13th International Conference on Applied Informatics
Eszterházy Károly Catholic University
Eger, Hungary, February 19–21, 2026

A Multimodal Learning Framework for Self-Supervised Off-Road Traversability Understanding

Ines Meliani[a], **Nour Elhouda Ben Saadi**[b], **Zoltán Istenes**[c]

[a]Eötvös Loránd University
dbpgo8@inf.elte.hu

[b]Eötvös Loránd University
nfs@inf.elte.hu

[c]Eötvös Loránd University
istenes@inf.elte.hu

Abstract

Abstract. Accurate terrain traversability assessment is critical for safe and efficient autonomous navigation, particularly in challenging off-road environments where traditional geometric analysis may be insufficient. Existing methods often rely on hand-crafted features or output scalar traversability costs, failing to capture the full dynamic interaction between the robot and the terrain. This paper introduces a novel two-stage self-supervised learning framework that leverages monocular visual input to predict both a quantitative traversability score and the expected future proprioceptive response (specifically, the IMU time-series encompassing acceleration and gyroscope readings) experienced by the robot. Our approach addresses the scarcity of labeled traversability data by utilizing abundant unlabeled image-IMU pairs collected during routine robot operation. In the first stage, V-ICE (Vision-IMU Contrastive Encoder), a vision encoder is trained using a contrastive learning objective (InfoNCE) [1] to align its latent representation with that of a synchronized IMU encoder. This process grounds the visual features directly in the robot's physical interaction with the environment, learning to associate visual terrain cues (e.g., texture, elevation change) with corresponding proprioceptive signatures (e.g., vibrations, impacts) without requiring manual labels. In the second stage, DHTP (Dual-Head Traversability Predictor), two lightweight prediction heads are attached to the pretrained

vision encoder to simultaneously regress a scalar roughness score and predict the full future IMU waveform from a single image. Crucially, V-ICE can be trained on large-scale simulation datasets where perfect image-IMU synchronization is guaranteed, eliminating the need for extensive, potentially noisy real-world labeling efforts for this foundational step. The subsequent DHTP, which requires only limited labeled real-world data for the final prediction tasks, benefits significantly from the robust, physically-grounded representation learned in V-ICE, achieving high accuracy with minimal labeled examples. This framework further enables a practical adaptation strategy: the pretrained V-ICE encoder can be paired with newly trained DHTP heads, learned specifically on limited data from a *different* robot platform. This allows the system to adapt to the unique physical characteristics (e.g., mass, size, suspension, sensor placement) of the new robot without requiring complex modeling or recalibration, leveraging the generalizable visual-physical understanding acquired in V-ICE. The dual output given by quantitative roughness score and predicted IMU waveform, provides richer proprioceptive foresight than scalar-only methods, enabling direct applications in path planning algorithms for selecting smoother routes and for proactive speed adjustment based on anticipated terrain-induced vibrations or impacts. We demonstrate the effectiveness of our approach using data from the TartanAir v2 dataset, showing that the self-supervised pretraining significantly improves the sample efficiency and generalization capability of the subsequent supervised prediction tasks. Our work contributes a new paradigm for vision-based traversability that integrates proprioceptive forecasting and offers a practical pathway for deploying traversability models across diverse robot platforms with minimal data requirements for adaptation. Traditional traversability analysis often relies on geometric features extracted from dense 3D maps (e.g., slope, roughness, step height) [7], including methods leveraging point cloud and sensor fusion [7]. Learned approaches have gained traction, mapping visual features directly to traversability scores using supervised learning [3]. However, obtaining large, accurate ground-truth labels (e.g., "passable", "roughness level") is labor-intensive and often subjective, limiting scalability. Our approach circumvents this need for manual labels in the crucial first stage. Self-supervised learning has emerged as a powerful technique to leverage unlabeled data for representation learning. Methods like STERLING [6] use robot experience for terrain representation, while V-STRONG [4] employs vision foundation models guided by IMU-derived pseudo-labels. These works typically predict scalar traversability costs. Approaches like [2] use fast visual traversability estimation for long-distance navigation. However, none explicitly predict the full proprioceptive time-series from vision alone, which offers richer predictive power. Our work advances this area by explicitly predicting the IMU waveform, providing detailed proprioceptive forecasting. Furthermore, our two-stage approach, separating representation learning (V-ICE) from task-specific prediction (DHTP) using self-supervision in the first stage, is a distinct contribution aimed at improving data efficiency and adaptability. Cross-modal learning, exemplified by frameworks like CLIP [5] for vision-text, demonstrates the power of contrastive objectives in aligning representations across modalities. Applying such principles

to vision-proprioception for robotics is a promising but under-explored area, which our contrastive learning approach in V-ICE directly addresses. Our method's novelty lies in: (1) The explicit prediction of the full IMU time-series from vision, offering richer feedback than scalar scores. (2) The two-stage design where the foundational visual-physical understanding (V-ICE) is learned self-supervisedly, potentially on simulation data, significantly reducing reliance on labeled real-world data for the core representation. (3) The resulting framework's adaptability, allowing the pretrained V-ICE vision encoder to be coupled with newly trained DHTP heads on minimal data from a different robot, bypassing the need for explicit physical modeling or recalibration for each new platform. This contrasts with many existing methods that require retraining or significant adaptation for new robots. In summary, we present a novel framework for vision-based traversability assessment that uniquely combines self-supervised learning for robust representation acquisition with the ability to predict detailed proprioceptive futures. This dual capability, combined with improved data efficiency and cross-platform adaptability, advances the state of the art in robot perception for autonomous navigation. The predicted outputs directly support advanced navigation tasks like path planning and proactive speed control, enhancing the robot's ability to navigate challenging terrains safely and efficiently.

References

[1] T. Chen, S. Kornblith, M. Norouzi, G. Hinton: *A simple framework for contrastive learning of visual representations*, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607.

[2] J. Frey, M. Mattamala, N. Chebrolu, C. Cadena, M. Fallon, M. Hutter: *Fast traversability estimation for wild visual navigation*, arXiv preprint arXiv:2301.07381 (2023).

[3] Y. Gao, T. D. Barfoot, S. McLaughlin: *Terrain traversability analysis using multi-scale convolutional neural networks*, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2017, pp. 5373–5380.

[4] V. Kumar, A. Valverde, S. Singh: *V-STRONG: Visual self-supervised traversability learning for off-road navigation*, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024, pp. 8901–8908.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever: *Learning transferable visual models from natural language supervision*, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.

[6] A. Schmid, L. Teixeira, R. Siegwart: *STERLING: Self-supervised terrain representation learning for unstructured environments*, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2023, pp. 4321–4328.

[7] H. Zou, C. Zhou, H. Li, X. Wang, Y. Wang: *Application of 3D point cloud and visual-inertial data fusion in Robot dog autonomous navigation*, PLoS ONE (2025).