




# Where Should I Walk? Predicting Terrain Properties from Images via Self-Supervised Learning

**Journal Article****Author(s):**

[Wellhausen, Lorenz](#) ; [Dosovitskiy, Alexey](#); [Ranftl, René](#); [Walas, Krzysztof](#); [Cadena, Cesar](#) ; [Hutter, Marco](#) 

**Publication date:**

2019-04

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000323783>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

IEEE Robotics and Automation Letters 4(2), <https://doi.org/10.1109/lra.2019.2895390>

# Where Should I Walk?

## Predicting Terrain Properties from Images via Self-Supervised Learning

Lorenz Wellhausen<sup>1</sup>, Alexey Dosovitskiy<sup>2</sup>, René Ranftl<sup>2</sup>, Krzysztof Walas<sup>3</sup>, Cesar Cadena<sup>4</sup>, and Marco Hutter<sup>1</sup>

**Abstract**—Legged robots have the potential to traverse diverse and rugged terrain. To find a safe and efficient navigation path and to carefully select individual footholds, it is useful to be able to predict properties of the terrain ahead of the robot. In this work, we propose a method to collect data from robot-terrain interaction and associate it to images. Using sparse data acquired in teleoperation experiments with a quadrupedal robot, we train a neural network to generate a dense prediction of the terrain properties in front of the robot. To generate training data, we project the foothold positions from the robot trajectory into on-board camera images. We then attach labels to these footholds by identifying the dominant features of the force-torque signal measured with sensorized feet. We show that data collected in this fashion can be used to train a convolutional network for terrain property prediction as well as weakly supervised semantic segmentation. Finally, we show that the predicted terrain properties can be used for autonomous navigation of the ANYmal quadruped robot.

**Index Terms**—Semantic Scene Understanding; Visual-Based Navigation; Visual Learning

### I. INTRODUCTION

ROBOT navigation through natural environments poses numerous challenges not present in indoor and other man-made environments. Perceived terrain geometry cannot be assumed to be rigid without severely restricting operational capabilities, for instance in the presence of vegetation. Even flat ground might not be negotiable without the risk of failure in the presence of bodies of water, sand and other challenging terrain types. Moreover, terrain properties change depending on the environmental conditions such as precipitation (e.g. dry vs. wet sand) and temperature (e.g. water vs. ice). Because these properties typically cannot be directly measured remotely, they need to be estimated from sensor streams in order to enable safe and efficient navigation and individual foothold selection.

Previous methods largely focus on purely geometric environment models for traversability estimation [1], [2], [3].

Manuscript received: September, 10th, 2018; Revised December, 10th, 2018; Accepted January, 8th, 2019.

This paper was recommended for publication by Editor Tamim Asfour upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by Intel Network on Intelligent Systems. It has been conducted as part of ANYmal Research, a community to advance legged robotics.

<sup>1,4</sup>Authors are with the Robotics Systems Lab<sup>1</sup> and Autonomous Systems Lab<sup>4</sup>, ETH Zürich [authors@mavt.ethz.ch](mailto:authors@mavt.ethz.ch)

<sup>2</sup>Authors are with Intel Labs

<sup>3</sup>Author is with the Institute of Control, Robotics and Information Engineering, Poznan University of Technology

Digital Object Identifier (DOI): see top of this page.

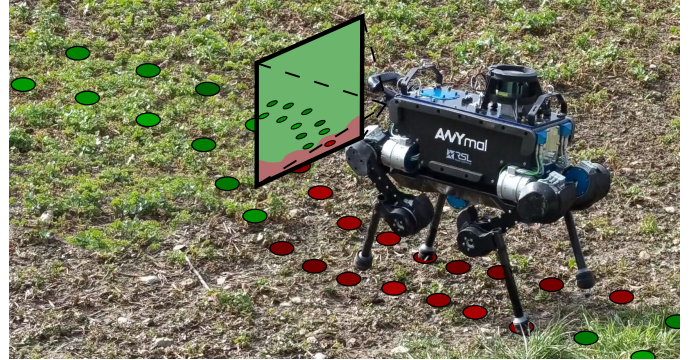


Fig. 1: Robot footholds are projected into camera images to automatically obtain label values. They are used to train a convolutional network to predict dense terrain properties from images.

Unfortunately, information provided by these approaches is not sufficiently detailed to support effective locomotion in complex natural environments. To overcome these limitations, terrain classification is widely employed for more detailed terrain perception [4], [5]. However, these methods are typically limited to a manually pre-defined fixed set of terrains and usually do not account for intra-class property variations. Additionally, ground interaction dynamics depend on robot morphology and locomotion mode and are difficult to impossible to obtain via hand-annotation, without terrain interaction.

### A. Contribution

In this paper, we take a step towards fully automated self-supervised learning and prediction of navigation-relevant terrain properties. We approach the problem by associating terrain information obtained from robot operation with camera images. To this end, we project robot footholds to the camera frame. To obtain image labels associated with the footholds, we estimate the terrain properties automatically and without any human intervention, by measuring the interaction during locomotion using sensorized feet.

We deploy and test the proposed system on the ANYmal quadruped [6]. Based on recordings of terrain interaction, we derive a *ground reaction score*, which serves as a measure for the difficulty of terrain negotiation. To accurately regress this score from images, we then train a convolutional network (CNN). Moreover, we show that semantic terrain segmentation labels for over 70,000 images can be obtained with less than an hour of manual work and can be used to train an accurate segmentation model. Finally, navigation trials that use

a 2D map of the local ground reaction score as a basis for path planning, exhibit intuitive navigation behavior without manually specifying preferred terrain types.

To summarize, our main contributions in this work are:

- 1) An automated system for pixel-accurate image annotation using foothold positions and proprioceptive sensor readings.
- 2) Definition of ground reaction score as a metric of terrain negotiation difficulty.
- 3) Adaptation of a CNN to facilitate learning on sparse and noisy labels.
- 4) Experimental verification of terrain property prediction for robot navigation.

### B. Related Work

Traditional path planning approaches for mobile robots use a geometric representation of the environment, like 2D height-maps [1], [2], point clouds [3], and signed distance fields [7]. They show good performance in environments with rigid terrain and obstacles, but do not capture deformable obstacles and non-rigid surfaces. This is not sufficient for ground robots, which by definition interact with the ground and need to exert forces onto the environment to move forward. For semantic-aware navigation, additional and richer sensor streams like grayscale [4], RGB [8], [5], [9], [10], [11], NIR [12], [5] and thermal cameras [13] as well as RADAR [14] have been used.

The approach based on thermal inertia [13] requires observation of terrain over longer periods and is difficult to apply in use-cases other than space exploration, while modern RADARs are typically bulky and heavy and exceed the payload capacity of our target platform. Most camera-based approaches either classify different terrain types and then attach a value to every class [4], [5], [9], [12] or directly predict a manually defined traversability value [8], [10], [11]. Approaches based on classifying the terrain can reach impressive performance in navigating natural terrain [5], but are limited by the necessity to specify explicit terrain classes. Additionally, human experts who define the segmentation might be unaware of the actual interaction dynamics with the environment. This can be sufficient for use-cases with constant and known environmental conditions [4], [9], but does not scale for varying weather and terrain conditions. If for example sand is given as a terrain class, its properties can vary strongly depending on factors like humidity and compression, necessitating the creation of sub-classes.

Weakly and self-supervised learning has increasingly been investigated to obtain scalable navigation solutions, obtaining training data from additional sensors [9], [10] or robot state [8], [11]. Current methods learn either traversability or terrain class, but do not estimate terrain properties directly. Of these works only Barnes et al. [10] make dense, pixel-wise predictions.

Terrain identification via haptic interaction is an active area of research. Automatic class discovery has been achieved via clustering of terrain on a full legged system while walking [15]. Estimating physical ground properties has, to the

best of our knowledge, only been demonstrated on a single-leg test bench [16] and on full legged systems using an active probing motion [17], [18], but not during locomotion.

Inspired by the aforementioned works, our approach aims to take terrain property prediction a step further. We identify terrain properties from haptic interaction during locomotion and associate them with images to facilitate self-supervised learning of dense pixel-accurate prediction of terrain properties from RGB images.

## II. APPROACH

The developed framework for self-supervised terrain property learning is illustrated in Figure 2. It comprises three main components: The *foothold projection module* maps the foothold positions to the camera plane. The *labeling module* provides a label for each foothold. The *prediction module*, implemented by a CNN, is trained to estimate the labels from RGB images.

### A. Foothold Projection

To be able to project the 3D positions of footholds into camera images, we need to compute the positions of robot foot contact closures and the camera trajectory in a common coordinate frame. For tracking the camera pose we use a visual SLAM system ORB-SLAM2 [19]. Thanks to loop closure detection, it enables accurate camera pose tracking even on long trajectories. In our experiments we provided RGB-D input to the SLAM system to ensure robustness and precision, although we have found RGB-only SLAM to work well too.

We now consider a legged robot with point-feet, but the proposed technique would also apply to other ground robots. For each contact point, we project a circle centered at this point to all camera views. To this end, we first compute the positions of the contact point and two points on the circle circumference relative to the current camera view, using joint-encoder angles, robot kinematics, and an estimate of the ground plane based on the recent contact points. We then project these three points to each of the other camera views and use them to draw an ellipse in each. We neglect possible occlusions of foothold positions caused by vertical geometry. In future work, occluded footholds could be filtered out using a 3D reconstruction of the terrain. Specifically, we obtain the coordinates  $\mathbf{x}^{i,t}$  of the three points projected from foothold  $i$  to frame  $t$  as follows:

$$[\mathbf{x}_c^{i,t} \quad \mathbf{x}_1^{i,t} \quad \mathbf{x}_2^{i,t}] = \mathbf{K} \cdot \mathbf{T}_{CW}^t \cdot [\mathbf{p}_c^i \quad \mathbf{p}_1^i \quad \mathbf{p}_2^i]. \quad (1)$$

Here  $\mathbf{p}_c^i$ ,  $\mathbf{p}_1^i$ , and  $\mathbf{p}_2^i$  are the center and the two circumference points of foothold  $i$ , expressed in the fixed global coordinate frame of the SLAM system.  $\mathbf{K}$  is the intrinsic camera calibration matrix and  $\mathbf{T}_{CW}^t$  is the transformation matrix from the global coordinates into the camera-fixed frame at time  $t$ . By computing  $\mathbf{x}^{i,t}$  for every foothold  $i \in [1, N]$  and every image  $t \in [1, T]$ , we label all recorded images while taking into account the entire robot trajectory.

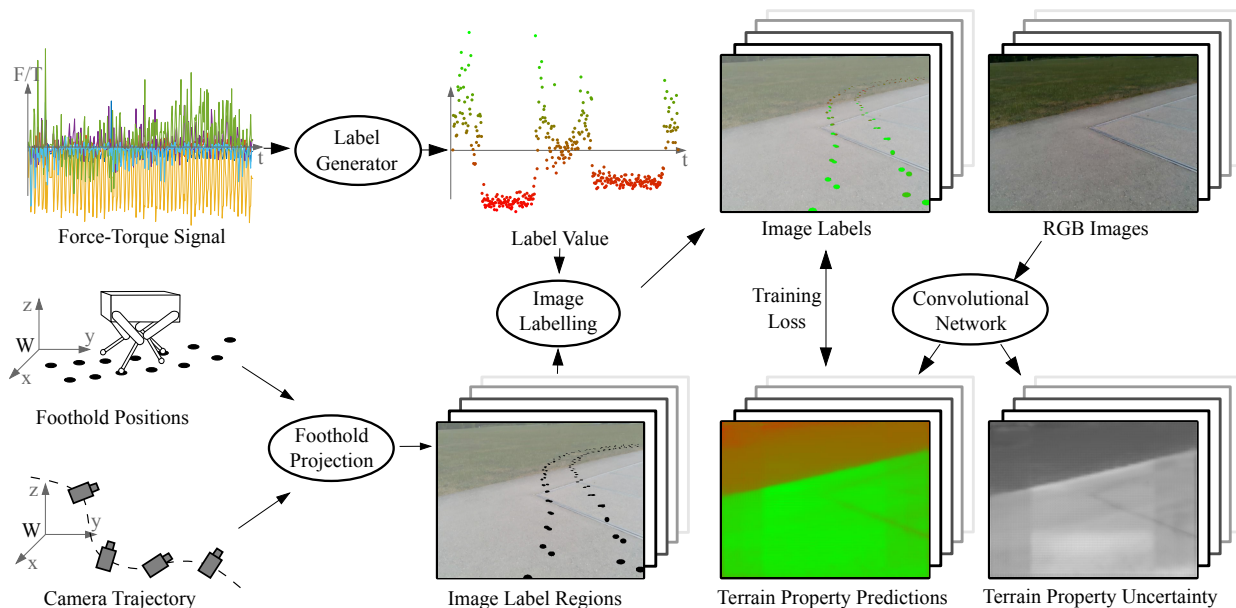


Fig. 2: Self-supervised terrain property learning pipeline. Foothold positions are projected into images using information about the robot camera trajectory and fused with label values generated from foot-mounted force-torque sensors. The resulting labels are used to train a convolutional network to predict terrain properties from camera images.

### B. Label Generation

We have experimented with two ways of generating the labels for footholds. The first one allows to learn semantic terrain segmentation with weak human-provided annotation. The second one derives the labels fully automatically from proprioceptive robot measurements and allows to predict a learned terrain property that we call *ground reaction score*. We now describe both in detail.

1) *Weakly Supervised Semantic Segmentation*: Our foothold projection system allows to annotate semantic classes in the image simply by assigning a semantic label to each time step in the sequence. A human annotator only needs to replay the recorded image stream and mark each transition between terrain types with a time stamp and a terrain type. Given this weak annotation, the terrain class labels are attached to every recorded robot state and the corresponding foothold positions in the time interval between terrain transitions. Since in practice transitions between terrains are relatively rare (on average every 86 seconds in our recordings), annotation of a video can be performed with very low effort. Examples of resulting annotated frames are shown in Figure 3a and in the supplementary video. The labels are sparse and may not be perfectly aligned with terrain transitions, but in what follows we show that it is still possible to learn a well-performing terrain segmentation system from these.

2) *Self-Supervised Ground Reaction Score Regression*: While weakly supervised manual annotation can allow for good segmentation results, it comes with downsides: the need for manual effort, the introduced ambiguity of labels on regions of terrain transitions, and the lack of actual terrain characteristics associated with the human-annotated classes. Measures derived from the internal state of the robot, e.g. cost-of-transport (CoT), are difficult to accurately associate with a single foothold, which leads to inaccurate labels along terrain

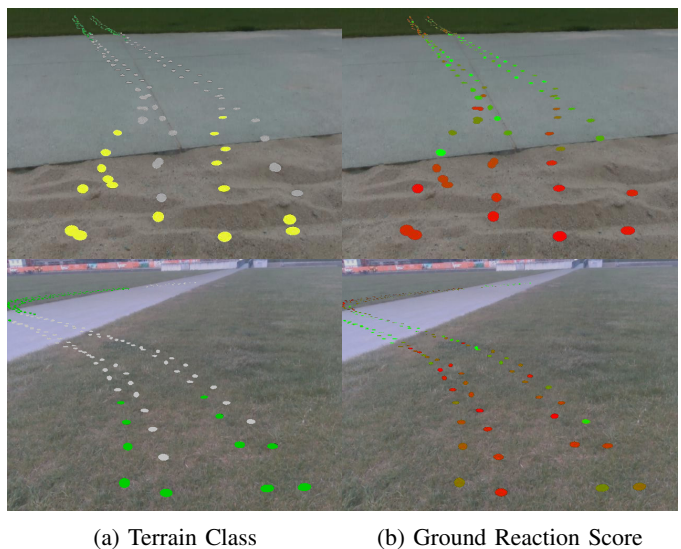


Fig. 3: Labels of both types overlaid onto the input image. Color code is explained in Figure 6. Note how ground reaction score annotation is better aligned with the terrain borders.

borders, similar to what we experienced with terrain class labels (see Figure 3a). As discussed in Section I-B, measuring physical properties during locomotion is currently not feasible. For this reason, we define an empirical terrain property, which we can measure while walking. To this end, we recorded data from six-axis force-torque (F/T) sensors mounted on the robot feet which perform the terrain interaction. This provides a time series of measurements for each leg of the robot. In order to use these for labeling, we need to extract low-dimensional navigation-relevant values from the raw measurement stream.

We start by segmenting the signal of every stance phase, including touch-down and lift-off, into segments of equal length. To each of these segments we apply the continuous wavelet transform using the Morse wavelet. This generates a

13860-dimensional feature vector, similar to that used by [20]. Because of asymmetries introduced by imperfections in the calibration of the robot limbs, we treat the signal of every leg’s sensor separately. Similarly, since the ground reaction forces while walking forward show different features than while walking backwards, we separate stance phase samples based on the longitudinal body velocity sign to account for this. We end up with eight sets of feature vectors: two per leg, split by the longitudinal direction of motion. Treating these signals separately is crucial to isolate and identify the effects of the terrain on the ground reaction forces.

We now aim to extract from this high-dimensional vector a low-dimensional physically meaningful value. On each of 8 subsets we perform principal component analysis (PCA) and rescale the principal components to have zero-mean and unit variance. We find that the first principal component of every set explains on average 37.5% of the signal variance. We select this first principal component as a measure of the terrain properties, refer to it as “ground reaction score” in what follows, and learn to regress it per-pixel from images. By construction, this score summarizes a variety of terrain properties relevant for generating ground reaction forces during the stance phase, however there is no guarantee that is strictly measuring any physical value. We confirmed in our experiments that this ground reaction score encodes terrain properties relevant for navigation and thus it can be used for trajectory planning.

### C. Network Training

Our network architecture is based on ERFNet [21], which achieves good performance in semantic segmentation while running in real time on mobile computational hardware. The network architecture is illustrated in Figure 4. Compared to the original ERFNet, we change the number of output layers to match the number of classes in our classification task or our regression task, respectively. Additionally, we introduce skip connections between intermediate decoder and encoder layers, to improve performance along terrain transitions, by adding two skip connection blocks in the decoder. In a skip connection block the tensors from encoder and decoder are concatenated, one 1D-non-bottleneck layer [21] is applied, followed by a convolution for dimensionality reduction.

For the classification task, we employ the standard cross-entropy loss function, only applied where labels are available. For the regression task, we predict both the mean  $m_i$  of the ground reaction score as well as its variance  $\sigma_i$  for each pixel  $i$  and employ a negative Gaussian log-likelihood loss [22], averaged over valid pixels  $i \in \mathcal{V}$  for which ground truth  $m_i^{gt}$  is available:

$$\mathcal{L} = \sum_{i \in \mathcal{V}} \frac{(m_i - m_i^{gt})^2}{2\sigma_i^2} + \log \sigma_i. \quad (2)$$

The image labels obtained through foothold projection are sparse and cover between 0.1% and 10.9% of image pixels. Moreover, these labels are concentrated around the center and bottom of images, owing to the mostly longitudinal motion of the robot and the parallax effect. Due to the limited receptive field of every pixel, this leads to left, right, and top areas of

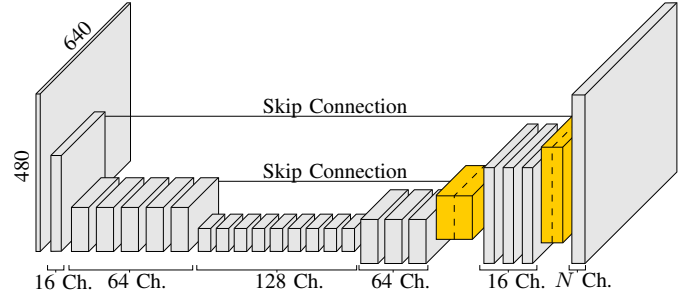


Fig. 4: Network architecture based on ERFNet [21]. Grey blocks are identical to the original architecture, yellow-colored blocks were added to facilitate skip connections.

the images having very small gradients. In order to alleviate this issue, we employ a semi-supervised learning technique called Mean Teacher [23]. We maintain two copies of the network and feed the same image to both, albeit with different image augmentations, and then enforce the output of both to be identical using an  $L2$  loss. This way we obtain a training response for all parts of the image, including the unlabeled regions.

## III. EXPERIMENTS

The proposed framework was extensively tested in various field experiments using the ANYmal quadruped robot [6]. We evaluate the terrain property prediction and demonstrate its use for path planning.

### A. Dataset

We generated a dataset by teleoperating the robot through different environments. To minimize the effect of the locomotion controller, we used the same trotting gait at all times. We recorded the data in an urban park, forest, and on farmland, under varying lighting and weather conditions. The dataset includes multiple terrain types, which are shown in Figure 5 and further explained in Section III-C. The robot is equipped with four *BOTA Rokubi Mini 1.00 USB* [24] six-axis F/T sensors, one mounted on each foot. Image data was recorded using a forward-facing *Intel Realsense ZR300* mounted at a pitch angle of  $22.3^\circ$ . The position of the feet relative to the camera was measured using joint encoders at 400 Hz. The total dataset length is 167 minutes with F/T data recorded at 100 Hz and image data recorded at 10 Hz at a resolution of  $640 \times 480$ . Out of 23 total robot sorties, we selected 4 as a validation set, while taking care that all terrain types are present in both the training and the validation set and that the same piece of terrain is not present in both. This results in the training set containing 70 822 images and the validation set containing 15 134 images. We test our approach in real robot navigation experiments.

### B. Training Details

Since images are recorded from a camera stream at 10Hz on a platform with a speed around  $0.3 \frac{m}{s}$ , they are highly correlated in time. We therefore sample every tenth image for training to reduce the dataset size.



Fig. 5: Example images of different terrain types in the dataset. *Asphalt* (a), *gravel path and grass* (b), *dirt* (c), *sand* (d)

To improve the generalization performance, we perform image augmentation on the input images. As mentioned in Section II-C, we employ the Mean Teacher learning strategy [23] where two input images with different augmentations are fed to the network and deviations between the outputs are punished. The following geometric augmentations are applied to both input images with the same parameters so that the assumption that outputs should be identical still holds:

- Horizontal flip with probability 0.5
- Random rotation uniformly sampled from  $[-5^\circ, 5^\circ]$
- Random crop to image scale uniformly sampled from  $[0.6, 1.0]$

After applying the geometric augmentation to both input images we apply individually sampled color jitter in brightness, contrast, saturation and hue to the images.

### C. Weakly Supervised Semantic Segmentation

We consider five terrain types in the classification task: *asphalt*, gravel path (hereafter referred to as *path*), *grass*, *dirt*, and *sand*. Example images of all terrains are shown in Figure 5. Note that some of the terrains look very similar in the relatively low-quality images recorded by the robot’s camera, and cannot be discriminated with naive approaches, such as color thresholding. Training labels are generated by replaying the recorded dataset and manually noting the time stamps of transitions between terrains, as described in Section II-B. With this approach, we were able to label the entire dataset in under one hour.

We select the ERFNet model with skip connections trained using the Mean teacher approach, based on the performance on the regression task, which will be discussed in the next section. We evaluate it by computing the per-class classification accuracy for each class and the mean per-class accuracy, which is the average of these per-class accuracies. The results, evaluated on the validation set, are shown in Table I. Mean per-class accuracy is 90.6%, indicating the overall high performance of the system, especially given the very sparse labels provided for training. Lower accuracy on the softer terrain types – *sand* and *dirt* – is caused by the fact that these terrain types are rare and therefore underrepresented in the dataset. A larger training dataset or re-weighting the loss for underrepresented classes would likely lead to higher accuracy.

Qualitative results of semantic terrain segmentation are shown in Figures 6b and 6c. We show the predicted class and the confidence, measured by the probability assigned by the network to the dominant class. The overall prediction quality is high, but the predictions are noisy around borders between

different terrain types. This is because of the relatively small size of the dataset and the noise in the human-annotated terrain transition timestamps. Confusion between *asphalt* and *path* terrain types is captured by low confidence, visible for instance in the center image of Figure 6c.

TABLE I: Terrain classification accuracy.

<i>asphalt</i>	<i>path</i>	<i>grass</i>	<i>dirt</i>	<i>sand</i>	Mean
93.1%	97.0%	97.3%	85.2%	80.3%	90.6%

### D. Self-Supervised Ground Reaction Score Regression

We have obtained the ground reaction score on the collected dataset as described in Section II-B2. To gain an understanding of the physical meaning of the ground reaction score, we plot the first two principal components of the terrain response in Figure 7a (left) and color each point according to its terrain class, as in Section III-C. There is a clear correlation between the terrain type and the ground reaction score (the first principal component), while the second principal component does not appear to have a correlation with the terrain type. This is further reinforced by Figure 7a (right) that demonstrates a quantitative difference in ground reaction score between different terrain types: there is a clear progression of per-terrain values, and the means are well separated in most cases. Both the mean and the variance increase with the rigidity of the terrain. A possible explanation for the increasing variance is that on hard terrain, high end-effector speeds at touchdown induce a higher peak reaction force than on soft terrain, where the impact is dampened. At lower impact speeds damping has less influence which would explain the relative proximity of ground reaction score of hard and soft terrain on the lower end. It is important to note that we only use the terrain types in Figure 7a to illustrate that the ground reaction score contains meaningful information about terrain properties, but it is learned in fully self-supervised fashion. The ground reaction score correlates with both the CoT, shown in Figure 7b, as well as the maximum z-velocity during stance phase, shown in Figure 7c. This indicates that we can use it to derive a planning cost, which encapsulates energy-efficiency as well as expected locomotion disturbances.

Qualitative results of ground reaction score prediction are shown in Figures 6d and 6e. Terrain boundaries are noticeably sharper than for the classification task and gradual terrain transitions are perceived as such. We believe this is thanks to the better alignment of the automatically annotated labels with the terrain transitions.

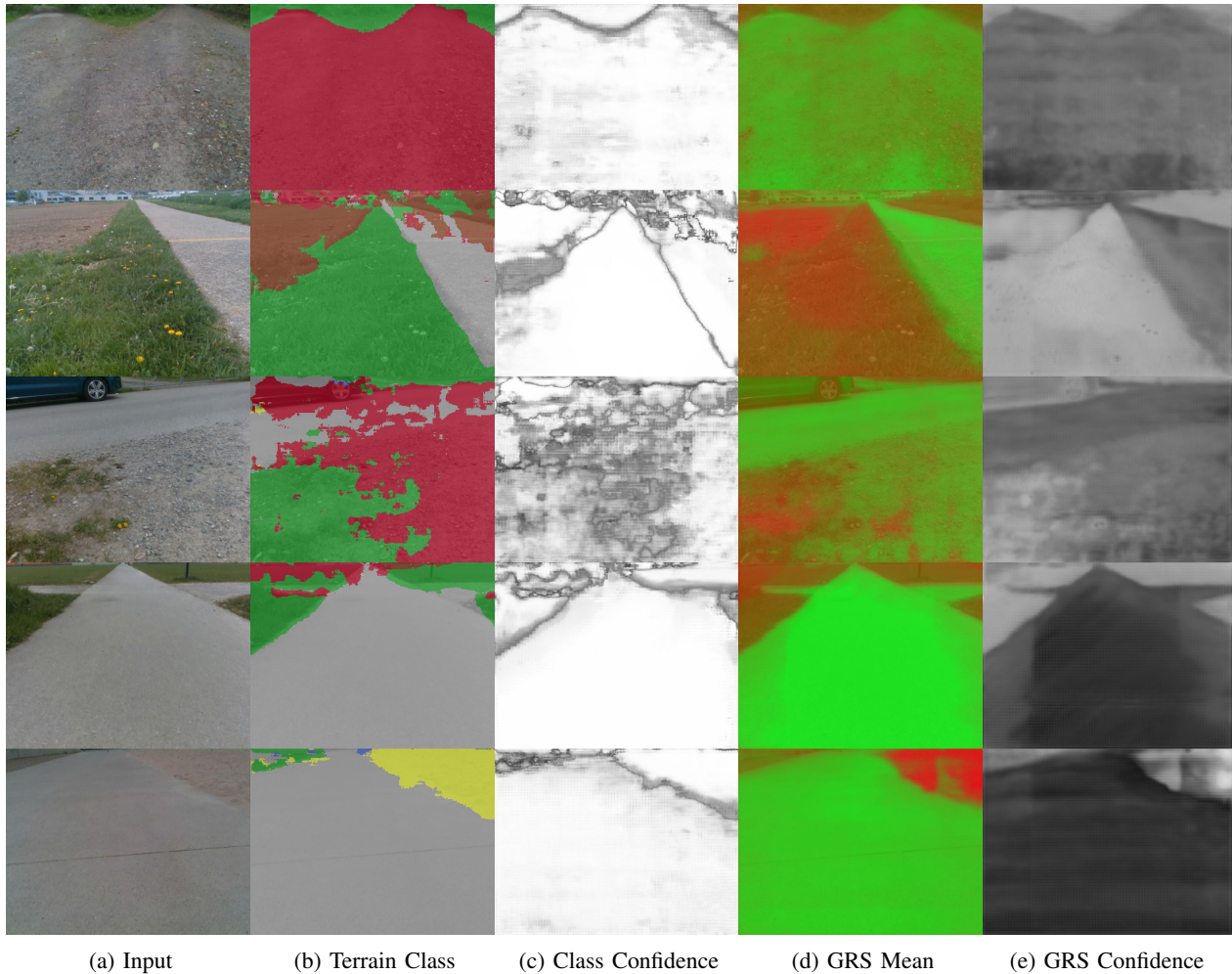


Fig. 6: Representative qualitative prediction results of weakly supervised semantic segmentation and self-supervised ground reaction score (GRS) regression. Class color code: *asphalt* - gray, *path* - red, *grass* - green, *dirt* - brown, *sand* - yellow. Ground reaction score color code: low - red, high - green. Confidence color code: low - black, high - white.

For a quantitative evaluation, we measure the effect of network modifications – skip connections and Mean Teacher – on the prediction quality. Figure 8 shows validation curves on the regression task. Both modifications improve the performance of the network in terms of final loss value and the stability of training.

#### E. Hardware Path Planning Experiments

Finally, we verify the applicability of our approach to robot navigation tasks and compare it to a geometric navigation approach [1].

1) *Experimental Setup*: Experiments were performed on the quadrupedal robot ANYmal. A *Nvidia Jetson TX2* was used for inference of the neural network, running the *ERFNet* + skip connection regression network at 6.0Hz. The planning algorithm was run on a PC equipped with a *Intel i7-4600U* CPU. One experiment was performed in a park with asphalt paths, grass, and sand pits, while another was performed on a forest path. The former is in the same general environment used for data collection while the latter is different from both training and validation locations. The experiments were conducted during late fall in gloomy lighting conditions, dif-

ferent from the sunny and overcast weather during initial data collection in spring. The terrain is flat, allowing us to focus on the terrain properties rather than geometric characteristics.

2) *Path Planning*: Path planning is performed on a 2D grid with a cell size of 10cm which is updated continuously. Its size is expanded automatically if measurements fall outside of the grid while keeping the same resolution. Every cell maintains an estimate of its ground reaction score mean and uncertainty. Ground reaction score prediction images are projected into 3D space using depth information from the depth camera and the pose estimate of the robot odometry. Depth values are clipped at 7m due to decreasing depth accuracy of the sensor. For every projected pixel a Kalman filter update step is performed in the cell it is projected into. The ground reaction score prediction is used as measured value, the uncertainty prediction is used as measurement variance. A process noise of  $\frac{0.001}{\text{Hz}}$  is applied to account for robot state estimation drift.

Motivated by Figure 7b, the negative ground reaction score is used as cost, rescaled such that all costs are positive, and the Dijkstra’s algorithm is used to find the optimal path. Unobserved grid regions have a medium cost which falls between empirically observed costs of *grass* and *sand*. We do

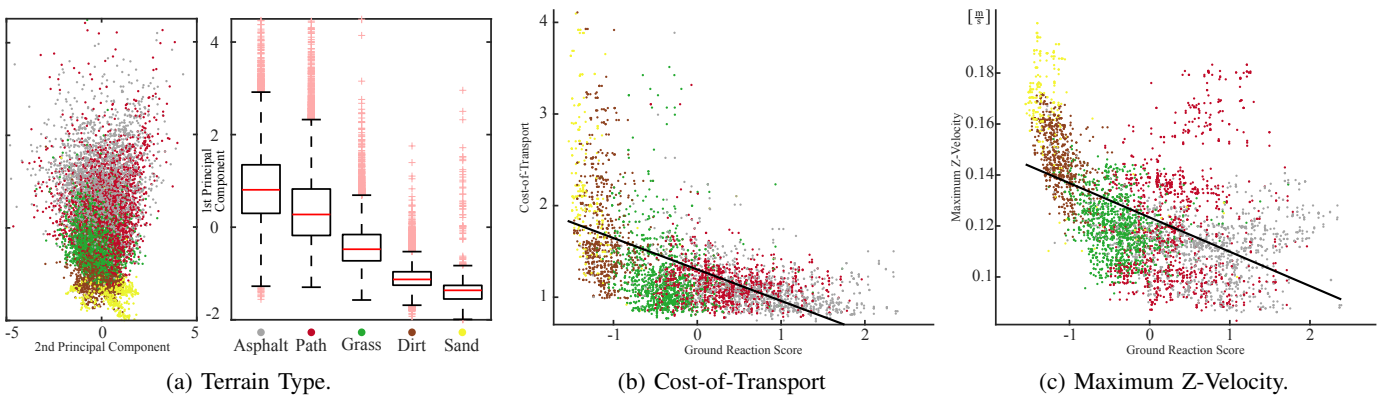


Fig. 7: Ground reaction score plotted against various indicators for terrain properties. Despite the significant intraclass variance, the per-class means of the ground reaction score for different terrains in Figure (a) are clearly separated. The 2nd principal component does not show any correlation with terrain type. Samples of Figures (b) and (c) were median filtered in time to reduce noise. Ground reaction score and CoT have a Pearson correlation coefficient of  $-0.503$ , ground reaction score and maximum z-velocity have a Pearson correlation coefficient of  $-0.516$ .

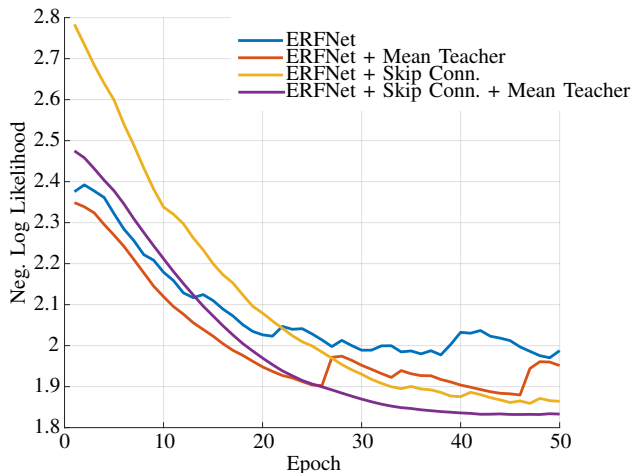


Fig. 8: Smoothed validation loss plots for *ERFNet* on ground reaction score regression data using combinations of skip connections and Mean Teacher.

not set a lethal cost because we can, by the nature of our data collection method, traverse any terrain we predict a ground reaction score for. The path is continuously replanned at 1Hz

We compare our path planner to the work of Wermelinger et al. [1], which computes a geometric traversability measure. Traversability is computed as the weighted average of terrain slope, step height and terrain variance in a local patch. The same 2D grid as for our approach with the same 10cm resolution is used.

3) *Results*: In the first task the robot had to successively navigate from one goal point to the next. After a goal point was reached, the robot was turned towards the next goal point, before resuming navigation, such that it is visible to the forward-facing RGB-D camera. The robot preferred asphalt over grass and avoided entering the sand pit, unless the goal point required it to (Points 4, 5, 6, 9), as shown in Figure 9 (left). Interestingly, in these cases, the planner does not strictly minimize the path length in sand but finds a trade-off between total path length and path length in sand. A notable behavior is also the slight detour over asphalt between points 7 and 8 instead of walking directly over grass.

The geometric planner plans direct paths from start to goal

positions, as show in Figure 9 (right). Minor steps on the border of the sand pit are the only perceived obstacle but are rightfully judged as traversable. Terrain height variance of the sand and short grass is not perceived at the grid size of 10cm. While the robot is able to follow every path planned by the geometric planner due to its advanced mobility, it is typically less safe than to follow the path along asphalt planned by our approach, due to the higher risk of failure on sand. These results qualitatively demonstrate that our approach, to learn ground reaction properties via experience of robot locomotion, enables navigation where geometric planners fail, without the need to specify arbitrary terrain classes.

In the second navigation task, the robot was set on a curvy gravel path, in a location completely different from where the training set was recorded, bordered by a grassy field on one side and a forest on the other, pictured in Figure 10. The goal position was commanded 10m in front of the robot and continuously updated to stay at this relative position. When the robot approaches a bend, the commanded pose falls on grass, but the optimal path towards it lies on the lower cost gravel path along the grass border. When the robot starts to turn to follow the path, the next goal position lies closer to the gravel than the previous one, leading the robot to eventually completing the turn. This experiment demonstrates emerging path following behavior without having an explicit notion of what a path is.

#### IV. CONCLUSION

In this work we proposed an approach for weakly supervised and self-supervised learning of terrain properties and have shown its value for prediction of terrain characteristics from RGB images which is used for robot navigation. The proposed approach opens up multiple avenues for future work. First, advanced weakly supervised learning methods could be employed to better deal with sparse image annotation provided by our labeling technique. Second, analyzing the proposed ground reaction score in detail and examining which exact terrain properties are encoded in this metric could give more insight into robot-terrain interaction during locomotion. Moreover, if other properties are derived from the F/T readings,



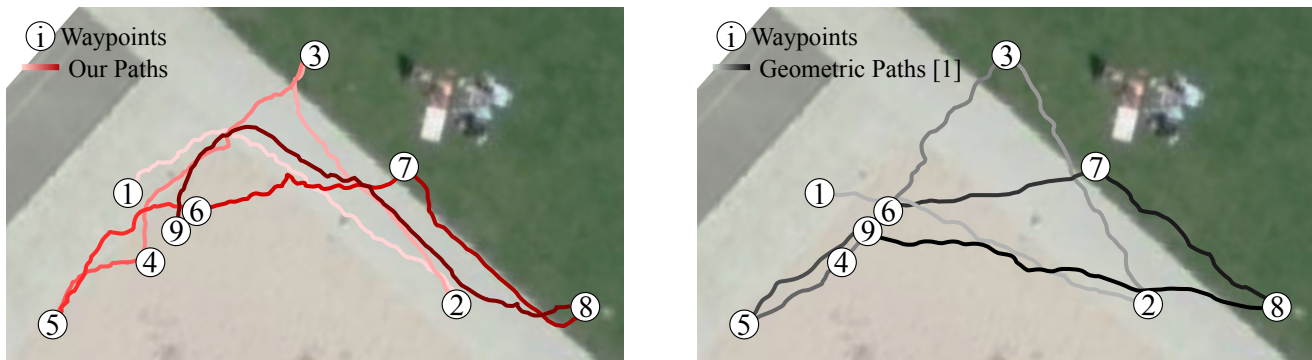


Fig. 9: Path planning over *sand*, *asphalt*, and *grass* with waypoints commanded sequentially. The color gradient (light to dark) indicates successive path segments. Our approach, using the negative ground reaction score as cost (left) avoids sand whenever possible, whereas the geometric approach by Wermelinger [1] (right) plans a direct path.

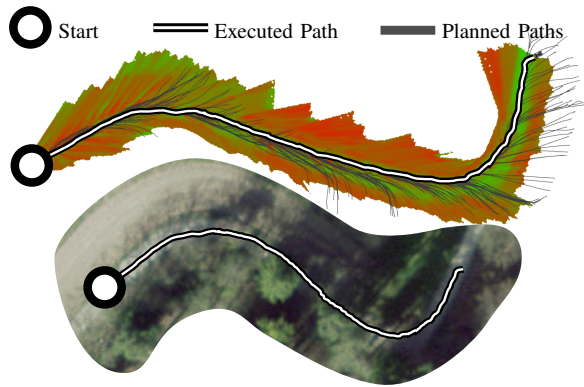


Fig. 10: Continuously commanded goal position 10m in front of the robot results in path following behavior. Cost map on top and aerial view on bottom.

they can be readily incorporated in the proposed framework. Third, extending foothold projection with depth information to respect occlusions and environment geometry could enable applications for beyond line-of-sight navigation and predicting traversable geometry, like vegetation. Fourth, employing safe exploration techniques for data collection could absolve the need for any human input to the system. Finally, while this work demonstrated the application of the approach to a legged robot, it could be generalized to other types of ground robots.

## REFERENCES

- [1] M. Wermelinger, P. Fankhauser, R. Diethelm, P. Krüsi, R. Siegwart, and M. Hutter, "Navigation planning for legged robots in challenging terrain," in *IROS*. IEEE, 2016, pp. 1184–1189.
- [2] R. O. Chavez-Garcia, J. Guzzi, L. M. Gambardella, and A. Giusti, "Image classification for ground traversability estimation in robotics," in *Int. Conf. on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 325–336.
- [3] P. Krüsi, P. Furgale, M. Bosse, and R. Siegwart, "Driving on point clouds: Motion planning, trajectory optimization, and terrain assessment in generic nonplanar environments," *Journal of Field Robotics*, vol. 34, no. 5, pp. 940–984, 2017.
- [4] B. Rothrock, R. Kennedy, C. Cunningham, J. Papon, M. Heverly, and M. Ono, "Spoc: Deep learning-based terrain classification for mars rover missions," in *AIAA SPACE 2016*, 2016, p. 5539.
- [5] D. M. Bradley, J. K. Chang, D. Silver, M. Powers, H. Herman, P. Rander, and A. Stentz, "Scene understanding for a high-mobility walking robot," in *IROS*. IEEE, 2015, pp. 1144–1151.
- [6] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch, *et al.*, "Anymal-a highly mobile and dynamic quadrupedal robot," in *IROS*. IEEE, 2016, pp. 38–44.
- [7] H. Oleynikova, Z. Taylor, M. Fehr, R. Siegwart, and J. Nieto, "Voxblox: Incremental 3d euclidean signed distance fields for on-board map planning," in *IROS*. IEEE, 2017, pp. 1366–1373.
- [8] D. Kim, J. Sun, S. M. Oh, J. M. Rehg, and A. F. Bobick, "Traversability classification using unsupervised on-line visual learning for outdoor robot navigation," in *ICRA*. IEEE, 2006, pp. 518–525.
- [9] K. Otsu, M. Ono, T. J. Fuchs, I. Baldwin, and T. Kubota, "Autonomous terrain classification with co-and self-training approach," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 814–819, 2016.
- [10] D. Barnes, W. Maddern, and I. Posner, "Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy," in *ICRA*. IEEE, 2017, pp. 203–210.
- [11] N. Hirose, A. Sadeghian, M. Vázquez, P. Goebel, and S. Savarese, "Gonet: A semi-supervised deep learning approach for traversability estimation," *arXiv preprint arXiv:1803.03254*, 2018.
- [12] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *ICRA*. IEEE, 2017, pp. 4644–4651.
- [13] C. Cunningham, W. L. Whittaker, and I. A. Nesnas, "Improving slip prediction on mars using thermal inertia measurements," *RSS*, 2017.
- [14] C. Ordóñez, R. Alicea, B. Rothrock, K. Ladyko, M. Harper, S. Karumanchi, L. Matthies, and E. Collins, "Modeling and traversal of pliable materials for tracked robot navigation," in *Unmanned Systems Technology XX*, vol. 10640. SPIE, 2018, p. 106400F.
- [15] P. Dallaire, K. Walas, P. Giguere, and B. Chaib-draa, "Learning terrain types with the pitman-yor process mixtures of gaussians for a legged robot," in *IROS*. IEEE, 2015, pp. 3457–3463.
- [16] L. Ding, H. Gao, Z. Deng, J. Song, Y. Liu, G. Liu, and K. Iagnemma, "Foot-terrain interaction mechanics for legged robots: Modeling and experimental validation," *Int. J. of Robotics Research*, vol. 32, no. 13, pp. 1585–1606, 2013.
- [17] W. Bosworth, J. Whitney, S. Kim, and N. Hogan, "Robot locomotion on hard and soft ground: Measuring stability and ground properties in-situ," in *ICRA*. IEEE, 2016, pp. 3582–3589.
- [18] H. Kolvenbach, C. Bärtschi, L. Wellhausen, R. Grandia, and M. Hutter, "Haptic inspection of planetary soils with legged robots," in *Submitted to: RA-L*. IEEE, 2018.
- [19] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
- [20] K. Walas, D. Kanoulas, and P. Kryczka, "Terrain classification and locomotion parameters adaptation for humanoid robots using force/torque sensing," in *Humanoids*. IEEE, 2016, pp. 133–140.
- [21] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convnet for real-time semantic segmentation," *Trans. on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2018.
- [22] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NIPS*, 2017.
- [23] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NIPS*, 2017, pp. 1195–1204.
- [24] "Bota systems rokubi mini 1.00 usb." [Online]. Available: <https://www.botasytems.com/products/>