

1 Abstract

With this project, we estimate ground traversability for a legged crocodile-like robot. We generate different synthetic grounds and let the robot walk on them in a simulated environment to collect its interaction with the terrain.

Then, we train a deep convolutional neural network using the data collected through simulation to predict whether a given ground patch can be traverse or not. Later, we highlight the strength and weakness of our method by using interpretability techniques to visualise the network's behaviour in interesting scenarios.

2 Introduction

Effective identification of traversable terrain is essential to operate mobile robots in every type of environment. Today, there two main different approaches used in robotics to properly navigate a robot: online and offline. The first one uses local sensors to map the surroundings "on the go" while the second equip the mobile robot with an already labeled map of the terrain.

In most indoor scenarios, specific hardware such as infrared or lidar sensors is used to perform online mapping while the robot is exploring, this is the case of the most recent vacuum cleaner able to map all the rooms in an apartment. With the recent breakthroughs of deep learning in computer vision, more and more cameras have been used in robotics. For example, self-driving cars utilize different cameras around the vehicle to avoid obstacle using object detection. Indoor scenarios share similar features across different places shifting the problem from which ground can be traversable to which obstacle must be avoided. For instance, the floor is always flat in almost all rooms due to is artificial design. Moreover, usually, traversability must be estimated on the fly due to the high number of possible obstacles and to the layout of the objects in each room may not be persistent in time.

On the other hand, outdoor scenarios may have less artificial obstacle but their homogeneous ground makes challenging to determine where the robot can properly travel. Moreover, a given portion of the ground may not be traversable by all direction due to the not uneven terrain. Fortunately, a height map of the ground can be obtained easily by using third-party services or special flying drones. Those maps are extremely valuable in robotics applications since they provide an efficient way to examine the features of terrains, such as bumps, holes, and walls.

These scenarios have different difficulties. In indoor environments, it is easier to move the robot on the ground since it is designed for humans, but harder to perform obstacle avoidance. While in outdoors scenario it is maybe more challenged to the first estimate where the robot can go due to the huge variety of ground features that may influence traversability. To learn where to move, an artificial controller must be trained to predict the robot interaction with the environment. Such a process requires to collect some data to train the controller.

However, this may not be a straight forward process. In indoors terrain, most of the times, data is collected by driving the robot directly in the environment by a human or an artificial controller. While in the outdoors scenario data is assembled using the simulation for convenience.

Our approach aims to estimate traversability of a legged robot krocodile-like robot called Krock. We generate different uneven grounds in form of height map and let the robot walk for a certain amount of time on each one of them while recording its interaction, position and orientation. After, for each stored robot position we crop the corresponding ground portion in which the robot was during the simulator, those patches composed the training dataset. We select a minimum space in a fixed amount of time that the robot must travel to successfully traverse a ground region and use it to label the dataset. Then, we fit a deep convolutional neural network to predict the traversability probability. Later, we evaluate it using different metrics using real-world terrains.

The report is organized as follow, the next chapter introduces the related work, Chapter 2 describes our approach, Chapter 3 talks in deep about the implementation details, Chapter 4 shows the results and Chapter 5 discuss conclusion and future work.

3 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 [**ugur2010traversability**]. Visual perception is known to be 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 animals to extract geometric information of the surrounding and plan a safe path through it.

Different methodologies have been proposed to collect the data and then learn to correctly navigate the environment. Most of the methodologies rely on supervised learning, where first the data is gathered and then a machine learning algorithm is trained sample to correctly predict the traversability of those samples. Among the huge numbers of methods proposed, there are two categories based on the input 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 [**sofman2006improving**], outdoor robot navigation can benefit from the availability of overhead imagery. For this reason, elevation data has also been used to extract geometric information. Chavez-Garcia et al. [**omar2018traversability**], proposed a framework to estimate traversability using only elevation data in the form of height maps.

Elevation data can also be estimated by flying drones. Delmerico et al. [**delmerico2016active**] 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 convolutional neural network to segment the terrain in different traversable classes.

Whereas appearance methods, to a greater extent related to camera images processing and cognitive analyses, have the objective of recognising colours and patterns not related to the common appearance of terrains, such as grass, rocks or vegetation. Those images can be used to directly estimate the traversability cost.

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 [**ugur2010traversability**] or Gaussian models [**sofman2006improving**]. Those techniques rely on texture descriptors, for example, Local Binary Pattern [**ojala2002multiresolution**], 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. [**giusti2015amachine**] 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 also be used together to train a traversability classifier. Delmerico et al. [**delmerico2017onthespot**] extended the previous work [**delmerico2016active**] 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. [**tobias2017anytime**] proposed a locomotion planning that is learned in a simulation environment.

On other major distinction we can made 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. [**omar2018traversability**].

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. [**wermelinger2016navigation**] uses 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. [**wagner2016foot**] 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 automatise the features extraction process has been proposed recently. Lorenz et al. [wellhausen2019where] use local sensing to train a deep convolutional neural network to predict 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

Add mirko's paper

4 Model

This section gives a high-level overview of the steps in our approach. We will first describe the data gathering, then the post-processing pipeline used to generate the dataset. After, we will introduce the model used to fit the data and then show the results with different test environments.

4.1 Robot: Krock

Our task was to apply the approach proposed in [omar@traversability] with a legged robot developed at EPFL named *Krock*. Figure ?? shows the robot in the simulated environment.

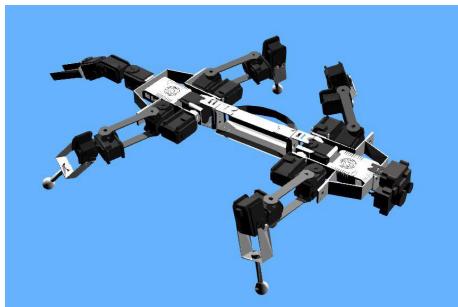


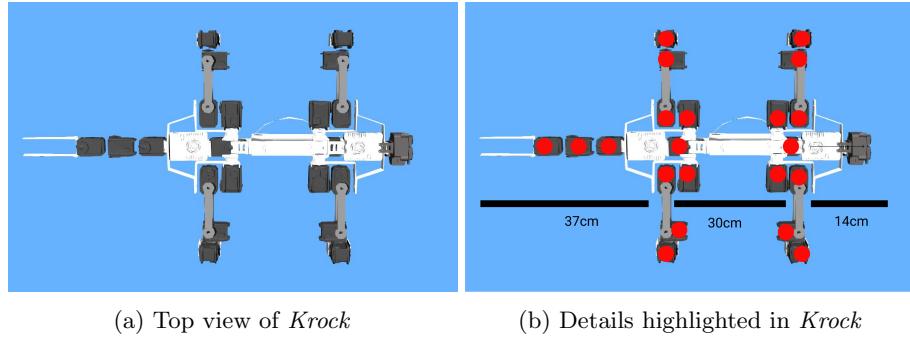
Figure 1: *Krock*

Krock has four legs, each one of them is equipped with three motors in order to rotate in each axis. The robot is also able to raise itself in three different

??

Figure 4: *Krock* different gait configuration.

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 and weights around 1.4kg. The next figure ?? 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 figure shows the robots going forward.

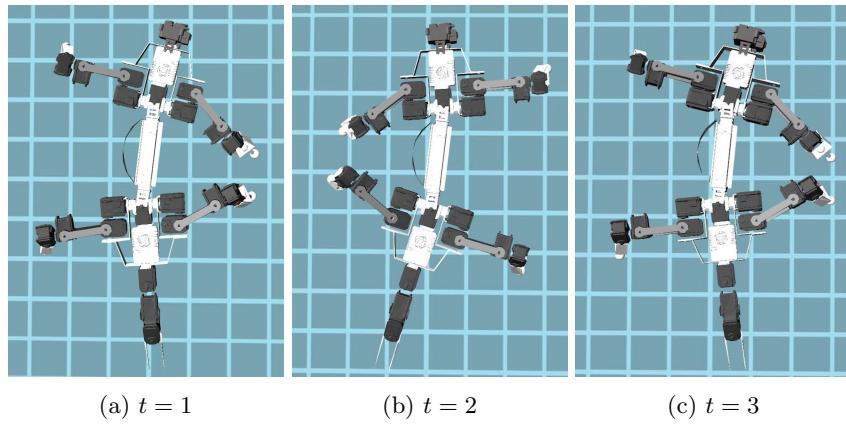


Figure 3: *Krock* moving forward.

Krock can also raise its own body in three different configurations. Figure Add Krock gait picture

4.2 Simulation

Our approach relies on simulations to collect the data needed to train the estimator. We used Webots [[webots](#)], a professional mobile robot simulator, as the backbone for our framework.

We generate fifteen synthetic 10×10 meters long heightmaps with different features, such as walls, bumps and slopes using the same techniques described in [[omar2018traversability](#)]. A heightmap is a gray image, formally a 2D array, where each element, pixel, represents a height value. Since each image has a range from $[0, 255]$, we also need to scale them when needed. For this reason, we associated a height scaling value for each height map that is used to increase or decrease the steepness. The next figure shows some of the maps used.

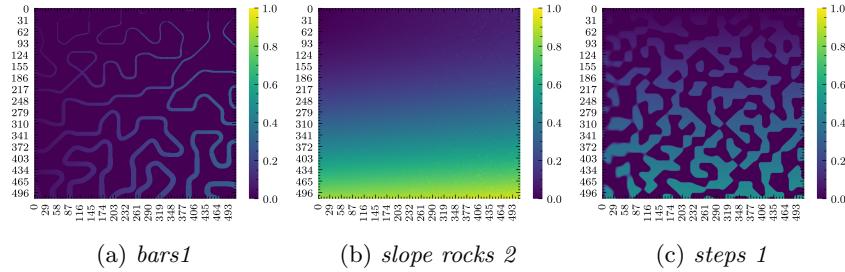


Figure 5: Some of the heightmaps used in the simulation.

The following figure shows the 3d rendered version of the bars heightmap.



Figure 6: *Bars1* 3d map

To generate the train data, each map is loaded into the simulator, then the robot is spawned in the world and we let it walk forward for a certain amount of time or until it reaches the edge of the map. While moving, we store his pose, that contains position and orientation, with a rate of 50hz.

We fix the robot's gait in its standard configuration and run fifteen simulations for each map, where one simulation one robot spawns and death, with a height scale factor of one. We also decide to include steepest training samples by using one of the maps with a slope with heights factors from [1, 7]. The following pictures show this specific map with different scaling factors.

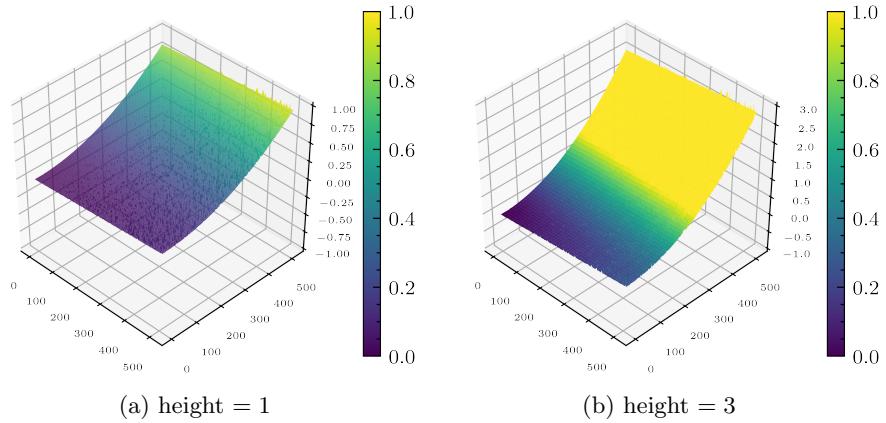


Figure 7: *Slope_rocks2* map with two different height scaling factor.

Krock is equipped with a controller implemented with the Robotic Operating System (ROS) software that is used as a communication channel between our framework and the simulator. Basically, ROS exposes nodes, called *topics*, in which we can listen for incoming messages. Thus, *Krock* is able to stream its pose to all the clients connected to its pose topic.

We also generate the test set, by running the robot in real-world environment, such as *Quarry* a cave map, that we later use to test the fitted model.

Update this image, maybe a blender render?

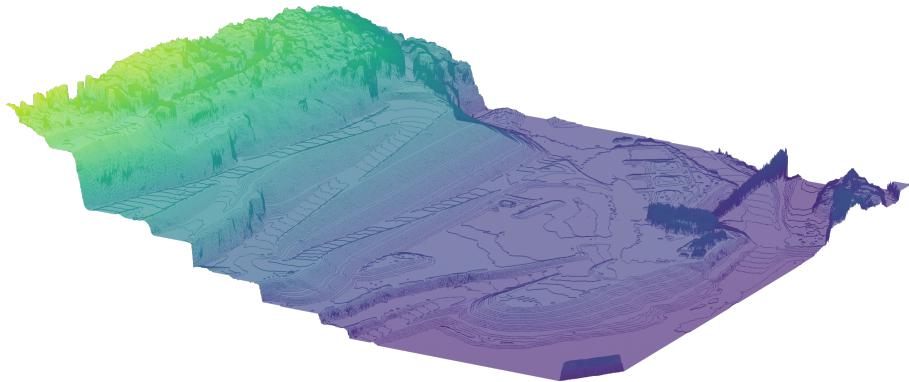


Figure 8: *Quarry* map.

By observing the simulations we notice two problems. First, *Krock*'s tail sometime gets under the body, second, if a random spawn strategy is used, in certain maps the agent may directly fall on an obstacle.

To overcome the first problem, we decide to remove the tail after we ensure by asking the developers that this will not compromise the traversability. We immediately observe less noisy data.

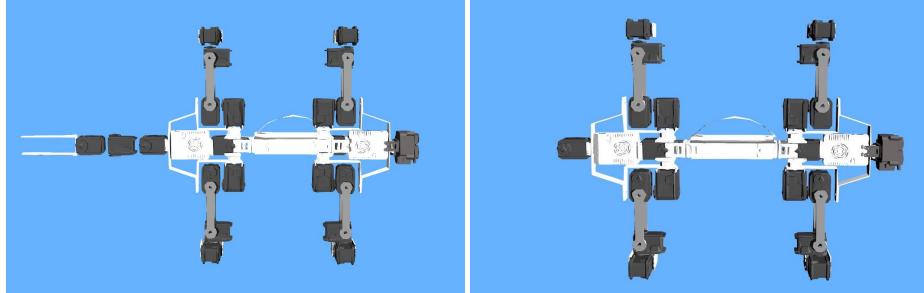


Figure 9: *Krock* with and without tail

The second issue was more challenging. If the robot spawns on an obstacle then it will be stuck for all the simulation compromising the quality of the data. This is very important for the *bars* map, where there are lots of wall of different sizes and the probability to fall on one of them is very high. So, for some maps, we guarantee that *Krock* does not directly spawn into obstacles by only spawning the robot in the flat ground first. The following figures show the result of this strategy on the *bars1* map where the robot was spawned where there is no obstacle. The following image shows different *Krock*'s spawn position in the simulator on the *bars1* map, the reader can notice that the robot is always placed where there are no obstacles under it.



Figure 10: *Krock*'s spawns on the *bars1* map.

We run a total of storing the results using ROS's .bag files for a total of .

add number
of simula-
tions

add size of
all bags.

4.3 Postprocessing

In order to create the training dataset, we need to post process the data stored during the simulation. To train our classifier we need to first compute the advancement for each pose at each time for a specific time window, $pose_{it}$. Then, we can use it to crop a patch from the heightmap that was used in the simulation of the correct size. Such patch must include the whole *Krock* footprint and the maximum possible advancement in a given time window. By keeping this final goal in mind, we quickly describe each step taken to reach it. Figure 11 shows each step applied in the pipeline.

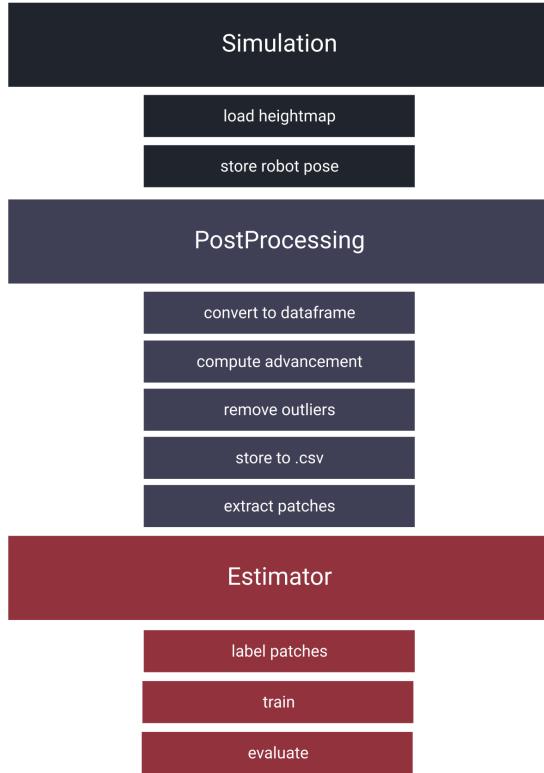


Figure 11: Pipeline

We implemented an easy to use API to define a cascade stream of function that is applied one after the other using a multi-thread queue to speed-up the process. First, we convert each *.bag* file into a *.csv* file and store it since this operation must be done only once. Then, we define a time window t and for each entry in one simulation run we compute the future advancement using booth position and angle from the pose, we also store the resulting data frame into a *.csv* file. During the process, we also clean the data by removing the samples in which the robot was upside down or out of the map.

link to the project

Finally, we crop from each heightmap the patches corresponding to each data point. To compute the patch size need to know the maximum possible advancement of *Krock* for a specific time window. In order to find it, we run *Krock* on a flat ground map and we take the mean across all of them. We observe a mean advancement of 33cm per second.

For a given advancement, we ensure the whole *Krock* footprint is included on the left part to take into account the situations in which an obstacle is directly under the robot. To clearly show the process, we used a simulated environment in which *krock* has a big wall in front of him and one small under his legs. The following picture shows the terrain.

[add figure with the two walls](#)

Clearly, assuming *Krock* is able to traverse 33cm of flat ground per second, in one second it won't be able to reach the obstacle, this is shown in the first picture. Increasing the time window, the taller wall appears while the small one under the robot is always correctly included.

[add figure with the patch computation](#)

4.4 Estimator

To classify the patches we need a class associated to each one of them, thus we must label each patch as *traversable* or *not traversable*. So, we selected a threshold, tr , and labeled each patch by $patch_{advancement} > tr$. In other words, if *Krock* was able to advance in a patch more than the threshold, then that patch is labeled as *traversable* and vice-versa. The pair of patches and labels represent the final dataset.

[add an image with a trivial traversable and not traversable patch](#)

We used a deep convolutional neural network to fit the training set. Convolutional Neural Network is able to map images into a features space and then to target class. Before feeding the patches into the model, we subtract in each patch the value in the middle in order to normalize them by making height invariant. Visually,

[add figure with a patch not centered and centered](#)

We evaluate different models with different hyperparameters to select the best one using the metrics score on the test set as guidelines. We then visually evaluate the network by running it on the *Quarry* map for different orientations. In order to do so, we extract patches from the ground using a sliding window and feed them to the model. Then, we create a texture where each pixel represents the traversability probability, the brighter the higher.

5 Implementation

5.1 Tools

The most important tools and libraries used in our work were:

- ROS Melodic
- Numpy
- Matplotlib
- Pandas
- OpenCV
- PyTorch
- FastAI
- imgaug
- Blender

The framework was entirely developed on Ubuntu 18.10 with Python 3.6.

5.1.1 ROS Melodic

The Robot Operating System (ROS) [ROS] 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 robust and complex robot behavior regardless of the platforms. ROS work by generating a peer-to-peer connection where each *node* is to communicate between the others by exposing sockets endpoints to stream data called *topics*.

Each *node* can subscribe to received the incoming messages or publish new data on a specific *topic*. In our case, *Krock* exposes different topics, for example */pose*, in which we can subscribe in order to get the real time informations about the state of the robot. Unfortunately, ROS does not natively support Python3, so we had to compile it by hand. Since it was not difficult and time consuming operation we decided to share the ready-to-go binaries as docker image.

5.1.2 Numpy

Numpy is a fundamental packages for any scientific use. Thanks to its powerful N-dimensional array object with the sophisticated broadcasting functions, it is possible to express efficiently any matrix operation. We utilised *Numpy* to manipulate matrices in an expressive and efficient way.

scrape some text from the previous section

where
should I
place the
link to
docker

5.1.3 Matplotlib

Matplotlib is a widely used Python 2D plotting library which generates high quality figures in a variety of hardcopy formats and interactive environments across platforms. It provides a similar functional interface to MATLAB and a deep ability to customise every region of the figure. Almost every figures made in this report were produced using Matplotlib. It is worth citing *seaborn* a data visualization library that we inglobate in our work-flow. It is based on Matplotlib and it provides a high-level interface for drawing attractive and informative statistical graphics.

5.1.4 Pandas

Pandas is a Python library providing fast, flexible, and expressive data structures in a tabular form. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Today, it one of the most flexible open source data manipulation tool available. 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. It provides to two primary data structures, **Series** and **DataFrame** for representing 1 dimensional and 2 dimensional data respectively.

Generally, pandas does not scale well and it is mostly used to handle small dataset while relegating big datato 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.

5.1.5 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. With a huge community of more than fourtyseven thousand people, the library is a perfect choice to handle image data. In our framework, OpenCV is used to pre and post-process the heightmaps and the patches.

5.1.6 PyTorch

PyTorch is Python open source deep learning framework. It provides Tensor computation (like NumPy) with strong GPU acceleration and Deep neural networks built on a tape-based autograd system. Due to its *Python-first* philosophy it has a deep integration into Python allows popular libraries and packages to be used, such as *OpenCV* or *Pillow*.

Due to its simply yet expressive and beautiful object oriented API it has been adopted be a huge number of researches and enthusiasts all around the world creating a flourishing community. Its main advantage over other mainstream frameworks such as TensorFlow are a cleaner API structure, better debugging,

cite TF

code shareability and enormous number of high quality packages. All the neural network proposed in this project are built using Pytorch.

5.1.7 FastAI

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

5.2 imgaug

Image augmentation (imgaug) is a python library to perform image augmenting operations on images. They provide a variety of methodologies, affine transformations, perspective transformations, contrast changes and gaussian noise, to build sophisticated pipelines. They support images, heatmaps, segmentation maps, masks, keypoints/landmarks, bounding boxes, polygons and line strings.

5.2.1 Blender

Blender is the free and open source 3D creation suite. It supports the entirety of the 3D pipeline modeling, rigging, animation, simulation, rendering, compositing and motion tracking, even video editing and game creation. We used Blender to render some of the 3D terrain used to evaluate the trained model.

5.3 Data Gathering

5.4 Postprocessing

5.5 Estimator

5.5.1 Vanilla Model

5.5.2 ResNet

We decide to use a Residual Network, ResNet [**he2015deep**], variant. Residual network 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 identify map to make the addition possible. This allows a stronger gradient flows and mitigates the degradation problem. A Residual Units is composed by a two *3x3 Convolution, BatchNorm [ioffe2015batch]* and a *Relu* blocks. Formally, it is defined as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + h(\mathbf{x}) \quad (1)$$

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

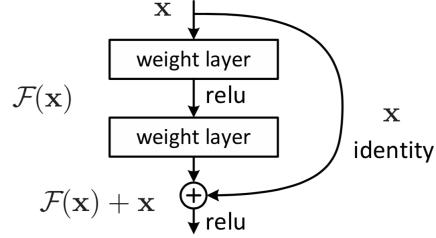


Figure 12: *Resnet* building block.

iiiiii HEAD When the input and output shapes mismatch, the *identity map* is applied to the input as 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.

Following the recent work of He et al. [he2015identity] we adopt *pre-activation* in each block. Easily, *Pre-activation* reverse the order of the operation in a block. ===== When the input and output shapes mismatch, the *residual map* is 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.

Following the recent work of He et al. [he2015identity] we adopt *pre-activation* in each block. *Pre-activation* just reverse the order of the operation in a block. iiiiiiii master

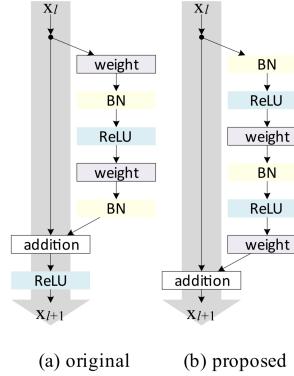


Figure 13: *Preactivation*

Finally, we also used the *Squeeze and Excitation* (SE) module [hu2017squeeze]. A form of attention that weights the channel of each convolutional operation by learnable scaling factors. The next figure visualises the SE module.

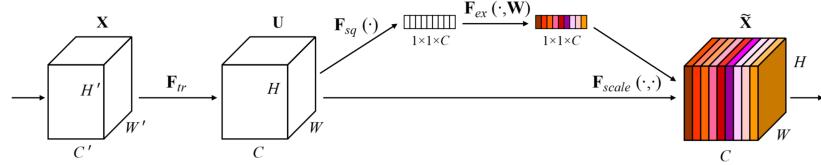


Figure 14: *Preactivation*

iiiiiii HEAD

5.5.3 Data Augmentation

===== We select *ResNet* architecture similar to the ones originally proposed to fit the *CIFAR10* [[cifar](#)] dataset. We used a initial features size of 64 with three blocks of depth $n = 1$. The final model has [i.i.i.i.i.i.9292b346900ba8f12cdaac1cbc](#)

[add model picture](#)

[add learnable parameters](#)

6 Results

In the section we show and evaluate the models results. We will start by presenting to the reader the networks score on each metric, then we will use the best models to predict the traversability of real world terrain. Finally, we will use handcrafted patches, for example a wall of a certain height in a specific position, to test the robustness of the network by trying to highlighting its behaviour.