TECHNICAL UNIVERSITY OF CRETE

DOCTORAL DISSERTATION

Visual localization in unstructured environments through deep learning

---

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

by

Georgios Petrakis

Chania, October 2023

Examination Committee:

Panagiotis Partsinevelos (Supervisor)
Professor, Technical University of Crete

Michail Lagoudakis
Professor, Technical University of Crete

Georgios Chalkiadakis
Professor, Technical University of Crete

Stelios Mertikas
Professor, Technical University of Crete

Anastasios Doulamis
Associate Professor, Technical University of Athens

Nikolaos Geroliminis
Professor, Ecole Polytechnique Fédérale de Lausanne

Charalampos Ioannidis
Professor, Technical University of Athens

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development, Education and Lifelong Learning
Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

The implementation of the doctoral thesis was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities

Abstract

Scene understanding, localization and mapping, play a crucial role in computer vision, robotics and geomatics, providing valuable knowledge through a vast and increasing number of methodologies and applications. However, although the literature flourishes with related studies in urban and indoor environments, far fewer studies concentrate in unstructured environments.

The main goal of this dissertation is to design and develop a visual localization framework based on deep learning that aims to enhance scene understanding and the potential of autonomous navigation in challenging unstructured scenes and develop a precise positioning methodology, for characteristic point localization in GNSS-denied environments. The dissertation can be divided in five different parts: (a) design of the training and evaluation datasets, (b) implementation and improvement of a keypoint detection and description neural network for unstructured environments (c) implementation and development of a lightweight neural network for visual localization focused on unstructured environments and integration of the trained model in a SLAM (Simultaneous Localization and Mapping) system as a feature extraction module (d) development of a lightweight encoder-decoder architecture for lunar ground segmentation (e) development of a precise positioning and mapping alternative for GNSS-denied environments.

Regarding the first part of the dissertation, two datasets were designed and created for the training and evaluation of keypoint detectors and descriptors. The training dataset includes 48 000 of FPV (First-Person-View) images with wide range of variations in landscapes, including images from Earth, Moon and Mars while the evaluation dataset includes about 120 sequences of planetary-(like) scenes where each sequence contains the original image and five different generated representations of the same scene, in terms of illumination and viewpoint.

In the second part of this dissertation, a self-supervised neural network architecture called SuperPoint was implemented and modified, investigating its efficiency in keypoint detection and description applied in unstructured and planetary scenes. Three different SuperPoint models were produced: (a) an original SuperPoint model trained from scratch, (b) an original fine-tuned SuperPoint model, (c) an optimized SuperPoint model trained from scratch. The experimentation proved that the optimized SuperPoint model provides superior performance, compared with the original SuperPoint models and handcrafted keypoint detectors and descriptors.

Concerning the third part of the dissertation, a multi-task deep learning architecture is developed for keypoint detection and description, focused on poor-featured unstructured and planetary scenes with low or changing illumination while the training and evaluation processes were conducted using the proposed datasets. Moreover, the trained model was integrated in a visual SLAM (Simultaneous Localization and Maping) system as a feature extraction module, and tested in two

feature-poor unstructured areas. Regarding the results, the proposed architecture provides increased accuracy in terms of keypoint description, outperforming well-known handcrafted algorithms while the proposed SLAM achieved superior results in areas with medium and low illumination compared with the ORB-SLAM2 algorithm.

In the fourth part of the dissertation, a lightweight encoder-decoder neural network (NN) architecture is proposed for rover-based ground segmentation on the lunar surface. The proposed architecture is composed by a modified MobilenetV2 as encoder and a lightweight U-net decoder while the training and evaluation process were conducted using a publicly available synthetic dataset with lunar landscape images. The proposed model provides robust segmentation results, achieving similar accuracy with the original U-net and U-net-based architectures which are 110 - 140 times larger than the proposed architecture. This study, aims to contribute in lunar ground segmentation utilizing deep learning techniques, while it proves a significant potential in autonomous lunar navigation ensuring a safer and smoother navigation on the moon.

Regarding the fifth part of the dissertation, a precise positioning alternative was developed aiming to localize fiducial markers and characteristic points of the scene, providing their local coordinates in 3D space under a high level of accuracy. At first, the fiducial markers are placed in the scene where one of them is used as the origin marker, while the target markers represent the characteristic points or features. Subsequently, the proposed SLAM algorithm enables an RGB-Depth camera to map the desired area and localize itself in an unknown and challenging environment, while in combination with geometrical transformations, localization and optimization techniques, the present methodology estimates the coordinates of target markers and an arbitrary point cloud which approximates the structure of the environment.

It is clear that the use of deep learning in unstructured and planetary environments in terms of scene recognition, localization and mapping provides a significant potential for the future applications, reinforcing crucial topics such as autonomous navigation in hazardous and unknown environments. This dissertation aspires to encourage the investigation and development of AI models and datasets, focused on planetary exploration missions and especially on high and low-level scene understanding using computationally efficient equipment and methods, reducing the economic and energy costs of robotic systems.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Prof Panagiotis Partsinevelos, for his guidance, support, encouragement and cooperation during this challenging journey. He inspired me with his ideas and solutions to issues that I encountered while he was always available to discuss my considerations and research questions.

I would like to acknowledge the other members of my supervisory committee, Prof. Michail Lagoudakis and Prof. Georgios Chalkiadakis, for their support throughout my studies. Moreover, I sincere thank Prof. Stelios Mertikas, Prof. Anastasios Doulamis, Prof. Nikolaos Geroliminis and Prof Charalampos Ioannidis for accepting to participate as members of my examination committee.

Also, I would like to acknowledge the SenseLab laboratory for providing me all the equipment that I needed in my experiments. Moreover, I thank my colleagues Achilles Tripolitsiotis, Angelos Antonopoulos and Zisis Charokopos for their help during the experimentation.

Finally, I would like to sincere thank my family. My beloved wife Chrysa for her daily support and love and my treasures Konstantinos and Anastasia who help me to comprehend the meaning of my life.

Contents

List of Figures

List of Tables

List of Abbreviations

AI: Artificial Intelligence
CC: Creative Commons
CNN: Convolutional Neural Network
DEM: Digital Elevation Model
DL: Deep Learning
FPV: First-Person-View
GNSS: Global Navigation Satellite System
IMU: Inertial Measurement Unit
IoT: Internet of Things
MLC: Multi-line Convergence method
NN: Neural Network
PA: Plane Alignment
PCA: Principal Component Analysis
PPP: Precise Point Positioning
ReLU: Rectified-Linear-Unit
RGB: Red Green Blue
SLAM: Simultaneous Localization and Mapping
SVM: Support Vector Machine
VO: Visual Odometry

Dedicated to my beloved children: Konstantinos & Anastasia

Chapter 1

Introduction

A key factor in all mobile robotic systems is autonomous navigation, which enables the robotic system to sense, perceive, interpret its surroundings, act, and ultimately complete its mission without any human intervention or control (Bagnell *et al.* 2010). Autonomous navigation enhances the functionality of robotic systems, allowing them to perform demanding, hazardous, or even impossible tasks for humans in a wide range of environments, while also increasing safety in the workspace and reducing the risk of accidents and injuries (Cheng *et al.* 2018).

Scene understanding, localization and mapping constitute significant scientific tasks within autonomous navigation, incorporating numerous novel methodologies, applications, and services. These applications range from self-driving cars and search and rescue operations to industrial automation and inspection, and they are deployed in various environments, each presenting unique challenges. However, despite the abundance of literature on autonomous navigation in urban and indoor settings, there are notably far fewer studies that focus on unstructured environments (Guastella & Muscato 2021). As Brock *et al.* (2016) stated, “unstructured environments are the environments that have not been modified specifically to facilitate the execution of a task by a robot”.

Unstructured environments present unique challenges for autonomous robotic systems, especially for ground-based vehicles that must adapt and operate within uncertain and demanding conditions (Wang *et al*. 2017). Navigation and localization over uneven terrain in completely unknown or GNSS (Global Navigation Satellite System)-denied scenes, including planetary environments, while adapting to changes in illumination or weather conditions, can significantly reduce the effectiveness of most state-of-the-art algorithms, which may fail to provide robust and accurate results.

Furthermore, planetary landscapes feature sand dunes, large rocks, boulders, craters, and harsh surfaces, further complicating the safe navigation and operation of a rover. Several incidents during Mars exploration underscore the importance of effective rover navigation: The Opportunity rover was stuck within a sand dune on Meridiani Planum (Cowen 2005) for five weeks, the Spirit rover was trapped in soft soil in an area called "Troy," leading to the mission's termination in 2011, and Curiosity's wheels suffered damage due to the harsh Martian terrain.

Nonetheless, for nearly two decades, planetary rovers have continued to explore Mars, while the Artemis program's primary objective is to establish a human presence on the Moon, with the aim to further improve the required technology for upcoming space missions (Dunbar 2019).

Ground-based autonomous vehicles can also play a crucial role in other types of unstructured environments, such as construction sites, which are considered hazardous due to heavy machinery and the presence of moving construction vehicles. Therefore, autonomous navigation has the potential to enhance worker and equipment safety by automating tasks that are currently performed manually, including material handling, excavation, and site inspection (Xu et al. 2020).

However, despite the importance of autonomous navigation in unstructured environments, there is a scarcity of specialized studies focused on these terrains, along with a lack of datasets compared to urban and indoor environments (Schubert *et al*. 2018, Geiger *et al*. 2012, Burri *et al*. 2016, Huang *et al*. 2018). Hence, there is a need to explore and develop new approaches, methodologies, and datasets that emphasize the autonomous navigation capabilities in unstructured and planetary environments.

The motivation behind this dissertation stems from the need to address a significant gap in the literature concerning scene understanding, localization, and mapping in GNSS-denied, unstructured environments. Specifically, two main driving forces fuel this research: (i) the necessity to investigate autonomous navigation techniques in challenging terrains using specialized learning-based architectures and (ii) the need to investigate and develop precise localization techniques based on visual sensors in unstructured environments. Following these motivations, this dissertation proposes a visual localization framework tailored to unstructured environments, capable of providing both low and high levels of scene understanding and precise localization of characteristic points with high accuracy.

1.1 Objectives

The main objective of this dissertation is to design and develop a visual localization framework based on deep convolutional neural networks that aims to enhance scene understanding and the potential of autonomous navigation in challenging unstructured environments and develop a precise positioning methodology, for characteristic point localization in GNSS-denied environments.

The principal objective of this dissertation that is described above, can be divided in five major objectives which analyzed below.

1. Creation of datasets for training and evaluation of the learning-based architectures

The backbone of the learning-based visual localization framework is the training and evaluation datasets which are used to feed the deep learning architectures. However, these methods use image datasets which include scenes by urban, indoor, or vegetated environments (Weyand *et al.* 2020, Lin *et al.* 2014, Cao *et al.* 2021) while there is a lack of publicly available datasets for unstructured environments.

In this dissertation, a dataset specialized on unstructured environments focused on rocky and sandy scenes is created aiming to be used for the training and evaluation processes of deep learning architectures for the feature extraction. The dataset will contain about 50 000 images by planetary-like and real-planetary scenes.

Moreover, an evaluation dataset is designed inspired by the structure of HPatches dataset (Balntas *et al.* 2017), using about 120 representative images from rocky and sandy terrains. For each image, a sequence of five different image representations of the same scene will be generated, aiming to evaluate the proposed architectures in terms of illumination and viewpoint changes.

2. Implementation, development and improvement of deep learning models for accurate feature extraction in unstructured environments

Feature extraction which is composed by interest-keypoint detectors and descriptors, is a fundamental building block in low-level scene understanding and autonomous navigation while multiple conventional and learning-based approaches have been proposed, focused on urban or indoor environments (Xin *et al.* 2019). In this dissertation, an investigation of deep convolutional neural networks (CNNs) in feature extraction is conducted, while the selected architectures will be implemented and improved aiming to export characteristic features in unstructured environments with high accuracy and robustness.

Initially, the selected architectures will be explored in different training processes, parameterization, and fine-tuning, using the aforementioned proposed dataset, while afterwards the architectures will be modified aiming to improve their efficiency.

Subsequently, the models will be evaluated in terms of their robustness and accuracy in different lighting conditions and geometric transformations, while they will be compared with the initial pre-trained architectures and other well-known keypoint-detectors and descriptors, utilizing the proposed evaluation dataset.

3. Development of a visual SLAM (Simultaneous Localization and Mapping) algorithm focused on completely unknown and unstructured environments

Visual SLAM is a fundamental component of autonomous navigation since it allows a robotic system to build a map of its surroundings while simultaneously estimating its own position within that map using only camera sensors. One of the main components of a SLAM system is the feature extraction module, since it provides the external information from the observed environment.

However, the effectiveness of traditional SLAM systems, based on handcrafted feature extraction algorithms can be highly reduced in challenging environments with limited texture, repetitive patterns, or dynamic scenes. On the other hand, the SLAM systems which utilize learning-based approaches in feature extraction, are based on NNs which are focused on urban or indoor environments.

In this dissertation, a visual SLAM system is proposed in which an optimized, trained and fine-tuned NN for feature extraction, focused on unstructured environments, is integrated. Thus, during the SLAM process, the feature extraction will be performed by the integrated NN, increasing the robustness and accuracy in challenging unstructured environments, with lack of visual cues, illumination changes and limited texture. Regarding the evaluation of the proposed SLAM system, an extended experimentation will be conducted in order to estimate the accuracy of self-localization during the SLAM process.

4. Development of a learning-based precise positioning methodology in GNSS-denied environments

The precise positioning is the main research field of geodesy and topograhy while it is crucial in many applications and services including navigation, target tracking, search and rescue, inspection, etc, (Queralta *et al.* 2020). Conventional geodetic equipment provides high accuracy in positioning but requires extensive human effort in the field while GNSS signal coverage can be decreased or denied in several environments including heavy vegetated or intense rocky areas, construction sites or other planets (Trigkakis *et al.* 2020).

In this dissertation, a precise positioning methodology is developed, in order to localize fiducial markers in unknown scenes with a centimeter-level of accuracy. The methodology is based on the proposed SLAM system focused on unstructured environments while it utilizes an RGB camera with a depth sensor in order to calculate the scale of the features. After the mapping process, the methodology uses the extracted point cloud, the camera poses and the detected markers aiming to estimate the target locations while transforms the estimations in a generated local coordinate system. Regarding the testing process, the methodology will be evaluated using a geodetic total station which will provide the ground truth measurements.

5. Development of a deep learning model for scene understanding using semantic information in rocky unstructured environments

Semantic segmentation determines a significant potential in visual localization and autonomous navigation, allowing the machines to derive deeper insights from the scene, enhancing their ability to perceive and interpret their surroundings. Semantic information reinforces the identification and recognition of objects, scenes, and landmarks in pixel level, enriching the low-level information of mere visual cues to high-level scene understanding (Bowman *et al.* 2017).

However, one of the most challenging tasks of semantic segmentation architectures is their efficiency in performance-time which is inadequate for low-resources computing systems. Most of the semantic segmentation architectures includes a large number of parameters, increasing the required computing power, while lightweight architectures have to sacrifice the segmentation precision, affecting the overall quality of the results

(Mo *et al.* 2022). Although several studies investigate semantic segmentation for real-time processing, they are focused on self-driving cars applications in urban environments (Wang *et al.* 2019, Zhou *et al.* 2019, Zhang *et al.* 2022). Nevertheless, it is crucial for an autonomous robotic system in a rocky unstructured scene to detect and classify semantic features, especially the boulders and rocks which can harm itself and the attached equipment.

In this dissertation, an encoder-decoder semantic segmentation neural network will be implemented and improved aiming to interpret and recognize a scene in a rocky unstructured environment. Moreover, the proposed NN will be modified in order to increase its efficiency for real-time applications, without severe decrease of its accuracy. Due to the lack of publicly available datasets for unstructured environments, the proposed NN will have to be capable of training with limited-size of datasets.

The objectives described above are complementary and compose a framework of methodologies focused on scene understanding localization and mapping in unstructured environments. More specifically, the objective 2 is based on the objective 1 due to the designed datasets while the objective 3 utilizes the extracted models of objective 2. The objective 4 is based on the proposed SLAM of objective 3, while the objective 5 enriches the functionality of the framework increasing the level of scene understanding (fig. 1.1).

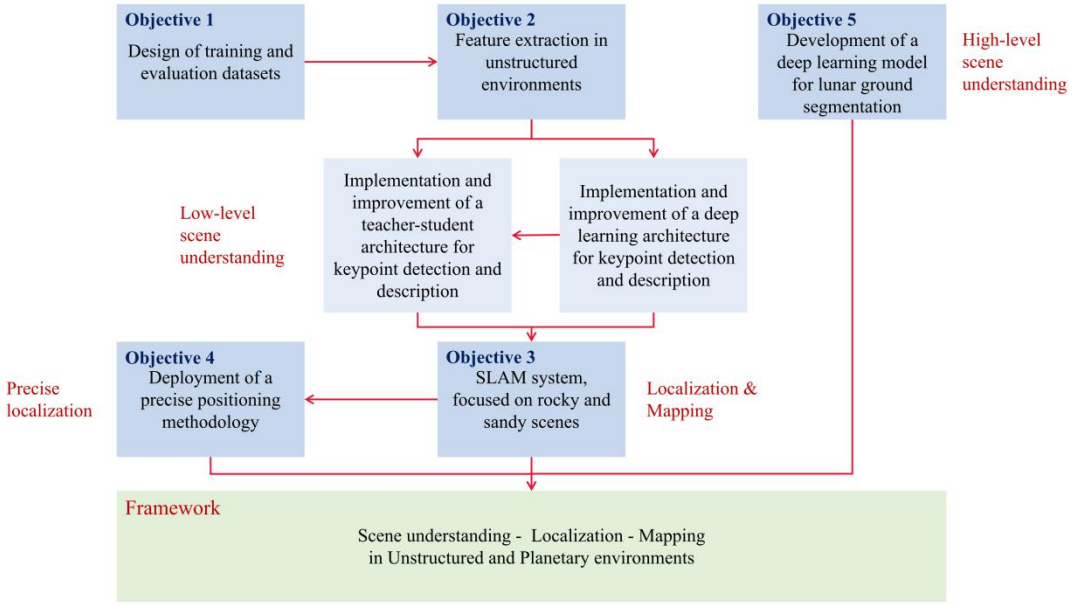


Figure 1.1 Structure of the dissertation objectives

1.2 Contribution and originality

In this dissertation, a visual localization framework is developed, including optimized techniques and novel methodologies for visual localization through deep learning, focused on unstructured and challenging environments. To the best of the author's knowledge, there is no similar framework which studies the potential of artificial

intelligence in autonomous navigation and localization techniques, focused on challenging, unstructured and GNSS-denied environments.

More specifically, the contributions of the dissertation are stated below:

- Investigation of feature extraction potential, in challenging and unstructured environments, including planetary scenes, using handcrafted and deep learning keypoint detectors and descriptors.
- Design of a training dataset for unknown and unstructured environments using FPV (First-Person-View) images from planetary environments.
- Implementation of an evaluation dataset for unknown and unstructured environments including sequences with original and transformed images, aiming to test the efficiency of the algorithms in terms of illumination and viewpoint. To the best of the author's knowledge, this evaluation dataset is the only publicly available dataset for testing handcrafted and deep learning-based algorithms in unstructured environments
- Implementation and optimization of a self-supervised CNN-based architecture for keypoint detection and description in unstructured environments.
- Implementation and development of a teacher-student CNN-based architecture for visual localization in unstructured environments.
- Implementation and development of a SLAM system, specialized in unstructured and challenging environments. It utilizes the aforementioned teacher-student CNN-based model as feature extraction module, in order to accurately interpret the external information in unstructured feature-poor scenes, with intense lighting changes.
- Implementation and development of a lightweight semantic segmentation architecture for high-level scene understanding in rocky environments. The model will provide high efficiency in performance-time with limited size of datasets without reducing the segmentation accuracy.
- Creation of a benchmark dataset for visual SLAM evaluation using sequences of RGB-depth data and the corresponding ground truth, in rocky unstructured environments
- Development of a precise positioning methodology for GNSS-denied environments which can be adapted and optimized in specialized environments through deep learning. The methodology combines a proposed deep learning-based SLAM algorithm with localization and geometric transformation techniques aiming to optimize the location of fiducial targets with high level of accuracy.
- The proposed precise positioning methodology utilizes computer vision within a topographic approach, providing a potential for autonomous accurate mapping in challenging and harsh environments.

1.3 General background

Intelligent machines and robotic systems, are able to communicate with humans and interact with the environment providing valuable knowledge (Alkendi *et al.,* 2021) through a vast and increasing number of methodologies and applications (Rubio et al. 2019). For instance, humanoid robots demonstrate a significant potential in manufacturing, education, retail, healthcare, companionship, etc (Appel *et al.* 2020) which notes a growing demand in several workspaces (Ajoudani *et al.* 2018). On the other hand, aerial robotic systems, provide significant value in search and rescue applications (Ajith & Jolly, 2020), precision agriculture (Velusamy *et al.* 2022), military surveillance (Gupta *et al* 2022), construction (Tatum & Liu, 2017) etc, while underwater vehicles play a vital role in marine engineering through hydrographic surveys, hull inspection, oil and gas exploration (Yang *et al.* 2021, Wynn *et al.* 2014). Self-driving cars are envisaged as the future of transportation gathering the research interest of academia and industry (Badue *et al.* 2021, Simoni *et al.* 2019) while planetary rovers are crucial for collecting data in space exploration missions (Zhang *et al.* 2019).

One of the most crucial components of mobile robotic systems is the self-localization, the ability of a machine to continuously track its position and orientation in the scene during a mission (Panigrahi & Bisoy, 2022). Self-localization typically relies on modules and sensors attached in the robotic system and can be divided in two main categories: (a) the GNSS (Global Navigation Satellite System)-based self-localization and (b) the sensor-based self-localization

Regarding the GNSS-based self-localization, GNSS receivers are electronic devices that are attached to a robotic system and through the timing of the received satellite signals, are able to triangulate and define the robotic system's position. Although the accuracy of a conventional GNSS receiver is limited, RTK (Real-time-kinematics) and PPP (Precise Point Positioning) methods are able to enhance the GNSS receiver, providing centimeter-level of accuracy (Huang *et al.* 2023). However, GNSS receivers include several constraints that can affect the performance of a robotic system. The main issue of a GNSS receiver is the signal degradation in dense urban or vegetated environments due to signal blockage or multipath effect, an effect that occurs when GNSS module receives signals at different times due to the signal reflection by different surfaces (Partsinevelos *et al.* 2020). Moreover, in some environments including covered-structures (e.g tunnels), indoor environments or other planets, GNSS signal does not exist.

On the other hand, the sensor-based self-localization techniques, utilize the attached sensors (Mohamed *et al.* 2019) and the movement of a robotic system in order to track its location, avoiding the dependence on external resources. These techniques are called odometry techniques in robotics terminology and include the visual odometry, wheel odometry, inertial odometry, laser odometry radar-based odometry, and sonar-ultrasonic odometry.

Visual odometry is able to analyze visual input using one or more on-board cameras, monitoring the movement of visual features over time, estimating the mobile robot's position and orientation. The wheel odometry, utilizes rotary encoders aiming to calculate the number of wheel's rotations while inertial odometry uses the IMU (Inertial Measurement Unit) sensor which contains gyroscopes and accelerometers for measuring the linear acceleration and angular velocity of the robotic system. On the other hand, laser, radar and sonar-ultrasonic odometry techniques, perform self-localization by tracking laser speckle patterns, radio and sound waves respectively, reflected by the surroundings (Mohamed *et al.* 2019). Moreover, hybrid odometry approaches have been proposed aiming to increase the efficiency of self-localization combining more odometry techniques and sensors including visual-laser odometry, visual-inertial odometry, visual-radar odometry, radar-inertial odometry etc (Huang *et al.* 2019, Usenko *et al* 2016, Doer & Trommer 2020, Mostafa *et al.* 2018).

Although, the odometry techniques provide valuable information about the position and orientation of a robotic system, several constraints are able to decrease their efficiency. For instance, visual odometry can be affected by feature-poor environments with low illumination. Wheel odometry provides decreased accuracy in slippery environments with / or uneven terrains due to wheel slippage (Mohamed *et al.* 2019) while the inertial odometry is affected by errors of the accelerometer and gyroscope measurements (Solin *et al* 2018). Laser odometry is unsuitable for robotic systems with limited computing resources due to large amounts of data processing generated by the LiDAR sensor (Aqel *et al.* 2016), radar odometry is affected by outliers and uneven terrains (Quist *et al.* 2016) while sonar odometry provides limited range compared with LiDAR or radar, reducing its efficiency in long distances (Burguera *et al.* 2007).

However, the main issue of odometry is the error accumulation over time which significantly affects the self-localization accuracy. This issue is encountered by SLAM (Simultaneous Localization and Mapping), a super-set of odometry that is able to estimate the robot's location constructing a dynamically generated local map simultaneously while a back-end optimization algorithm minimizes the errors between the predicted map features and the initially observed features. Moreover, a loop-closure module, detects a re-visit of a previously explored area performing further optimization of the whole scene, increasing the accuracy of mapping and camera trajectory estimations (Taketomi *et al.* 2017).

Nevertheless, the last decade, an evolution is observed in visual-based localization and mapping (Poddar *et al.* 2019) due to the modern image interpretation techniques and the cost-effective required equipment (e.g cameras, conventional computing systems). The main idea of localization and mapping using visual data is the extraction of distinct points from the camera frames which are called "keypoints". Keypoints are recognized by keypoint detectors through the analysis of local intensity changes aiming to detect edges and corners with increased contrast, compared with their local neighbourhood. Afterwards, a process called "keypoint description", establishes the keypoints aiming to be recognizable in the neighboring frames,

regardless of the viewpoint or scale changes while a matching process utilizes the keypoint locations and descriptions, matching the common features among the neighboring frames (Liu *et al.* 2021). The detection and description outputs, which compose the low-level scene understanding, aid the pose estimation algorithms to compute the position and orientation of the camera during its motion in the scene. However, several factors including intensive illumination changes, extreme weather conditions, occlusions or poor information in the scene, are able to reduce the robustness and accuracy of feature extraction (Naseer *et al.,* 2018).

Several studies investigate the use of deep learning in feature extraction and SLAM (Li *et al.* 2018, Arshad et al. 2021, Duan et al. 2019, Xiao et al. 2019, Li et al. 2018) aiming to increase the efficiency of self-localization processes. Convolutional neural network (CNN)-based architectures, are trained in order to detect interest-points which are included in edges, corners, shadows and color changes without the need of pixel-level image processing techniques (Kazerouni *et al.* 2022). Instead of traditional feature extraction algorithms, the CNN-based models are able to predict features with increased accuracy and robustness in challenging environments due to the following factors (Martins *et al.* 2021, Tang *et al.* 2019, Mokssit *et al.* 2023):

- Robustness in variations: CNN-based models are able to provide robust results in challenging environments and conditions with variations in scale, rotation, and illumination
- Multi-scale processing: CNN-based models are able to perform image processing in multiple scales, increasing the accuracy of keypoint detection in different image resolutions
- Specialized scenarios: CNN-based models can be re-trained or fine-tuned aiming to recognize features in specialized scenarios, conditions or environments.
- Flexibility in inference-time: deep learning models can be modified in order to reduce the inference-time, a crucial factor for embedded systems with limited resources.

Another significant use of deep learning is the semantic segmentation. Semantic segmentation models are trained to recognize semantic objects or features e.g. buildings, roads, trees etc, increasing the level of scene understanding which is crucial for autonomous navigation, especially in dynamic or completely unknown environments (Naseer *et al.* 2017). The predicted semantic features, can aid the robotic systems to navigate with increased efficiency and safety, since many features including trees, large rocks buildings etc that could harm a robotic system (Lai 2022) are localized and classified by the semantic segmentation model. Moreover, semantic segmentation reinforces the robustness of the scene understanding in cases of seasonal variations or extreme weather conditions which are able to deform the low-level information of the scene (Larsson *et al.* 2019) while is able to provide increased effectiveness in environments suffered by occlusions, identifying and recovering the occluded features (Qin *et al.* 2022).

1.4 Structure of the dissertation

The structure of this dissertation is as follows:

- In chapter 2, the literature review is analyzed and divided in three main sections (a) feature extraction in challenging environments (b) Scene understanding using semantic segmentation in unstructured environments and (c) Precise positioning and mapping in GNSS-denied environments
- Chapter 3, focuses on the methods and techniques that were developed and implemented for the proposed framework. At first, two approaches for visual localization in unstructured environments are presented: (a) a CNN-based self-supervised architecture for keypoint detection and description, and (b) a lightweight CNN-based teacher-student architecture which is able to extract keypoints, local and global descriptors while a SLAM algorithm which uses the aforementioned feature extractor is proposed. Subsequently, two datasets for unstructured environments are designed and utilized in order to be used in the training and evaluation processes. Then, a semantic segmentation architecture for planetary environments is proposed, capable of being used in systems with low computing resources, providing high efficiency with a limited size of dataset while finally, a precise positioning and mapping alternative in GNSS-denied environments is presented, designed for point positioning in feature-poor scenes with low illumination.
- In chapter 4, an extended description about the implementation and technical details of the proposed framework are analyzed, including the feature extraction approaches, the proposed SLAM, the semantic segmentation architecture and the precise positioning and mapping alternative, while afterwards the corresponding evaluation and results are presented.
- Chapter 5, discusses and analyzes the results while all the proposed architectures and algorithms are compared with the corresponding state-of-the-art implementations, highlighting the capabilities and limitations of the proposed framework.
- Finally, in chapter 6, the conclusions of each component of the proposed framework are presented while afterwards a summation about the achievement of the thesis' objectives is presented.

Chapter 2

Literature Review

In this section, the state-of-the-art approaches that compose the research fields of this dissertation, are analyzed with respect to three main pillars:

1. Feature extraction in challenging environments
2. Scene understanding using semantic segmentation in unstructured environments
3. Precise positioning and mapping in GNSS-denied environments

In the following sections of this chapter, the literature review and theoretical concepts of the three pillars above are critically discussed aiming to clearly describe the gap of knowledge and the research directions of this dissertation.

2.1 Feature extraction in challenging environments

Feature extraction from visual data, plays a critical role in computer vision and robotics, offering valuable low-level information for real-time self-localization and scene understanding. The first step of feature extraction is the keypoint detection which refers to distinctive and invariant image locations that represent important features including unique patterns, corners, edges etc while subsequently, the keypoint description encodes each detected keypoint's relevant information, enabling the robust matching among the neighboring frames and scene recognition (Liu *et al.* 2021). Feature extraction process is the backbone of many advanced computer vision applications and tasks including autonomous navigation and 3D reconstruction, with several handcrafted algorithms and deep learning architectures proposed in the literature.

2.1.1 Handcrafted keypoint detectors and descriptors

Handcrafted keypoint detectors rely on designed filters or mathematical operations that are based on gradient-based or intensity-based techniques, while attempt to maintain their reliability in scale, rotation, and viewpoint changes (Isık *et al.* 2015). There are several widely used keypoint detectors including Harris (Harris & Stephens 1988), Shi-Tomasi (Shi & Tomasi 1993), FAST (Rosten & Drummond 2006), and AKAZE (Alcantarilla *et al.* 2013) and keypoint descriptors such as ORB (Rublee *et al.* 2011), SIFT (Lowe 2004) and SURF (Bay *et al.* 2008).

Harris (Harris & Stephens 1988) is a widely used algorithm for identifying corners in an image. It aims to localize regions with significant intensity changes in different

directions, which are indicative of corner points. The algorithm begins by calculating the gradient of the image while afterwards computes the autocorrelation matrix for each pixel by convolving the image gradients with a Gaussian window. From the autocorrelation matrix, the algorithm derives a corner response function by taking into account the eigenvalues of the matrix. High eigenvalues indicate corners, while low eigenvalues correspond to edges or flat regions (fig 2.1). The corners are then identified by applying a threshold to the corner response values or by selecting local maxima in the response map.

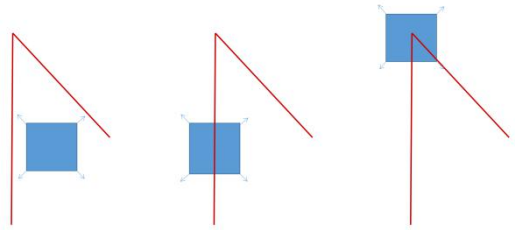


Figure 2.1 (left) flat region (blue square): Region without change in all directions, (middle) edge: No change along the edge direction, (right) corner: change in all directions

The Shi-Tomasi (Shi & Tomasi 1993) detector, is an extension of the Harris corner detector, which instead of using the corner response function based on eigenvalues, it selects keypoints based on a score computed from the smallest eigenvalue of the autocorrelation matrix. This modification allows the algorithm to prioritize the detection of sharper corners, ranking the keypoints based on their scores and selects the most important, up to a specified number or based on a minimum threshold.

FAST (Features from Accelerated Segment Test) (Rosten & Drummond 2006) algorithm is a popular keypoint detector known for its computational efficiency. It aims to identify keypoints by comparing the pixel intensities in a circular neighborhood (Fig 2.2). The algorithm examines a set of contiguous pixels and evaluates whether a pixel within the set is significantly brighter or darker than the central pixel. By finding a contiguous set of at least 12 such consecutive pixels, the algorithm can classify the central pixel as a keypoint. This rapid classification process makes the FAST detector well-suited for real-time applications.

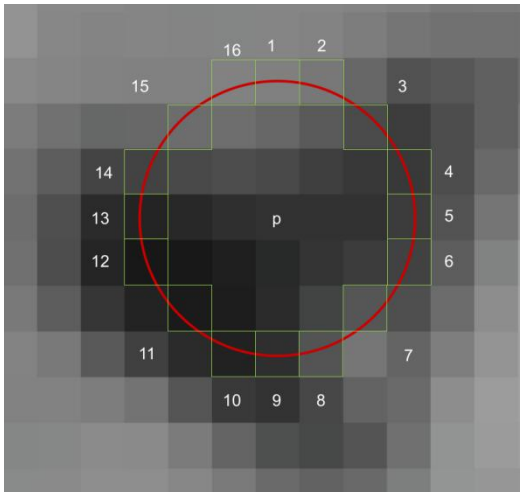


Figure 2.2 The circular neighborhood that the FAST algorithm uses. The highlighted pixels (in green squares) are the set of contiguous pixels which are evaluated and compared with the central pixel (p) in order to classify a corner

The AKAZE (Accelerated-KAZE) (Alcantarilla et al. 2013) algorithm is a keypoint detector and descriptor, designed to be robust to various image transformations, including rotation, scale changes, and affine transformations. It operates by extracting multiscale nonlinear diffusion filtering responses from the input image, capturing the image's structural information (e.g edges) across different scales. Subsequently, AKAZE computes the gradient magnitude and orientation using the nonlinear scale space representation, utilized to generate a binary descriptor, which encodes the local features around each keypoint. The descriptor is designed to be highly distinctive and robust to image variations, allowing for accurate matching of keypoints.

ORB (Oriented FAST and Rotated BRIEF) (Rublee *et al.* 2011) combines the efficiency of the FAST keypoint detector with the robustness of the BRIEF (Binary Robust Independent Elementary Features) (Calonder et al. 2010) descriptor. Initially, ORB builds a pyramid which is a multi-scale representation of a single image while subsequently identifies keypoints using the FAST algorithm, comparing the intensities of pixels in a circular neighborhood. Afterwards, it computes a binary feature vector for each keypoint using the BRIEF descriptor, which encodes the relative intensities of pixel pairs. To enhance invariance in rotations, ORB additionally calculates an orientation for each keypoint based on the intensity distribution around it, used to reinforce the rotational invariance of BRIEF descriptor. ORB has gained popularity due to its balance between speed and accuracy, making it suitable for real-time applications.

SIFT (Scale-Invariant Feature Transform) (Lowe 2004) is a widely used keypoint descriptor. Initially, SIFT constructs a scale-space representation of the input image by applying Gaussian blurring at multiple scales, while subsequently it locates potential keypoints as local maxima in images filtered by difference-of-Gaussian (DoG), highlighting regions with significant changes in terms of intensity across scales. The descriptor part, generates a total of 128 bin values for each keypoint, by extracting a 16x16 pixel neighborhood around each identified feature and subsequently subdividing the region into blocks. Although SIFT is a quite accurate

keypoint descriptor with increased robustness in scale, rotation, and affine transformations, it requires high computational cost.

SURF (Bay *et al.* 2008) relies on scale-space analysis for keypoint detection in different scales while employs the concept of box filters which are approximations of second-order Gaussian derivatives aiming to accelerate the computation of image features. It utilizes integral images to efficiently calculate the sum of pixel intensities within rectangular regions allowing for fast feature extraction and matching. Regarding the keypoint description, it encodes the local image information using gradient orientations and magnitudes. SURF is designed to be both distinctive and invariant to changes in scale, rotation, and affine transformations while provides low computational cost.

The handcrafted keypoint detectors and descriptors encounter significant tasks of computer vision with respectable accuracy, however they are quite susceptible and error prone in high level of noise, complex backgrounds, image artifacts or low level of illumination, which are able to decrease their efficiency (Liu *et al.* 2021).

2.1.2 Keypoint detectors and descriptors based on deep learning

Before the advent of deep learning (DL), several studies investigated the use of machine learning techniques such as SVM (Support Vector Machine), PCA (Principal Component Analysis) and decision tree in handcrafted algorithms, aiming to improve the feature extraction accuracy (Ke *et al.* 2004, Rosten *et al.* 2006, Ma *et al.* 2021). However, the last decade the research interest has been shifted in the feature extraction architectures based on deep learning and more specifically, CNNs.

CNN-based feature extraction architectures are often composed by a detection and description part in order to extract both keypoint predictions and their corresponding descriptions. Initially, DL architectures creates response maps aiming to detect interest points while subsequently learn representations of each keypoint using either local patches centered on the predicted keypoints or the entire image using the pixel-level keypoint locations (Ma *et al.* 2021).

In keypoint detection and description process, the DL architectures can be divided in three main categories in terms of the learning process:

- Supervised learning
- Self-supervised learning
- Unsupervised learning

In supervised learning, the DL architecture uses annotated images during the training process in order the model to learn feature extraction in a scene or object while in unsupervised learning, the architecture utilizes only the non-annotated image dataset and several transformation techniques constructing a learning process, based on the comparison between the original and transformed images. On the other hand, in self-supervised learning, initially, the DL architecture predicts the feature maps from the

training data using the weights from a previous training while afterwards, it utilizes the predicted feature maps as ground truth for the main training process.

LIFT (Learned Invariant Feature Transform) (Yi *et al.* 2016) is a DL architecture for keypoint detection and description which is based on supervised learning. It is composed of three CNN-based components: The detector, the orientation estimator and the descriptor while it uses image patches in order to feed the architecture instead of the entire image, due to scalability in the learning process without information losses. Regarding the training process, the architecture is not trained with an end-to-end manner because of different aspects that the individual components try to optimize. Instead, the descriptor is trained first, then the orientation estimator and finally the detector based on the weights of the previous training processes. The ground truth which the LIFT uses for the training process, are generated by a Structure-from-Motion (SfM) algorithm which is performed on images captured under various illumination conditions and viewpoints.

LF-net (Ono *et al.* 2018) is another approach of a supervised DL architecture for keypoint detection and description (fig. 2.3). It is composed of two main components: The first component is a fully CNN which predicts local keypoint locations (x_i, y_i) combined with the corresponding scale (s_i) and orientation (θ_i) on entire images while simultaneously, it uses an optimization technique called "softargmax" for sub-pixel accuracy in keypoint detection. The K most important interest points including locations, scale map and orientation map, combined with image patches cropped around of the selected keypoints using a differentiable sampler (STN), feed the second component which is a CNN-based descriptor. The descriptor is composed of three convolutional filters followed by two fully connected layers while the final output (D_i) is normalized using the L-2 norm.

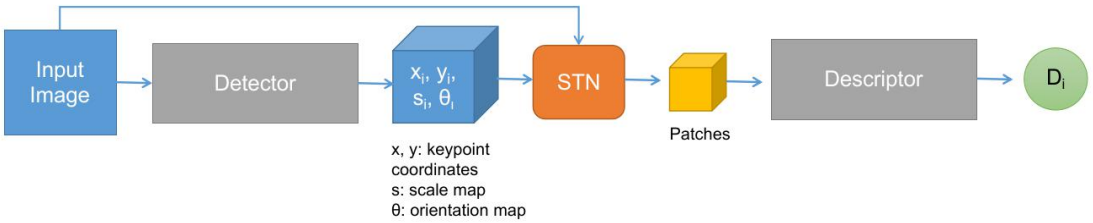


Figure 2.3 LF-net end-to-end architecture

Superpoint (DeTone *et al.* 2018) is a fully-CNN for keypoint detection and description trained in a self-supervized manner operating on full-sized images. Regarding its architecture, Superpoint is composed of a shared encoder which is based on VGG (Simonyan & Zisserman 2015) NN reducing the input image dimensions and two decoder branches where the first one learns to detect interest keypoints and the second one, the corresponding descriptions. The training process of Superpoint can be divided in three different stages:

- Interest-point pre-training: In this stage, the detector part of SuperPoint, called MagicPoint, is trained using labeled synthetic data which are simple geometric shapes with well-defined keypoints, representing the labels of the images.
- Interest-point self-training: The detector is trained using the weights from the pre-training process using unlabeled real-world images. In this stage, a technique called "homographic adaptation" transforms the input images in terms of scale and viewpoint aiming to train and optimize the detector to automatically label real-world images.
- Joint training: In this stage, the SuperPoint architecture is trained using the labels extracted by the trained MagicPoint model of the previous step, aiming to predict both interest points and the corresponding descriptions in real-world images.

It's worth noting that Superpoint is quite efficient in illumination and viewpoint changes. Detailed information about SuperPoint is referred in chapter 3.

Unsuperpoint (Christiansen *et al.* 2019), is a CNN-based architecture inspired by Superpoint, for keypoint detection and description which is trained in an unsupervised manner operating on entire images. Similar to Superpoint, UnsuperPoint is composed of a shared encoder based on VGG neural network which provides a downsampled feature map while three decoders focus on learning to extract keypoint locations, the corresponding scores and descriptors respectively, for each image. Instead of supervised and self-supervised-based NNs, Unsuperpoint doesn't require ground-truth based on SfM algorithm such as LIFT or LF-net nor automatic labeling using a pre-trained detector as SuperPoint. Unsuperpoint utilizes a siamese network of Unsuperpoint twins (twin A and twin B) where the first one (twin A) processes original images and the second one (twin B), the corresponding warped images transformed by a transformation matrix T. Afterwards, the twin A transforms the detected keypoints by the transformation matrix T and the loss functions track the keypoint correspondences from both twins (fig 2.4). Although Unsuperpoint seems a reliable and straightforward solution of feature extraction, it has not evaluated in real-time applications.

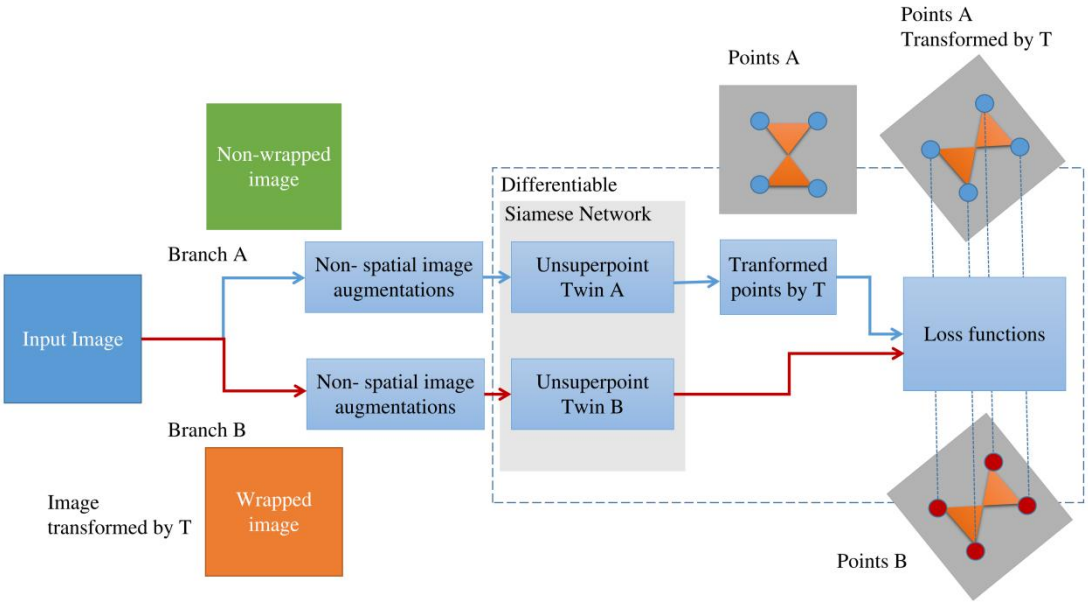


Figure 2.4 Unsupervised learning process of Unsuperpoint

HF-net (Hierarchical Feature network) (Sarlin *et al.* 2019), is a lightweight CNN-based architecture focused on visual localization operating in full-sized images. HF-net architecture comprises a shared encoder and three sub-modules which focus on keypoint detection scores, local and global description. The shared encoder processes the input image using a pre-trained MobilenetV2 (Sandler *et al.* 2018) NN, a quite popular architecture designed for mobile devices, while the global descriptor branch is composed by a NetVLAD layer (Arandjelovic *et al.* 2016), located on top of the last output of the shared encoder. The local descriptor sub-module uses a decoder of Superpoint, which undertakes the extraction of keypoint scores and local descriptors. Regarding the training process, the HF-net is based on a teacher-student architecture in order to reduce the complexity of an end-to-end multi-task NN during training. More specifically, HF-net utilizes two NN as teachers during the training process, feeding the main HF-net architecture with the corresponding teachers' predictions which represent the ground-truth. Superpoint is the teacher for keypoint detection and local description while NetVLAD is the teacher for global description. The teacher-student HF-net architecture is presented in the following figure:

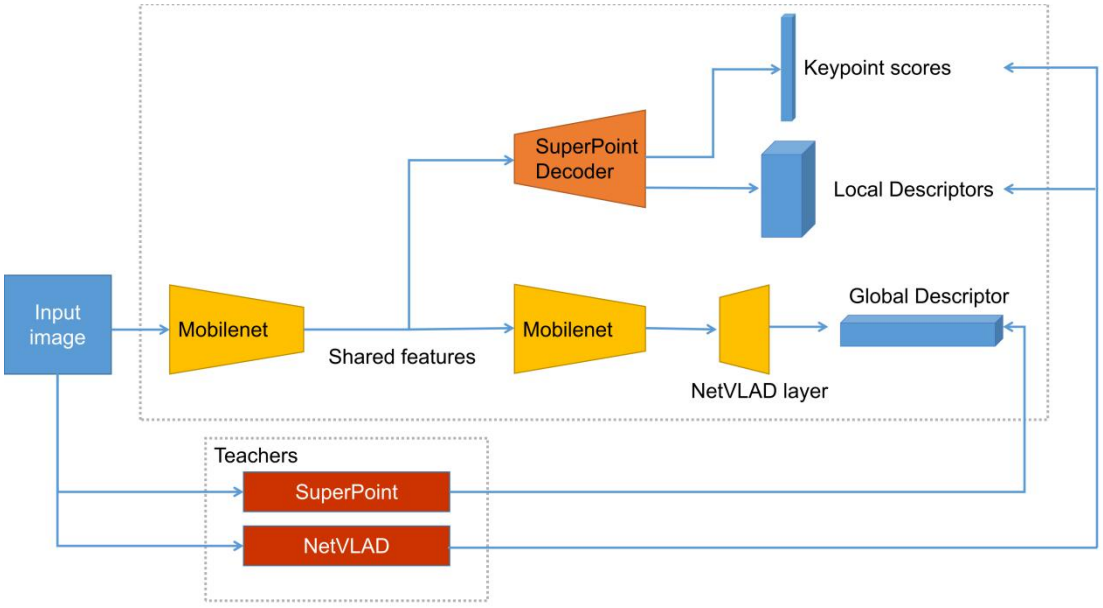


Figure 2.5 HF-net architecture

2.1.3 Feature extraction and SLAM in challenging environments

The handcrafted feature extraction algorithms have been widely investigated and used in multiple applications of computer vision and photogrammetry for more than three decades, while recently several studies utilize the DL-based architectures aiming to optimize more complicated tasks such as visual localization and autonomous navigation.

The literature abounds with studies which focus on autonomous driving (Lategahn *et al* 2011, Singandhupe *et al.* 2019), indoor navigation (Zou *et al.* 2022), 3D reconstruction (Inzerillo *et al* 2018) or inspection (Jordan *et al.* 2018) in urban and industrial scenes. However, there are far fewer studies which concentrate in autonomous navigation-based techniques applied on unstructured environments including rocky, vegetated and underwater scenes or even on planetary environments such as Mars and Moon.

Regarding the Earth-based environments, in Aulinas *et al.* 2011, a visual feature extraction methodology is proposed for underwater environments aiming to improve the subsea scene understanding using an optical sensor. The authors propose a methodology which is based on a SLAM algorithm in order to retrieve 3D spatial information, combined with a semantic segmentation technique which uses traditional image processing algorithms for contextual object identification (e.g rock) and the SIFT (Lowe 2004) and SURF (Bay et al. 2008) algorithms for feature extraction on the segmented images. In Jung et al. 2022, a technique for coloring 3D point clouds using visual data in subsea environments is proposed, aiming to reinforce the handcrafted 3D feature extraction algorithms, since visual data is suffered by low illumination and noise in subsea scenes. In Guo *et al.* 2018, a methodology for place recognition using LiDAR intensity is presented, tested in large-scale vegetated scenes. Authors utilize a 3D local descriptor called ISHOT (Intensity Signature of Histograms of Orientations) aiming to match features in a pre-built 3D LiDAR-based map, while a probabilistic place voting technique aid to bring out the most likely place candidate, from the global database in the scene.

Concerning the planetary-based scenes, several studies investigate feature extraction methodologies in extraterrestrial terrains using conventional algorithms. In Oelsch *et al.* 2017 and Wan *et al.* 2017, an evaluation of handcrafted feature extraction algorithms, in Mars-like environment is presented. In Oelsch *et al.* 2017, authors compare the algorithms' performance in terms of location recognition, using Devon Island dataset (Furgale *et al.* 2012), concluding that SURF (Bay *et al.* 2008) achieves the highest accuracy in non-vegetated and rocky unstructured environments. On the other hand, Wan *et al.* 2017 evaluates the efficiency of the algorithms in terms of several metrics including repeatability and precision, using simulated images generated with the aid of the DEM (Digital Elevation Model) and DOM (Digital Orthophoto Map) of a Mars region, proving that SIFT (Lowe 2004) scores the highest overall efficiency. In Wu *et al.* 2018, an improved version of SIFT (Lowe 2004) for high-resolution remote sensing images from Mars and Moon is proposed, aiming to reinforce the invariance of SIFT in differences due to illumination. Initially, authors apply feature extraction using SIFT while afterwards a Gaussian suppression function is used to evenly distribute the histogram which is highly biased due to the solar azimuth angle and finally, the suppression function is performed to the extracted descriptors. The improved-SIFT technique. provides 40 - 60% increased accuracy, based on the total number of correct matches.

Several studies investigate methodologies aiming to improve the autonomous navigation in feature-poor environments such as planetary scenes. In Otsu *et al.* 2018, authors attempt to solve this issue presenting a system called VOSAP (VO-aware sampling-based planner) which explores the rich-featured paths which are available in the scene, achieving increased performance in localization accuracy, tested in simulated Mars-like surfaces. In Kostavelis *et al* 2014, a complete vision system for autonomous rover-based Mars exploration is proposed aiming to increase the accuracy in terms of performance-time and self-localization. The system is based on two stereo cameras, for navigation and localization while the SURF (Bay *et al.* 2008) algorithm is selected for feature extraction and matching. Although authors managed to highly increase the performance-time of the system, accumulated errors are expected in visual odometry and localization for long-distances. In Giubilato *et al.* 2021, a stereo SLAM system, focused on the loop-closure refinement using elevation information of the terrain in poor-featured environments is presented, using a technique called Gaussian Process Gradient Maps. The authors use Moon and Mars-analogous terrains for the experimentation of the system while compare it with state-of-the-art SLAM algorithms including ORB-SLAM2 (Mur-Artal & Tardos 2017) and VINS-MONO (Qin *et al.* 2017) determining higher efficiency especially in loop closure. In Hong *et al.* 2021, a stereo SLAM system for highly detailed 3D point-cloud mapping in planetary environments is proposed. The authors combine traditional front-end and back-end SLAM components in order to produce a sparse map, with a self-supervized deep learning architecture which generates disparity maps aiming to dense the 3D scene information.

Datasets are crucial for visual localization, SLAM and autonomous navigation, since are used for evaluation of feature extraction algorithms such as HPatches (Balntas *et al.* 2017), for evaluation of SLAM systems such as TUM (Schubert *et al.* 2018), KITTI (Geiger *et al.* 2012), EuRoC (Burri *et al.* 2016), or for training in case of deep learning-based architectures including COCO (Lin *et al.* 2014), Berkeley Deep Drive (Fisher *et al.* 2018), Google Landmarks (Noh *et al.* 2017), etc. However, datasets which present completely unknown, vegetation-free and unstructured environments are quite few compared with datasets which was captured in urban and indoor scenes. In Driver *et al.* 2023, authors propose a dataset with remote sensing images for training deep learning architectures to extract keypoints and descriptors in small celestial bodies aiming in autonomous navigation for the future spacecrafts. Concerning the FPV (first-person-view) or rover-based datasets, most of these studies, focus on SLAM evaluation designing datasets in Moon and Mars-like environments using cameras, IMUs and LiDAR data combined with GNSS-based ground truth (Meyer *et al.* 2021, Furgale *et al.* 2012, Giubilato *et al* 2022, Hewitt *et al.* 2018). However, to the best of our knowledge there is no publicly available dataset for training deep learning architectures for keypoint detection and description in FPV images, nor a feature extraction evaluation dataset.

In this study, two different approaches for feature extraction based on deep learning in unstructured and challenging environments are proposed:

- A Keypoint detection and description model architecture based on SuperPoint (DeTone *et al.* 2018), trained and evaluated with two proposed training and benchmark datasets for unstructured environments and planetary scenes,
- A Lightweight teacher-student architecture for feature extraction, trained and evaluated using two proposed training and benchmark datasets for unstructured environments and planetary scenes while a SLAM system which uses the aforementioned deep learning model as a feature extractor, is proposed and tested in planetary-like environments.

Regarding the first approach, a SuperPoint neural network (DeTone, 2018) is implemented, optimized and trained in order to accurately conduct feature extraction in unstructured environments, focused on rocky and sandy scenes. For the training process, a dataset of 48 000 images was utilized (Petrakis & Partsinevelos 2023) representing unstructured and planetary scenes from Earth, Moon and Mars. Concerning images from Earth, they were captured from construction sites, mountainous areas, sandy beaches and a quarry, while the images from Mars were collected by a publicly available dataset of NASA which includes rover-based images, captured by Mars Science Laboratory (MSL, Curiosity) rover. Regarding the lunar dataset, it dataset includes artificial images, created by Keio University in Japan. Concerning the learning process, the standalone part of SuperPoint detector, was trained in three phases, one time with synthetic data and two times using the aforementioned dataset enabling homographic adaptation, a technique to increase the efficiency of the architecture in geometric transformations. Finally, the SuperPoint neural network was trained based on the weights of the standalone detector in order to fine-tune the keypoint detector and train the descriptor. Three different models were produced using the aforementioned dataset: (a) an original SuperPoint model, trained from scratch, (b) an original fine-tuned SuperPoint model, (c) an optimized model, trained from scratch. The models were evaluated using a benchmark dataset (Petrakis & Partsinevelos, 2023), designed for unstructured environments including earthy and planetary scenes, aiming to test the accuracy in illumination and viewpoint changes. The experimentation proves that the optimized SuperPoint model provides satisfactory results in keypoint detection and description, compared with the original SuperPoint and popular handcrafted detectors and descriptors.

Concerning the second approach, although several studies propose feature extraction and SLAM-based techniques in unstructured environments, there is no study that investigates the potential of a deep learning in keypoint detection and description, focused and fine-tuned for unstructured environments or planetary scenes. In this study, a teacher-student CNN-based architecture is developed for keypoint detection and description in unstructured and challenging environments with feature-poor scenes and low or changing illumination. The proposed architecture was trained using the aforementioned dataset including images from Earth, Mars and Moon while for the evaluation process, the aforementioned benchmark dataset (Petrakis & Partsinevelos, 2023) was utilized testing the model in terms of illumination and viewpoint changes. Moreover, the trained model, was integrated in a visual SLAM system as a feature extraction module, aiming to investigate the potential of a deep

learning-based SLAM system, focused on unstructured and planetary scenes. The proposed architecture and SLAM system provide high accuracy and superior results compared with several well-known algorithms. The main contributions of this study can be described as follows:

- An investigation of the feature extraction potential through deep learning in unstructured environments was conducted
- A training and evaluation dataset for keypoint detection and description focused on unstructured and planetary scenes were designed and implemented. To the best of the author's knowledge, the benchmark dataset is the only publicly available dataset for testing handcrafted and deep learning-based algorithms in unstructured environments
- A deep learning model for keypoint detection and description focused on unstructured and challenging environments was developed
- A visual SLAM that is aware of unstructured and planetary scenes using the proposed deep learning model was implemented.

2.2 Scene understanding using semantic segmentation in unstructured environments

Semantic segmentation is a powerful technique for scene understanding, that enables machines to comprehend and interpret visual information in a more detailed and meaningful way. By dividing an image into several regions and assigning semantic labels to each pixel, semantic segmentation algorithms or architectures, can accurately identify and classify different objects and structures within the scene. Semantic segmentation is a distinct task that differs from similar computer vision tasks such as classification and object detection. Classification involves assigning a single label or category to an entire input image or object, aiming to identify the presence of a specific class within the image but does not provide any information about the location or extent of the object. On the other hand, object detection is not only limited in classifying the objects with single labels but also it localizes their positions within the image simultaneously. Object detection algorithms encompass the objects using bounding boxes, indicating their spatial extent, allowing for the detection and classification of multiple objects within a single image. Semantic segmentation, aims to classify and assign a class label to each pixel in an image, effectively segmenting the image into different regions based on their semantic meaning. Instead of providing bounding boxes, semantic segmentation provides a pixel-level understanding of the scene, enabling precise localization and detailed segmentation of objects and their boundaries. This level of understanding goes beyond traditional image recognition, as machines are able to comprehend the contextual relationships and spatial layout of the scene. (Guo *et. al* 2018). Contemporary semantic segmentation architectures, provide a significant potential for even more intelligent and sophisticated systems that can interact with and comprehend the world around them, reinforcing applications such as autonomous driving, surveillance systems, augmented reality, under-water and planet exploration, etc (Garcia *et al.* 2017*)*. Pixel-wise semantic or image segmentation can

be performed through traditional image processing or machine learning algorithms while the last decade, deep learning architectures have shown a huge potential in this field (Hao *et al* 2020).

2.2.1 Traditional and machine learning -based segmentation approaches

Traditional algorithms primarily rely on handcrafted features (e.g. edges, corners) and conventional image processing techniques to extract relevant information and classify pixels into different semantic classes. Most of the traditional segmentation algorithms are based on pixel color using several color spaces such as RGB, YcBcr, HSL, HSI, YIQ e.t.c (Cohen *et al.* 2015, Kasson & Plouffe 1992) and on texture analysis, utilizing algorithms including histogram of oriented gradients (HoG) (Dalal &Triggs 2005), canny edge detector (Canny 1986), laplacian of gaussian (LoG), BoW (Bag of Words) (Csurka *et al.* 2004) etc.

On the other hand, supervized machine learning-based approaches are able to classify each pixel in an image into specific semantic features. Initially, relevant features, based on color, texture, shape, or other visual attributes, are extracted by the input images while subsequently, the extracted features are utilized as input to a classification algorithm, such as decision trees (Quinlan 1986), random forests (Breiman 2001), support vector machines (SVM) (Cortes & Vapnik 1995) etc. Afterwards, during the training process, the algorithms using a dataset with pairs of original and annotated images, create a classification model that can accurately assign class labels to pixels in unknown images.

Significant role in traditional supervised semantic segmentation is the unsupervized segmentation which composes a complementary approach, aiming to gather refined information about the classes and their consistency. Some of the most widely used unsupervized segmentation algorithms in the literature are the k-means clustering (Hartigan *et al.* 1979), random walker (Grady 2006), active contour models (Kass *et al.* 1988) and watershed segmentation (Roerdink & Meijster 2000). Although unsupervized segmentation algorithms don't identify semantic features, however, they are utilized to investigate the potential of the data and the desired classes before performing supervized semantic-segmentation.

2.2.2 Deep learning-based segmentation approaches

Deep learning-based architectures, enables a superior scene understanding, compared with the traditional and machine learning approaches, predicting more accurate and robust pixel-level segmentation maps (Sehar & Naseem 2022). Deep learning architectures, particularly convolutional neural networks (CNNs), have revolutionized semantic segmentation by automatically learning intricate features and capturing complex spatial relationships leveraging their ability to extract hierarchical features from raw image data, allowing them to accurately distinguish between different object

classes and delineate their boundaries. The training process involves feeding large datasets with pairs of original and annotated images into a CNN-based neural network, enabling it to learn and generalize patterns effectively (Lateef & Ruichek 2019).

The tricky part of the CNN-based semantic segmentation models, is the spatial awareness and accurate localization of semantic features' outline, due to dimensionality reduction and max-pooling operations that are performed by a typical CNN model. This issue is commonly encountered by an encoder-decoder architecture, utilized in an end-to-end training process, in order to combine the capturing of low and high-level contextual information in the image content with the corresponding regions of the desired semantic features. The encoder component is responsible for capturing hierarchical representations of the input image, gradually extracting high-level features through a series of convolutional and pooling layers, encoding essential information about the image's content. On the other hand, the decoder component, utilizes these encoded features to generate a segmentation map that matches the spatial dimensions of the original input. The decoder, applies transposed convolutions or upsampling operations aiming to reconstruct the spatial details and refine the segmentation map (Ye & Sung 2019) while at the end, an activation function (e.g softmax) exports the predicted classes (Banerjee *et al.* 2020) (fig 2.6).

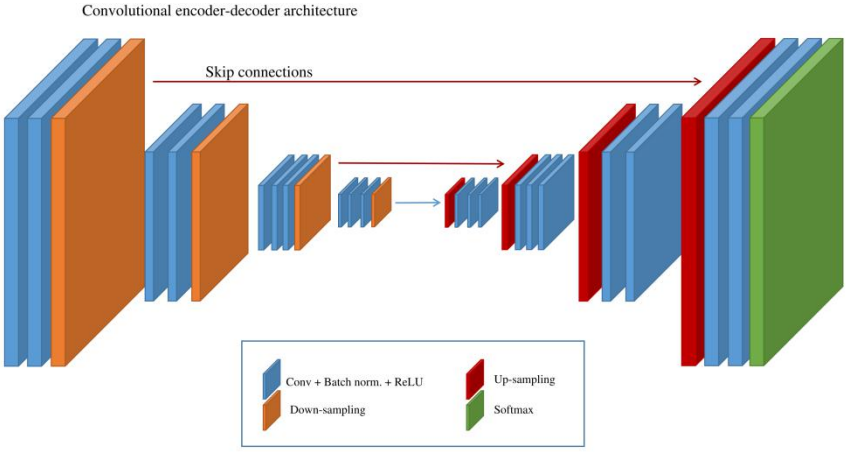


Figure 2.6 Typical example of an encoder-decoder architecture

There are several widely used encoder-decoder architectures for semantic segmentation including FCN (Fully Convolutinoal Network) (Long *et al.* 2015), U-net (Ronneberger *et al.* 2015), Segnet (Badrinarayanan *et al.* 2015), PSPNet (Zhao *et al.* 2017), DeeplabV3+ (Chen *et al.* 2018), etc.

FCN (Long *et al.* 2015), is a pioneering architecture which introduced the concept of fully convolutional networks for semantic segmentation, replacing the fully connected layers with convolutional layers, allowing to utilize image datasets of arbitrary sizes. The encoder extracts features through convolutional layers and downsampling techniques, while afterwards the decoder upsamples the feature maps to the original image size. Skip connections reinforce the communication between the encoder and

decoder aiming the extracted by the encoder features to correlate with the corresponding spatial information, providing each semantic feature with its segmentation boundary.

U-Net (Ronneberger *et al.* 2015) is a fully convolutional architecture for semantic-segmentation which initially proposed for biomedical image segmentation, however it has been widely used in a wide range of segmentation tasks. It is composed by an encoder-decoder architecture with skip connections preserving spatial information with high accuracy. The encoder captures contextual information and learns high-level representations, while the decoder recovers spatial information and generates the segmentation mask. U-Net is able to provide satisfactory accuracy after training with small training datasets, a fact that makes U-net suitable for applications with limited size of datasets.

SegNet (Badrinarayanan *et al.* 2015), is a lightweight encoder-decoder architecture for semantic segmentation which maintains the spatial information of the encoder, by saving pooling indices in the max-pooling layers, performing efficient upsampling in the decoder. Although SegNet provides decreased accuracy in capturing feature's boundaries compared with other segmentation architectures such as U-net, it can efficiently upsample feature maps without requiring additional parameters. In other words, Segnet is lighter and requires less memory than U-net which instead of pooling indices, it uses the entire feature maps in order to perform upsampling operations in the decoder part.

PSPNet (Zhao *et al.* 2017) is composed of three main parts: a convolutional neural network which extracts a tuning number of feature maps, a "pyramid pooling module" which reinforces the accurate multi-scale contextual information detection and an upsampling module, which reconstructs the segmented image, using bi-linear interpolation or transposed convolution methods. The pyramid pooling module aggregates information at different scales by pooling feature maps with different kernel sizes, allowing PSPNet to have a global semantic understanding, preserving detailed information.

DeepLabV3+ (Chen *et al.* 2018) is the latest version of the DeepLab family and is composed of an encoder-decoder architectures aiming to perform image segmentation with efficiency and accuracy. In the encoder part, DeepLabV3+ utilizes a technique called "separable atrous convolution" which separate the process of convolution in two main components: the "depthwise convolution" and "pointwise convolution" aiming to apply and combine different filters to each input channel. Regarding the decoder, it is composed of upsampling operations in order to reconstruct the segmented images achieving accurate semantic features' localization.

In summary, semantic segmentation based on deep learning proves a significant potential compared with traditional and machine learning algorithms in several fields including, medical, remote sensing, self-driving cars, etc, FCN introduced the encoder-decoder architecture while U-net utilizes sophisticated skip connections in

order to predict semantic features' boundaries with high accuracy without the need of large datasets. On the other hand, Segnet focuses on high inference time due to the lightweight "pooling indices" technique, while PSPNet and DeepLabV3+ introduced several new methods for accurate multi-scale segmentation such as "pyramid pooling module" and "separable atrous convolution" respectively.

2.2.3 Semantic segmentation in unstructured environments

Semantic segmentation plays a significant role in unstructured and planetary scene understanding, offering to a robotic system or a planetary rover valuable knowledge about its surroundings (Swan *et al.* 2021). Through terrain semantic segmentation, robotic systems are able to analyze images or videos and accurately detect and classify multiple features or regions within their environments, allowing superior comprehension and spatial awareness. More specifically, robotic systems are capable of identifying and differentiating various elements including boulders, rocks, craters, soil, or even potential obstacles and hazards, reinforcing the accurate path planning and enabling the robotic system to navigate in challenging landscapes with safety. Moreover, accurate semantic segmentation is able to identify potential mineral deposits or geological formations, contributing to scientific research for planet exploration.

Several studies investigate semantic segmentation in unstructured and planetary scenes using traditional algorithms without learning-based processes including George *et al.* 2000, Howard & Seraji 2001, Gong & Liu 2012, Di *et al.* 2013, and machine learning algorithms such as Song & Shan 2006, Dunlop *et al.* 2007, Fujita & Ichimura 2011 and Lu & Oij 2017. However, the last five years, terrain semantic segmentation based on deep neural networks dominates the literature (Kuang *et al.* 2022).

Regarding the earthy unstructured scenes, in Baheti *et al.* 2020, authors propose a semantic segmentation methodology focused on self-driving in unstructured environments, using a modified DeepLabV3+ (Chen *et al.* 2018) neural network with dilated Xception (Chollet 2016) as a backbone network while Baheti *et al.* 2020, proposes a U-net neural network which utilizes the EfficientNet (Mingxing & Le 2019) as the encoder. Both models were trained and evaluated with IDD (Indian Driving dataset) dataset due to its high diversity achieving satisfactory results using mIoU (mean Intersection over Union) metric. In Guan *et al.* 2022, authors propose a lightweight neural network for terrain semantic segmentation focused on unstructured environments which is capable of merging multi-scale visual features, in order to efficiently group and classify different types of terrains while a reinforcement learning algorithm, is able to utilize the predicted segmentation maps aiming to plan and guide a robot in paths with high safety. Similarly, in Guan *et al.* 2021, a real-time terrain mapping method for autonomous excavators is presented, which is able to provide semantic and geometric information for the terrain using RGB images and 3D point cloud data, while a dataset which includes images from construction sites is designed and utilized. Regarding the datasets for earthy unstructured environments, in

Metzger *et al.* 2021 and Wigness *et al.* 2019, two publicly available datasets were developed for semantic segmentation deep learning models, focusing on self-driving in semi-unstructured or dense-vegetated environments. In Metzger *et al.* 2021, the dataset designed, for accurate comprehension in scenes with high coverage in grass, asphalt, soil and sand, while authors in Wigness *et al.* 2019, targeted more on dense-vegetated and rough terrain scenes for off-road self-driving scenarios.

Concerning the planetary environments, several methodologies have been proposed for feature detection and terrain or scene segmentation aiming to reinforce and improve planet exploration tasks including landing, rover-based path planning, localization or planet surface investigation. In Furlan et al. 2019, a modified-U-net architecture (Ronneberger et al. 2015) for rock segmentation on the martian surface is proposed, which was trained and tested with a Mars-like dataset (Furgale et al. 2012) captured on Devon Island, achieving satisfactory accuracy while in Furlan et al. 2020, authors conduct a performance evaluation in rock detection for Mars-like environments using an original and modified versions of SSD (Single-Shot-Detector) (Liu *et al.* 2016) neural network, trained with the aforementioned dataset (Furgale et al. 2012). In Kuang *et al.* 2021, a modified Unet++ architecture (Zhou *et al.* 2018) for rock segmentation in planetary-like environments is proposed where two rounds of training are performed for the learning process. In the pre-training stage, the proposed architecture is fed by a synthetic dataset, created by a proposed algorithm while in the fine-tuning stage, the architecture is trained using a limited part of the Katwijk beach planetary rover dataset (Hewitt *et al.* 2018). In Tomita *et al.* 2020 and Claudet *et al.* 2022, authors conduct a benchmark analysis in Hazard Detection (HD) for planetary landing using several state-of-the-art semantic segmentation models compared with a replicated HD algorithm from NASA's Autonomous Landing Hazard Avoidance Technology (ALHAT) project while for the training process, realistic and noisy DEMs (Digital Elevation Models) with hazardous features were generated. The results proved that the segmentation architectures provide high efficiency on hazard detection outperforming the ALHAT algorithm in performance time and accuracy.

The sky and ground segmentation in planetary environments is investigated by several studies since it is able to refine the scene understanding (Kuang *et al.* 2021, Ebadi *et al.* 2022). In Kuang *et al.* 2021, an architecture for sky and ground segmentation for planetary scenes is proposed using a neural network inspired by U-net (Ronneberger et al. 2015) and NiN (Network In Network) (Lin *et al.* 2013), trained for two rounds with SkyFinder (Mihail *et al.* 2016) and Katwijk beach planetary rover datasets (Hewitt *et al.* 2018) respectively, while in Ebadi *et al.* 2022, skyline contour identification in Martian environment is conducted, using DeeplabV3+ (Chen *et al.* 2018) for sky and ground segmentation aiming to estimate the rover's global position.

A significant limitation of deep learning methods in planetary environments, is the lack of qualitative available datasets real or even synthetic, compared with datasets for urban or indoor environments which abound in the literature (Müller et al. 2021). In Müller *et al.* 2021, authors propose a simulator which is able to construct valuable synthetic scenes for planetary environments including rich metadata while

furthermore it is capable of generating multi-level semantic labels based on pre-defined materials. On the other hand, in Swan *et al.* 2021, authors propose a large-scale dataset called AI4MARS for terrain semantic segmentation of Mars, aiming to reinforce autonomous navigation on the martian surface. AI4Mars includes about 35K annotated images captured by Curiosity, Opportunity and Spirit rovers while the labeling conducted by experts with the aid of crowdsourcing using a web-based annotation tool. The proposed dataset was utilized to train DeeplabV3 neural network, proving that is a valuable dataset for Martian terrain semantic segmentation.

A crucial use of terrain classification in planetary environments is the path planning optimization (Chiodini *et al.* 2020). In Huang *et al.* 2021, a deep learning model for terrain segmentation is proposed using PSPNet (Zhao *et al.* 2017) model, trained by real rover-based images from Mars and artificial images generated by the Unity3D software, aiming to automate a path planning algorithm on the Martian surface while in Chiodini *et al.* 2021, authors propose a methodology for path rerouting using imagery data, depth maps and a CNN-based neural network trained with Katwijk beach planetary rover dataset (Hewitt *et al.* 2018), aiming to detect and avoid obstacles such as rocks and boulders.

Although several studies investigate rover-based scene recognition in Martian surface or planetary environments in general, quite few investigate similar tasks for the lunar surface. Lunar topography includes several features including rocks, boulders and craters, while the terrain in many areas is quite uneven with mounds and valleys. Although several studies propose methodologies for crater (Jia *et al.* 2020, Hashimoto *et al.* 2019, Hu *et al.* 2021) or hazard (Moghe & Zanetti 2020) detection and segmentation, they use remote sensing images (not rover-based) and focus on safe landing while there is a deficiency in rock and boulder identification during the rover navigation; a quite important issue for the smooth and trouble-free navigation.

In this study, a lightweight encoder-decoder neural network (NN) architecture is proposed for rover-based ground segmentation on the lunar surface. The proposed architecture is based on U-net (Ronneberger *et al.* 2015) and MobilenetV2 (Sandler *et al.* 2018) while the training and evaluation process were conducted using a synthetic dataset with lunar landscape images. The proposed model provides robust results, achieving similar accuracy with original U-net and U-net-based architectures which are 100 - 150 times larger than the proposed architecture. This study aims to contribute in lunar ground segmentation utilizing deep learning techniques, while it proves a significant potential in autonomous lunar navigation ensuring a safer and smoother navigation on the moon. To the best of our knowledge, this is the first study that proposes a lightweight semantic segmentation architecture for the lunar surface, focused on rover navigation.

2.3 Precise positioning and mapping in GNSS-denied environments

Geospatial technology, based on higher geodesy, remote sensing and geographical information science has change the way that scientists, engineers and citizens study or interact with their environment. This fact results in fundamental advances in various topics of geomatics such as location-based applications, spatial data infrastructures, navigation or geodetic equipment. Nevertheless, conventional surveying, although is the most accurate and robust method of applied geodesy, it remains a time-consuming process with significant human effort (Carrera-Hernández *et al.,* 2020). On the other hand, GNSS (Global Navigation Satellite System) positioning method, provides unsatisfactory accuracy in several cases including dense urban and high vegetated areas or remote unstructured environments due to the degraded GNSS signal coverage (Chiang *et al*., 2019) while in planetary exploration missions GNSS doesn't exist.

More specifically, urban, vegetated or rocky areas pose challenges to precise GNSS positioning because of signal interference, multipath effect or line of sight occlusion, factors which do not necessarily decrease over time during the measurement (Bastos *et al.* 2013). Typically, even a satellite signal blockage of short duration can significantly degrade performance in navigation systems. On the other hand, non-GNSS surveying alternatives, including total stations and laser scanners, involve time consuming practices in the field and/or costly equipment. There are cases where typical surveying cannot be substituted at the moment from GNSS, while in other cases classic surveying remains impractical. In between, there is an unresolved set of circumstances, where the need of cost-effective rapid mapping in GPS-denied environments remains crucial (Trigkakis *et al.* 2020).

Several studies attempt to improve the GNSS signal in challenging environments with respect to independent system analysis (Panigrahi *et al.,* 2015) while other methodologies propose techniques including angle approximation (Tang *et al.,* 2015), shadow matching (Urzua *et al.,* 2017), multipath estimations using 3D models (Zahran *et al,* 2018) and statistical models (Romero-Ramirez *et al.,* 2018, Partsinevelos *et al.,* 2020).

Alternative solutions to this problem result from methods which work on improving GNSS positioning performance by introducing information from other modalities. In (Panigrahi *et al.,* 2015), the authors use a dead reckoning sensor, which is a spatial sensor comprising of 3-axis gyros, 3-axis accelerometers, 3-axis magnetometers, temperature and barometric altimeter to extrapolate a trajectory when there is no GNSS signal. Similarly, Kim & Sukkarieh 2005, integrate a simultaneous localization and mapping (SLAM) system to a GNSS/ Inertial Navigation System (INS) fusion filter. INS calculates the position of a moving object by dead reckoning without external references. In Javanmardi *et al.* 2017 and Jende *et al.* 2018, authors use post and real-time processing of high-resolution aerial images utilizing statistical techniques in order to georeference mobile mapping data extracted from GNSS-denied areas, while in Heng et al. 2019, a platform for autonomous vehicle is proposed, which performs real-time 3d mapping without the need of GNSS using a

360-degree camera system, multi-view geometry and fully convolutional neural networks. In Bobbe *et al.* 2017, authors combine an aerial (UAV) and a ground (computing processing) platform to georeference aerial data extracted by online visual SLAM based on ORB-SLAM2 algorithm utilizing photogrammetric techniques while authors in Mostafa *et al.* 2018, use visual odometry and extended Kalman filters in order to improve on inertial measurement unit (IMU) measurements. Xu *et al.* 2019, developed an ORB-SLAM method for real-time locating systems in indoor GNSS-denied environments in order to detect characteristic points achieving an accuracy which ranges from 0.39–0.18 m aided by a depth sensor to acquire the scale information of the scene. Tang et al. 2015 claim SLAM performance in featureless environments is poor and opt for a method without SLAM while in Munguía *et al.* 2016, a visual-based SLAM system for navigation using a UAV, a monocular camera, an orientation sensor (AHRS) and a position sensor (GNSS receiver) is proposed. The system performs SLAM processing for navigation but it fuses GNSS measurements during initialization period to estimate the metric scale of the scene.

Several studies have been conducted for accurate and / or rapid mapping with the use of mobile mapping systems (MMS), Photogrammetry and image processing techniques. In Kalacska *et al*. 2020, authors follow the approach of Structure-from-Motion (SfM) with multi-view stereo technique of Photogrammetry to produce ortho-images and 3D surfaces without the use of ground control points (GCPs) using UAVs equipped with GNSS receivers and optical sensors. In Pinto & Matos 2020, dense 3D information in underwater environments is constructed through the fusion of multiple light stripe range (LSR) and photometric stereo (PS) methods outperforming the corresponding conventional methods in terms of accuracy while in Bañón *et al.* 2019, aerial images and ground control points (GCPs) are used in order to produce a 3D model in a coastal region through SfM. The characteristic points are measured using a GNSS receiver for the validation of the methodology with a vertical RMSE error of 0.12 meters. Tomaštík J. *et al*. 2017, evaluate the positional accuracy of forest rapid - mapping, using point cloud data created by UAV images and the Agisoft software with an accuracy level of 20 cm.

Various studies are referred to localization and detection methods employing MMS equipped with stereo sensors. Haque *et al.* 2020, propose an unmanned aerial system (UAS) which is able to find its location in a 3D CAD model of a pre-defined environment. The UAS with a stereo-depth camera, maps the area using ORB-SLAM2 algorithm, detects and extracts vector features with the aid of a convolutional neural network (CNN) and rectifies its location comparing the SLAM mapping area with the 3D CAD model. In Li *et al.* 2017, authors propose a pose estimation methodology based on mobile accelerometers, visual markers and stereo vision fusion, achieving a centimeter level of accuracy while in (Vrba & Saska 2020; Vrba *et al.,* 2019), a methodology that detects a micro aerial vehicle (MAV) is proposed, utilizing machine learning techniques and an RGB-Stereo depth camera with an average RMS error of 2.86 meters. In Zhang C. et al, 2019, a real-time obstacle avoidance method is developed with the aid of a stereo camera, a GNSS receiver and an embedded system mounted on a UAV in order to detect obstacles and follow an alternative, obstacle-

free path. In Ma *et al.* 2021, authors utilize a UAV with two cameras and a GNSS receiver in order to detect and geographically localize insulators in power transmission lines based on the bounding box of the detected insulators. Moreover, depth-based cameras have been used in UAVs for autonomous landing in GNSS-denied environments, where a UAV is able to detect, locate and land on an unmanned ground vehicle (UGV) making use of information from a multi-camera system and deep learning algorithms (Yang *et al.,* 2018; Animesh *et al.,* 2019).

As referred above, the literature abounds of positioning methodologies for GNSS-denied areas, rapid mapping solutions using photogrammetric techniques or localization systems based on SLAM and detection. Although most of the studies propose alternative localization solutions, none of them focus on surveying or traditional topography combined with computer vision and multi-view geometry algorithms.

In this study, a cost-effective, rapid and efficient surveying solution for GNSS-denied and challenging environments is proposed where a RGB-Depth sensor and a computing system are enough to map an area of interest. The methodology combines the following three proposed methods aiming to localize specific points which are represented by specialized fiducial markers in the scene:

- A SLAM system based on deep learning, focused on environments with poor-featured information and intense illumination changes
- A localization method called multi-line convergence method (MLC) and
- An optimization method called Plane alignment (PA)

The study is an extended and improved approach of two previous studies (Trigkakis *et al,* 2020, Petrakis *et al,* 2023) which utilized the MLC and PA methods with a traditional SLAM system in order to localize the desired fiducial markers. More specifically, in Trigkakis *et. al,* 2020, an implementation based on SLAM, point cloud and image processing techniques, localizes characteristic points in a local coordinate system using a monocular camera in combination with a fiducial marker. Although the main issue of the monocular setup approaches is the scale estimation (Sahoo *et al.,* 2021), the study controlled this issue by using the MLC method achieving an accuracy level of 50 cm. In Petrakis *et al,* 2023, the aforementioned methodology was extended using a stereo camera instead of a single sensor and validated, conducting various indoor and outdoor experiments on dense and sparse urban and vegetated scenes. In Petrakis *et al,* 2023, the methodology is capable of mapping an unknown area, providing refined estimations of point coordinates in a local 3D coordinate system fusing stereo SLAM, image processing techniques and coordinate system transformations, increasing the accuracy level in about 10 cm.

In this study, the methodology was further extended aiming to be used in challenging environments with poor information in visual cues and features or/and intense illumination changes including unstructured scenes. To achieve this goal, a proposed SLAM system based on a lightweight neural network is utilized, where combined with the aforementioned MLC and PA methods, is able to provide robust results and

maintain satisfactory accuracy in challenging and unstructured GNSS-denied environments. The main improvement of this study compared with the two previous approaches is that can be specialized in specific environments and scenes by re-training and fine-tuning of the neural network for each specific scene, an extremely useful approach, since it is potentially capable utilizing in several environments including glacial scenes, dense-canopy areas, factory facilities, dense-urban environment etc, maintaining the rapid mapping and robust results.

To the best of our knowledge, there is no similar solution that makes use of a visual SLAM algorithm based on deep learning, an RGB-depth camera and a fiducial marker in order to provide a 3D local coordinate system in challenging environments with high level of accuracy. Unlike the similar localization methods, the coordinate estimations were not extracted in a software-based reference system but in a reference system which is well-defined in the scene. The main contributions of the study are as follows:

- An alternative surveying solution was developed using a deep learning-based SLAM, multi-view geometry and coordinate system transformations.
- The methodology can be performed with minimum and cost-effective equipment since an RGB-depth camera and at least one fiducial marker are enough to map an unknown environment localizing characteristic points in a 3D local coordinate system.
- All coordinate estimations are transferred and exported in a local reference system which is well-defined in the scene, using the plane and the pose of a fiducial marker.
- The proposed solution can be specialized in specific challenging environments by re-training and fine-tuning the neural network which is integrated on the SLAM system.

To sum up, in this chapter, a literature review was conducted regarding the feature extraction and semantic segmentation in unstructured environments following with the precise positioning methods in GNSS-denied environments. In chapter 3, the methodological approaches that were developed in each pillar of this dissertation are analyzed, aiming to fulfill the gaps that emerged by the literature review.

Chapter 3

Methodological Approach

In this chapter, the methodology employed in this dissertation is presented, which aimed to explore and analyze the datasets, methods, techniques and architectures that were utilized in order to develop the proposed framework. Initially, two different approaches for feature extraction in challenging environments are presented while the integration and use of the second architecture in a SLAM system, is analyzed. Afterwards, a methodology for scene understanding using semantic segmentation is analyzed while finally, the architecture and techniques that were developed for the precise positioning and mapping alternative in GNSS-denied environments, are described.

3.1 Visual localization in challenging environments

Visual localization either as a technique to estimate the pose of a robotic system in a predefined map or as a SLAM sub-module performed in unknown environments, includes several challenges when applied in unstructured environments including planetary scenes. More specifically, in rocky and sandy environments or even Moon and Mars-like scenes two are the main challenges that this study attempts to encounter:

- Poor information in visual cues and features
- Intense changes in lighting conditions

In this study, two deep learning architectures were implemented aiming to deal with the above challenges. Initially, a CNN-based self-supervized architecture is implemented and optimized for keypoint detection and description while afterwards a lightweight CNN-based teacher-student architecture is developed for feature extraction and SLAM-based navigation. Both architectures were trained, fine-tuned and evaluated using datasets that were designed for unstructured and planetary-based environments.

3.1.1 Keypoint detection and description model architecture

To deal with the issues of visual cues lack and intense lighting changes in unstructured environments, SuperPoint (DeTone et al. 2018), a state-of-the-art methodology which outperforms handcrafted and deep learning feature extractors (Bojanic *et al.* 2019, Liu *et al.* 2022) was implemented and improved.

3.1.1.1 SuperPoint architecture

Superpoint is a fully convolutional neural network, composed by an encoder-decoder architecture which is performed using full-sized images as input. At first, a shared encoder, based on VGG neural network (Simonyan & Zisserman 2015) is utilized aiming to reduce the image dimensionality using three max-polling operations, extracting image cells in a size of $H_c = H / 8$ and $W_c = W / 8$ where H and W are the height and width of an image. The extracted tensor is imported in two decoders, one of which acts as a keypoint detector and the other one as a descriptor (fig 3.1).

Concerning the keypoint detector decoder, it undertakes the reconstruction of the full-sized image, extracting the probability of a keypoint existence in each pixel. Initially, it forms a tensor $X \in \mathbb{R}^{H_c W_c \times 65}$ where 65 channels is composed by 64 non-overlapping 8x8 pixel cells and an extra cell, called "no interest point dustbin" (DeTone *et al.* 2018). Subsequently, this tensor is imported to a "softmax function" where the dustbin cell is removed while the resulted tensor is reshaped to a full-sized image output ($\mathbb{R}^{HxW}$) after a "reshape operation". It's worth noting that the detector decoder doesn't upsample the full resolution of the image using transposed convolution techniques such as Unet due to high demands on computing resources while according to DeTone et al. 2018, these upsampling techniques are able to introduce checkerboard artifacts. Instead a"sub-pixel convolution" (Shi et al. 2016) is utilized, which doesn't include training parameters, aiming to reduce the computation process.

Regarding the descriptor decoder, it computes a tensor $\mathbb{R}^{H_c W_c \times D}$ where D is the descriptor length equal to 256 while via two convolutional layers, it extracts fixed feature maps in a shape of $I_{desc}{}^{H_c W_c \times D}$. The feature maps are reconstructed to the full-sized dimensions through a bi-linear interpolation while afterwards, the L2 norm operation is performed aiming to calculate the unit length of the descriptors. It's worth noting that the original SuperPoint architecture utilizes bi-cubic interpolation instead of bi-linear. However, in case of unstructured environments, it was observed that bi-linear interpolation provided similar accuracy while reducing the computation process compared with bi-cubic interpolation.

SuperPoint utilizes a unified loss function which is composed by the loss function of keypoint detector ($\mathcal{L}_p$) and the loss function of the descriptor ($\mathcal{L}_d$). SuperPoint uses pairs of wrapped images with the predicted keypoint locations and the corresponding transformation matrices or homography, utilized as ground truth. The unified loss function is presented in equation (3.1):

$$\mathcal{L}(X, X', D, D'; Y, Y', S) = \mathcal{L}_p(X, Y) + \mathcal{L}_p(X', Y') + \lambda \mathcal{L}_d(D, D', S) \quad (3.1)$$

Where $\mathcal{L}_p(X, Y)$ and $\mathcal{L}_p(X', Y')$ are the keypoint detector loss function for the original and a wrapped image respectively, defined as follows:

$$\mathcal{L}_p(X, Y) = \frac{1}{H_c W_c} \sum_{\substack{h=1 \\ w=1}}^{H_c W_c} l_p(x_{hw}; y_{hw}) \quad (3.2)$$

$$\text{with: } l_p(x_{hw}; y) = -\log\left(\frac{\exp(x_{hwy})}{\sum_{k=1}^{65} \exp(x_{hwk})}\right), \quad (3.3)$$

where $x_{hw} \in X$ are pixel cells of the input image while $y_{hw} \in Y$ the corresponding labels.

The loss function of the descriptor can be defined below:

$$\mathcal{L}_d(D, D', S) = \frac{1}{(H_c W_c)^2} \sum_{\substack{h=1 \\ w=1}}^{H_c W_c} \sum_{\substack{h'=1 \\ w'=1}}^{H_c W_c} l_d(\|d_{hw}\|_2, \|d_{h'w'}\|_2; s_{hwh'w'}), \quad (3.4)$$

Where: $\|d_{hw}\|_2$ and $\|d_{h'w'}\|_2 \in D$ are the normalized descriptor cells from the original and wrapped image respectively while $s_{hwh'w'}$ is a binary variable which presents the homography correspondence between (h, w) and (h', w') cells.

Moreover, the parameter λ_d was added, aiming to reinforce the balance between negative and positive correspondences while the hinge loss is used (3.5):

$$l_d(d, d'; s) = \lambda_d * s * \max(0, m_p - d^T d') + \\ + (1 - s) * \max(0, d^T d' - m_n), \quad (3.5)$$

where m_p and m_n are the positive and negative margins (Rosasco *et al.* 2004).

It's worth noting that, in the original SuperPoint, the descriptor cells ($\|d_{hw}\|_2$, $\|d_{h'w'}\|_2$) are not normalized. However, it was observed that the normalized descriptors, tuning the factor λ (3.1) and the weighting term λ_d (3.5) accordingly, produced more accurate results in unstructured environments (see section 4.1.2).

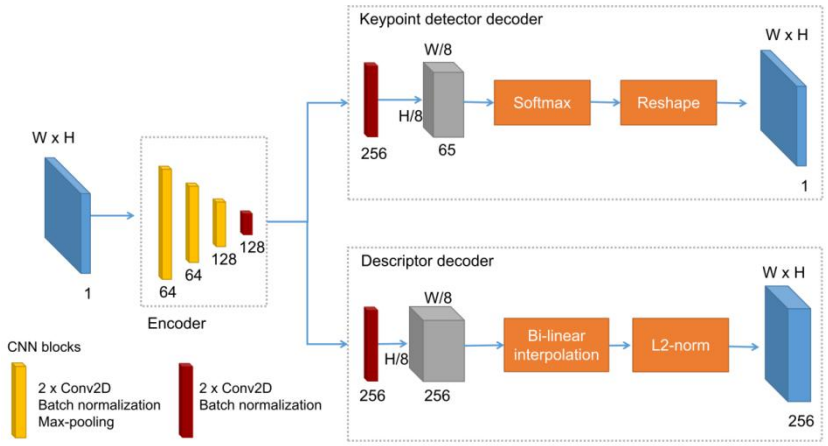


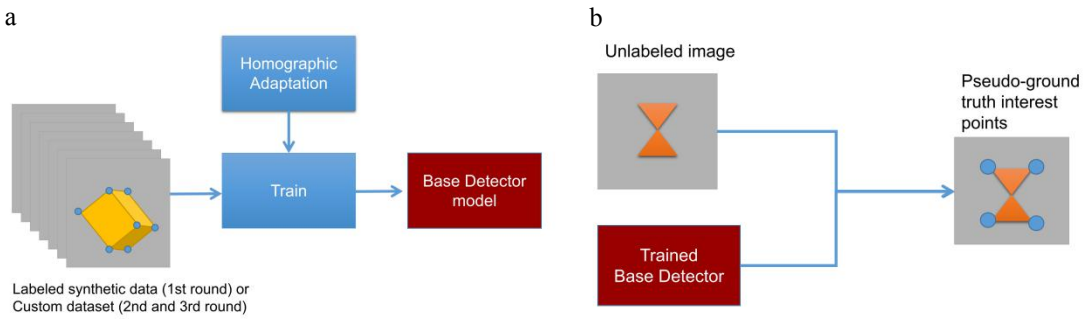
Figure 3.1 SuperPoint architecture

3.1.1.2 Self-supervized training of SuperPoint

The self-supervized training process of SuperPoint is conducted in several rounds aiming to increase the accuracy of feature detection. At first, the standalone keypoint detector, called MagicPoint (DeTone *et al.* 2018) is trained using a generated synthetic dataset which includes 2D geometric shapes such as lines, ellipses, triangles etc. During the training process, homographic adaptation is performed, which combines multiple random homographies of the input image and the keypoint predictions of the model, aiming to reinforce the efficiency in geometric transformations (fig 3.2a).

After the first round of training, the trained model is used in order to extract pseudo-ground truth of the desired dataset (fig 3.2b) while afterwards, the MagicPoint is re-trained using the desired dataset and the extracted labels while the homographic adaptation is utilized also. It's worth noting that, the MagicPoint training with the desired dataset can be repeated for two or three rounds using the optimized pseudo-ground truth each time, in order to further improve the detector's accuracy.

Finally, the SuperPoint including detector and descriptor is trained using the desired dataset and the optimized psudo-ground truth (fig. 3.2c).



c

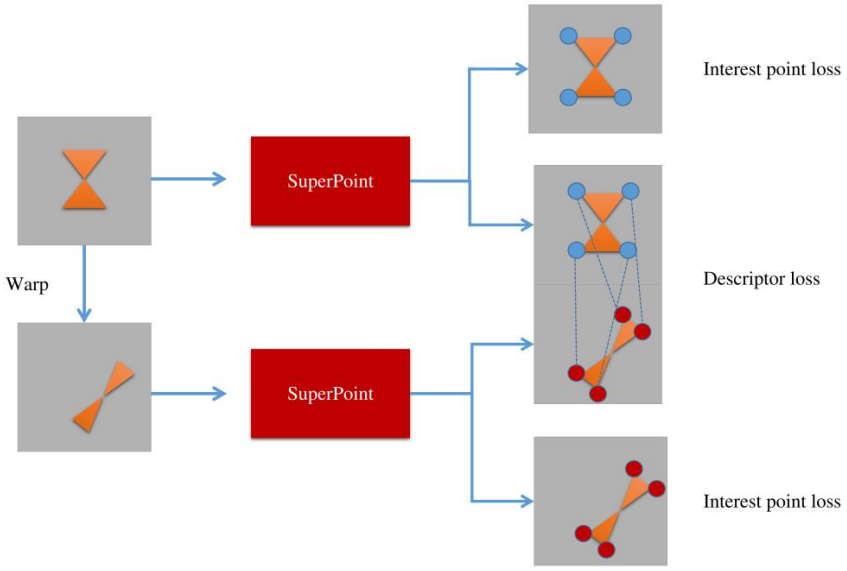


Figure 3.2 SuperPoint training process: (a) MagicPoint training using homographic adaptation, (b) Pseudo-ground truth prediction based on the trained model, (c) SuperPoint training and fine-tuning

3.1.2 Lightweight feature extraction model architecture

In this section, an alternative and lightweight approach for keypoint detection and description through deep learning in unstructured environments is presented. More specifically, a multi-task encoder-decoder architecture, based on HF-net (Sarlin *et al.* 2019) was implemented, focused on visual localization, being able to predict keypoint locations, local and global descriptors. The proposed NN, which can be called HF-net2, utilizes a teacher - student architecture (fig.3.3) in order to increase efficiency in terms of performance-time without decreasing the accuracy and reliability, while being capable of using in real-time applications.

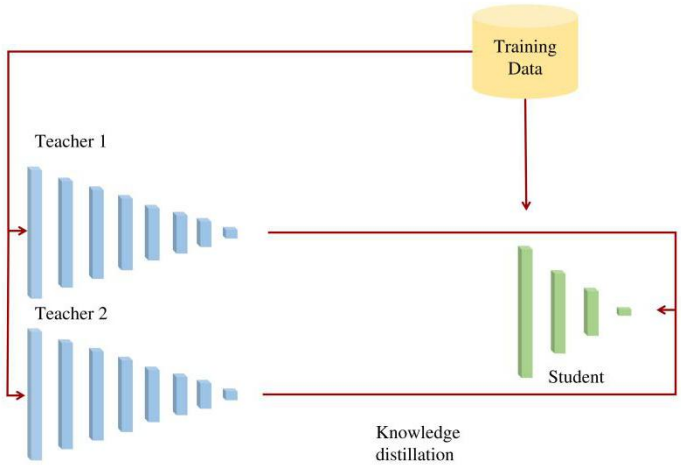


Figure 3.3 A multi-teacher-student architecture

Initially, the training dataset feeds two pre-trained models which represent the teachers, while the distilled knowledge plays the role of ground truth for the student during the training process. The one teacher utilizes the SuperPoint architecture (DeTone et al. 2018) extracting keypoint locations and local descriptors while the second one extracts global descriptors using the NetVLAD architecture (Arandjelovic et al. 2016).

The student architecture is composed of a shared encoder and three different sub-modules which focus on: (a) keypoint detection (b) local description (c) global description. For the shared encoder, the MobilenetV3-large (Howard *et al.* 2019) is utilized instead of MobilenetV2 (Sandler *et al.* 2018) which is used on the original HF-net while a decoder based on SuperPoint extracts the keypoint scores and local descriptors. Simultaneously, on top of the last feature map of MobilenetV3-large a NetVLAD layer predicts the global descriptor of each entire image (fig. 3.4).

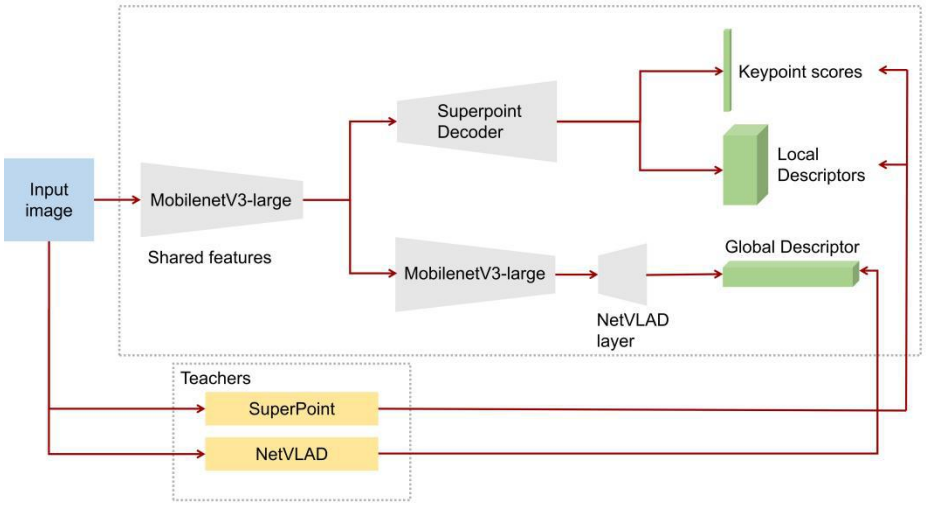


Figure 3.4 HF-net2 architecture

MobilenetV3-large utilizes Neural Architecture Search (NAS) (Elsken *et al.* 2019) and the non-linearity activation function called hard-Swish (Ramachandran *et al* 2017) which combines the Swish activation function (3.6) (Ramachandran *et al* 2017) with the piece-wise alternative of the sigmoid function $\frac{ReLU6(x+3)}{6}$, described by the equation (3.7).

$$\text{swish}[x] = x\,\sigma(x) \quad (3.6)$$

$$\text{h}-\text{swish}[x] = x\frac{\text{ReLU6}(x+3)}{6} \quad (3.7)$$

Where σ(x) is the sigmoid function and ReLU6 is a modification of the well-known "rectified linear unit" activation function.
Thus, MobilenetV3-large achieves increased efficiency in terms of performance time and accuracy compared with MobilenetV2.

As a result, the proposed NN, is able to combine multi-task prediction, utilizing knowledge distillation, achieving a flexible end-to-end training process with high efficiency in low-resources computing systems. The proposed NN was trained and fine-tuned using a dataset which includes FPV images from Earth, Mars and Moon, aiming to increase its robustness in unstructured and challenging environments. More information about the dataset is referred in the section 3.1.4.

3.1.3 SLAM system for unstructured environments

The proposed NN, trained and optimized for unstructured environments was integrated in a SLAM system aiming to increase the efficiency of autonomous navigation in challenging conditions and completely unknown environments. In other words, the proposed SLAM system focuses on unstructured environments with visual cues lack and intense lighting conditions, using the trained and fine-tuned model for keypoint detection, local and global description.

The proposed SLAM system is based on ORB-SLAM2 (Mur-Artal & Tardos 2017) where instead of using ORB descriptor to extract features, the HF-net2 model is utilized. The proposed SLAM uses RGB images with the corresponding depth information aiming to be scale-aware while is divided in three different modules which are performed simultaneously in separated threads: (a) tracking, (b) local mapping and (c) loop closing.

The tracking module processes the RGB-Depth data while the integrated HF-net2 model predicts keypoint locations, local and global descriptors. The keypoints and local descriptors are used for the camera pose prediction while when the SLAM system detects multiple features in neighbored frames, it extracts a new keyframe using the camera pose predictions. The keyframes which are treated as landmarks combined with the keypoints, aid the local mapping module to map the surroundings.

The loop-closing module is based on DXSLAM (Li *et al.* 2020) and combines the Fast BoW (Bag-of-Words) algorithm (Munoz-Salinas & Medina-Carnicer, 2020) which converts each keyframe in a vector of words using the pre-trained vocabulary tree, with the images representation extracted by the global descriptors predicted by the proposed HF-net2 model. When a loop is detected, the system optimizes the camera trajectory and point cloud using full bundle adjustment technique (fig 3.5).

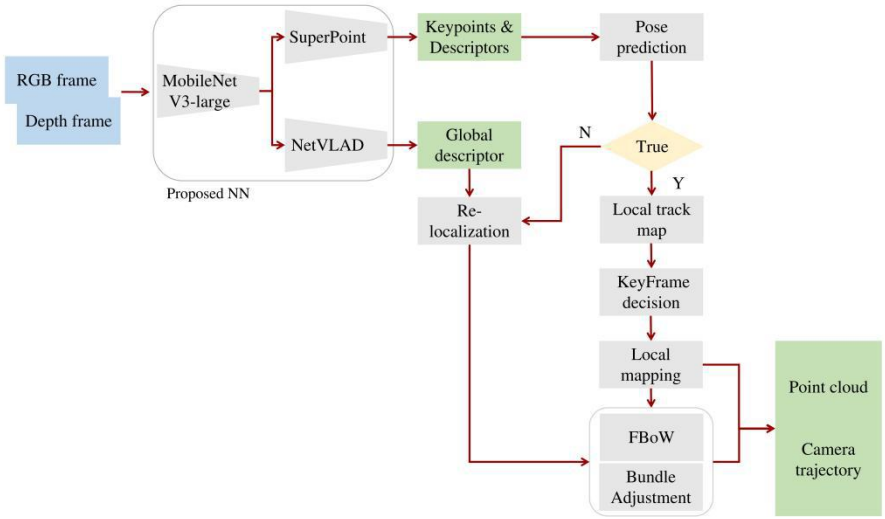


Figure 3.5 SLAM architecture based on the proposed NN

3.1.4 Datasets

Several studies investigate feature extraction in FPV images in urban and indoor environments using deep learning architectures. However, those architectures are trained with general-purpose datasets such as COCO (Lin *et al.* 2014) while the best of author's knowledge, there is not any deep learning model for feature extraction, focused on unstructured and planetary scenes.

Thus, two different datasets were design aiming to train and evaluate the proposed deep learning methodologies:

- A training dataset
- An evaluation dataset

3.1.4.1 Training dataset

The training dataset includes 48 000 of FPV (First-Person-View) or rover-based images with wide range of variations in landscapes, including images from Earth, Moon and Mars.

Regarding the Earth, dataset contains 26 000 RGB images, captured from construction sites, mountainous areas, sandy beaches and a quarry from the area of Crete, Greece (fig 3.6a). The images were taken in scenes with various lighting and weather conditions in day and nighttime while the camera was located 1.5 meter from the ground in a direction of 10 and 60 degrees from the horizon (fig. 3.7).

The images from Mars were collected by a publicly available dataset of NASA which includes about 13 000 images captured by Mars Science Laboratory (MSL, Curiosity) rover using three instruments: Mastcam Right eye Mastcam Left eye, and MAHLI

(Lu 2023) (fig 3.6b). Concerning the Moon's dataset, includes about 9 000 artificial rover-based images which generated and released with CC (Creative Commons) license by Keio University in Japan (fig 3.6c). The dataset was created using the Moon LRO LOLA digital elevation model which is based on the Lunar Orbiter Laser Altimeter (Smith et al. 2010) combined with the simulation software Terragen of Planetside Software.

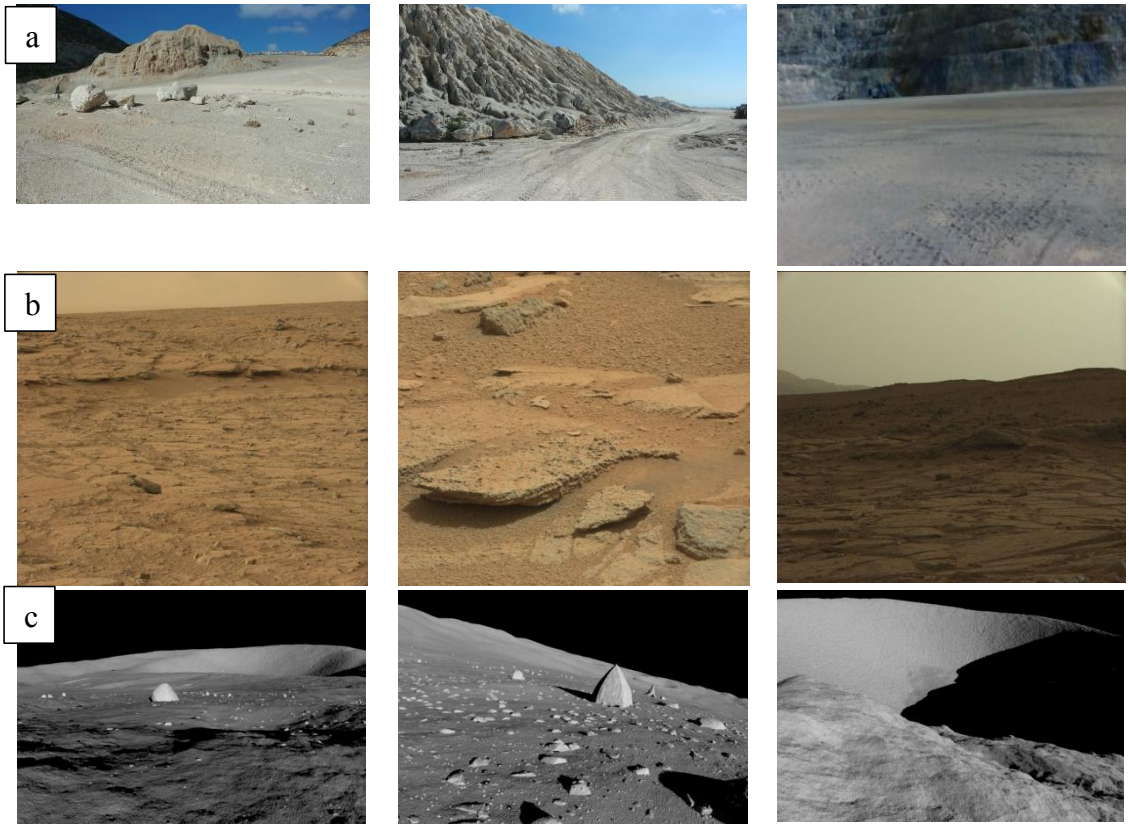


Figure 3.6: A sample of training dataset. (a) images from Earth, (b) images from Mars (c) images from artificial lunar surface

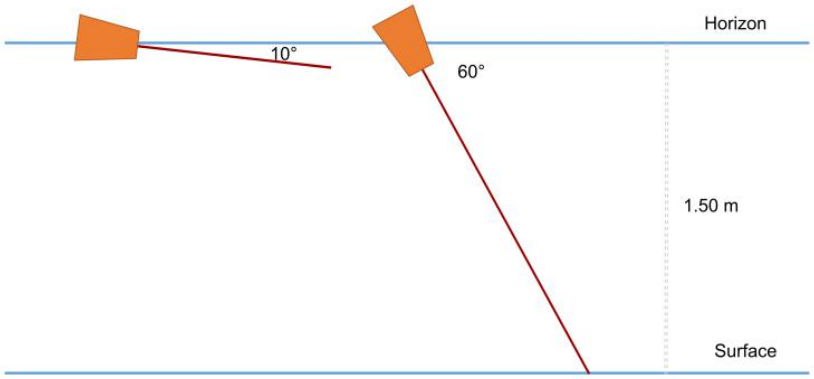


Figure 3.7 Camera's direction relative to the horizon

3.1.4.2 Evaluation dataset

For the evaluation of feature extraction architectures and the comparison with other widely used handcrafted algorithms, an evaluation dataset were designed for unstructured and planetary scenes inspired by HPatches dataset (Balntas *et al.* 2017), one of the most popular datasets for keypoint detection and description evaluation in general-interest images.

The proposed dataset contains 120 sequences of images from Earth, Mars and Moon which not included in the training dataset. Each sequence is composed of the original image and five different representations of the original image in terms of illumination and viewpoint. More specifically, 60 out of 120 sequences includes the original image and five generated images with intense illumination changes while the remaining 60 sequences contains the original image and five generated images with various viewpoint changes (fig. 3.8, fig. 3.9). In each sequence, five transformation matrices determine the ground truth between the original image and each of the five representations. The sequences with illumination changes contains identity matrices, since the only difference among the representations is the illumination.

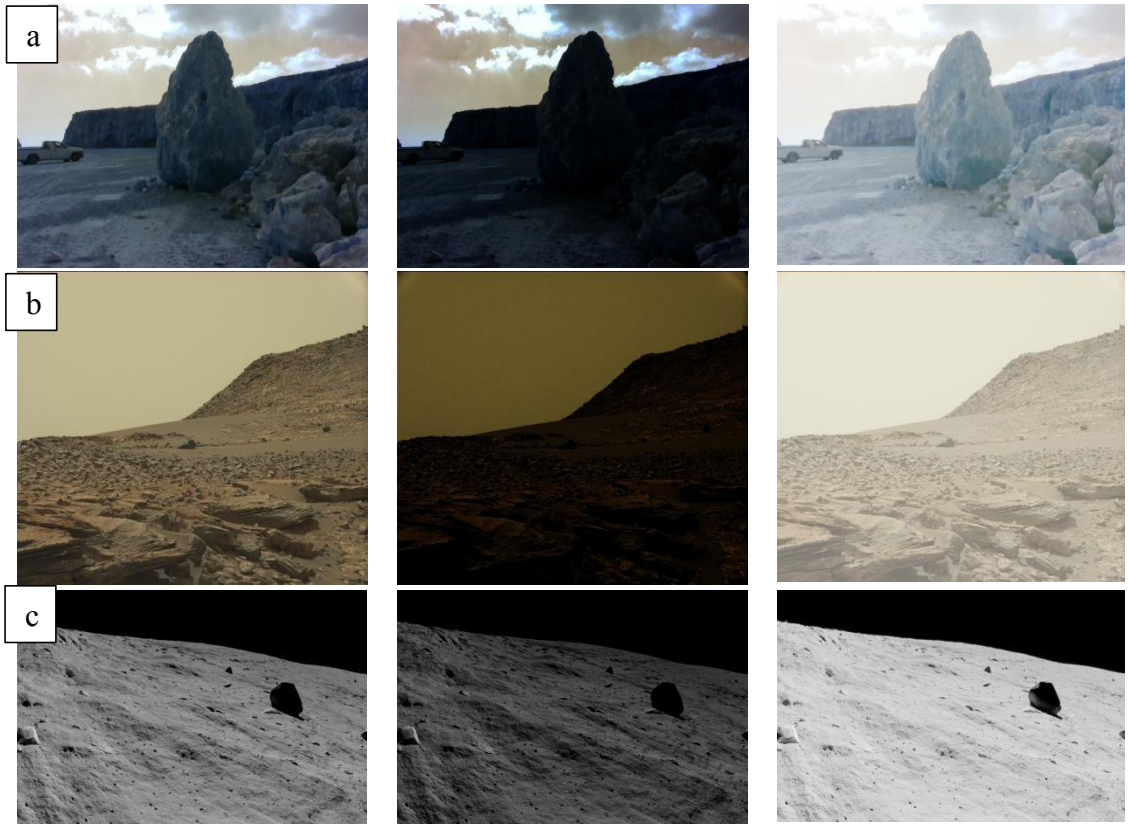


Figure 3.8: A sample of illumination-part evaluation dataset. (a) sequence from Earth, (b) sequence from Mars (c) sequences from artificial lunar surface

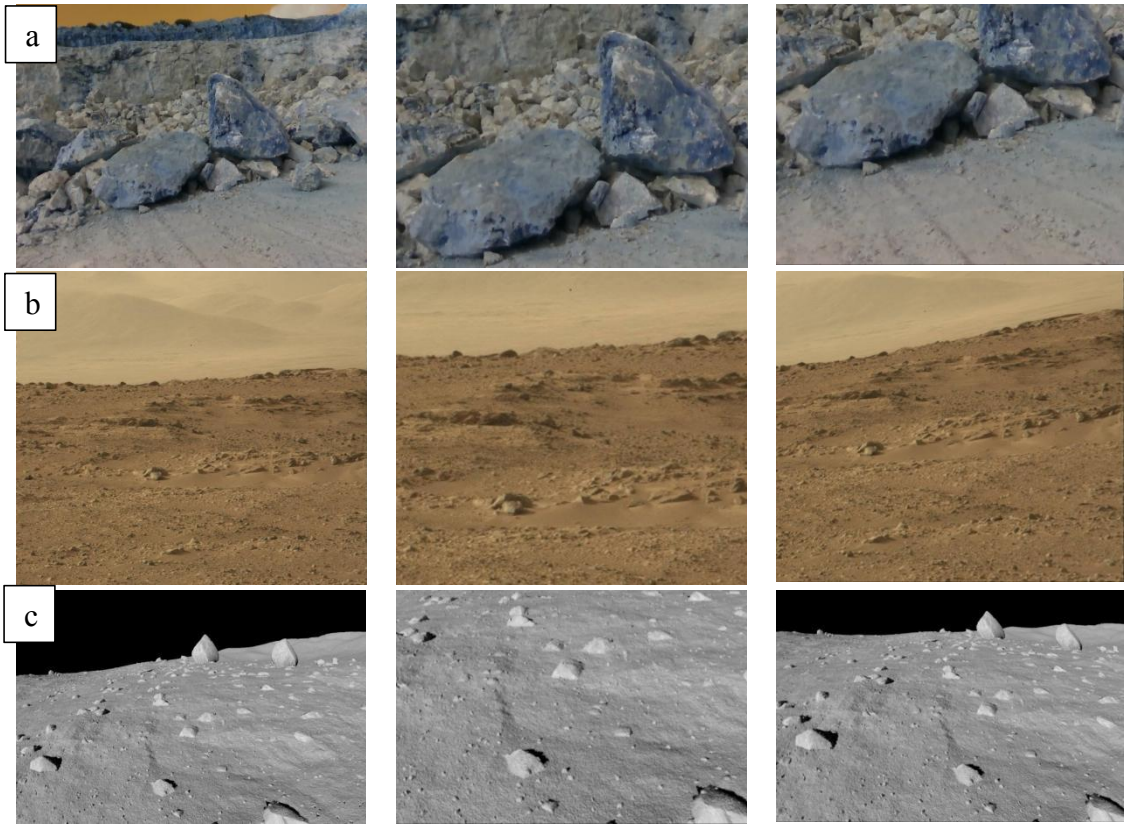


Figure 3.9: A sample of viewpoint-part evaluation dataset. (a) sequence from Earth, (b) sequence from Mars (c) sequences from artificial lunar surface

3.2 Semantic segmentation in unstructured environments

Semantic information in unstructured environments provides a contextual understanding of objects and their relationships within an image, enabling machines to recognize and categorize features semantically, reinforcing crucial tasks including autonomous navigation in unknown planetary scenes. Although the literature includes several studies focused on terrain segmentation in unstructured scenes, there are two main gaps, that the dissertation attempts to fill:

- Semantic segmentation specialized in the lunar surface, since most of the studies investigate scene understanding through semantic segmentation in earthy unstructured environments or in the martian surface
- A lightweight semantic segmentation model, capable of being used in systems with low computing resources, providing high efficiency after training with a limited size of dataset

In other words, the scope of this study, is the development of a lightweight semantic segmentation model which is able to provide increased accuracy with potential use in real-time tasks during a rover-based mission on the lunar surface. Two challenges have to be encountered. The first one is the lack of valuable rover-based datasets for the lunar surface, compared with Mars where several datasets have been implemented.

The second challenge is the size of the model, since most of the semantic segmentation architectures are computationally expensive.
To address these challenges above, a modified U-net architecture is proposed, since U-net is an efficient and accurate neural network in terms of accuracy which doesn't require large datasets (Eui-ik *et al.* 2021, Chhabra *et al.* 2022).

More specifically, the proposed architecture, is composed by an encoder-decoder architecture where a modified version of MobileNetV2 neural network (Sandler *et al.* 2018) is used as an encoder and a lighter decoder of U-net is utilized for the segmentation stage. To speed up, the learning process, the MobileNetV2 has been trained with ImageNet, a well-known image dataset which includes millions of general-purpose photographs, so as during the training process, to "transfer" its earned "experience" to the model, encountering the issue of the limited size of lunar surface dataset.

3.2.1 Modified U-net architecture

3.2.1.1 U-net architecture

As referred above, the proposed architecture is based on U-net (Ronneberger *et al.* 2015), a well-known architecture for semantic segmentation which initially proposed for medical applications. The U-shaped model of U-net can be separated in two main components: (a) the encoder, which reduces the image dimensions, increasing the feature maps while learns to classify the desired features, and (b) the decoder, which reconstructs the image dimensions, decreasing the feature maps and performs precise segmentation of the detected features. U-net decoder, is able to segment the detected features retrieving the topology of the image content through four skip connections among different levels of the encoder which transfer information to the decoder in order to maintain the spatial details of images with the aim to reconstruct them (fig. 3.10).

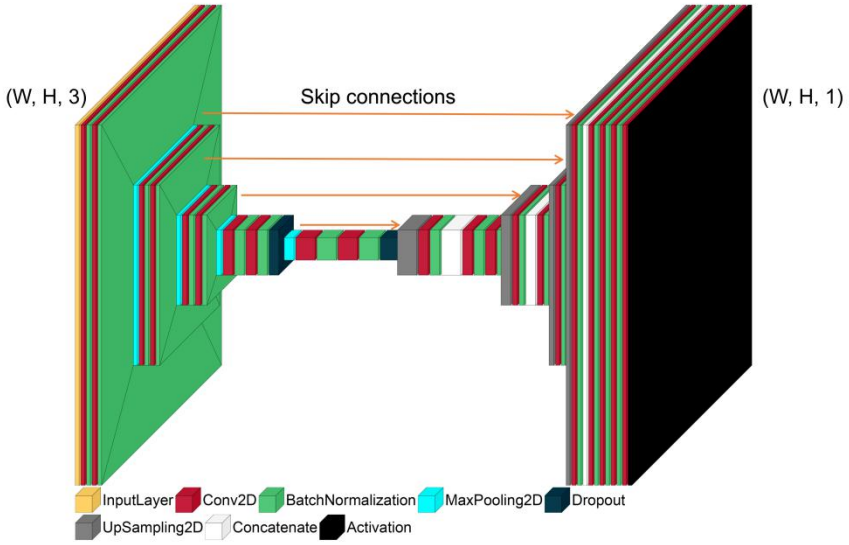


Figure 3.10 U-net architecture

U-net is mainly composed by convolutional (Conv2D) and "BatchNormalization" layers. Regarding the encoder-decoder functionality, the encoder downsamples the image through the "MaxPooling2D" layer, and the decoder upsamples the image using the UpSampling2D layer while the "Concatenate" layer creates the skip connections between the encoder and decoder part. At the end, "softmax" (3.8) which is the activation function is utilized in order to export the segmentation map for each input image.

$$\sigma(\overrightarrow{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (3.8)$$

Where $\vec{z}$ is the input vector and z_i are the elements of the input vector while $\sum_{j=1}^{K} e^{z_j}$ is a normalization term with K classes which ensures that the output of the function will sum to one and each output value will be in a range of (0, 1).

3.2.1.2 U-net with MobileNetV2 as encoder

Although U-net is an accurate semantic segmentation architecture, it provides high performance-time while it requires a time-consuming training process with much experimentation in fine-tuning, since it includes about 31,000,000 trainable parameters. In order to accelerate the training process, "transfer learning" technique is utilized, using a pre-trained (with ImageNet dataset) MobileNetV2 (Sandler *et al.* 2018) as the encoder (fig. 3.11).

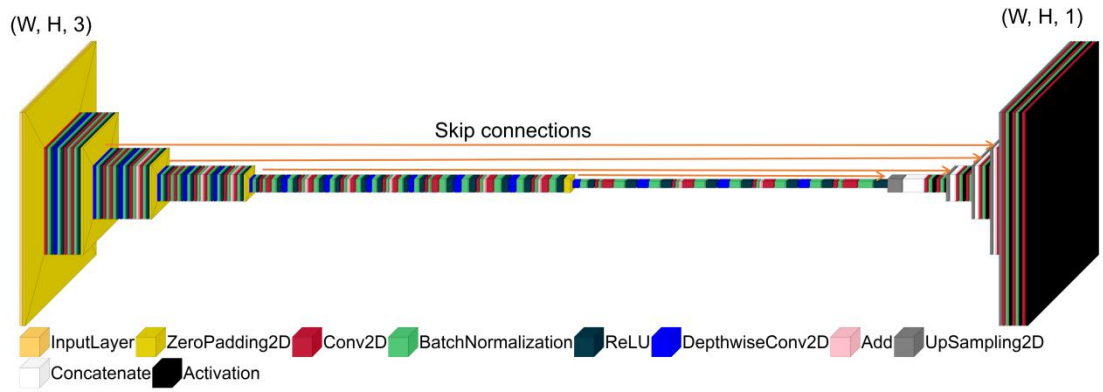


Figure 3.11 Architecture of U-net with MobilenetV2 as encoder

MobileNetV2 (Sandler *et al.* 2018) is a CNN-based architecture designed for providing high efficiency in mobile devices while has been utilized in multiple tasks of computer vision including classification, semantic segmentation, object detection, etc. The main MobilenetV2 architecture is composed by 19 residual bottleneck layers where each bottleneck is based on inverted residual block. The inverted residual block is based on a narrow-wide-narrow approach using a point-wise convolution with Relu6, followed by a depth-wise convolution with Relu6, followed by a linear point-wise convolution, while a skip connection, merges the input of the block with the output through the "Add" layer (fig 3.12). This approach reduces the extracted parameters and computation compared with conventional convolution layers while according to Sandler *et al.* 2018 when the kernel k=3 for 3x3 depth-wise convolution, the computational cost is about 9 times smaller compared with traditional convolution without significant reduction in accuracy.

More specifically, if the input of a traditional convolution is $h_i \times w_i \times d_i$ where h and w, the image dimensions and d, the depth or channels while the output is $h_i \times w_i \times d_j$, then the computational cost is calculated as $h_i \times w_i \times d_i \times d_j \times k \times k$, where k, the kernel size, while the corresponding computational cost of an inverted residual block will be: $h_i \times w_i \times d_i (k^2 + d_j)$.

Figure 3.12 Inverted residual block architecture

The combination of the original pre-trained MobileNetV2 as an encoder with U-net decoder, provides a more lightweight architecture including about 8,000,000 trainable parameters which quite less than U-net which includes about 31,000,000, while is able to accelerate the training process. However, this architecture remains unsuitable for applications which require high efficiency in inference-time especially for real-time tasks.

3.2.1.3 A lightweight version of U-net with MobileNetV2 as encoder

To deal with low-performance time without reducing the accuracy, a more lightweight architecture is proposed based on a modified MobileNetV2 encoder and a lightweight U-net decoder.

Regarding the modified MobileNetV2, is composed by an initial fully convolution layer followed by 13 residual bottleneck layers, instead of the original MobileNetV2 which includes 19, since right after the block 13, the parameters are highly increased from about 92,000 to 155,000 in the original architecture. Moreover, to further reduce the computational cost, the depth-multiplier which is a positive factor that multiplies the channels through the depth-wise convolution, was defined with a value of 0.35 instead of 1.0 which is the default value aiming to decrease the output channels of the depth-wise convolution layers. It's worth mentioning that for depth-multiplier values less than 1.0, the depth-multiplier is applied to all layers except the last convolution layer.

Concerning the U-net decoder, all the filters of the convolution layers were divided by the factor of 2 aiming to accelerate the segmentation stage while the four skip connections connects the input image, the block 1, block 3 and block 6 of the encoder respectively.

The proposed architecture includes about 220,000 trainable parameters which are far fewer than the 31,000,000 and 8,000,000 trainable parameters of U-net and original MobileNetV2/U-net respectively.

The proposed architecture with detailed representation of the layers is presented in figure 3.13 while a more abstract representation is depicted in figure 3.14.

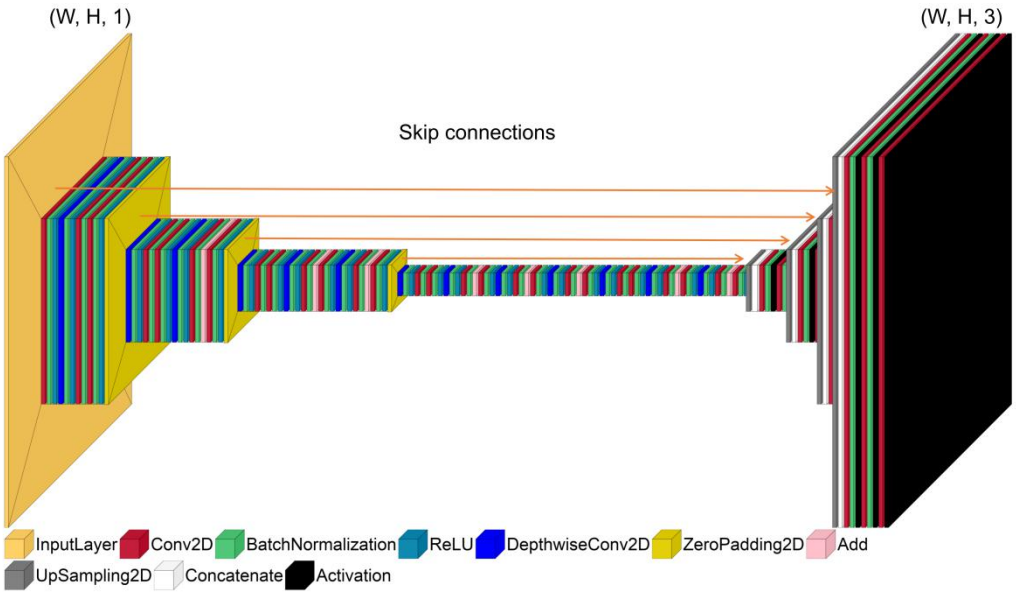


Figure 3.13 Proposed architecture

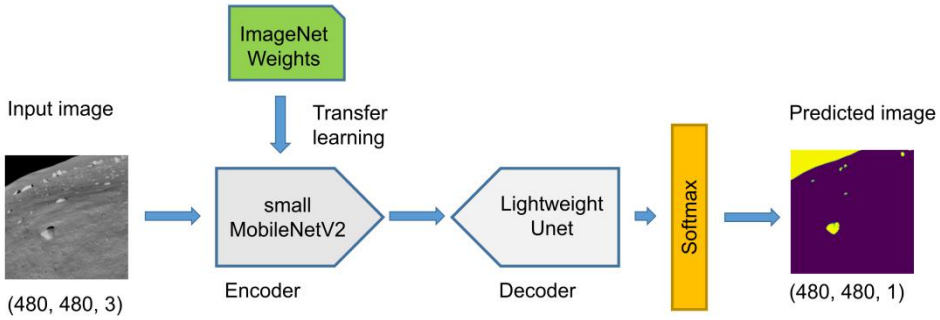


Figure 3.14 Proposed architecture for lunar terrain segmentation

The proposed architecture deals the first aforementioned challenge for a semantic segmentation in unstructured environments, since it doesn't require large datasets due to U-net capabilities and transfer learning technique, while it is suitable for systems with low computing resources since it includes only 220,000 parameters which are about 140 times less than the original U-net. Moreover, as the experimentation proves (see section 4) the accuracy is not reduced, providing robust and satisfactory results.

3.2.2 Dataset

As referred above, there is a lack in datasets for lunar surface segmentation while to the best of author's knowledge, there is not rover-based image dataset which depicts the real lunar landscapes instead of Mars where several rover-based datasets have been proposed.

Thus, for training and validation of the proposed architecture in lunar environment, a dataset with artificial rover-based images which depict lunar landscapes was utilized, created by the Space Robotics Group of Keio University in Japan. The images generated using Planetside Software's Terragen and a real DEM (Digital Elevation Model, which is based on data from the Lunar Orbiter Laser Altimeter on NASA (Smith *et al.* 2010). It includes about 9,700 artificial images and the corresponding annotated masks taking the following four classes into account (fig 3.15):

- Large rocks
- small rocks
- Sky
- Ground (backgrand)

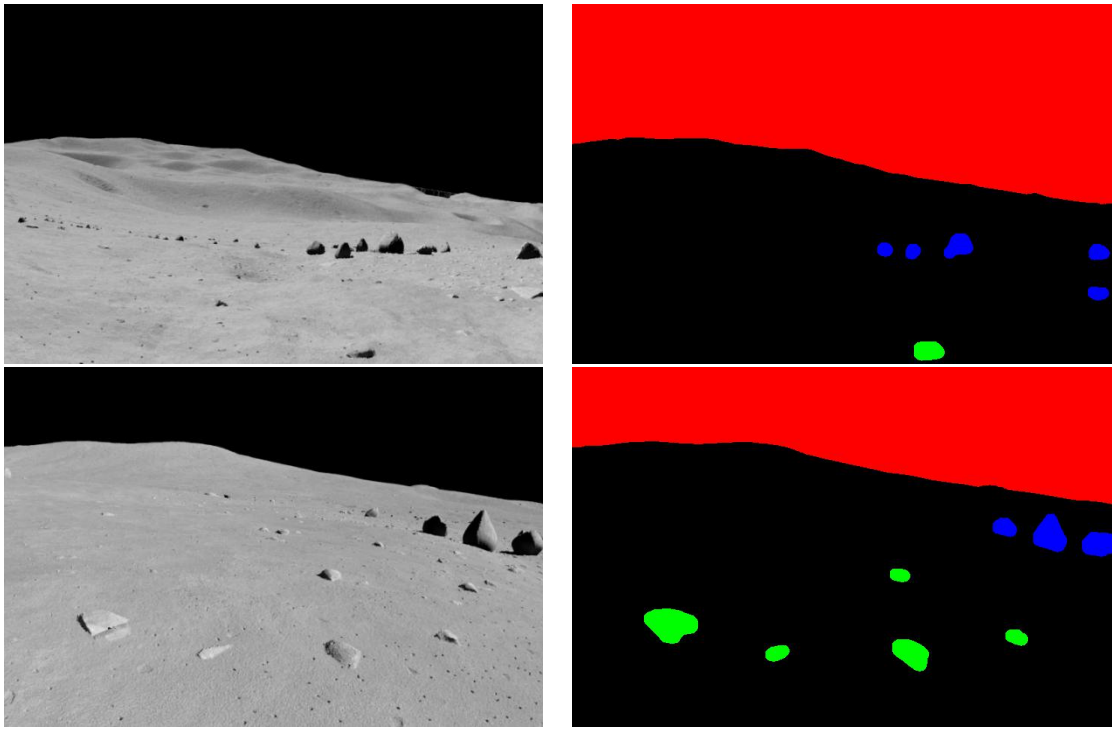


Figure 3.15 Dataset of lunar surface for semantic segmentation bySpace Robotics Group of Keio University in Japan. The artificial images are presented in the left column while the corresponding masks in the right column

Several drawbacks are included in the dataset, such as the decreased accuracy in feature segmentation and the lack of balance between the classes of large rocks and small rocks, since the examples of small rocks are by far more than the examples of large rocks. To deal with the imbalanced classes, the two classes of rocks were merged in one class. Thus, the new dataset includes the following classes:

- Rocks
- Sky
- Ground (background)

Nevertheless, since this is the only publicly available dataset for the lunar surface focused on semantic segmentation, it was utilized in order to train and validate the proposed architecture, aiming to provide a lightweight model for potential use in systems with low computing resources during the rover navigation, on the lunar surface.

3.3 Precise positioning and mapping in GNSS-denied environments

The main goal of this study is the localization of fiducial markers and characteristic points of the scene, providing their local coordinates in 3D space under a high level of accuracy, using minimal equipment. In other words, presented methodology maps an area of interest, by extracting the pose estimation of pre-defined fiducial markers and

a point cloud in a local coordinate system using an RGB-Depth camera. At first, the fiducial markers are placed in the scene where one of them is used as the origin marker while the target markers represent the characteristic points or features. Subsequently, the proposed SLAM (see section 3.1.3), enables the RGB-Depth camera to map the desired area and localize itself in an unknown and challenging environment, while in combination with geometrical transformations, localization and optimization techniques, the present methodology estimates the coordinates of target markers and an arbitrary point cloud which approximates the structure of the environment.

3.3.1 System Architecture

The system architecture is presented in the following figure:

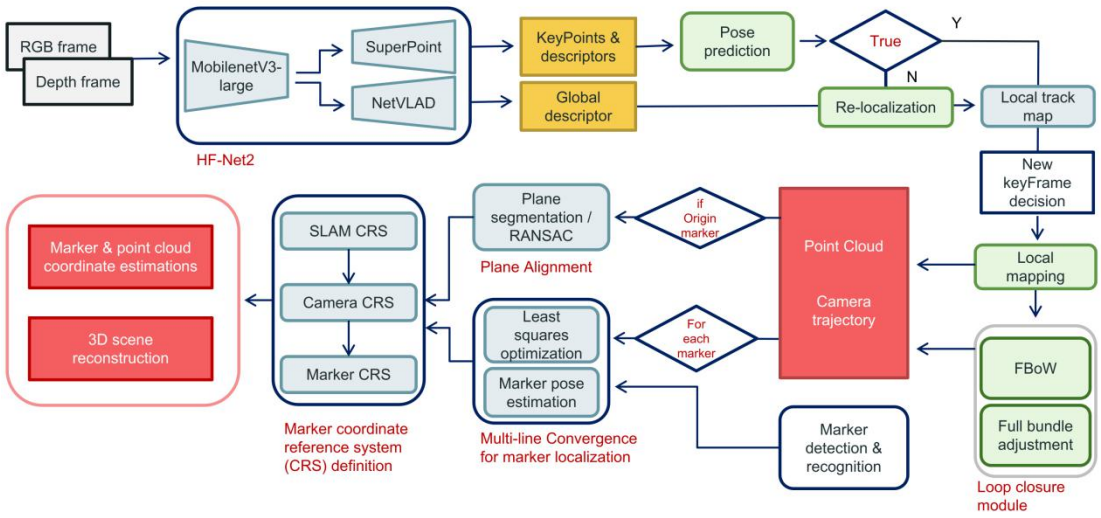


Figure 3.16 Overall architecture of precise positioning and mapping methodology

Initially, a fiducial marker which is defined as the origin of the local coordinate system is placed in the area of interest, while a number of fiducial markers which are defined as targets represent natural or artificial features. Afterwards, the user can record a video, passing through all the desired characteristic points capturing RGB and depth information. During the data processing, the RGB-Depth image frames are inserted to the proposed HF-Net2 neural network (please see section 3.1.2) in order to extract keypoints and local / global descriptors of the scene (feature extraction module). The system using the internal parameters of the camera and the local keypoints and descriptors, predicts the camera pose while utilizes the global descriptors in case of a prediction failure. If it observes groups of features in multiple sequential frames, it stores a keyframe. Based on the process above, the SLAM algorithm outputs multiple keyframes which are treated as landmarks since, in combination with the keypoints, are necessary for the local mapping, the loop closure detection and for the re-localization of the camera. For the optimization of the camera's pose prediction, local mapping and loop closure detection, SLAM algorithm

utilizes the bundle adjustment (BA) algorithm using the Levenberg-Marquardt method (Mur-Artal & Tardo 2017).

After the end of the SLAM process, it outputs a point cloud and a trajectory of the scene while traditional image processing techniques such as adaptive and Otsu thresholding (Otsu 1979) provide the identifications of target markers. Subsequently, through the multi-line convergence method (Trigkakis *et. al,* 2020), the locations of the markers are estimated while the pose of the origin marker is optimized with the utilization of plane alignment method (Trigkakis *et. al,* 2020). Finally, the coordinate estimations are transferred in a local coordinate system, defined by the pose of the origin marker. After the end of the process, the user is able to study the mapping area and conduct measurements using a 3D point cloud, a camera trajectory and the marker estimations, which are defined in the local coordinate system with origin, the origin marker.

3.3.2 Coordinate system definition

A core component of the methodology for the final coordinate estimations and 3D scene reconstruction is the coordinate system definition. The first coordinate system is defined and established by the proposed SLAM system using the first frame of the captured video. The x and y axes in this initial coordinate system, follow the right and top directions of the frame respectively while the z axis is equivalent with the camera direction towards the landscape of the area. Subsequently, the calibration data and the camera pose (retrieved by camera trajectory information) along with the target marker coordinates which are calculated by marker detection module, extract the vectors of rotation and translation that are utilized in transformation of the initial reference system to the camera reference system. Finally, the reference system definition module, calculates the translation vector and rotation matrix from the orientation and translation of the origin marker and defines the final reference system based on the origin marker's pose. The x and y axes of the marker reference system follow the right and top direction of the marker while the z axis follows the zenith direction (fig. 3.17).

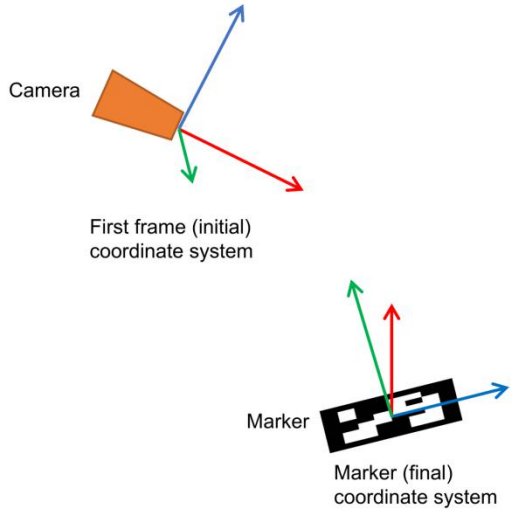


Figure 3.17 Initial and final coordinate systems. The initial coordinate system defined by SLAM is formed by the first recorded frame of the camera while the final coordinate system is defined by a single marker (the origin marker). Both coordinate systems are visualized with x axis in blue, y axis in green and z axis in red

3.3.3 Multi-line convergence (MLC) and Plane Alignment (PA) methods

Multi-line convergence method (MLC), is a method for the marker location definition that is based on the observation that the extended line segments which connect each marker pose estimation with the corresponding camera position, converge in an area that corresponds to the location of the marker in the 3D scene (fig 3.18). The method defines the optimized point that the extended line segments converge, using pseudo-inverse least squares optimization (Samuel 2004, Eldén 1982). The method can be described by the following equation:

$$p = S^{+} \cdot C, \quad (3.9)$$

Where p is the minimized distance of the theoretical convergence point from all the lines while S+ is the pseudo-inverse matrix of S which is defined in eq 3.10. C is defined in eq. 3.11.

$$S = \sum_i [n_i n_i^T - I], \quad (3.10)$$

$$C = \sum_i [n_i n_i^T - I]\, a_i, \quad (3.11)$$

Where each line is defined with “i”, “α_i” is the starting point of line “i” and “n_i” is the direction of line i while “I” is an identity matrix.

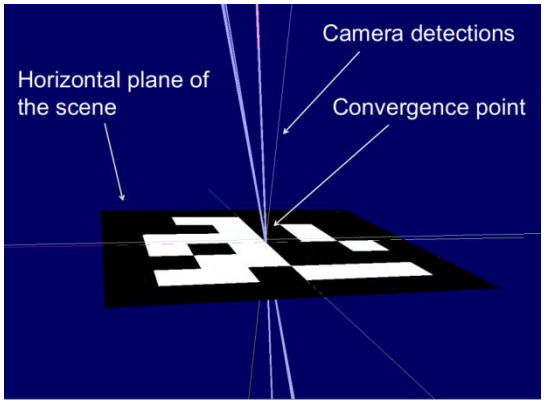


Figure 3.18 The detection of a fiducial marker implies that its location lies on a line connecting the camera's location with the center of the fiducial marker. By obtaining multiple such detections, it is determined the point where the lines intersect, or at least, are close to intersecting

Regarding the Plane alignment method (P.A.), it is performed to correct the translation and rotation errors of the origin marker that defines the final coordinate system of the scene. This step is important because any pose estimation error in the origin marker is transferred in every target marker and point cloud data of the scene. With the PA method, the pose and rotation of the origin marker is corrected leading to reliable measurements and an accurate definition of the origin coordinate system.

More specifically, initially, plane segmentation is performed on the point cloud, aiming to produce part of the point cloud that matches a plane while it gives access to the plane coefficients, in the form of ax+by+cz+d = 0. At the same time, it is able to obtain the normal vector from the plane coefficients, forming n = [a, b, c].

For the marker alignment with the point cloud, a rotation is performed that when applied on the pose normal vector, it aligns it with the plane normal vector. By expressing this rotation based on the normal vectors, it is able to apply it as a rotation matrix to the pose rotation matrix (through matrix multiplication), and define all three rotation angles at the same time:

Using the above formulas, a transformation matrix U is obtained (eq. 3.17):

$$G = \begin{pmatrix} A \cdot B & -\|A \times B\| & 0 \\ \|A \times B\| & A \cdot B & 0 \\ 0 & 0 & 1 \end{pmatrix}, (3.12)$$

$$u = \frac{(A \cdot B)A}{\|(A \cdot B)A\|} = A, (3.13)$$

$$\upsilon = \frac{B - (A \cdot B)A}{\|B - (A \cdot B)A\|}, (3.14)$$

$$\omega = B \times A, (3.15)$$

$$F = (u \quad \upsilon \quad \omega)^{-1} = \left(A \quad \frac{B - (A \cdot B)A}{\|B - (A \cdot B)A\|} \quad B \times A\right)^{-1}, (3.16)$$

$$U = F^{-1} \cdot G \cdot F, \quad (3.17)$$

The multiplication of U with a vector v expresses the rotation from A to B where A and B are the normal vectors of the solution. Instead of multiplying U with a vector, it can be multiplied with the rotation matrix corresponding to the pose of the marker to form a new pose. Then the normal of the marker will align with the plane normal.

The matrix U, when multiplied with the marker's rotation matrix (U x R) forms a new rotation matrix that is aligned to the plane normal. In order for the algorithm to match enough points to obtain an accurate normal, a procedure of plane segmentation is performed on the entire point cloud while then, the procedure is applied again, but only locally (1 meter radius). Finally, the minimum distance between the most significant 1-meter radius plane normal and the set of normals from the entire point cloud is estimated. Thus, the plane normal is extracted from the entire point cloud, that matches the local plane normal best. By performing this procedure, it is possible to correct both the pose and rotation of a fiducial marker, leading to robust measurements, and a fine-tuned definition of the origin coordinate system, which is paramount to obtaining the final estimates for all markers that do not participate in defining the coordinate system.

As a conclusion, in this chapter the methodologies and architectures that were developed in order to reinforce the efficiency of feature extraction, SLAM, semantic segementation and precise positioning processes in unstructured environments were analyzed. In the next chapter, the implementation details and the extracted results derived by an extended experimentation of each methodology including the optimized SuperPoint model, the Hf-net2 with the proposed SLAM, the modified U-net and the precise positioning alternative are presented.

Chapter 4

Implementation and results

In this chapter, the technical details about the implementation and the design of experimentation for the proposed framework is analyzed while the results of each methodology are presented. The structure of this chapter follows the four main components of the framework described below:

- Implementation and results of SuperPoint model
- Implementation and results of HF-net2 architecture and the proposed SLAM
- Implementation and results of the proposed NN for semantic segmentation
- Implementation and results of the precise positioning and mapping in GNSS-denied environments

4.1 Implementation and results of SuperPoint model

In this section, the implementation and training procedure of the SuperPoint architecture are presented, while afterwards the evaluation and results of the extracted models are described.

4.1.1 SuperPoint implementation and training

SuperPoint was implemented using the TensorFlow (Abadi *et al.* 2015) deep learning platform and trained utilizing the proposed dataset (see 3.1.4.1), aiming to increase the SuperPoint's sensitivity in planetary and unstructured scenes.

As described in section 3.1.1, the original SuperPoint's architecture was improved applying the following two modifications:

- The bi-linear interpolation is utilized for feature maps reconstruction in full-sized images instead of bi-cubic interpolation, used by the original SuperPoint
- In the loss function of keypoint description, the descriptors of initial and wrapped images are L2 normalized while tuning the weighting parameters including λ and λ_d accordingly (see 3.1.1.1).

During the experimentation, three SuperPoint models were produced following the training approaches presented below:

- The original SuperPoint was trained from scratch, using the proposed dataset aiming to focus on planetary environments
- The original SuperPoint was trained using the proposed datastet, based on the weights extracted by the training of SuperPoint with COCO (Lin *et al.* 2014)

dataset (fine-tuning). This model aims to combine the general-purpose knowledge, with the specialized knowledge for unstructured environments acquired by the proposed dataset

- The optimized SuperPoint trained from scratch, using the proposed dataset, aiming to focus on planetary environments

Both, original and optimized SuperPoint models were trained under the same parameterization. For each model, the MagicPoint which is the standalone detector of SuperPoint, was trained for three rounds applying 18 000 iterations with batch size equal to 32 and homographic adaptation enabled. Subsequently, SuperPoint was trained for 250 000 iterations with batch size equal to 2 with homographic adaptation disabled due to high demands on computing resources. The Adam optimizer with default learning rate equal to 0.001 were utilized while the image input size that was used is 240 x 320 in grayscale.

Before each round of training, the weights from the last round are used to extract the pseudo-ground truth of the dataset which is subsequently used in the next round of training. It's worth noting that in the first round, the pseudo-ground truth is extracted using the weights based on a MagicPoint model, trained with the synthetic shapes dataset.
Regarding the computing resources, an Intel i7-4771 CPU with 3.50GHz × 8 combined with an NVIDIA GeForce GTX 1080 Ti GPU were utilized while an external hard drive of 3.5 inches and a size of 4TB was used for retrieving and storing data during the training.

4.1.2 Evaluation and results of SuperPoint models

In this section, the implemented SuperPoint models focused on unstructured environments, are evaluated in terms of keypoint detection and description, compared with well-known and widely used algorithms and the pre-trained SuperPoint model.

The evaluation is conducted using the benchmark dataset (3.1.4.2) designed for planetary and unstructured scenes while the repeatability and homography estimation metrics are utilized for the evaluation of keypoint detection and description respectively.

Regarding the evaluation of keypoint detection, the produced models are compared with the algorithms SHI (Shi & Tomasi 1993), Harris (Harris & Stephens 1988), and FAST (Rosten & Drummond 2006) implemented with OpenCV library (Bradski 2000) and the original SuperPoint model, trained with 80 000 general-purpose images of COCO dataset. The repeatability metric, which determines the efficiency of the model to detect the same keypoints in different image representations of the same scene, was estimated using 300 detected points as the maximum limit and threshold of correctness ε=3 pixels (table 4.1).

Concerning the evaluation of keypoint description, the produced models are compared with ORB (Rublee et al. 2011), SIFT (Lowe 2004) and the original SuperPoint pre-trained with COCO dataset. The homography estimation metric was utilized, based on nearest neighbor matching of keypoints and the corresponding descriptors between an original image and a different representation of same image while 1000 detected points were utilized with correctness threshold ε=3 (table 4.2).

Keypoint detectors	Rep. (i)	Rep. (v)
FAST	0.72	0.61
Harris	0.75	0.73
SHI	0.74	0.61
Original SuperPoint (Pre-trained)	0.85	0.63
Original SuperPoint (Trained from scratch)	0.83	0.65
Original SuperPoint (Fine-tuning)	0.83	0.65
Optimized SuperPoint (Trained from scratch)	0.82	0.66

Table 4.1 Evaluation of keypoint detectors based on illumination (i) and viewpoint (v) changes in planetary and unstructured environments, using repeatability metric with ε=3

Descriptors	Homography estimation (i)	Homography estimation (v)
ORB	0.82	0.53
SIFT	0.97	0.96
Original SuperPoint (Pre-trained)	0.98	0.81
Original SuperPoint (Trained from scratch)	0.99	0.85
Original SuperPoint (Fine-tuning)	0.98	0.84
Optimized SuperPoint (Trained from scratch)	0.99	0.87

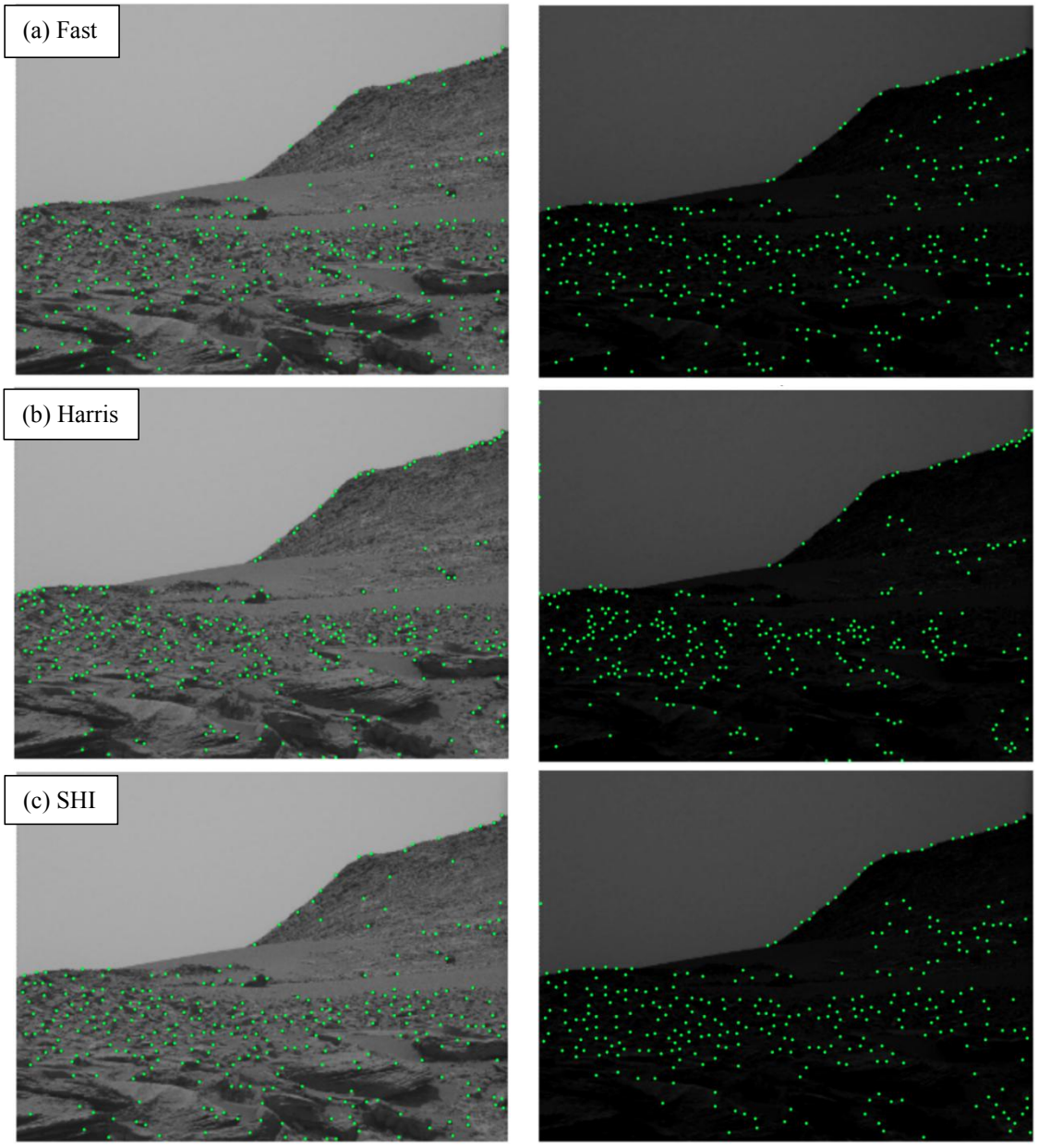
Table 4.2 Evaluation of keypoint descriptors based on illumination (i) and viewpoint (v) changes in planetary and unstructured environments, using homography estimation with ε=3

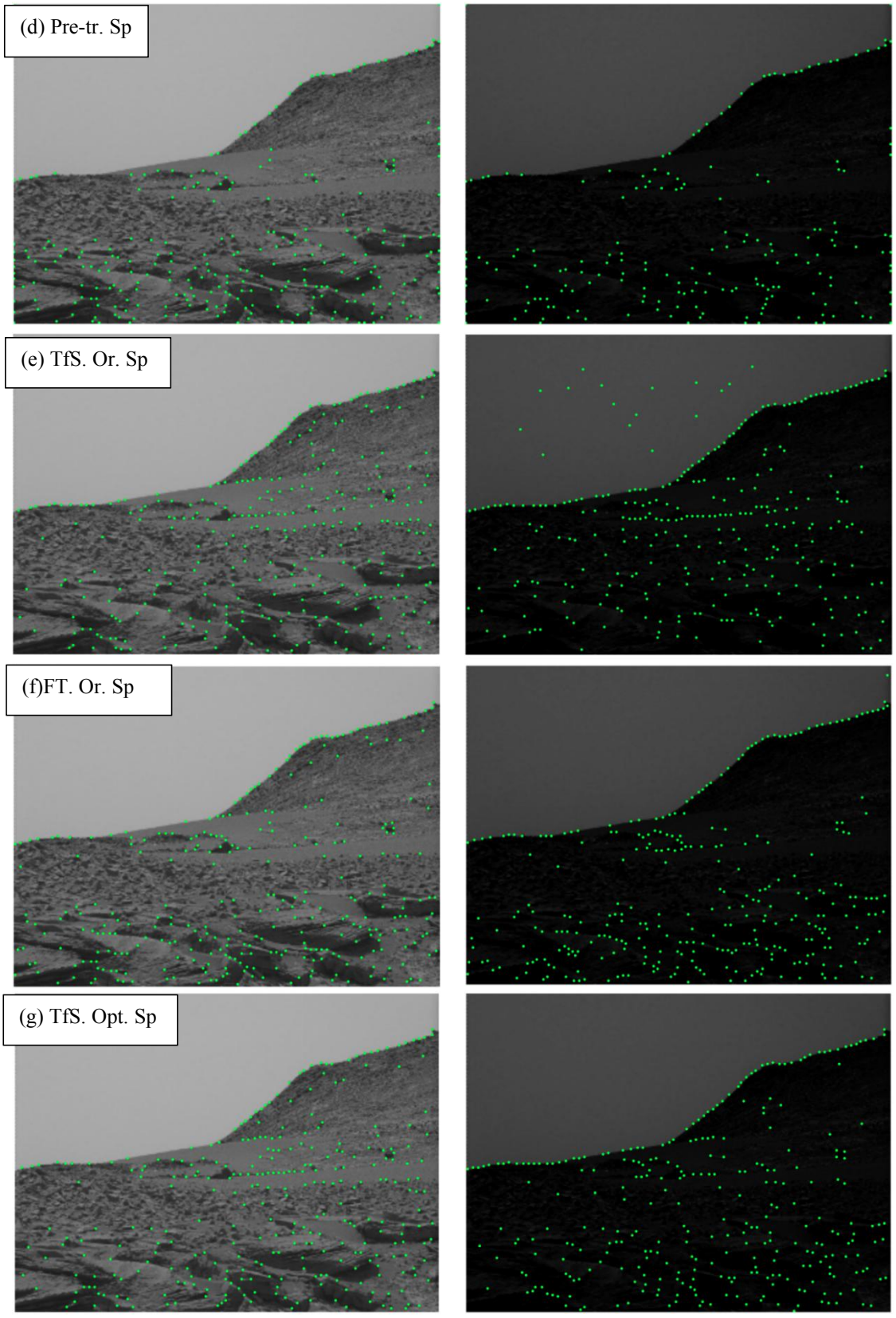
As presented in table 4.1, the optimized and original SuperPoint models trained and fine-tuned with the proposed dataset, provided similar repeatability of 0.82 and 0.83 respectively in terms of illumination changes, outperforming the SHI, Harris and FAST detectors, while the pre-trained SuperPoint model achieves the highest repeatability equal to 0.85. Instead, the optimized SuperPoint model outperforms SHI, FAST and all the original SuperPoint models, (the pre-trained model and trained from scratch with the proposed dataset) in terms of viewpoint changes, achieving a repeatability score equal to 0.66. It's worth noting that Harris detector provides the highest repeatability in terms of viewpoint changes, equal to 0.73.

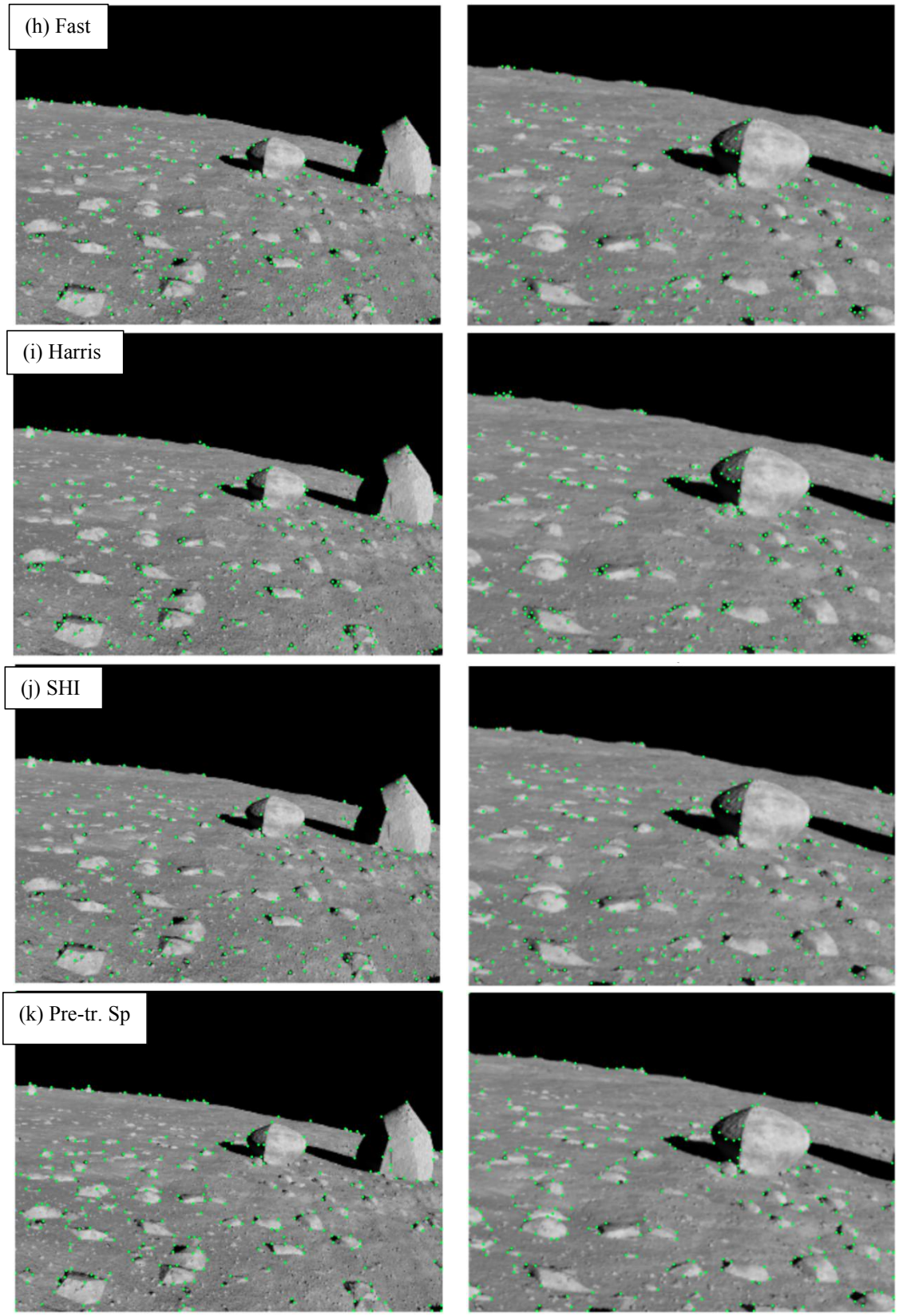
As presented in table 4.2, the optimized and original SuperPoint models provide the highest homography estimation in terms of illumination changes (0.99) outperforming the descriptors ORB, SIFT and the original pre-trained and fine-tuned SuperPoint

models. In terms of viewpoint changes, the SIFT algorithm provides high accuracy in a level of 0.95 while the optimized SuperPoint model, achieves homography estimation equal to 0.87, outperforming ORB and all the original SuperPoint models (the trained and fine-tuned with the proposed dataset models and the pre-trained model).

Qualitative results of keypoint detection and description evaluation, are depicted in figures 4.1, 4.2.




(d) Pre-tr. Sp
(e) TfS. Or. Sp
(f)FT. Or. Sp
(g) TfS. Opt. Sp


(h) Fast
(i) Harris
(j) SHI
(k) Pre-tr. Sp

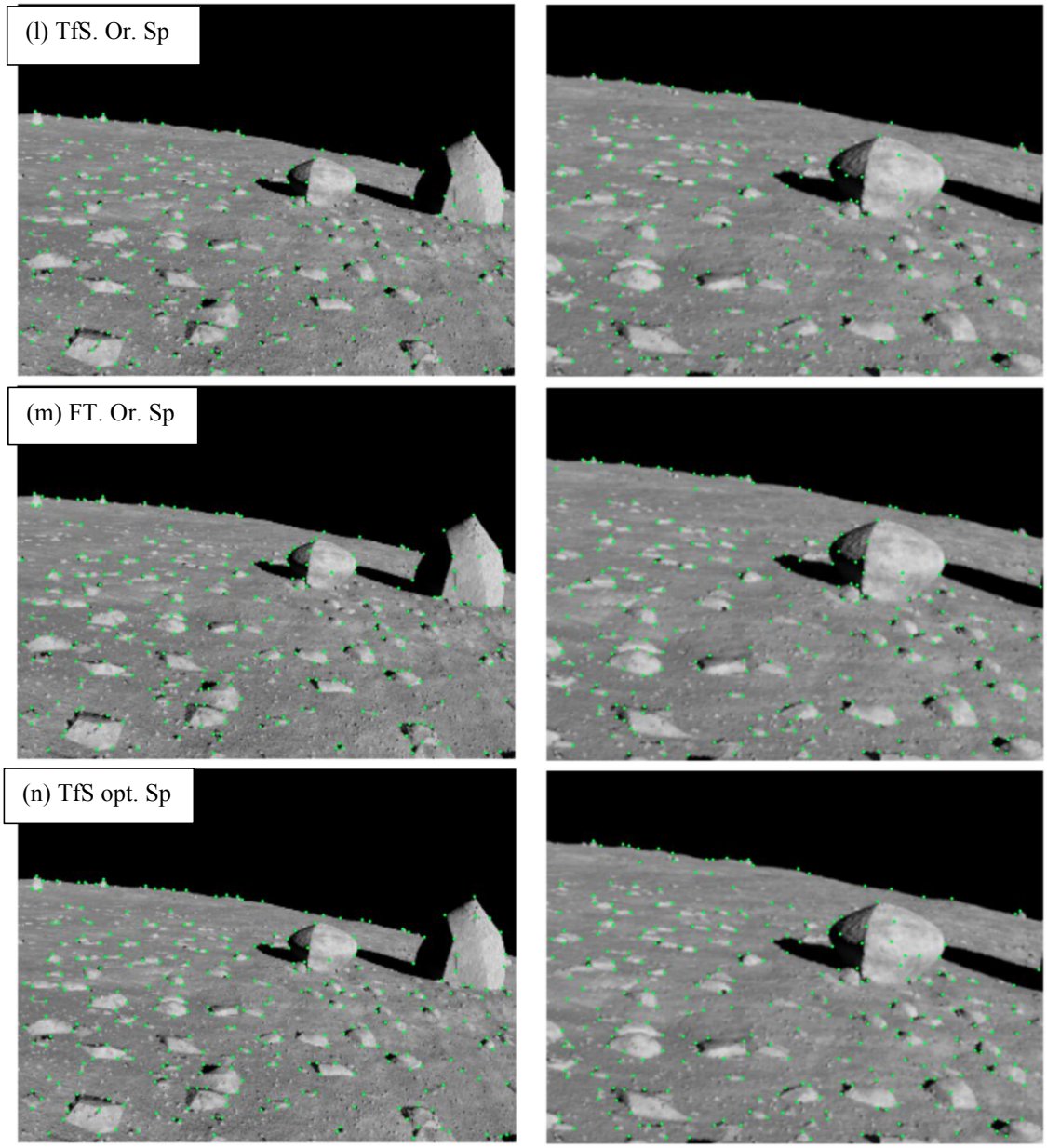
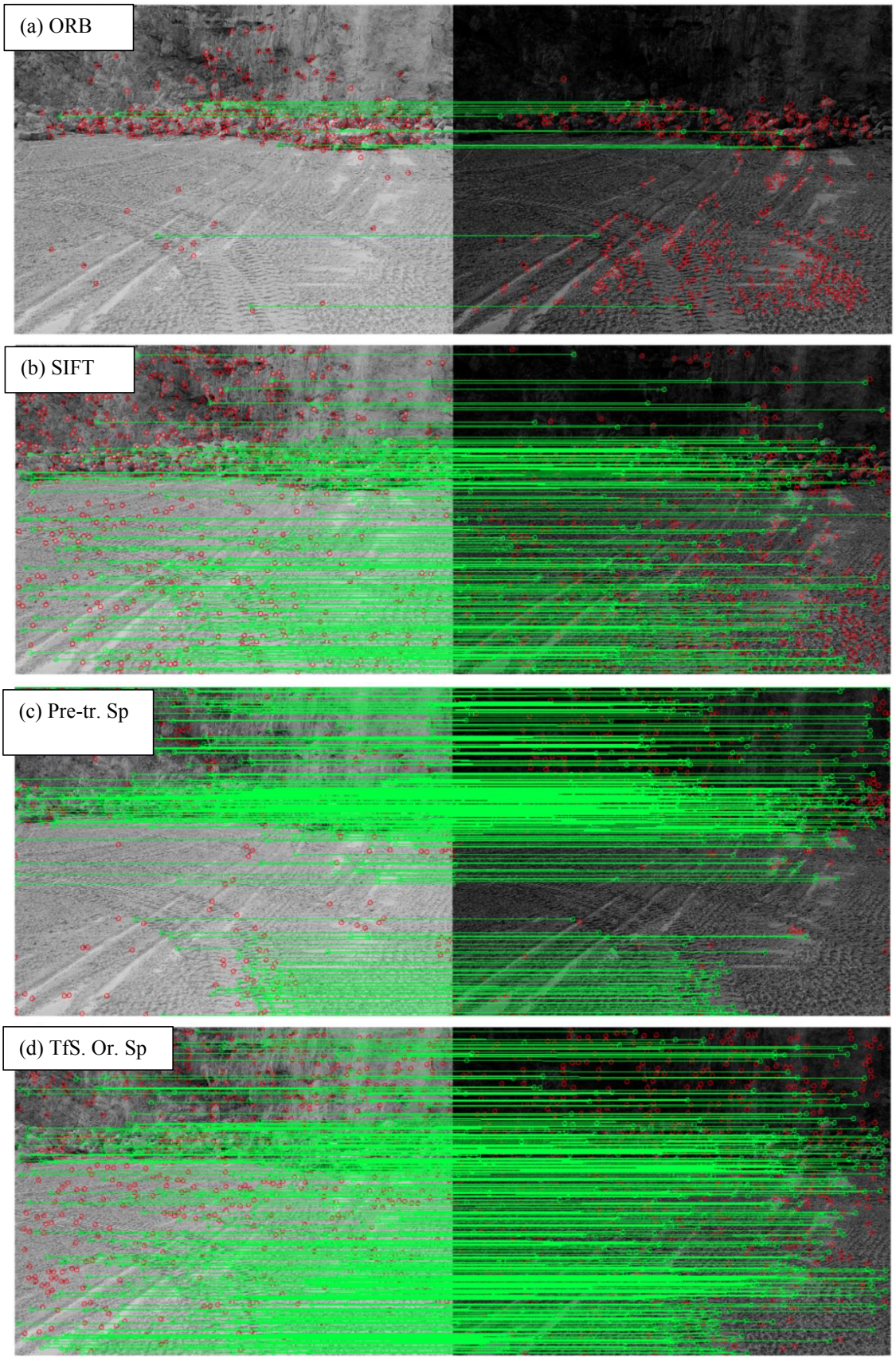
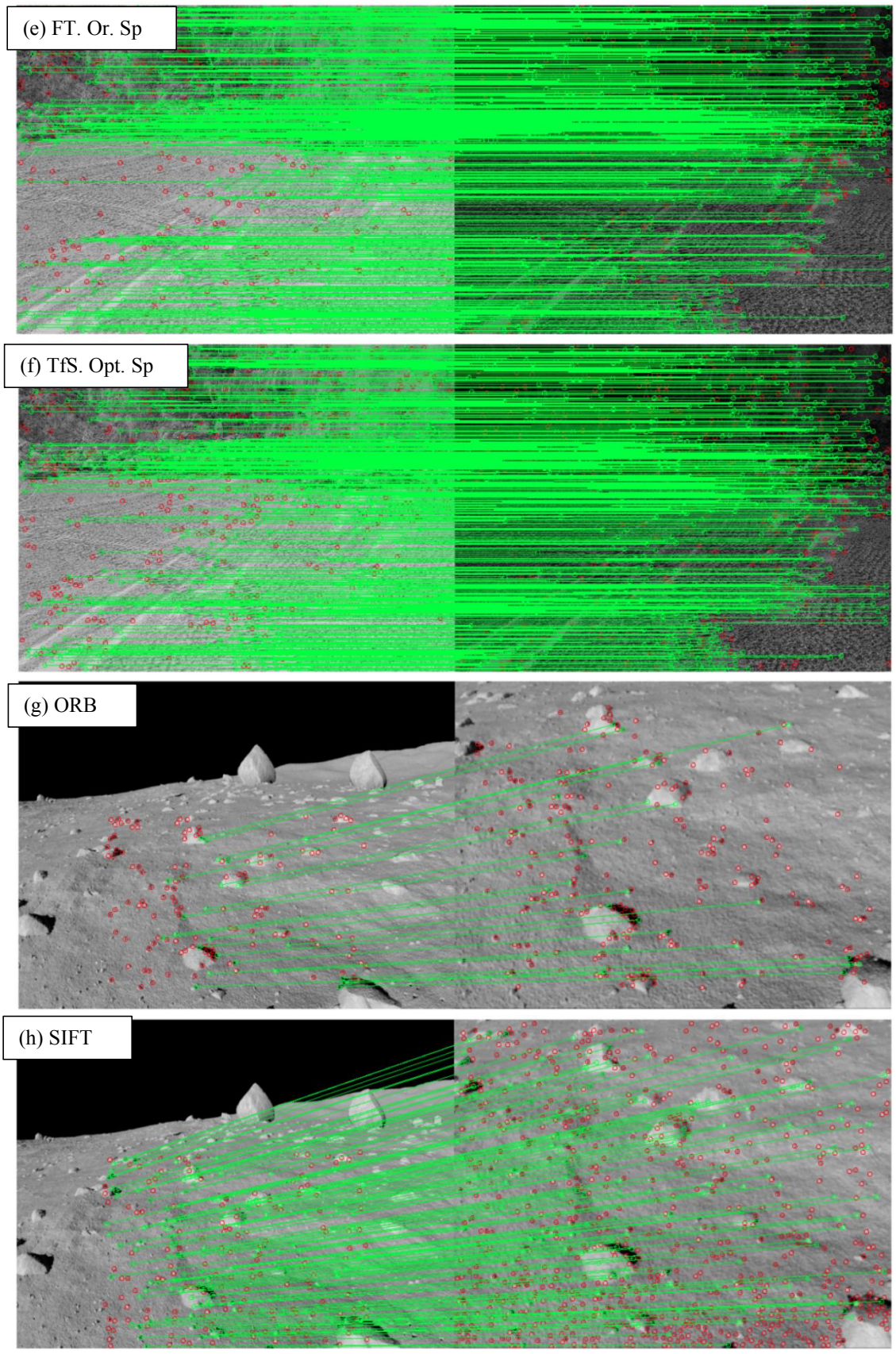


Figure 4.1 a - g: Detected keypoints in two images from a scene of Mars with different levels of illumination: (a) FAST, (b) Harris, (c) SHI, (d) Pre-trained SuperPoint, (e) original SuperPoint, trained from scratch with the proposed dataset, (f) original SuperPoint, fine-tuned with the proposed dataset, (g) optimized SuperPoint, trained from scratch with the proposed dataset. **h-n**: Detected keypoints in two images from the same scene of artificial lunar surface with different viewpoints: (h) FAST, (i) Harris, (j) SHI, (k) Pre-trained SuperPoint, (l) original SuperPoint, trained from scratch with the proposed dataset, (m) original SuperPoint, fine-tuned with the proposed dataset, (n) optimized SuperPoint, trained from scratch with the proposed datase


(a) ORB
(b) SIFT
(c) Pre-tr. Sp
(d) TfS. Or. Sp


(e) FT. Or. Sp
(f) TfS. Opt. Sp
(g) ORB
(h) SIFT

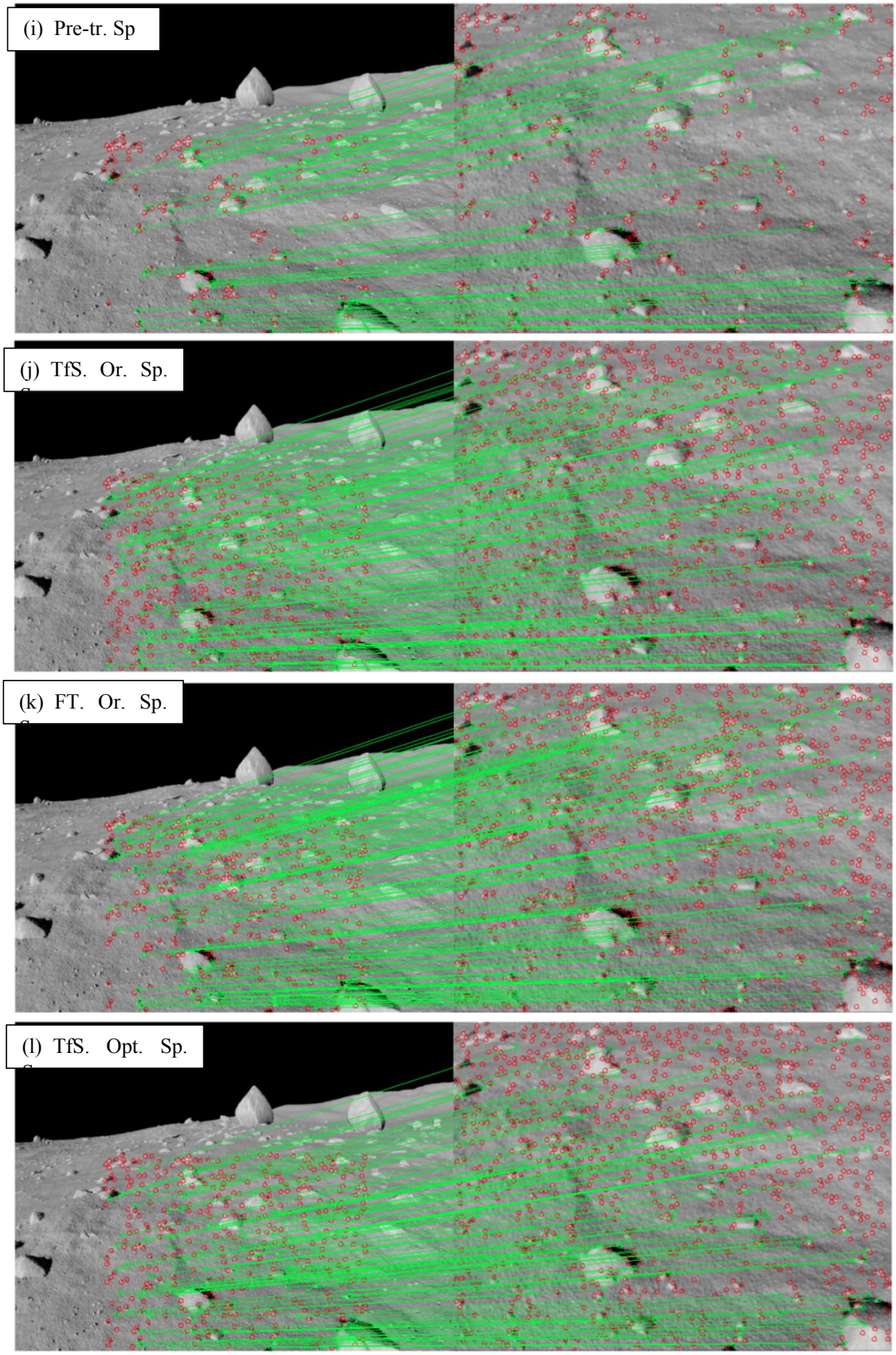


Figure 4.2 a - f: Keypoint matches in two images from an earthy scene in different levels of illumination: (a) ORB, (b) SIFT, (c) Pre-trained SuperPoint, (d) original SuperPoint, trained from scratch, (e) original fine-tuned SuperPoint, (f) optimized SuperPoint, trained from scratch. g-l: Keypoint matches in two images from the same lunar scene in different viewpoints: (g) ORB, (h) SIFT,

(i) Pre-trained SuperPoint, (j) original SuperPoint, trained from scratch, (k) original fine-tuned SuperPoint, (l) optimized SuperPoint, trained from scratch

As observed in both figures (4.1, 4.2), the trained from scratch original and optimized SuperPoint models, using the proposed dataset, provide high accuracy and sensitivity in feature-poor scenes with illumination and viewpoint changes, outperforming the handcrafted algorithms and the pre-trained SuperPoint model. The fine-tuned SuperPoint, provides refined results compared with the pre-trained SuperPoint but it is not as accurate as the trained from scratch models.

4.2 Implementation and results of HF-net2 architecture and SLAM

In this section, the implementation and results of HF-net2 architecture is analyzed while afterwards an extended experimentation of the proposed SLAM system is presented.

4.2.1 Training process

As referred in chapter 3, HF-net2 uses a multi-task distillation approach for the training process using SuperPoint and NetVLAD as the teachers for keypoint detection, local and global description respectively. Utilizing this self-distillation process, there is no need for labeled data, since the labeling of the dataset is implicitly conducted by the teachers which provide the corresponding ground truth to the student network.

During the experimentation, the following two models were produced:

- HF-net2: The proposed architecture was trained from scratch using the proposed dataset with the FPV images from Earth, Mars and Moon.
- HF-net: The original HF-net (Sarlin *et al* 2019) was trained from scratch with the proposed dataset.

Both models were trained for 30 000 iterations while, the RMSProp optimizer was utilized with learning rate in a range of 0.001 - 0.00001.

The training process of the multi-teacher-student architecture is conducted using the following loss function:

$$L = e^{-w_1}\left\|d_s^g - d_{t_1}^g\right\|_2^2 + e^{-w_2}\left\|d_s^l - d_{t_2}^l\right\|_2^2 + \\ + 2e^{-w_3}\mathrm{CrossEntropy}(p_s, p_{t_3}) + \sum_i w_i \tag{4.1}$$

Where $\left\|d_s^g - d_{t_1}^g\right\|_2$ is the L2 norm of student (s) and NetVLAD (t_1) global descriptors while $\left\|d_s^l - d_{t_2}^l\right\|_2$ is the L2 norm of student (s) and SuperPoint (t_2) local descriptors. The $w_{1,2,3}$ represent optimized variables while p_s and p_{t3} represent the keypoint scores of student (s) and SuperPoint (t_3) respectively.

For the training process, an Intel i7-4771 CPU with 3.50GHz × 8 combined with an NVIDIA GeForce GTX 1080 Ti GPU were utilized, while the implementation of the architecture was conducted using the TensorFlow (Abadi *et al.* 2015) deep learning platform.

4.2.2 Evaluation and Results of HF-net2

To evaluate HF-net2 model, the proposed evaluation dataset was utilized. As referred in section 3.1.4 the dataset includes 120 sequences of images from Earth, Moon and Mars, designed for evaluation in terms of illumination and viewpoint changes. The performance of HF-net2 in keypoint detection and description, was tested and compared with several well-known algorithms for keypoint detection and description.

Regarding the keypoint detection, the repeatability and mAP (mean Average Precision) metrics were utilized in image sequences which includes image representations with different illumination (i) or viewpoint (v) (table 4.3). The repeatability measure the percentage of keypoints that are repeatable in different image representations of the same scene while mAP utilizes the precision $\frac{\#\ correct\ matches}{\#\ matches}$ and recall $\frac{\#\ corrent\ matches}{\#\ correspondences}$ curve, aiming to form a reliable metric for the accuracy of the algorithms. Similarly in the description part, the matching score, which is the percentage of the correct matching points out of a pre-defined number of detected points (e.g 300), and the mAP are utilized, aiming to evaluate the proposed descriptor in terms of illumination (i) and viewpoint (v) changes (table 4.4).

Keypoint detectors	Rep. (i)	mAP (i)	Rep. (v)	mAP (v)
SIFT	0.48	0.24	0.54	0.26
FAST	0.65	0.46	0.61	0.38
Harris	0.71	0.55	0.77	0.57
SuperPoint	0.82	0.77	0.76	0.67
HF-net (original)	0.72	0.68	0.69	0.47
HF-net2 (proposed)	0.74	0.71	0.69	0.49

Table 4.3 Evaluation of HF-net2 as a keypoint detector in terms of intense illumination (i) and viewpoint (v) changes using repeatabilitity metric

Keypoint descriptors	Matching score (i)	mAP (i)	Matching score (v)	mAP (v)
SIFT	0.51	0.87	0.54	0.83
ORB	0.46	0.61	0.36	0.34
SuperPoint	0.81	0.99	0.71	0.99
HF-net (original)	0.72	0.98	0.58	0.94
HF-net2 (proposed)	0.74	0.98	0.63	0.95

Table 4.4 Evaluation of HF-net2 as a descriptor in terms of intense illumination (i) and viewpoint (v) changes using mAP an matching score metrics

As presented in table 4.3, regarding the illumination changes, SuperPoint which is the teacher of HF-net and HF-net2, achieves the highest repeatability and mAP with values 0.82 and 0.77 respectively while the proposed HF-net2 follows with the next most accurate results with repeatability and mAP with values 0.72 and 0.68 respectively. The original HF-net provides lower accuracy in terms of repeatability and mAPs compared with SuperPoint and HF-net2, while the non-learning algorithms noted significantly decreased accuracy. Concerning the viewpoint changes, Harris achieves the highest repeatability with a value of 0.77 while SuperPoint provides the highest overall accuracy with repeatability 0.76 and mAP 0.67. The HF-net and HF-net2 architectures provide respectable accuracy while the proposed HF-net2 achieves slightly higher mAP (0.49) than HF-net (0.47). SIFT and FAST noted significantly lower accuracy than Harris algorithm and learning-based architectures.

Regarding the evaluation of descriptors (table 4.4), SuperPoint achieves the highest matching score and mAP both in illumination and viewpoint changes while the proposed architecture provides the next most accurate results with matching score in a level of 0.75 and 0.65 and mAP 0.98 and 0.95, in illumination and viewpoint changes respectively, outperforming the original HF-net and the traditional descriptors SIFT and ORB. It's worth noting that the superiority of SuperPoint proves that it is a robust keypoint detection and description architecture, capable of being a teacher of HF-net during the training process.

Concerning the performance-time of the proposed model was tested in several frame frame resolutions including the following:

- 560 x 500 pixels
- 640 x 480 pixels
- 720 x 480 pixels
- 1920 x 1080 pixels

In the resolutions of 560x500, 640x480 and 720x480 pixels, that are quite common in sensors focused on robotics, the model achieves 16 ms (milliseconds) inference-time per frame and 62.5 FPS (Frames per Second) in a GPU-enabled machine while in the resolution of 1920 x 1080 pixels the inference time increases in 70 ms and 14.28 FPS. It's worth noting that SuperPoint and the handcrafted algorithms provide similar inference time but the HFnet model exports not only keypoint scores and local descriptions but also global descriptions in the same time. The aforementioned results in inference-time are considered satisfactory and prove that the model is adequate for real-time applications.

4.2.3 Evaluation of the proposed SLAM system

As referred in chapter 3, the proposed HF-net2, was integrated in a SLAM system as feature extraction module aiming to increase the sensitivity of the system in multiple and accurate keypoint detections and descriptions in order to encounter the issue of illumination changes and lack of visual cues in unstructured and planetary scenes. The proposed SLAM is based on ORB-SLAM2 (Mur-Artal & Tardos 2017) and is built on

C++ and Python programming languages while RGB images combined with depth information were utilized for accurate scale estimation.

An extended experimentation was conducted in two environments with rocky and sandy terrains respectively while in each experiment, videos and ground truth data were captured aiming to evaluate the accuracy of the predicted camera trajectory. Regarding the equipment, the RGB and depth sensor of Intel RealSense D435 camera was utilized, while the CHCNAV i73 RTK GNSS receiver was used in order to measure the coordinates of the camera trajectory in a geodetic reference system.

More specifically, two sources of data were captured in each experiment:

- A rosbag, a file format in ROS (Robot Operating System) (Stanford Artificial Intelligence Laboratory et al. 2018), which includes XYZ coordinates, orientation and optical center with respect to the world origin of the SLAM coordinate system, captured in a video of 30 FPS with resolution 848 x 480.
- GNSS data which contains XYZ coordinates in GGRS-87 geodetic coordinate system, captured with a frequency of one measurement per second.

Regarding the camera data, initially RGB and depth frames are extracted from the rosbag file and synchronized, so as each timestamp to correspond in a specific RGB and depth frame while afterwards the proposed SLAM estimates the camera trajectory in TUM format (Schubert *et al.* 2018) using the RGB-depth information. Concerning the GNSS data, are transferred in a local coordinate system with origin the starting point of the trajectory and formed in a TUM format. Subsequently, both data sources are synchronized aiming each timestamp to be related with a specific location in SLAM and GNSS data. Finally, the predicted and ground-truth trajectories are compared, estimating the accuracy of the predicted trajectory. The pipeline of the SLAM evaluation process is presented in fig. 4.3.

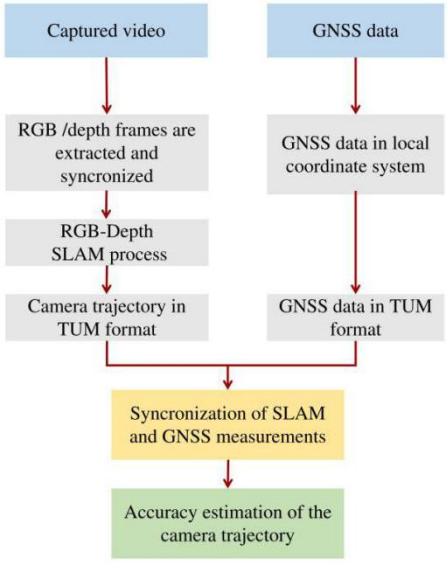


Figure 4.3 Pipeline of the SLAM evaluation process

4.2.3.1 Experiments and results

As referred above, the experiments were conducted in two different environments: a rocky and a sandy scene (fig 4.4) while different cases in terms of illumination were performed. The proposed SLAM system was compared with the ORB-SLAM2 (Mur-Artal & Tardos 2017), one of the most popular SLAM systems**,** aiming to evaluate the added value of the HF-net2 model trained in unstructured environments instead of a traditional keypoint detector and descriptor such as ORB.

Figure 4.4 Left image: Rocky scene, right image: sandy scene

More specifically, the experiments were performed in day and evening time using different trajectory paths and natural light, while an experiment was conducted with artificially low illumination with extremely fast lighting changes (table 4.5).

Experiments	Scene	Day-time	Illumination	Light
Square-based path	Rocky terrain	10:00 a.m	High	Natural
Square-based path	Rocky terrain	7:00 p.m	Medium to low	Natural
right-angle based path	Rocky terrain	5:00 p.m	Medium	Natural
Random path	Sandy terrain	10:00 a.m	High	Natural
Random path	Sandy terrain	--	Very low - Low	Artificial

Table 4.5 Experiments, performed in different scenes, trajectory paths and illumination conditions

The results of the experiments above, are presented in the tables 4.6 - 4.10. The metrics RMSE (Root-Mean-Squared-Error) and standard deviation with max and min errors are used, calculated using the GNSS-based data as a reference, aiming to determine the SLAM systems' accuracy. The corresponding visual results, with the predicted and GNSS-based trajectories, are presented in figure 4.5.

	RMSE (m)	Std. Dev. (m)	Max (m)	Min (m)
ORB-SLAM2	0.10	0.03	0.19	0.02
HF-net2-based SLAM	0.11	0.04	0.22	0.03

Table 4.6 Square-based path in rocky terrain with high illumination

	RMSE (m)	Std. Dev. (m)	Max (m)	Min (m)
ORB-SLAM2	0.15	0.06	0.32	0.03
HF-net2-based SLAM	0.12	0.05	0.19	0.03

Table 4.7 Square-based path in rocky terrain with medium to low illumination

	RMSE (m)	Std. Dev. (m)	Max (m)	Min (m)
ORB-SLAM2	0.09	0.05	0.20	0.01
HF-net2-based SLAM	0.08	0.04	0.16	0.008

Table 4.8 Right angle-based path in rocky terrain with medium illumination

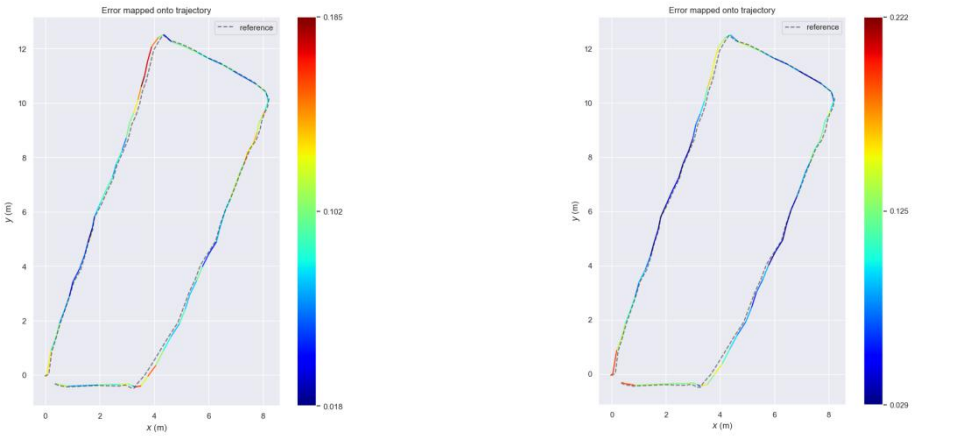
	RMSE (m)	Std. Dev. (m)	Max (m)	Min (m)
ORB-SLAM2	0.34	0.12	0.58	0.11
HF-net2-based SLAM	0.21	0.07	0.35	0.04

Table 4.9 Random path in sandy terrain with high illumination

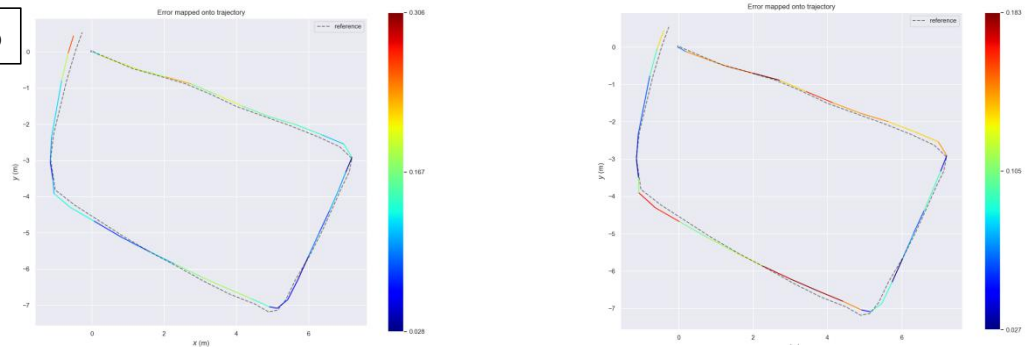
	RMSE (m)	Std. Dev. (m)	Max (m)	Min (m)
ORB-SLAM2	0.50	0.17	0.85	0.21
HF-net2-based SLAM	0.24	0.06	0.36	0.07

Table 4.10 Random path in sandy terrain with artificially quite low illumination which changes during the SLAM process with a range of extremely low to low lighting conditions

a



b



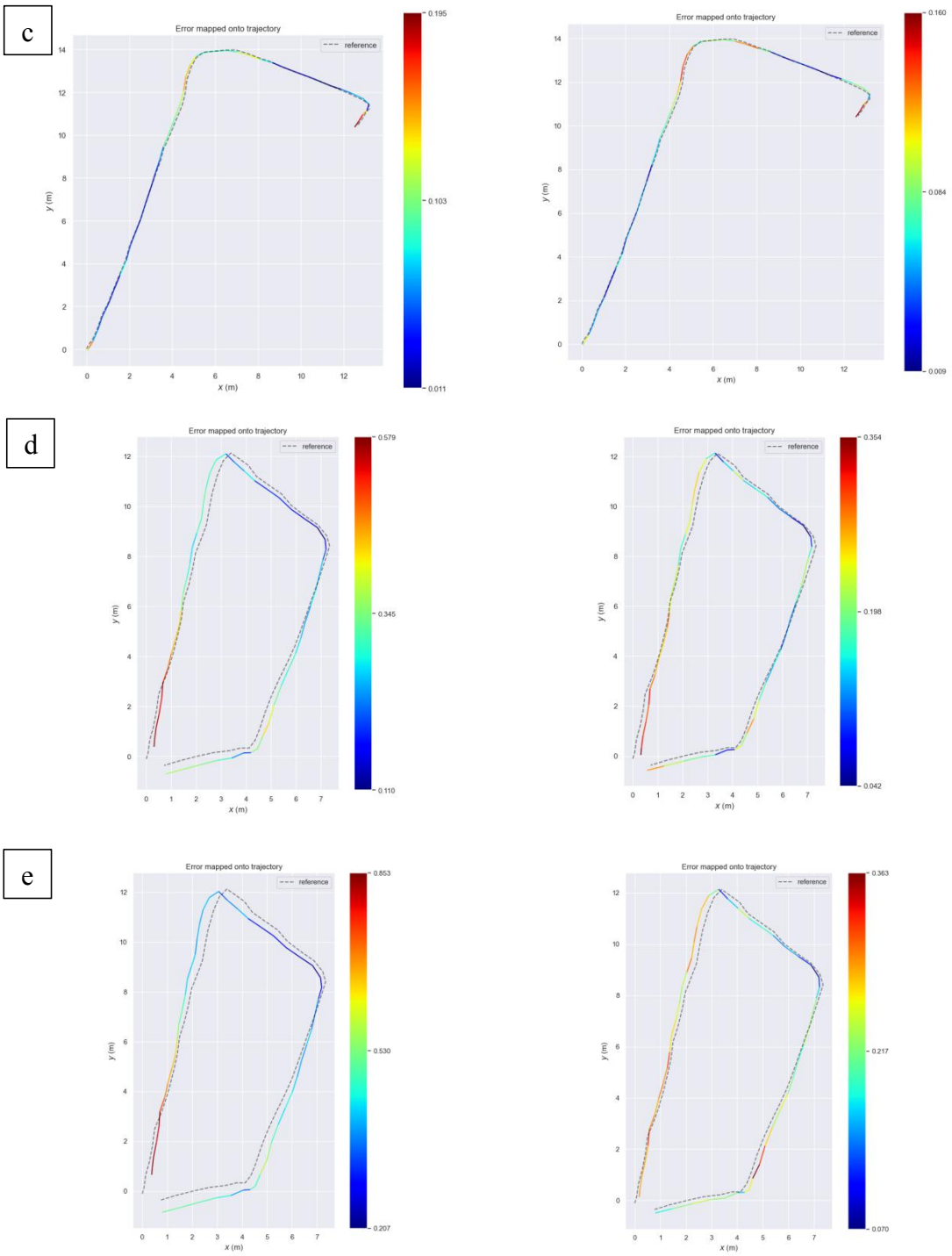


Figure 4.5 Predicted trajectories of the ORB-SLAM2 (left column) and proposed SLAM (right column) compared with the ground truth trajectory (presented as gray dashed line). (a) rocky terrain with high illumination, (b) rocky terrain with medium to low illumination, (c) rocky terrain with medium illumination, (d) sandy terrain with high illumination, (e) sandy terrain with artificially low illumination

Regarding the rocky scene, the proposed SLAM and ORB-SLAM2 provide similar accuracy under normal conditions as presented in tables 4.6 and 4.8, since ORB-SLAM2 slightly outperforms the proposed SLAM in square-based path with high illumination while the reverse occurs in the right-angle based path with medium

illumination (fig 4.5a, fig 4.5c). However, in square-based path with low illumination (table 4.7), the proposed SLAM provides higher accuracy with RMSE error in a value of 0.12 with maximum error equal to 0.19 m instead of ORB-SLAM2 which produced a maximum error equal to 0.32 m (fig 4.5b).

Concerning the sandy scene, the proposed SLAM provides significant higher accuracy than ORB-SLAM2 in high illumination with RMSE 0.21 m and standard deviation 0.07 instead of ORB-SLAM2 which provides RMSE 0.34 and standard deviation 0.12 respectively (table 4.9, fig. 4.5d). In the last experiment, the same data frames were processed using GAMMA correction (eq. 4.2) aiming to highly decrease the illumination in a specific range.

$$\text{Output} = \left(\frac{\text{I}}{255}\right)^{\frac{1}{\gamma}} 255 \quad (4.2)$$

Where I is the input pixel value and γ the gamma parameter which controls the image brightness. The gamma values below 1 (gamma < 1) produce darker images while gamma values above 1 (gamma > 1) produce brighter images than the original image. In this experiment all the recorded frames were processed aiming to generate frames with low illumination using uniformly random gamma values between 0.2 - 0.4 (fig 4.6).

As a result, the SLAM systems encounter a scene which lack of significant visual cues in a quite low illumination environment with changing lighting conditions in each frame. However, the proposed SLAM, maintained its accuracy with RMSE 0.24 m and standard deviation 0.06 m, instead of ORB-SLAM2 accuracy which is further decreased with RMSE 0.50 m and standard deviation 0.12 m. It's worth noting that the maximum and minimum errors of the proposed SLAM is 0.36 and 0.07 respectively while the corresponding errors of ORB-SLAM2 are 0.85 and 0.21 (table 4.10, fig. 4.5e).

Figure 4.6 Right: Original image, middle: darken image with gamma=0.4, right: darken image with gamma=0.2

4.3 Implementation and results of the proposed NN for semantic segmentation

In this section, the implementation of the proposed modified U-net architecture is described while afterwards, the evaluation and results of the model for lunar ground semantic segmentation, are presented.

4.3.1 Training process of modified U-net

The proposed architecture was implemented using Python and Keras / TensorFlow deep learning library (Chollet *et al.* 2015) while several Python libraries including NumPy (Harris *et al.* 2020), Matplotlib (Hunter 2007) and Scikit-learn (Pedregosa *et al.* 2011) were utilized.

The main goal of the architecture is to detect and localize rocks and boulders while in order to segment the whole scene, three classes are taken into account: rocks, sky and background. The training data which constitute the 70% of the lunar landscape dataset feed the modified U-net while the remaining 30% of the dataset is used for the validation and testing. The model was trained for 15 epochs using early stopping technique while the batch size was defined equal to 16. The categorical cross entropy loss function and Adam optimizer with a learning rate of $5x10^{-5}$ were utilized. Regarding the input size, the dimensions of 480 x 480 pixels was used, since it was observed that a larger image size provided more refined results than the widely used size of 256 x 256 pixels.

The training and validation process were conducted in a machine with Intel i7- 3.50 GHz x 8 cores of CPU, 16 Gb of RAM and NVIDIA GTX 1080 Ti of GPU with CUDA version 11.2 enabled.

4.3.2 Evaluation and Results of modified U-net

The proposed architecture was trained and validated using Dice-coefficient and Recall metrics which are defined with the following formulas:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.3)$$

$$\text{Dice} = \frac{\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \quad (4.4)$$

Where, TP stands for true positive while FN and FP stand for false negative and false positive. The results of dice coefficient and recall after the training process are presented in table 4.11 while the learning curves of loss function and dice coefficient are depicted in fig 4.7.

	Loss	Dice-coef	Recall
Training	0.07	0.79	0.98
Validation	0.06	0.78	0.98

Table 4.11 Loss function, dice-coefficient and recall after the training process

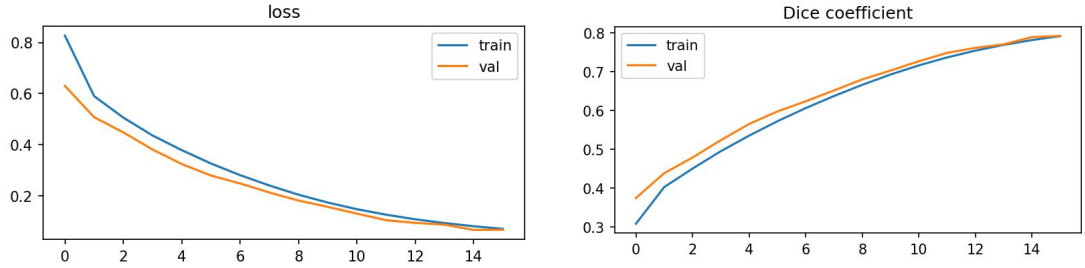
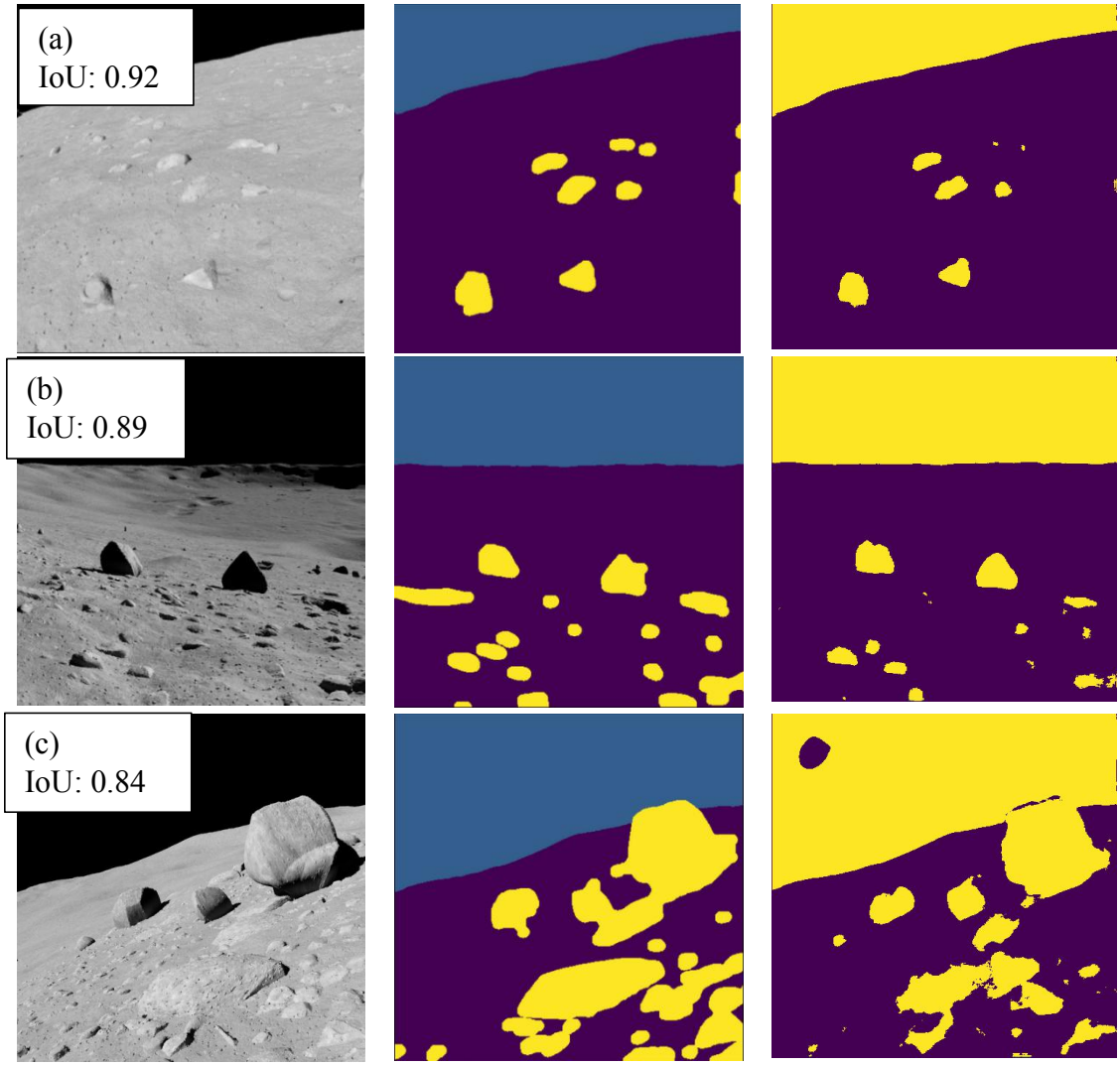


Figure 4.7 Loss and dice coefficient curves during training and validation

As observed in table 4.11, the value of loss function is below 0.1, the dice-coefficient is in a level of 0.80 while the recall is close to 1.0, indicating that the model will provide satisfactory results while in figure 4.7 the learning curves of the training and validation process for loss function and dice-coefficient are quite close after the sixth epoch without fluctuations proving that the model doesn't overfit.

After the training process, the proposed architecture was validated in testing data which are completely unknown for the model including images from the synthetic dataset and from real lunar landscape images while the corresponding qualitative results are presented in the figures 4.8 and 4.9.



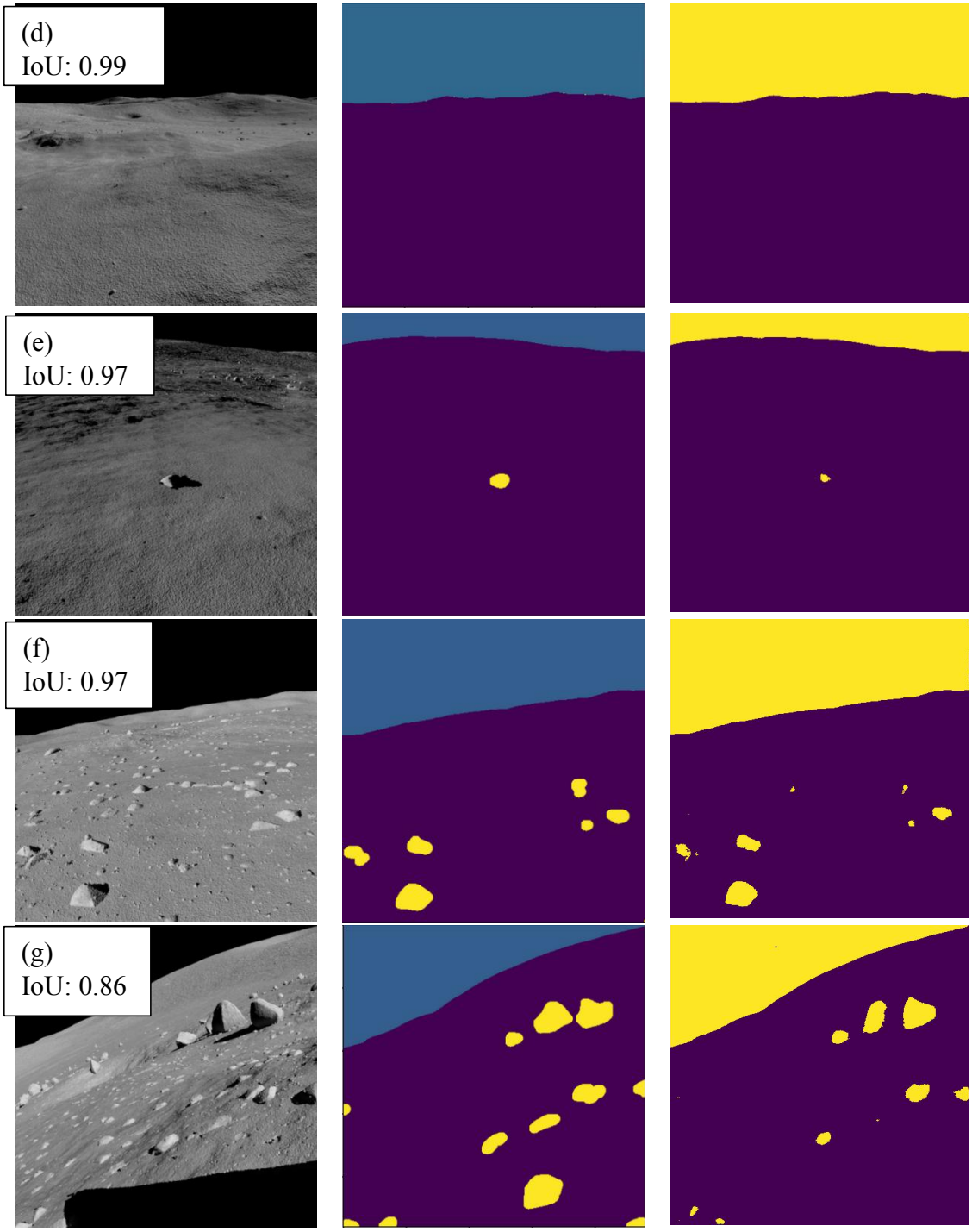
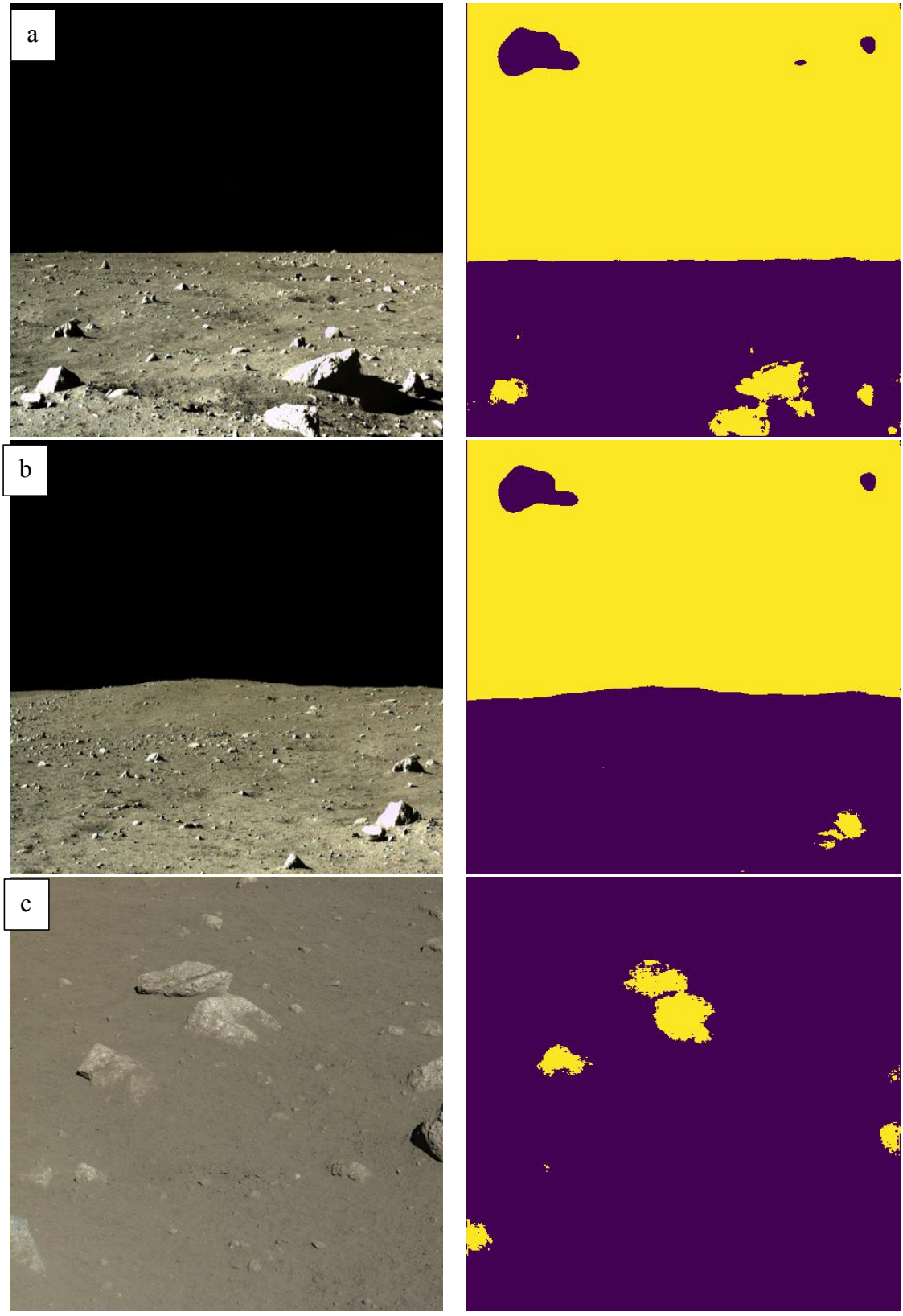


Figure 4.8 Left column: Original images from the synthetic lunar surface, (middle column) The corresponding annotated masks, (right column) Predictions of the proposed architecture. In each prediction (row) the IoU (Intersection over Union) metric is presented.


a
b
c

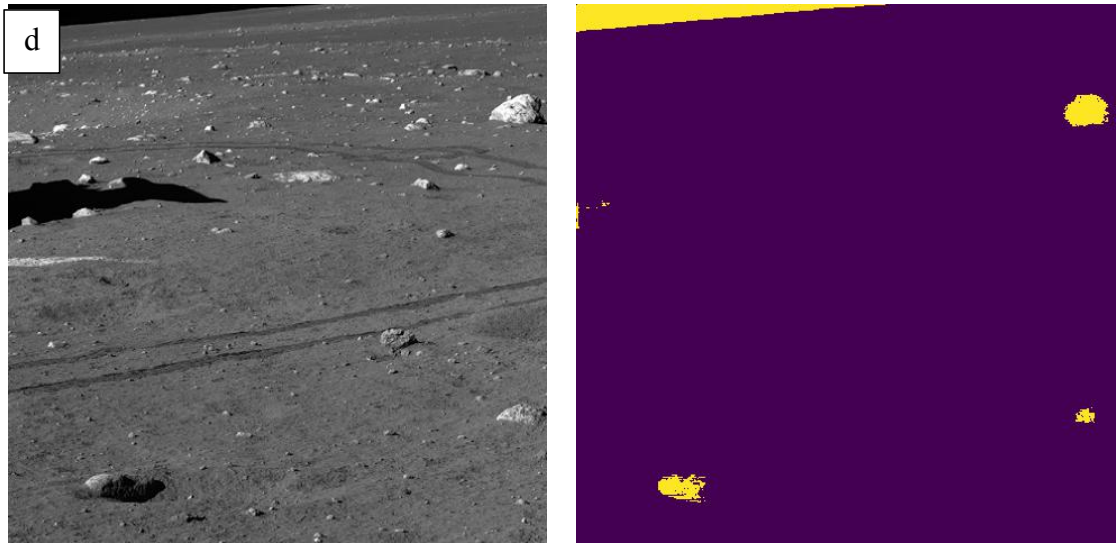


Figure 4.9 Left column: Real images from the lunar surface, (right column) Predictions of the proposed architecture. In each prediction (row) the IoU (Intersection over Union) metric is presented.

As observed in figure 4.8, the proposed architecture provides satisfactory results in testing data with synthetic images, achieving IoU (Intersection over Union) in a level of 0.85 or above. It is able to differentiate the sky from the ground region defining the horizon line with high accuracy while it precisely predicts the location of the small rocks and boulders on the lunar surface. It is not affected from the number of rocks that exist in the scene, since it is able to provide robust results in a scene without any or one rock (fig 4.8d, 4.8e) or with multiple small rocks and boulders (fig 4.8c).

Moreover, the proposed architecture achieves respectable results in real rover-based images (fig 4.9) which are quite different in terms of color and illumination compared with the training data while the model is not affected from the camera tilt, being efficient to identify rocks, either the camera targets on the horizon (fig 4.9ab) or on the ground (fig 4.9cd).

Regarding size of the model, it includes only 220,000 trainable parameters while the weights file size of the model is about 3.5 MB which is quite small for semantic segmentation models. The model was tested in terms of inference time for a set of images with a size of 480x480 pixels using three different computing setups: (a) a GPU-enabled conventional desktop machine, (b) CPU-only conventional desktop machine and (c) a CPU-only embedded system with quite low resources. The results are presented in table 4.12.

Inference time	Conventional machine /GPU-enabled		Conventional machine /CPU-only		Embedded system Rasp. Pi 4	
	ms	FPS	ms	FPS	ms	FPS
Proposed model	43	23.25	100	10	1080	0.92

Table 4.12 Inference time (in milliseconds and FPS) of the proposed model in a desktop GPU-enabled and CPU-only conventional desktop computer and in a CPU-only embedded system with low resources

As observed in the table 4.12, the model provides quite satisfactory inference time in the GPU-enabled machine achieving 43 ms inference time per image and about 23

FPS (Frames per second), while the model performs sufficiently without GPU (CPU-only) in the same machine, providing a performance time in a level of 100 ms per image and 10 FPS. The model was also tested on a Raspberry Pi 4 with 4 GB of RAM which is a CPU-only embedded system with quite low resources, providing inference time equal to 1080 ms and 0.92 FPS. Overall, the results are considered respectable taking into account that the image segmentation tasks require high-end GPU-enabled machines and prove that the model is able to be utilized in GPU-enabled or CPU-only conventional machines and embedded systems with low computing resources.

4.4 Implementation and results of the precise positioning and mapping in GNSS-denied environments

In this section, the implementation and pipeline of the precise positioning methodology is presented while afterwards the experimentation and results are analyzed.

4.4.1 System implementation

The methodology is a combination of SLAM, image processing and geometric transformations, implemented on C++ and Python programming languages. The system is able to receive as input ROS (Robot Operating System)-based files (rosbag files) (Stanford Artificial Intelligence Laboratory et al. 2018) since in case of real-time processing, ROS (Robot Operating System) is the leading open-source ecosystem for the robotic systems. Thus, initially, the system processes the rosbag files, extracting and synchronizing the RGB and depth frames.

Regarding the SLAM part, the proposed SLAM system (see 3.1.3) is used, while the integrated HF-net2 model (see 3.1.2), utilized as a feature extractor, has been trained and evaluated by TensorFlow (Abadi *et al.* 2015) deep learning library. Concerning the image processing for fiducial marker calculation, the ArUco (Romero-Ramirez *et al.* 2018, Garrido-Jurado *et al.* 2016) and OpenCV (Bradski 2000) libraries are utilized, in order to detect and identify the fiducial markers while regarding the geometric transformations, the PCL library (Rusu & Cousins 2011) is utilized for plane segmentation.

The overall procedure of mapping consists of three main stages. At first, python scripts extract the data from inside a Robot Operating System (ROS) bag file which is captured by an RGB-depth camera during the video recording process. The camera itself is used to obtain the calibration information from its factory settings while the image data streams (RGB and depth) are separated into frames and stored in two separate folders. The following step is the SLAM processing using the extracted frames as input, relating the image content with a camera trajectory and a point cloud. Finally, the fiducial markers of the scene are detected and the coordinate system with

origin, the origin marker is generated, producing the markers and point cloud coordinate estimations (fig 4.10).

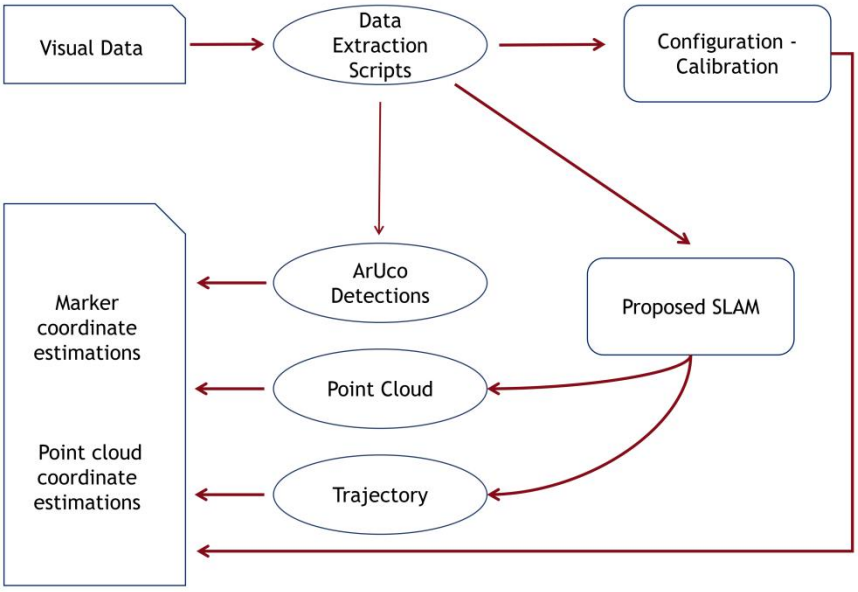


Figure 4.10 Pipeline of the overall end-to-end methodology

4.4.2 Equipment setup

The main equipment components include an RGB-depth camera, the ArUco markers, utilized as fiducial markers and a computing machine (embedded or conventional). Concerning the RGB-depth-camera, the Intel Realsense D435 camera was used which includes two infrared sensors (left and right), an RGB sensor and an infrared projector for the depth information (fig 4.11a). In the present study, only the RGB and depth sensor were used. The resolution of the RGB sensor is 1920 x 1080, the depth output is 1280 x 720, the focal length is 1.93 mm, while the format is 10-bit RAW.

The origin and the target markers are 30 x 30 cm in size while they are installed in a custom-made adjustable stand. This stand is able to stabilize the marker pose in a horizontal reference plane with the aid of two stainless steel threaded rods and a leveler (fig. 4.11b).

For validation purposes and ground-truth measurements, a Topcon GPT 3000 geodetic total station was used with ± (3mm+2ppm × D) mean square error (MSE) measurement accuracy where D is the measured distance between the total station and the prism (fig 4.11c).

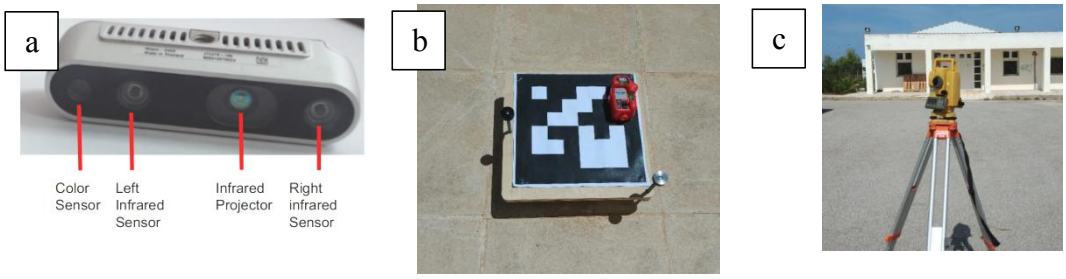


Figure 4.11. (a) the Intel Realsense D435 camera (b) The origin marker located on a custom-made adjustable stand which is able to stabilize the marker pose in a horizontal reference plane using two stainless steel threaded rods and a leveler. (c) The GTP-3000 geodetic total station

4.4.3 Experimentation and results

To validate the present methodology a set of experiments was performed in the following areas (fig. 4.12):

- an unstructured urban area (university campus)
- a sandy area
- a rocky area

All areas lack of feature-rich information while most of the experiments were conducted with high, medium and low lighting conditions.

Figure 4.12 (a) Unstructured urban area (university campus) (b) sandy area (c) rocky area

For the evaluation process, a geodetic total station was utilized in order to measure the reference coordinates of the visual markers and several characteristic points. The origin of the local coordinate system was defined using the center of the origin marker with the coordinates X=0, Y=0 and Z=0. It's worth mentioning that the videos were recorded at 30fps using 848 × 480 resolution.

For the evaluation of the experiments the absolute error ($|X_{meas} - X_{est}|$) between the measured coordinates of X, Y, Z and the corresponding estimations is used while the horizontal error ($\sqrt{X_{err}^2 + Y_{err}^2}$) is also calculated.

In each experiment, a fiducial marker which represents the origin of local coordinate system and one or three markers which represent the targets are located to the scene and measured with the total station for ground truth information. Afterwards, the RGB-depth camera, is guided through a desired trajectory path in order to identify the markers and maps the surroundings.

The experiments were designed aiming to simulate a real-case scenario of surveying a plot or a field in which traditional land surveying techniques and equipment are

utilized. More specifically, the main field-work of a surveyor is to measure the coordinates of a few points that form the borders of the mapping area while in most of the cases, the path that the surveyor follows can be approached with right-angle and squared-based paths.

Thus, in the present experimentation, the methodology was tested utilizing the commonly-used paths that referred above while the fiducial markers which represent the characteristic points of the path were placed on locations aiming to form the shape of each path similarly to a real-case scenario. For instance in a surveyed area with square shape, the fiducial markers are placed in the four corners of the square.

4.4.3.1 Experiments

Square path - Unstructured urban area (University campus)

The first experiment was conducted in the university campus (fig 4.12a) in a sunny day with high illumination. Four markers were used, one for the origin and three for the targets in a distance of about 5 meters while the camera followed a square path as presented in figure 4.13. The results of this experiment are presented in table 4.13.

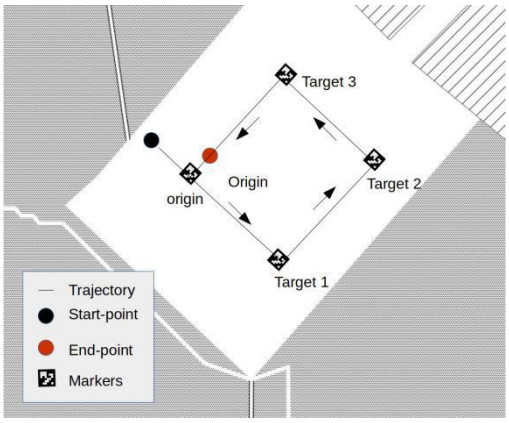


Figure 4.13 Square trajectory path and camera direction

Target marker	X (cm)	Y (cm)	Z (cm)
Target 1: Ground Truth	0	486	0
Target 1: Estimations	7	49	1
Target 1: Error	7	4	**1**
Target 1: XY error	**8.06**		-
Target 2: Ground truth	-492	486	0
Target 2: Estimation	-491	493	-0.3
Target 2: Error	1	7	**0.3**
Target 2: XY Error	**7.07**		-
Target 3: Ground truth	-497	0	0
Target 3: Estimation	-486	-1.4	-8
Target 3: Error	11	1.4	**8**
Target 3: XY Error	**11.08**		-

Table 4.13 Estimations of square-path experiment in unstructured urban area

As observed in table 4.13, the horizontal error of all the targets are in a level of 10 cm or below while the vertical error is in a range of 1 cm (target 1) to 8 cm (target 3). It's worth mentioning that the errors in Y and Z axes of target 1 which are 4 cm and 1 cm respectively, the errors in X and Z axes of target 2 which are 1 cm and 0.3 cm respectively and the error in Y axis of target 3 which is 1.4 cm, are quite close to the ground truth measured with a geodetic total station.

Right-angle path - Sandy area

The second experiment was conducted in a sandy area (fig 4.12b) with high illumination. Three markers were used, one for the origin and two for the targets in a distance of about 6 meters while the camera followed a right-angle path as presented in figure 4.14. The results of this experiment are presented in table 4.14.

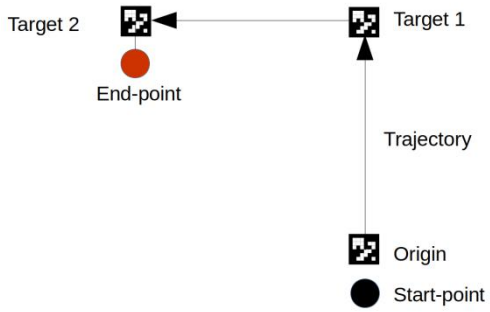


Figure 4.14 Right-angle path in sandy area

Target marker	X (cm)		Y (cm)	Z (cm)
Target 1: Ground Truth	0		600	0
Target 1: Estimations	0.07		605	-0.85
Target 1: Error	0.07		5	**0.85**
Target 1: XY error		**5**		-
Target 2: Ground truth	-600		600	0
Target 2: Estimation	-602		596.8	15
Target 2: Error	2		3.2	**15**
Target 2: XY Error		**3.8**		-

Table 4.14 Estimations of right-angle path experiment in the sandy area

As observed in table 4.14, the horizontal error of all the targets provides high accuracy since the error is in a level of 5 cm. Regarding the vertical errors, the error in target 1 is 0.85 cm which is almost equal to the reference while the error of target 2 is quite larger (15 cm). This error was possibly generated due to temporary SLAM inadequacy in performing local mapping close to target 2, combined with the lack of loop-closure which could further optimize the results.

Square path - Rocky area

The third experiment was conducted in a rocky area (fig 4.12c) with medium illumination. Four markers were used, one for the origin and three for the targets in a distance of about 6 meters while the camera followed a square path as presented in figure 4.15a. The results of this experiment are presented in table 4.15.

Target marker	X (cm)		Y (cm)	Z (cm)
Target 1: Ground Truth	7		598	2
Target 1: Estimations	9		589	0.60
Target 1: Error	2		9	**1.4**
Target 1: XY error		**9.22**		-
Target 2: Ground truth	582		529	2
Target 2: Estimation	568		526	1.3
Target 2: Error	14		3	**0.7**
Target 2: XY Error		**14.32**		-
Target 3: Ground truth	479		-47	1
Target 3: Estimation	473		-35	-8
Target 3: Error	6		12	**9**
Target 3: XY Error		**13.41**		-

Table 4.15 Estimations of square path experiment in the rocky area

As presented in table 4.15, the horizontal error in target 1 is about 9 cm while in targets 2 and 3 the error is increased in a level of 15cm. The vertical error is quite low in the targets 1 and 2 (1.4 and 0.7 cm respectively) while in the third target, the error is equal to 9 cm. The overall accuracy in this experiment, is slightly lower than the sandy and unstructured urban areas, however the results are considered satisfactory taking into account the medium illumination during the experiment and the poor-feature information of the rocky area.

Right-angle path - Rocky area

This experiment was conducted in the rocky area (fig 4.12c) with medium illumination due to cloudy weather. Three markers were used, one for the origin and two for the targets forming a right-angle path. The first two targets were positioned on a straight line from the origin in a distance of about 4 and 9 m respectively while the third target, positioned on the perpendicular line, in a distance of about 11 meters from the origin (fig 4.15b). The results of this experiment are presented in table 4.16.

Target marker	X (cm)		Y (cm)	Z (cm)
Target 1: Ground Truth	34		414	3
Target 1: Estimations	21		412	-0.5
Target 1: Error	13		2	**3.5**
Target 1: XY error		**13.15**		-
Target 2: Ground truth	71		922	1.6
Target 2: Estimation	41		916	1.3
Target 2: Error	30		6	**0.3**
Target 2: XY Error		**30.6**		-
Target 3: Ground truth	576		879	2.1
Target 3: Estimation	543		883	-19
Target 3: Error	33		4	**21.1**
Target 3: XY Error		**33.2**		-

Table 4.16 Estimations of right-angle path experiment in the rocky area with high illumination

As observed in table 4.16, the horizontal error in the target 1 is about 13 cm while the targets 2 and 3 is in a level of 30 cm. The vertical error of the targets 1 and 2 is about

3.5 cm or below and the error of the target 3 is in a level of 20 cm. The decreased accuracy of this experiment compared with the aforementioned experiments is reasonable due to two reasons: At first, the trajectory path of the camera and the distance of the targets from the origin are quite larger (11 meters instead of 6) while at second the SLAM algorithm doesn't perform loop closure in the right-angle path, a significant optimization step for the accuracy of the SLAM results which affect the coordinate estimations.

For further experimentation, the RGB frames of the aforementioned experiment were processed in order to artificially reduce the illumination aiming to test the methodology in quite low illumination (nighttime). The results of this experiment are presented in the table below.

Target marker	**X (cm)**		**Y (cm)**	**Z (cm)**
Target 1: Ground Truth	34		414	3.2
Target 1: Estimations	23		412	0.7
Target 1: Error	11		2	**2.5**
Target 1: XY error		**11.18**		
Target 2: Ground truth	71		922	2
Target 2: Estimation	46		915	9
Target 2: Error	25		7	**7**
Target 2: XY Error		**26**		
Target 3: Ground truth	577		879	2
Target 3: Estimation	546		886	-9
Target 3: Error	31		7	**11**
Target 3: XY Error		**31.8**		

Table 4.17 Estimations of right-angle path experiment in the rocky area with very low illumination (night time)

As presented in table 4.17 the horizontal error in the first target is about 11 cm while the targets 2 and 3 is in a level of 30 cm while the vertical error of the target 1 is 2.5 cm and the errors of targets 2 and 3, are 7 cm and 11 cm respectively. It's worth noting that beyond the vertical error of target 2 which is increased, all the other horizontal and vertical errors are decreased instead of increasing due to the low illumination. This is possibly due to the training process of the HFnet2 with the proposed dataset (see 3.1.4) which includes thousands of images captured in nighttime but also the high contrast where some regions had, because of the white color of several features (fig. 4.16).

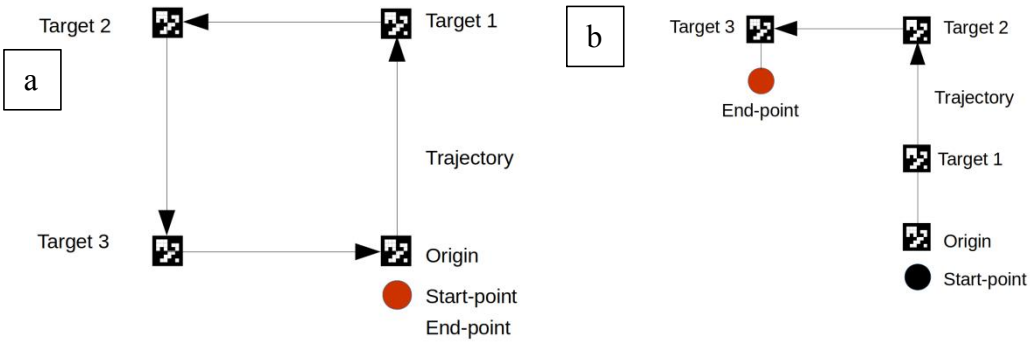


Figure 4.15 Trajectory paths in rocky area: (left) Square path experiment (right) right-angle path experiment

Figure 4.16 Rocky area (left) physical illumination, (right) artificially low illumination

To sum up, in this chapter, the implementation of each methodology including the required programming languages, libraries and platforms was described, while the experimentation details and the extracted results were presented. In the next chapter, the results for each methodology and architecture are further analyzed, while comparisons with similar well-known and state-of-the-art algorithms are presented, aiming to further evaluate the proposed framework.

Chapter 5

Discussion

In this chapter, the interpretation of the results, described and analyzed in chapter 4 are presented, following the structure below:

- Discussion about the results of the optimized SuperPoint architecture
- Discussion about the results of HF-net2 architecture and proposed SLAM
- Discussion about the results of the proposed NN for semantic segmentation
- Discussion about the results of the precise positioning methodology in GNSS-denied environments

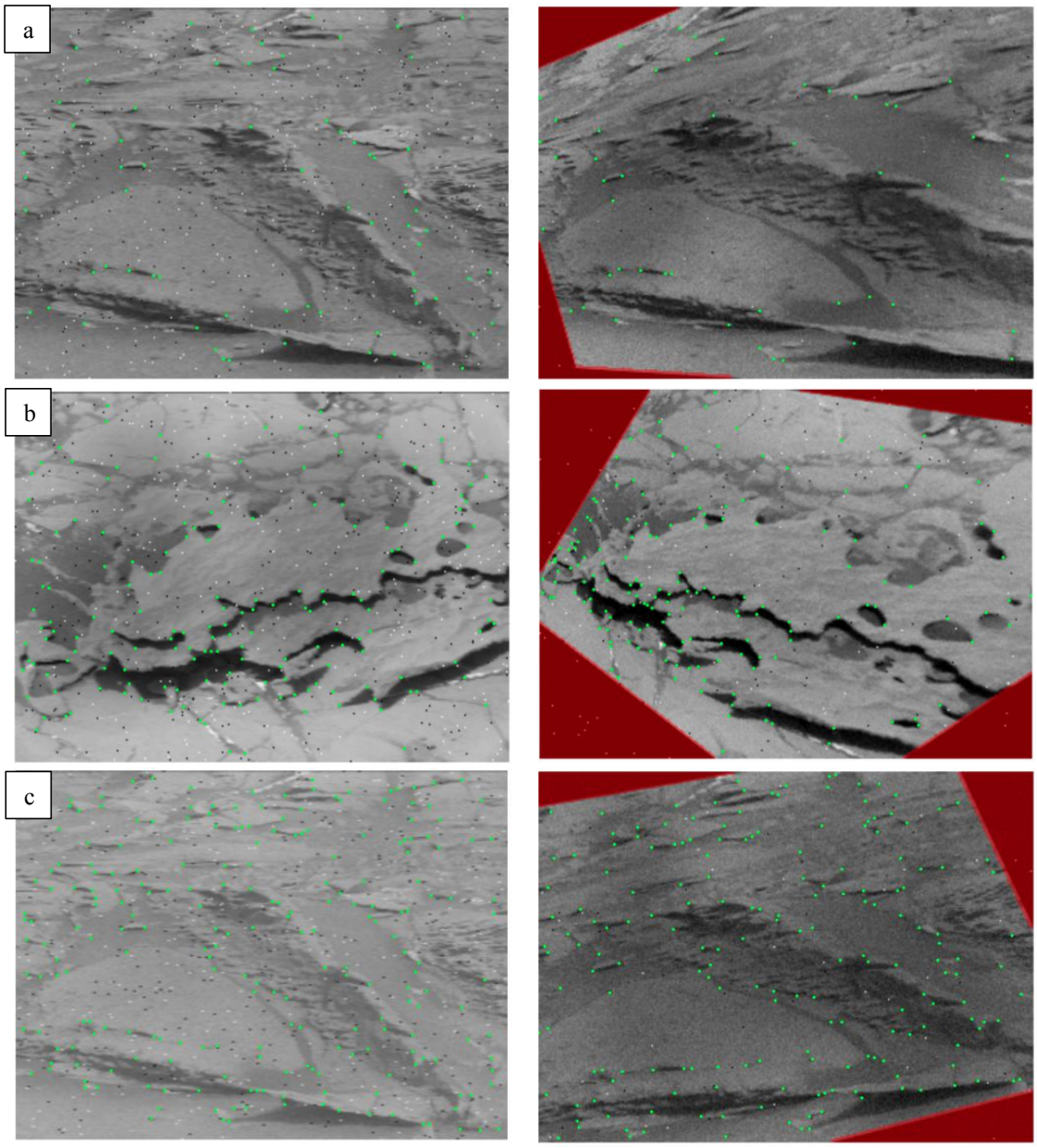
5.1 Discussion about the results of the optimized SuperPoint architecture

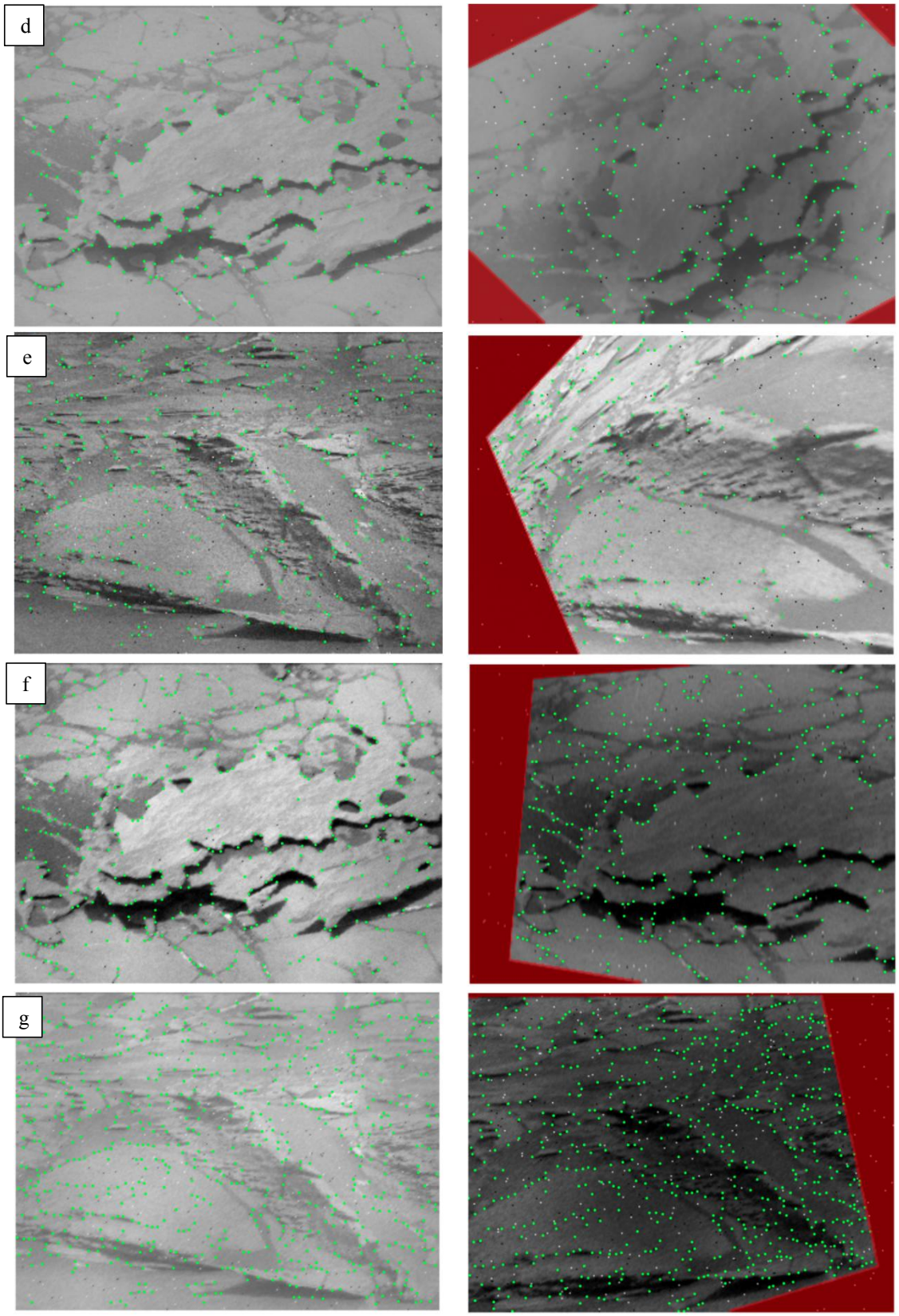
In this study, an investigation of SuperPoint architecture's efficiency in keypoint detection and description applied to unstructured and planetary scenes was conducted. Two modifications in the original SuperPoint architecture including the use of bi-linear instead of bicubic interpolation in the descriptor decoder and the normalization of the descriptors in the descriptor's loss function, were implemented, aiming to increase the accuracy of the model in unstructured environments. The original and an optimized architecture of SuperPoint, were trained with the proposed dataset, producing three different models: (a) an original SuperPoint model trained from scratch, (b) an original fine-tuned SuperPoint model, (c) an optimized SuperPoint model, trained from scratch with the same parametarization as the corresponding original model. The models were evaluated using the designed benchmark dataset while the repeatability and homography estimation metrics were utilized in order to evaluate the produced models and compared with the pre-trained SuperPoint model, trained with COCO dataset and several popular keypoint detectors and descriptors.

Regarding the evaluation of keypoint detectors in terms of illumination changes, although the original and optimized models perform respectable results outperforming the handcrafted algorithms, the pre-trained SuperPoint provides the highest score in repeatability. This is reasonable, since the COCO dataset which has been utilized for the pre-trained SuperPoint model, includes thousands of images with increased variance in lighting conditions, instead of the proposed dataset which includes limited variance in illumination changes. On the contrary, the optimized SuperPoint model achieves the highest repeatability in terms of viewpoint changes, proving that enriching the dataset with high variance in illumination changes, the overall accuracy of the optimized SuperPoint will be enhanced outperforming the pre-trained SuperPoint in both illumination and viewpoint changes. Concerning the evaluation of

descriptors, the optimized SuperPoint model outperforms all the original SuperPoint models and ORB algorithm in illumination and viewpoint changes while SIFT achieves the highest score in overall homography estimation.

In figure 5.1, the progress of the SuperPoint's learning process is presented through the visualization of detected features in two scenes from Mars. Initially, in fig. 5.1a, 5.1b the features are detected using the MagicPoint (the standalone detector of SuperPoint) model trained with the synthetic shapes dataset, while afterwards the results of the MagicPoint models produced by two rounds of MagicPoint training with the proposed dataset (fig. 5.1c, 5.1d, 5.1e, 5.1f), prove the increased sensitivity in feature-poor planetary scenes. Finally, in fig 5.1g, 5.1h, the superiority of the final SuperPoint model is presented through the multiple detected features which describe the content of each scene with quite higher detail than the aforementioned MagicPoint models.




d
e
f
g

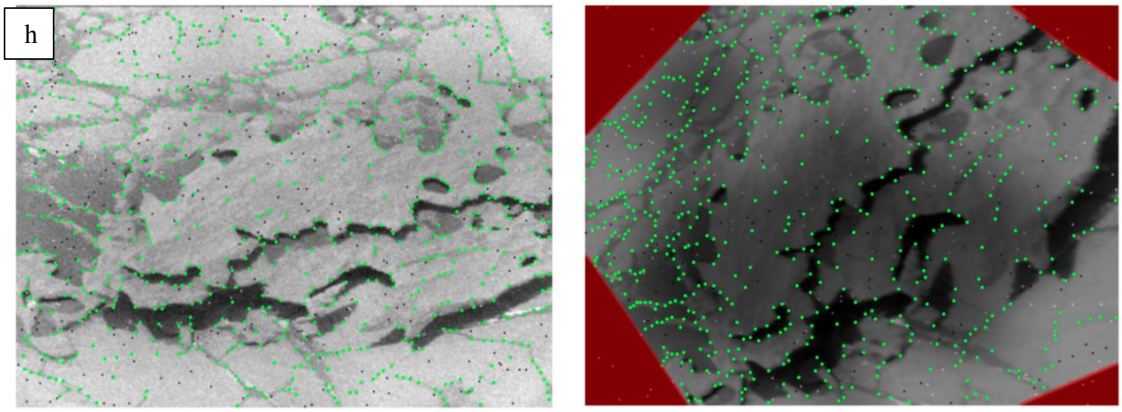


Figure 5.1 (a, b) MagicPoint model trained with synthetic shapes dataset (c, d) first round of MagicPoint training with the proposed dataset, (e, f) second round of MagicPoint training with the proposed dataset, (g, h) SuperPoint model, trained after two rounds of MagicPoint training

It's worth mentioning that most of the studies which utilize feature extractors based on deep learning, use models that have been trained with general-purpose datasets such as COCO, regardless of the environments that are applied. The superiority of SuperPoint models, trained for unstructured environments, compared with the pre-trained SuperPoint, proves that the feature extractors based on deep learning, trained for a specialized and completely different environment, are able to provide increased efficiency compared with a model trained with general-purpose datasets including images from urban, indoor, or vegetated scenes.

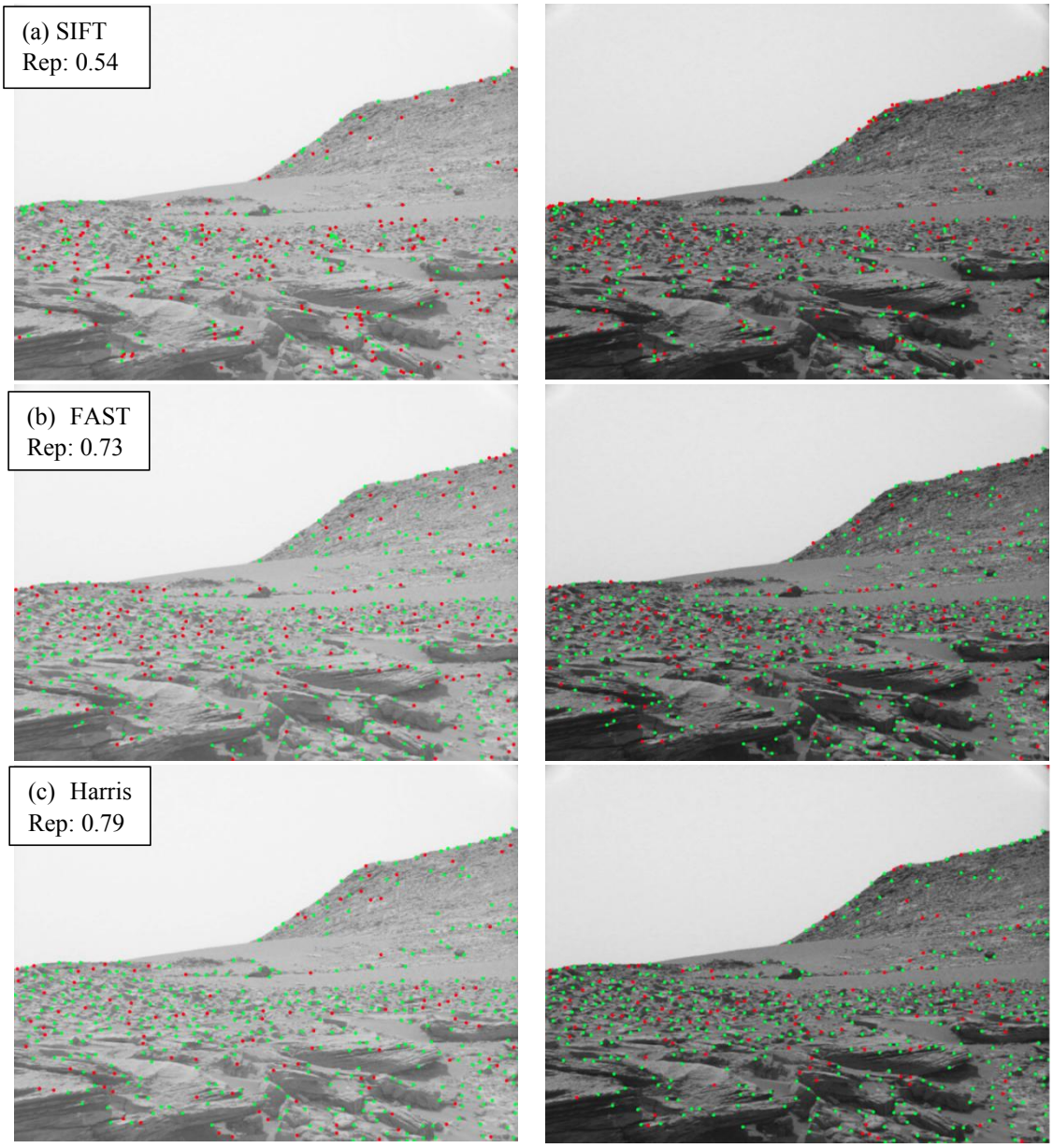
5.2 Discussion about the results of HF-net2 architecture and proposed SLAM

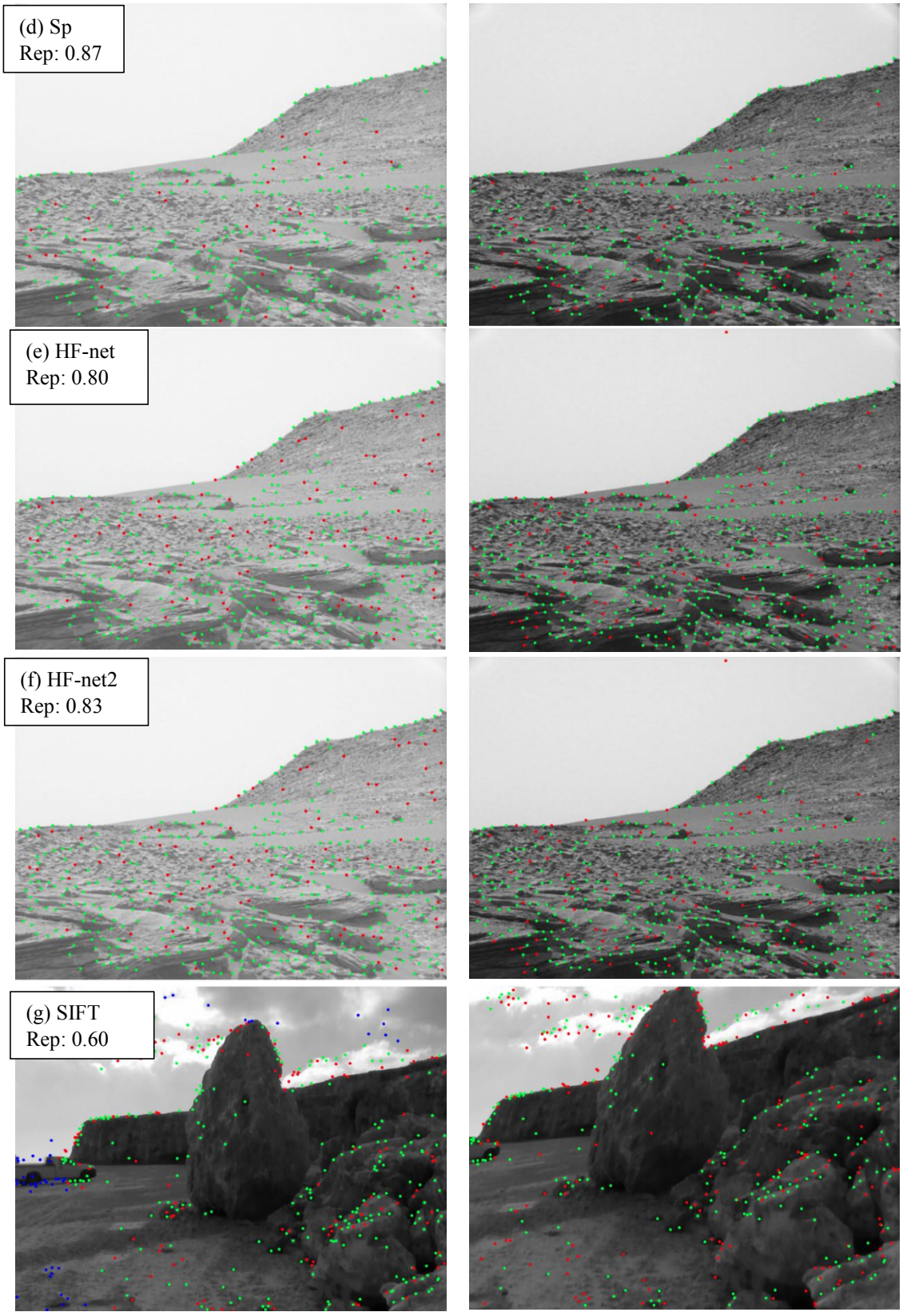
In this study, HF-net2, a multi-task teacher student architecture for keypoint detection and description is proposed, aiming to be used in computer vision tasks including autonomous navigation in challenging unstructured environments. Regarding its architecture, SuperPoint and NetVLAD neural networks are used as teachers for extracting keypoint locations, local and global descriptors aiming of labeling the dataset while in the main HF-net2 architecture, MobilenetV3-large is utilized as a shared encoder and three different sub-modules, a keypoint detector, a local descriptor and a global descriptor represent the multi-task decoder part or the architecture. The HF-net2, was trained using a dataset composed of 48 000 captured and selected images from Earth, Mars and Moon aiming to learn accurate feature extraction in unstructured and planetary environments, while the trained model was integrated in a SLAM system as a feature extraction module, for further evaluation in unstructrured scenes. The main goal of the proposed feature extraction model is to efficiently deal with two significant challenges in unstructured and planetary environments (a) the lack of visual cues (b) and the intense illumination changes.

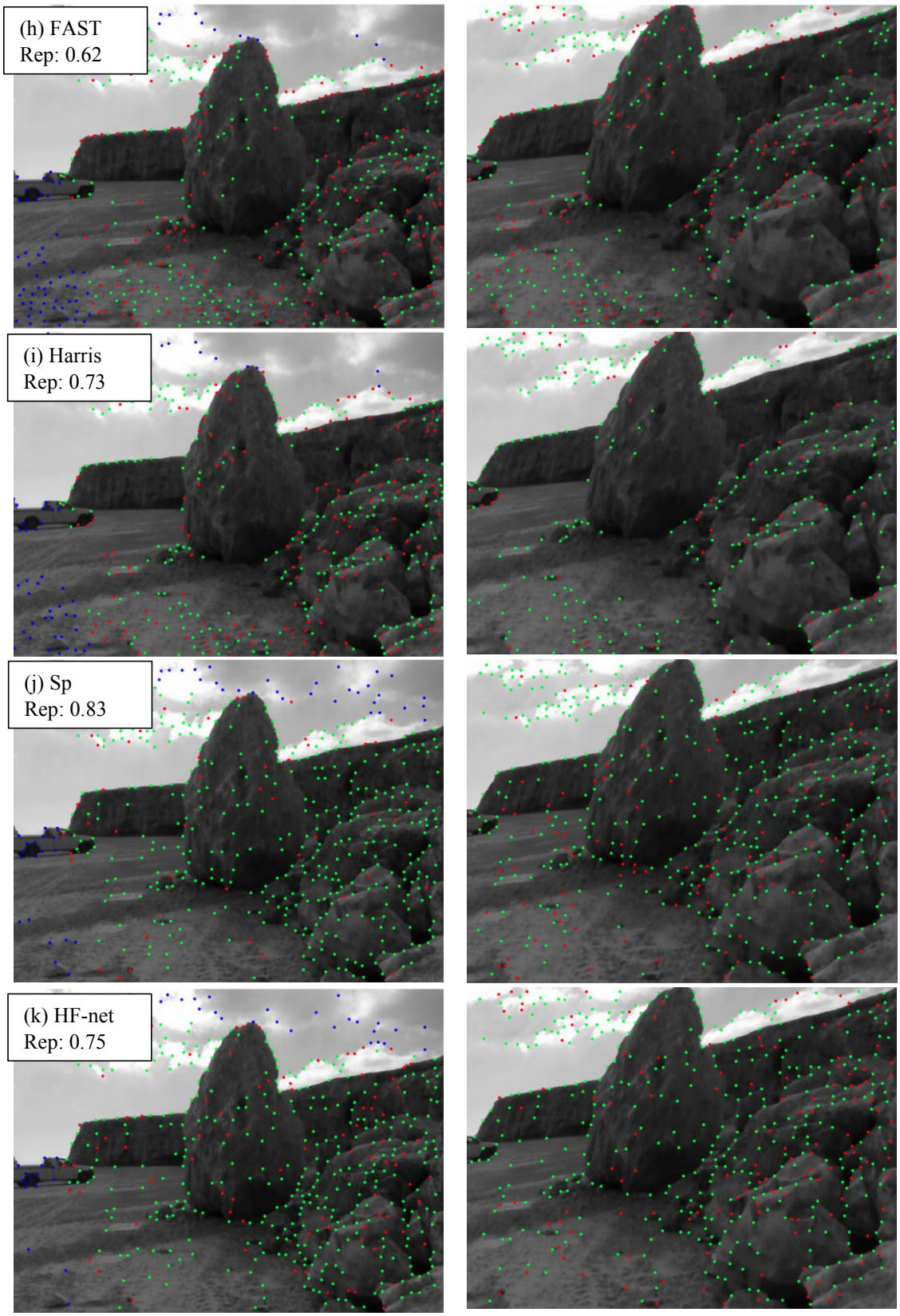
To evaluate the model, a benchmark dataset was created which contains image sequences from Earth, Mars and Moon, aiming to evaluate the detector and descriptor of the model in terms of illumination and viewpoint changes. Regarding the keypoint

detection, HF-net2 achieves the highest repeatability and mAP (0.74 and 0.71 respectively), in terms of illumination changes after the Superpoint, outperforming the original HF-net trained under the same dataset and parameters and several well-known keypoint detectors including SIFT, FAST and Harris. In terms of viewpoint changes, HF-net2 provides respectable accuracy which is slightly higher than original HF-net while provide increased overall accuracy compared with the SIFT and FAST algorithms. Regarding the keypoint description, the proposed model outperforms the original HF-net, SIFT and ORB descriptors in terms of both illumination and viewpoint changes, achieving the highest matching score and mAP after the SuperPoint.

Qualitative results in keypoint detection and description, can also prove the superiority of the proposed architecture and its robustness in scenes with lack of visual cues (fig. 5.2, fig. 5.3).




(d) Sp
Rep: 0.87
(e) HF-net
Rep: 0.80
(f) HF-net2
Rep: 0.83
(g) SIFT
Rep: 0.60


(h) FAST
Rep: 0.62
(i) Harris
Rep: 0.73
(j) Sp
Rep: 0.83
(k) HF-net
Rep: 0.75

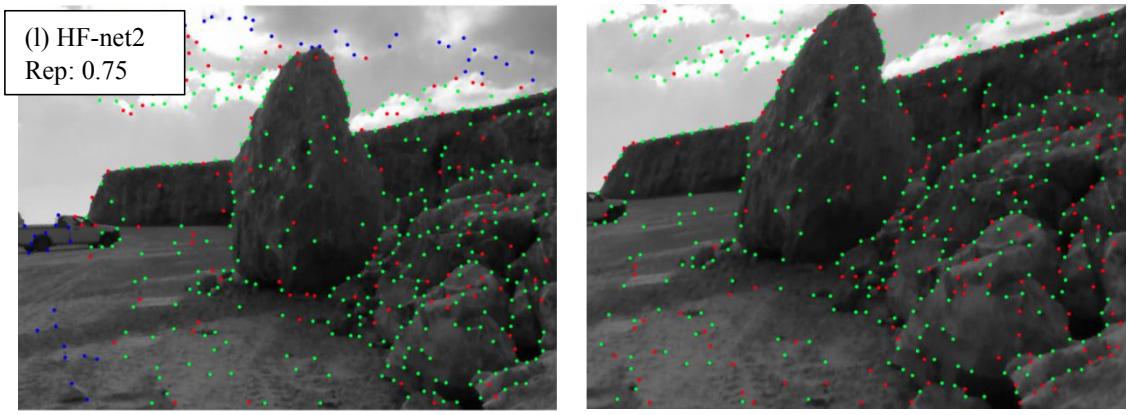
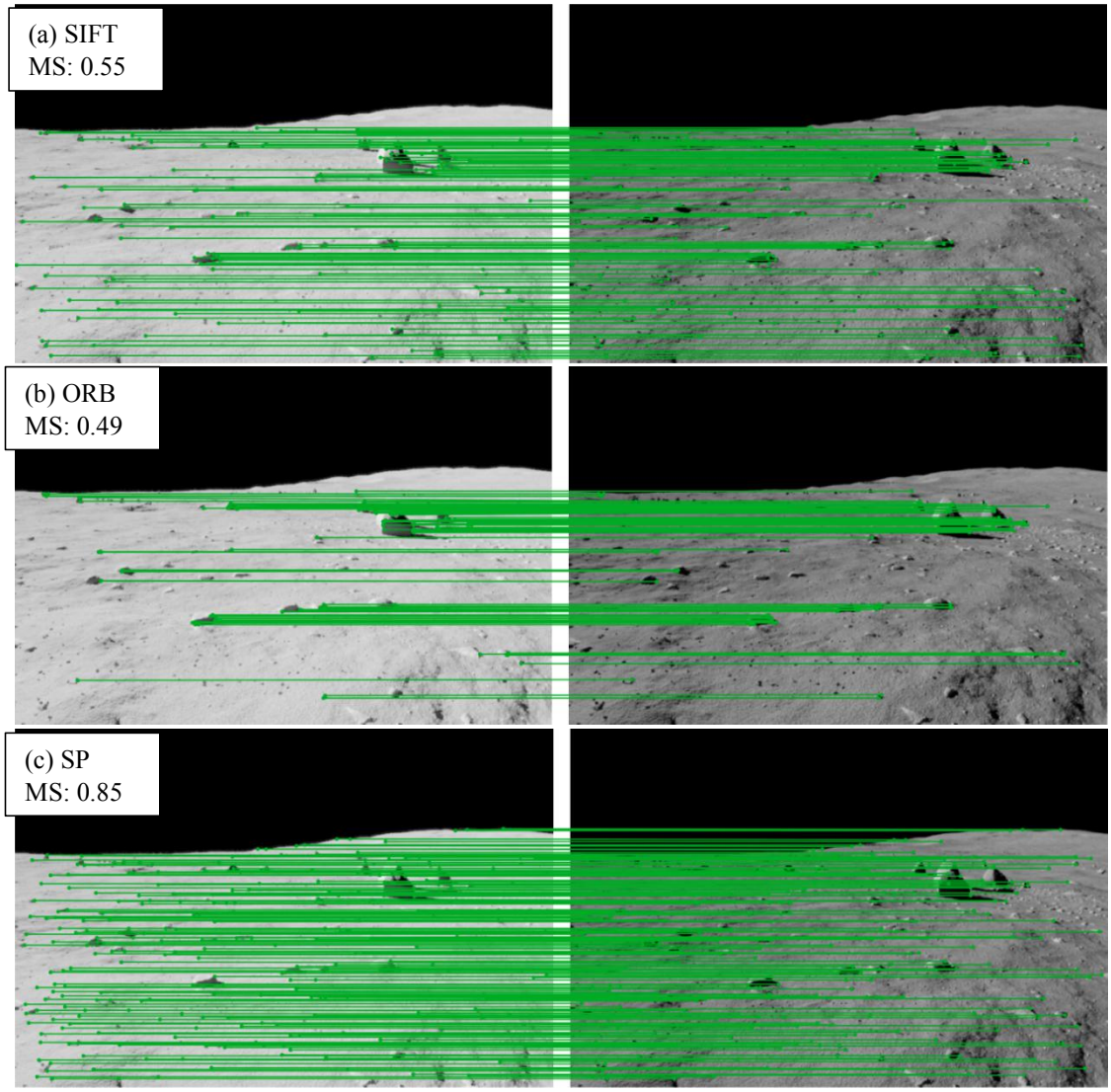
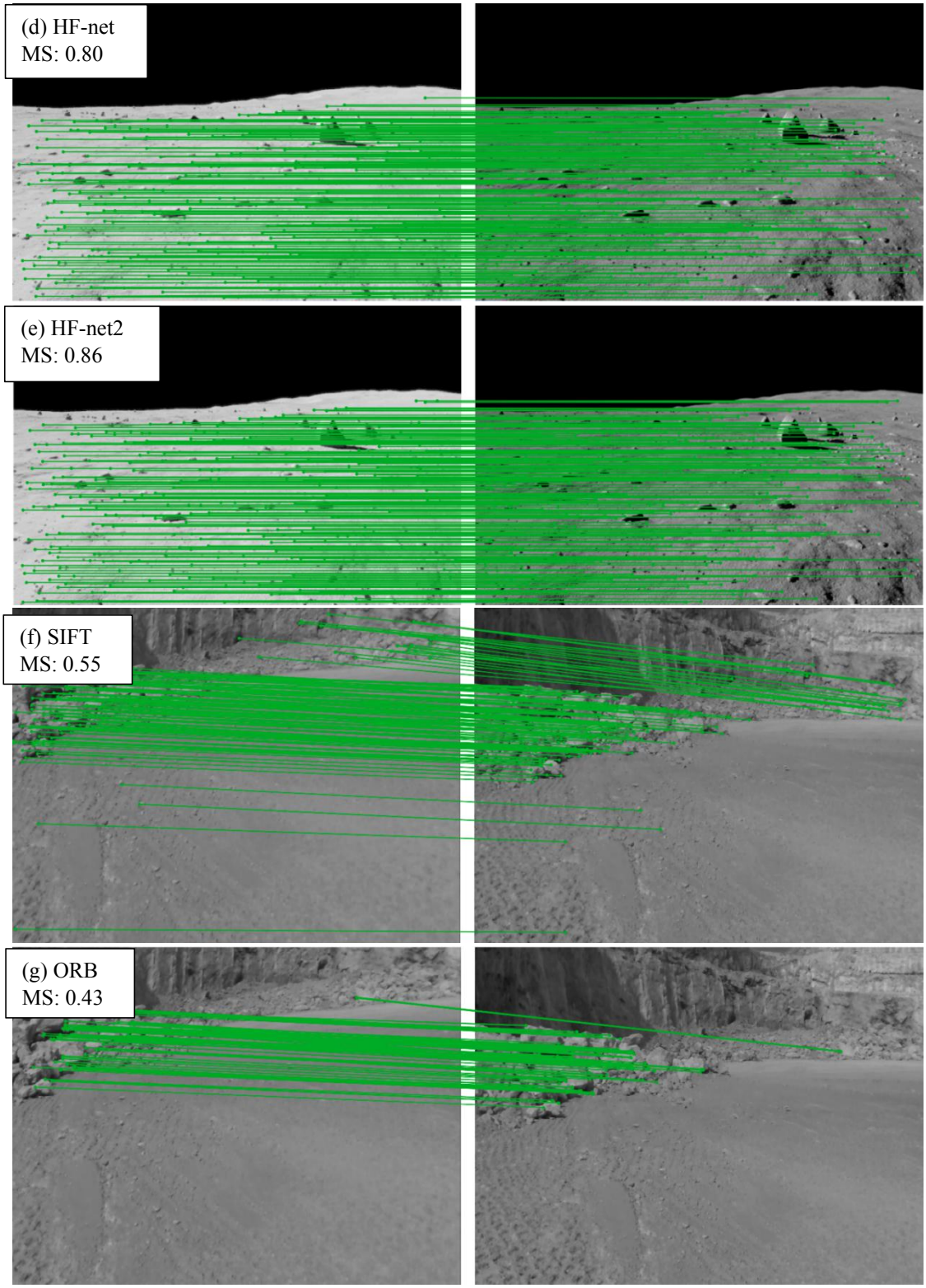


Figure 5.2: Keypoint locations and repeatability scores for SIFT, FAST, Harris, SuperPoint, original HF-net and HF-net2. Two images from the evaluation dataset are presented: (a - f) scene from Mars testing illumination changes, (g-l) earthy scene testing viewpoint changes. The green dots are points that were detected in both images while the red dots are detected points in one image only. The blue points are not depicted in both images due to different viewpoint




(d) HF-net
MS: 0.80
(e) HF-net2
MS: 0.86
(f) SIFT
MS: 0.55
(g) ORB
MS: 0.43

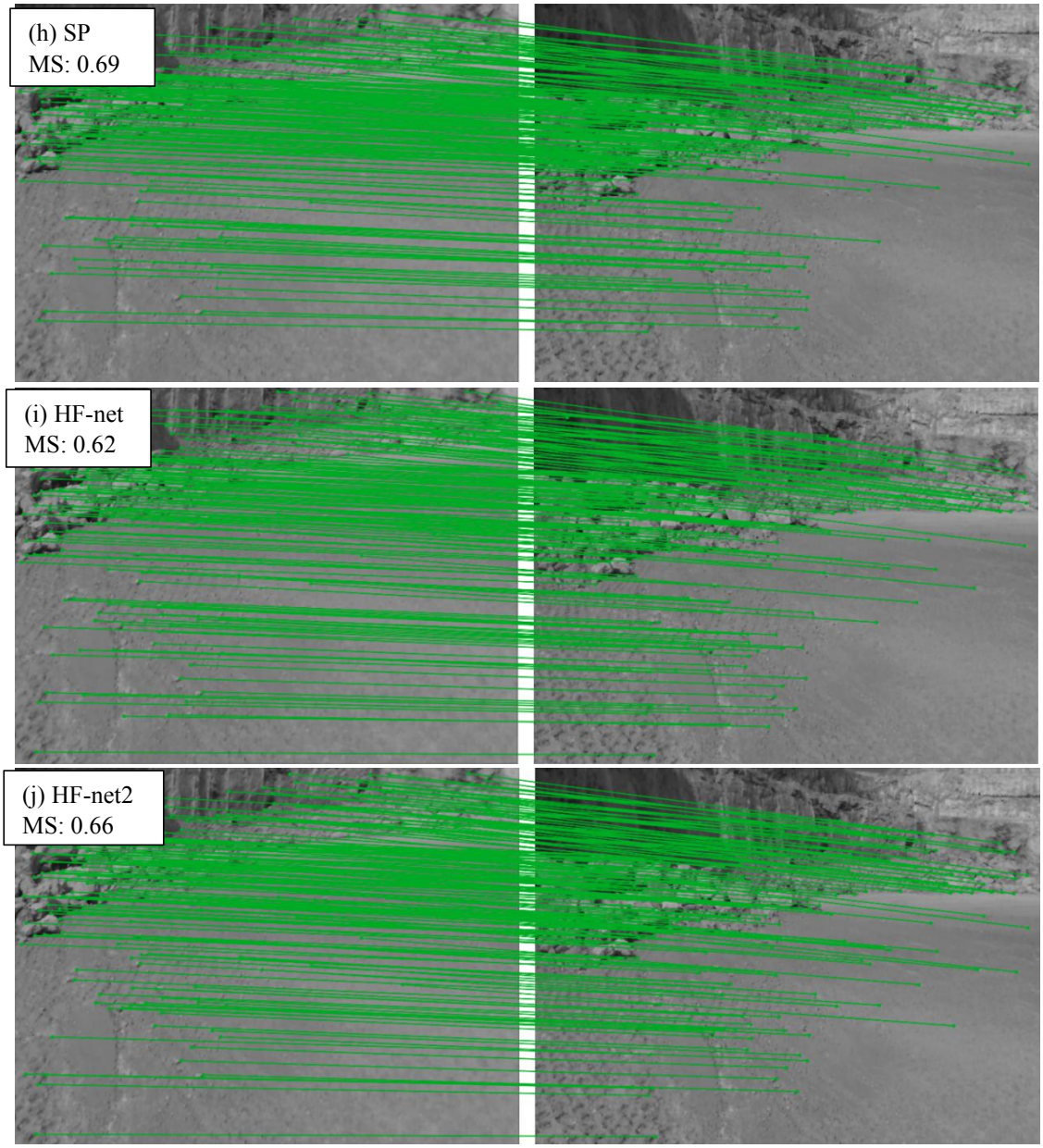


Figure 5.3: Matching scores of the SIFT, ORB, SuperPoint, original HF-net and proposed HF-net2 descriptors. Two images from the evaluation dataset are presented: (a - e) lunar scene testing illumination changes, (f - j) earthy scene testing viewpoint changes

As presented in figure 5.2, the green dots extracted from SuperPoint, HF-net2 and HF-net, which are repeatable points in both (right and left) images of each row, are localized on more meaningful features which better describe the scene, than the detected keypoints of SIFT, FAST and Harris in terms of illumination. Moreover, the HF-net2 achieves the highest repeatability after the SuperPoint with a value of 0.83 while in terms of viewpoint, HF-net2 and HF-net achieves a repeatability score with a value of 0.75 outperforming SIFT, FAST and Harris. As presented in figure 5.3, which visualizes the matching points in two different scenes where suffer from lack of visual cues, the proposed architecture provides outstanding results (fig 5.3e, 5.3j) compared with SIFT and ORB and increased matching score compared with the original HF-net. It's worth noting that in fig 5.3a-e which represents an artificial lunar

scene, the proposed architecture outperforms even SuperPoint which was its teacher during training.

The HF-net2 model was further evaluated as an integrated feature extraction module in a SLAM system based on ORB-SLAM2, while an extended experimentation was performed in a rocky and a sandy scene in different day-time, using an RGB-Depth camera, while a GNSS receiver was utilized for ground truth. Regarding the experimentation in the rocky scene, beyond the first experiment of a square-based path with high illumination where the proposed SLAM provides similar results compared with ORB-SLAM2, the proposed SLAM outperforms ORB-SLAM2 in the experiments with a square-based path and a right-angle path with low and medium illumination respectively. Moreover, in the second square-based path experiment (fig. 4.5b) the error of ORB-SLAM2 in the end of the path is in a level of 30 cm instead of the proposed SLAM with an error in a level of 12 cm.

Concerning the sandy scene, the first experiment, was performed using a random-based path in a sunny day with high illumination, while in the second experiment the illumination of the first experiment was artificially decreased using different levels of quite low illumination in each frame aiming to evaluate the proposed SLAM system in extremely challenging conditions. The proposed SLAM proved its robustness achieving an RMSE error in a level of 0.25 and standard deviation in a level of 0.05 in both experiments instead of ORB-SLAM2 where in the first experiment noted an RMSE in a level of 0.35 and in the second experiment in a level of 0.50 with standard deviation 0.12 and 0.17 respectively.

The experimentation of the SLAM systems, proves that the proposed SLAM provides higher accuracy than ORB-SLAM2 in unstructured environments with medium and low illumination while in extremely challenging scenes either due to poor-featured information or to extremely low illumination, the proposed SLAM extracts significant higher and robust results compared with ORB-SLAM2.

5.3 Discussion about the results of the proposed NN for semantic segmentation

In this study, a deep learning architecture for semantic segmentation is proposed based on U-net neural network, which is able to understand semantically the scene, focused on detecting and classifying rocks and boulders on the lunar surface. The main goal of this study is a lightweight deep learning model with potential of real-time use, in order to increase the safety of rover navigation during a mission on the moon.

Thus, an encoder-decoder architecture was developed which is composed by a modified MobileNetV2 neural network as encoder and a lightweight U-net decoder. Regarding the MobileNetV2 architecture, it includes a fully convolution layer followed by 13 residual bottleneck layers while the depth-multiplier factor was

defined in a value of 0.35 instead of the original MobileNetV2 which includes 19 residual bottleneck layers and the default depth-multiplier factor is equal to 1.0. Concerning the segmentation stage, all the filters of U-net decoder were divided by the factor of 2 while the skip connections transfer information related with the spatial content of each image from the initial input, the block 1, the block 3 and the block 6 of the encoder part.

As presented in section 4.3, the proposed architecture provides robust results achieving IoU in a level of 0.80 or above, detecting and classifying rocks and boulders with satisfactory accuracy in both synthetic and real rover-based images from the lunar surface. To further validate the proposed architecture, it was compared with three similar and widely used encoder-decoder architectures based on U-net:

- The original U-net
- The U-net with VGG16 (Simonyan & Zisserman 2015) as encoder
- The U-net with the original MobileNetV2 as encoder

The architectures above, were trained and tested under the same parametrization so as a fair and proper evaluation to be conducted.

The trainable parameters of the proposed architecture are about 220,000 while the corresponding trainable parameters of U-net, VGG16/U-net and MobileNetV2/U-net are about 31,000,000, 24,000,000 and 8,000,000 respectively while the weights file sizes are about 370 MB for U-net, 285 MB for VGG16/U-net and 97 MB for MobileNetV2/U-net while the corresponding weights file size of the proposed architecture is about 3.5 MB (table 5.1).

Architecture Encoder / Decoder	Total params	Trainable params	Non-trainable params	Model file size (MB)
U-net	31,061,416	31,047,712	13,704	373.1
VGG16 / U-net	23,752,708	23,748,676	4,032	285.4
MobV2 / U-net	8,047,876	8,011,780	36,096	97.3
Proposed architecture	228,588	221,724	6,864	3.5

Table 5.1 Parameters and model size of the U-net, VGG16/U-net, MobV2/U-net and the proposed architecture

In figure 5.4, qualitative results from the alternative and the proposed architectures are depicted while in table 5.2 the corresponding IoU score is presented. It's worth noting that original MobileNetV2/U-net could not converge with this specific parametrization, thus in the results below the proposed architecture is compared with original U-net and VGG16/U-net.

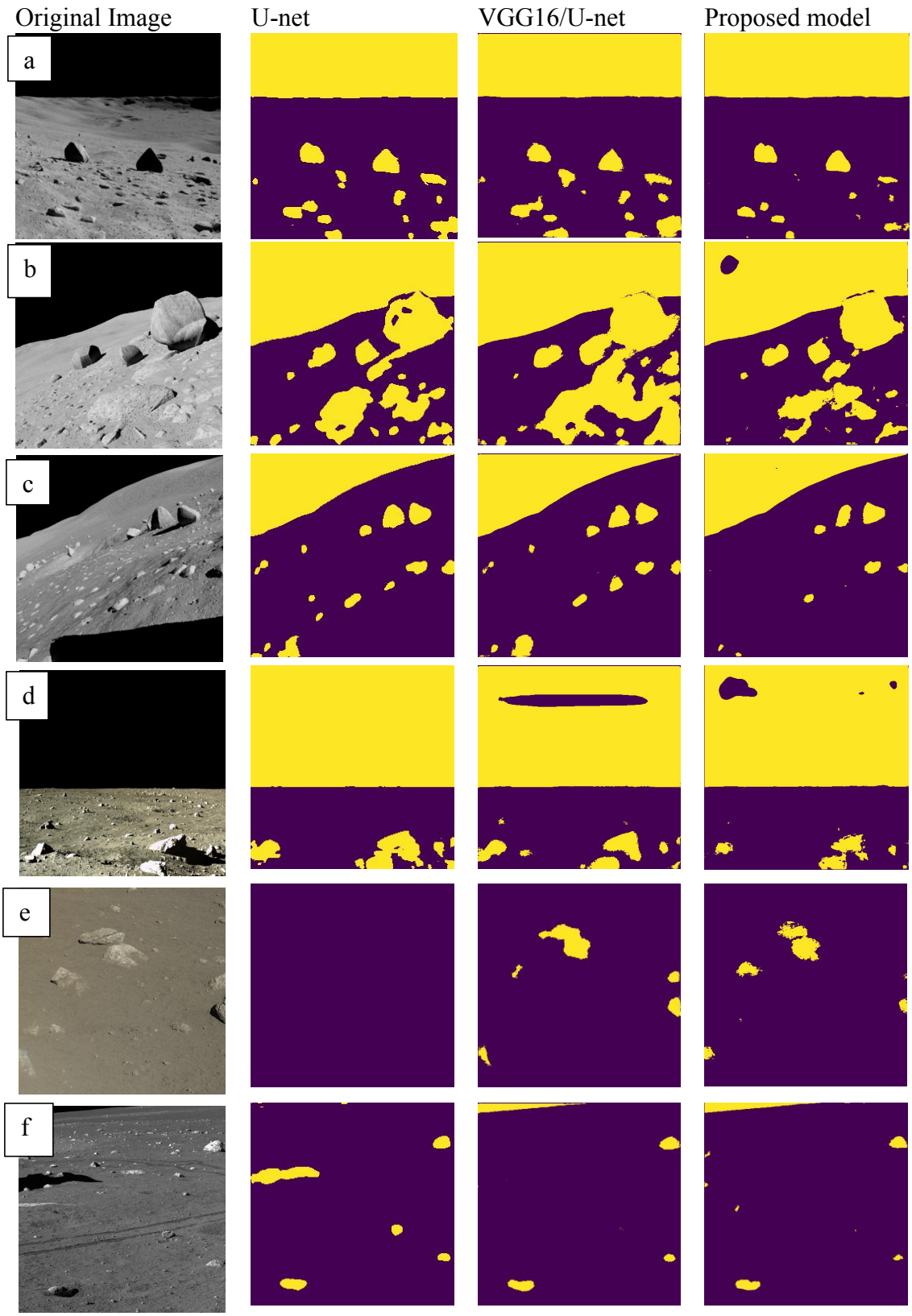


Figure 5.4 First column: original synthetic (a, b, c) and real (d, e, f) lunar images, second column: original U-net model predictions, third column: VGG16/U-net model predictions, fourth column: proposed architecture

Architecture Encoder / Decoder	IoU
U-net	0.86
VGG16 / Unet	0.82
Proposed model	0.84

Table 5.2 IoU score in testing data of the original U-net, VGG16/U-net and the proposed model, trained with the same dataset and parametrization

As observed in figure 5.4, all the models produce respectable segmentation results. In figures 5.4a, 5.4b and 5.4c, 5.4d the proposed model provides similar accuracy in rocks segmentation compared with the original U-net and VGG16/U-net, predicting all the important rocks and boulders that could harm a rover during navigation. On the other hand, in figures 5.4e, 5.4f which are real images from lunar surface, the proposed architecture provides refined segmentation results compared with the alternative models. For instance, in figure 5.4e, the proposed model precisely segments the two main rocks on the ground instead of original U-net which fails to predict them while VGG16/U-net falsely unifies them in a bigger rock. Similarly, in figure 5.4f, the proposed model and VGG16/U-net produce quite close results while original-U-net falsely predicts a large shadow as a rock.

Regarding the evaluation of the models on testing data in terms of intersection over union (IoU), the proposed architecture provides an IoU score of 0.84 (table 5.2) outperforming the VGG16/U-net while is close to IoU of U-net which is equal to 0.86. The results above, determines the superiority of the proposed architecture since, it is about 110 times and about 140 times smaller than the VGG16/U-net and the original U-net respectively while provides similar segmentation predictions in both alternative architectures.

It’s worth noting that, although all the models provide robust results in sky segmentation defining the horizon line, they are unable to classify the sky as separate class. This is due to dataset's lack of color variety on the ground features while in addition, many images include large black areas which represent shadows on the ground, but because the images are synthetic, there is no a meaningful difference between the sky and the large black shadows. Nevertheless, a refined synthetic rover-based dataset or a dataset with real lunar landscape images, could solve this issue, improving the classification results of all the models.

Regarding the inference-time, the models were tested on a large set of images with a size of 480x480 in three different computing setups: (a) a GPU-enabled conventional desktop machine, (b) CPU-only conventional desktop machine and (c) a CPU-only embedded system with quite low resources. The corresponding results are presented in the table 5.3 and figure 5.5.

Inference time per image	Conventional machine /GPU-enabled		Conventional machine / CPU-only		Embedded system / Rasp. Pi 4	
	ms	FPS	ms	FPS	ms	FPS
U-net	100	10	850	1.17	19680	0.05
VGG16/U-net	52	19.23	640	1.56	11120	0.09
Proposed model	43	23.25	100	10	1080	0.92

Table 5.3 Comparison in terms of inference time (in milliseconds and FPS) of the original U-net, VGG16/U-net and the proposed model in a desktop GPU-enabled and CPU-only conventional desktop computer and in a CPU-only embedded system with low resources

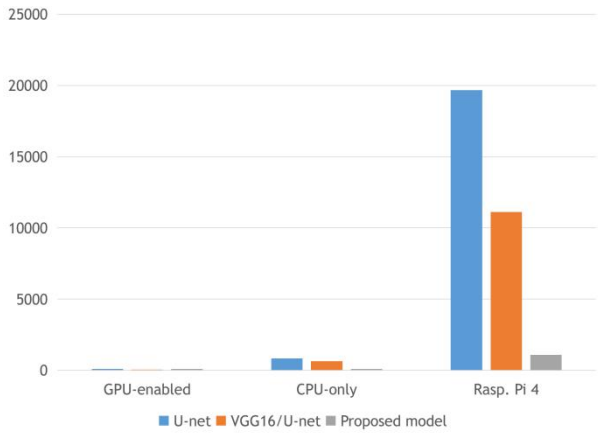


Figure 5.5 Inference time in millisecond (ms) of the U-net, VGG16 / U-net and the proposed model for the GPU-enabled machine, the CPU-only machine, and the Raspberry Pi 4 embedded system

As observed in table 5.3 and figure 5.5, the proposed model achieves quite less inference time compared with the U-net and VGG16 / U-net while the difference in performance-time is increased among the models when the computing resources are reduced. Regarding the GPU-enabled machine, the proposed model achieves 43 ms and 23.25 FPS, while the VGG16/U-net provides 52 ms (19.23 FPS) of inference-time and U-net about 100 ms (10 FPS) which is twice the time compared with the proposed model. In the CPU-only machine, the proposed model provides inference-time in a level of 100 ms (10 FPS) while the VGG16 / U-net and U-net models perform predictions with 640 ms (1.56 FPS) and 850 ms (1.17 FPS) inference-time respectively, six and nine times more than the proposed model. Concerning the Raspberry Pi 4 with 4GB of RAM embedded system, the proposed model achieves an inference-time about 1080 ms (0.92 FPS) which is quite satisfactory since to the best of our knowledge, this embedded system provides the lowest computing resources on the market, especially in deep learning. Instead, the VGG16/U-net and U-net models provide 11120 ms (0.09 FPS) and 19680 ms (0.05 FPS) inference-time, proving that the proposed model is about 11 and 20 times faster in the Raspberry Pi 4 embedded system compared with the VGG16/U-net and U-net models respectively.

5.4 Discussion about the precise positioning and mapping methodology in GNSS-denied environments

In this study, a precise positioning methodology is proposed, which estimates characteristic points and arbitrary point locations in 3D space of an unstructured and challenging GNSS-denied environment, with centimeter-level of accuracy, using at least a fiducial marker and an RGB-Depth camera. At first, the camera is guided through a desired path, identifies the markers and maps the surroundings. In a second

step, a local coordinate system is created for the scene with an origin defined by the initial marker that the camera comes across, while the coordinates of the target markers and the arbitrary points of the point cloud are calculated based on the origin marker. In order to evaluate the methodology, different sets of experiments were performed in terms of study area, the number and location of markers, the trajectory paths and the illumination conditions. Although the methodology's accuracy is highly correlated with some factors that will be further mentioned, it achieved quite satisfactory horizontal and vertical accuracy considering the poor-feature scenes and in some cases the low illumination.

As presented in the section 4.4, the proposed methodology achieved satisfactory results with low horizontal and vertical error, in the first three experiments conducted with medium and high illumination in unstructured urban, sandy and rocky areas with values in a range of about 4 to 14.50 cm for the horizontal error and about 0.30 to 15 cm for the vertical error. However, in the last two experiments of the right-angle path conducted in the rocky area with medium and artificially low illumination respectively, the horizontal and vertical errors increased, providing errors in a range of 13 cm to 33 cm (horizontal error) and 0.3 to 21 cm (vertical error) for the experiment with medium illumination and in a range of 11 to 32 cm (horizontal error) and 2.5 to 11 cm (vertical error) for the experiment with low illumination. This increase of the errors is due to several factors including the large distance of the target 2 and target 3 from the origin marker, the lack of loop closure which could further optimize the results and the feature-poor environment of the rocky area. It's worth noting that the results in the experiment with low illumination is similar and slightly refined than the results of the same experiment with medium illumination. This fact, proves the importance of HF-net2 model in the SLAM for the mapping process on challenging and unstructured environments in terms of illumination changes.

To further validate the importance of deep learning and especially HF-net2 neural network in the proposed methodology, all the experiments described in the section 4.4 were performed also using the ORB-SLAM2. The horizontal and vertical errors of both approaches are presented in the tables (5.4 - 5.8):

	Proposed with HF-net2-SLAM		**Proposed with ORB-SLAM2**	
Target marker	**XY error(cm)**	**Z error (cm)**	**XY error(cm)**	**Z error (cm)**
Target 1	8.06	1	28.6	5.8
Target 2	7.07	0.3	27.3	1
Target 3	11.08	8	28.6	4
Mean	8.74	3.1	28.17	3.6

Table 5.4 Horizontal and vertical errors of the methodology based on HF-net2 and ORB-SLAM2. Experiment in unstructured urban area (University campus) with high illumination - square path

	Proposed with HF-net2-SLAM		**Proposed with ORB-SLAM2**	
Target marker	**XY error(cm)**	**Z error (cm)**	**XY error(cm)**	**Z error (cm)**
Target 1	5	0.85	5.9	32.4
Target 2	3.8	15	26.5	80
Mean	4.4	7.9	16.2	56.2

Table 5.5 Horizontal and vertical errors of the methodology based on HF-net2 and ORB-SLAM2. Experiment in sandy area with high illumination - right-angle path

	Proposed with HF-net2-SLAM		Proposed with ORB-SLAM2	
Target marker	**XY error(cm)**	**Z error (cm)**	**XY error(cm)**	**Z error (cm)**
Target 1	9.22	1.4	18.63	0.7
Target 2	14.32	0.7	18.60	0.85
Target 3	13.41	9	7.12	1.5
Mean	12.32	3.7	14.8	1.01

Table 5.6 Horizontal and vertical errors of the methodology based on HF-net2 and ORB-SLAM2. Experiment in rocky area with medium illumination - square path

	Proposed with HF-net2-SLAM		Proposed with ORB-SLAM2	
Target marker	**XY error(cm)**	**Z error (cm)**	**XY error(cm)**	**Z error (cm)**
Target 1	13.15	3.5	15.7	7.8
Target 2	30.6	0.3	26	35.7
Target 3	33.2	21.1	33	26.2
Mean	25.65	8.3	24.9	23.2

Table 5.7 Horizontal and vertical errors of the methodology based on HF-net2 and ORB-SLAM2. Experiment in rocky area with medium illumination - large right-angle path

	Proposed with HF-net2-SLAM		Proposed with ORB-SLAM2	
Target marker	**XY error(cm)**	**Z error (cm)**	**XY error(cm)**	**Z error (cm)**
Target 1	11.18	2.5	17.8	34.1
Target 2	26	7	46.7	104.3
Target 3	31.8	11	49.6	0.72
Mean	23.0	6.83	38.03	46.37

Table 5.8 Horizontal and vertical errors of the methodology based on HF-net2 and ORB-SLAM2. Experiment in rocky area with artificially low illumination - large right-angle path

In table 5.4 which present the results from the experiment in the unstructured urban area (university campus), the proposed methodology outperforms the ORB-SLAM2 based methodology since it provides a mean horizontal error (MHE) in a level of 9 cm and mean vertical error (MVE) equal to 3.1 cm instead of ORB-SLAM2-based methodology which provides a MHE in a level of 28 cm and a MVE equal to 3.6 cm. It's worth noting that the horizontal error of ORB-SLAM2-based methodology is above three-times larger than the proposed methodology.

In table 5.5 which represent the results from the experiment with right-angle path in the sandy area, a quite feature-poor scene, the proposed methodology provides satisfactory results with MHE 4.4 cm and MVE equal to 7.9 cm instead of the ORB-SLAM2-based methodology which exports a MHE equal to 16.2 cm and the MVE equal to 56.2 cm which is about 7 times larger than the corresponding MVE of the proposed methodology.

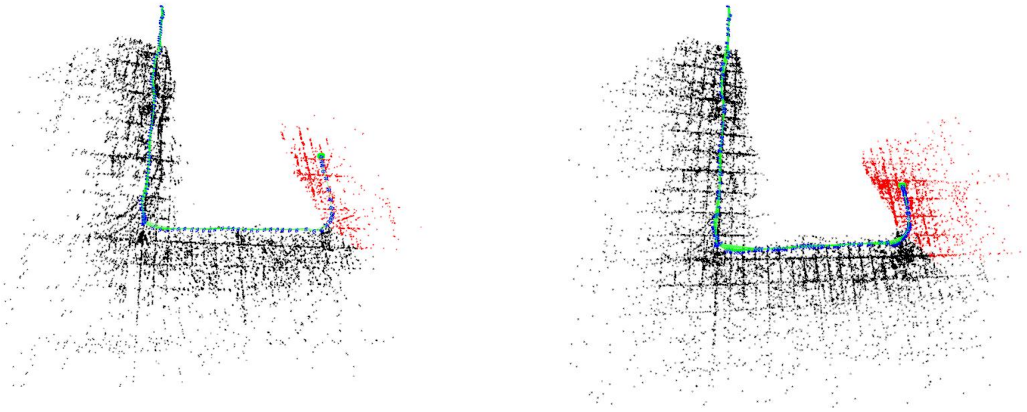
Table 5.6, represents the results from a square-path experiment conducted in the rocky area with medium illumination, an area with more features than the unstructured urban and sandy areas. In this experiment the two methodologies provide similar results since the proposed methodology achieves a MHE in a level of 12 cm instead of the ORB-SLAM2-based methodology with MHE in a level of 15 cm while the MVE

of the proposed methodology is in a level of 4 cm instead of the ORB-SLAM2-based methodology which is in a level of 1 cm.

The last two experiments presented in tables 5.7 and 5.8, were performed in the rocky area following a larger right-angle path. In the first experiment, the physical medium illumination of the scene (table 5.7) was utilized while in the second experiment (table 5.8), artificially decreased in terms of illumination RGB frames were used. Regarding the first experiment (table 5.8), the two methodologies provide similar MHE in a level of 25 cm while the proposed methodology outperforms the ORB-SLAM2-based methodology in terms of vertical error, producing a MVE in a level of 8 cm instead of 23 cm. In the second experiment, where the mapping process was conducted using RGB frames with quite low illumination, the proposed methodology maintains its accuracy compared with the first experiment providing a MHE equal to 23 cm and MVE in a level of 7 cm while the ORB-SLAM2-based methodology produces quite increased errors with MHE in a level of 40 cm and MVE equal to 46.37 cm (table 5.8). It's worth noting that the target 2 vertical error of the ORB-SLAM2-based methodology is in a level of 1 m which is inadequate even for rapid mapping applications.

As the experimentation proves, the proposed methodology is able to accurately map and localize fiducial markers in feature-poor scenes or environments with low illumination, outperforming the ORB-SLAM2-based methodology. This is due to the feature extraction process of HF-net2 which is capable of detecting and describing features with a more sophisticated fashion than the ORB algorithm. For instance, the figure below (fig. 5.6), presents the captured features of ORB-SLAM2 (a) and HFnet2-SLAM (b) during the SLAM process of the unstructured urban area where the only features are the pavement joints. It is clear that the HFnet2-SLAM captures with a more refined way the pavement joints forming clearer squares, than the ORB-SLAM2.

(a) (b)



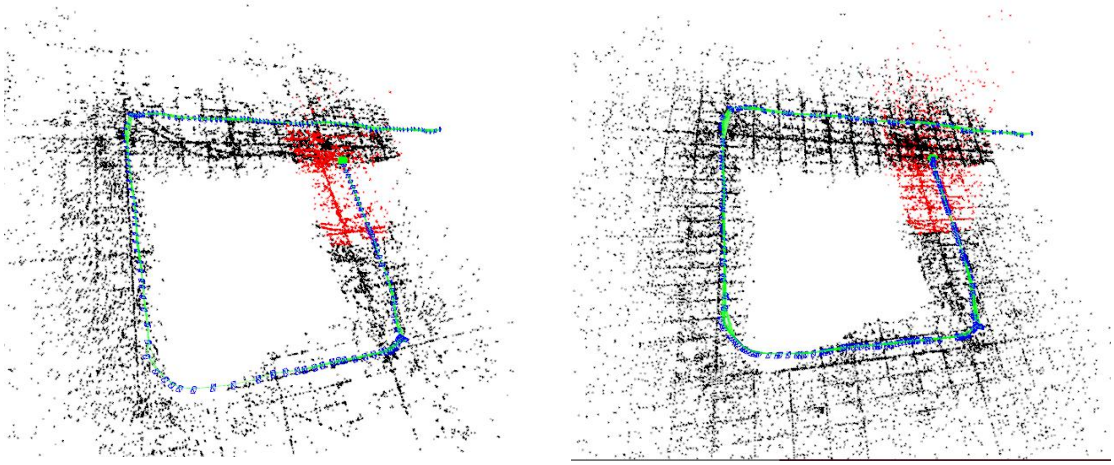


Figure 5.6 Square-path experiment in unstructured urban area (University campus) Column (a): mapping using ORB-SLAM2, (b) mapping using HFnet2-SLAM

Regarding the mapping process of methodology on the field, a certain way of mapping has to be followed in order the methodology to provide qualitative results. The camera has to follow a trajectory that begins a few meters before the origin marker and then proceed, approaching close and overtaking all the markers (origin and targets) maintaining a non-complicated path. This technique, provides a reasonable sequence of frames to the system, aiding the feature extraction process, camera pose estimations and consequently marker localization using the M.L.C and P.A methods. The sudden unreasonable camera movements or a complicated camera trajectory where the camera doesn't approach the markers directly maintaining a steady direction, will significantly decrease the accuracy.

Moreover, the error of each target is dependent only on the location of the origin marker and not affected by the other markers. Nevertheless, the increased distance between the origin and target markers can affect the accuracy. This issue is a limitation of the methodology which can be encountered with the integration of a life-long SLAM architecture (as referred by the computer vision community) which constitutes one of the main future goals of the proposed methodology.

As a conclusion, in this chapter an extended analysis of each part of the proposed framework was conducted, evaluating its efficiency using state-of-the-art algorithms and architectures. In the next chapter the final thoughts and conclusions regarding each pillar of the framework are presented.

Chapter 6

Conclusions

In this chapter, the conclusions of the visual localization framework specialized for unstructured environments are presented, following the structure below:

- Conclusions of the optimized SuperPoint architecture
- Conclusions of the HF-net2 architecture and proposed SLAM
- Conclusions of the modified U-net for unstructured scenes
- Conclusions of the precise positioning and mapping methodology in GNSS-denied environments

6.1 Conclusions of the optimized SuperPoint architecture

In summary, a SuperPoint architecture was utilized aiming to develop a model for keypoint detection and description, with increased sensitivity and accuracy in unstructured environments and planetary scenes. SuperPoint was implemented and trained using the proposed training dataset which includes planetary-like and real-planetary scenes. During experimentation, three different models were produced using the aforementioned dataset: (a) an original SuperPoint model, trained from scratch, (b) an original fine-tuned SuperPoint model, (c) an optimized model, trained from scratch. The models were evaluated using a proposed benchmark dataset, designed for unstructured environments including earthy and planetary scenes, testing their accuracy in illumination and viewpoint changes. The experimentation proves that the optimized SuperPoint model provides satisfactory results in keypoint detection and description, compared with the re-trained SuperPoint models, the original SuperPoint model trained with COCO dataset and several popular handcrafted detectors and descriptors.

However, the lack of samples with high variance in illumination changes, is a significant issue for the methodology which affects the optimized and original SuperPoint models, providing slightly lower repeatability in keypoint detection compared with the pre-trained SuperPoint. A second issue is that SIFT descriptor outperforms the optimized SuperPoint although the proposed model outperforms all the original SuperPoint models and ORB.

Thus, the future work of the optimized model can be focused on two main improvements. At first, the proposed dataset could be enriched including real or artificial images with high lighting changes, while the difficulty of homographic adaptation during the training process could be increased through more examples and extremely transformed representations. The aforementioned improvements, will

further escalate the efficiency of the model in illumination and viewpoint changes, providing refined results in both illumination and viewpoint changes.

As a conclusion, the optimized SuperPoint model, is a promising solution for accurate keypoint detection and description in unstructured and planetary scenes, which could be an inspiration for the computer vision community, increasing the potential for accurate autonomous navigation in completely unknown and unstructured scenes.

6.2 Conclusions of the HF-net2 architecture and proposed SLAM

To sum up, a multi-task distillation-based architecture, called HF-net2 was developed aiming to implement a keypoint detector and descriptor which focuses on unstructured environments and completely unknown planetary scenes. The model was trained with a specialized image dataset from Earth, Mars and Moon and evaluated using a proposed benchmark dataset, compared with several keypoint detectors and descriptors, testing its accuracy in illumination and viewpoint changes. HF-net2 proved its robustness achieving the highest overall accuracy after the SuperPoint which was its teacher during the training process.

Moreover, the HF-net2 model was integrated in a visual SLAM system based on ORB-SLAM2 while an extended experimentation was conducted in two unstructured scenes, using an RGB-depth camera and an RTK-GNSS receiver, utilized for ground truth. The experimentation, which performed in two different areas with several illumination conditions, proved that the proposed SLAM provides satisfactory accuracy in unstructured feature-poor environments with illumination changes, outperforming the ORB-SLAM2.

The future work of this study is two-fold. At first, the proposed architecture will be further improved and fine-tuned by enriching the training and evaluation dataset with more rover-based data, while secondly, the proposed SLAM system will be further optimized in terms of a loop closing module utilizing only the global descriptor instead of a BoW algorithm which can fail in completely unknown environments (Garcia-Fidalgo *et al.* 2018).

As a conclusion, this study proved that the use of deep learning architectures in feature extraction provides a crucial potential in autonomous navigation on unstructured environments which can reinforce the planetary exploration missions, acquiring extremely valuable knowledge for the future of humanity.

6.3 Conclusions of the modified U-net for unstructured scenes

In summary, an encoder-decoder deep learning architecture for lunar ground segmentation was developed, aiming to reinforce the potential of rover safety during a mission. The main goal of this study was the implementation of a semantic

segmentation model with low requirements in computing resources and large training datasets.

To achieve this goal, a deep learning architecture based on U-net neural network was developed, since U-net is able to provide respectable results, trained with limited size of datasets (Ronneberger *et al.* 2015). To reduce the computational cost of U-net, a modified MobileNetV2 neural network was used as the encoder, while a lighter version of U-net decoder was implemented in order to accelerate the segmentation stage. The proposed architecture was trained with a publicly available dataset with artificial lunar scenes, which is the only available training dataset for the Moon environment.

As a result, the proposed model achieves satisfactory accuracy in scene segmentation, not only in testing data of synthetic images but also in real rover-based images of the lunar surface while it includes significantly less trainable parameters than U-net based alternatives. The proposed architecture was evaluated compared with the original U-net, the VGG16/U-net and the original MobileNetV2/Unet neural networks which were trained under the same parametrization. The proposed architecture is about 140 times smaller than the original U-net, 110 times than the VGG16/U-net and 36 times smaller than the original MobilenetV2/U-net while it provides similar accuracy with the original U-net and outperforms the U-net based alternatives. Moreover, the models were tested in three different computing setups, two conventional machines (GPU-enabled and CPU-only) and an embedded system with low computing resources, proving that the proposed model is quite faster than the U-net-based alternatives in all computing systems and especially in the embedded system.

However, the proposed model could be further improved especially in classification and segmentation tasks adding more classes such as, sandy regions, bedrocks, craters, etc, using a refined dataset with synthetic and real lunar images. Given that a qualitative dataset from the lunar surface will be available in the near future due to the planned missions of NASA's Artemis program, the proposed architecture is able to provide a significant potential in lunar exploration, ensuring safe and precise navigation.

6.4 Conclusions of the precise positioning and mapping methodology in GNSS-denied environments

The present study, proposes an alternative mapping methodology which focuses on point localization in challenging GNSS-denied environments in terms of feature-poor information and low illumination. The main contribution of this study is that solves the issue of characteristic points' precise localization in a few-centimeter level of accuracy using only an RGB-depth camera, a conventional computing system and at least one fiducial marker while a novel deep learning-based SLAM method, the MLC

and P.A methods compose a pipeline of algorithmic processing in order to provide the desired coordinate estimations.

In other words, instead of similar computer vision systems, this study focuses on the accurate point localization using a coordinate system defined in the scene, based on the pose of a physical marker. This fact, makes the methodology completely comparable with the traditional surveying process in which the measurements are conducted using a geodetic total station and the coordinate system is defined in the scene, using the internal geometry of the total station on a reference point. However, while the traditional surveying requires significant human effort and a quite costly equipment, the proposed methodology is conducted with just following a specific path of the scene, being able to be used not only by humans but also by mobile robotic systems. It's worth noting that the proposed methodology doesn't aspires to replace the traditional surveying, since it is a robust and well-established methodology with the highest accuracy in point localization, instead the present study could be utilized as an alternative in GNSS-denied environments, that is hard or impossible of using topographic equipment.

Although the present study provides several novel advances compared with the conventional localization methods, some limitations are still under research. Although the mapping process is quite straightforward, is affected by factors including long distance between the origin and the targets, the complicated paths and the high speed of the camera which are able to decrease the accuracy of the results. However, knowing the factors that decrease the accuracy, is setting the basis for the future research of the further methodology development. For instance, the optimization of the SLAM algorithm in order to maintain its accuracy in larger distances, or effectively match and export keyframes with high camera speed, could increase even more the efficiency of this study.

Taking into account all the above constraints, the main goal of this study for the future is to achieve accuracy in a level of 2-5 cm in feature-poor scenes with low illumination and trajectory paths larger than 20 m. This level of accuracy under the aforementioned challenging conditions in combination with the cost-effective equipment is possible to change the way of mainstream point localization introducing a precise positioning alternative with potential use in autonomous robotic systems.

6.5 Summary of conclusions

In this dissertation, a visual localization framework was presented, which is composed of deep learning-based methodologies for critical computer vision tasks including feature extraction, SLAM, semantic segmentation and point positioning, with main goal to reinforce the potential of scene understanding, localization and mapping in unstructured and planetary environments.

One of the main challenges of the dissertation was the lack of training and benchmark datasets focused on unstructured environments, compared with urban, vegetated and indoor environments where multiple datasets are publicly available in order to train and evaluate deep learning architectures, algorithms and SLAM implementations. Thus, two datasets were designed, including images from planetary-like and real-planetary scenes, for the training and evaluation processes of the proposed deep learning models and the related state-of-the-art algorithms.

The second challenge was to investigate the potential of deep learning-based feature extractors in planetary environments while afterwards to modify and improve the selected approaches. Regarding the selected architectures, SuperPoint is a self-supervized convolutional neural network, which is able to extract keypoints and descriptors while the pre-trained model, provided by the authors of DeTone *et al.* 2018, was trained using the general-purpose dataset COCO. Through the experimentation and evaluation process, the architecture was further improved providing satisfactory results in unstructured environments, compared with the pre-trained and re-trained original SuperPoint models and several conventional keypoint detectors and descriptors. Concerning the HF-net neural network, it is an encoder-decoder CNN-basesd architecture which is trained through a teacher-student approach using the SuperPoint and NetVLAD as teachers for local and global feature extraction respectively. The HF-net was further improved replacing the shared encoder MobilenetV2 with the MobilenetV3-large due to its increased efficiency in terms of performance time and accuracy. The new version of HF-net, called HF-net2 was trained and evaluated using the proposed training and evaluation datasets aiming to extract accurate keypoint detectors and descriptors in unstructured and planetary environments. The experimentation proved that the HF-net2 outperforms the original HF-net model and several handcrafted algorithms in terms of illumination and viewpoint changes while its efficiency was quite close to the SuperPoint model which is the teacher of both HF-net and HF-net2 models.

The third challenge was the development of a visual SLAM system focused on unstructured environments aiming to improve the autonomous navigation in feature-poor scenes under intense lighting changes or low illumination conditions. Thus, a SLAM algorithm based on ORB-SLAM2 was proposed, which utilizes the HF-net2 model as feature extractor instead of the ORB algorithm, specializing the SLAM system in unstructured and planetary scenes. For the evaluation of the proposed SLAM, an extended experimentation was conducted in two different study areas with feature-poor information (a) a rocky scene and (b) a sandy scene, in different illumination conditions and trajectory paths. The equipment of experimentation includes an RGB-depth camera, a conventional laptop for the video recording and an RTK-GNSS receiver for the ground truth measurements. The proposed SLAM proved that is able to provide robust and accurate results maintaining its accuracy in high, medium and low illumination outperforming the ORB-SLAM2 especially in challenging environments with low illumination.

The fourth challenge was the development of a deep learning architecture for semantic segmentation focused on lunar environment aiming to recognize rocks and boulders which could harm the attached equipment of an autonomous rover. The architecture should be trained with a limited size of dataset while being efficient in computing systems with low resources. Thus, a semantic segmentation architecture was developed based on U-net while a modified MobilenetV2 was implemented, aiming to reduce the training parameters of the proposed architecture. The present architecture was trained and tested using a dataset with simulated lunar scenes while three different architectures, an original MobilenetV2/U-net, a VGG16/U-net and an original U-net architecture were also trained and evaluated under the same parametrization. The results proved that the proposed architecture outperforms the original MobilenetV2/U-net and the VGG16/U-net and provides similar accuracy of the U-net while it contains only 220,000 parameters instead of the original MobilenetV2/U-net, the VGG16/U-net and the U-net which include about 8,000,000, 24,000,000 and 31,000,000 parameters respectively. The models were also evaluated in terms of performance-time using three computing setups: (a) a GPU-enabled machine, (b) a CPU-only machine and (c) a Raspberry Pi 4 embedded system. The results proved the superiority of the proposed architecture, since it was 2 times faster than U-net in a GPU-enabled machine, nine times faster in a CPU-only machine and 20 times faster in a Raspberry Pi 4 embedded machine.

The fifth challenge was the development of a precise positioning methodology for challenging GNSS-denied environments. The methodology should estimate the coordinates of specific points with centimeter-level of accuracy in several scenes of unstructured environments with feature-poor information and medium to low illumination. Thus, a precise positioning methodology was developed inspired by the traditional surveying using computer vision and deep learning techniques, with limited requirements in terms of equipment. More specifically, the methodology could divided in two main procedures. (a) the field work and (b) the data processing. Regarding the field work, the user or a robotic system is able to map the study area using only an RGB and depth sensor, a conventional machine (laptop) and at least one fiducial marker while after placing the targets on the ground, the sensors need only to cross a path which follows the arrangement of the targets in the scene. After the field work, the recorded video feeds the algorithmic pipeline which is a combination of the proposed SLAM system with target detection, localization and optimization techniques. The experimentation proved that the methodology is able to provide centimeter-level of accuracy in three different study areas including an unstructured urban, a sandy and a rocky scene with several conditions of illumination. Moreover, the use of a deep learning model focused on unstructured environments seems to significantly improve the results, since the proposed methodology with use of deep learning-based SLAM, outperforms the proposed methodology with use of ORB-SLAM2, especially in the environments with low illumination.

It is clear that the use of deep learning in unstructured and planetary environments in terms of scene recognition, localization and mapping provides a significant potential for the future applications, reinforcing crucial topics such as autonomous navigation

in harsh environments. The scalability of deep learning-based methodologies is extremely important since the proposed framework could be further optimized or implemented in different environments using enriched or specialized datasets. For instance, after the landing of a rover on the Moon during the NASA Artemis mission, new rover-based data will be available, enriching the proposed dataset with real images from the lunar surface, improving the efficiency of the proposed feature extraction and semantic segmentation models. Moreover, beyond the planetary environments, there is a need of unmanned vehicles and robotic systems in several other hazardous environments such as glacial scenes or smoky areas, where the proposed framework could be utilized using different datasets for the fine-tuning process or/end more sensors such as laser range-finders. Thus, this dissertation aspires to encourage the investigation and development of AI models and datasets using computationally efficient methods and equipment, aiming to reinforce the autonomous navigation focused on challenging environments, improving the future society and well-being.

References

Abadi M, Agarwal A, Barham P et al, TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org

Ajith V S, Jolly, KG, Unmanned aerial systems in search and rescue applications with their path planning: a review, J. Phys.: Conf. Ser. 2115 01-2020, 10.1088/1742-6596/2115/1/012020

Ajoudani A, Zanchettin A.M, Ivaldi S, Albu-Schäffer A, Kosuge K, Khatib O, Progress and prospects of the human–robot collaboration. Auton Robot 42, 957–975, 2018, https://doi.org/10.1007/s10514-017-9677-2

Alcantarilla P, Nuevo J, Bartoli A, Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces, In British Machine Vision Conference (BMVC), Bristol, UK, September 2013

Alkendi Y, Seneviratne L, Zweiri Y, State of the Art in Vision-Based Localization Techniques for Autonomous Navigation Systems, in IEEE Access, vol. 9, pp. 76847-76874, 2021, doi: 10.1109/ACCESS.2021.3082778

Animesh S, Harsh S,, Mangal K, Autonomous Detection and Tracking of a High-Speed Ground Vehicle using a Quadrotor UAV, In Proceedings of the AIAA Scitech Forum 2019, San Diego, California, 7-11 January

Appel M, Izydorczyk D, Weber S, Mara M, Lischetzke T, The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers, Computers in Human Behavior, Volume 102, 2020, Pages 274-286, DOI: 10.1016/j.chb.2019.07.031

Aqel M.O.A, Marhaban M.H, Saripan M.I, Napsiah bt. I, Review of visual odometry: types, approaches, challenges, and applications. SpringerPlus 5, 1897, 2016 DOI: 10.1186/s40064-016-3573-7

Arandjelovic R, Gronat P, Torii A, Pajdla T, Sivic J, NetVLAD: CNN Architecture for Weakly Supervised Place Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297-5307

Arshad S, Kim G-W. Role of Deep Learning in Loop Closure Detection for Visual and Lidar SLAM: A Survey. Sensors. 2021, 21(4):1243, DOI: 10.3390/s21041243 (loop closure)

Aulinas J, Carreras M, Llado X, Salvi J, Garcia R, Prados R, Petillot Y, Feature extraction for underwater visual SLAM, OCEANS 2011 IEEE - Spain, Santander, Spain, 2011, pp. 1-7, doi: 10.1109/Oceans-Spain.2011.6003474

Badrinarayanan V, Kendall A, Cipolla R, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, ArXiv 2015

Badue C, Guidolini R, Carneiro R, Azevedo P, Cardoso V, Forechi A, Jesus L, Berriel R, Paixão T, Mutz F, Veronese L, Oliveira-Santos T, De Souza A, Self-driving cars: A survey, Expert Systems with Applications, Volume 165, 2021, DOI: 10.1016/j.eswa.2020.113816

Bagnell J, Bradley D, Silver D, Sofman B, Stentz A, Learning for Autonomous Navigation, in IEEE Robotics & Automation Magazine, vol. 17, no. 2, pp. 74-84, 2010, doi: 10.1109/MRA.2010.936946

Baheti B, Innani S, Gajre S, Talbar S, Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment, In Proceedings of CVPRW, Seattle, WA, USA, 2020, pp. 1473-1481, DOI: 10.1109/CVPRW50498.2020.00187

Baheti B, Innani S, Gajre S, Talbar S, Semantic scene segmentation in unstructured environment with modified DeepLabV3+, Pattern Recognition Letters, Volume 138, 2020, Pages 223-229, DOI: 10.1016/j.patrec.2020.07.029.

Balntas V, Lenc K, Vedaldi A, Mikolajczyk K, HPatches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5173-5182

Banerjee K, Prasad V, Gupta RR, Vyas K, Anushree H, Mishra B, Exploring Alternatives to Softmax Function, ArXiv 2020

Bastos A, Hasegawa H, Behavior of GPS Signal Interruption Probability under Tree Canopies in Different Forest Conditions. Eur. J. Remote Sens. 2013, 46, 613–622. DOI:10. 5721/EuJRS20134636.

Bay H, Ess A, Tuytelaars T, Gool L, Speeded-Up Robust Features (SURF), Computer Vision and Image Understanding, Volume 110, Issue 3, 2008, Pages 346-359, DOI: 10.1016/j.cviu.2007.09.014

Bañón L, Pagán J.I, López I, Banon C, Aragonés L, Validating UAS-Based Photogrammetry with Traditional Topographic Methods for Surveying Dune Ecosystems in the Spanish Mediterranean Coast. J. Mar. Sci. Eng., 2019, 7, 297. DOI: 10.3390/jmse7090297

Bobbe M, Kern A, Khedar Y, Batzdorfer S, Bestmann U, An Automated Rapid Mapping Solution Based on ORB SLAM 2 and Agisoft Photoscan API, International Micro Air Vehicle Conference and Flight Competition (IMAV), Toulouse, France, Sept 18 - 22, 2017

Bojanic D, Bartol K, Pribanic T, Petkovic T, Donoso Y, Mas J, On the Comparison of Classic and Deep Keypoint Detector and Descriptor Methods, 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), Dubrovnik, Croatia, 2019, pp. 64-69, DOI: 10.1109/ISPA.2019.8868792

Bowman S, Atanasov N, Daniilidis K, Pappas G, Probabilistic data association for semantic SLAM, 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 1722-1729, DOI: 10.1109/ICRA.2017.7989203.

Bradski G, The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000
Breiman L, Random Forests. Machine Learning 45, 5–32, 2001, DOI: 10.1023/A:1010933404324

Brock O, Park J, Toussaint M, Mobility and manipulation. In Springer Handbook of Robotics, Springer: Cham, Switzerland, 2016, pp. 1007–1036

Brosh E, Friedmann M, Kadar I, Lavy L, Levi E, Rippa S, Lempert Y, Fernandez-Ruiz B, Herzig R, Darrell T, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 0-0

Burguera A, Gonzalez Y, Oliver G, Probabilistic Sonar Scan Matching for Robust Localization, Proceedings 2007 IEEE International Conference on Robotics and Automation, Rome, Italy, 2007, pp. 3154-3160, DOI: 10.1109/ROBOT.2007.363959

Burri M, Nikolic J, Gohl P, Schneider T, The EuRoC micro aerial vehicle datasets, The International Journal of Robotics Research, 2016, DOI:10.1177/0278364915620033

Calonder M, Lepetit V, Strecha C, Fua P, BRIEF: Binary Robust Independent Elementary Features, ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg, DOI: 10.1007/978-3-642-15561-1_56

Canny J, A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, (6), 679–698, 1986

Cao Y, He Z, Wang L, Wang W, Yuan Y, Zhang D, Zhang J, Zhu P, Gool L, Han J, Hoi S, Hu Q, Liu M, VisDrone-DET2021: The Vision Meets Drone Object Detection Challenge Results, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 2847-2854

Carrera-Hernández JJ, Levresse G, Lacan P, Is UAV-SfM surveying ready to replace traditional surveying techniques?, International Journal of Remote Sensing, 2020 41:12, 4820-4837, DOI: 10.1080/01431161.2020.1727049

Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully

Connected CRFs, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 2018, DOI: 10.1109/TPAMI.2017.2699184.

Cheng J, Cheng H, Meng M, Zhang H, Autonomous Navigation by Mobile Robots in Human Environments: A Survey, 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), Kuala Lumpur, Malaysia, 2018, pp. 1981-1986, DOI: 10.1109/ROBIO.2018.8665075

Chhabra S, Rohilla R, A Comparative Study on Semantic Segmentation Algorithms for Autonomous Driving Vehicles, Ijraset Journal For Research in Applied Science and Engineering Technology, 2022, DOI: 10.22214/ijraset.2022.44511

Chiang KW, Tsai GJ, Chang HW, Joly C, EI-Sheimy N, Seamless navigation and mapping using an INS/GNSS/grid-based SLAM semi-tightly coupled integration scheme, Information Fusion, 2019, 50, 181-196, DOI: 10.1016/j.inffus.2019.01.004

Chiodini S, Pertile M, Debei A, Occupancy grid mapping for rover navigation based on semantic segmentation, ACTA IMEKO, 2021, 10.21014/acta_imeko.v10i4.1144
Chiodini S, Torresin L, Pertile M, Debei S, Evaluation of 3D CNN Semantic Mapping for Rover Navigation, ArXiv 2020

Chollet F et al, Keras, 2015 GitHub. Retrieved from https://github.com/fchollet/keras

Chollet F, Xception: Deep Learning with Depthwise Separable Convolutions, ArXiv, 2016

Christiansen P, Kragh M, Brodskiy Y, Karstoft H, UnsuperPoint: End-to-end Unsupervised Interest Point Detector and Descriptor, arXiv 2019

Claudet T, Tomita K, Ho K, Benchmark Analysis of Semantic Segmentation Algorithms for Safe Planetary Landing Site Selection, in IEEE Access, vol. 10, pp. 41766-41775, 2022, DOI: 10.1109/ACCESS.2022.3167763

Cohen A, Rivlin E, Shimshoni I, Sabo E, Memory based active contour algorithm using pixel-level classified images for colon crypt segmentation, Computerized Medical Imaging and Graphics, Volume 43, 2015, Pages 150-164, DOI: 10.1016/j.compmedimag.2014.12.006.

Cortes C, Vapnik V, Support-vector networks. Mach Learn 20, 273–297, 1995, DOI: 10.1007/BF00994018

Cowen R.: Opportunity rolls out of purgatory. Science News 167, 2005

Csurka G, Dance C, Fan L, Willamowski J, Bray C, Visual categorization with bag of keypoints, In Proceedings of the European Conference on Workshop on Statistical Learning in Computer Vision, Prague, The Czech Republic, 2004

Dalal N, Triggs B, Histograms of Oriented Gradients for Human Detection. In proceedings of CVPR, 2005

DeTone D, Malisiewicz T, Rabinovich A, SuperPoint: Self-Supervised Interest Point Detection and Description, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 224-236

Di K, Yue Z, Liu Z, Wang S, Automated rock detection and shape analysis from mars rover imagery and 3D point cloud data. J. Earth Sci. 24, 125–135, 2013, DOI: 10.1007/s12583-013-0316-3

Ding, L., Deng, Z., Gao, H. et al. Planetary rovers' wheel–soil interaction mechanics: new challenges and applications for wheeled mobile robots. Intel Serv Robotics 4, 17–38 (2011). https://doi.org/10.1007/s11370-010-0080-5

Doer C, and Trommer F, An EKF Based Approach to Radar Inertial Odometry, 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Karlsruhe, Germany, 2020, pp. 152-159, DOI: 10.1109/MFI49285.2020.9235254.

Driver T, Skinner K, Dor M, Tsiotras P, AstroVision: Towards autonomous feature detection and description for missions to small bodies using deep learning, Acta Astronautica, 2023, DOI: 10.1016/j.actaastro.2023.01.009

Duan C, Junginger S, Huang J, Jin K, Thurow K, Deep Learning for Visual SLAM in Transportation Robotics: A review, Transportation Safety and Environment, Volume 1, Issue 3, 12 December 2019, Pages 177–184, DOI: 10.1093/tse/tdz019

Dunbar B, What is Artemis?, NASA. Archived from the original on August 7, 2019. Retrieved May 13, 2023

Dunlop H, Thompson D.R, Wettergreen D, Multi-Scale Features for Detection and Segmentation of Rocks in Mars Images. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007, DOI: 10.1109/CVPR.2007.383257.

Ebadi K, Coble K, Atha D, Schwartz R, Padgett C, Hook JV, Semantic mapping in unstructured environments: Toward autonomous localization of planetary robotic explorers, IEEE Aerospace Conference, 2022

Eldén L, A weighted pseudoinverse, generalized singular values, and constrained least squares problems, BIT, 1982, 22, 487–502, DOI: 10.1007/BF01934412

Elsken T, Metzen J, Hutter F, Neural Architecture Search: A Survey, arXiv 2019

Eui-ik J, Sunghak K, Soyoung P, Juwon K, Imho C, Semantic segmentation of seagrass habitat from drone imagery based on deep learning: A comparative study, Ecological Informatics, 66, 2021, DOI: 10.1016/j.ecoinf.2021.101430

Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv:1805.04687, 2018

Fujita K, Ichimura N, A Terrain Classification Method for Planetary Rover Utilizing Dynamic Texture. In Proceedings of the AIAA Guidance, Navigation, and Control Conference, American Institute of Aeronautics and Astronautics, Reston, VA, USA, 8 August 2011; pp. 1–13, DOI: 10.2514/6.2011-6580

Furgale P T, Carle P, Enright J, and Barfoot T D, The Devon Island Rover Navigation Dataset, International Journal of Robotics Research, 2012

Furgale P T, Carle P, Enright J, and Barfoot T D, The Devon Island Rover Navigation Dataset, International Journal of Robotics Research, 2012

Furlan F, Rubio E, Sossa H, Ponce V, CNN Based Detectors on Planetary Environments: A Performance Evaluation, Front. Neurorobot.,14, 2020, DOI: 10.3389/fnbot.2020.590371

Furlan F, Rubio E, Sossa H, Ponce V, Rock Detection in a Mars-Like Environment Using a CNN. In proceedings of MCPR 2019, Springer, DOI:10.1007/978-3-030-21077-9_14

Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J, A Review on Deep Learning Techniques Applied to Semantic Segmentation, ArXiv 2017

Garcia-Fidalgo E, Ortiz A, ibow-lcd: An appearance-based loop-closure detection approach using incremental bags of binary words. IEEE Robot. Autom. Lett. 2018, 3:3051–3057. DOI: 10.1109/LRA.2018.2849609

Garrido-Jurado S, Muñoz Salinas R, Madrid-Cuevas FJ, Medina-Carnicer R, Generation of fiducial marker dictionaries using mixed integer linear programming, Pattern Recognition, 2016, 51, 481-491, DOI: 10.1016/j.patcog.2015.09.023

Geiger A, Lenz P, Urtasun R, Are we ready for autonomous driving? The KITTI vision benchmark suite, IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3354-3361, DOI: 10.1109/CVPR.2012.6248074.

George D.A, Privitera C.M, Blackmon T.T, Zbinden E, Stark L.W, Segmentation of Stereo Terrain Images. In proceedings of Human Vision and Electronic Imaging V, Bellingham, WA, USA, 2000, Volume 3959, pp. 669–679, DOI: 10.1117/12.387204

Giubilato R, Gentil C, Vayugundla M, Schuster M, Vidal-Calleja T, Triebel R, GPGM-SLAM: a Robust SLAM System for Unstructured Planetary Environments with Gaussian Process Gradient Maps, ArXiv 2021

Giubilato R, Stürzl W, Wedler A, Triebel R, Challenges of SLAM in Extremely Unstructured Environments: The DLR Planetary Stereo,Solid-State LiDAR, Inertial Dataset, IEEE Robotics and automation letters vol.7, 2022

Gong X, Liu J, Rock detection via superpixel graph cuts, In Proceedings of ICIP, 2149-2152, 2012 DOI:10.1109/ICIP.2012.6467318

Grady L, Random Walks for Image Segmentation, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 11, pp. 1768-1783, 2006, DOI: 10.1109/TPAMI.2006.233.

Guan T, He Z, Song R, Manocha D, Zhang L, TNS: Terrain Traversability Mapping and Navigation System for Autonomous Excavators, ArXiv 2021

Guan T, Kothandaraman D, Chandra R, Sathyamoorthy A.J, Weerakoon K, Manocha D, GA-Nav: Efficient Terrain Segmentation for Robot Navigation in Unstructured Outdoor Environments, in IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 8138-8145, July 2022, DOI: 10.1109/LRA.2022.3187278

Guastella DC, Muscato G, Learning-Based Methods of Perception and Navigation for Ground Vehicles in Unstructured Environments: A Review. Sensors. 2021, 21(1):73, DOI: 10.3390/s21010073

Guo J, Borges P, Park C, Gawel A, Local Descriptor for Robust Place Recognition using LiDAR Intensity, arXiv 2018

Guo Y, Liu Y, Georgiou T, Leu M, A review of semantic segmentation using deep neural networks. Int J Multimed Info Retr 7, 87–93, 2018 DOI: 10.1007/s13735-017-0141-z

Gupta, P., Pareek, B., Singal, G., Rao, D.V., (2022) Edge device based Military Vehicle Detection and Classification from UAV. Multimed Tools Appl 81, 19813–19834, https://doi.org/10.1007/s11042-021-11242-y

Hao S, Zhou Y, Guo Y, A Brief Survey on Semantic Segmentation with Deep Learning, Neurocomputing, Volume 406, 2020, Pages 302-321, DOI: 10.1016/j.neucom.2019.11.118

Haque A, Elsaharti A, Elderini T, Elsaharty M.A, Neubert J, UAV Autonomous Localization Using Macro-Features Matching with a CAD Model. Sensors, 2020, 20, 743, DOI: 10.3390/s20030743

Harris C, Stephens M, A Combined Corner and Edge Detector, Alvey Vision Conference. Vol. 15, 1988

Harris C.R, Millman K.J, van der Walt S.J, et al. Array programming with NumPy. Nature 585, 357–362, 2020, DOI: 10.1038/s41586-020-2649-2.

Hartigan J. A, and Wong M.A, Algorithm AS 136: A K-Means Clustering Algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, 1979, pp. 100–08. DOI: 10.2307/2346830

Hashimoto S, Mori K, Lunar Crater Detection based on Grid Partition using Deep Learning, In Proceedings of SACI, Timisoara, Romania, 2019, pp. 75-80, DOI: 10.1109/SACI46893.2019.9111474

Heng L, Choi B, Cui Z, Geppert M, Hu S, Kuan B, Liu P, Nguyen R, Yeo Y, Geiger A, et al. Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. IEEE International Conference on Robotics and Automation (ICRA), Montreal, Canada, May 20-24, 2019

Hewitt R, Boukas E, Azkarate M, Pagnamenta M, Marshall J, Gasteratos A, Visentin G, The Katwijk beach planetary roverdataset, The international Journal of Robotics research, 2018, DOI: 10.1177/0278364917737153

Hong S, Bangunharcana A, Park J-M, Choi M, Shin H-S, Visual SLAM-Based Robotic Mapping Method for Planetary Construction. Sensors, 2021, 21(22):7715 DOI: 10.3390/s21227715

Howard A, Sandler M, Chu G, Chen L, Chen B, Tan M, Wang W, Zhu Y, Pang R, Vasudevan V, Le Q, Adam H, Searching for MobileNetV3, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314-1324

Howard A, Seraji H, An intelligent terrain-based navigation system for planetary rovers, in IEEE Robotics & Automation Magazine, vol. 8, no. 4, pp. 9-17, Dec. 2001, DOI: 10.1109/100.973242

Hu Y, Xiao J, Liu L, Zhang L, Wang Y, Detection of Small Impact Craters via Semantic Segmenting Lunar Point Clouds Using Deep Learning Network, Remote Sensing. 2021, 13(9):1826, DOI: 10.3390/rs13091826

Huang G, Du S, & Wang D, GNSS techniques for real-time monitoring of landslides: a review. Satell Navig 4, 5, 2023 DOI: 10.1186/s43020-023-00095-5

Huang G, Yang Li, Cai Y, Zhang D, Terrain classification-based rover traverse planner with kinematic constraints for Mars exploration, Planetary and Space Science, 209, 2021, DOI: 10.1016/j.pss.2021.105371

Huang K, Xiao J, Stachniss C, Accurate Direct Visual-Laser Odometry with Explicit Occlusion Handling and Plane Detection, 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 2019, pp. 1295-1301, DOI: 10.1109/ICRA.2019.8793629

Huang X, Cheng X, Geng Q, Cao B, Zhou B, Wang P, Lin Y, Yang R, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 954-960

Hunter J.D, Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Inzerillo L, Mino G, Roberts R, Image-based 3D reconstruction using traditional and UAV datasets for analysis of road pavement distress, Automation in Construction, Volume 96, 2018, Pages 457-469, DOI: 10.1016/j.autcon.2018.10.010

Isık S, Ozkan K, A Comparative Evaluation of Well-known Feature Detectors and Descriptors . International Journal of Applied Mathematics Electronics and Computers, 2015 3 (1) , 1-6 . DOI: 10.18100/ijamec.60004

Javanmardi M, Javanmardi E, Gu Y, Kamijo S, Towards High-Definition 3D Urban Mapping: Road Feature-Based Registration of Mobile Mapping Systems and Aerial Imagery. Remote Sens. 2017, 9, 975, DOI: 10.3390/rs9100975

Jende P, Nex F, Gerke M, Vosselman G, A fully automatic approach to register mobile mapping and airborne imagery to support the correction of platform trajectories in GNSS-denied urban areas. ISPRS J. Photogramm. Remote Sens. 2018, 141, 86–99, DOI:10.1016/j.isprsjprs.2018.04.017

Jia Y, Wan G, Liu L, Wu Y, Zhang C, Automated Detection of Lunar Craters Using Deep Learning, In Proceedings of ITAIC, Chongqing, China, 2020, pp. 1419-1423, DOI: 10.1109/ITAIC49862.2020.9339179.

Jordan S, Moore J, Hovet S, Box J, Perry J, Kirsche K, Lewis D, Tse H, State-of-the-art technologies for UAV inspections. IET Radar Sonar Navig., 12: 151-164, 2018 DOI: 10.1049/iet-rsn.2017.0251

Jung K, Hitchcox T, Forbes J, Performance Evaluation of 3D Keypoint Detectors and Descriptors on Coloured Point Clouds in Subsea Environments, ArXiv 2022

Kalacska M, Lucanus O, Arroyo-Mora J.P, Laliberté É, Elmer K, Leblanc G, Groves A. Accuracy of 3D Landscape Reconstruction without Ground Control Points Using Different UAS Platforms. Drones, 2020, 4, 13. DOI: 10.3390/drones4020013

Kass M, Witkin A, Terzopoulos D, Snakes: Active contour models, International journal of computer vision, vol. 1, no. 4, pp. 321–331, 1988, DOI: 10.1007/BF00133570

Kasson J. M. & Plouffe W, An analysis of selected computer interchange color spaces, ACM Transactions on Graphics (TOG), vol. 11, no. 4, pp. 373–405, 1992.

Kazerouni I, Fitzgerald L, Dooly G, Toal D, A survey of state-of-the-art on visual SLAM, Expert Systems with Applications, Volume 205, 2022, DOI: 10.1016/j.eswa.2022.117734

Ke Y, Sukthankar R, PCA-SIFT: a more distinctive representation for local image descriptors, Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., Washington, DC, USA, 2004, pp. II-II, DOI: 10.1109/CVPR.2004.1315206

Kim J, Sukkarieh S, 6DOF SLAM aided GNSS/INS navigation in GNSS denied and un-known environments, J. Glob. Position. Syst. 2005, 4, 120–128. DOI:10.5081/jgps. 4.1.120

Kostavelis I, Nalpantidis L, Boukas E, Rodrigalvarez A, Stamoulias I, Lentaris G, Diamantopoulos D, Siozios K, Soudris D, Gasteratos A, SPARTAN: Developing a Vision System for Future Autonomous Space Exploration Robots. J. Field Robotics, 31: 107-140, 2014, https://doi.org/10.1002/rob.21484

Kuang B, Gu C, Rana ZA, Zhao Y, Sun S, Nnabuife SG. Semantic Terrain Segmentation in the Navigation Vision of Planetary Rovers—A Systematic Literature Review, Sensors. 2022, 22(21):8393. DOI: 10.3390/s22218393

Kuang B, Rana ZA, Zhao Y, Sky and Ground Segmentation in the Navigation Visions of the Planetary Rovers, Sensors, 2021, 21(21):6996 DOI: 10.3390/s21216996

Kuang B, Wisniewski M, Rana ZA, Zhao Y, Rock Segmentation in the Navigation Vision of the Planetary Rovers, Mathematics, 2021, 9(23):3048. DOI: 10.3390/math9233048

Lai T, A Review on Visual-SLAM: Advancements from Geometric Modelling to Learning-based Semantic Scene Understanding, arXiv 2022

Larsson M, Stenborg E, Toft C, Hammarstrand L, Sattler T, Kahl F, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 31-41

Lateef F, Ruichek Y, Survey on semantic segmentation using deep learning techniques, Neurocomputing, Volume 338, 2019, Pages 321-348, DOI: 10.1016/j.neucom.2019.02.003

Lategahn A, La H, A Review of SLAM Techniques and Security in Autonomous Driving, 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 2019, pp. 602-607, DOI: 10.1109/IRC.2019.00122.

Lategahn H, Geiger A, Kitt B, Visual SLAM for autonomous ground vehicles, 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 2011, pp. 1732-1737, DOI: 10.1109/ICRA.2011.5979711.

Li D, Shi X, Long Q, Liu S, Yang W, Wang F, Wei Q, Qiao F, DXSLAM: A Robust and Efficient Visual SLAM System with Deep Features, arXiv preprint arXiv:2008.05416, 2020

Li J, Besada J.A, Bernardos A.M, Tarrío P, Casar J.R, A novel system for object pose estimation using fused vision and inertial data, Information Fusion, 2017, 33, 15-28, DOI: 10.1016/j.inffus.2016.04.006

Li R, Wang S, Gu D, Ongoing Evolution of Visual SLAM from Geometry to Deep Learning: Challenges and Opportunities. Cogn Comput 10, 875–889, 2018, DOI: 10.1007/s12559-018-9591-8

Li R, Wang S, Long Z, Gu D, UnDeepVO: Monocular Visual Odometry Through Unsupervised Deep Learning, 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 2018, pp. 7286-7291, DOI: 10.1109/ICRA.2018.8461251

Lin M, Chen Q, Yan S, Network in Network, arXiv 2013

Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick L, Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. DOI: 10.1007/978-3-319-10602-1_48

Liu C, Xu J, Wang F, A Review of Keypoints' Detection and Feature Description in Image Registration, Hindawi, Scientific programming, 2021

Liu C, Xu J, Wang, F. A Review of Keypoints' Detection and Feature Description in Image Registration. Sci. Program. 2021, 2021, 8509164

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC, SSD: Single Shot MultiBox Detector, In Proceedings of ECCV 2016, Springer, Cham. DOI: 10.1007/978-3-319-46448-0_2

Liu Y, Li J, Huang K, Xiangting L, Xiuyuan Q, Chang L, Long Y, Zhou J, MobileSP: An FPGA-Based Real-Time Keypoint Extraction Hardware Accelerator for Mobile VSLAM, in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 12, pp. 4919-4929, Dec. 2022, DOI: 10.1109/TCSI.2022.3190300

Long J, Shelhamer E, Darrell T, Fully Convolutional Networks for Semantic Segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431-3440

Lowe D, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision. 60 (2): 91–110, 2004 CiteSeerX 10.1.1.73.2924

Lu S, Oij S.L, Horizon Detection for Mars Surface Operations. In Proceedings of the 2017 IEEE Aerospace Conference, Big Sky, MT, USA, 4–11 March 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8, DOI: 10.1109/AERO.2017.7943975

Lu, S, Mars surface image (Curiosity rover) labeled data set version 1, updated January 2023

Ma J, Jiang X, Fan A, Jiang J, Yan J, Image Matching from Handcrafted to Deep Features: A Survey. Int J Comput Vis 129, 23–79, 2021 DOI: 10.1007/s11263-020-01359-2

Ma Y, Li Q, Chu L, Zhou Y, Xu C, Real-Time Detection and Spatial Localization of Insulators for UAV Inspection Based on Binocular Stereo Vision. Remote Sensing, 2021, 13, 230, DOI: 10.3390/rs13020230

Mair E, Hager D, Burschka D, Suppa M, Hirzinger G, Adaptive and generic corner detection based on theaccelerated segment test.In:Computer Vision ECCV 2010,Springer, pp. 183–196

Martins H, Bruno S, Colombini E, LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method, Neurocomputing, Volume 455, 2021, Pages 97-110, DOI: 10.1016/j.neucom.2021.05.027

Metzger K, Mortimer P, Wuensche JH, A Fine-Grained Dataset and its Efficient Semantic Segmentation for Unstructured Driving Scenarios, ArXiv 2021

Meyer L, Smísek M, Villacampa F, Maza L, Medina D, Schuster M, Steidle F, Vayugundla M, Muller M, Rebele B, Wedler A, Triebel R, The MADMAX data set for visual-inertial rover navigation on Mars, J Field Robotics, 38, 833– 853, 2021, DOI: 10.1002/rob.22016

Mihail R.P, Workman S, Bessinger Z, Jacobs N, Sky segmentation in the wild: An empirical study, In Proceedings of WACV, Lake Placid, NY, USA, 7–10 March 2016

Mingxing T, Le Q, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, ArXiv 2019

Mo Y, Wu Y, Yang X, Liu F, Liao Y, Review the state-of-the-art technologies of semantic segmentation based on deep learning, Neurocomputing, Volume 493, 2022, Pages 626-646, DOI: 10.1016/j.neucom.2022.01.005

Moghe R, Zanetti R, A Deep Learning Approach to Hazard Detection for Autonomous Lunar Landing. J Astronaut Sci 67, 2020, DOI: 10.1007/s40295-020-00239-8

Mohamed S, , Haghbayan M, Westerlund T, Heikkonen J, Tenhunen H, Plosila J, A Survey on Odometry for Autonomous Navigation Systems, in IEEE Access, vol. 7, pp. 97466-97486, 2019, DOI 10.1109/ACCESS.2019.2929133

Mokssit S, Licea D, Guermah B, Ghogho M, Deep Learning Techniques for Visual SLAM: A Survey, in IEEE Access, vol. 11, pp. 20026-20050, 2023, DOI: 10.1109/ACCESS.2023.3249661

Mostafa M, Zahran S, Moussa A, El-Sheimy N, Sesay A, Radar and Visual Odometry Integrated System Aided Navigation for UAVs in GNSS Denied Environment. Sensors 2018, 18, 2776. DOI:10.3390/s18092776

Mostafa M, Zahran S, Moussa A, El-Sheimy N, Sesay A. Radar and Visual Odometry Integrated System Aided Navigation for UAVS in GNSS Denied Environment. Sensors. 2018; 18(9):2776. DOI: 10.3390/s18092776

Munguía R, Urzua I, Bolea Y, Grau A, Vision-Based SLAM System for Unmanned Aerial Vehicles. Sensors 2016, 16, 372. DOI:10.3390/s16030372

Munoz-Salinas R, Medina-Carnicer R, UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers, Pattern Recognition, p. 107193, 2020, DOI: 10.1016/j.patcog.2019.107193

Mur-Artal R, Tardos J, ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras, ArXiv 2017

Müller MG, Durner M, Gawel A, Stürzl W, Triebel R, Siegwart R, A Photorealistic Terrain Simulation Pipeline for Unstructured Outdoor Environments, In Proceedings of IROS, Prague, Czech Republic, 2021, pp. 9765-9772, DOI: 10.1109/IROS51168.2021.9636644

Naseer T, Burgard W, and Stachniss C, Robust Visual Localization Across Seasons, in IEEE Transactions on Robotics, vol. 34, no. 2, pp. 289-302, April 2018, doi: 10.1109/TRO.2017.2788045

Naseer T, Oliveira G, Brox T, Burgard W, Semantics-aware visual localization under challenging perceptual conditions, 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017, pp. 2614-2620, DOI: 10.1109/ICRA.2017.7989305

Noh H, Araujo A, Sim J, Weyand T, Han B, Large-scale image retrieval with attentive deep local features. In CVPR, 2017

Oelsch M, Opdenbosch V, Steinbach E, Survey of Visual Feature Extraction Algorithms in a Mars-like Environment, 2017 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 2017, pp. 322-325, DOI: 10.1109/ISM.2017.58

Ono Y, Trulls E, Fua P, Yi K, LF-Net: Learning Local Features from Images. On NIPS, 2018.

Otsu K, Agha-Mohammadi A, Paton M, Where to Look? Predictive Perception With Applications to Planetary Exploration, in IEEE Robotics and Automation Letters, vol. 3, no. 2, pp. 635-642, April 2018, DOI: 10.1109/LRA.2017.2777526

Panigrahi N, Doddamani SR, Singh M, Kandulna BN, A method to compute location in GNSS denied area, IEEE International CONECCT, 2015, 1-5, DOI: 10.1109/CONECCT.2015.7383907

Panigrahi P, Bisoy S, Localization strategies for autonomous mobile robots: A review, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 8, Part B, 2022, Pages 6019-6039, DOI: 10.1016/j.jksuci.2021.02.015

Partsinevelos P, Chatziparaschis D, Trigkakis D, Tripolitsiotis A. A Novel UAV-Assisted Positioning System for GNSS-Denied Environments. Remote Sensing. 2020; 12(7):1080. DOI: 10.3390/rs12071080

Partsinevelos P, Chatziparaschis D, Trigkakis D, Tripolitsiotis, A, A Novel UAV-Assisted Positioning System for GNSS-Denied Environments. Remote Sens., 2020, 12, 1080. DOI: 10.3390/rs12071080

Pedregosa et al, Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011

Petrakis G, Antonopoulos A, Tripolitsiotis A, Trigkakis D, Partsinevelos P, Precision mapping through the stereo vision and geometric transformations in unknown environments. Earth Sci Inform, 2023, 16, 1849–1865. DOI: 10.1007/s12145-023-00972-2

Petrakis G, Partsinevelos P, Keypoint detection and description in unstructured environments through deep learning, Robotics 2023, MDPI, 12, 137, 10.3390/robotics12050137

Pinto AM, Matos A.C, MARESye: A hybrid imaging system for underwater robotic applications, Information Fusion, 2020, 55, 16-29, DOI: 10.1016/j.inffus.2019.07.014

Poddar S, Kottath R, Karar V, Motion Estimation Made Easy: Evolution and Trends in Visual Odometry. In: Hassaballah, M., Hosny, K. (eds) Recent Advances in Computer Vision. Studies in Computational Intelligence, vol 804. Springer, 2019 DOI: 10.1007/978-3-030-03000-1_13

Qin G, Li Y, Wu L, Xiong J, OPR-SLAM: A Semantic SLAM with Occluded Point Recovery for Dynamic Environments, 2022 41st Chinese Control Conference (CCC), Hefei, China, 2022, pp. 6458-6463, DOI: 10.23919/CCC55666.2022.9902235

Qin T, Li P, Shen S, INS-Mono: A Robust and Versatile MonocularVisual-Inertial State Estimator,ArXiv 2017

Queralta J, Almansa M, Schiano F, Floreano D, Westerlund T, UWB-based System for UAV Localization in GNSS-Denied Environments: Characterization and Dataset, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 4521-4528, DOI: 10.1109/IROS45743.2020.9341042

Quinlan J.R, Induction of decision trees. Mach Learn 1, 81–106, 1986, DOI: 10.1007/BF00116251

Quist E, Niedfeldt P, Beard R, Radar odometry with recursive-RANSAC, in IEEE Transactions on Aerospace and Electronic Systems, vol. 52, no. 4, pp. 1618-1630, August 2016, DOI: 10.1109/TAES.2016.140829

Ramachandran P, Zoph B, Le Q, Searching for Activation Functions, arXiv 2017

Roerdink J, Meijster A, The Watershed Transform: Definitions, Algorithms and Parallelization Strategies, Fundamenta Informaticae, vol. 41, no. 1-2, pp. 187-228, 2000, DOI: 10.3233/FI-2000-411207

Romero-Ramirez F.J, Muñoz-Salinas R., Medina-Carnicer R, Speeded up detection of squared fiducial markers, Image and Vision Computing, 2018, 76, 38-47, DOI: 10.1016/j.imavis.2018.05.004

Romero-Ramirez FJ, Muñoz-Salinas R, Medina-Carnicer R, Speeded up detection of squared fiducial markers, Image and Vision Computing, 2018 76, 38-47, DOI: 10.1016/j.imavis.2018.05.004

Ronneberger O, Fischer P, Brox T, U-Net: Convolutional Networks for Biomedical Image Segmentation, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. DOI: 10.1007/978-3-319-24574-4_28

Rosasco L, Vito E, Caponnetto A, Piana M, Verri A, Are loss functions all the same? Neural Computation. 16 (5), 2004, DOI:10.1162/089976604773135104

Rosten E, Drummond T, Machine Learning for High-speed Corner Detection, Computer Vision – ECCV 2006. Lecture Notes in Computer Science. Vol. 3951. pp. 430–443. doi:10.1007/11744023_34. ISBN 978-3-540-33832-1. S2CID 1388140

Rosten E, Drummond T, Machine Learning for High-Speed Corner Detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg, DOI: 10.1007/11744023_34

Rubio F, Valero F, Llopis-Albert C. A review of mobile robots: Concepts, methods, theoretical framework, and applications. International Journal of Advanced Robotic Systems. 2019;16(2). doi:10.1177/1729881419839596

Rublee E, Rabaud V, Konolige K, Bradski G, ORB: An efficient alternative to SIFT or SURF, 2011 International Conference on Computer Vision, Barcelona, Spain, 2011, pp. 2564-2571, DOI: 10.1109/ICCV.2011.6126544.

Rusu RB, Cousins S, 3D is here: Point Cloud Library (PCL), IEEE International Conference on Robotics and Automation (ICRA), May 9-13, 2011, Shanghai, China
Sahoo B, Biglarbegian M, Melek W. Monocular Visual Inertial Direct SLAM with Robust Scale Estimation for Ground Robots/Vehicles. Robotics. 2021, 10(1):23. DOI: 10.3390/robotics10010023

Samuel B, Introduction to inverse kinematics with Jacobian transpose, pseudoinverse and damped least squares methods, IEEE Transactions in Robotics and Automation, 2004, 17

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L, MobileNetV2: Inverted Residuals and Linear Bottlenecks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L, MobileNetV2: Inverted Residuals and Linear Bottlenecks Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520

Sarlin P, Cadena C, Siegwart R, Dymczyk M, From Coarse to Fine: Robust Hierarchical Localization at Large Scale, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12716-12725

Schubert D, Goll T, Demmel N, Usenko V, Stückler J, Cremers D, The TUM VI Benchmark for Evaluating Visual-Inertial Odometry, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 1680-1687, DOI: 10.1109/IROS.2018.8593419.

Sehar U, Naseem M.L, How deep learning is empowering semantic segmentation. Multimed Tools Appl 81, 30519–30544, 2022, DOI: 10.1007/s11042-022-12821-3

Shi J, Tomasi C, Good features to track. Technical report, 1993 Cornell University
Shi W, Caballero J, Huszár F, Totz J, Aitken A, Bishop R, Rueckert D, Wang Z, Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network, CVPR, 2016

Simoni M, Kockelman K, Gurumurthy K, Bischoff J, Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios, Transportation Research Part C: Emerging Technologies, Volume 98, 2019, DOI: 10.1016/j.trc.2018.11.002

Simonyan K, Zisserman A, Very Deep Convolutional Networks for Large-Scale Image Recognition, ArXiv 2015, DOI 10.48550/arXiv.1409.1556

Singandhupe A, La H, A Review of SLAM Techniques and Security in Autonomous Driving, Third IEEE International Conference on Robotic Computing (IRC), 2019

Smith E, Zuber T, Jackson B, et al. The Lunar Orbiter Laser Altimeter Investigation on the Lunar Reconnaissance Orbiter Mission. Space Sci Rev 150, 209–241, 2010, DOI: 10.1007/s11214-009-9512-y

Solin A, Cortes S, Rahtu E, Kannala J, Inertial Odometry on Handheld Smartphones, 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 2018, pp. 1-5, DOI: 10.23919/ICIF.2018.8455482

Song Y, Shan J, A Framework for Automated Rock Segmentation from the Mars Exploration Rover Imagery. In Proceedings of the ASPRS 2006 Annual Conference, Reno, NV, USA, 1–5 May 2006.

Stanford Artificial Intelligence Laboratory et al, Robotic Operating System, 2018, Retrieved from https://www.ros.org

Swan RM, Atha D, Leopold HA, Gildner M, Oij S, Chiu C, Ono M, AI4MARS: A Dataset for Terrain-Aware Autonomous Driving on Mars, In proceedings of CVPRW, Nashville, TN, USA, 2021, pp. 1982-1991, DOI: 10.1109/CVPRW53098.2021.00226

Taketomi T, Uchiyama H, Ikeda S, Visual SLAM algorithms: a survey from 2010 to 2016. IPSJ T Comput Vis Appl 9, 16, 2017 DOI: 10.1186/s41074-017-0027-2

Tang J, Chen Y, Niu X, Wang L, Chen L, Liu J, Shi C, Hyyppä J, LiDAR Scan Matching Aided Inertial Navigation System in GNSS-Denied Environments. Sensors, 2015, 15, 16710-16728. DOI: 10.3390/s150716710

Tang J, Ericson L, Folkesson J, Jensfelt P, GCNv2: Efficient Correspondence Prediction for Real-Time SLAM, in IEEE Robotics and Automation Letters, vol. 4, no. 4, pp. 3505-3512, Oct. 2019, DOI: 10.1109/LRA.2019.2927954

Tatum M., Liu J., (2017), Unmanned Aircraft System Applications in Construction, Procedia Engineering, Volume 196, Pages 167-175, ISSN 1877-7058, https://doi.org/10.1016/j.proeng.2017.07.187.

Tomaštík J, Mokroš M, Saloň Š, Chudý F, Tunák D, Accuracy of Photogrammetric UAV-Based Point Clouds under Conditions of Partially-Open Forest Canopy. Forests, 2017, 8, 151. DOI: 10.3390/f8050151

Tomita K, Skinner K, Iiyama K, Jagatia B, Nakagawa T, Ho K, Hazard Detection Algorithm for Planetary Landing Using Semantic Segmentation, AIAA 2020-4150. ASCEND 2020, DOI: 10.2514/6.2020-4150

Trigkakis D, Petrakis G, Tripolitsiotis A, Partsinevelos P, Automated Geolocation in Urban Environments Using a Simple Camera-Equipped Unmanned Aerial Vehicle: A Rapid Mapping Surveying Alternative? ISPRS Int. J. Geo-Inf, 2020, 9, 425, DOI: 10.3390/ijgi9070425

Trigkakis D, Petrakis G, Tripolitsiotis A, Partsinevelos P. Automated Geolocation in Urban Environments Using a Simple Camera-Equipped Unmanned Aerial Vehicle: A Rapid Mapping Surveying Alternative? ISPRS International Journal of Geo-Information, 2020, 9(7):425. DOI: 10.3390/ijgi9070425

Urzua S, Munguia R, Grau A, Vision-based SLAM system for MAVs in GPS-denied environments. In: International Journal of Micro Air Vehicles, 2017, 283–296, DOI: 10.1177/1756829317705325

Usenko V, Engel J, Stückler J, Cremers D, Direct visual-inertial odometry with stereo cameras, 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 2016, pp. 1885-1892, DOI: 10.1109/ICRA.2016.7487335

Velusamy P, Rajendran S, Mahendran RK, Naseer S, Shafiq M, Choi J-G. Unmanned Aerial Vehicles (UAV) in Precision Agriculture: Applications and Challenges. Energies. 2022, 15(1):217. https://doi.org/10.3390/en15010217

Vrba M, Heřt D, Saska M, Onboard Marker-Less Detection and Localization of Non-Cooperating Drones for Their Safe Interception by an Autonomous Aerial System, in

IEEE Robotics and Automation Letters, 2019, 4, 3402-3409, DOI: 10.1109/LRA.2019.2927130

Vrba M, Saska M, Marker-Less Micro Aerial Vehicle Detection and Localization Using Convolutional Neural Networks, in IEEE Robotics and Automation Letters, 2020, 5, 2459-2466, DOI: 10.1109/LRA.2020.2972819

Wan W, Peng M, Xing Y, Wang Y, Liu Z, Di K, Teng B, Mao X, Zhao Q, Xin X, Jia M, A Performance comparison of feature detectors for planetary rover mapping and localization, ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLII-3/W1, 2017, pp.149-154, DOI: 10.5194/isprs-archives-XLII-3-W1-149-2017

Wang C, Meng L, She S, Mitchell I, Li T, Tung F, Wan W, Meng M, Silva, C, Autonomous mobile robot navigation in uneven and unstructured indoor environments, 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 2017, pp. 109-116, DOI: 10.1109/IROS.2017.8202145

Wang W, Zhu D, Wang X, Hu Y, Qiu Y, Wang C, Hu Y, Kapoor A, Scherer S, TartanAir: A Dataset to Push the Limits of Visual SLAM, 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020, pp. 4909-4916, DOI: 10.1109/IROS45743.2020.9341801

Wang Y, Zhou Q, Liu J, Xiong J, Gao G, Wu X, Latecki L, Lednet: A Lightweight Encoder-Decoder Network for Real-Time Semantic Segmentation, 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 2019, pp. 1860-1864, DOI: 10.1109/ICIP.2019.8803154.

Weyand T, Araujo A, Cao B, Sim J, Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2575-2584

Wigness M, Eum S, Rogers JG, Han D, Kwon H, A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments, In Proceedings of IROS, Macau, China, 2019, pp. 5000-5007, DOI: 10.1109/IROS40897.2019.8968283

Wu B, Zeng H, Hu H, Illumination invariant feature point matching for high-resolution planetary remote sensing images, Planetary and Space Science, Volume 152, 2018, Pages 45-54, DOI: 10.1016/j.pss.2018.01.007

Wynn R, Huvenne V, Le Bas T, Murton B; Connelly D, Bett B, Ruhl H, Morris K, Peakall J, Parsons D, Sumner E, Darby S, Dorrell R, Hunt J, Autonomous Underwater Vehicles (AUVs): Their past, present and future contributions to the

advancement of marine geoscience. Marine Geology, 352, 2014 https://doi.org/10.1016/j.margeo.2014.03.012

Xiao L, Wang J, Qiu X, Rong Z, Zou X, Dynamic-SLAM: Semantic monocular visual localization and mapping based on deep learning in dynamic environment, Robotics and Autonomous Systems, Volume 117, 2019, Pages 1-16, DOI: 10.1016/j.robot.2019.03.012

Xin X, Jiang J, Zou W, A review of Visual-Based Localization. In Proceedings of the 2019 International Conference on Robotics, Intelligent Control and Artificial Intelligence, 2019 DOI: 10.1145/3366194.3366211

Xu L, Feng C, Kamat VR, Menassa CC. An Occupancy Grid Mapping enhanced visual SLAM for real-time locating applications in indoor GPS denied environments, Autom. Constr. 2019, 104, 230–245, DOI:10.1016/j.autcon.2019.04.011

Xuaand X,García de Sotoa B, On-site Autonomous Construction Robots:A review of Research Areas, Technologies, and Suggestions for Advancement, ISARC2020, DOI:10.22260/ISARC2020/0055

Yang T, Ren Q, Zhang F, Xie B, Ren H, Li J, Zhang Y, Hybrid Camera Array-Based UAV Auto-Landing on Moving UGV in GPS-Denied Environment. Remote Sensing, 2018 10, 1829, DOI: 10.3390/rs10111829

Yang Y, Xiao Y, Li T, A Survey of Autonomous Underwater Vehicle Formation: Performance, Formation Control, and Communication Capability, in IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 815-841, Secondquarter 2021, doi: 10.1109/COMST.2021.3059998

Ye J, Sung W, Understanding Geometry of Encoder-Decoder CNNs Proceedings of the 36th International Conference on Machine Learning, PMLR 97:7064-7073, 2019

Yi M, Trulls E, Lepetit V, Fua P, LIFT: Learned Invariant Feature Transform, Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9910. Springer, DOI: 10.1007/978-3-319-46466-4_28

Zahran S, Moussa A, Sheimy N El, Enhanced UAV navigation in GNSS denied environment using repeated dynamics pattern recognition, IEEE/ION PLANS, 2018, 1135–1142. DOI : 10.1109/PLANS.2018.8373497

Zhang C, He T, Zhan Q, Hu X, Visual Navigation Based on Stereo Camera for Water Conservancy UAVs, In Proceedings of the ICIST 2019, Hulunbuir, China, DOI: 10.1109/ICIST.2019.8836851

Zhang J, Xia Y, Shen G, A novel learning-based global path planning algorithm for planetary rovers, Neurocomputing, Volume 361, 2019, Pages 69-76, DOI: 10.1016/j.neucom.2019.05.075.

Zhao H, Shi J, Qi X, Wang X, Jia J, Pyramid Scene Parsing Network, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6230-6239, DOI: 10.1109/CVPR.2017.660

Zhou Q, Wang Y, Liu J, Jin X, Latecki J, An open-source project for real-time image semantic segmentation, 2019, Science China, Information Sciences

Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J, UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of DLMIA, 2018, Springer, DOI: 10.1007/978-3-030-00889-5_1

Zou Q, Sun Q, Chen L, Nie B, Li Q, A Comparative Analysis of LiDAR SLAM-Based Indoor Navigation for Autonomous Vehicles, in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 7, pp. 6907-6921, July 2022, DOI: 10.1109/TITS.2021.3063477