Autonomous Robots (2025) 49:19
https://doi.org/10.1007/s10514-025-10202-x

Wild visual navigation: fast traversability learning via pre-trained models and online self-supervision

Matias Mattamala[1] · Jonas Frey[2,3] · Piotr Libera[2] · Nived Chebrolu[1] · Georg Martius[3,4] · Cesar Cadena[2] · Marco Hutter[2] · Maurice Fallon[1]

Received: 10 February 2024 / Accepted: 12 May 2025

Abstract

Natural environments such as forests and grasslands are challenging for robotic navigation because of the false perception of rigid obstacles from high grass, twigs, or bushes. In this work, we present Wild Visual Navigation (WVN), an online self-supervised learning system for visual traversability estimation. The system is able to continuously adapt from a short human demonstration in the field, only using onboard sensing and computing. One of the key ideas to achieve this is the use of high-dimensional features from pre-trained self-supervised models, which implicitly encode semantic information that massively simplifies the learning task. Further, the development of an online scheme for supervision generator enables concurrent training and inference of the learned model in the wild. We demonstrate our approach through diverse real-world deployments in forests, parks, and grasslands. Our system is able to bootstrap the traversable terrain segmentation in less than 5 min of in-field training time, enabling the robot to navigate in complex, previously unseen outdoor terrains.

1 Introduction

Traversability estimation is a core capability needed to allow robots to autonomously navigate in field environments. It is understood as the *affordance* (Gibson, 1979) necessary for a robot to navigate within its environment, i.e to understand which areas can be accessed and navigated through and at what cost. While the topic has been widely studied for wheeled or flying robots under the idea of occupancy mapping (Moravec & Elfes, 1985), the development of new platforms with advanced mobility skills, such as legged robots, prompts a reconsideration of current definitions of traversability, as new and more complex types of natural terrain can be traversed (Miki et al., 2022).

Existing approaches, which build upon deep neural models for semantic segmentation (Maturana et al., 2017) or anomaly detection (Wellhausen et al., 2020), have demonstrated navigation in off-road environments; however there are recurring problems with the collection and labeling of large amounts of relevant training data. Self-supervised systems have addressed this challenge by generating labeled datasets from past robot deployments, using classification carried out in hindsight (Wellhausen et al., 2019) or using predictions of the robot motion (Gasparino et al., 2022). Nevertheless, these previous methods are still trained on robot-specific datasets and subsequently deployed without further adaptation. The Learning Applied to Ground Vehicles (LAGR) program (Kim et al., 2006; Hadsell et al., 2009) was a first effort towards systems able to adapt in the field, where robots generated its own supervision signals during deployment, facilitating the training of machine learning models within off-road environments.

In this work, we present WVN, a system inspired by the aforementioned approaches to achieve self-supervised,

Matias Mattamala and Jonas Frey have contributed equally to this work.

✉ Matias Mattamala
matias@robots.ox.ac.uk

✉ Jonas Frey
jonfrey@ethz.ch

[1] Dynamic Robot Systems Group, University of Oxford, 23 Banbury Road, OX2 6NN Oxford, Oxfordshire, UK

[2] Robotic Systems Lab, ETH Zurich, Leonhardstrasse 21, 8092 Zurich, Zurich, Switzerland

[3] Autonomous Learning Group, Max Planck Institute for Intelligent Systems, Max-Planck-Ring 4, 72076 Tübingen, Baden-Württemberg, Germany

[4] Computer Science Department, University of Tübingen, Maria-von-Linden-Strasse 6, 72076 Tübingen, Baden-Württemberg, Germany

Published online: 18 July 2025

online traversability estimation, solely requiring a few minutes of demonstrations in the field. It combines visual input and proprioceptive information to generate supervision signals while the robot operates, enabling it to simultaneously train the traversability model and use it for online inference (Fig. 1). A key idea in WVN is exploiting high-dimensional features from pre-trained models. This simplifies the learning task while also exploiting the semantic correspondences implicitly learned by these models via offline self-supervised training on large datasets.

This article extends the system presented by Frey et al. (2023), addressing some of the limitations raised in the original formulation and introducing additional features for system integration and field deployment. The contributions are:

1. **An online, multi-camera self-supervision pipeline** that extends the original approach by enabling the use of multiple vision sources for supervision and inference.
2. **The use of pre-trained models** as feature extraction backbones, namely DINO-ViT Caron et al. (2021) and, in addition to our previous work, STEGO Hamilton et al. (2022). We demonstrate that this eases the overall traversability prediction training process.
3. **A feature sub-sampling strategies** to efficiently process pixel-wise high-dimensional features. We present additional strategies to SLIC Achanta et al. (2012) used in our previous work, which further exploit the semantic priors already encoded in the features.
4. **Real-world experiments**, demonstrating hardware integration and onboard execution of WVN with one and multiple cameras, achieving real-world navigation tasks after minutes of training.
5. **Open-source implementation** of the WVN system with ROS Quigley et al. (2009) integration, with a set of baseline model weights trained on diverse environments.

2 Related work

2.1 Traversability from geometry

Classical approaches for traversability estimation analyze the geometry of the environment using 3D sensing (Moravec & Elfes, 1985). Solutions from the DARPA SubT Challenge Chung et al. (2023), used representations such as point clouds and meshes to evaluate navigational metrics like risk or stepping difficulty (Cao et al., 2022; Fan et al., 2021).

However, a purely geometric analysis has proven insufficient and data-driven methods have bridged this gap by learning platform-specific traversability from real data or simulations. Chavez-Garcia et al. (2018) used simulations of a ground robot moving on an elevation map. Yang et al. (2021) extended this approach for legged platforms, capturing the risk of failure, energy cost and time required for navigation. Recently, Frey et al. (2022) expanded this approach to volumetric data and massive parallelization in data collection from simulation. Nevertheless, using geometry-only could be insufficient to represent natural growth such as high grass, branches or bushes.

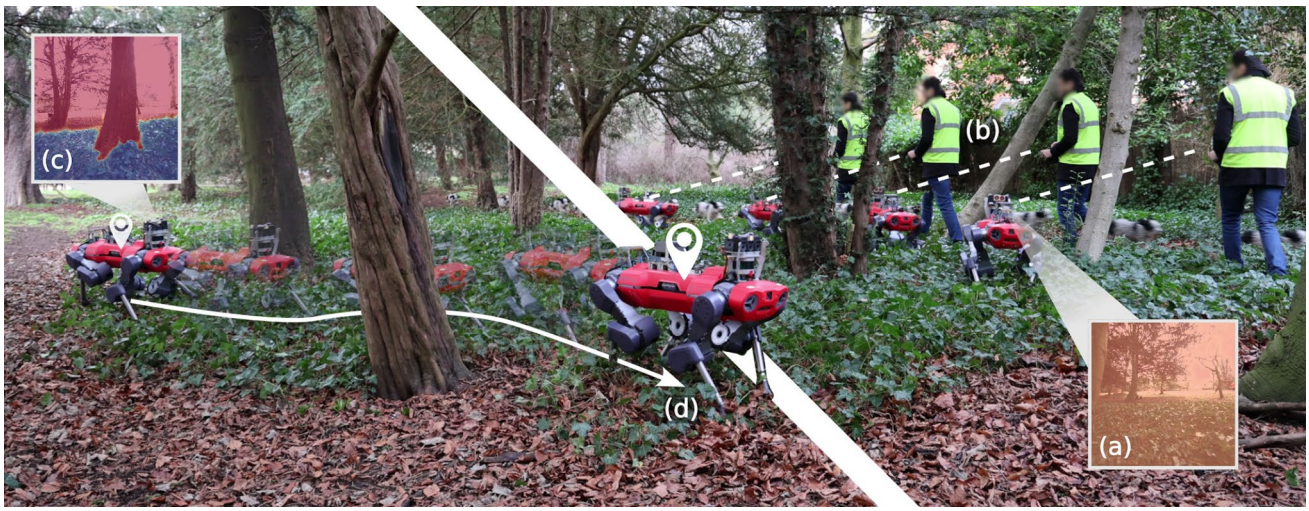

Fig. 1 Wild visual navigation (WVN) learns to predict traversability from images via online self-supervised learning. Starting from a randomly initialized traversability estimation network without prior assumptions about the environment (**a**), a human operator drives the robot around areas that are traversable for the given platform (**b**). After a few minutes of operation, WVN learns to distinguish between traversable (blue filled square) and untraversable (red filled square) areas (**c**), enabling the robot to navigate autonomously and safely within the environment (**d**)

2.2 Traversability from semantics

Semantic segmentation methods aim to address the aforementioned challenges by assigning semantic classes to the representations, with different navigation costs. Bradley et al. (2015) presented a scene understanding system for a legged platform, trained and evaluated using geographically diverse data. Maturana et al. (2017) demonstrated autonomous off-road navigation using semantics projected onto 3D map around a wheeled platform. Schilling et al. (2017) used semantically segmented features that were classified into fixed classes using a random forest classifier. Belter et al. (2019) developed a semantic terrain analysis module to guide a whole-body planner in a multi-legged platform. Recently, Shaban et al. (2022) presented an approach for off-road navigation that learns a dense traversability map from sparse point-clouds, while Cai et al. (2022) mapped terrain semantics to vehicle speed profiles as a proxy for traversability.

Most of these methods rely on pre-trained or fine-tuned semantic segmentation models with pre-defined class labels. In this work we exploit the advances in self-supervised models, such as DINO-ViT (Caron et al., 2021), to determine semantically similar regions without manual supervision.

2.3 Traversability from self-supervision

Self-supervised methods address the challenges of predefined classes and costs by using past robot experiences (Kim et al., 2006; Bajracharya et al., 2009). Modern methods rely on deep neural networks trained from weakly supervised data, and the supervision depends on the robot platform. Wellhausen et al. (2019) used the reprojected footholds from a legged robot to provide supervision of walkable areas; Zürn et al. (2021) exploited sounds produced by a wheeled robot moving on different terrain as a proxy for supervision; Gasparino et al. (2022) instead used the receding-horizon trajectory of a Model Predictive Controller (MPC).

BADGR Kahn et al. (2021) predicted future robot states and events from images, including its position and crash probability, which can be interpreted as traversability. TerraPN (Sathyamoorthy et al., 2022) used odometry and IMU signals as supervision to learn a traversability model in 25 min—including data collection and learning. Guaman Castro et al. (2023) predicted traversability based on IMU supervision conditioned on the velocity of the robot. Recently, Jung et al. (2024) presented a system that shares with WVN the use of pre-trained models for self-supervision.

While WVN follows similar self-supervision strategies, we aim for concurrent supervision signal generation and learning achieving online adaptation in the field.

2.4 Traversability from anomalies

Anomaly detection methods are motivated by the imbalance of positive and negative samples in self-supervised methods. Instead of training a discriminative model of traversability, they focus on learning generative models of the traversed terrain. This distribution is used as a proxy to set out-of-distribution (OOD) inputs as untraversable.

Richter and Roy (2017) trained an autoencoder to predict OOD scenes from images, switching to safer navigation behaviors when traversing novel environments. Wellhausen et al. (2020) used multi-modal sensing from haptics, vision and depth to identify anomalies such as flames and water reflections. Schmid et al. (2022) show-case the effectiveness of anomaly detection for identifying safe terrain from vision in an off-road driving scenario. Further, Ji et al. (2022) formulated a proactive anomaly detection approach that evaluated candidate trajectories for local planning depending on their probability of failure.

While we do not explicitly use anomalies to determine traversability, we do use it as a confidence metric to leverage the sparse supervision signals, as it has also been recently explored by Seo et al. (2023).

3 Methods

3.1 System overview

The objective of this work is to design a navigation system that estimates dense traversability scores of the terrain from RGB images. We use a neural network model—Multi-Layer Perceptron (MLP)—trained online, in a self-supervised manner, from supervision signals generated by a robot interacting with its environment. The system should require only a brief demonstration from a human operator for data collection and learning.

WVN is implemented as a two-processes system that run at different rates, as shown in Fig. 2. The *feature extraction & inference process* processes images from different cameras, extracts visual features, and performs inference to predict the traversability score pixel-wise. The *online learning process* estimates traversability scores from proprioception, generates the supervision signals from hindsight information, and executes an inner training loop to update the traversability prediction model. While the former supplies visual features for training as images are processed, the latter provides the most updated learned model at a fixed time rate.

The main definitions used in the rest of the paper are summarized in Table 1, and the technical details of each process are provided as follows.

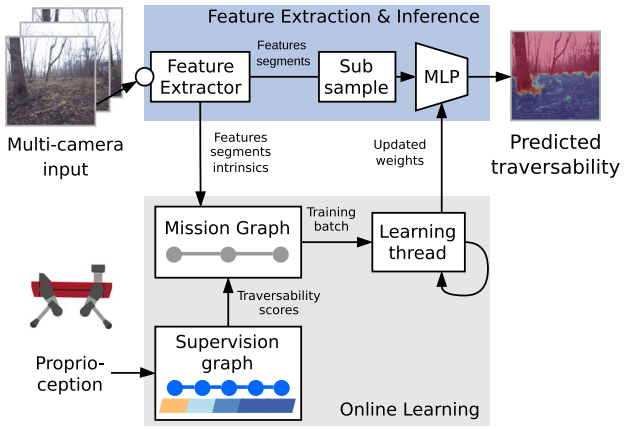

Fig. 2 System overview: WVN only requires monocular RGB images, odometry, and proprioceptive data as input, which are processed to extract features and supervision signals used for online learning and inference of traversability (see Sect. 3)

Table 1 Main definitions used in this work

Symbol	Definition
I	RGB image with height H and width W
F	Feature map with dim. $E \times H \times W, E = 90$ or 384
M	Weak segmentation mask with height H and width W
S	Reprojected supervision with dim. $H \times W \in [0, 1]$
τ	Traversability score $\in [0, 1]$
f_n	Per-segment embedding of dim. $E = 90$ or 384
τ_n	Per-segment traversability score

3.2 Feature extraction and inference

3.2.1 Multi-camera processing

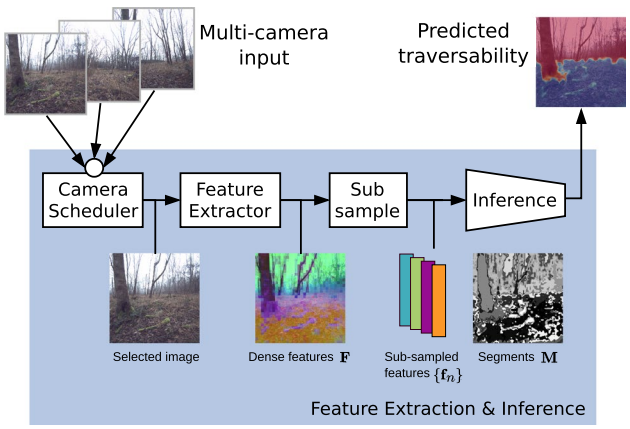

Fig. 3 Feature extraction and inference process: the camera scheduler module (Sect. 3.2.1) selects one camera from the available pool, and provides the RGB image to the feature extractor module (Sect. 3.2.2). This extracts dense visual features F using pre-trained models. Next, the sub-sample module produces a reduced set of embeddings $\{\mathrm{f}_n\}$ using a subsampling strategy based on a weak segmentation system (Sect. 3.2.3). Lastly, the inference module predicts traversability from the image using the embeddings

While the original WVN was designed for single-camera processing, it presented limitations during navigation, constraining it to motions within the camera Field of View (FoV).

We enabled multi-camera operation by developing a camera scheduler based on the weighted round-robin algorithm Katevenis et al. (1991). This ensured that the system only processes a single camera at a time, depending on priorities specified by the cameras being used for training and inference, or inference-only (Fig. 3).

3.2.2 Feature extraction

After a camera is selected in each cycle of the scheduler, the following steps are camera-agnostic. Given an RGB image I, we extract dense, pixel-wise visual feature maps (*embeddings*) F. In contrast to previous works based on fine-tuned Convolutional Neural Networks (CNNs), we rely on recent self-supervised network architectures to generate high-dimensional features that encode meaningful semantics without requiring labels.

In our implementation, we integrated the self-supervised DINO-ViT (Caron et al., 2021), which provides 384-dimensional pixel-wise feature embeddings. We additionally considered STEGO (Hamilton et al., 2022), which uses a DINO-ViT backbone with additional layers trained with contrastive learning, providing 90-dimensional features and segmentation mask. Before extracting the features, we resize the input images to a resolution of 224×224. The resulting dense features F are too large to be stored in GPU memory for online training. Hence, we introduced feature sub-sampling strategies to reduce the dimensionality of F.

3.2.3 Feature sub-sampling

We implemented different sub-sampling strategies to reduce the number of pixel-wise embeddings from 224×224 to a reduced set of $\sim$100 embeddings $\{\mathrm{f}_n\}$. The strategies use a weak segmentation system to partition the image into a set of segments M, and then average the embeddings within each segment:

- **SLIC:** In our original implementation and inspired by Lee et al. (2017), we explored the use of SLIC (Achanta et al., 2012) to reduce the dimensionality to $\sim$100 segments per image. While they are fast to compute, they have the disadvantage of being based on texture only, not necessarily grouping pixels by semantics.
- **STEGO:** It provides class-free segments, which implicitly encode semantic affinity, which directly defines the segments M.

- **Random:** We randomly select a set of 100 embeddings from the feature map. In this case, we have no segments but the feature locations only.

The original STEGO implementation considers the task of semantic segmentation based on a fixed set of classes across a full dataset. To assign each pixel to a semantic class, the authors compute prototype feature vectors across a training dataset offline. A pixel can then be assigned to a semantic class based on its cosine-similarity embedding with respect to the identified prototype feature vectors. This is not applicable in our scenario, where we do not consider a fixed set of classes, nor can identify suitable prototypes vectors before the mission, given that these would strongly depend on the deployment environment. Instead, we compute a fixed number of prototype features per image using KNN-clustering, which guarantees a fixed number of segments per image. Figure 4 illustrates qualitative examples of the different segments produced by the SLIC and STEGO methods, and in Sect. 5.3.2, we experimentally test these different strategies. We open-sourced the full re-implementation of the modified STEGO version.[1]

After this sub-sampling step, the subset of embeddings $\{\mathrm{f}_n\}$ and their image locations, segments M and camera intrinsics are shared with the *online learning* process for training (Sect. 3.3).

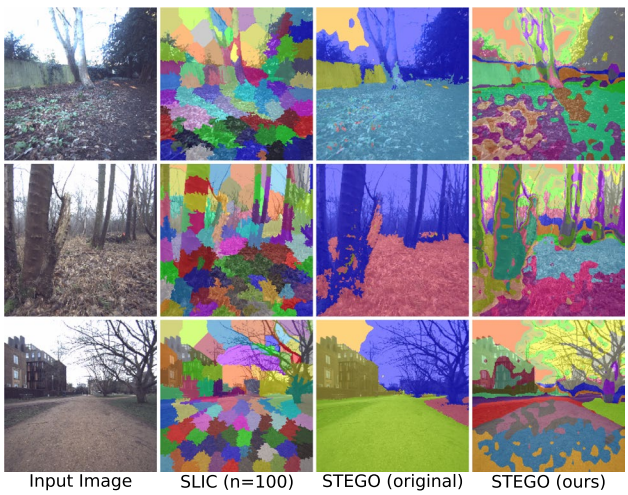

Fig. 4 Comparison feature segmentation methods for 3 example images. SLIC over-segments the image, but fails to construct semantically coherent segments (e.g. top row merging fence and ground into a single segment). The STEGO segmentation aligns with the semantics, but the computation of prototype vectors across a full dataset limits the number of semantic classes, leading to merging of two semantic classes into a single segment (grass and walkway, bottom row). Our modified version of STEGO, over-segments the image but still provides semantically meaningful segments without pre-setting prototype vectors before deployment

[1] GitHub: https://github.com/leggedrobotics/self_supervised_segmentation.

3.2.4 Inference

Lastly, this process provides predictions of traversability for all the incoming images from the different cameras. This is achieved by inferencing the MLP model, which is updated at a fixed rate by the online learning process (Sect. 5.1). The model predicts traversability from the embeddings $\{\mathrm{f}_n\}$ using two different approaches:

- **Segment-wise inference:** This is the approach implemented originally Frey et al. (2023), which predicts a traversability score τ_n for each embedding f_n, and assigns the same score for all the pixels corresponding to the given segment.
- **Pixel-wise inference:** Alternatively, we predict fine-grained traversability from the dense features F, given that the MLP forward pass can be executed with low-latency for a batch of features.

Section 5.3.2 provides qualitative examples of the improvements that each method provides when the system is deployed in different natural environments.

3.3 Online learning

3.3.1 Traversability score generation

Defining which terrain is traversable or not depends on the capabilities of the specific platform. We define a continuous *traversability score* $\tau \in [0, 1]$, where 0 is untraversable and 1 is fully traversable. We use the terrain *traction* Cai et al. (2023), which measures the discrepancy between the robot's current linear velocity as estimated by the robot (v_x, v_y), and the reference velocity command $(\bar{v}_x, \bar{v}_y)$ given by an external human operator or planning system.

We define the mean squared velocity error as:

$$v_{\text{error}} = \frac{1}{2}\left((\bar{v}_x - v_x)^2 + (\bar{v}_y - v_y)^2\right) \in \mathbb{R} \tag{1}$$

We smooth v_{error} with a 1-D Kalman Filter before passing it through a sigmoid function to obtain a valid traversability score:

$$\tau = \text{sigmoid}\left(-k\left(v_{\text{error}} - v_{\text{thr}}\right)\right) \tag{2}$$

with k the steepness of the sigmoid, and v_{thr} the midpoint of the sigmoid that assigns a traversability score of 0.5. These values are calibrated depending on the motion specifications of each platform and determine how the velocity error is stretched to the [0, 1] interval.

3.4 Supervision and mission graphs

The system generates supervision signals by accumulating information in hindsight, during operation. Our approach is inspired by graph-based SLAM pipelines that leverage both local and global graphs to integrate measurements: we maintain a *Supervision Graph* to store short-horizon traversability data, and a global *Mission Graph* which stores the generated training data during a mission, shown in Fig. 5.

3.4.1 Supervision graph

The supervision graph stores within its nodes information about the current time, robot pose, and estimated traversability score (Sect. 3.3.1). This graph is implemented as a ring buffer, which only keeps a fixed number of nodes N_{sup}, separated from each other by a distance d_{sup}.

The stored information is a footprint track with traversability scores τ. It is used to associate traversability scores with features by projecting the footprint track into the previous camera viewpoints.

3.4.2 Mission graph

The mission graph stores all the information required for online training. The mission nodes are added to the graph after feature extraction if the distance with respect to the last added node is larger than d_{mis}. Each mission node contains the RGB image I, the weak segmentation mask M and per-segment features f_n with their corresponding traversability supervision τ_n.

3.4.3 Supervision generation

When a new mission node is added, we update the supervision labels τ_n by reprojecting the footprint track and corresponding traversability scores τ onto all the images of the mission nodes within a fixed range (Fig. 5b).

Each mission node then has an auxiliary image with the reprojected path, S. We use the weak segmentation mask M to assign per-segment traversability supervision values τ_n by averaging the score over each segment. Segments that do not overlap with the reprojected footprint track are set to zero (i.e untraversable). The outcome are pairs of per-segment features f_n and traversability score τ_n for each mission node, used for training.

3.5 Traversability and anomaly learning

We train a small neural network in an online fashion that determines the feature traversability score τ_n from a given segment feature f_n. This reduces the visual traversability estimation problem to simple regression task. Further, we model the uncertainty about the unvisited (and hence, unlabeled) areas by using anomaly detection techniques to bootstrap a confidence estimate.

First, we elaborate on how a confidence score for a feature is obtained, then we describe the traversability estimation learning task.

3.5.1 Confidence estimation

To obtain a segment-wise confidence estimate, we aim to learn the distribution over all traversed segment features f_n. An encoder-decoder network $f^{\theta_r}_{\text{reco}}$ is trained to compress the segment feature f_n into a low dimensional latent space and reconstruct the original input features f_n. The reconstruction loss is given by the Mean Squared Error (MSE) between the predicted features and the original feature compute over all channels E:

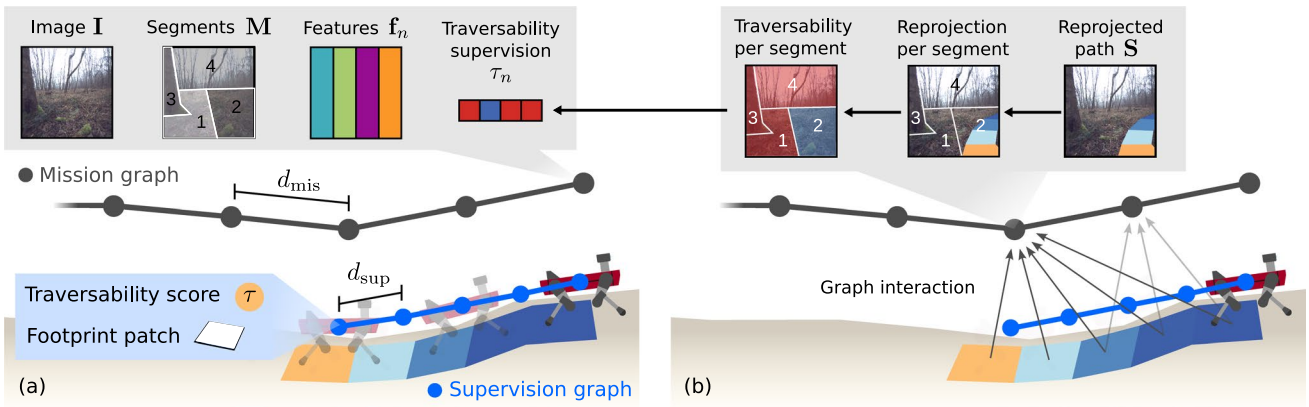

Fig. 5 Supervision and mission graphs: **a** Information stored in each graph over the mission. While the Supervision Graph only stores temporary information about the robot's footprint in a sliding window, the Mission Graph saves the data required for online learning over the full mission. The color of the footprint patches indicates the generated traversability score. **b** The interaction between graphs updates the traversability in the mission nodes by reprojecting the robot's footprint and traversability scores

$$\mathcal{L}_{\text{reco}}(\mathrm{f}_n) = \begin{cases} \frac{1}{E}\sum_e \|f^{\theta_r}_{\text{reco}}(\mathrm{f}_{n,e}) - \mathrm{f}_{n,e}\|^2 & \text{if traversed,} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This ensures that the network only learns to reconstruct the embeddings that are labeled, in an anomaly detection fashion. Consequently, the trained network reconstructs known (*positive*) feature embeddings, i.e. similar to the traversable segments, with small reconstruction loss; feature embeddings of unknown (*anomalous*) segments the network was never tasked to reconstruct, such as trees or sky, induce a high reconstruction loss.

The unbounded reconstruction loss $\mathcal{L}_{\text{reco}}$ for a segment is mapped to a confidence measure $c(\mathcal{L}_{\text{reco}}) \in [0, 1]$ by first identifying the mode of the traversed segment losses. For this we fit a Gaussian distribution $\mathcal{N}(\mu_{\text{pos}}, \sigma_{\text{pos}})$ over the reconstruction losses per batch of the traversed segments (i.e, positive samples):

$$\mu_{\text{pos}} = \frac{1}{n_{\text{trav}}} \sum_{n \in \mathcal{T}} \mathcal{L}_{\text{reco}}(\mathrm{f}_n), \tag{4}$$

$$\sigma_{\text{pos}} = \sqrt{\frac{1}{n_{\text{trav}}} \sum_{n \in \mathcal{T}} \left(\mathcal{L}_{\text{reco}}(\mathrm{f}_n) - \mu_{\text{pos}}\right)^2} \tag{5}$$

with $\mathcal{T}$ being the set of segments that were traversed, i.e. have a valid traversability score τ_n computed from robot sensing data, and n_{trav} is the total number of traversed segments. We set the segment confidence to 1 if the loss of the segment is smaller than μ_{pos} and otherwise we set it by evaluating the unnormalized Gaussian likelihood:

$$c(\mathcal{L}_{\text{reco}}(\mathrm{f}_n)) = \exp\left(\frac{(\mathcal{L}_{\text{reco}}(\mathrm{f}_n) - \mu_{\text{pos}})^2}{2(\sigma_{\text{pos}}\, k_\sigma)^2}\right), \tag{6}$$

where we introduce the tuning parameter k_σ, which allows to scale the confidence.

3.5.2 Traversability estimation

We train a small network $f^{\theta_t}_{\text{trav}}$ with a single channel output to regress on the provided segment traversability score τ. For the untraversed segments with unknown traversability score, we follow a conservative approach setting $\tau = 0$ but using the confidence score to scale their overall contribution. The loss for traversability estimation is computed using the confidence-weighted MSE:

$$\mathcal{L}_{\text{trav}}(\mathrm{f}) = \underbrace{\sum_{n \in \mathcal{T}} \|f^{\theta_t}_{\text{trav}}(\mathrm{f}_n) - \tau_n\|^2}_{\text{Contribution of traversed } (labeled) \text{ segments}} \tag{7}$$

$$+ \underbrace{\sum_{n \in \mathcal{T}^C} (1 - c(\mathrm{f}_n))\, \|f^{\theta_t}_{\text{trav}}(\mathrm{f}_n) - 0\|^2}_{\text{Contribution of untraversed segments}}, \tag{8}$$

with $\mathcal{T}$ the set of traversed segments; $\mathcal{T}^C$ is the complement set of untraversed segments. This formulation enables the learning process to "overwrite" previously unknown samples as new data is used for training:

- If the segment n was traversed: it will contribute to the loss using the assigned traversability score: $\mathcal{L}_{\text{trav}}(\mathrm{f}_n) = \|f^{\theta_t}_{\text{trav}}(\mathrm{f}_n) - \tau_n\|^2$
- If the segment n was untraversed and it does not resemble a positive sample: its confidence will be low $c(\mathrm{f}_n) \to 0$ and $\mathcal{L}_{\text{trav}}(\mathrm{f}_n) \to \|f^{\theta_t}_{\text{trav}}(\mathrm{f}_n) - 0\|^2$
- If the segment n was untraversed but it does resemble a positive sample: its confidence $c(\mathrm{f}_n) \to 1$ and $\mathcal{L}_{\text{trav}}(\mathrm{f}_n) \to 0$, effectively not contributing to the loss anymore. This motivates the network to learn the traversability score measured by physically interacting with the segment as a opposed to being too pessimistic.

As we aim to provide the estimated traversability as input for a local planning system, we automatically define a threshold to determine the traversable and untraversable areas. We select a traversability threshold τ_{thr} by measuring the current performance of the system in a self-supervised manner. We compute the Receiver Operating Characteristic (ROC) throughout training by classifying all segments with confidence under 0.5 as negative and traversed segments as positive labels. Then, we decide on the traversability threshold only by setting the desired False Positive Ratio (FPR).

3.5.3 Implementation details

We implemented $f^{\theta_r}_{\text{reco}}$ and $f^{\theta_t}_{\text{trav}}$ as a two-layer MLPs with [256, 32] unit dense layers and ReLU non-linear activation functions. Both networks share the weights of the hidden layers. $f^{\theta_r}_{\text{reco}}$ has a reconstruction head with E output neurons and $f^{\theta_t}_{\text{trav}}$ a single channel traversability head followed by a sigmoid activation. The 32-channel hidden layer functions as the bottleneck of the encoder-decoder structure. The total loss per segment during training is given by:

$$\mathcal{L}_{\text{total}}(\mathrm{f}) = w_{\text{trav}}\mathcal{L}_{\text{trav}}(\mathrm{f}) + w_{\text{reco}}\,\mathcal{L}_{\text{reco}}(\mathrm{f}). \tag{9}$$

with w_{trav} and w_{reco} allowing to weigh the traversability and reconstruction loss respectively. We used Adam (Kingma & Ba, 2015) to jointly train the networks with a fixed constant learning rate of 0.001. For a single update step, 8 valid mission nodes are randomly chosen to form a data batch, where we defined a node as valid if at least a single segment

of the node has non-zero traversability score. For all our experiments we set $k_\sigma = 2$, $w_{\text{trav}} = 0.03$, $w_{\text{reco}} = 0.5$ and use a maximum FPR of 0.15 to determine the traversability threshold. Please refer to our previous publication (Frey et al., 2023) for ablation studies of the different parameter and design choices.

4 Closed-loop integration

We integrated the learned traversability estimate into a standard navigation pipeline to achieve autonomous navigation with a quadrupedal platform. The details are explained as follows.

4.1 Local terrain mapping

In order to use the predicted traversability for navigation tasks, we used an open-source terrain mapping framework (Miki et al., 2022; Erni et al., 2023) that produced a robot-centric 2.5D elevation map from the onboard depth cameras and LiDAR sensing. The framework enabled the fusion of the predicted traversability images via raycasting, taking into account the occlusions with the terrain, as well as temporal fusion of the traversability information via exponential averaging.

4.2 Local planning

We used the projected visual traversability as a cost map for a reactive local planner (Mattamala et al., 2022) to generate a $\mathrm{SE}(2)$ twist command to drive the robot towards a goal while avoiding untraversable terrain. The twist command was the input to a robust learning-based locomotion controller (Miki et al., 2022), which is able to traverse rough terrain typically inaccessible to wheeled robots.

4.3 Autonomous path following

Lastly, we implemented a navigation strategy to guide the robot in path-like environments. The method continuously spawned new goals in front of the robot by finding the furthest traversable position in the local terrain map, within the FoV of the front-facing camera. This strategy was used to motivate simple autonomous navigation and exploration without requiring a global planner and a large-scale representation of the environment.

5 Experiments

5.1 Platform description

For our experiments we used the ANYbotics ANYmal C and D legged robots. In both configurations the robots were equipped with an additional NVidia Jetson Orin AGX. We used the manufacturer's state estimator to obtain $\mathrm{SE}(3)$ pose and body velocity measurements. The LiDAR and depth cameras available on the robots were only used for the local terrain mapping module (Sect. 4.1).

For the ANYmal C experiments, we used a single global shutter, wide FoV camera from the Sevensense Alphasense Core unit. For the ANYmal D experiments, we used the RGB images from the integrated front and rear wide-angle cameras.

WVN was implemented in pure Python code using PyTorch Paszke et al. (2019) and ROS 1 Quigley et al. (2009). Both processes ran on the Jetson Orin, and were implemented as separate ROS nodes. Inter-process communication was implemented using ROS publisher-subscriber paradigm, while the trained model weights were shared via write-read operations every 5 s for simplicity.

5.2 Real-world deployments

We executed different deployments to validate WVN in real environments in terms of adaptation to new scenes, the advantages of the visual traversability estimation compared to purely geometric, and autonomous navigation demonstrations.

5.2.1 Fast adaptation on hardware

Our first experiment involved teleoperating the ANYmal C robot around 3 loops in University Parks, Oxford, UK, to evaluate the fast adaptation capabilities of WVN when walking on grass and dirt, on open areas, and around trees.

Figure 6 illustrates the main outcomes of the experiment, showing that the system learned to predict robot-specific traversability over the 3 loops while running onboard. Section (a) shows how the robot starts with a very poor segmentation after 9 steps of training (21 s) but this greatly improves after 800 steps (2 min), where it can correctly segment the dirt as traversable terrain while keeping the tree untraversable. Similar behavior occurs in section (b) in which the segmentation is conservative at the beginning but it extends across the other grass patches in later iterations. Section (c) also illustrates some issues related to the SLIC segmentation, as some segments of the wooden wall (step 1186) are incorrectly clustered with patches of the grass, which is not observed in the other captures.

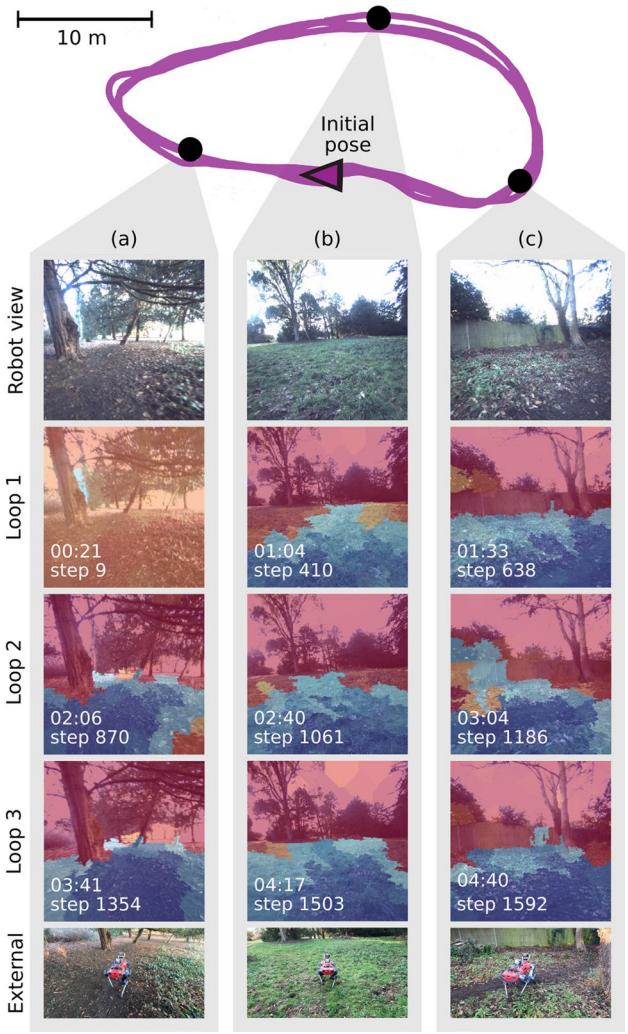

Fig. 6 Adaptation on real hardware: We tested the online adaptation capabilities of our system by teleoperating the robot to complete 3 loops in a park (top, route shown in purple filled square). The columns show different parts of the loop (**a**–**c**); each row displays the improvement of the traversability estimate over time and training steps

5.2.2 Benefits of visual traversability versus geometric methods

Our second experiment aimed to illustrate the advantages of visual traversability estimation in challenging natural environments. We teleoperated the ANYmal C robot around high grass, loose branches, and bushes in Wytham Woods, Oxford, UK. Figure 7, bottom right, shows a representative shot of the experiment, the forward-facing camera image and WVN's prediction.

To compare the different traversability methods, we used the terrain mapping module (Sect. 4.1), as it allowed us to compare geometry-only and visual traversability. We compared against two geometric methods that are real-time capable and have been used in previous works:

- Geometric method based on heuristics such as height and slope of the terrain Wermelinger et al. (2016).
- Geometric method based on a learned model of traversability, which is part of the terrain mapping system Miki et al. (2022).
- Visual traversability provided by WVN, raycasted onto the terrain map.

The geometric methods only require an elevation representation of the surface to determine traversability from the 2.5D geometry. For WVN we executed a training procedure by driving the robot around the environment for a few minutes, only using images from the forward-facing camera.

Figure 7 illustrates the output *traversability map* obtained by all the methods (bottom), as well as the corresponding SDFs generated from them (top). The geometric methods correctly determine the trees as untraversable areas. Our system is also able to successfully discriminate the trees, confirming the findings observed in Sect. 5.2.1. Furthermore, the advantages of WVN are observed in high-grass areas, which are represented as elevation spikes in the map that are classified as untraversable by the geometric methods but not by our visual traversability estimate. These differences become more evident in the SDFs where all the areas with low traversability scores become obstacles.

5.2.3 Point-to-point autonomous navigation between trees

We executed closed-loop navigation tasks to demonstrate that WVN can easily adapt to a new environment, and the learned traversability estimate can be used to deploy the robot autonomously.

We taught the ANYmal C robot to navigate in a woodland area containing dirt, high grass, and trees. A human operator drove the robot for 2 min through loose dirt and grass—an area that can be easily traversed by the legged platform. Then we commanded the local planner to execute autonomous point-to-point navigation avoiding obstacles, only using the visual traversability for closed-loop planning Sect. 4.

Figure 8 illustrates the scene used for the experiment and the trajectories used for training and testing autonomous navigation. The robot successfully managed to reach 8 out of 8 goals, where the human operator deliberately chose targets behind trees to challenge the system. This was achieved even though neither geometry nor any additional assumptions about the environment were used during training.

We also show some examples of the SDFs generated during operation used by the local planner in **(b)**, which indicate the trees as obstacles. Lastly, in post-processing we fused the predicted traversability measures into a complete

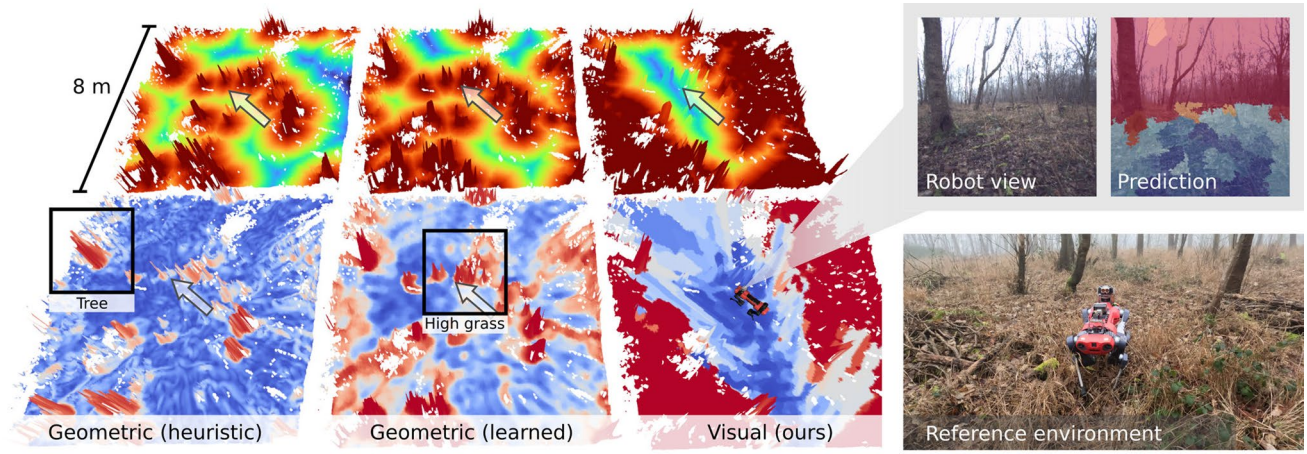

Fig. 7 Visual vs geometric traversability: Illustration of traversability map (bottom row) and corresponding signed distance field (SDF) (top row) for three different traversability estimation methods applied to the same terrain patch. Our visual traversability estimate provides clear advantages for local planning compared to geometric methods, where the latter get heavily affected by traversable high grass or branches (bottom row). This is evident when comparing the SDF's, where geometry-based methods are more sensitive to the spikes produced by high grass areas (top row)

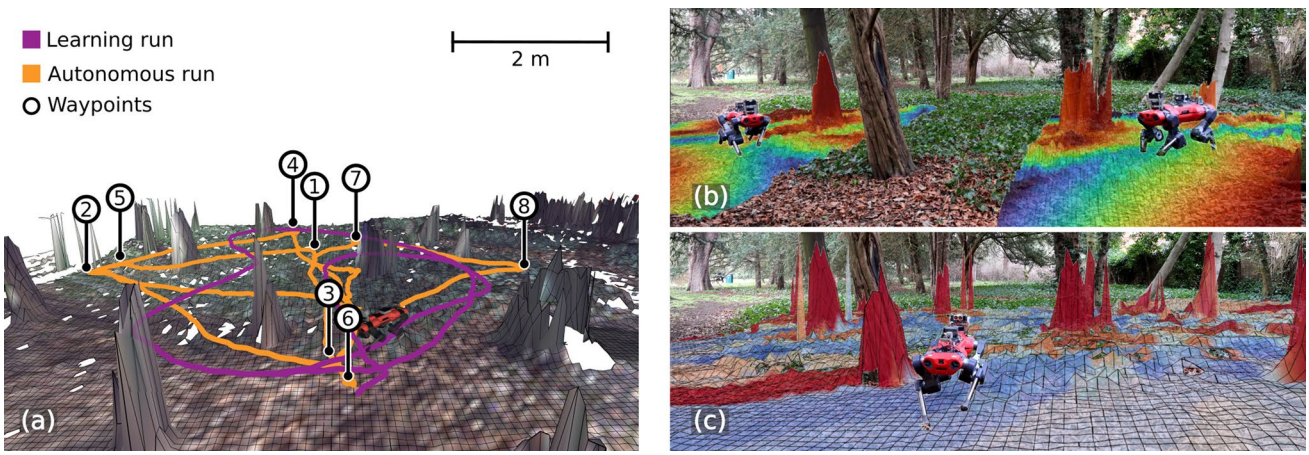

Fig. 8 Point-to-point autonomous navigation: **a** After teleoperating the robot for 2 min (path shown in purple filled square), we successfully achieved autonomous navigation in a woodland environment (path shown in orange filled square). **b** Some of the SDFs generated from the predicted traversability during autonomous operation. **c** Global 2.5D reconstruction of the testing area and predicted traversability, generated in post-processing to illustrate the capabilities of our approach

map in **(c)**, which correctly aligned with the trees. However, given that in this experiment we used the SLIC segmentation method from our previous work, we observed some obstacle artifacts. This limitation is addressed in Sect. 5.2.5, where we deploy our multiple-camera setup and the novel segmentation and pixel-wise prediction method.

5.2.4 Kilometer-scale autonomous navigation in the park

We demonstrated that WVN enabled preference-aware path-following behavior as a result of the human demonstrations and the online learning capabilities of the system. This experiment was also executed with the ANYmal C platform.

We executed 3 runs in a footpath at University Parks, Oxford, UK. Similarly to our previous experiments, we trained the system for less than 2 min along the footpath. We then disabled the learning process to ensure that the predicted traversability strictly mimics the human preference during the demonstration run. The autonomous path following system from Sect. 4.3 was used to guide the robot forward along the path.

In the 3 runs the robot was able to follow the path for hundreds of meters—mostly staying in the center of the path, avoiding grass, bushes, benches, and pedestrians. Figure 9 shows the trajectories followed in each run, starting from different points in the footpath. For runs 1 and 3 we used the same parameters, $k_\sigma = 2$ and FPR$= 0.15$. In run 2 we relaxed the parameters to $k_\sigma = 3$ and FPR$= 0.3$, producing a less conservative behavior that drove the robot to other visually similar areas in the park (mud patches) requiring

manual intervention to correct the heading. When the robot approached an intersection we adjusted, if necessary, the heading to follow the desired footpath.

Overall, we achieved autonomous behavior that would have been difficult to achieve using only geometry, as the path boundaries were often geometrically indistinguishable. On the other hand, instead of training and using a semantic segmentation system to learn *all* the possible traversable classes in the park (pavement, gravel path, roadway or grass), we showed that this short teleoperated demonstration of the gravel footpath was sufficient for WVN to generate semantic cues to achieve the desired path following behavior.

5.2.5 Multi-camera deployment from indoor to outdoor environments

This last experiment demonstrates the adaptation capabilities of WVN and the new multi-camera integration on the ANYmal D quadruped. The deployment was executed at the Max Planck Institute in Tübingen, Germany. We deployed WVN using STEGO segmentation and features during training and perform the inference pixel-wise. Throughout the 7 min teleoperated session, we provide snapshots of the environment, the traversability predictions, as well as the visual and geometric traversability, illustrated in Fig. 10.

The deployment started within a laboratory setting. Upon covering $\sim 8\,\mathrm{m}$ **(a)**, WVN correctly identified the floor as traversable. Transitioning to a corridor after $\sim 40\,\mathrm{m}$ **(b)**, the visual traversability accurately classified windows and closed glass doors as impassable, which the geometric traversability erroneously report as traversable. When exiting the building, WVN correctly predicted the paved walkways as traversable, which can be seen within the courtyard **(c)** at $\sim 86\,\mathrm{m}$, outdoor walkway **(d)** at $\sim 127\,\mathrm{m}$, and the paved road **(e)** at $\sim 148\,\mathrm{m}$. When entering a small grass area with sparse vegetation after walking for a few minutes, ours correctly identified the field as traversable, while the geometric traversability fails to distinguish between trees and penetrable vegetation **(f)** at $\sim 260\,\mathrm{m}$.

The integration of traversability estimates from both cameras enabled us to update the traversability to the front and the back of the robot. This allowed to overcome the restricted field of view limitations shown in Fig. 7. Multiple cameras also allow for more reactive behavior in dynamic environments, where it is crucial for planning to update the belief about the environment constantly.

5.3 Offline analysis

We executed two offline experiments to assess the differences of the feature sub-sampling and inference methods. These experiments were executed in post-processing using logs of previous real-world experiments, on an Nvidia Quadro T2000 Laptop GPU with Intel i7-10875 H CPU.

5.3.1 Segments versus pixel-wise inference

First, we compared the visual traversability prediction differences when performing segmen-wise and pixelwise inference. We ran WVN in post-processing, on the recorded logs from the Sects. 5.2.2 and 5.2.1 experiments.

Figure 12 shows some examples of the traversability predictions when using STEGO and SLIC segments, as well as pixel-wise segmentation. We observed consistencies between the segment-wise and pixel-wise predictions,

Fig. 9 Kilometer-scale navigation: We deployed our system to learn to segment the footpath of a park after training for a few steps. We executed 3 runs starting from different points in the park: orange filled square *run 1* (0.55 km), purple filled square *run 2* (0.5 km), and blue filled square *run 3* (1.4 km). Minor interventions were applied to guide the robot in intersections; major interventions ($\star$) were required for some areas when the robot miss-classified muddy patches for the path

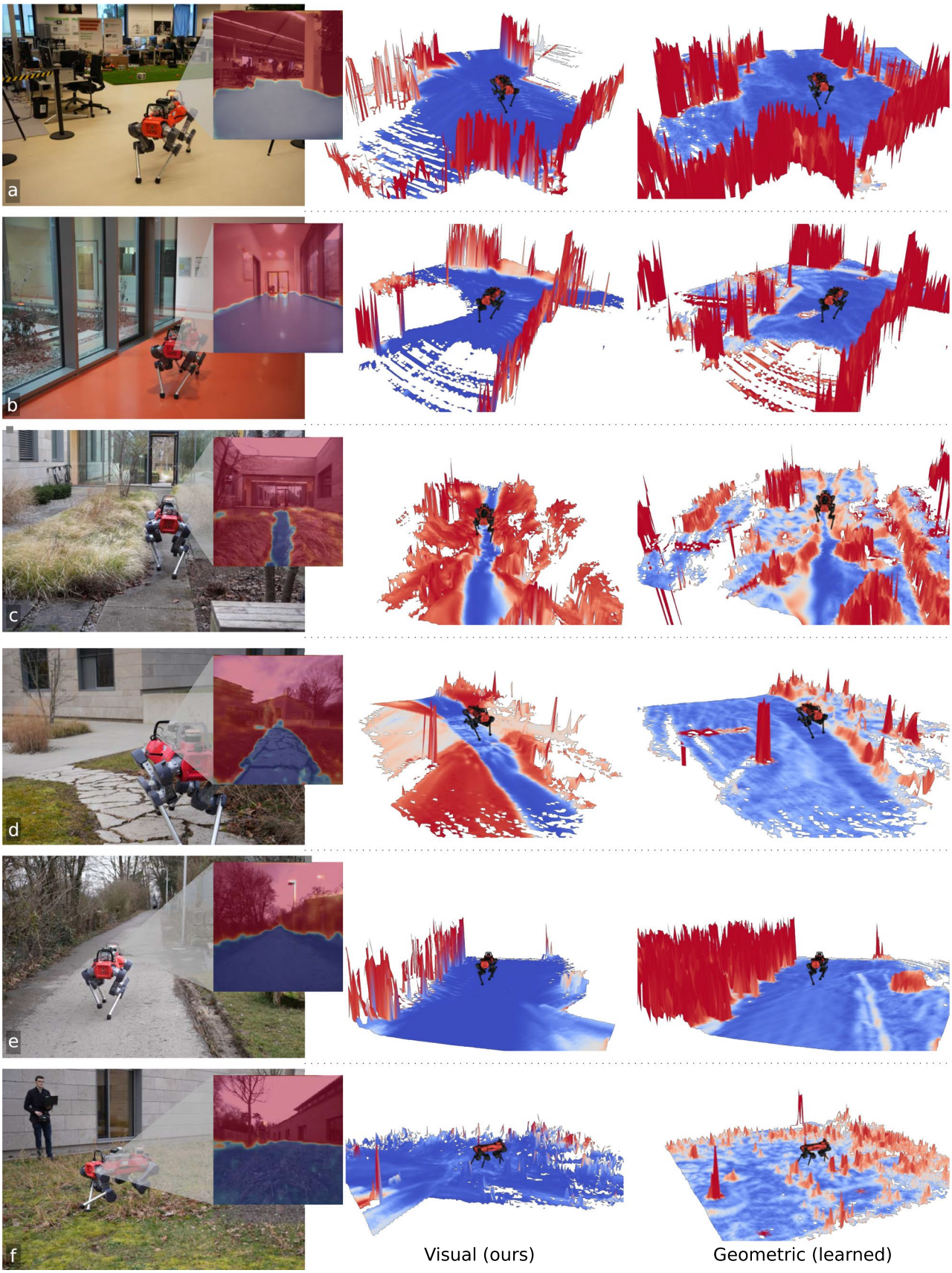
a
b
c
d
e
f
Visual (ours)
Geometric (learned)

Fig. 10 Deployment with multi-camera setup. Left: Real scene and visual traversability prediction. Center: Visual traversability projected on the local terrain map. Right: Geometric traversability computed from elevation map. The robot was teleoperated throughout the experiment. Each example **a**–**f** is sequentially and it is discussed in detail in Sect. 5.2.5

which we explain due to the intrinsic properties of the features (discussed in Sect. 5.3.2). However, pixel-wise inference shows advantages in providing fine-grained traversability predictions, disregarding the artifacts that weak-segmentation methods such as SLIC induce, and seen in Fig. 11b, c on the tree trunk.

STEGO did not produce significant differences in terms of the output, consistently segmenting the traversed areas for both inference approaches. However, we did observe problems over segmenting certain areas, such as the plants in row (d), which suggest that both the features and the segments 'agreed' on the object being semantically similar to the other traversed areas.

5.3.2 Feature subsampling

This second experiment compared different sub-sampling strategies presented in Sect. 3.2.3 in terms of traversability prediction and training. Our methodology involved re-running WVN in post-processing using the recorded signals from the *Autonomous Navigation in the Park* sequence (Sect. 5.2.4). We executed five runs for each case, training the traversability prediction model from scratch without any pre-trained weights. We recorded the produced traversability predictions as well as the learning curves. Figure 12 shows the training loss, averaged over the five runs with 2σ confidence bounds, and qualitative examples of the traversability predictions when using the pixel-wise inference method.

We observed the most benefit when using the STEGO segments and features, which enabled rapid adaptation in terms of segmenting the footpath as traversable (Fig. 12). This was also reflected in the corresponding loss curves, which achieved faster convergence and lower training loss than the other methods.

Regarding the two other sub-sampling methods, random and SLIC, we did not observe significant differences in the predicted traversability. This can be explained by the use of the same DINO-ViT features, which suggests that most of the expressive power is already encoded in the features, and the contribution of the sampling and mean averaging does not considerably affect the predictions. The main difference is the slightly improved training stability reflected on the lower confidence bounds of SLIC compared to random sub-sampling.

6 Conclusion

We presented Wild Visual Navigation (WVN), a system that leverages the latest advances in pre-trained self-supervised networks with a scheme to generate supervision signals while a robot operates, to achieve online, onboard visual traversability estimation. The fast adaptation capabilities of our system allowed us to easily deploy robots for navigation tasks in new environments after just a few minutes of learning from human demonstrations.

We demonstrated WVN through different real-world experiments and offline analyses, illustrating its fast adaptation capabilities, the consistency of its traversability prediction for local planning, and closed-loop navigation experiments, in both indoor and natural scenes. The experiments validated the key idea behind our approach of exploiting the semantic priors from pre-trained models, enabling fast generalization and adaptation in unseen scenarios from small data collected during demonstrations *in the wild*.

Regarding the limitations, the use of traction as the traversability score metric, as well as the closed-loop integration with the local terrain map via raycasting are the main aspects to investigate. These are some of the main open scientific and engineering questions for WVN.

Lastly, to foster further research in the field, we provide the community with the codebase and pre-trained models for different environments as baselines.

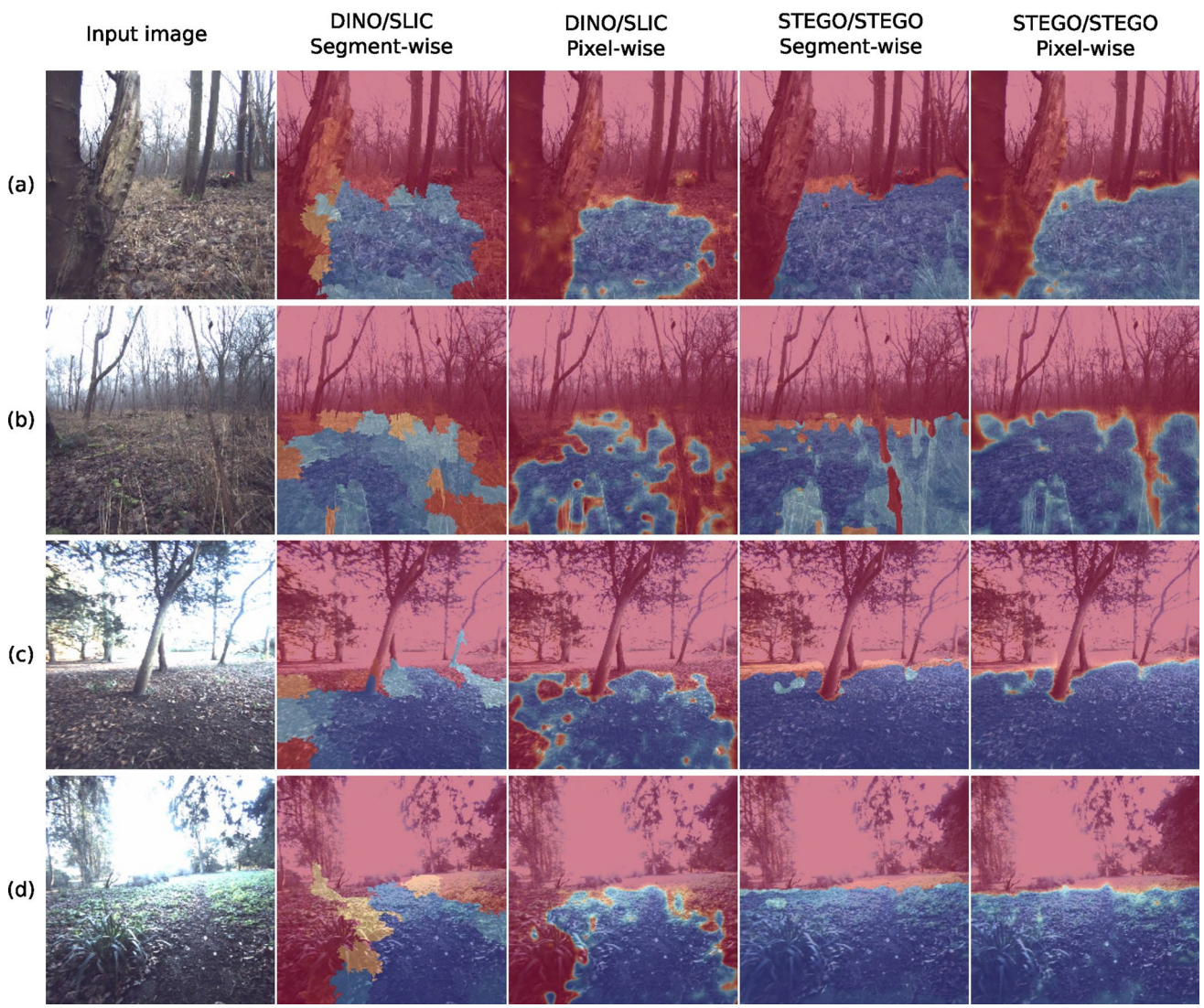

Fig. 11 Inference approaches: We qualitatively compared segment-wise and pixel-wise inference using pre-trained DINO and STEGO features. We observed advantages in executing the inference in a pixel-wise manner, which provided a fine-grained prediction regardless of the pre-trained features

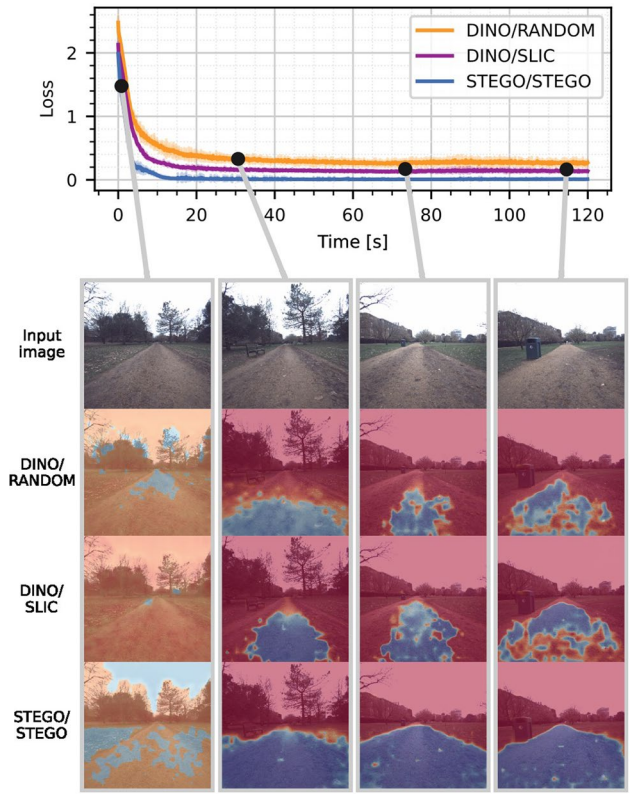

Fig. 12 Feature Sub-sampling: We tested the different sub-sampling methods in the recorded path-following sequence from Sect. 5.2.4. We observed that STEGO provides significant improvements for the path-following task in both traversability prediction fidelity and training stability

Acknowledgements The authors also thank Benoit Casseau for technical support, and Pía Cortés-Zuleta for critically proofreading the manuscript.

Author contributions M.M. and J.F developed the system, integrated it on hardware, designed experiments, and wrote the paper. P.L. developed the modified STEGO model. N.C and M.F. supported the field experiments. All authors reviewed the manuscript.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. Open access funding provided by Swiss Federal Institute of Technology Zurich. Open access funding provided by Swiss Federal Institute of Technology Zurich. This work was supported by the Swiss National Science Foundation (SNSF) through project 188596, the National Centre of Competence in Research Robotics (NCCR Robotics), the European Union's Horizon 2020 research and innovation program under Grant Agreement Nos. 101016970, 101070405, and 852044, and an ETH Zurich Research Grant. Jonas Frey is supported by the Max Planck ETH Center for Learning Systems. Matias Mattamala is supported by the National Agency for Research and Development (ANID) / DOCTORADO BECAS CHILE/2019—72200291 and NCCR Robotics. Maurice Fallon is supported by a Royal Society University Research Fellowship.

Data availability Sample datasets are available alongside the Code: https://bit.ly/498b0CV—Project page: https://bit.ly/3M6nMHH.

Declarations

Conflict of interest The authors declare no competing interests.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence,34*(11), 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Bajracharya, M., Howard, A., Matthies, L. H., Tang, B., & Turmon, M. (2009). Autonomous off-road navigation with end-to-end learning for the LAGR program. *Journal of Field Robotics,26*(1), 3–25. https://doi.org/10.1002/ROB.20269

Belter, D., Wietrzykowski, J., & Skrzypczynski, P. (2019). Employing natural terrain semantics in motion planning for a multi-legged robot. *Journal of Intelligent & Robotic Systems,93*(3–4), 723–743. https://doi.org/10.1007/S10846-018-0865-X

Bradley, D. M., Chang, J. K., Silver, D., Powers, M., Herman, H., Rander, P., & Stentz, A. (2015). Scene understanding for a high-mobility walking robot. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 1144–1151). https://doi.org/10.1109/IROS.2015.7353514

Cai, X., Ancha, S., Sharma, L., Osteen, P. R., Bucher, B., Phillips, S., Wang, J., Everett, M., Roy, N., & How, J. P. (2023). *EVORA: Deep evidential traversability learning for risk-aware off-road autonomy*. CoRR arXiv:2311.06234. https://doi.org/10.48550/ARXIV.2311.06234

Cai, X., Everett, M., Fink, J., & How, J. P. (2022). Risk-aware off-road navigation via a learned speed distribution map. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2931–2937). https://doi.org/10.1109/IROS47612.2022.9982200

Cao, C., Zhu, H., Yang, F., Xia, Y., Choset, H., Oh, J., & Zhang, J. (2022). Autonomous exploration development environment and the planning algorithms. In *International conference on robotics and automation (ICRA)* (pp. 8921–8928). https://doi.org/10.1109/ICRA46639.2022.9812330

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *International conference on computer vision (ICCV)*. https://doi.org/10.1109/ICCV48922.2021.00951

Chavez-Garcia, R. O., Guzzi, J., Gambardella, L. M., & Giusti, A. (2018). Learning ground traversability from simulations. *IEEE Robotics and Automation Letters,3*(3), 1695–1702. https://doi.org/10.1109/LRA.2018.2801794

Chung, T. H., Orekhov, V., & Maio, A. (2023). Into the robotic depths: Analysis and insights from the DARPA subterranean challenge.

Annual Review of Control, Robotics, and Autonomous Systems,6(1), 66. https://doi.org/10.1146/ANNUREV-CONTROL-062722-100728

Erni, G., Frey, J., Miki, T., Mattamala, M., & Hutter, M. (2023). MEM: Multi-modal elevation mapping for robotics and learning. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. https://doi.org/10.1109/IROS55552.2023.10342108

Fan, D. D., Otsu, K., Kubo, Y., Dixit, A., Burdick, J., & Agha-Mohammadi, A. (2021). STEP: Stochastic traversability evaluation and planning for safe off-road navigation. In *Robotics: Science and systems (RSS)*. https://doi.org/10.15607/RSS.2021.XVII.021

Frey, J., Hoeller, D., Khattak, S., & Hutter, M. (2022). Locomotion policy guided traversability learning using volumetric representations of complex environments. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. https://doi.org/10.1109/IROS47612.2022.9982190

Frey, J., Mattamala, M., Chebrolu, N., Cadena, C., Fallon, M., & Hutter, M. (2023). Fast traversability estimation for wild visual navigation. In *Robotics: Science and systems (RSS)*. https://doi.org/10.15607/RSS.2023.XIX.054

Gasparino, M. V., Sivakumar, A. N., Liu, Y., Velasquez, A. E. B., Higuti, V. A. H., Rogers, J., Tran, H., & Chowdhary, G. (2022). WayFAST: Navigation with predictive traversability in the field. *IEEE Robotics and Automation Letters,7*(4), 10651–10658. https://doi.org/10.1109/LRA.2022.3193464

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Guaman Castro, M., Triest, S., Wang, W., Gregory, J. M., Sanchez, F., Rogers, J. G., & Scherer, S. (2023). How does it feel? Self-supervised costmap learning for off-road vehicle traversability. In *International conference on robotics and automation (ICRA)* (pp. 931–938). https://doi.org/10.1109/ICRA48891.2023.10160856. IEEE.

Hadsell, R., Sermanet, P., Ben, J., Erkan, A., Scoffier, M., Kavukcuoglu, K., Muller, U., & LeCun, Y. (2009). Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics,26*(2), 120–144. https://doi.org/10.1002/ROB.20276

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., & Freeman, W. T. (2022). Unsupervised semantic segmentation by distilling feature correspondences. In *International conference on learning representations (ICLR)*. https://openreview.net/forum?id=SaKO6z6Hl0c

Ji, T., Sivakumar, A. N., Chowdhary, G., & Driggs-Campbell, K. (2022). Proactive anomaly detection for robot navigation with multi-sensor fusion. *IEEE Robotics and Automation Letters,7*(2), 4975–4982. https://doi.org/10.1109/LRA.2022.3153989

Jung, S., Lee, J., Meng, X., Boots, B., & Lambert, A. (2024). V-STRONG: Visual self-supervised traversability learning for off-road navigation. In *International conference on robotics and automation (ICRA)*.

Kahn, G., Abbeel, P., & Levine, S. (2021). BADGR: An autonomous self-supervised learning-based navigation system. *IEEE Robotics and Automation Letters,6*(2), 1312–1319. https://doi.org/10.1109/LRA.2021.3057023

Katevenis, M., Sidiropoulos, S., & Courcoubetis, C. (1991). Weighted round-robin cell multiplexing in a general-purpose ATM switch chip. *IEEE Journal on Selected Areas in Communications,9*(8), 1265–1279. https://doi.org/10.1109/49.105173

Kim, D., Sun, J., Oh, S. M., Rehg, J. M., & Bobick, A. F. (2006). Traversability classification using unsupervised on-line visual learning for outdoor robot navigation. In *IEEE international conference on robotics and automation (ICRA)* (pp. 518–525). https://doi.org/10.1109/ROBOT.2006.1641763

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*.

Lee, H., Kwak, K., & Jo, S. (2017). An incremental nonparametric Bayesian clustering-based traversable region detection method. *Autonomous Robots,41*(4), 795–810. https://doi.org/10.1007/S10514-016-9588-7

Mattamala, M., Chebrolu, N., & Fallon, M. (2022). An efficient locally reactive controller for safe navigation in visual teach and repeat missions. *IEEE Robotics and Automation Letters,7*(2), 2353–2360. https://doi.org/10.1109/LRA.2022.3143196

Maturana, D., Chou, P.-W., Uenoyama, M., & Scherer, S. (2017). Real-time semantic mapping for autonomous off-road navigation. In *Field and service robotics* (pp. 335–350). https://doi.org/10.1007/978-3-319-67361-5_22

Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V., & Hutter, M. (2022). Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*. https://doi.org/10.1126/SCIROBOTICS.ABK2822

Miki, T., Wellhausen, L., Grandia, R., Jenelten, F., Homberger, T., & Hutter, M. (2022). Elevation mapping for locomotion and navigation using GPU. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2273–2280). https://doi.org/10.1109/IROS47612.2022.9981507

Moravec, H., & Elfes, A. (1985). High resolution maps from wide angle sonar. *IEEE international conference on robotics and automation (ICRA)* (vol. 2, pp. 116–121). https://doi.org/10.1109/ROBOT.1985.1087316

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *International conference on neural information processing systems (NeurIPS)*.

Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., & Ng, A. (2009). ROS: An open-source robot operating system. In *IEEE international conference on robotics and automation (ICRA)*.

Richter, C., & Roy, N. (2017). Safe visual navigation via deep learning and novelty detection. In *Robotics: Science and systems (RSS)*. https://doi.org/10.15607/RSS.2017.XIII.064

Sathyamoorthy, A. J., Weerakoon, K., Guan, T., Liang, J., & Manocha, D. (2022). TerraPN: Unstructured terrain navigation using online self-supervised learning. In *IEEE/RSJ International conference on intelligent robots and systems (IROS)* (pp. 7197–7204). https://doi.org/10.1109/IROS47612.2022.9981942

Schilling, F., Chen, X., Folkesson, J., &Jensfelt, P. (2017). Geometric and visual terrain classification for autonomous mobile navigation. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 2678–2684). https://doi.org/10.1109/IROS.2017.8206092

Schmid, R., Atha, D., Schöller, F., Dey, S., Fakoorian, S., Otsu, K., Ridge, B., Bjelonic, M., Wellhausen, L., Hutter, M., & Aghamohammadi, A. (2022). Self-supervised traversability prediction by learning to reconstruct safe terrain. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 12419–12425). https://doi.org/10.1109/IROS47612.2022.9981368

Seo, J., Kim, T., Kwak, K., Min, J., & Shim, I. (2023). Scate: A scalable framework for self- supervised traversability estimation in unstructured environments. *IEEE Robotics and Automation Letters,8*(2), 888–895. https://doi.org/10.1109/LRA.2023.3234768

Shaban, A., Meng, X., Lee, J., Boots, B., & Fox, D. (2022). Semantic terrain classification for off-road autonomous driving. In Faust, A., Hsu, D., Neumann, G. (Eds.), *Conference on robot learning (CoRL). Proceedings of machine learning research* (vol. 164, pp. 619–629).

Wellhausen, L., Dosovitskiy, A., Ranftl, R., Walas, K., Cadena, C., & Hutter, M. (2019). Where should i walk? Predicting terrain

properties from images via self-supervised learning. *IEEE Robotics and Automation Letters,4*(2), 1509–1516. https://doi.org/10.1109/LRA.2019.2895390

Wellhausen, L., Ranftl, R., & Hutter, M. (2020). Safe robot navigation via multi-modal anomaly detection. *IEEE Robotics and Automation Letters*. https://doi.org/10.1109/LRA.2020.2967706

Wermelinger, M., Fankhauser, P., Diethelm, R., Krüsi, P. A., Siegwart, R., & Hutter, M. (2016). Navigation planning for legged robots in challenging terrain. In *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. https://doi.org/10.1109/IROS.2016.7759199

Yang, B., Wellhausen, L., Miki, T., Liu, M., & Hutter, M. (2021). Real-time optimal navigation planning using learned motion costs. In *IEEE international conference on robotics and automation (ICRA)* (pp. 9283–9289). https://doi.org/10.1109/ICRA48506.2021.9561861

Zürn, J., Burgard, W., & Valada, A. (2021). Self-supervised visual terrain classification from unsupervised acoustic feature learning. *IEEE Transactions on Robotics,37*(2), 466–481. https://doi.org/10.1109/TRO.2020.3031214

Matias Mattamala is a Postdoctoral Researcher in the Dynamic Robot Systems Group at the University of Oxford. He received his M.Sc in Electrical Engineering from the Universidad de Chile in 2018, and his Ph.D. in vision-based legged robot navigation from the University of Oxford in 2023. His research interests are in the foundations and systems for robot autonomy—representations, perception, action, and learning—with applications to field robotics.

Jonas Frey received his M.Sc. degree in Robotics, Systems, and Control from the Swiss Federal Institute of Technology (ETH Zürich), Switzerland, in 2021. He earned his Ph.D. in Robotics through joint research with the Robotic Systems Laboratory and the Max Planck Institute for Intelligent Systems. His research focuses on perception, navigation, and locomotion, with the goal of building intelligent robots that understand their surroundings and can be effectively deployed in real-world environments.

Piotr Libera received his M.Sc. degree in Robotics, Systems, and Control from the Swiss Federal Institute of Technology (ETH Zürich), Switzerland, in 2024. He earned his B.Sc. in Computer Science from the Warsaw University of Technology in 2022. His research interests focus on perception for robotics.

Nived Chebrolu is a Postdoctoral Research Associate at the Oxford Robotics Institute, University of Oxford, UK. He received his M.Sc. in Robotics from Ecole Centrale de Nantes (ECN), France, and the University of Genoa, Italy in 2015, and his Ph.D. from the University of Bonn in 2021, where he developed registration techniques for agricultural robotic applications. His research interests are in developing robust localization and mapping techniques for field robotics applications.

Georg Martius is a Full Professor in the Department of Computer Science at the University of Tübingen and a Max Planck Research Group Leader at the Max Planck Institute for Intelligent Systems. His research focuses on autonomous learning, self-organization, and adaptive behavior in robotics and artificial intelligence. He earned his Ph.D. from the University of Göttingen and the Bernstein Center for Computational Neuroscience in 2009, following a Diploma in Computer Science from the University of Leipzig. He has held postdoctoral positions at several Max Planck Institutes and was an Independent Fellow at the Institute of Science and Technology Austria.

Cesar Cadena is a Senior Scientist at the Institute of Robotics and Intelligent Systems at ETH Zurich, where he leads the Perception, Mapping, and Navigation team within the Robotic Systems Lab. He also serves as the Managing Director and Head of Education for the ETH RobotX initiative. Prior to this, he was a Senior Researcher with the Autonomous Systems Lab at ETH Zurich, and held research positions at the School of Computer Science at the University of Adelaide and the Department of Computer Science at George Mason University. He obtained his Ph.D. from the Department of Computer Science and Systems Engineering at the University of Zaragoza.

Marco Hutter received the M.Sc. and Ph.D. degrees in design, actuation, and control of legged robots from the Swiss Federal Institute of Technology (ETH Zürich), Zürich, Switzerland, in 2009 and 2013, respectively. He is currently an Associate Professor of robotic systems and the Director of the Center for Robotics, ETH Zürich. He is a Co-Founder of several ETH startups, such as ANYbotics, Zurich, Switzerland, and Gravis Robotics, Zurich. He is also the Director of Boston Dynamics AI Institute Zurich Office, Zurich. He is the Principal Investigator of the NCCRs robotics, automation, and digital fabrication. His research interests are in the development of novel machines and actuation concepts together with the underlying control, planning, and machine learning algorithms for locomotion and manipulation. He was a recipient of the ERC Starting Grant and the winner of the DARPA SubT Challenge.

Maurice Fallon received the B.Eng. degree in electrical engineering from University College Dublin, Dublin, Ireland, in 2004 and the Ph.D. degree in acoustic source tracking from the University of Cambridge, Cambridge, U.K., in 2008. From 2008 to 2012, he was a Postdoc and a Research Scientist with MIT Marine Robotics Group working on SLAM. Later, he was the Perception Lead of MIT's team in the DARPA Robotics Challenge. Since 2017, he has been a Royal Society University Research Fellow and an Associate Professor with the University of Oxford, Oxford, U.K. He leads the Dynamic Robot Systems Group, Oxford Robotics Institute. His research interests include probabilistic methods for localization, mapping, multisensor fusion, and robot navigation.