
Traversability Estimator for Legged Robot

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics
Artificial Intelligence

presented by
Francesco Saverio Zuppichini

under the supervision of
Prof. Student's Alessandro Giusti
co-supervised by
Prof. Student's Co-Advisor

June 2019

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Francesco Saverio Zuppichini
Lugano, 28 June 2019

Abstract

Effective identification of traversable terrain is essential to operate mobile robots in different environments. Historically, to estimate traversability, texture and geometric features were extracted before train a traversability estimator in a supervised way. However, with the recent deep learning breakthroughs in computer vision, terrain features may be learned directly from raw data, images or heightmaps, with higher accuracy.

We implement a full pipeline to estimate traversability with high accuracy using only data generated through simulation using a deep convolutional neural network. The method is based on the framework proposed by Chavez-Garcia et al. [5] and was originally tested on two wheels small robot. Collecting data in a simulation environment yields several advantages over real world. Ground can be easily generated to include any features and can be loaded on the fly into the simulator. Multiple simulations can be run in parallel reducing the cost for real world hardware.

The dataset was generated using thirty synthetic surfaces with different features such as slopes, holes, bumps, and walls. Then, we let the robot walk each map on for a certain amount of time while storing its interactions. Later, precisely crop a patch from each simulation trajectory's point that includes the whole robot's footprint and the amount of ground it should reach without obstacles in a decided time window. Then, we label each patch as traversable or not traversable based on a minimum advancement depending on the robot used. We open source the framework and the same methodology can be used with any type of robots. We test the methodology on a legged crocodile-like robot that presents challenging locomotion to be learned in order to correctly predict its traceability.

We select and test different networks architecture and select the best performing one. We quantify shows its performance with different numeric metrics on different real-world terrains. In addition, we qualitatively evaluate the network by showing the predicted traversability probability on different grounds to better visualize the acquired knowledge.

Later, we utilize model interpretability techniques to understand which patches confuse the network. We visualize the ground regions that the network fails to classify and find which part of the inputs was responsible for the wrong predictions. Finally, we test the model strength and robustness by comparing its prediction on custom patches composed by crafted features, such as a patch with a wall ahead, to the ground truth obtained by running again Krock on that ground in the simulator. The results suggest that the model was able to match the ground truth in different situations.

Contents

Contents	iii
1 Introduction	1
2 Related Work	3
3 Methodology	5
4 Implementation	9
4.1 Robot	9
4.2 Data Gathering	10
4.2.1 ground generation	11
4.2.2 Real world maps	13
4.3 Simulation	13
4.3.1 Setup	13
4.3.2 Spawning	13
4.3.3 Interactions	15
4.4 Postprocessing	15
4.4.1 Parse trajectories	15
4.4.2 Advancement computation	18
4.4.3 Patch extraction	18
4.4.4 Final dataset	18
4.5 Label patches	18
4.6 Estimator	20
4.6.1 Vanilla Model	20
4.6.2 ResNet	24
4.6.3 Preactivation	24
4.6.4 Squeeze and Excitation	25
4.6.5 MicroResNet	25
4.7 Normalization	25
4.8 Data Augmentation	28
4.8.1 Dropout	28
4.8.2 Coarse Dropout	28
4.8.3 Simplex Noise	28
4.9 Tools	30
4.9.1 Software for simulation	30

4.9.2	Data processing	30
4.9.3	Visualization	31
5	Results	33
5.1	Hardware	33
5.2	Experimental validation	33
5.2.1	Metrics	33
5.3	Training setup	33
5.4	Quantitative results	34
5.4.1	Model selection	34
5.5	Final results	34
5.5.1	Classification	34
5.5.2	Regression	34
5.6	Qualitative results	36
5.6.1	Quarry	36
5.6.2	Bars	36
5.6.3	Small village	39
6	Interpretability	41
6.0.1	Features separability	41
6.0.2	Train set	42
6.0.3	Test set	42
6.1	Grad-CAM	44
6.2	Quarry dataset	44
6.2.1	Best	47
6.2.2	Worst	49
6.2.3	False Negative	50
6.2.4	False Positive	51
6.3	Robustness	52
6.3.1	Untraversable wall ahead at increasing distance	52
6.3.2	Increasing height walls ahead	55
6.3.3	Increasing height and distance walls ahead.	57
6.3.4	Tunnel	57
6.3.5	Ramps	60
7	Conclusions	63

Chapter 1

Introduction

All living beings, including humans, need to traverse ground to eat, sleep and breed. Animals through millennia of evolution have developed different techniques to classify ground regions. Usually, local sensors are used to map new terrain to effectively navigate the environment. Humans, for instance, are equipped with a powerful traversability estimator, called vision, that is able to extract features from a local region, a patch, such as elevation, size, and steepness and combine them with the brain to produce a local planner able to find a suboptimal path to cross.

Similarly, ground robots need to walk in order to fulfill their tasks. Taking inspiration from nature, robots must be able to correctly aggregate terrain features to predict their traversability. In this project, we consider the task to estimate the traversability of 3D terrain for a specific ground robot. Traversability can be estimated by 3D geometric, appearance features or both [17]. The first one utilizes purely geometric features to label a patch, while the second relies on categories, glass, rocks, etc, with different traversability probabilities extract mostly on camera's images. In both cases, those samples are used in supervised learning to train a traversability estimator. While appearance data must be collected directly from a real or simulated environment, geometric features do not directly depend on the robot's interactions with the ground.

In most indoor scenarios, data is collected through specific hardware, such as infrared sensors, cameras or LIDAR, during a robot exploration of the environment. Due to their artificial design, indoor environments shared similar features across the world, the flatness of the floor, stairs, and ceiling. So, even if an agent is trained with samples collected in one location, due to physical limitations, the learned mapping can be effective in different locations. Moreover, in indoor environments, traversability estimation is mostly solved with obstacle avoidance since, by design, the floor alone do not include bumps, ramps or holes. In addition, while the robot operates there may be other agents interacting with the environment, such as humans, therefore a correct function to avoid collision must properly learn to effectively move safely the robot.

On the other hand, outdoor grounds may have less artificial obstacle but they have a not homogeneous ground making challenging to estimate where the robot can properly travel. Also, a given patch with a specific shape may not be traversable by all direction due to the not uniform features of the terrain. On top of that, using real-world data may be unfeasible due to the time required to move the robot on different grounds and the possibility to introduce bias if the data samples are not varied enough. Fortunately, the ground can be generated artificially and used in a simulated environment to let the robot walk in it while storing its interactions. Moreover, grounds maps can be easily obtained nowadays by using third-party services such as google maps or flying drone equipped with

mapping technologies.

We proposed a full pipeline to estimate traversability tested on a legged crocodile-like robot on uneven terrain based entirely on data gathered through simulation. Simulation offers several advantages than real-world data collection. First, we can include a rich array of different terrains without the need to physically move them. Data is collected faster and multiple simulations can be launch in parallel keeping the cost of the real world hardware low. We first generated thirty synthetic maps encoded as heightmaps with different features: bumps, walls, slopes, steps, and holes. Then, we run the robot on each one of them for a certain amount of time while storing its interactions with the environment, position, and orientation. Later, we generate a training dataset by cropping a patch for each stored data point in such way that includes the robot's footprint and the maximum possible amount of ground that can traverse in a selected time window. After, we label each patch by traversable or not traversable if the robot's advancement computed in that time using the same time window is less or greater than a threshold. The threshold depends on the robot's locomotion and is calculated by spawning it in front of the different obstacle and observing its advancement, we set the value for the tested robot to twenty centimeters. Those patches are used to fit a deep convolutional neural network to prediction the traversability leaving to the network the task to extract important ground's features.

The report is organized as follow, the next chapter 2 introduces the related work, chapter 3 gives a high-level overview of our approach, chapter 4 talks in deep about the implementation details showing the dataset generation procedure and describes different convolutional neural network architectures. Chapter 5 describes the model selection process, shows the quantitatively and qualitatively results for the best performing architecture. Chapter 6, we test the model's predictions robustness using different techniques. First, we proved the network's ability to learn meaningful features from the inputs, section 6.0.1. Then, 6.2 we visualize different inputs to understand which grounds' regions caused the predictions. We adopted a technique that allows highlighting the part of the terrain contributes the most to the network's output. Lastly, in section 6.3, we tested the model's robustness by creating different patches with unique characteristics. We compared the estimator's predictions to the ground truth gather the simulator.

Chapter 2

Related Work

The learning and perception of traversability is a fundamental competence for both organisms and autonomous mobile robots since most of their actions depend on their mobility [23]. Usually, visual perception is used in most all animals to correctly estimate if an environment can be traversed or not. Similar, a wide array of autonomous robots adopt local sensors to mimic the visual properties of beings to extract meaningful information of the surrounding and plan a safe path through it. This chapter gives an overview of some of the existing methods in the literature to solve this task.

Most of the methodologies to estimate traversability rely on supervised learning, where first the data is gathered and then a machine learning algorithm is trained to correctly predict the traversability of those samples. Most of the approaches can be clustered into two categories, based on the inputs data: geometric and appearance based methods.

Geometric methods aim to detect traversability using geometric properties of surfaces such as distances in space and shapes. Those properties are usually slopes, bumps, and ramps. Since nearly the entire world has been surveyed at 1 m accuracy [22], outdoor robot navigation can benefit from the availability of overhead imagery, for this reason, recently, elevation data has been widely used to generate datasets. Chavez-Garcia et al. [5], proposed a framework to learn traversability using elevation data as input in the form of height maps. They trained a convolutional neural network to predict from a dataset, entirely generated through simulation, the traversability probability of a small ground region around the robot.

Elevation data can also be estimated by flying drones. Delmerico et al. [7] proposed a collaborative search and rescue system in which a flying robot that explores the map and creates an elevation map to guide the ground robot to the goal. They train on the fly a CNN to segment the terrain in different traversable classes. This map is later used to plan a path for the mobile robot.

Whereas appearance methods, to a greater extent related to camera images processing and cognitive analyses, have the objective of recognizing colors and patterns not related to the common appearance of terrains, such as grass, rocks or vegetation. Historically, the collected data is first preprocessed to extract texture features that are used to fit a classic machine learning classified such us an SVM [23] or Gaussian models [22]. Those techniques rely on texture descriptors, for example, Local Binary Pattern [16], to extract features from the raw data images obtained from local sensors such as cameras. With the rise of deep learning methods in computer vision, deep convolution neural network has been trained directly on the raw RGB images bypassing the need to define characteristic features.

One recent example is the work of Giusti et al. [3] where a deep neural network was training

on real-world hiking data collected using head-mounted cameras to teach a flying drone to follow a trail in the forest. Geometric and appearance methods can be used together, one example is the work of Delmerico et al.[6] extended the previous work [7] by proposing the first on-the-spot training method that uses a flying drone to gather the data and train an estimator in less than 60 seconds.

Data can also extract in simulations, where an agent interacts in an artificial environment. Usually, no-wheel legged robot able to traverse harder ground, can benefits from data gathering in simulations due to the high availability. For example, Klamt et al. [12] introduced a locomotion planner that is learned in a simulation environment.

On other major distinction, is between different types of robots: wheel and no-wheel. We will focus on the later since we adopt a legged crocodile-like robot to extend the existing framework proposed by Chavez-Garcia et al. [5].

Legged robots show their full potential in rough and unstructured terrain, where they can use a superior move set compared to wheel robots. Different frameworks have been proposed to compute safe and efficient paths for legged robots. Wermelinger et al. [24] use typical map characteristics such as slopes and roughness gather using onboard sensors to train a planner. The planner uses a RRT* algorithm to compute the correct path for the robot on the fly. Moreover, the algorithm is able to first find an easy local solution and then update its path to take into account more difficult scenarios as new environment data is collected.

Due to uneven shape rough terrain, legged robots must be able to correctly sense the ground to properly find a correct path to the goal. Wagner et al. [13] developed a method to estimate the contact surface normal for each foot of a legged robot relying solely on measurements from the joint torques and from a force sensor located at the foot. This sensor at the end of a leg optically determines its deformation to compute the force applied to the sensor. They combine those sensors measurement in an Extended Kalman Filter (EKF). They showed that the resulting method is capable of accurately estimating the foot contact force only using local sensing.

While the previous methods rely on handcrafted map's features extraction methods to estimate the cost of a given patch using a specific function, new frameworks that automatize the features extraction process has been proposed recently. Lorenz et al. [15] use local sensing to train a deep convolutional neural network to predict the terrain's properties. They collect data from robot ground interaction to label each image in front of the robot in order to predict the future interactions with the terrain showing that the network is perfectly able to learn the correct features for different terrains. Furthermore, they also perform weakly supervised semantic segmentation using the same approach to divide the input images into different ground classes, such as glass and sand, showing respectable results.

Chapter 3

Methodology

We developed a framework to learn traversability directly on ground patches for any robot using purely simulated data. The core idea, originally proposed by Chavez-Garcia et al. [5], is simple yet elegant, a robot is let travel forward with a fix controller into a simulator on different synthetic terrains. While moving it stores its interactions with the environment, namely pose and orientation. Then we crop a region of ground, a patch, around each traveled position and directly train a neural network on them to predict traversability. Figure 3.1 shows the framework's steps.

Collecting data through simulation has several main advantages such as variety, cost, and speed. With simulations, we can easily increase the number of data samples at any time by simply incorporate more maps or run more experiments. Contrarily, a real robot requires to first identify the suitable ground and then physically transport the machine there and finally let it walk on it. Also, real robots cannot completely explore all grounds due to the possibility to be damaged, especially in outdoors scenarios. This prevents the full exploration of the environment. On the other hand, a virtual robot can be destroyed and regenerate indefinitely. This allows a total exploration of all not traversable samples, essential to generate a quality dataset. Of course, a realistic model controller must be implemented to simulate the real robot. Furthermore, we are not limited by real constraint and we can design the ground to maximize the variety of robot's interactions. For instance, we can artificial design maps to include any kind of situations and challenges. Clearly, this also reduces the time required to collect the robot's interactions with the terrain. The time could be further minimized by running different simulations in parallel bypassing the demand for more physical hardware. To create a quality dataset, we must generate a series of various surfaces and ensure their diversity. Intuitively, those maps should be small enough to allow fast exploration but big enough to include several features. Modern ground generation techniques allows to generate a rich array of terrains with different characteristics such us bumps, ramps, slopes in different levels and sizes. We adopted the same ground generation procedured proposed in the original work [5]. Figure 3.2 shows a small collection of synthetic terrains with different features we created. We represented the terrains with heightmaps, where the height component of the ground is stored as pixels in a gray image. Thus, the image is a top-view representation of the terrain where the brightness describes the height.

Robots interacts with the maps when it is loaded on the simulator. To ensure a correct exploration of each terrain, we randomly spawn the robot multiple times on the same map and let it walk for a fixed amount of time. Random spawn reduces the possibility to repeatedly visiting the same spot and it maximizes the robot's exploration.

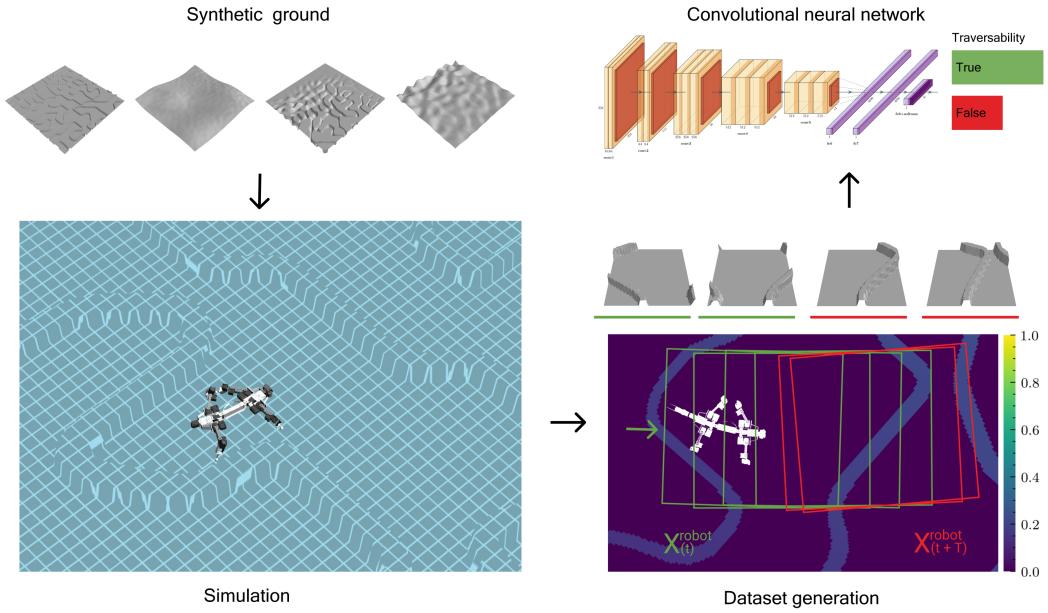


Figure 3.1. The framework’s main building block in counter-clockwise order. The robot is a legged robot crocodile-like. First, we generated meaningful synthetic grounds. Then, the robot randomly spawns and walks forward with a fix controller on the maps inside a simulated environment. While moving, it stores its interactions. Then, we crop a region of ground, a patch, for each simulation trajectory around the robot. The patch’s size is calculated according to its locomotion. We labeled those images using a defined threshold. Finally, we trained a deep convolutional neural network to predict traversability.

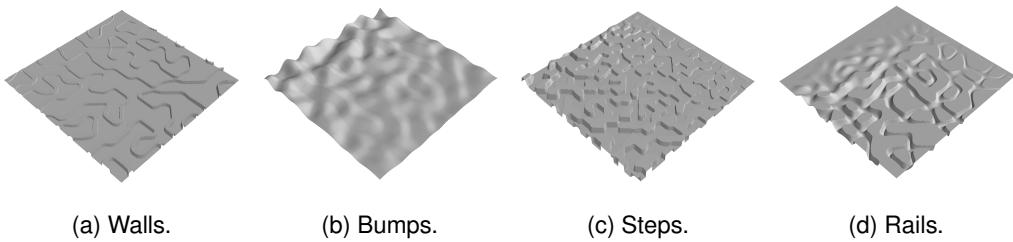


Figure 3.2. Some artificially generated terrains, $10 \times 10\text{m}$, with one specific feature each.

The robot position is tracked and stored during simulation to extract a small portion of ground, a patch, around each of its trajectory's position. The patches are cropped from the heightmap using a specific size based on the robot's footprint and its velocity. Each resulting image is labeled using a minimum advancement based on the robot's characteristic. We tested our framework on a legged crocodile-like robot called *Krock*. Due to its four legs, the robot has very unique locomotion allowing it to overcome different obstacles making estimate traversability more challenging. Figure 3.3c helps to visualize the procedure. The dataset generation process is detailed explained in section 4.2.

The dataset generated is used to train a deep convolutional neural network. The model estimates traversability given an heightmap, an image, representing a small region of terrain as input. We improved the original architecture by adopting a ResNet variant based on recent works [8] [9] [10]. We methodologically tested different variations of the new architecture to finally generate a model with three-time fewer parameters than the network proposed by the original paper [5] with comparable performance. Due to its smaller size, the network has a lower prediction time than the original one. Thus, it is can be easily deployed using less demanding hardware. We appraised the models' performance booth quantitatively, with numerical metrics, and qualitatively, showing the traversability estimation on real-world terrains.

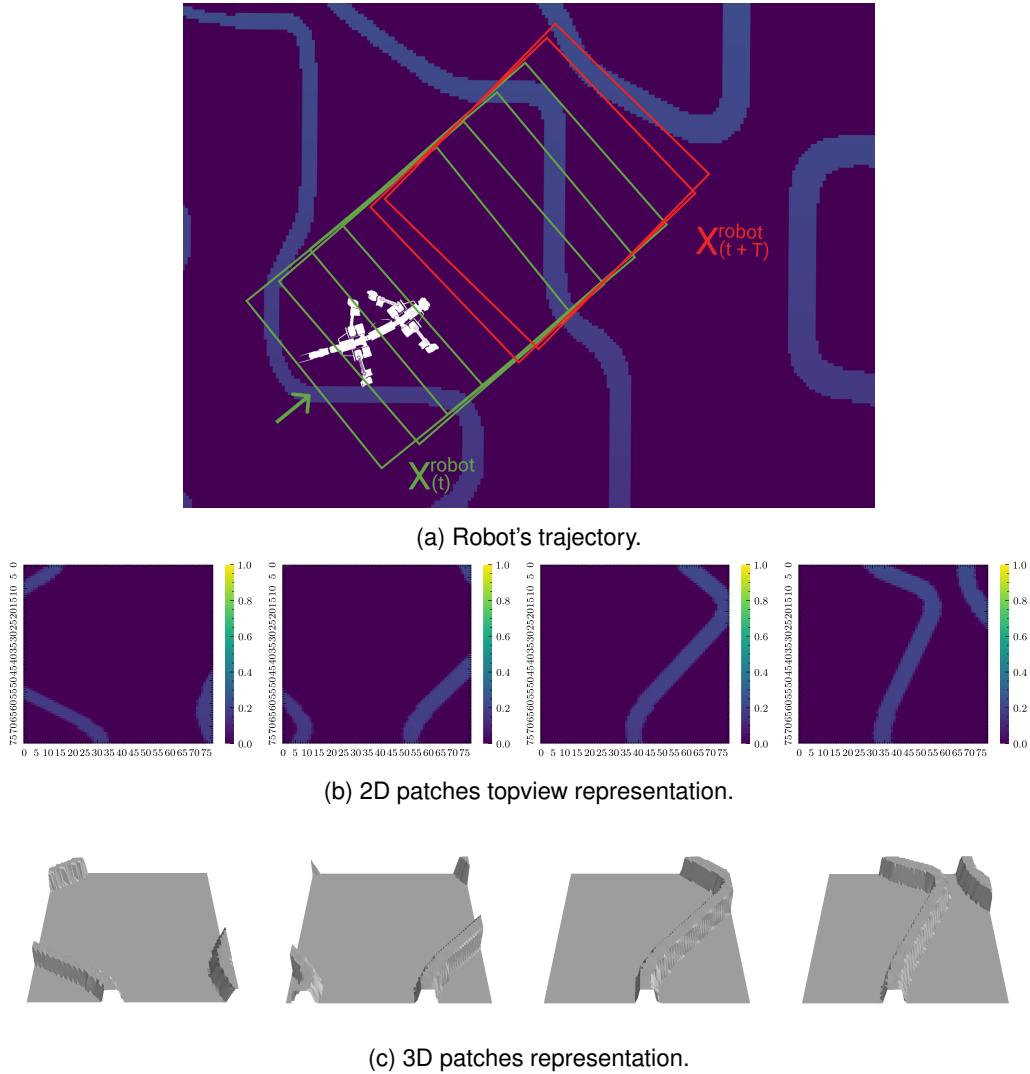


Figure 3.3. Example of a robot's trajectory (a) extracted during training on a map with different walls. Robot's initial position is shown by its white silhouette. Patches borders are labeled with greed if traversable and red if not and showed as 2D (b) and 3D(c) rendered images. The robot traverses the patches from left to right.

Chapter 4

Implementation

This section talks in detail about the implementation of the framework. We will first talk about from the utilized robot, then describe each step in the dataset generation, illustrate the utilized estimators and finish by listing the employed software.

4.1 Robot

We tested our framework on a legged crocodile-like robot called *Krock* developed at EPFL. The next figure shows the robot in real life with a coat and in a simulated environment.



(a) Krock in real life.

(b) Krock in the simulator.

Figure 4.1. *Krock*

Krock robot was created for the purpose of monitoring real crocodiles, hence it must walk and behave realistically enough to fool the real crocodiles. *Krock* has four legs, each one of them is equipped with three motors in order to rotate in each axis. The robot can raise itself in three different configurations, gaits, using the four motors on the body connected to the legs. In addition, there are another set of two motors in the inner body part to increase *Krock*'s move set. The tail is composed by another set of three motors and can be used to perform a wide array of tasks. The

robot is 85cm long, weights around 1.4kg and in its standard gait configuration its distance from the ground is 16cm. The next figure 4.2 shows *Krock* from the top helping the reader understanding its composition and the correct ratio between its parts. Also, each motor is highlighted with a red marker. *Krock*'s moves by lifting and moving forward one leg after the other. The following

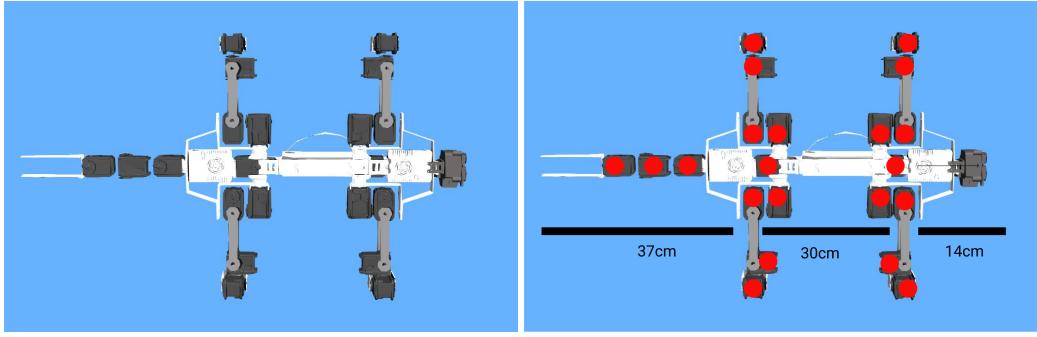
(a) Top view of *Krock*(b) Details highlighted in *Krock*

Figure 4.2

figure shows the robots going forward. In our framework we fixed the gait configuration to normal,

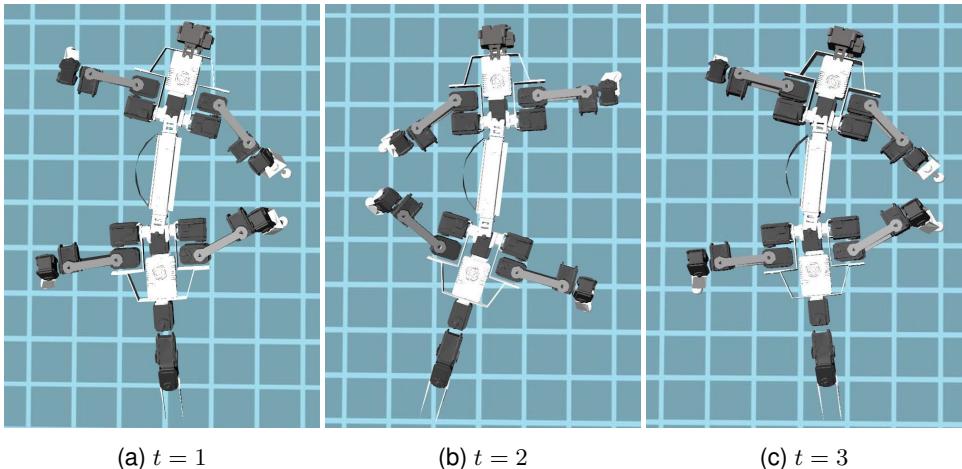
(a) $t = 1$ (b) $t = 2$ (c) $t = 3$

Figure 4.3. *Krock* moving forward. In this gait configuration, the robot distance from the ground is 16cm.

showed in figure ??, where the body is at the same legs' motors height.

4.2 Data Gathering

This section describe in detail how we created, collected and processed the synthetic dataset used to train the traversability estimator in supervised way.

4.2.1 ground generation

We created thirty rom each trajectory, what you did in simulation was to command the robot to move forward for 2 secs, i.e. created a trajectory from the spawning pose to the pose reached at the end of those 2 secs of moving forward. Ergo, you have to extract the patches at each pose of the trajectory generated at simulation timelution of 0.02cm/pixel using 2D simplex noise variant of Perlin noise [1], a widely used technirom each trajectory, what you did in simulation was to command the robot to move forward for 2 secs, i.e. created a trajectory from the spawning pose to the pose reached at the end of those 2 secs of moving forward. Ergo, you have to extract the patches at each pose of the trajectory generated at simulation timelitterature. We divide the maps into five main categories of terrains to cluster differents ground features: *romeachtrajectory, whatyoudidinsimulationwastocommandtherobottomoveforwardfor2secs, i.e.createdatrajec* and *holes*.

Bumps: We genered four different maps with increasing bumps' height using simplex noise.

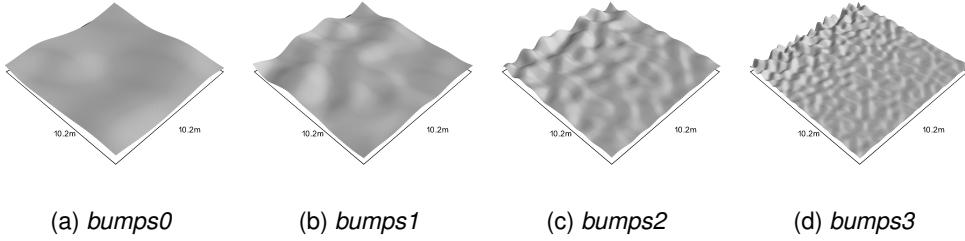


Figure 4.4. Bumps maps.

Bars: In these maps there are wall with different shapes and heights.

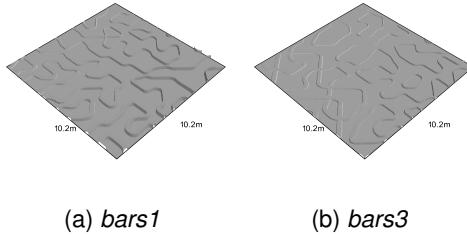


Figure 4.5. Bars maps.

Rails: Flat grounds with slots.

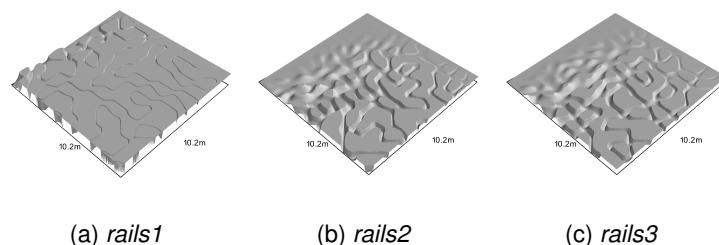


Figure 4.6. Rails maps.

Steps: These are maps with various steps at increasing distance and frequency.

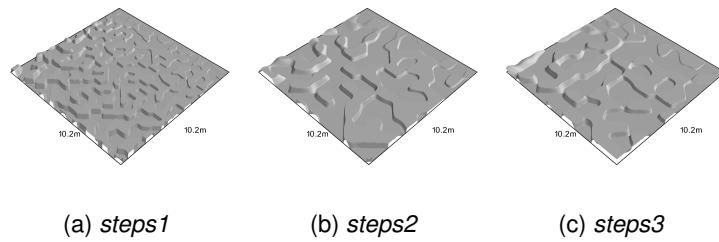


Figure 4.7. Steps maps.

Slopes/Ramps: Maps composed by uneven terrain scaled by different height factors from 3 to 5 used to include samples where *Krock* has to climb.

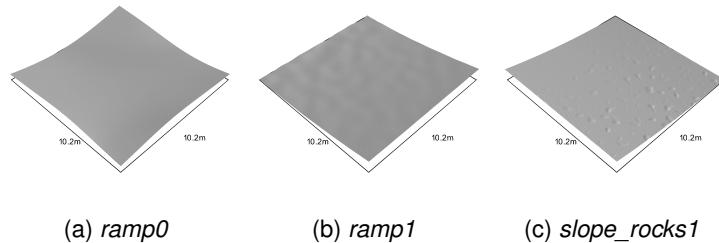


Figure 4.8. Slopes maps.

Holes We also included a map with holes

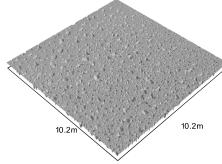
(a) *holes1*

Figure 4.9. Holes map.

4.2.2 Real world maps

We also included real word terrains. We gather two heightmaps produced by ground mapping flying drones from sensefly’s dataset, a quarry and a small village. Figure 4.10 shows the original and a 3D render of the heightmap for each terrain

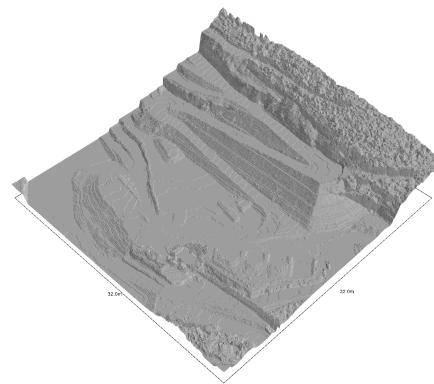
4.3 Simulation

4.3.1 Setup

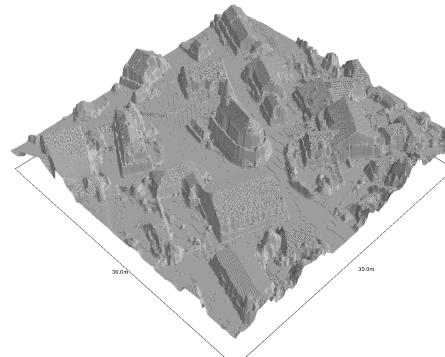
We used Webots as the simulation platform for the robot model and the generated terrains. The controller was implemented using the Robotic Operating System, and send data using a special nodes called topics. The controller implements a ROS’ node to publish Krock status including its pose at a rate of 250hz. We decided to reduce it to 50hz by using ROS build it `throttle` command.

4.3.2 Spawning

To collect *Krock*’s interaction with the environment, we spawn the robot on the ground and let it move forward for t seconds. We repeat this process n times per each map. Unfortunately, spawning the robot is not a trivial task. In certain maps, for instance, *bars1*, a map with tons of walls, we must avoid spawning *Krock* on an obstacle otherwise the run will be ruined by *Krock* getting stuck at the beginning introducing noise in the dataset. To solve the problem, we defined two spawn strategies, a random spawn and a flat ground spawn strategy. The first one is used in most of the maps without big obstacles such as *slope_rocks*. This strategy just spawns the robot on random position and rotation. While, the flat ground strategy first selects suitable spawn positions by using a sliding window on the heightmap of size equal *Krock*’s footprint and check if the mean pixel value is lower than a small threshold. If so, we store the center coordinates of the patch as a candidate spawning point. Intuitively, if a patch is flat then its mean value will be close to zero. Since there may be more flat spawning positions than simulations needed, we have to reduce the size of the candidate points. To maintain the correct distribution on the map to avoid spawning the robot always in the same cloud of points, we used K-Means with k clusters where k is equal to the number of simulations we wish to run. By clustering, we guarantee to cover all region of the map removing any bias. Picture ?? shows this strategy on *bars1*.



(a) Quarry



(b) A small village.

Figure 4.10. Real world maps obtained from sensefly's dataset. The left images shows the real world location while on the right a render in 3D of the corresponding heightmaps. Both images have a maximum height of 10m.

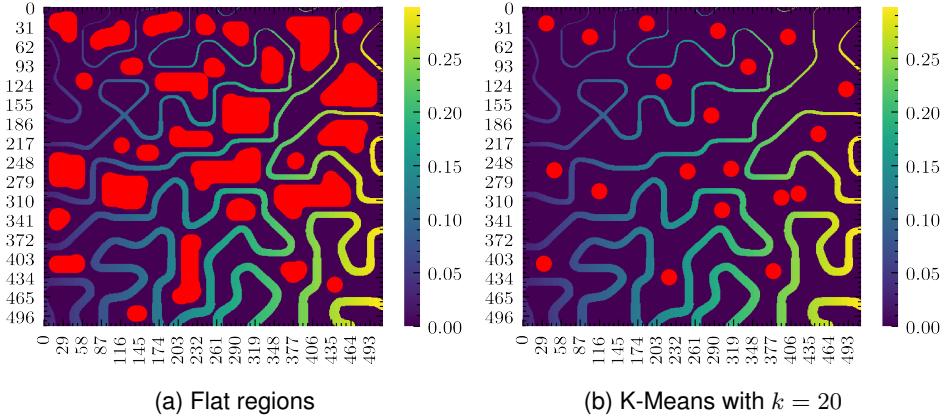


Figure 4.11. Examples of the spawning selection process (marked as red blobs) for the map *bars1*

4.3.3 Interactions

In each simulation, the robot is spawned using the two spawning strategies and it walks forward for a fixed amount of time. We selected ten or twenty seconds depending on the map. Every five simulations we completely regenerate the robot’s body to counter possible damages during the hard traversable ground. We moved the robot only forward with a fixed controller since we cannot estimate if the robot can traverse a patch while moving sideways. Table ?? shows the maps configuration used in the simulator. For some map we add three different rocky textures to create more complicated situations. For example, adding some rocks in the ramps map allowed the robot to anchor the legs to the rocks and climb them easily.

4.4 Postprocessing

To generate the dataset, we need to extract we had to extract the patches at each robot’s pose of the trajectory generated at simulation time. Figure ?? shows each step in the postprocess pipeline.

4.4.1 Parse trajectories

First, we converted the robot’s trajectories, stored as *.bag* files, to a more convenient data structure, pandas’ dataframes. Then, we cache them into *.csv* files. This operation is done only once. After, we loaded the stored dataframes with the respective terrains and start the data cleaning process. We removed the the first second of the simulation time to account for the robot spawning time. Then, we eliminated all the entries where a part of Krock was outside the edges of the map. After the data cleaning process, we converted quaternion rotation to Euler notation using the `tf` package from ROS. Then, we extracted the sin and cos from the Euler orientation last component and stored them in a column. In the simulator, the position (0, 0) correspond to the center of the map, while on the heightmap, like all images, (0, 0) is the top left corner. To later crop the the correct region from the map, we needed to convert the robot’s position into the heightmap’s coordinates. This part is also run once and it is cached to speed up latter runs.

Map	Height(m)	Spawn	Texture	Simulations	Max time(s)	Dataset
<i>bumps0</i>	2	random	-			Train
			rocks1	50	10	Train
			rocks2			Train
<i>bumps1</i>	1	random	-			Train
			rocks1	50	10	Train
			rocks2			Train
<i>bumps2</i>	1	random	-			Train
			rocks1	50	10	Train
	2		rocks2			Train
<i>bumps3</i>	1	random	-			Train
			rocks1	50	10	Train
			rocks2			Train
<i>steps1</i>	1	random	-	50	10	Train
<i>steps2</i>	1	flat	-	50	10	Train
<i>steps3</i>	1	random	-	50	10	Train
<i>rails1</i>	1	flat	-	50	20	Train
<i>rails2</i>	1		flat	-	10	Train
<i>rails3</i>	1		flat	-	10	Train
<i>bars1</i>	1	flat	-			Train
	2		-	50	10	Train
<i>bars3</i>	1	flat	-			Train
<i>ramp0</i>	1	random	rocks1	50	10	Train
			rocks2	50		Train
	3					Train
<i>ramp1</i>	4	random	-	50	10	Train
	5					Train
<i>slope_rocks1</i>	3					Train
	4	random	-	50	10	Train
	5					Train
<i>holes1</i>	1	random	-	50	10	Train
<i>quarry</i>	10	random	-	50	10	Test
<i>arc rocks</i>	1	random	-	50	10	Evaluation
Total: 1650						

Table 4.1. Maps configuration used in the simulator.

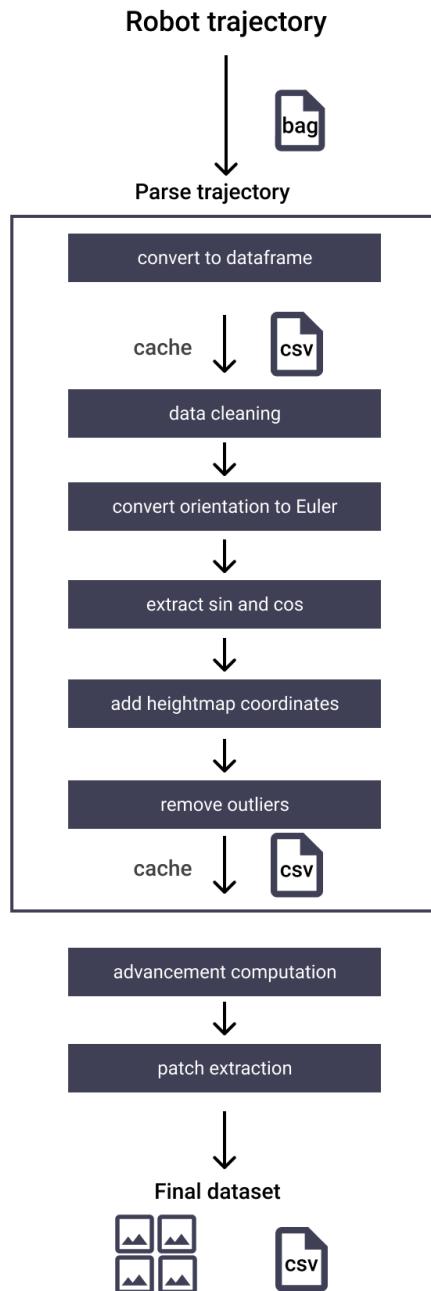


Figure 4.12. Postprocessing pipeline flow graph, starting from the top.

4.4.2 Advancement computation

To compute the robot's advancement we needed to define a time window, Δt , and consider how much the robot traveled between the current pose, $X(t)$ and the future one $X(t+\Delta t)$. We observed from the simulator than the robot moves each leg every second. Thus, we selected a time window of two seconds to allow Krock to move both legs.

4.4.3 Patch extraction

Each patch must contain both Krock's footprint, to include the situations where the obstacle is under the robot, and certain amount of ground region in front of it. Intuitively, we want to include in each patch the correct amount of future informations according to the selected time window. Thus, we should add the maximum possible ground that Krock could traverse.

To find out the correct value, we must compute the maximum advancement on a flat ground for the Δt and use it to calculate the final size of the patch. We compute it by running some simulations of *Krock* on flat ground and averaging the advancement getting a value of 70cm in our $\Delta t = 2s$.

Each patch must include Krock's footprint and the maximum possible distance it can travel is a Δt . So, since Krock's pose was stored from IMU located in the juncture between the head and the legs, we have to crop from behind its length, 85cm minus the offset between the IMU and the head, 14cm. Then, we have to take 70cm, the maximum advancement with a $\Delta t = 2s$ plus the removed offset. The following figure visualizes the patch extraction process. Lastly, we create a final dataframe containing the map coordinates, the advancement, and the patches paths for each simulation and store them to disk as .csv files.

4.4.4 Final dataset

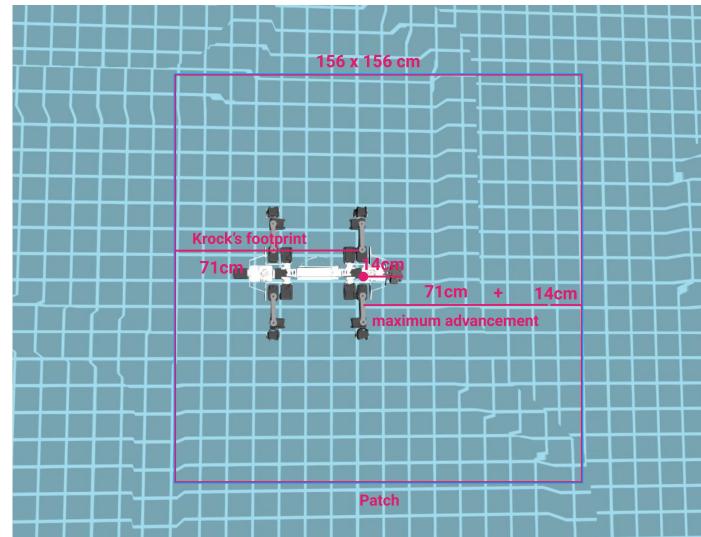
The final dataset is composed by $\approx 450K$ and $\approx 37K$ patches for the train and test set respectively. We also generated $\approx 38K$ patches using the *arc rocks* map to further evaluate the model. The whole pipeline took less than one hour to run the first time with 16 threads, and, once it is cached, less than fifteen minutes to extract all the patches. To completeness, we plotted in figure ?? the mean advancement over each map in the training set. As sanity check, we visualized some of the patches from the train set ordered by advancement, to conclude the data generation process was correct. Figure 4.15 shows a sample of forty patches.

4.5 Label patches

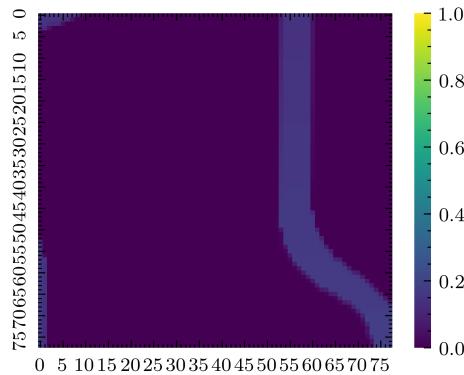
To classify the patches we first need to label them. We defined a threshold to represent and label each patch traversable if the advancement is greater or not traversable if it is lower. Ideally, the threshold should be small enough to include as less as possible false positive and big enough to cover all the cases where Krock gets stuck. We empirically compute the threshold's value by spawning Krock in front of a bump and a ramp and let it walk.

Bumps We spawned the robot on the *bumps3* map close to the end and let it go for 20 seconds. Figure 4.16 shows Krock in the simulated environment.

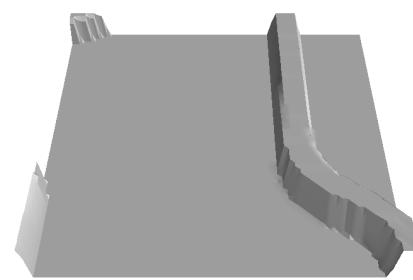
Due to its characteristic locomotion, Krock tried to overcome the obstacle using its legs to move itself to the top but it fell back producing a spiky advancement where first it is positive and then



(a) Robot in the simulator.



(b) Cropped patch in 2d.



(c) Cropped patch in 3d.

Figure 4.13. Patch extraction process for $\Delta t = 2\text{s}$.

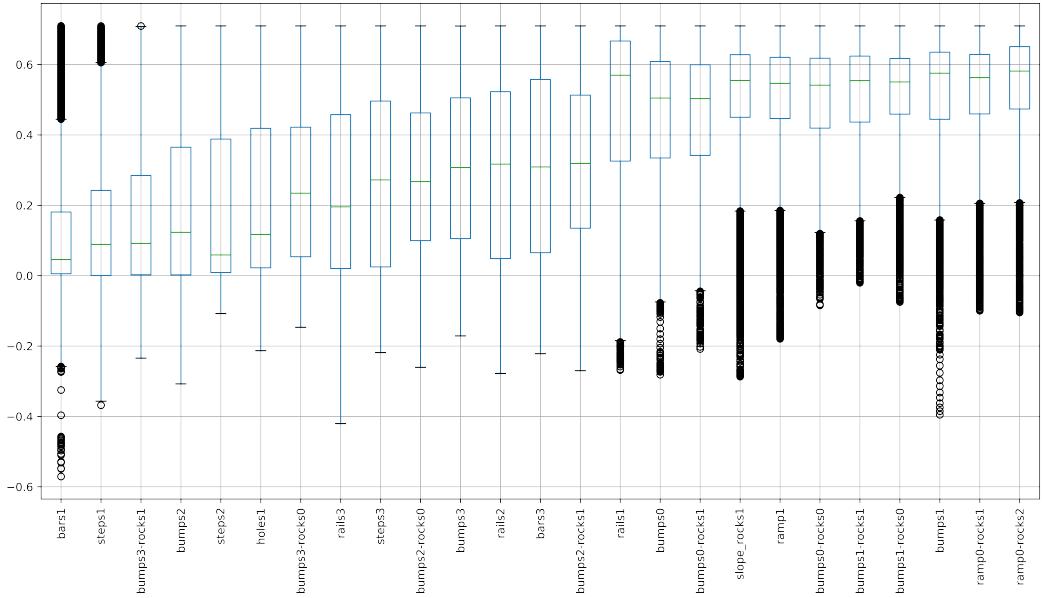


Figure 4.14. Advancement on each map with a $\Delta t = 2\text{s}$ in ascendent order.

negative. Figure 4.17 shows the advancement on this map with $\Delta t = 2$, we can notice the noisy output.

A wheel robot, on the other hand, will not produce such graph since it cannot free itself easily from an obstacle. Image a wheel robot moving forward, once it hits an obstance it stops moving, if we plot its advancement in this situation it will look more like a step than a series of spikes.

Ramps The trehsold should be small enough to not create any false positive, patches that are traversable but were classified as not. So, we let the robot walk up hill on the *slope_rocks1* map with a height scaling factor set to 5. We knew from empirical experiment that the robot is able to climb this steep ramp, the advancement is showed in figure 4.18. The mean $\approx 0.4\text{cm}$ over a $\Delta t = 2$, an impressive value considering the steepest of the surface. We used this information combined with the previous experiment to choose a threshold that is beetwen the upper bound of *bumps* and beetwen the lower bound of *slope_rocks1*. This ensured to minimize the false positive. A good value is $tr_{\Delta t=2\text{s}} = 20\text{cm}$.

4.6 Estimator

In this section we described the five convolution neural networks architecture we tested.

4.6.1 Vanilla Model

The original model proposed by Chavez-Garcia et all [5] is a CNN composed by a two 3×3 convolution layer with 5 filters; 2×2 Max-Pooling layer; 3×3 convolution layer with 5 filters; a fully connected layer with 128 outputs neurons and a fully connected layers with two neurons. Figure 4.19 visualizes the architecture.

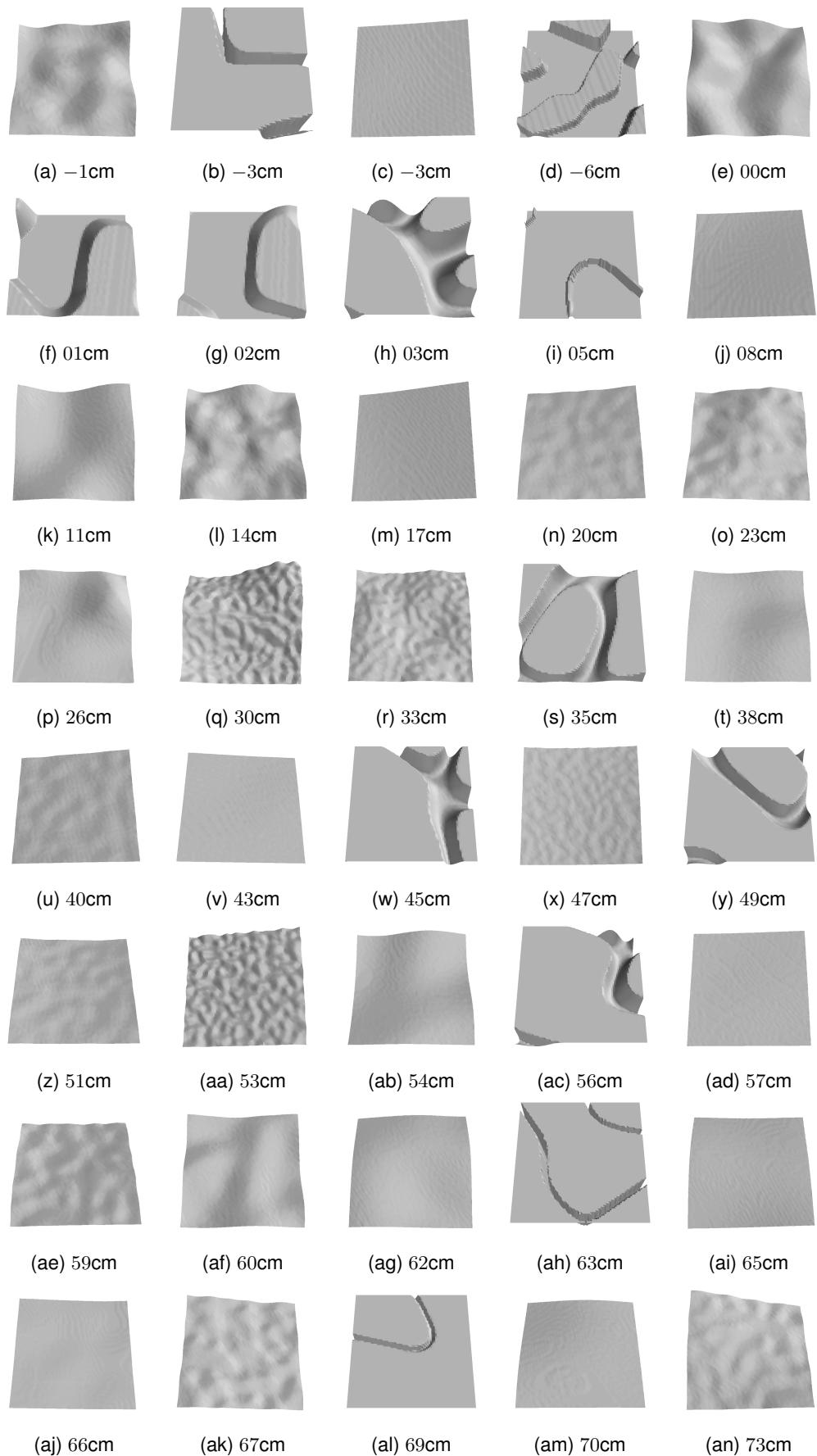


Figure 4.15. Sample of extracted patches from the train set ordered by advancement. The first patches have mostly obstacle close to the robot's head, while the latter have smoother surfaces.

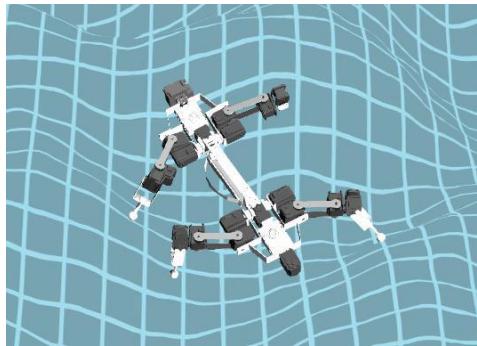


Figure 4.16. Krock tries to overcome an obstacle in the *bumps3* map.

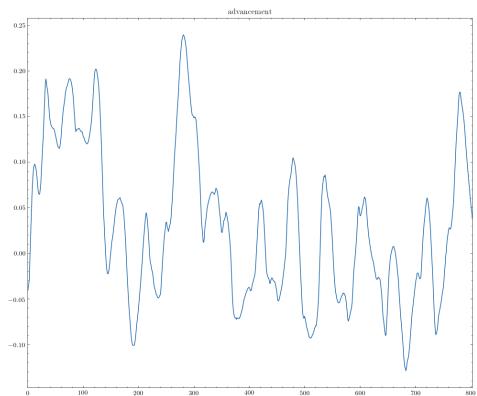


Figure 4.17. Advancement over time with $\Delta t = 2$ on *bumps3*.

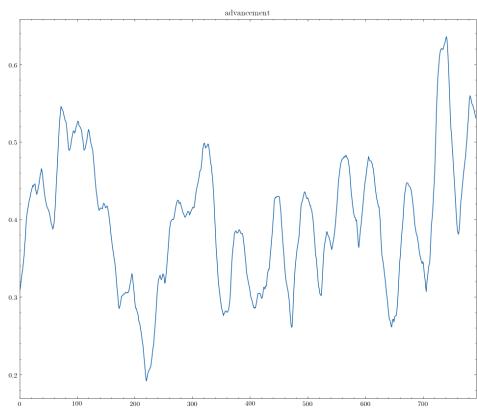


Figure 4.18. Advancement over time with $\Delta t = 2s$ on *slope_rocks3*.

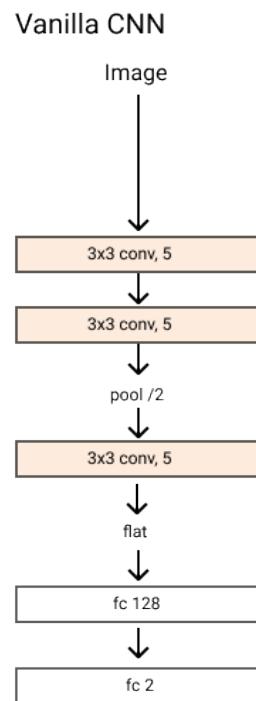


Figure 4.19. Vanilla model proposed by Chavez-Garcia et al. [5]. The rectangle represent the building block of each layer. Convolutiol layers are described by the kernel size, the number of filters and the stride.

4.6.2 ResNet

We adopt a Residual Network, ResNet [8], variant. Residual networks are deep convolutional networks consisting of many stacked Residual Units. Intuitively, the residual unit allows the input of a layer to contribute to the next layer's input by being added to the current layer's output. Due to possible different features dimension, the input must go through and identity map to make the addition possible. This allows a stronger gradient flows and mitigates the degradation problem. A Residual Unit is composed by a two 3×3 Convolution, Batchnorm [11] and a Relu blocks. Formally defined as:

$$y = \mathcal{F}(x, \{W_i\}) + h(x) \quad (4.1)$$

Where, x and y are the input and output vector of the layers considered. The function $\mathcal{F}(x, \{W_i\})$ is the residual mapping to be learned and h is the identity mapping. The next figure visualises the equation. When the input and output shapes mismatch, the *identity map* is applied to the input as

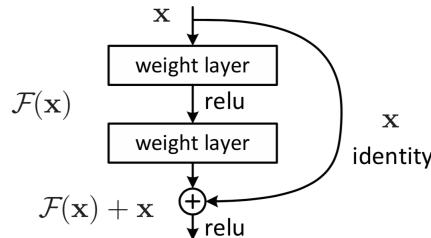


Figure 4.20. Resnet block [8]

a 3×3 Convolution with a stride of 2 to mimic the polling operator. A single block is composed by a 3×3 Convolution, Batchnorm and a Relu activation function.

4.6.3 Preactivation

Following the recent work of He et al. [9] we adopted *pre-activation* in each block. *Pre-activation* works by just reversing the order of the operations in a block.

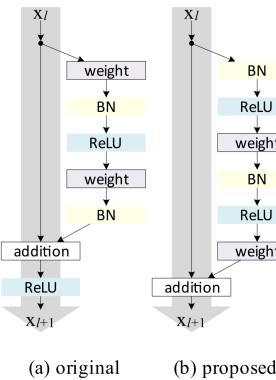


Figure 4.21. Preactivation [9]

4.6.4 Squeeze and Excitation

Finally, we also used the *Squeeze and Excitation* (SE) module [10]. It is a form of attention that weights the channel of each convolutional operation by learnable scaling factors. Formally, for a given transformation, e.g. Convolution, defined as $F_{tr} : \mathbf{X} \mapsto \mathbf{U}$, $\mathbf{X} \in \mathbb{R}^{H' \times W' \times C'}$, $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$, the SE module first squeeze the information by using average pooling, F_{sq} , then it excites them using learnable weights, F_{ex} and finally, adaptive recalibration is performed, F_{scale} . The next figure visualises the SE module.

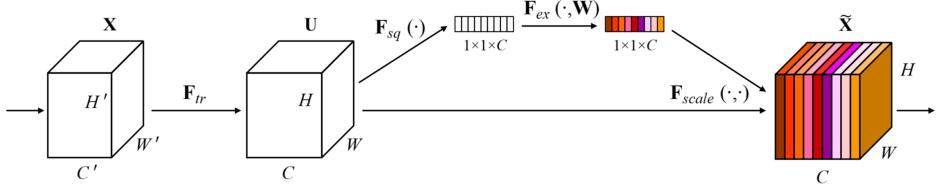


Figure 4.22. *Squeeze and Excitation* [10]

4.6.5 MicroResNet

The proposed network's architecture is composed by n ResNet blocks and channel incrementing factor of 2. Originally, ResNet was used to classify RGB images with 224×224 pixels and used an aggressive convolution and max-polling operation in the first layer to reduce the images's size. We decided to adopt a less aggressive convolution motivated by the smaller size of our inputs. We tested two kernel sized: 7×7 and 3×3 with stride of 2 and 1 respectively. Lastly, we used LeakyReLU [4] with a negative slope of 0.1 instead of ReLU to allow a better gradient flow during backpropagation. LeakyRelu is defined as follows

$$\text{LeakyRelu}(x) = \begin{cases} x & \text{if } x > 0 \\ 0.1x & \text{otherwise} \end{cases} \quad (4.2)$$

We called this model architecture *MicroResNet*. We proposed four networks layouts: with $n = 1$ and two fist layers setups, 3×3 stride = 1 and 7×7 , with and without the Squeeze and Excitation module. Those architectures can be used both for regression and classification by only chaning the output size of the fully connected later, 2 for classification and 1 for regression. All others layers are equals. Figure 4.23 add additional informations. Our models have approximately 35 times less parameters than the smalles ResNet model, ResNet18 [8], that has 11M parameters. To simplicity we used the following notation to describe each architecture variant: MicroResNet-3x3/7x7-/SE.

4.7 Normalization

Before feeding the data to the models, we need to make the patches height invariant to correctly normalize different patches taken from different maps with different height scaling factor. To do so, we subtract the height of the map corresponding *Krock*'s position from the patch to correctly center it. The following figure shows the normalization process on the patch with the square in the middle.

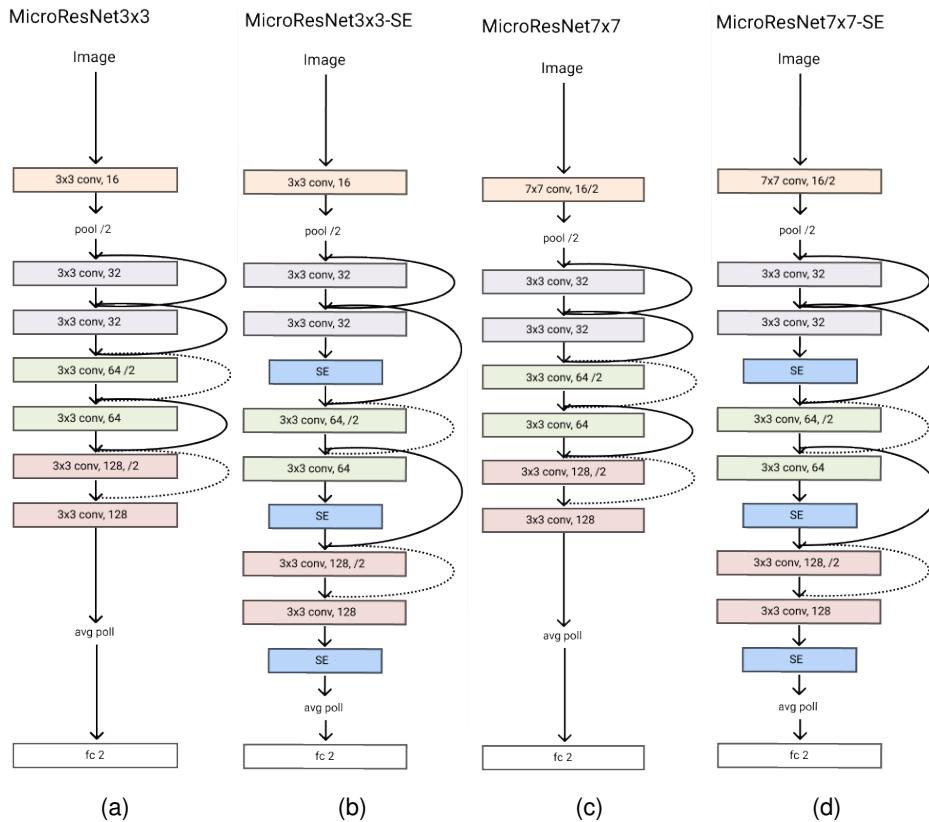


Figure 4.23. MicroResNet architectures for classification. All models have an initial features size of 16 and $n = 1$ with and without the Squeeze and Excitation module. The rectangle represent the building block of each layer. Convolutiol layers are described by the kernel size, the number of filters and the stride. Lines and dashed lines between layers represent the residual operations and the shortcut respectively.

Input	(1, 78, 78)			
	$3 \times 3, 16$ stride 1		$7 \times 7 16$ stride 2	
	2×2 max-pool			
	$3 \times 3, 16$	x 1	$3 \times 3, 32$	
Layers	SE	-	SE	-
	$3 \times 3, 32$		x 1	
	SE	-	SE	-
	$3 \times 3, 64$	x 1	$3 \times 3, 128$	
	SE	-	SE	-
	average pool, 1-d fc, softmax			
Parameters	313,642	302,610	314,28	303,250
Size (MB)	5.93	5.71	2.41	2.32

Table 4.2. MicroResNet architecture. First layer is on the top. Some architecture's blocks are equal across models, this is shows by sharing columns in the table.

??

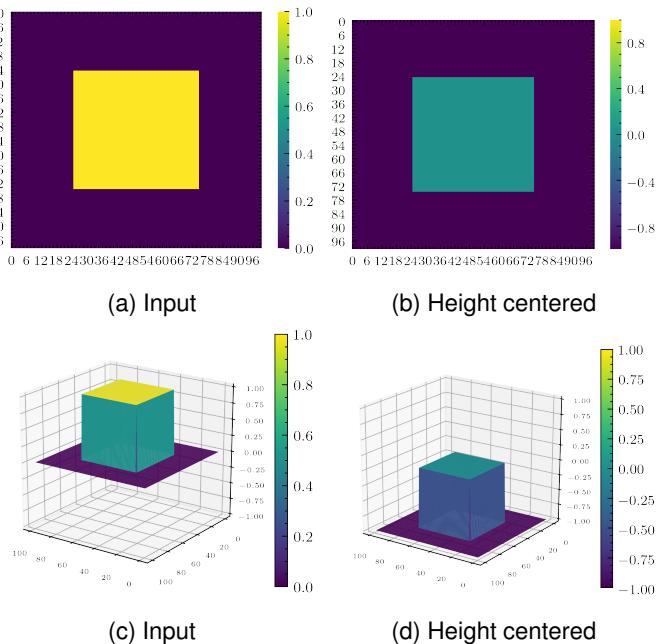


Figure 4.24. Normalization process. Each patch is normalized by subtracting the height value corresponding to the robot position to center its height.

4.8 Data Augmentation

Data augmentation is used to change the input of a model in order to produce more training examples. Since our inputs are heightmaps we cannot utilize the classic image manipulations such as shifts, flips, and zooms. Imagine that we have a patch with a wall in front of it, if we random rotate the image the wall may go in a position where the wall is not facing the robot anymore, making the image now traversable with a wrong target. We decided to apply dropout, coarse dropout, and random simplex noise since they are traversability invariant. To illustrate those techniques we are going to use the same square patch showed before 4.24.

4.8.1 Dropout

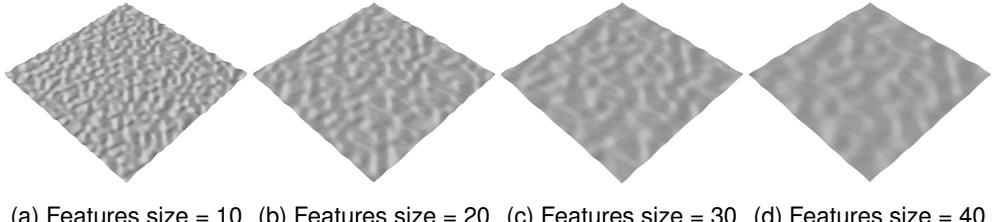
Dropout is a technique to randomly set some pixels to zero, in our case we flat some random pixel in the patch.

4.8.2 Coarse Dropout

Similar to dropout, it sets to zero random regions of pixels with defined boundaries.

4.8.3 Simplex Noise

Simplex Noise is a form of Perlin noise that is mostly used in ground generation. Our idea is to add some noise to make the network generalize better since lots of training maps have only obstacles in flat ground. Since it is computationally expensive, we randomly fist apply the noise to five hundred images with only zeros. Then, we randomly scaled them and add to the input image.



(a) Features size = 10 (b) Features size = 20 (c) Features size = 30 (d) Features size = 40

Figure 4.25. Simplex noise on flat ground. The feature size corespond to the amount of noise we introduce in the surface. A lower value produces noisy grounds, while a bigger one smoother terrains.

Figure ?? shows the tree data augmentation techniques used applied the input image. To ensure we do create untraversable patches from traversable patch we limited the coarse dropout region to only few centiments and we apply a smoother simplex noise on traversable patches. In detail, in all the traning epochs, we augmented 80% of the input images with dropout and coarse dropout. Dropout has a probability between 0.05 and 0.1 and coarse dropout has a probability of 0.02 and 0.1 with a size of the lower resolution image from which to sample the dropout between 0.6 and 0.8. Simplex noise was applied on the 70% of the training data samples with a feature size between 1 and 50 with a random scaling factor between 15–25 and 5–15 from traversable and not traversable patches respectively.

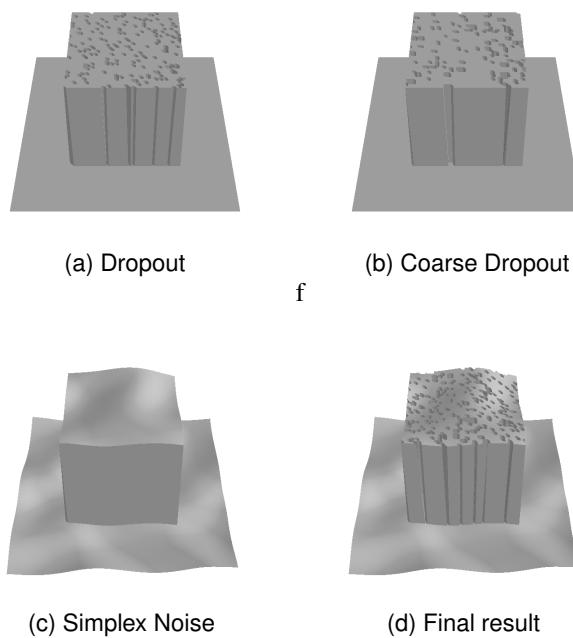


Figure 4.26. Data augmentation applied on a patch. First we randomly set to flat some small regions of the ground. Then, we slightly mutate the surface using simplex noise.

4.9 Tools

We quickly list the most important tools and libraries adopted in this project. The framework was entirely developed on Ubuntu 18.10 with Python 3.6.

4.9.1 Software for simulation

ROS Melodic The Robot Operating System (ROS) [18] is a flexible framework for writing robot software. It is *de facto* the industry and research standard framework for robotics due to its simple yet effective interface that facilitates the task of creating a robust and complex robot behavior regardless of the platforms. ROS works by establishing a peer-to-peer connection where each *node* is to communicate between the others by exposing sockets endpoints, called *topics*, to stream data or send *messages*. Each *node* can subscribe to a *topic* to receive or publish new messages. In our case, *Krock* exposes different topics on which we can subscribe in order to get real-time information about the state of the robot. Unfortunately, ROS does not natively support Python3, so we had to compile it by hand. Because it was a difficult and time-consuming operation, we decided to share the ready-to-go binaries as a docker image.

4.9.2 Data processing

Numpy Numpy is a fundamental package for any scientific use. It allows to express efficiently any matrix operation using its broadcasting functions. Numpy is used across the whole pipeline to manipulate matrices.

Pandas To process the data from the simulations we rely on Pandas, a Python library providing fast, flexible, and expressive data structures in a tabular form. Pandas is well suited for many different kinds of data such as handle tabular data with heterogeneously-typed columns, similar to SQL table or Excel spreadsheet, time series and matrices. We take advantages of the relational data structure to perform custom manipulation on the rows by removing the outliers and computing the advancement.

Generally, pandas does not scale well and it is mostly used to handle small dataset while relegating big data to other frameworks such as Spark or Hadoop. We used Pandas to store the results from the simulator and inside a Thread Queue to parse each .csv file efficiently.

OpenCV Open Source Computer Vision Library, OpenCV, is an open source computer vision library with a rich collection of highly optimized algorithms. It includes classic and state-of-the-art computer vision and machine learning methods applied in a wide array of tasks, such as object detection and face recognition. We adopt this library to handle image data, mostly to pre and post-process the heatmaps and the patches.

CNN training

Pytorch PyTorch is a Python open source deep learning framework. It allows Tensor computation (like NumPy) with strong GPU acceleration and Deep neural networks built on a tape-based auto grad system. Due to its Python-first philosophy, it is easy to use, expressive and predictable it is widely used among researchers and enthusiasts. Moreover, its main advantages over other

mainstream frameworks such as TensorFlow [2] are a cleaner API structure, better debugging, code shareability and an enormous number of high-quality third-party packages.

FastAI FastAI is library based on PyTorch that simplifies fast and accurate neural nets training using modern best practices. It provides a high-level API to train, evaluate and test deep learning models on any type of dataset. We used it to train, test, and evaluate our models.

imgaug Image augmentation (imgaug) is a python library to perform image augmenting operations on images. It provides a variety of methodologies, such as affine transformations, perspective transformations, contrast changes and Gaussian noise, to build sophisticated pipelines. It supports images, heatmaps, segmentation maps, masks, key points/landmarks, bounding boxes, polygons, and line strings. We used it to augment the heatmap, details are in section 4.8

4.9.3 Visualization

matplotlib Almost all the plots in this report were created by Matplotlib, is a Python 2D plotting library. It provides a similar functional interface to MATLAB and a deep ability to customize every region of the figure. All It is worth citing *seaborn* a data visualization library that we inglobate in our work-flow to create the heatmaps. It is based on Matplotlib and it provides an high-level interface.

mayavi Mayavi is a scientific data visualization and plotting in python library. We adopt it to rendered in 3D all the surfaces used in our framework.

Chapter 5

Results

In this section, we evaluate five convolutional neural networks: the model proposed by Chavez-Garcia et all [5] and the four MicroResnet presented in the last chapter. We will select the best performing one based on the classification score on the test set. We will show the results of the same architecture trained on regression instead of classification and prove the latter yields better results. Finally, we qualitatively evaluate the best model by predicting traversability in real-world terrains.

5.1 Hardware

We run all the experiment on a Ubuntu 18.10 work station equipped with a Ryzen 2700x, a powerful CPU with 8 cores and 16 threads, and a NVIDIA 1080 GPU with 8GB of dedicated RAM.

5.2 Experimental validation

We selected as *validation* set ten percent of the training data. Since we stored each run of Krock as a *.csv* file, validation and train set do not overlap.

5.2.1 Metrics

Classification: To evaluate the model's classification performance we used two metrics: *accuracy* and *AUC-ROC Curve*. Accuracy scores the number of correct predictions made by the network while the AUC-ROC Curve represents degree or measure of separability, informally it tells how much model is capable of distinguishing between classes. For each experiment, we selected the model with the higher AUC-ROC Curve during training to be evaluated on the test set.

Regression: We used the Mean Square Error to evaluate the model's performance.

5.3 Training setup

Initially, to train the models we first used Standard Gradient Descent with momentum set to 0.95, weight decay to $1e - 4$ and an initial learning rate of $1e - 3$ as was originally proposed in He

et al. [8]. However, we later switched Leslie Smith’s 1cycle policy [21] that allowed us to train the network faster and with higher accuracy. The average training time is ≈ 10 minutes. We minimized the Binary Cross Entropy (BCE) for the classifier and the Mean Square Error (MSE) for the regressor.

5.4 Quantitative results

5.4.1 Model selection

We compared the four different MicroRenset architectures ?? and the Vanilla CNN 4.6.1 proposed in the original work [5] on the test set using as metric the ROC AUC. Ideally, the second MicroResnet model should be able to extrapolate more features from the patch thanks to the big kernel size. We evaluated those models using a time window of two seconds, a threshold of 20cm and we applied data augmentation as we described in section 4.8. We run five experiments per architecture and we selected the best performing network. Table 5.1 shows the results for each architecture. The best performing model was MicroResNet7x7-SE that obtained a ROC AUC of 0.896 on the test set. Moreover, the Squeeze and Excitation operator improved the performance of both MicroResNet3x3 and MicroResNet7x7 but only latter gained a huge boost of 0.021.

Model	ROC AUC		Params
	Mean	Top	
Vanilla	0.892	0.890	974,351
MicroResnet3x3	0.881	0.883	302,610
MicroResnet3x3-SE	0.883	0.888	313,64
MicroResnet7x7	0.867	0.875	303,250
MicroResnet7x7-SE	0.888	0.896	314,28

Table 5.1. Comparison of the four MicroResnet architectures 4.23 and the original convolutional neural network (Vanilla CNN) proposed in the original work [5] on the test set using as metrics the ROC AUC. The best performing model is MicroResNet7x7-SE.

5.5 Final results

5.5.1 Classification

Table 5.2 table shows in deep the score of the best network for each dataset. We obtained 95% accuracy and 0.96 AUC on the validation set and 88% accuracy and 0.89 AUC on the test set.

5.5.2 Regression

We regressed on the advancement using the best model selected previously, MicroResNet7x7-SE. Table 5.3 table shows in deep the score of the best network for each dataset. The regressor can be used to classify patches using different thresold by just see if the predicted advancement is greater or lower than a threshold. However, we always need to fix a threshold to estimate traversabilit.

Dataset			MicroResNet		Size(m)	Resolution(cm/px)
Type	Name	Samples	ACC	AUC		
Synthetic	Training	429312	-	-	10 × 10	2
	Validation	44032	95.2 %	0.961	10 × 10	2
	Arc Rocks	37273	85.5 %	0.888	10 × 10	2
Real evaluation	Quarry	36224	88.2 %	0.896	32 × 32	2

Table 5.2. Classification results on different datasets.

Dataset			MicroResNet		Size	Resolution(cm/px)
Type	Name	Samples	MSE			
Synthetic	Training	-	-	-	10 × 10	2
	Validation	44032	0.006	0.006	10 × 10	2
	Arc Rocks	37273	0.020	0.020	10 × 10	2
Real evaluation	Quarry	36224	0.022	0.022	32 × 32	2

Table 5.3. Classification results on different datasets.

Once the value is fixed, the same model trained on the same dataset to classify patches drastically outperforms the regressor. Table 5.4 shows the accuracy of two MicroResNet7x7-SE on classification and regression after a threshold of twenty centimeters is selected. For this reason, we choose the MicroResNet7x7-SE trained on the classification task to be evaluated.

	Regression	Classification
	ACC	
Top	72.8%	88.2%
Mean	73.6%	87.8%

Table 5.4. Regression and classification accuracy for the same model, MicroResNet7x7-SE, once a threshold of 20cm is fixed. The classifier outperforms the regressor by 15%.

5.6 Qualitative results

We qualitatively evaluated the classifier predictions by plotting the traversability probability on different maps in 3D. We used a sliding window to extract the patches from the tested maps to collect the model's predictions. Then, we created a texture based on the traversable probability, we have colored it using a colormap and finally applied on the 3D model of each map. We evaluated four rotations, from bottom to top, top to bottom, left to right and right to left since those are the most human understandable. By using those angles, we can optimize the cropping process since to avoid rotate each patch based on the Krock's head position. Instead, we can just rotate the entire map beforehand.

5.6.1 Quarry

The first map we evaluate is Quarry, this map is 32×32 m long and has a maximum height of 10m. We can use some of the terrain interesting features, such as are three bis slopes and the rocky ground on the top, to evaluate the model. For instance, we expect the trail on the slopes to be traversable at almost any rotation, especially from left to right and vice-versa. The top part should be hard to traverse in almost any case. Figure 5.1 shows the traversability probability directly on the map.

some where place the colormap bar

Correctly, the lower part of the map, composed by flat regions, was labeled with high confidence as traversable in all rotations. On the other hand, the traversability of the slopes and the bumps on the top region depends on the robot orientation.

5.6.2 Bars

Bars is a map composed of different heights wall, thus we expected similar probabilities with different orientation. Figure ?? shows the traversability probabilities on the terrain.

Definitely, this was a hard map for the robot due to the high number of not traversable walls. Interesting, we can identify a tunnel near the bottom center of the maps that shows how the model correctly label those patches depending on the orientation. Figure 5.3 highlights this detail.

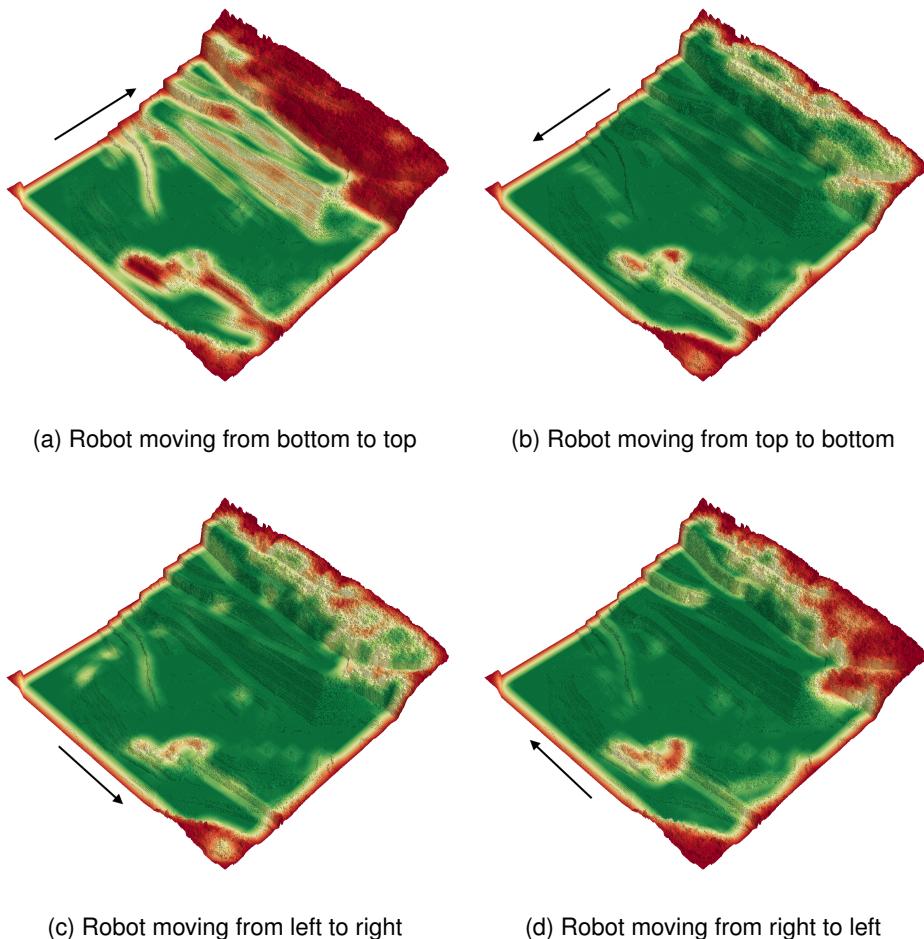


Figure 5.1. Traversability probability on the Quarry map, $32 \times 32\text{m}$, for different Krock's rotation. The values are obtained by sliding a window on the map to create the patches to predict their traversability.

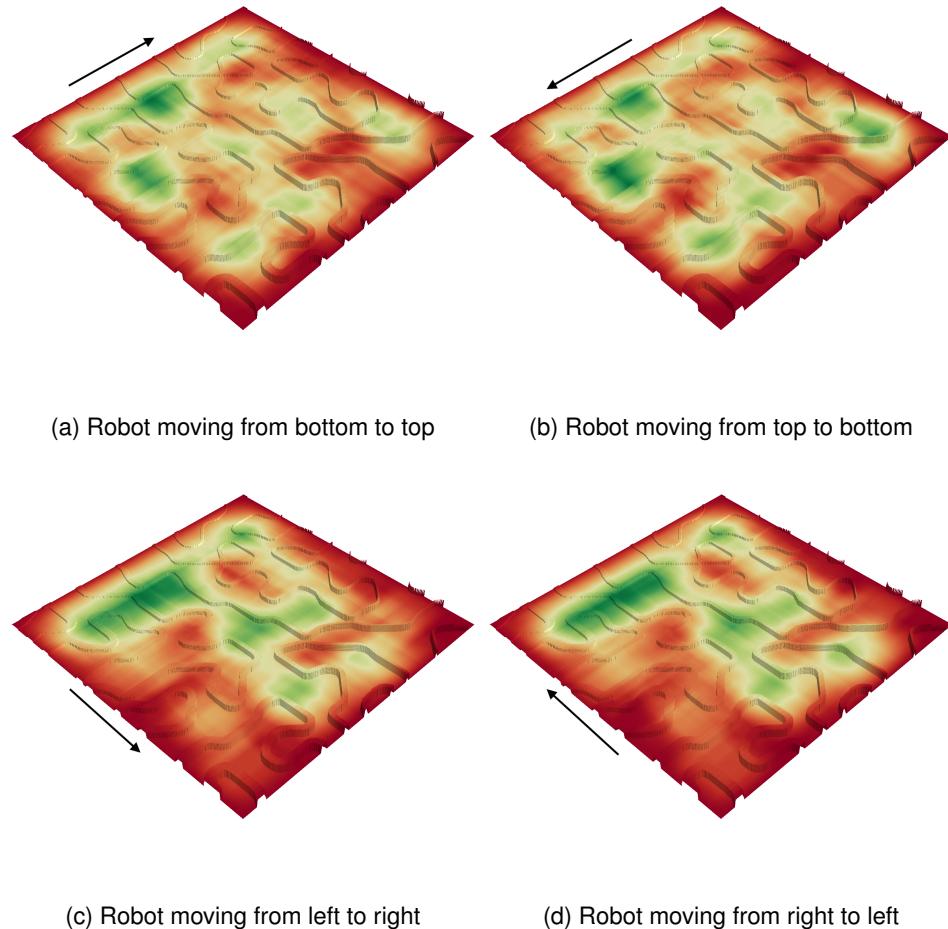


Figure 5.2. Traversability probability on the bars map, a $10 \times 10\text{m}$, for different Krock's rotation. The values are obtained by sliding a window on the map to create the patches to predict their traversability.

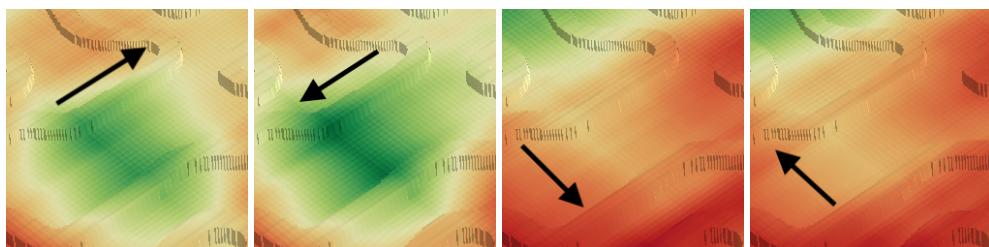


Figure 5.3. Detail of a region in the bars map where there are two walls forming a tunnel. Correctly, when the robot is travel following the trail the region is label as traversable.

5.6.3 Small village

We applied the same procedure to also a $10 \times 10\text{m}$ small village map. Figure 5.4 describes its traversability for four different rotations.

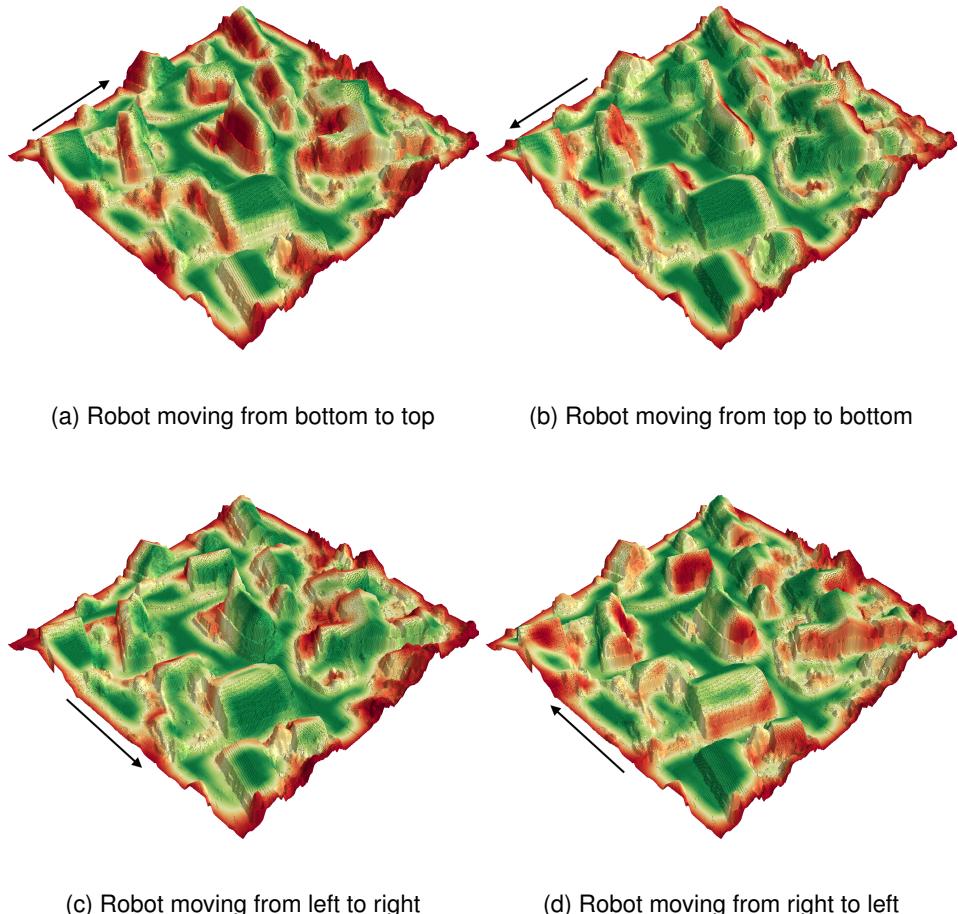


Figure 5.4. Traversability probability on the map of a small village for different Krock's rotation. The surface covers $30 \times 30\text{m}$ and has a maximum height of 10m. The values are obtained by sliding a window on the map to create the patches to predict their traversability.

All the streets were labeled as traversable, while the church's roof traversability depends on the orientation. For example, most steep roofs are not traversable when walking uphill. On the other hand, if Krock walks side by side they can be traversed. Figure 5.5 shows this behavior on the church's roof.

This chapter is under development!

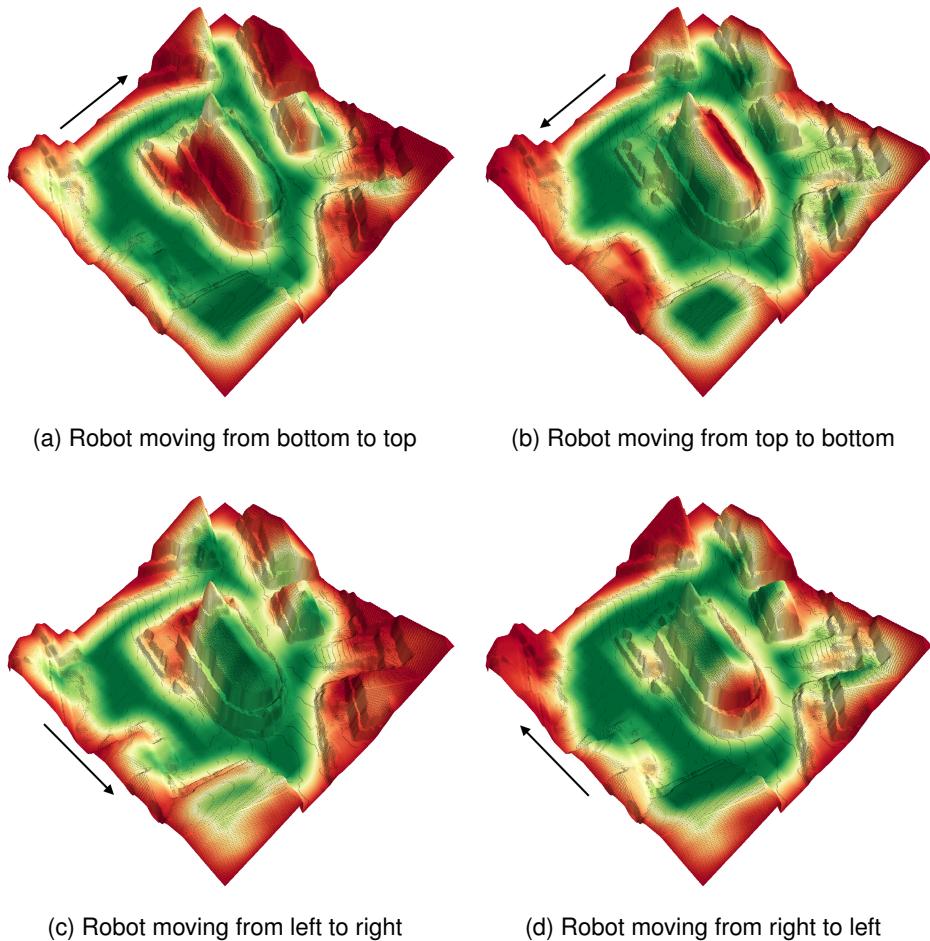


Figure 5.5. Detail of the church in the small village map for different Krock's rotation. All the fours images are correctly labeled accordingly to the robot orientation. In the first two images only downhill part of the roof is traversable. While in the last two the robot is able to travel till the end.

Chapter 6

Interpretability

Understanding the model's strength, robustness and, limitations is crucial to gain a better understanding of its outputs. This is even more important when working with controllers that can be deployed in real scenarios. In this section, we evaluated the quality of our traversability estimator with different methodology. First, we showed that the model has correctly learned grounds' features and was able to separable terrains based on them. Second, we utilized the challenging Quarry dataset to visualize the most traversable, the least traversable and the misclassified patches. Utilizing a special method, we determined that the model always looked at the correct features in the ground even if when it fails. Lastly, we crafted patches with different unique features, such as walls, bumps, etc, to test the robustness of the model by comparing its outputs to the real data gathered from the simulator.

6.0.1 Features separability

In general, convolutional neural networks learn to encode images by applying filters of increasing size at each layer. Usually, the first layers learn basic features, such as edges, while the final one encodes complex shapes. The final convolutional layer's outputs are usually referred to as features space, consequently, a feature vector is just the output of the last layer for a given image. Those last features are combined and mapped to the correct classes by one or more fully connected. The following image help to visualize the different features learned at each layer. It was generated by plotting the learned features for different categories at different layers by Lee et al. [14]. So, a correctly trained network should be able to separate those features based on predicted classes. Intuitively, given two classes \mathcal{A} and \mathcal{B} , for example, *cat* and *dog*, the high-level features for each class should not be the same, otherwise, the model may misclassify the input due to the overlap of different classes' features. One technique to discover the degree of separability of our network is to directly visualize the inputs features vectors for each class. Unfortunately, like most models, our network, MicroResNet, has a high dimensional feature space. Each patch is mapped to a $[128, 3, 3]$ vector. We cannot directly visualize a 128 dimension space, we reduced each feature vector dimension to a two-dimension by applying Principle Analysis Component (PCA) [20] to visualize it. We investigate the features space of the model booth in the train and test set.

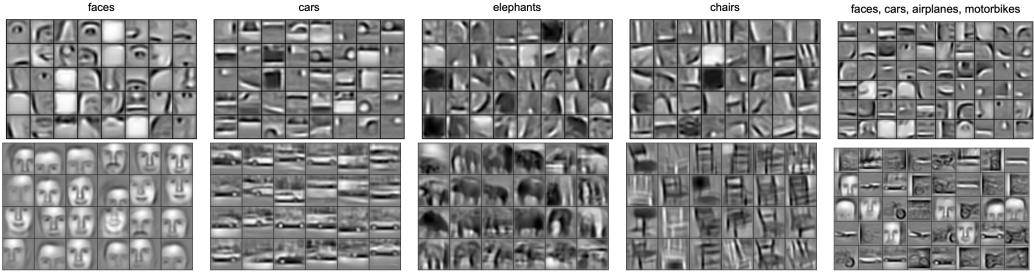


Figure 6.1. Figure from Lee et al. [14] paper where they showed for different classes the low-level features (up) and the high-level features (down) learned by a convolution neural network.

6.0.2 Train set

The following figures shows the features of 11K images sampled from the train set labeled with their classes, *traversable* and *not traversable*. We can clearly recognize two main clusters based on the labels' color, one on the left and one of the right. Those points are easily separable, even by human eyes, meaning that the model was able to learn meaning features from the dataset. To be sure the center of each class' point cloud is not overlapping we plotted the density of each cluster. Clearly, there is some distance between the centers. Furthermore, we can directly plot the patch corresponding to each feature vector to identify clusters of patches based on their position. Intuitively, if similar inputs are close to each other in the features space then the model also learned to effectively encode terrains features. We decided to not show all images on the same plot to avoid overcrowding the image. Instead, we clustered the points using K-Means with $k = 200$ clusters and then we took the patch that corresponded to the center point in each cluster. In this way, even if we are showing only a few inputs, we included all the meaningful features. The following image shows the result. Definitely, patches with similar features are close to each other yielding a quality features encoding. On the left-top side, we can distinguish highly untraversable patches with walls/bumps in front of the robot. Going down, we encounter patches with smaller obstacles. On the plateau, there are traversable patches with small obstacles such as light bumps. Importantly, those patches are the closest ones to the not traversable ones, so they were the hardest to separate, thus, to classify. Going up on the right side, we see some grounds with small steps. Finally, on the top, we find all the downhill patches, the simplest ones to traverse.

6.0.3 Test set

We can apply the same procedure on the test set. Since it is a real world quarry, this dataset is harder than the train set and present challenging situations for the robot. The following image shows the features space after reducing its dimension to two using PCA. Interesting, the traversable patches are very near to each other, while the others span a very big surface. This suggests that there are many not traversable terrains with different features. The traversable points are clustered near the center, this implies that most of them share similar features. We plotted the density for each class to better understand where the most points are mapped. The two centers are really close to each other, making those samples harder to separate and some not traversable points are mixed up with the traversable ones. This explains the elevated number of false negative that lower down the AUC score on this dataset. We can also visualize the patches by plotting them using their features

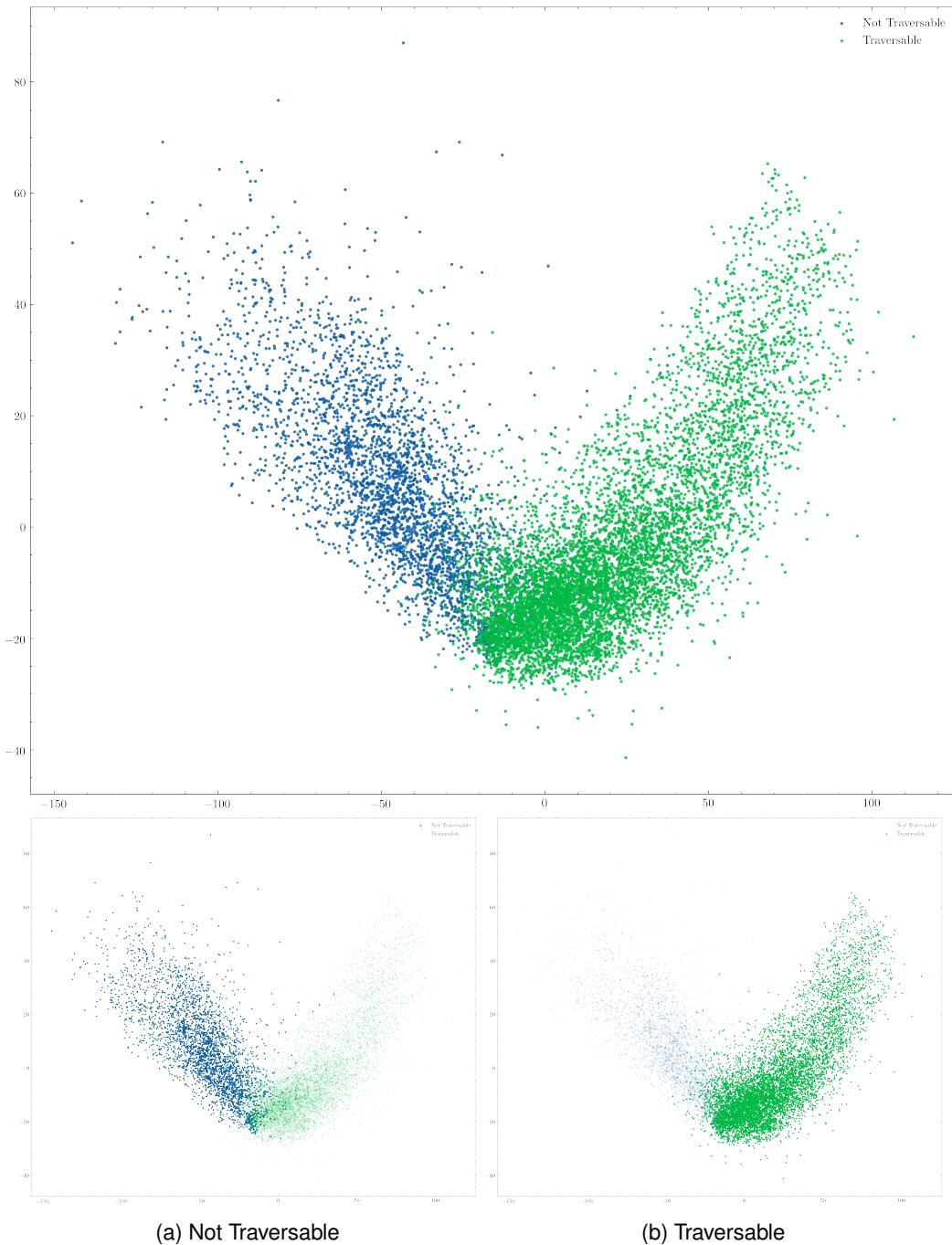


Figure 6.2. Principal Component Analysis on the features space computed using the features from the last convolutional layers on the training dataset.

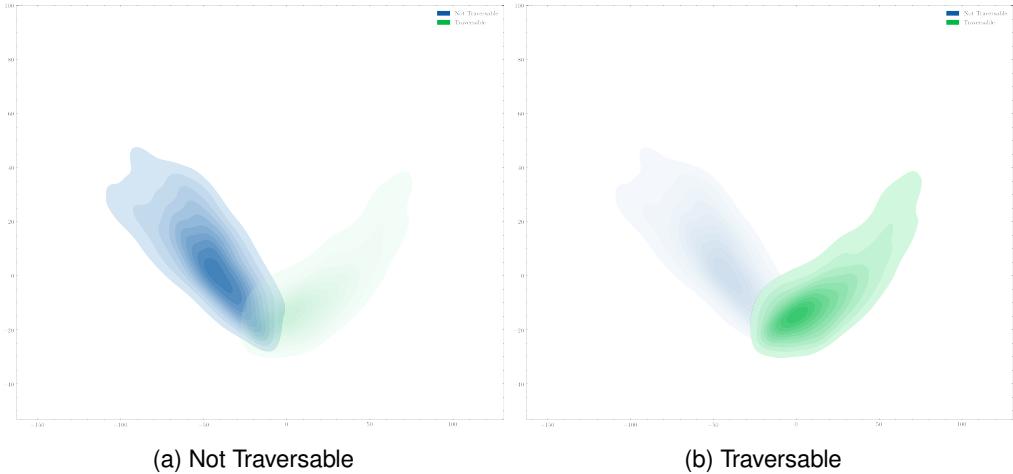


Figure 6.3. Density plot for the points sampled from the training dataset in the features space. The centers of the cluster are not overlapping yielding a good separability and correct learning.

coordinates On the top left, from the not traversable cloud, we can see patches with a high level of bumps. Going down we find surfaces with huge walls in front of the robot while going close to the center we start to see all the traversable patches. Those samples have not too steep slopes. If we move to the density center, green double shown in figure ??, we encounter lots of flat patches with little obstacles. Going up on the right branch we find downhill and on the top there are falls.

In the following section, we will take a deep look at the test set to find which patches confuse the model. Most probably, those samples will be located between the two clusters center where the difference between classes' features is minimum.

6.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [19] is a technique to produce visual explanations for convolutional neural networks. It highlights the regions of the input image that contribute the most to the predictions. In detail, the output with respect to a target class is back-propagated while storing the gradient and the output in the last convolution. Then, a global average is applied to the saved gradient keeping the channel dimension in order to get a 1-d tensor, this will represent the importance of each channel in the target convolutional layer. After, each element of the convolutional layer outputs is multiplied with the averaged gradients to create the grad cam. This whole procedure is fast and it is architecture independent.

6.2 Quarry dataset

After showing the model's capability of correctly separate classes' features we utilized Grad-CAM to visualize some patches inside the quarry dataset. The aim of this section is to show how the model looks at meaningful features in each input to make the prediction even when it fails. For instance, imagine we feed to the model a not traversable patch with an obstacle and the network

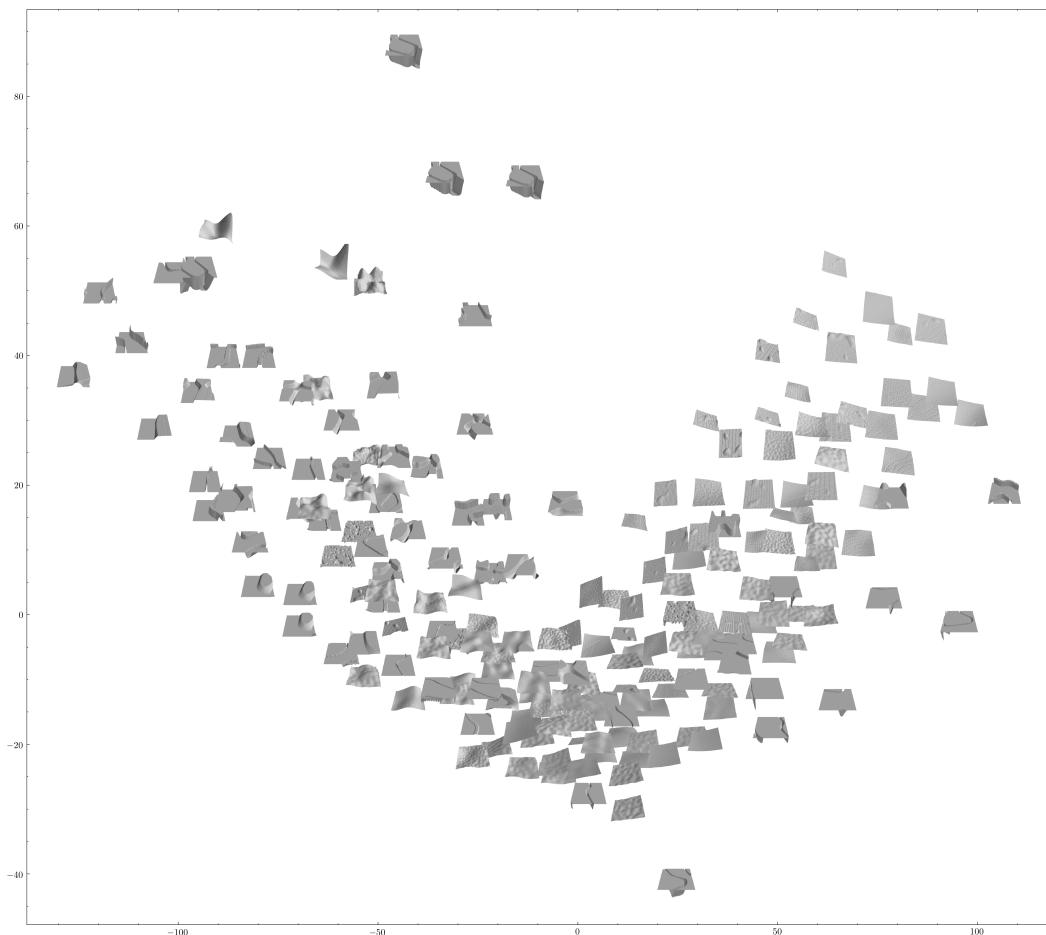


Figure 6.4. Patches plotted using the coordinate of the features vector obtained from the last convolutional layer's output and then reduced using PCA to a two-dimensional vector. Similar grounds are close to each other.

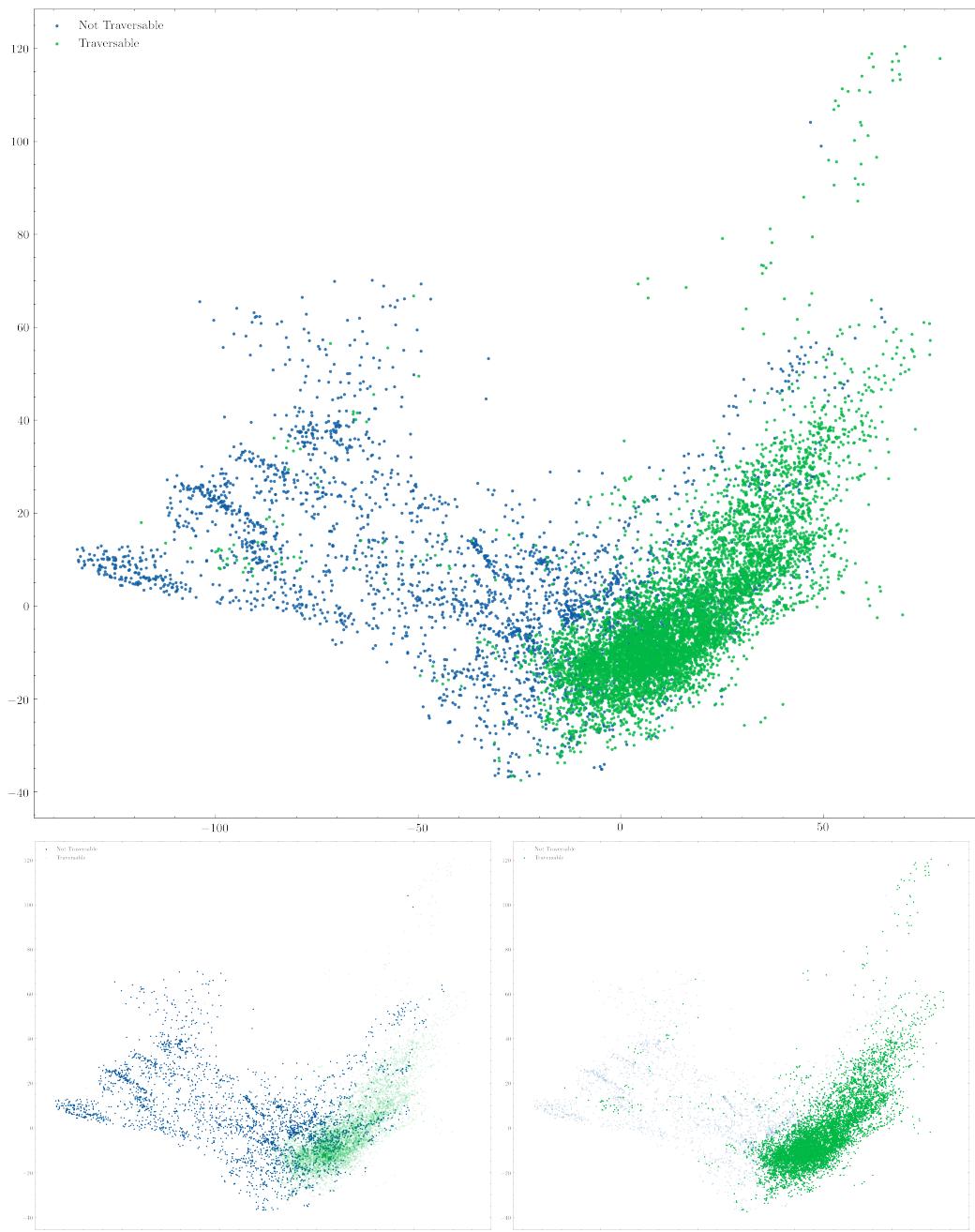


Figure 6.5. TODO

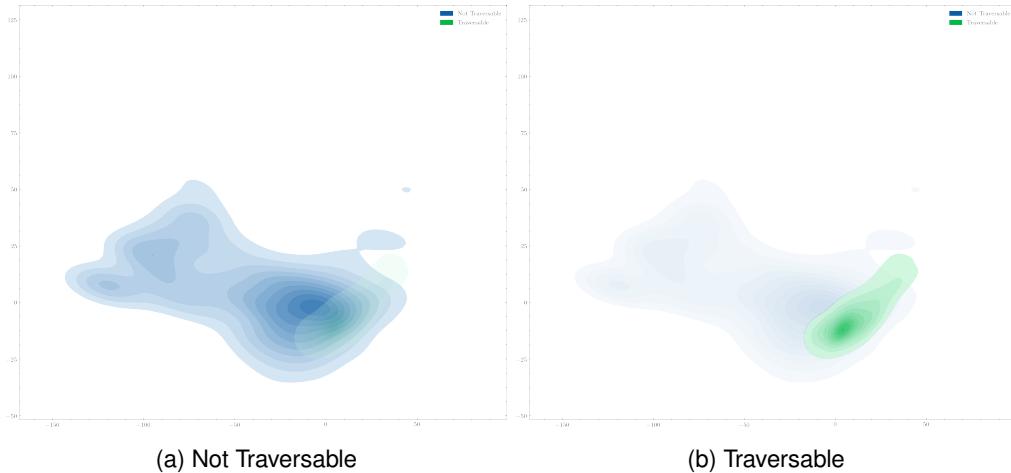


Figure 6.6. Density plot for the points sampled from the test dataset in the features space. The centers of the cluster are not overlapping yielding a good separability and correct learning.

label is as traversable. Clearly, the output is wrong but the model’s error may be caused by two different situations. The first one, the model could just have ignored the obstacle and looked away, meaning that it was not even able to understand there was an obstacle in the first place. Second, the network could have correctly look at the obstacle but thought that obstacle is traversable, showing the correct ability to find and use important features in the map. In the following sections, we showed that, even when the predictions are wrong, our model always look at the most important features of each input to determine its traversability.

We divided the dataset patches into four classes based on the model’s performance: worst, best, false positive and false negative. Then, we took twenty inputs from those sets and applied Grad-CAM to generate a texture that we placed on 3D render to better visualize which region of the inputs caused the prediction.

6.2.1 Best

The best patches are the inputs labeled and predicted traversable. We plotted twenty different images sampled uniformly according to the robot advancement to include as many interesting situations as possible. We recognized two main clusters of images: flat and slopes. The models is mostly interested in the left part of the patches with uneven ground on the left region, 6.9a, 6.9c, 6.9d, 6.9e, 6.9p, 6.9q, 6.9r, 6.9s. This is due to the fact that an obstacle in that reason will block the rear legs and prevent the robot to advance.

Moreover, in other patches 6.9b, 6.9i, 6.9j, 6.9t, the model's attention is ahead of the robot. In those situations, the robot is able to properly move at the beginning so the network must evaluate the possibility of obstacles ahead. There are two obvious cases, 6.9i and 6.9t. The first one is a totally flat surface, so the model will look as far as possible to check if there are obstacles. Similarly, in the second, a surface with a bump in the end, the network verifies that spot. So, rightly, the network analysis the first region of the patch that may contain an untraversable obstacle.

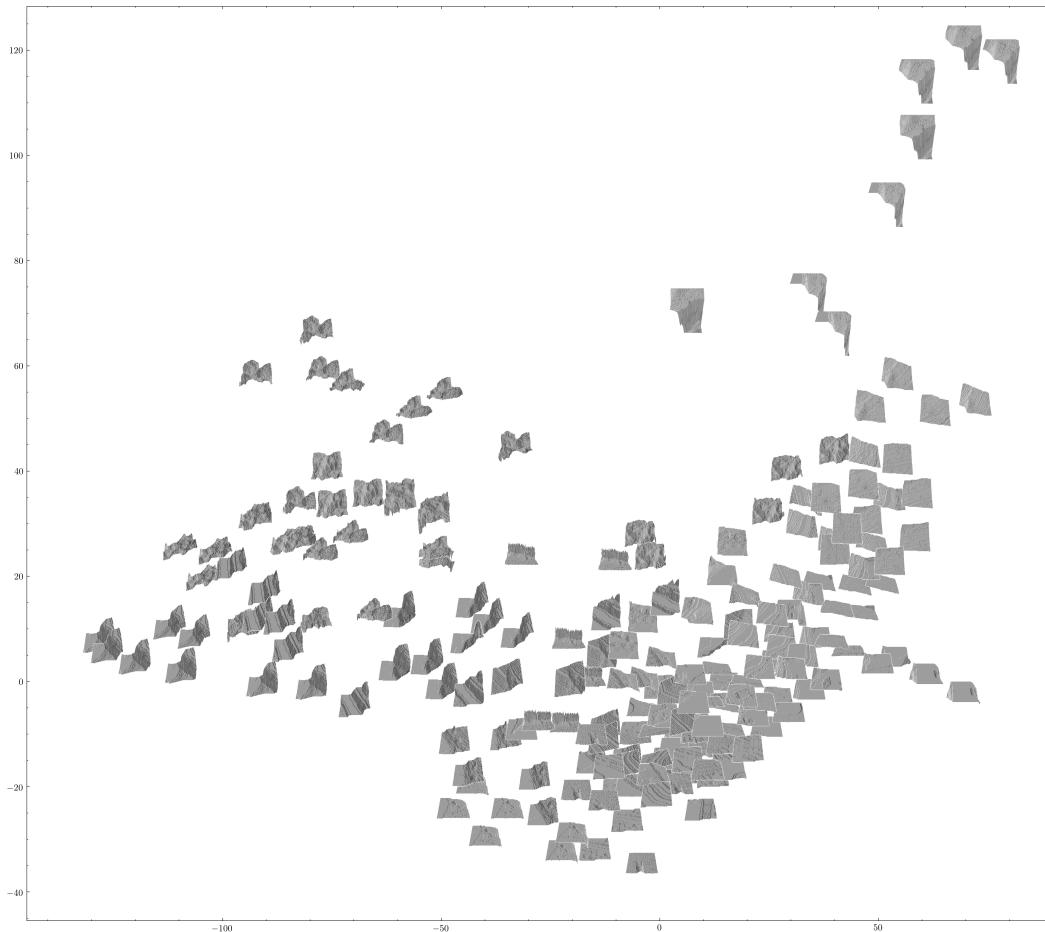


Figure 6.7. Patches that correspond to coordinates in the features space of the last convolutional layers on the test dataset. Similar grounds are close to each other.

??

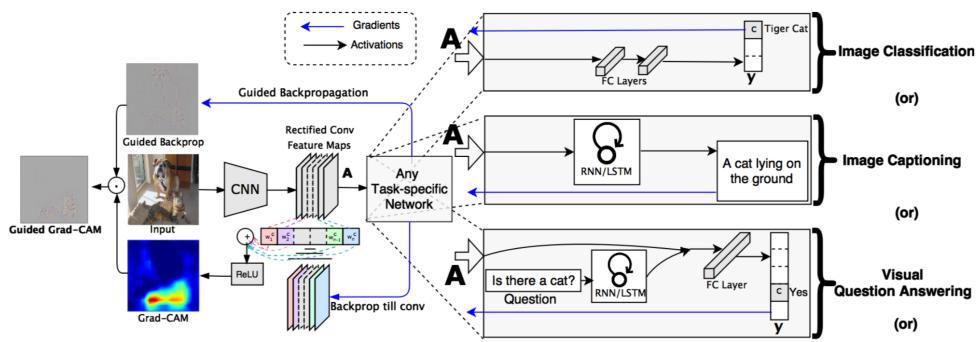


Figure 6.8. Grad-CAM procedure on an input image. Image from the original paper [19].

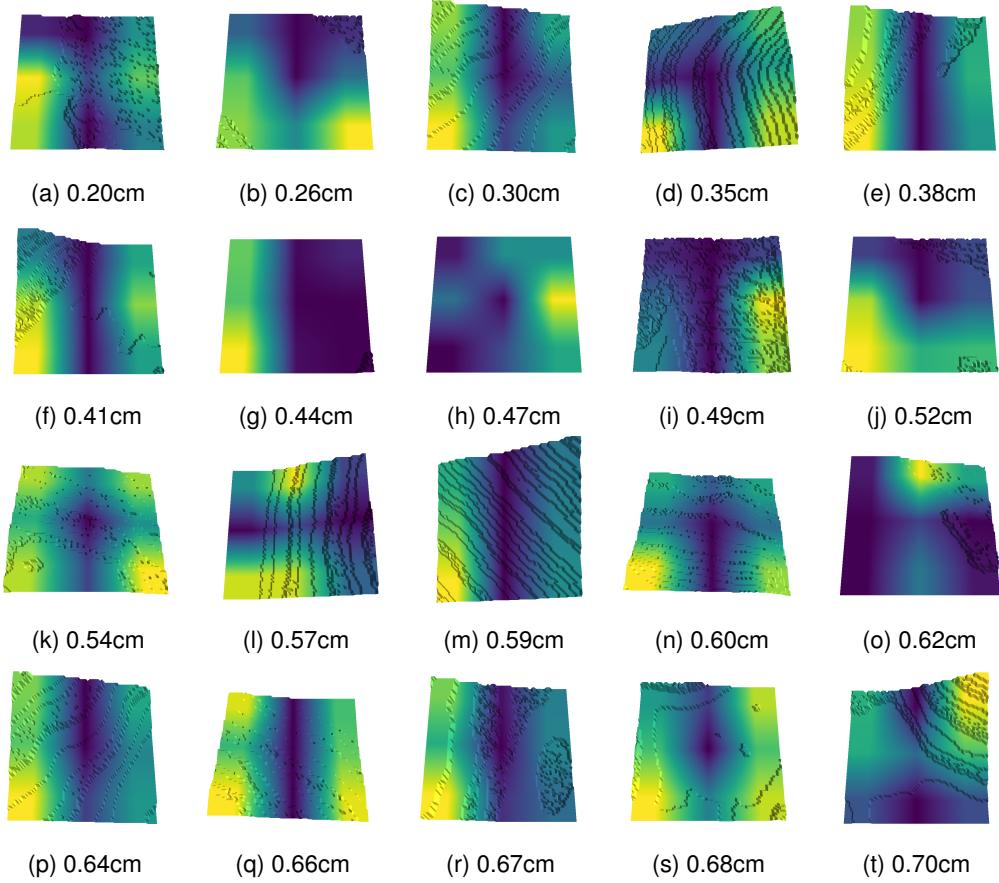


Figure 6.9

6.2.2 Worst

The worst patches are the inputs labeled and predicted not traversable. In this case, the region highlighted by the Grad-CAM are the features in the ground that contribute the most to make the patch not traversable. For instance, in some patches, 6.10b, 6.10f, 6.10t, the most not traversable features is directly under the robot legs, similar as before. While, for the others samples, 6.10a, 6.10c, 6.10d, 6.10e, ..., the attention of the network is on the obstacles in front of the robot. This is evident in 6.10c, 6.10c, 6.10e, where the big wall on the right is almost totally highlighted. So, without any doubts, also on this samples, the model is able to perfectly understand where to look to make the correct prediction

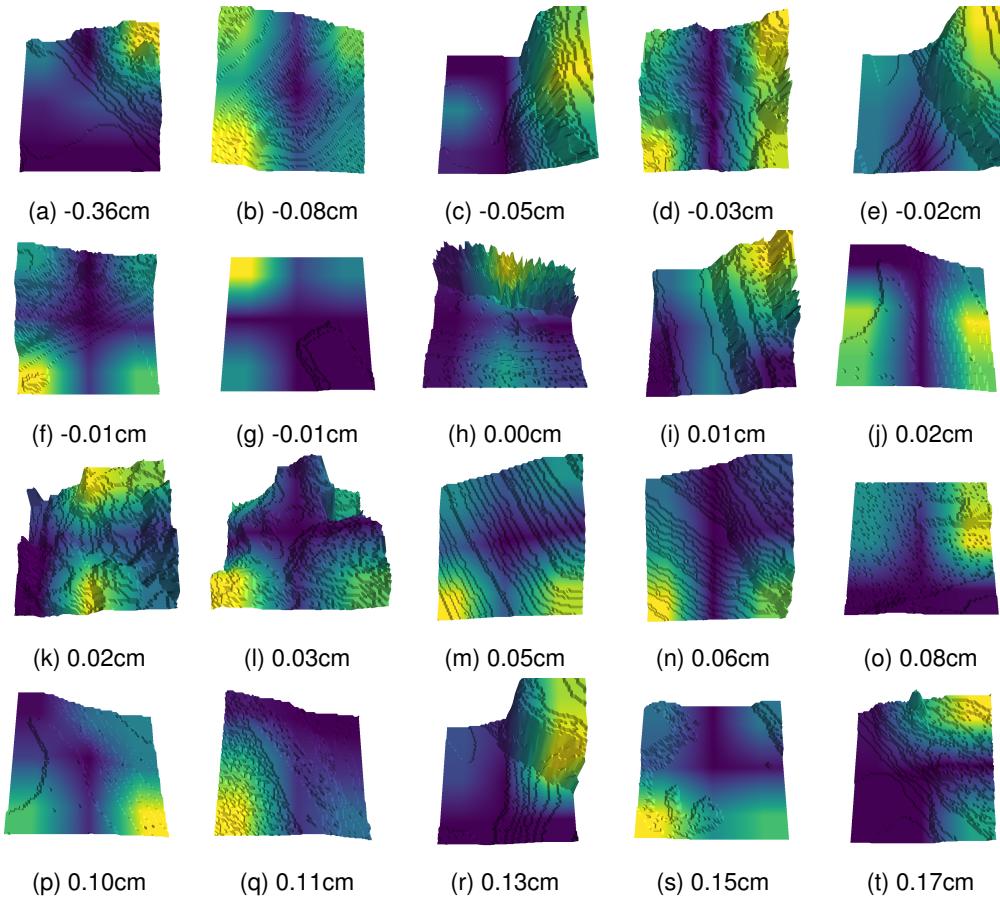


Figure 6.10

6.2.3 False Negative

Those are the inputs that were labeled as negative but predicted as positive. There are lots of interesting cases, we can cluster those patches in three groups: obstacles ahead, slopes and wall parallel to the robot. In figures 6.11a, 6.11f, 6.11n, the model is looking at the obstacle in front of the robot. Slopes, figures 6.11b, 6.11i, 6.11l, 6.11o, 6.11p, 6.11t, are not completely smooth, making harder to classify them. The last cluster of inputs are the ones with a wall parallel to the robot, figures 6.11n and 6.11t. In those cases the model looks at the initial position of the robot, left, and the final part of the patch. Intuitively, it is trying to understand if the robot fits on the trail.

In general, most of the region of interest on those patches are located on the left, close to the position of the robot's leg. Correctly, even if the prediction is wrong, the model looks at the first region of the surface, located near the legs, that can cause non traversability. This shows a correct behaviour even with wrong prediction meaning that the network is always looking in the correct spot.

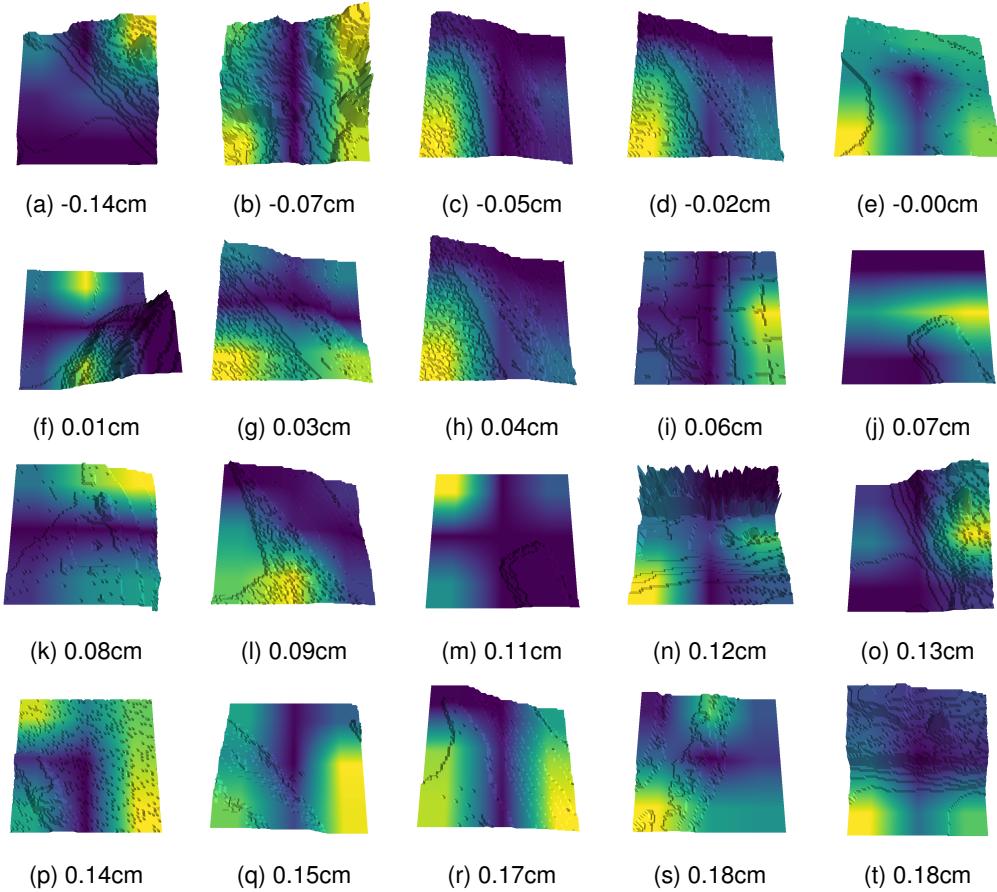


Figure 6.11

6.2.4 False Positive

Those are the most interesting ones, they are the samples labeled as traversable but predicted as not. They include different types of patches, some with obstacles ahead, slopes and flat regions. In each case, the model looks at meaningful features that can effect traversability. In the slopes, 6.12i and ??, the first part of the surface is highlighted. Similar to before, the model is looking if the rear legs will be able to move and thought that those small steps will cause the robot to not advance enough. The same consideration can be done for figure 6.12f. On the other hand, some inputs. definitely confused the model. For instance, in figures 6.12h and 6.12s the model is more interested in the big obstacle on the right part. This lead to classify those patches as not traversable, even if the obstacle is more distant than the threshold, in this case, the network was totally confused. Interesting, figure 6.12n shows an almost identical situation in which the network correctly looks at the initial part of the ground. We can deduce that even if in almost all cases the network's prediction is caused by the correct part of the inputs, it can be confused by a big obstacle near the end.

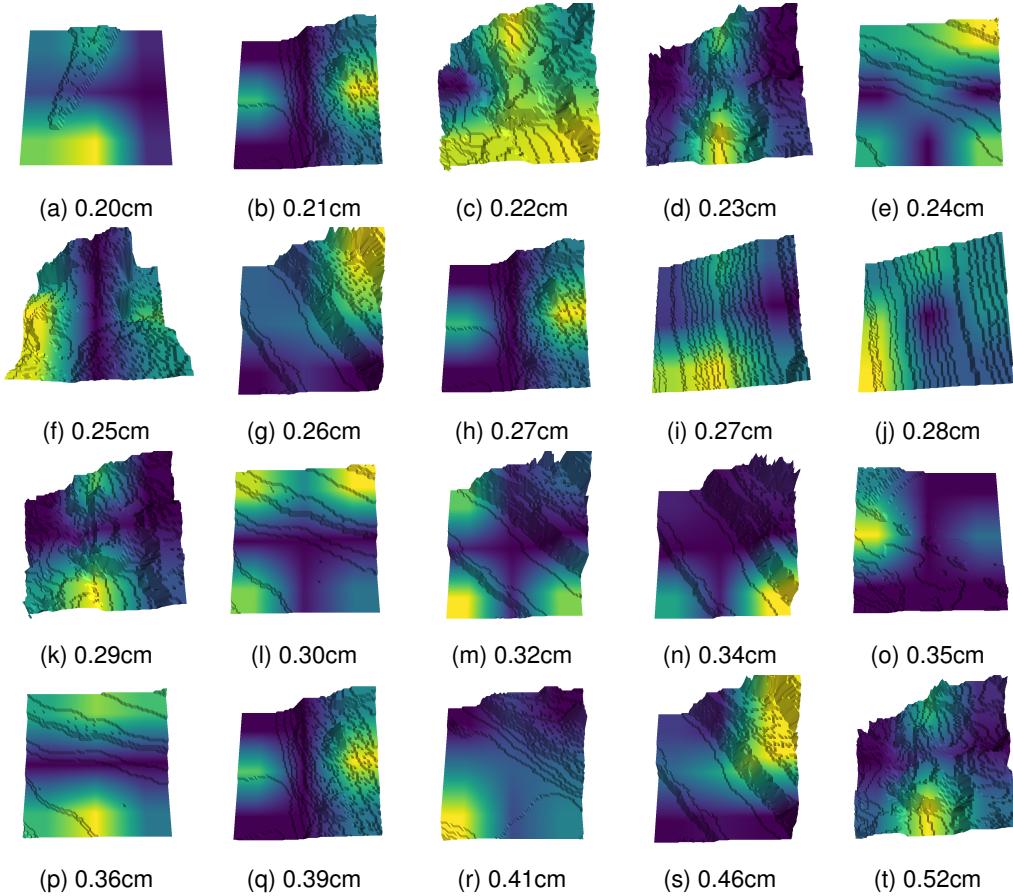


Figure 6.12

6.3 Robustness

To conclude the investigation of the network's abilities, we tested the model's robustness on patches with different features, walls, bumps, ramps, and examined the model's predictions against the real robot advancement obtained from the simulator. In general, the model's outputs matched the ground truth. When it fails, on some edge cases, it showed a certain degree of uncertainty. According to the previous experiments, we used a threshold of twenty centimeters and a time window of two seconds.

6.3.1 Untraversable wall ahead at increasing distance

We first test we performed was to place a not traversable wall in front of *Krock* at gradually moving towards the end. At a certain distance, we expected the model's prediction to be traversable even if the wall itself is too tall. Why? Because the robot will be able to travel more than the threshold before being stopped by the obstacle.

We created fifty-five different patches by first placing wall 16cm long exactly in front of the robot and then move it by 1cm at the time towards the end. Figure ?? shows some of the inputs

patches ordered by distance from the robot. We remind the reader that Krock traverse the patch from left to right. The model's predictions are displayed in figure 6.14. We can see how the two

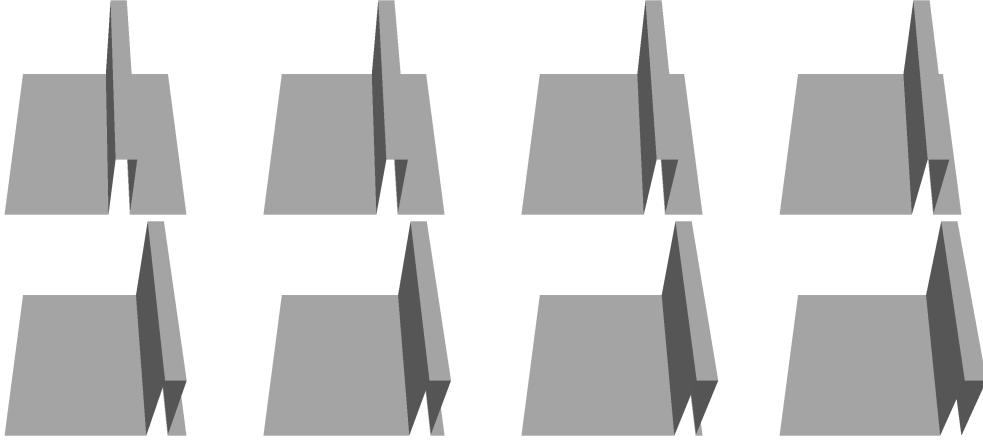


Figure 6.13. Some of the tested patches with a non traversable walls at increasing distance from the robot.

classes invert their values around 20cm. Moreover, the predictions, are consistent and do not change multiple times. Intuitively, if a wall is traversable from a certain distance, it will still be if we place even further. To be sure the results are correct, we run the last not traversable and the first traversable

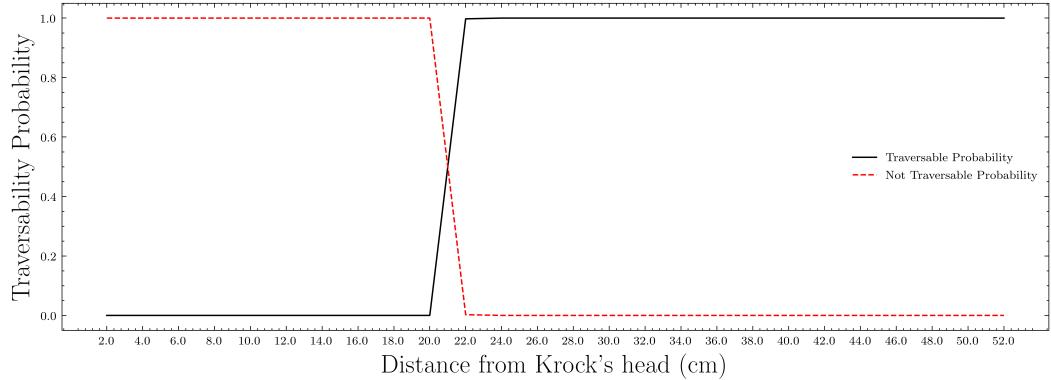


Figure 6.14. Traversability probabilities for patches with a non traversable walls at increasing distance from the robot.

patch on the simulator to get real advancement. In the simulator, Krock advances 18.3cm on the first, not traversable patch ?? where the wall is at 20cm from the robot's head. While, on the first traversable patch, figure reffig :walls-traversable-b, with a wall at 22cm, the robot was able to travel for 20.2cm. Correctly, the network's predictions are supported by the ground truth obtained from the simulation. Even more important, the model understood that the distance from the obstacle is more relevant than its height. Furthermore, we increased the wall size of the first traversable patch, figure ??, to 10 and to 50m to see if the model will be confused. Accurately, the robot was not confused by the enormous wall and the predictions did no change and those patches were still

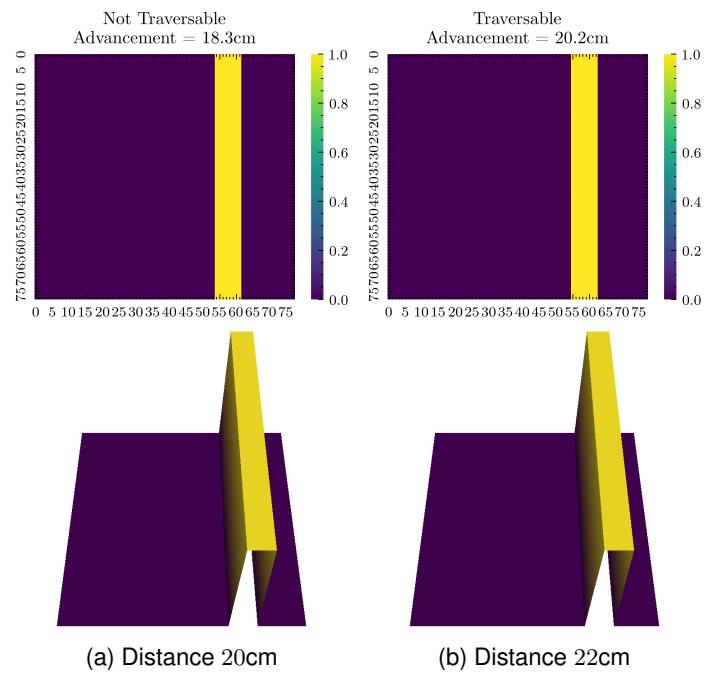


Figure 6.15. We run the last and first not traversable and traversable patch labeled by the model. Correctly, the real advancement is greater than the threshold when the distance between the robot is also greater.

labeled as traversable. Figure 6.16 visualize those patches.

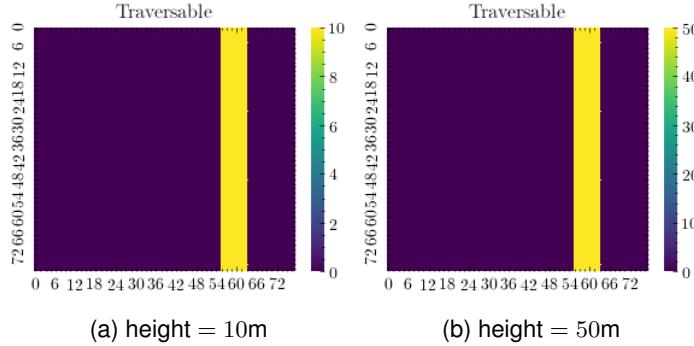


Figure 6.16. Two patches with a very tall wall at a distance major than the threshold. The model labels them as traversable without been confused by the huge obstacle.

6.3.2 Increasing height walls ahead

After we tested the distance from Krock's and a wall, we decided to fix the obstacle position but increase its heights. We run forty patches in the simulator with a wall place exactly in front of the robot with height from 1cm to 20cm. Figure ?? shows some of the inputs.

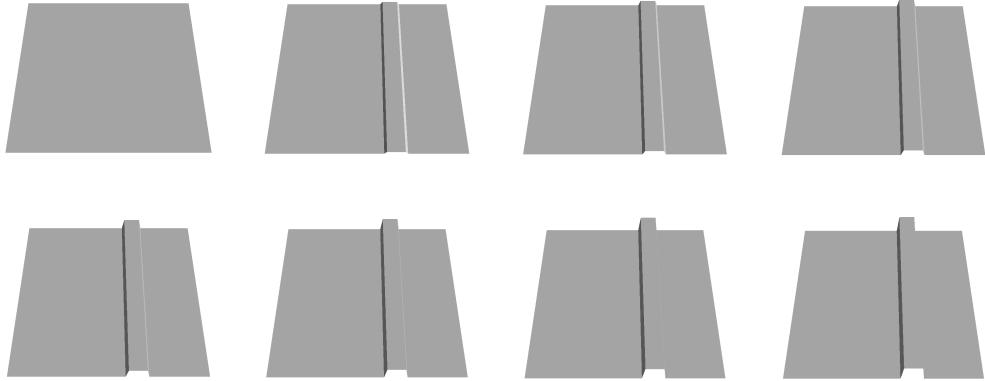


Figure 6.17. Some of the tested patches with a walls at increasing height ahead of Krock.

The models predicted that the walls under 10cm are traversable. We plotted the classes probabilities in figure 6.18 in which we see that in the edge case, $\approx 10\text{cm}$, the model's prediction change smoothly revealing a degree of uncertainty, positively the estimator outputs to change smoothly. We compared the model's prediction with the advancement computed in the simulator using the same approach from the last section. Figure ?? shows the results from the last traversable patch and the first non traversable.

In the first case, the simulator outputted and advancement of 21.2cm meaning that Krock was able to overcome the obstacle, while it failed in the second case. Correctly, the predictions matched the real data.

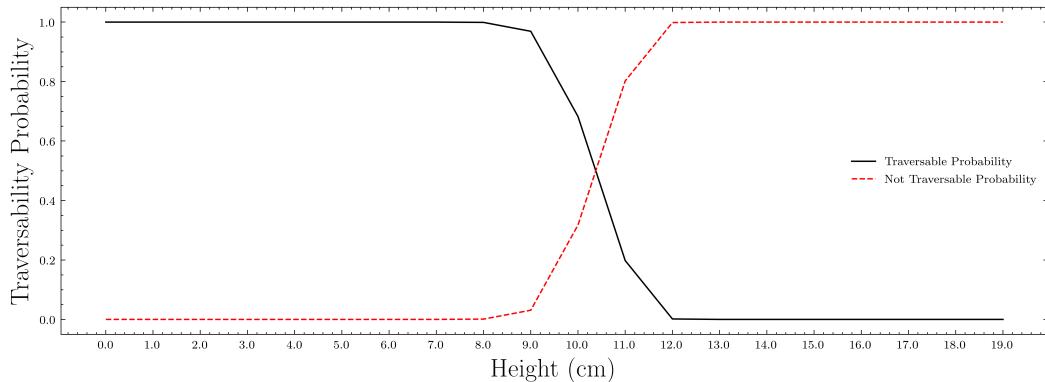


Figure 6.18. Traversability probabilities against walls height in front of Krock.

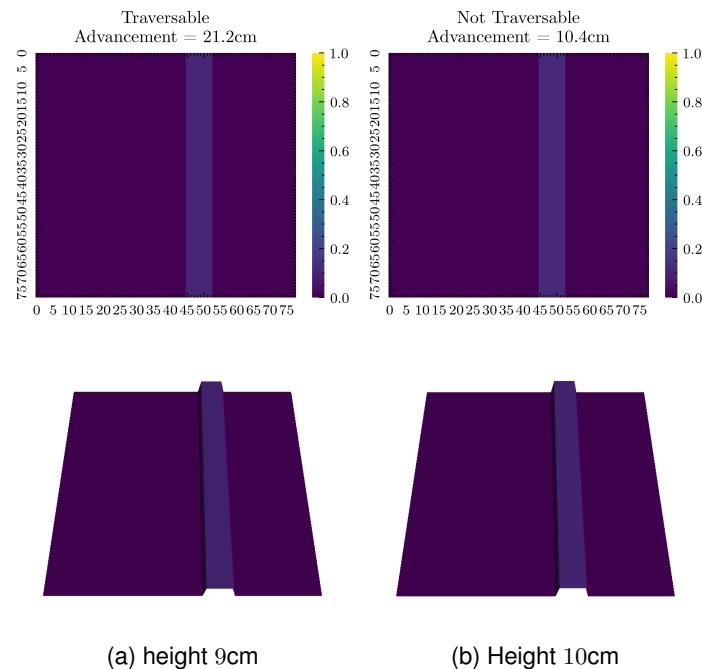


Figure 6.19. We run the last and first traversable and not traversable patch labeled by the model. Correctly the model's prediction matches the advancement from the simulator.

6.3.3 Increasing height and distance walls ahead.

We combined the previous experiments and tested the model predictions against the ground truth for each height/distance combination. To reduce the number of samples and improve readability, we limited to consider only patches with a wall tall between 5cm and 20cm, we know from previous sections patches with a value smaller and bigger obstacle are traversable and not traversable respectively. Similar, we set the wall's distance from Krock's head between 1cm to 30cm for the same reasons. To evaluate the model's prediction, we run all the patches several times on the simulator and average the results. Figure ?? shows the models outputs. We highlighted the false positive and the false negative by red and blue respectively. Since we spawned the robot directly on the patch inside the simulator, the outputs may change across different runs. Sometimes, two runs on the same patch can produce slightly different advancement due to some really small changes in the initial position of the robot caused the spawning procedure or some lag. Figure 6.20b displays the real advancements average across six different runs for all the patches. Additionally, for completeness, we also display in figure ?? the variance between all simulation's runs to highlighted cases where the advancement changes the most across different runs:

We obtained an overall accuracy of 91%. The model failed to predict some of the edge cases. The false positives are located in two regions: on the bottom left and on the top center. The first ones are the patches with a wall just ahead Krock of heights between $\approx 8 - 11$. The second region appears when the wall is at ≈ 20 cm, the threshold. But, even if the model failed to classify those inputs, it shows a correct degree of uncertainty. For example, most of predictions' probability for the patches at 10cm close to the robot are less than 0.9, some of them even close to 0.5.

The false negative, in the middle, are the inputs at distance close to the threshold and with a wall height between traversable and not traversable. Even if the model wrongly classified some of the inputs, all those errors are in the edge cases where the predicted classes' probability is not maximum. Moreover, in most cases, the model's showed uncertainty, especially on the false negative. Also, the prediction changes smoothly without any spikes accordingly to the features of the terrain. This shows a correct degree of understanding of the surface inputs. For instance, If the model outputs not traversable at height of 10cm and at a distance of 16cm, then all the taller wall are correctly labeled as not traversable showing consistency and predictability.

6.3.4 Tunnel

We also tested the model against tunnels. We expected the model to not perform well since in the train set samples with tunnels. The only two maps in the all dataset with possible tunnels, depending on the robot's orientation, *bars1* and *bars2*. Also, by looking at the different patches plotted in the features space, figure 6.4, there only a few patches with tunnels. We generated ten different terrains starting from two walls at 30cm from the borders. Figure visualize the inputs 6.21. The model predicted that all patches expect the last two are traversable, figure ???. Unfortunately, after comparing the real data from the simulator, the model scored an accuracy of just 40%. Only the first two patches are traversable, while all the others are not. When the distance between the walls reach a certain point, the robot was stuck and it was not able to move. Interesting, all the patches with a space inside the tunnel smaller than Krock's footprint, the last four samples, represented an impossible situation, since in real life the robot cannot go inside an obstacle. But, even some of those inputs were classified as traversable by the model. We utilized Grad-CAM to further investigate what the model is looking in those unrealistic patches. Figure 6.23 visualize the Grad-CAM directly on them. In the first two, were the model predicted traversable but the distance

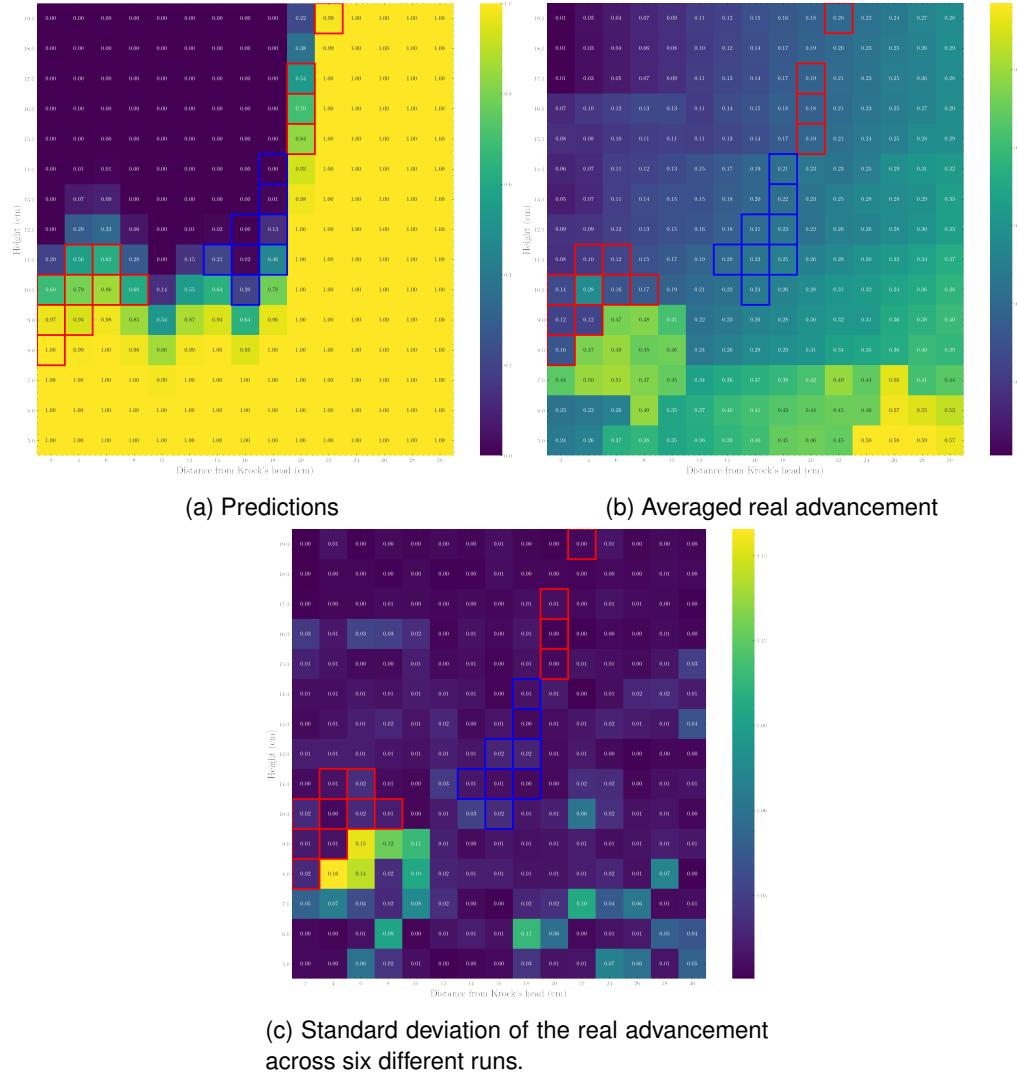


Figure 6.20. Results for all the combination of patches with a wall's distance from 2 – 30cm and heights between 5 – 19cm. False negatives are labeled with red while false positive with blue. The model failed to classify some edges cases when the wall is very close to Krock's head and when the wall's distance is near the threshold. The overall accuracy is 91%.

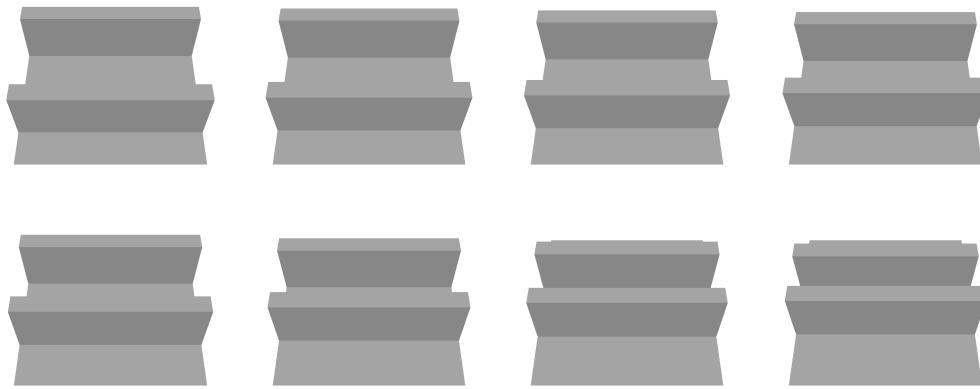


Figure 6.21. Some of the tested patches with tunnel at different distances.

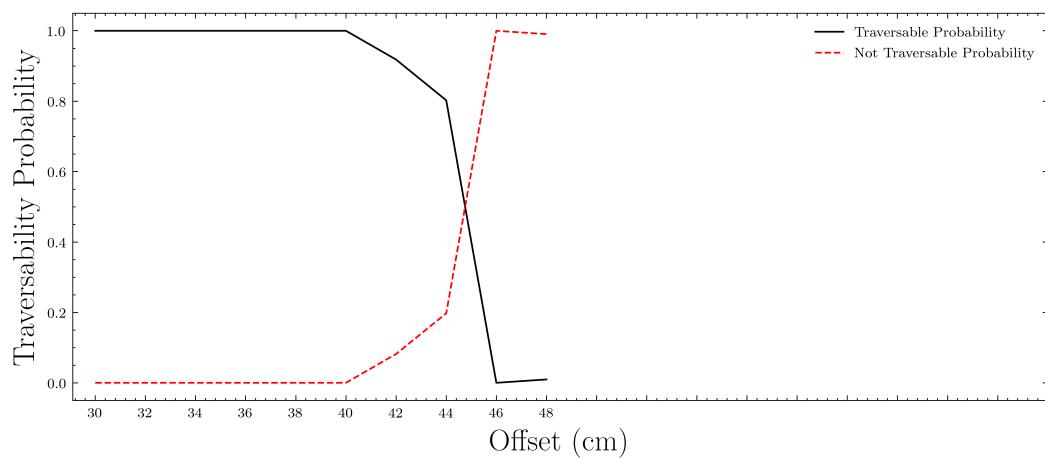


Figure 6.22. Traversability probabilities against tunnel offset from the top borders..

between the walls is less than the robot's footprint, the front region was highlighted meaning that the network did not consider the distance between the walls, while in the last two patches, the spot corresponding to the robot placement is highlighted. However, the model was clearly confused. This can be solved by incorporating in the train set samples with tunnels.

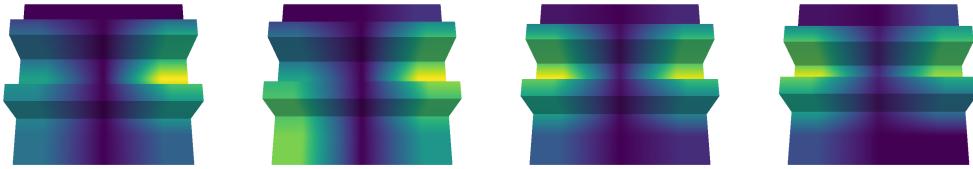


Figure 6.23. Fours patches with the distance between the wall smaller than Krock's footprint making impossible for the robot traverse them. Only the last to surfaces were correct label as not traversable.

6.3.5 Ramps

Using the same methodology from the last sections, we generated twenty ramps with a maximum height from 0.25m to 4m. Figure 6.24 shows some of the inputs. The model labeled as traversable the surfaces with a maximum height less than 1m and not traversable the others, figure ?? plots the traversability probabilities against the maximum height of each ramp. Then, we tested the last traversable patch and the first not traversable with the real advancement gather from the simulator. Figure ?? shows that the model's predictions are confirmed by the ground truth values.

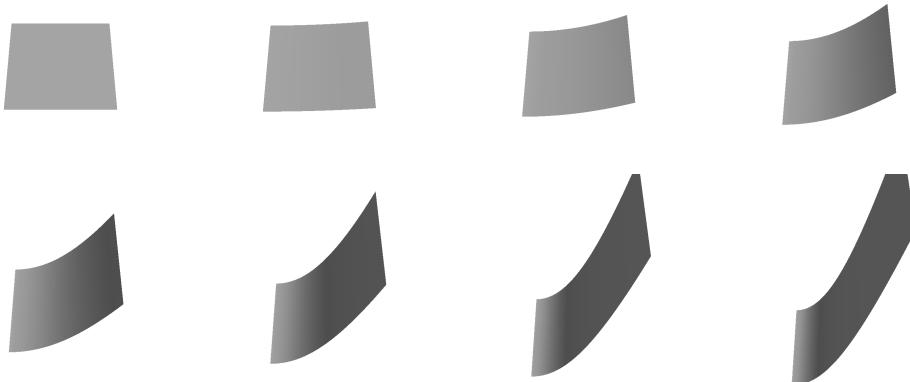


Figure 6.24. Some of the tested patches with steep ramps.

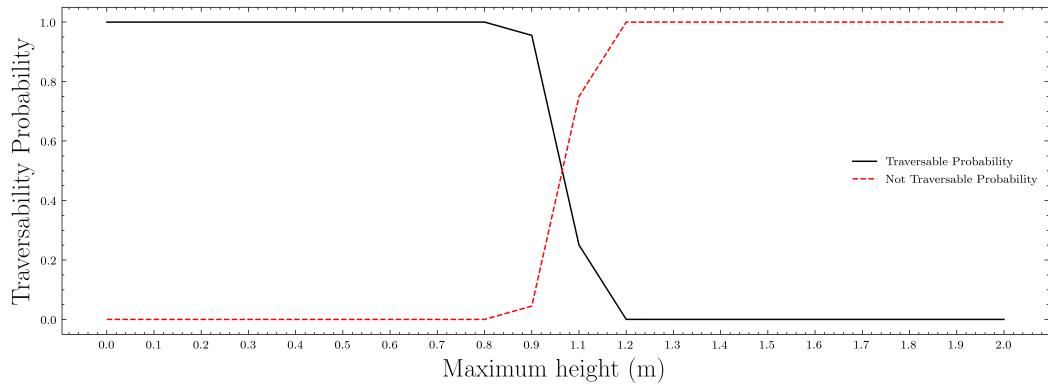


Figure 6.25. Traversability probabilities against maximum height of each ramp.

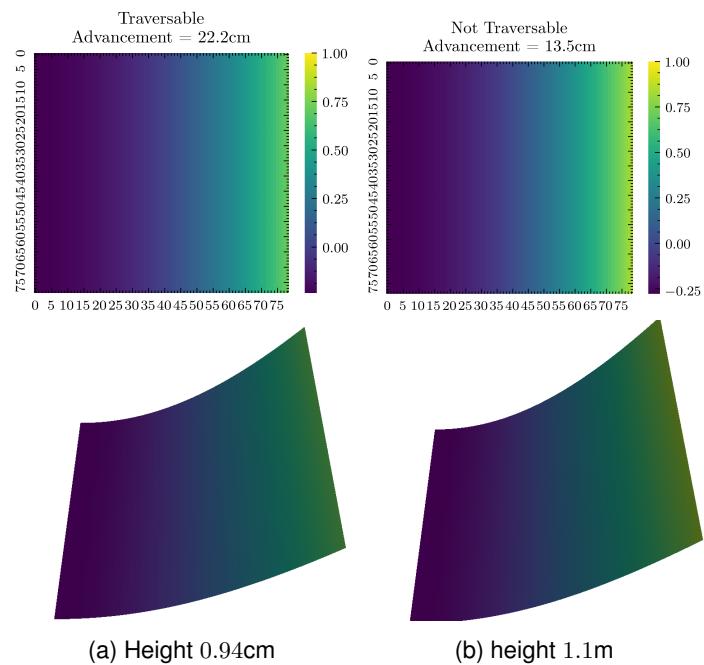


Figure 6.26. The last traversable and the first non traversable patches with a steep ramp ahead of Krock.

Chapter 7

Conclusions

We developed a complete framework for traversability estimation based entirely on simulation data and applicable to any ground robot. Simulation data allowed cheap, fast, quality and massive data gathered since 1) no real robot was used 2) simulations can be run faster than real-time and parallelize 3) maps can be engineered to maximize robots exploration and interactions with specific features 4) simulations can be run for any amount of time.

We trained a classifier to learn the robot's locomotion by supervised learning directly on ground patches. We tested the pipeline on a legged crocodile-like robot. The robot has unique locomotion that is able to overcome more obstacles than normal wheels robots making the estimation more challenging. We utilized a deep convolutional neural network to regress and classify the patches. We have shown both methods to be effective and we selected the latter due to its better accuracy once a threshold is fixed. We quantitatively measured the model's performance using different numeric metrics. Then, we visualized the traversability probability on real-world terrain to qualitatively evaluate the network. Later, we tested the estimator's capability of correctly extracting meaningful features from the patches. First, we proved that features were correctly aggregate and separate based on their classes. Then, we explored different samples from the train set by visualizing with part of the surfaces caused the model's prediction. We discovered that the model always looks at the correct spot. Finally, we stressed the network's strength and robustness by comparing different custom patches with the real advancement computed in the simulator. The results showed that the model expressed a correct degree of uncertainty in some edge cases while been able to properly classify most samples.

weakness of the method

Bibliography

- [1] e. a. A. Lagae. "a survey of procedural noise functions". *Comp. Graphics Forum*, 29(8):2579–2600, 2010.
- [2] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning, 2016.
- [3] J. G. Alessandro Giusti, D. C. Ciresan, F.-L. He, J. P. Rodríguez, F. Fontanam, M. Faessler, C. Forster, J. Schmidhuber, G. D. Caro, D. Scaramuzza, and L. M. Gambardella. A machine learning approach to visual perception of forest trails. 2016.
- [4] A. Y. N. Andrew L. Maas, Awni Y. Hannun. Rectifier nonlinearities improve neural network acoustic models.
- [5] R. O. Chavez-Garcia, J. Guzzi, L. M. Gambardella, and A. Giusti. Learning ground traversability from simulations. 2017.
- [6] J. Delmerico, A. Giusti, E. Mueggler, L. M. Gambardella, and D. Scaramuzza. On-the-spot training" for terrain classification in autonomous air-ground collaborative teams. 2017.
- [7] J. Delmerico, E. Mueggler, J. Nitsch, and D. Scaramuzza. Active autonomous aerial exploration for ground robot path planning. 2016.
- [8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks, 2016.
- [10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks, 2017.
- [11] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.
- [12] T. Klamt and S. Behnke. Anytime hybrid driving-stepping locomotion planning. 2017.
- [13] M. B. L. Wagner, P. Fankhauser and M. Hutter. Foot contact estimation for legged robots in rough terrain. 2016.
- [14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations.

- [15] W. Lorenz, D. Alexey, R. Ren, W. Krzysztof, C. L. Cesar, and H. Marco. Where should i walk? predicting terrain properties from images via self-supervised learning. 2019.
- [16] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. 2002.
- [17] P. Papadakis. Terrain traversability analysis methods for unmanned ground vehicles: A survey, 2013.
- [18] M. Quigley, B. Gerkey, K. Conley, T. F. Josh Faust, J. Leibs, E. Berger, R. Wheeler, and A. Ng. Ros: an open-source robot operating system. 2009.
- [19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2016.
- [20] J. Shlens. A tutorial on principal component analysis, 2014.
- [21] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, 2018.
- [22] B. Sofman, E. Lin, J. A. Bagnell, N. Vandapel, and A. Stentz. Improving robot navigation through self-supervised online learning. 2006.
- [23] E. Ugur and E. Sahin. Traversability: A case study for learning and perceiving affordances in robots. 2010.
- [24] M. Wermelinger, P. Fankhauser, R. Diethelm, P. Krusi, R. Siegwart, and M. Hutter. Navigation planning for legged robots in challenging terrain. 2016.