Tartan IMU: A Light Foundation Model for Inertial Positioning in Robotics Supplementary Material

Shibo Zhao[1†*], Sifan Zhou[1†*], Raphael Blanchard[1], Yuheng Qiu[1], Wenshan Wang[1], Sebastian Scherer[1*]
[1]Carnegie Mellon University

1. Summary

In this appendix, we present a comprehensive analysis of our approach from multiple perspectives:

- **Performance Evaluation:** We first showcase the performance of the Tartan IMU model, supported by figures and videos.
- **Qualitative Analysis:** We then analyze key design choices, including heterogeneous pre-training, the comparison between local and global coordinate systems, and the effectiveness of multi-head versus single-head architecture.
- **Dataset and Model Details:** Next, we provide an in-depth overview of our diverse training datasets and the foundation model's architecture, including detailed specifications of the training data used in its development.
- **Related Work:** Finally, we extend the discussion on related work, providing additional context and insights.

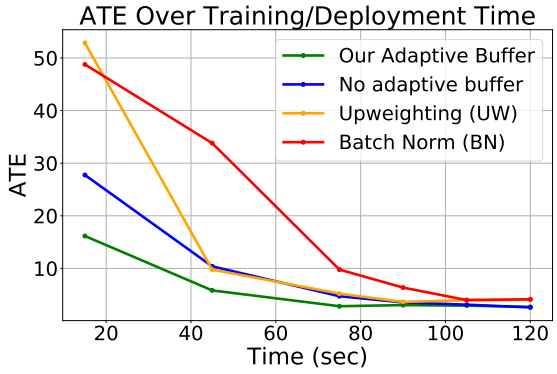

Fig. 1. Memory buffer ablation study. Our dynamic memory buffer selection allows for faster adaptation than known baselines.

2. Online Adaptation Evaluation

Figure 1 demonstrates the effectiveness of our adaptive selection strategy in improving real-time adaptation performance. While the baseline model performs well without adaptive selection, our approach significantly enhances both training speed and accuracy while utilizing only 75% of the available data. These results emphasize the critical role of maintaining a compact yet diverse dataset in achieving optimal performance. To further validate the performance of our online adaptation pipeline, we conduct an additional experiment on a circle-like trajectory (see Figure 2). From left to right, it is evident that our model can provide more accurate trajectory predictions within 90 seconds.

3. More Ablation Study

In this section, we did more experiments to show the effectiveness of our Tartan IMU model including heterogeneous pre-training, coordinate selection, and multi-head design.

3.1. The Discussion of Heterogeneous Pretraining

Heterogeneity, as previously defined, encompasses the varied characteristics and formats of data from different sources or platforms. Previous results indicate that training on a diverse range of robotic data significantly enhances model performance, emphasizing the critical role of large-scale, high-quality, and heterogeneous datasets in improving the generalization capabilities of the IMU model.

To further demonstrate the effectiveness of our heterogeneous pretraining strategy, we provide a qualitative comparison of results. As illustrated in Figure 3, the model trained on a heterogeneous dataset—including IMU data from both wheeled vehicles and quadruped robots (left column)—achieves significantly higher position estimation accuracy than the model trained on a single motion pattern (right column). This trajectory analysis highlights the advantages of heterogeneous pretraining, offering compelling evidence of its effectiveness in enhancing pose estimation.

3.2. The Effectiveness of Coordinate Selection

As previously discussed, our approach estimates position by regressing velocity within a local coordinate system. Compared to directly predicting absolute position in a global frame, this velocity-based method mitigates the risk of overfitting to specific trajectories. This design enhances the model's generalization ability and facilitates more robust feature learning.

*Corresponding author. †Equal contribution with alphabetical order.

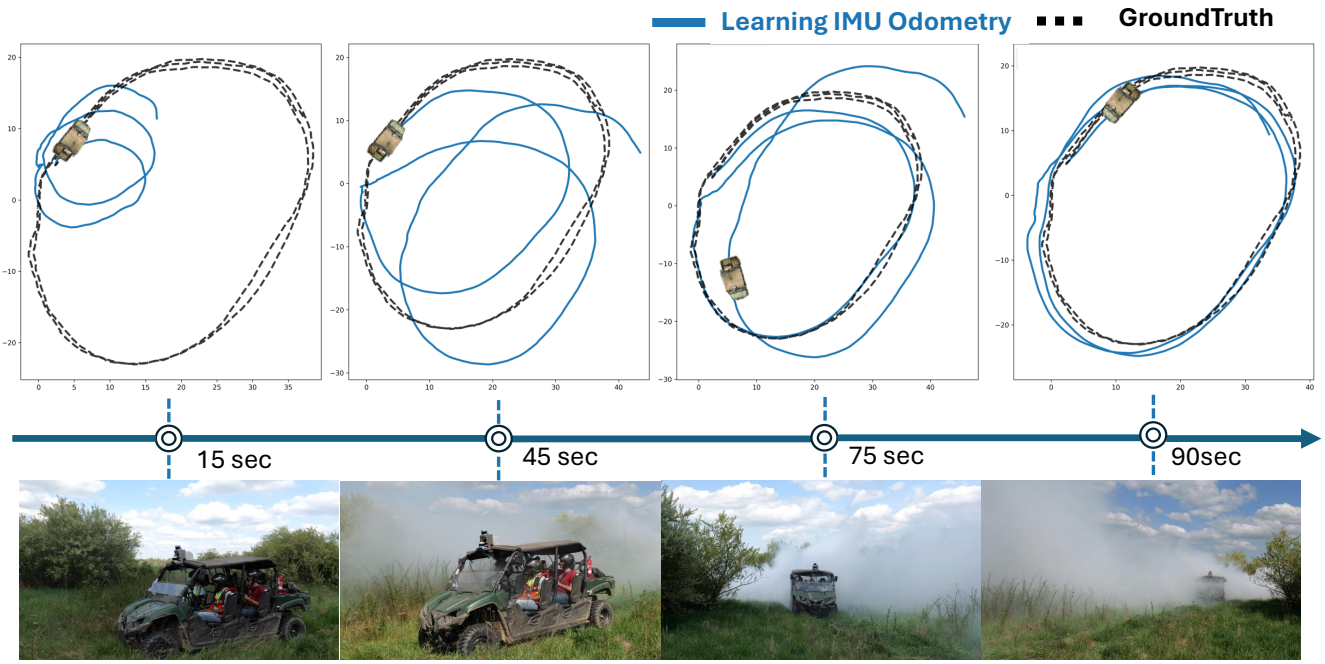

Fig. 2. Performance of Online Adaptation on Unseen Trajectory. The Tartan IMU model progressively learns unseen circular patterns through incremental training data. It can be seen that our model can learn new motion patterns within 90 seconds.

To illustrate the impact of our approach, we present a visual comparison of trajectory predictions using different regression targets, as shown in Figure 4. Predicting velocity within a local coordinate system consistently yields more accurate trajectory estimates. In contrast, while global position predictions preserve relative trajectory scales, they exhibit significant estimation errors. These results confirm the effectiveness of our chosen regression target—velocity in local coordinates—demonstrating its advantages across different coordinate frameworks.

3.3. The Effectiveness of Multi-head Design

Our approach employs distinct regression heads to predict velocities specific to different robotic platforms. As shown in Figure 5, using a single, coupled prediction head for both wheeled and quadruped robots leads to a noticeable increase in absolute trajectory error (ATE) over time (Right). In contrast, decoupling velocity predictions into separate regression heads significantly improves trajectory accuracy. These results highlight the effectiveness of our robot-decoupled multi-head design, which reduces interference between distinct motion patterns and enhances adaptability across diverse robotic platforms.

4. Dataset and Implementation Details

4.1. Training Dataset

As detailed in Tab. 1, we trained the Tartan IMU model on a large-scale, heterogeneous dataset comprising over **100** hours of real-world IMU data. Those datasets encompasses a diverse mix of autonomous driving and urban navigation behaviors from 10 distinct robotic platforms with diverse dynamics. These platforms span multiple categories, including custom-built all-terrain vehicles, quadruped robots, handheld human-operated devices, and unmanned aerial vehicles.

The dataset features trajectories with a broad range of dynamic behaviors and top speeds, ranging from 1.0 to 15 m/s, across varied operational environments. These environments include office buildings, suburban areas, university campuses, and high-speed indoor racing settings. All data were sourced from publicly available datasets or contributed by researchers from previous projects.

4.2. Data Augmentation

Data augmentation is a widely used technique to increase data diversity and mitigate overfitting. Given that different IMU devices exhibit varying levels of Gaussian white noise and bias, we introduce randomized noise perturbations during training to enhance adaptability across diverse IMU hardware. Specifically, we randomly superimpose Gaussian white noise and bias onto each input sample. The acceleration and gyroscope biases are sampled from a uniform distribution as a 1×6 vector, while two sets of Gaussian noise with $10 \times 200 \times 3$ are applied to each input sample of size $10 \times 200 \times 6$.

We implement the model using PyTorch [16] and train it with the Adam optimizer, starting with an initial learning rate of 0.0001. Following the training strategy in [5], we

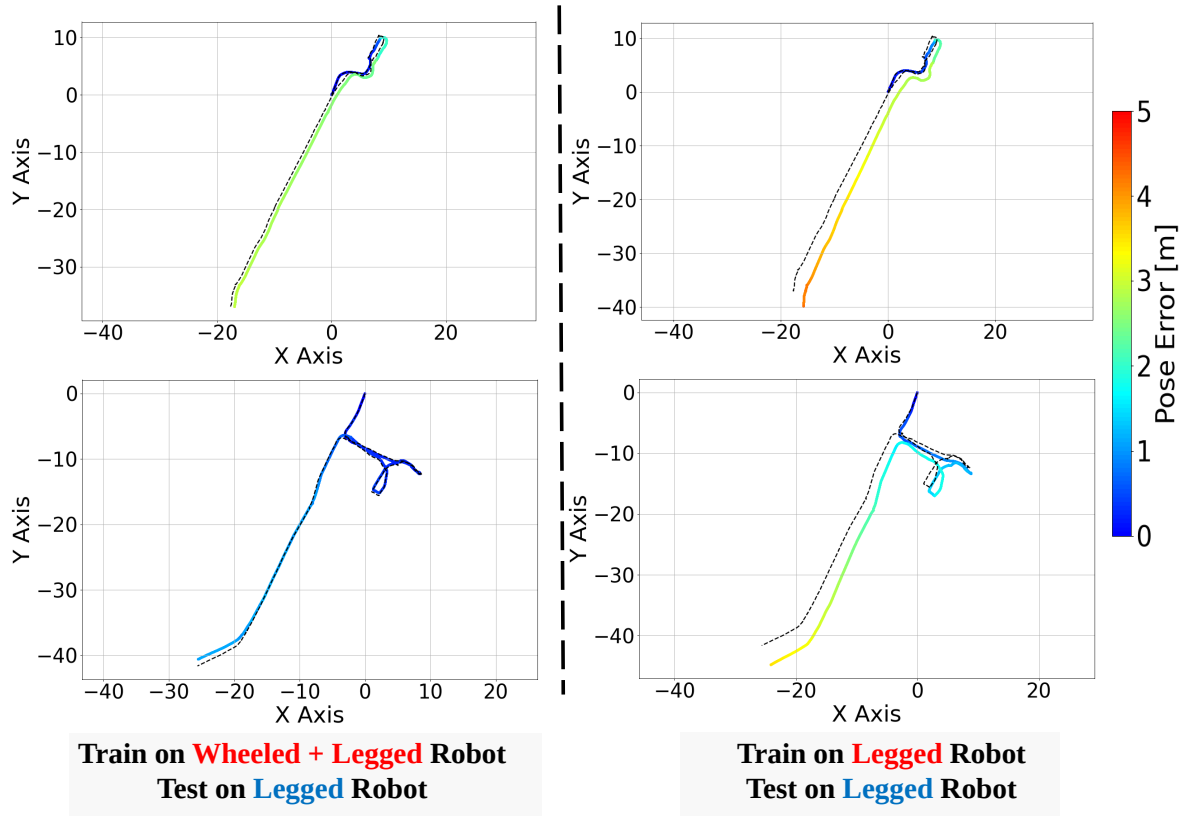

Fig. 3. Qualitative Results of Heterogeneous Pretraining. The model on the left, pretrained on a diverse dataset incorporating IMU data from both wheeled vehicles and quadruped robots, achieves significantly higher position estimation accuracy than the model on the right, which was trained on a single motion pattern. This trajectory comparison underscores the advantages of our heterogeneous pretraining approach in enhancing model performance.

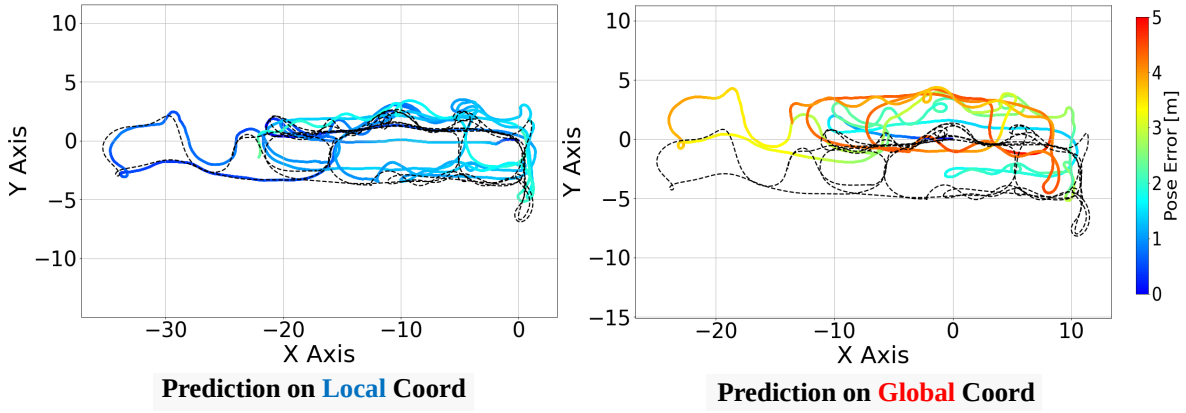

Fig. 4. Qualitative Results of Different Regression Targets (Different Coordinate Systems). The left side illustrates velocity estimation within a local coordinate system, while the right side depicts direct global position prediction. Estimating local velocity results in significantly more accurate trajectory predictions.

first use MSE loss until convergence, then switch to negative log-likelihood (NLL) loss for further training. The model requires approximately 30 hours of training on an NVIDIA 4090 GPU, with the model achieving the lowest validation loss selected for testing.

4.3. Implementation Details.

For network input, we extract samples from each data sequence using a sliding window at a fixed sampling fre-

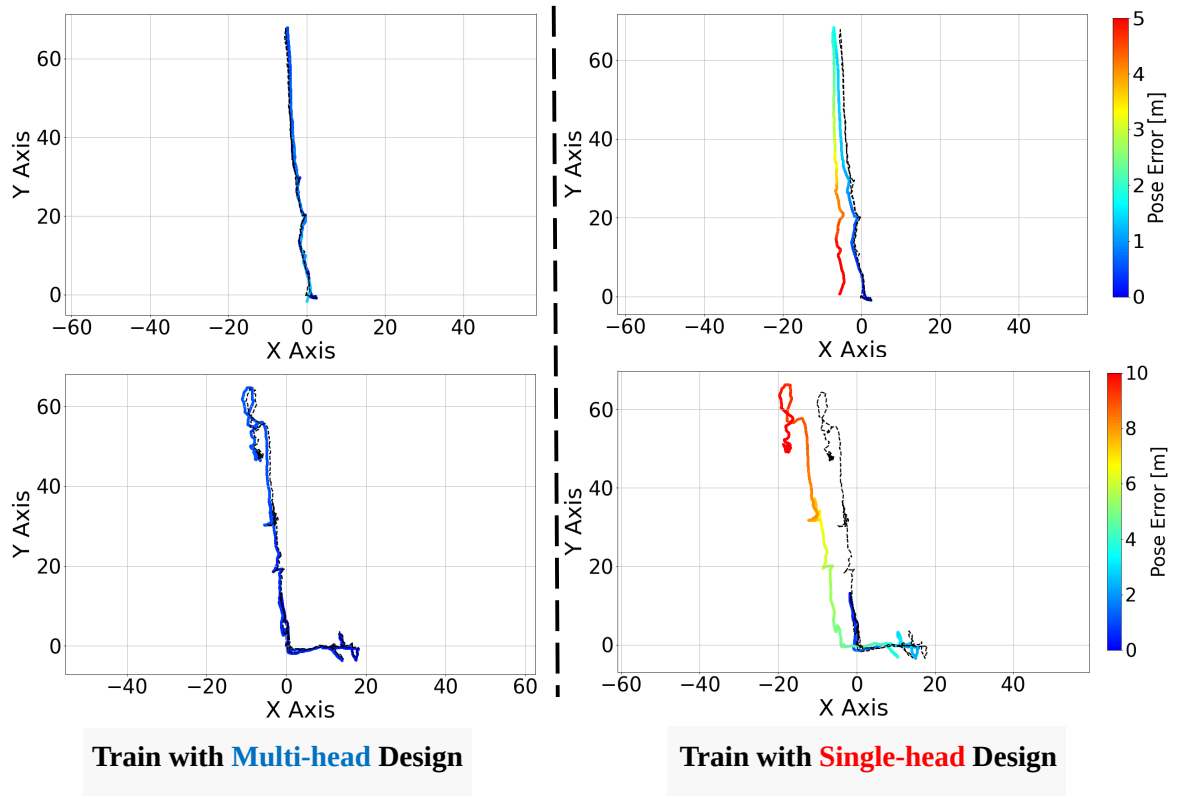

Fig. 5. Qualitative Results of Multi-head Design. The left side illustrates the robot-decoupled multi-head design, while the right side represents the coupled single head architecture. The predictions based on multi-head yield much better trajectory prediction results.

	Dataset	Platform	Speed	Frequency	Total Hrs.	Hrs. Used	Environment
1	SubT-MRS [24]	QR/UGV/Handheld/UAV	2m/s	200 Hz	500h	80h	In/Outdoors/Urban
2	IDOL [19]	Handheld	1.4m/s	200 Hz	20h	10h	In/Outdoors
3	RNIN-VIO [5]	Handheld	1.2m/s	100 Hz	7h	7h	In/Outdoors
4	BlackBird [1]	UAV	7m/s	100 Hz	10h	5h	Indoors
5	UZH-FPV Drone Racing [6]	UAV	12.8m/s	500 Hz	1h	0.5h	Indoors
	Total			-	538h	102.5h	

Table 1. **The TartanIMU training dataset** contains over 100 hours of IMU data in challenging indoor, outdoor, and off-road environments across 10 different robots of varying sizes, speeds, and capabilities. ATV: All-terrain vehicle. Quadruped(Legged) robot. QR: Quadruped(Legged) robot. UAV: Unmanned Aerial Vehicle.

quency. Each window includes N IMU samples, resulting in an input dimension is $N \times 6$. To standardize the input, we linearly interpolate all sequences to 200 Hz. With a window size of 200 (represents 1.0s), and an LSTM with 10 time steps, each input sample has a shape of $10 \times 200 \times 6$. The supervision signal is the relative position of the window within the body coordinate system over the window. We implement the model using Pytorch [16] and train it with the Adam optimizer at an initial learning rate of 0.0001. Following the training strategy in [5], we first use MSE loss to train until convergence, and then switch to negative log-likelihood (NLL) loss until convergence. Training requires approximately 30 hours on an NVIDIA 4090 GPU. The model yielding the lowest validation loss is selected for testing.

4.4. Metrics

Following previous methods [2, 5, 19], we evaluate our method on Absolute Trajectory Error (ATE), Time-Relative Trajectory Error(T-RTE), and Distance Relative Trajectory Error (D-RTE) metrics to demonstrate the effectiveness.

5. More Related Work

5.1. Existing Foundation Models

Recent large language models (LLMs) like GPT and LLaMA have excelled in various domains but struggle with non-text data. Vision-language models (VLMs) address this by integrating multiple modalities, including images and videos, to enhance data understanding. As LLMs evolve, foundational models for vision and language have emerged, with some focusing on open-vocabulary VLMs for state estimation in robotics [8, 10, 11, 14, 17]. Models like LEXIS [10] and FM-Loc [14] utilize CLIP features for indoor localization and mapping tasks, enhancing room-level scene recognition. However, these models do not fully explore the broader applicability of foundational features. AnyLoc [11] integrates dense foundational features for state-of-the-art place recognition, while FoundLoc [8] combines AnyLoc with Visual-Inertial Odometry (VIO) for GNSS-denied environments, demonstrating VLM effectiveness on UAVs and embedded hardware. ViNT [17] adapts general-purpose pre-trained models for vision-based robotic navigation, outperforming specialized models.

Despite these advances, there remains a gap in the use of IMU data for robotics state estimation within existing foundational models, and no benchmark currently supports the development and evaluation of such models. This paper addresses this gap by introducing a foundation model that effectively incorporates IMU data for state estimation.

5.2. IMU Dataset in Different Robot Platforms

In the field of vehicle navigation, the KITTI dataset [7] serves as a widely adopted benchmark. The sensors are rigidly mounted on the car chassis and a high-precision GPS/IMU system, providing ground truth with 100 Hz inertial data and 10 Hz GPS/images. Besides, several large-scale IMU datasets like Rellis 3D [9], SubT-MRS [24], MADMAX [13], M2DGR [23], TartanDrive 1.0/2.0 [18, 21] support vehicle localization across diverse indoor/outdoor scenarios and motion patterns.

For quadrupedal robots, datasets like SubT-MRS [24], LegKILO [15], and ETH-Legged [20] capture legged robot motion across different environments. Pedestrian datasets, such as OxIOD [4] and SIMD [12], focus on human motion patterns. Drones are represented by datasets like EuRoC MAV [3], Tartanair [22], BlackBird [1], and UZH-FPV [6] for state estimation in flight.

While these datasets span multiple robotic platforms, they often prioritize visual or LiDAR-based navigation, overshadowing IMU contributions. This work aims to reposition IMUs as critical for state estimation, urging the community to reassess their potential alongside other sensors.

References

[1] Amado Antonini, Winter Guerra, Varun Murali, Thomas Sayre-McCord, and Sertac Karaman. The blackbird uav dataset. *The International Journal of Robotics Research*, 0 (0):0278364920908331, 0. 4, 5

[2] Martin Brossard, Axel Barrau, and Silvère Bonnabel. Ai-imu dead-reckoning. *IEEE Transactions on Intelligent Vehicles*, 5(4):585–595, 2020. 4

[3] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35 (10):1157–1163, 2016. 5

[4] Changhao Chen, Peijun Zhao, Chris Xiaoxuan Lu, Wei Wang, Andrew Markham, and Niki Trigoni. Deep-learning-based pedestrian inertial navigation: Methods, data set, and on-device inference. *IEEE Internet of Things Journal*, 7(5): 4431–4441, 2020. 5

[5] Danpeng Chen, Nan Wang, Runsen Xu, Weijian Xie, Hujun Bao, and Guofeng Zhang. Rnin-vio: Robust neural inertial navigation aided visual-inertial odometry in challenging scenes. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 275–283, 2021. 2, 4

[6] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the UZH-FPV drone racing dataset. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019. 4, 5

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 5

[8] Yao He, Ivan Cisneros, Nikhil Keetha, Jay Patrikar, Zelin Ye, Ian Higgins, Yaoyu Hu, Parv Kapoor, and Sebastian Scherer. Foundloc: Vision-based onboard aerial localization in the wild. *arXiv preprint arXiv:2310.16299*, 2023. 5

[9] Peng Jiang, Philip Osteen, Maggie Wigness, and Srikanth Saripalli. Rellis-3d dataset: Data, benchmarks and analysis. 2020. 5

[10] Christina Kassab, Matias Mattamala, Lintong Zhang, and Maurice Fallon. Language-extended indoor slam (lexis): A versatile system for real-time visual scene understanding, 2024. 5

[11] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9 (2):1286–1293, 2024. 5

[12] Feng Liu, Hongyu Ge, Dan Tao, Ruipeng Gao, and Zhang Zhang. Smartphone-based pedestrian inertial tracking: Dataset, model, and deployment. *IEEE Transactions on Instrumentation and Measurement*, 73:1–13, 2024. 5

[13] Lukas Meyer, Michal Smíšek, Alejandro Fontan Villacampa, Laura Oliva Maza, Daniel Medina, Martin J Schuster, Florian Steidle, Mallikarjuna Vayugundla, Marcus G Müller, Bernhard Rebele, et al. The madmax data set for visual-inertial rover navigation on mars. *Journal of Field Robotics*, 38(6):833–853, 2021. 5

[14] Reihaneh Mirjalili, Michael Krawez, and Wolfram Burgard. Fm-loc: Using foundation models for improved vision-based localization, 2023. 5

[15] Guangjun Ou, Dong Li, and Hanmin Li. Leg-kilo: Robust kinematic-inertial-lidar odometry for dynamic legged robots. *IEEE Robotics and Automation Letters*, 9(10):8194–8201, 2024. 5

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 2, 4

[17] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023. 5

[18] Matthew Sivaprakasam, Parv Maheshwari, Mateo Guaman Castro, Samuel Triest, Micah Nye, Steve Willits, Andrew Saba, Wenshan Wang, and Sebastian Scherer. Tartandrive 2.0: More modalities and better infrastructure to further self-supervised learning research in off-road driving tasks. *arXiv preprint arXiv:2402.01913*, 2024. 5

[19] Scott Sun, Dennis Melamed, and Kris Kitani. Idol: Inertial deep orientation-estimation and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6128–6137, 2021. 4

[20] Marco Tranzatto, Takahiro Miki, Mihir Dharmadhikari, Lukas Bernreiter, Mihir Kulkarni, Frank Mascarich, Olov Andersson, Shehryar Khattak, Marco Hutter, Roland Siegwart, and Kostas Alexis. Cerberus in the darpa subterranean challenge. *Science Robotics*, 7(66):eabp9742, 2022. 5

[21] Samuel Triest, Matthew Sivaprakasam, Sean J. Wang, Wenshan Wang, Aaron M. Johnson, and Sebastian Scherer. Tartandrive: A large-scale dataset for learning off-road dynamics models, 2022. 5

[22] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 5

[23] Jie Yin, Ang Li, Tao Li, Wenxian Yu, and Danping Zou. M2dgr: A multi-sensor and multi-scenario slam dataset for ground robots. *IEEE Robotics and Automation Letters*, 7(2): 2266–2273, 2021. 5

[24] Shibo Zhao, Yuanjun Gao, Tianhao Wu, Damanpreet Singh, Rushan Jiang, Haoxiang Sun, Mansi Sarawata, Yuheng Qiu, Warren Whittaker, Ian Higgins, et al. Subt-mrs dataset: Pushing slam towards all-weather environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22647–22657, 2024. 4, 5