DEVELOPMENT OF A PROBABILISTIC PERCEPTION SYSTEM FOR CAMERA-LIDAR SENSOR FUSION

JOHAN SAMIR OBANDO CERON
2181732

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE DE AUTOMÁTICA Y ELECTRÓNICA
MAESTRIA EN INGENIERIA DE DESARROLLO DE PRODUCTOS
SANTIAGO DE CALI
2021

DEVELOPMENT OF A PROBABILISTIC PERCEPTION SYSTEM FOR CAMERA-LIDAR SENSOR FUSION

JOHAN SAMIR OBANDO CERON

MAGISTER EN INGENIERIA DE DESARROLLO DE PRODUCTOS
(MASTER OF SCIENCE IN PRODUCT DEVELOPMENT ENGINEERING)
Modalidad Proyecto de grado

Director
VICTOR ROMERO CANO
Ph.D. in Field Robotics
School of Aerospace, Mechanical and Mechatronic Engineering, University of Sydney.

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE
FACULTAD DE INGENIERÍA
DEPARTAMENTO DEAUTOMÁTICA Y ELECTRÓNICA
MAESTRIA EN INGENIERIA DE DESARROLLO DE PRODUCTOS
SANTIAGO DE CALI
2021

Nota de aceptación:

Aprobado por el Comité de Grado en cumplimiento de los requisitos exigidos por la Universidad Autónoma de Occi-dente para optar al título de Magister en Ingeniería de Desarrollo de Productos.

Juan Carlos Perafan Villota

Jurado

Alejandro Parada-Mayorga

Jurado

Santiago de Cali, Jun 02, 2021

ACKNOWLEDGMENTS

CONTENIDO

pág.

LISTA DE FIGURAS

LISTA DE TABLAS

pág.

ABSTRACT

Multi-modal depth estimation is one of the key challenges for endowing autonomous machines with robust robotic perception capabilities. There has been an outstanding advance in the development of uni-modal depth estimation techniques based on either monocular cameras, because of their rich resolution or LiDAR sensors due to the precise geometric data they provide. However, each of them suffers from some inherent drawbacks like high sensitivity to changes in illumination conditions in the case of cameras and limited resolution for the LiDARs. Sensor fusion can be used to combine the merits and compensate the downsides of these two kinds of sensors. Nevertheless, current fusion methods work at a high level. They processes sensor data streams independently and combine the high level estimates obtained for each sensor. In this thesis, I tackle the problem at a low level, fusing the raw sensor streams, thus obtaining depth estimates which are both dense and precise, and can be used as a unified multi-modal data source for higher level estimation problems.

This work proposes a Conditional Random Field (CRF) model with multiple geometry and appearance potentials that seamlessly represents the problem of estimating dense depth maps from camera and LiDAR data. The model can be optimized efficiently using the Conjugate Gradient Squared (CGS) algorithm. The proposed method was evaluated and compared with the state-of-the-art using the commonly used KITTI benchmark dataset. In addition, the model is qualitatively evaluated using data acquired by the author of this work.

Palabras clave:

Probabilistic Graphical Model (PGM), Conditional Random Field (CRFs), Sensor Fusion, LiDAR, Monocular Camera, Perception and Robot Operating System (ROS).

RESUMEN

La estimación de profundidad usando diferentes sensores es uno de los desafíos clave para dotar a las máquinas autónomas de sólidas capacidades de percepción robótica. Ha habido un avance sobresaliente en el desarrollo de técnicas de estimación de profundidad unimodales basadas en cámaras monoculares, debido a su alta resolución o sensores LiDAR, debido a los datos geométricos precisos que proporcionan. Sin embargo, cada uno de ellos presenta inconvenientes inherentes, como la alta sensibilidad a los cambios en las condiciones de iluminación en el caso de las cámaras y la resolución limitada de los sensores LiDAR. La fusión de sensores se puede utilizar para combinar los méritos y compensar las desventajas de estos dos tipos de sensores. Sin embargo, los métodos de fusión actuales funcionan a un alto nivel. Procesan los flujos de datos de los sensores de forma independiente y combinan las estimaciones de alto nivel obtenidas para cada sensor. En este projecto, abordamos el problema en un nivel bajo, fusionando los flujos de sensores sin procesar, obteniendo así estimaciones de profundidad que son densas y precisas, y pueden usarse como una fuente de datos multimodal unificada para problemas de estimación de nivel superior.

Este trabajo propone un modelo de campo aleatorio condicional (CRF) con múltiples potenciales de geometría y apariencia que representa a la perfección el problema de estimar mapas de profundidad densos a partir de datos de cámara y LiDAR. El modelo se puede optimizar de manera eficiente utilizando el algoritmo Conjugate Gradient Squared (CGS). El método propuesto se evalua y compara utilizando el conjunto de datos proporcionado por KITTI Datset. Adicionalmente, se evalua cualitativamente el modelo, usando datos adquiridos por el autor de esté trabajo.

INTRODUCTION

More than 50 years have passed since the first industrial robot began service in a car assembly line [1]. Since that time, robotics has been studied extensively and improvements in the field have created opportunities for new robotics applications. Approximately twenty years ago, Autonomous Vehicles (AV) and Driver Assistance Systems (DAS) have gained remarkable attention. It is undeniable that these technologies could drastically reduce the number of car accidents and fatalities. These technologies are ranked among the most transformative public-health initiatives in human history.

Since DARPA Grand Challenge (2004, 2005) and Urban Challenge (2007), many researchers have become interested in this area, as well as huge companies in tech and the auto industry. Waymo, Tesla Motors, Volkswagen Group, Mercedes Benz, GM Cruise, Ford, and NVIDIA are among the growing list of companies who are currently testing their autonomous vehicles on California roads, according to reports submitted to the Department of Motor Vehicles (DMV). Fully autonomous vehicles, that is, being in the automation levels 4-5, have a much more complicated mission and yet not fully available. They must be extremely reliable, safe, and robust to unseen phenomenon given their highly sensitive role. A fully autonomous vehicle must execute all operations, monitor itself, and be able to handle all unprecedented events and conditions, like roads without markings, unexpected objects and debris on the road, unseen environments, adverse weather, etc. There are myriads of examples of complicated scenarios where robots do not work very well. This problem calls for better sensors, better processing power, and better algorithms before fully autonomous robots start working close to humans.

One of the ongoing debates among driver-less car designers is about the choice of exteroceptive sensors, their mounting position and the high cost they have. Two of the most used sensors are the Light Detection and and Ranging (LiDAR) and the monocular camera. The LiDAR is a remote sensing device that uses a set of rotating laser to measure distances. Pulses of light are emitted from a laser scanner, and when the pulse hits a target, a portion of its photons are reflected back to the scanner. Due to the location of the scanner, the directionality of the pulse, and the time between pulse emission and return are known, the 3D location (XYZ coordinates) from which the pulse reflected is calculated [2].

Mass-producing LiDAR sensors has been challenging, mainly due to its relatively high price. Some companies have pushed the idea that similar to humans, cars can perceive and navigate using just the eyes. It is to say that they want to use just cameras in their robotic platforms to perceive the environment. Although multiple view reconstruction provides an attractive alternative due to a near instantaneous gathering of dense 3D data, leading to dense scene reconstructions from image data alone [3][4], unfortunately, stereo reconstruction fidelity is limited in range by the baseline and the image resolution. This seriously impedes accurate reconstruction beyond a few meters from the camera. Hence, this idea has not been so efficient.

LiDAR technology is growing with remarkable investments, and sensor prices are dramatically dropping. Thus, most companies are planning to actually make use of multiple LiDARs. However, the precision of operation of an autonomous vehicle is, thus, limited by the reliability of the associated sensors. Each type of sensor has its own limitations, for example, LiDAR sensor readings are often affected by weather phenomena such as rain, fog or snow [7]. Multimodal sensing is necessary, because a single modality cannot usually capture complete knowledge of a rich natural phenomena.

Most autonomous car manufacturers are designing multi-module systems to model the environment and make decisions. Currently, they are working on platforms that use different sensors modalities. Most of the research on sensor fusion has been devoted to studying how to combine information given by camera and LiDAR sensor in order to get a better environment representation. Developing robust perception systems is one of the most important research endeavors in the robotics field. Perception is understood as the task of extracting semantic information from sensory data that can come from multiple sources, such as cameras, laser sensors, radars, etc. Limited perception capabilities and underdeveloped processing techniques using different sensor modalities are a common problem in intelligent systems. These problems do not allow creating a good 3D reconstruction under challenging environmental conditions. 3D reconstruction is important to perceive the world, and it is an imperative prerequisite for autonomous navigation.

Applications in which it is necessary to perceive the environment automatically go beyond the urban scenario. Perception systems are very important for getting information that is used in the description of objects found in outdoor scenarios, such as trees or plants, fruits to be harvested, among others. The work iN [8] presents a solution to the detection problem of apple trees using data from a LiDAR sensor. However, jobs that extend this type of technology using different sensors for unstructured environments such as forests are extremely rare.

As it was mentioned before, data about an environment can be obtained from different types of sensors. Acquiring sensing data using heterogeneous acquisition mechanisms is referred to as multimodal sensing [9]. Data fusion is the process by which multimodal data streams are jointly analyzed to capture knowledge of a certain environment. Lahat et al., identifies several challenges that are imposed by multimodal data. These challenges can be broadly categorized in two segments:

1. Challenges at acquisition level and
2. Challenges due to uncertainty in the data sources.

Challenges due to problems at the data acquisition level include differences in physical units of measurement, differences in sampling resolutions, and differences in spatio-temporal alignment. The uncertainty in data sources also pose challenges that include noise such as calibration errors, quantization errors or precision losses, differences in reliability of data sources, inconsistent data and missing values [10].

It is important thus to develop robust perception systems that are able to combine the strengths of different sensors in order to obtain more reliable and meaningful data in the face of environmental limitations, while still overcoming physical and technological limitations . Hence, the primary contribution of this project will be the development of a deep regression model for fusing sparse LiDAR and dense monocular image data. The problem will be modelled as a conditional random field (CRF) that takes both a sparse set of depth samples and RGB images as input, and infers a range value for every pixel in the image. These kinds of approaches allow for dramatic reductions in the pre-processing's complexity of subsequent tasks such as object recognition, instance segmentation, among others. Additionally, these tasks can be more robustly executed due to the increased quality of the input data.

1. PROBLEM STATEMENT

The intelligent management of agricultural areas is set to increase further over the coming years. Autonomous vehicles will be an increasingly common sight on farm fields. This kind of vehicles have become more common in agriculture, and many of the challenges identified in other scenes such as in the cities still persist.

For any autonomous vehicle whether to be used on the streets or on farms, the perception system is one of the primary modules. Nowadays there are a lot of challenges related to the perception system and it is worth delving into them. One of these challenges is developing robust and inexpensive multi-sensory perception systems.

A practical problem is described later, without ignoring the fact that this sensor fusion system might be used in general autonomous cars applications. It is important to highlight that the development of a low-cost and robust sensory fusion system is the core of this thesis. It is important to clarify why and where a multi-sensory perception system can contribute significantly. The importance of a perception system in self-driving cars was previously discussed, but it is not the only practical example.

Now the practical problem I would like to discuss is in the agro-industrial area since the sensory fusion system is very pertinent to tackle several tasks in this field as well.

One of the most basic needs to guide the definition of urban, agro-industrial and territorial management policies is to have a digital topographic representation or map of cities, crops and forests. These maps should ideally be created from multiple sensors whose responses are complementary (color information, for example, complements the returns of a LiDAR sensor in the presence of rain or low reflective objects). Once a topographic representation has been constructed, it can be used to produce and geo-localize higher-level estimates (e.g., location and classification of different trees and plants, crop density, location, and types of pests).

Data can be collected using both aerial and terrestrial unmanned vehicles equipped with hyper-spectral cameras, stereo cameras and LiDAR (Light Detection And Ranging) sensors. The processing of the acquired data can be used to generate a digital forest model (DFM), where each tree is a single object in a geo-spatial database that provides not only a greater knowledge about the forest (or crop) composition but also macroeconomic aspects like biomass volume, spectral vegetation indexes and growth rate.

Thus, DFM will support forest planners in making multi-criteria decisions (MCDA) when planning harvesting operations in the case of crops or coordinating preservation policies (among others) in the case of forests, while taking into account all possible infrastructural and geomorphological constraints. However creating a DFM, or the map of a city, require a highly accurate and dense point cloud of the environment at hand.

Motivated for building 3D reconstructions from which representations of different vegetation features of an environment can be obtained with high quality and precision. A robust perception system is proposed. It is known that cameras provide near instantaneous capture of the workspace's appearance such as texture and color, but from a single view, little geometrical information. On the other hand, laser readings may be so sparse that significant information about the surface is missing.

The considerations above motivate the formulation of this work's research question: How to develop a perception system for fusing a laser scan with a RGB image in order to produce a higher-resolution range?

2. OBJECTIVES

2.1 GENERAL OBJECTIVE

Develop a three-dimensional dense reconstruction system that allows characterizing external and internal environments in adverse environmental conditions, at a low cost, and from the data provided by LiDAR and camera sensors.

2.2 SPECIFIC OBJECTIVES

- Develop a library that allows capturing and visualizing the information of the laser and camera sensors.
- Develop the architecture of a fusion system that allows reconstructing 3D environments from two sensory modalities such as camera and laser.
- Implement the fusion system.
- Evaluate the performance of the 3D reconstruction system.

3. JUSTIFICATION

This thesis is mainly concerned with how a robot extracts useful information for 3D reconstruction, robust automatic navigation or other applications from dissimilar sensors, and how this information can be related to each other for getting better representations of the environment. In static environments, the registration of data is fairly easy. However, in real-world applications, the environment is generally dynamic, and the data captured on different days can appear significantly different.

Drastic changes in environmental appearance due to changing seasons, lighting conditions, and dynamical objects make the task of visual perception extremely difficult. The image of a location captured on a bright summer day can appear significantly different from the image of the same location captured on a gray snow-covered winter day [11]. Therefore, if we use the image data alone to localize on a snowy winter day within an a priori map collected on a sunny summer day, then it might not be possible. However, if we use data from LiDAR sensor as well, then we can boost the registration process by using complementary information provided by different sensing modalities.

On the contrary, using just a LiDAR sensor might be seen as a good alternative, as it collects data in an easy way with high accuracy. Nevertheless, it is ineffective during heavy rain or low hanging clouds; it is also sparse and the sensor, extremely expensive. Hence, developing a technique for low-level data fusion between laser and monocular camera to perceive the environment can ensure that the estimated information about the environment is both accurate and dense in spite of the environmental conditions. The perception system to develop should take into consideration also the most important data imperfections, such as incomplete data, a complex background with different colors and textures, some obscure objects, different illumination conditions, and the blurriness associated with the camera's low resolution.

4. RELATED TO WORK

4.1 IMPORTANCE OF 3D RECONSTRUCTION

Having highly detailed 3D models of natural phenomena, urban and rural scenes, industrial sites, among others, opens new horizons for applications. In the domain of autonomous driving, an accurate 3D model could be used by vehicles to navigate themselves or to reason about the scene. On the other hand, the objects reconstruction technology has a very important role in archeology and ancient history; these disciplines often require simple, robust and cheap methods for scanning objects, so these can be studied.

Image-based modeling methods which aim to provide fast and accurate have gained significant attention due to the advances in camera technology and the flexible and economic data acquisition possibilities they provide [14].

The majority of these applications do not require any color information within the 3D model. Whereas extending these models with color information can be an added value to the application. For example, during inspection of industrial companies a photo-realistic point cloud is used to detect damaged sections.

Today, most mapping systems integrate multiple sensors in order to make a colored and detailed 3D model. This results in the need for data fusion at sensor-level, for example 2D-3D registration between LiDAR sensors and cameras.

4.2 PERCEPTION

Every robotic system can be seen as composed by three modules: sensing, planning and acting. In the majority of the cases, the robot will be interacting with the real environment. Machines just have access to this world through the measures provided by sensors. Interpreting the measured values in order to make decisions about their future actions is what perception is all about. Without perception, machines would not be able to make decisions, and fill their purpose [15].

The perception system converts the sensors raw data into consistent and useful information. The quality and richness of the obtained information will have a direct effect over the performances of the control and planning modules.

One robotics application that illustrates the main features of a state-of-the-art perception system is SLAMMOT: Simultaneous Localization, Mapping and Moving Object Tracking. It allows a robot to construct a consistent map of the visited places, localize itself in such maps and track moving objects in its sensors' coverage region. This feature serves as the basis for scene understanding, which is a key prerequisite for making a robot truly autonomous [16].

4.3 MEASURING THE SURROUNDINGS

Sensors are the basic element of perception. They transform physical signals into signals that are understandable by the machine. The choice of the sensors for a particular task is non trivial; criteria for their selection are cost, precision, range of measurements, energy consumption and effect over the environment. These criteria make them suitable for particular tasks under specific conditions. Hence, it is very important to determine the characteristics of sensors to use. In our case, our sensors are camera and LiDAR, and their corresponding strengths and weaknesses will be presented.

4.3.1 Camera

Most driving functionalities rely heavily on receiving and processing signals in the visible light spectrum. All mobile robots benefit from a camera-based setup, e.g. monocular vision and stereo vision [18] [19]. Inexpensive cost, color perception, high resolution (as opposed to conventional LiDAR) and the potential of obtaining rich semantic information are among their advantages. They are by far the most employed sensor for detecting traffic signs, pedestrians, vehicles and road markings. 3D localization and tracking, however, require depth information that would only be available with a multi-camera system. Cameras are, in general, sensitive to lighting and weather conditions.

4.3.2 LiDAR

Light Detection and Ranging (LiDAR) is a light-based ranging system which transmits 600 1000nm invisible laser. Reflections from the environment are sensed using a photodetector, and range values are estimated based on time-of-flight or phase-shift. The result is a semi-dense 3D point cloud of the environment. Since LiDARs use invisible light, they do not interfere with ambient light and work equally well under different lighting conditions. LiDAR data can be used for directly estimating the presence of obstacles along with their position [20] [21].

LiDARs provide accurate and direct structure measurements on scene geometry. Unlike stereo systems, these measurements come with no post-processing cost and do not rely on feature-matching algorithms. However, they have a maximum sensing range of 70 to 100m, relatively low refresh rates, and sensing problems in adverse weather such as rain, snow, fog, or dust. One other issue with LiDAR is presence of dark or low-reflective obstacles.

There has been remarkable effort from the industry side over the past couple of years to improve LiDAR technology and make it more affordable. This is in part due to the demand for more reliable active sensors from the auto industry.

4.4 MULTI SENSOR DATA

To increase the robustness and improve the overall estimation capability of a sensory system, a lot of research effort has recently been devoted for combining data from complementary sensors, also known as sensor fusion. It has become a necessary approach to overcome the disadvantages of single-sensor perception architectures providing a richer description of a dynamic environment.

Sensors are designed to provide specific data extracted from the environment. For instance, LiDAR provides depth measurements that are useful for estimating features such as the position and shape (lines) of obstacles within its field of view. On the contrary, cameras provide visual characteristics that can be used to infer information about the appearance of obstacles. Intelligently combining these features from sensors may give a complete view of the objects and the environment around the intelligent vehicle and it is the objective of a fusion architecture [22].

Fig. 1 illustrates the cover of area, field of view, and typical operating ranges, for both a human-driven vehicle as well as a hypothetical autonomous vehicle.

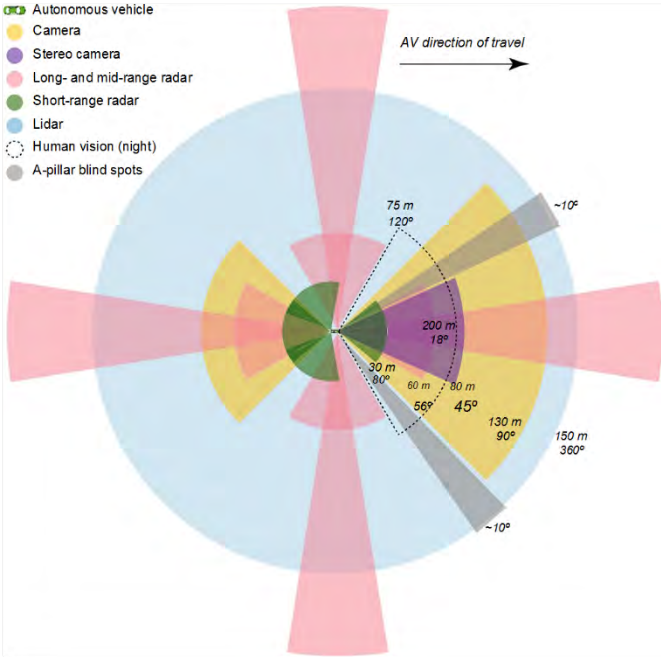

***Fig.* 1**. Example illustration of the various sensors, with reasonable estimates of coverage area. [119]

Traditionally, sensor fusion methods are defined in three different levels, according to the stage at which the sensor fusion is performed. Low-level fusion utilizes the raw sensor data. High-level fusion integrates high-level estimates obtained from each sensor modalities individually. Between low and high level fusion, there is another method called hybrid fusion or feature level fusion. This project will focus on low-level fusion.

4.5 POINT CLOUD REGISTRATION

Registration is the process of aligning several shapes (two or more) in a common coordinate system. It is generally applied to overlapping pairs of 2D images or 3D point cloud models. Point cloud alignment is a fundamental problem for many robotics and computer vision applications. Typical registration tasks generally require 3D point cloud alignment.

The mathematical mapping is expressed by the transformation relationship (e.g., scale, rotation, translation and shape deformation) between the coordinate systems of the two datasets. To assemble point clouds into more comprehensive ones requires aligning them along distinctive shapes that are common to the scans. Aligning point clouds to find their relative motion is a critical component for any mobile platform that uses a LIDAR sensor to navigate.

The research community has developed different techniques to deal with the point cloud registration problem [23] [24]. With the possibility to register multiple scans into more comprehensive 3D models, it is no surprise that part of the research community has focused on developing and using registration techniques to map urban indoor and outdoor spaces in 3D [25] [26]. An example of registering one point cloud to one another such that a comprehensive 3D map is generated is illustrated in Fig. 2.

Modern depth cameras commonly produce pairs of depth and color images. Many industrial 3D scanners are also equipped with synchronized color cameras and provide software that associates color information with the 3D scans. Multi-view stereo pipelines reconstruct colored point clouds from image collections. Considering color along with the geometry can increase the accuracy of point cloud registration [27] [28].

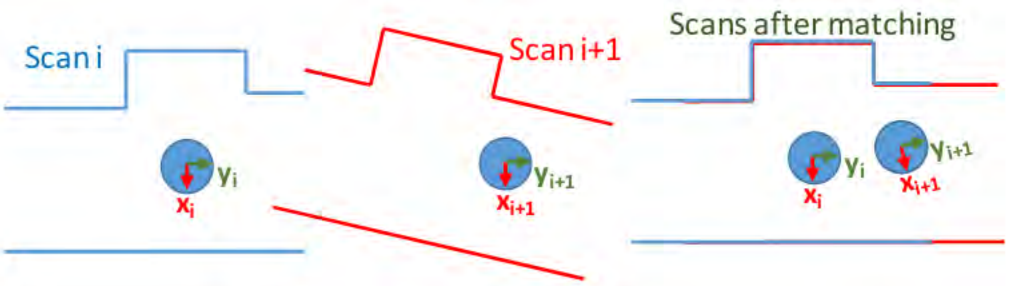

***Fig.* 2**. 3D mapping with LiDAR, registering multiple scans (shown here for 2D scans from atop view) [120].

4.6 SIMPLE LINEAR ITERATIVE CLUSTERING

With the advent of self-driving cars, there has been in the robotics community an increasing interest in the development of robust perception systems that provide correct estimates even under different and challenging environmental conditions. Although some approaches to robust perception resort to statistical methods for dealing with data outliers [29], the work presented in this paper belongs to the group that tackles the robust-perception problem by leveraging the complementary nature of pasive and active sensor modalities.

Multi-sensor approaches to robotic perception, can be categorised according to the level at which the data from each sensing modality is fused in order to obtain the estimate of interest. According to [30], data fusion can be made at the level of symbolic estimates or high level fusion, at the level of features or medium level fusion, or at the level of raw data or low level fusion.

Low level fusion methods on the other hand explore the complementary relationships between passive and active sensors at the pixel level. The approaches in [31],[32],[33] follow this intuition but require the fused modalities to have similar coverage densities. Additionally, low level fusion fosters the development of recognition approaches that use an improved and unified version the multi-modal information in the object recognition task [34]. The proposed framework provides a procedure for fusing lidar and image data independently of the lidar data's density. In previous papers I show how this low level fusion (at the pixel or superpixel level) improves the object recognition task in indoor environments [35].

In order to build the framework proposed in this project, range measurements are first projected on the image space. These sparse depth measurements are locally extended using Simple Linear Iterative Clustering (SLIC) [36]. SLIC is a simple and parallelisable method, based on k-means clustering, for decomposing an image into a regular grid of visually homogeneous regions or so-called super-pixels. As a result, SLIC super-pixels provide a regular grouping of image pixels according to their distance both spatially and in the colour space. I use a python library, scikit-image [37], to generate the super-pixel segmentation in order to assign depth values to all of the pixels within super-pixels with at least one range measurement. The image segmented and the boundaries are shown in Fig. 3 and Fig. 4.

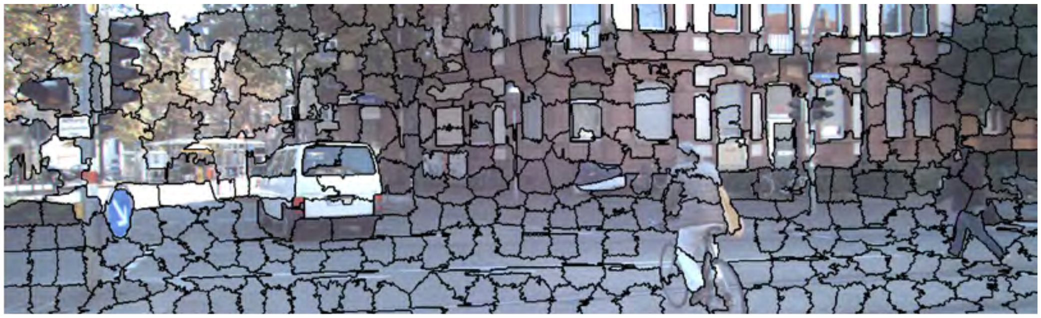

***Fig.* 3**. Image segmented using SLIC into superpixels.

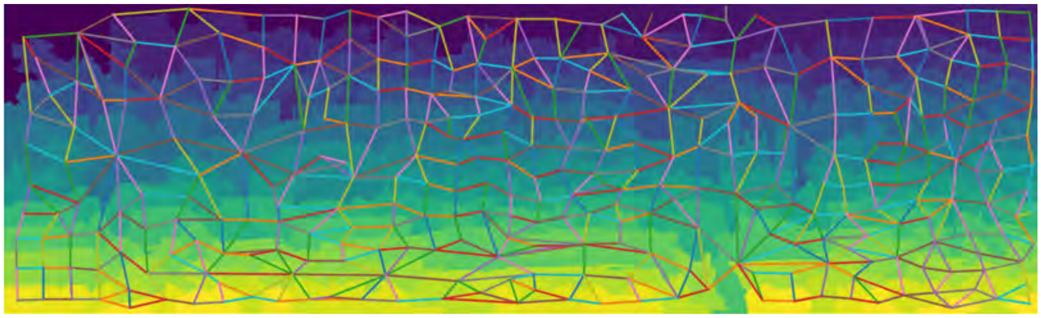

***Fig.* 4**. Superpixels: boundary neighbors and centroids.

4.7 DATA PROJECTION

In this section, we give a brief introduction to the alignment of image and LIDAR point cloud. As presented in [38], the Velodyne HDL-64E LIDAR and a camera are mounted on the roof of a vehicle and they are synchronized by a hardware trigger. Once the rolling LIDAR is facing forward, the camera gets triggered. The camera and LIDAR are cross-calibrated so that the point cloud can be aligned with the image by projecting the LIDAR points onto the image plane. The projection of a 3D point $\mathbf{x} = (x, y, z, 1)^T$ in rectified (rotated) camera coordinates to a point $\mathbf{y} = (u, v, 1)^T$ in the i'th camera image is given as

$$\mathbf{y} = \mathbf{P}^{(i)}_{\text{rect}}\mathbf{x}$$

with

$$\mathbf{P}^{(i)}_{rect} = \begin{pmatrix} f^{(i)}_u & 0 & c^{(i)}_u & -f^{(i)}_u b^{(i)}_x \\ 0 & f^{(i)}_v & c^{(i)}_v & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

the i’th projection matrix. Here, $b^{(i)}_x$ denotes the baseline (in meters) with respect to reference camera. Note that in order to project a 3D point $\mathbf{x}$ in reference camera coordinates to a point y on the i’th image plane, the rectifying rotation matrix of the reference camera $\mathbf{R}^{(0)}_{rect}$ must be considered as well:

$$\mathbf{y} = \mathbf{P}^{(i)}_{rect}\mathbf{R}^{(0)}_{rect}\mathbf{x}$$

Here, $\mathbf{R}^{(0)}_{rect}$ has been expanded into a 44 matrix by appending a fourth zero-row and column, and setting $\mathbf{R}^{(0)}_{rect}(4, 4) = 1$. We have registered the Velodyne laser scanner with respect to the reference camera coordinate system. The rigid body transformation from Velodyne coordinates to camera coordinates is given by,

$$\mathbf{T}^{cam}_{velo} = \begin{pmatrix} \mathbf{R}^{\text{cam}}_{velo} & \mathbf{t}^{cam}_{velo} \\ 0 & 1 \end{pmatrix}$$

a 3D point x in Velodyne coordinates gets projected to a point y in the i’th camera image as

$$\mathbf{y} = \mathbf{P}^{(i)}_{rect}\mathbf{R}^{(0)}_{rect}\mathbf{T}^{cam}_{velo}\mathbf{x}$$

Subsequently, as a preprocessing step, points with a negative Z value are removed. Then the remaining points can be projected onto the image plane with the projection matrix given by:

$$[x'y'z']^T = \mathbf{y}\,[x_p\ y_p\ z_p\ 1]^T$$

The projected pixel coordinates of the LIDAR points can be obtained by:

$$[x, y] = \left[\frac{x'}{z'}, \frac{y'}{z'}\right]$$

Once the 3D points are projected on a 2D plane corresponding to that of the camera, the projected points are filtered. This filter is executed to use only 3D information that is within the dimensions of the image. Therefore, information that is outside the range of the camera will not be considered for the depth estimation.

In Fig. 5 and Fig. 6 the raw point cloud projected on the image, and the average of depth measurements assigned for each superpixel.

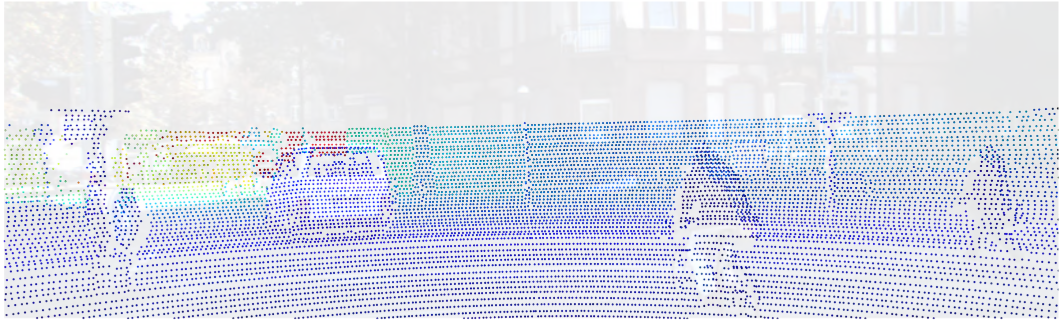

***Fig.* 5**. Sparse raw depth measurement projected onto the image.

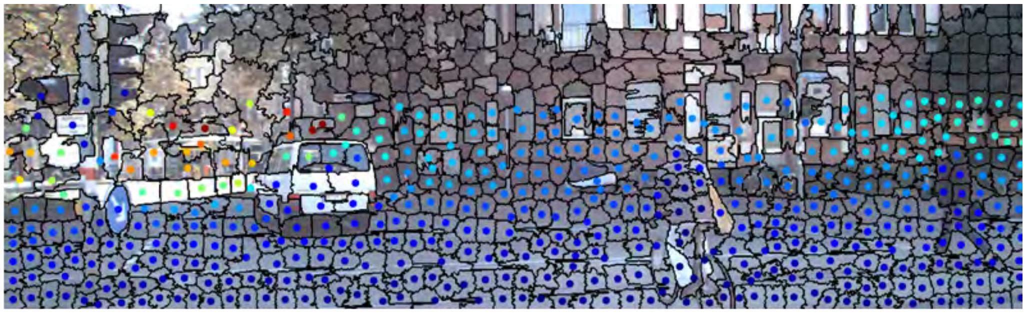

***Fig.* 6**. Depth assignment: Each pixel inside the superpixels has the same depth value.

Fig. 6 shows the depth assignment for each superpixel's centroid. This depth value considers the information of all the 3D points that are projected within each single superpixel. In this way, to determine the depth value of the centroid, an average is calculated between all the depth values which are inside of the superpixel's segment. This value changes in each scene (frame) and depends on the number of superpixels selected and the resolution of the lidar sensor used.

4.8 THE ROBOT OPERATING SYSTEM

The Robot Operating System (ROS) is a flexible framework for writing robot software. It is a collection of tools, libraries, and conventions that aim to simplify the task of creating complex and robust robot behavior across a wide variety of robotic platforms [39], [40].

ROS processes are represented as nodes in a graph structure, connected by edges called topics. ROS nodes can pass messages to one another through topics, make service calls to other nodes or provide a service for other nodes. ROS's core functionality allows developers to visualize and record data, easily navigate the ROS package structures, and create scripts automating complex configuration and setup processes [42].

4.9 MATHEMATICAL BACKGROUND

4.9.1 Elements of probability

This section presents a review of some basic concepts of probability theory, including random variables, joint probability distribution, bayes rule and probabilistic graphical models. The definitions introduced in this section are instrumental to the formualtion of the sensor fusion model presented in further sections.

4.9.1.1 Random Variables

A random variable is a numerical description of the outcome of a random phenomenon. A random variable that might assume only a finite number or an infinite sequence of values is said to be discrete. One the contrary, a random variable that assume any value in some interval on the real number line is said to be continuous.

For instance, a random variable representing the number of automobiles sold at a particular dealership on one day would be discrete, while a random variable representing the weight of a person in kilograms (or pounds) would be continuous [43], [44].

For example, suppose we have a random variable Grade that reports the final grade of a student, then the representation of the statement is $P(Grade = A)$. We usually use uppercase letters X, Y, Z to denote random variables. In discussing generic random variables, we often use a lowercase letter to refer to a value of a random variable.

4.9.1.2 Joint probability distribution

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. For a discrete random variable, X, the probability distribution is defined by a probability mass function, denoted by $f(X)$. This function provides the probability for each value of the random variable. The joint probability mass function is the function:

$$f_{XY}(x, y) = P(X = x, Y = y)$$

In the development of the probability function for a discrete random variable, two conditions must be satisfied:

(1) $f(x)$ must be nonnegative for each value of the random variable,

(2) the sum of the probabilities for each value of the random variable must equal one.

A continuous random variable may assume any value in an interval on the real number line or in a collection of intervals. Since there is an infinite number of values in any interval, it is not meaningful to talk about the probability that the random variable will take on a specific value; instead, the probability that a continuous random variable will lie within a given interval is considered.

In the continuous case, the counterpart of the probability mass function is the probability density function, also denoted by $f(x)$ [47]. For a continuous random variable, the probability density function provides the height or value of the function at any particular value of x; it does not directly give the probability of the random variable taking on a specific value. However, the area under the graph of $f(x)$ corresponding to some interval, obtained by computing the integral of $f(x)$ over that interval, provides the probability that the variable will take on a value within that interval. The joint probability density function [47] for the continuous random variable in any region R of 2-D space is:

$$P((X, Y) \in R) = \iint_R f_{XY}(x, y) dx dy$$

A probability density function must satisfy two requirements:

(1) $f(x)$ must be nonnegative for each value of the random variable

(2) the integral over all values of the random variable must equal one.

4.9.1.3 Marginal distributions

Often when confronted with the joint probability of two random variables, we wish to restrict our attention to the value of just one or the other. I can calculate the probability distribution of each variable separately in a straightforward way, if I simply remember how to interpret probability functions.

These separated probability distributions are called the marginal distributions of the respective individual random variables. This contrasts with a conditional distribution, which gives the probabilities contingent upon the values of the other variables [48]. Marginalisation refers to the process of 'removing' the influence of one or more events from a probability.

If X and Y are discrete random variables and $f(x, y)$ is the value of their joint probability distribution at (x, y), the marginal distribution functions ?? ?? are given by:

$$f_X(x) = \sum_y f_{XY}(x, y)$$

$$f_Y(y) = \sum_x f_{XY}(x, y)$$

Here, we have 'removed' either x or y. Given two continuous random variables X and Y whose joint distribution is known, then marginal probability density function can be obtained by integrating the joint probability distribution over Y, and vice versa. The marginal distributions [49] for two continuous random variables are given by:

$$f_X(x) = \int_y f_{XY}(x, y)dy$$
$$f_Y(y) = \int_x f_{XY}(x, y)dx$$

where:

$$x \in [a, b], \text{ and } y \in [c, d]$$

4.9.1.4 Conditional probability

Sometimes the computation of the probability of an event is changed by the knowledge that a related event has occurred or by some additional conditions established on the experiment. This new probability is declared as a conditional probability, because we have some prior information about conditions under which the experiment is going to be performed [50].

Conditional probability answers the question, how does the probability of an event change if we have extra information? The conditional probability $P(E|F)$ is the probability that E happens, given that F has happened. F is the new sample space. This conditional probability [51] is written as follow:

$$P(E|F) = \frac{P(EF)}{P(F)}$$

We can visualize conditional probability as follows. Think of $P(A)$ as the proportion of the area of the whole sample space taken up by A. For $P(A|B)$ we restrict our attention to B. That is, $P(A|B)$ is the proportion of area of B taken up by A. It is shown in the following Fig. 7.

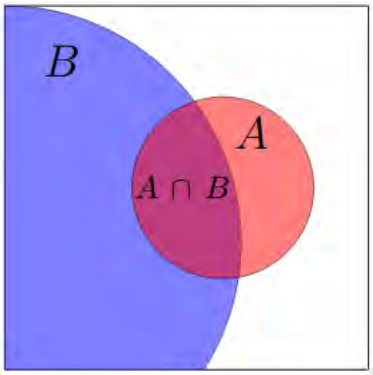

***Fig.* 7**. Conditional probability [x].

4.9.1.5 Chain Rule

This format is particularly useful in situations when I know the conditional probability, but I am interested in the probability of the intersection. Sometimes I have conditional distributions but want the joint distribution.

Intuitively it states that the probability of observing events E and F is the probability of observing F, multiplied by the probability of observing E, given that you have observed F. From the definition of the conditional distribution, I immediately see that:

$$P(E, F) = P(E|F)P(F)$$

More generally, if $E1, E2..$ are events, then I can write:

$$P\left(E_1, E_2, \ldots, E_n\right) = P\left(E_1\right) P\left(E_2|E_1\right) \ldots P\left(E_n|E_1 E_2 \ldots E_{n-1}\right)$$

Thus, can always write any joint distribution as an incremental product of conditional distributions:

$$P\left(E_1, E_2, \ldots x_n\right) = \prod_i P\left(E_i|E_1 \ldots x_{E-1}\right)$$

In other words, I can express the probability of a combination of several events in terms of the probability of the first, the probability of the second given the first, and so on [52]. It is important to notice that I can expand this expression using any order of events.

4.9.1.6 Bayes Rule

The concept of conditional probability was introduced previously. We noted that the conditional probability of an event is a probability obtained with the additional information that some other event has already occurred.

The conditional probability of B given A can be found by assuming that event A has occurred and, working under that assumption, calculating the probability that event B will occur [53]. In this section we extend the discussion of conditional probability to include applications of Bayes' theorem or Bayes' rule, which we use for revising a probability value based on additional information that is later obtained.

One key to understanding the essence of Bayes' theorem is to recognize that we are dealing with sequential events, whereby new additional information is obtained for a subsequent event, and that new information is used to revise the probability of the initial event. In this context, the terms prior probability and posterior probability are commonly used [54].

A prior probability is an initial probability value originally obtained before any additional information is obtained. On the other hand, a posterior probability is a probability value that has been revised by using additional information that is later obtained [42].

Bayes' theorem is a result in probability theory that relates conditional probabilities. The two conditional probabilities $P(A|B)$ and $P(B|A)$ are in general different.

Bayes theorem gives a relation between $P(A|B)$ and $P(B|A)$. An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.

The power of Bayes' rule is that in many situations where we want to compute $P(A|B)$ it turns out that it is difficult to do so directly, yet we might have direct information about $P(B|A)$. Bayes' rule enables us to compute $P(A|B)$ in terms of $P(B|A)$.

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

It is common to think of Bayes rule in terms of updating our belief about a hypothesis A in the light of new evidence B. Specifically, our posterior belief $P(A|B)$ is calculated by multiplying our prior belief $P(A)$ by the likelihood $P(B|A)$ that B will occur if A is true.

4.9.2 Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are a rich framework for encoding probability distributions over complex domains, joint (multivariate) distributions over large numbers of random variables that interact with each other [56], [57].

Two branches of graphical representations of distributions are commonly used. The first one is Bayesian networks also known as directed graphical models, in which the links of the graphs have a particular directionality indicated by arrows. And the other main class of graphical models are Markov Random Fields, also known as undirected graphical models, in which the links do not carry arrows and have no directional significance [58] as is shown in Fig. 8 .

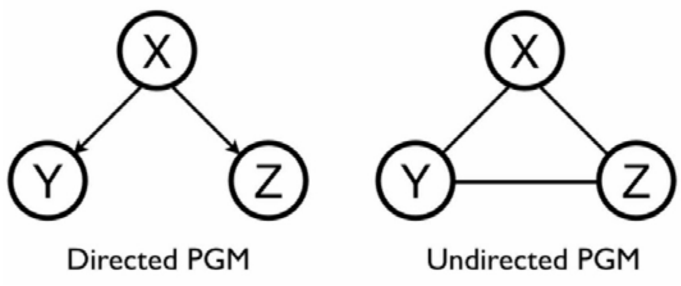

***Fig.* 8**. Probabilistic Graphical Models: a directed PGM (left, a.k.a. Bayesian Network) and an undirected PGM (right, a.k.a. Markov Random Field).

Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are better suited to expressing soft constraints between random variables. Both families encompass the properties of factorization and independences, but they differ in the set of independences they can encode and the factorization of the distribution that they induce.

A graph comprises nodes (also called vertices) connected by links (also known as edges or arcs). In a probabilistic graphical model, each node represents a random variable (or group of random variables), and the links express probabilistic relationships between these variables [59].

The graph then captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors each depending only on a subset of the variables.

4.9.2.1 Markov Random Fields

Bayesian networks are a class of models that can compactly represent many interesting probability distributions. However, some distributions may have independence assumptions that cannot be perfectly represented by the structure of a Bayesian network [60].

There exists, however, another technique for compactly representing and visualizing a probability distribution that is based on the language of undirected graphs. This class of models (known as Markov Random Fields or MRFs) can compactly represent independence assumptions that directed models cannot represent [62].

A Markov Random Field (MRF) is a probability distribution p over variables $x1, \ldots, xn$ defined by an undirected graph G in which nodes correspond to variables xi [63]. The probability p has the form:

$$p\left(x_1, \ldots, x_n\right) = \frac{1}{Z} \prod_{c \in C} \phi_c\left(x_c\right)$$

where C denotes the set of cliques (i.e. fully connected subgraphs) of G, and each factor ϕ_c is a nonegative function over the variables in a clique. The partition function:

$$Z = \sum_{x_1, \ldots, x_n} \prod_{c \in C} \phi_c\left(x_c\right)$$

is a normalizing constant that ensures that the distribution sums to one.

4.9.2.2 Conditional Random Fields

The use of discriminative models for classification tasks has become popular [64]. CRFs offer a lot of advantages over the generative approaches by directly modeling the conditional probability $P(Y|X)$. Thus, the relations between the input variables do not need to be explicitly represented. Since no assumption is made about the underlying structure of the observations X, the model is able to incorporate a rich set of non-independent overlapping features of the observations.

In the context of images, various authors have successfully applied CRFs for classification tasks and have reported significant improvement over the MRF based generative models [65], [66].

A Conditional Random Field (CRF) is an undirected graphical model in which edges represent conditional dependencies between random variables at the nodes. The distribution of each random variable yi is conditioned on an input sequence x. The conditional dependency of the random variables on x is defined by using feature functions with some associated weights. Together, they can be used to determine the probability of each yi. Dependencies among the input variables x do not need to be represented because the model is conditional, affording the use of complex and rich features of the input. Thus, CRFs are discriminative models, that is, they model $p(y|x)$.

Formally, a CRF is a Markov network over variables $X \cup Y$ which specifies a conditional distribution [67],

$$P(y|x) = \frac{1}{Z(x)} \prod_{c \in C} \phi_c \left(x_c, y_c\right)$$

with partition function

$$Z(x) = \sum_{y \in \mathcal{Y}} \prod_{c \in C} \phi_c \left(x_c, y_c\right)$$

Note that in this case, the partition constant now depends on x (therefore, we say that it is a function), which is not surprising: $p(yx)$ is a probability over y that is parametrized by x, in other words it encodes a different probability function for each x. In that sense, a conditional random field results in an instantiation of a new Markov Random Field for each input x.

4.10 BACKGROUND AND STATE OF ART

Depth estimation from monocular images is a long-standing problem in computer vision. Early works on depth estimation using RGB images usually relied on handcrafted features and inference on probabilistic graphical models. Classical methods include shape-from-shading [68] and shape-from-defocus [69]. Newer methods treat depth estimation as a machine learning problem, most recently using deep artificial neural networks [70] [71]. For instance, Saxena et al. [72] estimated the absolute scales of different image patches and inferred a depth image using a Markov Random Field model. Eigen et al. [73], [74] used a multiscale convolutional network to regress from color images to depths. Roy et al. [75] combined shallow convolutional networks with regression forests to reduce the need for large training sets. In [76] the proposed attention model is seamlessly integrated into a CRF, allowing end-to-end training of the entire architecture. This approach benefits from a structured attention model which automatically regulates the amount of information transferred between corresponding features at different scales.

The approach of Li et al. [77] combines deep learning features on image patches with hierarchical CRFs defined on a superpixel segmentation of the image. They use pretrained AlexNet [78] features of image patches to predict depth at the center of the superpixels. Liu et al. [79] also propose a deep structured learning approach that avoids hand-crafted features. In this paper is presented a deep structured learning scheme which learns the unary and pairwise potentials of continuous CRF in a unified deep CNN framework. Liu et al. [80] proposed a discrete-continuous CRF model to take into consideration the relations between adjacent superpixels, e.g., occlusions.

Recent works have also shown the benefit of adopting multi-task learning strategies, e.g. for jointly predicting depth and performing semantic segmentation, ego-motion estimation or surface normal computation [81], [82], [83]. Some recent papers have proposed unsupervised or weakly supervised methods for reconstructing depth maps [84] [85].

With the rapid development of deep neural networks, monocular depth estimation based on deep learning and computer vision tecniques has been widely studied recently and achieved promising performance in accuracy [86]. However, not considering information from other sensors makes the estimate not so robust. The aforementioned works use different kinds of network frameworks, loss functions, and training strategies with just one sensory modality. The architecture proposed in this thesis uses two sensory modalities.

Fusing data coming from multiple sensors has the potential of improving the robustness of depth estimates. Ma et al. [87] uses RGB images together with sparse depth information to train a bottleneck network architecture. Other works [88], [89], [90], [91] are able to fuse the information from both sources to significantly improve the resolution of low quality and sparse range images.

Wang et al. [92] proposed a multi-scale feature fusion method for depth completion using sparse LIDAR data. Ma et al. [93], [94] proposed two methods, a supervised method for depth completion using a ResNet based architecture and a self-supervised method which uses the sparse LiDAR input along with pose estimates to add additional training information based on depth and photometric losses.

Although recent methods have achieved impressive progress in evaluation metrics such as the pixel-wise relative error, most of them neglect the geometric constraints in the 3D space. This component is considered in our CRF model which makes this approach different from previous fusion methods.

Providing strong cues on surface information is relevant for improving depth prediction accuracy [95]. Recently, Zhang et al. [96] proposed to predict surface normals and occlusion boundaries using a deep network and further utilized them to help depth completion in indoor scenes. Qiuet al. [97] propose an end-to-end deep learning system to produce dense depth from sparse LiDAR data and a color image taken from outdoor on-road scenes leveraging surface normal as the intermediate representation. Zhang et al. [98] predicted surface normals by leveraging RGB data, leading to a better prior for depth completion. The model developed in thesis takes advantage of the surface normals to improve the performance of the proposed model.

In contrast, to other models [99], [100], [101], [102], the model proposed here does not rely on a stereo matching algorithm that tend to be computationally costly.

5. CONSTRUCTING THE MULTI-SENSOR DEPTH PREDICTION MODEL

This thesis will apply probabilistic graphical models to the problem of fusing low- resolution depth images with high-resolution camera images to enhance the resolution and accuracy of the depth image. Specifically, a Conditional Random Field (CRF) [103] method will be proposed for integrating both data sources.

CRFs are popular graphical models used for structured prediction. While extensively studied in classification (discrete) domains, CRFs have been less explored for regression (continuous) problems. One of the pioneering works on continuous CRFs can be attributed to [104], in which they were proposed for global ranking in document retrieval.

Different from the previous efforts, the perception system to develop will have the ability of utilizing a unary potential, different pair wise potentials and a higher order potentials defined on super pixels, without relying on any geometric priors nor any extra information, other than range measurement coming from a LiDAR sensor. Our approach will perform this data fusion using a multi-resolution CRF, which ties together image and range data, and fast optimization techniques such as a conjugate gradient algorithm, for the CRF inference problem.

5.1 PROPOSED ARCHITECTURE

This thesis proposes a CRF model to predict dense depth images from a single camera and a scanning laser. We make the common assumption that an image is composed of small homogeneous regions called superpixels. The aim is to assign each image superpixel with a range value using both image appearance and sparse laser data. However, this framework will be flexible. It might work on pixels or superpixels, this parameter will determine the resolution of the algorithm.

We formulate the energy function as a typical combination of unary potentials and pairwise potentials over the nodes and edges of the image. These potentials are built based on multiple geometry and appearance information that seamlessly represents the problem of estimating dense depth maps from camera and LiDAR data.

In our case, the unary term aims to regress the depth value from a single superpixel and the pairwise term encourages neighboring superpixels with similar appearances to take similar depths.

5.1.1 CRF-based camera-LIDAR fusion fordepth estimation

In this thesis, depth estimation is formulated as a superpixel-level inference task on a modified Conditional Random Field (CRF). Our proposed model is a multi-sensor extension to the classical pairwise CRF. In this section, I introduce the CRF model proposed. I show how to fuse the information of an image and a sparse LIDAR point cloud with our novel CRF framework.

5.2 OVERVIEW

The Conditional Random Field (CRF) is a type of undirected probabilistic graphical model which is widely used for solving labeling problems. Formally, let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ be a set of discrete random variables to be inferred from an observation or input tensor $\mathbf{Y}$, which in turn are composed of the observation variables c_i and y_i, where i is an index over superpixels. For each super pixel i, the variable c_i corresponds to an observed three-dimensional colour value; y_i is an observed range measurement.

The goal of our framework is to infer the depth of each pixel in a single image depicting general scenes. I make the common assumption that an image is composed of small homogeneous regions (superpixels) and consider the graphical model composed of nodes defined on superpixels. Note that our framework is flexible and can estimate depth values on either pixels or superpixels.

The remaining question is how to parameterize this undirected graph. Intuitively, I want the model captures the affinities between the depth estimates among superpixels in a given neighborhood. These affinities can be captured as follows: Let $\tilde{P}(X, Y)$ be an unnormalized Gibbs joint distribution parameterized as a product of factors Φ, where

$$\Phi = \{\phi_1(D_1), \ldots, \phi_k(D_k)\},$$

and

$$\tilde{P}(X, Y) = \prod_{i=1}^{m} \phi_i(D_i).$$

We can then write a conditional probability distribution of the depth estimates X given the observations Y, using the previously introduced Gibbs distribution as follows:

$$Pr(X|Y) = \frac{P(X,Y)}{Z(Y)}$$

where

$$Z(Y) = \sum_X \tilde{P}(X,Y).$$

$Z(Y)$, also known as the partition function, works as a normalizing factor which marginalises X from $\tilde{P}(X,Y)$, allowing the calculation of the probability distribution $P(X|Y)$:

$$P(X|Y) = \frac{1}{\sum_X \tilde{P}(X,Y)} \tilde{P}(X,Y).$$

Therefore, similar to conventional CRFs, I model the conditional probability distribution of the data with the following density function:

$$\mathrm{P}(\mathbf{X}|\mathbf{y}) = \frac{1}{\mathrm{Z}(\mathbf{Y})} \exp(-E(\mathbf{X},\mathbf{Y}))$$

where E is the energy function; Z is the partition function defined as:

$$\mathrm{Z(Y)} = \int_{\mathrm{Y}} \exp\{-E(\mathrm{X},\mathrm{Y})\}\mathrm{dY}.$$

Since Z is continuous, the integral equation above can be analytically calculated. This is different from the discrete case, in which approximation methods need to be applied. To predict the depths of a new image, we solve the following maximum a-posteriori (MAP) inference problem:

$$\mathbf{x}^{\star} = \operatorname*{argmax}_{\mathbf{x}} \mathrm{P}(\mathbf{X}|\mathbf{Y}).$$

To simplify the solution to the energy function, one can take the negative logarithm of the left hand side and right side of the equation of the probability distribution $P(X|Y)$, and the problem of maximizing the conditional probability becomes an energy minimization problem.Therefore, maximizing the probability distribution $P(X|Y)$ is equivalent to minimizing the corresponding energy function:

$$\mathbf{x}^{\star} = \arg \underset{\mathbf{x}}{\text{mín}}\, E(\mathbf{X}, \mathbf{Y}).$$

I formulate the energy function as a typical combination of unary potentials U and pairwise potentials V over the nodes (superpixels) N and edges S of the image x:

$$E(\mathbf{X}, \mathbf{Y}) = \sum_{p \in \mathcal{N}} U\left(x_p, \mathbf{y}\right) + \sum_{(p,q) \in \mathcal{S}} V\left(x_p, x_q, \mathbf{y}\right)$$

The unary term U aims to regress the depth value from a single superpixel. The pairwise term V encourages neighbouring superpixels with similar appearances to take similar depths [105], [106].

Fig. 9 illustrates the modules of the proposed model. On the top left is the fused view of the image and LIDAR point cloud on superpixels. On the top right are the normal surface map and RGB inputs used in the pairwise potentials. On the top middle is the graph structure of the CRF: The yellow nodes represent the centroid of image superpixels and the green branches the connection between them. The outputs of unary part and the pairwise part are then fed to the CRF structured loss layer, which minimizes the corresponding energy function. On the bottom left is the probabilistic output, a dense depth map and uncertainty estimation map, of the method proposed.

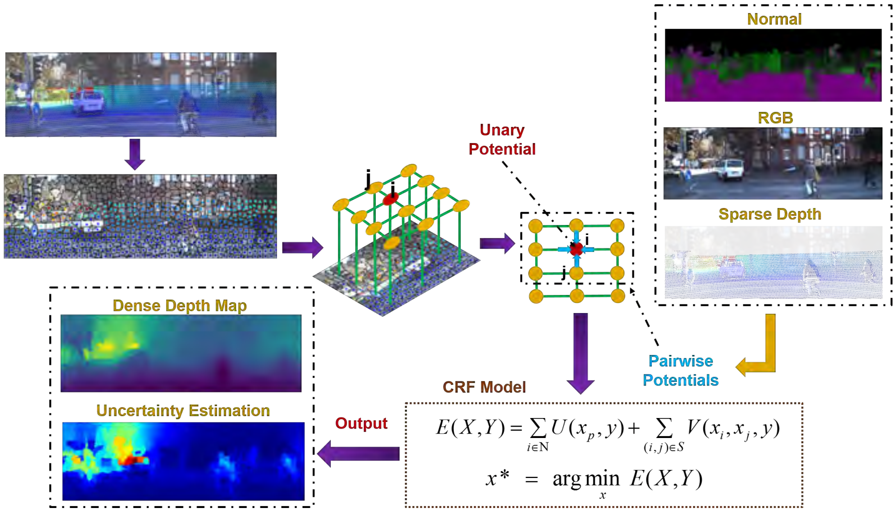

***Fig.* 9**. Illustration of the proposed model.

5.3 POTENTIAL FUNCTIONS

The proposed multi-modal depth estimation model is composed of unary and pairwise potentials. For an input image, which has been over-segmented into n superpixels, we define a unary potential for each superpixel. The pairwise potentials are defined over the four-neighbor vicinity of each superpixel.

The unary potentials are built by aggregating all LiDAR observations inside each superpixel. The pairwise part is composed of similarity vectors, each with K components, that measure the agreement between different features of neighbouring superpixel pairs. Therefore, we explicitly model the relations of neighbouring superpixels through pairwise potentials. In the following, we describe the details of the potentials involved in the energy function.

5.3.1 Unary potential

The unary potential is constructed from the LiDAR sensor measurements by considering the least square loss between estimated x_i and observed y_i depth values:

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i \in L} \sigma_i \left(x_i - y_i\right)^2$$

$$\Phi(\mathbf{x}, \mathbf{y}) = \|\mathbf{W}(\mathbf{x} - \mathbf{z})\|^2$$

where L is the set of indexes for which a depth measurement is available, and σ_i is a constant weight placed on the depth measurements. This potential measures the quadratic distance between the estimated range X and the measured range Z, where available. Finally, in order to write the unary potential in a more efficient matrix form, we define the diagonal matrix W with entries:

$$\mathbf{W}_{i,i} = \begin{cases} \sigma_i & \text{if } i \in L \\ 0 & \text{otherwise} \end{cases}$$

5.3.2 Colour pairwise potential

We construct a pairwise potential from K types of similarity observations, each of which enforces smoothness by exploiting colour consistency features of neighbouring superpixels. This pairwise potential is expressed as:

$$\Psi^c(\mathbf{x}, \mathbf{I}) = \sum_{i} \sum_{j \in N(i)} e_{i,j} \left(x_i - x_j\right)^2$$

$$\Psi^c(\mathbf{x}, \mathbf{I}) = \|\mathbf{S}\mathbf{x}\|^2$$

where I is a RGB image, $N(i)$ is the set of horizontal and vertical neighbours of i, and each row of S represents weighting factors for pairs of adjacent range nodes. As the edge strength between nodes we use an exponentiated L_2 norm of the difference in pixel appearance.

$$e_{i,j} = \exp -\frac{\|\mathbf{c}_i - \mathbf{c}_j\|^2}{\sigma_d^2}$$

where $\mathbf{c}_i$ is the RGB colour vector of pixel i and σ_d is a tuning parameter. A small σ_d value increases sensitivity to changes in the image.

5.3.3 Surface-normal pairwise potential

The mathematical formulation of this potential is similar to the previous colour potential, however, the surface-normal potential considers surface normal similarities instead of colour. The weighting factors $nr_{i,j}$ for this case, are formulated using the cosine similarity, which is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1°, and it is less than 1 for any angle in the interval $(0, (pi)]$ radians. It is thus a measurement of orientation instead magnitude [107].

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\|\|\mathbf{B}\| \cos\theta$$

Therefore, the cosine similarity is expressed like:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{t=1}^{n} A_i B_i}{\sqrt{\sum_{t=1}^{n} A_i^2}\sqrt{\sum_{t=1}^{n} B_i^2}}$$

where A_i and B_i are the components of vectors A and B respectively. Finally, we define our surface normal potential in the following equation.

$$\Psi^n(\mathbf{x}, \mathbf{In}) = \sum_i \sum_{j \in N(i)} nr_{i,j} \left(x_i - x_j\right)^2$$

$$\Psi^n(\mathbf{x}, \mathbf{In}) = \|\mathbf{Px}\|^2$$

$$nr_{i,j} = \frac{\sum_{t=1}^{n} In_i In_j}{\sqrt{\sum_{t=1}^{n} In_i^2}\sqrt{\sum_{t=1}^{n} In_j^2}}$$

5.3.4 Depth pairwise potential

This pairwise potential encodes a smoothness prior over depth estimates which encourages neighboring superpixels in the image to have similar depth. Usually, pairwise potentials are only related to the color difference between pairs of superpixels, however depth smoothness is a valid hypotheses which can potentially enhance depth inference.

For enforcing depth smoothness, a distance-aware Potts model was adopted. The neighboring points with smaller distance are considered to be more likely to have the same depth.

The mathematical formulation of this potential is similar to the colour pairwise potential, as it follows the Potts model:

$$\Psi^d(\mathbf{x}, \mathbf{D}) = \sum_i \sum_{j \in N(i)} e_{i,j} \left(x_i - x_j\right)^2$$

and the weighting factor $dp_{i,j}$ for this case is formulated as:

$$dp_{i,j} = \exp - \frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{\sigma_p^2}$$

where $\mathbf{p}_i$ is the 3D location vector of the LiDAR point i, and σ_p is the parameter controlling the strength of enforcing the close points to take similar depth values.

5.3.5 Uncertainty potential:

Depth uncertainty estimation is important for refining depth estimation [108], [109], and in safety critical systems [110], it allows an agent to identify unknowns in an environment in order to reach optimal decisions. Our method provides pixel-wise depth uncertainties estimates by taking into account the amount of LiDAR points present for each superpixel.

The uncertainty potential is similar to the unary potential. It is constructed from the number of LiDAR points projected on a superpixel and calculating the following least square loss:

$$U^c(\mathbf{x}, \mathbf{y}) = \sum_{i \in L} \sigma_i \left(u_i - unc_i\right)^2$$

$$U^c(\mathbf{x}, \mathbf{y}) = \|\mathbf{W}(\mathbf{u} - \mathbf{unc})\|^2$$

where the vector $\mathbf{unc}$ is defined as follows:

$$\mathbf{unc}_{i,i} = \begin{cases} \sigma_i & \text{if P projected on SPx is 0} \\ \psi i & \text{if P projected on SPx is >0 and <2} \\ mean & \text{otherwise} \end{cases}$$

where P is a 3D point and SPx is a superpixel. In locations with accurate and sufficient LiDAR points, the model will produce depth predictions with a high confidence. This uncertainty estimation provides a measure of how confident the model is about the depth estimation.This results in an overall better performance, since uncertain estimates with high uncertainty can be neglected by higher level tasks that use the estimated depth maps as an input.

5.4 OPTIMIZATION

With the unary and the pairwise potentials defined, we can now write the energy function as:

$$E(\mathbf{X}, \mathbf{Y}) = (\alpha)\, \Phi(\mathbf{x}, \mathbf{y}) + (\beta)\Psi^c(\mathbf{x}, \mathbf{I}) \ldots$$

$$+ \ \ldots (\gamma)\Psi^n(\mathbf{x}, \mathbf{In}) + (\delta)\Psi^d(\mathbf{x}, \mathbf{In}) \qquad (1)$$

The scalars α, β, γ, $\delta \in$ [0,1] are weightings between the four terms. We may further expand the unary and pairwise potentials to the form:

$$\Phi(\mathbf{x},\mathbf{y}) = \alpha(\mathbf{x}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x} - 2\mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{x} + \mathbf{z}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\mathbf{z}) \tag{2}$$

$$\Psi^{c}(\mathbf{x},\mathbf{In}) = \beta(\mathbf{x^{T}S^{T}S_{X}}) \tag{3}$$

$$\Psi^{n}(\mathbf{x},\mathbf{In}) = \gamma(\mathbf{x^{T}P^{T}P_{X}}) \tag{4}$$

$$\Psi^{d}(\mathbf{x},\mathbf{In}) = \delta(\mathbf{x^{T}D^{T}D_{X}}) \tag{5}$$

We shall pose the problem as one of finding the optimal range vector $\mathbf{x}^{*}$ such that:

$$\mathbf{x}^{*} = \underset{\mathbf{x}}{\operatorname{argmin}}\,\{E(\mathbf{X},\mathbf{Y})\}$$

Substituting equations 2, 3, 4 and 5 into 1 and solving for x reduces the problem to: $Ax = b$ where,

$$\mathbf{A} = \alpha(\mathbf{W^{T}W}) + \beta(\mathbf{S^{T}S}) + \gamma(\mathbf{P^{T}P}) + \delta(\mathbf{D^{T}D})$$

$$b = \alpha(\mathbf{W^{T}Wz})$$

If the uncertainty potential is added to the model, the mathematical formulation is the following:

$$\mathbf{A} = \left(\begin{smallmatrix} \alpha\mathbf{W^{T}W} & 0 \\ 0 & \mathbf{W^{T}W} \end{smallmatrix}\right) + \left(\begin{smallmatrix} \beta\mathbf{S^{T}S} & 0 \\ 0 & 0 \end{smallmatrix}\right) + \left(\begin{smallmatrix} \gamma\mathbf{P^{T}P} & 0 \\ 0 & 0 \end{smallmatrix}\right) + \left(\begin{smallmatrix} \delta\mathbf{D^{T}D} & 0 \\ 0 & 0 \end{smallmatrix}\right)$$

$$b = \left(\begin{smallmatrix} \alpha\mathbf{W^{T}Wz} & 0 \\ 0 & \mathbf{W^{T}Wunc} \end{smallmatrix}\right)$$

All I need to perform the optimization is to solve a large sparse linear system. The methods for solving sparse systems are distinguished in two categories: direct and iterative. Direct methods are robust but require large amounts of memory, as the size of the problem grows. On the other hand, iterative methods provide better performance but may exhibit numerical problems [11]. In this paper, the fast algorithm Conjugate Gradient Squared proposed by Hestenes and Stiefel [112], [113] is employed to solve the energy minimization problem.

6. EXPERIMENTS

6.1 THE KITTI ODOMETRY DATASET

In this work I use the odometry dataset, which includes both camera and LiDAR measurements . The odometry dataset consists of 22 sequences. Among them, 5 sequences are used to calculate the hyperparameters and 10 for evaluation. I use a random subset of 50 images from the test sequences for the final evaluation. We use the right RGB camera. The Velodyne LiDAR measurements are projected onto the RGB images.

For testing only the bottom crop (912 X 228) is used, since the LiDAR returns no measurement to the upper part of the images. The KITTI dataset for depth completion and prediction is: http://www.cvlibs.net/datasets/kitti/eval_depth_all.php

6.2 IMPLEMENTATION DETAILS

I implement the framework on a desktop Core i7, 8GB memory RAM with an NVIDIA GeForce GT 720M. Processing one image takes around 7m with 1800 superpixels. Take around 1 day to find the hyperparameters of the probabilistic model proposed. These values were found using random search.

6.3 EVALUATION METRICS

I evaluate the accuracy of our method in depth prediction using the 3D laser ground truth on the test images. I use the following depth evaluation metrics root mean squared error (RMSE), mean absolute error (MAE) and mean absolute relative error (REL), among which RMSE is the most important indicator and chosen to rank submissions on the leader-board since it measures error directly on depth and penalizes on further distance where depth measurement is more challenging. These metrics were used by [115], [116], [117] to estimate the accuracy of monocular depth prediction.

$$RMSE = \sqrt{\frac{1}{|T|} \sum_{d \in T} \|\hat{d} - d\|^2}$$

$$MAE = \frac{1}{T} \sum_{d \in T} \|\hat{d} - d\|^2$$

$$REL = \frac{1}{T} \sum_{d \in T} \left(\frac{\left| \hat{d} - d \right|}{\hat{d}} \right)$$

where d is the ground truth depth, $\hat{d}$ is the estimated depth, and T denotes the set of all points in the test set images. In order to compare the results with Eigen et al. [98] and Godard et al. [104], I crop the image to the evaluation crop applied by Eigen et al. I also use the same resolution of the ground truth depth image and cap the predicted depth at 80 m.

7. RESULTS AND DISCUSSION

I evaluate our approach on the raw sequences of the KITTI benchmark, which is a popular dataset for single image depth map prediction. The sequences contain stereo imagery taken from a driving car in an urban scenario. The dataset also provides 3D laser measurements from a Velodyne laser scanner that I use as ground-truth measurements (projected into the stereo images using the given intrinsics and extrinsics in KITTI). This dataset has been used to train and evaluate the state-of-the-art methods and allows for quantitative comparison.

First, I evaluate the prediction accuracy of our proposed method with different potentials in Section 7.1. Second, in Section 7.2 I explore the impact of the number of the number of superpixels on the depth estimation performance. Third, Section 7.3 compares my approach to state-of-the-art methods on the KITTI dataset. Finally, in Sections 7.4 and 7.5, I demonstrate two use cases of the proposed algorithm, one for creating LiDAR super-resolution from sensor data provided by the KITTI dataset and another one for a dataset collected in the context of this work.

7.1 ARCHITECTURE EVALUATION

This section presents an empirical study on the impact of different potential functions and hyparameter choices on the depth prediction accuracy. In a first experiment, I compare the impact of sequentially adding my proposed pairwise potentials. I first evaluate a model with only unary and colour pairwise potentials. Then I added the surface-normal pairwise potential, and finally the depth pairwise potential is included. As shown in Table I, the RMSE is improved after adding each pairwise potential.

TABLE I. Method performance after adding pairwise potentials (lower is better)

Algorithm	Potential functions	RMSE
Ours	I	865.31
Ours	II	854.24
Ours	III	849.39

Fig. 10 shows the qualitative evaluation of the impact of the pairwise potentials. In row order: 1st: Pairwise potential I, penalizes dissimilar depth estimates of neighboring pixels which take similar colors in the RGB image, 2nd: pairwise potential II, penalizes the depth differences between neighboring superpixels whose normal surface vectors have large cosine similarities, 3rd: pairwise potential III, penalizes neighboring superpixels with large observed depth differences.

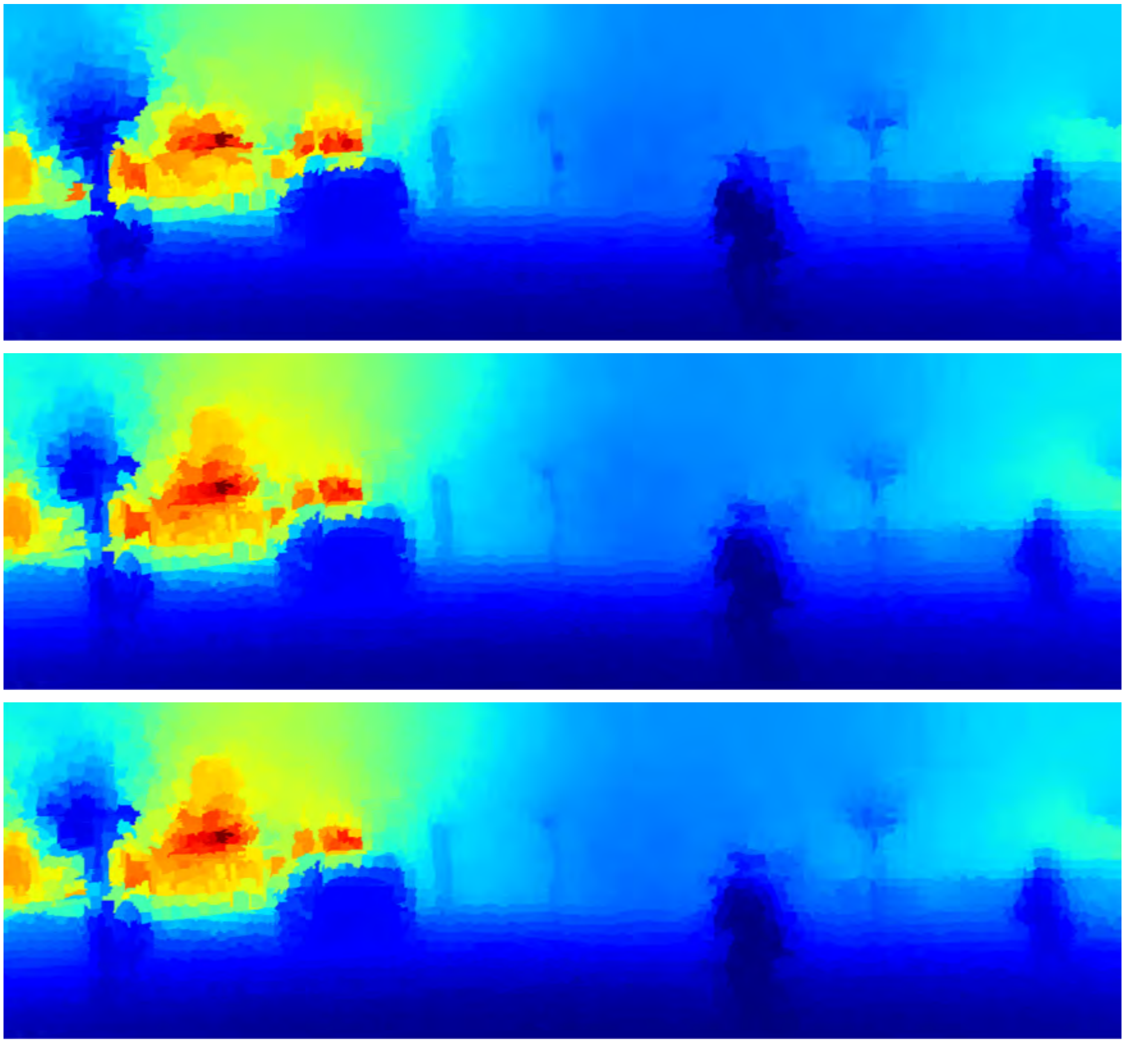

***Fig.* 10**. Qualitative evaluation of the impact of the pairwise potentials defined as CRF terms.

7.2 ON NUMBER OF SUPERPIXELS

In this section, I explore the relation between the prediction accuracy and the number of superpixels.

As displayed in Fig. 11, a greater number of superpixels yields better results in error measurements. Although the more number of superpixels improves the quality of depth map, this makes the computational cost to increase.

TABLE II. Comparison on the number of super-pixels (lower is better)

Algorithm	#Superpixels	RMSE	Time x frame
Ours	1200	1370.27	5 m
Ours	2400	1050.55	11 m
Ours	5500	848.84	26 m

Additionally, Table II shows the time to estimate depth using all the potentials. As can be seen, the execution time varies according to the number of superpixels set. It is very important to highlight that the proposed model is not computationally efficient. However, this can be solved using parallel programming or others approaches.

7.3 ALGORITHM EVALUATION FOR DEPTH COMPLETION

The KITTI odometry dataset is more challenging dataset as compared to other dataset for depth estimation. Since, the distances in the KITTI dataset are larger than other datasets as in NYU-Depth-V2 dataset.

The value of error metrics have been taken from the respective research articles. The quantitative results of the proposed method here and other existing methods on KITTI dataset are shown in Table III. From Table III, I can observe that the proposed method outperforms other existing methods. Qualitative results are shown in Fig. 12.

TABLE III. Depth completion errors by different methods on the test set of KITTI depth completion benchmark (lower is better)

Algorithm	RMSE	MAE
Schneider et al. [105]	2312.57	605.47
Cheng et al. [106]	1019.64	279.46
Huang et al. [107]	841.77	253.47
Hambarde et al. [108]	830.57	247.85
Ma et al. [109]	814.73	249.95
Ours	849.39	263.31

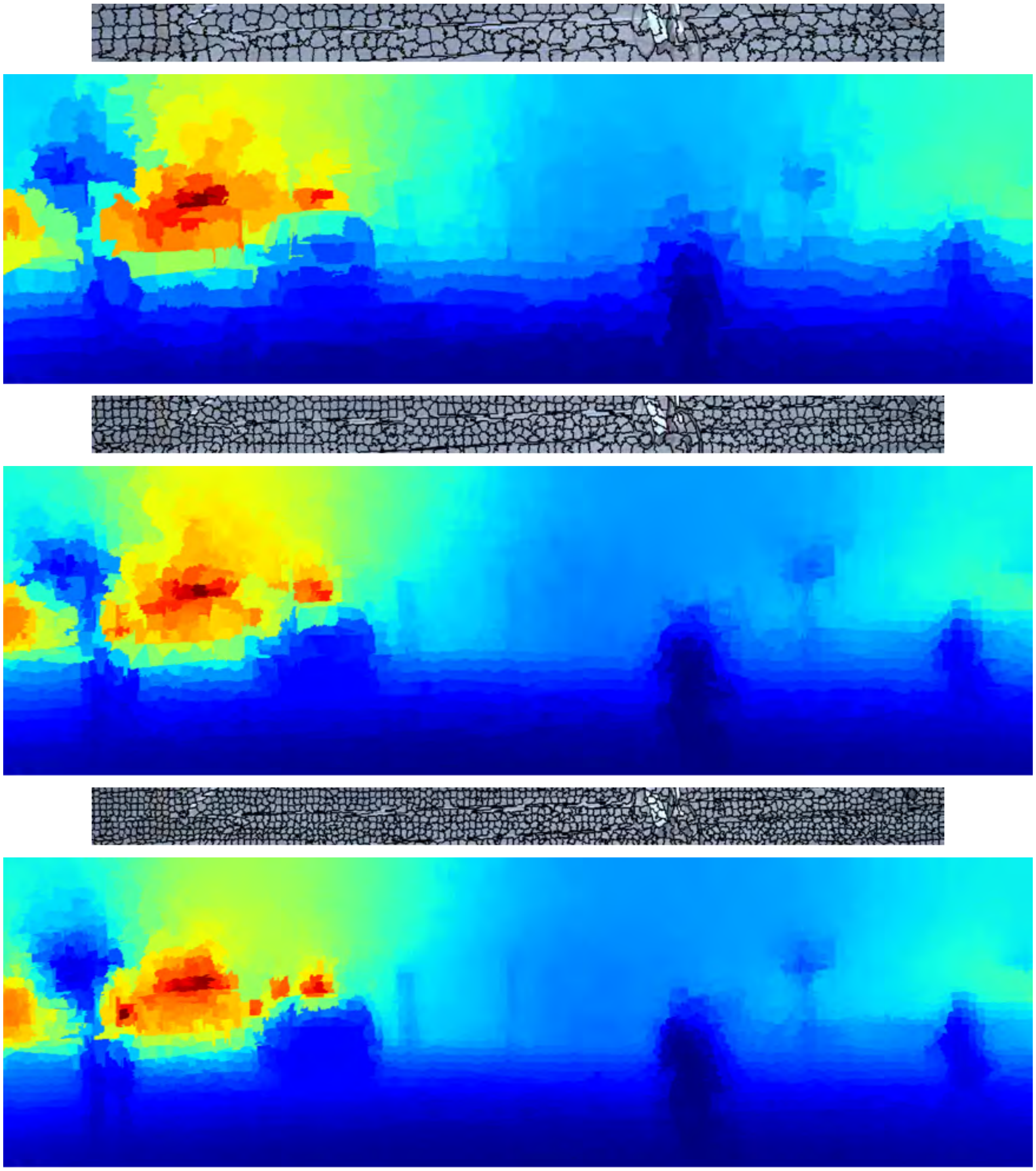

***Fig.* 11**. Visual comparison of dense depth maps produced by the CRF framewor proposed varying superpixel size. From top to bottom, 1200, 2400 and 5500 superpixels.

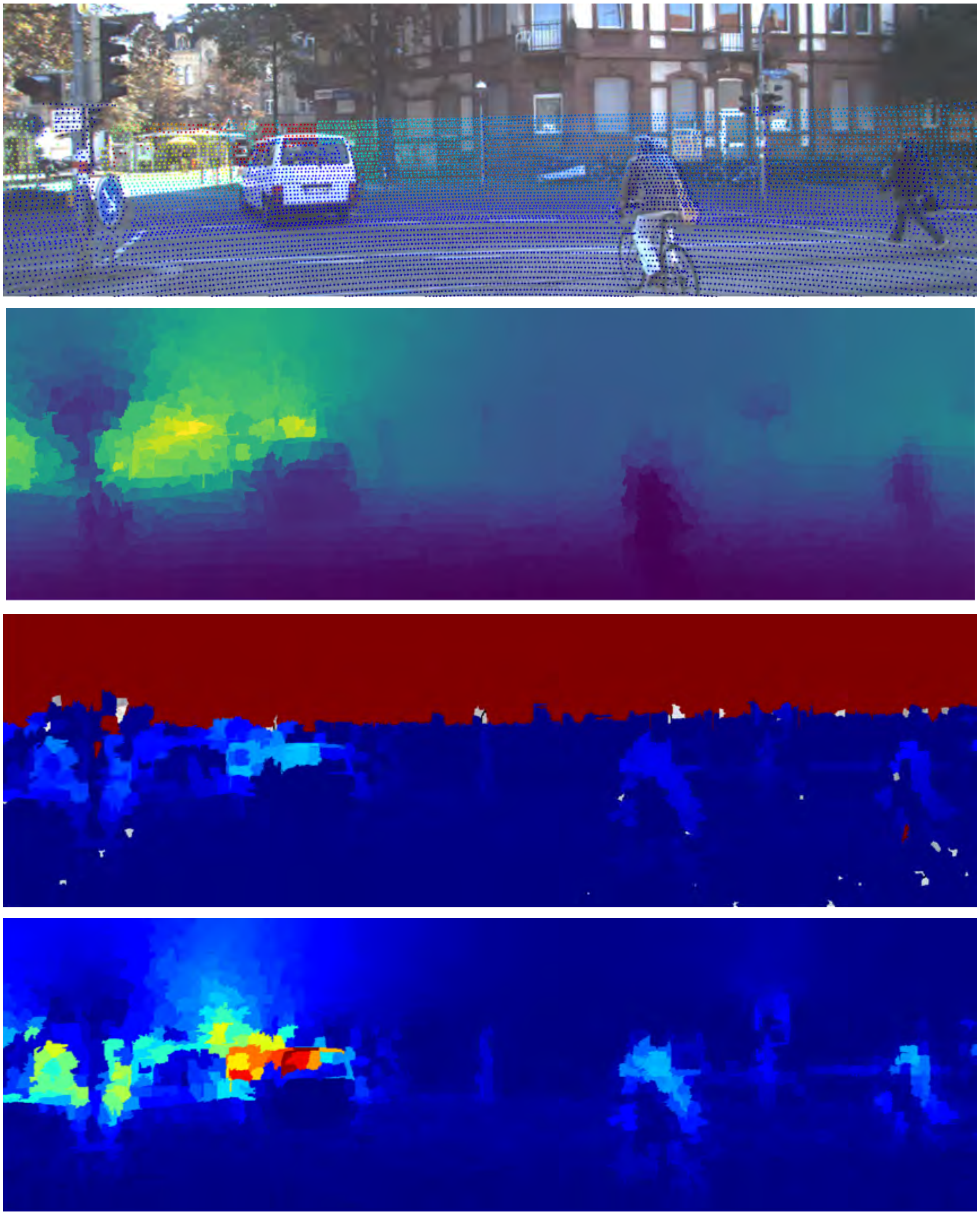

***Fig.* 12**. Depth completion and uncertainty estimates of our approach on the KITTI raw test set. From top to bottom: RGB and raw depth projected onto the image; high-resolution depth map; raw uncertainty; and estimated uncertainty map.

7.4 ALGORITHM EVALUATION FOR LIDAR SUPER-RESOLUTION

I present another demonstration of the method in super-resolution of LiDAR measurements. 3D LiDARs have a low vertical angular resolution and thus generate a vertically sparse point cloud. I use all measurements in the sparse depth image and RGB images as input to the framework.

On the other hand, starting from LiDAR Super-Resolution map I can generate a 3D reconstruction of the scene. Reconstruction of three-dimensional (3D) scenes has many important applications, such as autonomous navigation, environmental monitoring [112] and other computer vision tasks [110].

Therefore, a dense and accurate model of the environment is crucial for autonomous vehicles. In fact, imprecise representations of the vehicle's surrounding may lead to unexpected situations that could endanger the passengers. In this paper, the 3D modeling is generated using the combination of image and range data is a sensor fusion approach that takes strength from each in order to overcome their limitations. Images normally have higher resolution and more visual information than range data, and range data are noisy, sparse, and have less visual information, but already contain 3D information.

The qualitative and quantitative results presented here suggest that our system provides 2D depth map, which can be converted into a 3D point cloud, of reasonable quality. The qualitative results for LIDAR Super-Resolution task are shown in Fig. 13.

Following [114], I use a random subset of images from the test sequences for evaluation. Specifically, I take the bottom part 912×228 due to no depth at the top area, and only evaluate the pixels with ground truth. The performance of the proposed approach and state-of-the-art depth completion methods are recorded in Table IV.

TABLE IV. Depth estimation errors by different methods on the test set of KITTI depth estimation benchmark (lower is better)

Algorithm	RMSE	REL	Log10
Cadena et al. [113]	7.14	0.179	-
Liao et al. [114]	4.51	0.113	0.049
Fu et al. [115]	3.67	0.072	-
Ma et al. [116]	3.37	0.073	-
Cheng et al. [117]	3.24	0.059	-
Hambarde et al. [118]	3.11	0.069	0.038
Ours	3.59	0.072	0.041

Table III and Table IV show that the approach proposed in this thesis achieves good

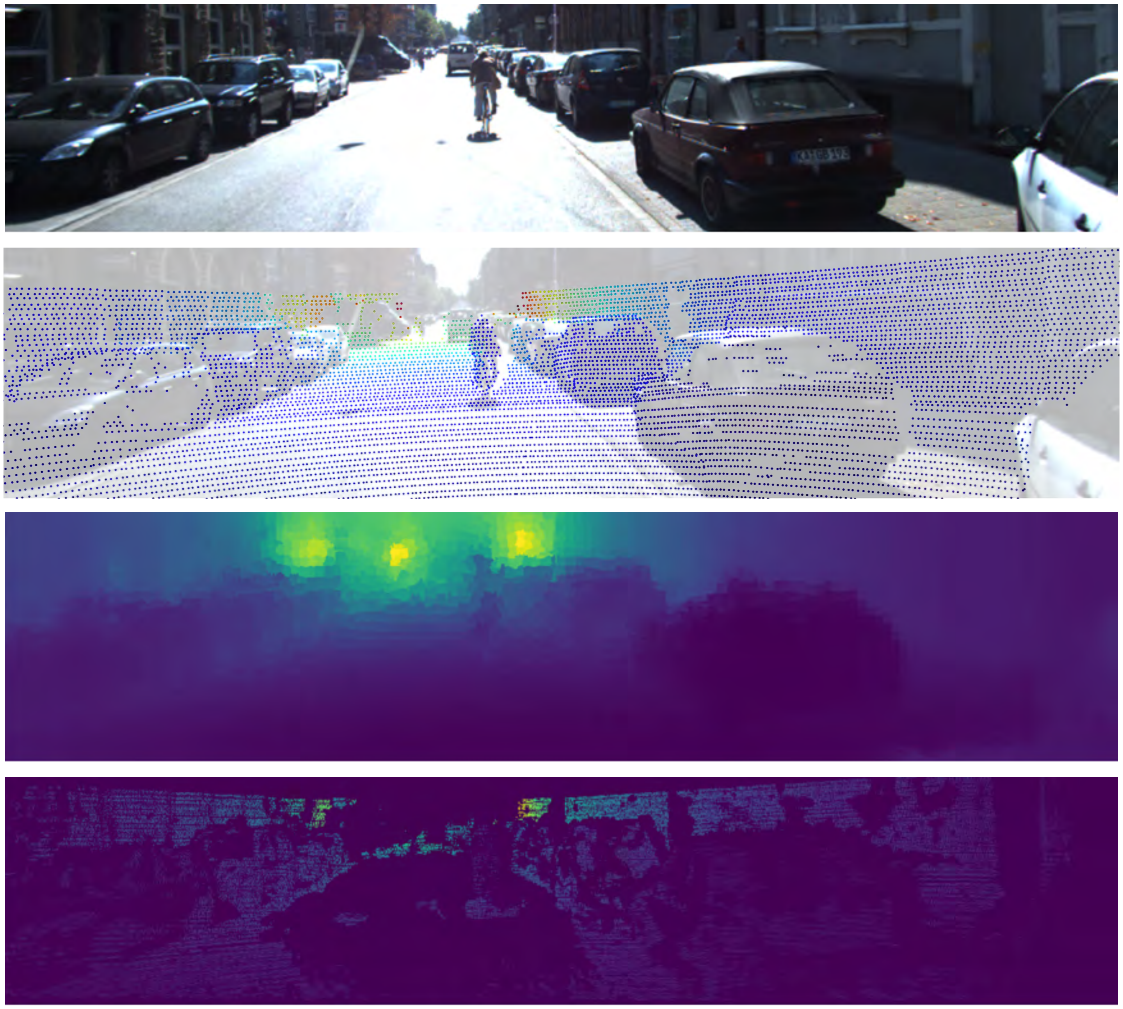

***Fig.* 13**. LiDAR super-resolution. Creating dense point clouds from sparse raw measurements. From top to bottom: RGB image, raw depth map, predicted depth and ground truth depth map. Distant cars are almost invisible in the raw depth map, but are easily recognizable in the predicted depth map performance in single image depth map prediction on the popular KITTI dataset. It is able to predict detailed depth maps on thin and distant objects. It also estimates reasonable depth in image parts in which there is no ground-truth available for supervised learning.

The qualitative results presented here suggest that the system provides 2D depth map of reasonable quality. Nevertheless, it is instructive to consider how the accuracy of the approach depends on the density of laser measurements and the number of superpixeles selected.

Although recent methods have achieved impressive progress in evaluation metrics, as shown in the Tables III and IV, most of them neglect the geometric constraints in the 3D space, or use stereo cameras, infer depth estimation from only cameras or require dataset to train their deep learning models.

All those components translate into disadvantages. For example, using stereo cameras increase the cost of system development. To use deep learning models, it is necessary to have large computational sources and in most of the cases a dataset with its corresponding depth labels.

Those components were considered in the CRF model which makes this approach different from previous fusion methods, being more robust and of low cost.

7.5 APPLICATION: UAO DRIVING LIDAR-RGB DATASET

Thus far, I have sampled depth from high-quality LiDAR depth maps, but in practice sparse depth inputs may come from less reliable sources. Therefore I provide a qualitative evaluation of this model on my own well-calibrated LiDAR and RGB dataset. I use a VLP-16 LiDAR along with a Stereo Labs Zed Mini camera of 1280×720 resolution. The robotic platform used to gather the dataset is shown in Fig. 14.

***Fig.* 14**. Jackal platform, monocular camera and lidar.

This dataset enables us to prove the stability and robustness of the proposed model in particularly challenging scenarios. The scenes were recorded with low resolution of camera and the LiDAR sensor in comparison of the KITTI benchmark.

Notably, the algorithm proposed is able to estimate a dense depth map of indoor and outdoor environments using color and sparse depth data. Experimental results are shown in Fig.15, Fig.16 and Fig.17. The dark red reflects farther distances and the dark blue reflects closer distances.

Despite of the lower number of LiDAR channels, the proposed method has provided accurate depth information even under challenging outdoor conditions, as shown in Fig.17. In this scene there is lot of variability in terms of light and shadows generated by the environment and the weather itself.

After a close look at Fig.15, Fig.16 and Fig.17, it is noticeable that no depth observations from the LiDAR are available at the top and bottom locations of the colour image. After inference, the depth estimates, shown in the bottom images, at the above locations is consistent with the information provided by the image. Therefore I can conclude that the framework proposed here is reliable to work in the depth prediction task. Additionally, it also solves the depth completion problem, as it is able to deal with highly sparse input point clouds projected on the image space.

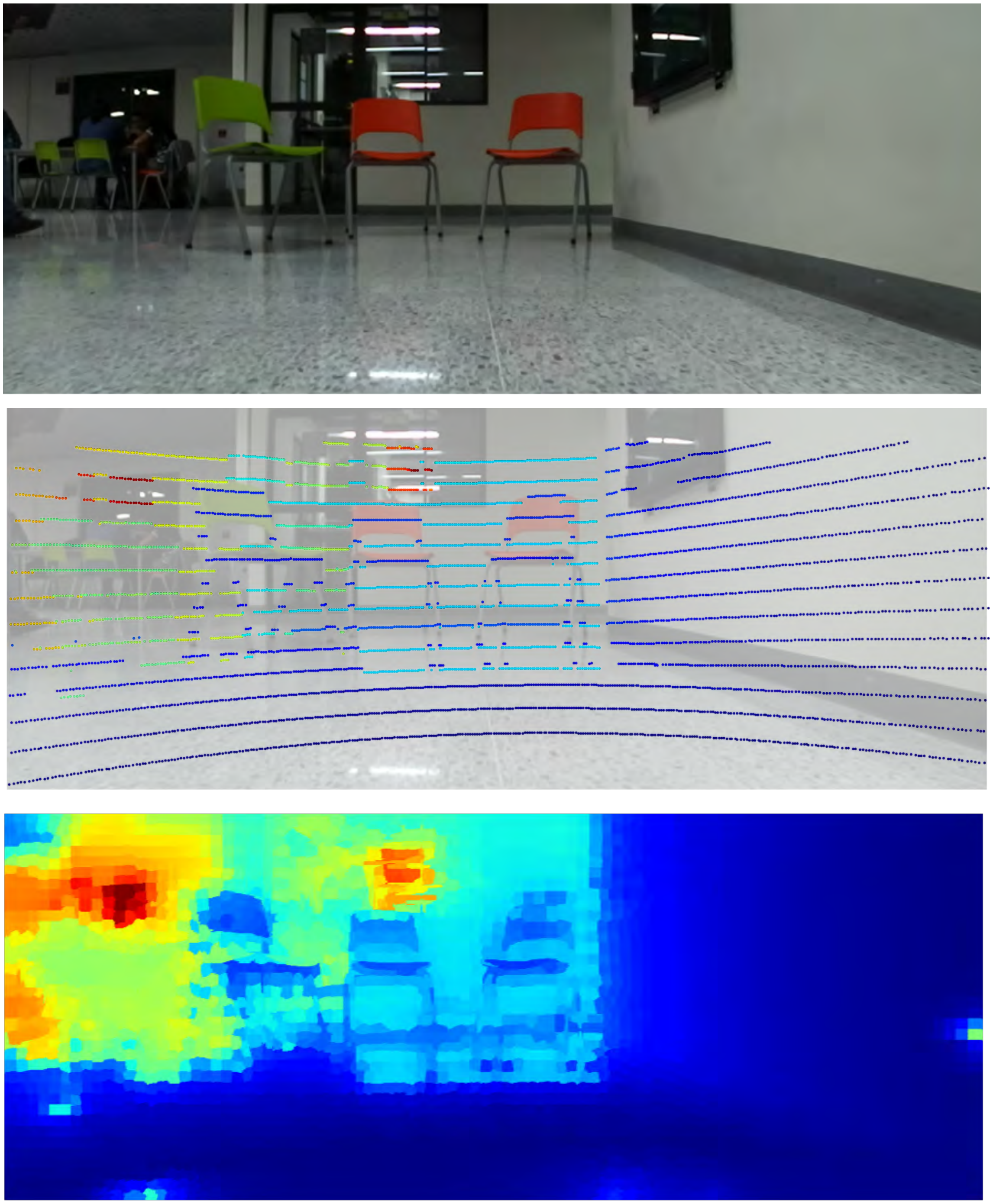

***Fig.* 15**. Indoors: LiDAR super-resolution. Creating dense point clouds from sparse raw measurements and color. From top to bottom: RGB image, raw depth map and predicted depth

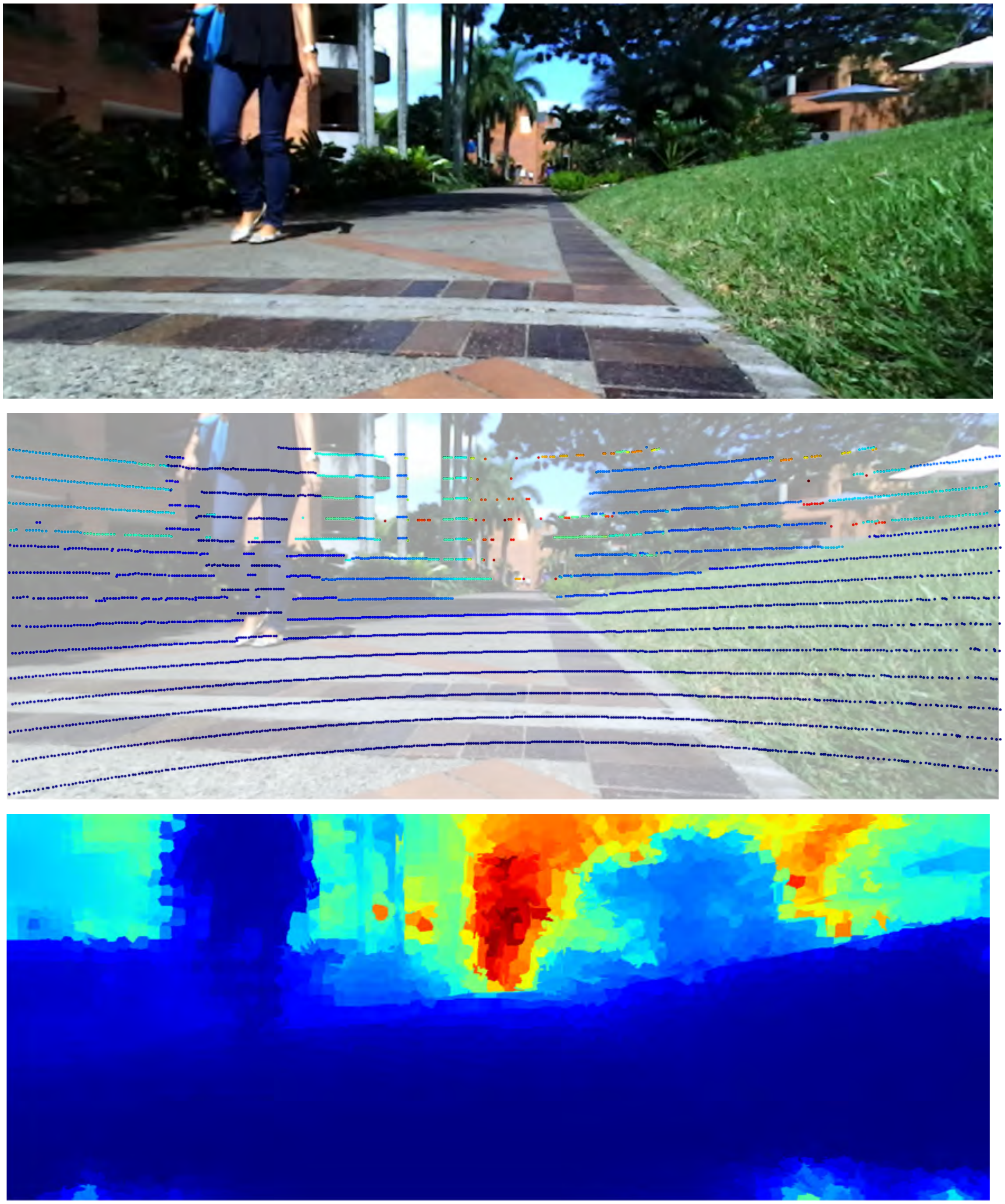

***Fig.* 16**. Outdoors: LiDAR super-resolution. Creating dense point clouds from sparse raw measurements and color. From top to bottom: RGB image, raw depth map and predicted depth.

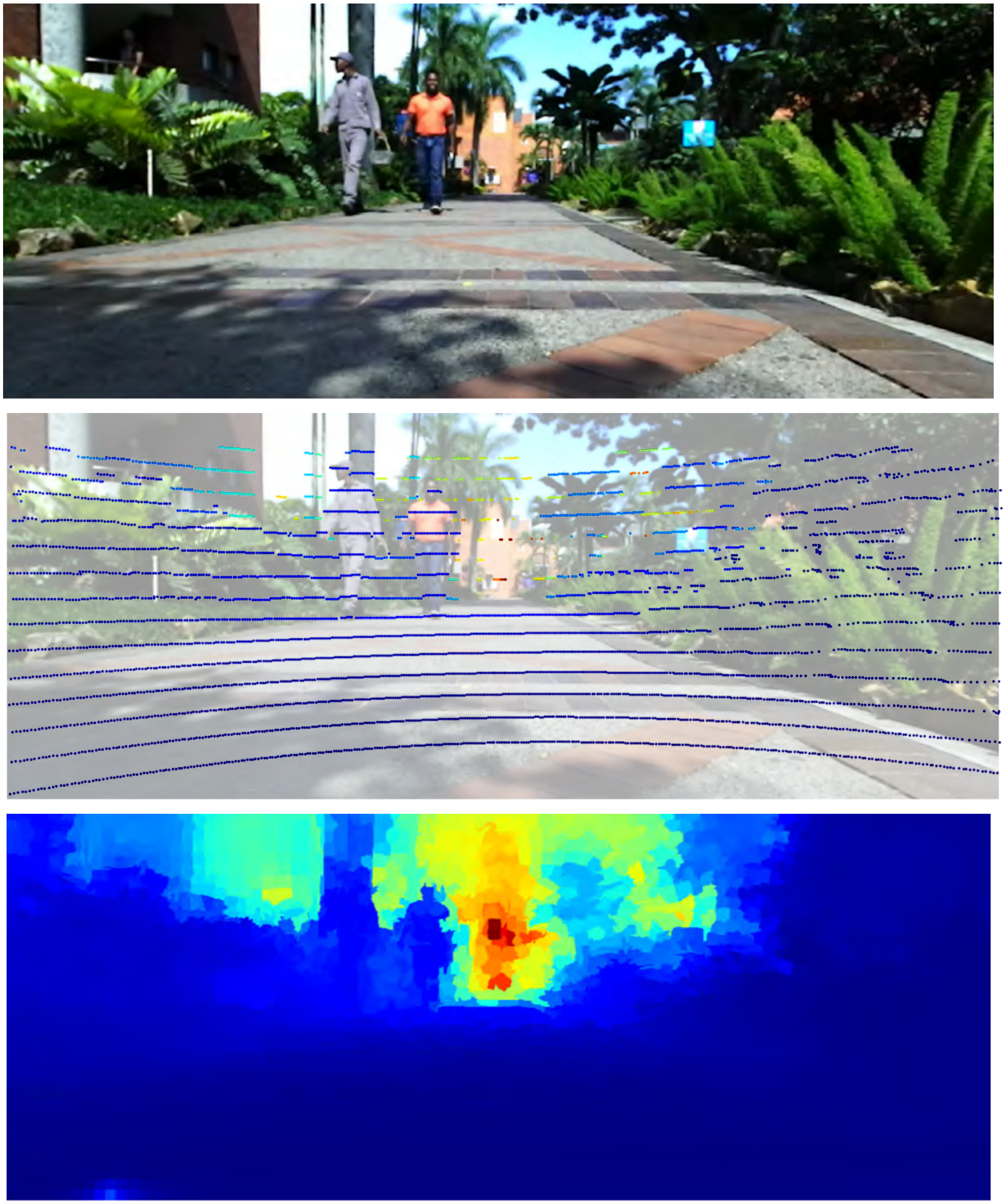

Fig. **17**. Outdoors: LiDAR super-resolution. Creating dense point clouds from sparse raw measurements and color. From top to bottom: RGB image, raw depth map and predicted depth.

8. CONCLUSIONS

In this document, I described an innovative approach to fuse information from different sensor modalities, like cameras and lidar, in order to probabilistically estimate a dense point cloud.

The approach achieves good performance in single image depth map prediction on the popular KITTI dataset. It is able to predict detailed depth maps on thin and distant objects. It also estimates reasonable depth in image parts in which there is no ground-truth available for supervised learning.

The qualitative and quantitative results presented here suggest that our system provides 2D depth maps of reasonable quality. Nevertheless, it is instructive to consider how the accuracy of our approach depends on the density of laser measurements and the number of superpixeles selected. The method proposed in this thesis works well even with sparse point clouds. The computational demand required to run the framework at the level of pixel is extremely high. However, there is room for improvement.

This method opens up an important avenue for research into multi sensor fusion and the more general 3D perception problems, which might benefit substantially from sparse depth samples.

9. APPLICATIONS AND FUTURE WORK

The automotive industry is rapidly evolving on the technology front and this growth is primarily attributed to changes in consumer preferences, and an equal boost from legislative bodies. Technological advances include improving vehicle performance, passenger safety, communication skills, and driving comfort, among others. The demand for safe and luxury vehicles has increased, so automakers have started to focus on improving road safety and accident prevention.

The framework proposed in this work enables safer driving and autonomous navigation by continuously monitoring the surrounding space, avoiding collisions by measuring the distance to objects with high confidence.

On the other hand, agriculture is probably one of the most traditional and longest existing trades. However, the sector is well advised to adopt innovative technologies and benefit from the opportunities offered by increasing automation. One of these advances is the autonomous operation of agricultural vehicles and precision agriculture. The objective of this last concept is to reduce expenditure and significantly increase yields. The model proposed in this thesis is timely. The system improve safety and productivity in the agriculture sector.

According to the qualitative results in outdoor environment, I consider the perception system proposed in this project, which use two sensory modalities such as LiDAR and Monocular camera, can significantly support agriculture in increasing yields and using land more efficiently. For example, this system could be used to identify obstacles and avoid them during scene mapping, or generate a 3D reconstruction of each fruit to be analyzed. The use of these technologies represents an important step in preparing the industry for the future.

REFERENCES

[1] R. Alterovitz, S. Koenig, and M. Likhachev, “Robot Planning in the Real World: Research Challenges and Opportunities”, AIMag, vol. 37, no. 2, pp. 76-84, Jul. 2016.

[2] M. Melin, A. C. Shapiro, and P. Glover-Kapfer, “LIDAR for ecology and conservation,” WWF Conserv. Technol. Ser., vol. 1, no. 3, p. 40, 2017, doi: 10.13140/RG.2.2.22352.76801.

[3] D. Scharstein, R. Szeliski and R. Zabih, .A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,"Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001), 2001, pp. 131-140, doi: 10.1109/SMBV.2001.988771.

[4] R. Hartley and A. Zisserman, Multiple View Geometry in Computer Vision, 2nd ed. Cambridge: Cambridge University Press, 2004.

[5] Rasshofer, R. H., Spies, M., and Spies, H., “Influences of weather phenomena on automotive laser radar systems”, Advances in Radio Science, vol. 9, pp. 49–60, 2011. doi:10.5194/ars-9-49-2011.

[6] S. Bargoti, J. P. Underwood, J. I. Nieto and S. Sukkarieh, .A pipeline for trunk detection in trellis structured apple orchards", Journal of Field Robotics, vol. 32, no. 8, pp. 1075-1094, 2015.

[7] D. Lahat, T. Adali, and C. Jutten, “Multimodal data fusion: an overview of methods, challenges, and prospects,” Proceedings of the IEEE, vol. 103, no. 9, pp. 1449–1477, 2015.

[8] V. DeSilva, J. Roche and A. Kondoz, "Fusion of LiDAR and camera sensor data for environment sensing in driverless vehiclesïn arXiv:1710.06230v2, 2018, [online] Available: https://arxiv.org/abs/1710.06230v2.

[9] Pandey, Gaurav. “An Information Theoretic Framework for Camera and Lidar Sensor Data Fusion and its Applications in Autonomous Navigation of Vehicles”, 2014.

[10] D. Ceylan, N. J. Mitra, Y. Zheng and M. Pauly, Çoupled structure-from-motion and 3d symmetry detection for urban facades", ACM Trans. Graph., vol. 33, no. 1, pp. 2:1-2:15, Feb. 2014.

[11] R. Benenson, "Perception for driverless vehicles: design and implementation", Ph.D. Thesis École Nationale Supérieure des Mines de Paris, no. 2008, [online] Available: https://pastel.archivesouvertes. fr/pastel-00005327.

[12] S. Petti, "Safe navigation within dynamic environments: a partial motion planning approach", Ph.D. dissertation, École des Mines de Paris, 2007

[13] K. Yamaguchi, D. McAllester and R. Urtasun, .Efficient joint segmentation occlusion labeling stereo and flow estimation", Proc. Eur. Conf. Comput. Vis., pp. 756-771, 2014.

[14] A. Vatavu, R. Danescu and S. Nedevschi, "Stereovision-Based Multiple Object Tracking in Traffic Scenarios Using Free-Form Obstacle Delimiters and Particle Filters,ïn IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 1, pp. 498-511, Feb. 2015, doi: 10.1109/TITS.2014.2366248.

[15] A. Petrovskaya and S. Thrun, "Model Based Vehicle Tracking for Autonomous Driving in Urban Environments", Proceedings of Robotics: Science and Systems, Jun. 2008.

[16] T. Vu, O. Aycard and F. Tango, .Object perception for intelligent vehicle applications: A multi-sensor fusion approach,"2014 IEEE Intelligent Vehicles Symposium Proceedings, 2014, pp. 774-780, doi: 10.1109/IVS.2014.6856588.

[17] R. Omar Chavez-Garcia, "Multiple Sensor Fusion for Detection Classication and Tracking of Moving Objects in Driving Environments", HAL, 2014.

[18] R. B. Rusu and S. Cousins, "3D is here: Point Cloud Library (PCL),"2011 IEEE International Conference on Robotics and Automation, 2011, pp. 1-4, doi: 10.1109/ICRA.2011.5980567.

[19] A. Knapitsch, J. Park, Q.-Y. Zhou and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction", ACM Trans. Graph., vol. 36, no. 4, 2017.

[20] A. Delaunoy and M. Pollefeys, "Photometric bundle adjustment for dense multiview 3d modeling", Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.

[21] Y. Furukawa and C. Hernandez, "Multi-View Stereo: A Tutorial", Foundations and Trends in Computer Graphics and Vision Series Now Publishers Incorporated, 2015.

[22] Q.-Y. Zhou, J. Park and V. Koltun, .^{O}pen3D: A modern library for 3D data processing", 2018.

[23] J. Park, Q.-Y. Zhou and V. Koltun, Çolored point cloud registration revisited", Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 143-152, 2017.

[24] G. Agamennoni, P. Furgale and R. Siegwart, "Self-tuning M-estimators,"2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 4628-4635, doi: 10.1109/ICRA.2015.7139840

[25] F. Castanedo, .^{A} review of data fusion techniques", Sci. World J., vol. 2013, 2013.

[26] P. Piniés, L. M. Paz and P. Newman, "Too much TV is bad: Dense reconstruction from sparse laser with non-convex regularisation,"2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 135-142, doi: 10.1109/ICRA.2015.7138991

[27] V. Romero-Cano and J. I. Nieto, "Stereo-based motion detection and tracking from a moving platform,"2013 IEEE Intelligent Vehicles Symposium (IV), 2013, pp. 499-504, doi: 10.1109/IVS.2013.6629517.

[28] M. Tanner, P. Piniés, L. M. Paz and P. Newman, "What lies behind: Recovering hidden shape in dense mapping,"2016 IEEE International Conference on Robotics and Automation (ICRA), 2016, pp. 979-986, doi: 10.1109/ICRA.2016.7487230.

[29] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition,"2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 681-687, doi: 10.1109/IROS.2015.7353446.

[30] J. Obando, V. Romero, N. Llanos, W. Toro. ProbabilisticPerception System for Object Classification Based on Camera -LiDAR Sensor Fusion. LatinX in AI Research at ICML 2019, Jun 2019, Long Beach, United States.

[31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods,ïn IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 11, pp. 2274-2282, Nov. 2012, doi: 10.1109/TPAMI.2012.120.

[32] S. van der Walt et al., "scikit-image: Image processing in Python", PeerJ, vol. 2, pp. e453, Jun. 2014, [online] Available: http://dx.doi.org/10.7717/peerj.453.

[33] A. Geiger, P. Lenz, C. Stiller and R. Urtasun, "Vision meets robotics: The KITTI dataset", Int. J. Robot. Res., vol. 32, no. 11, pp. 1231-1237, 2013.

[34] A KOUBAA, Robot operating system (ROS) [J]", Studies in computational intelligence, vol. 1, no. 6, pp. 342-348, 2016.

[35] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. B. Foote, J. Leibs, et al., -OS: An open-source robot operating system", Proc. ICRA Open-Source Softw. Workshop, 2009.

[36] M. Quigley, E. Berger and A. Ng, "Stair: Hardware and software architecture", Proc. AAAI Robot. Workshop, 2007.

[37] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, USA, MA, Cambridge:MIT Press, 2009.

[38] C. M. Bishop, Pattern Recognition and Machine Learning, New York, NY, USA:Springer, 2011..

[39] K. A. Stroud and D. J. Booth, Advanced Engineering Mathematics, New York, NY, USA:Palgrave Macmillan, 2003.

[40] G. Roussas, An Introduction to Probability and Statistical Inference, San Francisco, CA, USA:Academic, 2003.

[41] F. M. Dekking, A Modern Introduction to Probability and Statistics, U.K., London:Springer-Verlag, pp. 313-328, 2005.

[42] M. R. Spiegel, J. J. Schiller, R. A. Srinivasan and M. LeVan, Probability and Statistics, New York, NY, USA:McGraw-Hill, 2009.

[43] W. Mendenhall and R. J. Beaver. Introduction to Probability and Statistics-9th edition. Duxbury Press, International Thomson Publishing, Pacific Grove, CA, 1994.

[44] A First Course in Probability, New York:Collier Macmillan, 1976.

[45] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques, Cambridge, MA, USA:MIT Press, 2009.

[46] C.M. Bishop, Pattern Recognition and Machine Learning., Aug. 2006.

[47] J.E. Freund, Mathematical Statistics with Applications, pp. 524-529, 2004.

[48] P. Bessière, E. Mazer, J. M. Ahuactzin-Larios, and K. Mekhnacha, Bayesian Programming. CRC Press, Dec. 2013. [Online]. Available: http://hal.inria.fr/hal-00905797.

[49] S. Lauritzen, Graphical Models, New York:Clarendon, 1996.

[50] L. E. Sucar, "Probabilistic graphical modelsïn Advances in Computer Vision and Pattern Recognition, London, U.K.:Springer, pp. 978, 2015.

[51] K. P. Murphy, Machine learning: a probabilistic perspective. MIT, 2012

[52] C. R. Rao, E. J. Wegman and J. L. Solka, Handbook of Statistics, Wiley, vol. 24, 2005.

[53] P. Perez, "Markov Random Fields and Images", CWI Quarterly, vol. 11, no. 4, pp. 413-437, 1998.

[54] J. Lafferty, A. McCallum and F. Pereira, Çonditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", Proc. Int'l Conf. Machine Learning, pp. 282-289, 2001.

[55] X. He, R. S. Zemel and M. Carreira-Perpindn, "Multiscale conditional random fields for image labeling", Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., pp. 695-703, 2004..

[56] S. Kumar and M. Herbert, "Discriminative Fields for Modeling Spatial Dependencies in Natural Images", Proc. 18th Ann. Conf. Neural Information Processing Systems, 2004.

[57] P. Awasthi, A. Gagrani and B. Ravindran, Ïmage modeling using tree structured conditional random fields", Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), pp. 2060-2065, 2007.

[58] Ruo Zhang, Ping-Sing Tsai, J. E. Cryer and M. Shah, "Shape-from-shading: a survey,ïn IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 8, pp. 690-706, Aug. 1999, doi: 10.1109/34.784284.

[59] S. Suwajanakorn and C. Hernandez, "Depth from focus with your mobile phone", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[60] D. Eigen, C. Puhrsch and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network", Proc. Advances Neural Inf. Process. Syst.,

pp. 2366-2374, 2014.

[61] J. Xie, R. Girshick and A. Farhadi, "Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks", Proc. Eur. Conf. Comput. Vis., pp. 842-857, 2016.

[62] A. Saxena, S. H. Chung and A. Y. Ng, "Learning depth from single monocular images", Proc. Adv. Neural Inf. Process. Syst., pp. 1161-1168, 2005.

[63] D. Eigen, C. Puhrsch and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network", Adv. Neural Inf. Proc. Syst., 2014.

[64] D. Eigen and R. Fergus, "Predicting depth surface normals and semantic labels with a common multi-scale convolutional architecture", Proc. IEEE Int. Conf. Comput. Vis., pp. 2650-2658, 2015.

[65] A. Roy and S. Todorovic, "Monocular depth estimation using neural regression forest", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR (CVPR), 2016.

[66] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation", CVPR, 2018.

[67] B. Li, C. Shen, Y. Dai, A. V. den Hengel and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs", Proc. Conf. Computer Vision and Pattern Recognition (CVPR), pp. 1119-1127, 2015.

[68] A. Krizhevsky, I. Sutskever and G. Hinton, ÏmageNet Classification with Deep Convolutional Neural Networks", Proc. Neural Information and Processing Systems, 2012.

[69] F. Liu, C. Shen and G. Lin, "Deep convolutional neural fields for depth estimation from a single image", Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5162-5170, 2015.

[70] M. Liu, M. Salzmann and X. He, "Discrete-continuous depth estimation from a single image", Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 716-723, 2014.

[71] D. Eigen and R. Fergus, "Predicting depth surface normals and semantic labels with a common multi-scale convolutional architecture", 2014.

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

[73] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price and A. Yuille, "Towards unified depth and semantic prediction from a single image", Proc. IEEE Conf. Comput. Vis. Pattern Recog., pp. 2800-2809, 2015.

[74] C. Godard, O. Mac Aodha and G. J. Brostow, Ünsupervised monocular depth estimation with left-right consistency", Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[75] Y. Kuznietsov, J. Stückler and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6647-6655, 2017.

[76] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, “Monocular depthestimation based on deep learning: An overview,”Sci. China Technol.Sci., vol. 63, no. 9, pp. 1612–1627, Sep. 2020.

[77] F. Ma and S. Karaman, "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image,"2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 4796-4803, doi: 10.1109/ICRA.2018.8460184.

[78] Q. Yang, R. Yang, J. Davis and D. Nister, "Spatial-Depth Super Resolution for Range Images,"2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383211.

[79] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In NIPS, 2005. 1, 4, 5, 7.

[80] H. Andreasson, R. Triebel and A. Lilienthal, "Vision-based interpolation of 3d laser scans", Proceedings of the 2006 IEEE International Conference on Autonomous Robots and Agents (ICARA 2006), 2006.

[81] B. Wang, Y. Feng and H. Liu, "Multi-scale features fusion from sparse lidar data and single image for depth completion", Electronics Letters, 2018.

[82] F. Ma and S. Karaman, "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image,"2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 4796-4803, doi: 10.1109/ICRA.2018.8460184.

[83] F. Ma, G. V. Cavalheiro and S. Karaman, "Self-Supervised Sparse-to-Dense: Self-Supervised Depth Completion from LiDAR and Monocular Camera,"2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 3288-3295, doi: 10.1109/ICRA.2019.8793637.

[84] B. -U. Lee, H. -G. Jeon, S. Im and I. S. Kweon, "Depth Completion with Deep Geometry and Context Guidance,"2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 3281-3287, doi: 10.1109/ICRA.2019.8794161.

[85] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 175-185, 2018.

[86] W. Maddern and P. Newman, Real-time probabilistic fusion of sparse 3D LIDAR and dense stereo,"2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2016, pp. 2181-2188, doi: 10.1109/IROS.2016.7759342.

[87] K. Nickels, A. Castano and C. Cianci, "Fusion of lidar and stereo range for mobile robots", Int. Conf. on Advanced Robotics, 2003.

[88] A Harrison and P Newman Image and sparse laser fusion for dense scene reconstruction In Proc of the Int Conf on Field and Service Robotics (FSR), Cambridge, Massachusetts, July 2009.

[89] K. Park, S. Kim and K. Sohn, "High-Precision Depth Estimation with the 3D LiDAR and Stereo Fusion,"2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 2156-2163, doi: 10.1109/ICRA.2018.8461048.

[90] J. Lafferty, A. McCallum and F. Pereira, Çonditional random fields: Probabilistic models for segmenting and labeling sequence data", Proc. 18th Int. Conf. Mach. Learn., pp. 282-289, 2001.

[91] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang and H. Li, "Global ranking using continuous conditional random fields", Proc. Adv. Neural Inf. Process. Syst., pp. 1281-1288, 2008.

[92] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In NIPS, 2005.

[93] A Harrison and P Newman Image and sparse laser fusion for dense scene reconstruction In Proc of the Int Conf on Field and Service Robotics (FSR),

Cambridge, Massachusetts, July 2009.

[94] G. Sidorov, A. Gelbukh, H. Gómez-Adorno and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model", Computación y Sistemas, vol. 18, no. 3, pp. 491-504, 2014.

[95] S. Walz, T. Gruber, W. Ritter and K. Dietmayer, Üncertainty depth estimation with gated images for 3D reconstruction,"2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), 2020, pp. 1-8, doi: 10.1109/ITSC45102.2020.929457.

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

[97] A. Kendall and Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, 2017.

[98] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. arXiv preprint arXiv:1406.2283, 2014.

[99] Q. Yang, R. Yang, J. Davis and D. Nister, "Spatial-Depth Super Resolution for Range Images,"2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383211.

[100] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang and H. Li, "Global ranking using continuous conditional random fields", Proc. Adv. Neural Inf. Process. Syst., pp. 1281-1288, 2008.

[101] Y. Kuznietsov, J. Stuckler and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction", arXiv preprint arXiv:1702.02706, 2017.

[102] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, et al., "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image", CoRR, vol. abs/1812.00488, 2018.

[103] W. Van Gansbeke, D. Neven, B. De Brabandere and L. Van Gool, "Sparse and Noisy LiDAR Completion with RGB Guidance and Uncertainty,"2019 16th International Conference on Machine Vision Applications (MVA), 2019, pp. 1-6, doi: 10.23919/MVA.2019.8757939.

[104] C. Godard, O. Mac Aodha and G. J. Brostow, Ünsupervised monocular depth estimation with left-right consistency", Conference on Computer Vision and Pat-

tern Recognition (CVPR), 2017.

[105] N. Schneider, L. Schneider, P. Pinggera, U. Franke, M. Pollefeys and C. Stiller, "Semantically guided depth upsampling", German Conference on Pattern Recognition, pp. 37-48, 2016.

[106] X. Cheng, P. Wang and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network", arXiv preprint arXiv:1808.00150, 2018.

[107] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang and H. Li, "HMS-Net: Hierarchical Multi-Scale Sparsity-Invariant Network for Sparse Depth Completion,ïn IEEE Transactions on Image Processing, vol. 29, pp. 3429-3441, 2020, doi: 10.1109/TIP.2019.2960589.

[108] P. Hambarde and S. Murala, "S2DNet: Depth Estimation From Single Image and Sparse Samples,ïn IEEE Transactions on Computational Imaging, vol. 6, pp. 806-817, 2020, doi: 10.1109/TCI.2020.2981761.

[109] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari and N. Navab, "Deeper Depth Prediction with Fully Convolutional Residual Networks,"2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 239-248, doi: 10.1109/3DV.2016.32.

[110] C. Mallet and C. Bretar, "Full waveform topographic lidar: State-of-the-art", ISPRS J. Photogramm. Remote Sens., vol. 64, no. 1, pp. 1-16, Jan. 2009.

[111] R. Horaud, M. Hansard, G. Evangelidis and C. Ménier, .ᴬn overview of depth cameras and range scanners based on time-of-flight technologies", Mach. Vis. Appl., vol. 27, no. 7, pp. 1005-1020, Oct. 2016.

[112] F. Ma and S. Karaman, "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image,"2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 4796-4803, doi: 10.1109/ICRA.2018.8460184.

[113] C. Cadena, A. Dick and I. D. Reid, "Multi-modal auto-encoders as joint estimators for robotics scene understanding", Proc. Robot.: Sci. Syst. Conf., pp. 377-386, 2016.

[114] Y. Liao, L. Huang, Y. Wang, S. Kodagoda, Y. Yu and Y. Liu, "Parse geometry from a line: Monocular depth estimation with partial laser observation,"2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 5059-5066, doi: 10.1109/ICRA.2017.7989590.

[115] C. Fu, C. Mertz and J. M. Dolan, "LIDAR and Monocular Camera Fusion: On-road Depth Completion for Autonomous Driving,"2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019, pp. 273-278, doi: 10.1109/ITSC.2019.8917201.

[116] F. Ma and S. Karaman, "Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image,"2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 4796-4803, doi: 10.1109/ICRA.2018.8460184.

[117] X. Cheng, P. Wang and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network", arXiv preprint arXiv:1808.00150, 2018.

[118] P. Hambarde and S. Murala, "S2DNet: Depth Estimation From Single Image and Sparse Samples,ïn IEEE Transactions on Computational Imaging, vol. 6, pp. 806-817, 2020, doi: 10.1109/TCI.2020.2981761.

[119] B. Schoettle, "Sensor fusion: A comparison of sensing capabilities of human drivers and highly automated vehicles", Aug. 2017.

[120] A. Anas, "Methods of Point Cloud Alignment with Applications to 3D Indoor Mapping and Localization", Mar. 2017.