the sun from a single image, making sun-aided navigation possible without additional sensors or a specially-oriented camera (Clement et al., 2017). Some of these methods have been based on hand-crafted illumination cues such as shadows and variation in sky brightness (Lalonde et al., 2011; Clement et al., 2017), while others have attempted to learn such cues from data using deep Convolutional Neural Networks (CNNs) (Ma et al., 2016).

Convolutional Neural Networks (CNNs) have been applied to a wide range of classification, segmentation, and learning tasks in computer vision (LeCun et al., 2015). Recent work has shown that CNNs can learn orientation information directly from images by modifying the loss functions of existing discrete classification-based CNN architectures into continuous regression losses (Ma et al., 2016; Kendall et al., 2015; Kendall and Cipolla, 2016). Despite their success in improving prediction accuracy, most existing CNN-based models do not report uncertainty estimates, which are important in the context of data fusion.

For classification, it is possible to restrict CNN model outputs to a certain range (e.g., using a softmax function) and interpret these values as the model's confidence in its output. As Gal (2016) noted, however, this can be misleading because these values can be unjustifiably large for test points far away from training data. To address this, Gal and Ghahramani (2016b) showed that it is possible to achieve covariance outputs that better quantify model uncertainty for classification and regression tasks, with only minor modifications to existing CNN architectures. An early application of this uncertainty quantification was presented by Kendall and Cipolla (2016) who used it to improve their prior work (Kendall et al., 2015) on camera pose regression.

We build on previous work by Clement et al. (2017), who demonstrated empirically that techniques for single-image sun estimation based on hand-crafted models (Lalonde et al., 2011) and Convolutional Neural Networks (CNNs) (Ma et al., 2016) could be incorporated into a stereo visual odometry pipeline to reduce estimation error in the manner of Lambert et al.

(2012). We also build on the work of Peretroukhin et al. (2017), who presented preliminary experimental results comparing Sun-BCNN against the method of Lalonde et al. (2011) and its VO-informed variant (Clement et al., 2017) as well as the Sun-CNN of Ma et al. (2016) on the KITTI odometry benchmark (Geiger et al., 2013), both in terms of raw measurement accuracy and in terms of their impact on VO accuracy.

While our method is similar in spirit to the work of Ma et al. (2016), who built a CNN-based sun sensor as part of a relocalization pipeline, our model makes three important improvements: 1) in addition to a point estimate of the sun direction, we output a principled covariance estimate that is incorporated into our estimator; 2) we produce a full 3D sun direction estimate with azimuth and zenith angles that is better suited to 6-DOF robot pose estimation problems (as opposed to only the azimuth angle and 3-DOF estimator used by Ma et al. (2016)); and 3) we incorporate the sun direction covariance into a VO estimator that accounts for growth in pose uncertainty over time (unlike Clement et al. (2017)). Furthermore, our Bayesian CNN includes a dropout layer after every convolutional and fully connected layer (as outlined by Gal and Ghahramani (2016b) but not done by Kendall and Cipolla (2016)).

## 5.4 Sun-Aided Stereo Visual Odometry

We adopt a sliding window sparse stereo VO technique (adopted based on the frame-to-frame pipeline described in Chapter 3) that has been used in a number of successful mobile robotics applications (Cheng et al., 2006; Furgale and Barfoot, 2010; Geiger et al., 2011; Kelly et al., 2008). Our task is to estimate a window of SE(3) poses  $\{\mathbf{T}_{k_1,0}, \mathbf{T}_{k_1+1,0}, \dots, \mathbf{T}_{k_2-1,0}, \mathbf{T}_{k_2,0}\}$  expressed in a base coordinate frame  $\underline{\mathcal{F}}_0$ , given a prior estimate of the transformation  $\mathbf{T}_{k_1,0}$ . We accomplish this by tracking keypoints across pairs of stereo images and computing an initial guess for each pose in the window using frame-to-frame point cloud alignment, which we then refine by solving a local bundle adjustment problem over the window. In our experiments we choose a window size of two, which we observed to provide good VO accuracy at low computational cost. We select the initial pose  $\mathbf{T}_{1,0}$  to be the first GPS ground truth pose such that  $\underline{\mathcal{F}}_0$  is a local East-North-Up (ENU) coordinate system with its origin at the first GPS position.

### 5.4.1 Observation Model

We assume that incoming stereo images have been undistorted and rectified in a pre-processing step, and model the stereo camera as a pair of perfect pinhole cameras with focal lengths  $f_u, f_v$  and principal points  $(c_u, c_v)$ , separated by a fixed and known baseline  $b$  (see Section 3.2.2).

If we take  $\mathbf{p}_0^j$  to be the homogeneous 3D coordinates of keypoint  $j$ , expressed in our chosen base frame  $\underline{\mathcal{F}}_0$ , we can transform the keypoint into the camera frame at pose  $k$  to obtain  $\mathbf{p}_k^j = \mathbf{T}_{k,0}\mathbf{p}_0^j = \begin{bmatrix} p_{k,x}^j & p_{k,y}^j & p_{k,z}^j & 1 \end{bmatrix}^T$ . Our observation model  $\mathbf{g}(\cdot)$  (defined with disparity, unlike the function  $\mathbf{f}(\cdot)$  in Section 3.2.2) can then be formulated as

$$\mathbf{y}_{k,j} = \mathbf{g}(\mathbf{p}_k^j) = \begin{bmatrix} u \\ v \\ d \end{bmatrix} = \begin{bmatrix} f_u p_{k,x}^j / p_{k,z}^j + c_u \\ f_v p_{k,y}^j / p_{k,z}^j + c_v \\ f_u b / p_{k,z}^j \end{bmatrix}, \quad (5.1)$$

where  $(u, v)$  are the keypoint coordinates in the left image and  $d$  is the disparity in pixels.

### 5.4.2 Sliding Window Bundle Adjustment

Like with PROBE, we use the open-source `viso2` package (Geiger et al., 2011) to detect and track keypoints between stereo image pairs. Based on these keypoint tracks, a three-point Random Sample Consensus (RANSAC) algorithm (Fischler and Bolles, 1981) generates an initial guess of the interframe motion and rejects outlier keypoint tracks by thresholding their reprojection error. We compound these pose-to-pose transformation estimates through our chosen window and refine them using a local bundle adjustment, which we solve using the nonlinear least-squares solver Ceres (Agarwal et al., 2016). The objective function to be minimized can be written as

$$\mathcal{J} = \mathcal{J}_{\text{reprojection}} + \mathcal{J}_{\text{prior}}, \quad (5.2)$$

where

$$\mathcal{J}_{\text{reprojection}} = \sum_{k=k_1}^{k_2} \sum_{j=1}^J \mathbf{e}_{\mathbf{y}_{k,j}}^T \mathbf{R}_{\mathbf{y}_{k,j}}^{-1} \mathbf{e}_{\mathbf{y}_{k,j}} \quad (5.3)$$

and

$$\mathcal{J}_{\text{prior}} = \mathbf{e}_{\hat{\mathbf{T}}_{k_1,0}}^T \mathbf{R}_{\hat{\mathbf{T}}_{k_1,0}}^{-1} \mathbf{e}_{\hat{\mathbf{T}}_{k_1,0}}. \quad (5.4)$$

The quantity  $\mathbf{e}_{\mathbf{y}_{k,j}} = \hat{\mathbf{y}}_{k,j} - \mathbf{y}_{k,j}$  represents the reprojection error of keypoint  $j$  for camera pose  $k$ , with  $\mathbf{R}_{\mathbf{y}_{k,j}}$  being the covariance of these errors. The predicted measurements are given by  $\hat{\mathbf{y}}_{k,j} = \mathbf{g}(\hat{\mathbf{T}}_{k,0}\hat{\mathbf{p}}_0^j)$ , where  $\hat{\mathbf{T}}_{k,0}$  and  $\hat{\mathbf{p}}_0^j$  are the estimated poses and keypoint positions in base frame  $\underline{\mathcal{F}}_0$ .

The cost term  $\mathcal{J}_{\text{prior}}$  imposes a normally distributed prior  $\check{\mathbf{T}}_{k_1,0}$  on the first pose in the current window, based on the estimate of this pose in the previous window. The error in the current estimate  $\hat{\mathbf{T}}_{k_1,0}$  of this pose compared to the prior can be computed via the SE(3) matrix logarithm as  $\mathbf{e}_{\hat{\mathbf{T}}_{k_1,0}} = \log(\check{\mathbf{T}}_{k_1,0}^{-1} \hat{\mathbf{T}}_{k_1,0})^\vee \in \mathbb{R}^6$ . The  $6 \times 6$  matrix  $\mathbf{R}_{\hat{\mathbf{T}}_{k_1,0}}$  is the covari-

ance associated with  $\check{\mathbf{T}}_{k_1,0}$  in its local tangent space, and is obtained as part of the previous window's bundle adjustment solution. This prior term allows consecutive windows of pose estimates to be combined in a principled way that appropriately propagates global pose uncertainty from window to window, which is essential in the context of optimal data fusion.

## 5.5 Orientation Correction

In order to combat drift in the VO estimate produced by accumulated orientation error, we adopt the technique of [Lambert et al. \(2012\)](#) to incorporate absolute orientation information from the sun directly into the estimation problem. We assume the initial camera pose and its timestamp are available from GPS and use them to determine the global direction of the sun  $\mathbf{s}_0$ , expressed as a 3D unit vector, from ephemeris data. We define the world frame  $\mathcal{F}_0$  to be a local ENU coordinate system with the initial GPS position as its origin. At each timestep we update  $\mathbf{s}_0$  by querying the ephemeris model using the current timestamp and the initial camera pose, allowing our model to account for the apparent motion of the sun over long trajectories.

By transforming the global sun direction into each camera frame  $\mathcal{F}_k$  in the window, we obtain predicted sun directions  $\hat{\mathbf{s}}_k = \hat{\mathbf{T}}_{k,0}\mathbf{s}_0$ , where  $\hat{\mathbf{T}}_{k,0}$  is the current estimate of camera pose  $k$  in the base frame. We compare the predicted and estimated sun directions to introduce an additional error term into the bundle adjustment cost function (cf. Equation (5.2)):

$$\mathcal{J} = \mathcal{J}_{\text{reprojection}} + \mathcal{J}_{\text{prior}} + \mathcal{J}_{\text{sun}}, \quad (5.5)$$

where

$$\mathcal{J}_{\text{sun}} = \sum_{k=k_1}^{k_2} \mathbf{e}_{\mathbf{s}_k}^T \mathbf{R}_{\mathbf{s}_k}^{-1} \mathbf{e}_{\mathbf{s}_k}, \quad (5.6)$$

and  $\mathcal{J}_{\text{reprojection}}$  and  $\mathcal{J}_{\text{prior}}$  are defined in Equations (5.3) and (5.4), respectively. This additional term constrains the orientation of the camera, which helps limit drift in the VO result due to orientation error ([Lambert et al., 2012](#)).

Since  $\mathbf{s}_k$  is constrained to be unit length, there are only two underlying degrees of freedom. We therefore define  $\mathbf{f}(\cdot)$  to be a function that transforms a 3D unit vector in camera frame  $\mathcal{F}_k$  to a zenith-azimuth parametrization:

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = \mathbf{f}(\mathbf{s}_k) = \begin{bmatrix} \cos(-s_{k,y}) \\ \text{atan2}(s_{k,x}, s_{k,z}) \end{bmatrix} \quad (5.7)$$

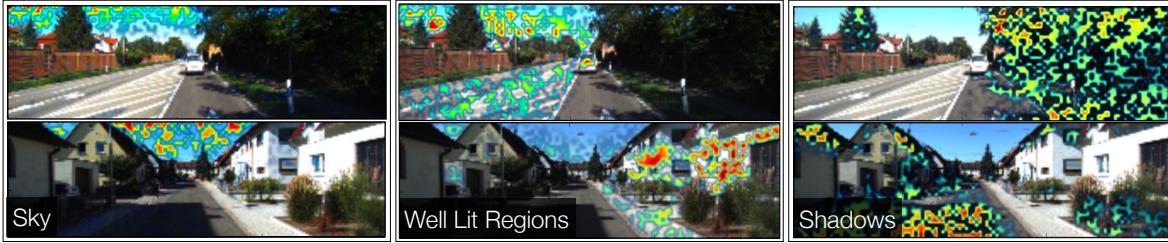


Figure 5.4: Three conv1 layer activation maps superimposed on two images from the KITTI odometry benchmark (Geiger et al., 2013) 00 and 04 for three selected filters. Each filter picks out salient parts of the image that aid in sun direction inference.

where  $\mathbf{s}_k = \begin{bmatrix} s_{k,x} & s_{k,y} & s_{k,z} \end{bmatrix}^T$ . We can then define the term  $\mathbf{e}_{\mathbf{s}_k} = \mathbf{f}(\hat{\mathbf{s}}_k) - \mathbf{f}(\mathbf{s}_k)$  to be the error in the predicted sun direction, expressed in azimuth-zenith coordinates, and  $\mathbf{R}_{\mathbf{s}_k}$  to be the covariance of these errors. While  $\mathbf{R}_{\mathbf{s}_k}$  would generally be treated as an empirically determined static covariance, in our approach we use the per-observation covariance computed using Equation (5.12), which allows us to weight each observation individually according to a measure of its intrinsic quality. In practice, we also attempt to mitigate the effect of outlier sun predictions by applying a robust Huber loss to the sun measurements in our optimizer.

## 5.6 Indirect Sun Detection using a Bayesian Convolutional Neural Network

We use a Bayesian Convolutional Neural Network (BCNN) to infer the direction of the sun and an associated uncertainty, and refer to our model as Sun-BCNN. We motivate the choice of a deep model through the empirical findings of Clement et al. (2017) and Ma et al. (2016), who demonstrated that a CNN-based sun detector can substantially outperform hand-crafted models such as that of Lalonde et al. (2011) both in terms of measurement accuracy and in its application to a VO task.

We choose a deep neural network structure based on GoogLeNet (Szegedy et al., 2015) due to its use in past work that adapted it for orientation regression (Kendall and Cipolla, 2016; Kendall et al., 2015). Unlike Ma et al. (2016), we choose to transfer weights trained on the MIT Places dataset (Zhou et al., 2014) rather than ImageNet (Deng et al., 2009). We believe the MIT Places dataset is a more appropriate starting point for localization tasks than ImageNet since it includes outdoor scenes and is concerned with classifying physical locations rather than objects.

### 5.6.1 Cost Function

We train Sun-BCNN by minimizing the cosine distance between the unit-norm target sun direction vector  $\mathbf{s}_k$  and the predicted unit-norm sun direction vector  $\hat{\mathbf{s}}_k$ , where  $k$  indexes the images in the training set:

$$\mathcal{L}(\hat{\mathbf{s}}_k) = 1 - (\hat{\mathbf{s}}_k \cdot \mathbf{s}_k). \quad (5.8)$$

Note that in our implementation, we do not formulate the cosine distance loss explicitly, but instead minimize half the square of the tip-to-tip Euclidian distance between  $\mathbf{s}_k$  and  $\hat{\mathbf{s}}_k$ , which is equivalent to Equation (5.8) since both vectors have unit length:

$$\begin{aligned} \frac{1}{2} \|\hat{\mathbf{s}}_k - \mathbf{s}_k\|^2 &= \frac{1}{2} (\|\hat{\mathbf{s}}_k\|^2 + \|\mathbf{s}_k\|^2 - 2(\hat{\mathbf{s}}_k \cdot \mathbf{s}_k)) \\ &= 1 - (\hat{\mathbf{s}}_k \cdot \mathbf{s}_k) \\ &= \mathcal{L}(\hat{\mathbf{s}}_k). \end{aligned}$$

We ensure that our network output,  $\hat{\mathbf{s}}_k$ , has a unit norm by appending a normalization layer to the network.

### 5.6.2 Uncertainty Estimation

Following recent work on Bayesian Convolutional Neural Networks (BCNNs) ([Gal and Ghahramani, 2016a,b](#); [Gal, 2016](#)), we modify our model architecture to enable the computation of principled covariance estimates associated with each predicted sun direction. To achieve computationally tractable Bayesian inference with a CNN architecture, BCNNs exploit a connection between stochastic regularization (e.g., dropout, a widely used technique in deep learning to mitigate overfitting) and approximate variational inference of a Bayesian Neural Network. We outline the technique here briefly, and refer the reader to [Gal and Ghahramani \(2016a\)](#) for more details.

The method begins with a prior  $p(\mathbf{w})$  on the weights in a deep neural network and attempts to compute a posterior distribution  $p(\mathbf{w}|\mathbf{X}, \mathbf{S})$  given training inputs  $\mathbf{X} = \{\mathbf{x}_k\}$  and targets  $\mathbf{S} = \{\mathbf{s}_k\}$ . This posterior can be used to compute a predictive distribution for test samples but is generally intractable. To overcome this, the BCNN approach notes that CNN training with stochastic regularization can be viewed as variational inference if we define a variational distribution  $q(\mathbf{w})$  as:

$$q(\mathbf{w}_i) = \mathbf{M}_i \text{diag} \left\{ \{b_j^i\}_{j=1}^{K_i} \right\}, \quad (5.9)$$

$$b_j^i \in \text{Bernoulli}(p_i). \quad (5.10)$$

Here,  $i$  indexes a particular layer in the neural network with  $K_i$  weights,  $\mathbf{M}$  are the weights to be optimized,  $b_j^i$  are Bernoulli distributed binary variables, and  $p_i$  is the dropout probability for weights in layer  $i$ .

With this variational distribution  $q(\mathbf{w})$ , training a CNN with dropout is analogous to minimizing  $\text{KL}(p(\mathbf{w}|\mathbf{X}, \mathbf{S}) \parallel q(\mathbf{w}))$ , the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior. At test time, the first two moments of the predictive distribution are approximated using Monte Carlo integration over the weights  $\mathbf{w}$ :

$$\mathbb{E} [\hat{\mathbf{s}}^*]_k = \hat{\mathbf{s}}_k^* \approx \frac{1}{N} \sum_{n=1}^N \hat{\mathbf{s}}_k^*(\mathbf{x}_k^*, \mathbf{w}^n) \quad (5.11)$$

$$\begin{aligned} \mathbb{E} \left[ \hat{\mathbf{s}}_k^* \hat{\mathbf{s}}_k^{*T} \right] &\approx \tau^{-1} \mathbf{1} + \frac{1}{N} \sum_{n=1}^N \hat{\mathbf{s}}_k^*(\mathbf{x}_k^*, \mathbf{w}^n) \hat{\mathbf{s}}_k^*(\mathbf{x}_k^*, \mathbf{w}^n)^T \\ &\quad - \hat{\mathbf{s}}_k^* \hat{\mathbf{s}}_k^{*T}, \end{aligned} \quad (5.12)$$

where  $\mathbf{1}$  is the identity matrix, and  $\mathbf{w}^n$  is a sample from  $q(\mathbf{w})$  (obtained by sampling the network with dropout). The model precision,  $\tau$ , is computed as

$$\tau = \frac{pl^2}{2M\lambda}, \quad (5.13)$$

where  $p$  is the dropout probability,  $l$  is the characteristic length scale,  $M$  is the number of samples in the training data, and  $\lambda$  is the weight decay.

Following Gal and Ghahramani (2016a), we build our BCNN by adding dropout layers after every convolutional and fully connected layer in the network. We then retain these layers at test time to sample the network stochastically, following the technique of Monte Carlo Dropout, and obtain the relevant statistical quantities using Equations (5.11) and (5.12).

### 5.6.3 Implementation and Training

We implement our network in Caffe (Jia et al., 2014), using the L2Norm layer from the Caffe-SL fork<sup>1</sup> to enforce a unit-norm constraint on the final output. We train the network using stochastic gradient descent, setting all dropout probabilities to 0.5, performing 30,000 iterations with a batch size of 64, and setting the initial learning rate to be between  $10^{-3}$  and  $10^{-4}$ . Training requires approximately 2.5 hours on an NVIDIA Titan X GPU. Interestingly, Figure 5.4 shows that some convolutional filters learned by Sun-BCNN on the KITTI dataset appear to correspond to illumination variations reminiscent of the visual cues designed by

---

<sup>1</sup><https://github.com/wanji/caffe-sl>

[Lalonde et al. \(2011\)](#).

### Data Preparation & Transfer Learning

We resize images from their original size to  $[224 \times 224]$  pixels to achieve the image size expected by GoogleLeNet. We experimented with preserving the aspect ratio of the original image and padding zeros to the top and bottom of the resized image, but found that preserving the vertical resolution (as done by [Ma et al. \(2016\)](#)) results in better test-time accuracy. We do not crop or rotate the images, nor do we augment the dataset in any other way.

### Model Precision

We find an empirically optimal model precision  $\tau$  (see Equation (5.13)) by optimizing the Average Normalized Estimation Error Squared (ANees) across the entire test set for each dataset. While this hyperparameter should in principle be tuned using a validation set, we omit this step to keep our training procedure consistent with that of [Ma et al. \(2016\)](#). We note that the BCNN uncertainty estimates are affected by two significant factors: 1) variational inference is known to underestimate predictive variance ([Gal, 2016](#)); and 2) we assume the observation noise is homoscedastic. As noted by [Gal \(2016\)](#), the BCNN can be made heteroscedastic by learning the model precision during training, but this extension is outside the scope of this work.

### Data Partitioning

We partition our data into training and testing sets using a leave-one-out approach based on temporally disjoint sequences of images. That is, given  $N$  sequences, the model tested on sequence  $i$  is trained with sequences  $\{1, 2, \dots, N\} \setminus i$ . This process varies based on the dataset, and we discuss the specifics in the experimental discussion corresponding to each. In contrast to randomly holding out a subset of the data, this method minimizes the similarity of training and testing data for temporally correlated image streams.

## 5.7 Simulation Experiments

We assess the benefit of incorporating sun observations of varying quality by conducting a series of simulation experiments consisting of a stereo camera moving along loopy trajectories of varying shapes through a simulated field of point landmarks, with a single static directional landmark representing the sun. Figure 5.5 shows several such loopy trajectories.

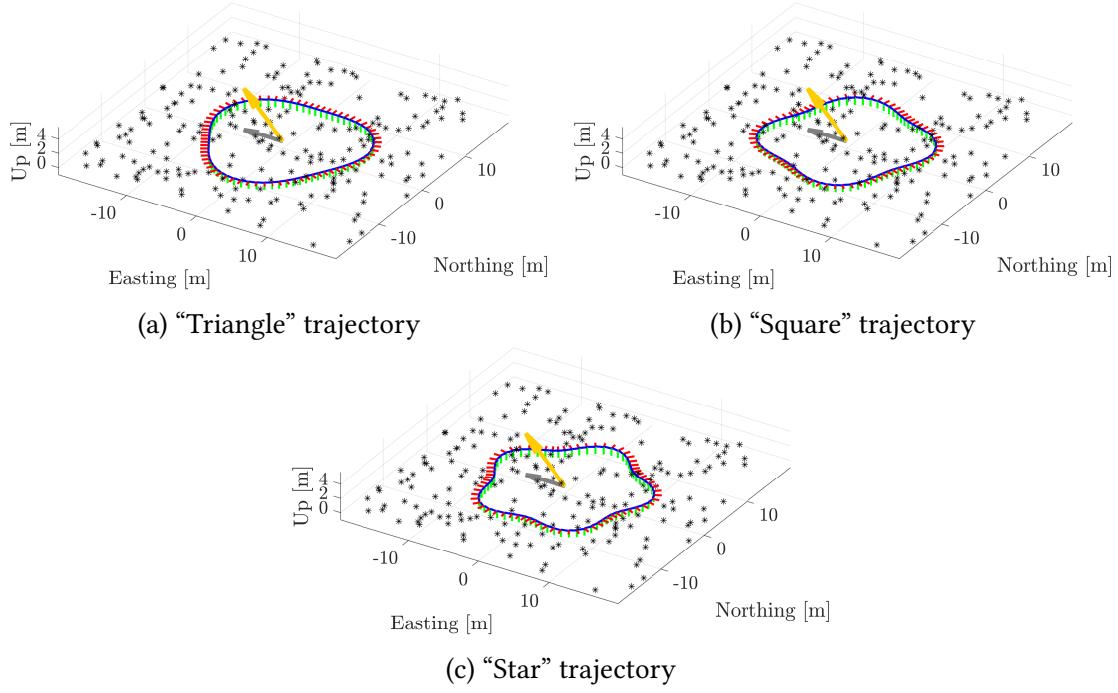


Figure 5.5: One loop of the “Triangle”, “Square”, and “Star” trajectories, consisting primarily of translation and yaw rotation. Landmarks are shown as black asterisks, and the simulated sun direction is indicated with a yellow arrow along with its projection, in grey, on the EN-plane.

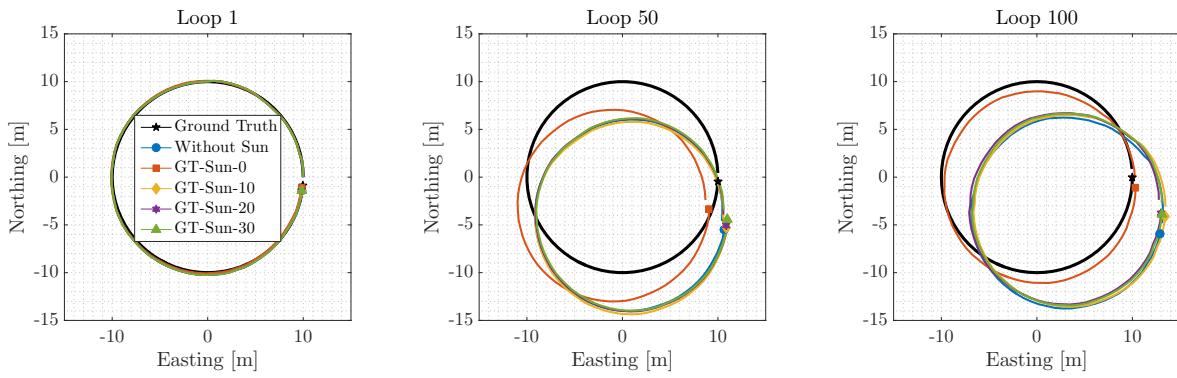


Figure 5.6: Selected segments of a 100-loop “Circle” trajectory, without sun corrections, and with sun corrections corrupted by varying levels of artificial Gaussian noise. The effect of VO drift can be clearly seen, as well as the benefit of incorporating observations of a directional landmark such as the sun.

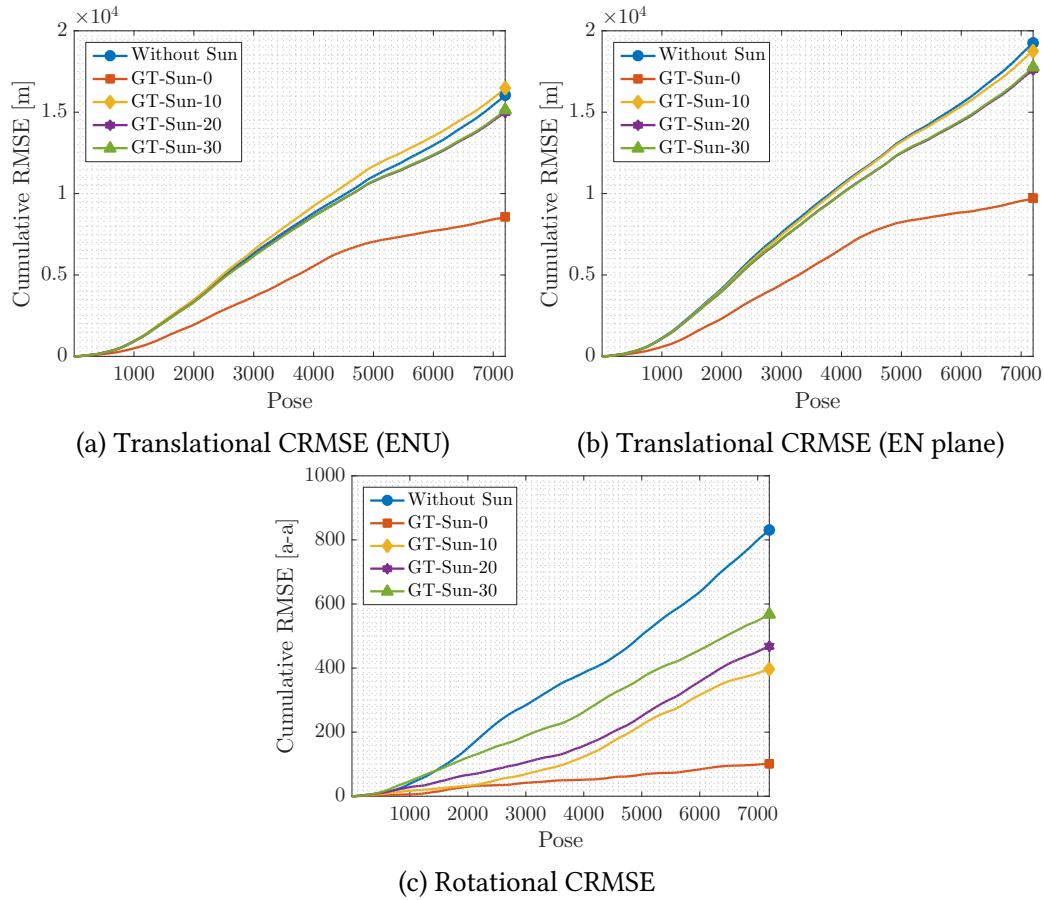


Figure 5.7: Cumulative root mean squared error (CRMSE) of a simulated 100-loop circular trajectory, without sun corrections, and with sun corrections corrupted by varying levels of artificial Gaussian noise. The accumulated estimation error is greatly reduced by incorporating observations of the sun, and the benefit decreases as these observations become noisier.

Table 5.1: Comparison of translational and rotational average root mean squared errors (ARMSE) on simulated sequences.

<b>Loop Shape</b>	Circle	Triangle	Square	Star
<b># Loops</b>	100	100	100	100
<b>Trans. ARMSE [m]</b>				
Without Sun	2.22	2.00	2.33	1.41
GT-Sun-0	1.19	1.62	2.13	0.75
GT-Sun-10	2.29	2.07	2.05	1.32
GT-Sun-20	2.08	2.12	2.31	1.33
GT-Sun-30	2.10	1.95	2.16	1.38
<b>Trans. ARMSE (EN-plane) [m]</b>				
Without Sun	2.67	1.88	2.57	1.10
GT-Sun-0	1.34	1.89	2.56	0.83
GT-Sun-10	2.61	2.04	2.26	0.99
GT-Sun-20	2.44	2.03	2.57	0.88
GT-Sun-30	2.46	2.00	2.35	1.25
<b>Rot. ARMSE (<math>\times 10^{-3}</math>) [axis-angle]</b>				
Without Sun	115.32	144.56	107.27	111.19
GT-Sun-0	14.10	113.58	59.21	30.69
GT-Sun-10	55.22	115.03	75.62	39.17
GT-Sun-20	65.02	121.11	80.41	49.75
GT-Sun-30	78.73	145.22	100.91	72.39

We simulate the sun at  $45^\circ$  of zenith and an arbitrary azimuth angle, and corrupt observations of the ground truth sun vector with artificial noise such that the mean angular distance (a non-negative quantity) between the observed and true sun direction is  $0^\circ$ ,  $10^\circ$ ,  $20^\circ$ , and  $30^\circ$ , labeling these conditions *GT-Sun-0*, *GT-Sun-10*, *GT-Sun-20*, and *GT-Sun-30*, respectively. In our experiments, we treated the measurement noise as an additive quantity sampled from a zero-mean isotropic 3D Gaussian distribution, and renormalized the resulting vectors to enforce the unit-norm constraint.

We simulate the sun at  $45^\circ$  of zenith and an arbitrary azimuth angle, and corrupt observations of the ground truth sun vector with artificial noise such that the mean angular distance (a non-negative quantity computed from the dot product) between the ground truth and noisy sun vectors is  $0^\circ$ ,  $10^\circ$ ,  $20^\circ$ , or  $30^\circ$ . We label these conditions *GT-Sun-0*, *GT-Sun-10*, *GT-Sun-20*, and *GT-Sun-30*, respectively. We generated these noisy measurements by first sampling 3-vectors from an isotropic zero-mean multivariate Gaussian distribution, then adding these vectors to the ground truth sun vector, and finally normalizing the result to unit length. We chose the covariance of this distribution to yield the desired average angular distance in each case. Note that although the distribution from which we sample noise vectors is zero-mean, the average angular distances will not be zero-mean because angular distance is non-negative.

Our choice to add noise in  $\mathbb{R}^3$  and re-normalize was motivated by the fact that this process yields approximately Gaussian error distributions over the azimuth and zenith error angles, which is an important property assumed by our VO pipeline to produce maximum likelihood motion estimates based on the fusion of multiple data sources. We note that these distributions are less Gaussian-like for larger covariances (due to the geometry of the unit 2-sphere) and for ground truth vectors near singularities (e.g., zero zenith).

We also experimented with sampling simulated measurements from a Von Mises-Fisher distribution (Fisher, 1953), which is approximately analogous to an isotropic Gaussian distribution that respects the geodesics on the unit 2-sphere. However, we observed that the resulting distributions on azimuth and zenith error were severely non-Gaussian, which violated the assumption of zero-mean Gaussian noise in our VO pipeline and interfered with our VO experiments.

Since our VO pipeline does not incorporate loop closures, the effects of drift in the VO solution can be clearly seen by examining individual loops in the camera trajectory. Figure 5.6 shows three loops from the “Circle” trajectory, demonstrating that the VO solution drifts significantly from the true trajectory by the 100th loop. Figure 5.7 plots the translational and rotational cumulative root mean squared error (CRMSE) for this trajectory, which measures the growth in total estimation error over time. Figure 5.7c in particular highlights the significant effect of sun sensing on rotational error, where we see a clear progression in estimation

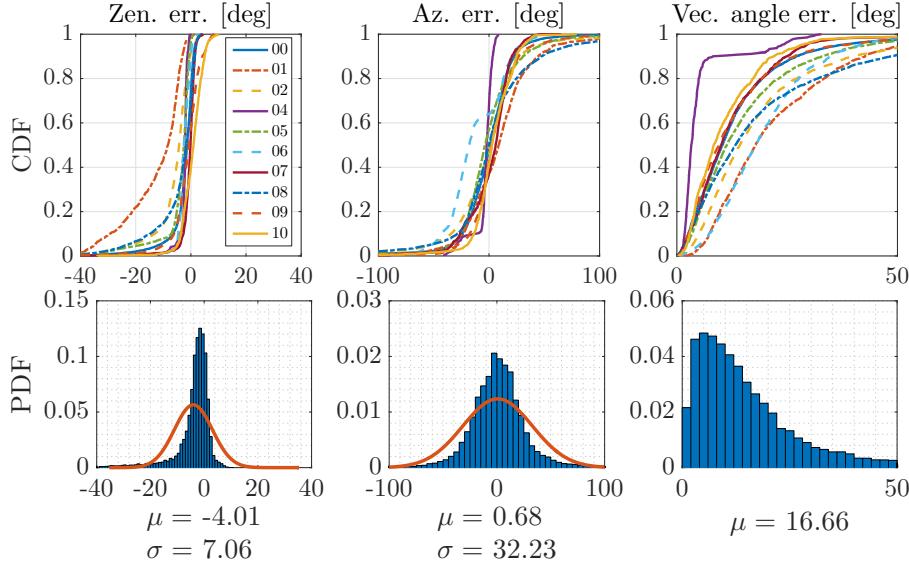


Figure 5.8: Distributions of azimuth error, zenith error, and angular distance for Sun-BCNN compared to ground truth over each test sequence in the KITTI dataset. *Top row:* Cumulative distributions of errors for each test sequence individually. *Bottom row:* Histograms and Gaussian fits of aggregated errors.

error as the sun direction observations become more noisy.

Table 5.1 shows that while all four simulation trajectories display consistent and predictable reductions in rotational average root mean squared error (ARMSE), this is not always the case for translational ARMSE. This is because translational errors are only partially induced by rotational errors, with the remainder made up of ‘sliding’ motions orthogonal to the direction of travel. These non-rotational errors are highly dependent on the specific trajectory, where more or less of the observed feature tracks can be explained by a sliding motion instead of a rotation. Due to the coupling of translational and rotational errors, correcting for rotational error in such cases may actually worsen the translational error (e.g., on the “Triangle” sequence).

While we do not implement this in our work, we speculate that incorporating an appropriate motion model into our VO formulation would significantly mitigate the impact of these errors by, for example, imposing a nonholonomic constraint on a ground vehicle or accounting for the dynamics of a quadcopter.

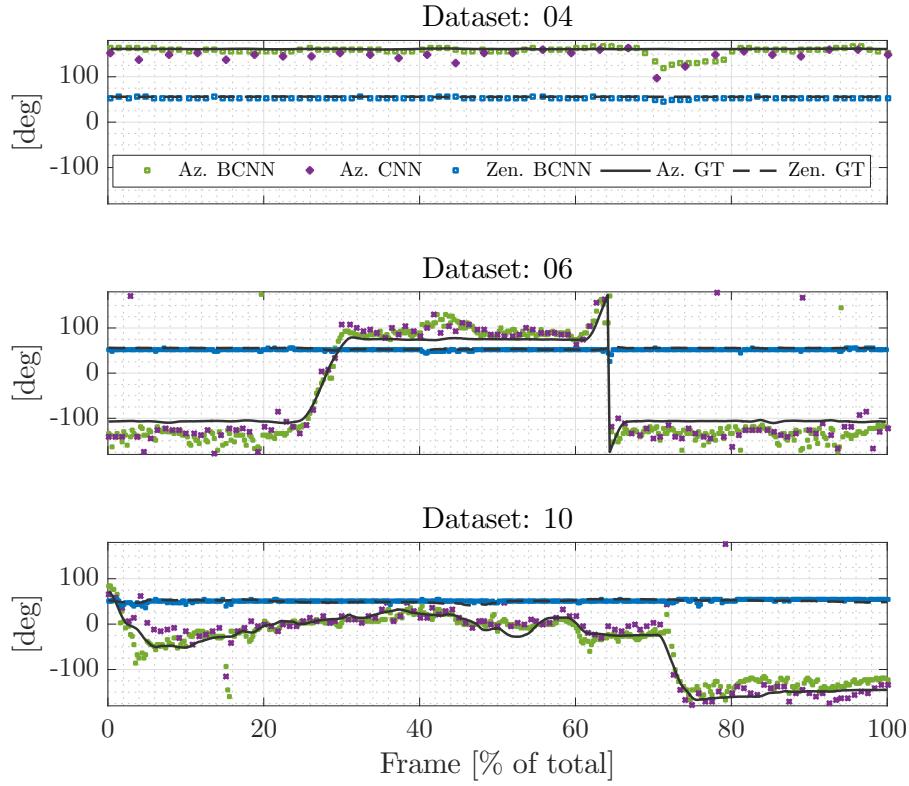


Figure 5.9: Azimuth (Sun-CNN and Sun-BCNN) and zenith (Sun-BCNN only) predictions over time for KITTI test sequences 04, 06 and 10. Sun-CNN is trained and tested on every tenth image, whereas Sun-BCNN is trained and tested on every image. In our VO experiments, we use the Sun-BCNN predictions of every tenth image to make a fair comparison.

Table 5.2: Test Errors for Sun-BCNN on KITTI odometry sequences with estimates computed at every image.

<b>Sequence</b>	<b>Zenith Error [deg]</b>			<b>Azimuth Error [deg]</b>			<b>Vector Angle Error [deg]</b>			<b>ANEE<sup>1</sup></b>
	Mean	Median	Stdev	Mean	Median	Stdev	Mean	Median	Stdev	
00	-2.59	-1.37	5.15	-0.33	0.81	25.61	13.56	10.31	13.14	1.00
01	-12.53	-8.31	10.33	8.95	8.83	33.67	22.16	17.85	15.00	1.38
02	-6.13	-4.26	7.38	-1.03	0.74	37.61	19.69	14.32	18.25	1.40
04	-2.42	-2.11	1.64	-3.89	-2.18	9.14	5.33	3.29	6.44	0.30
05	-4.31	-2.51	6.18	-0.74	-3.80	29.81	15.66	11.33	14.80	1.05
06	-2.48	-2.52	2.27	-12.22	-17.86	25.78	19.78	17.72	11.35	1.93
07	-0.69	-0.16	3.26	1.25	5.98	20.27	12.44	10.05	9.97	0.97
08	-4.46	-1.61	8.14	3.66	-0.14	41.73	19.90	13.30	19.59	1.04
09	-1.35	-0.75	5.60	4.78	2.36	23.84	13.09	9.48	12.66	0.73
10	0.59	0.95	3.90	3.64	2.61	19.15	11.23	8.34	9.83	1.08
All	-4.01	-2.26	7.06	0.68	0.53	32.23	16.66	12.08	15.91	-

<sup>1</sup> We compute Average Normalized Estimation Error Squared (ANEE<sup>1</sup>) values with all sun directions that fall below a cosine distance threshold of 0.3 (relative to ground truth) and set  $\tau^{-1} = 0.015$ .

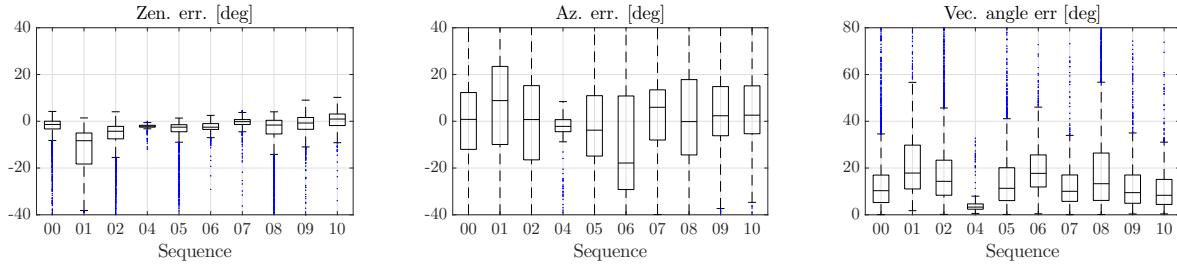


Figure 5.10: Box-and-whiskers plot of final test errors on all ten KITTI odometry sequences (c.f. Table 5.2).

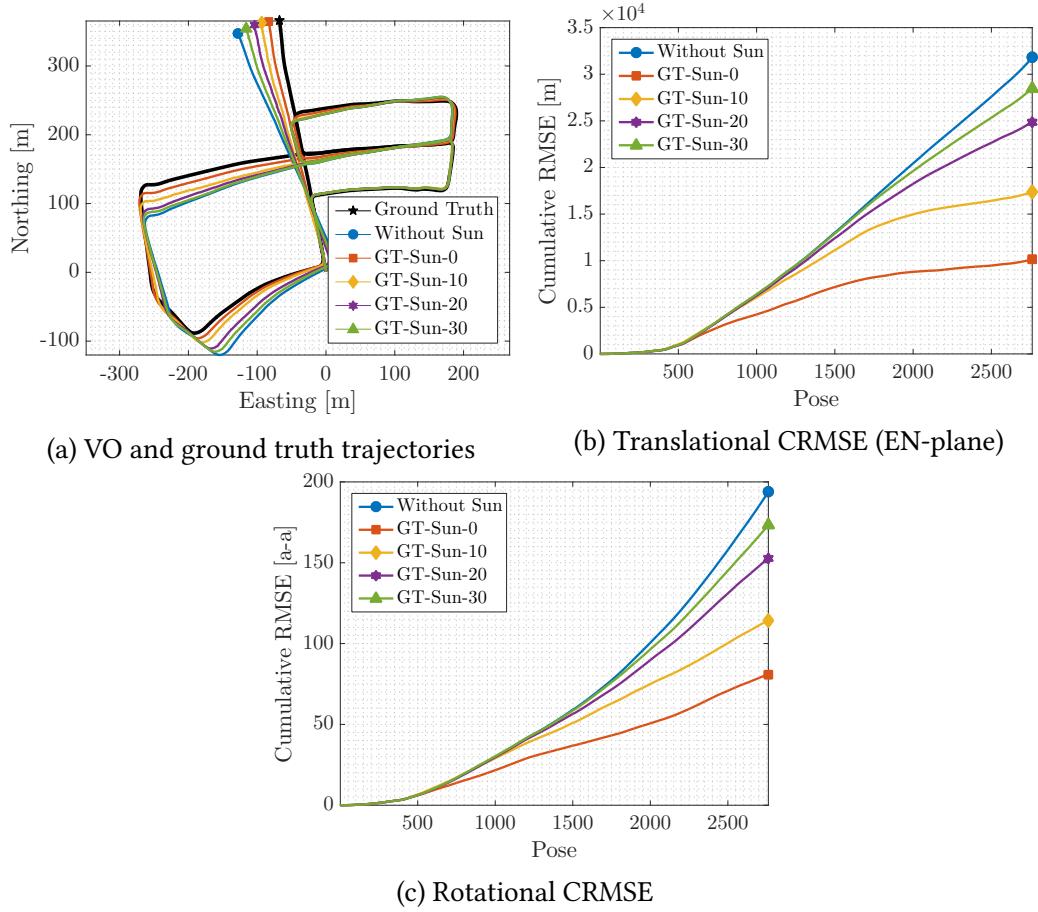


Figure 5.11: VO results for KITTI odometry sequence 05 using simulated sun measurements at every tenth pose. We observe a clear progression in cumulative root mean squared error (CRMSE) in translation and rotation as noise in the simulated sun measurements increases.

## 5.8 Urban Driving Experiments: The KITTI Odometry Benchmark

We investigated the performance of Sun-BCNN on the KITTI odometry benchmark training set (Geiger et al., 2013), which consists of 21.6 km of urban driving data<sup>2</sup>. Importantly, the dataset includes 6-DOF ground truth poses obtained from an accurate GPS/INS tracking system, as well as calibrated transformations between this sensor and the colour stereo pair we use for sun estimation and VO in our experiments. This allows us to create a training set of ground truth sun vectors for each image by querying the solar ephemeris model at each ground truth pose and rotating the resulting vector from the GPS/INS frame  $\mathcal{F}_0$  (which is an ENU coordinate system) into the camera coordinate frame  $\mathcal{F}_k$ . For each of our experiments, we trained Sun-BCNN on nine benchmark sequences and tested on the remaining sequence. This procedure is consistent with that of Ma et al. (2016), against whose Sun-CNN we directly compare, and allows us to evaluate each sequence using the maximum amount of training data.



Figure 5.12: Sun BCNN predictions and associated ground truth sun directions on the KITTI sequence 05. *Top two rows:* Sun BCNN produces accurate predictions in a variety of azimuth values. *Bottom row:* Poor results occur rarely due to shadow ambiguities.

---

<sup>2</sup>Because we rely on the first pose reported by the GPS/INS system, we used the raw (rectified and synchronized) sequences corresponding to each odometry sequence. However, the raw sequence 2011\_09\_26\_drive\_0067 corresponding to odometry sequence 03 was not available on the KITTI website at the time of writing, so we omit sequence 03 from our analysis.

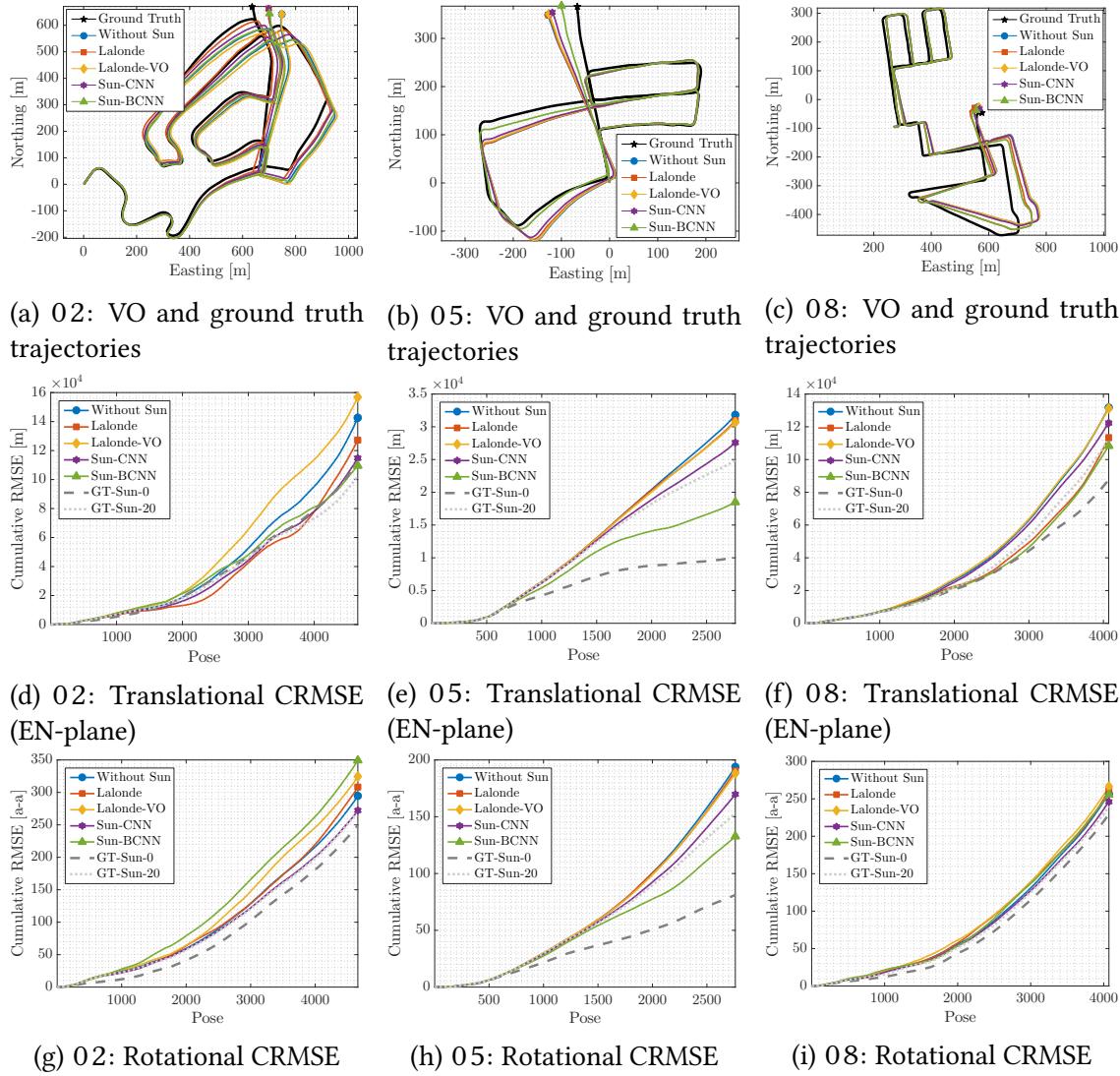


Figure 5.13: VO results for KITTI odometry sequences 02, 05, and 08 using estimate sun directions at every tenth pose. *Top row*: Estimated and ground truth trajectories in the Easting-Northing (EN) plane. *Middle row*: Translational cumulative root mean squared error (CRMSE) in the EN-plane. *Bottom row*: Rotational CRMSE. Sun-BCNN significantly reduces the estimation error on sequence 05, while the Lalonde (Lalonde et al., 2011), Lalonde-VO (Clement et al., 2017), and Sun-CNN (Ma et al., 2016) methods provide modest reductions in estimation error. The remaining sequences are less clear, but Sun-BCNN generally provides some benefit.

Table 5.3: Comparison of translational and rotational average root mean squared error (ARMSE) on KITTI odometry sequences with and without sun direction estimates at every tenth image. The best result (excluding simulated sun sensing) is highlighted in bold.

Sequence <sup>1</sup>	00	01 <sup>2</sup>	02	04	05	06	07	08	09	10
Length [km]	3.7	2.5	5.1	0.4	2.2	1.2	0.7	3.2	1.7	0.9
<b>Trans. ARMSE [m]</b>										
Without Sun	4.33	198.52	28.59	2.48	9.90	3.35	4.55	28.05	10.44	5.54
GT-Sun-0	5.40	114.69	23.83	2.23	4.84	3.50	1.58	31.55	8.21	3.67
GT-Sun-10	4.85	123.84	25.34	2.45	5.84	2.80	2.94	28.47	8.65	4.81
GT-Sun-20	4.78	136.60	22.33	2.46	8.16	3.03	3.90	27.54	8.68	5.45
GT-Sun-30	4.83	157.14	27.30	2.48	8.93	3.44	4.62	26.73	10.10	5.28
Lalonde	<b>3.81</b>	200.34	28.13	<b>2.47</b>	9.88	3.36	4.61	29.70	10.49	<b>5.48</b>
Lalonde-VO	4.87	199.03	29.41	2.48	9.74	<b>3.30</b>	4.52	27.82	11.06	5.59
Sun-CNN	4.36	192.50	<b>26.58</b>	2.48	8.92	3.38	4.30	<b>26.99</b>	10.15	5.58
Sun-BCNN	4.44	<b>188.46</b>	26.89	2.48	<b>8.50</b>	4.10	<b>4.21</b>	27.71	<b>10.13</b>	5.61
<b>Trans. ARMSE (EN-plane) [m]</b>										
Without Sun	4.53	230.73	30.66	1.81	11.50	3.68	5.44	32.37	11.65	5.95
GT-Sun-0	3.41	136.76	24.12	1.46	3.67	3.96	1.80	21.51	7.77	3.71
GT-Sun-10	5.05	149.36	24.79	1.79	6.29	2.73	3.51	22.41	8.90	5.09
GT-Sun-20	5.14	164.37	22.04	1.80	9.01	3.13	4.66	27.58	8.86	5.81
GT-Sun-30	5.12	188.61	22.65	1.83	10.31	3.83	5.50	27.65	11.16	5.58
Lalonde	<b>3.95</b>	232.66	27.30	<b>1.81</b>	11.20	3.70	5.52	27.84	11.41	<b>5.87</b>
Lalonde-VO	5.38	231.33	33.68	1.82	11.13	<b>3.61</b>	5.42	32.24	12.41	6.00
Sun-CNN	4.56	224.91	24.65	1.82	9.99	3.74	5.16	30.09	11.21	5.99
Sun-BCNN	4.68	<b>220.54</b>	<b>23.58</b>	1.82	<b>6.70</b>	4.78	<b>5.05</b>	<b>26.59</b>	<b>10.97</b>	6.03
<b>Rot. ARMSE (<math>\times 10^{-3}</math>) [axis-angle]</b>										
Without Sun	23.88	185.30	63.18	12.97	70.18	23.24	49.96	63.13	26.77	21.54
GT-Sun-0	11.20	38.82	53.48	11.75	29.38	17.66	20.37	56.39	17.00	12.60
GT-Sun-10	17.05	64.51	58.78	12.86	41.47	18.90	34.05	54.89	19.71	14.26
GT-Sun-20	18.84	94.65	58.03	12.91	55.39	19.67	43.34	58.82	20.99	25.87
GT-Sun-30	23.40	121.21	57.79	13.01	62.73	23.96	49.92	56.74	25.63	20.15
Lalonde	<b>21.10</b>	188.06	66.02	<b>12.96</b>	69.00	23.27	50.49	64.22	26.27	<b>20.49</b>
Lalonde-VO	27.91	185.52	69.52	12.98	68.09	<b>22.79</b>	49.74	65.35	28.82	22.10
Sun-CNN	24.05	177.45	<b>58.32</b>	13.00	61.48	23.34	47.77	<b>60.55</b>	<b>26.19</b>	21.99
Sun-BCNN	26.96	<b>175.21</b>	75.02	13.00	<b>47.96</b>	23.80	<b>47.57</b>	62.85	26.29	20.85

<sup>1</sup> Because we rely on the timestamps and first pose reported by the GPS/INS system, we use the raw (rectified and synchronized) sequences corresponding to each odometry sequence. However, the raw sequence 2011\_09\_26\_drive\_0067 corresponding to odometry sequence 03 was not available on the KITTI website at the time of writing, so we omit sequence 03 from our analysis.

<sup>2</sup> Sequence 01 consists largely of self-similar, corridor-like highway driving which causes difficulties when detecting and matching features using libviso2. The base VO result is of low quality, although we note that including global orientation from the sun nevertheless improves the VO result.

### 5.8.1 Sun-BCNN Test Results

Once trained, we analyzed the accuracy and consistency of the Sun-BCNN mean and covariance estimates. We obtained the mean estimated sun vector by evaluating Equation (5.11) with  $N = 25$  and then re-normalized the resulting vector to preserve unit length. To obtain the required covariance on azimuth and zenith angles, we sampled the vector outputs, converted them to azimuth and zenith angles using Equation (5.7), and then applied Equation (5.12). We investigate the impact of this parametrization (as opposed to working in azimuth and zenith coordinates directly) later in this paper. As shown in Table 5.2, we chose a value for the model precision  $\tau$  such that the Average Normalized Estimation Error Squared (ANees) of each test sequence is close to one (i.e., the estimator is consistent).

Figures 5.8 and 5.10 plot the error distributions for azimuth, zenith, and angular distance for all ten KITTI odometry sequences, while Figure 5.9 shows three characteristic plots of the azimuth and zenith predictions over time. We see that the errors in azimuth and zenith are strongly peaked around zero and are reasonably well described by a Gaussian distribution, which are important properties assumed by our VO pipeline to produce maximum likelihood motion estimates based on the fusion of multiple data sources. Note that the error distribution in zenith is slightly biased towards negative values due to the presence of a long tail on the negative side of the mean. This is an artifact of the azimuth-zenith parameterization when the sun zenith is small (i.e., when the sun is high in the sky), since zenith angles are defined on  $[0, \pi]$ . In practice, we attempt to reduce the influence of the long negative tail by imposing a robust Huber loss on the sun measurement errors in our optimization problem.



Figure 5.14: GPS track and sample images from the Devon Island traverse, with the start of each sequence highlighted. The Devon Island dataset is conducive to visual sun sensing due to the presence of strong environmental shadows, reflective surfaces such as mud and water, occasionally visible sun, and self-shadowing by the sensor platform. (Map data: Google, DigitalGlobe)

Table 5.2 summarizes the Sun-BCNN test errors numerically. Sun-BCNN achieved median vector angle errors of less than 15 degrees on every sequence except sequence 01 and 06, which were particularly difficult in places due to challenging lighting conditions. It is interesting to note that sequences 00 and 06 also have higher than average ANees values,

which indicates that the estimator is overconfident in its estimates despite their low quality. We suspect this behaviour stems from the assumption of homoscedastic noise in the BCNN, which treats all input images as being equally amenable to sun estimation across the entire sequence.

### 5.8.2 Visual Odometry Experiments

We evaluated the influence of the estimated sun directions and covariances obtained from Sun-BCNN on the KITTI odometry benchmark using the sun-aided VO pipeline previously described. To place these results in context, we compare them against the results obtained using simulated sun measurements with varying levels of noise, the method of [Lalonde et al. \(2011\)](#) and its VO-informed variant ([Clement et al., 2017](#)), and the Sun-CNN of [Ma et al. \(2016\)](#).

#### Simulated Sun Sensing

In order to gauge the effectiveness of incorporating sun information in each sequence, and to determine the impact of measurement error, we constructed several sets of simulated sun measurements by computing ground truth sun vectors and artificially corrupting them with varying levels of zero-mean Gaussian noise. We obtained these ground truth sun vectors by transforming the ephemeris vector into each camera frame using ground truth vehicle poses. Using the same convention as our experiments with simulated trajectories, we created four such measurement sets with  $0^\circ$ ,  $10^\circ$ ,  $20^\circ$ , and  $30^\circ$  mean angular distance from ground truth.

Figure 5.11 shows the results we obtained using simulated sun measurements on sequence 05, in which the basic VO suffers from substantial orientation drift.<sup>3</sup> Incorporating absolute orientation information from the simulated sun sensor allows the VO to correct these errors, but the magnitude of the correction decreases as sensor noise increases, consistent with the results of our simulation experiments. As shown in Table 5.3, which summarizes our VO results for all ten sequences, this is typical of sequences where orientation drift is the dominant source of error.

While the VO solutions for sequences such as 00 do not improve in terms of translational ARMSE, Table 5.3 shows that rotational ARMSE nevertheless improves on all ten sequences when low-noise simulated sun measurements are included. This implies that the estimation errors of the basic VO solutions for certain sequences are dominated by non-rotational effects, and that the apparent benefit of the Lalonde method on translational ARMSE in sequence 00 is likely coincidental.

---

<sup>3</sup>In order to make a fair comparison to the Sun-CNN of [Ma et al. \(2016\)](#), who compute sun directions for every tenth image of the KITTI odometry benchmark, we subsample the sun directions obtained through each other method to match.

## Vision-based Sun Sensing

Figure 5.12 illustrates the behaviour of Sun-BCNN on four characteristic images from test sequence 05 by overlaying the Sun-BCNN predictions and associated ground truth sun directions for each image. The two frames in the top row both contain strong shadows which typically result in very accurate sun predictions. Conversely, the bottom row highlights two examples of rare situations where ambiguous shadows lead to very inaccurate predictions. As previously mentioned, we mitigate the influence of these outlier measurements by imposing a robust Huber loss on the sun measurement errors in our optimizer.

Figure 5.13 shows the results we obtained for sequences 02, 05, and 08 using the Sun-CNN of Ma et al. (2016), which estimates only the azimuth angle of the sun, our Bayesian Sun-BCNN which provides full 3D estimates of the sun direction as well as a measure of the uncertainty associated with each estimate, and the method of Lalonde et al. (2011) in its original and VO-informed (Clement et al., 2017) forms, which provide 3D estimates of the sun direction without reasoning about uncertainty. A selection of results using simulated sun measurements are also displayed for reference. All four sun detection methods succeed in reducing the growth of total estimation error on this sequence, with Sun-BCNN reducing both translational and rotational error growth significantly more than the other three methods. Both Sun-CNN and Sun-BCNN outperform the two Lalonde variants, consistent with the results of Ma et al. (2016) and Clement et al. (2017).

Table 5.3 shows results for all ten sequences using each method. With few exceptions, the VO results using Sun-BCNN achieve improvements in rotational and translational ARMSE comparable to those achieved using the simulated sun measurements with between 10 and 30 degrees average error. As previously noted, sequences such as 00 do not benefit significantly from sun sensing since rotational drift is not the dominant source of estimation error in these cases. Nevertheless, these results indicate that CNN-based sun sensing is a valuable tool for improving localization accuracy in VO and an improvement that comes without the need for additional sensors or a specially oriented camera.

## 5.9 Planetary Analogue Experiments: The Devon Island Rover Navigation Dataset

In addition to urban driving, we further investigate the usefulness of Sun-BCNN in the context of planetary exploration using the Devon Island Rover Navigation Dataset (Furgale et al., 2012), which consists of various sensor data collected using a mobile sensor platform traversing a 10 km loop on Devon Island in the Canadian High Arctic (Figure 5.14). The rugged

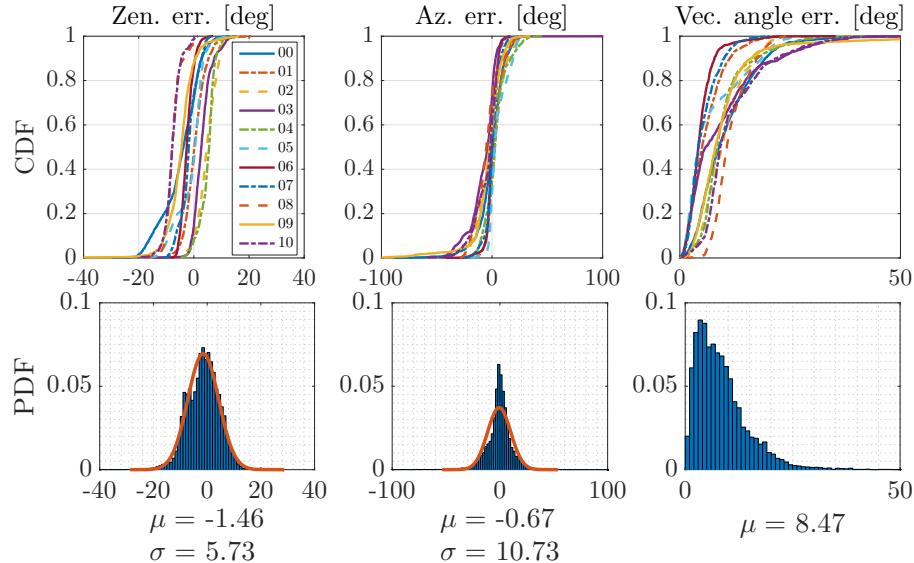


Figure 5.15: (Devon Island) Distributions of azimuth error, zenith error, and angular distance for Sun-BCNN compared to ground truth over each test sequence. *Top row:* Cumulative distributions of errors for each test sequence individually. *Bottom row:* Histograms and Gaussian fits of aggregated errors.

landscape of Devon Island (Figure 5.14) is a significant departure from the structured urban environment of Karlsruhe. Unlike the KITTI odometry benchmark, the Devon Island dataset provides ground truth vehicle orientations for only a small number of images, which means that our previous method of generating ground truth sun vectors using ground truth poses is not applicable. However, the sensor platform used to collect the dataset was equipped with a hardware sun sensor and inclinometer, both of which were used by [Lambert et al. \(2012\)](#) to correct VO drift. For our purposes, we ignore the inclinometer and use the sun sensor measurements as training targets for Sun-BCNN.

The Devon Island environment contains many features one might expect to be amenable to visual sun detection. As shown in Figure 5.14, the dataset contains strong environmental shadows, stretches of wet terrain featuring reflective mud and water, and some self-shadowing from the sensor platform itself. At times the sun is partially visible to the camera, although these images tend to be saturated and do not immediately allow for accurate localization of the sun in the image.

For the purposes of our experiments, we partition the dataset into 11 sequences of approximately 1 km each, chosen such that the full pose of the vehicle at the beginning of each sequence is available from the ground truth data (see Figure 5.14). In aggregate, the sequences contain 13257 poses with associated sun sensor measurements. We apply a similar training and testing procedure as for the KITTI dataset, with the exception that we now withhold one

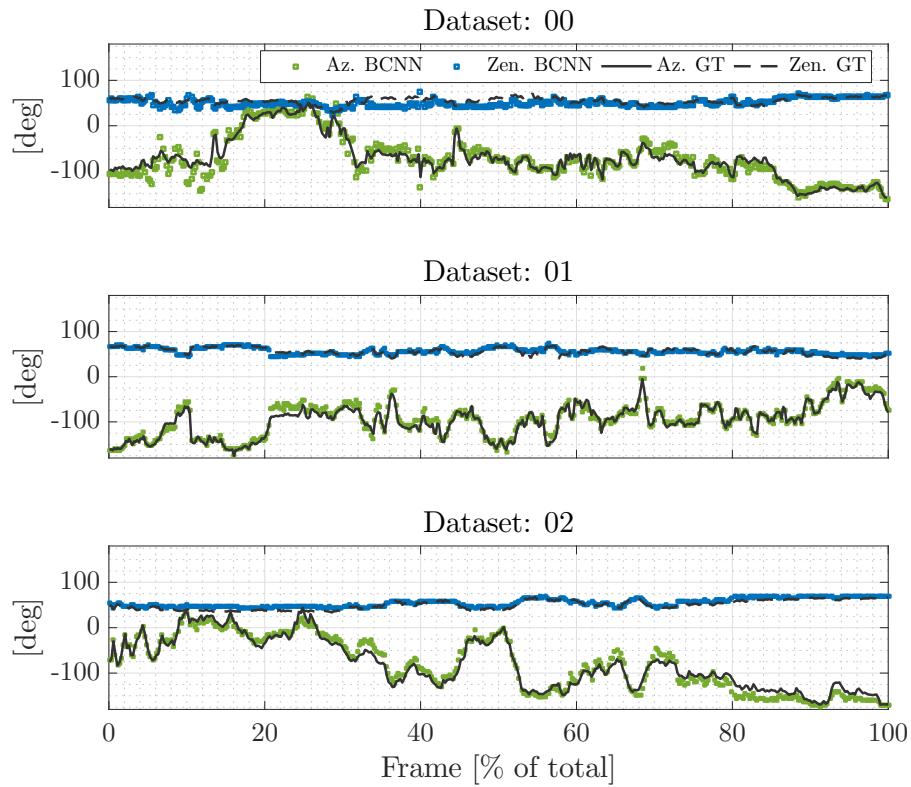


Figure 5.16: Azimuth (Sun-BCNN azimuth and zenith predictions over time for Devon Island test sequences 00, 01 and 12. Sun-BCNN is trained and tested on all frames (in our VO experiments, we use the Sun-BCNN predictions of every tenth image to make a fair comparison).

Table 5.4: Test Errors for Sun-BCNN on Devon Island odometry sequences with estimates computed at every image.

<b>Sequence</b>	<b>Zenith Error [deg]</b>			<b>Azimuth Error [deg]</b>			<b>Vector Angle Error [deg]</b>			<b>ANEE<sup>2</sup></b>
	Mean	Median	Stdev	Mean	Median	Stdev	Mean	Median	Stdev	
00	-4.77	-3.77	6.82	-0.65	0.69	12.41	10.48	8.86	6.96	1.27
01	0.47	0.21	3.91	2.96	2.31	7.01	5.97	5.06	4.01	0.59
02	4.66	4.68	3.52	-0.72	-1.32	11.78	10.02	9.51	4.76	1.37
03	3.09	2.70	3.41	-7.47	-4.03	12.88	9.39	5.83	8.75	1.11
04	4.93	5.53	2.90	3.27	2.72	10.09	9.78	8.41	5.60	0.89
05	-1.01	0.46	4.97	5.26	2.46	8.23	7.19	4.15	6.60	0.92
06	-2.45	-2.58	2.23	-0.23	-0.30	5.07	4.72	4.17	3.16	0.31
07	-1.80	-1.87	3.28	0.47	0.20	6.45	5.23	4.25	3.38	0.41
08	-7.46	-7.88	2.85	-4.93	-5.14	10.30	11.61	10.63	3.96	1.33
09	-4.72	-4.46	5.27	-3.91	-2.13	14.61	9.90	8.02	8.56	0.86
10	-7.69	-7.82	2.92	-4.81	-1.54	10.80	11.79	9.19	7.52	0.91
All	-1.46	-1.23	5.73	-0.67	-0.14	10.73	8.47	7.15	6.31	-

<sup>1</sup> We compute Average Normalized Estimation Error Squared (ANEE<sup>2</sup>) values with all sun directions that fall below a cosine distance threshold of 0.3 (relative to ground truth) and set  $\tau^{-1} = 0.01$ .

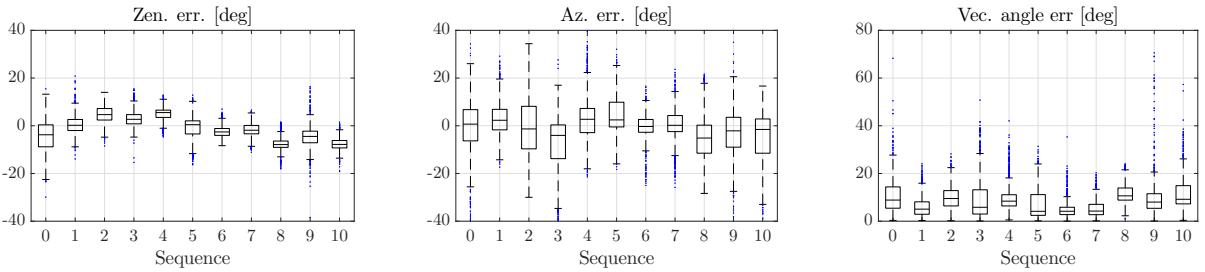


Figure 5.17: Box-and-whiskers plot of final test errors on Devon Island odometry sequences (c.f. ??).

sequence for validation and hyper-parameter tuning in addition to the sequence withheld for testing. This leaves nine sequences remaining to form the training sets for each test and validation pair.

### 5.9.1 Sun-BCNN Test Results

As in our experiments with the KITTI odometry benchmark, we obtained the mean estimated sun vector by evaluating Equation (5.11) with  $N = 25$  and re-normalizing the resulting vector to preserve unit length. To obtain the required covariance on azimuth and zenith angles, we again sampled the vector outputs, converted them to azimuth and zenith angles using Equation (5.7), and then applied Equation (5.12). As shown in Table 5.4, we chose a value for the model precision  $\tau$  such that the Average Normalized Estimation Error Squared (ANEE<sup>2</sup>) of each test sequence is close to one (i.e., the estimator is consistent).

Figures 5.15 and 5.17 plot the error distributions for azimuth, zenith, and angular distance for all 11 Devon Island odometry sequences, while Figure 5.16 shows three characteristic plots of the azimuth and zenith predictions over time. We see that the errors in azimuth and zenith are strongly peaked around zero and are better described by a Gaussian distribution than in the case of KITTI (c.f. Figure 5.8), which as we previously mentioned are important properties assumed by our VO pipeline to appropriately fuse data. The distribution of zenith errors in the Devon Island dataset does not exhibit the same bias and long tail we observed in the KITTI dataset. This is likely because the sun is much lower in the sky (i.e., the zenith angle is further from zero) in the Devon Island dataset than in the KITTI dataset, so there is no clipping of the distribution near zero zenith.

Table 5.4 summarizes the test errors and ANEES of each sequence numerically, while Figures 5.15 and 5.17 plot the error distributions for azimuth, zenith, and angular distance for each sequence. Figure 5.16 shows three characteristic plots of the azimuth and zenith predictions over time. Sun-BCNN achieved median vector angle errors of less than 10 degrees on every sequence except sequence 08. Consistent with the results we observed in the KITTI experiments, the sequences with the highest median vector angle error (sequences 02 and 08) also have the highest ANEES values, again indicating that the homoscedastic noise assumption is perhaps ill suited to this environment.

### 5.9.2 Visual Odometry Experiments

As in our KITTI benchmark experiments, we compare visual odometry results on each of our 11 test sequences both with sun-based orientation corrections and without. Notably, we do not report results using simulated sun measurements since we are unable to generate these measurements without ground truth vehicle orientations for every image. We also do not report results using the Sun-CNN of Ma et al. (2016) since we do not have access to their model. However, we do compare the results obtained using Sun-BCNN to those obtained using the hardware sun sensor as well as the Lalonde (Lalonde et al., 2011) and Lalonde-VO (Clement et al., 2017) methods.

Figure 5.18 shows sample VO results on three sequences from the Devon Island dataset using no sun measurements, the hardware sun sensor, Sun-BCNN, and the Lalonde variants. While the Lalonde methods struggle in this environment, Sun-BCNN yields significant improvements in VO accuracy, nearly on par with those obtained using the hardware sun sensor.

Table 5.5 summarizes these results numerically for all 11 sequences in the dataset. While the addition of sun sensing using either the hardware sensor or Sun-BCNN generally results in significant reductions in error, we note that in certain cases (e.g., sequence 05), sun sensing

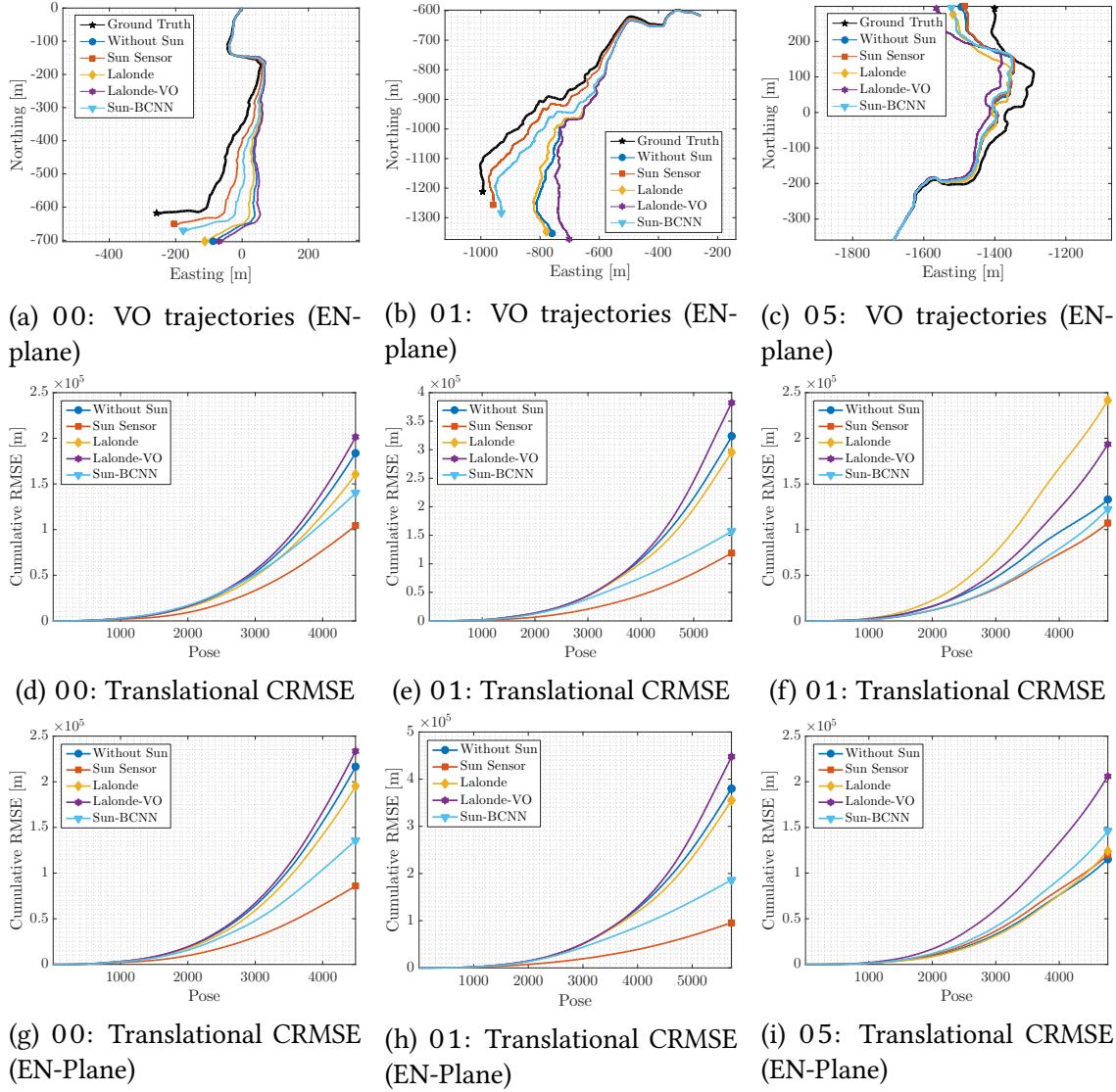


Figure 5.18: VO results for Devon Island sequences 00, 01, and 05 using estimated sun directions. *Top row*: Estimated and ground truth trajectories in the EN-plane. *Bottom rows*: Translational cumulative root mean squared error (CRMSE). Sun-BCNN significantly reduces the estimation error on sequences where the sun sensing has an impact (c.f. ??).

Table 5.5: Comparison of average root mean squared error (ARMSE) on Devon Island sequences with and without sun direction estimates using both a hardware sun sensor and vision-based methods. The best result using a vision-based method is bolded.

Sequence	00	01	02	03	04	05	06	07	08	09	10
Length [km]	0.9	1.1	1.0	1.0	0.9	1.0	1.1	1.0	0.9	0.7	0.6
<b>Trans. ARMSE [m]</b>											
Without Sun	40.93	56.51	41.58	42.04	30.52	27.82	58.91	40.04	47.22	11.39	12.94
Hardware Sun Sensor	23.26	20.79	9.79	22.03	30.79	22.47	24.14	29.59	47.97	6.26	8.50
Lalonde	35.77	51.74	53.32	47.00	39.55	50.70	94.77	59.37	<b>45.78</b>	10.03	16.23
Lalonde-VO	44.83	66.91	44.17	59.84	42.87	40.62	52.16	36.04	50.52	11.34	16.74
Sun-BCNN	<b>31.17</b>	<b>27.45</b>	<b>16.00</b>	<b>26.02</b>	<b>29.34</b>	<b>25.70</b>	<b>33.43</b>	<b>32.25</b>	50.80	<b>4.27</b>	<b>14.92</b>
<b>Trans. ARMSE (EN-plane) [m]</b>											
Without Sun	48.20	66.49	43.58	45.92	31.08	24.23	43.01	22.33	40.85	9.30	15.59
Hardware Sun Sensor	19.13	16.74	8.99	21.18	28.27	25.08	29.27	21.76	28.89	5.14	9.70
Lalonde	43.45	62.03	36.21	49.44	<b>20.13</b>	<b>26.13</b>	53.22	<b>18.10</b>	35.62	6.01	18.45
Lalonde-VO	52.05	78.26	40.20	59.09	50.12	43.28	53.62	42.71	49.99	11.74	20.17
Sun-BCNN	<b>30.28</b>	<b>32.65</b>	<b>9.62</b>	<b>14.32</b>	33.26	30.62	<b>36.44</b>	23.18	<b>13.53</b>	<b>4.45</b>	<b>14.75</b>

has little or no impact on the VO result. We suspect that the translation errors in these cases are dominated by non-rotational effects, similarly to those observed in our experiments with the KITTI dataset, although it is difficult to be certain in the absence of rotational ground truth. As previously mentioned, the incorporation of a motion prior in the VO estimator would likely reduce the impact of these errors.

## 5.10 Sensitivity Analysis

In this section we analyze the sensitivity of our model to cloud cover, investigate the possibility of model transfer between urban and planetary analogue environments, and examine the impact of different methods for computing the mean and covariance of a norm-constrained vector on the accuracy and consistency of the estimated sun directions.

### 5.10.1 Cloud Cover

Given that both the KITTI and Devon Island datasets were collected in sunny conditions, it is natural to wonder whether and to what extent Sun-BCNN is affected by cloud cover. As shown in Figure 5.4, Sun-BCNN relies in part on shadows and other local illumination variations to estimate the direction of the sun. Since the diffuse nature of daylight in cloudy conditions tends to soften shadows and other shading variations, one might expect Sun-BCNN to perform worse in cloudy conditions. Accordingly, we investigated the effect of cloud cover

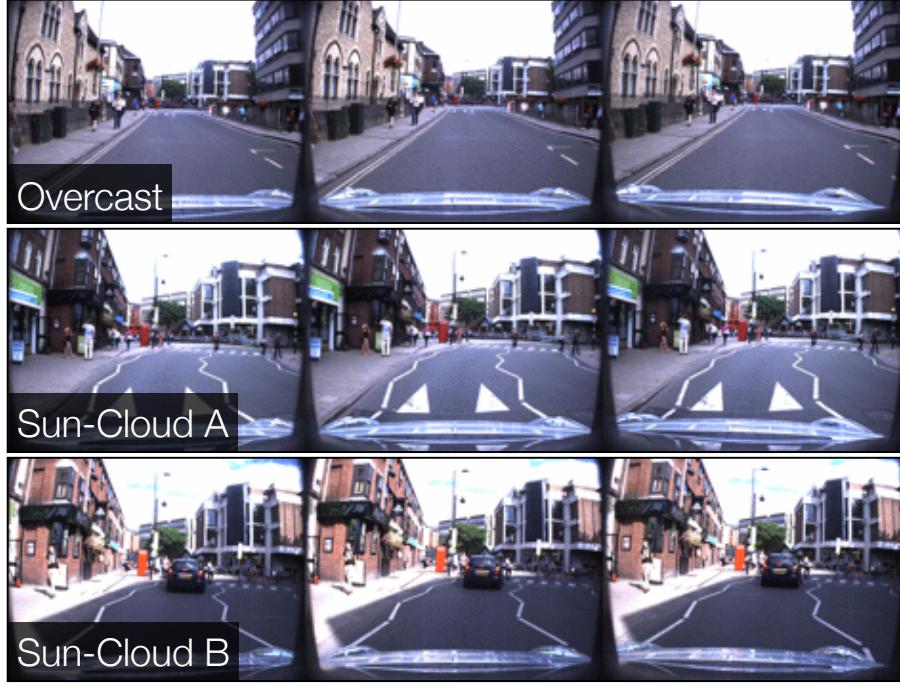


Figure 5.19: Sample images of approximately the same location taken from three different Oxford Robotcar sequences we used to investigate the effect of cloud cover on Sun-BCNN.

on Sun-BCNN using selected sequences from the Oxford Robotcar Dataset (Maddern et al., 2016), which consists of 1000 km of urban driving along a consistent route but in varying weather conditions and at varying times over the course of a year.

### Procedure

We selected three sequences collected within a two hour period on the same day (namely 2014-07-14-14-49-50, 2014-07-14-15-16-36, and 2014-07-14-15-42-55), which consist of the same route observed under different lighting conditions. Figure 5.19 presents sample images from each of these sequences, which we label *Overcast*, *Sun-Cloud A*, and *Sun-Cloud B*, respectively. To evaluate the performance of Sun-BCNN in each of these conditions, we partition each sequence into a randomly selected set of training (80%), validation (10%) and test (10%) images, and then train and test Sun-BCNN on each of the nine train-test permutations.

### Results

Figure 5.20 shows the results of these experiments with box and whisker plots for azimuth, zenith and vector angle errors while Table 5.6 summarizes the results numerically. We obtained the most accurate test predictions using the model trained on *Sun-Cloud B*, the se-

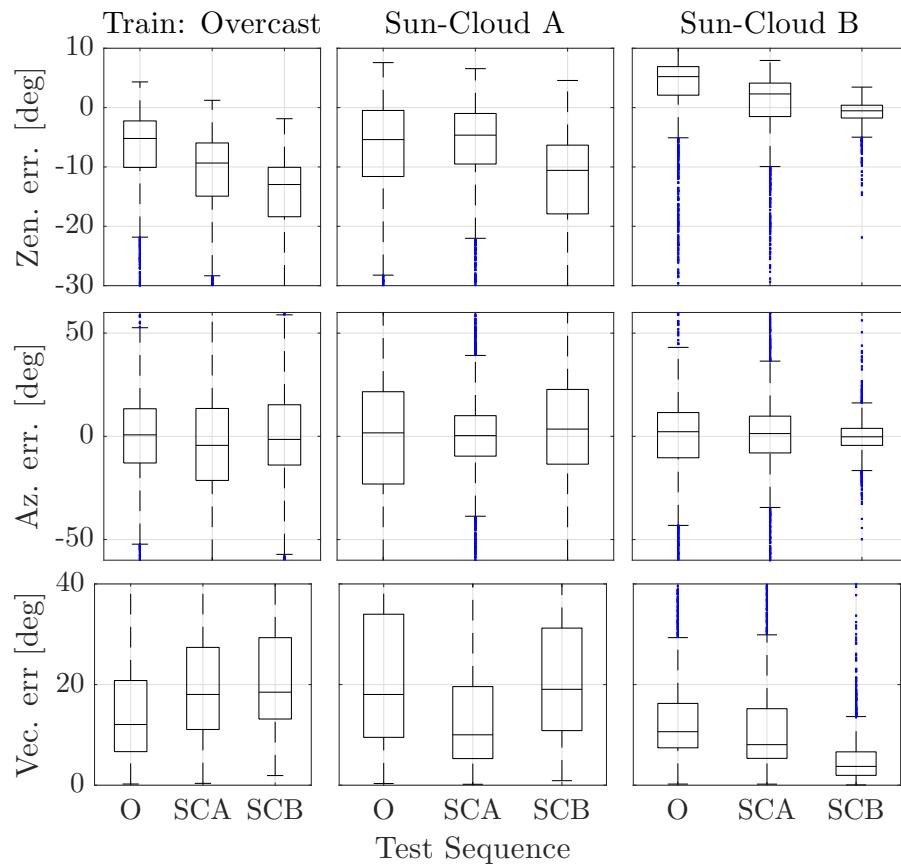


Figure 5.20: Box-and-whiskers plot for zenith, azimuth and vector angle errors for nine different combinations of train-test sequences taken from the Oxford Robotcar dataset. Each column corresponds to a different training sequence, and each plot contains three different test sequences. In the bottom legend, we use the labels O: *Overcast*, SCA: *Sun-Cloud A*, SCB: *Sun-Cloud B*.

quence with the least amount of cloud cover. Notably, this model produced vector angle errors on the *Overcast* test set that were lower than those trained with its own *Overcast* training set. Moreover, we note that the *Sun-Cloud A* model achieved similar test errors when applied to the *Sun-Cloud B* test set as when applied to the *Overcast* test set. Similarly, the *Sun-Cloud B* model achieved similar test errors when applied to the *Sun-Cloud A* test set as when applied to the *Overcast* test set. From this we can conclude the following: 1) that Sun-BCNN can still perform well in the presence of cloud cover; and 2) that training in environments illuminated by strong directional light (i.e., sunny conditions) can significantly improve sun estimation accuracy in different test conditions.

### 5.10.2 Model Generalization

It may also be natural to ask how well a Sun-BCNN model trained in an urban environment performs in a planetary analogue environment and vice versa. This would provide some indication of whether the model generalizes to new environments or if a philosophy of place-specific excellence (e.g., the place-specific visual features of [McManus et al. \(2014\)](#)) is more appropriate for the task of illumination estimation.

#### Procedure

We attempted to answer this question by creating three larger datasets from combinations of the sequences used in our previous experiments:

1. KITTI odometry sequences 00 - 10;
2. Devon Island sequences 00 - 10; and
3. the previously discussed *Overcast*, *Sun-Cloud A*, and *Sun-Cloud B* sequences from the Oxford Robotcar dataset.

We randomly partitioned each dataset into training (90%) and test (10%) sets. We then trained three separate Sun-BCNN models on each training set, and evaluated each trained model on each of the three test sets.

#### Results

Figure 5.21 shows the results of these experiments with box and whisker plots for azimuth, zenith and vector angle errors while Table 5.6 summarizes the results numerically. We see that none of the three models generalize well to environments other than the one in which they were trained, yielding large and significantly biased test errors. We note, however, that

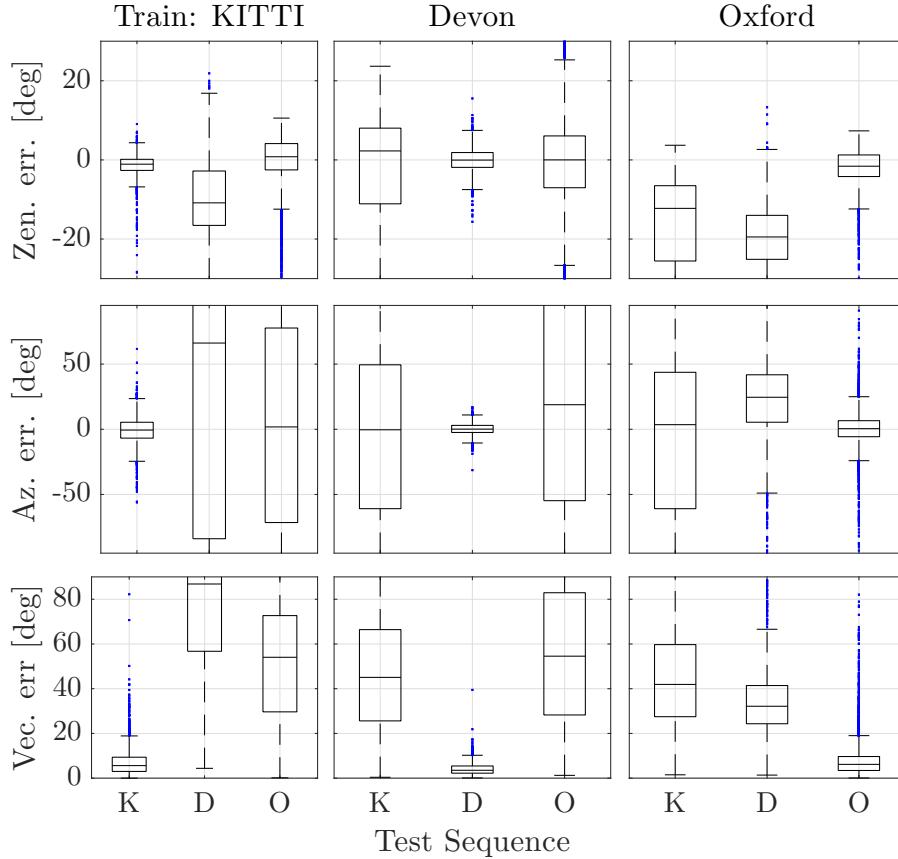


Figure 5.21: Box-and-whiskers plot for zenith, azimuth and vector angle errors for nine different combinations of train-test datasets. Each column corresponds to a different training sequence, and each plot contains three different test sequences. In the bottom legend, we use the labels K: KITTI, D: Devon Island, O: Oxford. All three models produce large biased errors when applied to other datasets, likely due to variations in optical properties and parameter settings across cameras.

the Oxford model was the least egregious offender, and speculate that this may be because the Oxford sequences contain significantly more training images than the other two datasets (approximately 3 times as many as the KITTI odometry benchmark and 5 times as many as the Devon Island dataset).

A possible explanation for the poor generalization of these models is the fact that each dataset was collected using different cameras with different optical properties and parameter settings. We believe these differences affect Sun-BCNN's ability to recover an accurate estimate of a three dimensional direction vector, since metrically important quantities such as the principal point and focal length of the sensor can vary significantly from camera to camera. Furthermore, differences in dynamic range may also significantly affect the ability of Sun-BCNN to treat shading variations consistently.

Table 5.6: Test Errors for Sun-BCNN on three different Oxford Robotcar sequences collected on the same day with different lighting conditions.

<b>Train</b>	<b>Test</b>	<b>Zenith Error [deg]</b>			<b>Azimuth Error [deg]</b>			<b>Vector Error [deg]</b>		
		Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.
Overcast <sup>1</sup>	Overcast	-7.12	-5.20	7.04	-0.66	0.72	29.36	15.22	12.06	11.73
	Sun-Cloud A	-11.58	-9.34	7.94	-5.71	-4.37	37.21	21.19	18.03	14.07
	Sun-Cloud B	-15.23	-12.96	8.00	0.05	-1.49	38.83	23.36	18.49	15.05
Sun-Cloud A <sup>2</sup>	Overcast	-7.17	-5.39	9.05	-0.67	1.68	51.27	23.66	18.03	18.11
	Sun-Cloud A	-6.49	-4.64	7.88	0.29	0.35	27.42	14.31	10.02	12.75
	Sun-Cloud B	-12.89	-10.58	8.94	1.87	3.51	40.41	23.45	19.06	16.75
Sun-Cloud B <sup>3</sup>	Overcast	3.34	5.22	6.46	-0.32	2.24	26.07	13.95	10.63	11.32
	Sun-Cloud A	-0.14	2.30	7.36	-1.08	1.34	28.54	13.76	8.06	14.60
	Sun-Cloud B	-0.84	-0.54	2.07	-0.36	-0.22	9.00	5.11	3.73	5.13

<sup>1</sup> 2014-07-14-14-49-50    <sup>2</sup> 2014-07-14-15-16-36    <sup>3</sup> 2014-07-14-15-42-55

Table 5.7: Test Errors for Sun-BCNN on different training and test datasets.

<b>Train</b>	<b>Test</b>	<b>Zenith Error [deg]</b>			<b>Azimuth Error [deg]</b>			<b>Vector Error [deg]</b>		
		Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.
KITTI	KITTI	-1.49	-1.08	2.99	-0.64	-0.60	11.46	7.16	5.61	6.23
	Devon Island	-9.27	-10.86	9.97	26.78	66.15	113.23	81.32	86.82	33.48
	Oxford	-0.02	0.80	6.59	-0.44	1.81	91.30	52.39	54.05	29.46
Devon Island	KITTI	-2.37	2.27	14.30	-5.58	-0.38	78.01	48.16	45.06	27.85
	Devon Island	-0.08	-0.05	3.20	0.20	0.12	5.52	4.24	3.52	2.96
	Oxford	-1.35	0.00	11.57	17.12	18.85	96.86	55.52	54.55	29.88
Oxford	KITTI	-17.05	-12.25	13.19	-6.94	3.55	77.70	44.66	41.91	23.00
	Devon Island	-20.07	-19.47	9.81	20.92	24.56	45.52	35.16	32.15	16.07
	Oxford	-1.96	-1.59	4.60	0.19	0.48	15.08	8.08	6.16	7.68

Table 5.8: A comparison of prediction errors from different mean estimation methods.

Sequence	Mean Type	Zenith Error [deg]			Azimuth Error [deg]			Vector Error [deg]		
		Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.
KITTI	Method I	-1.50	-1.06	2.96	-0.56	-0.47	11.52	7.16	5.52	6.27
	Method II	-1.06	-0.76	2.44	-0.30	-0.37	30.18	11.49	5.95	18.60
Devon	Method I	-0.07	0.02	3.18	0.19	0.27	5.76	4.22	3.55	3.04
	Method II	0.04	0.09	3.17	1.11	0.26	24.62	9.19	4.05	20.22
Oxford	Method I	-1.97	-1.66	4.59	0.20	0.51	15.31	8.12	6.10	7.74
	Method II	-1.45	-1.27	3.95	-1.58	0.11	34.46	13.18	6.76	19.24

### 5.10.3 Mean and Covariance Computation

In our formulation, Sun-BCNN outputs a sampling of unit-norm 3D vectors. Due to the unit-norm constraint, it is not immediately clear how to apply Equations (5.11) and (5.12) to calculate the mean and covariance of these samples. In this section we present and empirically evaluate two possible procedures for each computation using the previously discussed combined datasets for KITTI, Devon Island, and Oxford.

#### Mean computation

**Procedure** We investigated two different methods for computing the mean of the sampled sun vectors, which we refer to as *Method I* and *Method II*.

1. In *Method I* (used in this work), we first evaluate Equation (5.11) directly on the constrained unit vectors produced by  $N$  stochastic passes through the BCNN. We then renormalize the resulting mean vector to enforce unit length, and convert it to azimuth and zenith angles using Equation (5.7).
2. In *Method II*, we first convert each of the  $N$  unit vectors produced through stochastic passes through the BCNN to azimuth and zenith angles using Equation (5.7). We then evaluate Equation (5.11) on the angles themselves to obtain the mean in azimuth-zenith coordinates.

We evaluated both methods using the same combined datasets and partitioning scheme as in the transfer learning experiment previously presented.

**Results** Table 5.8 presents the azimuth, zenith and vector errors for the two mean computation methods. *Method I* produces lower vector errors and smaller standard deviations in azimuth and zenith on all three datasets.

Table 5.9: A comparison of ANEES values for different mean and covariance propagation methods.

<b>Sequence</b>	<b>Covariance Type</b>	<b>Mean Type</b>	<b>ANEES</b>
KITTI	Method I	Method I	0.95
		Method II	5.10
	Method II	Method I	1.40
		Method II	0.87
Devon	Method I	Method I	1.29
		Method II	10.05
	Method II	Method I	0.50
		Method II	0.85
Oxford	Method I	Method I	1.50
		Method II	2.14
	Method II	Method I	1.30
		Method II	0.89

## Covariance Computation

**Procedure** We further investigated two different covariance computation methods, which we also refer to as *Method I* and *Method II*.

1. In *Method I*, we first evaluate Equation (5.12) directly on the constrained unit vectors produced by  $N$  stochastic passes through the BCNN, yielding a  $3 \times 3$  covariance. We then compute a  $2 \times 2$  covariance on azimuth and zenith by propagating the  $3 \times 3$  covariance through a linearized Equation (5.7).
2. In *Method II* (used in this work), we first convert each of the  $N$  unit vectors produced by stochastic passes through the BCNN to azimuth and zenith angles, and then evaluate Equation (5.12) on the angles themselves.

We once again re-used the transfer learning datasets with the same partitioning scheme, and evaluated covariances on the test sets corresponding to each of the three models. To control for the effect of tuning the model precision  $\tau$ , we replace the diagonal elements of each covariance matrix with the diagonal elements of the empirical covariance corresponding to the entire test set (computed based ground truth azimuth and zenith errors). We then compared the consistency of the cross-correlations of each method (i.e., the off-diagonal components of the covariance matrix) by computing ANEES values over the each model’s corresponding test set using both mean computation methods.

**Results** Table 5.9 lists the ANEES values produced by each method of covariance computation when paired with each mean computation method. *Method I* covariances produced better ANEES values when paired with *Method I* mean estimation, but *Method II* covariances paired well with either mean estimation scheme.

## 5.11 Summary

In summary, with Sun-BCNN, we applied learned *pseudo-sensors* to the problem of illumination direction in outdoor environments. Sun-BCNN presented

1. the application of a Bayesian CNN to the problem of sun direction estimation, incorporating the resulting covariance estimates into a visual odometry pipeline;
2. an empirical demonstration that a Bayesian CNN with dropout layers after each convolutional and fully-connected layer can achieve state-of-the-art accuracy at test time;
3. a loss function that incorporated a 3D unit-length sun direction vector, appropriate for full 6-DOF pose estimation;
4. experimental results on over 30 km of visual navigation data in urban (Geiger et al., 2013) and planetary analogue (Furgale et al., 2012) environments;
5. an investigation into the sensitivity of the Bayesian CNN-based sun estimate to cloud cover, camera and environment changes, and measurement parameterization; and
6. open-source software<sup>4</sup>.

---

<sup>4</sup><https://github.com/utiasSTARS/sun-bcnn-vo>.

# Chapter 6

## Learned Pose Corrections

Life can only be understood backwards;  
but it must be lived forwards.



---

Søren Kierkegaard

### 6.1 Introduction

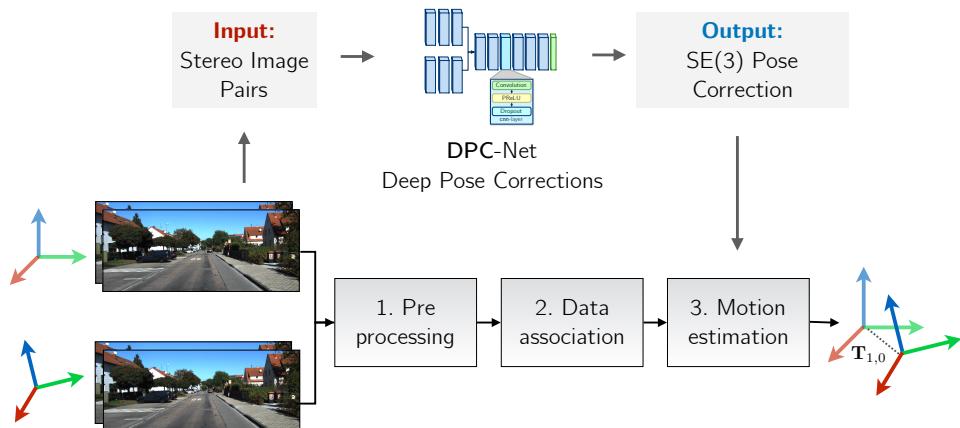


Figure 6.1: DPC-Net (Deep Pose Corrections) is a pseudo-sensor that learns to correct a ‘classical’ VO pipeline.

If a pseudo-sensor like Sun-BCNN can extract useful orientation information directly from images, can we create a more general approach that can more directly learn SE(3) corrections to an existing VO estimate? In this chapter, we explore this by training what we call a *deep pose correction* (DPC) network to correct a classical pipeline.

DPC is associated with one publication:

- Peretroukhin, V. and Kelly, J. (2018). DPC-Net: Deep pose correction for visual localization. *IEEE Robotics and Automation Letters*, 3(3):2424–2431.

## 6.2 Motivation

Deep convolutional neural networks (CNNs) are at the core of many state-of-the-art classification and segmentation algorithms in computer vision (LeCun et al., 2015). These CNN-based techniques achieve accuracies previously unattainable by classical methods. In mobile robotics and state estimation, however, it remains unclear to what extent these deep architectures can obviate the need for classical geometric modelling. In this chapter, we focus on improving visual odometry (VO): the task of computing the egomotion of a camera through an unknown environment with no external positioning sources. Visual localization algorithms like VO can suffer from several systematic error sources that include estimator biases (Peretroukhin et al., 2014), poor calibration, and environmental factors (e.g., a lack of scene texture). While machine learning approaches can be used to better model specific subsystems of a localization pipeline (e.g., the heteroscedastic feature track error covariance modelling of Peretroukhin et al. (2016, 2015a)), much recent work (Costante et al., 2016; Clark et al., 2017; Kendall et al., 2015; Melekhov et al., 2017; Oliveira et al., 2017) has been devoted to completely replacing the estimator with a CNN-based system.

We contend that this type of complete replacement places an unnecessary burden on the CNN. Not only must it learn projective geometry, but it must also understand the environment, and account for sensor calibration and noise. Instead, we take inspiration from recent results (Peretroukhin et al., 2017) that demonstrate that CNNs can infer difficult-to-model geometric quantities (e.g., the direction of the sun) to improve an existing localization estimate. In a similar vein, we propose a system (Figure 6.2) that takes as its starting point an efficient, classical localization algorithm that computes high-rate pose estimates. To it, we add a Deep Pose Correction Network (DPC-Net) that learns low-rate, ‘small’ corrections from training data that we then fuse with the original estimates. DPC-Net does not require any

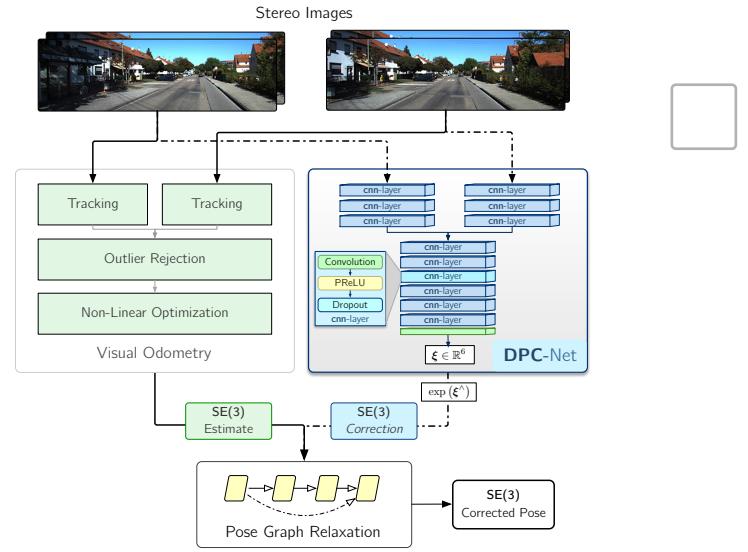


Figure 6.2: We propose a Deep Pose Correction network (DPC-Net) that learns  $SE(3)$  corrections to classical visual localizers.

modification to an existing localization pipeline, and can learn to correct multi-faceted errors from estimator bias, sensor mis-calibration or environmental effects.

Although in this work we focus on visual data, the DPC-Net architecture can be readily modified to learn SE(3) corrections for estimators that operate with other sensor modalities (e.g., lidar). For this general task, we derive a pose regression loss function and a closed-form analytic expression for its Jacobian. Our loss permits a network to learn an unconstrained Lie algebra coordinate vector, but derives its Jacobian with respect to SE(3) geodesic distance.

## 6.3 Related Work

Visual state estimation has a rich history in mobile robotics. We direct the reader to ([Scaramuzza and Fraundorfer, 2011](#)) and ([Cadena et al., 2016](#)) for detailed surveys of what we call *classical*, geometric approaches to visual odometry (VO) and visual simultaneous localization and mapping (SLAM).

In the past decade, much of machine learning and its sub-disciplines has been revolutionized by carefully constructed deep neural networks (DNNs) ([LeCun et al., 2015](#)). For tasks like image segmentation, classification, and natural language processing, most prior state-of-the-art algorithms have been replaced by their DNN alternatives.

In mobile robotics, deep neural networks have ushered in a new paradigm of end-to-end training of visuomotor policies ([Levine et al., 2016](#)) and significantly impacted the related field of reinforcement learning ([Duan et al., 2016](#)). In state estimation, however, most successful applications of deep networks have aimed at replacing a specific sub-system of a localization and mapping pipeline (for example, object detection ([Yang et al., 2016](#)), place recognition ([Sunderhauf et al., 2015](#)), or bespoke discriminative observation functions ([Haarnoja et al., 2016](#))).

Nevertheless, a number of recent approaches have presented convolutional neural network architectures that purport to obviate the need for classical visual localization algorithms. For example, Kendall et al. ([Kendall et al., 2015; Kendall and Cipolla, 2017](#)) presented extensive work on PoseNet, a CNN-based camera re-localization approach that regresses the 6-DOF pose of a camera within a previously explored environment. Building upon PoseNet, Melekhov ([Melekhov et al., 2017](#)) applied a similar CNN learning paradigm to relative camera motion. In related work, Costante et al. ([Costante et al., 2016](#)) presented a CNN-based VO technique that uses pre-processed dense optical flow images. By focusing on RGB-D odometry, Handa et al. [Handa et al. \(2016\)](#), detailed an approach to learning relative poses with *3D Spatial Transformer* modules and a dense photometric loss. Finally, Oliveira et al. ([Oliveira et al., 2017](#)) and Clark et al. ([Clark et al., 2017](#)) described techniques for more general sensor fusion and mapping. The former work outlined a DNN-based topometric localization

pipeline with separate VO and place recognition modules while the latter presented VINet, a CNN paired with a recurrent neural network for visual-inertial sensor fusion and online calibration.

With this recent surge of work in end-to-end learning for visual localization, one may be tempted to think this is the only way forward. It is important to note, however, that these deep CNN-based approaches do not yet report state-of-the-art localization accuracy, focusing instead on proof-of-concept validation. Indeed, at the time of writing, the most accurate visual odometry approach on the KITTI odometry benchmark leaderboard<sup>1</sup> remains a variant of a sparse feature-based pipeline with carefully hand-tuned optimization (Cvišić and Petrović, 2015).

Taking inspiration from Sun-BCNN (Chapter 5) that shows that CNNs can be used to inject global orientation information into a visual localization pipeline, and leveraging ideas from recent approaches to trajectory tracking in the field of controls (Li et al., 2017a; Punjani and Abbeel, 2015), we formulate a system that learns *pose corrections* to an existing estimator, instead of learning the entire localization problem *ab initio*.

## 6.4 System Overview: Deep Pose Correction

We base our network structure on that of (Handa et al., 2016) but learn SE(3) *corrections* from stereo images and require no pre-training. Similar to (Handa et al., 2016), we use primarily 3x3 convolutions and avoid the use of max pooling to preserve spatial information. We achieve downsampling by setting the stride to 2 where appropriate (see Figure 6.3 for a full description of the network structure). We derive a novel loss function, unlike that used in Kendall and Cipolla (2017); Melekhov et al. (2017); Oliveira et al. (2017), based on SE(3) geodesic distance. Our loss naturally balances translation and rotation error without requiring a hand-tuned scalar hyper-parameter. Similar to Costante et al. (2016), we test our final system on the KITTI odometry benchmark, and evaluate how it copes with degraded visual data.

Given two coordinate frames  $\mathcal{F}_i$ ,  $\mathcal{F}_{i+\Delta_p}$  that represent a camera's pose at time  $t_i$  and  $t_{i+\Delta_p}$  (where  $\Delta_p$  is an integer hyper-parameter that allows DPC-Net to learn corrections across multiple temporally consecutive poses), we assume that our visual localizer gives us an estimate<sup>2</sup>,  $\hat{\mathbf{T}}_{i,i+\Delta_p}$ , of the true transform  $\mathbf{T}_{i,i+\Delta_p} \in \text{SE}(3)$  between the two frames. We

---

<sup>1</sup>See [http://www.cvlibs.net/datasets/kitti/eval\\_odometry.php](http://www.cvlibs.net/datasets/kitti/eval_odometry.php).

<sup>2</sup>We use  $(\hat{\cdot})$  to denote an estimated quantity throughout the paper.

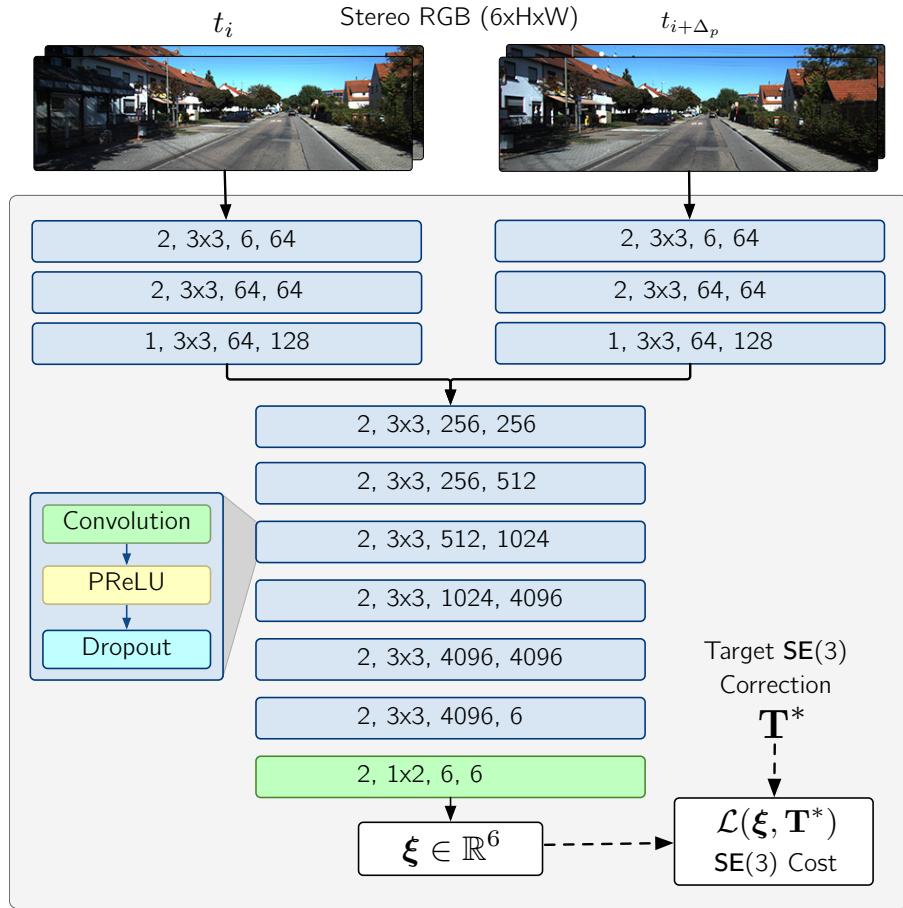


Figure 6.3: The Deep Pose Correction network with stereo RGB image data. The network learns a map from two stereo pairs to a vector of Lie algebra coordinates. Each darker blue block consists of a convolution, a PReLU non-linearity, and a dropout layer. We opt to not use MaxPooling in the network, following [Handa et al. \(2016\)](#). The labels correspond to the stride, kernel size, and number of input and output channels of for each convolution.

aim to learn a target correction,

$$\mathbf{T}_i^* = \mathbf{T}_{i,i+\Delta_p} \hat{\mathbf{T}}_{i,i+\Delta_p}^{-1}, \quad (6.1)$$

from two pairs of stereo images (collectively referred to as  $\mathbf{I}_{t_i,t_{i+\Delta_p}}$ ) captured at  $t_i$  and  $t_{i+\Delta_p}$ <sup>3</sup>. Given a dataset,  $\{\mathbf{T}_i^*, \mathbf{I}_{t_i,t_{i+\Delta_p}}\}_{i=1}^N$ , we now turn to the problem of selecting an appropriate loss function for learning SE(3) corrections.

#### 6.4.1 Loss Function: Correcting SE(3) Estimates

One possible approach to learning  $\mathbf{T}^*$  is to break it into constituent parts and then compose a loss function as the weighted sum of translational and rotational error norms (as done in [Melekhov et al. \(2017\)](#); [Oliveira et al. \(2017\)](#)). This however does not account for the possible correlation between the two losses, and requires the careful tuning of a scalar weight.

In this work, we instead choose to parametrize our correction prediction as  $\mathbf{T} = \exp(\boldsymbol{\xi}^\wedge)$ , where  $\boldsymbol{\xi} \in \mathbb{R}^6$ , a vector of Lie algebra coordinates, is the output of our network (similar to [Handa et al. \(2016\)](#)). We define a loss for  $\boldsymbol{\xi}$  as

$$\mathcal{L}(\boldsymbol{\xi}) = \frac{1}{2} g(\boldsymbol{\xi})^T \Sigma^{-1} g(\boldsymbol{\xi}), \quad (6.2)$$

where

$$g(\boldsymbol{\xi}) \triangleq \log \left( \exp(\boldsymbol{\xi}^\wedge) \mathbf{T}^{*-1} \right)^\vee. \quad (6.3)$$

Here,  $\Sigma$  is the covariance of our estimator (expressed using unconstrained Lie algebra coordinates), and  $(\cdot)^\wedge, (\cdot)^\vee$  convert vectors of Lie algebra coordinates to matrix vectorspace quantities and back, respectively. Given two stereo image pairs, we use the output of DPC-Net,  $\boldsymbol{\xi}$ , to correct our estimator as follows:

$$\hat{\mathbf{T}}^{\text{corr}} = \exp(\boldsymbol{\xi}^\wedge) \hat{\mathbf{T}}, \quad (6.4)$$

where we have dropped subscripts for clarity.

#### 6.4.2 Loss Function: SE(3) Covariance

Since we are learning estimator *corrections*, we can compute an empirical covariance over the training set as

---

<sup>3</sup>Note that the visual estimator does not necessarily compute  $\hat{\mathbf{T}}_{i,i+\Delta_p}$  directly.  $\hat{\mathbf{T}}_{i,i+\Delta_p}$  may be compounded from several estimates.

$$\Sigma = \frac{1}{N-1} \sum_{i=1}^N \left( \xi_i^* - \bar{\xi}^* \right) \left( \xi_i^* - \bar{\xi}^* \right)^T, \quad (6.5)$$

where

$$\xi_i^* \triangleq \log(\mathbf{T}_i^*)^\vee, \quad \bar{\xi}^* \triangleq \frac{1}{N} \sum_{i=1}^N \xi_i^*. \quad (6.6)$$

The term  $\Sigma$  balances the rotational and translation loss terms based on their magnitudes in the training set, and accounts for potential correlations. We stress that if we were learning poses directly, the pose targets and their associated mean would be trajectory dependent and would render this type of covariance estimation meaningless. Further, we find that, in our experiments,  $\Sigma$  weights translational and rotational errors similarly to that presented in ? based on the diagonal components, but contains relatively large off-diagonal terms.

### 6.4.3 Loss Function: SE(3) Jacobians

In order to use Equation (6.2) to train DPC-Net with back-propagation, we need to compute its Jacobian with respect to our network output,  $\xi$ . Applying the chain rule, we begin with the expression

$$\frac{\partial \mathcal{L}(\xi)}{\partial \xi} = g(\xi)^T \Sigma^{-1} \frac{\partial g(\xi)}{\partial \xi}. \quad (6.7)$$

The term  $\frac{\partial g(\xi)}{\partial \xi}$  is of importance. We can derive it in two ways. To start, note two important identities Barfoot (2017). First,

$$\exp((\xi + \delta\xi)^\wedge) \approx \exp((\mathcal{J}\delta\xi)^\wedge) \exp(\xi^\wedge), \quad (6.8)$$

where  $\mathcal{J} \triangleq \mathcal{J}(\xi)$  is the left SE(3) Jacobian. Second, if  $\mathbf{T}_1 \triangleq \exp(\xi_1^\wedge)$  and  $\mathbf{T}_2 \triangleq \exp(\xi_2^\wedge)$ , then

$$\begin{aligned} \log(\mathbf{T}_1 \mathbf{T}_2)^\vee &= \log(\exp(\xi_1^\wedge) \exp(\xi_2^\wedge))^\vee \\ &\approx \begin{cases} \mathcal{J}(\xi_2)^{-1} \xi_1 + \xi_2 & \text{if } \xi_1 \text{ small} \\ \xi_1 + \mathcal{J}(-\xi_1)^{-1} \xi_2 & \text{if } \xi_2 \text{ small.} \end{cases} \end{aligned} \quad (6.9)$$

See Barfoot (2017) for a detailed treatment of matrix Lie groups and their use in state estimation.

### Deriving $\frac{\partial g(\xi)}{\partial \xi}$ , Method I

If we assume that only  $\xi$  is ‘small’, we can apply Equation (6.9) directly to define

$$\frac{\partial g(\xi)}{\partial \xi} = \mathcal{J}(-\xi^*)^{-1}, \quad (6.10)$$

with  $\xi^* \triangleq \log(\mathbf{T}^*)^\vee$ . Although attractively compact, note that this expression for  $\frac{\partial g(\xi)}{\partial \xi}$  assumes that  $\xi$  is small, and may be inaccurate for ‘larger’  $\mathbf{T}^*$  (since we will therefore require  $\xi$  to be commensurately ‘large’).

### Deriving $\frac{\partial g(\xi)}{\partial \xi}$ , Method II

Alternatively, we can linearize Equation (6.3) about  $\xi$ , by considering a small change  $\delta\xi$  and applying Equation (6.8):

$$g(\xi + \delta\xi) = \log \left( \exp \left( (\xi + \delta\xi)^\wedge \right) \mathbf{T}^{*-1} \right)^\vee \quad (6.11)$$

$$\approx \log \left( \exp \left( (\mathcal{J}\delta\xi)^\wedge \right) \exp \left( \xi^\wedge \right) \mathbf{T}^{*-1} \right)^\vee. \quad (6.12)$$

Now, assuming that  $\mathcal{J}\delta\xi$  is ‘small’, and using Equation (6.3), Equation (6.9) gives:

$$g(\xi + \delta\xi) \approx \mathcal{J}(g(\xi))^{-1} \mathcal{J}(\xi) \delta\xi + g(\xi). \quad (6.13)$$

Comparing this to the first order Taylor expansion:  $g(\xi + \delta\xi) \approx g(\xi) + \frac{\partial g(\xi)}{\partial \xi} \delta\xi$ , we see that

$$\frac{\partial g(\xi)}{\partial \xi} = \mathcal{J}(g(\xi))^{-1} \mathcal{J}(\xi). \quad (6.14)$$

Although slightly more computationally expensive, this expression makes no assumptions about the ‘magnitude’ of our correction and works reliably for any target. Note further that if  $\xi$  is small, then  $\mathcal{J}(\xi) \approx 1$  and  $\exp(\xi^\wedge) \approx 1$ . Thus,

$$g(\xi) \approx \log(\mathbf{T}^{*-1})^\vee = -\xi^*, \quad (6.15)$$

and Equation (6.14) becomes

$$\frac{\partial g(\xi)}{\partial \xi} = \mathcal{J}(-\xi^*)^{-1}, \quad (6.16)$$

which matches *Method I*. To summarize, to apply back-propagation to Equation (6.2), we use Equation (6.7) and Equation (6.14).

#### 6.4.4 Loss Function: Correcting SO(3) Estimates

Our framework can be easily modified to learn SO(3) corrections only. We can parametrize a similar objective for  $\phi \in \mathbb{R}^3$ ,

$$\mathcal{L}(\phi, \mathbf{C}_*) = \frac{1}{2} f(\phi)^T \Sigma^{-1} f(\phi), \quad (6.17)$$

where

$$f(\phi) \triangleq \log (\exp (\phi^\wedge) \mathbf{C}_*^{-1})^\vee. \quad (6.18)$$

Equations (6.8) and (6.9) have analogous SO(3) formulations:

$$\exp ((\phi + \delta\phi)^\wedge) \approx \exp ((\mathbf{J}\delta\phi)^\wedge) \exp (\phi^\wedge), \quad (6.19)$$

and

$$\begin{aligned} \log (\mathbf{C}_1 \mathbf{C}_2)^\vee &= \log (\exp (\phi_1^\wedge) \exp (\phi_2^\wedge))^\vee \\ &\approx \begin{cases} \mathbf{J}(\phi_2)^{-1} \phi_1 + \phi_2 & \text{if } \phi_1 \text{ small} \\ \phi_1 + \mathbf{J}(-\phi_1)^{-1} \phi_2 & \text{if } \phi_2 \text{ small,} \end{cases} \end{aligned} \quad (6.20)$$

where  $\mathbf{J} \triangleq \mathbf{J}(\phi)$  is the left SO(3) Jacobian. Accordingly, the final loss Jacobians are identical in structure to Equation (6.7) and Equation (6.14), with the necessary SO(3) replacements.

#### 6.4.5 Pose Graph Relaxation

In practice, we find that using camera poses several frames apart (i.e.  $\Delta_p > 1$ ) often improves test accuracy and reduces overfitting. As a result, we turn to pose graph relaxation to fuse low-rate corrections with higher-rate visual pose estimates. For a particular window of  $\Delta_p + 1$  poses (see Figure 6.4), we solve the non linear minimization problem

$$\{\mathbf{T}_{t_i, n}\}_{i=0}^{\Delta_p} = \underset{\{\mathbf{T}_{t_i, n}\}_{i=0}^{\Delta_p} \in \text{SE}(3)}{\operatorname{argmin}} \mathcal{O}_t, \quad (6.21)$$

where  $n$  refers to a common navigation frame, and where we define the total cost,  $\mathcal{O}_t$ , as a sum of visual estimation and correction components:

$$\mathcal{O}_t = \mathcal{O}_v + \mathcal{O}_c. \quad (6.22)$$

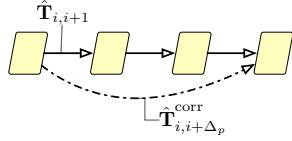


Figure 6.4: We apply pose graph relaxation to fuse high-rate visual localization estimates ( $\hat{\mathbf{T}}_{i,i+1}$ ) with low-rate deep pose corrections ( $\hat{\mathbf{T}}_{i,i+\Delta_p}^{corr}$ ).

The former cost sums over each estimated transform,

$$\mathcal{O}_v \triangleq \sum_{i=0}^{\Delta_p-1} \mathbf{e}_{t_i, t_{i+1}}^T \Sigma_v^{-1} \mathbf{e}_{t_i, t_{i+1}}, \quad (6.23)$$

while the latter incorporates a single pose correction,

$$\mathcal{O}_c \triangleq \mathbf{e}_{t_0, t_{\Delta_p}}^T \Sigma_c^{-1} \mathbf{e}_{t_0, t_{\Delta_p}}, \quad (6.24)$$

with the pose error defined as

$$\mathbf{e}_{1,2} = \log \left( \hat{\mathbf{T}}_{1,2} \mathbf{T}_2 \mathbf{T}_1^{-1} \right)^\vee. \quad (6.25)$$

We refer the reader to [Barfoot \(2017\)](#) for a detailed treatment of pose-graph relaxation.

## 6.5 Experiments

To assess the power of our proposed deep corrective paradigm, we trained DPC-Net on visual data with localization estimates from a sparse stereo visual odometry (S-VO) estimator. In addition to training the full  $SE(3)$  DPC-Net, we modified the loss function and input data to learn simpler  $SO(3)$  rotation corrections, and simpler still, yaw angle corrections. For reference, we compared S-VO with different DPC-Net corrections to a state-of-the-art dense estimator. Finally, we trained DPC-Net on visual data and localization estimates from radially-distorted and cropped images.

### 6.5.1 Training & Testing

For all experiments, we used the KITTI odometry benchmark training set [Geiger et al. \(2013\)](#). Specifically, our data consisted of the eight sequences 00,02 and 05-10 (we removed sequences 01, 03, 04 to ensure that all data originated from the ‘residential’ category for training and test consistency).

For testing and validation, we selected the first three sequences (00,02, and 05). For each

test sequence, we selected the subsequent sequence for validation (i.e., for test sequence 00 we validated with 02, for 02 with 05, etc.) and used the remaining sequences for training. We note that by design, we train DPC-Net to predict corrections for a specific sensor and estimator pair. A pre-trained DPC-Net may further serve as a useful starting point to fine-tune new models for other sensors and estimators. In this work, however, we focus on the aforementioned KITTI sequences and leave a thorough investigation of generalization for future work.

To collect training samples,  $\{\mathbf{T}_i^*, \mathbf{I}_{t_i, t_i + \Delta_p}\}_{i=0}^N$ , we used a stereo visual odometry estimator and GPS-INS ground-truth from the KITTI odometry dataset<sup>4</sup>. We resized all images to [400, 120] pixels, approximately preserving their original aspect ratio<sup>5</sup>. For non-distorted data, we use  $\Delta_p \in [3, 4, 5]$  for training, and test with  $\Delta_p = 4$ . For distorted data, we reduce this to  $\Delta_p \in [2, 3, 4]$  and  $\Delta_p = 3$ , respectively, to compensate for the larger estimation errors.

Our training datasets contained between 35,000 and 52,000 training samples<sup>6</sup> depending on test sequence. We trained all models for 30 epochs using the Adam optimizer, and selected the best epoch based on the lowest validation loss.

## Rotation

To train rotation-only corrections, we extracted the  $\text{SO}(3)$  component of  $\mathbf{T}_i^*$  and trained our network using Equation (6.17). Further, owing to the fact that rotation information can be extracted from monocular images, we replaced the input stereo pairs in DPC-Net with monocular images from the left camera<sup>7</sup>.

## Yaw

To further simplify the corrections, we extracted a single-degree-of-freedom yaw rotation correction angle<sup>8</sup> from  $\mathbf{T}_i^*$ , and trained DPC-Net with monocular images and a mean squared loss.

<sup>4</sup>We used RGB stereo images to train DPC-Net but grayscale images for the estimator.

<sup>5</sup>Because our network is fully convolutional, it can, in principle, operate on different image resolutions with no modifications, though we do not investigate this ability in this work.

<sup>6</sup>If a sequence has  $M$  poses, we collect  $M - \Delta_p$  training samples for each  $\Delta_p$ .

<sup>7</sup>The *stereo* VO estimator remained unchanged.

<sup>8</sup>We define yaw in the camera frame as the rotation about the camera's vertical  $y$  axis.



Figure 6.5: Illustration of our image radial distortion procedure. Left: rectified RGB image (frame 280 from KITTI odometry sequence 05). Middle: the same image with radial distortion applied. Right: distorted, cropped, and scaled image.

### 6.5.2 Estimators

#### Sparse Visual Odometry

Our baseline estimator is based on the reprojection-error based maximum likelihood approach detailed in Chapter 3. As before, we rely on the open-source `viso2` package [Geiger et al. \(2011\)](#), to detect and track sparse stereo image key-points,  $\mathbf{y}_{i,c_1}$  and  $\mathbf{y}_{i,c_0}$ , between stereo image pairs (assumed to be undistorted and rectified). We model reprojection errors (due to sensor noise and quantization) as zero-mean Gaussians with a known covariance,  $\mathbf{R}$ :

$$\mathbf{e}_{i,t} = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})) \quad (6.26)$$

$$\sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad (6.27)$$

where  $\mathbf{f}(\cdot)$  is the stereo camera projection function. To generate an initial guess and to reject outliers, we use three point Random Sample Consensus (RANSAC) based on stereo reprojection error. Finally, we solve for the maximum likelihood transform,  $\mathbf{T}_t^*$ , through a Gauss-Newton minimization:

$$\mathbf{T}_t^* = \underset{\mathbf{T}_t \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_{i,t}^T \mathbf{R}^{-1} \mathbf{e}_{i,t}. \quad (6.28)$$

#### Sparse Visual Odometry with Radial Distortion

Similar to [Costante et al. \(2016\)](#), we modified our input images to test our network's ability to correct estimators that compute poses from degraded visual data. Unlike [Costante et al. \(2016\)](#), who darken and blur their images, we chose to simulate a poorly calibrated lens model by applying radial distortion to the (rectified) KITTI dataset using a plumb-bob distortion model. The model computes radially-distorted image coordinates,  $x_d, y_d$ , from the normalized coordinates  $x_n, y_n$  as

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = (1 + \kappa_1 r^2 + \kappa_2 r^4 + \kappa_3 r^6) \begin{bmatrix} x_n \\ y_n \end{bmatrix}, \quad (6.29)$$

where  $r = \sqrt{x_n^2 + y_n^2}$ . We set the distortion coefficients,  $\kappa_1$ ,  $\kappa_2$ , and  $\kappa_3$  to  $-0.3, 0.2, 0.01$  respectively, to approximately match the KITTI radial distortion parameters. We solved Equation (6.29) iteratively and used bilinear interpolation to compute the distorted images for every stereo pair in a sequence. Finally, we cropped each distorted image to remove any whitespace. Figure 6.5 illustrates this process.

With this distorted dataset, we computed S-VO localization estimates and then trained DPC-Net to correct for the effects of the radial distortion and effective intrinsic parameter shift due to the cropping process.

### Dense Visual Odometry

Finally, we present localization estimates from a computationally-intensive keyframe-based dense, direct visual localization pipeline (Clement and Kelly, 2018) that computes relative camera poses by minimizing photometric error with respect to a keyframe image. To compute the photometric error, the pipeline relies on an inverse compositional approach to map the image coordinates of a tracking image to the image coordinates of the reference depth image. As the camera moves through an environment, a new keyframe depth image is computed and stored when the camera field-of-view differs sufficiently from the last keyframe.

We used this dense, direct estimator as our benchmark for a state-of-the-art visual localizer, and compared its accuracy to that of a much less computationally expensive sparse estimator paired with DPC-Net.

#### 6.5.3 Evaluation Metrics

To evaluate the performance of DPC-Net, we use three error metrics: mean absolute trajectory error, cumulative absolute trajectory error, and mean segment error. For clarity, we describe each of these three metrics explicitly and stress the importance of carefully selecting and defining error metrics when comparing *relative* localization estimates, as results can be subtly deceiving.

##### Mean Absolute Trajectory Error (m-ATE)

The mean absolute trajectory error averages the magnitude of the rotational or translational error<sup>9</sup> of estimated poses with respect to a ground truth trajectory defined within the same

---

<sup>9</sup>For brevity, the notation  $\log(\cdot)^\vee$  returns rotational or translational components depending on context.

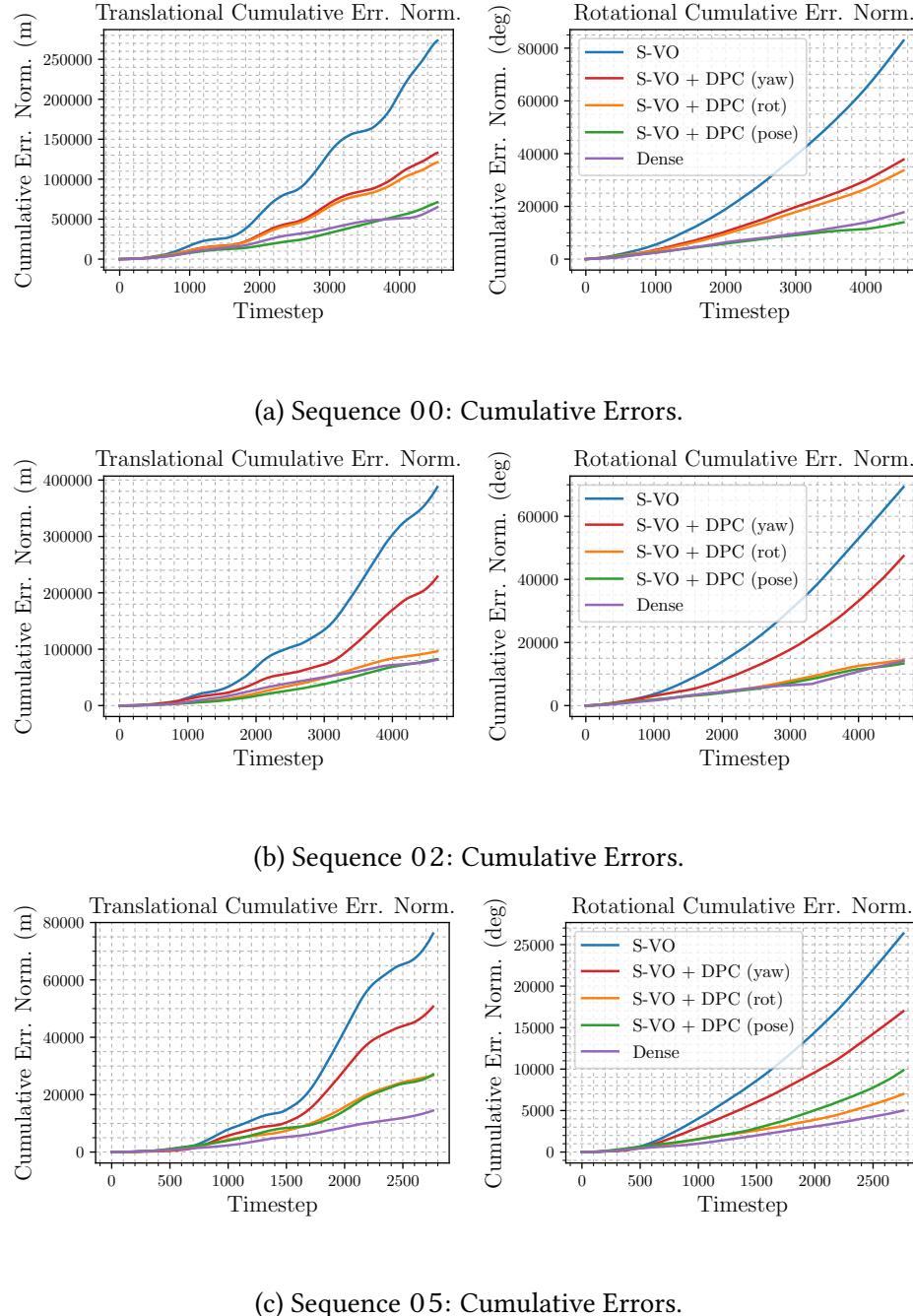


Figure 6.6: c-ATE for S-VO with and without DPC-Net.

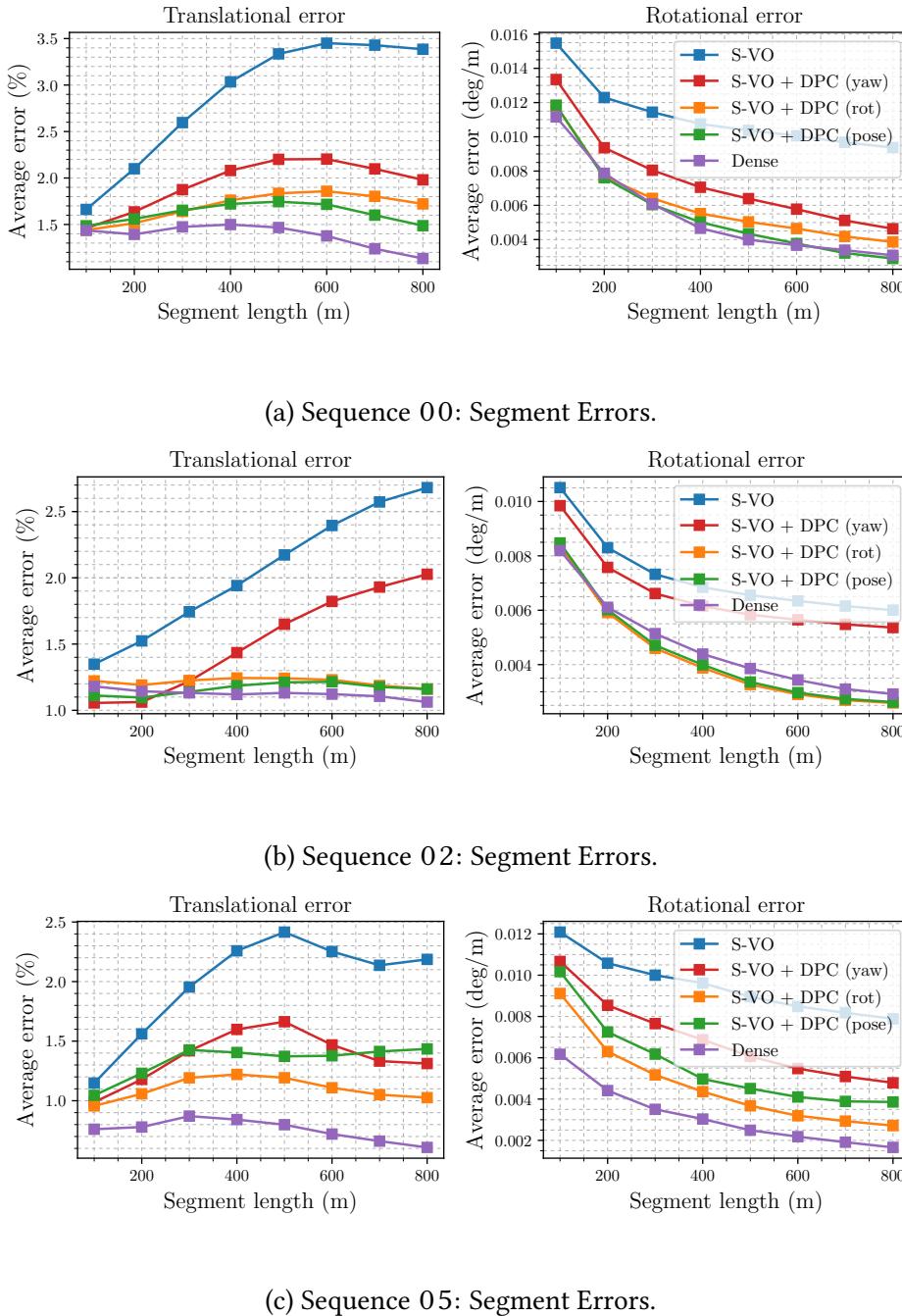


Figure 6.7: Mean segment errors for S-VO with and without DPC-Net.

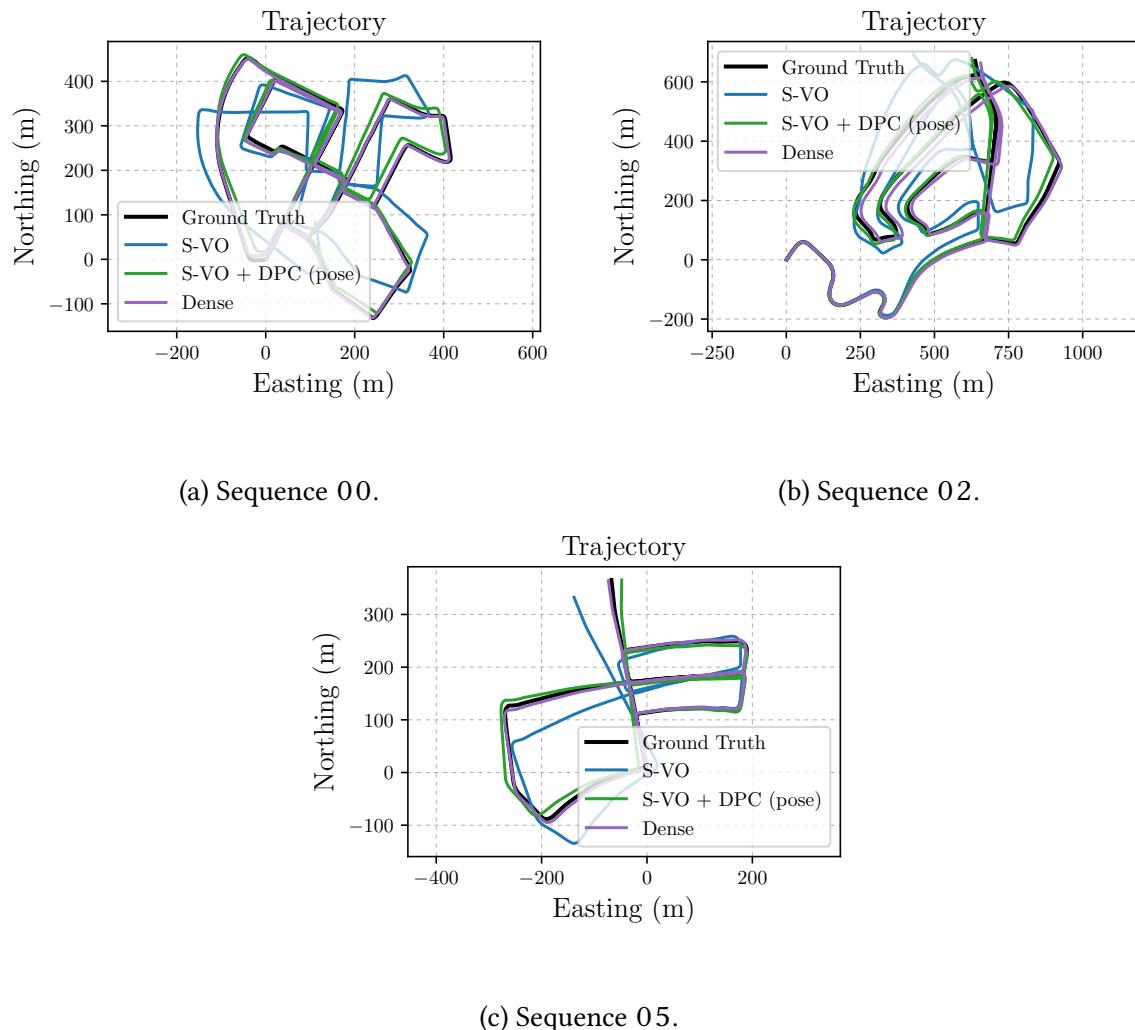


Figure 6.8: Top down projections for S-VO with and without DPC-Net.

navigation frame. Concretely,  $\mathbf{e}_{\text{m-ATE}}$  is defined as

$$\mathbf{e}_{\text{m-ATE}} \triangleq \frac{1}{N} \sum_{p=1}^N \left\| \log \left( \hat{\mathbf{T}}_{p,0}^{-1} \mathbf{T}_{p,0} \right)^\vee \right\|. \quad (6.30)$$

Although widely used, m-ATE can be deceiving because a single poor relative transform can significantly affect the final statistic.

### Cumulative Absolute Trajectory Error (c-ATE)

Cumulative absolute trajectory error sums rotational or translational  $\mathbf{e}_{\text{m-ATE}}$  up to a given point in a trajectory. It is defined as

$$\mathbf{e}_{\text{c-ATE}}(q) \triangleq \sum_{p=1}^q \left\| \log \left( \hat{\mathbf{T}}_{p,0}^{-1} \mathbf{T}_{p,0} \right)^\vee \right\|. \quad (6.31)$$

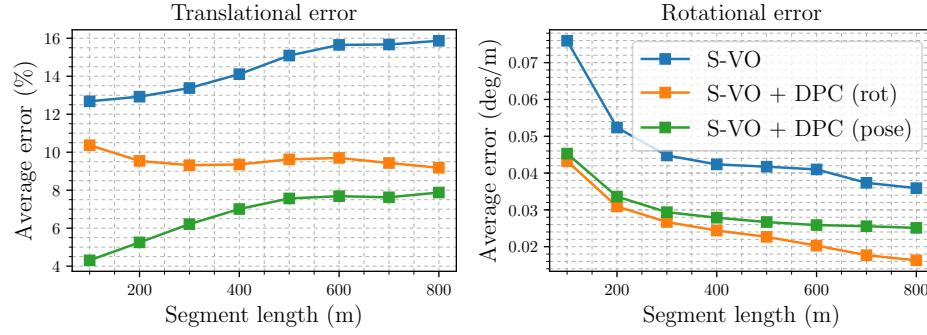
c-ATE can show clearer trends than m-ATE (because it is less affected by fortunate trajectory overlaps), but it still suffers from the same susceptibility to poor (but isolated) relative transforms.

### Segment Error

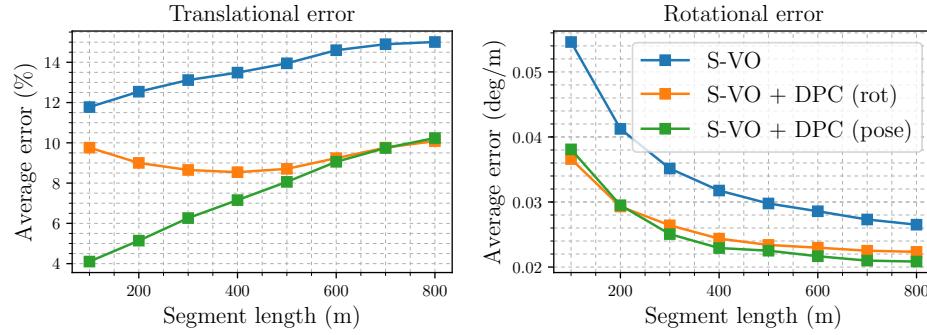
Our final metric, segment error, averages the end-point error for all the possible segments of a given length within a trajectory, and then normalizes by the segment length. Since it considers multiple starting points within a trajectory, segment error is much less sensitive to isolated degradations. Concretely,  $\mathbf{e}_{\text{seg}}(s)$  is defined as

$$\mathbf{e}_{\text{seg}}(s) \triangleq \frac{1}{s N_s} \sum_{p=1}^{N_s} \left\| \log \left( \hat{\mathbf{T}}_{p+s_p,p}^{-1} \mathbf{T}_{p+s_p,p} \right)^\vee \right\|, \quad (6.32)$$

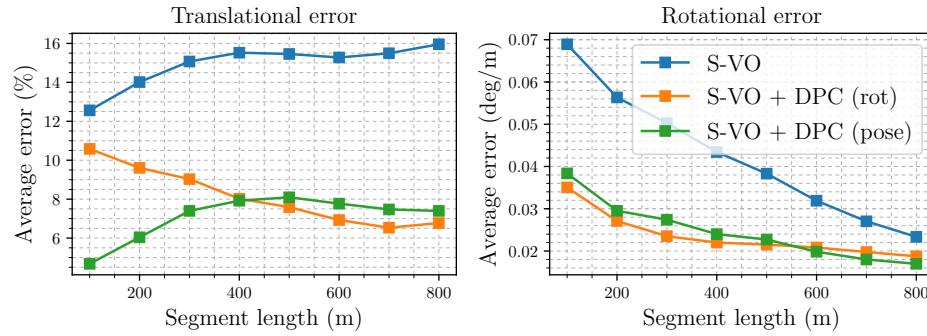
where  $N_s$  and  $s_p$  (the number of segments of a given length, and the number of poses in each segment, respectively) are computed based on the selected segment length  $s$ . In this work, we follow the KITTI benchmark and report the mean segment error norms for all  $s \in [100, 200, 300, \dots, 800]$  (m).



(a) Sequence 00 (Distorted): Segment Errors.



(b) Sequence 02 (Distorted): Segment Errors.



(c) Sequence 05 (Distorted): Segment Errors.

Figure 6.9: c-ATE and segment errors for S-VO with radially distorted images with and without DPC-Net.

Table 6.1: m-ATE and Mean Segment Errors for VO results with and without DPC-Net.

Sequence (Length)	Estimator	Corr. Type	m-ATE		Mean Segment Errors	
			Translation (m)	Rotation (deg)	Translation (%)	Rotation (millideg / m)
00 (3.7 km) <sup>1</sup>	S-VO	—	60.22	18.25	2.88	11.18
	Dense	—	<b>12.41</b>	<b>2.45</b>	<b>1.28</b>	<b>5.42</b>
	S-VO + DPC-Net	Pose	15.68	3.07	1.62	5.59
		Rotation	26.67	7.41	1.70	6.14
		Yaw	29.27	8.32	1.94	7.47
02 (5.1 km) <sup>2</sup>	S-VO	—	83.17	14.87	2.05	7.25
	Dense	—	<b>16.33</b>	3.19	1.21	4.67
	S-VO + DPC-Net	Pose	17.69	<b>2.86</b>	<b>1.16</b>	4.36
		Rotation	20.66	3.10	1.21	<b>4.28</b>
		Yaw	49.07	10.17	1.53	6.56
05 (2.2 km) <sup>3</sup>	S-VO	—	27.59	9.54	1.99	9.47
	Dense	—	<b>5.83</b>	<b>2.05</b>	<b>0.69</b>	<b>3.20</b>
	S-VO + DPC-Net	Pose	9.82	3.57	1.34	5.62
		Rotation	9.67	2.53	1.10	4.68
		Yaw	18.37	6.15	1.37	6.90

<sup>1</sup> Training sequences 05, 06, 07, 08, 09, 10. Validation sequence 02.

<sup>2</sup> Training sequences 00, 06, 07, 08, 09, 10. Validation sequence 05.

<sup>3</sup> Training sequences 00, 02, 07, 08, 09, 10. Validation sequence 06.

<sup>4</sup> All models trained for 30 epochs. The final model is selected based on the epoch with the lowest validation error.

## 6.6 Results & Discussion

### 6.6.1 Correcting Sparse Visual Odometry

Figures 6.6 and 6.7 plot c-ATE and mean segment errors for test sequences 00, 02 and 05 for three different DPC-Net models paired with our S-VO pipeline. Table 6.1 summarize the results quantitatively, while Figure 6.8 plots the North-East projection of each trajectory. On average, DPC-Net trained with the full SE(3) loss reduced translational m-ATE by 72%, rotational m-ATE by 75%, translational mean segment errors by 40% and rotational mean segment errors by 44% (relative to the uncorrected estimator). Mean segment errors of the sparse estimator with DPC approached those observed from the dense estimator on sequence 00, and outperformed the dense estimator on 02. Sequence 05 produced two notable results: (1) although DPC-Net significantly reduced S-VO errors, the dense estimator still outperformed it in all statistics and (2) the full SE(3) corrections performed slightly worse than their SO(3) counterparts. We suspect the latter effect is a result of motion estimates with predominantly rotational errors which are easier to learn with an SO(3) loss.

In general, coupling DPC-Net with a simple frame-to-frame sparse visual localizer yielded a final localization pipeline with accuracy similar to that of a dense pipeline while requiring significantly less visual data (recall that DPC-Net uses resized images).

Table 6.2: m-ATE and Mean Segment Errors for VO results with and without DPC-Net for distorted images.

Sequence (Length)	Estimator	Corr. Type	m-ATE		Mean Segment Errors	
			Translation (m)	Rotation (deg)	Translation (%)	Rotation (millideg / m)
00-distorted (3.7 km)	S-VO	—	168.27	37.15	14.52	46.43
	S-VO + DPC	Pose	114.35	28.64	<b>6.73</b>	29.93
		Rotation	<b>84.54</b>	21.90	9.58	<b>25.28</b>
02-distorted (5.1 km)	S-VO	—	335.82	51.05	13.74	34.37
	S-VO + DPC	Pose	<b>196.90</b>	<b>23.66</b>	<b>7.49</b>	<b>25.20</b>
		Rotation	269.90	53.11	9.25	25.99
05-distorted (2.2 km)	S-VO	—	73.44	12.27	14.99	42.45
	S-VO + DPC	Pose	<b>47.50</b>	<b>10.54</b>	<b>7.11</b>	24.60
		Rotation	71.42	13.10	8.14	<b>23.56</b>

## 6.6.2 Distorted Images

Figure 6.9 plots mean segment errors for the radially distorted dataset. On average, DPC-Net trained with the full SE(3) loss reduced translational mean segment errors by 50% and rotational mean segment errors by 35% (relative to the uncorrected sparse estimator, see Table 6.2). The yaw-only DPC-Net corrections did not produce consistent improvements (we suspect due to the presence of large errors in the remaining degrees of freedom as a result of the distortion procedure). Nevertheless, DPC-Net trained with SE(3) and SO(3) losses was able to significantly mitigate the effect of a poorly calibrated camera model. We are actively working on modifications to the network that would allow the corrected results to approach those of the undistorted case.

## 6.7 Summary

In summary, DPC-Net learns six degree-of-freedom corrections for a given egomotion pipeline and environment. Our contributions include

1. the formulation of a novel deep corrective approach to egomotion estimation,
2. a novel cost function for deep SE(3) regression that naturally balances translation and rotation errors, and
3. an open-source implementation of DPC-Net in PyTorch<sup>10</sup>.

<sup>10</sup>See <https://github.com/utiasSTARS/dpc-net>.

# Chapter 7

## Learned Probabilistic Rotations

I shall be telling this with a sigh  
Somewhere ages and ages hence:  
Two roads diverged in a wood, and I—  
I took the one less traveled by,  
And that has made all the difference.

Robert Frost

### 7.1 Introduction

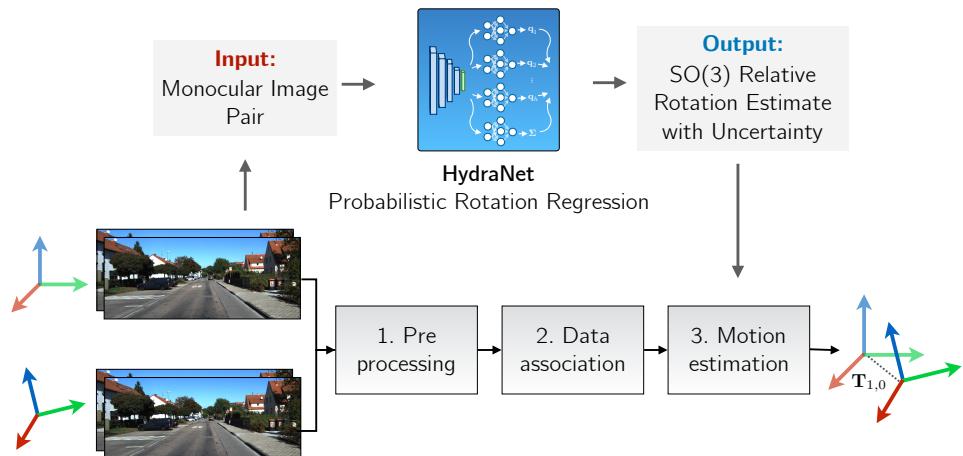


Figure 7.1: We use a network structure we name *HydraNet* to learn rotations with associated uncertainty.

Finally, we take the lessons of Sun-BCNN and DPC-Net and present a method to learn probabilistic estimates of rotation. We show that this pseudo-sensor can extract relative ro-

tation estimates which can be fused with a classical visual odometry egomotion estimate through pose graph optimization. To do this, we develop a network structure we call *HydraNet* that can account for both epistemic and aleatoric sources of uncertainty and adapt it to the problem of estimating elements of  $\text{SO}(3)$ . This work is associated with one publication:

- Peretroukhin, V., Wagstaff, B., and Kelly, J. (2019). Deep probabilistic regression of elements of  $\text{SO}(3)$  using quaternion averaging and uncertainty injection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19) Workshop on Uncertainty and Robustness in Deep Visual Learning*, pages 83–86, Long Beach, California, USA.

## 7.2 Motivation

Accounting for position and orientation, or pose, is at the heart of computer vision. Many algorithms in image classification and feature tracking, for example, are explicitly concerned with output that is robust to camera orientation. Conversely, algorithms like visual odometry, structure from motion, and SLAM use visual sensors to estimate and track the pose of a camera as it moves through some environment.

Recent work (Clark et al., 2017; Melekhov et al., 2017; Kendall et al., 2015) has attempted to transfer the success of deep neural networks in many areas of computer vision to the task of camera pose estimation. These approaches, however, can produce arbitrarily poor pose estimates if sensor data differs from what is observed during training (i.e., it is ‘out of training distribution’) and their monolithic nature makes them difficult to debug. Further, despite much research effort, classical motion estimation algorithms, like stereo visual odometry, still achieve state-of-the-art performance in nominal conditions<sup>1</sup>. Nevertheless, the representational power of deep regression algorithms makes them an attractive option to complement classical motion estimation when these latter methods perform poorly (e.g., under diverse

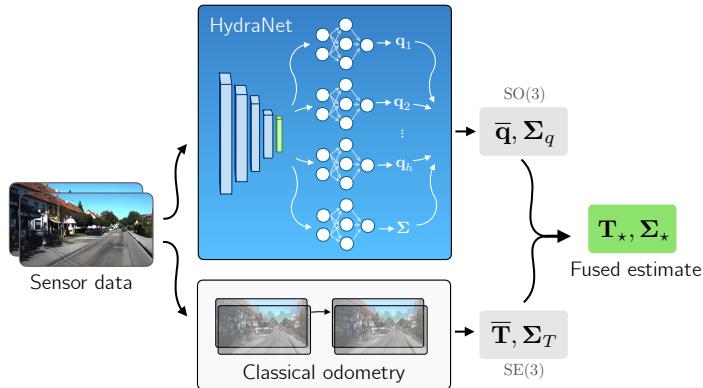


Figure 7.2: We improve classical pose estimation by fusing it with deep probabilistic models.

<sup>1</sup>Based on the KITTI odometry leaderboard (Geiger et al., 2013) at the time of writing.

lighting conditions or low scene texture). By endowing deep regression models with a useful notion of uncertainty, we can account for out-of-training-distribution errors and fuse these models with classical methods using probabilistic factor graphs. In this work, we choose to focus on rotation regression, since many motion algorithms are sensitive to rotation errors (Peretroukhin et al., 2018), and good rotation initializations can be critical to robust optimization.

## 7.3 Related work

Much recent work in the literature has been devoted to replacing classical localization algorithms with deep network equivalents. Some approaches (Clark et al., 2017; Kendall et al., 2015; Kendall and Cipolla, 2017; Melekhov et al., 2017) learn poses directly, while others learn them indirectly as the spatial transforms that result in minimal loss defined over some other domain (e.g., pixel or depth space) (Byravan and Fox, 2017; Handa et al., 2016).

Despite this surge of research in neural-network-based replacements, some authors have nevertheless used deep networks to augment classical state estimation algorithms. Deep networks have been trained as pose correctors whose corrections can be fused with existing estimates through pose graph relaxation (Peretroukhin and Kelly, 2018) (see Chapter 6), and as depth prediction networks that can be incorporated into a classical monocular pipelines to provide an initial estimate for metric scale (Yang et al., 2018). Our work is perhaps closest in spirit to (Haarnoja et al., 2016) which fuses deep probabilistic observation functions with classical models using a Kalman Filter, but focuses on unconstrained targets and does not investigate uncertainty quantification.

In the robotics community, there has been significant effort to leverage the tools of matrix Lie groups to handle poses and associated uncertainty (Solà et al., 2018; Barfoot and Furgale, 2014). In parallel, the computer vision community has developed a rich literature of rotation averaging (Hartley et al., 2013) which focuses on principled ways to combine elements of  $\text{SO}(3)$  based on different metrics defined over the group.

Finally, ensembles of networks have been shown to be a scalable way to extract uncertainty for deep regression and classification (Lakshminarayanan et al., 2017), while multi-headed networks have been proposed in the context of ensemble learning (Lee et al., 2015) and for bootstrapped uncertainty in reinforcement learning (Osband et al., 2016).

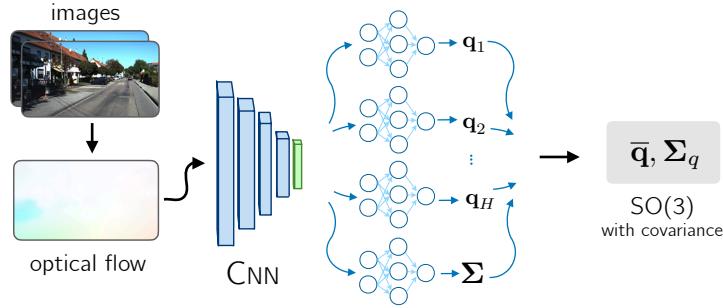


Figure 7.3: The HydraNet structure. Input data (in this case, pre-processed optical flow images) is passed through a main body and then through a number of heads. Outputs are combined to produce an average and an uncertainty.

## 7.4 Approach

We develop our method for probabilistic  $SO(3)$  regression in three steps. First, we motivate why learning elements of  $SO(3)$  is particularly germane to field of egomotion estimation. Second, we present a multi-headed network that can regress unconstrained targets and produce consistent uncertainty estimates. Toward this end, we present a one-dimensional regression experiment, validating prior works ([Lakshminarayanan et al., 2017](#); [Osband et al., 2016](#)) that suggest a bootstrap-inspired approach provides better calibrated uncertainties than one based on stochastic sampling. Finally, we extend these results to targets that belong to  $SO(3)$  by defining a rotation average using the quaternionic metric, and show how we can compute anisotropic uncertainty on four-dimensional unit quaternions.

### 7.4.1 Why Rotations?

We focus our attention on learning rotations for a number of reasons. First, rotations can be learned without reference to scale, using monocular images without the need for metric depth estimation. These images can come from cheap, light-weight imaging sensors that can be found on many ground and aerial vehicles. Furthermore, many depth-equipped sensors like stereo cameras and RGB-D cameras have limited depth range and produce poor depth estimates in large-scale outdoor environments. Second, many egomotion estimation techniques, like visual odometry or visual SLAM, are particularly sensitive to rotation estimates as small early errors have a large influence on final pose estimates. Finally, the constrained nature of rotations presents several difficulties for optimization algorithms. Indeed, if rotations are known, the general problem of pose graph relaxation becomes a linear least squares problem that can be solved with no initial guess for translations ([Carlone et al., 2015b](#)).

### 7.4.2 Probabilistic Regression

In one dimension, given an input  $x$ , with a target output  $y_t$ , we desire a probabilistic estimate

$$\bar{y}, \sigma^2, \quad (7.1)$$

where  $\sigma^2$  captures some notion of model uncertainty (owing to the central limit theorem, we will often make the assumption of Gaussian likelihood).

#### HydraNet

One possible way to obtain  $\bar{y}$  is to train a deep neural network,  $g(x)$ . To endow this network with uncertainty, we present a network structure we call HydraNet (see Figure 7.4). HydraNet is composed of a large, main ‘body’ with multiple heads that each output a prediction,  $g_i(x)$ . To compute  $\bar{y}$ , we compute the arithmetic mean of the outputs,

$$\bar{y} = \frac{1}{H} \sum_{i=1}^H g_i(x). \quad (7.2)$$

The head structure, however, provides several key advantages toward the goal of estimating consistent uncertainty. Namely, it allows us to define the overall uncertainty in terms of two sources, *epistemic* ( $\sigma_e$ ) and *aleatoric* ( $\sigma_a$ ):

$$\sigma^2 = \sigma_e^2 + \sigma_a^2. \quad (7.3)$$

The former,  $\sigma_e$ , is also sometimes referred to as model uncertainty; it is a measure of how close a particular test sample is to known training samples. The latter,  $\sigma_a$ , is inherent to the observation of the target itself. Even if the model can localize a test sample exactly in some salient input space, the aleatoric uncertainty will prevent exact regression due to physical processes like sensor noise.

To account for aleatoric uncertainty, we follow prior work (Haarnoja et al., 2016; Lakshminarayanan et al., 2017) and dedicate one head of the network to regressing a variance directly through a negative log likelihood loss under the assumption of Gaussian likelihood.

To capture epistemic uncertainty, we train each head with random weight initializations and apply losses independently during training. During test time, we compute a sample covariance over the different outputs. This approach is inspired by the method of the statistical bootstrap (Osband et al., 2016), which predicts population statistics by computing statistics over subsets of a sample chosen with replacement. Unlike (Osband et al., 2016), we do not train each head of the network with a bootstrapped sample, but instead rely on the random

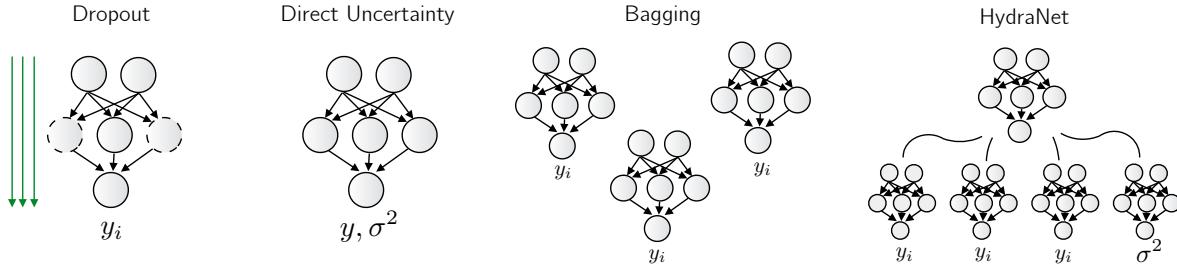


Figure 7.4: Different scalable approaches to neural network uncertainty.

initializations of their parameters and the method of dropout to introduce sufficient stochasticity into their outputs. Unlike (Lakshminarayanan et al., 2017), we do not require numerous trained models that can incur high computational cost for complex regression tasks.

### One-dimensional experiment

To build intuition for the advantages of HydraNet over other methods of extracting uncertainty (e.g., uncertainty through dropout (Gal and Ghahramani, 2016b)), we constructed an experiment similar to that presented in (Osband et al., 2016). We compared HydraNet to four other approaches: (1) direct aleatoric variance regression where the network outputs a second variance parameter that is constrained to be positive, (2) uncertainty through dropout at test time (Gal and Ghahramani, 2016b), (3) bootstrap aggregation (or bagging) of multiple independent models, and (4) HydraNet with no aleatoric uncertainty output.

For each method, we trained a four-layer fully-connected network to regress the output of a one-dimensional function:

$$y_i = x_i + \sin(4(x_i + \omega)) + \sin(13(x_i + \omega)) + \omega, \quad (7.4)$$

where  $w \sim \mathcal{N}(\mu = 0, \sigma^2 = 3^2)$ . Our training set consisted of 1000 samples randomly drawn from  $x \in [0.0, 0.6] \cup [0.8, 1.0]$ , while the test set consisted of 100 samples uniformly drawn from  $x \in [-2, 2]$ . The function and the train/test samples are shown in Figure 7.5a. More details about this experiment can be found in the supplementary materials.

The direct aleatoric uncertainty regression and HydraNet methods were trained using a negative log likelihood loss under the assumption of Gaussian likelihood, while the other methods were trained to minimize mean squared error. We repeated training 100 times, and recorded the test-time negative log likelihood for each method at each repetition. We summarize the results in Figure 7.6. Figure 7.5 presents representative samples from the 100 repetitions for each method. Typically, direct uncertainty regression and dropout are overconfident in the out-of-distribution regions. We replicated the findings of (Osband et al., 2016) who find

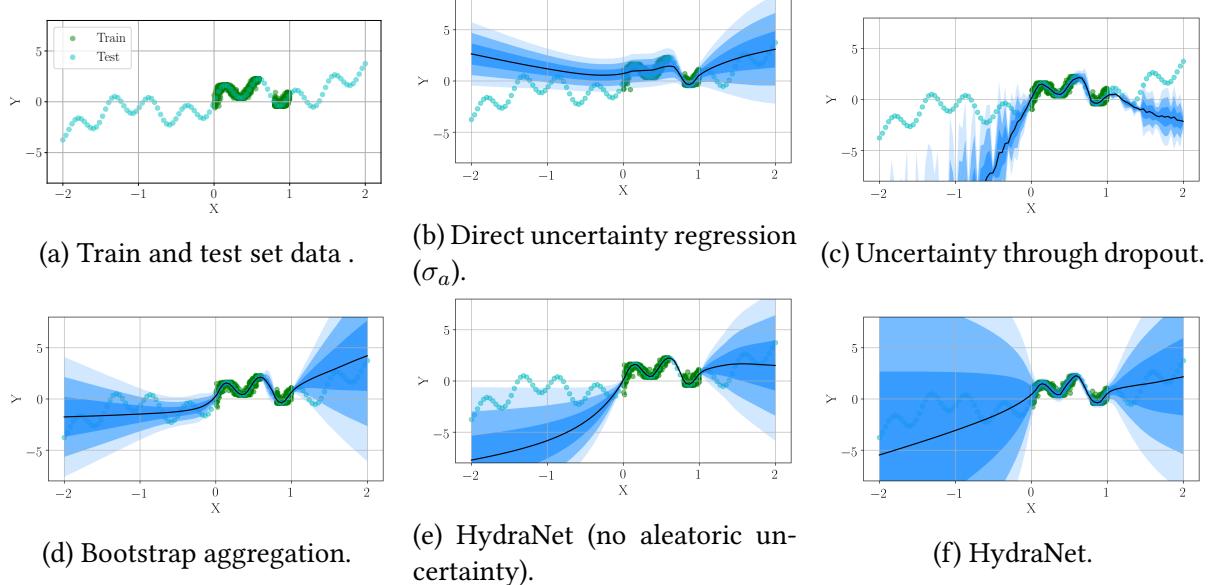


Figure 7.5: A comparison of different ways to extract uncertainty from deep networks. Each shade of blue represents one standard deviation  $\sigma$  produced by the model.

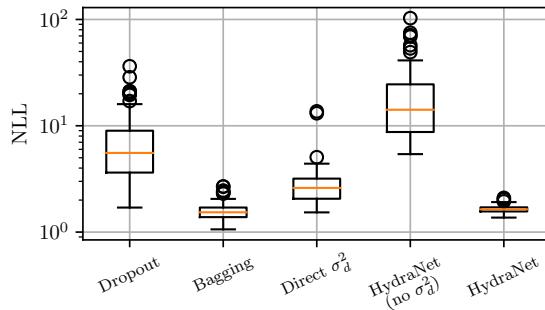


Figure 7.6: Negative log likelihood statistics of 100 repetitions of five neural-network-based uncertainty estimators. HydraNet performs similarly to bagging.

that uncertainty with dropout does not vary smoothly and can collapse outside of the training distribution. HydraNet combined with direct aleatoric uncertainty learning, however, produced similar excellent likelihoods to bootstrap aggregation without requiring multiple models.

### 7.4.3 Deep Probabilistic SO(3) Regression

In order to extend the ideas of HydraNet to the matrix Lie group SO(3), we consider different ways to regress and combine several estimates of rotations. Given a network,  $g(\cdot)$ , and an input  $\mathcal{I}$ , we consider how to extend the ideas of HydraNet to process several outputs,  $g_i(\mathcal{I})$ , and combine them into an estimate of a ‘mean’ rotation,  $\bar{\mathbf{R}}$ , and an associated  $3 \times 3$  covariance

matrix,  $\Sigma$ . To produce estimates of rotation for a given HydraNet head, we consider two options. First if  $g(\mathcal{I}) \in \mathbb{R}^3$ , then we can use the matrix exponential to produce a rotation matrix,

$$\mathbf{R} = \text{Exp}(g(\mathcal{I})). \quad (7.5)$$

Since the capitalized exponential map  $\text{Exp}(\cdot)$  is surjective (Barfoot, 2017; Solà et al., 2018), this approach can parametrize any valid rotation matrix. Alternatively, if  $g(\mathcal{I}) \in \mathbb{R}^4$ , we can normalize it to produce a unit quaternion that resides on  $S^3$ ,

$$\mathbf{q} = \frac{g(\mathcal{I})}{\|g(\mathcal{I})\|}. \quad (7.6)$$

Unit quaternions are a double cover of  $\text{SO}(3)$ , and can represent any rotation. We choose to use this latter parametrization because of its simple analytic mean expression that we describe below.

### Rotation Averaging

To produce a mean of several  $\text{SO}(3)$  elements (i.e., to evaluate Equation (7.2) for rotations), we turn to the field of rotation averaging (Hartley et al., 2013). Given several estimates of a rotation, we define the mean as the rotation which minimizes some squared metric defined over the group<sup>2</sup>,

$$\bar{\mathbf{R}} = \underset{\mathbf{R} \in \text{SO}(3)}{\operatorname{argmin}} \sum_{i=1}^n d(\mathbf{R}_i, \mathbf{R})^2. \quad (7.7)$$

There are three common choices for a bijective metric (Hartley et al., 2013; Carlone et al., 2015b) on  $\text{SO}(3)$ . The angular, chordal and quaternionic:

$$d_{\text{ang}}(\mathbf{R}_a, \mathbf{R}_b) = \|\text{Log}(\mathbf{R}_a \mathbf{R}_b^T)\|_2, \quad (7.8)$$

$$d_{\text{chord}}(\mathbf{R}_a, \mathbf{R}_b) = \|\mathbf{R}_a - \mathbf{R}_b\|_F, \quad (7.9)$$

$$d_{\text{quat}}(\mathbf{q}_a, \mathbf{q}_b) = \min(\|\mathbf{q}_a - \mathbf{q}_b\|_2, \|\mathbf{q}_a + \mathbf{q}_b\|_2), \quad (7.10)$$

where  $\text{Log}(\cdot)$ , represents the capitalized matrix logarithm (Solà et al., 2018), and  $\|\cdot\|_F$  the Frobenius norm. In the context of Equation (7.7), using the angular metric leads to the *Karcher mean*, which requires an iterative solver and has no known analytic expression. Applying the chordal metric leads to an analytic expression for the average but requires the use of

---

<sup>2</sup>Although this is a natural formulation for the rotation mean, it is possible to define other means in terms of absolute errors - see (Hartley et al., 2013).

Singular Value Decomposition. Using the quaternionic metric, however, leads to a simple, analytic expression for the rotation average as the normalized arithmetic mean of a set of unit quaternions (Hartley et al., 2013),

$$\bar{\mathbf{q}} = \underset{\mathbf{R}(\mathbf{q}) \in \text{SO}(3)}{\operatorname{argmin}} \sum_{i=1}^H d_{\text{quat}}(\mathbf{q}_i, \mathbf{q})^2 = \frac{\sum_{i=1}^H \mathbf{q}_i}{\left\| \sum_{i=1}^H \mathbf{q}_i \right\|}. \quad (7.11)$$

This expression is simple to evaluate numerically, and if necessary, can be easily differentiated with respect to its constituent parts. For these reasons, we opt to construct our SO(3) HydraNet using unit quaternion outputs, and evaluate the rotation average using the quaternionic metric.

### SO(3) Uncertainty

There are several ways to approach uncertainty on SO(3). One method (Carlone et al., 2015a) is to define a probability density directly on the group via the isotropic von Mises-Fisher density. This approach has two downsides: (1) it is isotropic and cannot account for dominant degrees of freedom (e.g., vehicle yaw during driving), and (2) estimating the concentration parameter requires approximations or iterative solvers (?).

Instead, we opt to parametrize uncertainty over SO(3) by injecting uncertainty onto the manifold (Forster et al., 2015; Barfoot and Furgale, 2014; Barfoot, 2017) from a local tangent space about some mean element,  $\bar{\mathbf{q}}$ ,

$$\mathbf{q} = \text{Exp}(\boldsymbol{\epsilon}) \otimes \bar{\mathbf{q}}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad (7.12)$$

where  $\otimes$  represents quaternion multiplication. In this formulation,  $\Sigma$  provides a  $3 \times 3$  covariance matrix that can express uncertainty in different directions. Further, given a mean rotation,  $\bar{\mathbf{q}}$ , and samples,  $\mathbf{q}_i$ , we use the logarithmic map to compute a sample covariance matrix,

$$\Sigma_e = \frac{1}{H-1} \sum_{i=1}^H \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T, \quad \boldsymbol{\phi}_i = \text{Log}(\mathbf{q}_i \otimes \bar{\mathbf{q}}^{-1}). \quad (7.13)$$

#### 7.4.4 Loss Function

As with one-dimensional HydraNet, we train a direct regression of covariance through a parametrization of positive semi-definite matrices using a Cholesky decomposition<sup>3</sup> (Hu and

---

<sup>3</sup>Note that in all the experiments presented in this paper, we omit the off-diagonal components of this covariance and only learn a diagonal matrix with non-negative components.

Kantor, 2015; Haarnoja et al., 2016)). Given the network outputs of a unit quaternion  $\mathbf{q}$ , and a positive semi-definite matrix  $\Sigma$ , we define a loss function as the negative log likelihood of a given rotation under Equation (7.12) (see (Forster et al., 2015)) for a given target rotation,  $\mathbf{q}_t$ , as

$$\mathcal{L}_{\text{NLL}}(\mathbf{q}, \mathbf{q}_t, \Sigma_a) = \frac{1}{2} \boldsymbol{\phi}^T \Sigma_a^{-1} \boldsymbol{\phi} + \frac{1}{2} \log \det(\Sigma_a), \quad (7.14)$$

where  $\boldsymbol{\phi} = \text{Log}(\mathbf{q} \otimes \mathbf{q}_t^{-1})$ . Combining the sample covariance, with the learned covariance, we extend Equation (7.3) to

$$\Sigma_t = \Sigma_e + \Sigma_a. \quad (7.15)$$

This covariance estimate is designed to grow for out-of-training-distribution errors (and account for *domain shift* (Lakshminarayanan et al., 2017)) while still accounting for uncertainty within the training set. We note that unlike Bayesian methods, we do not interpret each head as a *sample* from a posterior distribution<sup>4</sup>. Indeed, we note that in our 1D experiments, the heads have very small variance within the training distribution. The multi-headed structure and rotating averaging serves simply as a way to model epistemic uncertainty when the model encounters inputs that differ from those seen during training. We summarize our training and test procedures in Algorithm 4 and Algorithm 5 respectively.

---

**Algorithm 4** Supervised training for SO(3) regression

---

**Require:** Training data  $\mathcal{T}$ , training targets  $\mathbf{q}_t$ , untrained model  $g_\theta(\cdot)$  with parameters  $\theta$  and  $H + 1$  heads

**Ensure:** Probabilistic regression model  $g_\theta(\cdot)$

```

1: function TRAINHYDRANET( $\mathcal{T}$ )
2:   for each mini-batch  $\mathcal{T}_i$  do
3:     Output  $\Sigma_a$  ▷ 1st head, Chol. decom.
4:     for heads 2...( $H + 1$ ) in  $g$  do
5:       Output  $\mathbf{q}_h$  ▷ Equation (7.6)
6:       Evaluate NLL loss ▷ Equation (7.14)
7:     end for
8:     Backprop, update  $\theta$ 
9:   end for
10:  return  $g(\cdot)$ 
11: end function

```

---

<sup>4</sup>Notably, this means we do not scale our direct uncertainty when averaging as  $\frac{1}{H} \Sigma_a$ .

---

**Algorithm 5** Testing of SO(3) regression

---

**Require:** Test sample  $\mathcal{I}_j$ , trained model  $g_\theta(\cdot)$

**Ensure:** Test prediction  $\mathbf{q}$ , covariance  $\Sigma_t \succcurlyeq 0$

```

1: function TESTHYDRANET( $\mathcal{I}_j, g_\theta(\cdot)$ )
2:   Output  $\Sigma_a$                                       $\triangleright$  1st head, Chol. decom.
3:   for heads 2...( $H + 1$ ) in  $g$  do
4:     Output  $\mathbf{q}_h$                                  $\triangleright$  Equation (7.6)
5:   end for
6:   Compute  $\bar{\mathbf{q}}$                              $\triangleright$  Equation (7.11)
7:   Compute  $\Sigma_e$                                  $\triangleright$  Equation (7.13)
8:   return  $\bar{\mathbf{q}}, \Sigma_e + \Sigma_a$ 
9: end function

```

---

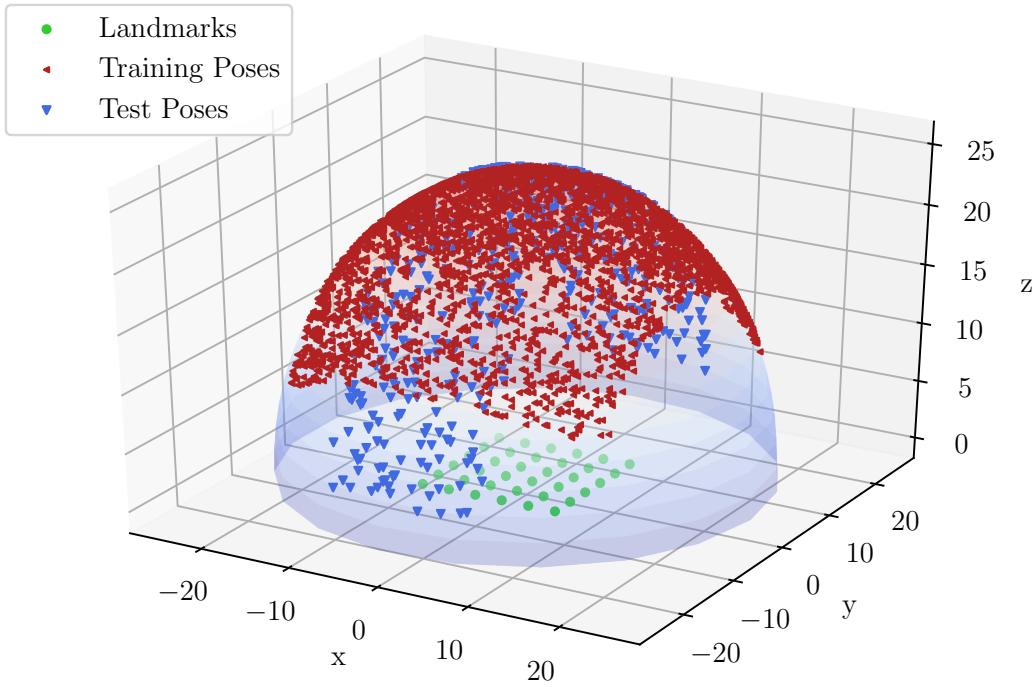
## 7.5 Experiments

### 7.5.1 Uncertainty Evaluation: Synthetic Data

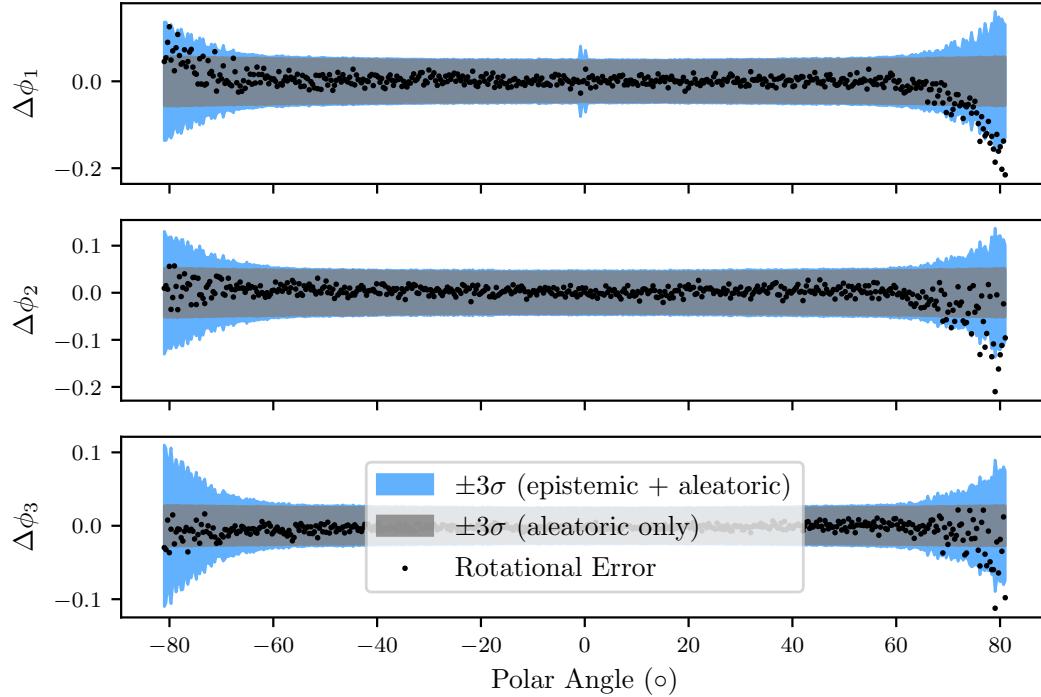
Before we embarked on training with real data, we analyzed our proposed HydraNet structure on a synthetic world. Our goal was to produce probabilistic estimates of camera orientation based on noisy pixel coordinates of a set of fixed point landmarks. To accomplish this, we simulated a monocular camera observing a planar grid of evenly spaced (see Figure 7.7a) landmarks from a hemisphere surrounding the grid. We aligned the monocular camera's optical axis with the centre of the hemisphere so that all landmarks were visible in every camera pose. At each pose, we computed noisy pixel locations of the projection of every landmark, and stacked these 2D locations as an input vector. We generated 15000 training samples with poses that were randomly sampled from the hemisphere in the polar angle range of  $[-60, 60]$  degrees. For testing, we sampled 500 poses in the range of  $[-80, 80]$  degrees, purposely widening the range to include orientations that were not part of training.

To regress the camera orientation, we constructed a five layer residual network and attached 26 heads (25 + 1 for direct uncertainty learning) to regress a probabilistic estimate of  $\mathbf{q}_{c,w}$ , the orientation of the camera with respect to the world frame. Refer to the supplementary materials for full experiment details.

Figure 7.7b plots rotational errors  $\phi = \text{Log}(\mathbf{q} \otimes \mathbf{q}_t^{-1})$  along with 3 sigma bounds based on both the total covariance,  $\Sigma_t$ , and the direct covariance  $\Sigma_a$ . The final regression estimates have consistent uncertainty, composed of a static aleatoric uncertainty and an epistemic uncertainty (Equation (7.13)) that grows when the test samples come from unfamiliar input data.



(a) Synthetic world used to illustrate our method. A monocular camera observes a  $6 \times 6$  grid of point landmarks from poses sampled on a semi-sphere. The test set includes poses that are outside the training distribution.



(b) Rotation estimation errors for a deep network trained using our HydraNet approach on synthetic data (noisy pixel locations of 36 landmarks). We note that outside of the training distribution, our epistemic uncertainty ( $\Sigma_e$ ) grows, as expected.

Figure 7.7: Synthetic experiments of probabilistic rotation regression with HydraNet.

Table 7.1: HydraNet regression results for the 7scenes dataset compared to results reported in (Kendall and Cipolla, 2017). We report mean angular errors and the negative log likelihood (lower is better).

<b>Scene</b>	<b>Error (deg)</b>		<b>NLL</b>	
	HydraNet	PoseNet	HydraNet	PoseNet
💡	Chess	6.3	4.5	-6.0
	Fire	14.9	11.3	-3.6
	Heads	14.3	13.0	-3.9
	Office	8.6	5.6	-5.4
	Pumpkin	9.0	4.8	-5.0
	Kitchen	8.8	5.4	-5.0
	Stairs	11.8	12.4	-4.7

### 7.5.2 Absolute Orientation: 7-Scenes

Next, we used HydraNet to regress absolute orientations from RGB images from the 7-Scenes dataset (Glocker et al., 2013). Our goal was to achieve similar errors to other regression techniques (Kendall and Cipolla, 2017) but augment them with consistent covariance estimates. For this experiment, we used `resnet34` (He et al., 2016) (pre-trained on the ImageNet dataset) for the body of HydraNet and attached 25 HydraNet heads, each consisting of two fully connected layers. We cropped and resized all RGB images to match the expected ImageNet size and omitted the depth channel. The supplementary material provides more details about our experimental parameters. Table 7.1 presents the mean angular errors and negative log likelihoods achieved by our method. The HydraNet-based network produces similar angular errors to other regression methods (Kendall and Cipolla, 2017) but with additional benefit of consistent three-degree-of-freedom uncertainty. Note that we spent little time optimizing the network itself, and note that state-of-the art errors can be achieved using more sophisticated pixel-based losses (Brachmann and Rother, 2018). However, the general HydraNet structure and loss can be used whenever a probabilistic rotation output is required. Further, our results show that our covariance formulation can be used for ‘large’ rotation elements, where techniques (e.g., (Peretroukhin and Kelly, 2018)) that assume ‘small’ corrections may fail.

### 7.5.3 Relative Rotation: KITTI Visual Odometry

Finally, to show the benefit of fusing deep probabilistic estimates with classical estimators, we trained a network to estimate relative frame-to-frame rotations on the KITTI visual odometry (VO) benchmark. To regress relative rotations, we use the HydraNet-based network described

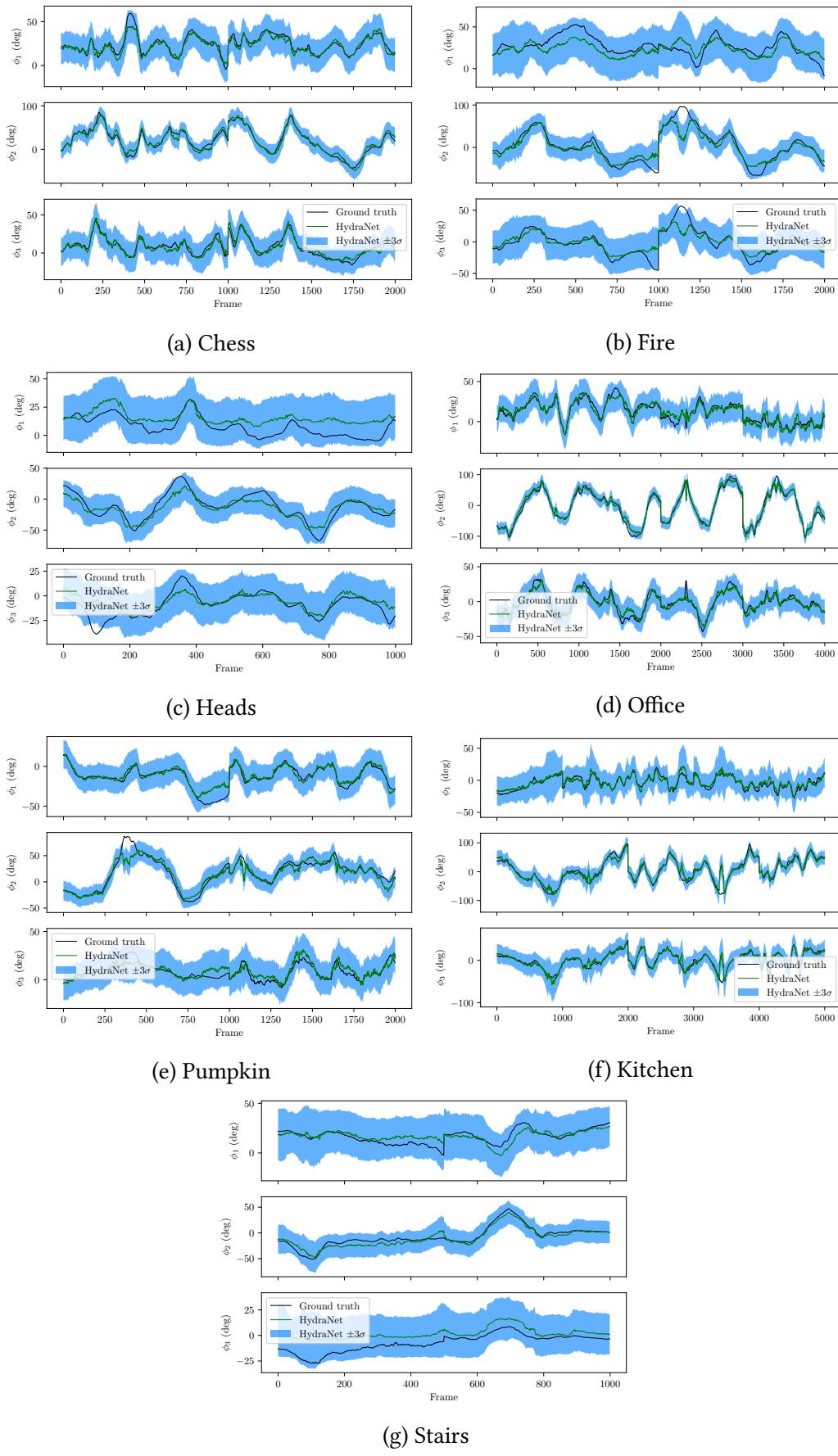


Figure 7.8: Probabilistic regression plots for all seven datasets from the 7-Scenes dataset.

Table 7.2: Results of fusing HydraNet relative rotation regression with classical stereo visual odometry.

<b>Sequence (Length)</b>	<b>Estimator</b>	<b>m-ATE</b>		<b>Mean Segment Errors</b>	
		Translation (m)	Rotation (°)	Translation (%)	Rotation (°/100m)
00 (3.7 km)	DeepVO (Wang et al., 2017)	—	—	—	—
	SfMLearner (Zhou et al., 2017)	—	—	65.27	6.23
	UnDeepVO (Li et al., 2017b)	—	—	4.14	1.92
	viso2-s	27.91	6.25	1.96	0.81
	viso2-s + HydraNet	9.86	2.83	1.34	0.63
	Keyframe Direct VO	12.41	2.45	1.28	0.54
02 (5.1 km)	DeepVO	—	—	—	—
	SfMLearner	—	—	57.59	4.09
	UnDeepVO	—	—	5.58	2.44
	viso2-s	64.67	8.45	1.47	0.56
	viso2-s + HydraNet	50.19	6.51	1.47	0.63
	Keyframe Direct VO	16.33	3.19	1.21	0.47
05 (2.2 km)	DeepVO	—	—	2.62	3.61
	SfMLearner	—	—	16.76	4.06
	UnDeepVO	—	—	3.40	1.50
	viso2-s	23.72	8.10	1.79	0.79
	viso2-s + HydraNet	9.85	3.23	1.38	0.60
	Keyframe Direct VO	5.83	2.05	0.69	0.32



Table 7.3: HydraNet regression results for the KITTI odometry dataset. We report mean angular errors and the negative log likelihood (lower is better).

<b>Sequence</b>	<b>Mean Angular Error (°)</b>	<b>NLL</b>
00	0.199	-16.84
02	0.138	-18.44
05	0.109	-19.31

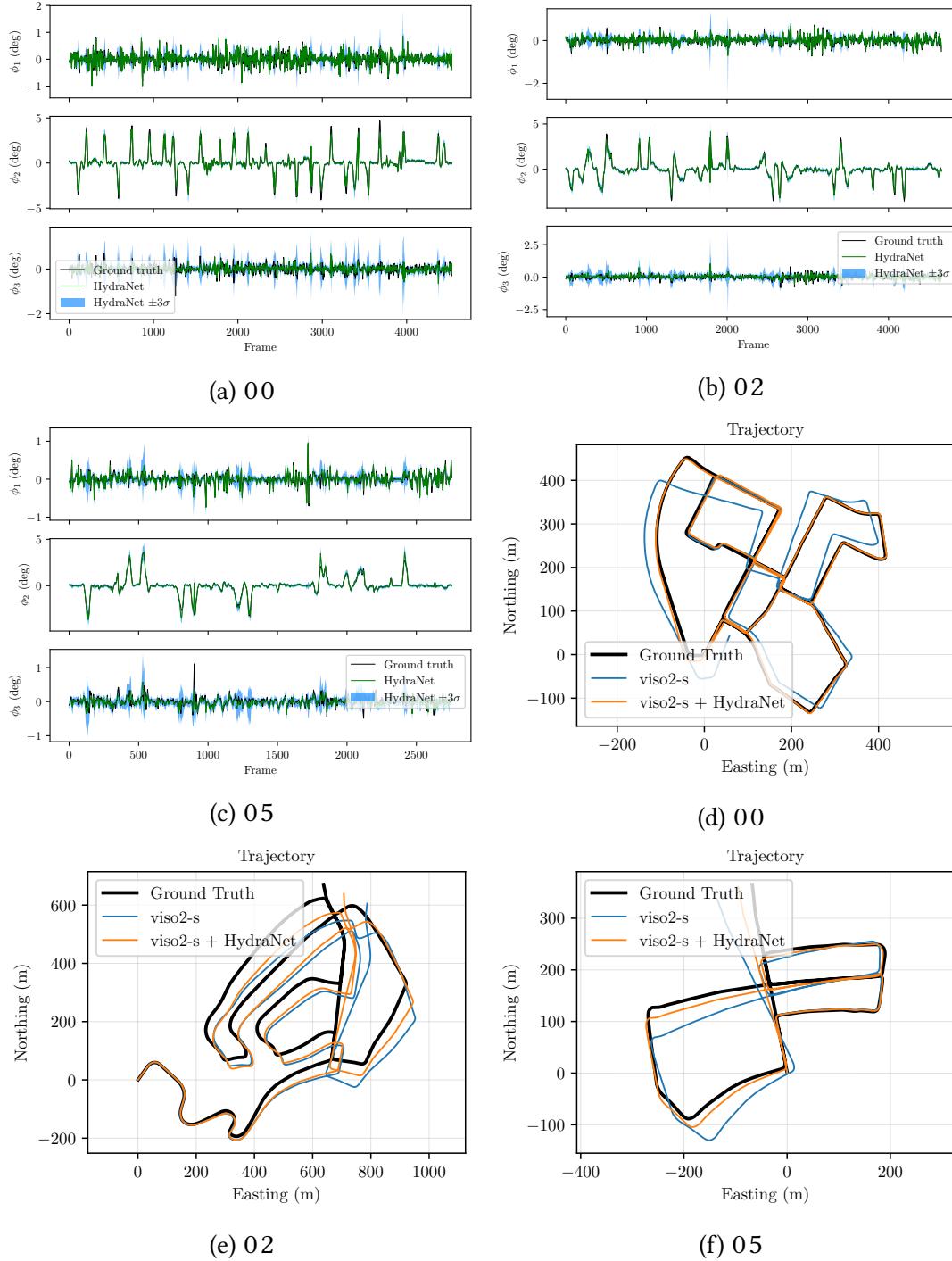


Figure 7.9: KITTI frame-to-frame rotation probabilistic regression for sequences 00, 02 and 05. Top-down trajectory plots show localization improvements after fusion with a classical stereo visual odometry pipeline.

in Figure 7.3. For each pair of poses, we process two RGB images (taken from the left RGB camera) into a two channel dense optical flow image using a fast classical algorithm ([Farnebäck](#),

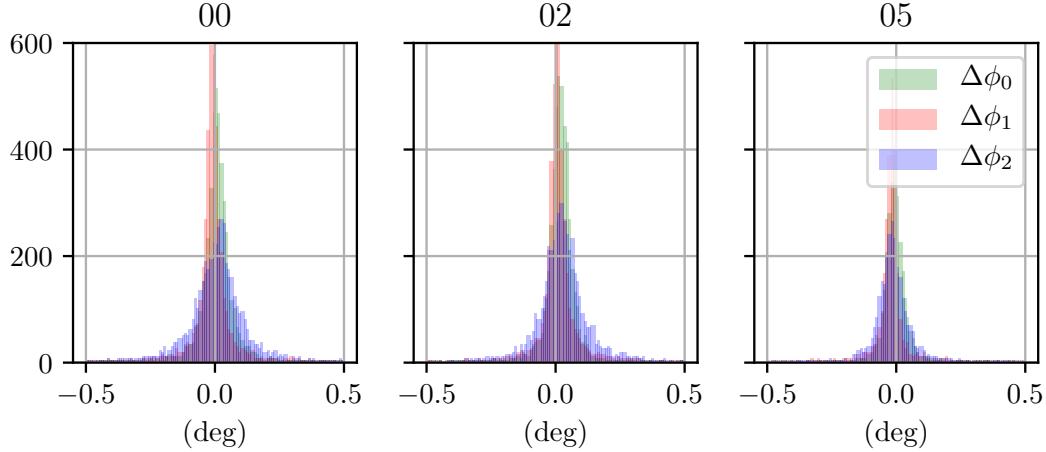


Figure 7.10: Error histograms for test KITTI sequences 00, 02, and 05 on three rotational axes.

2003). Compared to using raw images, we found that using the optical flow pre-processing greatly improved training robustness and rotation accuracy. Since we use two-channel flow images, the body of the network is not pre-trained and instead contains an eight layer convolutional network (see supplemental materials for additional details). We maintained the same head structure as the 7-Scenes experiment. Table 7.3 and Figure 7.10 detail the mean test error and negative log likelihood for KITTI odometry sequences 00, 02 and 05 (chosen for their complexity and length). For each sequence, we trained the model on the remaining sequences in the benchmark. We found our model produced mean errors of approximately 0.1 degrees on all three test sequences. The covariance produced by HydraNet was consistent, spiking during yawing motions when the largest errors occurred (see Figure 7.9). Despite its consistency, the network covariance was dominated by  $\Sigma_a$ . We suspect that unlike the synthetic data,  $\Sigma_e$  remained small throughout the tests sets due to a more constrained input space (RGB or flow images, compared to pixel locations), but leave a thorough investigation to future work.

### Classical VO

For the classical visual odometry estimator, we use a similar pipeline to that used in the prior two chapters based on the open-source `viso2` package (Geiger et al., 2011) to detect and track sparse stereo image key-points. In brief, our pipeline modelled stereo re-projection errors,  $e_{l,t_i}$ , as zero-mean Gaussians with a known static covariance,  $\mathbf{R}$ . To generate an initial guess and to reject outliers, we used three point Random Sample Consensus (RANSAC) based on stereo re-projection error. Finally, we solved for the maximum likelihood transform,  $\mathbf{T}_t^*$ ,

through a Gauss-Newton minimization of

$$\mathbf{T}_t^* = \underset{\mathbf{T}_t \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_i^T \mathbf{R}^{-1} \mathbf{e}_i. \quad (7.16)$$

After convergence, we approximate the frame-to-frame transformation uncertainty as ([Barfoot, 2017](#)):

$$\Sigma_{\text{vo}} \approx \left( \sum_{l=1}^{N_t} \mathbf{J}_{\mathbf{e}_i}^T \mathbf{R}^{-1} \mathbf{J}_{\mathbf{e}_i} \right)^{-1}, \quad (7.17)$$

where  $\mathbf{J}_{\mathbf{e}_i}$  refers to the Jacobian of each reprojection error.

## Fusion via Graph Relaxation

To fuse these estimates with classical VO, we used pose graph relaxation. We describe our method briefly and refer the reader to ([Barfoot, 2017](#)) for a more detailed treatment. For every two poses, we defined a loss function based on a contribution from the estimator and from the network, weighed by their respective covariances:

$$\mathbf{T}_{1,w}^*, \mathbf{T}_{2,w}^* = \underset{\mathbf{T}_{1,w}, \mathbf{T}_{2,w} \in \text{SE}(3)}{\operatorname{argmin}} \mathcal{L}(\hat{\mathbf{T}}_{2,1}, \hat{\mathbf{C}}_{2,1}) \quad (7.18)$$

$$= \delta \boldsymbol{\xi}_{1,2}^T \Sigma_{\text{vo}}^{-1} \delta \boldsymbol{\xi}_{1,2} + \delta \boldsymbol{\phi}_{1,2}^T \Sigma_{\text{hn}}^{-1} \delta \boldsymbol{\phi}_{1,2} \quad (7.19)$$

where  $\delta \boldsymbol{\xi}_{1,2} = \text{Log} \left( (\mathbf{T}_{2,w} \mathbf{T}_{1,w}^{-1}) \hat{\mathbf{T}}_{2,1}^{-1} \right)$  and  $\delta \boldsymbol{\phi}_{1,2} = \text{Log} \left( (\mathbf{C}_{2,w} \mathbf{C}_{1,w}^T) \hat{\mathbf{C}}_{2,1}^T \right)$ . The estimates  $\hat{\mathbf{T}}_{2,1}$ ,  $\Sigma_{\text{vo}}$  and  $\hat{\mathbf{C}}_{2,1}$ ,  $\Sigma_{\text{hn}}$  are provided by our classical estimator and the HydraNet network respectively.

Table 7.2 summarizes the results when we perform this fusion - and Figure 7.9 shows the final effect on the trajectory for sequence 00. We found that fusing deep rotation regression with classical methods results in motion estimates that significantly out-perform other methods that rely on deep regression alone. However, we note that even with consistent estimates, a small bias can affect the final fused estimates (e.g., sequence 05) and removing bias is an important avenue for future work. Further, the KITTI dataset contains few deleterious effects that negatively affect classical algorithms, and therefore we expect that this fusion would produce even more pronounced improvements on more varied visual data.

## 7.6 Summary

In this chapter, we described a method (our final *pseudo-sensor*) to regress probabilistic estimates of rotation using a deep multi-headed network structure. We used the quaternionic metric on  $\text{SO}(3)$  to define a rotation average, and extracted anisotropic covariances by modelling uncertainty through noise injection on the manifold.

Our novel contributions were

1. a deep network structure we call *HydraNet* that builds on prior work ([Lakshminarayanan et al., 2017](#); [Osband et al., 2016](#)) to produce meaningful uncertainties over unconstrained targets,
2. a loss formulation and mathematical framework that extends HydraNet to means and covariances of the rotation group  $\text{SO}(3)$ ,
3. and open source code for  $\text{SO}(3)$  regression<sup>5</sup>.

---

<sup>5</sup>[https://github.com/utiasSTARS/so3\\_learning](https://github.com/utiasSTARS/so3_learning)

# Chapter 8

## Conclusion

Of what a strange nature is knowledge!  
It clings to a mind when it has once  
seized on it like a lichen on a rock.

---

MARY SHELLEY, *Frankenstein; or, The Modern Prometheus*

And thus, we have reached the end. This dissertation has dealt with the development of a general framework for improving the performance of model-based visual odometry pipelines through learned probabilistic pseudo-sensors that extract difficult-to-model latent information. We presented four examples of such *pseudo-sensors*. We close with a final summary of the contributions and list of publications, a discussion of potential future work, and some concluding remarks.

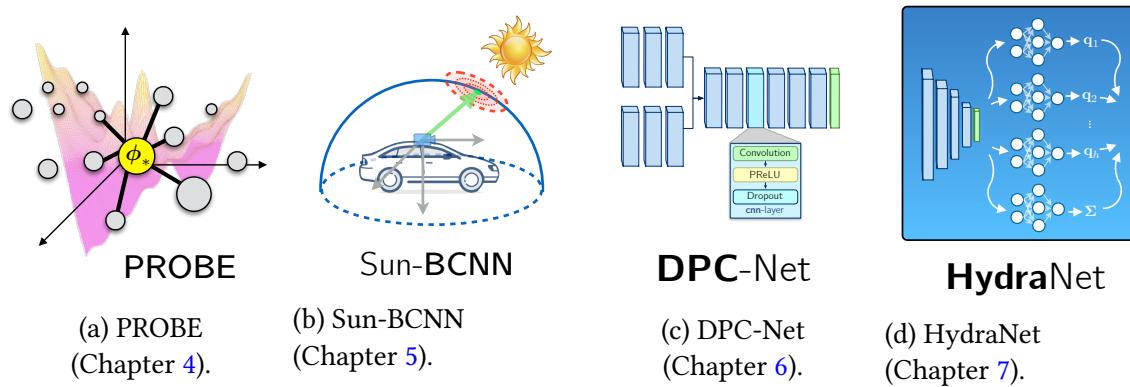


Figure 8.1: Four pseudo-sensors presented in this dissertation.

## 8.1 Summary of Contributions

### 8.1.1 Predictive Robust Estimation

We began with a pseudo-sensor that used a heteroscedastic noise model to enable predictively robust estimation. PROBE contributed

1. a probabilistic model for indirect stereo visual odometry, leading to a predictive robust algorithm for inference on that model,
2. two different approaches to constructing the robust algorithm: one based on k-nearest neighbours, and one based on Generalized Kernel (GK) estimation,
3. a procedure for training our model using pairs of stereo images with known relative transforms, and
4. an iterative, expectation-maximization approach to train our GK model when the relative ground truth egomotion was unavailable.

### 8.1.2 Sun BCNN

With Sun-BCNN, we applied learned *pseudo-sensors* to the problem of illumination direction in outdoor environments. In sum, the novel contributions were:

1. the application of a Bayesian CNN to the problem of sun direction estimation, incorporating the resulting covariance estimates into a visual odometry pipeline;
2. an empirical demonstration that a Bayesian CNN with dropout layers after each convolutional and fully-connected layer can achieve state-of-the-art accuracy at test time;
3. a loss function that incorporated a 3D unit-length sun direction vector, appropriate for full 6-DOF pose estimation;
4. experimental results on over 30 km of visual navigation data in urban (Geiger et al., 2013) and planetary analogue (Furgale et al., 2012) environments;
5. an investigation into the sensitivity of the Bayesian CNN-based sun estimate to cloud cover, camera and environment changes, and measurement parameterization; and
6. open-source software<sup>1</sup>.

---

<sup>1</sup><https://github.com/utiasSTARS/sun-bcnn-vo>.

### 8.1.3 Deep Pose Corrections

Next, we generalized the results of Sun-BCNN to learn full six degree-of-freedom corrections for a particular egomotion pipeline and a given environment with DPC-Net. Our contributions included

1. the formulation of a novel deep corrective approach to egomotion estimation,
2. a novel cost function for deep  $\text{SE}(3)$  regression that naturally balances translation and rotation errors, and
3. an open-source implementation of DPC-Net in PyTorch<sup>2</sup>.

### 8.1.4 Deep Probabilistic Inference of $\text{SO}(3)$ with HydraNet

Finally, we applied the lessons of DPC-Net and Sun-BCNN to learning only rotation estimates through a network structure that incorporated both aleatoric and epistemic uncertainty which can be fused with classical pipelines through pose graph optimization. With this work, we contributed

1. a deep network structure we call *HydraNet* that built on prior work [Lakshminarayanan et al. \(2017\)](#); [Osband et al. \(2016\)](#) to produce meaningful uncertainties over unconstrained targets,
2. a loss formulation and mathematical framework that extends HydraNet to means and covariances of the rotation group  $\text{SO}(3)$ ,
3. and open source code for  $\text{SO}(3)$  regression<sup>3</sup>.

## 8.2 Future Work

There are many avenues for future work. For instance, although the fusion of pipelines with pseudo-sensors can significantly improve localization performance in a given environment, there are few guarantees that the final estimates are accurate and consistent.

A potential thread of future work addresses this deficiency by developing more tightly-coupled perception systems that can be ‘certified’ to produce globally-optimal solutions while still being robust to adverse environmental effects. To do this, we propose that another way to fuse learned models with classical optimization techniques. Instead of treating the learning as

---

<sup>2</sup>See <https://github.com/utiasSTARS/dpc-net>.

<sup>3</sup>[https://github.com/utiasSTARS/so3\\_learning](https://github.com/utiasSTARS/so3_learning).

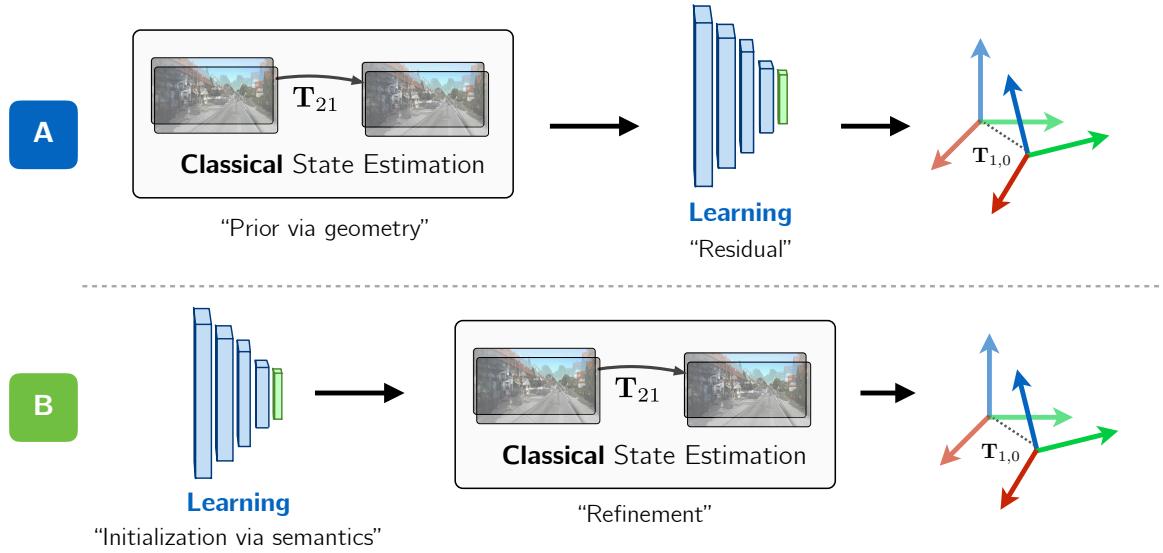


Figure 8.2: Two different ways to incorporate learning with classical pipelines. The first is by using pipelines as a *prior* which can then be corrected by learned approaches, while the second is use learning as an initialization which can then be *refined* by classical techniques.

a way to incorporate corrective residuals, we propose that we use learning as initializations, or priors, that are more invariant to effects like large view-point changes and have the ability to generalize to large-scale environments (Figure 8.2). By incorporating these learned ‘priors’ into an optimization framework, we can leverage recently-developed theory on convex relaxations (e.g., that presented in Rosen et al. (2019)) to ensure that the final localization and mapping results are ‘optimal’, in the sense that they are the global minima of a maximum-likelihood-based loss, and ‘safe’, in the sense that they provide consistent uncertainty estimates even in the presence of unmodelled effects.

## 8.3 Final Remarks

The German Philosopher Georg Hegel had a particular dialectic method that involved a triad: a thesis, antithesis and synthesis. Somewhat curiously, this *thesis* has been an attempt at a *synthesis* of the thesis posed by classical visual egomotion, and the antithesis posed by data-driven end-to-end learning techniques. My hope is that the synthesis proposed by the paradigm of learned *pseudo-sensors* will prove to be a fruitful one within the field of state-estimation, and for the broader robotics community at large.

## 8.4 Coda: In Search of Elegance

Machinery that gives us abundance has left us in want. Our knowledge has made us cynical. Our cleverness, hard and unkind. We think too much, and feel too little. More than machinery, we need humanity.

---

CHARLIE CHAPLIN, *The Great Dictator*

Looking back on my academic journey, I see a path from a fascination with the possible applications of autonomous systems to a fascination with autonomy *in-and-of-itself*. As a budding researcher, I saw robust, accurate perception as a means towards an end. An end which entailed truly autonomous systems ‘perceiving’ and ‘interpreting’ their surroundings with the goal of exploring distant planets and navigating busy urban streets. Now, however, I see ‘perception’ as an end in itself with a plethora of fascinating mathematical, philosophical and ethical challenges that can be tackled in light of, but not subservient to, the potential goals of some grander autonomous system. Throughout this transition, I have become more interested not only in the flesh and blood of perception systems, but also in the *spirit* of them. If perception is one of the bridges we must build to reach the land of autonomy, I am concerned not only with the structural integrity that lets us cross it today, but also with an elegance and rigour that lets it serve as model for posterity.

With this in mind, I want to address a concerning shift that has occurred in the research community throughout my academic career. Many researchers who work on algorithms that enable autonomy (not only in perception, but also in planning and controls) have given up on the dream of modelling the world with the tools of Euclid, Newton and Euler in favour of methods that rely on exemplary data to ‘train’ arbitrarily complex predictive black-box models. In my estimation, this shift has brought with it a certain sense of resignation to the overwhelming complexity of the world. Researchers (including yours truly) are often content to use vague notions of complexity as reason to avoid building analytic models. Instead, we turn to high-parameter data-efficient crude surrogates of our own brains to model what we do not want to. This, I believe, is a tempting mistake. Although these solutions may serve as useful tools to temporarily bridge gaps in our understanding of the world, we will inevitably deplete the low-hanging empirical fruits that they can bear, and we will be left with a deep sense of dissatisfaction that only elegance and simplicity can fill.

I am certainly not the first or the last person who has taken issue with data-driven meth-

ods. Noam Chomsky gives the following critique of purely statistical approaches to science<sup>4</sup>. Consider the study of bee colonies. In order to extract interpretable models of their behaviour (e.g., there is a queen bee, there are worker bees, etc.), one has to observe these colonies meticulously over generations. So why not avoid that entire endeavour and use a data-driven approach? We could set up a camera to observe a bee colony and collect data over several years. By tracking each bee, we could use the tools of modern machine learning to construct and train a large parametric model of each of their positions. Once complete, we could then query this model with a new image from our camera and recover, with extreme precision, the predicted location of each bee. This may allow us to improve honey production, but what have we learned? Is this an elegant model of bee behaviour? Have we not just transformed the problem of understanding the bees into one of understanding this surrogate model? Now consider doing the same with celestial objects—Kepler be damned!

Some may argue that the entire goal of science is predicting the future states of nature, so elegance is irrelevant. I vehemently disagree. I would rather have an interpretable model that is wrong during specific situations (where I can verify that certain assumptions are violated), rather than an obfuscated model which has vague limits to its predictive power.

If history is any judge, the models that stand the test of time are ones that are born out of our meticulous labour and enlightened insight to extract salient principles out of the complexity of the world. The hope that this labour can be replaced with black-box surrogates that indirectly learn these same principles is troubling and, in my view, unnecessary. No matter how much anthropomorphic language we use to describe these surrogates (endowing them with ‘understanding’, ‘attention’, and ‘forgetfulness’), they will always be limited by our own ability to collect sufficient data, and by our ability to craft them in such a way as to consume significant amounts of training exemplars without ‘overfitting’ to them. What’s more, if these models have any interaction with the world, they will also affect the world, and we are committing ourselves to an endless game of cat-and-mouse. Although it may seem that our time is best spent crafting ever-more-clever surrogates, we will soon reach a point where we would be better off using the time and resources towards studying a particular problem more directly.

I do not want to cast aspersions flippantly. The transition to data-driven approaches in computer vision happened for good reason and with much hesitation. The elegance of analytic models has historically only been exceeded by their inability to model the often inelegant ‘real world’. As late as 2006, roboticists were joking that the dirty secret of computer vision is that it doesn’t work. Recent efforts into combining the connectionist ideas of the 20th century with the computational power accessible in the 21st have undoubtedly created systems that do

---

<sup>4</sup><http://norvig.com/chomsky.html>

attain impressive empirical results, and there is a constant stream of new theoretical insights into the types of networks and optimization methods that work well in a given domain.

However, as the world becomes more connected and complex with every passing day, I think it is of utmost importance that autonomy researchers are not tempted to focus solely on empirical results at the cost of elegant solutions. It is now well-accepted that data-driven methods are not the panacea (like it might have seemed for a brief moment a few years ago) to all problems in autonomy. However, this passive agreement may not be enough. Instead, we need to actively suppress the urge to try and solve a problem first through general ‘learning’ methods that are becoming more and more easy to implement and less and less easy to understand. We do not need to relinquish the dream of understanding the world and relegate ourselves to simply predicting it by any means possible. We can instead strive to simplify it and interpret it. If after significant effort we fail at that goal, and only then, should we turn to data-driven learned models to fill in the gaps in our understanding.

# **Appendices**

# Appendix A

## PROBE: Isotropic Covariance Models through K-NN

### A.1 Introduction

This appendix presents our initial exploratory work on isotropic covariance modelling for image features through k-nearest-neighbours. Unlike the work we presented in Chapter 4, here we opt to learn a scalar weight,  $\beta_i$ , that represents the *quality* of image features. We use  $\beta_i$  to scale the covariance of each feature during non-linear optimization. We learn this quality ( $\beta_i$ ) *indirectly* by computing an egomotion estimate with a small subset of visual features and then storing the resulting position error in a prediction space. We repeat this for a fixed amount of iterations. During testing, we map features into this prediction space and compute a  $\beta_i$  based on the  $K$  nearest errors in this prediction space. Our framework is flexible enough that we do not require ground truth at every image and we can, potentially, learn the model based on a single loop closure error.

This initial work is associated with the following publication:

- Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*, pages 3668–3675, Hamburg, Germany.

#### A.1.1 Theory

To solve for the relative egomotion,  $\mathbf{T}_t$ , between two camera frames,  $\underline{\mathcal{F}}_{c_0}$  and  $\underline{\mathcal{F}}_{c_1}$ , we follow the technique described in Section 3.2.3 to convert stereo observations into point-clouds and then solve for the maximum likelihood SE(3) transformation. We associate with each match

$\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}$  a vector of *predictors*,  $\phi_{i,t}$ . We compute the covariance as a function of these predictors, so that  $\mathbf{R}_{i,c_0} = \mathbf{R}_{i,c_1} = \mathbf{R}_{i,t} = \mathbf{R}(\phi_{i,t})$ , and we use the same covariance function for features in both frames<sup>1</sup>,

$$\mathbf{y}_{i,c_0} \sim \mathcal{N}(\bar{\mathbf{y}}_{i,c_0}, \mathbf{R}_{i,t}) = \mathcal{N}(\bar{\mathbf{y}}_{i,c_0}, \mathbf{R}(\phi_{i,t})) \quad (\text{A.1})$$

$$\mathbf{y}_{i,c_1} \sim \mathcal{N}(\bar{\mathbf{y}}_{i,c_1}, \mathbf{R}_{i,t}) = \mathcal{N}(\bar{\mathbf{y}}_{i,c_1}, \mathbf{R}(\phi_{i,t})). \quad (\text{A.2})$$

This then builds the following weighted least squares objective,

$$\mathbf{T}_t^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \sum_{i=1}^{N_t} \mathbf{e}_i(\mathbf{T}_t)^T \Sigma_{i,t}^{-1} \mathbf{e}_i(\mathbf{T}_t). \quad (\text{A.3})$$

where  $\Sigma_{i,t}$  is now given by,

$$\Sigma_{i,t} = \mathbf{D} \mathbf{G}_{i,c_1} \mathbf{R}(\phi_{i,t}) \mathbf{G}_{i,c_1}^T \mathbf{D}^T + \mathbf{D} \mathbf{T}_t \mathbf{G}_{i,c_0} \mathbf{R}(\phi_{i,t}) \mathbf{G}_{i,c_0}^T \mathbf{T}_t^T \mathbf{D}^T \quad (\text{A.4})$$

We build a model for  $\mathbf{R}(\phi_{i,t})$  as,

$$\mathbf{R}(\phi_{i,c}) = \beta(\phi_{i,t}) \bar{\mathbf{R}}, \quad (\text{A.5})$$

with

$$\beta(\phi_{i,c}) = \left( \frac{1}{\epsilon_{\text{avg}} K} \sum_{k=1}^K \epsilon_k \right)^\gamma, \quad \epsilon_k \in \text{k-NN}(\phi_{i,t}), \quad (\text{A.6})$$

where  $\{\epsilon_k\}_{k=1}^K$  are  $K$  egomotion errors that are ‘nearest’ to  $\phi_{i,c}$ ,  $\epsilon_{\text{avg}}$  is an average error,  $\bar{\mathbf{R}}$  is a baseline *nominal* covariance, and  $\gamma > 1$  is a hyper-parameter designed to exaggerate the effect of small changes in position error.

### A.1.2 Training

Training proceeds by traversing the training path, selecting a subset of visual features at each step, and using them to compute an incremental position estimate. By comparing the estimated position to the ground truth position, we compute the translational Root Mean Squared Error (RMSE),  $\epsilon$ , and store it at each feature’s position in the prediction space. The full algorithm is summarized in Algorithm 6.

---

<sup>1</sup>We conjecture that this is reasonable in a VO setup, where images change minimally between consecutive frames.

---

**Algorithm 6** Train PROBE based on a dataset ( $\mathcal{D}$ ) of pairs of input sensor data ( $\mathcal{I}_s$ ) and ground truth egomotion ( $\mathbf{T}_s$ ).

---

```

function BUILDPROBEMODEL( $\mathcal{D}$ )
  for  $l \leftarrow [1, \dots, N_{iter}]$  do
    for all  $\mathcal{I}_s, \mathbf{T}_s$  in  $\mathcal{D}$  do
       $\mathcal{F} \leftarrow \text{EXTRACTFEATURES}(\mathcal{I}_s)$ 
       $\{f_1, \dots, f_J\} \leftarrow \text{SAMPLE}(\mathcal{F})$ 
       $\hat{\mathbf{T}} \leftarrow \text{COMPUTETRANSFORM}(\{f_1, \dots, f_J\})$ 
       $\epsilon \leftarrow \text{ERROR}(\hat{\mathbf{T}}, \mathbf{T}_s)$ 
       $\{\phi_{s,1}, \dots, \phi_{s,J}\} \leftarrow \text{PREDICTOR}(\{f_1, \dots, f_J\})$ 
      Insert  $\{\phi_{s,1}, \dots, \phi_{s,J}\}$  into  $\mathcal{M}$  and store  $\epsilon$  at all  $J$  locations
    end for
  end for
  return  $\mathcal{M}$ 
end function

```

---

### A.1.3 Testing

To use the PROBE model in a test environment, we compute the location of each observed visual feature in our prediction space, and then compute its relative weight  $\beta_i$  as a function of its  $K$  nearest neighbours in the training set. For efficiency, the  $K$  nearest neighbours are found using a  $k$ -d tree. The final scaling factor  $\beta_i$  is a function of the mean of the  $\alpha$  values corresponding to the  $K$  nearest neighbours, normalized by  $\epsilon_{avg}$ , the mean  $\alpha$  value of the entire training set.

---

**Algorithm 7** Compute scalar covariance factors,  $\beta_i$ , for a set of stereo feature tracks (and IMU data),  $\mathcal{F}$ , given a PROBE model  $\mathcal{M}$ .

---

```

function USEPROBE( $\mathcal{M}, \mathcal{F}, \gamma$ )
   $\epsilon_{avg} \leftarrow \text{AVERAGEERROR}(\mathcal{M})$ 
  for all  $f_i$  in  $\mathcal{F}$  do
     $\phi_i \leftarrow \text{PREDICTOR}(f_i)$ 
     $\epsilon_1, \dots, \epsilon_K \leftarrow \text{FINDKNN}(\phi_i, K, \mathcal{M})$ 
     $\beta_i \leftarrow \left( \frac{1}{\epsilon_{avg} K} \sum_{k=1}^K \epsilon_k \right)^\gamma$ 
  end for
  return  $\beta = \{\beta_i\}$ 
end function

```

---

The value of  $K$  can be determined through cross-validation, and in practice depends on the size of the training set and the environment. The computation of  $\beta_i$  is designed to map small differences in learned  $\alpha$  values to scalar weights that span several orders of magnitude. An appropriate value of  $\gamma$  can be found by searching through a set range of candidate values



Figure A.1: Three types of environments in the KITTI dataset, as well as 2 types of environments at the University of Toronto. We use one trial from each category to train and then evaluate separate trials in the same category.

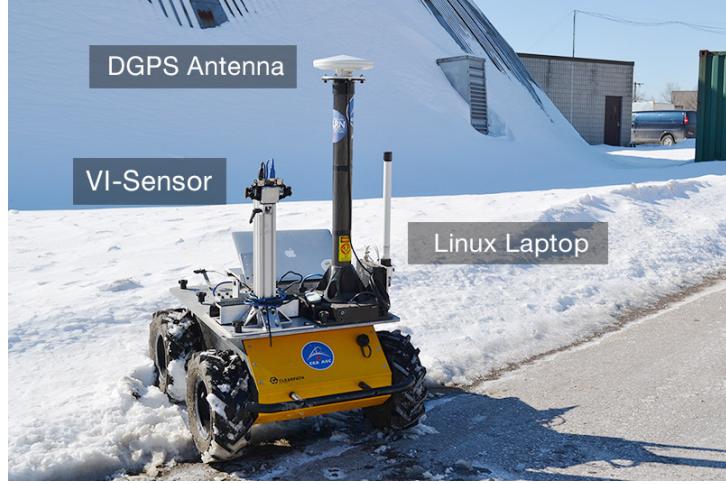


Figure A.2: Our four-wheeled skid-steered Clearpath Husky rover equipped with Skybotix VI-Sensor and Ashtech DGPS antenna used to collect the outdoor UTIAS dataset.

and choosing the value that minimizes the average RMSE (ARMSE) on the training set.

## A.2 Experiments

We trained and evaluated this version of PROBE in two sets of experiments. The first set of experiments made use of 4.5 km of data from the City, Residential, and Road categories of the KITTI dataset (Geiger et al., 2013). In the second set of experiments, we collected indoor and outdoor datasets at the University of Toronto Institute for Aerospace Studies (UTIAS) using a Skybotix VI-Sensor mounted on an Adept MobileRobots Pioneer 3-AT rover and a Clearpath Husky rover, respectively (Figure A.2). In both cases, the camera recorded stereo images at 10 Hz while the IMU operated at 200 Hz. The outdoor dataset consisted of a 264 m training run followed by a 302 m evaluation run, with ground truth provided by RTK-corrected GPS. The indoor dataset consisted of a 32 m training run and a 33 m evaluation run through a room with varying lighting and shadows. For the indoor dataset, no ground truth was available, so we trained PROBE using only the knowledge that the training path should form a closed loop.

Table A.1: Comparison of translational Average Root Mean Square Error (ARMSE) and Final Translational Error on the KITTI dataset.

Trial	Type	Path Length	Nominal RANSAC (99% outlier rejection)		Aggressive RANSAC (99.99% outlier rejection)		PROBE	
			ARMSE	Final Error	ARMSE	Final Error	ARMSE	Final Error
26_drive_0051	City <sup>1</sup>	251.1 m	4.84 m	12.6 m	3.30 m	8.62 m	3.48 m	8.07 m
26_drive_0104	City <sup>1</sup>	245.1 m	0.977 m	4.43 m	0.850 m	3.46 m	1.19 m	3.61 m
29_drive_0071	City <sup>1</sup>	234.0 m	5.44 m	30.3 m	5.44 m	30.4 m	3.03 m	12.8 m
26_drive_0117	City <sup>1</sup>	322.5 m	2.29 m	9.07 m	2.29 m	9.07 m	2.76 m	9.08 m
30_drive_0027	Residential <sup>1,†</sup>	667.8 m	4.22 m	12.2 m	4.30 m	10.6 m	3.64 m	4.57 m
26_drive_0022	Residential <sup>2</sup>	515.3 m	2.21 m	3.99 m	2.66 m	6.09 m	3.06 m	4.99 m
26_drive_0023	Residential <sup>2</sup>	410.8 m	1.64 m	8.20 m	1.77 m	8.27 m	1.71 m	8.13 m
26_drive_0027	Road <sup>3</sup>	339.9 m	1.63 m	8.75 m	1.63 m	8.65 m	1.40 m	7.57 m
26_drive_0028	Road <sup>3</sup>	777.5 m	4.31 m	16.9 m	3.72 m	13.1 m	3.92 m	13.2 m
30_drive_0016	Road <sup>3</sup>	405.0 m	4.56 m	19.5 m	3.33 m	14.6 m	2.76 m	13.9 m
UTIAS Outdoor	Snowy parking lot	302.0 m	7.24 m	10.1 m	7.02 m	10.6 m	6.85 m	6.09 m
UTIAS Indoor	Lab interior	32.83 m	—	0.854 m	—	0.738 m	—	0.617 m

<sup>1</sup> Trained using sequence 09\_26\_drive\_0005. <sup>2</sup> Trained using sequence 09\_26\_drive\_0046. <sup>3</sup> Trained using sequence 09\_26\_drive\_0015.

† This residential trial was evaluated with a model trained on a sequence from the city category because of several moving vehicles that were better represented in that training dataset.

We compare PROBE to what we call a *nominal* visual-inertial navigation pipeline (or VINS, a pipeline based on the point-cloud-error-based VO described in Chapter 3, but with an initial estimate of rotation given by integrating angular velocities from an IMU), as well as a VINS with an aggressive RANSAC routine. In the nominal pipeline, we use RANSAC with enough iterations to be 99% confident that we select only inliers when as many as 50% of the features are outliers. In the aggressive case, we increase the confidence level to 99.99%. When PROBE is used, we apply a pre-processing step that makes use of the rotational estimate from the IMU to reject any egregious feature matches by thresholding the cosine distance between pairs of matched feature vectors. We assume small translations between frames and typically set the threshold to reject feature vectors that are separated by more than five degrees.

For feature extraction, matching, and sparse optical flow calculations, we use the open source vision library libviso2 (Geiger et al., 2011). For all prediction space calculations, we use features in the left image of the stereo pair.

To evaluate PROBE, we run a nominal VINS pipeline on a given test trial, tune the RANSAC threshold to achieve reasonable translation error (< 5% final drift), then repeat the trial with the aggressive RANSAC procedure. Finally, we run this procedure again, this time disabling RANSAC completely and applying our trained PROBE model (with pre-processing) to each observed feature. Table A.1 compares the performance of each trained PROBE model to that of the nominal VINS pipeline and aggressive-RANSAC VINS.

# Appendix B

## Visual Odometry Implementation Details

This appendix presents some implementation details (including all Jacobians) for a standard VO pipeline. You can find a python implementation of this in `pyslam`. The math is largely based on [Barfoot \(2017\)](#).

### B.1 Overview

1. Set the initial camera pose,  $\mathbf{T}_{cw}$ , to a known transform  $\mathbf{T}_{c_0w}$  that defines a world coordinate frame of the first camera frame,  $\underline{\mathcal{F}}_{c_0}$ , with respect to world frame  $\underline{\mathcal{F}}_w$ .
2. Match a set of  $N_t$  landmarks that have been tracked between two stereo camera poses,  $\underline{\mathcal{F}}_{c_0}$  and a subsequent frame  $\underline{\mathcal{F}}_{c_1}$ , and have the image coordinates  $\{\mathbf{y}_{i,c_0}, \mathbf{y}_{i,c_1}\}_{i=1}^{N_t}$ . This step is non-trivial in general, but for our work, we largely rely on the framework `viso2` to provide these putative correspondences.
3. Perform RANSAC using the method of [Umeyama \(1991\)](#) and described in [Barfoot \(2017\)](#) to reject outlying tracks.
4. Select initial guess of the relative transform,  $\mathbf{T}_{c_1c_0} = \mathbf{T}_t$ . Solve the M-estimation problem:

$$\mathbf{T}_{c_1c_0}^* = \mathbf{T}_t^* = \operatorname{argmin}_{\mathbf{T} \in \text{SE}(3)} \sum_{i=1}^N \rho \left( \sqrt{\mathbf{e}_i^T \Sigma_i^{-1} \mathbf{e}_i} \right) = \operatorname{argmin}_{\mathbf{T} \in \text{SE}(3)} \sum_{i=1}^N \rho(\epsilon_i), \quad (\text{B.1})$$

where  $\rho(\cdot)$  can be selected based on empirical performance,  $\mathbf{e}_i$  is the reprojection error,  $\mathbf{e}_i(\mathbf{T}_t) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$ , and  $\mathbf{f}$  is the stereo camera model.

5. Update the current camera pose as:

$$\mathbf{T}_{cw} \leftarrow \mathbf{T}_t \mathbf{T}_{cw} \quad (\text{B.2})$$

## B.2 Solution with Robust Loss

In order to solve Equation (B.1), we use IRLS, and solve the problem,

$$\mathbf{T}^* = \underset{\mathbf{T} \in \text{SE}(3)}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^N \mathbf{e}_i^T \mathbf{M}_i \mathbf{e}_i \quad (\text{B.3})$$

where  $\mathbf{M}_i(\mathbf{T}) \triangleq \frac{1}{\epsilon_i} \frac{\partial \rho}{\partial \epsilon_i} \Sigma_i^{-1}$ . To solve this, we begin at an operating point, we linearize about an operating point  $\mathbf{T}_t^{(n)}$ , as:

$$\delta \boldsymbol{\xi}^* = \underset{\delta \boldsymbol{\xi} \in \mathbb{R}^6}{\operatorname{argmin}} \mathcal{L}(\delta \boldsymbol{\xi}) = \frac{1}{2} \sum_{i=1}^{N_t} (\mathbf{e}_i - \mathbf{J}_i \delta \boldsymbol{\xi})^T \mathbf{M}_i (\mathbf{e}_i - \mathbf{J}_i \delta \boldsymbol{\xi}) \quad (\text{B.4})$$

where  $\mathbf{e}_i = \mathbf{e}_i(\mathbf{T}_t^{(n)})$ ,  $\mathbf{J}_i = \mathbf{J}_i(\mathbf{T}_t^{(n)})$  and  $\mathbf{M}_i = \mathbf{M}_i(\mathbf{T}_t^{(n)})$ . This is solved by the normal equations,

$$\delta \boldsymbol{\xi}^* = \left( \sum_{i=1}^{N_t} \mathbf{J}_i^T \mathbf{M}_i \mathbf{J}_i \right)^{-1} \left( \sum_{i=1}^{N_t} \mathbf{J}_i^T \mathbf{M}_i \mathbf{e}_i \right). \quad (\text{B.5})$$

Given an optimal we update the operating point using:

$$\mathbf{T}_t^{(n+1)} = \text{Exp}(\delta \boldsymbol{\xi}^*) \mathbf{T}_t^{(n)} \quad (\text{B.6})$$

## B.3 Deriving the Necessary Jacobians

We can compute  $\mathbf{J}_i$  by beginning with  $\mathbf{e}_i(\mathbf{T}_t) = \mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{T}_t \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))$  and making the following first order approximations:

$$\mathbf{T}_t \approx (\mathbf{1} + \delta \boldsymbol{\xi}^\wedge) \mathbf{T}_t^{(n)} \quad (\text{B.7})$$

$$\mathbf{f}(\mathbf{p}) \approx \mathbf{f}(\mathbf{p}^{(n)}) + \mathbf{J}_p(\mathbf{p}^{(n)}) \delta \mathbf{p}. \quad (\text{B.8})$$

where  $\mathbf{J}_p(\mathbf{p}^{(n)}) = \frac{\partial \mathbf{f}}{\partial \mathbf{p}} \Big|_{\mathbf{p}^{(n)}}$  and  $\mathbf{p}^{(n)} = \mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})$ . Next we note that,

$$\mathbf{p}^{(n)} + \delta\mathbf{p} \approx (\mathbf{1} + \delta\xi^\wedge) \mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}) \quad (\text{B.9})$$

$$= \underbrace{\mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})}_{\mathbf{p}^{(n)}} + \delta\xi^\wedge \underbrace{\mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})}_{\mathbf{p}^{(n)}} \quad (\text{B.10})$$

$$= \mathbf{p}^{(n)} + \delta\xi^\wedge \mathbf{p}^{(n)} \quad (\text{B.11})$$

$$= \mathbf{p}^{(n)} + (\mathbf{p}^{(n)})^\odot \delta\xi \quad (\text{B.12})$$

which allows us to write  $\delta\mathbf{p} = (\mathbf{p}^{(n)})^\odot \delta\xi$ . Here we have used the notation from [Barfoot \(2017\)](#), to define:

$$\mathbf{p}^\odot = \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix}^\odot = \begin{bmatrix} \mathbf{1} & -\mathbf{p}^\wedge \\ \mathbf{0}^T & \mathbf{0}^T \end{bmatrix}. \quad (\text{B.13})$$

Combining everything, we have:

$$\mathbf{e}_i(\delta\xi) \approx \underbrace{\mathbf{y}_{i,c_1} - \mathbf{f}(\mathbf{p}^{(n)})}_{\mathbf{e}_i(\mathbf{T}_t^{(n)})} - \mathbf{J}_p \delta\mathbf{p} \quad (\text{B.14})$$

$$= \mathbf{e}_i(\mathbf{T}_t^{(n)}) - \mathbf{J}_p(\mathbf{p}^{(n)}) (\mathbf{p}^{(n)})^\odot \delta\xi \quad (\text{B.15})$$

$$= \mathbf{e}_i(\mathbf{T}_t^{(n)}) - \underbrace{\mathbf{J}_p(\mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0})) (\mathbf{T}_t^{(n)} \mathbf{f}^{-1}(\mathbf{y}_{i,c_0}))^\odot}_{\mathbf{J}_i} \delta\xi \quad (\text{B.16})$$

which concludes our derivation. For a stereo camera model defined with the origin in the left camera frame,

$$\mathbf{y}_{i,c} = \begin{bmatrix} u_l \\ v_l \\ d \end{bmatrix} = \mathbf{f}(\mathbf{p}_{i,c}) = \mathbf{f}\left(\begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}\right) = \mathbf{M} \frac{1}{z} \mathbf{p}_{i,c}, \quad (\text{B.17})$$

where

$$\mathbf{M} = \begin{bmatrix} f & 0 & c_u & 0 \\ 0 & f & c_v & 0 \\ 0 & 0 & 0 & fb \end{bmatrix}, \quad (\text{B.18})$$

the Jacobian,  $\mathbf{J}_p$  is given by

$$\mathbf{J}_p = \frac{\partial \mathbf{f}}{\partial \mathbf{p}} = \begin{bmatrix} \frac{f}{z} & 0 & -f \frac{x}{z^2} & 0 \\ 0 & \frac{f}{z} & -f \frac{y}{z^2} & 0 \\ 0 & 0 & -f \frac{b}{z^2} & 0 \end{bmatrix}. \quad (\text{B.19})$$

# Bibliography

- Agarwal, S., Mierle, K., et al. (2016). Ceres solver.
- Alcantarilla, P. F. and Woodford, O. J. (2016). Noise models in feature-based stereo visual odometry.
- Altmann, S. L. (1989). Hamilton, rodrigues, and the quaternion scandal. *Math. Mag.*, 62(5):291–308.
- Barfoot, T. D. (2017). *State Estimation for Robotics*. Cambridge University Press.
- Barfoot, T. D. and Furgale, P. T. (2014). Associating uncertainty with three-dimensional poses for use in estimation problems. *IEEE Trans. Rob.*, 30(3):679–693.
- Brachmann, E. and Rother, C. (2018). Learning less is more-6d camera localization via 3d surface regression. In *Proc. CVPR*, volume 8.
- Byravan, A. and Fox, D. (2017). SE3-nets: Learning rigid body motion using deep neural networks. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 173–180.
- Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., and Leonard, J. J. (2016). Past, present, and future of simultaneous localization and mapping: Toward the Robust-Perception age. *IEEE Trans. Rob.*, 32(6):1309–1332.
- Carlone, L., Rosen, D. M., Calafiore, G., Leonard, J. J., and Dellaert, F. (2015a). Lagrangian duality in 3D SLAM: Verification techniques and optimal solutions. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 125–132.
- Carlone, L., Tron, R., Daniilidis, K., and Dellaert, F. (2015b). Initialization techniques for 3D SLAM: A survey on rotation estimation and its use in pose graph optimization. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4597–4604.

- Cheng, Y., Maimone, M. W., and Matthies, L. (2006). Visual odometry on the mars exploration rovers - a tool to ensure accurate driving and science imaging. *IEEE Robot. Automat. Mag.*, 13(2):54–62.
- Clark, R., Wang, S., Wen, H., Markham, A., and Trigoni, N. (2017). Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem.
- Clement, L. and Kelly, J. (2018). How to train a CAT: learning canonical appearance transformations for direct visual localization under illumination change. *IEEE Robotics and Automation Letters*, 3(3):2447–2454.
- Clement, L., Peretroukhin, V., and Kelly, J. (2017). Improving the accuracy of stereo visual odometry using visual illumination estimation. In Kulic, D., Nakamura, Y., Khatib, O., and Venture, G., editors, *2016 International Symposium on Experimental Robotics*, volume 1 of *Springer Proceedings in Advanced Robotics*, pages 409–419. Springer International Publishing, Berlin Heidelberg. Invited to Journal Special Issue.
- Costante, G., Mancini, M., Valigi, P., and Ciarfuglia, T. A. (2016). Exploring representation learning with CNNs for Frame-to-Frame Ego-Motion estimation. *IEEE Robotics and Automation Letters*, 1(1):18–25.
- Crete, F., Dolmiere, T., Ladret, P., and Nicolas, M. (2007). The blur effect: perception and estimation with a new no-reference perceptual blur metric. In *Human vision and electronic imaging XII*, volume 6492, page 64920I. International Society for Optics and Photonics.
- Cvišić, I. and Petrović, I. (2015). Stereo odometry based on careful feature selection and tracking. In *Proc. European Conf. on Mobile Robots (ECMR)*, pages 1–6.
- Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition, (CVPR)*, pages 248–255.
- DeTone, D., Malisiewicz, T., and Rabinovich, A. (2016). Deep image homography estimation.
- Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *Proc. Int. Conf. on Machine Learning, ICML’16*, pages 1329–1338.

- Eisenman, A. R., Liebe, C. C., and Perez, R. (2002). Sun sensing on the mars exploration rovers. In *Aerosp. Conf. Proc.*, volume 5, pages 5–2249–5–2262 vol.5. IEEE.
- Engel, J., Stuckler, J., and Cremers, D. (2015). Large-scale direct SLAM with stereo cameras. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 1935–1942.
- Farnebäck, G. (2003). Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer.
- Fischler, M. and Bolles, R. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395.
- Fisher, R. (1953). Dispersion on a sphere. In *Proc. Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 217, pages 295–305. The Royal Society.
- Fitzgibbon, A. W., Robertson, D. P., Criminisi, A., Ramalingam, S., and Blake, A. (2007). Learning priors for calibrating families of stereo cameras. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 1–8.
- Florez, S. A. R. (2010). *Contributions by vision systems to multi-sensor object localization and tracking for intelligent vehicles*. PhD thesis.
- Forster, C., Carlone, L., Dellaert, F., and Scaramuzza, D. (2015). IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation.
- Forster, C., Pizzoli, M., and Scaramuzza, D. (2014). SVO: Fast semi-direct monocular visual odometry. In *Proc. IEEE Int. Conf. Robot. Automat.(ICRA)*, pages 15–22. IEEE.
- Furgale, P. (2011). *Extensions to the Visual Odometry Pipeline for the Exploration of Planetary Surfaces*. PhD thesis.
- Furgale, P. and Barfoot, T. D. (2010). Visual teach and repeat for long-range rover autonomy. *J. Field Robot.*, 27(5):534–560.
- Furgale, P., Carle, P., Enright, J., and Barfoot, T. D. (2012). The devon island rover navigation dataset. *Int. J. Rob. Res.*, 31(6):707–713.
- Furgale, P., Enright, J., and Barfoot, T. (2011). Sun sensor navigation for planetary rovers: Theory and field testing. *IEEE Trans. Aerosp. Electron. Syst.*, 47(3):1631–1647.

- Furgale, P., Rehder, J., and Siegwart, R. (2013). Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286.
- Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.
- Gal, Y. and Ghahramani, Z. (2016a). Bayesian convolutional neural networks with Bernoulli approximate variational inference. In *Proc. Int. Conf. Learning Representations (ICLR), Workshop Track*.
- Gal, Y. and Ghahramani, Z. (2016b). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, pages 1050–1059.
- Garg, R., Carneiro, G., and Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *European Conf. on Comp. Vision*, pages 740–756. Springer.
- Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.*, 32(11):1231–1237.
- Geiger, A., Ziegler, J., and Stiller, C. (2011). StereoScan: Dense 3D reconstruction in real-time. In *Proc. IEEE Intelligent Vehicles Symp. (IV)*, pages 963–968.
- Geman, S., McClure, D. E., and Geman, D. (1992). A nonlinear filter for film restoration and other problems in image processing. *CVGIP: Graphical models and image processing*, 54(4):281–289.
- Glocker, B., Izadi, S., Shotton, J., and Criminisi, A. (2013). Real-time rgb-d camera relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 173–179.
- Grewal, M. S. and Andrews, A. P. (2010). Applications of kalman filtering in aerospace 1960 to the present [historical perspectives]. *IEEE Control Syst. Mag.*, 30(3):69–78.
- Haarnoja, T., Ajay, A., Levine, S., and Abbeel, P. (2016). Backprop KF: Learning discriminative deterministic state estimators. In *Proc. Advances in Neural Inform. Process. Syst. (NIPS)*.
- Handa, A., Bloesch, M., Pătrăucean, V., Stent, S., McCormac, J., and Davison, A. (2016). gvnn: Neural network library for geometric computer vision. In *Computer Vision – ECCV 2016 Workshops*, pages 67–82. Springer, Cham.

- Hartley, R., Trumpf, J., Dai, Y., and Li, H. (2013). Rotation averaging. *Int. J. Comput. Vis.*, 103(3):267–305.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6(9):813–827.
- Hu, H. and Kantor, G. (2015). Parametric covariance prediction for heteroscedastic noise. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 3052–3057.
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Int. Conf. Multimedia (MM)*, pages 675–678.
- Kelly, J., Saripalli, S., and Sukhatme, G. S. (2008). Combined visual and inertial navigation for an unmanned aerial vehicle. In *Proc. Field and Service Robot. (FSR)*, pages 255–264.
- Kendall, A. and Cipolla, R. (2016). Modelling uncertainty in deep learning for camera relocalization. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 4762–4769.
- Kendall, A. and Cipolla, R. (2017). Geometric loss functions for camera pose regression with deep learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6555–6564.
- Kendall, A., Grimes, M., and Cipolla, R. (2015). PoseNet: A convolutional network for Real-Time 6-DOF camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946.
- Kerl, C., Sturm, J., and Cremers, D. (2013). Robust odometry estimation for RGB-D cameras. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 3748–3754.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc.

- Lalonde, J.-F., Efros, A. A., and Narasimhan, S. G. (2011). Estimating the natural illumination conditions from a single outdoor image. *Int. J. Comput. Vis.*, 98(2):123–145.
- Lambert, A., Furgale, P., Barfoot, T. D., and Enright, J. (2012). Field testing of visual odometry aided by a sun sensor and inclinometer. *J. Field Robot.*, 29(3):426–444.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., and Batra, D. (2015). Why M heads are better than one: Training a diverse ensemble of deep networks.
- Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., and Furgale, P. (2015). Keyframe-based visual–inertial odometry using nonlinear optimization. *Int. J. Rob. Res.*, 34(3):314–334.
- Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.*
- Li, Q., Qian, J., Zhu, Z., Bao, X., Helwa, M. K., and Schoellig, A. P. (2017a). Deep neural networks for improved, impromptu trajectory tracking of quadrotors. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5183–5189.
- Li, R., Wang, S., Long, Z., and Gu, D. (2017b). UnDeepVO: Monocular visual odometry through unsupervised deep learning.
- Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI’81, pages 674–679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Ma, W.-C., Wang, S., Brubaker, M. A., Fidler, S., and Urtasun, R. (2016). Find your way by observing the sun and other semantic cues.
- MacTavish, K. and Barfoot, T. D. (2015). At all costs: A comparison of robust cost functions for camera correspondence outliers. In *Proc. Conf. on Comp. and Robot Vision (CRV)*, pages 62–69.
- Maddern, W., Pascoe, G., Linegar, C., and Newman, P. (2016). 1 year, 1000 km: The oxford RobotCar dataset. *Int. J. Rob. Res.*
- Maimone, M., Cheng, Y., and Matthies, L. (2007). Two years of visual odometry on the mars exploration rovers. *J. Field Robot.*, 24(3):169–186.

- Mayor, A. (2019). *Gods and Robots*. Princeton University Press.
- McManus, C., Upcroft, B., and Newman, P. (2014). Scene signatures: Localised and point-less features for localisation. In *Proc. Robotics: Science and Systems X*.
- Melekhov, I., Ylioinas, J., Kannala, J., and Rahtu, E. (2017). Relative camera pose estimation using convolutional neural networks. In *Proc. Int. Conf. on Advanced Concepts for Intel. Vision Syst.*, pages 675–687. Springer.
- Nilsson, N. J. (1984). Shakey the robot. Technical report, SRI International.
- Oliveira, G. L., Radwan, N., Burgard, W., and Brox, T. (2017). Topometric localization with deep learning. *arXiv preprint arXiv:1706.08775*.
- Olson, C. F., Matthies, L. H., Schoppers, M., and Maimone, M. W. (2003). Rover navigation using stereo ego-motion. *Robot. Auton. Syst.*, 43(4):215–229.
- Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In *Proc. Advances in Neural Inform. Process. Syst. (NIPS)*, pages 4026–4034.
- Peretroukhin, V., Clement, L., Giamou, M., and Kelly, J. (2015a). PROBE: Predictive robust estimation for visual-inertial navigation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS’15)*, pages 3668–3675, Hamburg, Germany.
- Peretroukhin, V., Clement, L., and Kelly, J. (2015b). Get to the point: Active covariance scaling for feature tracking through motion blur. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Scaling Up Active Perception*, Seattle, Washington, USA.
- Peretroukhin, V., Clement, L., and Kelly, J. (2017). Reducing drift in visual odometry by inferring sun direction using a bayesian convolutional neural network. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA’17)*, pages 2035–2042, Singapore.
- Peretroukhin, V., Clement, L., and Kelly, J. (2018). Inferring sun direction to improve visual odometry: A deep learning approach. *International Journal of Robotics Research*, 37(9):996–1016.
- Peretroukhin, V. and Kelly, J. (2018). DPC-Net: Deep pose correction for visual localization. *IEEE Robotics and Automation Letters*, 3(3):2424–2431.

- Peretroukhin, V., Kelly, J., and Barfoot, T. D. (2014). Optimizing camera perspective for stereo visual odometry. In *Canadian Conference on Comp. and Robot Vision*, pages 1–7.
- Peretroukhin, V., Vega-Brown, W., Roy, N., and Kelly, J. (2016). PROBE-GK: Predictive robust estimation using generalized kernels. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 817–824.
- Peretroukhin, V., Wagstaff, B., and Kelly, J. (2019). Deep probabilistic regression of elements of SO(3) using quaternion averaging and uncertainty injection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’19) Workshop on Uncertainty and Robustness in Deep Visual Learning*, pages 83–86, Long Beach, California, USA.
- Punjani, A. and Abbeel, P. (2015). Deep learning helicopter dynamics models. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 3223–3230.
- Redfield, S. (2019). A definition for robotics as an academic discipline. *Nature Machine Intelligence*, 1(6):263–264.
- Rosen, D. M., Carlone, L., Bandeira, A. S., and Leonard, J. J. (2019). SE-Sync: A certifiably correct algorithm for synchronization over the special euclidean group. *Int. J. Rob. Res.*, 38(2-3):95–125.
- Scaramuzza, D. and Fraundorfer, F. (2011). Visual odometry [tutorial]. *IEEE Robot. Autom. Mag.*, 18(4):80–92.
- Sibley, G., Matthies, L., and Sukhatme, G. (2007). Bias reduction and filter convergence for long range stereo. In *Robotics Research*, pages 285–294. Springer Berlin Heidelberg.
- Sola, J. (2017). Quaternion kinematics for the error-state kalman filter. *arXiv preprint arXiv:1711.02508*.
- Solà, J., Deray, J., and Atchuthan, D. (2018). A micro lie theory for state estimation in robotics.
- Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., and Milford, M. (2015). On the performance of ConvNet features for place recognition. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, pages 4297–4304.
- Sunderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. (2015). Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proc. Robotics: Science and Systems XII*.

- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, pages 1–9.
- Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.-E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A., and Mahoney, P. (2006). Stanley: The robot that won the DARPA grand challenge. *J. Field Robotics*, 23(9):661–692.
- Tsotsos, K., Chiuso, A., and Soatto, S. (2015). Robust inference for visual-inertial sensor fusion. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5203–5210.
- Umeyama, S. (1991). Least-Squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(4):376–380.
- Vega-Brown, W. and Roy, N. (2013). CELLO-EM: Adaptive sensor models without ground truth. In *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pages 1907–1914.
- Vega-Brown, W. R., Doniec, M., and Roy, N. G. (2014). Nonparametric Bayesian inference on multivariate exponential families. In *Proc. Advances in Neural Information Proc. Syst. (NIPS) 27*, pages 2546–2554.
- Wang, S., Clark, R., Wen, H., and Trigoni, N. (2017). DeepVO: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2043–2050.
- Yang, F., Choi, W., and Lin, Y. (2016). Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *Proc. IEEE Int. Conf. Comp. Vision and Pattern Recognition (CVPR)*, pages 2129–2137.
- Yang, N., Wang, R., Stueckler, J., and Cremers, D. (2018). Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry. In *European Conference on Computer Vision (ECCV)*. accepted as oral presentation, arXiv 1807.02570.
- Zhang, G. and Vela, P. (2015). Optimally observable and minimal cardinality monocular SLAM. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, pages 5211–5218.
- Zhou, B., Krähenbühl, P., and Koltun, V. (2019). Does computer vision matter for action?

- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Inform. Process. Syst. (NIPS)*, pages 487–495.
- Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. (2017). Unsupervised learning of depth and Ego-Motion from video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6612–6619.