Universitat Autònoma
de Barcelona

Visual Scene Understanding for Autonomous Vehicles: Understanding Where and What

A dissertation submitted by **Germán Ros Sánchez** at Universitat Autònoma de Barcelonato fulfil the degree of **Doctor of Philosophy**.

Bellaterra, February 16, 2017

Co-Director	**Dr. Angel Sappa** Centre de Visió per Computador
Co-Director	**Dr. Julio Guerrero** Department of Applied Mathematics Universidad de Murcia
Co-Director	**Dr. Antonio López Peña** Dept. Ciències de la computació & Centre de Visió per Computador
Thesis committee	**Dr. Mathieu Salzmann** Computer Vision Laboratory École Polytechnique Fédérale de Lausanne, EPFL **Dr. Pedro Pinies** Department of Engineering Science University of Oxford **Dr. Joost van de Weijer** Centre de Visió per Computador
International evaluators	**Dr. Adrien Gaidon** Computer Vision Research Group Xerox Research Center Europe, XRCE **Dr. Pablo F. Alcantarilla** iRobot Research London, United Kingdom

This document was typeset by the author using LaTeX 2$_\varepsilon$.

ISBN: 978-84-945373-1-8

Printed by Ediciones Gráficas Rey, S.L.

Live as if you were to die tomorrow.
Learn as if you were to live forever.
— Mahatma Gandhi

Those that are firm in their will mold the world to themselves.
— Johann von Goethe

To my family, friends and Anna...

Acknowledgements

This stage as a PhD student has been one of the most rewarding, complete and hard of my life. It has been a stage of challenges, fun and discovery. I am fully aware of the evolution that this stage has produced in me, and I am no less aware of the influence that many people have had on me. To get where I am and to perceive the world as I perceive it today. Shall this serve as a humble recognition of all those who contributed to this work. I hope not to forget anyone.

Let me start with those academically close, *i.e.*, my supervisors, without their support this thesis would have never seen the light. Angel, thank you for your unconditional support, discussions and the freedom you gave me. Antonio, thanks for letting me be part of something bigger, for considering me for the toughest projects and for your help in the worst moments. Julio, I would need too many pages to express you my gratitude. Your involvement and support in this thesis have been enormous. Thank you for having walked this path by my side and for teaching me so much; some of the best moments have been in front of your digital whiteboard. I would also like to thank Daniel Ponsa, for his opinions and ideas, which have helped to improve this project.

I also thank those who welcomed me into their research labs. Prof. Christoph Stiller at MRT, KIT. Jose Manuel Alvarez and Richard Hartley for the opportunity they gave me at NICTA. Pablo F. Alcantarilla and Bjorn Stenger for sharing their time with me at Toshiba Research Europe. Watanabe-san and Okada-san for giving me the opportunity to participate in the Visconti project. I also must thank, Andreas Geiger, Henning Lategahn, Simon Stent and Riccardo Gherardi, who during my stays abroad opened my eyes to new ideas and shared their time with me.

My deepest gratitude to my colleagues at the Computer Vision Center with whom I shared many moments. Allow me to begin with those who I had the pleasure to closely work with, such as David Vazquez, whose tenacity has inspired us all; Biel, Jordi, Sergi, Laura and Joanna, who allowed me to teach a little of what I know in exchange of how much they taught me. And of course, thanks to all the SYNTHIA team, especially to Laura, Fran, Elias and Marc, for their efforts and for helping me during this journey.

I must also thank administration staff of CVC for all the help and sympathy provided over the years. To Mari Carmen, Ana, Claire, Montse, Eva, Silvia, Mireia, Alexandra and Meritxell, thank you very much. I thank those who made my way through the CVC memorable. Thank you for providing hospitality and humanity: Marco Pedersoli, Pep Gonfaus, Javier Marin, Jordi Gonzalez, Xu Hu, Adela Barbulescu, Camp Davesa and Ivet Rafegas. Without Felipe, Hana, Yaxing, Dena, Victor

and Bojana lunch breaks would not be the same. And of course I especially thank to my closest circle, Onur, Gemma, Ariel, Francesco, and Arash, who have been there with me, sharing good and bad times. You know well that I would never be able to pay back all what they have done for me. Your friendship has been the best of this stage.

I leave my most inner circle to the end. Thanks to all my family, my parents, Jose Angel and Maria Dolores, my sister Marina and my grandparents, Lola, Carmen, Manuel and Pedro. Thank you for all the strength and courage you gave me. And my deepest gratitude and love to Anna, who has given me all her unconditional support and love.

Abstract

Making Ground Autonomous Vehicles (GAVs) a reality as a service for the society is one of the major scientific and technological challenges of this century. The potential benefits of autonomous vehicles include reducing accidents, improving traffic congestion and better usage of road infrastructures, among others. These vehicles must operate in our cities, towns and highways, dealing with many different types of situations while respecting traffic rules and protecting human lives. GAVs are expected to deal with all types of scenarios and situations, coping with an uncertain and chaotic world. Therefore, in order to fulfil these demanding requirements GAVs need to be endowed with the capability of understanding their surrounding at many different levels, by means of affordable sensors and artificial intelligence. This capacity to understand the surroundings and the current situation that the vehicle is involved in, is called scene understanding. In this work we investigate novel techniques to bring scene understanding to autonomous vehicles by combining the use of cameras as the main source of information—due to their versatility and affordability—and algorithms based on computer vision and machine learning. We investigate different degrees of understanding of the scene, starting from basic geometric knowledge about *where* is the vehicle within the scene. A robust and efficient estimation of the vehicle location and pose with respect to a map is one of the most fundamental steps towards autonomous driving. We study this problem from the point of view of robustness and computational efficiency, proposing key insights to improve current solutions. Then we advance to higher levels of abstraction to discover *what* is in the scene, by recognizing and parsing all the elements present on a driving scene, such as roads, sidewalks, pedestrians, etc. We investigate this problem known as semantic segmentation, proposing new approaches to improve recognition accuracy and computational efficiency. We cover these points by focusing on key aspects such as: (i) how to leverage computation moving semantics to an offline process, (ii) how to train compact architectures based on deconvolutional networks to achieve their maximum potential, (iii) how to use virtual worlds in combination with domain adaptation to produce accurate models in a cost-effective fashion, and (iv) how to use transfer learning techniques to prepare models to new situations. We finally extend the previous level of knowledge enabling systems to reasoning about *what has change* in a scene with respect to a previous visit, which in return allows for efficient and cost-effective map updating.

Key words: *autonomous driving, computer vision, machine learning, applied mathematics*

Resumen

Hacer de los Vehículos Autónomos Terrestres una realidad al servicio de la sociedad supone uno de los mayores retos científicos de este siglo. Los beneficios potenciales de estos vehículos incluyen reducir accidentes, mejorar el tráfico y un mejor aprovechamiento de las infraestructuras. Los vehículos autónomos deben ser capaces de moverse en nuestras ciudades y autopistas, haciendo frente a cualquier tipo de situación mientras respetan las reglas de tráfico y protegen vidas humanas. Se requiere que estos vehículos se desenvuelvan en cualquier escenario, lidiando con información incierta y un mundo caótico. Para cumplir con esto,W los vehículos autónomos deben estar dotados de la capacidad para entender el entorno a diferentes niveles de complejidad, mediante el uso de sensores asequibles y de la inteligencia artificial. La capacidad de entender el medio en el que operan estos vehículos se conoce como entendimiento de la escena. En este trabajo se investigan nuevos técnicas para dotar a los vehículos de la capacidad para entender la escena, mediante la combinación de cámaras (debido a su versatilidad y bajo coste) y algoritmos de visión artificial y aprendizaje automático. Investigamos diferentes grados de entendimiento de la escena, comenzando por el conocimiento geométrico sobre *dónde* está el vehículo con respecto a la escena. La estimación robusta y eficiente de la posición y pose del vehículo con respecto a un mapa es uno de los puntos fundamentales para la conducción autónoma. Por ello estudiamos el problema desde el punto de vista de la robustez y la eficiencia computacional, proponiendo ideas clave para mejorar las soluciones actuales. Tras ello, avanzamos hacia niveles de abstracción más altos, para descubrir *qué* hay en la escena, reconociendo y segmentando todos los elementos de la misma, tales como: carreteras, aceras, peatones, etc. Investigamos este problema, conocido como segmentación semántica, proponiendo nuevas soluciones para mejorar el grado de acierto y la eficiencia del sistema. Para conseguir estos puntos nos centramos en aspectos clave, tal y como: (i) reducir el coste computacional estimando la semántica en un proceso *offline*, (ii) cómo entrenar modelos neuronales compactos para extraer su máximo potencial, (iii) cómo usar mundos virtuales junto con técnicas de adaptación de dominio para producir modelos precisos y de forma más asequible, y (iv) cómo usar técnicas de transferencia de conocimiento para que los modelos puedan trabajar en nuevos entornos. Finalmente, extendemos las capacidades del sistema, capacitándolo para razonar sobre *qué cosas han cambiado* en la escena con respecto a un tiempo anterior, lo que a cambio posibilita la actualización eficiente y barata de los mapas.

Palabras clave: *vehículos autónomos, visión artificial, aprendizaje automático*

Resum

Fer dels Vehicles Autònoms Terrestres una realitat al servei de la societat suposa un dels majors reptes científics i tecnològics d'aquest segle. Els beneficis potencials d'aquests vehicles inclouen reduir accidents, millorar el trànsit i un millor aprofitament de les infraestructures, entre molts d'altres. Els vehicles autònoms s'han de poder moure a les nostres ciutats, pobles i autopistes, enfrontant-se a qualsevol tipus de situació mentre respecten la normativa de trànsit i protegeixen vides humanes. Es requereix que aquests vehicles es desenvolupin en qualsevol escenari, bregant amb informació incerta i un món caòtic. Per complir aquests objectius els vehicles autònoms han d'estar dotats de la capacitat d'entendre l'entorn a diferents nivells de complexitat, mitjançant l'ús de sensors assequibles i de la intel·ligència artificial. La capacitat d'entendre el medi en el qual operen aquests vehicles es coneix com a enteniment de l'escena. En aquest treball s'investiguen noves tècniques per dotar els vehicles de la capacitat per entendre l'escena, mitjançant la combinació de càmeres (gràcies a la seva versatilitat i el seu baix cost) i algoritmes de visió artificial i d'aprenentatge automàtic. Vam investigar diferents graus d'enteniment de l'escena, començant pel coneixement geomètric d'on és el vehicle dins aquesta escena. L'estimació robusta i eficient de la posició i posi del vehicle respecte d'un mapa és un dels punts fonamentals per a la conducció autònoma. Per això estudiem el problema des dels punts de vista de la robustesa i de l'eficiència computacional, proposant idees clau per a millorar les solucions actuals. Després d'això, avancem cap a nivells d'abstracció més alts, per descobrir què hi ha a l'escena, reconeixent i segmentant tots els elements de la mateixa, com ara: carreteres, voreres, vianants, etc. Vam investigar aquest problema, conegut com a segmentació semàntica, proposant noves solucions per a millorar el grau d'encert i l'eficiència del sistema. Per aconseguir aquests punts ens centrem en aspectes clau, tals com: (i) reduir el cost computacional calculant la semàntica en un procés offline, (ii) entrenar models neuronals compactes per extreure el seu màxim potencial, (iii) fer servir móns virtuals juntament amb tècniques d'adaptació de domini per produir models precisos i de forma més assequible, i (iv) utilitzar tècniques de transferència i de coneixement perquè els models puguin treballar en nous entorns. Finalment, estenem les capacitats del sistema, instruint-lo per raonar sobre quines coses han canviat a l'escena respecte d'un temps anterior, fet que permet una actualització eficient i de baix cost dels mapes.

Paraules clau: *vehicles autònoms, visió artificial, aprenentatge automàtic*

Contents

List of Figures

List of Tables

1 Introduction

1.1 Ground Autonomous Vehicles

Figure 1.1 – Examples of Ground Autonomous Vehicles.

Autonomous Vehicles (AVs) and more specifically their ground counterpart, Ground Autonomous Vehicles (GAVs), have become one of the most revolutionary technologies of the beginning of this century. These vehicles are expected to cause a large social impact in the near future, providing us with a reliable and affordable source of transportation, reducing road fatalities, causing a more steady flow of traffic, reducing fuel consumption and noxious emissions, as well as improving driver comfort and enhance mobility for elderly and handicapped persons [40]. GAVs can be seen as the natural evolution of Advanced Driver Assistance Systems (ADAS), which aim to improve traffic safety by assisting drivers via warnings and performing controlled counteractive measures in dangerous situations.

The principal representative of GAVs is the autonomous car, which is the name that robotics cars receive. Autonomous cars and GAVs in general are robotic platforms endowed with a set of sensors to perceive their environment (*e.g.* cameras, radars, lasers, etc.), actuators to interact with the environment (*e.g.* automatic

steering, acceleration, brake, etc.) and processing hardware to run the algorithms required to understand the environment and make decision throughout these sensors and actuators. The combination of all these systems endows GAVs with certain degree of autonomy to navigate. Such levels of autonomy along with the capacities and limitations of autonomous cars have been defined by different organizations. The National Highway Traffic Safety Administration (NHTSA) proposes the following formal classification [39]:

- Level 0: The vehicle is totally controlled by a human driver
- Level 1: Some basic controls are automated *e.g.* braking
- Level 2: Several controls are automated and can be used together to create an ADAS system, *e.g.* parking assistance
- Level 3: The car can operate autonomously controlling all safety-critical functions in certain conditions (*e.g.* highway scenarios), but the driver must supervise the driving and retake control when the vehicle requests it. The vehicle must provide a sufficiently comfortable "reconnection time" for the driver.
- Level 4: The vehicle is fully autonomous, performing all safety-critical functions 100% of the time. The driver is not expected to retake control of the vehicle at any time.

An alternative taxonomy has been published by the Society of Automotive Engineers (SAE):

- Level 0: Vehicle with no automated control, but capable of issuing warnings
- Level 1: Vehicle endowed with several automated controls to create ADAS systems, Adaptive Cruise Control (ACC), Parking Assistance with automated steering, and Lane Keeping Assistance (LKA). Driver must be ready to take control at any time
- Level 2: Driver is on charge of detecting objects and other hazards, responding accordingly when the automated system fails to respond. The automated system executes accelerating, braking, and steering. The automated system can deactivate immediately upon takeover by the driver.
- Level 3: The vehicle can operate with full autonomy and no human intervention within specific environments, *e.g.* highway

- Level 4: The vehicle can operate in full autonomy in all but few conditions, such as severe weather. The driver is responsible to activate the system only when it is safe to do so (his/her attention then would not be required)
- Level 5: Driver sets the destination and starts the system, the vehicle is fully autonomous and requires no intervention

The automotive industry has progressively started moving from level 0 to higher levels of autonomy, fuelled by the strong motivation of eliminating —or at least palliating— some of the most challenging problems of our era, aiming for:

- Reducing accidents and fatalities (ideally to 0)
- Decreasing congestion in urban areas
- Improving efficiency in road usage
- Increasing human efficiency
- Improving energetic needs for global transportation

The plan is to progressively improve on these points on our way to mature NHTSA-level-4 and SAE-level-5 systems. However, developing GAVs with the aforementioned capabilities presents several technological challenges that are yet to be solved. For vehicles to acquire this level of autonomy it is required a very sophisticated comprehension of the surrounding scene; in other words, through its sensors and software a level-4 GAV must be able to process and understand which are the current entities present in the scene, *i.e.*cars, pedestrians, how many of them, etc.; its own situation within the scene, *e.g.* , type of lane in which the vehicle is located; and its situation with respect to other entities of the scene, considering factors such as speed, trajectory and intentions. The process to comprehend all these critical factors is commonly referred to as scene understanding. By definition, scene understanding refers to the task (or set of tasks) enabling an agent to gain a full interpretation of a scene through video, still images or other media, leading to a human-like interpretation and inference of general principles, rules, behaviours and arbitrary situations. In other words, scene understanding is the fundamental capacity of creating a helpful interpretation of our environment; a mental model that represents the first step towards smart decision making. This problem definition is so broad that in practice scene understanding can be seen as a field containing several problems instead of as an atomic entity; and this field largely intersects with the field of computer vision.

The knowledge of a scene is hierarchical, *i.e.*, a scene can be described at different levels of abstraction: it can be described according to the geometry of the scene

(*e.g.* , a flat region with a whole in the middle), to respect to the objects it contains (*e.g.* pedestrians, buildings, etc.), according to how these objects are placed on the scene (*e.g.* streets, indoor environment, etc.), in terms of the type of scene, in terms to the actions performed by the objects/entities of the scene (*e.g.* a traffic jam, a fight, etc.), among many other levels [1]. These levels bring different information about a scene and can be addressed as different problems, *e.g.* , localization, object detection, place recognition, semantic segmentation, activity recognition, etc. In this way, the scene understanding task, which is usually considered as a long-term goal, can be decomposed in sub-tasks that simplify and bound its scope.

The different levels of abstraction in scene understanding and the different sub-tasks that they spawn can be viewed according to their level of complexity. In this way, finding our own position on a scene would be at the bottom of the tasks to solve in order to fully understand the scene, followed by the knowledge on how to "parse" the elements of the scene considering their associated context (*i.e.*, image parsing or semantic segmentation), and then giving an interpretation of what kind of scene is that (place recognition) and how are their elements interacting (activity recognition). It is therefore, logical to establish a level of maturity for the task of scene understanding according to the degree of maturity obtained when addressing the different levels of abstraction. In this thesis we study how to address these first level of scene understanding, dealing with (ego) localization of a vehicle—*i.e.*, where is the vehicle in the scene?—and semantic segmentation—*i.e.*, what are the objects around me?.

In order to perform these tasks it has proven to be critical to use the right type of sensors. In this way, within the context of driving scenes, visual inputs coming from cameras have become the core source of information to reason about the current situation of traffic, vehicles and other agents. Cameras are a low-cost alternative to active sensors like Lidars and have proven to work well for several scene understanding sub-tasks, as for instance pedestrian detection and obstacle avoidance, reason why they are currently incorporated in all modern vehicles. Thus, since cameras are already there, why not use them to their fullest? The versatility and potentials of cameras in the context of autonomous vehicles is also one of the main motivations of this thesis on visual scene understanding.

1.2 Visual Scene Understanding

Recently, the visual scene understanding ecosystem has lived an important boost due mainly to the triumph of deep learning on several key problems, such as object recognition and semantic segmentation [26]. New approaches based on deep neural networks have proven to achieve a previously never seen level of accuracy

and generalization, going beyond human capabilities in some cases [15]. It is believed that all the knowledge that we have acquired on the creation of new models, datasets and hardware will contribute to the realization of visual scene understanding. However, the blooming of deep learning technologies will take years to be considered in a mature enough state so it can be fully deployed in GAVs.

To compensate for the lack of maturity of visual scene understanding, autonomous driving technology has focused on a technological surrogate that helps to simplify the problem: strong priors in the form of fully-annotated high-definition maps. Instead of trying to understand the current scene from scratch, GAVs rely on highly detailed maps of an area, that usually include 3D data, and the semantics associated to the static scene, among others. In this way, "understanding" becomes a retrieval process of pre-acquired and pre-annotated information when the precise localization of the vehicle with respect to one of these maps is known. In this way, critical information, as for instance, the limits of navigable regions, driving speeds and dangerous areas can be obtained through a manual or a semi-supervised process offline, which drastically simplifies the cognitive requirements of GAVs. Given this pre-baked knowledge, GAVs can focus on identifying new dynamic events, such as obstacles, pedestrians, other vehicles, the state of traffic lights, etc. These tasks are then carried out by a simplified visual scene understanding module.

In practice, mapping technology is considered mature enough to serve as a solid starting point for developing GAV systems, what motivates that a big part of the automotive industry is designing autonomous cars around this concept. On the other hand, those systems dealing with more abstract levels of understanding are still in an embryonary state, just covering partial functionalities such as pedestrian detection and traffic sign recognition. However, the interest in this type of systems is increasing, since they are considered the next leading technology in the development of GAVs. Currently, these technologies are seen as complementary, but in the future when this part of scene understanding becomes mature enough, both system would be able to operate independently, offering a rich source of redundancy when combined.

In this thesis we show several contributions to improve important aspects of two scene understanding sub-tasks, namely, visual localization, and semantic parsing of the scene. To this end, this document has been divided in two parts: part I covers the contributions we have done in visual localization, which resides on the side of the spectrum of those methods that are map-dependent. Part II covers our contributions in scene semantic parsing, with techniques that are map independent.

1.3 Objectives and Scope

The aim of this PhD dissertation is to develop new approaches to understand driving scenes by using computer vision, applied mathematics and machine learning. To this end we set the focus of this thesis on visual scene understanding for driving scenarios, proposing new solutions to understand (part I) "where" is our autonomous vehicle in the scene and (part II) "what" are the objects that surround our vehicle.

With this structure in mind, first we focus on addressing the "where", focusing on the task of robust visual localization. To this end, we study and extend geometric methods to produce faster and more reliable localization approaches to operate in real environments. We ask the following questions:

- Is it possible to exploit compressed regression to create more efficient localization methods?
- How can we embed robustness into compressed regression techniques?
- Can we improve localization robustness using robust manifold optimization?
- Is it possible to exploit rank constraints and sparsity to boost localization accuracy?

The second part of this thesis deals with the "what", focusing on semantic segmentation of the scene and semantic change recognition. Here we propose to use learning-based method and show how these can be combined with maps to improve the computational efficiency of the resulting system. In this part we deal with the following questions:

- How to exploit maps to embed semantic knowledge and reduce computational time?
- How to adapt pre-trained scene understanding models to operate in new unseen conditions?
- How to exploit virtual environments and domain adaptation to improve generalization and accuracy of scene understanding related problems?
- Can we create new training methods to deal with data multi-modality in deep learning models?
- How to compress a state-of-the-art model into an embedded device?
- Can deep learning and virtual worlds solve the problem of change detection in driving scenarios?

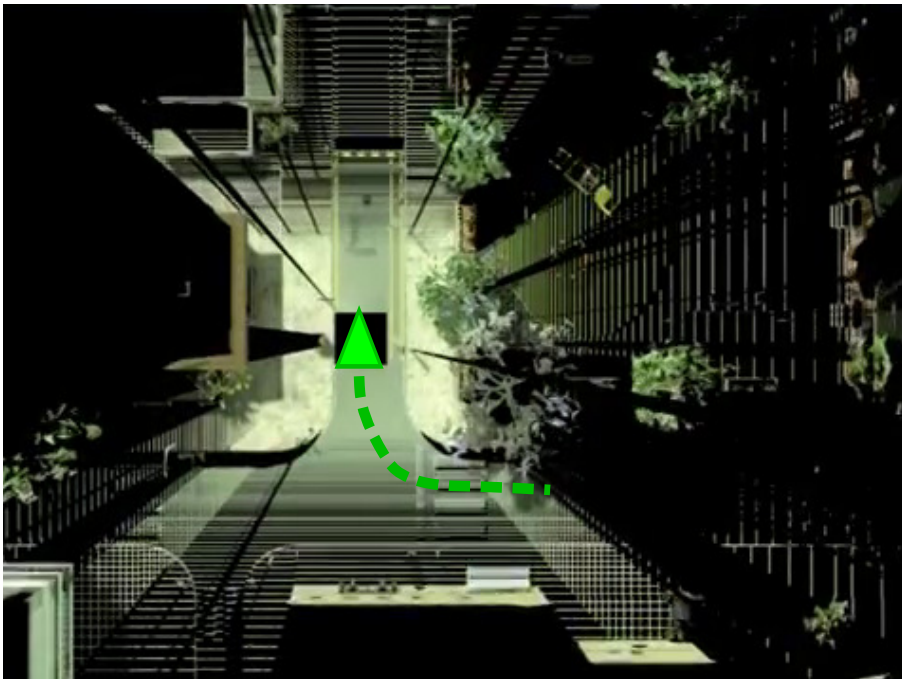
Bringing light to these questions from the point of view of computer vision and machine learning are the objectives of this PhD. Our aim is to contribute to the understanding and development GAVs and to produce mature systems that can understand their surroundings and make decisions accordingly.

1.4 Outline

This PhD thesis has been divided into three parts. The first part of this document, *Where am I on the Scene?*, deals with aspects of visual localization, with special emphasis in localization robustness and efficiency. This part starts with an introduction and motivation to the problem of visual localization in chapter 2. In chapter **??** we deal with embedding techniques and compressed regression to improve efficiency of ego-pose estimation for visual localization. These ideas are extended in chapter **??** where we study how to use data embeddings and robust manifold optimization to increase robustness for pose estimation in localization problems. Chapter **??** studies an alternative viewpoint to localization and pose estimation, making use of the inner low-rank and sparsity constraints of the problem to propose a Robust PCA approach. The methods proposed in chapter **??** are further improved in chapter **??** into an efficient Robust PCA method to deal with larger data volumes.

The second part of this document, *What is on the Scene?*, deals with aspects of semantic segmentation and change detection, with special emphasis to practical problems, such as compressing models to operate in embedded devices and adapting models to work in new conditions using transfer learning and domain adaptation. First, chapter 3 provides a general introduction to these problems. Then in chapter **??** we propose a new strategy based on the combination of mapping, localization and information retrieval to develop a real-time scene understanding approach. Chapter 4 deals with the problem of deploying semantic segmentation models in new unseen scenarios using online unsupervised adaptation. In chapter **??** we cover practical methods to train semantic segmentation models for driving environments, showing how to address data multi-modality issues and model compression for its use in embedded devices. Chapter **??** describes how the combination of virtual environments and domain adaptation can dramatically improve semantic segmentation approaches based on deep-learning. In chapter **??** we show that change detection and semantic change detection can be addressed as variations of semantic segmentation and how to make this possible by exploiting virtual worlds. Finally, in part III, *Clausula*, within chapter 5, we presents the conclusion of this PhD thesis along with a summary of scientific contributions, patents and related deliverables of this thesis.

Where am I on the Scene? Part I



Localization and trajectory estimation on a synthetic scene.

2 Geometric Methods for Localization and Visual Odometry

2.1 Localization and Visual Odometry

In this part of the thesis we deal with the task of localizing a vehicle within a scene using vision, which is one of the first levels of knowledge towards visual scene understanding. The task of localization or ego-localization consists of estimating your own position (and usually orientation) with respect to a scene, that is usually given in the form of a map. For this reason, the task of localization is intimately related to the task of mapping (map creation). In order to delve into these concepts we first need to clarify several key concepts. Firstly, it is important to know that in practice the tasks of localization and mapping can be solved as an atomic task—given rise to Simultaneous Localization and Mapping (SLAM)—, or independently, for instance assuming that a map has already been created and we are just interested in our ego-localization with respect to that map. In the context of ground autonomous vehicles, it is typical that the use of SLAM corresponds to a first stage in which a new area has to be mapped, while localization or re-localization is applied when we know a pre-existing map. The process of visual localization can consider information about the correlation between pre-stored visual references of an area (map) and new references to solve the association problem, but it can also use information coming from the relative motion of the agent (vehicle); something known as Visual Odometry. In many cases, both sources of information are combined.

Allow us to clarify that throughout this thesis we will be using the terms localization and Visual Odometry in a relaxed way, but always to refer to the task of estimating the position and orientation of a vehicle. It is also worth providing a proper definition to the concept of "map". In the context of this document we consider that a map is a '"conceptual" description of a region (or set of regions) of the space encoded in a digital and accessible format. We use the word "conceptual" to define the type of description of maps because such a description typically admits a wide range of information. There can be metric 3D maps, which contain the 3D geometry of an area in a metrically consistent fashion. It is also common to find topological maps, *i.e.*, representations of areas in terms of their connections but neglecting part or the totality of the geometric information of the scene. Also, the type of features used to represent maps could drastically vary. Some maps encode

geometric information as a sparse set of points, while others encode that information as a dense cloud of points. Some more specialized maps encode semantic information of the scene, such as the types of each object, something that we will show in further detail in chapter **??**.

2.2 A Brief History of Localization and Mapping

Visual Localization and Mapping tasks are defined as the tasks of estimating the trajectory of a moving robot (or vehicle) and to create a model of the environment, *i.e.*a map. This map can be formed by the references used to track the position of the robot or by any other piece of information considered useful to represent the environment, as for instance: dense 3D point clouds, representative objects like lane-marking and buildings, etc. Usually, the map itself only can be created after knowing the set of positions from which the information was taken (i.e., the robot trajectory). Here, the concept of "position" should be understood in a general sense. For topological maps, position may refer just to the relationship between a node and its neighbours, while for metrically-precise maps, position would mean the translational and angular values of the camera pose.

This chicken-and-egg problem dates back to the 80's, as described by Durrant-Whyte and Bailey in [8]. At the IEEE Robotics and Automation Conference of 1986, some researchers were trying to apply estimation methods to stochastic robot mapping and localization problems [8]. Among the first pioneers we should highlight Peter Cheeseman, Jim Crowley, Durrant-Whyte, Raja Chatila, Oliver Faugeras and Randal Smith, who actively participated in long discussions about the problem of consistent mapping. These discussions situated localization and mapping as fundamental problems within robotics, and showed some of the conceptual and practical problems that needed to be addressed.

After that, the next keystone was a series of works from different researches, such as Crowley, Chatila and Laumond, which made use of Kalman filter algorithms to perform the localization of the robot. However, the most important contribution was the work presented by Randall Smith, Matthew Self and Peter Cheeseman in [37], where the authors described a way of managing and estimating spatial relationships with their respective uncertainty. In this initial conception, the map —also called stochastic graph— was just a graph with some spatial constraints and the dependencies between estimations. The paper showed the fact that landmarks are correlated with each other due to the shared error in the estimation of the robot trajectory. In this way, it comes up that any consistent solution of the mapping problem would require the estimation of the vehicle pose, and on the other hand, the estimation of the vehicle pose would require the information about landmarks

positions. In other words, both problems were intrinsically connected.

After this, localization and mapping were widely studied by the robotics and the computer vision communities over more than two decades. This phenomenon was mainly occasioned by the natural capabilities of the formulation for integrating robot trajectories, the visual environment of agents and the uncertainty presented in the relations of these elements. At some point several successfully applications arose for robot navigation in indoor and outdoor scenarios [32], [30]. The interest on these problems helped to generate a remarkable amount of knowledge in this area, and made localization and mapping become mature technologies [34].

However, this maturity has to be understood in the context of the original problem, which imposes strong constraints about the environment and the scene. As a consequence the original definitions of these problems were extended to fit in many new cases that arose in more general contexts. This process also produced some new specializations, as for instance Visual Localization, and Visual Localization and Mapping, which consists of solving the SLAM problem with the sole use of cameras [5]. The use of these simple sensors allows the development of accurate autonomous systems at the same time that decreases costs and overall energy consumption.

One of these extensions consists of developing the technologies of localization and mapping to the context of Ground Autonomous Vehicles. However, the constraints arising here generate a problem that is much more complex than the original one. Autonomous vehicles operate in more general environments and have to consider the influence of external factors, such as other vehicles, pedestrians and abnormal alterations of the traffic (e.g., due to accidents or traffic jams). With the aim of solving these problems, the intelligent vehicles community adopted most of the tools and techniques produced by the robotics and computer vision fields. Trends that were originated within robotics were increasingly influencing new methods that arose in the intelligent vehicles community. Part of this heritage is found in the use of active sensors, such as lidar and radar to perform localization and mapping. Actually, within the intelligent vehicles literature, is very easy to find many successful approaches that make use of those sensors to acquire the desired data. Good examples of this are [29], [24] and [27], which describe practical approaches used in international competitions (e.g., the DARPA Grand Challenge, and the European Land Robot Trials), performing the first tests of these ideas in real conditions.

The use of active sensors can simplify the underlying estimation and mapping stages while producing remarkably good results. Such simplification is achieved by shifting part of the complexity from the core of localization and mapping to the acquisition stage, i.e., acquiring dense clouds of 3D points with lasers simplifies the remaining stages. However, developing these approaches solely based on active sen-

sors might be an important drawback, since there are already low-cost alternatives available, as the ones provided by cameras and vision-based algorithms. Cameras are becoming an essential component of modern cars. They are low-cost and are already there for other scene understanding applications, such as pedestrian detection and obstacle avoidance. These reasons led to the proposal of new approaches based on vision, such as the presented in [36], [28] and [19], where different authors show that this is a feasible technology.

2.3 On Robust and Efficient Localization

We have introduced the task of localization as one of the first levels of understanding in the context of GAVs, showing also how this task is intimately associated to the task of mapping. From a scientific point of view, the problem of localization is approachable from many different viewpoints, but in the context of this thesis we set our focus on the specific tasks of improving robustness and computational efficiency of ego-motion methods, a fundamental block for localization.

In the task of visual localization or visual ego-motion estimation there is uncertainty involved; the data used to perform the process of estimation is noisy and this may lead to the computation of wrong solutions and even produce the instability of the whole system. Thus, the proposal and development of new robust estimators and outlier detection techniques become critical for the reliability of the system. Furthermore, these solutions need to be computationally efficient, since they would have to run on embedded devices on-board of autonomous cars. Therefore, the challenge we address in this thesis is not just to build robust ego-pose estimator, but instead to create suitable approaches that fulfil with the balance between reliability and efficiency.

Taking these principles as our goals, we show how to address the problem of ego-motion estimation by using data embedding approaches and compressed regression methods to achieve extremely efficient solutions (chapter **??**). We also show how it is possible to operate in compressed regression spaces without neglecting the constraints imposed by the localization problem (in terms of pose and motion) and maintaining a high level of robustness (chapter **??**). In fact, an important part of our effort has been dedicated to exploit the constraints that arise when one deals with poses and motion. In this regard, through this first part of the document, we present several approaches to deal with these constraints by performing optimization on Lie-groups and Riemannian manifolds (chapters **??** and **??**). Overall, the leitmotiv of the first part of this thesis is to explore techniques to improve estimation robustness that are more efficient than the state-of-the-art, always setting localization for GAVs as our target application.

What is on the Scene? Part II

Semantic segmentation of an outdoor image.

3 Learning Methods for Scene Understanding

3.1 Understanding Driving Scenes

This part of the dissertation describes the work done trying to understand "what" is on the scene, *i.e.*, parsing driving scenes as a whole to recognize and segment all the present objects along with the different parts of the road infrastructure. This problem is known as semantic segmentation. Moreover, we extend and adapt the formulation of semantic segmentation to allow for the recognition of structural changes in the scene.

In chapter 1 we defined scene understanding to be composed by different levels of abstraction, what leads to the formulation of several sub-problems, such as object detection and semantic segmentation among others. We set our interest on parsing the elements of the scene using semantic segmentation, due to the need of a complete description of the scene for autonomous cars. First, we present a basic semantic segmentation pipeline following a classical pipeline composed of hand-crafted features, classifiers and a smoothing technique (Conditional Random Fields). We evaluate the performance of our approach and propose new techniques to improve its computational efficiency and its generalization capacity. Given the computational constraints of GAV systems, we propose a new strategy to leverage the computation of semantics to an offline stage, that is later retrieved after a re-localization process. In order to get around the generalization limitation we propose a caveat that consists of adapting pre-trained models on-the-fly to work in new situations.

From there, and in order to keep improving generalization, we resort to deep learning methods. During the last years the level of maturity associated to scene understanding problems has risen, thanks to the adoption of deep learning tools, proving that learning internal representations is a better practice than the use of hand-crafted features. In fact, Deep learning method have started to achieve super-human performance at tasks like image classification [14]. One of the most notorious results is the Deep Residual Networks [15], a deep learning based method that achieved 3.57% of error in the ILSVRC 2015 image classification task, surpassing

human capabilities for general image classification. In the context of driving scenes, another deep learning based approach achieved super-human performance at traffic sign recognition [17], achieving a 99.44% of accuracy. These results are an example of the maturity of image classification and object recognition problems, that have been due in part to the creation of large datasets, such as ImageNet [6], Microsoft COCO [25] and the blooming of deep learning [22]. However, the state of other problems like semantic segmentation are yet far from producing results of this quality.

One of the main causes for this has to do with the need of massive datasets required by deep learning; datasets that are not available for driving domains. We lack of equivalents to ImageNet and MS COCO. The process of acquiring a driving dataset is extremely challenging, since it requires to prepare a robotic platform equipped with different sensors such as lasers, cameras, IMUs, etc. All of them must be correctly calibrated (intrinsically and extrinsically) in order to know the relationship between different sensors. In addition to the complexity and cost associated to the development of the robotic platform for data acquisition, one needs to wait for suitable weather conditions for data collection. When collecting the dataset it is usually hard to sample corner cases, since these do not happen very often. In the end, it is easy to end up having a collection of redundant data. After the acquisition, data has to be annotated for the different tasks we would like to address. For semantic segmentation and similar pixelwise-output based problems, this task is extremely tedious and expensive. When dealing with a reduced number of classes, *e.g.* 12 categories, usually takes about 45 minutes. In addition to the large amount of time required to perform the annotation process and the high cost, human annotators tend to introduce noise and incoherence in the annotations. These problems become more evident when the level of abstraction of the task grows, *e.g.* when dealing with road curvature, object orientations, etc. Producing accurate ground truth for these tasks is still an open challenge.

One of the main contributions of this second part is to investigate different techniques to deal with data scarcity within semantic segmentation. We deal with data multi-modality, which arises when resorting to the combination of multiple datasets that are designed for different contexts under different constraints. Dealing with this multi-modality requires the design of new training methods. We also propose to deal with data scarcity by using synthetic data. To achieve this we propose SYNTHIA, a simulation environment of driving scenarios that provides precise ground truth for many scene understanding problems automatically. We investigate the use of virtual imagery in combination to domain adaptation techniques to create models capable of operating in real driving scenarios. During the training process we also consider the constraints imposed by driving scenarios and embedded systems, which limit the size and shape of deep learning architectures.

In these situations it becomes critical to produce compact models that can run on embedded hardware with limited memory. We show how to cope with this problem by exploiting transfer learning approaches, to produce models that perform like their state-of-the-art counterparts but require a fraction of the original memory.

At the end of this part we show how the tools constructed to perform semantic segmentation can be adapted to solve the task of (semantic) change detection, therefore enabling systems to understand the high-level knowledge behind human-made changes done to the scene. We propose to use this new tool to perform fast and cost-effective map updating.

4 Transferring Semantic Models to New Domains

The accuracy of semantic segmentation approaches is highly dependent on the training set being used and drops drastically when the statistic distribution of the test image does not match the expected distribution of the training set, a situation that will irremediably occur, as for instance, when the illumination changes from daytime to dusk. This issue is of critical importance in the context of GAVs. To address this problem we propose a fast unsupervised image transformation approach following a global color transfer strategy. Our proposal generalizes classical one-to-one color transfer schemes to the more suitable one-to-many scheme. In addition, our approach can naturally deal with the temporal consistency of video streams to perform a coherent transformation. We demonstrate the benefits of our proposal in two publicly available datasets using different state-of-the-art semantic labelling frameworks.

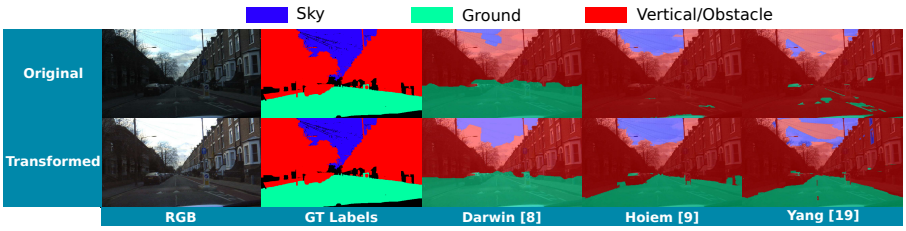


Figure 4.1 – Example results for semantic labelling on CamVid computed on the original images **(top)** and after the proposed transformation **(bottom).**

4.1 Motivation

In the context of driving scenes images are taken from a mobile platform in uncontrolled, cluttered environments, and the appearance of the objects greatly varies depending on elements such as daytime, weather, illumination and acquisition conditions [35]. Algorithms must be robust to the highly dynamic nature of the environment and, due to the time constraints of the applications, they require to be computationally efficient to achieve real-time capabilities. In this chapter, we focus on making semantic segmentation algorithms more tolerant to the aforementioned variations of the scene in an efficient and unsupervised fashion.

During the last decade, a number of approaches have focused on providing accurate semantic labelling [16, 23, 43, 46]. These approaches exploit the continuously increasing available data to learn more comprehensive models, able to deal with as many situations as possible. Unfortunately, dealing with general urban scenes involves a large number of different situations, making difficult to foresee a dataset large-enough (*i.e.*, complete) to cover all possible conditions. Furthermore, it is clearly infeasible to collect and learn in a single model all possible situations. As a consequence, in practise, classifiers are trained on relatively small datasets covering a limited number of conditions and provide promising results when they are evaluated on data from a similar distribution. However, as shown in Fig. 4.1 (**top**), there is a significant decrease in performance when these algorithms are applied to images coming from a different distribution. A common approach to solve this problem consists of retraining the classifier with labelled instances from each new environment (domain). More recent works deal with this problem adapting the classifier to each new domain transferring the knowledge by domain adaptation techniques [7, 38, 41]. However, these approaches require prior knowledge of the target domain and the collection of labelled instances which is time consuming and impractical for autonomous driving applications, where the domain changes dynamically over time.

In contrast to these methods, here, we propose an efficient unsupervised image transformation method following a global color transfer strategy (Fig. 4.1 (**bottom**)). In this way, test images are dynamically transformed according to a set of reference images by imposing their characteristic colors. This operation is applied following a novel one-to-many scheme, which generalizes classical methods such as Reinhard [31]. The proposed scheme overcomes the limitation of classical techniques when applied to dynamic environments, since the quality of the results depends on how similar are the images. Unlike classical methods, our algorithm only assumes that the context of the image is known (urban scene) and, thanks to the one-to-many transformation we can model the diversity of the reference images

(*i.e.*, one showing sunny conditions, the other showing bad weather, blooming, sky saturations, etc.), to increase the chances of having a reference and a test image with similar distributions. In addition, our approach can naturally deal with the temporal consistency of video streams to perform a coherent transformation. All these elements serve to improve the performance on the side of the semantic labelling method, avoiding a cumbersome retraining stage. Our algorithm does not require prior knowledge of the domain nor labelled information and is, therefore, very suitable for vehicle driving scenarios.

To demonstrate the benefits of our approach we conduct experiments on three different semantic labelling frameworks (*i.e.*, Hoiem [16], recent work by Yang [43] and Darwin [13]), on two publicly available datasets KITTI [11] and CamVid [3], along with a set of challenging images collected from the Internet. Our extensive set of experiments shows the effectiveness of our unsupervised image transformation approach as a step of semantic labelling at a negligible computational cost.

4.2 Related Work

Domain adaptation has been widely used to adjust the data distribution of the test set to the one in the training set to improve the performance of classifiers trained on limited data [7, 38, 41]. However, any domain adaptation method (supervised or semi-supervised/unsupervised [41]) requires prior knowledge of the target domain (test images) and therefore they are unfeasible for outdoor scenes with continuously changing acquisition conditions. In contrast, what we propose is an unsupervised preprocessing step that aims at correcting global mismatch between the color statistics in test image and the training set.

Color transfer (also referred as color mapping) is the process of altering the color of an image to impose color characteristics of another one (reference). This process is widely used in cinema or photo stitching applications where the same scene with the same contents is captured in slightly different instants and potentially with different cameras [45],[4],[18]. Therefore, by using color mapping we can impose the color characteristics of one of the images (reference) to the rest. As a result, all the images share the same color statistics [9]. A forerunner work in this area is the statistical analysis proposed in [31]. This approach aims at transferring the color distribution between two images using the mean and standard deviation of each color plane. The main benefit of this algorithm is its low computational cost. However, the algorithm uses an one to one mapping and assumes both images refer to the same scene (same contents). Therefore the quality of the transformation depends on the images similarity in composition. Different extensions have been proposed to consider richer representations of the color distributions in both

images (higher order moments), with the idea of more closely resemble both distributions. For instance, Hwang *et al.*in [18] proposed a method to preserve the gradients of the distribution. Although these methods provide better representations of the image contents than using only the mean and the standard deviation it is at the expense of a high computational cost and, therefore, they are not suitable for time-critical applications such as autonomous driving. In addition, all these methods are based on the same contents similarity assumption. What we propose in the next section is an algorithm that uses a one to many mapping to relax the contents similarity restriction. To this end, the algorithm creates a diversified set of reference images that will be used as a reference to transfer colors using first order statistics to keep a low computational cost.

4.3 The Proposed Approach

Our proposal of unsupervised image transformation is designed to run in real-time on images acquired in new environments, in which the visual condition might be severely different from those images used during the training stage of the semantic labelling framework. The two main criteria guiding the design of our approach are summarized as follows. First, image transformation must be transparent for semantic labelling methods and therefore these two processes must be decoupled. Second, image transformation must be very efficient, i.e., real-time capable, in order to be useful for urban semantic labelling. From these key points we propose a simple but effective pipeline consisting of two stages: **(i)** dictionary creation, where a set of representative images with enough variability are selected from a reference domain; and **(ii)** unsupervised image transformation, where a new image is first reconstructed on the reference dictionary and then the outcome of this transformation is used to perform color transfer based on a variation of Reinhard's method [31]. Both stages are detailed in the following subsections.

4.3.1 Dictionary creation

This initial stage creates a visual dictionary D_{ref} of well-known reference images. Novel images will be matched against this dictionary in order to perform the color correction. The construction of the dictionary is an offline process carried out just once.

The process starts by evaluating the full dataset in terms of image quality as proposed in [42], thus filtering out images of inappropriate illumination and selecting a subset of N images. Each of these N RGB images I^i_{RGB} is smoothed via a Gaussian filter of width $w = 5$ and $\sigma = 0.7$ to produce $\tilde{I}^i_{RGB}$. Then, each $\tilde{I}^i_{RGB}$ is mapped onto

the Lab color space [12] to perform an equalization of the lightness channel and to create a histogram H^i of $B = 300$ bins (100 bins per channel). The purpose of this preprocessing step is to account for some variations in the illumination and color of new images during the matching process against the reference set. We want to avoid as much as possible that two images representing places with similar objects (e.g., red buildings, white buildings, road, etc.), end up incorrectly matched due to changes in illumination.

The remaining is selecting a subset of K representative images that will serve as our reference. The best reference database would contain images used during the training stage of the semantic labelling framework, if those are available. Otherwise, it is possible to use a different database containing images presenting similar visual conditions to those in the training stage (see section 4.4.1). The subset of K representatives is chosen to maximize the visual variability, in order to form a diverse and rich visual dictionary. This process is summarized as follows. First, the obtained histograms $\{H^i\}_{i=1}^N$ are grouped into K clusters $\{H^i_{\text{ref}}\}_{i=1}^K$ using K-Means. These clusters are considered to be a good representation of the different modes of the original database, codifying different scenarios such as: urban areas with buildings, highways, green areas, etc. Finally, the reference dictionary $D_{\text{ref}} = \{\tilde{I}^i_{RGB}, H^i_{\text{ref}})\}_{i=1}^K$ is formed using the histograms corresponding to these clusters together with their associated smoothed RGB images. Regarding the selection of the number of clusters K, our experiments show that $K \approx 15$ is a good choice for the sequences under study. This number seems to be enough to represent the basic modes of an urban environment.

4.3.2 Unsupervised Image Transformation

In this stage, the new incoming images are corrected on-the-fly, in a fully unsupervised fashion. To this end we make use of the concepts proposed by Reinhard *et al.* [31] for global color correction. However, our approach extends [31] in order to generalize the matching process between test images and reference images from one-to-one to a more convenient one-to-K (one-to-many) matching. This allows us to consider the different modes of urban scenes and leads to a more accurate transfer and an improved semantic labelling, as shown in section 4.4.

Reference Image Selection

Once the reference dictionary is built, each incoming image I^j_{test} needs to be matched against the closer representative of the reference dictionary D_{ref}. To this end we analyse the histogram H^j_{test} of I^j_{test} in terms of the histograms of $D_{\text{ref}} = \{\tilde{I}^i_{RGB}, H^i_{\text{ref}})\}_{i=1}^K$. First, H^j_{test} is computed as described in sec. 4.3.1. Then, we recon-

struct H^j_{test} as a linear combination of the dictionary vectors $\{H^i\}_{i=1}^K$. This problem is cast as an ℓ_1-regularized least squares problem with nonnegativity constraints (RLS-NN) such as:

$$W^* = \underset{W}{\operatorname{argmin}} \|D_H W - H^j_{\text{test}}\|^2_{\ell_2} + \lambda \sum_{i=1}^{K} W_i \tag{4.1}$$
$$\textbf{subject to} \quad W_i \geq 0, \quad i = 1, \ldots, K.$$

Here, $D_H \in \mathbb{R}^{B \times K}$ stands for the K histograms of D_{ref} rearranged as a matrix and $W \in \mathbb{R}^K$ is the vector of sparse nonnegative weights for the reconstructed signal H^j_{test}. λ is a tunable parameter that controls the sparsity of the solution and, in our experiments, it has been fixed to $\lambda = 0.1$ through cross validation.

The solution of (4.1) is computed using [21], which implements a solution for large-scale ℓ_1-regularized least squares problems using a truncated Newton interior-point method [20]. Truncated Newton methods have the advantage of converging in a couple of iterations, while the cost of each iteration remains efficient for large systems.

Since W is sparse and nonnegative, we can directly interpret these weights as the relevance of each reference image in the reconstruction of $\tilde{I}^j_{\text{test}}$. The most similar reference image is the one that most contributed to the reconstruction, denoted as $\tilde{I}^{j*}_{RGB}$, with $j^* = \operatorname{argmax}_j W_j$. This image is chosen as the designated reference image for the matching.

Solving this task as an RLS-NN problem is computationally fast, and has the advantage that the obtained solution is specifically designed to be a "simple" combination of a few relevant terms. This is specially interesting for dealing with the transference of images coming from a video stream.

Dealing with videos, involves taking into account temporal consistency. Given the high correlation between consecutive frames, if the image I^j_{test} is transferred to the $k-$th reference image, I^{j+1}_{test} should most likely be transferred to the $k-$th reference too. Including this constraint can be tricky in methods like K-Nearest Neighbors. In contrast, in our formulation, it is straightforward to extend (4.1) into a Temporal Consistent Transference strategy (TCT) as follows:

$$W^*_{\text{Temp}} = \underset{W}{\operatorname{argmin}} \|\widehat{D^{\mathscr{L}}_H} W - \widehat{H}^{\mathscr{L}}_{\text{test}}\|^2_{\ell_2} + \lambda \sum_{i=1}^{K} W_i \tag{4.2}$$
$$\textbf{subject to} \quad x_i \geq 0, \quad i = 1, \ldots, K.$$

The difference with respect to (4.1) is that $\widehat{D^{\mathscr{L}}_H} = [D_H|\ldots|D_H]^T \in \mathbb{R}^{B\mathscr{L} \times K}$, is the dictionary stacked $\mathscr{L}$ times, where $\mathscr{L}$ stands for the length of a chosen temporal

(a) **(b)**

Figure 4.2 – Reinhard **(a)** vs Our approach **(b)** in terms of artefacts.

window (in our experiments $\mathscr{L} = 9$). Analogously, $\widehat{H}^{\mathscr{L}}_{\text{test}} = [H^{j-\mathscr{L}/2}_{\text{test}}|\ldots|H^{j}_{\text{test}}|\ldots|H^{j+\mathscr{L}/2}_{\text{test}}]^T \in \mathbb{R}^{B\mathscr{L}}$ is the stack of the histograms of $\mathscr{L}$ consecutive images. In this way, the resulting vector of weights $W \in \mathbb{R}^K$ already encodes the best assignment for the temporal window. Moreover, it is relevant to notice that for moderate values of $\mathscr{L}$, this method remains computationally efficient.

Transferring Colors

The last part of our approach performs the actual transference of colors between the matched images $[I^{j}_{\text{test}} \leftrightarrow I^{i}_{\text{ref}}]$. This is performed by using a linear transformation in the CIE Lab perceptual color space [12], inspired by Reinhard's seminal work [31]. CIE Lab color space minimizes the correlation between channels for most natural scenes. In [31], Reinhard *et al.* define a direct color transfer technique as a linear transformation along each of the three channels:

$$\begin{aligned} L' &= (L - \mu^L_t)\frac{\sigma^L_t}{\sigma^L_r} + \mu^L_r \\ a' &= (a - \mu^a_t)\frac{\sigma^a_t}{\sigma^a_r} + \mu^a_r \\ b' &= (b - \mu^b_t)\frac{\sigma^b_t}{\sigma^b_r} + \mu^b_r, \end{aligned} \tag{4.3}$$

where $\mu^i_t, i \in L, a, b$ are the channel means and $\sigma^i_t, i \in L, a, b$ are the channel standard deviations, calculated over all pixels of the image, for the test image I^{j}_{test}.

Similarly, μ_r^i and $\sigma_r^i, i \in L, a, b$ are the same statistics for the reference image I_{ref}^k.

Reinhard's method has the advantage of being simple and extremely fast while providing a fairly good transference of the colors. However, we noticed that for some visual conditions, this approach introduces color artefacts that end up affecting the semantic labelling results. An example of these artefacts is shown in Fig. 4.2**(a)**, where the sky become reddish. This phenomenon is an hallucination of the method that occurs when the L-channel of the test image is saturated and the method attempts to reduce it. This glitch can be easily fixed by updating (4.3) as follows:

$$L' = \begin{cases} (L - \mu_t^L)\frac{\sigma_t^L}{\sigma_r^L} + \mu_r^L, & \text{if } \mu_r^L - \mu_t^L > \tau \\ L, & \text{otherwise} \end{cases} \tag{4.4}$$

where τ is a tunable threshold, which in our experiments remains as $\tau = 50$. Finally, the transferred channels (L', a', b') are mapped back to the RGB color space to produce the adapted image I_{Trans}^j. At this point the resulting image is ready to be treated by a semantic labelling method. The performance benefits of including this transference process are presented in section 4.4 in terms of quantitative and qualitative results on different state-of-the-art semantic labelling frameworks and datasets.

4.4 Experimental Analysis

In this section, we conduct several experiments to demonstrate the benefits of incorporating our image transformation method in general state-of-the-art semantic segmentation tools. More precisely, we consider two different setups: (i) the training set used to create the semantic segmentation model is available and (ii) where the training set is not available. We also test the advantages of using a one-to-K matching. Experiments are conducted on three state-of-the-art urban datasets, where the visual conditions are largely different, using three different semantic labelling frameworks: Photo Pop-up (Hoiem *et al.* [16]), a recent work by Yang *et al.* [43], and Darwin (Gould *et al.* [13]). The first two approaches are existing algorithms for recovering the layout of a scene and are used as they are with no retraining. Both approaches were originally trained with generic images from generic environments. The last method is the framework introduced in [13]. In this case, we retrain the classifier with domain specific images (urban images) using a non-overlapping subsets of images from each dataset depending on the experiment.

(a) **(b)** **(c)**

Figure 4.3 – Example of urban images used in our experiments. a) Images from CamVid [3] dataset (day on the left and dusk on the right); b) Images from KITTI [10] road-dataset and; c) Collection of images retrieved from Google.

4.4.1 Datasets & classes

Experiments are conducted on three datasets. First, the Cambridge-driving Labeled Video Database (CamVid) [3], a publicly available collection of videos captured in the UK from the perspective of a driving automobile with ground truth labels that associate each pixel with one of 32 semantic classes. This dataset is divided in four sequences, i.e., 1TP, 6R0, 16E5 and Seq05SV. The first one consists of dark images, simulating dusk conditions, while the remaining are gathered in daytime conditions, as shown in Fig. 4.3**(a)**. In our experiments we transform the 32 semantic classes into 3 general classes: Sky (S), Ground plane (G) and Vertical obstacles (V).

Second, we use the also publicly available KITTI dataset [11]. KITTI also consists of images acquired from a moving car in Karlsruhe, Germany Fig. 4.3**(b)**. The ground

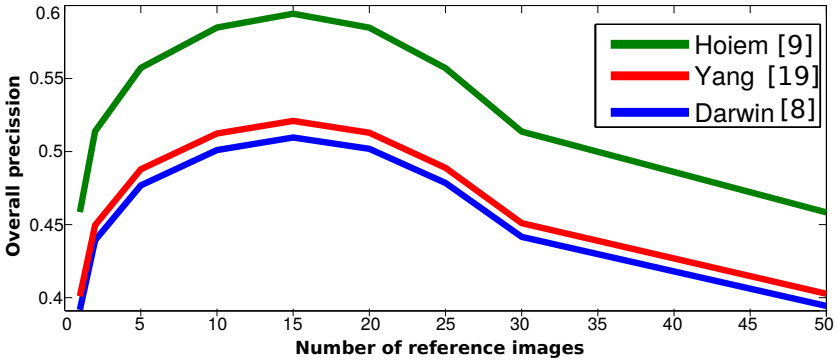


Figure 4.4 – Overall performance with respect to K, the number of reference images, on the CamVid dataset.

truth for this dataset has been generated by manually annotating the 323 images images of the KITTI-Road Benchmark [10]. Finally, we collect a set of images from Google showing particularly challenging scenarios such as night time and sunset situations as shown in Fig. 4.3**(c)**.

4.4.2 Quantitative and Qualitative Evaluation

First experiment: We start by assessing the variation in performance of the three semantic labellers, when the images are adapted following our approach. Quantitative evaluations in terms of Overall Precision (**OP**) and Per-class Average (**PC**) [23] on the KITTI and CamVid are summarized in Table 4.1 and Table 4.2, respectively.

As shown in Table 4.1, on the KITTI dataset, the impact of transferring images is not significant, even leading to some slight loss of performance for Yang's method and Darwin. This is mainly due to the good conditions of the images. However, as shown in Table 4.2, when the experiment is carried out on the CamVid dataset, the situation is quite different. For the dusk sequence (1TP) the accuracy of all three methods increases significantly, with more than 20 points of improvement for Hoiem and, 13 and 8 points for Yang and Darwin, respectively. Qualitative examples of this improvement are shown in Fig. 4.5**(a)**. For the remaining sequences, acquired in daytime, this improvement is reduced, and only Hoiem and Yang present some slight benefits. Again, this is mainly due to the already existing similarities in the color properties of the training and the testing images.

The impact of our method is now illustrated using the set of challenging images from Google Images showing night time and sunset situations. Qualitative results

Table 4.1 – Quantitative evaluations of three semantic labelling frameworks on the KITTI dataset. Adapted refers to performance when our image transformation is included in the pipeline.

	KITTI Road Dataset	
	OP	**PC**
Hoiem [16] Original	0.67	0.65
Hoiem [16] Transformed	0.67	0.65
Yang [43] Original	0.62	0.57
Yang [43] Transformed	0.61	0.57
Darwin [13] (Camvid) Original	0.67	0.62
Darwin [13] (Camvid) Transformed	0.64	0.62

Table 4.2 – Quantitative evaluations of three semantic labelling frameworks on the CamVid dataset. Adapted refers to performance when our image transformation is included in the pipeline.

	CamVid Dataset							
	1TP		6R0		16E5		05SV	
	OP	**PC**	**OP**	**PC**	**OP**	**PC**	**OP**	**PC**
Hoiem [16] Original	0.61	0.38	**0.85**	**0.65**	0.87	0.67	0.87	0.64
Hoiem [16] Transformed	**0.83**	**0.65**	0.84	0.63	**0.89**	**0.68**	0.87	0.64
Yang [43] Original	0.58	0.35	0.87	0.68	0.84	0.61	0.91	0.70
Yang [43] Transformed	**0.71**	**0.51**	0.87	0.68	**0.85**	**0.63**	0.91	**0.71**
Darwin [13] (KITTI) Original	0.68	0.61	0.84	0.77	**0.93**	**0.90**	**0.89**	**0.81**
Darwin [13] (KITTI) Transformed	**0.76**	**0.71**	0.84	0.77	0.91	0.84	0.85	0.73

of this experiment are presented in Fig. 4.5**(b)**. As shown, although the resulting labelling is far from perfect, the improvement is notorious. This is specially true for Hoiem and Darwin frameworks, which achieve the best results. From these quantitative and qualitative evaluations, we can conclude that including our image adaptation method improves the performance of semantic labellers specially in challenging situations not seen during the training stage.

Second experiment: The goal of this experiment is evaluating the influence of varying the parameter K (number of representative images) during the dictionary creation stage. To this end, we test the overall precision (OP) of the labellers with respect to different values of K on the CamVid dataset. As shown in Fig. 4.4, having several reference images for the transfer is important. Moreover, for these datasets, the optimal value is $K \approx 15$, for the three classifiers. From these results we can conclude that the proposed one-to-K transfer strategy outperforms the one-to-one

Table 4.3 – Influence of including temporal coherence (TCT) on the CamVid dataset. As shown, our approach for exploiting correlation between consecutive frames improves the performance of semantic labellers.

	CamVid Dataset							
	1TP		6R0		16E5		05SV	
	OP	**PC**	**OP**	**PC**	**OP**	**PC**	**OP**	**PC**
Hoiem [16] Transformed	0.82	0.64	0.83	0.63	0.89	0.68	0.87	0.64
Hoiem [16] Transformed-TCT	**0.83**	**0.65**	**0.84**	0.63	0.89	0.68	0.87	0.64
Darwin [13] (KITTI) Transformed	0.75	0.71	0.76	0.65	0.90	0.81	0.85	0.73
Darwin [13] (KITTI) Transformed-TCT	**0.76**	0.71	**0.84**	**0.77**	**0.91**	**0.84**	0.85	0.73

strategy by Reinhard.

Third experiment: The goal of this experiment is analyzing the impact of exploiting the temporal consistency in the transfer (TCT) process when dealing with video streams. Table 4.3 summarizes the performance variations when TCT is included on the CamVid dataset. From this results, we can conclude that considering temporal coherence outperforms still images. This improvement is specially relevant for Darwin in sequence 6R0, increasing the accuracy in 8 points.

4.5 Conclusions

In this chapter we have presented an unsupervised image transformation approach to improve outdoor semantic labelling in urban scenarios, addressing the problem of images coming from different color distributions. Our approach follows a global color transfer strategy based on a novel one-to-many matching process, which also considers temporal information present in video streams. We have demonstrated that incorporating the algorithm as a preprocessing step improves the performance of several state-of-the-art semantic labelling frameworks in publicly available challenging datasets.

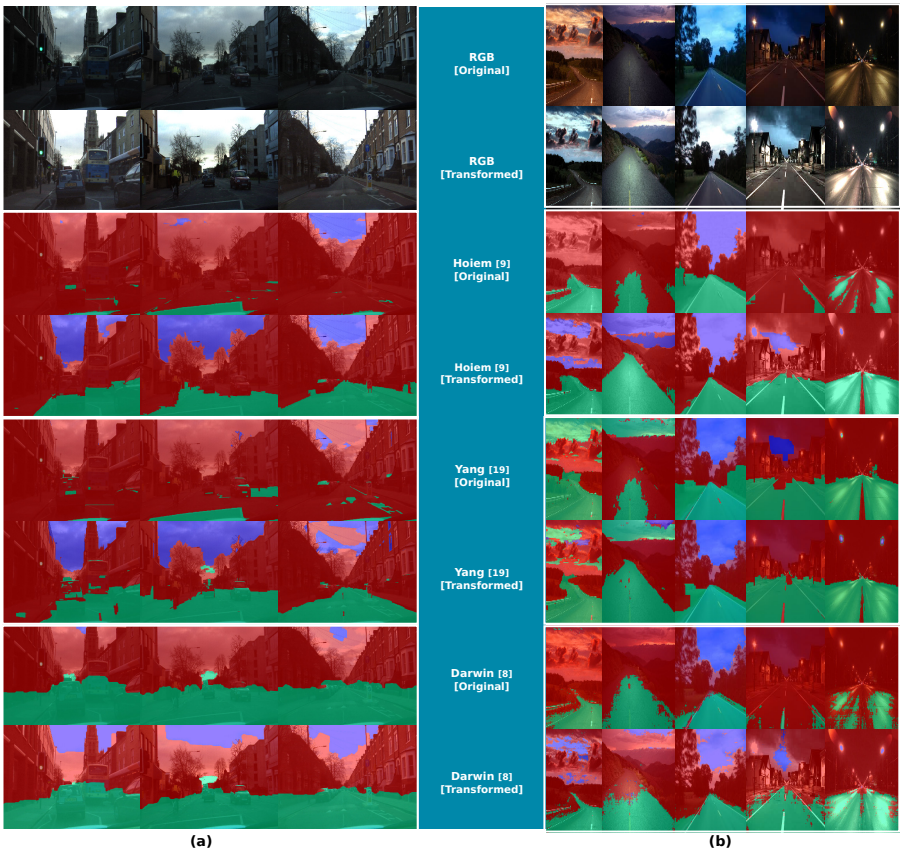


Figure 4.5 – **(a)** Qualitative results on the CamVid dusk dataset (1TP). As shown, labelling results are significantly better when images are adapted using the proposed method. **(b)** Qualitative results on the set of challenging images retrieved from Google. As shown, labelling results are significantly better when images are adapted using the proposed method.

Clausula Part III

Pictorial representation of the SYNTHIA virtual world.

5 Conclusions and Future work

5.1 Conclusions

In this PhD dissertation we have addressed the problem of understanding driving scenes from vision at different levels, from the "where am I in the scene?" to the "what is in my surroundings?"; from localization and mapping to the level of understanding involved in semantic segmentation and change detection. To achieve this we developed a strong tool-chain that included tools from robust statistics, such as L1 pose-averaging; tools from information theory, such as metric embedding and compressed regression; tools from Riemmanian geometry, such as manifold representation techniques and optimization; tools from compressive sensing, such as proximity operators, phase transitions and $\ell_0 \rightarrow \ell_1$ equivalences; tools from machine learning, such as the deconvolutional neural networks, the theory of domain adaptation and transfer learning; and tools from computer graphics, such as the realistic rendering of a city. Here, we take the opportunity to summarize the findings of this work.

In the first part of this dissertation we focused on visual localization, the "where", using geometric and algebraical techniques, addressing different versions of the Visual Odometry problem. Within this broad topic, we studied new formulations to improve robustness and computational efficiency of one of the core steps of Visual Odometry: ego-pose estimation. We proposed very fast alternatives based on oblivious embeddings and Lie-groups **??**. We then extended the robustness of our efficient alternatives, by reformulating the problem as a robust manifold averaging problem on a compressed evaluation space in chapter **??**. This was needed in order to compensate for the sensibility to noise of the approaches just based on ℓ_2 embeddings, and we proved that ℓ_2 embeddings could be used to achieve a fast and robust estimator in collaboration with averaging. In chapter **??**, we decided to face the problem of ego-pose estimation considering the intrinsic constraints arising from having a stereo rig moving through time. We modelled the main camera motion using rank and sparsity constraints, applying Robust PCA and compressed sensing techniques to discard outliers and obtain robust results.

The approach behind the Robust PCA formulation was improved in chapter **??**, to decrease computational time while extending the applicability to the technique to large-scale cases (*i.e.*, where larger amounts of data are involved). To this end we proposed a new formulation of Robust PCA as a specialized optimization on a product manifold. To optimize this manifold product space, we proposed an algorithm based on novel closed-form projectors with outstanding computational efficiency. We show that, in addition to ego-pose estimation, Robust PCA-like methods can be applied to a wide range of problems where outliers may be present and that our proposed approach has better theoretical and practical properties.

In the second part of this dissertation we focus on semantic segmentation and related problems, caring about "what" is in the scene. In chapter **??** we propose an automatic method to build maps that include rich semantic information of the scene as an offline approach. Then during an online stage visual localization is used as a proxy to recover the semantics associated to the static scene, which are further extended by detecting dynamic objects. This approach attempts to use the semantic map as a strong prior to speed the process of understanding up, leveraging computational resources to perform localization and detection of dynamic objects in real-time. However, the use of maps as strong priors brings several issues. First, using maps limits the level of autonomy of a vehicle, since it can only operate in those regions that haven been previously mapped. Secondly, those maps can be rendered useless very quickly, due to changes in the city. In addition, fast map updating is a complex and expensive process.

We deal with the first of these problems in chapter **??**, trying to become map independent by proposing a novel architecture for real-time semantic segmentation based on deep learning principles. This effort is extended in chapter **??**, where we address data scarcity for semantic segmentation and related scene understanding problems by proposing a new synthetic platform called SYNTHIA. Synthetic data is combined with domain adaptation techniques, showing a cost-effective philosophy to produce accurate models usable in real domains. The topic of the model adaptation to different domains is also covered in chapter 4, where we proposed an unsupervised method to adapt semantic segmentation models to drastic illumination changes.

In chapter **??**, we address fast map updating by reducing the bandwidth of the information to update, using change detection techniques. We derived a (semantic) change detection architecture out of our semantic segmentation model to produce a system capable of detecting and segmenting structural changes of the city due to man-made upgrades while ignoring visual seasonal changes.

The intention behind studying and developing methods related to localization and mapping—which use maps as strong priors—,and methods that are map-free—as the case of real-time semantic segmentation—, is to bring perception

redundancy to GAVs. We can conceive the combination of easy-to-update maps and real-time prior-less systems as the combination of two perception sensors. Such a combination would dramatically increase the reliability of autonomous cars, giving GAVs the power to operate in many different scenarios. For this reason we believe that the proposed approaches would have an important impact on the practical deployment of GAVs.

5.2 Future Perspective

From an academic point of view, driving scene understanding and its associated sub-tasks, such as semantic segmentation and change detection, have matured in a short period of time. As presented in this thesis, new models are showing very promising generalization capabilities and accurate results, when exposed to new unseen scenes. This has been possible due to the progress done in the machine learning techniques that are currently used to address these tasks and by the arrival of new and more challenging datasets such as SYNTHIA. However, there is yet a long path to explore in order to fulfil the expectations behind driving scene understanding.

First, we would like to rise a concern that has been around for a while: data scarcity and the data-versus-accuracy ratio. Currently, state-of-the-art techniques based on deep learning require an enormous amount of data with associated annotations. Semi-supervised and unsupervised alternatives place themselves as promising alternatives in a close future, but nowadays they can not compete against the results provided by supervised methods. The current trend of embracing synthetic data is an important step forward in order to address more sophisticated problems using cost-effective solutions. Nevertheless, the generation of the data (and its associated annotations) is still a process decoupled from the learning algorithm. In other words, first the data is generated (usually by a third party) and then a learning method is adapted to make use of the data. A large proportion of that data is usually redundant and therefore plays no important role during the learning process. This leads to a situation in which much of the learning time is wasted re-exploring already assimilated cases, while other critical cases are being ignored. These facts lead to the logical conclusion that a data generator should be at the core of learning pipelines. Data should be generated on demand according to the current requirements of the model (for instance analysing validation error). Following this philosophy, standard learning procedures would become active learning procedures. How to bring this concept to practice is something that we are currently exploring.

Related to the previous point, we would like to highlight the need of exploring

new problem representations in order to avoid unnecessary human-derived biases. This is probably one of the most relevant points discovered during the realization of this thesis. As previously described, we observed that slightly different ways of representing the ground truth information could lead to very different models, presenting a large gap in recognition accuracy. In other words, given a learning algorithm there are representations of a problem that are more suitable than others. This phenomenon seems to be occasioned by ambiguity introduced by humans and human biases. A possible approach to solve this issue could be to study the internal representations found in models trained to solve decision-making problems (*e.g.* , making decisions on which action perform next from a list of possible actions) [2], instead of understanding problems. To solve decision-making problems in an accurate fashion, agents need to acquire a good notion of how the scene is represented, in an indirect way. Could we use the information associated to the scene representation to avoid our biases?

A second important issue is to design new neural architectures specialized for scene understanding problems, *i.e.*, including the constraints present on a given domain. We have already seen important improvements at this regard, with proposals such as the Recurrent Instance Segmentation [33] and the Dilated Convolutions [44]. In both cases, architectures are tailored around the main flaw of semantic segmentation, sub-segmentation due to a limited receptive field. At the light of the promising results offered by these approaches we would like to dedicate further efforts on effective ways of including domain constraints into deep learning architectures.

We must also consider the gap between academically acceptable solutions and industrially acceptable solutions; a gap that is wide and still hard to bridge. In order validate a scene understanding system to achieve the level of quality requested by industry standards we may need to re-think some of the current assumptions. Current scene understanding models may not offer the required accuracy when moving to different geographic locations, *i.e.*, they could suffer from generalization insufficiency. However, a practical alternative may be to accept that fact and compensate it by applying cost-effective domain adaptation for each new location. This alternative requires to keep exploiting and developing effective domain adaptation techniques for the different scene understanding problems, but would help to move these systems to production in a much shorter time-frame.

These, among many others, are the current open problems in scene understanding for driving scenarios. We hope that this summary and discussion could serve to motivate researchers to take some of these challenges, with the final goal of bringing autonomous driving a step closer.

5.3 Contributions

In this PhD dissertation we have made both practical and theoretical contributions to autonomous driving from the point of view of visual localization-and-mapping techniques and visual scene understanding. In the scope of visual localization and mapping we have studied and propose novel approaches for the following problems:

- Fast and accurate compressed regression for pose-estimation in Visual Odometry
- Robust optimization methods on Lie-groups and robust manifold averaging for Visual Odometry
- New Manifold formulation as a triple direct product of Stiefel×SPD×Stiefel manifolds
- Robust PCA and fast optimization on the Low-rank-Sparse manifold
- Addressing Stereo Visual Odometry as a Low-rank-Sparse manifold

During the scene understanding part of this dissertation we have dealt with the problematic behind semantic segmentation of driving scenarios and (semantic) change detection. To this end we have propose new architectures based on deep deconvolutional networks along with new training methods and datasets. These contributions are summarized as follow:

- Efficient offline-online pipeline to perform real-time scene understanding via localization and retrieval
- Novel training methods to improve deconvolutional neural network in the task of semantic segmentation of driving scenes
- Novel training methods to improve change detection and semantic change detection results using deconvolutional neural networks

We have also put a special emphasis on practical problems when applying semantic segmentation in real scenarios, proposing new approaches based on unsupervised transfer learning and domain adaptation. To complement these results, we have also provided a method to compress state-of-the-art deconvolutional neural network to run on embedded devices. These contributions are summarized as follow:

- Efficient unsupervised transfer-learning approach to adapt semantic segmentation models on-the-fly to new illumination conditions
- Novel domain adaptation methods to improve accuracy when transferring from Virtual-to-Real scenes
- Novel training methods to deal with data multi-modality in driving semantic segmentation
- A novel technique to perform compression of devoncolutional neural networks to use them in embedded contexts

5.4 Patents

- Network compression via Transfer-Learning and Optimization methods for Embedded Contexts and Autonomous Vehicles. Filed 2016, Toshiba Research Corporation, main intellectual author.

5.5 Scientific Articles

This dissertation has led to the following communications:

5.5.1 Submitted Journals

- **German Ros**, Jose Alvarez and Julio Guerrero. Motion estimation via robust decomposition with constrained rank. IEEE Transactions on Intelligent Vehicles, 2016
- **German Ros**, Simon Stent, Pablo F. Alcantarilla, and Tomoki Watanabe. Training constrained deconvolutional networks for road scene semantic segmentation. IEEE Transactions on Intelligent Transportation Systems, 2016
- **German Ros**, Julio Guerrero, Angel Sappa, Daniel Ponsa, and Antonio Lopez. Fast and robust fixed-rank matrix recovery. International Journal on Computer Vision, 2016

5.5.2 Book Chapters

- **German Ros**, Laura Sellart, Gabriel Villalonga, Elias Maidanik, Francisco Molero, Marc Garcia, Adriana Cedeo, Francisco Perez, Didier Ramirez, Eduardo Escobar, David Vazquez, and Antonio M. Lopez. Semantic Segmenta-

tion of Urban Scenes via Domain Adaptation of SYNTHIA. Srpinger editorial, Domain Adaptation for Computer Vision Applications, 2016

5.5.3 International Conferences and Workshops

- **German Ros**, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proc. of Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA (**short-oral**), 2016.

- Pablo Alcantarilla, Simon Stent, **German Ros**, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. In Proc. of Robotics: Science and Systems (RSS). Michigan, USA, (**nominated to best systems paper award**), 2016

- **German Ros** and Jose Alvarez. Unsupervised image transformation for outdoor semantic labelling. In Proc. of IEEE Intelligent Vehicles Symposium (IV). Seoul, Korea, 2015

- Alejandro Gonzalez, Gabriel Villalonga, **German Ros**, David Vazquez, and Antonio Lopez. 3D-guided multiscale sliding window for pedestrian detection. In Proc. Iberian Conference on Pattern Recognition and Image Analysis (Ibpria). Santiago de Compostela, Spain, 2015

- **German Ros**, Sebastian Ramos, Manuel Granados, Amir H. Bakhtiary, David Vazquez, and Antonio Lopez. Vision-based offline-online paradigm for autonomous driving. In Proc. IEEE Winter Conference on Applications of Computer Vision (WACV). Hawaii, USA, 2015

- **German Ros**, Julio Guerrero, Angel Sappa, Daniel Ponsa, and Antonio Lopez. Fast and robust l1-averaging-based pose estimation for driving scenarios. In Proc. British Machine Vision Conference (BMVC)., Bristol, UK, 2013

- **German Ros**, Julio Guerrero, Angel Sappa, Daniel Ponsa, and Antonio Lopez. VSLAM pose initialization via Lie-groups and Lie-algebras optimization. In Proc. IEEE International Conference on Robotics and Automation (ICRA). Karlsruhe, Germany, 2013

- **German Ros**, Angel Sappa, Daniel Ponsa, and Antonio Lopez. Visual slam for driverless cars: A brief survey. In Proc. IEEE Workshop on Navigation, Perception, Accurate Positioning and Mapping for Intelligent Vehicles. Alcala de Henares, Spain, 2012

5.6 Contributed Code and Datasets

- **SYNTHIA virtual dataset**: A virtual dataset for driving scene understanding task, with focus on semantic segmentation, instance segmentation, change detection and localization and mapping. http://synthia-dataset.net
- **The Multi-Domain Road Scene Semantic Segmentation (MDRS3)**: a dataset for semantic segmentation of urban scenes where multiple domains are combined extended and improved from public datasets. http://www.toshiba.eu/eu/Cambridge-Research-Laboratory
- **The KITTI Semantic Dataset**: A collection of semantic annotations for part of the KITTI Visual Odometry dataset. http://adas.cvc.uab.es/s2uad
- **MatConvNet-DeconvNet**: A branch of the MatConvNet deeplearning framework, improved and specialized to work with deconvolutional neural networks for several problems, such as semantic segmentation, optical flow, depth estimation, pose estimation, etc. https://github.com/germanRos/MatConvNet-DeconvNet
- **Chainer-deconv**: A branch of the popular Chainer deep learning framework, improved and specialized to work with deconvolutional neural networks for several problems, such as semantic segmentation, change detection, depth estimation and network compression. https://github.com/germanRos/chainer-deconv
- **Yet Another Semantic Segmentator in 3D (YASS3D)**: A general semantic segmentation framework over 3D point-clouds, using 3D features and standard linear SVM and CRFs. https://github.com/germanRos/YASS3D
- **Fixed-Rank Alternated Direction Method with Augmented Lagrange Multiplier (FRADM)**: A general algorithm to perform Robust PCA via Manifold optimization with a new formulation on a product manifold. https://github.com/germanRos/FRADM
- **Robust L1-Averaging for Visual Odometry**: A method to perform robust and very fast model averaging on the $\mathbb{SE}(3)$ manifold using compressed regression with application to Visual Odometry. https://github.com/germanRos/l1avgvo
- **Ego-motion $\mathbb{SE}(3)$ optimization by Lie-group optimization (LieOpt)**: A method to perform compressed optimization on Lie-groups using algebraical cost functions, with application to pose estimation. https://github.com/dehabu/Lieopt

5.7 Scientific Dissemination

5.7.1 Invited Talks

- *Exploiting Virtual Worlds and Domain Adaptation for Driving Scene Understanding*, at Zoox, Palo Alto, USA, 2016
- *The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes*, International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, 2016
- *Semantic Segmentation for Driving Scenarios: On virtual worlds*, Industrial Robotics Institute, CSIC-UPC, Barcelona, Spain, 2016
- *Semantic Segmentation for Driving Scenarios: On Virtual Worlds and Embedded Platforms*, at Oxford University, Engineering Dept., Torr's Group, Oxford, UK, 2016
- *On Practical Semantic Segmentation for Autonomous Driving,* at NVIDIA Research, New Jersey, USA, 2016.
- *On Scene Understanding on Virtual Worlds*, at Xerox Research Center Europe, Grenoble, France, 2016
- *Vision-based offline-online paradigm for autonomous driving*, at the IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii, USA, 2015
- *Robust Matrix Decomposition with Fixed-Rank Constraint*, at UC Louvain in the ICTEAM Seminars in Mathematical Engineering, Belgium, 2014
- *On 3D Semantic Maps for Vision-based Offline-Online Autonomous Driving*, at NICTA, Canberra Research Lab, Canberra, Australia, 2013
- *Autonomous driving: when 3D mapping meets semantics*, at the Workshop of Computer Vision in Vehicle Technology: From Earth to Mars, in conjunction with the IEEE International Conference on Computer Vision, Sydney, Australia, 2013
- *3D Scene Understanding*, at the Department of Informatics and Systems, Universidad de Zaragoza,, Zaragoza, Spain, 2013

- *Visual slam for driverless cars: A brief survey*, at the IEEE Workshop on Navigation, Perception, Accurate Positioning and Mapping for Intelligent Vehicles, in conjunction with the Intelligent Vehicles Symposium, Alcala de Henares, Spain, 2012

5.7.2 Demos

- *SYNTHIA meets Virtual KITTI* at the International Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016

5.7.3 In the Media

- *Meet Synthia, the virtual driving school for autonomous cars*, Gizmag, June, 2016
- *Customizing your car is about to get more futuristic*, CNET.com, July, 2016
- *SYNTHIA is a massive virtual city where autonomous vehicles can safely learn how to drive*, Digitaltrends, June, 2016
- *Welcome to SYNTHIA City: Virtual world created so AI cars can learn to drive*, Daily Mail Online, June, 2016
- *Sim City Created for Self-Driving Cars*, Seeker, June, 2016
- *Investigadores de la UAB crean un videojuego para los coches autónomos*, La Vanguardia, Spain, June, 2016
- *Investigadores de Barcelona crean un videojuego para acelerar la llegada de los coches autónomos*, El Mundo, Spain, June, 2016
- *El videojuego para que los coches autónomos aprendan*, ABC, Spain, June, 2016
- *Mira, màquina*, magazine El Temps, Barcelona, July, 2016

Bibliography

[1] James R. Bergen and Edward H. Adelson. Early vision and texture perception. *Nature*, 333, 1988.

[2] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

[3] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 2008.

[4] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] B. Clipp, Jongwoo Lim, J.-M. Frahm, and M. Pollefeys. Parallel, real-time visual slam. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, oct. 2010.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Lixin Duan, Ivor W. Tsang, and Dong Xu. Domain transfer multiple kernel learning. *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, 34:465–479, 2012.

[8] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localisation and mapping (SLAM): Part i the essential algorithms. *IEEE Journal on Machine Robot and Applications*, 13(2), June 2006.

[9] H. S. Faridul, T. Pouli, C. Chamaret, J. Stauder, A. Tremeau, and E. Reinhard. A survey of color mapping and its applications. *Eurographics-State of the Art Reports*, pages 43–67, 2014.

[10] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research*, 32:1231–1237, 2013.

[12] R.C. Gonzalez and R.E. Woods. *Digital Image Processing (2nd Edition)*, chapter Section 10.4. Prentice Hall, 2002.

[13] Stephen Gould. Darwin: A framework for machine learning and computer vision research and development. *Journal of Machine Learning Research (JMLR)*, Dec 2012.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

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

[16] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *Intern. Journal of Computer Vision (IJCV)*, 75(1):151–172, 2007.

[17] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013.

[18] Y. B. Hwang, J. Y. Lee, I. S. Kwon, and S. J. Kim. Color transfer using probabilistic moving least squares. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[19] B. Kitt, A. Geiger, and H. Lategahn. Visual odometry based on stereo image sequences with RANSAC-based outlier rejection scheme. In *Proceedings of the IEEE IV Symposium*, San Diego, CA, USA, 2010.

[20] Kwangmoo Koh, Seung jean Kim, Stephen Boyd, and Yi Lin. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research*, 2007, 2007.

[21] Kwangmoo Koh, Seungjean Kim, and Stephen Boyd. l1 ls: A matlab solver for large-scale l1-regularized least squares problems. available from: https://web.stanford.edu/ boyd/l1_ls/, 2007.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[23] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Associative hierarchical random fields. *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, 2013.

[24] J. Levinson and S. Thrun. Robust vehicle localization in urban environments using probabilistic maps. In *Proc. IEEE Int. Conf. Robot. Automat.*, May 2010.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, 2014.

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] Thorsten Luettel, Michael Himmelsbach, Michael Manz, Andre Mueller, Felix von Hundelshausen, and Hans-Joachim Wuensche. Combining Multiple Robot Behaviors for Complex Off-Road Missions. In *Proc. IEEE Int. Conf. Intell. Transp. Systems*, Washington, DC, USA, October 2011.

[28] Christopher Mei, Gabe Sibley, Mark Cummins, Paul Newman, and Ian Reid. Rslam: A system for large-scale mapping in constant-time using stereo. *Int. J. Comput. Vision*, 94, Sep 2011.

[29] Michael Montemerlo, Jan Becker, Suhrid Bhat, Hendrik Dahlkamp, Dmitri Dolgov, Scott Ettinger, Dirk Haehnel, Tim Hilden, Gabe Hoffmann, Burkhard Huhnke, Doug Johnston, Stefan Klumpp, Dirk Langer, Anthony Levandowski, Jesse Levinson, Julien Marcil, David Orenstein, Johannes Paefgen, Isaac Penny, Anna Petrovskaya, Mike Pflueger, Ganymed Stanek, David Stavens, Antone Vogt, and Sebastian Thrun. Junior: The stanford entry in the urban challenge. *J. Field Robot.*, 25, Sep 2008.

[30] P. Newman, D. Cole, and K. Ho. Outdoor slam using visual appearance and laser ranging. In *Proc. IEEE Int. Conf. Robot. Automat.*, pages 1180 –1187, may 2006.

[31] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE CGA special issue on Applied Perception*, 21(5):34–41, 2001.

[32] Diego Rodriguez-Losada, Fernando Matia, Agustin Jimenez, and Ramon Galan. Consistency improvement for slamekf for indoor environments. In *Proc. IEEE Int. Conf. Robot. Automat.*, May 2006.

[33] Bernardino Romera-Paredes and Philip H. S. Torr. Recurrent instance segmentation. *CoRR*, abs/1511.08250, 2015.

[34] German Ros, Angel D. Sappa, Daniel Ponsa, and Antonio M. Lopez. Visual slam for driverless cars: A brief survey. In *Proc. IEEE Intelligent Vehicles Symposium; Workshop of Navigation, Perception, Accurate Positioning and Mapping for Intelligent Vehicles*, Alcalá de Henares, Spain, June 2012.

[35] T. Scharwachter, M. Enzweiler, U. Franke, and S. Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *European Conference on Computer Vision (ECCV)*, 2014.

[36] Gabe Sibley, Christopher Mei, Ian Reid, and Paul Newman. Vast-scale outdoor navigation using adaptive relative bundle adjustment. *Int. J. Robot. Res.*, 29, Jul 2010.

[37] Randall Smith, Matthew Self, and Peter Cheeseman. A stochastic map for uncertain spatial relationships. In *Proc. 4th Int. Symp. Robot. Research*, Cambridge, MA, USA, 1988. MIT Press.

[38] D. Vázquez, A.M. López, J. Marín, D. Ponsa, and D. Gerónimo. Virtual and real world adaptation for pedestrian detection. *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, 2013.

[39] Wikipedia. Autonomous car, 2016. [Online; accessed 22-August-2016].

[40] L.V. Woensel and G. Archer. Ten technologies which could change our lives. Technical report, EPRS - European Parlimentary Research Service, January 2015.

[41] J. Xu, S. Ramos, D. Vázquez, and A. M. López. Domain adaptation of deformable part-based models. *Transactions of Pattern Recognition and Machine Analyses (PAMI)*, 2014.

[42] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[43] Jimei Yang, Brian Price, Scott Cohen, and Ming-Hsuan Yang. Context-driven scene parsing with attention to rare classes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[44] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning representations (ICLR)*, 2016.

[45] L. Yuan and J. Sun. Automatic exposure correction of consumer photographs. In *European Conference on Computer Vision (ECCV)*, 2012.

[46] Chenxi Zhang, Liang Wang, and Ruigang Yang. Semantic segmentation of urban scenes using dense depth maps. In *European Conference on Computer Vision (ECCV)*, pages 708–721, 2010.