THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE

PREPAREE A L'UNIVERSITE DE BOURGOGNE

Ecole doctorale n° 37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Instrumentation et informatique de l'image

Par

Blanchon Marc

Polarization based urban scenes understanding

Thèse présentée et soutenue à Le Creusot, le 01 Octobre 2021

Composition du Jury :

CANU Stéphane	Professeur à Université de Rouen	Rapporteur
VASSEUR Pascal	Professeur à Université de Picardie Jules Verne	Rapporteur
AINOUZ Samia	Professeure à Université de Rouen	Examinateur
CHANUSSOT Jocelyn	Professeur à Université Grenoble Alpes	Président
MERIAUDEAU Fabrice	Professeur à Université Bourgogne Franche Comté	Directeur de thèse
SIDIBE Désiré	Professeur à Université Evry	Codirecteur de thèse
MOREL Olivier	MCF à Université Bourgogne Franche Comté	Codirecteur de thèse

Université Bourgogne Franche-Comté
École Doctorale SPIM

Doctoral Thesis

Polarization based urban scenes understanding

Author:
Marc Blanchon

Supervisors:
Prof. Fabrice Meriaudeau
Dr. Olivier Morel
Prof. Désiré Sidibé

Examiners:

Rapporteur - Stéphane Canu, Professor at Université de Rouen - LITIS
Rapporteur - Pascal Vasseur, Professor at Université de Picardie Jules Verne - MIS
Examinateur - Samia Ainouz, Professor at Université de Rouen - LITIS
Président - Jocelyn Chanussot, Professor at Université Grenoble Alpes - GIPSA-Lab

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Instrumentation and image processing
(Instrumentation et informatique de l'image)

in the

VIBOT EMR CNRS 6000
ImViA, EA 7535
Université de Bourgogne

October 7, 2021

Declaration of Authorship

I, Marc Blanchon, declare that this thesis titled, "Polarization based urban scenes understanding" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: **08/07/2021**

Date:

UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
ÉCOLE DOCTORALE SPIM

Abstract

VIBOT EMR CNRS 6000

ImViA, EA 7535

Université de Bourgogne

Doctor of Philosophy

Polarization based urban scenes understanding

by Marc Blanchon

Humans possess an innate ability to interpret scenes under any condition. Computer Vision tends to mimic these capabilities by implementing intelligent algorithms to address complex understanding problems. In this regard, we are interested in understanding outdoor urban scenes in various weather conditions. This thesis specifically addresses the problems arising from the presence of specularity in the scenes. To this end, we aim to take advantage of polarization indices to define such surfaces in addition to traditional objects. In terms of understanding, we aim to introduce polarization to the fields of computer vision and deep learning.

This thesis focuses on the following underlying challenges. First, the estimation of a semantic segmentation at the pixel level is investigated. We exploit polarization cues to define constraints upstream of the convolutional network and thus inject specularity understanding into the model. As DCNNs are data intensive, we propose the acquisition of a multimodal dataset allowing the comparison of the proposed method with RGB-centric methods. Moreover, to counteract the massive need for data, we establish a procedure to augment the polarimetric informations while maintaining the physical integrity of the information. In a second line of research, we address the problem of depth map estimation with a monocular image. Since the algorithms require a colorimetric information, we adapt the processes to an alternative type of imagery. This results in novel regularization terms that allow to accurately infer a depth map from a unique polarimetric image using deep learning. Constrained by the greedy aspect of DL, we build a loss function in accordance with the self-supervision principle. In this manner, we demonstrate the possibility to regularize the depth inference process using terms constraining the normals by relying on polarization. This approach allows us to reconstruct more accurately surfaces observing specular behavior or transparency phenomena.

Ultimately, our two lines of research show advances towards a more conventional use of polarization in modern computer vision.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Fabrice Meriaudeau, Désiré Sidibé, and Olivier Morel, for their great help and inspiration during my Ph.D. study. The experience gained through their guidance is invaluable and their accessibility has supported me enormously throughout these years. I am very grateful to Désiré, my first principal supervisor, for his patience, dedication, motivation, and knowledge that were instrumental to the success of this Ph.D. I valued his friendship, trust and his ability to understand me. I am very grateful to Fabrice, my main supervisor in a second time, for his involvement, his consideration, his trust and his rigor. He has been a huge support, bringing his numerous knowledge, advising and motivating me constantly. His friendship, his accessibility and his value of work have been undeniably driving forces towards the achievement of this thesis. I would also like to convey my sincere thanks to Olivier for his expertise, insightful comments, and valuable advice in all our exchanges. His professionalism and critical eye have undoubtedly led to the completion of this research work. I also address warm thanks to Ralph Seulin for his technical support, availability, and his great capacity of adaptation which largely counted in the success of this Ph.D.

I would also like to thank the members of the jury for their kindness in participating in my Ph.D. defense committee. Many thanks to Prof. Stéphane Canu and Prof. Pascal Vasseur for their careful reading of the thesis and their constructive remarks, and to Prof. Jocelyn Chanussot and Prof. Samia Ainouz for the investment of their valuable time to be both examiners of this thesis. Their involvement has undoubtedly broadened the understanding of the thesis and added new perspectives on our research problems.

Thanks to my host laboratory ImViA - EA 7535, to its director Franck Marzani as well as to all the collaborators of the CORES and IFTIM teams with whom I was able to share seminars, presentations and defenses which were instructive and which pushed me to enlarge my scientific culture.

I would also like to thank all my colleagues of the VIBOT team. During these three years, they allowed me to grow as a person and as a researcher. They have welcomed me with great kindness and have allowed me to progress. Cédric Demonceaux, Christophe Langlais, Christophe Stolz, David Fofi, Eric Fauvet, Lew Lew Yan Voon, Olivier Laligant, Nathalie Choffay, Olivier Aubreton, Omar Tahri, Raphael Duverne, I would like to thank all of you, each one of you, for all that you have brought to me individually or collectively. It has been a pleasure to work with you in a "human-sized" team and I am sure we will stay in touch. Some of you have known me for 8 years and have followed my evolution. I want you to know that these years have been enriching.

I would especially like to thank David and Omar, my friends, who taught me so much, were available at every moment and showed great patience. David, thank you for all the opportunities, for your constant support and for your kindness. I would also like to thank all my fellow Ph.D. Students with whom I have shared many moments and with whom I have enjoyed collaborating. Especially, Thibault and Daniel, together since the beginning, you have been a great support and I will keep excellent memories together, in the car for the events or talking about anything and everything for a break. Of course I don't forget David Strubel who was there at the beginning of my thesis and with whom we still exchange long hours on the phone, thank you. I would like my friendship with you to be repeated with all the people I could meet in the laboratory. All of you, colleagues, friends, I am not particularly expressive, but know that you all helped me without exception. Each one of you makes me who I am today and I thank you warmly.

Lastly, my thanks go to my family for their support and love through the trials and tribulations that have spaced the path to this thesis. You have supported me and had confidence in me through all these years. Also, I would like to express my love to my girlfriend Nathalie who has, without counting the cost, supported me and listened to me, helped me in many ways and especially to face the difficult moments. Their support has allowed me to move forward and finally accomplish my ambitions.

To my grandfather...

To my beloved mom...

Contents

List of Figures

List of Tables

List of Abbreviations

ASPP	**A**trous **S**patial **P**yramid **P**ooling
CEL	**C**ross **E**ntropy **L**oss
DCNN	**D**eep **C**onvolutional **N**eural **N**etwork
DL	**D**eep **L**earning
DoFP	**D**ivision **o**f **F**ocal **P**lane
HSL	**H**ue **S**aturation **L**uminance
IoU	**I**ntersection **o**ver **U**nion
MD	**M**ono**d**epeth
ML	**M**achine **L**earning
P2D	**P**olarimetry **2**(to) **D**epth
PwSS	**P**ixel–**W**ise **S**emantic **S**egmentation
RGB	**R**ed **G**reen **B**lue
SDI	**S**ørensen-**D**ice **I**ndex
SfX	**S**hape-**f**rom-**X**

Chapter 1

Introduction

1.1 Context and Motivation

For the past sixty years, computer vision has been one of the main areas of research towards intelligent machines. With the ambition of making machines "more human", many works tend to allow machines to see and understand similarly to the cognitive capabilities of humans. This interest took on even more meaning when learning applications became technologically accessible. The advent of Machine Learning (ML) methods has made it possible to develop increased cognitive abilities applicable to computers and thus to popularize high level domains like scene understanding.

In this thesis we address the problem of understanding scenes but, in addition, we focus on the use of an unconventional modality: polarimetry. Starting from the postulate that the vast majority of efficient methods depend on RGB images, it seems attractive to try to exploit the knowledge acquired in the RGB domain to export it and adapt it to new image modalities more suitable to certain situations. In this perspective, this thesis aims to use polarization to obtain a better scene understanding. From a global point of view, there are numerous approaches to scene understanding: classification, semantic segmentation, 3D reconstruction, etc. We will particularly tackle the tasks of segmentation and reconstruction.

Firstly, the semantic segmentation task corresponds to a pixel wise classification of an image. Allowing labels to be assigned at the pixel level as opposed to the image level for standard classification, this area focuses on the semantic aspect. Indeed, the overall objective is to differentiate but also to recognize. The uses of semantic segmentation are diverse and allow, for example, the accurate recognition of humans in an urban environment or the clear delineation of roads (all this in an autonomous vehicle context). Early approaches such as [110] allowed the delimitation of the domain as well as its distinction from other approaches of understanding such as classification [1], instance segmentation [44] etc. Despite a good number of pure image-processing methods, nowadays and in recent years, many methods are based on Deep Learning (DL) and Deep Convolutional Neural Networks (DCNNs). Indeed, the increased

learning capabilities and the approaches have evolved and significantly outperformed previously proposed methods. This being stated, the vast majority of contributions on semantic segmentation is now DL-based. Requiring a massive amount of data as well as significant computing power, learning architectures allow the translation from an image space to a feature space. As a result, networks are capable to learn from data redundancy, the attributes necessary for semantic segmentation. Traditionally, information provided to these processing modules are RGB images allowing an almost direct discrimination in the color space. However, it is notable that the constraints related to the modality are exported in the network. As follows, the events to which the modality implemented is not sensitive are by extension exported in the DCNN capabilities. Many methods have been developed to increase the abstraction capabilities of the networks to overcome data drawbacks. Networks have become deeper and deeper, the number of parameters has increased, features have become more abstract and as a consequence, the requirements in terms of data and computational capacity have increased. In this context of semantic segmentation, our intention was to introduce the possibility of using polarimetric information as a data source, but also to study whether the change of information could limit the use of greedy methods. Instead of relying only on the network to learn feature abstraction, the idea is to propose alternative approaches based on modality changes to reduce the complexity of such process by injecting prior knowledge and constraints.

In a second stage, the reconstruction of 3D scenes represents a radically different domain. The objective is to build a depth map corresponding to the distance between the lens and the objects in the scene. This sub-domain of computer vision is operated for multiple purposes ranging from augmented reality to autonomous navigation and surface analysis. Ultimately, this area of the vision is extremely demanding in terms of resources and pre-requisites. Whether it is a Shape from Motion method, stereovision or multiple view geometry, each of these approaches, despite their efficiency, are made complex by their acquisition constraints. More recently, novel approaches are based on monocular vision and the use of DCNN to estimate depth. In an end-to-end manner, the goal is to be able to infer from a simple image a precise depth map. Thus, the more robust these methods are, the less we rely on ground truths drawn by LiDaRs that become ineffective during weather alterations. Moreover, as these approaches have evolved, the constraints of supervision have been freed to finally obtain completely self-supervised methods. This has allowed a popularization of the field since it was no longer necessary to maintain an annotated database. Moreover, since depth annotation is generated by simple tools that can sometimes fail, some errors could be reduced by generalizing cost functions. Based on the statement of perspective geometry, the vast majority of approaches are RGB-centric since the visual

data are usual and straightforward (color imaging). This is also explained by the almost unique use of a few datasets like Kitti [56] and CityScapes [36], which are only available in RGB. As a consequence, despite similar formulations, non-conventional modalities have not been treated in this field probably due to lack of data. Regarding depth map estimation, our goal in this thesis is to incorporate polarization in state-of-the-art methods to take advantage of this image modality.

The overall objective being to operate these algorithms in complex urban areas and/or with altered weather conditions, the polarimetric modality is an excellent data candidate. Briefly, since polarimetry is by definition a modality sensitive to reflection phenomena and urban areas are prone to specularity (cars, windows, rain, puddles, etc.), polarization-driven processing units could show increased capabilities. Therefore, the global intention of this thesis is to merge polarization and current computational tools to benefit from the advantages of both parties. Thus, by moving away from RGB-centric methods, it would be possible to show improved approaches by the specific modality and therefore obtain segmentation and reconstruction methods robust to specularity and weather changes.

We strongly believe that polarization and its discriminative capabilities could improve many computer vision applications domains. Combining the knowledge acquired over time with this new data could provide new insights to some problems requiring physical understanding. In order to undertake a first step towards these polarization-centric systems, we propose to investigate two fundamental areas of computer vision: segmentation and scene reconstruction. These problems have been chosen as initiators to bring polarization closer to the actual computer vision domain. To such a degree, we aim at popularizing approaches based on polarization or any other physics-based vision by demonstrating the usefulness of such a component to traditional approaches.

1.2 ICUB Project

This thesis is founded by the French Agence Nationale de la Recherche (ANR) and is part of the project named *No conventional imagery for secure urban mobility* (ICUB) (ANR-17-CE22-0011). This project gathers two university structures namely University Bourgogne Franche Comté and INSA Rouen as well as two industrial actors Stellantis and StereoLabs. The overall concept is to establish methods that are robust to adverse weather conditions. Indeed, autonomous vehicles are subject to many phenomena that reduce or alter visibility. To counter these occurrences, we assume that the characterization of light reflection, i.e. polarization, may be a candidate to improve existing methods and thus be able to navigate safely using weather invariant

algorithms. Leveraging polarization information, the project is divided into two distinct axes, each of which is conducted by one of the two research teams. Thus, the different approaches are intended to be brought together to finally obtain a complete joint estimation pipeline using polarization to understand urban scenes in adverse weather conditions. LITIS in Rouen addresses detection problems by means of classification or bounding box estimation. ImViA's EMR VIBOT team at Université Bourgogne Franche Comté aims at building a scene understanding pipeline using segmentation or depth estimation at the pixel level. Finally, the combination of these approaches allows a complete multi-layer characterization of urban scenes.

1.3 Scope and Challenges

This thesis addresses the problem of understanding urban scenes exploiting polarization. Noting the vast majority of algorithms are RGB-centric and suffer from inabilities to categorize certain phenomena, we propose polarization as a discriminant factor to obtain an accurate understanding of urban scenes. By priorly characterizing the scenes in a different space, it is then possible to consider the problems according to various attributes. Consequently, we have investigated the possibilities of understanding urban scenes using polarization cues. Specifically, we explored different solutions like learning-based segmentation and deep learning depth reconstruction from a single view system.

1.3.1 Pixel-wise semantic segmentation

The problem of semantic segmentation has been established for a long time. Indeed, it is a question of differentiating the various regions of interest of an image. To address this problem, many methods have been developed. Whether they are image processing techniques, defining attractive frames and features to delimit the diverse objects of the scenes, or learning-based methods, all have demonstrated the ability of machines to "understand" these particular contexts. Most of these methods rely on the differentiation of textures through RGB-centric systems. This practice assumes texture is sufficiently discriminant to establish a robust segmentation. However, remarkably few methods consider phenomena like specularity. They are often neglected because they are not present in the databases. Indeed, these surfaces have the particularity to display textureless or saturation behavior. One can say that, traditionally, these issues are avoided by the absence of such cases in the images. Since the methods do not observe such occurrences, they are uncharacterized. Our method does not rely on the textural aspect of the surfaces but rather considers the interaction of the surfaces with light. It then becomes necessary to completely rebuild the segmentation pipeline to move from an RGB-centric to a polarization-centric approach. Consequently, we

address the problem of acquiring a reliable database including specular areas traditionally avoided. Recently, the use of deep learning approaches has increased the ability to describe visual scenes. We therefore propose implementing this similar kind of approach and, we investigate the possibility and viability of operating such models with physics-based vision. And since these greedy algorithms require a significant amount of image, we investigate the possibility to increase the size of the database using augmentation processes. In order not to alter the physical and crucial information to the characterization of specularity, we design transformations adapted to infer realistic and physically intact images.

Finally, we propose a complete pipeline, from data acquisition to pre-processing and learning for pixel-wise semantic segmentation.

1.3.2 Monocular depth inference

To understand the scenes, one can estimate the distance between objects and camera. Hence, many methods have emerged to reconstruct 3D scenes. Traditionally, from a monocular system, knowledge of the camera movement is required to estimate the coordinate of 3D points. Recent approaches have shown that, by using learning based approaches, many constraints can be avoided. For example, monodepth approaches have proven it is possible to infer a depth map from a static image. Similar to the field of segmentation, these methods are exclusively RGB-centric due to the availability of databases required for such algorithms. As a consequence, some visual phenomena are ignored due to the lack of understanding of these occurrences by the operated modality. Once again, the specularity being however omnipresent in urban areas, it is ignored and directly impacts the genericity of the algorithms. Therefore, we propose evaluating the possibilities of using polarization as a source modality for 3D reconstruction. With the objective of characterizing both diffuse and specular areas, we aim at establishing a set of constraints allowing to regularize the inference of a depth map. By infusing polarization cues into the deep learning model, we therefore seek a robust specularity-invariant algorithm, which will reconstruct urban scenes accurately. To this end, a novel depth estimation method named P2D is introduced to include these aspects in the monodepth domain. P2D considers both geometric and polarimetric cues to address the depth map estimation problem. The regularization terms can be further exploited to derive innovative approaches and move towards more and more robust methods.

1.4 Contributions

The next section details each work highlighting contributions and associated publications as author [14, 15, 16, 13, 9] or co-author [172, 173].

1.4.1 Peer-review publications

Pixel-wise Semantic Segmentation & Augmentation. The first major contribution of our work is presented in Chapter 4. Bringing together segmentation and augmentation, the chapter describes the scope of the research conducted to introduce polarimetry in the field of deep learning and also, more specifically, Pixel-wise semantic segmentation (PwSS). We therefore propose a first of its kind approach allowing to use polarization cues in segmentation which sets the state of the art of the domain. We also demonstrate the usefulness of such information by comparing it with approaches focused on color imaging. To operate these algorithms, we needed a set of polarization compatible transforms. We subsequently proposed a set of possible augmentations and their corresponding regularization.

Our contributions in this area are primarily motivated by a desire for more recurrent use of polarization in modern computer vision. We then aimed at demonstrating that the characterization of light interactions in a scene is sufficient to accurately determine object classes. Based on the observation that a majority of objects in urban environments are subject to reflection, we then defined a suitable representation to both differentiate objects and depict polarimetry faithfully. Due to a lack of data, a multimodal dataset was acquired allowing both training and a fair comparison with similar algorithms on other data. The data being never sufficient, we have also designed augmentation processes allowing to obtain new images respecting the physical integrity of the modality. Ultimately, a comparative study allowed us to estimate that polarization has an advantage over colorization for the segmentation of urban scenes.

ASSOCIATED PUBLICATIONS (SEGMENTATION):

- "Outdoor Scenes Pixel-wise Semantic Segmentation using Polarimetry and Fully Convolutional Network"
  Marc Blanchon, Olivier Morel, Yifei Zhang, Ralph Seulin, Nathan Crombez, Désiré Sidibé
  VISAPP 2019 - [14]

- "Utilisation de la polarimétrie pour la segmentation de scènes extérieures avec un réseau convolutif"
  Marc Blanchon, Olivier Morel, Yifei Zhang, Ralph Seulin, Nathan Crombez, Désiré Sidibé
  ORASIS 2019 - [15]

- "Exploration of Deep Learning-based Multimodal Fusion for Semantic Road Scene Segmentation."
  Yifei Zhang, Olivier Morel, **Marc Blanchon**, Ralph Seulin, Mojdeh Rastgoo, Désiré Sidibé
  VISAPP 2019 & ORASIS 2019 - [172, 173]

ASSOCIATED PUBLICATION (AUGMENTATION):

- "Polarimetric image augmentation"
  Marc Blanchon, Olivier Morel, Fabrice Meriaudeau, Ralph Seulin, Désiré Sidibé
  ICPR 2020 - [13]

ASSOCIATED COURSE:

- "Introduction to Polarization for Rendering and Vision"
  Kai Berger, **Marc Blanchon** - equal contribution
  SIGGRAPH Asia 2020 Courses - [9]

Monocular depth estimation using polarization cues. The second line of work, presented in Chapter 5, corresponds to the estimation of a depth map from a single polarimetric image. In this chapter we present, to our knowledge, the first approach inferring a depth map from polarization cues using deep learning. We examine the possibilities of loss term regularization to improve the depth maps traditionally deduced from RGB-centric algorithms. As follows, we aim to constrain the problem by formalizing an approach based on the relationship between surface normals and polarization. Since there is no available data, we also propose a polarimetric dataset of dynamic urban areas under different weather conditions. By this way, we aim to categorize the specular areas commonly neglected. Ultimately, we propose improvement possibilities since it has been found that our initial approach is not fully generic. This heads us to design different multimodal fusion methods that could be evaluated and should be made viable.

ASSOCIATED PUBLICATION:

- "P2D: a self-supervised method for depth estimation from polarimetry."
  Marc Blanchon, Désiré Sidibé, Olivier Morel, Ralph Seulin, Daniel Braun,

Fabrice Meriaudeau
ICPR 2020 - [16]

Towards scene understanding through polarization cues. Since we have proposed a number of methods to characterize urban scenes through polarization, we propose integrating all these works into a scene descriptor pipeline. The effort has been produced beforehand to force the use of single view systems in all cases of algorithm establishment. Thus, the fusion of these approaches is possible, starting from a polarimetric unique view, to infer a multidimensional descriptive image composed of segmentation, depth and polarization indices.

ASSOCIATED PUBLICATION:

- "Towards urban scenes understanding through polarization cues"
  Marc Blanchon, Désiré Sidibé, Olivier Morel, Ralph Seulin, Fabrice Meriaudeau
  Autonomous Robots - [17] - Submitted

1.4.2 Open source softwares / datasets

With the objective of promoting the dissemination and use of our algorithms and this particular data, we offer all of our work in open source:

- **Interpol** - A comprehensive list of interpolation method for polarization. Provides an integrated comparison tool.
  https://github.com/BlanchonMarc/InterPol
- **Pola_NewtonPolynomial** - Demosaicking DoFP images using Newton's polynomial interpolation python adaptation of the initial shared matlab code.
  https://github.com/BlanchonMarc/Pola_NewtonPolynomial
- **P_Augmentor** - Augmentation toolbox for polarization. Offer multimodal augmentation possibilities with transformation coherency.
  https://github.com/BlanchonMarc/P_Augmentor
- **AcquisitionFromTopics** - Multimodal synchronized acquisition through ROS.
  https://github.com/BlanchonMarc/Ros_AcquisitionFromTopics
- **PolaBot** - Multimodal RGB / NIR / Polarimetric dataset with segmentation annotation.
  https://vibot.cnrs.fr/polabot.html

- **Dense Alignment Toolbox** - A toolbox allowing for dense multimodal alignement
  https://github.com/BlanchonMarc/process-vibotorch/tree/master/Alignment
- **Vibotorch** - A pytorch wrapper allowing for reproduction of results for the segmentation part of this thesis. Embeding metrics, dataset management, etc.
  https://github.com/BlanchonMarc/vibotorch
- **Segmentation and P2D models**
  Models available on demand
- **P2D training and testing algorithms**
  Source code available on demand
- **Urban scenes under different weather conditions through polarization**
  Dataset available on demand

1.5 Organization

This dissertation is divided into the following chapters:
Chapter 2 introduces multiple concepts of either polarization and deep learning to avoid redundancies along this manuscript.
Chapter 3 provides a comprehensive overview of the related works by assessing both the learning-based segmentation field and the depth estimation domain.
Chapter 4 proposes solutions to segment complex urban scenes by using both polarization and deep learning. Due to the constraining framework, this Chapter also presents the respective dependencies of the algorithms such as: dataset construction, alignment, augmentation etc.
In addition, Chapter 4 provides a comprehensive range of evaluations for the different segmentation propositions.
Chapter 5 explores how to infer depth from a monocular polarimetric image. We propose a first-of-its-kind deep learning-based algorithm using polarization cues to derive depth maps. Additionally, capitalizing on the drawbacks of the previous method, we propose a comprehensive evaluation of fusion methods as well as a first step towards RGB and polarization fusion for accurate depth refinement.
Chapter 6 gives the final discussion of this thesis and ideas for future work on the presented problems.

Chapter 2

Background

This chapter is dedicated to the two main tools that will be widely used in this manuscript. To avoid redundancy and repetitive explanations, we propose defining Polarization in Section 2.1 and Deep Learning in Section 2.2. Thus, when general principles are needed for the understanding of the work, this chapter will serve as a reference.

2.1 Polarimetry

This section will be dedicated to polarimetry and will provide a general introduction to the field. We propose to describe this modality by defining it in three aspects. First, Section 2.1.1 introduce the general concept of polarization. Then, Section 2.1.2 will discuss the properties specific to polarimetric imaging from the sensor point of view. Finally, Section 2.1.3 will discuss the usual exploitation methods of this particular data.

2.1.1 General Principles

Polarimetry [34] is a particular modality that acquires the polarization state of objects in a scene. Briefly, polarization is a property that, in the case of image acquisition, concerns light. It is composed of two perpendicular waves called electric field $\vec{E}$ and magnetic field $\vec{B}$ that oscillate along the wave reflected to the sensor. A polarized wave is said to be elliptical but is ordinarily considered as the sum of linear and circular components. In general, in mobile acquisition systems, only the linear polarization is acquired since the circular polarization requires the mounting of a quarter wave plate and because, in nature, polarization is linear. It is possible to observe polarized light in nature with sunlight or observation of multiple non-natural light sources. A notable property is that if a light wave hits a surface and is reflected, then the wave becomes partially polarized. This property is particularly important since, unlike conventional modalities that focus on colors or textures, polarimetry reveals light behavior related to surfaces by using vectorial aspect of the light.

Consequently, it is possible to affirm that polarimetry measures changes in the state of light [154]. In addition, many principles are inspired or derived from Fresnel's equations [51]. This link is attributed to the close relationship between polarization and reflection of light. Therefore, it is notable that polarization has the potential capacity to infer, by light behavior, the properties of surfaces (i.e. refractive index, surface normals, etc.).

2.1.2 A Particular Sensor

Initially, polarization images were acquired using a manual or motorized polarizer rotated in front of the camera. This acquisition process was difficultly reliable in real conditions and was limiting the applications. Consequently, such cameras were conventionally used in controlled environment requiring lower frame rate. Nowadays, modern sensors allow acquisition of such data in real conditions with convenient frame rate. Indeed, it could be compared to RGB if we consider the Bayer matrix. As shown in Figure 2.1, a polarimetric camera contains a micro-grid of polarizers that allow the acquisition of different light orientations.

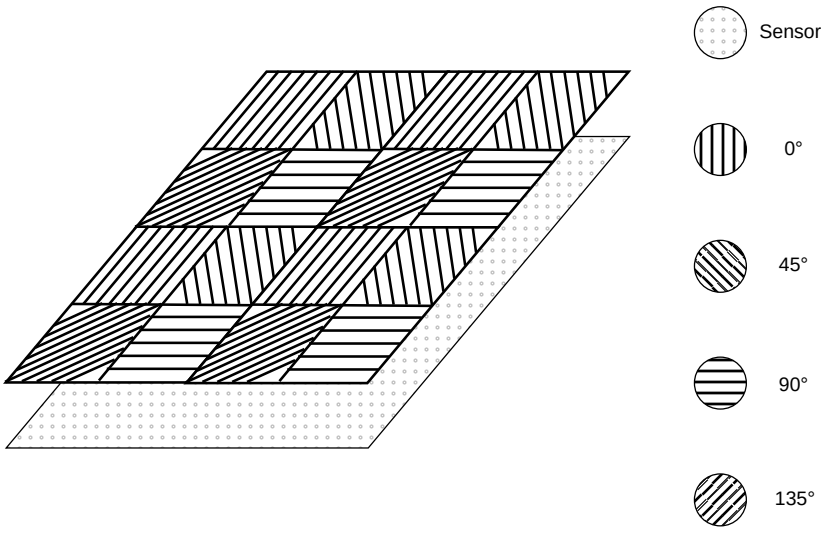


FIGURE 2.1: Illustration of a DoFP polarization sensor micro-grid.

It is then possible to observe a multitude of mini polarizers with four different orientations affixed to the sensor. This technology is called Division of Focal Plane (DoFP). With the DoFP, it is possible to acquire all the necessary information in a single shot. Hence, the sensors could be embedded to make dynamic acquisitions. Standardly, there are four angles $\{0, 45, 90, 135\}$ allowing calculations which will be discussed below. These different angles allow discriminating the components of the light and thus to separate the information orientation-wise. As schematized in Figure 2.2, thanks to these different orientations, the light can be filtered.

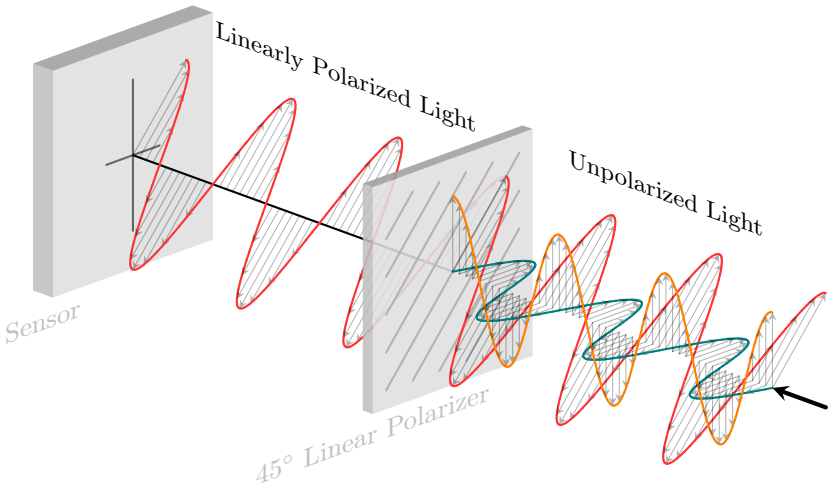


FIGURE 2.2: Illustration of unpolarized light filtration through 45° linear polarizer.

This filtering is reproduced with the different orientations of each micro-polarizer and thus allows the acquisition of a multitude of linearly polarized light components. Due to this sensor architecture, each pixel is independent since the information is influenced by the polarizer. Consequently, the intensity for each polarizer $P_{\{0,45,95,135\}}$ is "independent" and especially sparse. As shown in Figure 2.3, it is possible to recognize the microgrid by zooming in on an image.

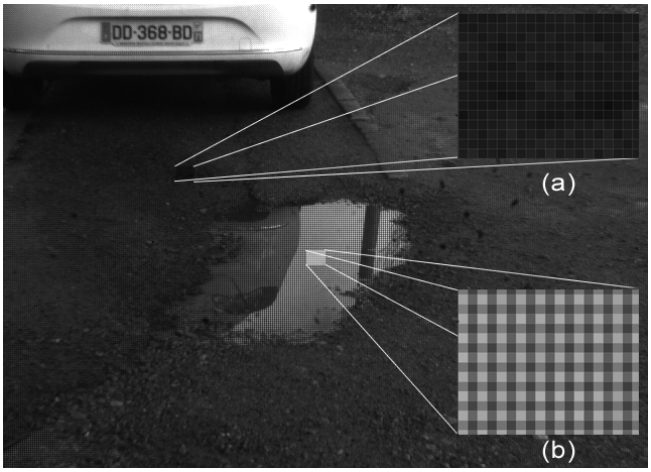


FIGURE 2.3: Zoom on a polarimetric image. (a) is a zoom on the non-polarized area, (b) is on a polarized area. Micro-grid effect can be observed on polarized area revealing the sensor architecture due to filtering.

Subsequently, it is possible to identify polarized areas using this pixelization phenomenon. Indeed, when an object reflects the light, the grid appears since the intensities per polarizer are different (consequence of the filtering).

2.1.3 Exploiting the Data

As previously stated, the intensities are sparse because of the sensor architecture. The images are acquired with a low resolution sensor, this can cause a significant problem since the image size is divided by two. In addition, on a low resolution, the shapes can be changed because of the correspondence between the real world and image space. Ultimately, it is necessary to acquire dense and aligned polarizer intensities otherwise the polarimetric information is incomplete.

To overcome this dimensionality problem, many approaches have been developed that allow the interpolation of polarimetric intensities. Principally used, Ratliff et al. [123] proposes to use a bilinear interpolation directly on the intensity images. More recently, more complex interpolation strategies have been investigated using the Newton polynomial [87] or the use of machine learning models based on sparse representation [170]. Despite the complexity of these algorithms, the best way to overcome the image sparsity problem is to operate a high-resolution camera with smaller pixels giving a more adequate real-world correspondence to image space. The density of the polarization images is crucial to calculate the Stokes parameters [138]. These parameters have been designed to describe the polarization state of light through a descriptive vector such as:

$$S = \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix} = \begin{pmatrix} P_H + P_V \\ P_H - P_V \\ P_{45} - P_{135} \\ P_R - P_L \end{pmatrix} = \begin{pmatrix} P_0 + P_{90} \\ P_0 - P_{90} \\ P_{45} - P_{135} \\ 0 \end{pmatrix} = \begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix}, \tag{2.1}$$

where P_H and P_V are respectively the horizontal and vertical polarization while P_R and P_L are respectively the power of right and left circularly polarized light. Since we do not acquire circular polarization, V remains null. I represents the total acquired intensity whereas U and Q are part of $L = Q + iU$ the straight polarization intensity, being a complex number that accounts for the tilt of the polarization direction θ.

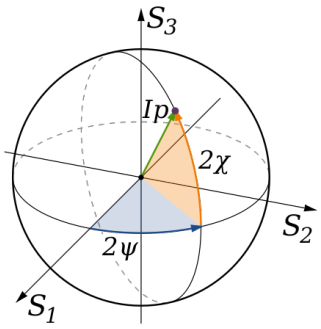


FIGURE 2.4: Spherical representation of Stokes vector on Poincaré Sphere.

As stated in Section 2.1.1, the polarization is said to be elliptical. Common representation projects the vector onto Poincaré sphere. Consequently, Stokes parameters can also be defined through 3 spherical coordinates I_p, 2ψ and 2χ shown in Figure 2.4. Equally, the parameters are:

$$\begin{cases} S_0 = I_p \\ S_1 = \rho I_p \cos 2\psi \cos 2\chi \\ S_2 = \rho I_p \sin 2\psi \cos 2\chi \\ S_3 = \rho I_p \sin 2\chi \end{cases}, \tag{2.2}$$

where I_p is the intensity of the beam, ρ the degree of polarization, ψ and ξ are defining factor for an ellipse being π and $\frac{\pi}{2}$ invariant. Consequently, one can define the intensity ι, the angle of polarization α as well as the degree of polarization ρ following (with no circular polarization acquired):

$$\begin{cases} \iota = S_0 = I_p \\ \alpha = \frac{1}{2} \arctan \frac{S_2}{S_1} = \psi \\ \rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0} \end{cases}. \tag{2.3}$$

The degree ρ and angle α of polarization are two very descriptive characteristics of polarization. ρ is analogous to the polarization strength and belongs to $[0, 1]$. This parameter quantifies the polarization light in a wave. Therefore, a completely specular wave will record $\rho = 1$.

$\alpha \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$ is the angle of the electric field $\vec{E}$ projected onto the image plane with respect to the reference. In other words, the polarization angle corresponds to the orientation of the polarization with regard to the incident plane.

2.1.4 Summary

This section reviews the fundamental principles of polarization imaging. It showed how this modality is particular and requires a singular processing. Indeed, we have first stated the particularities of the acquisition of the polarization state of light. This allowed us to establish a direct link between the surface and the light reflected by it. Next, we described the functioning of the sensors and their limits. To conclude, we have described general process making raw polarimetric images exploitable. This brief overview has, in addition to introducing the subject and establishing it, justified in large part the use of this particular information in our work.

2.2 Deep Learning

This section will be dedicated to the main notions related to Deep Learning. The recent advances in the field of learning and the increase in computing power have allowed this sub-branch of machine learning to emerge. Moreover, the availability of data, which is crucial for this kind of greedy algorithm, has allowed to guarantee a certain viability to these models. All in all, DL has become an inner part of the computer vision field by imposing itself as a powerful processing core for many approaches.

We propose detailing this area in several parts for reference throughout this manuscript. Starting with the general notions, we will describe the four basics in Section 2.2.1, namely data, network, loss and training. These generalities will then allow in Section 2.2.2 to deepen some key concepts for this thesis.

2.2.1 Basics

We propose a brief insight in the field of Deep Learning by defining the three main principles of the domain.

2.2.1.1 Data

Data is the critical point of greedy algorithms and specifically DL. Indeed, it is presumably the most sensitive subject in this field. Basically, the learning algorithm "feeds" on the data to learn how to optimize towards an objective set by the loss. In order for this task to be executed correctly, a coherent cost function is needed. But it is notable that, despite the consistency of the loss, if a database is unsuitable, then the learning will be unsuccessful. This observation leads to rules that broadly apply to any dataset. The data must be *unbiased*, *sufficient in number* and *representative*. Therefore, one must have a large number of images, and they must be suitable for the problem that is being addressed. As for the bias, it implies there must be a sufficient diversity to guarantee a robust learning.

FIGURE 2.5: Sample of MNIST Dataset [83]. It has been designed for handwritten digit recognition.

> Similar to a child who would be taught to recognize leaves. The teacher shows the child a leaf and describes him what it is. The child will not have enough examples to identify a leaf. *This is a quantity problem.*
>
> If a teacher instructs a student to identify a considerable number of things and rewards the student for each accurate answer. In all that was shown, there are many things but very few leaves. Even if they are present, they are largely in the minority. The teacher subsequently invites the student to describe each thing he identifies in front of him and will tend to neglect the leaves. *This is a problem of representativeness.*
>
> Ultimately, if the teacher shows the student green leaves. In everything he has learned, he has only seen green leaves and nothing else of that color. Then the student will say every green thing is a leaf. In another case, if the teacher shows a brown leaf, then the student will say it is not a leaf. *This is a bias problem.*
>
> In the end, this teacher/student example is an excellent analogy of Deep Learning that relies entirely on data. To such a degree, the teacher represents the loss, the student, the network. The data is the amount of information that the teacher has shown to the student. And the training is analogous to the framework defined by the teacher.

These three situations show how crucial data is. A significant number of responsibly designed RGB databases have been made available, including ImageNet [40] containing 14 million images, CityScapes [36] 25,000 images, MNIST [83] 60,000 images (shown in Figure 2.5), etc. As a result, the scientific community has been effective to address recurring computer vision issues through common information and benchmarks. These represent reliable utilities since there is a massive amount of data available.

One point that remains important to address is augmentation. Indeed, some tasks do

not maintain enough data to be learned, or it is necessary to counter bias in the data. In these cases, the traditional practice remains the augmentation which allows to avoid overfitting [114] or to obtain a consistent data. This established practice allows applying transformations to ensure the invariance of the models to certain phenomena such as rotation, flipping, etc. It consists in the application of operations resulting in photorealistic images. In such ways, the network will be efficient to learn from created data that have not been acquired by a sensor.

To conclude on the data, we allow ourselves some open questions justifying a majority of the work conducted during this thesis. Are the widely used datasets representative of all the problems to which they are attached? Are they representative enough to answer faithfully to the problems through colorimetry? And, is there a learning bias triggered by the use of colorization specifically when the modality is not sensitive to certain physical phenomena?

2.2.1.2 Network

Briefly, the network can be considered as the brain of the algorithm. It is a dynamic structure, composed of layers, which infers from the data a result from the product of its weights. A network is said to be deep if it includes at least one hidden layer, i.e. at least three layers. The trained weight dictionary is commonly defined as a model.

We will focus this section on Deep Convolutional Neural Network (DCNN) since it is the most widespread in the field of Computer Vision.

The DCNN concept is based on convolutional layers that act in the same way as standard convolution with a filtering operation. In short, from the convolution of a filter with an image derives an activation map named feature map. Convolutional layers sustain two key advantages compared to dense layers. One, convolutions use very few parameters in comparison since this structure forces the sharing of weights over the entire input. Two, they allow, by nature, to be position invariant, unlike the linear operations of dense layers which require priors on the neighboring pixels. In short, without knowledge of the data and interconnections between pixels, the convolution allows with fewer parameters to extract dense feature maps by operating filters whose weights are fixed through learning. As follows, we exploit the property of this operation to rely on the neighboring pixels to extract the activation map. Ultimately, a network is a succession of layers allowing the extraction of information without having to empirically fix the kernel weights. Not to forget an important property of convolutions, the dimensionality of the input information is reduced proportionally to the size of the kernels.

The networks are not only composed of convolution layers, but also of activations, sampling, pooling, etc. which will be addressed explicitly in Section 2.2.2 for the units used in the presented work.

In conclusion, the principle of the network is to transpose a source image into a desired space through the extraction of features regressed by the loss.

2.2.1.3 Loss

The loss is a minimizable cost function whose target is to attain an optimal objective. From a source image, one operates a forward-pass through the network resulting in a feature map. This map is evaluated through the loss function which means that the network is evaluated through this objective function. Thus, this measure influences the weights of the network through the back propagation.

Briefly on the back propagation [22], it allows to fine tune the weights of the network layers to minimize the global error. Thus, starting from the loss and using partial derivatives, the gradient goes up along the network to allow an adjustment of the layers.

To come back to the loss, it is important to note this aspect of Deep Learning has been re-evaluated in the last few years. While before, the consensus of the scientific community was to deepen DCNN, now there is a significant attraction towards coherent loss scaling. Thus, the Deep Learning competition has more or less shifted from a hardware challenge to a theoretical challenge. Nevertheless, there are some cost functions that have imposed themselves to answer some challenges. For example, the segmentation community tends to use the same regularization terms like Cross Entropy Loss or Intersection over Union [48]. In another domain, depth reconstruction rather addresses loss allowing self-supervision and building self-sufficient photometric comparison terms [58].

Indeed, there are two main types of losses. On the one hand, those that require ground truth and operate a comparison thanks to annotated datasets (the leading case of segmentation). On the other hand, the algorithms that cannot be compared to ground truth and that size the loss to avoid requiring external information (depth map inference). These two aspects explain the movement from supervision to self-supervision and the use of their respective losses. Indeed, tasks requiring supervision like classification and segmentation have been widely investigated and are almost taken for granted. On the other hand, unsupervised processes have become more and more recurrent due to the increasing complexity of the problems addressed.

2.2.2 Specific notions

2.2.2.1 Pooling

There are multiple types of pooling and two categories. The two categories are: indexed or non-indexed, while the types correspond rather to the applied operations. The pooling corresponds above all to a sampling whether it is up or down on a *kxk* windows size. The different operations commonly used are:

- **Max Pooling.** Retains the max value by eliminating the details.
- **Average Pooling.** Considers all important information and averages it.
- **Min Pooling.** Implies that "only the details count" and eliminates strong features.
- **Probabilistic Pooling.** Draws probabilities for each regions through activation normalization. Then, preserve the highest probability corresponding value.

Consequently, depending on the operations chosen, the impact on the feature maps is very different and involves various concepts. As illustrated in Figure 2.6, the result of these different approaches gives very distinct maps.

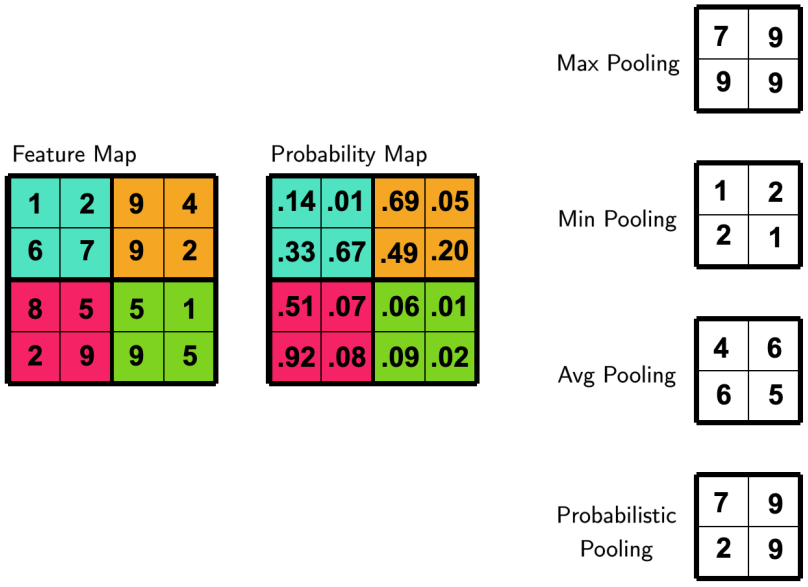


FIGURE 2.6: Illustration of different pooling strategies.

Then comes the concept of indexing. Indeed, hourglass algorithms do not only down-sample but also generally require a return to the original dimensionality. Thus, there are two approaches that imply two different behaviors:

- **Non-indexed.** Place the value in the upper left corner and fill the kernel with zeros.

- **Indexed.** Retrieves the index of the down-pooling layer to place the value at this same position. Fill the rest of the kernel with zeros.

Thus these two approaches observe respectively two different behaviors. While one considers that the spatial cue is not necessary for the integrity of the information, the other keeps the positioning and ensures its transmission. As shown in Figure 2.7, resulting maps differ due to the different techniques.

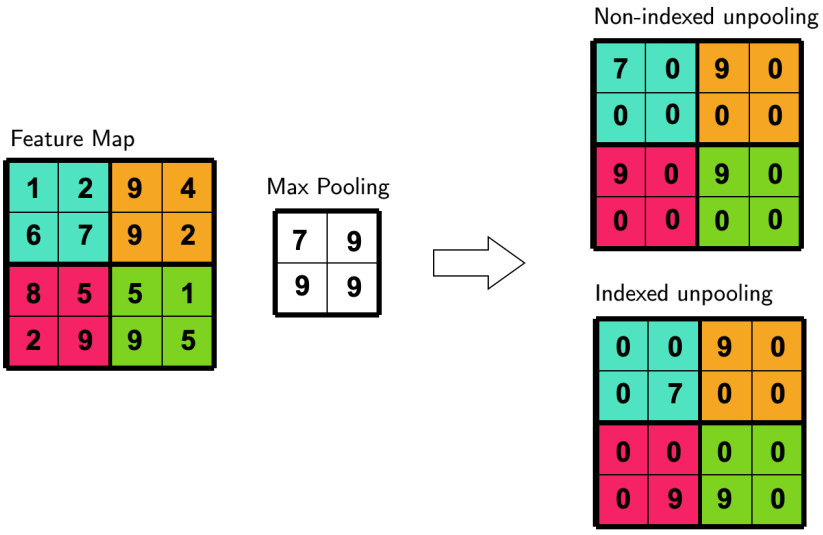


FIGURE 2.7: Illustration of indexed pooling.

Concluding, pooling is an essential operation that avoids the naive techniques of bilinear sampling. Although this allows to reduce the dimension of the images, through the different methods stated, it is possible to promote behaviors and therefore to infuse these layers with prior knowledge.

2.2.2.2 Atrous Spatial Pyramid Pooling (ASPP)

ASPP[1] is a concept created by Chen et al. [29] to define pooling strategy differently. The principle consists of several dilated convolutions, centered on the same pixel. Thus, one can define a pixel using a succession of convolutions to accumulate different receptive fields. Finally, it is not only a question of reducing the dimension of the image, but also of adding context to the remaining information. As a result of each convolution, a number of contextualized features are concatenated and processed by a dense 1x1 convolution. The resulting information will have been impacted by a panel of more or less neighboring information increasing the total impact on the image (receptive field added). A diagram in Figure 2.8 shows the organization of such a pooling architecture.

[1]This architecture is used in Section 4.5.2.

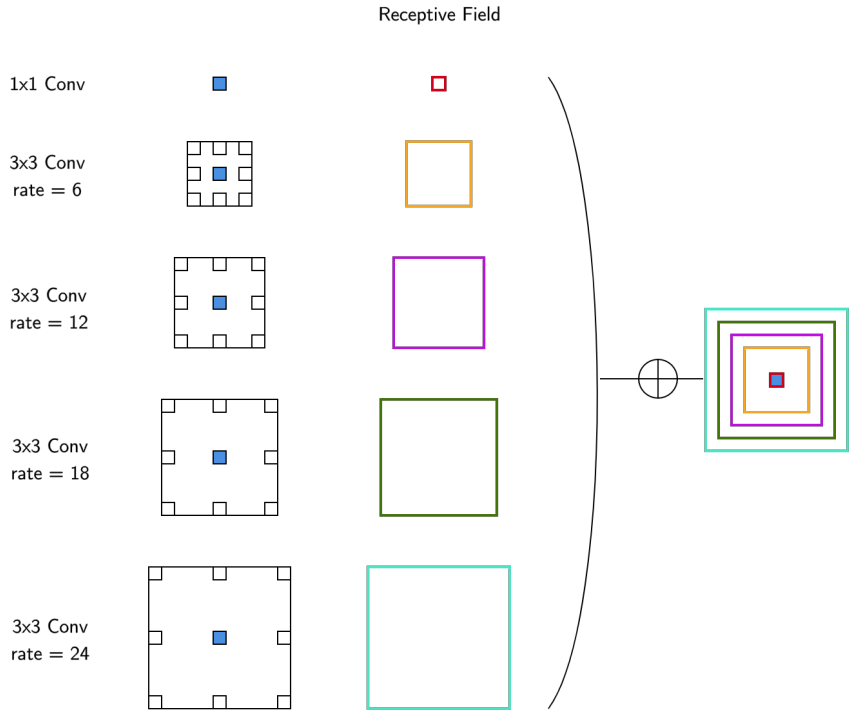


FIGURE 2.8: Diagram of ASPP strategy.

This technique has been validated in segmenting algorithms and shows that the use of such a process along with other complementary blocks allows a better understanding of the scene. By this, all indications are that this strategy favors the expansion of regions and the accountance of contours to define dense classes rather than classifying pixel by pixel without taking into account the neighbors.

2.2.2.3 Atrous Convolution

The concept of atrous convolutions has already been briefly expressed above. Indeed, this convolution is identical to a dilated convolution (shown in Figure 2.8). This operation increases the receptive field, i.e. the impact of the convolution on the original image. This has a particular effect which is to introduce context into the calculation. Instead of considering only the nearest neighbors, one can space the kernel and therefore consider distant pixels.

Thus the convolution atrous has a direct influence on the spatial importance. It is a question of finding a compromise between requiring context by strongly dilating the convolutions or forcing the importance of the localization through a tightened kernel.

2.2.3 Conclusion

This section allowed a visualization of the Deep Learning basics by introducing in turn the data, the network and the loss.

Through these three essential points, it has been possible to summarize the various key aspects of Deep Learning. In the same way, it was possible to expose the critical aspects of the work presented in this manuscript. Indeed, since data is a cornerstone of our methods, it is necessary to put open questions that can be addressed throughout this thesis. Moreover, it has been estimated by the community that the networks and their dimensioning depend largely on the task addressed. Finally, the cost function represents the critical element to sustain a safe and valid optimization. Once these three aspects are reviewed, it is possible to unpack some of the concepts that will be needed for the diverse applications proposed in this thesis. Thus, we have overviewed three specific notions recurrently used in our work for segmentation approaches.

2.3 Summary

This chapter reviewed general concepts of Polarimetry and Deep Learning. We first introduced the general principles of Polarimetry and described the modality as well as its particularities. Starting from the general concept, we decided to introduce the acquisition system by describing the sensor. Besides, we defined the framework for exploiting the data by stating the general equations governing the space deriving the characteristic images of the polarization state. Then, we proposed to briefly define Deep Learning. We emphasized on the different basics defining their importance in the field. Then, we defined different more detailed concepts that are necessary for a deeper understanding of several architectures. The two sections provide a broad introduction to the two key concepts widely used throughout this thesis.

Chapter 3

Literature Review

Computer vision includes many sub-domains allowing the understanding of the environment. Scene understanding is one of the most active areas since it allows machines to characterize scenes similar to human behavior. Among the primary applications, segmentation and depth estimation represent substantial challenges addressed by the community. The former allows pixel-level classification of images for semantic understanding, the latter estimates the proximity between the sensor and the scene by evaluating distances. In recent years, with the advent of learning-based approaches, these two domains have evolved, benefiting from the computing power of Deep Learning networks. A variety of studies have shown that recent algorithms largely outperform non-learning based approaches and that therefore these scene understanding problems are simplified by the abstraction possibilities inherent to machine learning. Thus, optimal estimates from an image can be generated through these modern techniques allowing robust and generic inferences.

This chapter provides an exhaustive review of the two domains that are segmentation and depth estimation. First, in the segmentation dedicated section, we highlight the different major contributions by positioning them with regards to CityScapes benchmark. In detail, the methods of semantic segmentations are summarized according to their backbone: VGG and ResNet. The different designs are discussed, aiming to provide a comprehensive understanding of the contributions. Describing the architectures and their proposals, this review also allows evaluating existing approaches as a preliminary work of Chapter 4.

Next, we dedicate a section to depth estimation methods by describing the domain. In detail, we propose in a first step to summarize the depth acquisition techniques by depicting their different advantages and drawbacks. We subsequently propose a review of learning-free multi-image depth estimation methods that have served as a step towards deep learning methods. Finally, we conduct a comprehensive survey of an extensive range of learning-based depth estimation methods. Describing the architectures and their particularities, this summary details the specifications of the

pioneering methods in the field to provide a comprehensive and heuristic view of deep learning based depth estimation. Moreover, we gather quantitative experimental results of depth estimation methods on KITTI Eigen benchmark. This analysis allows us to evaluate the performances and the robustness of the approaches. Besides, this survey allows, through the explanation of the contributions and their respective quantitative evaluation, to serve as preliminary work for Chapter 5.

3.1 Deep Learning-based Semantic Segmentation

In this section we will discuss different deep learning based segmentation methods. Through the leading approaches, we provide an overview of the field. In Section 3.1.1, an introduction as well as an overview of the research field will be proposed. Next, in Section 3.1.2, we will motivate the choice of these methods by displaying a panel of backbones and networks as well as their respective performances. Then, in Sections 3.1.3 and 3.1.4, we will expose methods based on VGG and ResNet respectively, since these are the two most widespread backbones in the domain. Finally, in Section 3.1.5, we will propose a conclusion and a discussion related to the domain.

3.1.1 Introduction

Segmentation remains a substantial area of computer vision. It is also historical since it is possible to find contributions proposed in the 70's [110]. The main idea of the domain is to be able to delimit and differentiate objects/areas. Thus, the segmentation can be assimilated to a transposition of an image space to an intermediate space, semantic or not. That is to say that instead of utilizing colorimetric information, the representation is changed to differentiate the areas of interest.

Multiple approaches can be used to perform image segmentation, apart from learning-based ones. Indeed, even before the advent of deep learning, it was possible to divide the different methods into four categories: region-based, feature clustering-based, edge-based, and model-based.

Region-based and edge-based are two types of methods that are significantly correlated [74, 33]. The principle is to divide the image using descriptors which have the effect of describing the image using contours from which regions can be deduced. For example, [72, 131] both proposed edge-based segmentation and [150, 107] proposed contour-based approaches. It is explicit that these two techniques are highly linked and seamlessly one can shift from one to the other.

The threshold-based techniques [11] are quite naive. Whether the level is deduced iteratively or empirically, this method consists in fixing a value at which the image

will be bounded and then binarized. Predominantly, this technique allows to separate areas like the foreground and background or to highlight recognizable objects.

Conversely, feature-based discrimination can be considered as more advanced or complex. It consists in clustering a group of features that have been previously extracted. This domain has been highly investigated [151, 86, 121]. Indeed, this is likely due to the vast amount of information extractors previously developed. As a consequence, researchers have been able to exploit the acquired knowledge and then derive robust segmentation algorithms.

Finally, model-based segmentation consists in adopting descriptive models to differentiate the objects in the image. Thus, models like the Markov Random Fields Model [99], active shape models [143] and active appearance models have been used to segment images [35].

In spite of all these possibilities and their respective robustness, the algorithms remain rather tied to the applications and unfortunately, they are rather rarely generalizable (viz. when they are designed for one task, they rarely adapt to others).

At present, the vast majority of algorithms have shifted into the field of deep learning. This is purely due to the performance observed with this approach. Thus DL usage allows to prevent the creation of a handmade feature space. In addition, DCNNs have proven their generalization abilities. The primary constraint is the amount of data but once this criterion is met, then the networks have such an abstraction capacity it is unnecessary to explicitly constrain the problem. In consequence, some algorithms provide knowledge into the network even if this practice is quite marginal. Preferably, researchers are seeking blocks or layers that will allow to abstract the data to higher degree and render it more understandable for the algorithms.

3.1.2 Major backbones and networks: a comparative evaluation

In the semantic learning-based segmentation landscape, a large number of methods are available. Indeed, the field is active and in constant evolution. However, it is possible to extract methods that stand out from the crowd. Moreover, among all the methods, a criterion allows to differentiate them: the backbone.

The backbone represents a term that designates the feature extraction structure. It can also be termed — erroneously — encoder and allows to transpose an image from its initial space to a latent space, usually a feature vector. The concept differentiating the terms encoder and backbone is that the backbone is predominantly used pre-trained and has been proven efficient for feature extraction and thus classification tasks. On the contrary, an encoder is a general term that only designates an architecture that reduces the dimensionality of the data by densifying the information.

From these feature extractors, a vast number of networks have been derived. As a reminder, a segmentation network can be summarized as a classification network with a secondary structure to recover the initial image dimension. It is also possible to call a segmentation network a classification network at pixel level.

To determine the most representative networks and their respective backbone, we propose in Table 3.1 a quantitative analysis of the results from significant networks on a common benchmark: CityScapes [36]. The comparison metrics are Intersections over Union (IoU) by classes or categories initially proposed by [48]. In addition, the CityScapes benchmark proposed to calculate the iIoU which corresponds to the instance-level IoU and is considered more representative of the actual results[1].

TABLE 3.1: Overview of major approaches highlighting performances on CityScapes benckmark [36] and backbones. Underlined and **bold** respresents respectively overall second best and best. *Italic* is best per backbone.

Backbone		Approach		Reported performances on CityScapes Benchmark [36]			
Name	Year	Name	Year	IoU class	iIoU class	IoU category	iIoU category
VGG [135]	2014	FCN [92]	2015	65.3	41.7	85.7	70.1
		SegNet [4]	2015	57.0	32.0	79.1	61.9
		DilatedNet [165]	2016	*67.1*	*42.0*	*86.5*	*71.1*
ReNet [146]	2015	ReSeg [145]	2016	*58.8*	-	-	-
ResNet [67]	2016	PSPNet [176]	2017	81.2	59.6	91.2	79.2
		RefineNet [89]	2016	73.6	47.2	87.9	70.6
		LKM [113]	2017	76.9	-	-	-
		EncNet [169]	2018	-	-	-	-
		DeepLab v2 [29]	2016	70.4	42.6	86.4	67.7
		DeepLab v3 [30]	2017	81.3	62.1	91.6	81.7
		DeepLab v3+ [28]	2018	***82.1***	***62.4***	***92.0***	***81.9***
		Mask-RCNN [66]	2017	-	-	-	-
ResNeXt [157]	2017	DShortcut [12]	2018	-	-	-	-
		ExFuse [175]	2018	-	-	-	-
MobileNet v1 [70]	2017	FSTSL [156]	2018	71.9	-	-	-
MobileNet v2 [130]	2018	LWRF [109]	2018	*72.1*	-	-	-
		Fast-SCNN [118]	2019	68.0	37.9	84.7	63.5

This evaluation table shows that mainly two backbones are used, namely VGG and ResNet. This is why the next two sections will be dedicated to these backbones and the networks that operate them.

3.1.3 VGG

VGG [135], standing for *Visual Geometry Group*, is an image recognition network that was proposed in 2015 by the VGG lab at Oxforfd university. It can be considered the pioneer in the field of classification and was, until the appearance of ResNet the only viable backbone. This encoding network is based on small 3x3 convolutions and

[1] Calculation of specific metrics and benchmark results available at: https://www.cityscapes-dataset.com/benchmarks/

is reasonably simple (compared to recent architectures). Similar to all encoders, it allows for dimensionality reduction of the input data and for its densification to obtain a representative vector, in particular it encodes an image $MxNx3$ into a feature vector $1x1x1000$. Initially and as shown in Figure 3.1, $M = 224$ and $N = 224$.

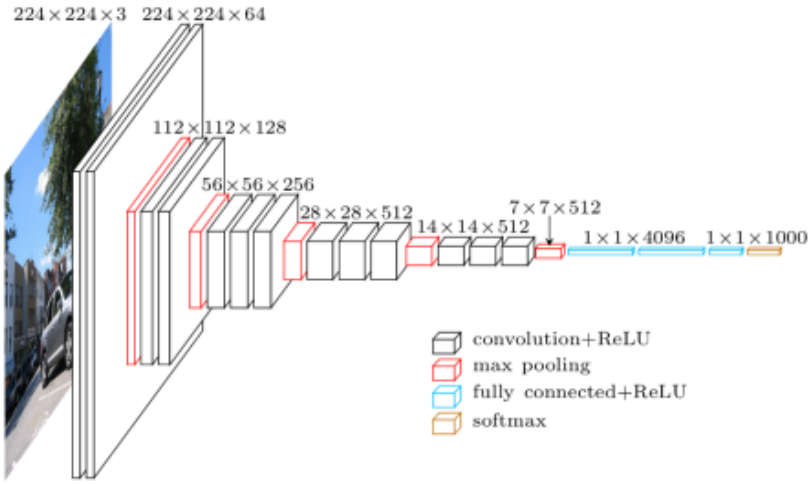


FIGURE 3.1: Simonyan and Zisserman's [135] VGG Architecture.

From its state-of-the-art performances, this network has allowed to derive multiple methods, three of which will be discussed below: FCN [92], SegNet [4] and DilatedNet [165].

3.1.3.1 FCN

Fully Convolutional Networks (FCN) [92], is an architecture proposed by Jonathan Long, Evan Shelhamer and Trevor Darrell in 2015. It is a highly recognized work to be the "first" pixel-wise semantic segmentation network (PwSS). With the objective of removing the image size constraint, the prevailing idea was to base the architecture solely on convolution layers as shown in Figure 3.2. Convolution effectively represents a valid technique ensuring the same image size at the output of the pipeline.

However, the challenge was predominantly based on this dimensionality. As follows, Long et al. proposed an approach based on pooling and deconvolution to retrieve the image dimension while keeping the semantic segmentation information across layers. While the encoder extracts the information and interprets the image, the decoder considers the task of increasing the dimension while keeping the localization aspect. Only, the effect of downsampling allows reducing the dimension but at the cost of the resolution and the details when proceeding to the upsampling. In response to this phenomenon, Long et al. introduced the concept of skip connection allowing an aggregation, in the decoder, of information coming directly from the encoder. By

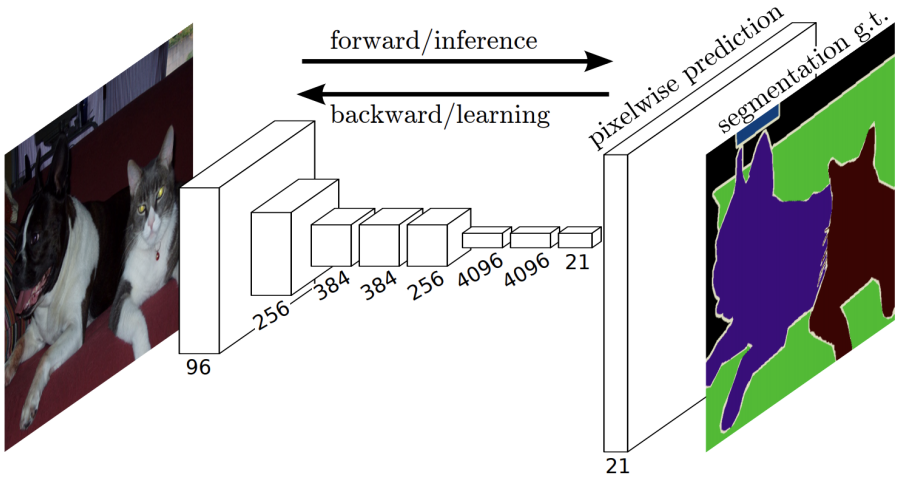


FIGURE 3.2: FCN Architecture from [92].

this approach, the model is efficient to infer fine-grained segmentation while taking advantage of the dimensionality reduction of the encoder.

3.1.3.2 SegNet

SegNet [4] is a network highly inspired by FCN but uses other innovative principles. While FCN proposed the skip connection concept, SegNet proposes a symmetric architecture with VGG. And, as a replacement to skip connections, authors develop the principle of indexed pooling. The encoder uses maxpooling operations to reduce the dimension and obtain a feature vector. From this pooling block, an information map of lower dimensionality and an index map is extracted similarly to [167]. This index map will then be used in the unpooling operations of the decoder to "reconstruct" the segmentation map with a proper positioning.

3.1.3.3 DilatedNet

In this approach, Yu et al. [165], instead of attempting sampling-based approaches, investigated the concept of dilated convolution. Thus, DilateNet is illustrated by its structure using the dilation properties of convolutions to reduce or increase the size of the maps. The idea is based on this concept of densification of feature maps but also on the increase of the range (i.e. the receptive field) of the convolutions. The advantage is that, for the same impact on the images, the use of dilation allows a much lesser use of parameters and thus the complexity of the networks is decreased by this way. As stated by the authors : *"the receptive field grows exponentially while the number of parameters grows linearly"*.

3.1.4 ResNet

ResNet [67] is another backbone and it can undoubtedly be considered as the most widespread in all domains. The creation of this architecture starts from an observation. Prior to this contribution, the community assumed that the deeper the network, the better it performed. This is legitimate and verifiable, but it is straightforward to observe that in reality, after a certain number of layers, the network loses resolution and performance. This fact comes from a recurrent problem in the learning task: the *vanishing gradient*. After an certain amount of operation, the results are largely approximated by the standard algorithmic methods such as rounding or floating point precision. This would suffer no impact if the training did not require back propagation. Thus, during the propagation, due to the approximation, the floating point estimate [62] or simply the chain rule [85] used during the calculation of the gradient, the gradient shrinks to zero. This vanishing gradient problem prevents the layers from updating and the network does not train anymore.

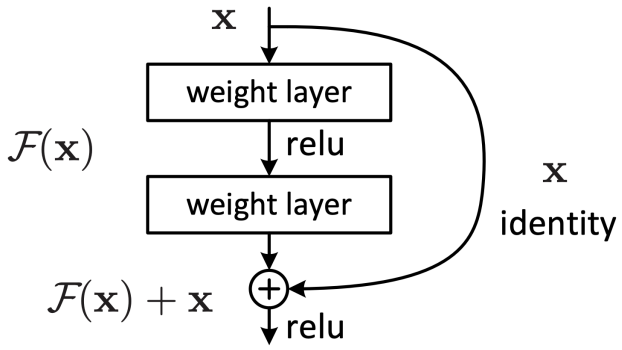


FIGURE 3.3: Illustration of ResNet [67] block.

ResNet overcomes this problem by implementing layer-wise skip connections that allow the gradient to be transmitted smoothly (shown in Figure 3.3). By adopting this strategy, it is then possible to use the starting postulate and increase the size of the networks. Once the vanishing gradient is eliminated, nothing prevents the densification of the networks since the skip connections allow the transmission of information.

3.1.4.1 PSPNET

Pyramid Scene Parsing Network (PSPNet) [176] is a pixel-wise semantic segmentation network derived from ResNet classification network. This contribution is focused on an innovative sampling method based on a multi-layer pyramid. The approach called *pyramid pooling* is an architecture in four levels (instead of one traditionally). Considered to keep the context of the images, the four stages of the pyramid consider in parallel the whole, half or portions of the base image. Ultimately, whether it is the

first feature extraction or the output of each layer of the pooling structure, all this information is adapted in dimensions and concatenated. All this followed by a dense convolution allows keeping both the local and global context of the image and thus to maintain the information across layers. Thus, the loss of information is reduced which allows to have sharp edges and accurate estimations of the classes according to the features extracted by the numerous layers of the network.

3.1.4.2 DeepLab

There are multiple versions of DeepLab. However, it is possible to summarize the ResNet-based contributions into some key concepts brought by Chen et al. [28]. First, the global architecture is based on dilated convolutions or atrous convolutions. As expressed for DilatedNet, this operation allows many advantages like dimensionality reduction or judicious management of the receptive field.

Subsequently, this method proposes the use of bilinear interpolation which seems to represent a naive approach to upsampling. In practice, the authors expressed that the use of another dimensionality recovery strategy is not especially necessary to obtain efficient results.

Ultimately, the contribution proposes the use of *fully connected CRF*. The principle is to consider the pixels as a node in a graph and that all these nodes are connected by different means. Thus, the idea is to enforce the assignment of similar labels to adjacent pixels. The usefulness of this kind of structure could be justified by the classification of areas instead of independent pixels. Thus, this architecture allows to keep a coherent structure and reduce the aberrations of segmentations with nested classes. In addition, fcCRFs shows great performances with regard to the edges and the object separation. The contribution of the CRF conveys enormous complexity to the network, but some demonstrations [80] has made it possible to remove certain constraints and thus make this type of structure viable. It is nevertheless considerable that this brings a great number of additional parameters and especially increases the complexity of the networks. To conclude on this architecture, it remains to this day (with its declensions) the state-of-the-art method for segmentation. Some networks highlight better class-wise performances but include a cost and a disproportionate complexity compared to the difference in efficiency.

3.1.5 Conclusion and discussion

After briefly explaining the major backbones and their main derived segmentation architectures, it was illustrated that learning-based segmentation has been a flagship field in computer vision. The recorded performances presented in the comparative study are far superior to observations prior to the advent of deep learning. However, it

is possible to see that this field is almost at the end of its course since the performances tend to stagnate and the research to move towards other processes of understanding. The emergence of instance segmentation or geometric understanding of scenes tends to reduce the amount of contribution in PwSS.

Despite this observation, it is also quite possible to note that the algorithms are subject to the same constraints as any deep learning algorithm, namely the need for data. Plus, an overwhelming majority classify colorimetric images and assume these data are thoroughly characteristic of the observed scenes. There is a considerable knowledge in the field of segmentation, but it has rarely been exported to other types of data to see if the performance on similar tasks can be improved by a strategic use of the data. Hence, the answer might not be to depend on the depth and number of layers of the network but to shift the data space to something more characteristic.

3.2 Depth Estimation

In this section of the literature review, various concepts related to depth estimation will be discussed. Starting from the principal acquisition technology in part 3.2.1, we will continue on the learning-less methods for depth estimation in part 3.2.2. Ultimately, we will more deeply discuss the concept of learning-based depth estimation from a single view in part 3.2.3 since this is the topic to which part of this manuscript is devoted.

3.2.1 Laser-based imaging

Laser-based imaging is a frequently used method for its accuracy and robustness. Indeed, it allows, regardless of the conditions of illumination or distance, to accurately estimate the distance between the sensor and an object using a laser projection. There is a wide variety of different sensors, and their performance is highly correlated with their price. Indeed, while some devices allow a smooth acquisition of several thousand points or at distances of several hundred meters, others are relatively limited in capacity. Despite this, the overall acquisition principle remains the same and is uninfluenced by this disparity in performance. The concept is considerably classic, a laser is projected on a surface that will then reflect it (back-scattering). As follows, it is possible to calculate the distance by measuring the time elapsed between the emission and reception. However, the whole principle is based on reflection. As a result, this approach can be inefficient especially when the target scenes include specular surfaces [71] (e.g. mirrors, glass, water, and other reflective surfaces or transparent/translucide ones). It also turns out that these sensors can be made deficient in adverse weather conditions. In particular when the projected laser can be refracted in rainy weather or altered in foggy weather. Conversely, few approaches are as robust in favorable

meteorological circumstances. Therefore, most datasets containing target depth maps use laser-based imaging like KITTI [56, 102, 52] and its LiDaR setup. But, the data contain a bias since they were acquired mainly in favorable conditions.

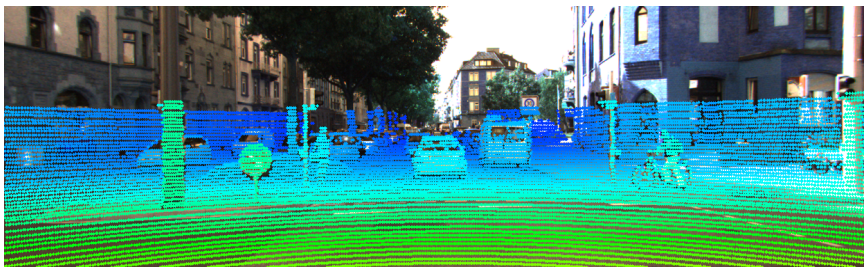


FIGURE 3.4: Illustration of LiDar point-cloud[2]. Sparsity can be observed as well as erroneous estimation on vehicle windows.

The LiDar, *laser imaging, detection, and ranging*, is a laser rangefinder commonly used since it is motorized and rotatable [148]. Rotating at 360° at a considerable speed, it allows to make panoramic acquisitions. Moreover, LiDaR are classified according to their number of slices, which gives them, not only a horizontal but also a vertical acquisition. Thus, the more slices, the more vertical points are acquired "semi-simultaneously". As shown in Figure 3.4, the generated point cloud can be affixed to the color image corresponding to the scene. This illustration also highlights the defects related to transparent surfaces, here, the vehicle windows.

It should be noted that this operation contains flaws of which two critical ones are identifiable. The acquisition of moving objects, especially if the relative speed between sensor and target is high, can be very inaccurate. The second drawback comes directly from the acquisition technology. A multitude of points is projected in the surrounding space. As previously discussed, even if it depends on the size of the pattern, only points are acquired [144]. As a consequence, the deduced point clouds are very sparse, and this factor is aggravated by the distance between an object and the sensor (visible in Figure 3.4). As a result, the depth maps are sparse or subject to interpolation to fill this sparsity.

In conclusion, LiDaR remains a robust tool for scene depth estimation. Regrettably, this device can be expensive or inefficient. As a result, the Computer Vision community tends to find estimation algorithms and consequently to avoid this scanner and its drawbacks.

3.2.2 Multi-image methods

Multi-image based depth estimation methods have been developed for several purposes. One of the main ones is the simple set-up of the acquisition system. Indeed,

[2]Image borrowed from henryzh47.github.io

standard cameras can be used to recover the depth. Thanks to the availability of such methods, many algorithms have been developed whom can be classified into two main groups: Multiple camera systems described in part 3.2.2.1, and Single camera systems in part 3.2.2.2.

3.2.2.1 Multiple camera systems

One of the best-known approaches in the field of multi-camera systems is stereovision. Highly inspired by human biology [98], this technique consists in the use of two cameras similarly to the eyes.

From this acquisition system, two different points of view are obtained and then allow the matching of points of interest. The concept is based on the projection of points from a three-dimensional space to a two-dimensional image plane. Hence, starting from a 3D point with homogeneous coordinates, it is possible to transpose it into the image frame such that:

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = kP \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}, \tag{3.1}$$

with $(x, y, 1)^T$ the 2D image coordinates and $(X, Y, Z, 1)^T$ the 3D world coordinates. Plus, k represents the scale factor and P the projection matrix.

P is a constraining matrix that contains the intrinsic parameters of the camera and the extrinsic parameters. These two parameters are respectively the proper properties of the camera independent of any external factor, and the information necessary for the positioning of the camera in the world frame like a 3x3 rotation matrix and a 3x1 translation vector. To find these two parameters, P is decomposed as follows:

$$P = \underbrace{\begin{pmatrix} f_x & 0 & x_0 \\ 0 & f_y & y_0 \\ 0 & 0 & 1 \end{pmatrix}}_{K} \underbrace{\begin{pmatrix} R_{00} & R_{01} & R_{02} & t_x \\ R_{10} & R_{11} & R_{12} & t_y \\ R_{20} & R_{21} & R_{22} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{E}. \tag{3.2}$$

In this manner, it is possible to retrieve the intrinsic parameters matrix K composed of: f_x and f_y the oriented focal distance in the image, and (x_0,y_0) the optical center coordinates in the image frame. In addition, we can recover the rotation matrix R as well as the translation vector t with respect to the world frame in the extrinsic

parameters matrix E. Note, the constraining aspect of P comes from the necessity of these parametric matrices $\{K, E\}$ since this implies a calibration of the system.

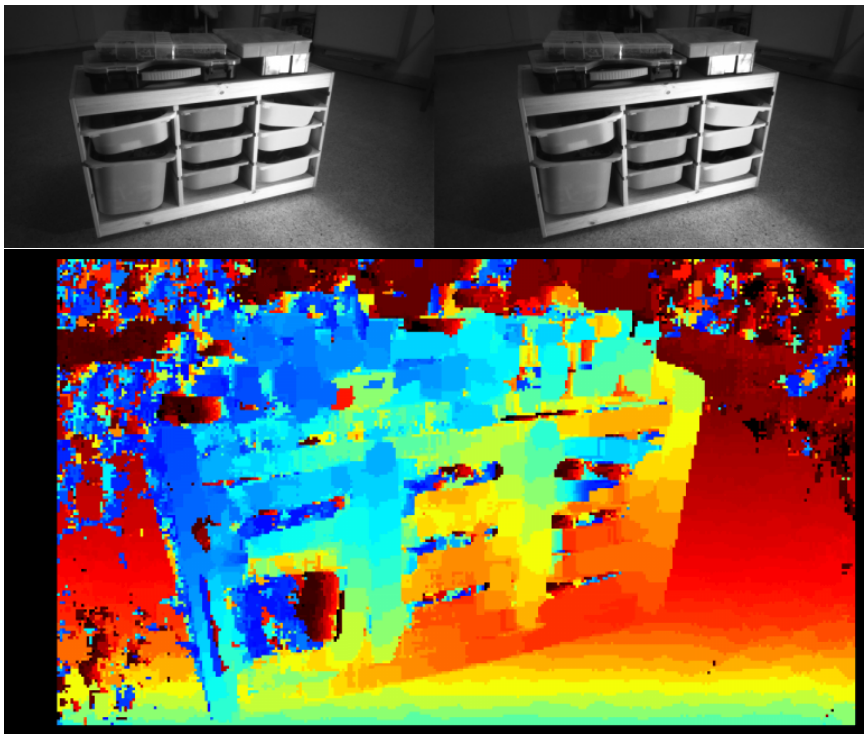


FIGURE 3.5: Intel RealSense acquisition and reconstruction example. Top row shows the rectified images from the camera, bottom row shows the obtained reconstruction.

Conversely, as soon as the intrinsic and extrinsic parameters are known, this system is solvable. This allows us to find the coordinates of 3D points as long as we can identify them in the two or N views. Hence, whether the system is binocular in stereovision as shown in Figure 3.5, or whether the cameras are multiplied in a multi-view system, the problem is reduced to the matching of points identifiable through different views. As proved in [8, 6, 132, 55] for stereovision, or in [64, 63, 24] for multiple view reconstruction, this domain has been thoroughly investigated. Over the years, the research community oriented these approaches towards the real time / online estimation through FPGA embedding [7], optimization improvements [79, 103, 126] or even speed enhancements [50].

Since the system is identified from the stage where it is calibrated, and thus the problem relies almost solely on matching, then many feature differentiation methods have emerged to advance the field.

Thus, different information extraction approaches such as: brightness-based [57], segment-based [140], feature-based [133] or segmentation-based [18], have contributed to the progress of multi-image reconstruction.

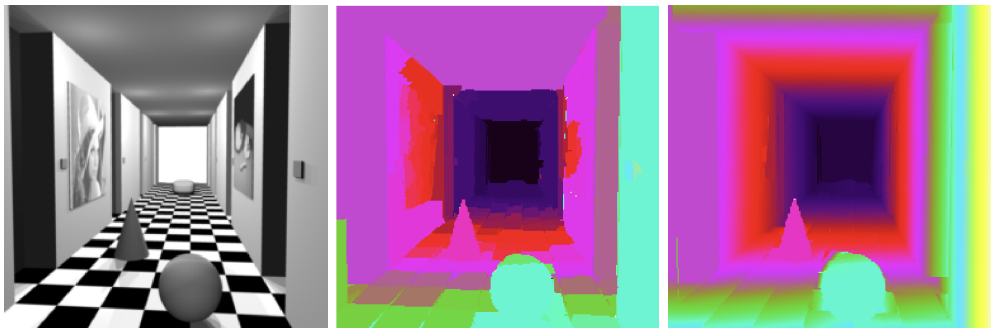


FIGURE 3.6: Illustration of stereovision depth estimation using [155] method. From left to right: reference image, first order reconstruction and second order reconstruction.

However, to highlight only one contribution, Woordford et al. [155] proposed new optimization approaches leading to leveraging graph cuts to relax surface-orientation major constraint. Therefore, while previous algorithms were struggling when addressing non-fronto-parallel surfaces, they succeeded into integrating second order derivatives leading to more accurate and reliable depth maps. Thus, this method allowed smoother and more precise surface reconstruction in realistic conditions. As shown in Figure 3.6, depth estimation became finer as [155] has been used.

Despite of all the numerous contributions on the multi-image concept and the considerable reconstructions accuracy, still some significant drawbacks subsist. First, as it was discussed above, the calibration aspect represents a key constraint for the acquisition system as well as its deployability. Secondly, as it was highlighted, even if a considerable amount of approaches focuses on features recognition, stereo matching by definition requires features. Unfortunately, the acquisition may observe textureless areas which lead to unsolvable problem. Last, the acquisition system requires two or more cameras, and this by essence could be problematic. In response, some methods lift this last constraint by operating a single camera.

3.2.2.2 Single camera systems

Depth estimation from a unique image prevents material constraints but raises other questions. Indeed, the use of such a setup prevents the use of triangulation made possible by the use of multiple sensors. Therefore, innovative approaches named Shape-from-X (SfX) have emerged. In this framework, X represents several possible cues like motion [23, 39, 32, 112], shading [69, 171], blur [49, 180], texture variation [2], polarization [104, 139] etc.

As a deepened example, Favaro and Soatto [49] proposed a original approach to benefit from de-focus phenomenon.

FIGURE 3.7: Estimate computed from Favaro and Soatto's method [49]. The two left images show different images from the focus point of view. The right image displays the result of their reconstruction approach.

The proposition is such that from two images with different camera settings, implying dissimilar focus, they propose computing either:

- an optimized inference when *point spread function* (PSF) [128] is known
- kernelized orthogonal operator through convolution

to estimate the 3D geometry of the scene. As shown in Figure 3.7, with two acquisition of a scene with different focus (two left images), a depth can be infered (right image) just by a blur-focused algorithm.

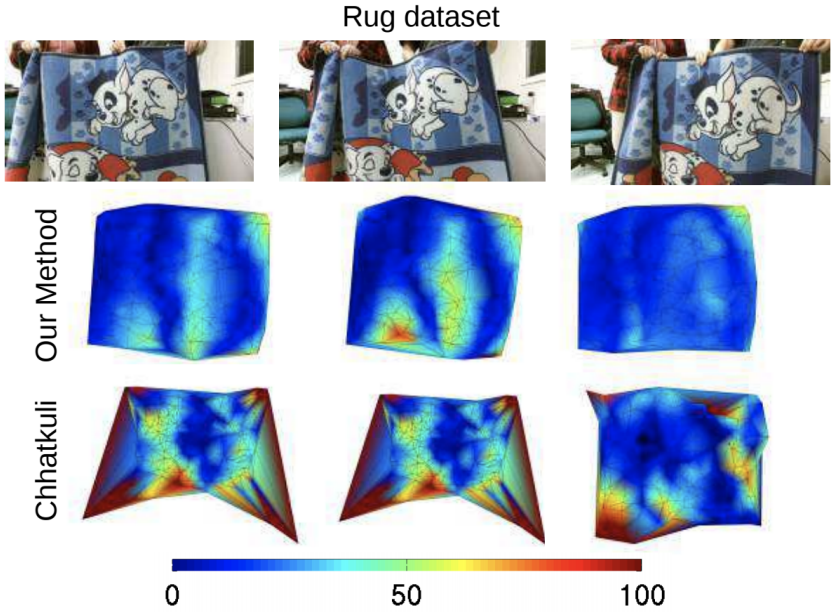


FIGURE 3.8: Example of estimates from Parashar et al. [112]. From top to bottom are: input images with deformable object wrapped over time, reconstruction error from [112] and reconstruction error from [32].

More recently, Parashar et al. [112] contributed in the field of shape-from-motion. They propose addressing *Isometric Non-Rigid Shape-from-Motion* which consists of reconstructing a non-rigid object observing shape variation over time. To accomplish

such a task, they introduce Riemmanian manifolds-based [84] representation of the deformable 3D surface. To such a degree, they succeed into modelizing the warps applied to the non-rigid object over time. As shown in Figure 3.8, the proposed method showed increased performances compared to [32].

Despite all these advanced methods, it is considerable that the computational complexity of such algorithms is extremely high. As the complexity of the algorithms increases, the performance and usability suffer.

In conclusion, despite the fact that the problem of single-camera depth estimation generates an infinite number of solutions and therefore the problem is considered to be ill-posed, the scientific community has been able to adopt innovative approaches to resolve the difficulties. Thus, at the cost of this increased complexity, the acquisition systems could be lightened and SfX field has proposed modern alternatives for depth estimation. However, the arrival of learning with the growth of computational power has allowed the development of learning-based approaches.

3.2.3 Learning-based monocular depth estimation

In the previous section, we saw it was possible, from several images, whether from one or more cameras, to estimate a depth map. However, each method based on "standard image processing" contains considerable drawbacks. With the advent of Deep Learning, many fields have changed and depth estimation is no exception. Indeed, in essence, learning should relax the constraints by taking advantage of a massive amount of data and computation.

These learning-based methods can be divided into two distinct parts, (semi-)supervised learning and self-supervised learning, which will be explained in parts 3.2.3.1 and 3.2.3.2 respectively.

3.2.3.1 (Semi-)supervised learning

This deep learning process[3] is excessively used as seen in section 3.1, and as for segmentation, for depth estimation, it requires a considerable amount of annotated data. Indeed, as segmentation requires label maps, DCNN-based depth estimation requires reference depth maps.

The problem is thus formulated differently. While previous methods required a characterization of the essential information to match across views, learning-based methods learn by themselves the necessary feature space, provided they utilize a consistent and representative amount of data.

[3]Here, the choice was made not to dissociate semi-supervised and supervised learning since they are based on the equivalent concept.

Among the first to investigate the benefit of a profusion of aligned RGB-D data, Eigen et al. [46] were able in 2014 to show that despite an ill-posed problem (as expressed in 3.2.2.2), it is possible to obtain sustainable results.

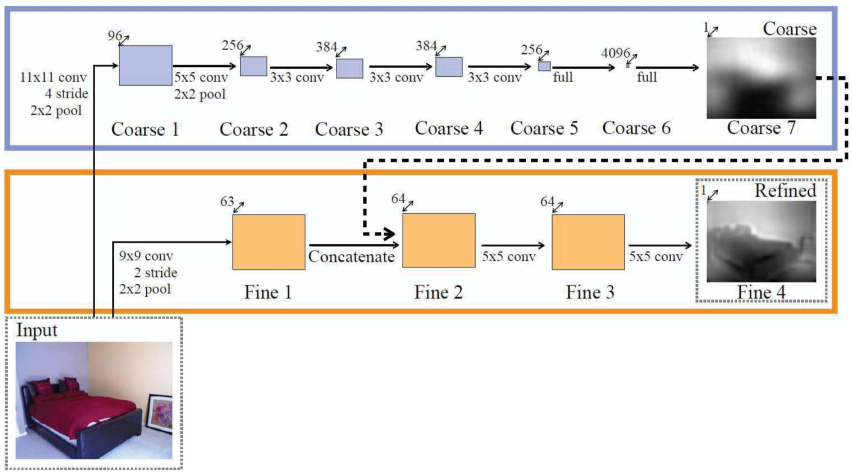


FIGURE 3.9: Eigen et al.[46] Architecture.

By implementing a biphasic network (shown in Figure 3.9) allowing the estimation of two depth maps, one coarse and the other refined, while proposing the use of a scale-invariant loss, they were capable to take advantage of rich annotated datasets such as NYU [108] and KITTI [56, 52, 102]. The contribution is twofold, firstly the cascade network of which the first one, from which the coarse map results, is not completely convolutional (last two layers dense and fully connected).

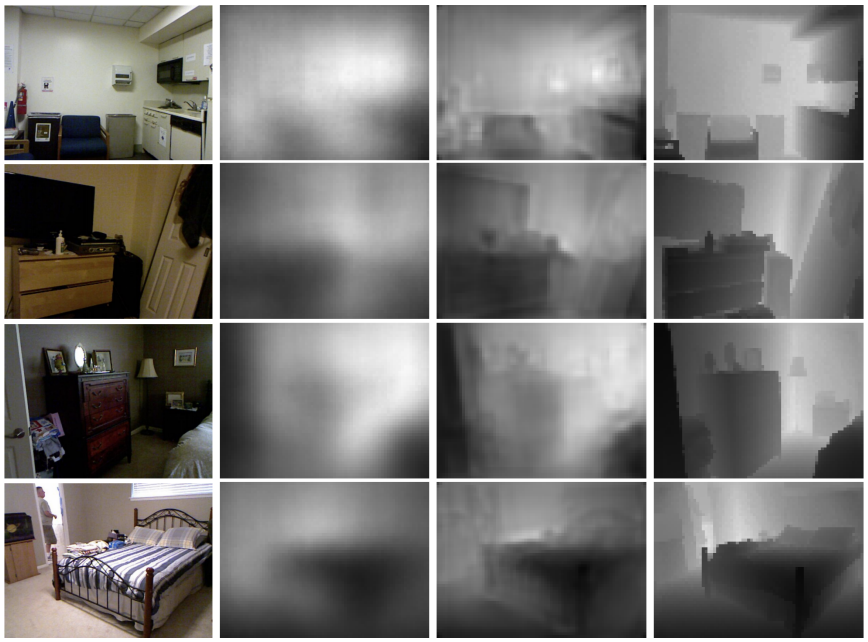


FIGURE 3.10: Eigen et al. Results on NYU dataset. From left to right: input image, coarse estimation, refined estimation and ground truth.

This fact implies the sacrifice of the categorization of local features in favor of a learning of global features. Consequently, this part removes the invariance pose to densify the intermediate latent space. In a second step, the scale-invariant error allows to introduce a further dimension to the learning. Rather than considering the pixel correspondence by simply using an absolute error, this function allows to partially force the depth relation between pixels. As a consequence, the proposed approach is no longer considered as a classification at the pixel level but as an estimation of the depth of the image. As shown in the Figure 3.10, the results obtained are impressive and above all will set the state of the art for the field by exceeding all previous performances regardless of the outdoor or indoor environment.

Following this major contribution, this first method has been improved. In 2015, Eigen and Fergus [45] added the semantic and normals to surface estimation by improving the previous method. Then, using multi-scale approach and by modifying the previous loss, they were able to improve the initial approach. Indeed, while adding a segmentation capability of estimation exceeding previous works, they also were capable to improve their previous outstanding performance.

Next, following these approaches, a multitude of contributions emerged. Liu et al. [90] proposed to investigate a joint collaboration between DCNN and continuous conditional random fields (CRF) showing great performance while having an unsmooth patch-like estimation due to super-pixel pre-segmentation for neighborhood relationship modeling. Another approach is Roy and Todorovic's [129] proposing CNN embedded neural regression tree enforcing the smoothness of output.

[158] investigate even more on CRF by introducing a multi-scale dimension to the estimation. Therefore, they created dedicated C-MF blocks which allows multi-scale fusion through the whole process. In consequence, this contribution proved an extensive performance exceeding all the previously cited works.

In addition, as an one-of-a-kind approach, [25] proposed to formulate the problem of depth estimation as a pixel-wise classification (can be assimilated as PwSS) problem. Indeed, they suggested the learning task is easier as formulated as a segmentation problem. As a finality, a CRF is applied to refine the map through local coherency reinforcement.

To cite a few other semi-supervised approaches, [61, 81, 53, 78], have all contributed to the growth of the field. Whether using a stereovision-based approach, adversarial learning, an ordinal regression formulation or a supervised SfM pre-computation, these methods require at least an intermediate depth estimation step.

A trend can be observed in the evolution of the field of (semi-)supervised depth estimation. Over the years, the algorithms have benefited from hardware support which has made it possible to evaluate such algorithms, despite their increasing complexity. In contrast, in addition to this prevailing tendency, each of these methods shares a common disadvantage. Indeed, they require a massive amount of annotated data and moreover, these datasets in addition to being consequent must be representative. Moreover, most datasets have annotations that depend on the acquisition process. As expressed in part 3.2.1, and to consider the example of KITTI, LiDaR is not providing perfect acquisition, especially in urban areas which contain many specular or refractive areas. As follows, when one supervises an algorithm with erroneous data, the model is necessarily influenced.

As a consequence, alternative self-supervised approaches have been developed to emancipate from annotation through a loss that does not require external information.

3.2.3.2 Self-supervised learning

Self-supervised depth estimation represents an extensively active domain in recent years. The concept comes from an observation concerning previous methods: the necessity of a ground truth is extremely restrictive. As mentioned before, in some cases the targets are unreliable enough and their acquisition is expensive. This field has even changed its purpose from another estimation method to ground truth generation. In this manner, the ambition of this kind of algorithm is to equal or even surpass the performance of sensors allowing a direct acquisition of this physical information that is the sensor-object distance. The primary concept is based on the same principles as standard self-supervised learning, i.e. a generalizing, discriminating and differentiating cost function, but also on the necessity of ground truth only for quantitative evaluation. Hence, the algorithms are not relying on external annotated data and can exclusively refer to the information they receive or produce to deduce the result by optimizing the loss function.

Fundamentally, the cost functions are similar across algorithms. They predominantly consist of the assembly of several terms of which two are extensively present: the reconstruction term and the smoothing term in the form of:

$$\Lambda = L_r + L_s + L_a, \tag{3.3}$$

where Λ is the overall loss, and L_r, L_s, L_a are respectively the reconstruction, the smoothing and the additional terms. In this section, we will address two diverse types of algorithms: self-supervised stereo-based supervision and self-supervised monocular-based supervision. In the repertoire presented here, we will consider that the method

is considered monocular if its inference is made monocularly. Consequently, some algorithms requiring two views for the learning period will remain valid for our selection criteria. In addition, a quantitative evaluation sub-part will display the different algorithms performances on the KITTI dataset's eigen split. This benchmark remains the reference in the domain as well as being freely available since it represents the comparison sample for the community. Last, the last sub-part will summarize and draw conclusions on the field to position the contributions of this manuscript in this research landscape.

Self-supervised stereo-based. This subpart of monocular depth estimation relies on a key point which is the training with a stereo image pair. These approaches are almost all based on the principles mentioned in section 1, with the difference that the processing core is a DL network and that consequently the loss and the images are sized for this computing node.

In 2016, Garg et al.[54] proposed an image reconstruction loss-based training procedure to infer depth following the equation:

$$L_r = \int_{\Omega} ||I_w(x) - I_1(x)||^2 dx, \tag{3.4}$$

which consists in a square error between the reconstructed image I_w and the left initial image I_1. Founded on an auto-encoder architecture, their approach predict an inverse-depth image which then derive an inverse warp allowing a photometric error between this synthesized image and the primary one.

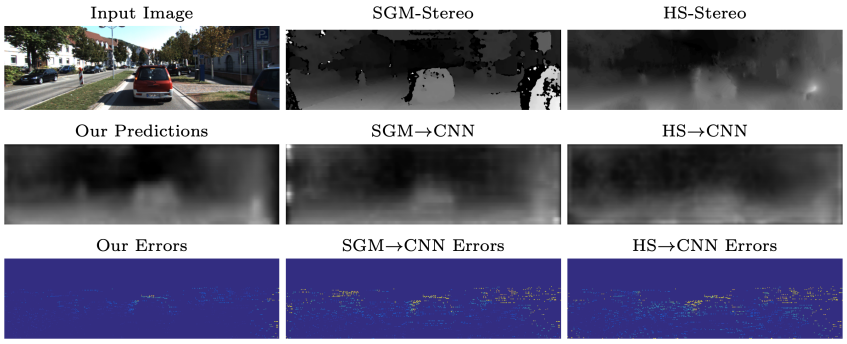


FIGURE 3.11: Results proposed by Garg et al. [54] highlighting error reduction compared to other methods.

Although this pioneering approach offers advantages, their image generation-dependent minimization method is undifferentiable. Consequently, at the cost of an increased optimization complexity, they require the use of Taylor approximation to linearize their deformed image and hence, allow the computation of a gradient. As shown in

Figure 3.11, this approach displayed accurate performances explicitly comparing to direct competitors semi-global matching [68].

One year after, [58] brings the left-right consistency concept as a term in the minimizable loss. Simplifying, the method consists in generating an opposite image. Using the spatial transformer network and its sampling blocks [73], this consists of reformulating the perspective geometry statement by recreating a projection.

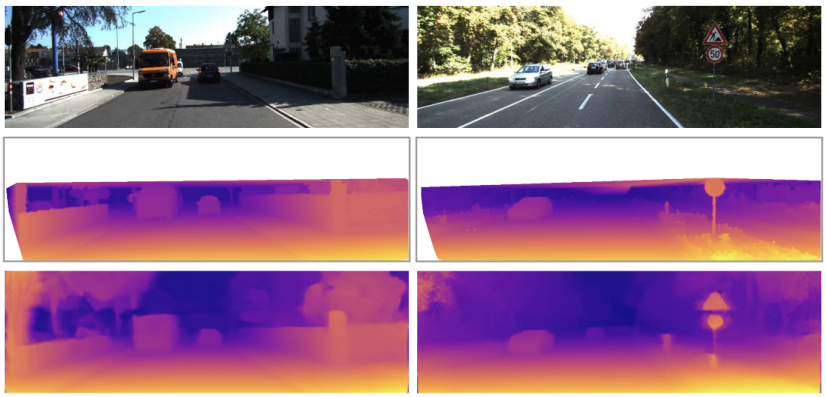


FIGURE 3.12: Godard et al. [58] qualitative results. Rows from top to bottom shows the input data, ground truth and estimates.

Moreover, Godard et al. propose to complexify the smoothness function to include an *edge-aware* effect that avoids edge blurring following:

$$L_s = \frac{1}{N} \sum_{i,j} |\delta_x d^l_{ij}| e^{-||\delta_x I^l_{ij}||} + |\delta_y d^l_{ij}| e^{-||\delta_y I^l_{ij}||}. \tag{3.5}$$

For information, the smoothness term allows attenuating the discontinuity in the estimation due to the gradient of the image. This contribution proposes the addition of a third term called *Left-Right Disparity Consistency Loss* ensuring equality between the first view and the second view projected with the deduced disparity by formulating C^l_{lr} as follows:

$$C^l_{lr} = \frac{1}{N} \sum_{i,j} |d^l_{ij} - d^r_{ij+d^l_{ij}}|, \tag{3.6}$$

where r and l stand for right and left respectively. C^l_{lr} is an additional L_a term regularizing the overall loss. As shown in the qualitative evaluation presented in 3.12, the predicted depth succeeds into inferring depth while preserving smooth transitions along distance and salient edges.

Investigating a radically different proposal, [100] suggests the possibility of generating depth maps using Generative *Adversarial Networks* (GAN) [60]. As shown in Figure

3.13, from an input RGB image, a generator will infer a depth map which then will be discriminated.

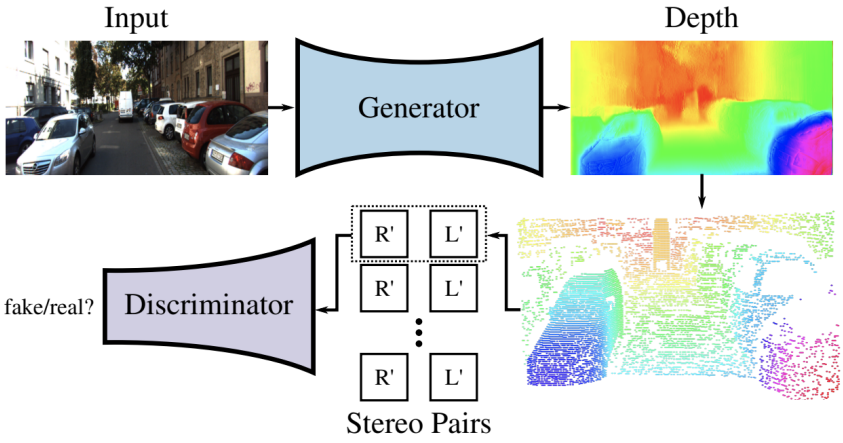


FIGURE 3.13: Mehta et al. [100] GAN Architecture.

The concept of GAN initially consists in the synthetization of photo-realistic images. Within the framework of this approach, the goal is to transpose the space of the colorimetry to the depth. Consequently, the discriminator is improved to ensure a valid estimation. To do so, the task of this network is to distinguish whether a generated view is plausible with reference to the real image collection. Once this network is valid, Mehta et al. [100] train the generator that will produce the depth images using a camera-transformation matrix provided. Finally, the competition of these two agent networks with the help of the associated losses largely inspired by [58] ultimately allow inferring an accurate depth map. It is notable that usually this kind of method based on photorealism generation tends to fail when it comes to generalization. In spite of this fact, [100] was able to show strong generalization capabilities which implicitly means that a physics norm has influenced the network.

To complete this review of various stereo-based methods, we propose a discussion of the trinocular assumption based system from Poggi et al. [117].

This approach is mainly inspired by the different drawbacks of the previously mentioned methods. Indeed, the stereo systems suffer from the same constraints as the acquisition system. Occlusion, object boundaries and left image borders represent a significant challenge for these algorithms. Counteracting, [117] proposes to emulate two views using the input as central image instead of one (for the stereo setup). As shown in Figure 3.14, the effect of such a procedure is, in addition to improving the accuracy, to eliminate the defects of the stereo system.

Many limitations are present due to the use of a stereo based training procedure. Some have been reported like inconsideration of the occlusion or erroneous estimates due

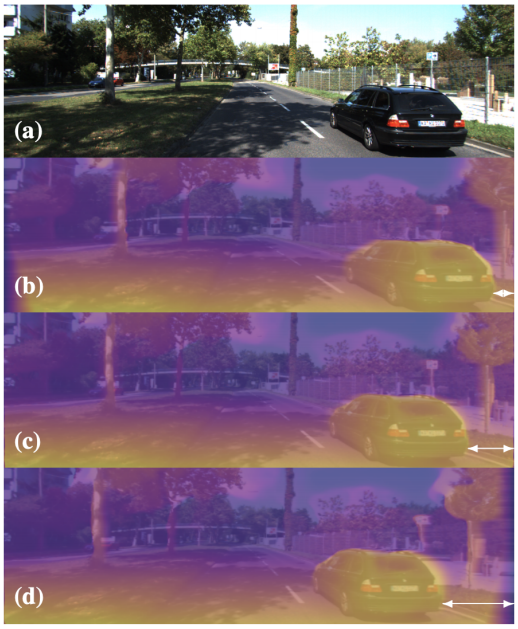


FIGURE 3.14: Poggi et al. qualitative evaluation. a) is the input image, b) and d) are respectively the left and right estimation while c) is the center (image-aligned) estimation.

to second view hallucination. In spite of this, these methods have reported results that have successively become references in the field. However, one factor remains crucial: the data. The main drawback of such methods remains the training requirements. The use of an image pair, in addition to introducing constraints, is heavy and not necessarily available. Also, the necessary acquisition system is necessarily more constraining. This is one of the major reasons why other methods requiring only one camera have been developed. Thus, the core concept of monocular estimation is no longer limited to the inference process but also to the training process.

Self-supervised monocular-based. Now that different stereo-based methods have been discussed, it is possible to shift towards monocular approaches which are nevertheless increasingly widespread. Indeed, the fundamental interest of such an approach remains the training with a single camera. Taking advantage of the use of the temporal dimension, these techniques are based on principles similar to depth estimation by pure image processing by considering the displacement between cameras to infer a disparity.

As a first outstanding contribution, [178] proposed to exploit the displacement between two successive images to infer depth.

As shown in Figure 3.15, they allowed an advance towards a unified framework composed of two networks. One network estimates the depth from a first view at time t. A second network then tries to determine the displacement between this first view I_t

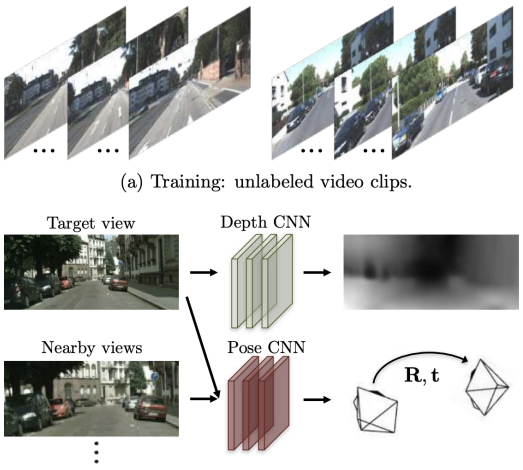


FIGURE 3.15: Zhou et al. [178] Architecture.

and a second one at time $t+1$ denoted I_{t+1}. Thus, similar to the no-learning-involved approach, this displacement $\{R, t\}$ allows the inverse deformation of the secondary image(s). Since a depth map and a camera pose are estimated, it is possible to project an image I_{t+1} on the source view I_t. Thus, this deformation allows computing a photometric error between the projected views and the target image following:

$$L_r = \sum_{<I_1,...I_N>\in\mathcal{S}} \sum_{p} |I_t(p) - \hat{I}_s(p)|, \tag{3.7}$$

with $\mathcal{S}$ the set of consecutive images in the video, p the indexes over pixel coordinates, I_t the original image and $\hat{I}_s$ the warped second image onto the original image.

Zhou et al. also emphasized some limitations of these architectures and proposed a set of overcomes. They determine that a recurrent and problematic situation is the low-texture regions. Their solution is then based on approaches making the same observation (e.g. [54],[58]). They then propose to use an explicit multi-scale approach by forcing it directly into the network. Then, they inject in the cost function a weighted smoothing term. From these two actions, the gradient errors emanating from textureless regions are reduced and allow a sustainable optimization. [178] also define a set of rules that allows the definition of operating cases. Thus, they determine that the scene must not present any moving object, not contain any (dis)occlusion and observing only Lambertian surfaces. These three assumptions guarantee a sane gradient. These rules are critical since these problems will either be addressed in the next contributions described below or in the manuscript presented here.

In the same year, Yang et al. [163] propose a *edge-aware* approach. Contrary to a large majority of approaches, they decide to use normals which are a derivative of the depth. They claim this step allows a more geometrically faithful reconstruction. To integrate

this concept, this contribution integrates layers specialized in this depth-to-normals conversion directly into the DCNN. As shown in Figure 3.16, their architecture allows a bi-directional availability of the normal fields and thus to regularize with it. This supplemental information can be derived from the depth implementing convolutions by considering the neighboring pixels.

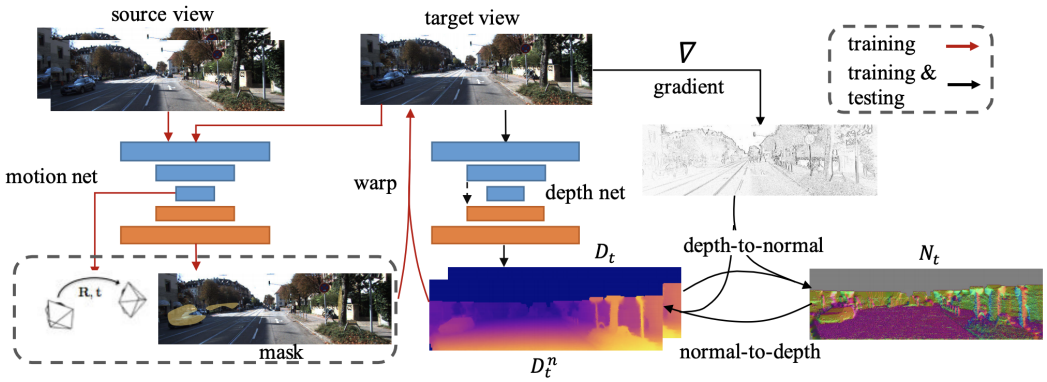


FIGURE 3.16: Framework proposed by Yang et al. [163].

To come back briefly on the aspect *edge-aware*, which constitute also a substantial part, they propose to modify the traditional smoothing function. Moreover, Yang et al. implement the use of the second order derivative allowing to eliminate ambiguities due to the surfaces but also to attempt reducing the bearing effect of the estimates.

In 2018, Mahjourian et al. [97] decide to focus on the geometric aspect by considering errors in 3D space following the additional term L_a:

$$L_a = ||T_t' - I||_1 + ||r_t||_1, \tag{3.8}$$

where I is the identity matrix, T_t' is the best transformation for mapping 3D points from warped view to the original view and r_t is the residual error from the 3D points mapping through ICP. Thus, they operate a loss function composed of discriminant errors in two dimensions, 2D and 3D. This method also introduces a novel game-changing approach allowing to filter the areas of interest and to mask in an innovative way the out-of-bounds pixels to eliminate the remanent errors. In conclusion, their method considers a four-terms loss composed of a 3D cloud point alignment error, a 2D reconstruction error, a smoothing term and a dissimilarity measure.

As shown in Figure 3.17, their qualitative study highlights performance exceeding previous results. [97] also demonstrates quantitatively, on the Kitti Eigen split benchmark, their performance is superior to previous approaches, despite a complexified loss.

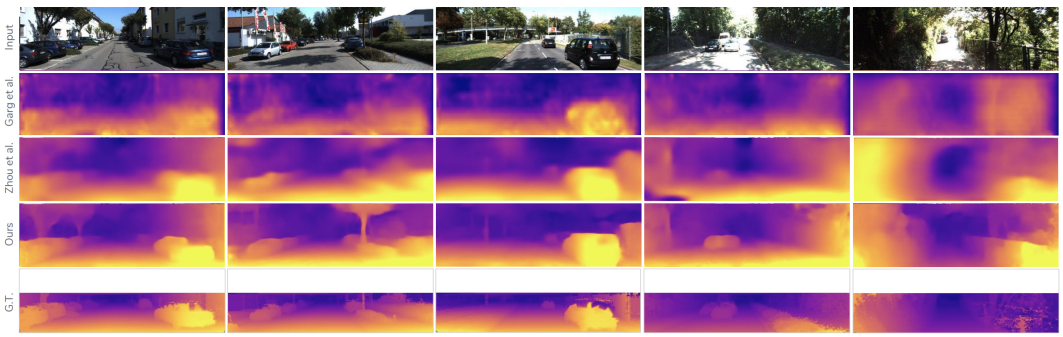


FIGURE 3.17: Mahjourian et al.[97] Qualitative evaluation compared to previous methods. From top to bottom are displayed input, then consecutively, results from [54], [178], [97] and the ground truth.

The same year, GeoNet [164] emerges and allows a triple estimation: the dense depth, the optical flow and the camera pose. To focus on the depth, the architecture named *rigid structure reconstructor* is remarkably similar with the difference that they add a third block named *non-rigid motion localizer* (see Figure 3.18).

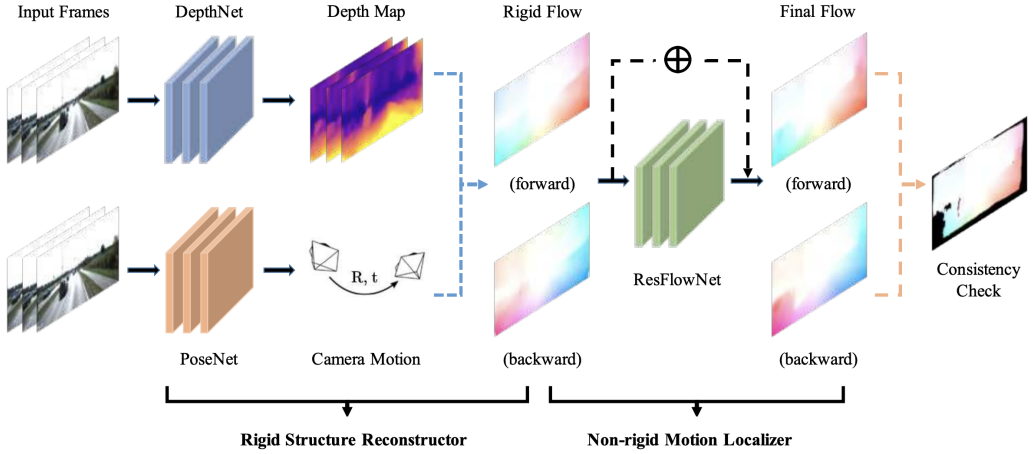


FIGURE 3.18: Geonet [164] Network.

Moreover, the authors attach a new importance to the choice of the loss terms and primarily consider two of the points stated by [178]. While their architecture *rigid structure reconstructor* suffers from the same shortcomings as the previous approaches, the addition of this expansion and the associated loss allows to consider occlusion and non-Lambertian surfaces. Indeed, this part allows estimating a consistency measure which improves the robustness regarding these phenomena. The main concept is then based on two terms L_r and L_s, respectively a photometric error and smoothness error following the equations:

$$L_r = \alpha \frac{1 - SSIM(I_t, \tilde{I}_s^{rig})}{2} + (1 - \alpha)||I_t - \tilde{I}_s^{rig}||_1, \tag{3.9}$$

with SSIM the structural similarity index [149], α a weighting coefficient determined through cross-validation, I_t the initial view and $\tilde{I}_s^{rig}$ the warped through rigid transformation second image. And the smoothness term L_s:

$$L_s = \sum_{p_t} |\Delta D(p_t)|.(e^{-|\Delta I(p_t)|})^T, \tag{3.10}$$

where Δ is the vector differentiation operator, T is the transpose of image gradient weighting and p_t indexes the pixel space.

Next, [147] proposes an originality by claiming that estimation does not necessarily require a learnable pose predictor. Indeed, this contribution proposes relying on the principles of *direct visual odometry* (DVO) to eliminate this sub-architecture to estimate the displacement between two views. Thus, the addition of a DVO [137] pose predictor is used to replace the usual PoseCNN. Since it does not require training, it provides an established relationship between the estimated pose and the depth map. Moreover, this addition does not require any additional effort since it can be derived directly from the reconstructed image which equally serves as a discriminant for the DCNN. This proof of concept opened up the field of possibilities by outlining the possibility to eliminate blocks deemed essential while obtaining precise performance.

Very similar to GeoNet[164], [181] proposes an approach of joint estimation of depth and optical flow. The approach is slightly distinct since the architecture is drastically different. As shown in Figure 3.19, where GeoNet requires only two dissociable pipelines, DF-Net requires four. To summarize, as traditionally, a map is generated using a PoseCNN and auto-encoders allowing the evaluation of a depth consistency loss. In a second estimation pipeline, the pose estimate and the two estimated maps are aggregated to derive two maps respectively forward flow and backward flow. Ultimately, a flow is estimated from the two initial views and these same two maps can be estimated using a FlowNet. Ultimately, flow maps are deduced from the region of interest masks.

All this information can be combined and compared using different terms. The *Forward-backward Depth Consistency Loss* is used to ensure the consistency of forward and backward estimates:

$$L_{FBDCL} = \sum_{p} ||D_t(p) - \bar{D}_t(p)||_1, \tag{3.11}$$

with $\bar{D}_t(p)$ is warped from D_{t+1} using rigid flow from t to $t+1$. A smoothing term imposes smooth transitions while preserving the object boundaries using the modelization proposed in [58]. A photometric error is based on ternary census transform [101,

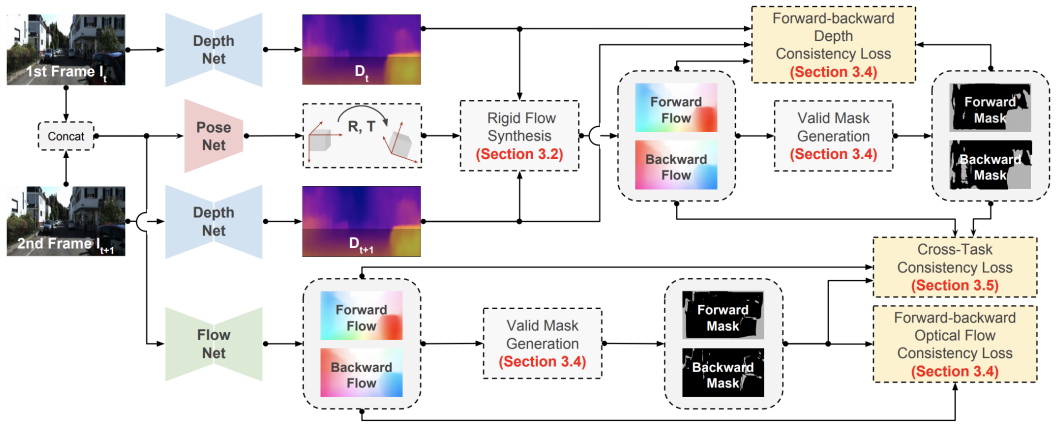


FIGURE 3.19: DF-Net [181] Architecture.

136, 166] to account for real world illumination conditions and allows an evaluation between initial view and projection:

$$L_r = \sum_p \rho(I_t(p), \bar{I}_t(p)|, \tag{3.12}$$

with ρ is the difference function between the initial view I_t and the warped view $\bar{I}_r$, here based on [101, 136, 166]. And finally, *Cross-task consistency* discriminates the differences between the optical flows estimated from the so-called rigid depth maps, and those estimated by the FlowNet:

$$L_{CTC} = \sum_p ||F_{rigid}(p) - F_{flow}(p)||_1, \tag{3.13}$$

where *rigid* and *flow* indices denote for respectively synthetized rigid flow and estimated flow [181], with all the tools deployed, allows to jointly and precisely estimate a refined depth map and an optical flow. On the other hand, it is considerable that the method described here represent a complicated version of GeoNet. However, the **Evaluation** sub-part will demonstrate Zou et al. [181] method is slightly more efficient in estimating a depth map.

Yang et al. [162] promoted the principle of edge learning with their method called LEGO as demonstrated by their designed smoothness term L_s:

$$L_s = \sum_p ||\Delta^2 D(p)||_1 e^{-\alpha|\Delta_d I(p_t)|}, \tag{3.14}$$

which corresponds to the second order smoothness with edge-aware behavior. The method is based on the observation that any planar surface does not have edges until they are at the surface's boundary of it which is similar to image processing based

approaches. This allows them to define surfaces based on the absence of visual cues - here the edges -. As a consequence, it is possible to force the normals to follow the same direction for a defined surface. Based on this concept, this contribution observes a similar architecture, in two blocks, one for the depth and the other for the pose, and adds a third decoder dedicated to the edges. Thus, employing their priors, the loss becomes a four-term minimizable function using in turn the boundary map, the depth map and the *fly-out mask* allowing to eliminate the pixels not remaining in the target view due to the displacement between acquisitions. An illustration of their architecture is available below in Figure 3.20.

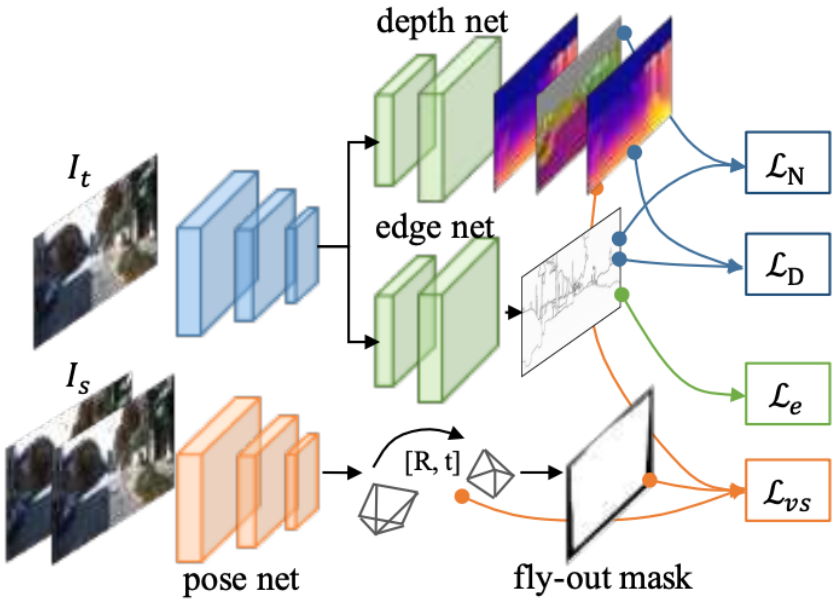


FIGURE 3.20: LEGO Architecture [162].

In 2019, Ranjan et al. [119] propose joint learning of four descriptive images: depth, optical flow, camera motion and motion segmentation. The principal interest of this method lies in the first remarkable use of Counting Collaboration. Like so, this framework allows the joint learning of several collaborative networks in a coordinated manner. This organization is ruled by a discrimination system on the pixels according to their displacement. This core allows an explicit differentiation between moving and static surfaces deriving all previously named descriptive images in a transparent way. The authors define a training procedure including two major steps, competition and collaboration. This ultimately allows them to obtain robust results but also to generate unique descriptive images as the combination of segmented flow in the moving regions and optical flow.

Following with another optical flow computation based approach, EPC++ [93] is an extensively competitive network since it allows many current methods to compare to a

very efficient network. The method starts from the fact that considering static scenes (as previously formulated) is unnecessary if we consider a network can understand the geometry of the scene as a whole. Thus, the authors propose learning jointly the 3D geometry per pixel and the motion.

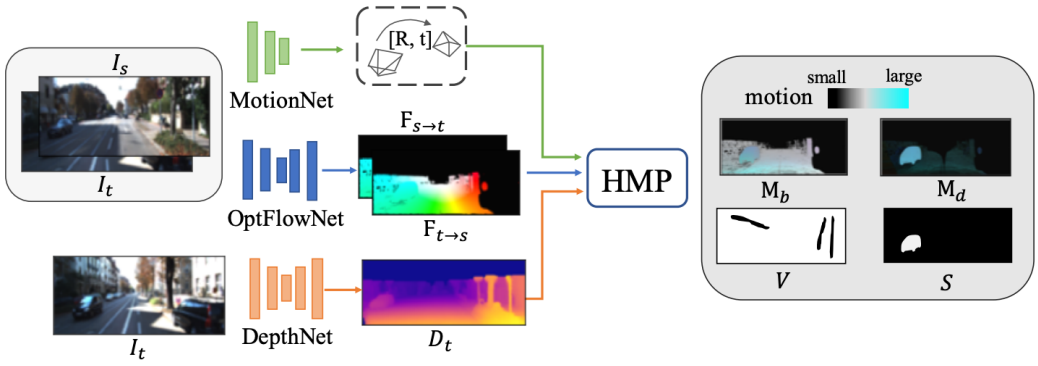


FIGURE 3.21: EPC++ Architecture [93].

As shown in Figure 3.21, this contribution opts for the use of three parallel networks each with a specific and distinguishable task. A MotionNet encoder deduces the pose and two auto-encoders estimate the depth and the optical flow respectively. These three pieces of information, once given to what the authors designate an *Holistic 3D Motion Parser*, allow to compute a segmentation mask for moving objects, an occlusion mask, and two 3D motion maps, one for the background, the other for the moving objects. This modeling allows to regress a precise depth map. Luo et al. also prove a network can learn to understand 3D motion at the pixel level, which emphasizes a high level of knowledge. The displayed performances show that a 3D geometry and motion concepts influenced networks outperforms all previous methods.

In the same year as Luo et al. [93], Casser et al. [26] proposed a method with an original feature. This concept is based on a question: "What if 3D motion was modeled and used to refine a network on the fly? ". This method is motivated by the use of such estimators in an autonomous system. The authors subsequently propose adapting the estimation model in operation.

This contribution is based on the same setup as the significant contributions discussed above. An image sequence is used to retrieve the depth by means of the pose and deformation of consecutive images. In addition, they model a 3D motion object predictor based on the same architecture as the ego-motion predictor. Using an instance segmentation mask, Casser et al. propose to learn on-the-fly the prediction of this instance motion in 3D space. Addressing the recurrent problem of objects changing scale over time, the authors propose to allow the model to learn this phenomenon, thus avoiding the estimation errors involved. At long last, the pipeline learns on

its own over time using short-training strategies on small sequences, hence reducing the discontinuity errors derived from the single-frame estimation. By this strategy, the model is refined as it observes more scenes. As shown in Figure 3.22, this tactic allows, despite a shallower/less complex ensemble, to obtain qualitative results. Moreover, the learning complexity is reduced while making the system widely usable as an autonomous system.

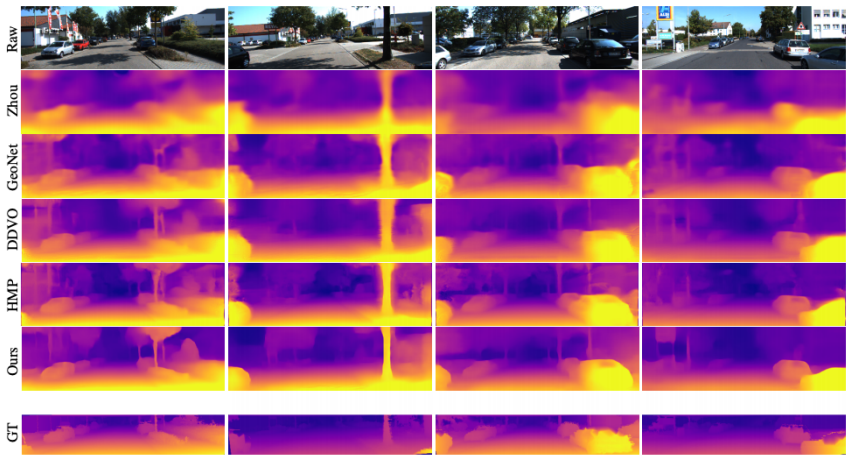


FIGURE 3.22: Quantitative evaluation of Casser et al. [26] method.

Neglecting this online approach, Monodepth v2 [59] is one of the best-known contributions in this field. In fact, it represents, even today, the state-of-the-art in terms of robust depth map estimation. Godard et al. [59] had in the past proposed an outstanding contribution that allowed a great advance. However, this previous method suffered from multiple flaws, one of which was the erroneous consideration of the occlusion. In reaction, the authors proposed Monodepth2, an increment of Monodepth v1. This approach introduces three game-changer concepts:

- A new design of reprojection loss to consider occlusion (*Reprojection*)
- A promising multi-scale UNet-based architecture (*Architecture*)
- A masking strategy removing camera motion-related errors (*Mask*)

Embedding a dissimilarity measure and a L1 distance, the *Reprojection* compare the target view with the warped second view. Although this approach is extremely similar to previous ones, a *Mask* is computed allowing the neglect of static pixels. As a considerable difference, while previous methods involved either optical flow or motion computation, this strategy consists solely into a per-pixel comparison of *Reprojections* computed with different setups. Indeed, the *Mask* is based on pixel-wise minimal photometric error between the target image and either the warped or the original second view. Ultimately, Godard et al. proposed to benefit from UNet [127] skip

connections as such *Architecture* allow for seamless multi-scale computation. Indeed, the authors proposed to extract depth maps at intermediate layers and upsample those. Ultimately, this loss aggregation and computation at full resolution reduce texture-copy artifacts. The upsampling strategy prior to error computation enforces a correct full resolution reconstruction and avoids "holes" in the maps as usually seen in other multi-scale methods. In Figure 3.23, a schematic detailing the key principles is displayed.

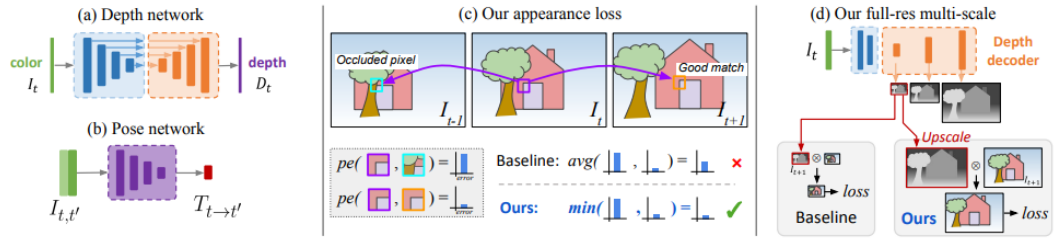


FIGURE 3.23: Monodepth v2 [59] key principles.

This approach was compared to all the methods previously explained. It clearly appears to be the most efficient, despite its reduced complexity. An absolutely remarkable fact in this contribution is the highlighting that the most important thing in a network is the objective function, especially in the self-supervised domain. Thus, a "simple" UNet can outperform other more dense methods by purely handling a properly dimensioned loss. Godard et al. have also proposed a method involving only two networks and requiring neither optical flow nor 3D motion. For all these reasons, this network remains the leading competitor in the race towards an accurate depth map. Therefore, this contribution will be used as a basis and benchmark for the methods developed for this manuscript. These methods will be explained in Chapter 5.

Very recently, Yang et al [160] proposed D3VO, an approach for joint learning of depth, pose and relative incertitude.

As improvement, some methods propose refining the depth map such as [152] with RoutedFusion, or to define a depth-related uncertainty [116]. Furthermore, some methods mobilize other further information to deduce the depth. Furthermore, many other features like semantics [31] or structured light pattern [125] seem to improve the estimation on an ad-hoc basis.

Quantitative Evaluation. This subpart proposes Table 3.2 to compare the different methods explained earlier. All methods have been evaluated on the Eigen Benchmark from KITTI dataset [56] using [46] metrics set.

This table allows us to highlight several facts. First, supervised methods seem to be the most efficient when comparing only the metrics. In fact, these methods are

robust only in the dataset used and are not very generic and therefore usable to a lesser extent in the real world. The single view and stereo based methods show an improvement over the years. In spite of the quantitative metrics being inferior to supervised methods, it is notable that they remain more generic and are therefore more adaptable to unobserved contexts. Overall, the methods combining both possibilities tend to have performances similar to those shown by supervised approaches. Intrinsically, these methods are more generic and thus the performance in real conditions is better, but more importantly, the difference between supervised and self-supervised is so insignificant that it is negligible. In the end, the best approaches rely on terms that do not require ground truth, and this explains why the community tends to create ground truths using these networks rather than using expensive acquisition systems. Ultimately, it is possible to estimate which method is the most efficient. From a metric point of view, D3VO [160] outperforms Monodepth2 [59] but the difference is negligible. The method of Godard et al. is undoubtedly considered as the state of the art since it is widely used as a benchmark since its release. D3VO has not yet been proven as much as Monodepth2 and the difference in performance is so meager that the choice of one method or the other will be subjective and depends largely on the availability of open-source network and weights.

Summary and conclusions. As it has been demonstrated, the scientific community attaches considerable importance to depth estimation. In recent years, researchers have turned to modern deep learning methods. Indeed, this allows less constrained approaches. Regrettably, the use of these techniques ordinarily requires strong assumptions but above all a massive amount of data. Moreover, it is notable that the models trained with certain data are linked to it, and consequently, when the modality used neglects certain phenomena such as specularity, so do the models.

3.3 Summary

In this chapter we have proposed a general literature review on the two global topics that the thesis addresses: segmentation and depth estimation. We have addressed the field of learning based segmentation by reviewing the different backbones and architectures developed over the years. Through a quantitative study we were able to evaluate the performances of the principal methods of the domain on the same dataset. Based on this evaluation, we further discussed the general concepts to explore the essentials of pixel-wise semantic segmentation through deep learning.

In a second step, we proposed an exhaustive review of depth estimation methods. First, we proposed overviewing the different acquisition possibilities and their inherent constraints. Next, we defined the problem of depth estimation exploiting several

images. Thus, this has allowed us to review different methods that are multi-camera or single camera to infer a depth map using consecutive images. Ultimately, we have deeply discussed the various learning-based methods by investigating the significant approaches in the field in chronological order. Through the evolution of architectures and losses, we have been able to exhaustively witness the major advances that have been proposed throughout the years. We have carried out a quantitative evaluation of the leading estimation methods from the beginning of the field to the most recent ones. Based on this analysis, we were able to deduce the most viable methods for real-world estimation.

In conclusion, deep learning has gained much attention in recent years, both for segmentation and depth estimation tasks. Scene understanding is one of the main topics in computer vision, and recent approaches addressing this problem are largely learning oriented. State-of-the-art methods efficiently uses data to infer different outcomes through deep learning. However, whether in segmentation or depth estimation, the methods depend largely on data acquired through color imaging. It is however not negligible that other information could benefit the networks to improve the performance of the diverse approaches. Thus, whether the exploitation of different modalities by deep learning methods can lead to better results remains an open question in needs of exploration.

TABLE 3.2: Quantitative evaluation. Comparison of depth estimation method to on KITTI 2015 [56] the Eigen split. Best results in each category are in **bold**; second best are underlined. D is for depth supervision, D* for auxiliary depth supervision. M and S corresponds respectively to mono and stereo self-supervision.

		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Approach	Train	lower is better				higher is better		
Eigen [46]	D	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu [91]	D	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Klodt [78]	D*M	0.166	1.490	5.9998	-	0.778	0.919	0.966
AdaDepth [81]	D*	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Kuznietsov [82]	DS	0.113	0.741	4.621	0.189	0.862	0.960	0.986
DVSO [161]	D*S	0.097	0.734	4.442	0.187	0.888	0.958	0.980
SVSM FT [94]	DS	0.094	0.626	4.525	0.177	0.891	0.965	0.984
Guo [61]	DS	0.096	0.641	4.095	0.168	0.892	0.967	0.986
DORN [53]	D	**0.072**	**0.307**	**2.727**	**0.120**	**0.932**	**0.984**	**0.994**
Zhou [178]	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [163]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [97]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [164]	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [147]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [181]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [162]	M	0.162	1.352	6.276	0.252	-	-	-
Ranjan [119]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [93]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth [26]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth2 w/o pretraining [59]	M	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2 [59]	M	**0.115**	**0.903**	**4.863**	**0.193**	**0.877**	**0.959**	**0.981**
Monodepth2 (1024 x 320) [59]	M	**0.115**	**0.882**	**4.701**	**0.190**	**0.879**	**0.961**	**0.982**
Garg [54]	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50 [58]	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
StrAT [100]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.970
3Net (R50) [117]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net (VGG) [117]	S	0.119	1.201	5.888	0.208	0.844	0.941	**0.978**
SuperDepth (1024 x 382) [115]	S	0.112	0.875	**4.958**	**0.207**	0.852	0.947	0.977
Monodepth2 w/o pretraining [59]	S	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Monodepth2 [59]	S	**0.109**	**0.873**	4.960	0.209	**0.864**	**0.948**	0.975
Monodepth2 (1024 x 320) [59]	S	**0.107**	**0.849**	**4.764**	**0.201**	**0.874**	**0.953**	0.977
UnDeepVO [88]	MS	0.183	1.730	6.570	0.268	-	-	-
Zhan FullNYU [168]	D*MS	0.135	1.132	5.585	0.229	0.820	0.933	0.971
EPC++ [93]	MS	0.128	0.935	5.0111	0.209	0.831	0.945	**0.979**
Monodepth2 w/o pretraining [59]	MS	0.127	1.031	5.266	0.221	0.836	0.943	0.974
Monodepth2 [59]	MS	0.106	0.818	4.750	0.196	0.874	0.957	**0.979**
D3VO uncertainty [160]	MS	0.101	0.772	4.532	0.190	0.884	0.956	0.978
D3VO ablation [160]	MS	0.105	0.791	4.650	0.193	0.878	0.957	**0.979**
D3VO full [160]	MS	**0.099**	**0.763**	**4.485**	**0.185**	**0.885**	**0.958**	**0.979**
Monodepth2 (1024 x 320) [59]	MS	**0.106**	**0.806**	**4.630**	**0.193**	**0.876**	**0.958**	**0.980**

Chapter 4

Deep Polarization-Based Semantic Segmentation

This chapter is dedicated to segmented map estimation from a polarimetric image. As described in Chapter 3, segmentation represents a rising field in computer vision. In essence, it is a pixel-level classification of images. Moreover, it is possible to add the semantic dimension to the problem to formulate interconnections of classes and to add a part of understanding.

In the framework of the presented work, we propose to use polarization as an input modality. While the vast majority of algorithms rely on so-called standard modalities, such as color imaging, we want to show that a change of space can modify, even simplify, the problem of semantic segmentation of complex urban scenes. As a consequence, we propose reviewing the whole pipeline from the database to the estimation via CNN through the creation of an adequate augmentation procedure. This process remains crucial due to the poor availability of polarization information assessing such scenes.

This chapter is constituted as follows. We briefly introduce our scenario and motivations in Section 4.1. In Section 4.2, we discuss the modality constraints as well as the choices made to make the polarization usable. Then, in Section 4.3, we focus on the augmentation process. Finally, Section 4.4 and Section 4.5 will respectively discuss the networks used and the experiments. Section 4.6 summarizes our work.

4.1 Introduction

Recently, pixel-wise semantic segmentation (PwSS) has achieved great success especially thanks to the increased computing capacity of modern machines. Yet, a less expensive solution is very little investigated, the moderate use of data and computing power. The formulation of the problem is then marginally different. Instead of feeding the networks with a vast amount of data and/or increasing their size, the problem is estimated upstream to discriminate the network. It is nevertheless notable that some

contributions are tackling the problem, such as few-shot learning [43, 134]. However, a very small minority transpose the data space to apply a constraint ahead of the network rather than impacting the processing core. This could be explained by the limited data availability of unconventional modality.

We aim to propose a complete PwSS pipeline to describe complex urban scenes based on polarimetric imagery. Here, "complex" defines the possibility of adverse weather conditions and scenes subject to specularity. We start with an observation: urban scenes are prone to specular reflections. Indeed, whether it is the metallic paint of cars, road signs, the presence of transparent windows or even the road under rainy conditions, these surfaces highly reflect the light. Rather than defining these reactions in a space that saturates in these cases, we propose to implement a modality in which these phenomena are defined. We then propose using polarization instead of color imaging which would export its constraints relative to these occurrences. We therefore rely on the polarization modality to simplify the segmentation problem in urban environments. Thus, we need deconstructing all the achievements in the PwSS domains to reconstruct a polarization-adapted pipeline: the dataset, the image representation, the augmentation and the estimation of the results.

In practice, the methods require massive datasets like KITTI [56] or Cityscape [36] that deliver both corresponding input image with a reference segmented map. These mass requirements are onerous to achieve when implementing a non-conventional modality, which motivates a more measured use of the data. A straightforward approach would be to acquire a large amount of data and then annotate it to enrich the generalization capacity of the network. Since our approach is alternative, we propose to acquire a limited number of images and use the augmentation to enrich the set. Although augmentation has been popularized for its advantages in terms of problem generalization and overfitting prevention, it is only viable for interpolable images. In short, as soon as an image is directly related to the physics of the scene, this process is not applicable since it will alter the validity of the observation. On the other hand, since the color information is interpolable and does not report any alteration as a result of it, then the augmentation is valid for this type of image. In line with the initial objective of the augmentation, we define a group of transformations applicable to polarimetric imaging and in particular, we focus on the physical adequacy of these. Since it is necessary to benchmark the proposed solution, we tackle the possibility of evaluating our method and comparing with the conventional approaches. Aiming this objective, we have aggregated our dataset taking into account that each polarimetric image and segmentation pair correspond to an RGB image and segmentation pair.

4.2 Modality-related constraints

As stated previously and more precisely in the Section 2.1, with polarization follows a set of constraints of its own. In this section, the numerous constraining items of the modality will be discussed. First, the formulation of the problem with respect to the data will be stated. Then, the representation aspect of the imagery will be addressed. This section will be concluded by the methods operated to aggregate and validate Polabot[1] dataset.

4.2.1 Formulation

With the objective of developing an algorithm relying on reflection to define urban scenes, it is necessary to formulate the problem. As mentioned before, we start from a statement: urban scenes are mostly subject to specular reflections. This phenomenon is even amplified when the weather conditions are unfavorable. Plus, although standard algorithms have acceptable performances in common environments, they tend to fail when they are in the presence of specular reflection, especially when observing puddles. Moreover, the methods are not helped by the usual datasets since they generally do not contain any of these occurrences.

As shown in the figure 4.1, algorithms that perform well in favorable cases tend to miss their estimate when they observe specularity. This kind of erroneous estimation can be problematic especially if an autonomous vehicle algorithm depends on these estimates.

However, it is necessary to bound the polarimetric images to obtain the desired effect: discriminate and simplify the problem. Unprocessed, polarization images are not really usable. As described in the Section 2.1, the images are sparse and not particularly descriptive. If the raw images are preserved, it is challenging to constrain the problem since the specularity is only defined by the pixel saturation. Although it is indeed characterized, it is necessary to transpose these images through the Stokes parameters to implicitly extract the informative part of the images. Consequently, to consider images exploitable by a CNN, it is necessary to decide on an exploitable representation image.

4.2.2 Image Representation

One critical point of machine understanding approaches is to have representative and understandable images. Image representation is omnipresent in the field of computer vision. On the other hand, it has become progressively transparent. The colorimetric images are subject to an image representation allowing to pass from the lowest level

[1] Available at: http://vibot.cnrs.fr/polabot.html

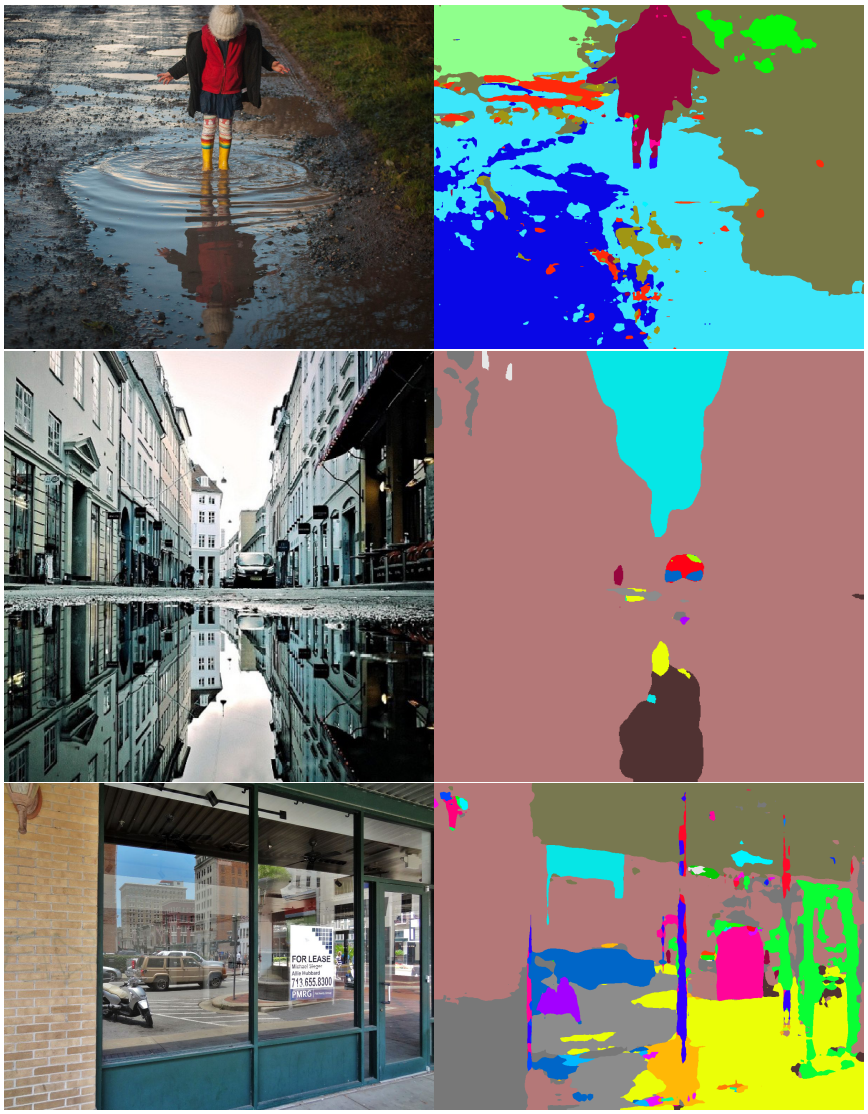


FIGURE 4.1: Segmentation results from [177] showing erroneous estimation on specular surfaces.

(raw) the Bayer matrix, to a composition in three channels, traditionally RGB. Polarization is not exempt from this, however, since the modality is unconventional and not as widely used, it is necessary to define advantageous rules to the future processing pipeline.

Our goal is to have a reliable and representative image while making it deep learning-friendly. Thus, according to our problem the images must intrinsically represent polarization and by extension specularity while preserving differentiable textures allowing the network to learn. Starting from the raw images, we recover the informative part of the images by computing the Stokes parameters:

$$S = \begin{pmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} P_H + P_V \\ P_H - P_V \\ P_{45} - P_{135} \\ P_R - P_L \end{pmatrix} = \begin{pmatrix} P_0 + P_{90} \\ P_0 - P_{90} \\ P_{45} - P_{135} \\ 0 \end{pmatrix}, \tag{4.1}$$

where s_n is the n^{th} Stokes parameter and P_Θ is the dense polarization image corresponding to the Θ angle oriented polarizer. It is notable that the P_θ images must be dense and therefore that between the raw imagery and these, it may be necessary to interpolate the images between pixels to densify them. This will depend mainly on the image resolution of the camera. Also, recall that in the absence of a quarter-wave plate, no circular polarization s_3 is acquired. From these, it is possible to easily find the three strong descriptors of the polarization, the intensity ι, the polarization angle α and the degree of polarization ρ:

$$\iota = \frac{P_0 + P_{45} + P_{90} + P_{135}}{2}, \tag{4.2}$$

$$\rho = \sqrt{\bar{s_1}^2 + \bar{s_2}^2}, \tag{4.3}$$

$$\alpha = \frac{1}{2}\tan^{-1}\left(\frac{s_1}{s_2}\right), \tag{4.4}$$

where $\bar{s_n}$ is the n^{th} Stokes parameter normalized by s_0.

Although there are a multitude of possibilities to combine these three images, it is necessary to consider their nature to obtain a representative result. While ι is related to the texture of the image, it is a standard grayscale image, α and ρ are two complementary images respectively the angle of polarization and its "strength". It is then necessary to aggregate these information to maintain the integrity and especially the interest of the polarization. Thus, the raw concatenation would strongly reduce the interest of the imaging for PwSS applications.

Also, to avoid imposing a relearning of the organization of the pixels for the network, but also to benefit from the advantages of the transfer learning, it is necessary to move towards a three-channels structure.

In this context, we propose representing these images in three HSL channels as proposed by [154] that will finally be transposed in RGB color space[2]. Indeed, this intermediate format allows infusing particular properties to the image, while keeping a simple transposability from HSL to RGB. As shown in the Figure 4.2, both modelization formats include a singular behaviour.

[2]Converting toolbix available at: https://github.com/BlanchonMarc/InterPol

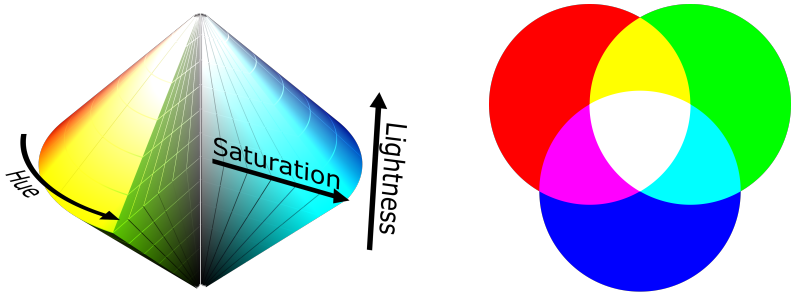


FIGURE 4.2: HSL and RGB color models.

The RGB model is widely operated since it is direct and additive. In contrast, Hue Saturation and Luminance (HSL) is somewhat different due to its cylindrical and cyclic nature. This model enjoys many advantages, but the prominent attraction for polarimetry is that its channels are neither bounded nor processed in the same way. They are therefore independent but complementary for visual representation. The key advantage of the HSL representation is its channel coding. Indeed, the three separate channels are quantized differently and allow a direct adaptation with the polarization. Hue represent a cyclic value between 0 and 360 which comfortably accommodates a 2π-periodic value and thus the polarization angle. The saturation represents the purity of the color indicated by the Hue. It is a percentile value that allows an analogy with the degree of polarization. The correspondence between the color and its intensity of this mode is convenient to use with polarization and the proper properties of α and ρ. Finally, L, the luminance, can easily accommodate the intensity value ι since its utility is very similar to the encoding of a texture. In conclusion, the polarimetric images will be mapped as:

$$H \longrightarrow 2 * \alpha, \quad S \longrightarrow \rho, \quad L \longrightarrow \frac{\iota}{255}. \tag{4.5}$$

As shown in the Figure 4.3, it is remarkable that the images are peculiar and that their hue is unnatural. Indeed, the texture is present but colors are affected according to the α orientation. Also, the intensity of the color is decided by ρ. In the end, the more a zone is colored, the more it is polarized. And according to the color, the various angles are observable in a transparent way.

After this representation, a three-channel image is obtained which could be used as input for a DCNN. Remarkably few models are trained with HSL-mode data, and the convention is until now RGB for the training task. Rather than having to end-to-end train a network by training the data encoding, it is more convenient to use RGB. Indeed, this will allow using pre-trained networks and thus to benefit from approved

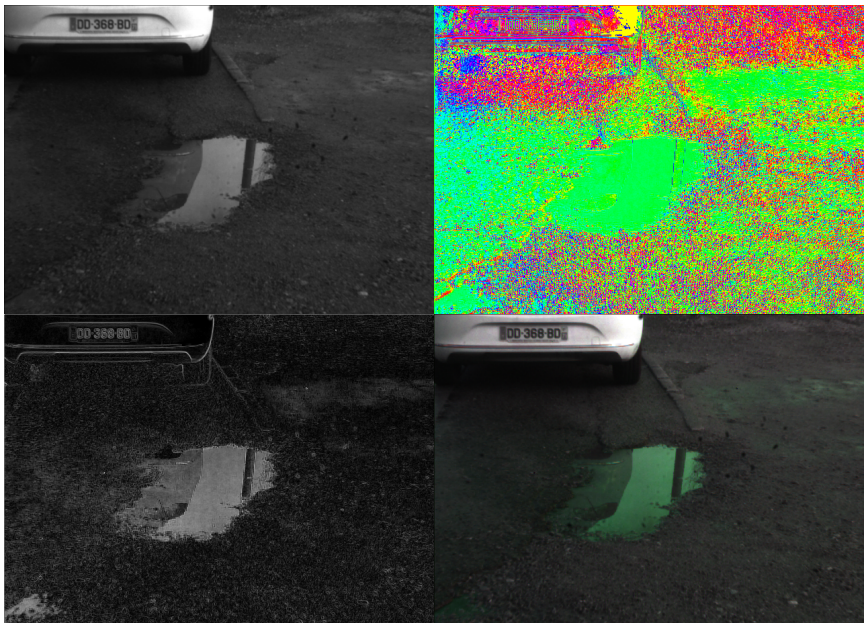


FIGURE 4.3: HSL representation of polarization. At the top left is the intensity ι, then at the top right the polarization angle α. At the bottom left is the polarization degree ρ and at the bottom right is the combination of the three informative images in HSL format and the equation 4.5.

transfer learning methods. As follows, the choice is considered to transpose the HSL images in RGB while keeping the particular display properties following the system:

$$C = (1 - |2L - 1|) \times S, \tag{4.6}$$

$$X = C \times (1 - \left| \frac{H}{60°} \mod 2 - 1 \right|), \tag{4.7}$$

$$m = L - \frac{C}{2}, \tag{4.8}$$

$$(R', G', B') = \begin{cases} (C, X, 0) & , 0° \leq H < 60° \\ (X, C, 0) & , 60° \leq H < 120° \\ (0, C, X) & , 120° \leq H < 180° \\ (0, X, C) & , 180° \leq H < 240° \\ (X, 0, C) & , 240° \leq H < 300° \\ (C, 0, X) & , 300° \leq H < 360° \end{cases}, \tag{4.9}$$

$$(R, G, B) = \Big((R' + m) \times 255, (G' + m) \times 255, (B' + m) \times 255 \Big). \tag{4.10}$$

After all these transformations to interpret the polarization, a polarimetric RGB-coded image is obtained.

4.2.3 Dataset

The dataset is a critical element when using machine learning algorithms. The scientific community agrees it is one of the most crucial points if not the most important. Conventionally, since DL-based methods are greedy, the more data the better. This is accommodating when using a widespread and easily acquired modality. On the other hand, when addressing non-conventional modalities, the major obstacle for the community is the need for a massive amount of data.

Our approach is somewhat alternative although also constrained by modality. The idea is to obtain viable results with the limited data available. First, it is necessary to acquire data in urban areas and under adverse conditions.

We constituted Polabot[3], a multimodal oriented dataset composed of 3 synchronized modalities: polarization, color and near infrared, acquired with the cameras referenced in Table A.1. As a final image collection, it is composed of 178 multimodal aligned, synchronized and annotated urban scene images with eight unique classes: unlabeled, sky, water, windows, road, car, building and none. Unlabeled corresponding to segmentation errors during manual annotation and none being the areas defined as irrelevant for our application. The different scenes propose several complex scenarios composed of puddles or buildings' windows which are often incorrectly estimated in usual methods. Since this collection is very restricted and insufficient to train a sustainable model, the augmentation presented in Section 4.3 represent an important requirement.

Yet, the acquisition of this dataset involves two major challenges prior to the augmentation: the synchronization of images from several sources and their alignment which will be respectively presented in 4.2.3.1 and 4.2.3.2.

4.2.3.1 Synchronization

One of the problems of acquisition of any multimodal and multifocal system remains the synchronization of images. To be specific, desynchronization may cause a detrimental effect when images are captured at high speed. A shift and the images do not exhibit sufficient correlation due to the distance traveled between two image triggers. In our case, the de-synchronization is impactful for the segmentation problem but also specifically for the cross-modality performance comparison. Indeed, as previously stated, this dataset allows us to train networks but also and especially to have a point of comparison to determine the advantage of a modality over another for

[3]Available at: http://vibot.cnrs.fr/polabot.html

our application. To prevent desynchronization errors, we have developed a rule-based system[4] to ensure the smallest possible cross-modality shift.

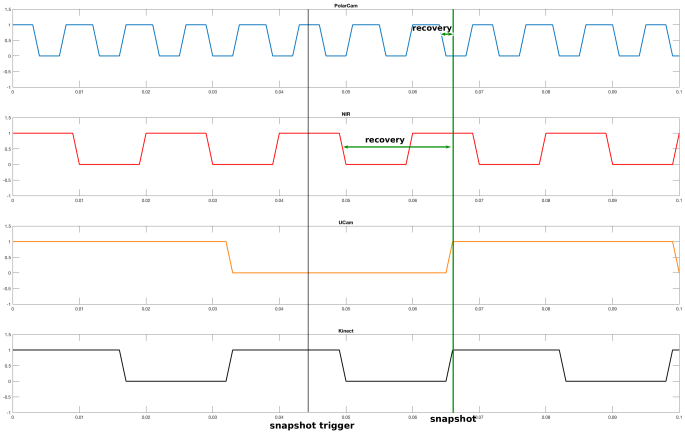


FIGURE 4.4: Recovery strategy-based synchronization.

The principle is to consider snapshot trigger signal. The cameras have a discontinuous image stream and this can be represented by square signals. As shown in the Figure 4.4, each rising edge corresponds to an image trigger (replacement of the previous image). The strategy is simple, at each artificial trigger, it is necessary to wait for the following acquisition of the slowest camera and then use a recovery strategy.

In conclusion, this simple rule-based strategy allows to drastically reduce the problems related to desynchronization and therefore to eliminate artifacts such as shift, motion blur or displacement between images.

4.2.3.2 Multimodal Alignment

Image alignment represents a well-known field. Very well defined, this problem is for the most part solved. However, when addressing the idea of multimodal alignment, known methods can be inefficient, especially when the images to be aligned are non-interpolatable. This is indeed the case handled in this section. The concept is to efficiently align an RGB image with a polarimetric image. It is significant to experience a precise framing since the correspondence at pixel level must be exact. Thus, a valid comparison can be performed between the segmentation results from independent networks addressed on different modalities.

Homography-based. Homography [3, 5] is one of the most popular methods to estimate the displacement between two images. Assuming that two images are on the same plane in space then they are linked by the homography. There is on top an underlying assumption that the two images must be in the same feature space.

[4] Available at: https://github.com/BlanchonMarc/Ros_AcquisitionFromTopics

This is not the case between colorimetric and polarimetric imaging. However, it is completely possible to find a common space in grayscale intensity. Then, just find and

FIGURE 4.5: Inacurate alignment through homography estimation.

map the corresponding feature points on each image and deduce the transformation. The result is R and t, the rotation matrix and the translation vector that allow to transform an image to align it with another one. This process is remarkably rapid and practical, but it suffers some disadvantages which are induced by the hypotheses stated above. Since the polarization is not interpolable, it is impossible without extra-computation to preserve the physical properties of the image. Therefore, it is mandatory to wrap only the colorimetric image and map it to the polarization. Secondly, it should be noted that the polarimetric and colorimetric intensities are supposed to be theoretically almost identical, but in practice, inconsistencies between them lead to erroneous estimates. For instance, the Figure 4.5, illustrates the case of an inaccurate polarization-RGB alignment.

Since there are inconsistencies between the spaces, it is then necessary to go one step further to allow an accurate alignment.

Homography-initialized dense alignment. Since a simple transformation through homography proves to be inefficient due to the nature of the images, it is possible to perform a dense alignment to refine it.

Directly inspired by image-based visual servoing methods, the proposed concept[5] allows to refine the parameters R and t iteratively. Starting from the reference polarimetric image I^* and the colorimetric image I_c, an error ϵ is computed such that:

$$\epsilon = I^* - I_c. \tag{4.11}$$

[5]Available at: https://github.com/BlanchonMarc/process-vibotorch

An interaction matrix L [27] can be determined as follows:

$$L = \begin{bmatrix} -1/Z & 0 & \Delta_x/Z & \Delta_x\Delta_y & -(1+\Delta_x^2) & \Delta_y \\ 0 & -1/Z & \Delta_y/Z & 1+\Delta_y^2 & -\Delta_x\Delta_y & -\Delta_x \end{bmatrix}, \tag{4.12}$$

with Z the distance between the plane and the camera, Δ_x and Δ_y the gradient along x and along y respectively. From the pseudo inverse of L and ϵ is derived the velocity vector v such that:

$$v = -\lambda L^{+}\epsilon, \tag{4.13}$$

with λ a scalar corresponding to the gain of the velocity. Thanks to the velocity vector, the Lie algebra and the exponential map [124], the increments $\hat{R}$ and $\hat{t}$ can be evaluated according to:

$$\hat{R}, \hat{t} = \text{exponential map}(v). \tag{4.14}$$

Finally, using the pose increments and an initial homography H, a new transformation matrix $\hat{H}$ can be estimated:

$$\hat{H} = K \times \left[(R\hat{R}) + (t + \hat{t}) \times \frac{n^T}{d} \right] \times K^{-1}, \tag{4.15}$$

with K the intrinsic parameters of the camera acquiring the image I_c.

FIGURE 4.6: Before and after dense alignment between polarimetric and RGB intensity images.

It is then possible to minimize the function through an iterative process or, in our case, to define an error target ϵ. Indeed, since the starting hypothesis has invalidated the possibility of photometrically comparing the intensity images, we must empirically set an error threshold that will satisfy the alignment needs. In addition, there is the parameter Z that must be estimated or evaluated using a grid search approach. Finally, as shown in the Figure 4.6, dense alignment seems to represent a viable solution. It is however important to mention the computational requirements are

high and that this kind of process is heavy. Despite its efficiency, it is a matter of obtaining a tradeoff between a potentially precise alignment and a computation time consistent with the application. In our dataset construction application, there is no online alignment which does not subject our pipeline to the slowness induced by this kind of process.

4.3 Augmentation

Augmentation has a major interest when dealing with a limited amount of data. This procedure allows to generalize the DCNN models and avoid overfitting [114]. However, this process has been principally designed to increase the population of interpolable modalities, but polarization cannot straightforwardly benefit from it by its nature.

Since the information acquired through a polarimetric sensor is intrinsically dependent on its pose, seeking a transformation without adaptation is invalid. For these reasons, we have explored the augmentation operations applicable to polarization under any conditions. Thus, we have estimated that the two simple adaptable operations are the rotation and the flipping respectively described in 4.3.1 and 4.3.2. Equally the translation can be operated but do not require any further processing while using standard lens. Finally, 4.3.3 will be dedicated to the final reproducible procedure to multiply the polarimetric images.

4.3.1 Rotation

Rotation remains a very common operation to increase the number of images in a dataset. In practice it is enough to apply a rotation to the image and a new one is obtained. However, since the polarization information is relative to the camera pose, it is necessary to modify the operation.

The prerequisite for augmentation is to create new realistic images. If we transpose an image rotation to the sensor point of view, this procedure is equivalent to rotating the camera. This is where the constraint to polarimetry comes from, since pivoting the sensor means changing the camera pose. Thus, altering the orientation of the camera changes the organization of the camera's pixel grid as shown in Figure 4.7. And the objective is to reorganize these pixels so that the polarization angle regains its physical integrity.

In Figure 4.7, the illustration on the left shows the initial polarizer grid, then directly on the right, the effect of a $90°$ rotation on this same grid.

The illustration on the far right shows the prerequisite for the image to be unaltered. A regularization operation is therefore necessary.

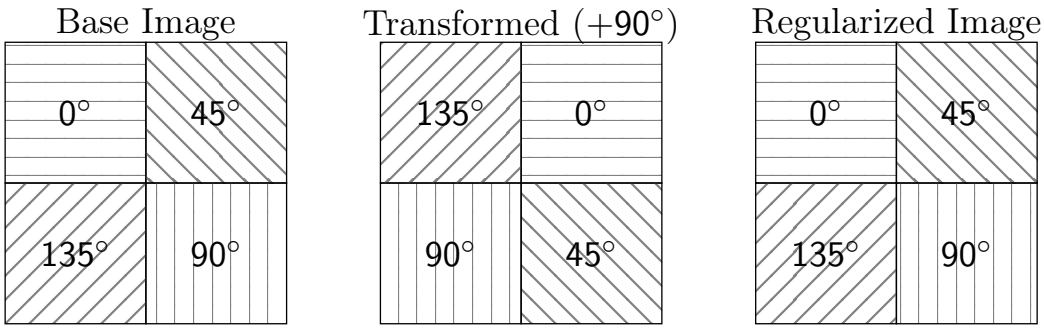


FIGURE 4.7: Illustration of the pixel grid of a rotated DoFP polarimetric camera.

We propose applying a rotation to the image, and, to regularize, to apply an inverse rotation to the polarization angle. Ultimately, it brings back both the physics of the scene while keeping this transformation to create a new valid image.

Let θ be the rotation angle applied to the camera, R_θ the rotation operation and H the hue channel of the image (which as a reminder corresponds to the angle of polarization α):

$$H_{\text{rotated}} = R_\theta(H_{prev} - 2 * \mathbf{1}\theta). \tag{4.16}$$

FIGURE 4.8: Step-by-step rotation applied to polarimetric image.

Since the general idea is to reobtain the viability of the scene physics, this regularization step reorganizing the pixels of the DoFP grid is mandatory. However, as shown in the Figure 4.8, this operation impacts the visual aspect of the image. As shown in the Section 4.2.2, the images are modulated on the HSL model. As a reminder, we indexed the α polarization angle on the Hue channel. Since the equation 4.16 applies a modification on this channel, then the global hue of the image changes. This is consistent with the fact that the change in pose implies a reorientation of all angles. To guarantee the periodicity of the angle but also to respect the HSL format, an

additional modulo step is necessary:

$$H_{final} = H_{transformed} \pmod{360} \tag{4.17}$$

As a partial conclusion, it is possible to deduce this operation is valid and respects the modality thanks to this visual indicator. In addition, the equation 4.17 allows a correct bounding as well as taking advantage of the HSL mode. Thus, any rotation is applicable without altering the information.

In addition, the rotation represents a physically verifiable operation. It does not imply an impossible transformation to the sensor and therefore it is possible to verify the viability of the regularization by performing acquisitions. As shown in the Figure 4.9, this manipulation was performed to physically verify if the polarization information maintained by our equation.

(A) Acquisition 0° (B) Acquisition 90° (C) Augmented 90°

FIGURE 4.9: Experimental validation of rotation process. Images (A) and (B) are acquisitions with physical rotations applied to the camera. Image (C) is the result of augmentation applied to image (A). Note the correct recovery of angle information.

The experimentation consists of two acquisitions: one image is captured with the camera oriented normally and a second one with a rotation of 90°. We can then perceive the change in hue that this rotation implies on polarized regions of interest. In addition, we can use the non-rotated image and transform it using our process. Thus, we obtain two images: one where the sensor has been physically rotated, and the other that has been artificially rotated. Subsequently, it is possible to see the shades are approximately the same. A difference is present, but this is due to the parallax effect (imperfect rotation around the depth axis) and since the sensor used is not square. Nevertheless, this experiment shows the validity of the rotation process applied to polarization.

In conclusion, this experimental proof validates the rotation process and its associated regularization equation.

4.3.2 Symmetry

Symmetry is another frequently used transformation for augmentation. It allows "disorienting" the algorithms so that they do not get used to positional criteria, much like rotation. The significant drawback of this transformation is that, unlike rotation, it is not physically verifiable. It goes without saying it is impossible to reverse the scene or to turn the sensor on itself. This is why we have developed a method to express the impact of this operation and thus validate it.

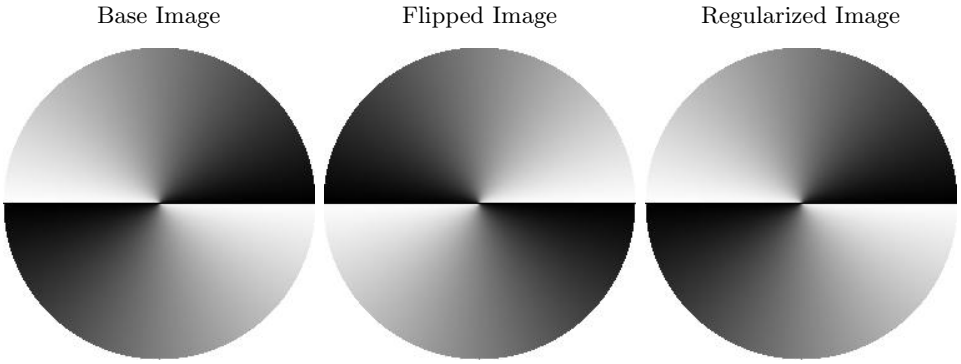


FIGURE 4.10: Illustration of the flipping procedure. From left to right, the base image, a flipped image and the regularized image

The figure 4.10 shows a circle separated into two halves. This illustration is specific because each point of the circle in the image represents the corresponding angle with the center as reference. This particular image can be considered as a synthetic image of the angle of polarization of the light reflected by a specular ball. Consequently, flipping operation should not alter this measure to preserve 3D coherency with the scene. If we consider only the top of the left circle, from left to right, we have a circular gradient ranging from 180 degrees to 0 degrees. Therefore, this image represents the full range of possible linear polarization angles in an image. Knowing that the image hue channel is periodic at 360 degrees, the reversal consists of inverting the axis selected for the transformation. Taking advantage of the periodicity of the angle and the selected format, then the transformation can be performed as follows:

$$H_{\text{flipped}} = -H_{prev}. \tag{4.18}$$

Clearly, the physical operation of flipping is operated on the image in addition to regularization. Thanks to the equation 4.17 in addition to equation 4.18, we possess the possibility to use the representation mode to our advantage. This operation makes the format valid and thus, the accumulation of these two manipulations is necessary to verify the simulation proposed in figure 4.10. Finally, the equation has this double

role of regularization for symmetry and buffer to guarantee the validity of the color space.

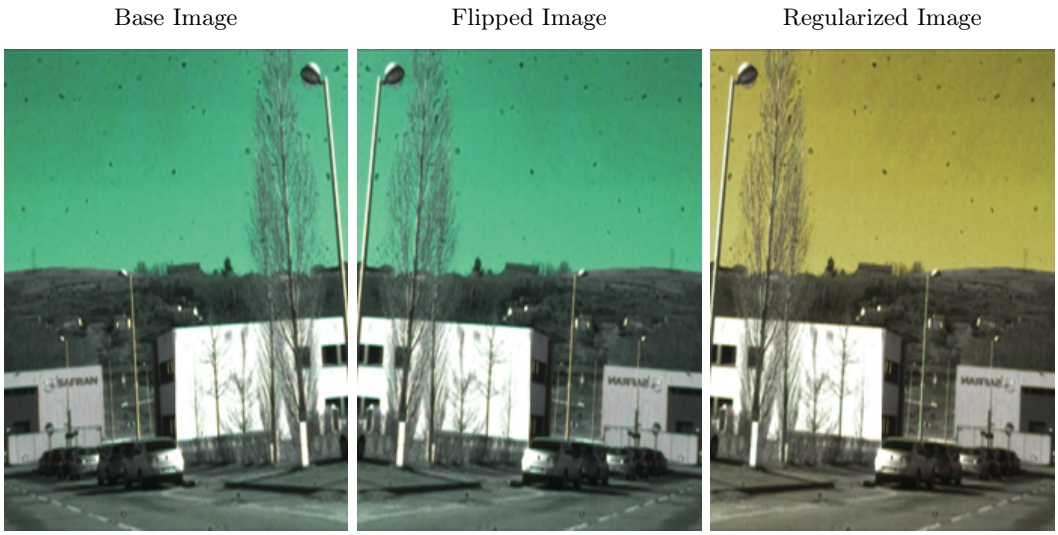


FIGURE 4.11: Flipping operation on real image and impact. From left to right, initial image, the flipped image and the regularized image.

It is then possible to perform these transformations on real images as shown in the figure 4.11. Since it is impossible to physically observe this kind of transformation, the visual hue indexer will be the sole indicator of an integral transformation. Despite this, it is equally possible to check if the hue is well "inverted" by checking on the HSL representation space (shown in figure 4.2) if the colors are properly symmetrical.

In conclusion, we note symmetry is another valid augmentation possibility to obtain realistic images of polarization thanks to our regularization process.

4.3.3 Final procedure

In this part the final procedure of augmentation will be explained, but first of all a point must be addressed: what happens to the other channels. Indeed, in the two previous parts on the different transformations, only the angle was modified. This can be explained by the invariance of the other two channels, the degree of polarization and the intensity, with respect to the pose. The observation is simple, the reflection strength or the texture does not change according to the pose, such as, a reflective object remains reflective regardless of the camera orientation. Similarly, the texture is unaffected by a change in pose similarly to usual colorimetric image processing. Thus, we conclude only the angle must undergo regularization operations while the other channels will only be affected by the rotation or flipping type transformation.

This observation allowing to conclude on the augmentation pipeline, it is then possible to define the final process to augment the dataset. Augmentation is typically composed of multiple operations performed simultaneously according to probabilistic

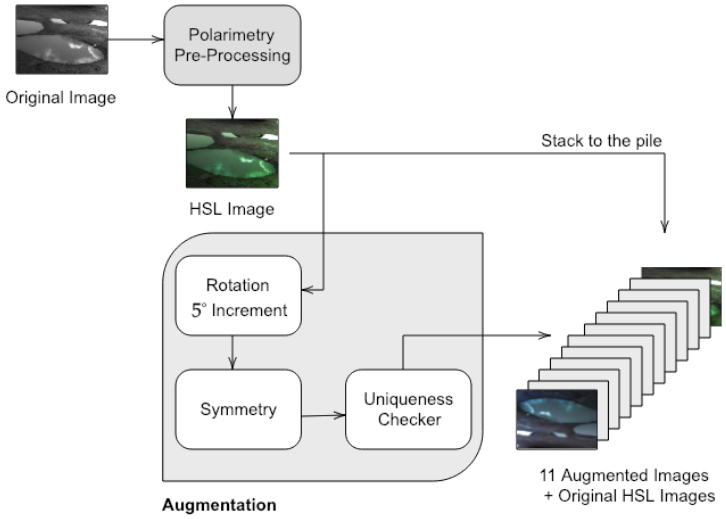


FIGURE 4.12: Illustration of the augmentation per image procedure. This process is repeated for each image in the original dataset to obtain a consistent large dataset. Then, the entire set of augmented images is shuffled.

or empirical rules. We therefore propose, with our transformations, to compose in the same way. Thus, we have the ability to cumulate rotation and flipping, at different increments or orientation. As shown in the figure 4.12, from a polarimetric image, we convert it to HSL as explained in Section 4.2.2. At that point in time, the image is randomly rotated by increments of 5° and flipped according to an empirically fixed probability. To complete the augmentation, our algorithm checks the uniqueness of the image to prevent redundancies and then converts these images in RGB format as expressed in Section 4.2.2.

At the end of this augmentation pipeline[6], starting from a unique image, it is possible to obtain N physically correct polarimetric images. Following our numerous experiments, we concluded that 11 images generated with a unique image (i.e. 12 images in total as shown in the illustration) was sufficient to obtain enough images.

Finally, implementing all these manipulations, we have proposed a new pipeline to perform augmentation operations on a physical-based modality, polarimetry.

4.4 Network architectures

As a usefulness proof of polarization for the understanding of urban areas, we propose performing a benchmark comparing color imaging and polarization imaging, but also to approve the augmentation method presented in Section 4.3.

[6]Available at: https://github.com/BlanchonMarc/P_Augmentor

Since one of the most widespread understanding applications is DL-based PwSS, we propose to perform this quantitative and qualitative study using proven architecture in literature.

4.4.1 SegNet

SegNet[4], shown in Figure 4.13, is a simple architecture composed of an encoder and a decoder. Designed for scene segmentation and widely used with "city/road" datasets, this network is not considered the state of the art but rather a pioneer in the field of PwSS. It is notable for its small number of layers and simple composition (shown in the table 4.1).

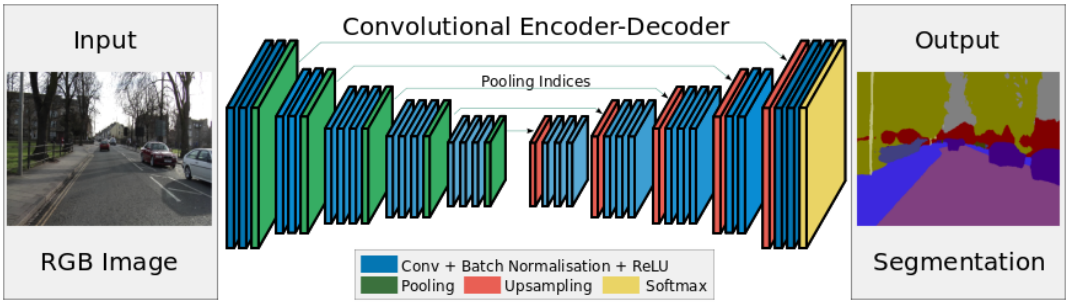


FIGURE 4.13: Illustration of SegNet architecture.

The main idea behind the emergence of this network as a point of comparison between RGB images and polarization images is that it is not necessary to use overly complex networks. Indeed, the approach here is not principally to observe efficient results but to compare in the most unbiased way the ability to appreciate the scenes through different modalities. Thus, through this network, it is possible, with a simple encoder-decoder architecture, to focus on the modality and not on the performances that the network could imply.

TABLE 4.1: Detailed SegNet architecture.

Encoder						Decoder					
	Type	**Kernel**	**Padding**	**Stride**	**Output Depth**		**Type**	**Kernel**	**Padding**	**Stride**	**Output Depth**
	conv1	3x3	1	1	Depth Image		conv14	3x3	1	1	512
Block1	conv2	3x3	1	1	64	**Block1d**	conv15	3x3	1	1	512
							conv16	3x3	1	1	512
	conv3	3x3	1	1	64		conv17	3x3	1	1	512
Block2	conv4	3x3	1	1	128	**Block2d**	conv18	3x3	1	1	512
							conv19	3x3	1	1	256
	conv5	3x3	1	1	128		conv20	3x3	1	1	256
Block3	conv6	3x3	1	1	256	**Block3d**	conv21	3x3	1	1	256
	conv7	3x3	1	1	256		conv22	3x3	1	1	126
	conv8	3x3	1	1	256		conv23	3x3	1	1	126
Block4	conv9	3x3	1	1	512	**Block4d**	conv24	3x3	1	1	64
	conv10	3x3	1	1	512						
	conv11	3x3	1	1	512		conv25	3x3	1	1	64
Block5	conv12	3x3	1	1	512	**Block5d**	conv26	3x3	1	1	Desired Depth
	conv13	3x3	1	1	512						

4.4.2 DeepLab V3+

DeepLab v3+[30], shown in Figure 4.14, is much more advanced compared to SegNet. Indeed, its complex design composed of atrous convolution and ASPP (these two concepts are explained in the Chapter 2) is much more powerful than the previous architecture.

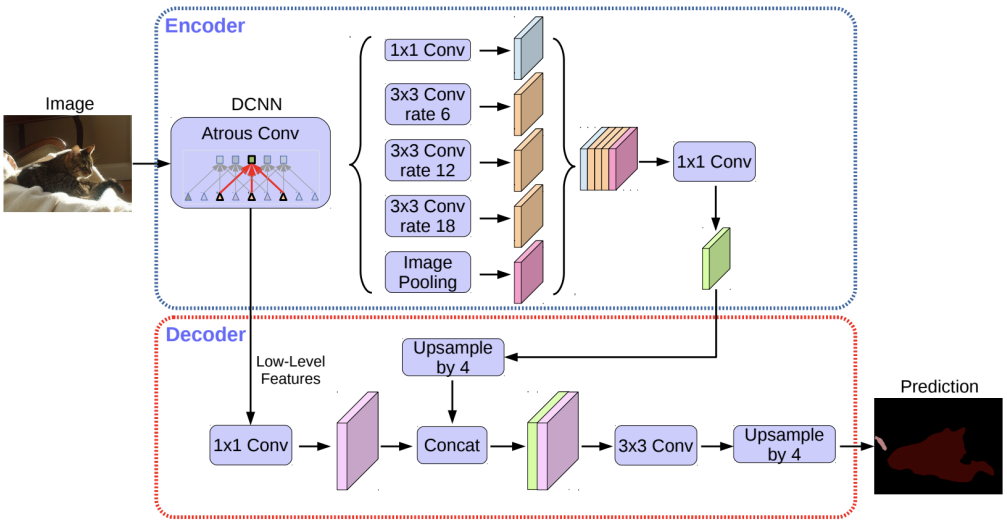


FIGURE 4.14: Illustration of DeepLab V3+ architecture. Schematic borrowed from [30].

The use of such a network is motivated by the need to quantify the utility of the augmentation in favorable situations. Indeed, by implementing a complex architecture, it is possible to benefit the learning of complex features and in this case, the physical validity of polarization imaging. In this manner, it will be possible to quantify the usefulness and/or the necessity of the augmentation.

4.5 Experiments

The experiments conducted allow two independent scopes.

On the one hand, a series of experiments is focused on the differences of modalities. The prime goal of this set is to demonstrate the interest of polarization over color imaging. The idea is to prove that with the same comparison method, a alternative form of information sustains a considerable interest for some fields. Therefore, without needing to unnecessarily increase the complexity of the processing cores, it is possible to reach satisfactory results.

On the other hand, we propose to investigate the usefulness of an augmentation adapted to a physics-based modality. Here, in order for the algorithm to grasp maximum advantage of the available information (valid or invalid), we use a state-of-the-art DCNN.

4.5.1 Modality-based Comparison

In this section, our proposal is to compare the modalities on the same basis, using aligned images, representing the same scenes. We manipulate the non-augmented images from PolaBot (see Section 4.2.3) and the network presented in Section 4.4.1.

Images presented to the network show urban scenes that are conducive to reflection. As a reminder, we assume a large number of objects are reflective or transparent in urban areas such as: cars, windows or wet roads. Thus, to highlight the respective capabilities of the networks trained jointly with colorization and polarization, the images were manually annotated with eight particular classes. These eight classes designed for autonomous robotics are: Sky, Water, Windows, Road, Cars, Building, None and Unlabeled. These classes were determined for their interest in complex areas but also because they can lead to confusion with standard sensors. Also, the None class represents all the areas that have been judged uninterestingly for our comparison (e.g. trees, sidewalks,...) and the Unlabeled class comes from manual annotation errors.

Therefore, a simple metric, allowing an explicit comparison, was chosen. Indeed, the accuracy per class has been defined as:

$$\text{Accuracy}_C = \frac{\sum P_C^p}{\sum P_C^g}, \tag{4.19}$$

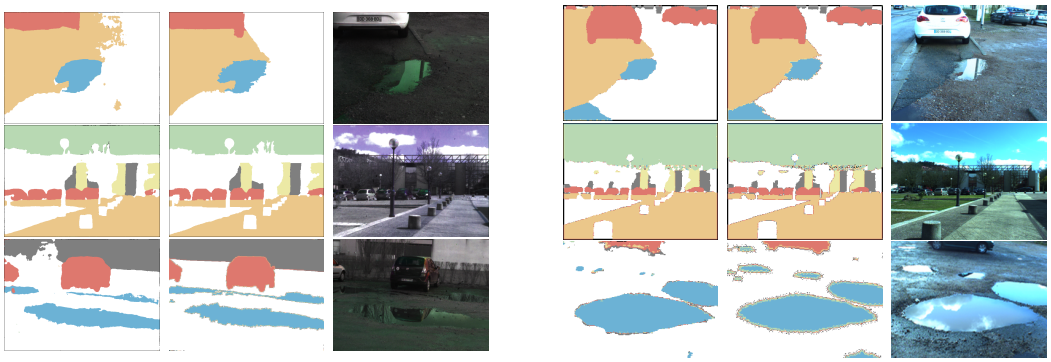
with C the class, P_C^p the correctly predicted pixel of the class C and P_C^g the ground truth pixel of class C.

For the training procedure, the construction of the dataset (explained in previous section) allows for pre-training and therefore the use of transfer learning [142, 111] to help convergence and benefit from efficient previous training. Therefore, both networks have been pre-trained using VGG16. This process, in addition to being advantageous, will allow us both to verify if the HSL to RGB approach is valid but also to see the adaptation capabilities of the network to a physics-based modality although pre-trained with color images.

4.5.1.1 Results

Since the training procedure was identical in the hyperparameters as well as in the order of appearance of the images, it is possible to quantitatively compare the results between the two modalities.

As shown in Figure 4.15, the qualitative results seem almost identical. However, when analyzing Table 4.2, a significant difference appears in favor of polarization.



(A) Polarimetry PwSS qualitative results.

(B) Color imaging PwSS qualitative results.

FIGURE 4.15: Qualitative results comparing colorization and polarization segmentations with identical scenes. For both the montages, respectively from left to right the images are: prediction, ground truth and input.

TABLE 4.2: PwSS quantitative results comparing polarimetry and RGB.

	@sky	@water	@windows	@road	@cars	@building	@none	mean
Model	higher is better							
Polarimetry	.753	.757	**.828**	.778	**.714**	**.876**	.789	**.785**
RGB	**.895**	**.786**	.445	**.784**	.484	.678	**.834**	.698

Indeed, the accuracy per class shows with a network trained with polarimetric information, windows, cars and buildings are recognized in a better way. Moreover, the mean column shows a significant improvement compared to the identical algorithm using colorization.

4.5.1.2 Discussion

After numerous experiments and despite a fair comparison, the polarization-based approach seems to be more appropriate for the PwSS of complex urban scenes compared to color imaging. Indeed, the use of polarization as a discriminant factor upstream of the network shows that for certain specificities such as reflective surfaces segmentation, this information is more robust. It should be noted that the dataset was designed to highlight these DCNN behaviors which may partially explain the increased performance. Nevertheless, despite these advantages, polarization has proved its value.

However, there is up until now an issue of data availability and consequently, the lack of consistent datasets.

This experiment also showed that DCNNs was able to adapt to supplementary information and that we could use transfer learning to rely only on fine-tuning. Moreover, the two methods compared having identical training, the learning on polarization emphasized a faster convergence.

On the other hand, polarization was supposed to have a significant advantage over the Sky class since it is polarized. One possible explanation is the color makes it easy to discern the blue of the sky. Indeed, after experimentation, the dataset was found to be biased by this fact and in every image showing sky, it was blue. Thus, this is in our opinion the most plausible way to explain this performance which exceeds that of the polarization-based network.

To conclude, it is significant to note that polarization offers a substantial advantage for specific applications and especially for recognizing areas prone to reflection. Regrettably, the data is relatively rare and therefore not widely available. Nevertheless, through these experiments, it has been shown that a network is not only capable of learning from physics-based images but also that it is beneficial. Even with limited data, the comparisons emphasized the increased capabilities when the learning was conducted with this unconventional data.

4.5.2 Augmentation-based Comparison

Now that it has been established, that polarization-based algorithms tend to show increased performance in complex urban areas; this section will allow a study of the augmentation and its impact. Indeed, the key idea is to show the interest of increasing the polarimetric image dataset by using the physics-friendly transforms established in section 4.3. Hence the use of an efficient network theoretically allowing a deeper use of the physical information of polarization. The concept is to check if the augmentation is valid and its impact on the results. Thus, the experiment will be carried out using three different approaches involving dataset: one will be not augmented, one augmented without using the regularization equations and a last implementing the augmentation presented in section 4.3. This approach allows a reliable comparison but also emphasizes the importance and even the necessity of an approach adapted to the modality.

Encouraging segmented areas' expansion. Usually, pixel segmentation techniques are discriminated using a simple loss chosen for its gradient behavior in logistic regression similar to the squared error loss for Linear regression. Indeed, the Cross

Entropy Loss (CEL) allows to classify using probabilities and compare an outcome with a reality. The multi-class CEL is defined as:

$$CEL(x, c) = -log\Big(\frac{exp(x[c])}{\sum_j exp(x[j])}\Big) = -x[c] + log\Big(\sum_j exp(x[j])\Big). \tag{4.20}$$

While this loss is efficient, it tends to encourage pixel accuracy but neglects object logic. The fact is that the areas to be segmented are often plain surfaces and -outside of occlusion- do not have different classes. Thus, some applications, particularly medical imaging, opt for the use of the Sørensen-Dice index (SDI). This one allows to encourage the propagation of a class and favors the fullness of the zones rather than considering only the pixel space. One could say SDI brings a more semantic dimension. This loss is defined as:

$$SDI = \frac{\sum_c^N 1 - \frac{2|X_c \cap Y_c|}{|X_c|+|Y_c|}}{N}, \tag{4.21}$$

with X the label, Y the prediction, c the class and N the number of classes. Since classes are unequally represented in the dataset, this metric allows an equal valuation of each of them unlike other losses. Therefore, we will use the Sørensen-Dice index for the training of the different networks since it allows us to obtain more satisfactory results semantically speaking.

Metrics. To effectively compare the different approaches and evaluate the impact of the different augmentation methods, a extensive range of metrics, as defined in Appendix B, is computed. Thus, the IoU was selected to quantify the overlap of the segmented areas with respect to the ground truth, the *recall* and the *precision* to measure respectively the consistency and the relevance of the segmentation. In addition, we chose to measure *specificity* since it quantifies the ability to select and remove bad classes. This last metric allows us to verify if the segmentation is efficiently performed by region and therefore SDI is meaningful for the application.

4.5.2.1 Results

As previously specified, three different DeepLab v3+ networks were trained with three different derivatives of the Polabot dataset. These three networks will be named:

- *None:* for Polabot without any augmentation.
- *Standard:* for Polabot augmented but without polarization angle regularization.
- *Regularized:* for Polabot augmented and aggregated with the regularization methods presented in Section 4.3.

The Standard or Regularized augmentation are both identical in the transformations applied on the images. This procedure, whether regularized or not, allows to obtain 2136 images from 178 (i.e. a 12/1 ratio). In addition, an evaluation was performed based on whether or not the backbones were pre-trained using provided model[7].

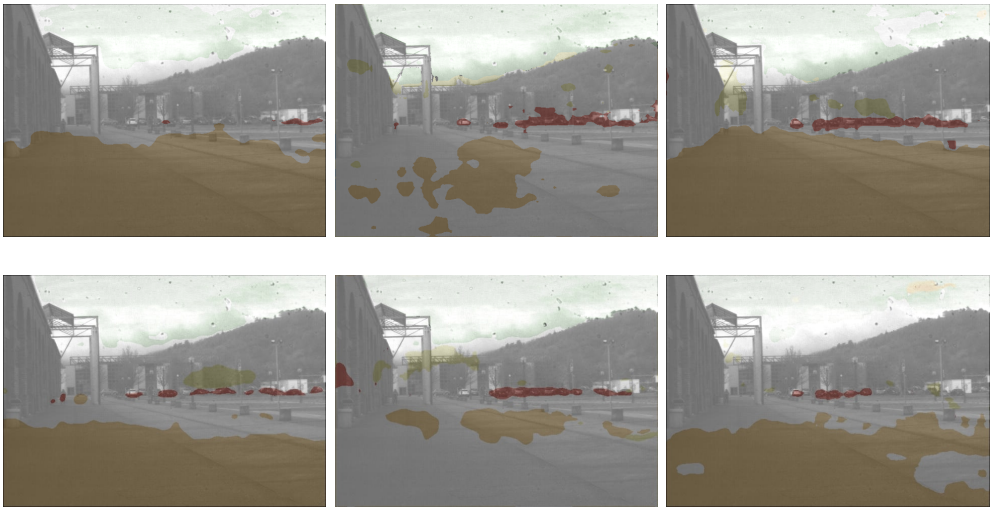


FIGURE 4.16: Illustration of segmentation results obtained according to different augmentation methods. The top line shows results for DeepLab v3+ network not pre-trained, and the bottom line the results with a pre-trained network. From left to right are presented predictions from networks trained with: un-augmented dataset, standarly augmented dataset and augmented dataset following our procedure. The four most representative segmented classes here are: *Road* (dark yellow/orange), *Cars* (red), *Sky* (light green) and *None* (light grey).

As shown in Figure 4.16, for a same scene, 6 estimates are then proposed by the benchmark.

It is then possible to extract a range of images from the test dataset which is a video sequence comprised of 8,049 images acquired at a frequency of 10Hz sharing many characteristics with the training dataset. Figure 4.17, shows a comprehensive panel of images representative of the results obtained from the various trainings.

TABLE 4.3: Quantitative evaluation of augmentation procedures. Impact of the augmentation procedure on DeepLabV3+ network. Specific classes have been highlighted in relation to the robotic application to witness the obstacle-wise performance. Due to the limited training, *Buildings* are almost undetected. For this reason, the averages denoted $\setminus B$ exclude the *Buildings* class from the calculation.

Augmentation	PreTraining	IoU (%)					Recall (%)					Precision (%)	Specificity (%)
		@water	@windows	@cars	Mean	Mean $\setminus B$	@water	@windows	@cars	Mean	Mean $\setminus B$		
None	No	40.0	20.6	20.8	30.5	32.2	35.2	15.8	22.5	**50.9**	50.0	50.0	89.6
	Yes	54.0	10.3	43.46	33.5	34.8	**42.4**	15.3	57.4	43.3	50.3	50.1	91.0
Standard	No	0.1	3.4	12.4	14.8	13.1	35.0	25.8	15.0	31.8	28.0	41.7	88.7
	Yes	10.2	3.0	19.7	21.8	20.0	35.2	22.9	23.4	37.0	33.4	41.2	91.2
Regularized	No	63.9	13.3	46.7	**43.4**	**50.3**	39.2	21.9	**60.8**	43.4	**50.5**	48.5	**91.3**
	Yes	**70.0**	**26.6**	**47.1**	37.8	38.5	35.0	**26.0**	48.0	42.0	38.5	**53.7**	90.7

[7] https://data.lip6.fr/cadene/pretrainedmodels/

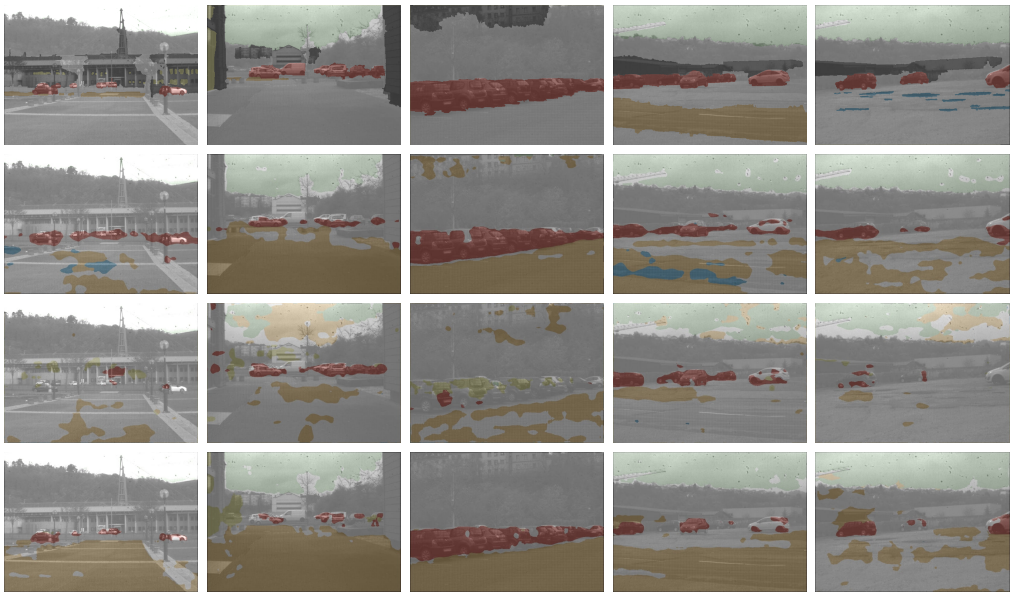


FIGURE 4.17: Examples of segmentation results according to the augmentation methods. From top to bottom are present the ground-truth, then consecutively the results from model with pre-training with: no augmentation, standard augmentation and regularized augmentation.

Finally, each metric described above has been computed in order to compare quantitatively the different processes. Thus, Table 4.3 shows the average performance of each pipeline.

4.5.2.2 Discussion

The results, both quantitative and qualitative, show increased performance when using our method's proposed pipeline. It is quite clear that augmenting the data without taking into account the physical dimension of the images disrupts the capabilities of the network. In this sense, it is notable that according to our results, it is better not to increase the data and to keep a small amount of image rather than to augment naively. However, the augmented and regularized data allowed us to observe the best overall results.

In spite of these performances, the pre-training seems to have a significant impact. Indeed, when it is performed, then the results are significantly better and it has a positive influence on the segmentation of the interest classes.

Finally, it is noticeable the adaptively augmented polarization allows a better recognition of the areas for which it has an advantage (i.e. the reflective areas). This is highlighted by the statistics calculated by interest class. This observation allows a second important conclusion. Networks, when using valid information, are able to take advantage of the physics and therefore can learn to interpret this new information.

The final important point to address is the augmentation procedure itself. When using conventional interpolable modalities, a broad extent of freedom is allowed. However, when using physics-based non-interpolable images, one must consider the factors defining the modality to size a customized augmentation. Thus, in addition to allowing for enhanced results, the augmentation, must be tailored specifically for the type of imaging used.

4.6 Summary

In this chapter, our work involving polarization and pixel-wise semantic segmentation has been defined. The polarimetric modality depending on many dependencies has also oriented our work on more meta domains such as: dataset construction, representation choice and establishment of an adapted augmentation procedure.

The two different axes proposed have demonstrated the usefulness of this unconventional modality for the understanding of complex urban scenes. A significant advantage was highlighted when color properties were put in competition with polarization properties. Therefore, our initial hypotheses on the ability to discern specular areas were verified by experimentation on our dataset designed for the occasion. In addition, the augmentation procedure was validated following a meticulous evaluation which allowed the leverage of many constraints on the image volume requirements. Moreover, we proved that this process allowed us to observe superior results. Ultimately, our benchmark on augmentation highlighted that it was preferable not to augment than to augment physics-based images in the usual way without taking into account the physical dimension of the data.

Furthermore, the effectiveness of deep learning-based polarization in urban scene understanding, which is a novel approach in the field, is demonstrated.

Ultimately, to appreciate the scenes from a geometric point of view, a polarization-based depth estimation pipeline will be elaborated in Chapter 5.

Chapter 5

Deep Polarization-Based Monocular Depth Estimation

This chapter is dedicated to depth estimation using a polarimetric monocular camera. We seek to infer, from a unique polarimetric view, a dense depth map while eliminating the recurrent problems of standard methods (saturation, specularity/reflection implied erroneous estimation, etc.). Indeed, since data from uncontrolled outdoor environments can be subject to numerous alterations, whether meteorological or due to the sensor, the algorithm should be robust and take into account all these characteristics. To this end, we propose a polarization-based approach of monodepth (MD). Contrarily to [159], [38] and [179] our intention is to propose a single modality DL-based method and which infer a depth map from a single polarization image. This is a crucial point since it alleviates constraints from acquisition setup and allow for real-world applications. Consequently, the aforementioned approaches relying on multiple view and/or modality, do not allow to solve the problem since they remain difficult to operate in dynamic urban environment. Deep learning-based single view estimation largely counteract previously stated flaws since it requires acquisition specificities only at training time while remaining a single image estimator at inference.

As detailed in Chapter 3, a remarkably large number of approaches have been used to infer a depth map from a single view. Due to the need for data representativeness, a vast majority of approaches rely on colorimetric images. On the contrary, it is noteworthy that all these methods show increased capabilities and allow consistent estimation. Consequently, MD represents a promising candidate and above all a significant baseline for altered data assessment. Among the numerous works in literature, it is possible to extract a method that has — in part — catalyzed the craze for MD, Godard et al.[58]. This approach is very promising since it uses two processes exportable to polarization: unsupervised learning and generically formulated loss through the statement of perspective geometry. To obtain more robust algorithms, the unsupervised approach is preferred since it negates the cost of data annotation.

Indeed, ground truth depth maps are tedious to acquire and often very inaccurate. As a result, the non necessity of them reduces the complexity of training but at the cost of a loss function of a different formulation. Secondly, the cost function shows adaptation and generalization capabilities. Borrowing the formulation of perspective geometry, it is therefore functional in all image spaces. It allows reconstructing a depth map from two views during training similarly to stereovision reconstruction algorithms. By all these aspects, [58] is a process exploitable with polarization since the prerequisites are not in conflict with the modality since the formulation is applicable to such space. In a second step, Godard et al. improved their method to include aspects of multi-scale, masking of immobile/occlusion zones. In [59], a field of possibilities is then opened, showing it is possible to limit the cost function and thus to add additional discriminant criteria to it. From this last contribution, we propose Polarimetry to Depth (P2D), a method adapted to polarimetric imaging. By extending the loss function to include a polarimetric term we propose a new network that will be subject to the constraints of perspective geometry — which are valid for any image pair — as well as sensitive to polarization data. Due to the prerequisites of such approaches, we propose a new data set that will be representative of road scenes. Plus, since one of the core objectives is to be invariant to weather changes, we have aggregated this set of images captured in diverse conditions. The proposed approach has been validated by the various experiments made possible by this dataset. Ultimately, since P2D showed some flaws in some experiments, we proposed other fusion-based methods. Indeed, P2D seems to be too relying on visual features, and thus showing weaknesses of genericity, we concluded it was necessary to use both polarimetric and colorimetric information. To accomplish this task and thus to group two very different types of images, we had to reverse-engineer multiple fusion methods and evaluate their validity in relation to the context. Ultimately, we propose a cascaded network method as well as a cascaded double loss to finally discriminate only the normals without passing specifically in the 3D space. With this approach we reduce the impact of inaccuracies when changing consecutive spaces and therefore address the problem in a more appropriate way.

5.1 Introduction

Scene reconstruction represents a major task in computer vision. It offers an observation of the three-dimensional world accurate in shape, size and geometric structure. This explains its great usefulness since it allows an increased understanding and thus a massive potential of application.

There are multiple possibilities to obtain a three dimensional scene reconstruction. Each method has its own constraints. A pair of cameras will subsequently allow the reconstruction by stereovision and operating a registration and a camera-to-camera

projection. A mobile 2D camera, on the other hand, will allow the reconstruction of the scene using registration through a sequence of images using Visual Odometry. In each of these cases, the methods suffer from the motion in the scenes and therefore produce artifacts. To be specific, the monocular method operating a visual odometry-based registration is valid only for static parts of the scenes.

To eliminate these drawbacks, some methods are DL-based and show improved capabilities but especially an abstraction of past constraints. Monocular supervised methods learn the direct correspondence between an input image and a ground truth image. These preliminary methods although reducing the constraints require a large dataset of images and, above all, their corresponding accurate reconstruction which makes them difficult to use, especially with a new data type.

To avoid aggregating heavy and imprecise datasets, unsupervised methods do not need ground truth at the cost of a more complete loss function. Indeed, a simple observation is that databases are quite unreliable. Specifically, when it is necessary to have an accurate depth map, the acquisition processes are not enough accurate. Essentially, for this kind of acquisition, a LiDaR is used. As follows, it projects a laser and measure the return time of the back-scattered beam. This is valid on one condition, that the laser is undiverted and therefore returns, unaltered, to the sensor. Under ideal conditions this method is extremely efficient. On the other hand, a substantial majority of approaches address the problem of reconstruction in urban scenes. One observation is that cars are highly reflective and therefore subject to laser erroneous reflection due to the specularity aspect of the surface. Also, in uncontrolled areas like these, there may be multiple surfaces modifying the light rays (windows, mirrors, ...). Ultimately, outdoor acquisitions are subject to climate change, rain, and therefore water accumulation on the tracks/objects which modifies the interaction of light with the surfaces. It is then possible to see the datasets are acquired only under certain conditions and that they periodically produce artifacts/imprecisions. As a reminder, the genericicty of the DCNN is highly dependent on the data distribution it was built on. Thus, from these observations, despite a non-supervised approach and consequently a loss supposed to bring this genericity, the approaches suffer from various weaknesses. Specifically, the estimation of specular surfaces is often incorrect and due to the nature of the observed scenes, these algorithms seem difficult to deploy in real conditions of use (for autonomous cars for example).

We propose using the light characteristics as an attachment point for our method. While previous approaches suffered from their lack of understanding of certain light phenomena we propose to use them to the benefit of our method. While specularity was mostly neglected and depth in these areas is poorly estimated in the color space, it

is particularly clearly defined in the polarimetric space. As follows, we aim at keeping the accurate estimates of the previous algorithms while aggregating the knowledge of surface-to-light interaction. Given a sufficient number of representative polarimetric images, we aim to produce high quality accurate reconstructions by exploiting both visual features and polarization data. A DL network will then be fed with polarimetric images and constrained by a particular loss integrating polarization specific terms. In a first step, we propose a method that is only dependent on polarimetric images. In a second step, we propose to investigate fusion methods that keep the viability of past methods while adding extra precision on specular areas.

5.2 Depth-to-polarization interconnections

Since the core idea is to keep the unsupervised aspect to the learning, a link between the input image (polarimetric) and the desired output image (depth) is needed. In the case of color imaging, the algorithms rely on the perspective geometry formulation. In other words, the approaches require visual features. It is possible to use this space in the polarization imaging but it will be restricted due to the absence of color. On the positive side, there is a direct link between the acquisition of the sensor and the depth thanks to the specularity.

5.2.1 Normals to angle of polarization

As detailed in Section 2.1, polarimetric acquisition is very versatile and allows deducing many characteristic images. One of the strengths of this modality is that we have a direct measurement of the polarization angle. This angle can be very discriminating as shown in Chapter 4, but it could also be very useful for linking the modality to a depth map.
As a reminder, the polarization angle α is calculated as follows:

$$\alpha = \frac{1}{2}\, arctan2(s_1, s_2). \tag{5.1}$$

As shown in Figure 5.1, it is equally possible to geometrically represent this component of polarimetric information mixing peculiar and perspective geometry.

There is a definite relationship between α inferred from acquisition and the plane normal. Provided that the surface is specular, then a relation exists between the reference and the direction of the electric field $\vec{E}$ projection. From this follows the possibility of deducing the normal $\vec{n}$ to the specular surface (i.e. a high degree of polarization ρ) since $\vec{E}$ is perpendicular to it. This statement is valid if and only if the surface is specular, which ensures perpendicularity between the two directions

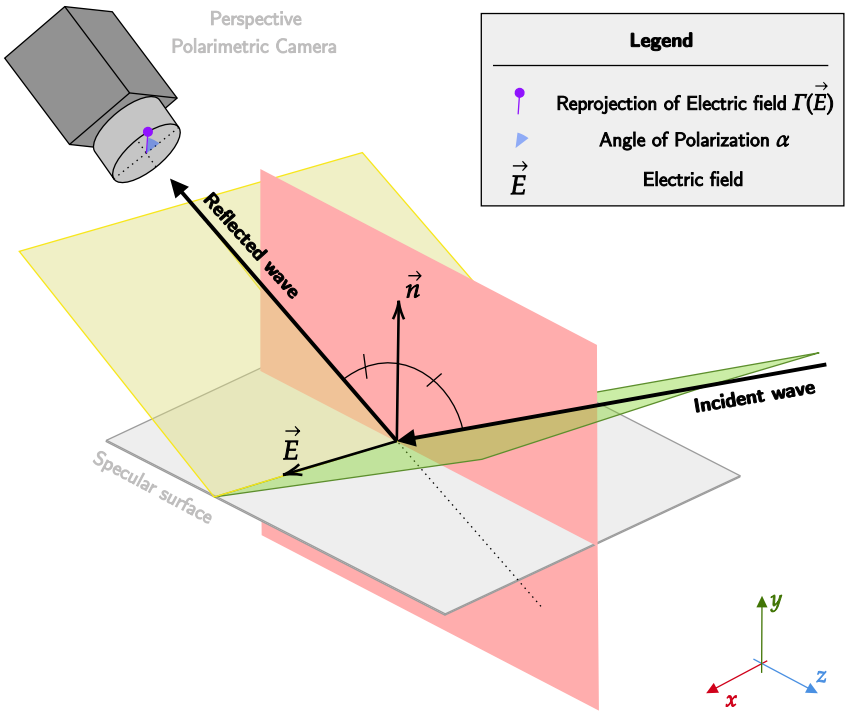


FIGURE 5.1: Illustration of imbrication of polarimetry peculiar and perspective geometry. Visual representation of angle of polarization measurement.

mentioned above. Otherwise, there would be no definite angle between $\vec{E}$ and $\vec{n}$ and therefore a regularization equation based on this principle would be neither differentiable nor optimizable. In order to guarantee the validity of the formulation, it is then necessary to discriminate using ρ as a criterion since this parameter allows a quasi equivalence to the specularity measurement.

5.3 Towards a unified polarization-based method for depth estimation

To infer a depth map from a single polarimetric image, we formulate a loss function supported by polarimetry-to-depth links. The accurate reconstruction is obtained by optimizing this function by a deep learning network in an unsupervised manner.

5.3.1 Defining the loss

Since the problem is unsupervised, the problem is not just to have the data and formulate a statistical function. Since we are not relying on ground truths, it is imperative to formulate an optimizable representative function that will take into account the polarimetric information. With the ambition of faithfully reconstructing the often neglected specular areas, we propose to build on similar work on reconstruction using polarization.

5.3.1.1 Prior polarimetric reconstruction error

The paper by Berger et al. [10] proposes an approach to minimize an error specific to polarimetric induced geometry. Drawing on the terms provided by Woodford et al. [155], the method consists in including a minimizable expression compelling a normal/polarization angle consistency. It is consequently shown that constraining a cost function involving a polarimetry-specific geometry is valid. Furthermore, this minimization approach is operable when optimizing a deep learning model since it depends on both the input and output of the processing pipeline and therefore could guarantees a self-supervision capability. Nevertheless, the acquisition setup as well as the problem formulation highly influence the error calculation. Indeed, [10] proposed an azimuth to acquired angle of polarization comparison. This approach is consistent under peculiar conditions implying restricted calibration of the camera or azimuth to angle of polarization specific link hypothesis. For this reason, our method proposes an alternative but similar approach allowing standard calibration and a generalized loss term releasing the constraints and allowing for easier use in real word applications.

5.3.1.2 Constraining the loss

Fundamentally, the function to minimize includes a reprojection error term and a smoothing term. Our method P2D employs the photometric error proposed in [59] since it has demonstrated its optimization and efficient convergence capabilities via deep learning. As a replacement for the edge aware first order smoothness, the second order derivative enhancement proposed by [155] is used to encourage fine transitions counterbalancing the discontinuities induced by the polarization parameters.
First, one penalizes the photometric reprojection error:

$$L_r = \min_{t'} \ pe(I_t, I_{t' \to t}), \tag{5.2}$$

with $t' \to t$ the pose transformation between two consecutive views and pe the reconstruction error:

$$pe(I_a, I_b) = \frac{\beta}{2}(1 - SSIM(I_a, I_b)) + (1 - \beta)||I_a - I_b||_1. \tag{5.3}$$

To comply with the specifications of a minimizable function, the reprojection error comprises the weighted combination of structural dissimilarity (DSSIM) and L1 difference penalizing the per-pixel deviation of the reprojection. As described in the original paper, $\beta = 0.85$ is used.

In a second step, a smoothing term is used to encourage a precise estimation of the planes while taking into account the edges:

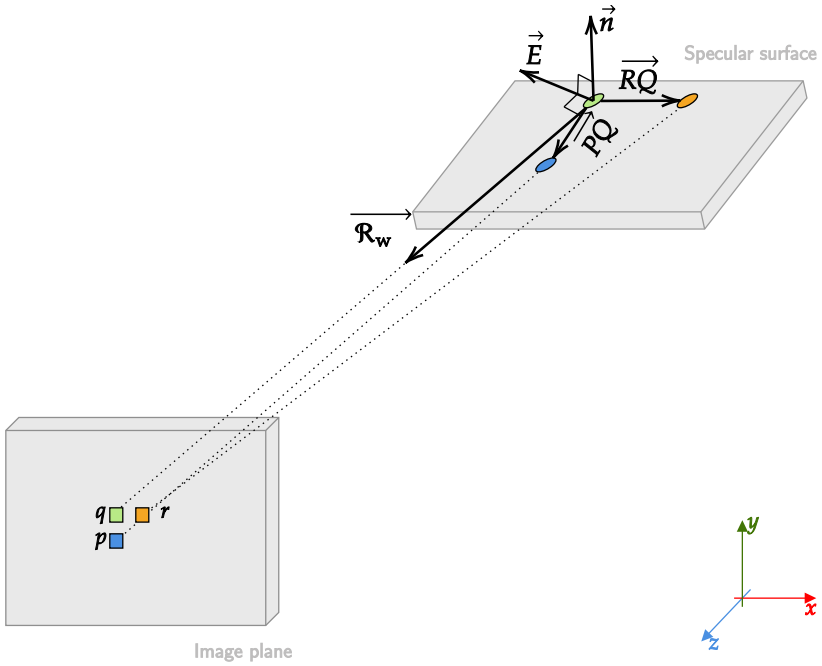


FIGURE 5.2: Illustration of the electric field estimation method.

$$L_s = |\delta_x^2 d_t^*| e^{-|\delta_x^2 I_t|} + |\delta_y^2 d_t^*| e^{-|\delta_y^2 I_t|}, \tag{5.4}$$

where $d_t^* = d_t / \bar{d}_t$ is the mean-normalized inverse depth enforcing the depth to be dense while reconstructing the planes [147] and the δ^2 operator is defined according to the second order prior smoothness term $\mathcal{S}(\{j,k,l\})$ [155]:

$$\mathcal{S}(\{j,k,l\}, d_t^*)_x = \delta_x^2 d_t^* = d_t^*(j) - 2 * d_t^*(k) + d_t^*(l), \tag{5.5}$$

with $\{j,k,l\}$ three neighboring pixels in the horizontal or vertical direction following the x-axis or y-axis orientation of the smoothing.

The weighted combination L_{diff} of these two terms then allows for a precise reconstruction of the non-reflective (diffuse) areas.

$$L_{\text{diff}} = \mu L_r + \lambda L_s, \tag{5.6}$$

with λ a scaling parameter set to $1e^{-3}$ and μ a binary mask defined in [59] taking occlusion and displacement of pixels along sequences into account.

Now, by drawing inspiration from and generalizing the contribution in [10], it is possible to include a third term into this loss to penalize poor reconstruction of reflective areas. By definition, polarization is defined by the orientation of the electric field. Consequently, it is possible to estimate the electric field orientation as a function of a normal derived from a plane.

Let us consider three neighboring pixels $\{p, q, r\}$ in the layout presented in the Figure 5.2. This arrangement is organized such that it removes fronto-parallel planes related uncertainties. Then, the projection of these three adjacent pixels into 3D results in three points of a plane, respectively $\{P, Q, R\}$. The local normal $\overrightarrow{n}$ is obtained via the cross-product of the two vectors $\overrightarrow{PQ}$ and $\overrightarrow{RQ}$ linking the points $\{P, Q, R\}$. By definition, the electric field $\overrightarrow{E(Q)}$ is perpendicular to the plane defined by the normal and the reflected wave when considering specular surfaces. Following the definition, $\overrightarrow{E(Q)}$ at 3D point Q can be deduced from the cross product between the local normal and $\overrightarrow{\mathcal{R}_w}$ at the point Q as follows:

$$\begin{aligned}\overrightarrow{E(Q)} = \Big[\Big(\Pi(p, D(p)) - \Pi(q, D(q))\Big) \times \\ \Big(\Pi(r, D(r)) - \Pi(q, D(q))\Big)\Big] \times \overrightarrow{\mathcal{R}_w},\end{aligned} \tag{5.7}$$

where $\Pi(x, D(x))$ is the 3D projection of pixel x relative to the disparity $D(x)$. In an optimal context, the polarization angle and the electric field maintain the same orientation and by extension the same angle relative to the reference as shown in the Figure 5.3. Conversely, when a depth map is incorrectly estimated, then the estimated local normal is inconsistent and consequently is the deduced polarization angle. Accordingly, we can add a term C_{pol} to the loss penalizing the deviation of the normal.

As shown in equation 5.8 and in Figure 5.3, to evaluate the deviation, it is necessary to back project the direction of the electric field onto the image plane and compare it with the angle of polarization α:

$$C_{\text{pol}}(q) = \rho(q) \left| \tan\left[\tan^{-1}\left(\Gamma(\overrightarrow{E(Q)})\right) - \alpha(q)\right]\right|, \tag{5.8}$$

where Γ is the back projection operator onto the image plane. Moreover, ρ allows the scaling of the loss reinforcing the necessity for correlation between α and ρ.

Since an angular differences is considered, and because this term will be combined with the reprojection term, the definition domains must be taken into account. The reprojection term clearly belongs to $[0, \infty[$ interval. To constrain the polarization term to the same interval, the absolute tangent is employed. As a result, the polarimetric loss term becomes:

$$L_{\text{pol}} = \frac{1}{N} \sum_{x \in \chi} C_{\text{pol}}(x), \tag{5.9}$$

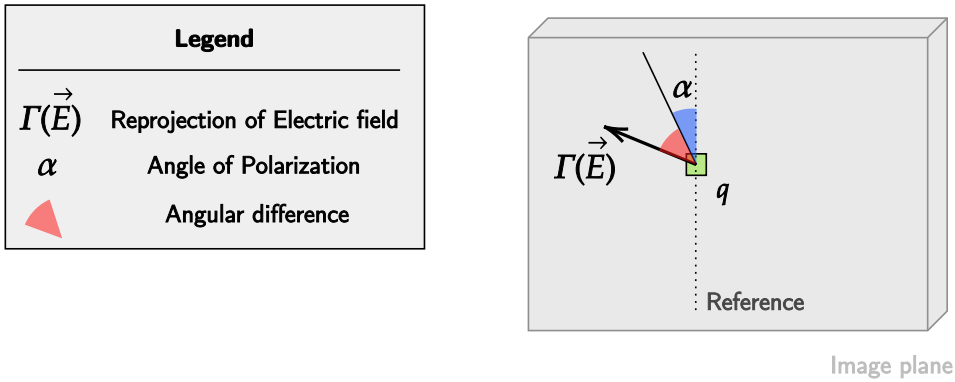


FIGURE 5.3: Angular difference visual representation. Here, the reference of the angle of polarization is vertical.

with N the number of pixels x in the set of reference image pixels χ. Finally, the loss used to train the network is defined by:

$$\Lambda = L_{\text{diff}} + \tau L_{\text{pol}}, \tag{5.10}$$

where τ is a binary mask derived from ρ such that:

$$\tau(x) = \begin{cases} 1, & \text{if } \rho(x) \geq 0.4 \\ 0, & \text{otherwise} \end{cases}. \tag{5.11}$$

The polarimetric term L_{pol} is taken into account only if the degree of polarization is relevant. Since, the relevance of both ρ and α are correlated, this mask ensure for a legitimate electric field orientation estimation. The final loss Λ is then just composed of reprojection error when the image area is unpolarized. Polarization components, when consistent, are taken into consideration and penalize the inaccurate reconstruction of specular surfaces.

5.3.2 Network architecture

Following [59], the network has an encoder-decoder architecture (a UNet with a ResNet 50 layout as shown in the Figure 5.4). It takes as input three-channel images obtained by concatenation of the intensity ι, the polarization angle α and the degree of polarization ρ. To overcome some inconsistencies related to the polarimetric modality and to consider exclusively areas with a minimum partial specularity, all degree of polarization values of lower than 0.4 are eliminated. This is justified by the fact that diffuse surfaces corresponding to low degree of polarization lead to a difference of $\pi/2$ between α and the electric field $\vec{E}$. When the disparity induced by the reprojection error is calculated, despite the accuracy of this calculation, the angular error will then tend towards $\pi/2$ leading the L_{pol} function to tend towards infinity and

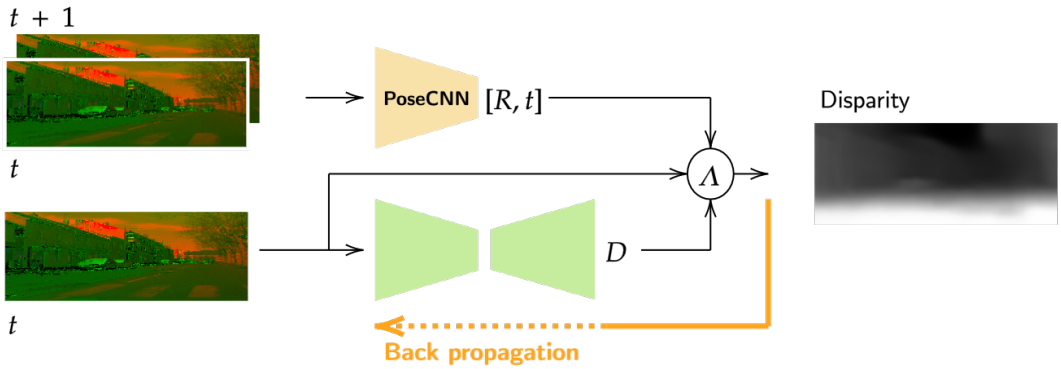


FIGURE 5.4: Illustration of the network as well as the loss calculation strategy and its back propagation. Drawing inspiration from [59], the depth estimation network is a UNet with a ResNet50 layout.

thus causing exploding gradient problems.

Similarly, a perfect ρ is physically unobservable which justifies an upper threshold. To combine a scale-factor effect and a regularization relative to physical property, the values are clipped to a maximum of 0.8.

5.3.3 Experiments

5.3.3.1 Implementation details

Datasets. The training dataset was acquired during both dry and rainy weather such that the experiments would highlight the capacity of polarimetric modality in diverse conditions. All acquisitions were made with an affordable polarimetric camera, the Basler Ace aca2440-75um POL, consisting of a Sony IMX250MZR sensor delivering a resolution of 2448 x 2048 pixels. The camera was mounted on board a driving car, recording a total of approximately 7,000 images per weather condition. The final training dataset is composed of 13,400 images. As for the evaluation dataset, it is composed of a completely independent set of 25 images, acquired separately from another view under mixed meteorological conditions.

Ground truth. Ground truth generation represents a critical point when it comes to addressing urban reconstruction problems. Because of specularity, accurate depth evaluation is difficult since ground truth generation commonly rely on LiDAR sensors which are occasionally unreliable for measuring specular surfaces geometry due to reflection or transparency. Indeed, it would be a prerequisite to spray matte coating over all the specular surfaces of a complex urban scene which is obviously unfeasible. To overcome such difficulties, the reference disparity has been pre-calculated using SGBM [68] and then refined by hand. It would have been possible to calculate the ground truth using a learning-based method. It should be considered the approach

presented here is to improve deep learning methods since they typically fail on specular surfaces. Moreover, it is notable the vast majority of networks are trained on the same database which consists of images in favorable weather conditions. For these reasons, the choice of a refined SGBM eliminates learning biases while providing ground truth taking into account specular surfaces. This approach is unconventional but permits to conceive a global idea of the reliability of the results while allowing the computation of metrics. In addition, the disparity remains a relative value and therefore the impact of manual refinement is minor.

Network training. The network was trained on a machine consisting of a Nvidia Titan Xp (12GB memory) GPU, 128GB of RAM and two CPUs accumulating a total of 24 physical cores. We use the following parameters for all the networks: a batch size of 12, a learning rate of 1e-3 and a maximum of 30 epochs. For a fast training, the images were downsampled without any interpolation method to maintain the physical properties. Following this routine, training with polarimetric images takes approximately 17 hours compared to 12 hours when training with intensity images only. The forward pass inference time is around 0.45 second per image in pure CPU processing.

Hyper-parameters. To define the set of hyper-parameters we conducted experiments to evaluate the final performance of each network. Table C.1 shows the study conducted to define whether pre-training was necessary and which empiric parameters were suitable.

Evaluation. We compared the results of our method P2D with the competitive state-of-the-art method described in [59]. Our P2D receives as input the polarization parameters by concatenation of the three channels $\{\iota, \alpha, \rho\}$. For the method in [59], we evaluate two versions. One version, G_{RGB}, using only intensity images and trained with the weights provided by the authors without fine-tuning. And, another version, G_{I}, trained in an end-to-end manner so that the network parameters are adapted to the intensity images at hand.

Metrics. The calculated metrics shown in the table 5.1 represents popular assessments within the reconstruction community that have been proposed by Eigen et al. [46]. They provide an unbiased and comprehensive measure of results. In particular, the δ values are calculated on the prediction/ground truth ratio and highlight an intrinsic precision of the reconstruction.

TABLE 5.1: Quantitative comparative results. For each network several metrics are computed neglecting the sky areas. In addition, we propose three different evaluations: on the *Raw* images at the output of the network, on the *Cropped* images to eliminate inconsistencies in the polarimetric network, and on the *Specular* areas only. G_{RGB} corresponds to the network presented in [59] without fine-tuning and taking $\{\iota, \iota, \iota\}$ the intensity concatenation as input, and G_I corresponds to the same network with fine-tuning. P2D corresponds to our method.

Type	**Network**	Abs_Rel	Sq_Rel	RMSE	RMSE_log	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
Raw	G_{RGB}	0.471	10.809	25.161	0.680	0.485	0.707	0.804
	G_I	0.482	9.144	22.332	0.617	0.431	0.695	0.838
	P2D	**0.322**	**4.504**	**20.651**	**0.484**	**0.537**	**0.801**	**0.896**
Cropped	G_{RGB}	0.533	14.050	29.312	0.780	0.449	0.658	0.771
	G_I	0.415	11.247	25.899	0.678	0.467	0.729	0.850
	P2D	**0.245**	**5.650**	**24.009**	**0.531**	**0.604**	**0.825**	**0.910**
Specular	G_{RGB}	0.341	8.249	7.236	0.306	0.666	0.808	0.896
	G_I	0.208	2.248	5.491	0.233	0.639	0.877	0.952
	P2D	**0.147**	**1.583**	**4.898**	**0.166**	**0.796**	**0.921**	**0.973**

At last, the sky reconstruction accuracy R_s is calculated as follows:

$$R_s = 1 - \left(\frac{\hat{y}_s}{y_s}\right), \tag{5.12}$$

where y_s is the sum of the binary masked pixels considered as sky in the ground truth and $\hat{y}_s$ the corresponding area in the prediction. This calculation is performed on the disparity, and one order of magnitude error deviation is considered acceptable. It focuses on the ability of the network to accurately estimate the sky and not propagate an erroneous evaluation in such areas. It is noteworthy this kind of precision is usually neglected since the reconstruction precision of these areas is removed from the frequent metrics. Ordinarily, sky zones are filtered out of the metrics beforehand. In this evaluation, these areas are also neglected while calculating Eigen et al. [46] metrics.

5.3.3.2 Results and discussion

Table 5.1 and Figure 5.5 allow for a quantitative and qualitative evaluation of the results. In addition, the quantitative results affiliated with the benchmark of the different hyper-parameters are shown in Appendix D. Analyzing the images in Figure 5.5, we can observe various responses of the networks. First, using the method in [59] with raw images (G_{RGB}), the results seem satisfactory at first glance. However, some characteristics of the images are altered. For example, specular areas, car windshields or bus stops are incorrectly detected. To be specific, car windshields are over-segmented into several parts rather than being detected as unique planar surface. In addition, the distance to reflective road lines is often under-estimated and farthest objects are ignored. However, as the weights of the network are not fine-tuned, its features representations have been learned exclusively from textures characteristics which limit the

performance of the method in specular or reflective areas. Nevertheless, the reconstruction is close enough to the ground truth which also shows the robustness of this approach and reinforces the initial idea of using it as a baseline method.

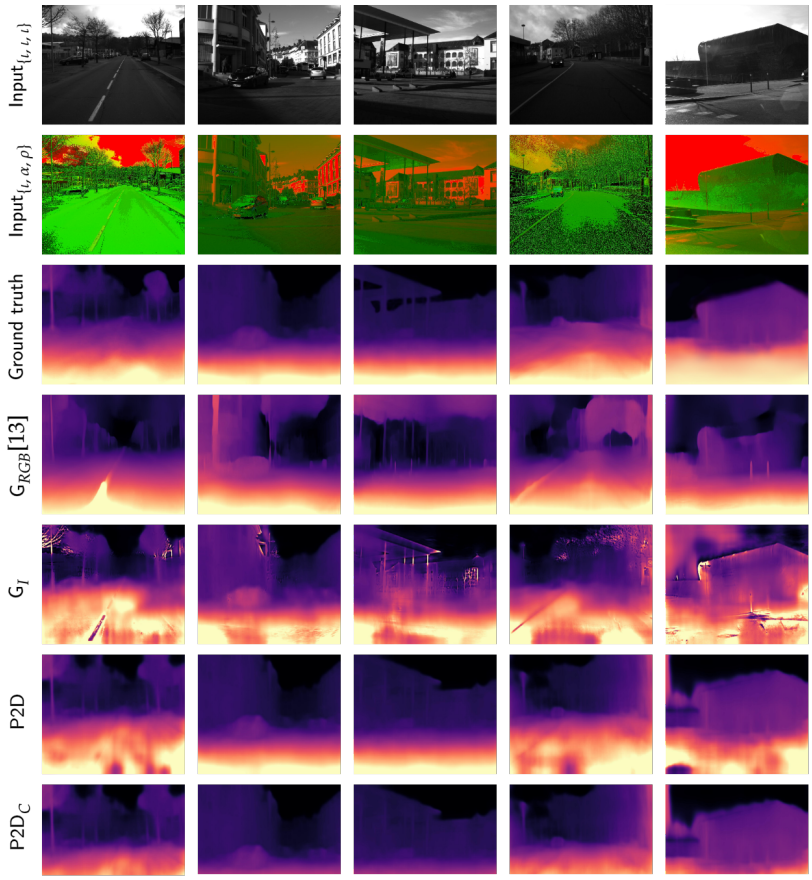


FIGURE 5.5: Illustration of results on five independent road scenes in mixed weather conditions. From top to bottom, the inputs to the two networks (scalar or polarimetric), the ground truth depth map, then the results of the three different networks: G_{RGB} corresponds to the network presented in [59] without fine-tuning, and G_I corresponds to the same network with fine-tuning. P2D corresponds to our method. The last row shows the crop version of the results from P2D to eliminate inconsistencies due to the modality and aberrations because of the camera position. Columns one and four correspond to acquisitions in light rainy weather, hence, the different road behaviour in polarimetric space. The other images are acquired under normal conditions.

When the network is trained end-to-end with polarimetric intensity images, the global impact of polarization is reduced but brings many significant constraints. Despite an accurate estimation of some areas like planar surfaces on cars and long distance objects (see row G_I in Figure 5.5), others areas are subject to some aberrations mainly on the reflective lines and polarized contours. This behaviour produces a direct impact on the network estimations. Hence, the addition of polarization-specific terms is necessary to improve the estimation of polarized areas.

When employing all the polarimetric information (ι, ρ and α) in our P2D network, we can observe a more accurate estimation of specular areas as well as sufficient reconstruction of diffuse areas.
We can however see that the results at limited distances, as shown in the P2D row of Figure 5.5, are occasionally incorrect. This is due to the fact polarimetric information varies according to the light and its reflection angle. Therefore, the position of the camera is primarily responsible for these erroneous estimates. Indeed, as explained earlier in the Datasets section, the images of the evaluation subsets were acquired with different camera poses. Consequently, the information from the images differs from the training set case leading the network to fail in estimating depth values at close distances. To have an estimation in favorable conditions, the choice of cropping the lower quarter of the images for a second evaluation is proposed. Note, this lower part corresponds to closer distances. Both the estimates and the ground truth depth maps are cropped. These results are shown in the P2D_C line of Figure 5.5 as well as the *Cropped* part of the Table 5.1. We can observe better performances especially when comparing errors with the δ metric. The most considerable improvements are obtained when looking at both $\delta > 1.25$ and $\delta > 1.25^2$ showing a respective 16% and 17% improvement compared to the G_{RGB} network.

Additionally, we perform an evaluation considering only the specular areas of the scenes. This is achieved using a rule-based naive system filtering polarization degrees higher than 0.4, hence keeping areas which are highly specular. The results shown in the last part of Table 5.1, exhibit improved performances for all the evaluated networks since the assessed pixel space is reduced. However, the largest improvement is obtained with our P2D method. Specifically, our method achieves 92% for $\delta > 1.25^2$ ratio, compared to 80% obtained by the state-of-the-art method. Consequently, we can see the polarimetric modality is beneficial for the reconstruction of urban scenes with many specular surfaces. Ultimately, to highlight the depth map reliability, sky reconstruction accuracy (eq. 5.12) has been computed in Table 5.2.

Table 5.2: Quantitative comparison of sky reconstruction accuracy.

Network	G_{RGB}	G_I	P2D
R_s	0.055	0.388	**0.532**

This specific metric has been computed since many evaluation metrics neglect such aspect which, however, could be informative, especially if one uses such an algorithm

for navigation. This ratio reveals P2D's ability to reconstruct slightly more than half of the sky correctly. It permits to demonstrate polarization imaging to be favorable also for such estimation.

5.4 Polarization and colorization fusion for accurate depth estimation: a proof of concept

Despite the innovative approach proposed in the previous section, it is notable that in the framework set, the problem is brought as an end-to-end learning problem. That is, it is necessary to reconstruct an image, starting only from the polarimetric priors and without further information from other modalities. This approach has demonstrated many qualities and highlights good performances. However, in some cases, this method tends to fail due to the lack of information or the annihilation of the perspective geometry term by the polarization. In brief, the joint learning of the perspective geometry and the polarization geometry remain a complex task. It is sometimes possible to observe, as in Figure 5.6, contradictions which lead the network to uncertain estimations.

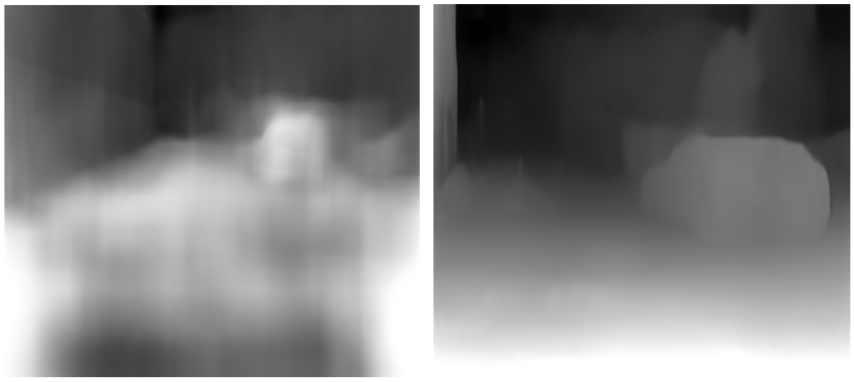


FIGURE 5.6: Case of estimation failure using P2D comparing to Monodepth v2. Left shows P2D estimation on polarimetric image. Right shows Monodepth v2 estimation on RGB similar image. Source images observe a car and both images were taken at the same time and under a color-favorable condition.

On the other hand, some color-based methods have been approved and are robust to many events. Although these approaches are directly related to the features present in the image, their learning with the guidance of massive databases allows strong abstraction capabilities.

The two principles contain contradictions but seem robust, hence, we propose to investigate the possibilities of merging the performances. As follows, the problem is no longer depth map estimation from polarization but a refinement of a map estimated from color images using polarization.

5.4.1 Estimating the appropriate fusion method

There are multiple methods of fusion [174]. However, a small population of approaches allow multimodal image fusion, especially when one is physics-based.

We propose estimating the different possibilities of fusion of polarization and color imaging. Thus, it will be possible to deduce which processes are applicable to theoretically obtain satisfactory results and especially to address the problem of depth map refinement.

5.4.1.1 Early fusion

Early fusion is a very common process since it is simple to implement. It basically consists of the image concatenation prior to the network. This technique requires perfectly aligned images and is frequently associated with 2.5D. Indeed, the RGB+D modality is extremely suitable since this information is complementary and can be aligned to the nearest pixel. Preliminary approaches like FuseNet [65] or MVCNet [95] have proposed segmentation methods using early fusion. This kind of approach would be completely adaptable to a self-supervised depth learning problem by adapting the loss function. A method close to our topic is RTFNet [141] which proposes a semantic segmentation from a fusion of RGB and thermal image. This contribution highlights that it is possible to merge via this process images in two different spaces and especially color-based and a physics-based image.

It has to be highlighted that a majority of the techniques are based on fusion applied to segmentation, although this is only an adaptation of the objective function to encounter a problem of depth estimation.

To assess whether this process is accessible to RGB polarimetric fusion, it is necessary to estimate the prerequisites for using such a method. The main idea is based on the concept of modality complementarity. Unfortunately, polarization and colorization share mutual information which would imply redundancies. It would indeed be possible to perform ablations to eliminate these redundancies or perform a thoughtful concatenation of the images. Thus, it would be possible to obtain a five-channel image composed of the three components of the RGB and the two complementary information of polarization, namely the angle and the degree of polarization.

As shown in Figure 5.7, such an approach have been schematized.

As an intermediate conclusion, this kind of approach already requires a massive amount of aligned information. In addition, this specific technique almost necessarily requires an end-to-end depth estimation. The problem formulation is therefore not compatible with such an approach. The key concept is to take advantage of the

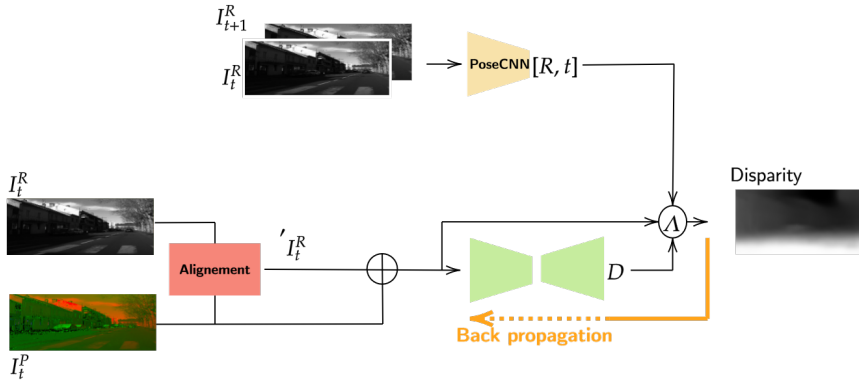


FIGURE 5.7: Early Fusion architecture illustration.

robustness of RGB approaches and the specificity of P2D. As a consequence, early fusion does not address the problem since it is potentially subject to the same flaws as P2D. Another possibility relies not on a fusion before the network but in its core.

5.4.1.2 Latent space fusion

Latent space fusion consists in examining a strategy to combine several feature vectors from two different modalities. The principle consists in having an encoder that extracts the main information from each channel and then accumulates it to mutualize the decoding process.

This concept can be based on statistical blocks, dense convolution layers or direct concatenation/addition to merge the information. This is based on the assumption that information is combinable and therefore the decoder allows, despite different modalities, to extract mutual data. However, most methods do not consider this point and assume that dimensionality remains the unique condition for fusion. Practice allows verifying this, but greedy algorithms such as deep learning are famous for their ability to reach an objective formulated by a function. This implies that whatever the type of data, valid or not, mergeable or not, the system will find an equilibrium point allowing to optimize the function.

In our case, we consider the latent space fusion to tend towards the estimation of an accurate depth map. We must not neglect the fact that the two modalities have similar points and that the polarization brings a unique information allowing to optimize further. Some methods [120, 47] have tried to merge these two data without deeply considering the impact of each modality and their influence for a semantic segmentation task. A look-alike latent space fusion based architecture is schematized in Figure 5.8. This schematic is adapted to the problematic of depth estimation which explains the PoseCNN network.

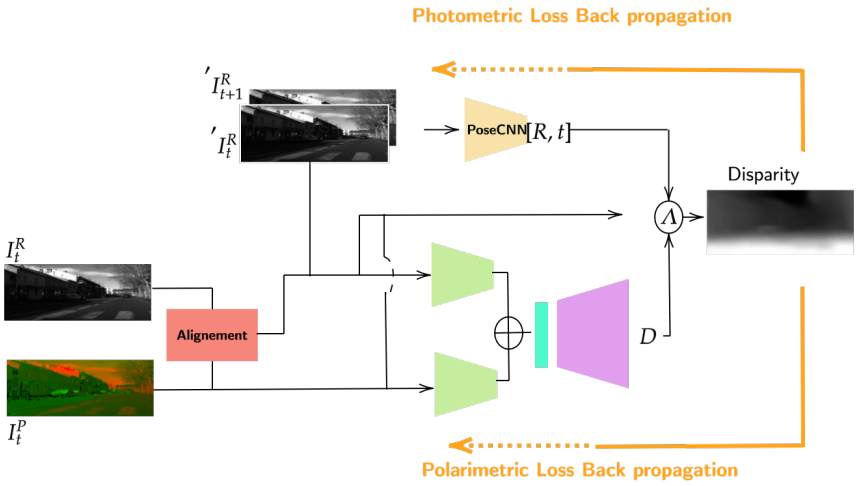


FIGURE 5.8: Latent space Fusion architecture illustration.

For our approach, we consider it would be convenient to force the distinguishability of non-mutual information. One of the possibilities we propose to investigate theoretically is the use of sparse coding. These networks based on sparse coding formulate the problem as a reconstruction of the initial image with a tolerance. The weights of the network subsequently become a descriptive dictionary that allows, from the feature vector, to reconstruct the initial image. This kind of network is traditionally trained by following a particular loss such that, given an input image X:

$$\Lambda = \frac{1}{2}||X - Cy||_2^2 + \lambda||y||_1, \tag{5.13}$$

with C the weight matrix, usually called dictionary in this field, and $\lambda > 0$, a weighting tradeoff parameter controlling the sparsity of the feature vector y. Thus, this Λ loss function aims to recover the initial image by penalizing the difference between the reconstruction and the initial image and ensure vector sparsity through L1 norm.

This kind of loss can allow for each modality to accurately describe the initial information. But in our case, the goal is to find a distinct space allowing to discriminate and separate the polarimetric information from the shared information. Thus, this kind of context can be delimited as follows:

$$y_p^p = y^p - y^r \quad with \quad y^p = y_m^p + y_p^p, \tag{5.14}$$

with y^p and y^r being respectively the feature vector from polarization and RGB image. Also, the indices m and p denotes for mutual and polarization information. Following this framework, it is possible to deduce two peculiar functions and more generally a joint loss allowing to reach this objective.

$$\gamma = \Lambda_p + \Lambda_r + ||C_r y_r - {}^{\{\iota\}} C_p y_p||_2^2 \tag{5.15}$$

with indiced Λ a modality related sparse coding loss fuction such that $\Lambda_p = \frac{1}{2}||X_p - C_p y_p||_2^2 + \lambda||y_p||_1$. In this formulation ${}^{\{\iota\}} C_p y_p$ represents the polarization intensity ι related reconstruction. In such way, the polarization parameters are discrimated and the function ensures for an accurate reconstruction of both mutual information shared through RGB and polarimetry.

Then, with an ensured distinguishable space between intensity and polarimetric parameters, it is possible to build an architecture allowing to take advantage of the supplementary information brought by the multi-modality without involving redundancy. Through this, the network can theoretically not be influenced as before by optimizing everything equally at once. The loss of photometric reconstruction as well as the estimation via PoseCNN can be operated exclusively on the mutual information while the polarimetry-related operations can be computed only with the polarimetric information. This kind of approach additionally allows to make clear connections between the input and output of the pipeline by making the spaces separable and distinct. Another considerable advantage of sparse coding is the economy of parameters. Through this kind of objective function, the general idea is to reconstruct a representative image at a more reasonable cost through the principle of sparsity.

In 2021, a promising approach [153] integrating sparse coding has emerged. Indeed, it extracts colorization and polarization information acquired with a bimodal sensor through a sparse coding optimizer deduced dictionary. This approach allows validating the hypothesis that these two information are separable. Moreover, the problem presented in the contribution is clearly more complex since the two pieces of information are, in the case of a bimodal sensor, physically merged. This confirms then the process described in this section can be exploited with two cameras with a prior image alignment or with one of the new cameras combining the two modalities into one. Moreover, it is still complex and costly to generate joint models allowing the modeling of vectors that share mutual information while keeping their differentiability for polarimetric information. Furthermore, this end-to-end training can lead to other constraints that make the optimization function fall into local minima. It would be more convenient to take advantage of already estimated depth maps and then refine them. In this sense, it is then possible to evaluate the possibilities of using a late fusion architecture.

5.4.1.3 Late fusion

Another possibility is based on a posteriori fusion from the networks. It consists in the encoding of two independent images towards a common space in which a fusion will be possible. Similar to latent space fusion, where the common space is examined during the intermediate step before decoding, here the common space is deduced at the end of the network for the two separate images. Next, it is a matter of finding a strategy to assemble these two transposed images to reach a final goal.

The general principle is based on two independent networks and then a final simple or multilayer network that will learn a fusion strategy.

This approach is totally suitable to the formulated problem, but the objective is shifted to the estimation of an objective function allowing the optimal fusion from the common space. In a first step, the intermediate objective can be set, namely to convert the source images. In our case of depth map refinement, the representation would then be a depth map specific to each modality. Thus, it would be possible to formulate two cost functions that would infuse the characteristics of each modality to obtain two accurate depth maps.

As shown in Figure 5.9, a preliminary per-modality depth estimation can be performed.

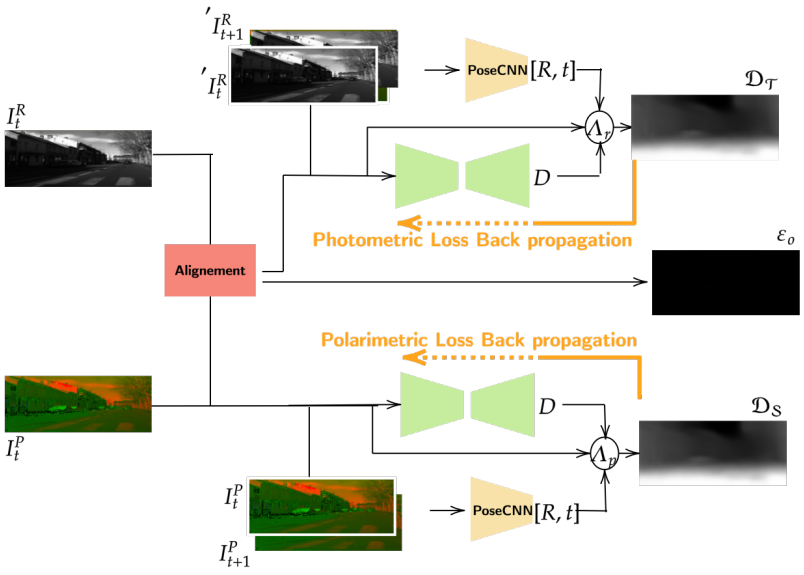


FIGURE 5.9: Late Fusion architecture illustration. Prior Estimations.

A first pipeline can consider the color modality by adopting the strategy of Godard et al. with its formulation of photometric error, smoothing and taking into account the occlusion and parallax parameters. A second pipeline addressing the polarization could only regularize the normals to match the polarization angle. This strategy would

allow obtaining sparse images but loaded with information where the polarization can be impactful. Discrimination with respect to the degree of polarization is then valid to consider only specular surfaces and to verify the equations relating normals and α. This approach would be a P2D relieved of the constraints of photometry-based perspective geometry.

At the end of these two pipelines, two depth maps would be obtained, one built using the perspective geometry formulation and the other verifying the validity with respect to polarimetry. Finally, these two maps could be aggregated using a fusion network based on raw uncertainty measurements. A strategy framework is proposed in Figure 5.10.

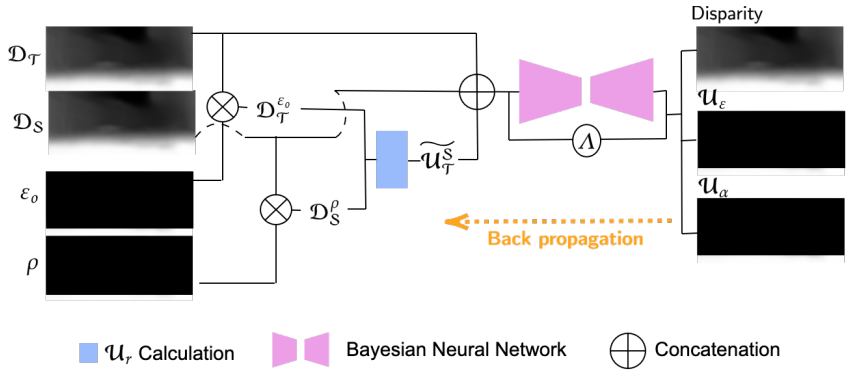


FIGURE 5.10: Late Fusion architecture illustration. Late refinement.

This uncertainty measure must integrate the errors inherent to the modalities or the cross modality alignment. We rectify the depth map according to the different errors that are available such as ε_o the alignment error and ρ the degree of polarization. While RGB image has been rectified, polarization cannot be affected by any transformation. Ultimately, the alignment error is impacted and due to the polarization modality. As for propagating the degree of polarization on the color-based estimation, this is necessarily due to the lack of modality knowledge. Color-based information do not observe any specularity awareness except by saturation. In highly specular area, it is a benefit to acknowledge such high impact data. In addition, most colorization-based monocular depth estimation approach are known to highly perform in textured/features regions and as contrary, to fail in different other cases (the saturation being one of those cases). In accordance, we scale inferred estimation by the inverse of ρ to manage those specific uncertainty as follows:

$$\begin{cases} \mathcal{D}_{\mathcal{T}}^{\varepsilon_o} = \mathcal{D}_{\mathcal{T}} \times \frac{1}{\rho} \\ \mathcal{D}_{\mathcal{S}}^{\rho} = \mathcal{D}_{\mathcal{S}} \times \varepsilon_o \end{cases} . \tag{5.16}$$

$\mathcal{D}_\mathcal{T}$ and $\mathcal{D}_\mathcal{S}$ are respectively the dispartities deduces from color and polarization. By substracting both the deduced depth maps, a new cross-modality estimation uncertainty can be computed such that:

$$\widetilde{\mathcal{U}_\mathcal{T}^\mathcal{S}} = |\mathcal{D}_\mathcal{S}^\rho - \mathcal{D}_\mathcal{T}^{\varepsilon\rho}|. \tag{5.17}$$

Another ultimate procedure to estimating uncertainty $\mathcal{U}$, although always relative since it is a student-teacher approach, is the method proposed by Poggi et al. [116]:

$$\widetilde{\mathcal{U}_\mathcal{T}^\mathcal{S}} = \frac{|\mu(\mathcal{D}_\mathcal{S}) - \mathcal{D}_\mathcal{T}|}{\sigma(\mathcal{D}_\mathcal{S})} + \log(\sigma(\mathcal{D}_\mathcal{S})) \tag{5.18}$$

The two previous methodologies have the advantage of requiring no tedious additional processing. However, as specified above, there is the notion of relativity which tends to imply the relative optimal is not the absolute optimal.

Now that we have three characteristic images, we need to consider a strategy to merge this information to refine a depth map. We propose a concept based on a Bayesian Neural Network (BNN) [41, 96]. Indeed, this kind of architecture allows modeling *epistemic* and *aleatoric* uncertainty [42]. As expressed in [75, 76] and [77], these two uncertainties allow to estimate respectively the robustness of the network and the impact of the data (mainly if it is subject to noise). Our singular goal is to refine a depth map and epistemic uncertainty modeling seems an excellent candidate to optimize in this direction. The overall idea would be to minimize the cross-modality uncertainty jointly with the epistemic uncertainty. This approach is almost necessary because there is no conspicuous domain where one can regularize two disparity measures. Such a method allows to have a differentiation and force the use of information coming from both modalities by the availability of $\widetilde{\mathcal{U}_\mathcal{T}^\mathcal{S}}$. Finally, as shown in Figure 5.10, it is possible to schematize this unsupervised architecture taking advantage of relative uncertainties but also of a global uncertainty modeled by the BNN.

In conclusion, this complex architecture would theoretically allow refining a depth map by taking advantage of both modality information and a complex uncertainty modelization through a BNN. On the other hand, this kind of setup requires a massive amount of data but especially a hardware architecture supporting an immense load of calculations. This constraint is directly attributed to the use of Bayesian neural network, simulating a large number of weights through Gaussian distributions. To reduce this charge and to make it feasible in the short term, a cascade modeling that would benefit from prior estimates could be advantageous.

5.4.1.4 Cascaded approach

The cascade approach is radically different from the previous ones since, in addition to not requiring end-to-end training of each pipeline, it is expected to take advantage of pre-learned components without altering the preliminarily observed results.

First, one takes the pre-trained Monodepth v2 network. The results inferred from an RGB image are almost optimal under favorable conditions but tend to deteriorate in the presence of specularity. We propose taking advantage of these estimates and to refine them in a second step using another network infused with polarization parameters. As shown in Figure 5.11, the cascade architecture is articulated in a single pipeline with two independent non-communicating estimation cores.

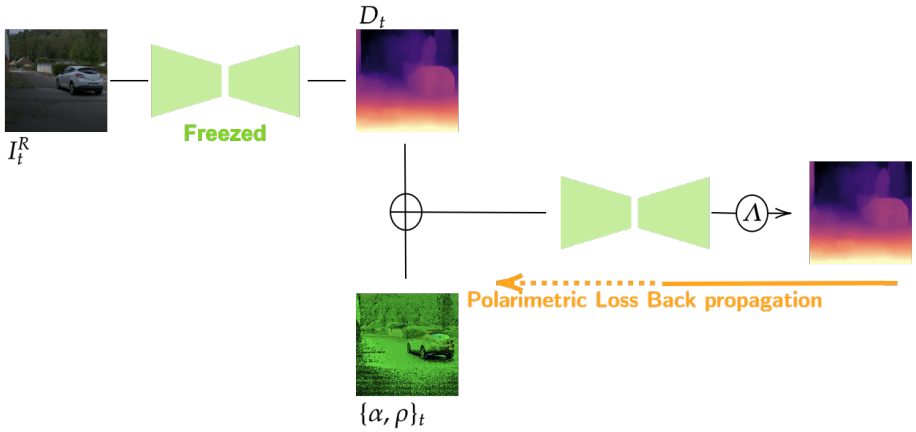


FIGURE 5.11: Cascaded Architecture illustration.

The main idea is to concatenate the pre-estimated depth map with the polarization parameters α and ρ. It is then possible to deduce a loss function which, using each of these channels, will in self-supervised manner, train the model to the unique locations where the RGB estimation fails. This objective function can be expressed as a function of the normals since the polarization angle allows to regularize them.

One can estimate a surface normal map from a depth map through oriented derivatives:

$$\Delta\mathcal{D} = \begin{bmatrix} g_x \\ g_y \end{bmatrix} = \begin{bmatrix} \frac{\delta\mathcal{D}}{\delta x} \\ \frac{\delta\mathcal{D}}{\delta y} \end{bmatrix}. \tag{5.19}$$

$$\vec{n} = \begin{bmatrix} -g_x \\ -g_y \\ 1 \end{bmatrix}. \tag{5.20}$$

As shown in Figure 5.12, from a depth map, one can compute the corresponding normals field.

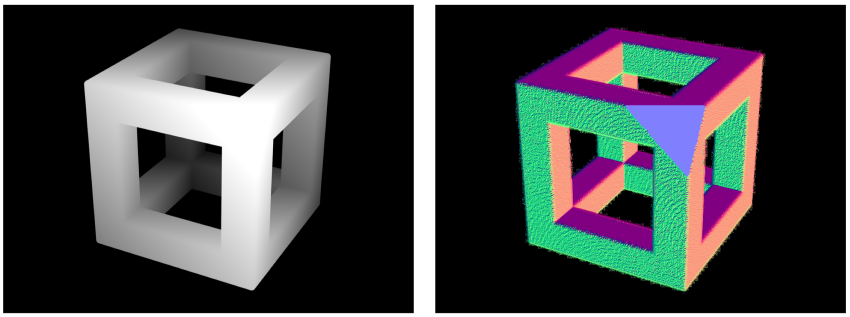


FIGURE 5.12: Illustration of depth to normals through oriented derivatives.

From this normals field , it is possible to calculate the angle in relation to the reference axis which will allow a simplified comparison with α:

$$\Theta = tan^{-a}\left(\frac{g_y}{g_x}\right). \tag{5.21}$$

Knowing the angle of polarization with normal orientation uncertainty and inspired from [37]:

$$A(\alpha, \Theta) = min(|\alpha - \Theta - \pi|, |\alpha - \Theta|, |\alpha - \Theta + \pi|), \tag{5.22}$$

with *min* operation allowing for accountance of polarimetry-related angle uncertainty with respect to the surface. Then, taking into account only specular surfaces and their related orientation uncertainty:

$$L_{pol} = \rho A(\alpha + \frac{\pi}{2}, \Theta). \tag{5.23}$$

Consequently, L_{pol} represents a self-supervision-compatible error term considering both the polarimetric information and the initial estimate from the color image. ρ is adequately used to consider only the pixels where specularity is observed and the relation α to $\vec{n}$ is verified.

This solution, although the others are potentially viable, seems to be the most appropriate and above all can allow for theory-testing prior experiments without the need for a large-scale dataset.

5.4.2 Prior experiments

To verify the cascade network theory, we propose a first evaluation of an image processing based method. This approach is in all respects similar to the principle expressed in Section 5.4.1.4.

Starting with a depth map estimate from Godard et al. [59], it is possible to deduce a normal map with equations 5.19 and 5.20.

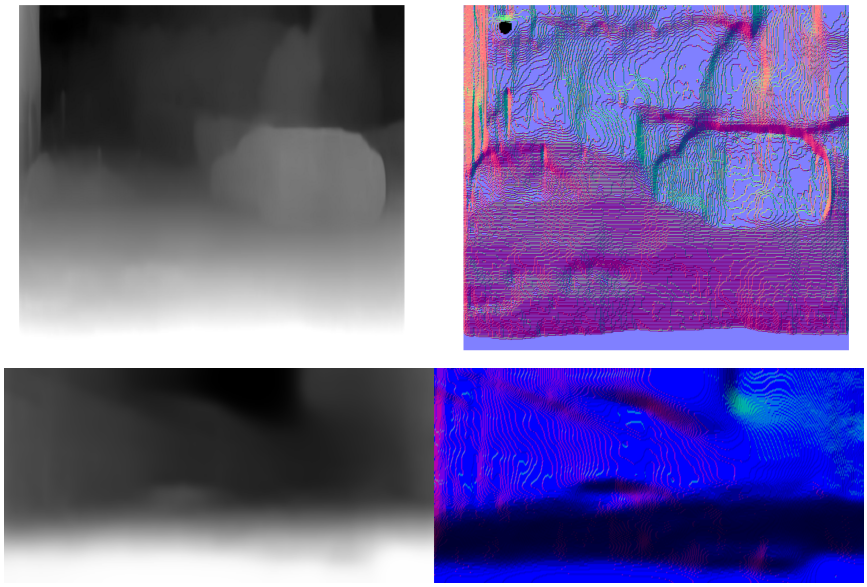


FIGURE 5.13: Normals estimation from Monodepth v2 estimate on two different scenes. Left column shows the depth image and right column the corresponding computed normals.

As shown in Figure 5.13, the resulting orientations are discussable. Especially in top row, normal vectors do not highlight properly the surfaces and tends to produce fictive shapes.

Assuming the normals field optimal (considering bottom row), one can compute the correspond angles following equation 5.21. A resulting grayscale image shown in Figure 5.14 can be displayed where $\Theta \in [0, \pi]$.

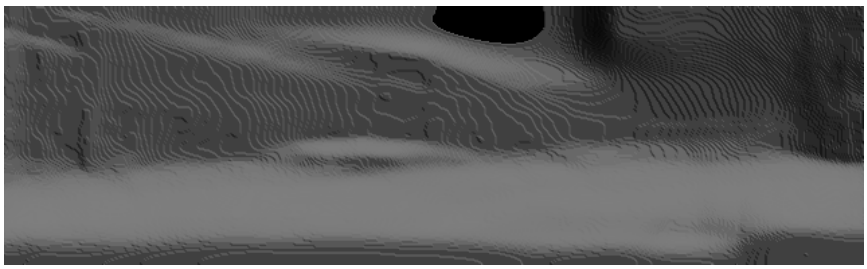


FIGURE 5.14: Angles computed from normals following equation 5.21.

In addition, the aligned angle of polarization α used for regularizing the previous angle map is shown in Figure 5.15.

Subsequently employing the two previous informative images, one can estimate the error through equation 5.23. ρ is subsequently enforcing the computation on specular surfaces and reducing the impact of polarization characteristics to only the desired

FIGURE 5.15: Aligned angle of polarization.

area. As shown in Figure 5.16, only few pixels are erronously estimated by the state-of-the-art colorization-based method. Consequently, this peculiar loss only focuses on these specific portions of the image that are specularity impacted.

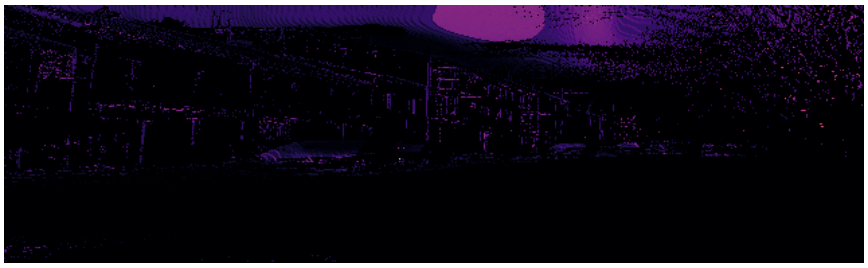


FIGURE 5.16: Resulting error image from normal angles comparison. The error is displayed following a colormap encouraging contrast. Thus, the brighter the higher.

To obtain a back propagation compatible error, the only missing element lie into either averaging or summing the image. The optimal results being full black image.

Thanks to the limited number of impacted pixels due to ρ filtering, the summation seems to represent the most ideal candidate avoiding vanishing gradient issues.

Ultimately, this technique while considering color information only involve the training of a polarization related network. This strategy is subsequently valid since the announced problem statement requires such a behaviour.

5.4.3 Deep learning based depth refinement through polarization cues

Since the pipeline has been validated by not involving any learning, it is possible to consider training a deep learning model on the same type of architecture. However, such experimentation requires making trade-offs. Indeed, data remains the major obstacle when addressing these bimodal learning issues. Therefore, to train an end-to-end depth map refinement architecture, it would be necessary to have an aligned RGB and polarization data set. An additional possibility is to use only polarimetric data and to "create" 3 channel images that can be interpreted by the first estimation network (see Figure 5.11).

Thus, since the polarization information is composed of three channels named intensity ι, angle α and degree ρ of polarization, an intensity stacking can be used as an input image for Godard et al. [59] similarly to the experiments conducted in Section 5.3. Taking advantage of the pre-trained architecture, we then obtain a robust depth estimation $\mathcal{D}$ that requires only refinements in the specular areas. Finally, $\mathcal{D}$ can be concatenated with the two other polarization parameters α and ρ to proceed to the training of the refinement network.

5.4.3.1 Loss adaptation

Initially, the cost function defined in equation 5.23 considered only the polarimetric part of the image. Indeed, this formulation only consists in identifying and quantifying the angular error between the normals deduced from the depth and those deduced from the polarization.

Unfortunately, this term is not restrictive enough and one of the optimal solutions towards which a network can tend is to infer a black image. Indeed, the impacted part of the images being limited due to the filtering by the degree of polarization, the global error is negligible. Moreover, no term enforces the network to infer a depth map. To counterbalance these two facts, we first restrict the network to retrieve a depth map by using a mean squared error (MSE):

$$L_d = \frac{1}{n}\sum_{x\in\chi}(\mathcal{D}(x) - \tilde{\mathcal{D}}(x))^2, \tag{5.24}$$

with L_d the depth retrieval term, χ indexing the pixels in the images, $\mathcal{D}$ the depth inferred from intensity stacking and $\tilde{\mathcal{D}}$ the final refined estimation. Note that L_d is not restrictive enough to solve the problem since the optimal solution consists in the reproduction of the initial map. However, this term is crucial to ensure the recovery of the valid depth pixels from Godard et al. [59] estimates.

As a reminder, our intention is to improve the areas where the polarization can be beneficial while keeping the integrity of the initial results of the state of the art method. To use the polarization, we propose to use jointly L_d and L_p from equation 5.23:

$$\Lambda = L_d + \beta L_p, \tag{5.25}$$

with Λ being the overall loss, L_d the depth recovery term, L_p the angle regularization from polarization term and β an empirically determinated scaling factor. This β factor is essential for a training involving both parts of the loss since, as previously stated, the error deduced from the polarization can be very limited due to the restrictions

imposed by ρ. In our case, we define that a factor $\beta = 50$ is adequate not to neglect the refinement part. Ultimately, the presented process can be assimilated to a pixel segmentation to which we add a regularization induced by the physics of polarization.

5.4.3.2 Experiments

As an initial step, one can validate one of the primary contraint consisting in refining only the pixels where the polarization information is meaningful. This "validation" step is performed by ρ filtering, which allows us to neglect all pixels where the degree of polarization is less than 0.4, or more simply, to neglect the non-specular / transparent surfaces.

In this first experiment shown in Figure 5.17, we feed the model a synthetic image $\{\mathcal{D}, \alpha, \rho\}$ which degree of polarization is set to zero.

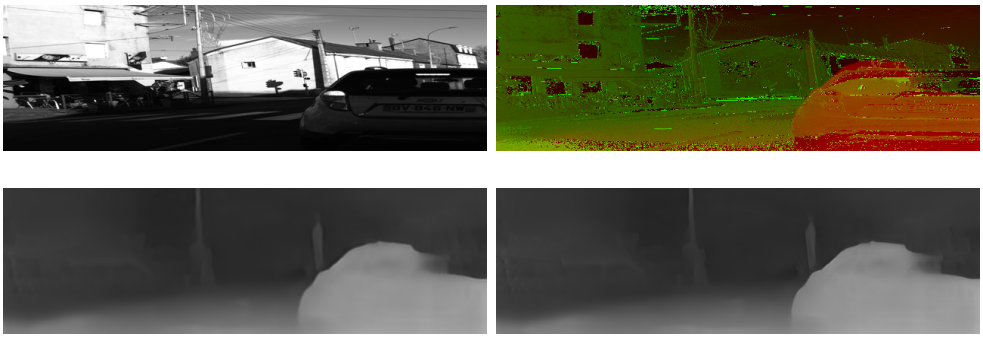


FIGURE 5.17: Depth refinement network validation without polarization. Top left is the Godard et al. [59] network input image, top right is the concatenation of initial depth $\mathcal{D}$, the angle of polarization α and the degree of polarization ρ. Bottom left is the output from the first network and bottom right is the final result from the cascade.

Once this step is validated, it is possible to provide the network the same constraints as P2D presented in Section 5.3. Thus, the database is identical, which allows a valid and unbiased evaluation to compare the networks.

5.4.3.3 Results and discussion

Figure 5.18 allow for a preliminary qualitative evaluation of the results. As preliminary results, images display the general behavior of the model. Indeed, the images show differences between the first estimation that is used as input of the second network and the final result forms the cascaded approach. Mainly, we can observe that most of the surfaces are preserved through the network with the exception of the sky and specular areas. This component was the key constraint for the method since the fusion approach should preserve unpolarized areas while refining only specular or transparent surfaces. In contrary, one can observe that the outputs are slightly deformed due to the formulation of the loss. Hence, one part account for preserving

the prior estimation from Godar et al. [59] method, the other quantify the normal angle error with respect to physically acquired information α. The mixture of both terms tends to produce approximate estimates and lead to erroneous estimates. As an example, the sky is badly reconstructed in the initial estimation and this kind of area contain polarization information. Despite the loss regularizing the estimations for specular surfaces, one can observe that the sky distance is still not correctly evaluated. In fact, the loss tends to produce smoothed estimates where there is contradiction between prior estimate and polarization-based loss term. This phenomenon can also be observed along transparent surfaces when looking for car windows / windshields. A straightforward solution would be to impact the weighting factor β to refine those estimates and make the network emphasize further on polarized areas. While in theory this technique can be used, $\beta = 50$ empirically fixed remains, at a preliminary stage, the best found coefficient. When $\beta < 50$, the network will optimize in a way that the specular areas will be neglected and the inference will produce identical results as the initial depth estimation network. When $\beta > 50$, the loss tends to diverge and will produce erroneous estimates with "waves" along specular surfaces. This coefficient is undoubtedly a key component of the loss and necessitates further experimentation to refine the results. However, we can deduce few facts from this deep learning based approach for depth refinement. First, the cascaded approach is valid and remain optimizable when addressing prior estimates refinement. Despite that this technique is questionably a fusion approach, prior experiments emphasize that, with suited loss, the network could be optimized to refine initial estimations. Next, manipulating the β factor highlighted that polarization-based term can highly affect the overall architecture output. In this configuration, a low factor will lead to trivial solution reproducing the input depth map, and a high factor will produce a divergent loss function causing erroneous estimates. This also highlighted that the loss function is fragile in this context. While, with P2D, a solely polarization image based loss was challenging to optimize due to the periodicity of the polarization information and the contradiction between photometric and polarimetric error, this cascaded architecture do not emphasize a less complex optimization procedure. Polarization-based term remain highly volatile and difficult to optimize, even with deep learning extensive capabilities for abstraction and simplification. Last, despite the simplification of the training procedure and the loss, prior experiments do not display evident capabilities for refining solely specular or transparent surfaces. As stated previously, since the method base its final estimation on a previously estimated map, the second network tends to find a trade-off without clear decision. That is to say, instead of totally re-evaluating the depth where the polarization allows it, the cascaded approach will tend to produce an

average between RGB-centric estimates and the error deduced from normal angle error. This last observation leads to the conclusion that a cascade approach is probably not adequate for this type of theme or that the loss is not properly established.

In conclusion, this first step towards depth map refinement using polarization and a cascade architecture remain encouraging. Whereas the first observations do not allow us to clearly estimate if this fusion approach is the most adequate, they have allowed us to delimit the constraints inherent to this kind of architecture. Experimentation with this loss needs to be deepened although it will probably be necessary to establish a new cost function that will optimize the specular surfaces while not altering the areas without polarization information. Although the other fusion approaches determined previously cannot be experimented with in the absence of massive RGB and Polarization aligned data, they had in common the ability to clearly separate the information channels. Future experiments based on these data could validate these architectures and thus prove whether or not it is possible to optimize a cost function that combines color-based photometric error and polarimetric error.

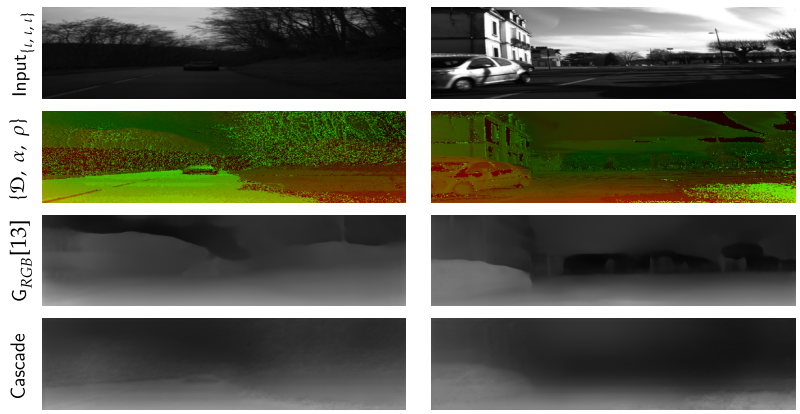


FIGURE 5.18: Qualitative evaluation of refinement method. Top row corresponds to first network input, second row the concatenation of depth estimated from first network and polarization parameters α and ρ. Third row corresponds to results from first network and last row to the final results of the refinement cascade. Specular details are correctly recovered and changed from initial estimates while unpolarized region mostly remain intact from the first to the second network output.

5.4.4 Conclusion

While P2D allowed a first step towards depth estimation by polarization alone, it observed aberrations in particular cases. In another domain, approaches requiring RGB information include specific sensitivities due to the modality used. Thus, specularity remains a notable weakness in this kind of algorithm.

To establish a generic polarimetric based method, some fusion architectures have been investigated. Starting with an estimation of the theoretical possibilities and needs, we have proposed a plausible panel of architectures that can combine the two modalities. This study has allowed to highlight the inability of some methodologies to face multimodality problems reliably. Since we had delimited the problem, no longer as an end-to-end estimation but as a refinement, many complex architecture were conceivable. This complexity and the prerequisites of these methods (namely latent space fusion and late fusion) indexed our study towards a cascade architecture. To prove the concept and since the process allowed it, we proposed to consider a no-learning experiment in real-world conditions. This proof of concept highlighted the possibility of operating such an algorithm. Finally, we have proposed a deep learning cascade refinement approach. Based on learning-free preliminary experiments, we used two networks, one of which is pre-trained with a state-of-the-art method. Mobilizing the knowledge acquired during the validation of fusion possibilities, it was possible to obtain preliminary results for RGB-based depth estimation refinement through polarization. Taking advantage of both the accurate estimates from an RGB-centric algorithm and the polarization information, this architecture displayed particular behavior, especially when addressing specular or transparent areas. First experiments showed optimization capabilities of such approach while not entirely validating the method. In brief, such approach requires further experimentation either by modulating the different component of the loss either by re-evaluate loss terms to suit the problem more adequately.

Nevertheless, other methods requiring much higher computational power are still plausible and remain candidates to perform an accurate depth estimation/refinement task from polarization.

5.5 Summary

This chapter addresses the problem of depth estimation from polarization. Such polarization-based and unique methods are undoubtedly at a preliminary stage. However, the multiple approaches have highlighted that polarization can allow, through different information, to infer depth maps in a whole new way. P2D is, to our knowledge, the first-of-its-kind method proposed to infer a depth map from a polarimetric single image and a DCNN. The results, although very promising, has displayed some weaknesses. In response to these erroneous estimates, we proposed examining the possibilities to go farther in the field by exploiting preliminary achievements in RGB. This study allowed us to evaluate the possibilities of fusion and eliminate the candidates which were unsuitable for the task. Additionally, we proposed, without learning involved, to determine if a fusion approach was viable. We conclude that apart from

Early Fusion, there is a vast possibility to improve the results of P2D and especially to make the approach more generic. Based on this study, we proposed a deep learning approach using a cascade of two network to propose a first step towards depth refinement through polarization. Thus, this approach was able to highlight another possibility to use polarization, this time in the form of fusion, to improve the performance of depth estimation of specular areas. Having no dataset with both color and polarization imaging aligned, the experiments were conducted under restricted conditions by depriving the first network of the colorization information.

However, the data barriers remain the principal obstacle, especially in the area of exploitation of specific unconventional data. We are confident that over time, as more polarimetric data become available, the greedy algorithms can be viably used to finally exploit these data. Due to this lack of data, some experiments couldn't be conducted for all the fusion approaches. Despite this, it is clear that these fusion strategies could be exploited as soon as a dataset emerges in the scientific community.

In conclusion, we have proven polarization remain a discriminative modality that, when used judiciously, could improve the performance of algorithms for geometric understanding of urban scenes.

Chapter 6

Conclusion, Perspectives and Future Work

6.1 Conclusion

The general aim of this thesis was to improve scene understanding algorithms by including polarization cues. In Chapter 2, we introduced the fundamental preliminary knowledge of polarimetry and deep learning. We in addition proposed brief explanations on specific concepts used through this thesis. In Chapter 3, we first explored existing segmentation methods using state-of-the-art deep learning techniques. In addition, we proposed a comprehensive review of the depth estimation field starting from sensor acquisition to deep learning based depth inference. Chapter 4 presented the first of two axis of the research conducted during this PhD thesis. Exploring the vast landscape of deep learning PwSS, this chapter emphasizes the complexity to introduce a non-conventional modality to the segmentation field. Thus, we proposed details on procedures ranging from dataset collection to singular augmentation to finalize on segmentation. In more detail, we have presented an end-to-end framework allowing exploitation of polarization cues through deep learning architectures. We subsequently manage to propose two polarimetric-influenced architecture emphasizing either the suitability of the data for such tasks or the necessity of modality-unique augmentation. Both our evaluation benchmarks highlighted extensive segmentation accuracy when addressing the problematic through polarisation prism. Therefore, proof has been made specularity understanding models perform equally or better than texture/color based. In Chapter 5, the second axis has been exposed. Depth estimation being a recurent problem, we attempted to use polarization to constraint monodepth approach. By aggregating sufficient data and designing a modality representation, we were capable to formalise a surface normal rectification term base on polarization cues. Constraining the problem through this bias showed extensive capabilities of reconstruction specular and transparent surfaces as well as sky delimitation. As a first approach, up to our knowledge, including such modality in the field,

we proposed a comparison with state-of-the-art method in the same conditions. Subsequently, this quantitative evaluation of P2D displayed better performances when the network is subject to polarimetric data. Extensive experimentation highlighted P2D being far from perfect generalization. That is to say, in some cases, the method tends to provide erroneous estimates. As a response, Chapter 5 equally provides suggestions of multiple multimodal fusion method. Aiming at combining advantages of both RGB and polarimetric modality, the problem has been formulated as a refinement instead of an end-to-end estimation. From this methods study, numerous strategies have been proposed to eliminate drawbacks of the previously designed method. Ultimately, we proposed a fusion process based on network cascade. This last experimentation underline the optimization viability of such fusion approaches to refine depth map while remaining a preliminary approach towards robust multimodal fusion for depth refinement.

All the proposed methods have shown excellent performance in both segmentation and depth estimation. We have aggregated specific data which have made accessible greedy algorithms like deep learning. It is undeniable that polarimetric data can be beneficial to a large number of approaches.

As a general conclusion, it is substantial to acknowledge the importance of data. We have proven that by wisely using unconventional data, algorithms can benefit from it. Despite the tendency to massively aggregate data or to make algorithms more complex, our methods have shown that with less, algorithms can be more accurate. The information provided to deep learning networks is frequently overlooked at the expense of robustness and genericity. Ultimately, a thoughtful choice of data can constrain the problem and simplify it. To such a degree, with less data, it is possible to observe excellent performances. It is impossible to consider autonomy themes if some recurrent phenomena are neglected (e.g. specularity or transparency). Through this thesis, in addition to the proposed methods, we aim to highlight the importance of data. Thus, we hope that estimation using unconventional modalities will be popularized. It would be attractive to move from color-based to physics-based vision and thereby feed collaborations allowing the provision of such data.

6.2 Perspective for polarization in modern computer vision

Polarization is a field that has been studied for a very long time. As it was shown in Chapter 2, this area has been active for many years. Unfortunately, this is a niche too rarely explored. Polarization offers a viable alternative to many computer vision algorithms as shown in [105, 104, 106, 37, 10, 122, 21, 20, 19]. It is notable this

modality allows to explore a new angle of vision and defines a completely different space. This unconventional information could be exploited to improve the performance of approaches for which color is not a sufficient discriminant. Finally, the lack of data remain the prime factor preventing the exploration of this domain in the long term with recent algorithms. Recently, the price of sensors has largely dropped, and technologies have advanced. While in the past a camera could cost several thousand euros for a minimal resolution, nowadays, these sensors are much more affordable (less than 1,000 euros) and observe a 4K resolution. In addition, some manufacturers have focused on combining multiple modalities by including polarizing filters on top of standard capture arrays. We believe that these advances are encouraging for the field and that these sensors could encourage a wider diffusion of polarization. Based on these facts, we propose some perspectives on polarization in modern computer vision. Indeed, this thesis has highlighted that polarimetry can represent an alternative modality for scene understanding. Integrating particular data, it has allowed, upstream of the network, to categorize previously neglected phenomena. The networks have proven their adaptability by revealing the capacity to learn from such information. Finally, a modality such as this one allows a direct link between physics and processing. Many approaches operate in environments prone to specularity, and there is every reason to believe that polarization could be used instead of color imaging. It would then be possible to estimate 3D motion from the polarization angle or the optical flow. Deducing the camera pose from the orientation changes observed in polarization (an idea derived from our augmentation process), segmenting specular instances, or indeed trying to alleviate Fresnel equations to trace the physical properties of materials. But also, many possibilities could emerge from the fusion procedures. Moreover, as we have proposed, it would be possible to aggregate the data to refine certain procedures. Depth estimation, panoptic segmentation, SLAM, all these domains could be based on light vectorial information in addition to texture.

Finally, thanks to deep learning networks, any information can be exploited as long as there is enough of it. The advent of these processing cores has extended the field of possibilities in vision. Their abstraction capacities have released many past constraints and allowed us to observe increased performances. Ultimately, the only thing missing in polarimetry is a community that collaborates to aggregate data to enlarge the number of interested parties. We hope that this thesis has helped to bring the field of polarization to the forefront by showing its appeal, and it will motivate others to contribute to this component of physics-based vision.

6.3 Future Work

Based on the work presented in this thesis, we give some recommendations for future research.

In Chapter 4 we proposed a segmentation approach. Although the procedure showed interesting performances, it is considerable that the amount of data was small. Also, the cases were particular and the dataset very focused. Thus, it would be attractive to aggregate a larger amount of data by integrating more cases and varied scenes. It would then be possible to compare much more globally the performances between polarization-centric and RGB-centric algorithms. As follows, it would be possible to really compare the generalization capabilities of the networks impacted by polarimetry. An ablation study could be performed to show the generalization and discrimination capabilities of such a method.

In Chapter 5 we proposed P2D, an approach for depth inference from a polarimetric monocular. Despite encouraging performances, the network showed a tendency to produce erroneous estimates in some contexts. In response to this, in this same Chapter, we proposed methods based on multimodal fusion to improve the results and refine the estimated depth maps using a state-of-the-art RGB-centric method. We have estimated several approaches, their respective counterparts and their associated losses. The possibility would be that, starting from these described architectures, a multimodal dataset would be acquired allowing the implementation of such approaches to finally verify the viability of the stated concepts.

In conclusion, many of the objectives were stated in Section 6.2. As deep learning becomes more able at providing reliable estimates, it would be interesting to benefit from it with an unconventional data. We expect these greedy algorithms to facilitate the use of complex modalities and thus offer valuable opportunities for innovation in the upcoming years.

Bibliography

[1] G. J. Agin, "Computer vision systems for industrial inspection and assembly.," *IEEE Computer*, vol. 13, no. 5, pp. 11–20, 1980.

[2] J. Aloimonos, "Shape from texture," *Biological cybernetics*, vol. 58, no. 5, pp. 345–360, 1988.

[3] E. Artin, *Geometric algebra, volume 3 of interscience tracts in pure and applied mathematics*, 1957.

[4] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *CoRR*, vol. abs/1505.07293, 2015. arXiv: 1505.07293. [Online]. Available: http://arxiv.org/abs/1505.07293.

[5] R. Baer, *Linear algebra and projective geometry*. Courier Corporation, 2005.

[6] J. Banks and P. Corke, "Quantitative evaluation of matching methods and validity measures for stereo vision," *The International Journal of Robotics Research*, vol. 20, no. 7, pp. 512–532, 2001.

[7] C. Banz, S. Hesselbarth, H. Flatt, H. Blume, and P. Pirsch, "Real-time stereo vision system using semi-global matching disparity estimation: Architecture and fpga-implementation," in *2010 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation*, IEEE, 2010, pp. 93–101.

[8] A. Bensrhair, P. Miché, and R. Debrie, "Fast and automatic stereo vision matching algorithm based on dynamic programming method," *Pattern Recognition Letters*, vol. 17, no. 5, pp. 457–466, 1996.

[9] K. Berger and M. Blanchon, "Introduction to polarization for rendering and vision," in *SIGGRAPH Asia 2020 Courses*, ser. SA '20, Virtual Event: Association for Computing Machinery, 2020, ISBN: 9781450381123. DOI: 10.1145/3415263.3419172. [Online]. Available: https://doi.org/10.1145/3415263.3419172.

[10] K. Berger, R. Voorhies, and L. H. Matthies, "Depth from stereo polarization in specular scenes for urban robotics," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, IEEE, 2017, pp. 1966–1973.

[11] K. Bhargavi and S. Jyothi, "A survey on threshold based segmentation technique in image processing," *International Journal of Innovative Research and Development*, vol. 3, no. 12, pp. 234–239, 2014.

[12] P. Bilinski and V. Prisacariu, "Dense decoder shortcut connections for single-pass semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6596–6605.

[13] M. Blanchon, O. Morel, F. Meriaudeau, R. Seulin, and D. Sidibé, "Polarimetric image augmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 7365–7371.

[14] M. Blanchon, O. Morel, Y. Zhang, R. Seulin, N. Crombez, and D. Sidibé, "Outdoor scenes pixel-wise semantic segmentation using polarimetry and fully convolutional network.," 2019.

[15] ——, "Utilisation de la polarimétrie pour la segmentation de scènes extérieures avec un réseau convolutif.," 2019.

[16] M. Blanchon, D. Sidibé, O. Morel, R. Seulin, D. Braun, and F. Meriaudeau, "P2d: A self-supervised method for depth estimation from polarimetry," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, 2021, pp. 7357–7364.

[17] M. Blanchon, D. Sidibé, O. Morel, R. Seulin, and F. Meriaudeau, "Towards urban scenes understanding through polarization cues," *arXiv preprint arXiv:2106.01717*, 2021.

[18] M. Bleyer and M. Gelautz, "A layered stereo algorithm using image segmentation and global visibility constraints," in *2004 International Conference on Image Processing, 2004. ICIP'04.*, IEEE, vol. 5, 2004, pp. 2997–3000.

[19] R. Blin, S. Ainouz, S. Canu, and F. Meriaudeau, "A new multimodal rgb and polarimetric image dataset for road scenes analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 216–217.

[20] R. Blin, S. Ainouz, S. Canu, and F. Meriaudeau, "Adapted learning for polarization-based car detection," in *Fourteenth International Conference on Quality Control by Artificial Vision*, International Society for Optics and Photonics, vol. 11172, 2019, p. 1 117 218.

[21] ——, "Road scenes analysis in adverse weather conditions by polarization-encoded images and adapted deep learning," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2019, pp. 27–32.

[22] A. E. Bryson, "A gradient method for optimizing multi-stage allocation processes," in *Proc. Harvard Univ. Symposium on digital computers and their applications*, vol. 72, 1961, p. 22.

[23] M. E. Caine, "The design of shape from motion constraints," MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, Tech. Rep., 1993.

[24] N. D. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla, "Using multiple hypotheses to improve depth-maps for multi-view stereo," in *European Conference on Computer Vision*, Springer, 2008, pp. 766–779.

[25] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[26] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8001–8008.

[27] F. Chaumette and S. Hutchinson, "Visual servo control. i. basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, 2006.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[29] ——, *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, 2017. arXiv: 1606.00915 [cs.CV].

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," pp. 801–818, 2018.

[31] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2624–2632.

[32] A. Chhatkuli, D. Pizarro, and A. Bartoli, "Non-rigid shape-from-motion for isometric surfaces using infinitesimal planarity.," in *BMVC*, 2014.

[33] C.-C. Chu and J. K. Aggarwal, "The integration of region and edge-based segmentation," in *Proceedings Third International Conference on Computer Vision*, IEEE Computer Society, 1990, pp. 117–118.

[34] E. Collett, "Field guide to polarization," Spie Bellingham, WA, 2005.

[35] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[36] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, "Polarimetric multi-view stereo," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1558–1567.

[38] Z. Cui, V. Larsson, and M. Pollefeys, "Polarimetric relative pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2671–2680.

[39] F. Dellaert, S. M. Seitz, C. E. Thorpe, and S. Thrun, "Structure from motion without correspondence," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, IEEE, vol. 2, 2000, pp. 557–564.

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[41] J. S. Denker and Y. LeCun, "Transforming neural-net output levels to probability distributions," in *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, 1990, pp. 853–859.

[42] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.

[43] N. Dong and E. P. Xing, "Few-shot semantic segmentation with prototype learning.," in *BMVC*, vol. 3, 2018.

[44] S. Edelman and T. Poggio, "Integrating visual cues for object segmentation and recognition," *Optics News*, vol. 15, no. 5, p. 8, 1989.

[45] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[46] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in neural information processing systems*, 2014, pp. 2366–2374.

[47] K. El Madawi, H. Rashed, A. El Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, IEEE, 2019, pp. 7–12.

[48] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[49] P. Favaro and S. Soatto, "A geometric approach to shape from defocus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 406–417, 2005.

[50] Y. Feng, P. Whatmough, and Y. Zhu, "Asv: Accelerated stereo vision system," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2019, pp. 643–656.

[51] A. Fresnel, *Oeuvres complètes d'augustin Fresnel*. Imprimerie impériale, 1868.

[52] J. Fritsch, T. Kuehnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[53] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[54] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, Springer, 2016, pp. 740–756.

[55] S. K. Gehrig, F. Eberli, and T. Meyer, "A real-time low-power stereo vision engine using semi-global matching," in *International Conference on Computer Vision Systems*, Springer, 2009, pp. 134–143.

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

[57] M. A. Gennert, "Brightness-based stereo matching," in *ICCV*, 1988.

[58] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

[59] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3828–3838.

[60] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[61] X. Guo, H. Li, S. Yi, J. Ren, and X. Wang, "Learning monocular depth by distilling cross-domain stereo networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 484–500.

[62] P. Gysel, M. Motamedi, and S. Ghiasi, "Hardware-oriented approximation of convolutional neural networks," *arXiv preprint arXiv:1604.03168*, 2016.

[63] R. Hartley, *A. zisserman multiple view geometry in computer vision*, 2000.

[64] R. I. Hartley, "Projective reconstruction and invariants from multiple images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 10, pp. 1036–1041, 1994.

[65] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*, Springer, 2016, pp. 213–228.

[66] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[68] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[69] B. K. Horn and M. J. Brooks, "The variational approach to shape from shading," *Computer Vision, Graphics, and Image Processing*, vol. 33, no. 2, pp. 174–208, 1986.

[70] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[71] S. Hrabar, "An evaluation of stereo and laser-based range sensing for rotorcraft unmanned aerial vehicle obstacle avoidance," *Journal of Field Robotics*, vol. 29, no. 2, pp. 215–239, 2012.

[72] G. Iannizzotto and L. Vita, "Fast and accurate edge-based segmentation with no contour smoothing in 2-d real images," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1232–1237, 2000.

[73] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *arXiv preprint arXiv:1506.02025*, 2015.

[74] H. G. Kaganami and Z. Beiji, "Region-based segmentation versus edge detection," in *2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, IEEE, 2009, pp. 1217–1221.

[75] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," *arXiv preprint arXiv:1511.02680*, 2015.

[76] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *2016 IEEE international conference on Robotics and Automation (ICRA)*, IEEE, 2016, pp. 4762–4769.

[77] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.

[78] M. Klodt and A. Vedaldi, "Supervising the new with the old: Learning sfm from sfm," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 698–713.

[79] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," in *European conference on computer vision*, Springer, 2002, pp. 82–96.

[80] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.

[81] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "Adadepth: Unsupervised content congruent adaptation for depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2656–2665.

[82] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6647–6655.

[83] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[84] J. M. Lee, *Riemannian manifolds: an introduction to curvature*. Springer Science & Business Media, 2006, vol. 176.

[85] G. W. Leibniz, "Reihe vii: Mathematische schriften vol.5 infinitesimal mathematik," in ser. Infinitesimalmathematik. -, 1674-1676, vol. 5, pp. 321–331.

[86] H.-Y. Li, H.-Q. Wang, and C.-B. Ding, "A new solution of automatic building extraction in remote sensing images," in *2006 IEEE International Symposium on Geoscience and Remote Sensing*, IEEE, 2006, pp. 3790–3793.

[87] N. Li, Y. Zhao, Q. Pan, and S. G. Kong, "Demosaicking dofp images using newton's polynomial interpolation and polarization difference model," *Optics express*, vol. 27, no. 2, pp. 1376–1391, 2019.

[88] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2018, pp. 7286–7291.

[89] G. Lin, A. Milan, C. Shen, and I. Reid, *Refinenet: Multi-path refinement networks for high-resolution semantic segmentation*, 2016. arXiv: 1611.06612 [cs.CV].

[90] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5162–5170.

[91] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[92] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[93] C. Luo, Z. Yang, P. Wang, Y. Wang, W. Xu, R. Nevatia, and A. Yuille, "Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2624–2641, 2019.

[94] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 155–163.

[95] L. Ma, J. Stückler, C. Kerl, and D. Cremers, "Multi-view deep learning for consistent semantic mapping with rgb-d cameras," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 598–605.

[96] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[97] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5667–5675.

[98] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 204, no. 1156, pp. 301–328, 1979.

[99] J. L. Marroquin, E. A. Santana, and S. Botello, "Hidden markov measure field models for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 11, pp. 1380–1387, 2003.

[100] I. Mehta, P. Sakurikar, and P. Narayanan, "Structured adversarial training for unsupervised monocular depth estimation," in *2018 International Conference on 3D Vision (3DV)*, IEEE, 2018, pp. 314–323.

[101] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[102] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[103] M. Michael, J. Salmen, J. Stallkamp, and M. Schlipsing, "Real-time stereo vision: Optimizing semi-global matching," in *2013 IEEE Intelligent Vehicles Symposium (IV)*, IEEE, 2013, pp. 1197–1202.

[104] O. Morel, M. Ferraton, C. Stolz, and P. Gorria, "Active lighting applied to shape from polarization," in *2006 International Conference on Image Processing*, IEEE, 2006, pp. 2181–2184.

[105] O. Morel, F. Meriaudeau, C. Stolz, and P. Gorria, "Polarization imaging applied to 3d reconstruction of specular metallic surfaces," in *Machine Vision Applications in Industrial Inspection XIII*, International Society for Optics and Photonics, vol. 5679, 2005, pp. 178–187.

[106] O. Morel, R. Seulin, and D. Fofi, "Catadioptric camera calibration by polarization imaging," in *Iberian Conference on Pattern Recognition and Image Analysis*, Springer, 2007, pp. 396–403.

[107] S. Mukherjee and S. T. Acton, "Region based segmentation in presence of intensity inhomogeneity using legendre polynomials," *IEEE Signal Processing Letters*, vol. 22, no. 3, pp. 298–302, 2014.

[108] P. K. Nathan Silberman Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[109] V. Nekrasov, C. Shen, and I. Reid, "Light-weight refinenet for real-time semantic segmentation," *arXiv preprint arXiv:1810.03272*, 2018.

[110] Y.-i. Ohta, T. Kanade, and T. Sakai, "An analysis system for scenes containing objects with substructures," in *Proceedings of the Fourth International Joint Conference on Pattern Recognitions*, 1978, pp. 752–754.

[111] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[112] S. Parashar, D. Pizarro, and A. Bartoli, "Isometric non-rigid shape-from-motion in linear time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4679–4687.

[113] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.

[114] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *CoRR*, vol. abs/1712.04621, 2017. arXiv: 1712.04621. [Online]. Available: http://arxiv.org/abs/1712.04621.

[115] S. Pillai, R. Ambruş, and A. Gaidon, "Superdepth: Self-supervised, super-resolved monocular depth estimation," in *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, 2019, pp. 9250–9256.

[116] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3227–3237.

[117] M. Poggi, F. Tosi, and S. Mattoccia, "Learning monocular depth estimation with unsupervised trinocular assumptions," in *2018 International conference on 3d vision (3DV)*, IEEE, 2018, pp. 324–333.

[118] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.

[119] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black, "Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 240–12 249.

[120] H. Rashed, A. El Sallab, S. Yogamani, and M. ElHelw, "Motion and depth augmented semantic segmentation for autonomous navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[121] E. Rashedi and H. Nezamabadi-Pour, "A stochastic gravitational approach to feature based color image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 4, pp. 1322–1332, 2013.

[122] M. Rastgoo, C. Demonceaux, R. Seulin, and O. Morel, "Attitude estimation from polarimetric cameras," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems-IROS*, 2018.

[123] B. M. Ratliff, C. F. LaCasse, and J. S. Tyo, "Interpolation strategies for reducing ifov artifacts in microgrid polarimeter imagery," *Optics express*, vol. 17, no. 11, pp. 9112–9125, 2009.

[124] C. Reutenauer, "Free lie algebras," in *Handbook of algebra*, vol. 3, Elsevier, 2003, pp. 887–903.

[125] G. Riegler, Y. Liao, S. Donne, V. Koltun, and A. Geiger, "Connecting the dots: Learning representations for active monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7624–7633.

[126] J. Rodrıéguez-Quiñonez, O. Sergiyenko, W. Flores-Fuentes, M. Rivas-Lopez, D. Hernandez-Balbuena, R. Rascón, and P. Mercorelli, "Improve a 3d distance measurement accuracy in stereo vision systems using optimization methods' approach," *Opto-Electronics Review*, vol. 25, no. 1, pp. 24–32, 2017.

[127] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[128] K. Rossmann, "Point spread-function, line spread-function, and modulation transfer function: Tools for the study of imaging systems," *Radiology*, vol. 93, no. 2, pp. 257–272, 1969.

[129] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5506–5514.

[130] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[131] A. D. Sappa, "Unsupervised contour closure algorithm for range image edge-based segmentation," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 377–384, 2006.

[132] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[133] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, IEEE, vol. 2, 2001, pp. 2051–2058.

[134] M. Siam, B. N. Oreshkin, and M. Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5249–5258.

[135] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[136] F. Stein, "Efficient computation of optical flow using the census transform," in *Joint Pattern Recognition Symposium*, Springer, 2004, pp. 79–86.

[137] F. Steinbrücker, J. Sturm, and D. Cremers, "Real-time visual odometry from dense rgb-d images," in *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, IEEE, 2011, pp. 719–722.

[138] G. G. Stokes, "On the composition and resolution of streams of polarized light from different sources," *Transactions of the Cambridge Philosophical Society*, vol. 9, p. 399, 1851.

[139] C. Stolz, M. Ferraton, and F. Meriaudeau, "Shape from polarization: A method for solving zenithal angle ambiguity," *Optics letters*, vol. 37, no. 20, pp. 4218–4220, 2012.

[140] Y. Sumi, Y. Kawai, T. Yoshimi, and F. Tomita, "3d object recognition in cluttered environments by segment-based stereo vision," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 5–23, 2002.

[141] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[142] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.

[143] B. Van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever, "Active shape model segmentation with optimal features," *IEEE transactions on medical imaging*, vol. 21, no. 8, pp. 924–933, 2002.

[144] J. Vaze and J. Teng, "High resolution lidar dem–how good is it," *Modelling and Simulation*, pp. 692–698, 2007.

[145] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, and A. Courville, "Reseg: A recurrent neural network-based model for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 41–48.

[146] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "Renet: A recurrent neural network based alternative to convolutional networks," *arXiv preprint arXiv:1505.00393*, 2015.

[147] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2022–2030.

[148] R. Wang, "3d building modeling using images and lidar: A review," *International Journal of Image and Data Fusion*, vol. 4, no. 4, pp. 273–292, 2013.

[149] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[150] M. A. Wani and B. G. Batchelor, "Edge-region-based segmentation of range images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 3, pp. 314–319, 1994.

[151] R. Watt, "Feature-based image segmentation in human vision," *Spatial Vision*, vol. 1, no. 3, pp. 243–256, 1986.

[152] S. Weder, J. Schonberger, M. Pollefeys, and M. R. Oswald, "Routedfusion: Learning real-time depth map fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4887–4897.

[153] S. Wen, Y. Zheng, and F. Lu, "A sparse representation based joint demosaicing method for single-chip polarized color sensor," *IEEE Transactions on Image Processing*, vol. 30, pp. 4171–4182, 2021.

[154] L. B. Wolff and A. G. Andreou, "Polarization camera sensors," *Image and Vision Computing*, vol. 13, no. 6, pp. 497–510, 1995.

[155] O. J. Woodford, "Global Stereo Reconstruction under Second Order Smoothness Priors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 12, pp. 2115–2128, 2008.

[156] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher-student learning," *arXiv preprint arXiv:1810.08476*, 2018.

[157] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[158] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5354–5362.

[159] L. Yang, F. Tan, A. Li, Z. Cui, Y. Furukawa, and P. Tan, "Polarimetric dense monocular slam," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3857–3866.

[160] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1281–1292.

[161] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 817–833.

[162] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 225–234.

[163] Z. Yang, P. Wang, W. Xu, L. Zhao, and R. Nevatia, "Unsupervised learning of geometry with edge-aware depth-normal consistency," *arXiv preprint arXiv:1711.03665*, 2017.

[164] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1983–1992.

[165] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[166] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*, Springer, 1994, pp. 151–158.

[167] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[168] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry

with deep feature reconstruction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 340–349.

[169] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[170] J. Zhang, H. Luo, R. Liang, A. Ahmed, X. Zhang, B. Hui, and Z. Chang, "Sparse representation-based demosaicing method for microgrid polarimeter imagery," *Optics letters*, vol. 43, no. 14, pp. 3265–3268, 2018.

[171] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, "Shape-from-shading: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 21, no. 8, pp. 690–706, 1999.

[172] Y. Zhang, O. Morel, M. Blanchon, R. Seulin, M. Rastgoo, and D. Sidibé, "Exploration of deep learning-based multimodal fusion for semantic road scene segmentation.," 2019.

[173] ——, "Exploration of deep learning-based multimodal fusion for semantic road scene segmentation.," 2019.

[174] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image and Vision Computing*, p. 104 042, 2020.

[175] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 269–284.

[176] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[177] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal on Computer Vision*, 2018.

[178] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[179] D. Zhu and W. A. Smith, "Depth from a polarisation+ rgb stereo pair," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7586–7595.

[180] S. Zhuo and T. Sim, "On the recovery of depth from a single defocused image," in *International Conference on Computer Analysis of Images and Patterns*, Springer, 2009, pp. 889–897.

[181] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 36–53.

Appendix A

Table of Cameras Characteristics

Table A.1: Characteristics of Cameras

	Kinect	UCam	NIR	Polarimetric
Constructor	Microsoft	IDS	Marlin	4D Technology
Reference	Kinect for Windows 2	USB 3 uEye CP	MAKO G G-131	PolarCam Model V
Interface	USB 3.0	USB 3.0	IEEE 802.3 1000BASE-T, IEEE 802.3af (PoE)	GigE Ethernet
Theoretical FPS	30	50	62	135
Resolution	1920 x 1080	1280 x 1024	640 x 480	640 x 460
Modality	RGB (D)	RGB	GS	Polarimetry (GS)

Appendix B

Segmentation Metrics

The *IoU* is defined as:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}. \tag{B.1}$$

The *recall* is defined as:

$$\text{S} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{B.2}$$

The *precision* is defined as:

$$\text{S} = \frac{\text{True Pisitives}}{\text{True Positives} + \text{False Positives}}. \tag{B.3}$$

The *specificity* is defined as:

$$\text{S} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}. \tag{B.4}$$

Appendix C

P2D - Hyper-parameters benchmark

Symbol	# Layer Resnet	Pre-Trained	Pose Estimator	Epoch	Lr	Clipping Low	Clipping High	Polarimetry Only	Smoothness	Loss
♣	50	No	posecnn	100	e-4	0.4	-	-	2nd e-2	No Rw
♢	50	No	posecnn	100	e-4	0.4	-	-	2nd e-2	No Rw
♡	18	Yes	posecnn	40	e-3	0.4	-	-	1st e-2	No Rw
♠	18	Yes	posecnn	18	e-3	0.4	-	-	1st e-2	No Rw
□	18	Yes	posecnn	30	e-3	0.4	-	-	1st e-3	No Rw
▽	18	Yes	shared	40	e-3	0.4	-	-	1st e-3	No Rw
△	18	Yes	shared	22	e-4	0.4	-	-	1st e-3	No Rw
◆	18	Yes	shared	40	e-4	0.4	-	-	1st e-3	No Rw
■	50	Yes	posecnn	40	e-4	0.4	-	-	1st e-3	No Rw
◀	50	Yes	posecnn	40	e-4	0.4	-	-	1st e-3	Rw
▶	50	Yes	posecnn	14	e-4	0.8	-	Yes	1st e-3	Rw
◯	50	Yes	posecnn	40	e-4	0.6	-	Yes	1st e-3	Rw
⊛	50	Yes	posecnn	40	e-4	0.4	-	Yes	1st e-3	Rw
◎	50	Yes	posecnn	2	e-4	0.4	0.8	Yes	1st e-3	Rw
⊖	50	Yes	posecnn	30	e-4	0.4	0.8	Yes	1st e-3	Rw
†	50	Yes	posecnn	40	e-5	0.4	0.8	Yes	1st e-3	Rw
⊤	50	Yes	posecnn	40	e-5	0.4	0.8	Yes	1st e-3	Rw
P2D	50	No	posecnn	40	e-4	0.4	0.8	No	2nd 5e-3	Rw

TABLE C.1: Hyper-parameters of Networks. No Rw stands for taking into account either the reflected wave or calculating a $\pi/2$ uncertainty. Polarimetry only consider passing only polarimetric characteristics without intensity.

Appendix D

Quantitative Evaluation - Hyper-paremeters search

Symbol	Abs rel	Sqr Rel	RMSE	RMSE Log	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
♣	0.55588762	5.26748077	18.44713169	0.6439612	0.36374533	0.61940697	0.77606769
♢	3.3686126	167.8684821	28.42702103	1.32727985	0.19212632	0.38062578	0.53130208
♡	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
♠	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
□	1.0303843	24.87593764	18.81575993	0.75462168	0.34022256	0.60203593	0.77666854
▽	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
△	0.96461063	7.12470944	20.88262871	1.02670804	0.15175467	0.30496211	0.46155345
♦	0.75546718	5.84624123	20.74055243	0.93557832	0.18929863	0.36592266	0.54174955
■	0.51232816	4.93071377	19.61147069	0.70184211	0.35845561	0.61102479	0.75587928
▲	0.38287898	6.01501411	19.09932481	0.55880806	0.43437571	0.74255633	0.88075857
▼	0.76403479	5.94209196	20.91100688	0.97697274	0.18841593	0.35870664	0.52645032
◯	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
⊛	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
◎	-	-	-	-	-	-	-
⊜	0.76443061	5.9460011	20.91314594	0.977776	0.18799568	0.35775111	0.52548628
†	0.48103555	5.26684022	20.25824758	0.71060804	0.35065238	0.63112678	0.79329612
⊤	0.41630519	4.92056124	20.35477976	0.67482096	0.4011475	0.66849616	0.81677506
P2D	**0.3557942**	**3.85713148**	**17.65362369**	**0.40850429**	**0.48569957**	**0.77988741**	**0.88819119**

Table D.1: Network Results for Raw Type

Symbol	Abs rel	Sqr Rel	RMSE	RMSE Log	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
♣	0.32341093	5.70662946	21.04086645	0.50855998	0.48195047	0.76779837	0.89512581
♢	2.69068501	156.0314041	31.71171169	1.1919541	0.29342888	0.51116519	0.63882835
♡	0.58345059	6.94698943	23.85342949	0.87476191	0.29946574	0.52921924	0.68527827
♠	0.58345059	6.94698943	23.85342949	0.87476191	0.29946574	0.52921924	0.68527827
□	0.57482394	19.95128894	21.35614534	0.58254845	0.49430758	0.74343469	0.86550555
▽	0.58345059	6.94698943	23.85342949	0.87476191	0.29948158	0.52921924	0.68527827
△	0.73416397	8.82097972	23.7691066	0.89847201	0.30261997	0.52410106	0.67418577
♦	0.58770072	6.87938775	23.58981014	0.83125732	0.30261937	0.52983251	0.69194762
■	0.41320904	6.054892	22.36034069	0.6548841	0.4032127	0.68462796	0.83656994
▲	0.39060907	7.93909456	22.06848248	0.58231546	0.48535972	0.74034814	0.86560412
▼	0.5838244	6.94570913	23.84958742	0.87428762	0.2991667	0.52907665	0.68526876
◯	0.58345059	6.94698943	23.85342949	0.87476191	0.29946574	0.52921924	0.68527827
⊛	0.58345059	6.94698943	23.85342949	0.87476191	0.29946574	0.52921924	0.68527827
◎	-	-	-	-	-	-	-
⊜	0.58345059	6.94698943	23.85342949	0.87476191	0.29946574	0.52921924	0.68527827
†	0.4712141	6.56367222	23.52546505	0.79981165	0.31860031	0.57912212	0.74810095
⊤	0.40326542	6.01842483	23.51694278	0.76680708	0.36243486	0.6294479	0.78649455
P2D	**0.28683322**	**4.65054994**	**20.44182853**	**0.47520682**	**0.52469207**	**0.7947532**	**0.91179457**

Table D.2: Network Results for Cropped Type

Symbol	Abs rel	Sqr Rel	RMSE	RMSE Log	$\delta > 1.25$	$\delta > 1.25^2$	$\delta > 1.25^3$
♣	0.18989838	0.54244479	1.34885349	0.23153979	0.76824878	0.91486633	0.95153378
♢	2.87024556	118.38922198	16.32300697	1.62632891	0.16620638	0.2791327	0.37254277
♡	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
♠	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
□	0.16731199	0.27777386	1.0451614	0.23296831	0.71147468	0.89252642	0.9723272
▽	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
△	0.24659982	0.6315535	1.16625939	0.28989203	0.69650318	0.86563945	0.93488497
♦	0.2224306	0.45180898	1.15844555	0.28439088	0.70842411	0.86194362	0.93311142
■	0.21036596	0.43420804	1.0714632	0.27133958	0.71419844	0.86713089	0.95214101
▲	0.15759349	0.28979545	1.05014808	0.21613899	0.7511274	0.93634769	0.97401999
▼	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
◯	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
⊛	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
◎	-	-	-	-	-	-	-
⊜	0.22216299	0.45082331	1.16179587	0.28487822	0.69452175	0.86103551	0.92367196
†	0.31576105	0.97961451	1.88119104	0.47258022	0.55352902	0.76281218	0.85789662
⊤	0.20751088	0.40243636	1.29060914	0.31456687	0.67543749	0.86465791	0.92402199
P2D	**0.14623286**	**0.23108592**	**0.89743302**	**0.19357311**	**0.76006202**	**0.94151676**	**0.98344542**

TABLE D.3: Network Results for Specular Type