robotics

MDPI

Article

Keypoint Detection and Description through Deep Learning in Unstructured Environments

Georgios Petrakis and Panagiotis Partsinevelos *

Spatial Information Systems Unit, School of Mineral Resources Engineering, Technical University of Crete, 73100 Chania, Greece; gpetrakis2@tuc.gr

* Correspondence: ppartsinevelos@tuc.gr; Tel.: +30-2821037628

Abstract: Feature extraction plays a crucial role in computer vision and autonomous navigation, offering valuable information for real-time localization and scene understanding. However, although multiple studies investigate keypoint detection and description algorithms in urban and indoor environments, far fewer studies concentrate in unstructured environments. In this study, 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. The proposed architecture was trained and evaluated using a training and benchmark dataset with earthy and planetary scenes. 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 a mAP (mean Average Precision) in a level of 0.95 in terms of keypoint description, outperforming well-known handcrafted algorithms while the proposed SLAM achieved two times lower RMSE error in a poor-featured area with low illumination, compared with ORB-SLAM2. To the best of the authors' knowledge, this is the first study that investigates the potential of keypoint detection and description through deep learning in unstructured and planetary environments.

Keywords: feature extraction; unstructured environments; visual SLAM; deep learning; autonomous navigation

check for updates

Citation: Petrakis, G.; Partsinevelos, P. Keypoint Detection and Description through Deep Learning in Unstructured Environments. *Robotics* **2023**, *12*, 137. https://doi.org/10.3390/robotics12050137

Academic Editor: Yugang Liu

Received: 31 August 2023
Revised: 27 September 2023
Accepted: 28 September 2023
Published: 30 September 2023

1. Introduction

Keypoint detection and description play a critical role in computer vision and robotics, offering valuable low-level information for the real-time self-localization and scene understanding. The feature extraction process is the backbone of many advanced computer vision applications and tasks including autonomous navigation and 3D reconstruction utilizing mainly handcrafted algorithms or deep learning architectures. 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 neighbouring frames and scene recognition [1]. The feature extraction process is crucial for the visual SLAM (Simultaneously Localization and Mapping) systems that estimate a robot's location and contructs a map of the surroundings [2] while being able to detect the loops of the camera trajectory, which aid to further optimize the robot's navigation [3].

The handcrafted and well-known keypoint detectors including Harris [4], Shi-Tomasi [5], FAST [6], AKAZE [7] and keypoint descriptors such as ORB [8], SIFT [9] and SURF [10] have been widely investigated and used in multiple applications of computer vision and photogrammetry for more than three decades, while recently, several studies have utilized the deep learning-based architectures aiming to optimize more complicated tasks such as visual localization and autonomous navigation [11].

However, although the investigation of feature extraction based on handcrafted or learning-based algorithms is an active topic for urban, indoor or vegetated environments, there is no study that investigates the potential of deep learning in keypoint detection and description focused on unstructured environments or planetary scenes. In this study, a multi-task CNN (Convolutional Neural Network)-based architecture is proposed focused on keypoint detection and description in unstructured and planetary environments. The proposed architecture follows a teacher–student approach aiming to efficiently speciallize in unstructured and challenging scenes with feature-poor information and low or changing illumination. The architecture was trained using a proposed dataset including images from Earth, Mars and the Moon and evaluated with a benchmark dataset designed for testing the proposed architecture in terms of illumination and viewpoint. 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 authors' 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. Related Work

The literature abounds with studies that focus on autonomous driving [12,13], indoor navigation [14], 3D reconstruction [15] or inspection [16] in urban and industrial scenes. However, there are far fewer studies that 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 [17], 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 that is based on a SLAM algorithm in order to retrieve 3D spatial information, combined with a semantic segmentation technique that uses traditional image processing algorithms for contextual object identification (e.g., rock) and the SIFT and SURF algorithms for feature extraction on the segmented images. In [18], 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 suffer from low illumination and noise in subsea scenes. In [19], a methodology for node-based autonomous navigation in subquatic environments is proposed in which an underwater vehicle is able to explore an unknown area through several revisits without any human intervention, while in each revisit, the vehicle attempts to optimize its path by searching for more features in the scene. In [20], authors propose a methodology for path planning in agricultural fields focused on vineyards that is able to be performed in two stages: at first, a learning-based algorithm such as support vector machines (SVM) detects the agricultural field's patterns using satellite images extracting occupancy grid maps, while subsequently, image processing and topologiocal methods generate maps through which the robotic system is able to design the path on the field. In [21], 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 helps 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 [22,23], an evaluation of handcrafted feature extraction algorithms, in a Mars-like environment is presented. In [22], authors compare the algorithms' performance in terms of location recognition, using Devon Island dataset [24], concluding that SURF achieves the highest accuracy in non-vegetated and rocky unstructured environments. On the other hand, study in [23] 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 scores the highest overall efficiency. In [25], an improved version of SIFT for high-resolution remote sensing images from Mars and the 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 utilized 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 [26], authors attempt to solve this issue presenting a system called VOSAP (VO-aware sampling-based planner) which explores the rich-featured paths available in the scene, achieving increased performance in localization accuracy, tested in simulated Mars-like surfaces. In [27], a stereo SLAM system is presented which is focused on the loop-closure refinement, using elevation information of the terrain in poor-feature environments, utilizing a technique called Gaussian Process Gradient Maps. The authors use Moon and Mars-analogous terrains for the experimentation of the system while comparing it with state-of-the-art SLAM algorithms including ORB-SLAM2 [28] and VINS-MONO [29], proving higher efficiency, especially in a loop closure. In [30], a stereo SLAM system for highly detailed 3D point-cloud mapping is proposed, focused on planetary environments. The authors combine traditional front-end and back-end SLAM components in order to produce a sparse map using a self-supervized deep learning architecture that generates disparity maps aiming to dense the 3D scene information.

Although several studies explore the use of feature extraction and SLAM-based techniques in unstructured environments, there is a lack of deep learning methods focused on unstructured and planetary scenes. To the best of the authors' knowledge, this is the first study that investigates the potential of keypoint detection and description through deep learning in unstructured environments.

In the following section, the CNN-based architecture and SLAM system are presented, while in Section 4, the implementation and results of the study are described. In Section 5, the results of the CNN-based architecture and SLAM are analyzed, while finally in Section 6, the conclusions and future work of this study are presented.

3. Materials and Methods

Visual keypoint detection and description include several challenges when applied in unstructured environments such as rocky or sandy scenes or completely unknown environments including planetary scenes. The main challenges of feature extraction in unstructured environments can be described as follows:

- Lack of visual cues and important features;
- Low and changing lighting conditions.

Both challenges complicate the feature extractors to detect and describe keypoints, especially the handcrafted algorithms, which cannot further be improved for specialized and challenging environments, unless significant modifications that require extended investigations are conducted. On the other hand, the learning-based approaches are trained in general-purpose datasets that require high performance computing resources for training in a specialized dataset. Moreover, most deep learning-based feature extractors are focused either only on extracting keypoints with local descriptors, which can be used in processes such as camera trajectory estimation and local mapping, or on global descriptors, which are able to match an image (query) with the most similar image from a database of images. However, both approaches are valuable for a SLAM system since the keypoint detection with local description are utilized for the construction of a 3D environment through local mapping and camera trajectory estimation while the global descriptors are able to aid the system for a loop detection.

Thus, because the main goal of this study is to develop a specialized and integrated solution for keypoint detection and description in challenging and unstructured environments with potential use for autonomous navigation, a deep learning architecture that is able to efficiently be trained for specific environments extracting both keypoints with local descriptors and global descriptors was implemented, modified and improved.

The scope of this study is to investigate the potential of a lightweight deep learning architecture focused on unstructured environments and planetary scenes aiming to deal with the challenges described above.

3.1. Model Architecture

The proposed architecture is based on HF-net [31], a lightweight architecture focused on visual localization extracting keypoint locations, and local and global descriptors. The proposed architecture, which can be called HF-net2, utilizes a multi-teacher–student architecture (Figure 1) in order to increase efficiency in the training process maintaining high performance time without decreasing the accuracy and reliability. The SuperPoint architecture [32] is used as a teacher for keypoint detection and local description and the NetVLAD architecture [33] as a teacher for global description.

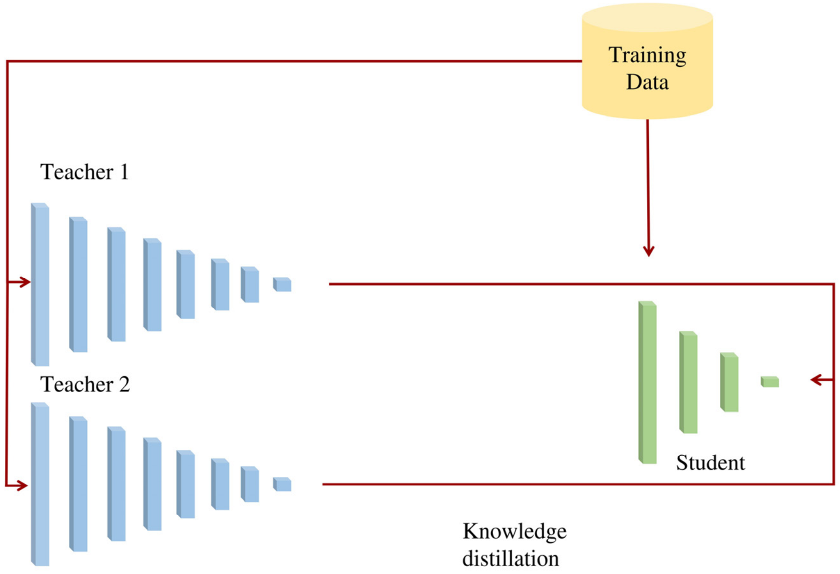


Figure 1. A multi-teacher-student architecture.

The student architecture is composed of a shared encoder and three different submodules that focus on (a) keypoint detection; (b) local description; (c) global description. For the shared encoder, the MobilenetV3-large [34] is utilized instead of MobilenetV2 [35], 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 (Figure 2).

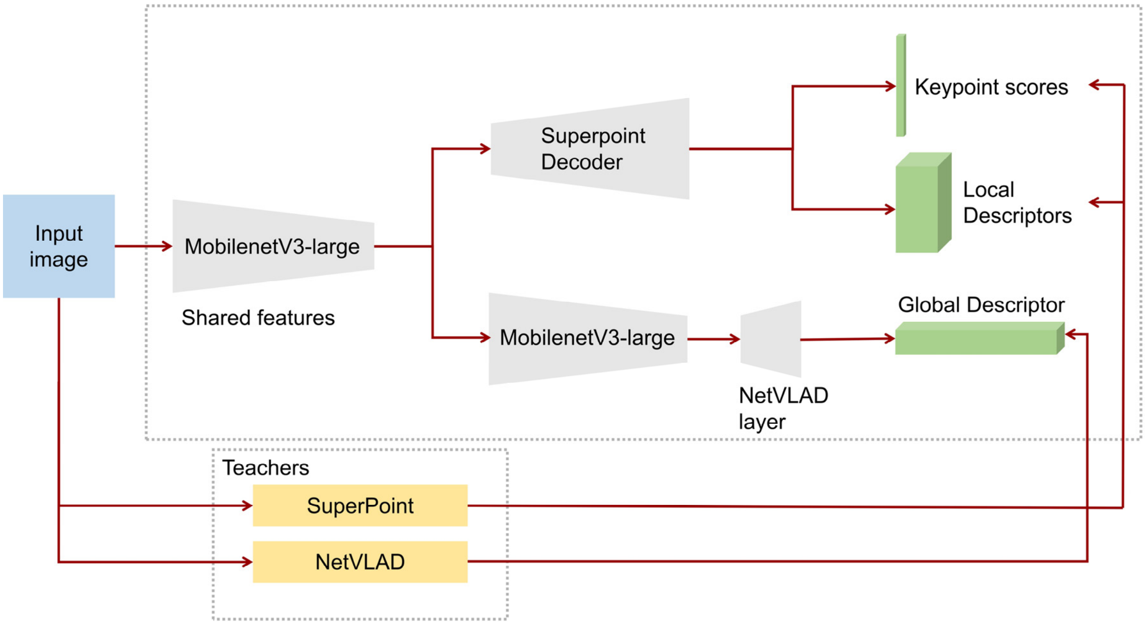


Figure 2. HF-net2 architecture.

More specifically, regarding the MobilenetV3-large, it receives and processes full-sized images as input, reducing its dimensionality aiming to learn low and high level features. It utilizes convolutional layers combined with inverted residual blocks with linear bottlenecks and depthwise separable convolutions in order to increase the efficiency while maintaining the accuracy [34]. MobilenetV3-large utilizes Neural Architecture Search (NAS) [36] and the non-linearity activation function called hard-Swish [37] which combines the Swish activation function (1) with the piece-wise alternative of the sigmoid function $\frac{\mathrm{ReLU6}(x+3)}{6}$ described by the Equation (2).

$$\mathrm{swish}[x] = x\,\sigma(x), \tag{1}$$

$$h - \mathrm{swish}[x] = x\frac{\mathrm{ReLU6}(x+3)}{6}, \tag{2}$$

where σ(x) is the sigmoid function and ReLU6 (Rectified Linear Unit) is a modification of the well-known ReLU activation function. Thus, MobilenetV3-large achieves increased efficiency in terms of performance time and accuracy compared with MobilenetV2.

Concerning the decoder parts of the architecture, the decoder of keypoint detection, produces the probability of keypoint existence in each pixel utilizing the "softmax" function while it reconstructs the image using an upsampling technique called "sub-pixel convolution" [38] instead of the transposed convolution technique, in order to reduce the computational cost. The decoder of local description includes two convolutional layers while by using bilinear interpolation and L2 normalization reconstructs the image dimensionality and exports the local descriptors with a fixed length. The NetVLAD decoder of global description utilizes the soft-assignment [33], which is a differentiable technique aiming to assign the weights of each local descriptor to a cluster (a group of similar local descriptors), while subsequently, the decoder aggregates the 1st-order statistics of the difference between local descriptors and the corresponding clusters based on the weights of the soft-assignment, resulting in a global descriptor.

Regarding the learning process, the training dataset initially feeds two pre-trained models that represent the teachers for keypoint detection, and local and global description (SuperPoint and NetVLAD), while the distilled knowledge plays the role of ground truth for the student. During the training process, the architecture tracks the output of the predicted keypoint scores by the student with the keypoints predicted by the SuperPoint (teacher) using the "Cross-entropy" cost function while it calculates the difference between the local and global descriptions exported by the student with the corresponding output of the SuperPoint and NetVLAD teachers, respectively, using the L2 norm operation. The unified loss function of the multi-teacher-student architecture is presented below [31]:

$$L = e^{-w_1}\|d_s^g - d_{t_1}^g\|_2^2 + e^{-w_2}\|d_s^l - d_{t_2}^l\|_2^2 + 2e^{-w_3}\mathrm{CrossEntropy}(p_s, p_{t_3}) + \sum_i w_i \quad (3)$$

where $\|d_s^g - d_{t_1}^g\|_2$ is the L2 norm of student (s) and NetVLAD (t_1) global descriptors while $\|d_s^l - d_{t_2}^l\|_2$ 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.

As a result, the proposed architecture is able to combine multi-task prediction, utilizing knowledge distillation, achieving an efficient and flexible end-to-end training process. The proposed architecture was trained and fine-tuned using a dataset that includes FPV (First Person View) images from Earth, Mars, and the Moon, aiming to increase its robustness in unstructured and challenging environments. More information about the dataset is referred to in Section 3.3.

3.2. SLAM Architecture

The proposed model 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 lack in visual cues and intense lighting conditions, using the trained and fine-tuned model for keypoint detection, and local and global description.

The proposed SLAM system is based on ORB-SLAM2, but instead of using ORB algorithm 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 being divided in three different modules which are performed simultaneously in separated threads: (a) tracking, (b) local mapping and (c) loop closing.

The tracking module is able to localize the camera position in each frame by searching for multiple matches on features detected in the local area while it utilizes motion-only bundle adjustment (BA) technique in order to improve the camera pose (rotation and translation) by minimizing the reprojection error between the matched points of 3D space (in world coordinates) and the detected keypoints. On the other hand, the local mapping module, is capable of processing and further improving the initial matched features of a local area using local BA which optimizes sets of multiple neighboured frames and all the points visible from those specific neighboured frames. The loop-closing module is based on DXSLAM [39] and combines the Fast BoW (Bag-of-Words) algorithm [40], which converts each keyframe in a vector of words using the pre-trained vocabulary tree, while each image representation is extracted by the global descriptors predicted by the proposed HF-net2 model. Furthermore, the system performs full BA in order to further optimize the camera trajectory and point locations.

Regarding the SLAM system pipeline, initially the RGB-Depth data are imported to the trained HF-net2 model which predicts keypoint locations, and local and global descriptors. The keypoints and local descriptors are initially used for the camera pose prediction while, when the camera pose is estimated with success, the SLAM system proceeds by detecting multiple features in neighboured frames, extracting keyframes using the estimated camera pose predictions. The keyframes that are treated as landmarks, combined with the keypoints, aid the local mapping module to map the surroundings. If there is a failure in camera pose prediction, a re-localization process re-estimates the camera location using the fast BoW and the global descriptors. If there is a loop during mapping, it is detected by the loop-closing module, while afterwards, the estimated trajectory and mapping are further optimized using full BA (Figure 3).

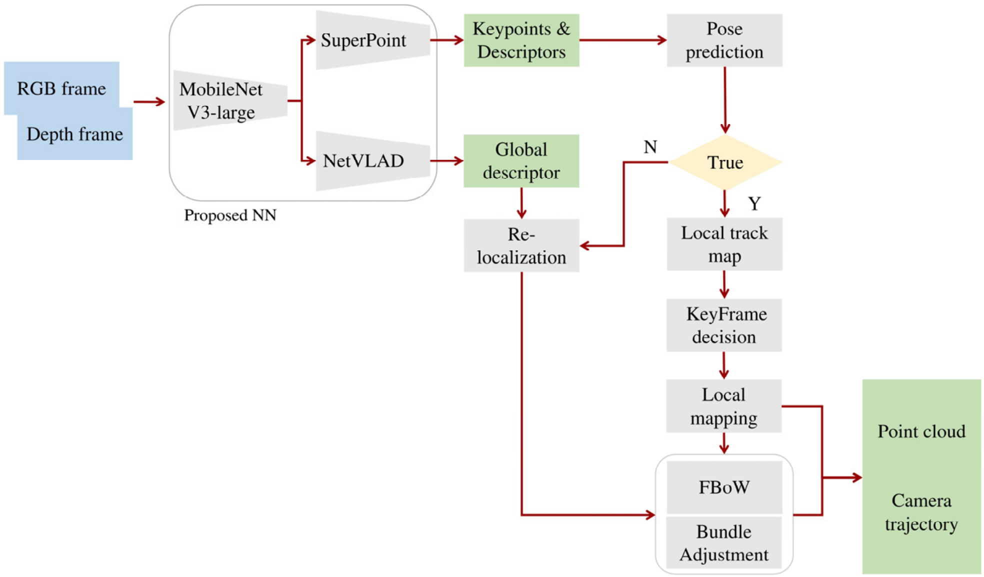


Figure 3. SLAM architecture based on the proposed NN.

3.3. 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 [41] while, to the best of the authors' knowledge, there is not any deep learning model for feature extraction, focused on unstructured and planetary scenes.

Thus, two different datasets were designed aiming to train and evaluate the proposed deep learning methodologies: (a) a training dataset and (b) an evaluation dataset.

3.3.1. Training Dataset

The training dataset includes 48,000 FPV or rover-based images with wide range of variations in landscapes, including images from Earth, the 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 (Figure 4a). The images were taken in scenes with various lighting and weather conditions in the day and night time, while the camera was located 1.5 m from the ground in a direction of 10 and 60 degrees from the horizon (Figure 5).

The images from Mars were collected by a publicly available dataset of NASA that includes about 13,000 images captured by Mars Science Laboratory (MSL, Curosity) rover using three instruments: Mastcam Right eye, Mastcam Left eye, and MAHLI [42] (Figure 4b). Concerning the Moon's dataset, it includes about 9000 artificial rover-based images which generated and released with CC (Creative Commons) license by Keio University in Japan (Figure 4c). The dataset was created using the Moon LRO LOLA digital elevation model which is based on the Lunar Orbiter Laser Altimeter [43] combined with the simulation software Terragen of Planetside Software (Version 4.6).

3.3.2. Evaluation Dataset

For the evaluation of the proposed architecture and the comparison with other widely used handcrafted algorithms, a benchmark dataset was designed for unstructured and planetary scenes inspired by HPatches dataset [44], one of the most popular evaluation datasets for keypoint detection and description in general-purpose images.

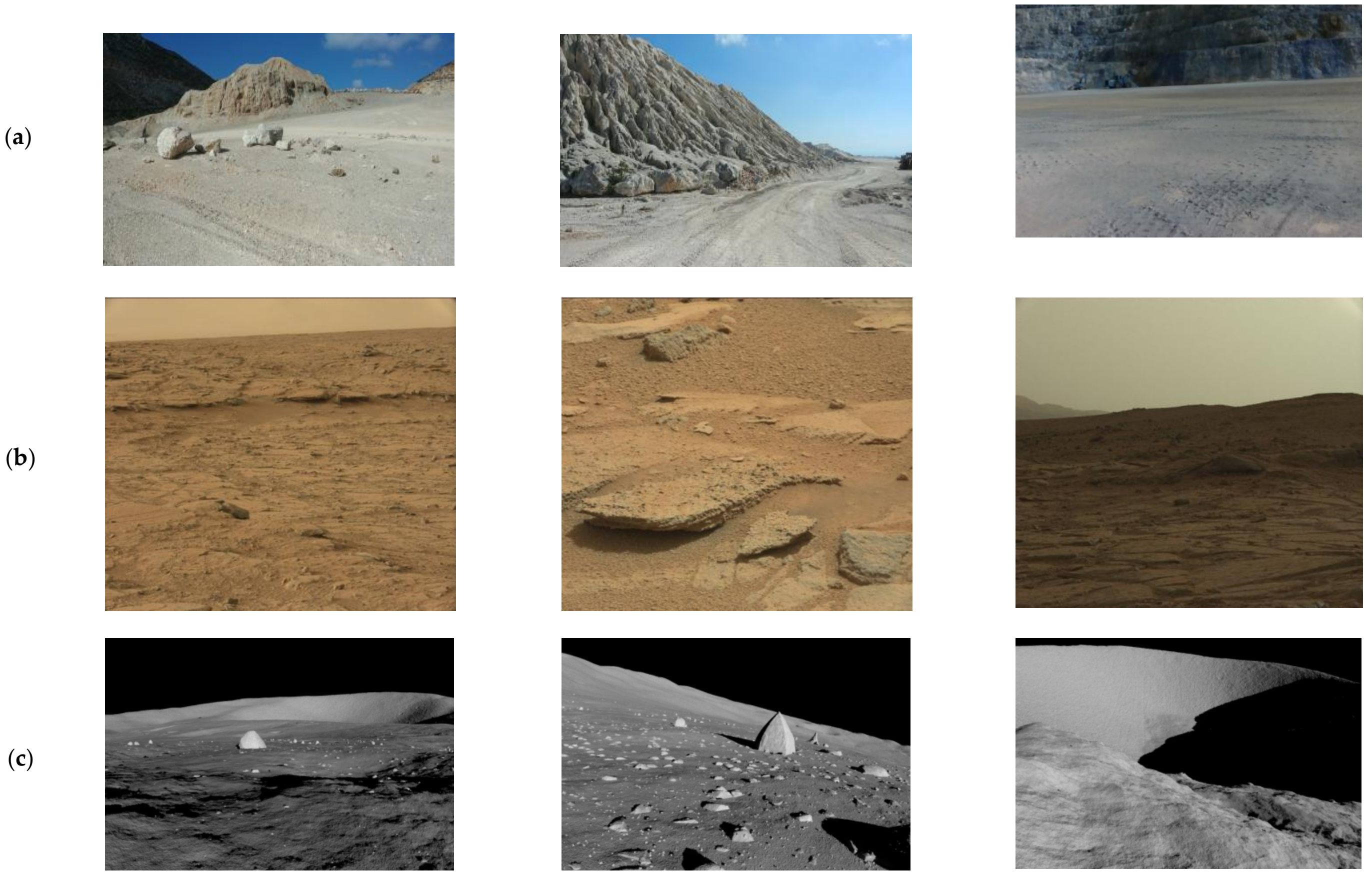


Figure 4. A sample of training dataset. (**a**) images from Earth, (**b**) images from Mars (**c**) images from artificial lunar surface.

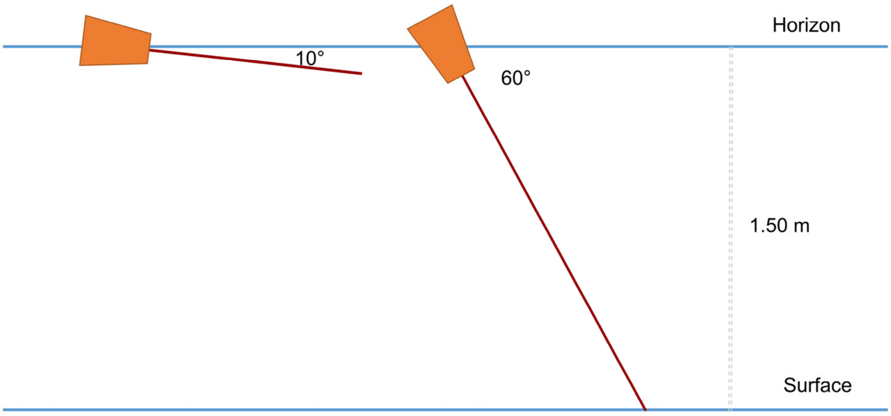


Figure 5. Camera's direction relative to the horizon.

The proposed dataset contains 120 sequences of images from Earth, Mars and the Moon, which are 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 (Figures 6 and 7). 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 contain identity matrices, since the only difference among the representations is the illumination.

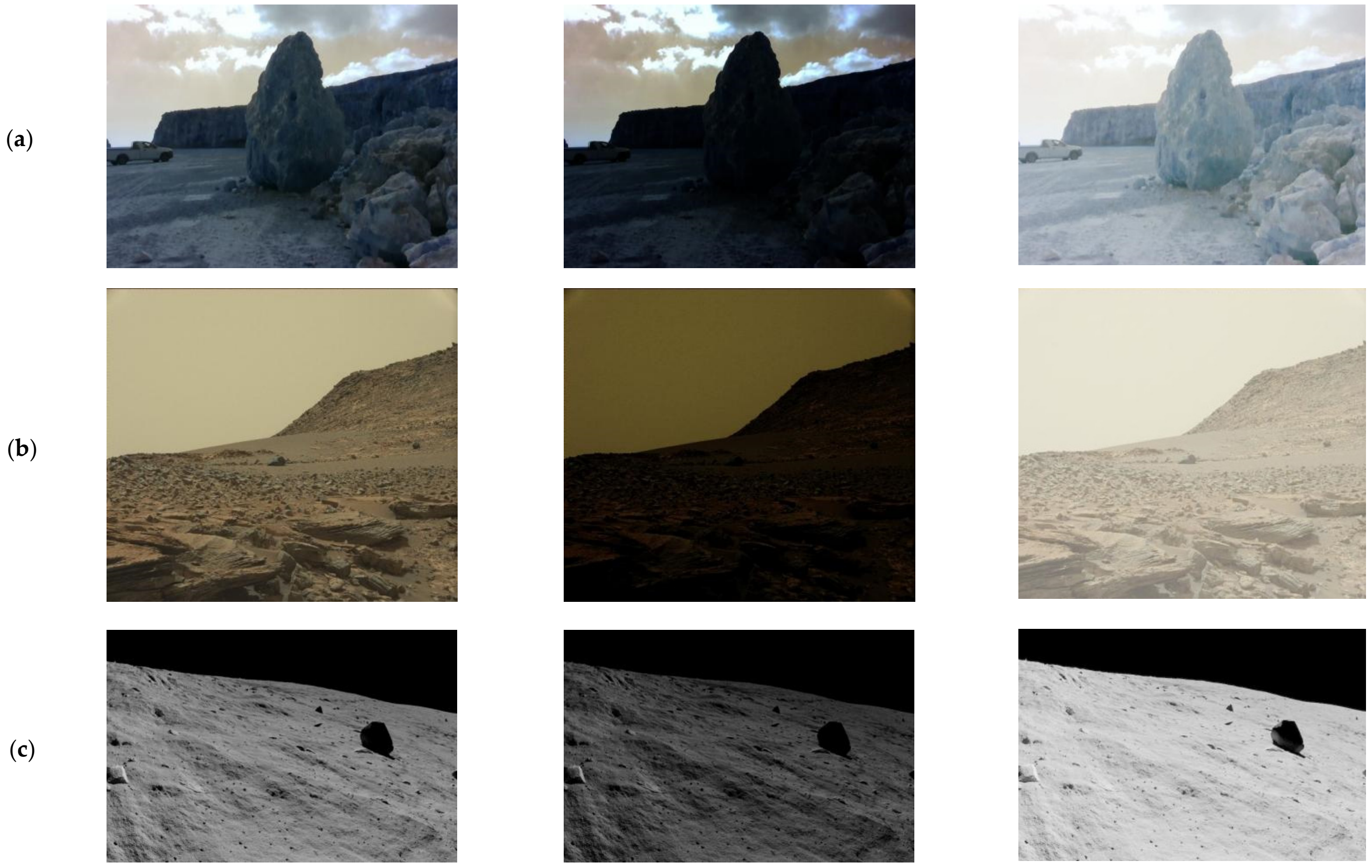


Figure 6. A sample of illumination-part evaluation dataset. (**a**) sequence from Earth, (**b**) sequence from Mars (**c**) sequence from artificial lunar surface.

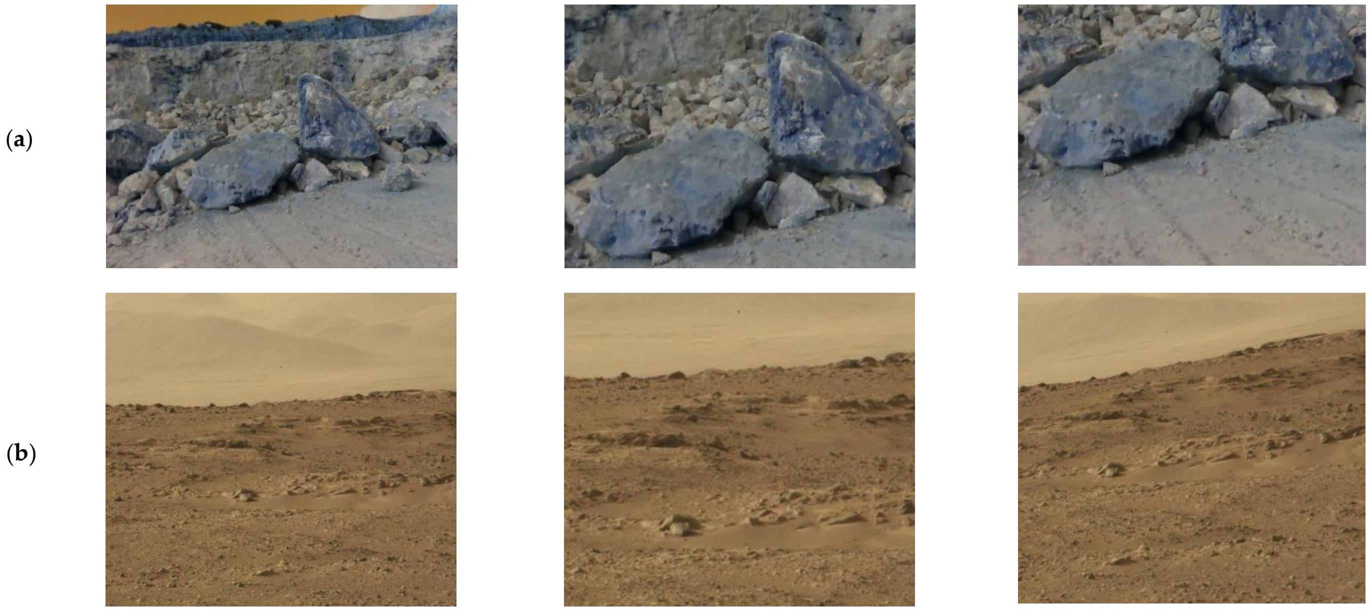


Figure 7. *Cont*.

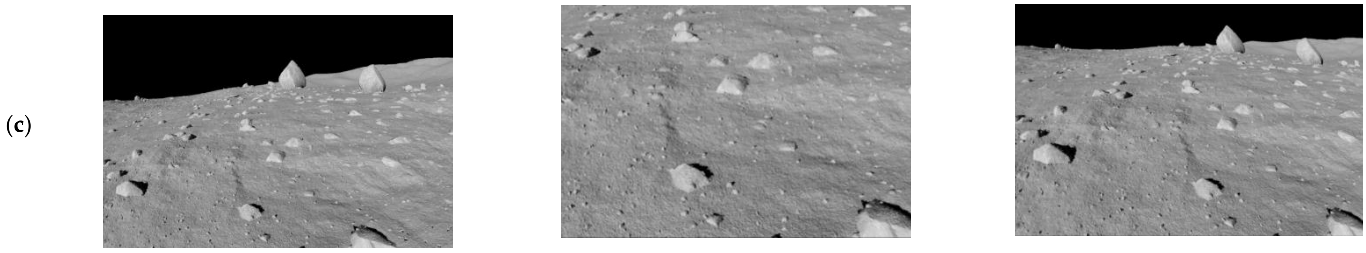


Figure 7. A sample of viewpoint-part evaluation dataset. (**a**) sequence from Earth, (**b**) sequence from Mars (**c**) sequence from artificial lunar surface.

4. Experimentation and Results

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

4.1. Training Process

As referred to in Section 3, HF-net2 uses a multi-task distillation approach for the training process using SuperPoint and NetVLAD as the teachers for keypoint detection, and local and global descriptions, 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 training dataset;
- HF-net: the original HF-net was trained from scratch using the proposed training dataset.

Both models were trained for 30,000 iterations while the RMSProp optimizer was utilized with a learning rate in a range of 0.001–0.00001. All the experimentation was conducted using an Intel i7-4771 CPU (Intel Corporation, Santa Clara, CA, USA) with 3.50 GHz × 8 combined with an NVIDIA GeForce GTX 1080 Ti GPU (Nvidia, Santa Clara, CA, USA) while the architecture was implemented using the TensorFlow [45] deep learning platform.

4.2. Evaluation and Results of the Proposed Architecture

To evaluate the HF-net2 model, the proposed benchmark dataset was utilized. As referred to in Section 3.3.2, the dataset includes 120 sequences of images from Earth, the Moon and Mars, designed for evaluation in terms of illumination and viewpoint changes.

Moreover, the efficiency of HF-net2 was tested and compared with several well-known handcrafted algorithms for keypoint detection and description including Harris [4], FAST [6], SIFT [9] and ORB [8]. Harris 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, and afterwards computes the autocorrelation matrix for each pixel by convolving the image gradients with a Gaussian window. FAST is able to identify keypoints by comparing the pixel intensities in a circular neighborhood in order to classify or not the central pixel as a keypoint. On the other hand, ORB initially builds a pyramid, which is a multi-scale representation of a single image and identifies keypoints using the FAST algorithm, and afterwards, it computes a binary feature vector for each keypoint using the BRIEF [46] descriptor, encoding the relative intensities of pixel pairs. SIFT constructs a scale-space representation of the input image by applying Gaussian blurring at multiple scales, and 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.

Regarding the metrics of keypoint detection, the repeatability and mAP (mean Average Precision) were utilized in aforementioned sequences which includes image representations with different illumination (i) or viewpoint (v) (Table 1). The repeatability measures the percentage of keypoints that are repeatable in different image representations of the same scene while mAP utilizes the precision $\frac{\#\text{ correct matches}}{\#\text{ matches}}$ and recall $\frac{\#\text{ corrent matches}}{\#\text{ correspondences}}$ 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 2).

Table 1. Evaluation of HF-net2 as a keypoint detector in terms of intense illumination (i) and viewpoint (v) changes using repeatability metric.

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

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

As presented in Table 1, 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 a repeatability of 0.74 and a mAP 0.71. The original HF-net provides lower accuracy in terms of repeatability and mAP 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 lower accuracy than Harris algorithm and learning-based architectures.

Regarding the evaluation of descriptors (Table 2), 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 at a level of 0.75 and 0.65 and a 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 is worth noting that the superiority of SuperPoint proves that it possesses robust keypoint detection and description architecture, and is capable of being a teacher of HF-net during the training process.

Concerning the performance-time of the proposed model, it achieves about 16 ms (milliseconds) per frame inference-time and 62.5 FPS (Frames per Second) in a GPU-enabled machine. It is worth noting that SuperPoint and the handcrafted descriptors 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 inference-time is quite satisfactory and adequate for real-time applications.

4.3. Evaluation of the Proposed SLAM System

As referred to in Section 3, the proposed HF-net2 was integrated in a SLAM system as a 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 a lack of visual cues in unstructured and planetary scenes. The proposed SLAM is based on ORB-SLAM2 and is built on C++ and Python programming languages.

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 Intel RealSense D435 camera (Intel Corporation, Santa Clara, CA, USA) was used, which includes two infrared sensors (left and right), an RGB and a depth sensor, while in this study, only the RGB and depth sensor were utilized for accurate scale estimation. The camera resolution of the RGB and depth sensor was 1920 × 1080 and 1280 × 720, respectively, the focal length was 1.93 mm and the format was 10-bit RAW. Moreover, the CHCNAV i73 RTK (Real Time Kinematics) (Shanghai Huace Navigation Technology Ltd., Shanghai, China) GNSS receiver was used in order to measure the coordinates of the camera trajectory in a geodetic reference system.

Two sources of data were captured in each experiment:

- A rosbag file, a ROS-based (Robot Operating System) file, which includes XYZ coordinates, orientation and an optical center with respect to the world origin of the SLAM coordinate system, captured in a video of 30 FPS with a resolution 848 × 480.
- GNSS data which contain 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 were extracted from the rosbag file and synchronized in order to correlate each RGB frame with the corresponding depth frame while afterwards the proposed SLAM estimates the camera trajectory in TUM (Technical University of Munich) format [47] using the RGB-depth information. Concerning the GNSS data, they 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 represent a specific location in the 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 Figure 8.

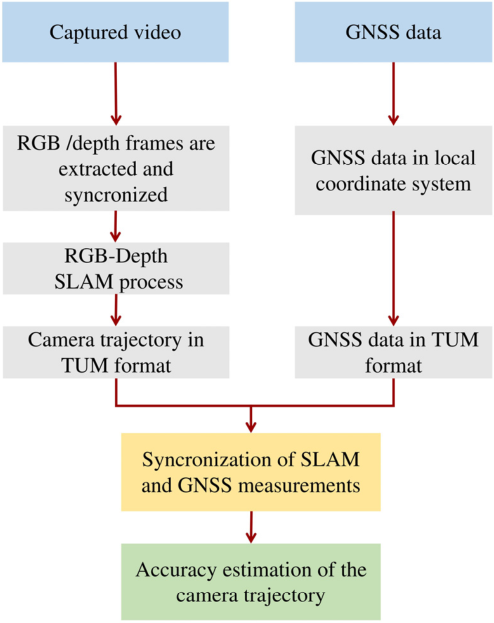

Captured video
GNSS data
RGB /depth frames are extracted and syncronized
GNSS data in local coordinate system
RGB-Depth SLAM process
Camera trajectory in TUM format
GNSS data in TUM format
Syncronization of SLAM and GNSS measurements
Accuracy estimation of the camera trajectory

Figure 8. Pipeline of the SLAM evaluation process.

4.4. Experiments and Results of the Proposed SLAM

The experiments were conducted in two different environments: a rocky and a sandy scene (Figure 9), while different cases in terms of illumination were performed. The proposed SLAM system was compared with the ORB-SLAM2, 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.

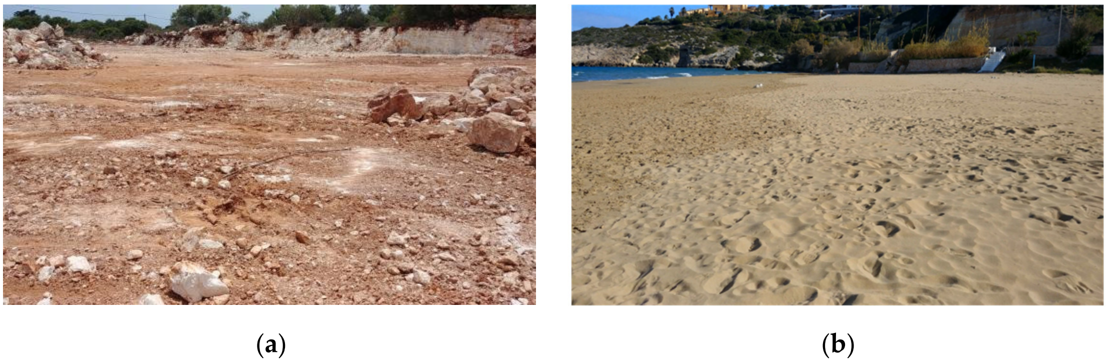


Figure 9. (**a**): Rocky scene, (**b**): sandy scene.

More specifically, the experiments were performed in day and evening time using different trajectory paths and natural light, while one experiment was conducted with artificially low illumination with extremely fast lighting changes (Table 3).

Table 3. Experiments, performed in different scenes, trajectory paths and illumination conditions.

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 to Low	Artificial

The results of the experiments above are presented in the Tables 4–8. 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 qualitative results, with the predicted and GNSS-based trajectories, are presented in Figure 10.

Table 4. Square-based path in rocky terrain with high illumination.

	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 5. Square-based path in rocky terrain with medium to low 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 6. Right angle-based path in rocky terrain with medium 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 7. Random path in sandy terrain with high 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 8. 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.

	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

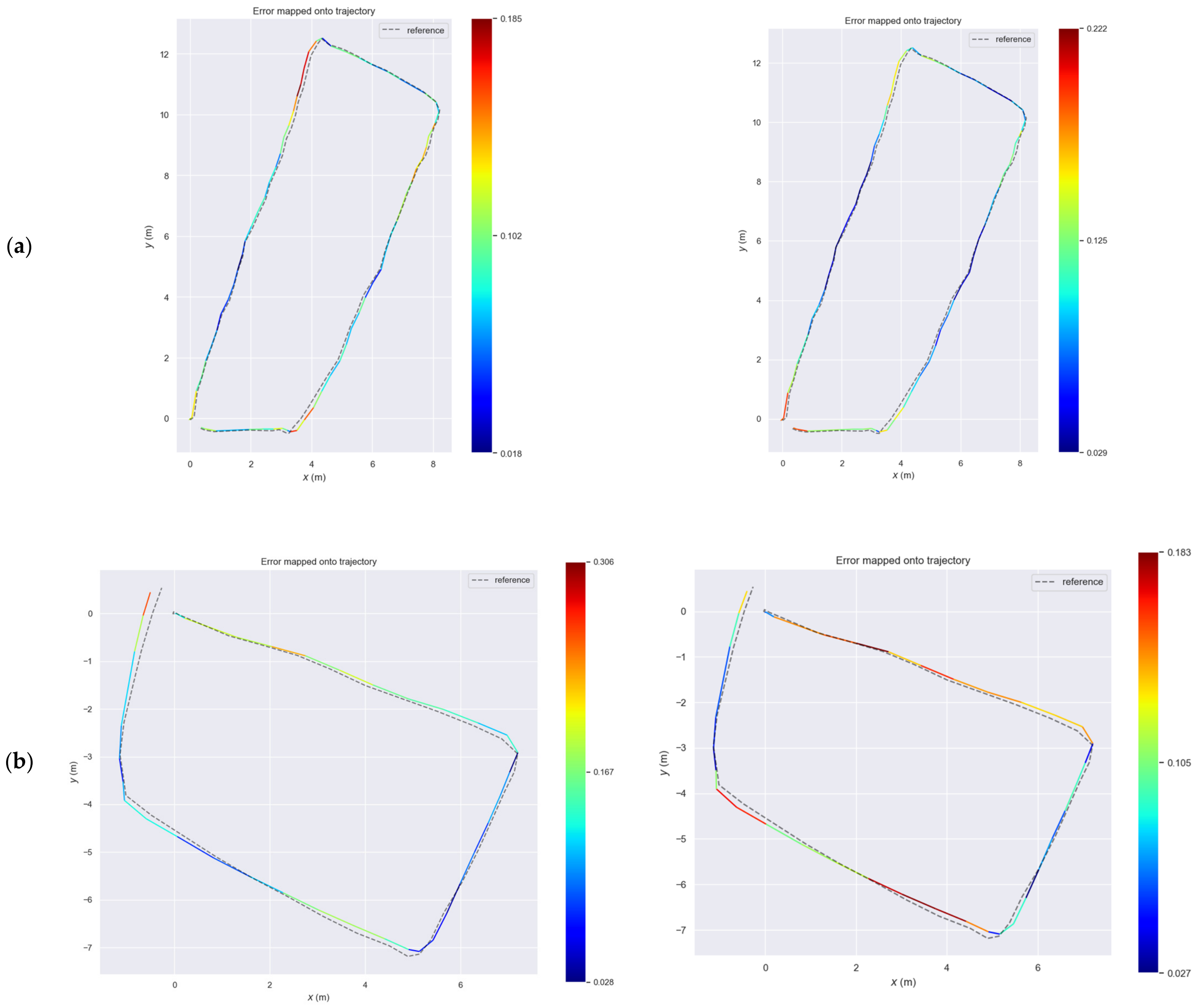


Figure 10. *Cont.*

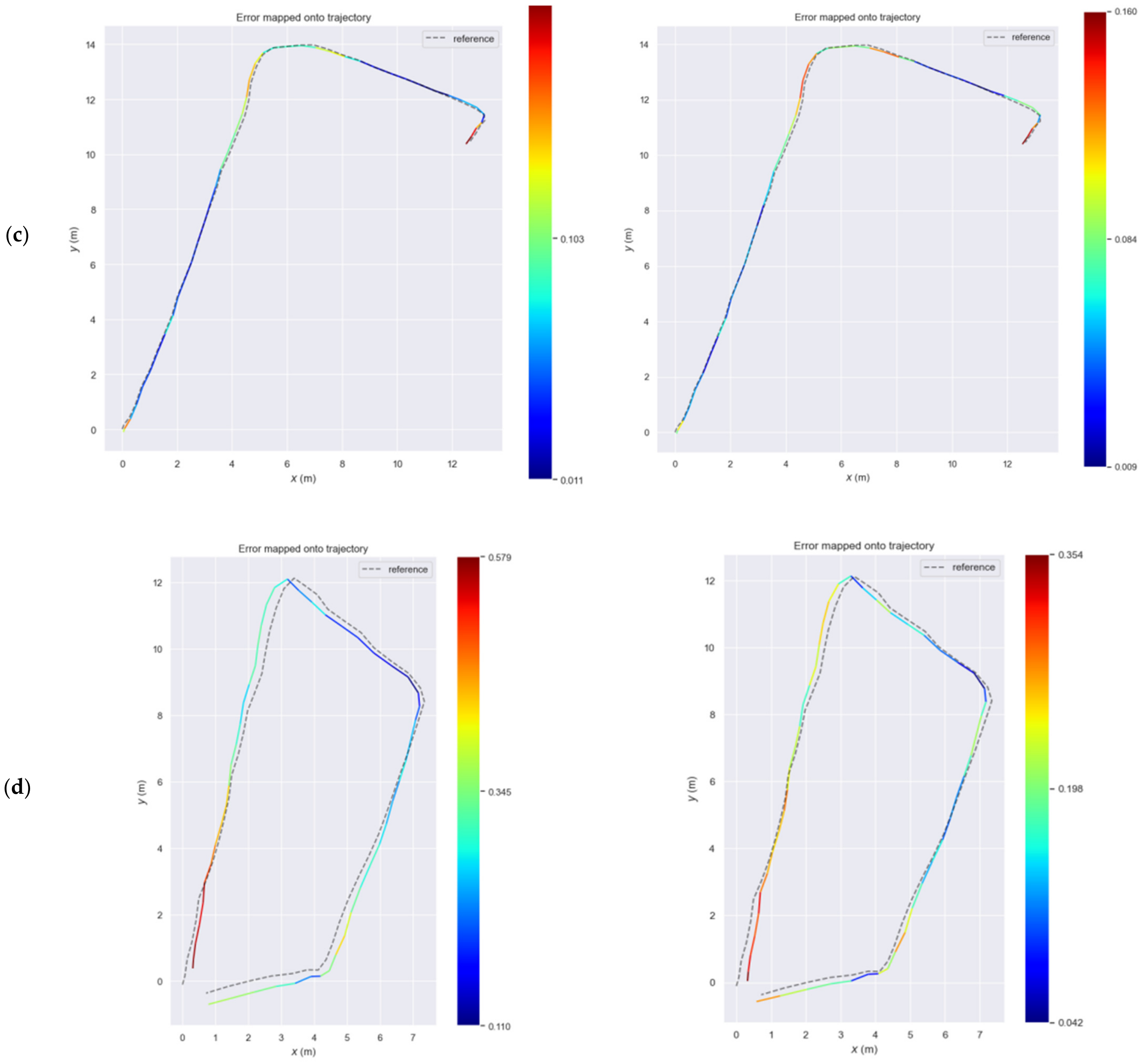


Figure 10. *Cont.*

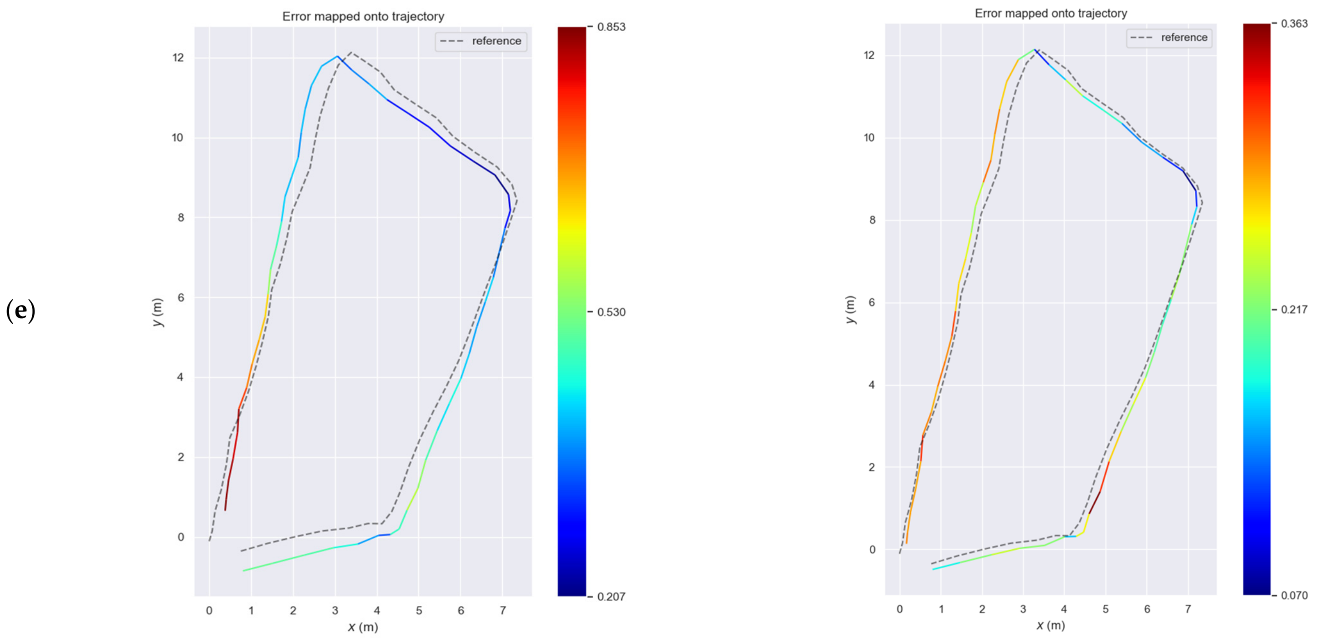


Figure 10. Predicted trajectories of the ORB-SLAM2 (left column) and proposed SLAM (right column) compared with the ground truth trajectory (presented as gray dashed line). (**a**) rocky terrain with high illumination, (**b**) rocky terrain with medium to low illumination, (**c**) rocky terrain with 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 and 6 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 (Figure 10a,c). However, in the square-based path with low illumination (Table 5), 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 the ORB-SLAM2, which produced an RMSE error at a level of 0.15 and a maximum error equal to 0.32 m (Figure 10b).

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 a standard deviation of 0.12, respectively. In the last experiment, the same data frames were processed using GAMMA correction (4) aiming to highly decrease the illumination in a specific range.

$$\text{Output} = \left(\frac{\text{I}}{255}\right)^{\frac{1}{\gamma}} 255 \tag{4}$$

where I is the input pixel value and γ is the gamma parameter which controls the image brightness. The gamma values below 1 (gamma < 1) produce darker images while the 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 and 0.4 (Figure 11).

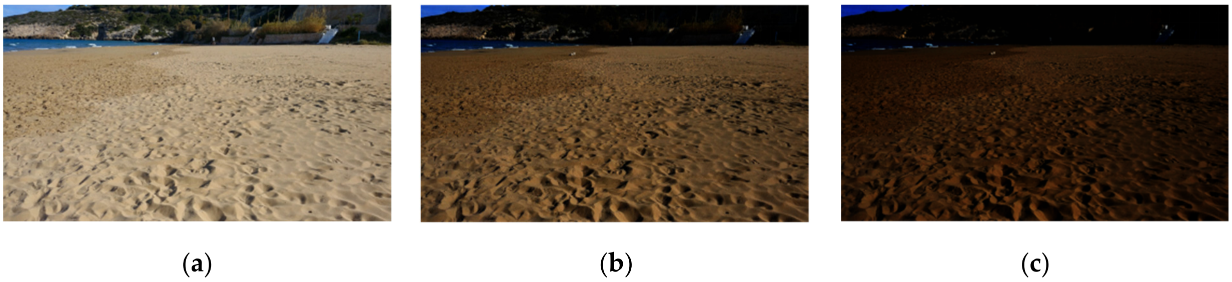


Figure 11. (**a**): Original image, (**b**): darken image with gamma = 0.4, (**c**): darken image with gamma = 0.2.

As a result, the SLAM systems encounter a scene which lack 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 is worth noting that the maximum and minimum errors of the proposed SLAM are 0.36 and 0.07, respectively, while the corresponding errors of ORB-SLAM2 are 0.85 and 0.21.

5. Discussion

In this study, 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 compose the multi-task decoder of the architecture. The HF-net2 was trained using a dataset composed of 48,000 captured and selected images from Earth, Mars and the 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 unstructured 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 the 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 (Figures 12 and 13).

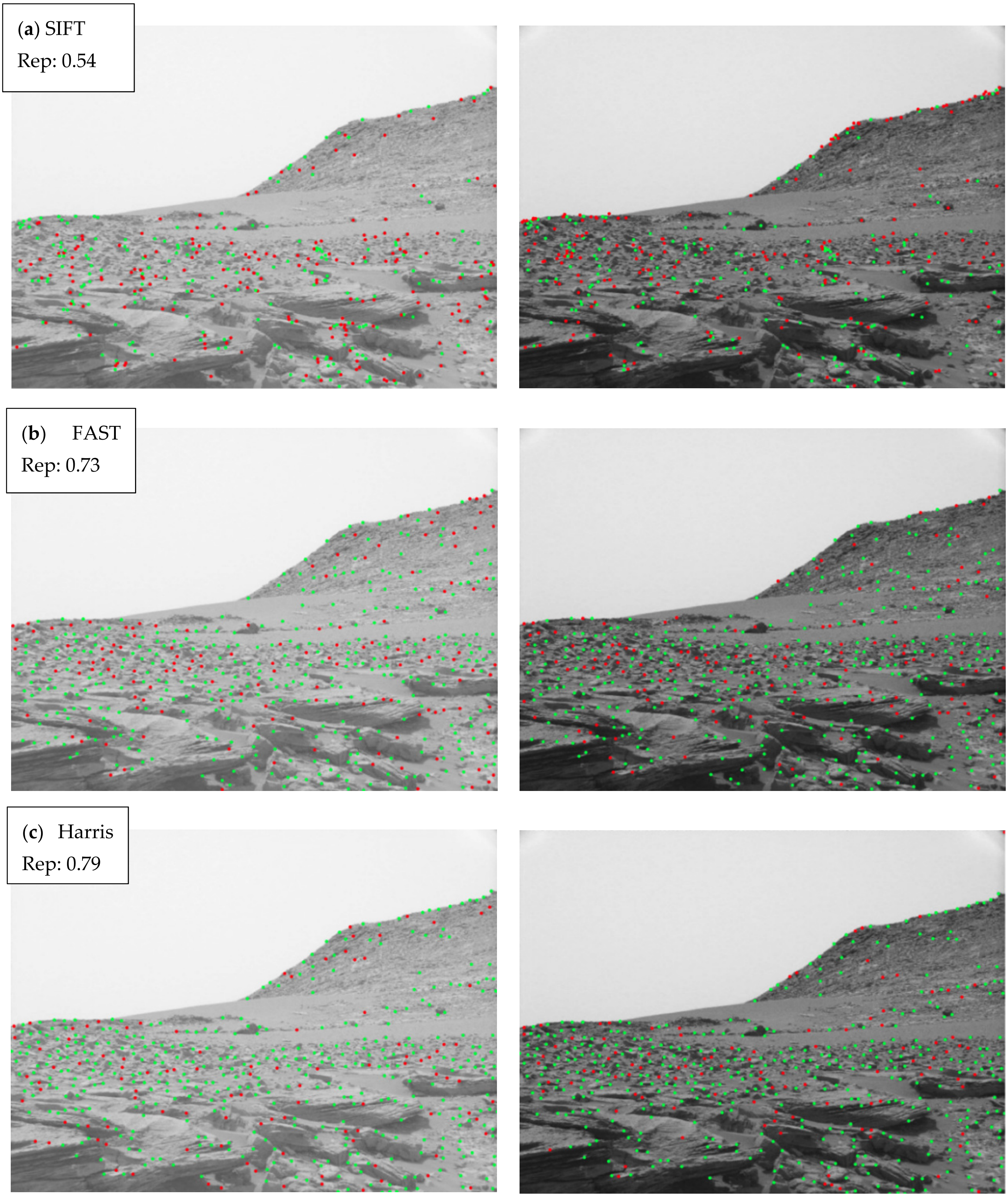


Figure 12. *Cont.*

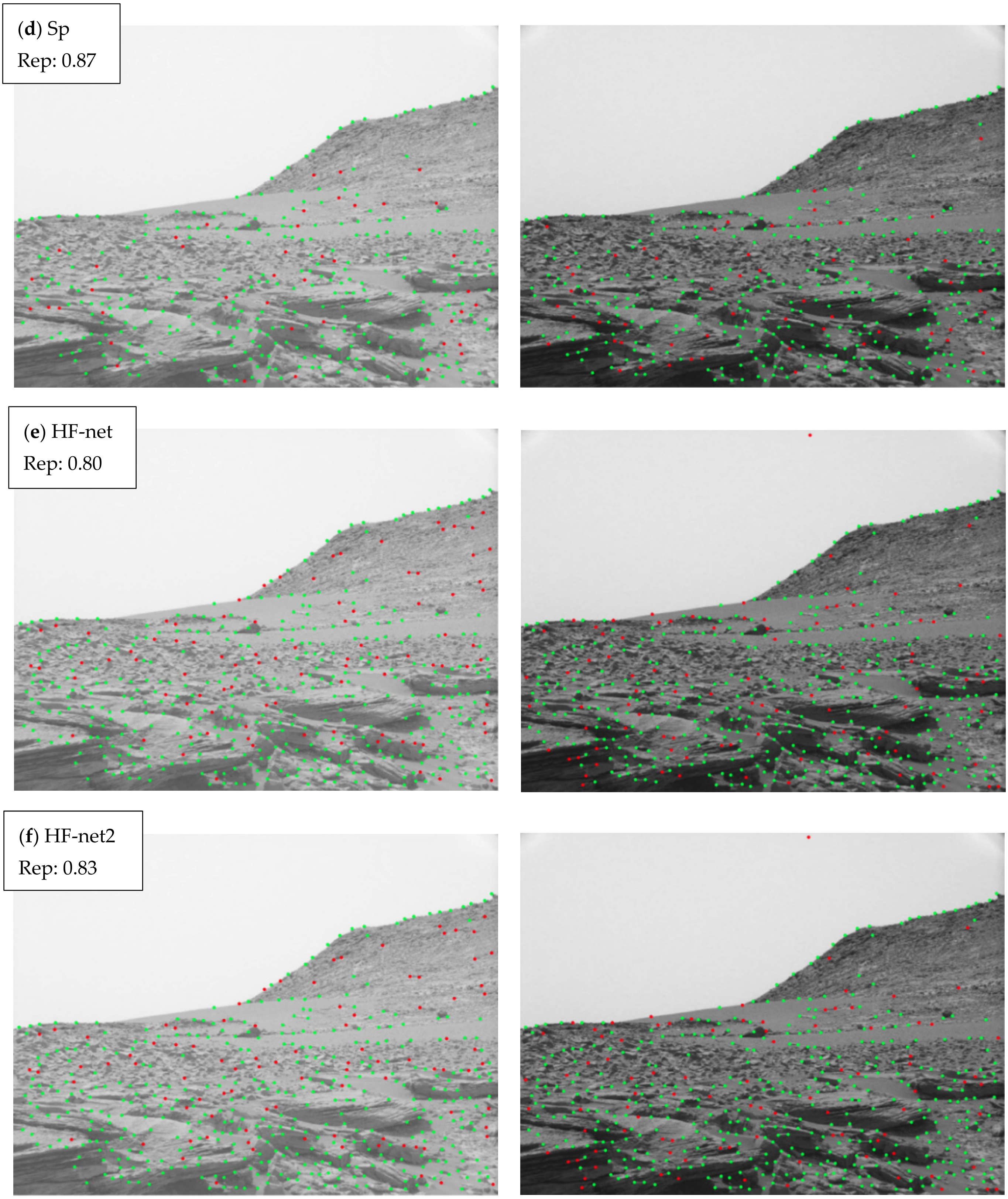


Figure 12. *Cont.*

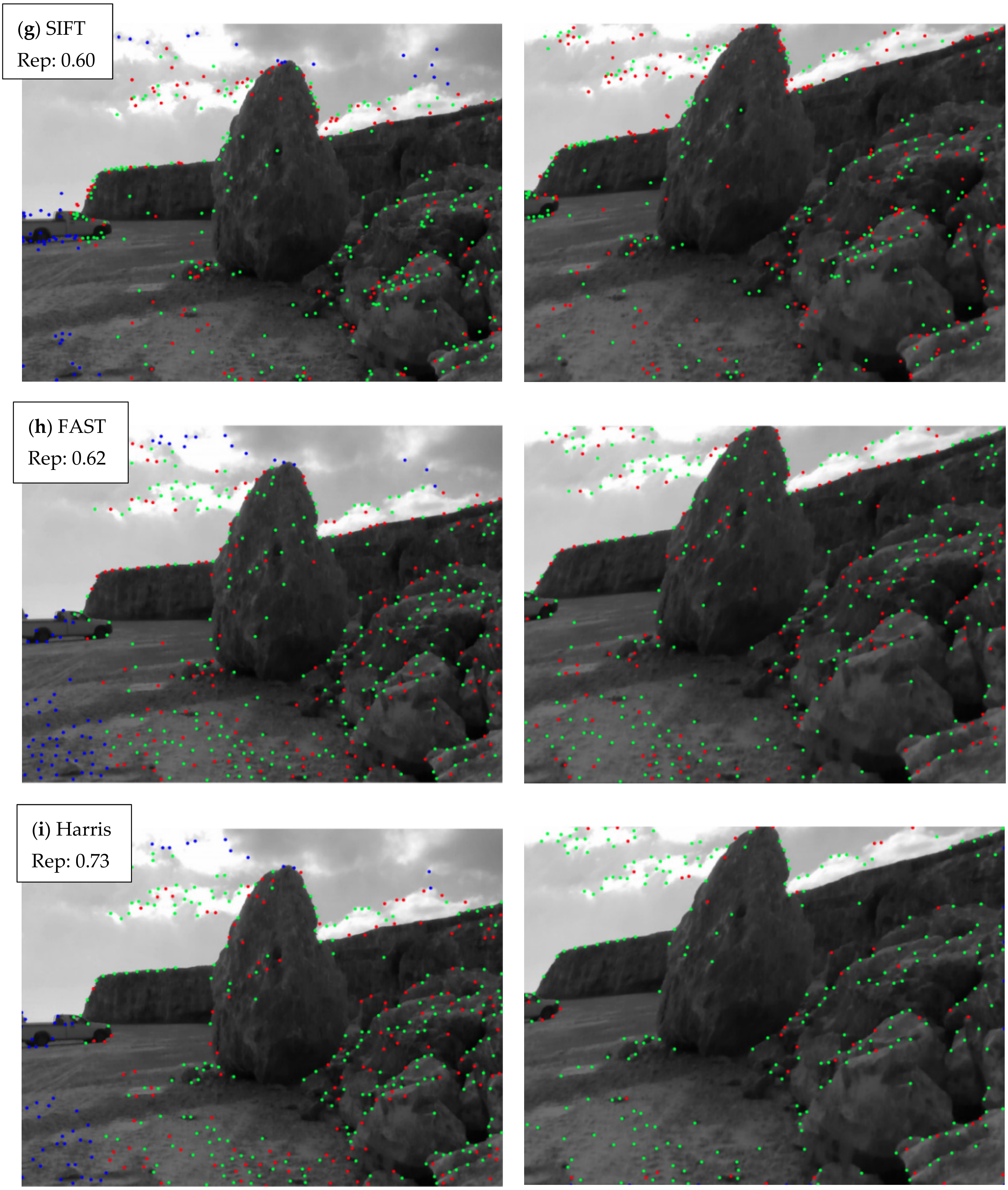
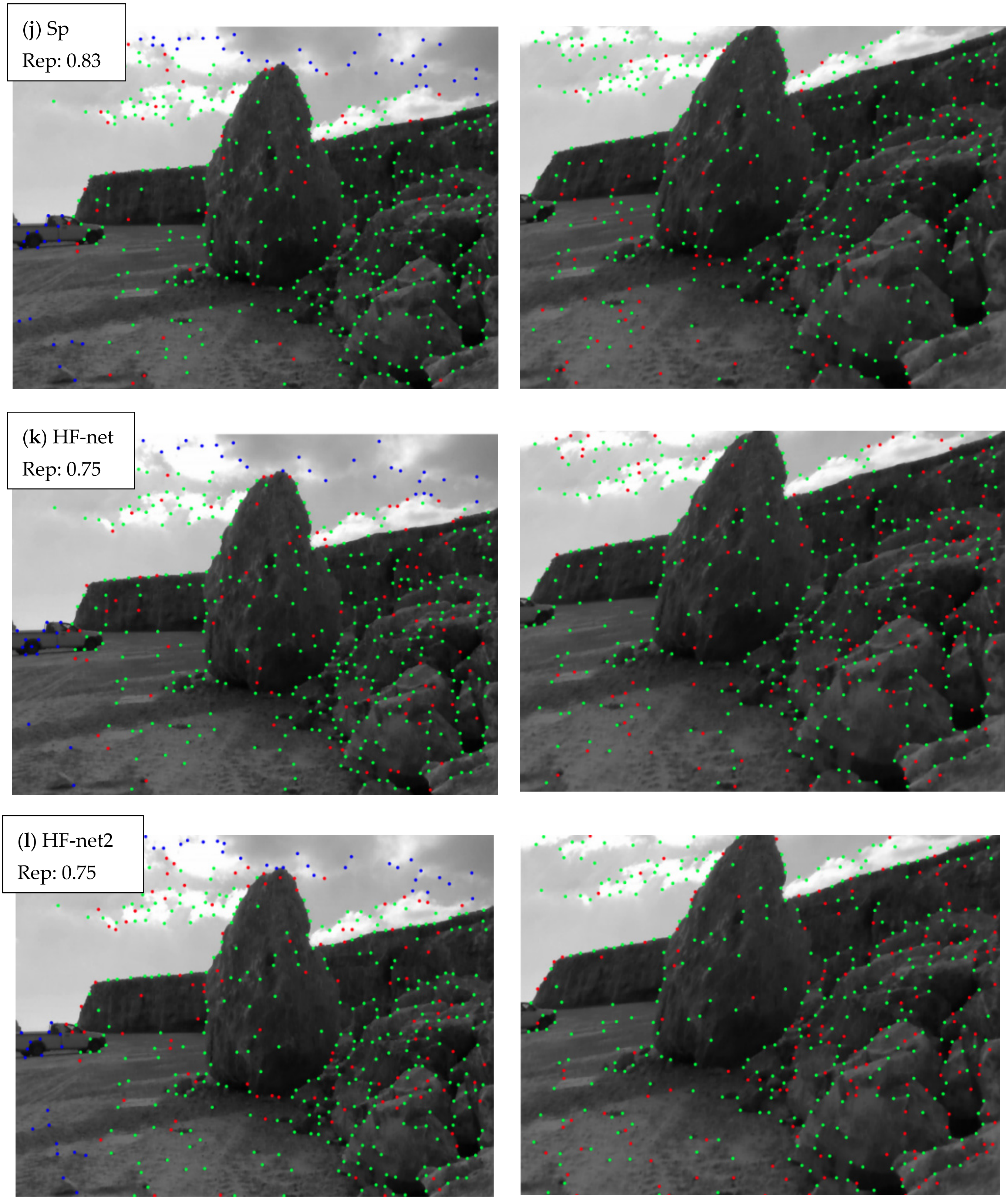


Figure 12. *Cont.*



(**j**) Sp
Rep: 0.83

(**k**) HF-net
Rep: 0.75

(**l**) HF-net2
Rep: 0.75

Figure 12. 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**) earthly 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 viewpoints.

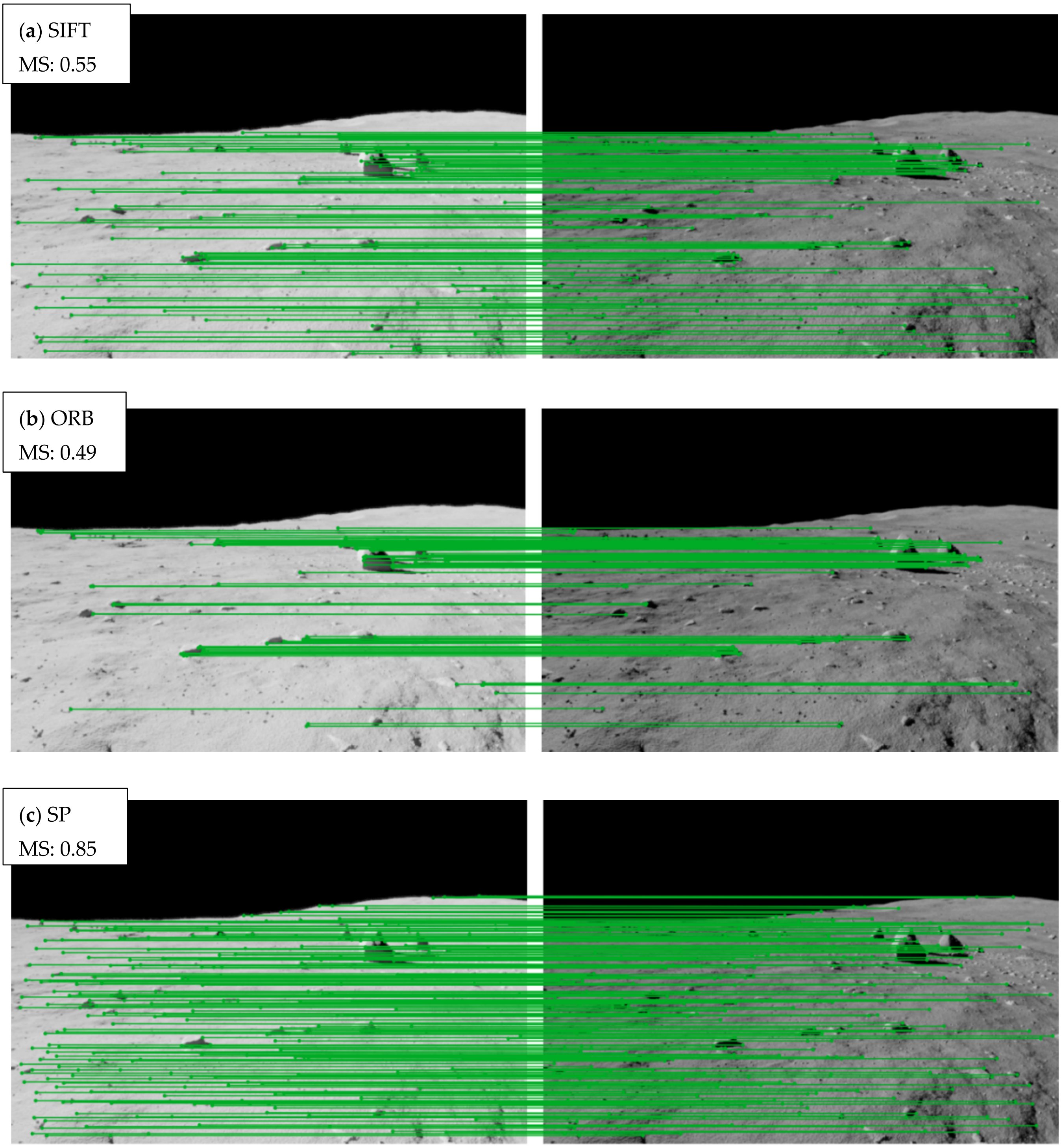


Figure 13. *Cont.*

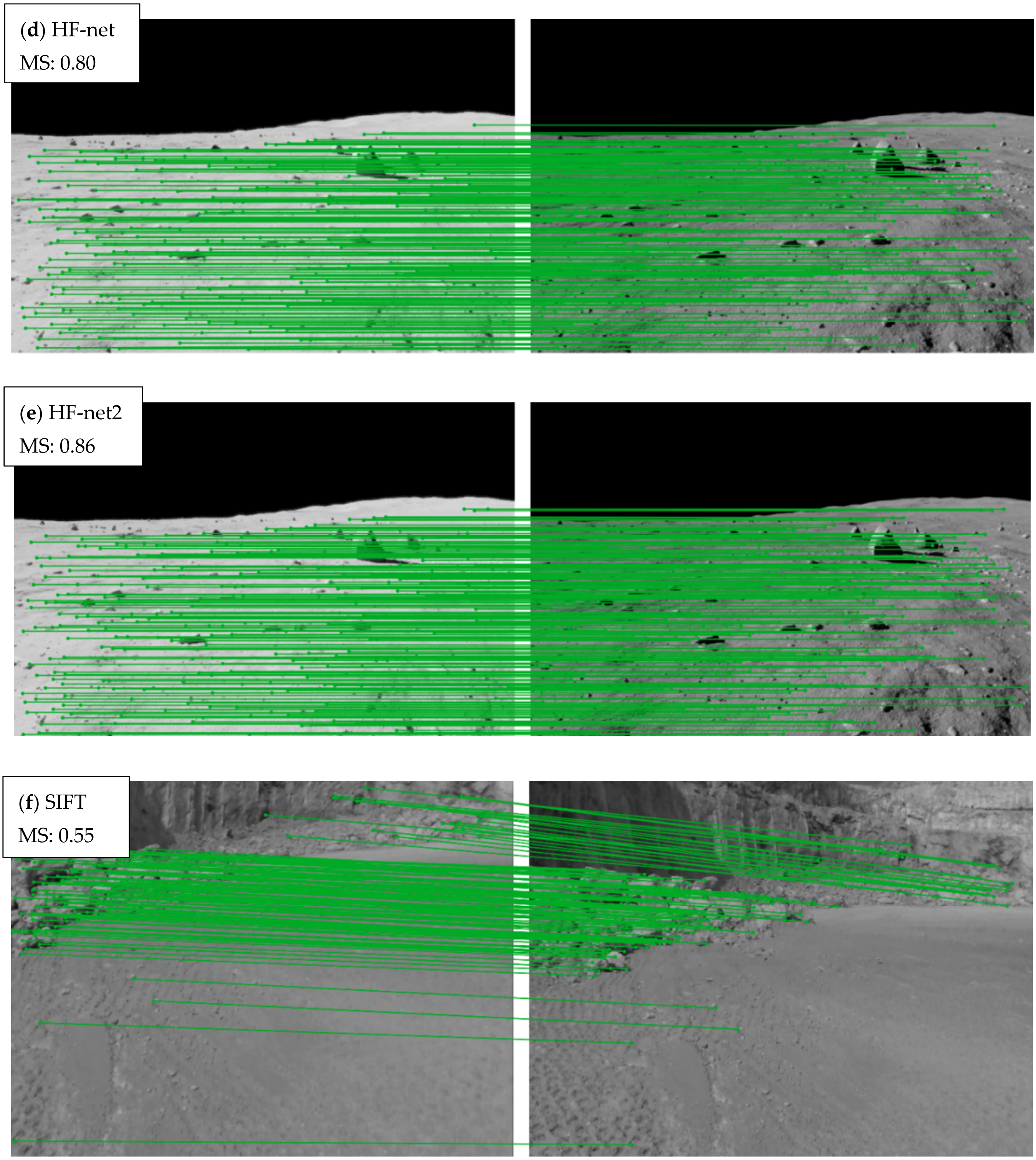


Figure 13. *Cont.*

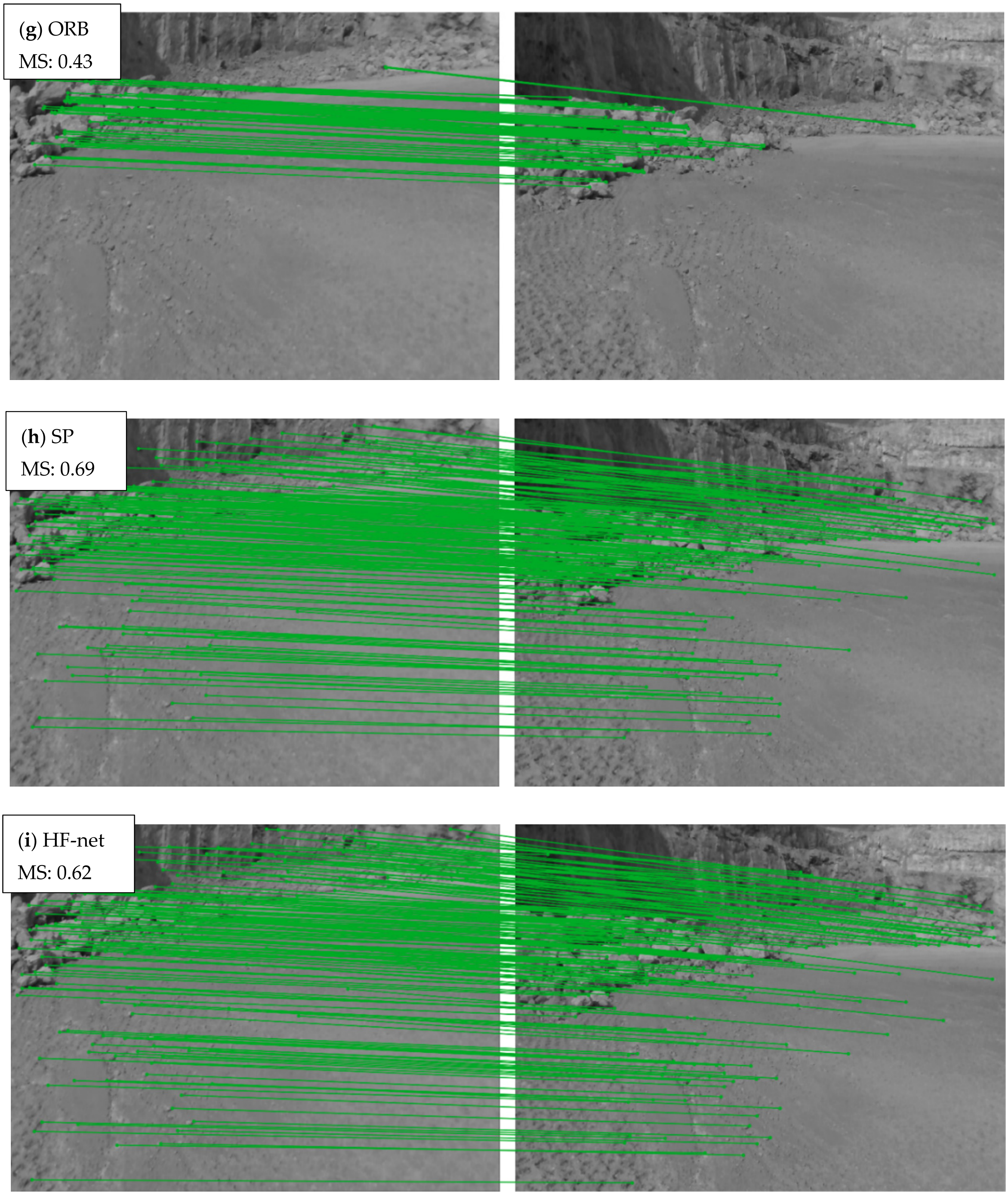


Figure 13. *Cont*.

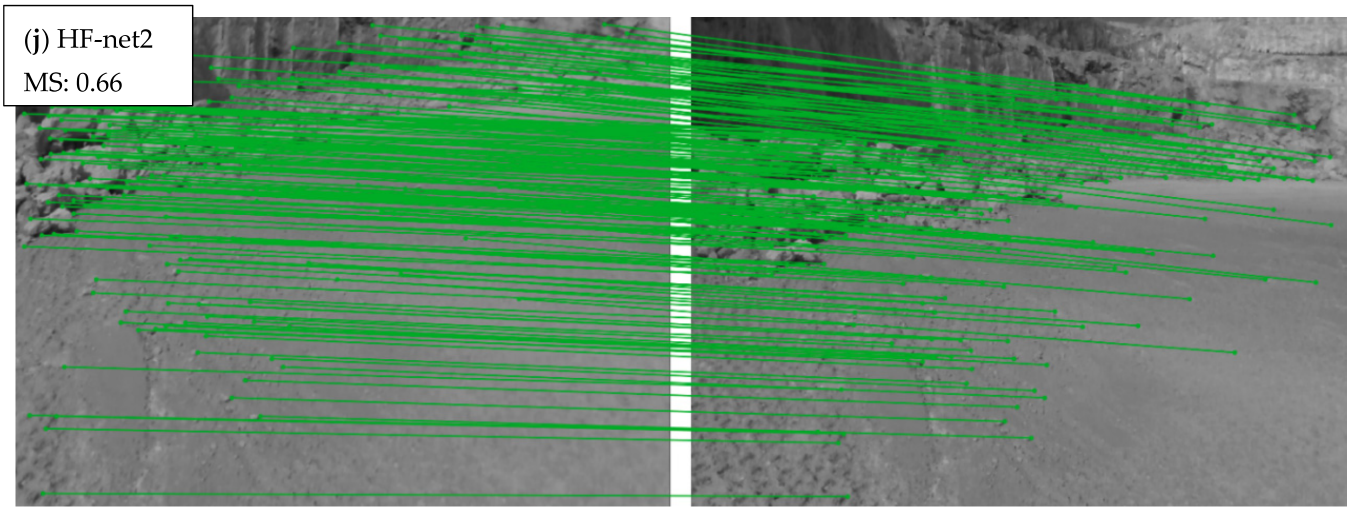


Figure 13. Matching scores of the SIFT, ORB, SuperPoint, original HF-net and proposed HF-net2 descriptors. Two images from the evaluation dataset are presented: (**a**–**e**) lunar scene testing illumination changes, (**f**–**j**) earthy scene testing viewpoint changes. The green lines connect the matching points between the two images.

As presented in Figure 12, the green dots extracted from SuperPoint, HF-net2 and HF-net, which are repeatable points in both the (right and left) images of each row, are localized on more meaningful features that better describe the scene than the detected keypoints of SIFT, FAST and Harris. 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 13, which visualizes the matching points in two different scenes that suffer from a lack of visual cues, the proposed architecture provides outstanding results (Figure 13e,j) compared with SIFT and ORB and an increased matching score compared with the original HF-net. It is worth noting that in Figure 13a–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 at different times of day, using an RGB-Depth camera, while an RTK 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 the ORB-SLAM2, the proposed SLAM outperforms the 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 (Figure 10b), the error of the ORB-SLAM2 in the end of the path is at a level of 30 cm instead of the proposed SLAM with an error at a level of 12 cm.

Concerning the sandy scene, the first experiment was performed using a random-based path on 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 at a level of 0.25 and a standard deviation at a level of 0.05 in both experiments, instead of the ORB-SLAM2 where, in the first experiment, the RMSE was noted at a level of 0.35, and at the second experiment, 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 the 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 significantly higher and more robust results compared with the ORB-SLAM2.

6. Conclusions

To sum up, a multi-task distillation-based architecture, called HF-net2, was developed aiming to implement a keypoint detector and descriptor which is focused on unstructured environments and completely unknown planetary scenes. The model was trained with a specialized image dataset with planetary scenes and evaluated in terms of illumination and viewpoint changes, using a proposed benchmark dataset. HF-net2 proved its robustness compared with several well-known feature extractors, achieving the highest overall accuracy after its teacher, SuperPoint.

Moreover, the HF-net2 model was integrated in a visual SLAM system based on ORB-SLAM2 aiming to develop a SLAM system that is specialized in unstructured environments. 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 [48].

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.

Author Contributions: Conceptualization, P.P. and G.P.; methodology, G.P.; software, G.P.; validation, G.P.; formal analysis, G.P.; investigation, G.P.; resources, G.P.; data curation, G.P.; writing—original draft preparation, G.P.; writing—review and editing, P.P. and G.P.; visualization, G.P.; supervision, P.P.; project administration, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The training dataset is publicly available on Kaggle in the following link: https://www.kaggle.com/datasets/georgepetrakis/unstructured-and-planetary-environments (accessed on 26 September 2023). The evaluation dataset is publicly available on Kaggle and Github in the following links: https://www.kaggle.com/datasets/georgepetrakis/ppatches-evaluation-dataset-for-planetary-scenes (accessed on 26 September 2023), https://github.com/Gpetrak/p-patches (accessed on 26 September 2023).

Acknowledgments: 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.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, C.; Xu, J.; Wang, F. A Review of Keypoints’ Detection and Feature Description in Image Registration. *Hindawi Sci. Program.* **2021**, *2021*, 8509164. [CrossRef]
2. Cadena, C.; Carlone, L.; Carrillo, H.; Latif, Y.; Scaramuzza, D.; Neira, J.; Reid, I.; Leonard, J.J. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age. *IEEE Trans. Robot.* **2016**, *32*, 6. [CrossRef]

3. Tsintotas, K.A.; Bampis, L.; Gasteratos, A. The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 11. [CrossRef]
4. Harris, C.; Stephens, M. A Combined Corner and Edge Detector. In Proceedings of the Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988. [CrossRef]
5. Shi, J.; Tomasi, C. *Good Features to Track*; Technical Report; Cornell University: Ithaca, NY, USA, 1993.
6. Rosten, E.; Drummond, T. Machine Learning for High-Speed Corner Detection. In Proceedings of the ECCV, Graz, Austria, 7–13 May 2006. [CrossRef]
7. Alcantarilla, P.; Nuevo, J.; Bartoli, A. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. In Proceedings of the BMVC, Bristol, UK, 9–13 September 2013. [CrossRef]
8. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011. [CrossRef]
9. Lowe, D. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
10. Bay, H.; Ess, A.; Tuytelaars, T.; Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
11. Li, R.; Wang, S.; Gu, D. Ongoing Evolution of Visual SLAM from Geometry to Deep Learning: Challenges and Opportunities. *Cogn. Comput.* **2018**, *10*, 875–889. [CrossRef]
12. Lategahn, H.; Geiger, A.; Kitt, B. Visual SLAM for autonomous ground vehicles. In Proceedings of the IEEE ICRA, Shanghai, China, 9–13 May 2011. [CrossRef]
13. Singandhupe, A.; La, H. A Review of SLAM Techniques and Security in Autonomous Driving. In Proceedings of the IRC, Naples, Italy, 25–27 February 2019. [CrossRef]
14. Zou, Q.; Sun, Q.; Chen, L.; Nie, B.; Li, Q. A Comparative Analysis of LiDAR SLAM-Based Indoor Navigation for Autonomous Vehicles. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 6907–6921. [CrossRef]
15. Inzerillo, L.; Mino, G.; Roberts, R. Image-based 3D reconstruction using traditional and UAV datasets for analysis of road pavement distress. *Autom. Constr.* **2018**, *96*, 457–469. [CrossRef]
16. 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.* **2018**, *12*, 151–164. [CrossRef]
17. Aulinas, J.; Carreras, M.; Llado, X.; Salvi, J.; Garcia, R.; Prados, R.; Petillot, Y. Feature extraction for underwater visual SLAM. In Proceedings of the OCEANS, Santander, Spain, 6–9 June 2011. [CrossRef]
18. Jung, K.; Hitchcox, T.; Forbes, J. Performance Evaluation of 3D Keypoint Detectors and Descriptors on Coloured Point Clouds in Subsea Environments. *arXiv* **2022**, arXiv:2209.12881.
19. Trabes, E.; Jordan, M.A. A Node-Based Method for SLAM Navigation in Self-Similar Underwater Environments: A Case Study. *Robotics* **2017**, *6*, 29. [CrossRef]
20. Santos, L.C.; Aguiar, A.S.; Santos, F.N.; Valente, A.; Petry, M. Occupancy Grid and Topological Maps Extraction from Satellite Images for Path Planning in Agricultural Robots. *Robotics* **2020**, *9*, 77. [CrossRef]
21. Guo, J.; Borges, P.; Park, C.; Gawel, A. Local Descriptor for Robust Place Recognition using LiDAR Intensity. *arXiv* **2018**, arXiv:1811.12646. [CrossRef]
22. Oelsch, M.; Opdenbosch, V.; Steinbach, E. Survey of Visual Feature Extraction Algorithms in a Mars-like Environment. In Proceedings of the ISM, Taichung, Taiwan, 11–13 December 2017. [CrossRef]
23. Wan, W.; Peng, M.; Xing, Y.; Wang, Y.; Liu, Z.; Di, K.; Teng, B.; Mao, X.; Zhao, Q.; Xin, X.; et al. A Performance comparison of feature detectors for planetary rover mapping and localization. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2017**, *XLII-3/W1*, 149–154. [CrossRef]
24. Furgale, P.T.; Carle, P.; Enright, J.; Barfoot, T.D. The Devon Island Rover Navigation Dataset. *Int. J. Robot. Res.* **2012**, *31*, 707–713. [CrossRef]
25. Wu, B.; Zeng, H.; Hu, H. Illumination invariant feature point matching for high-resolution planetary remote sensing images. *Planet. Space Sci.* **2018**, *152*, 45–54. [CrossRef]
26. Otsu, K.; Agha-Mohammadi, A.; Paton, M. Where to Look? Predictive Perception with Applications to Planetary Exploration. *IEEE Robot. Autom. Lett.* **2018**, *3*, 635–642. [CrossRef]
27. 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**, arXiv:2109.06596. [CrossRef]
28. Mur-Artal, R.; Tardos, J. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Trans. Robot.* **2016**, *33*, 1255–1262. [CrossRef]
29. Qin, T.; Li, P.; Shen, S. VINS-Mono: A Robust and Versatile MonocularVisual-Inertial State Estimator. *arXiv* **2017**, arXiv:1708.03852. [CrossRef]
30. Hong, S.; Bangunharcana, A.; Park, J.-M.; Choi, M.; Shin, H.-S. Visual SLAM-Based Robotic Mapping Method for Planetary Construction. *Sensors* **2021**, *21*, 7715. [CrossRef]
31. Sarlin, P.; Cadena, C.; Siegwart, R.; Dymczyk, M. From Coarse to Fine: Robust Hierarchical Localization at Large Scale. In Proceedings of the CVPR, Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
32. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]

33. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 26 June–1 July 2016.
34. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the ICCV, Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
35. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
36. Elsken, T.; Metzen, J.; Hutter, F. Neural Architecture Search: A Survey. *arXiv* **2019**, arXiv:1808.05377.
37. Ramachandran, P.; Zoph, B.; Le, Q. Searching for Activation Functions. *arXiv* **2017**, arXiv:1710.05941.
38. 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. In Proceedings of the CVPR, Las Vegas, NV, USA, 26 June–1 July 2016.
39. 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* **2020**, arXiv:2008.05416.
40. Munoz-Salinas, R.; Medina-Carnicer, R. UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers. *Pattern Recognit.* **2020**, *101*, 107193. [CrossRef]
41. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, L. Microsoft COCO: Common Objects in Context. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014. [CrossRef]
42. Lu, S. *Mars Surface Image (Curiosity Rover) Labeled Data Set Version 1*; Updated January 2023; NASA: Washington, DC, USA, 2023.
43. Smith, E.; Zuber, T.; Jackson, B.; Cavanaugh, J.F.; Neumann, G.A.; Riris, H.; Sun, X.; Zellar, R.S.; Coltharp, C.; Connelly, J.; et al. The Lunar Orbiter Laser Altimeter Investigation on the Lunar Reconnaissance Orbiter Mission. *Space Sci. Rev.* **2010**, *150*, 209–241. [CrossRef]
44. Balntas, V.; Lenc, K.; Vedaldi, A.; Tuytelaars, T.; Matas, J.; Mikolajczyk, K. H-Patches: A Benchmark and Evaluation of Handcrafted and Learned Local Descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2825–2841. [CrossRef]
45. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *arXiv* **2016**, arXiv:1603.04467.
46. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. *BRIEF: Binary Robust Independent Elementary Features*; ECCV 2010; Springer: Berlin/Heidelberg, Germany, 2010. [CrossRef]
47. Schubert, D.; Goll, T.; Demmel, N.; Usenko, V.; Stückler, J.; Cremers, D. The TUM VI Benchmark for Evaluating Visual-Inertial Odometry. In Proceedings of the IROS, Madrid, Spain, 1–5 October 2018. [CrossRef]
48. 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. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.