TartanDrive 1.5: Improving Large Multimodal Robotics Dataset Collection and Distribution

Matthew Sivaprakasam[1], Samuel Triest[1], Mateo Guaman Castro[1], Micah Nye[2], Mukhtar Maulimov[1], Cherie Ho[1], Parv Maheshwari[3], Wenshan Wang[1], and Sebastian Scherer[1]

***Abstract*— The complexity of the real world presents numerous challenges in robotics that must be overcome, such as handling complex physical interactions, learning novel tasks, and planning in unknown environments. Recently, large data-driven algorithms and deep learning models have been adopted and modified to solve these problems, but with the rise of these approaches there also arises a need for large amounts of diverse robotics data to train them on. In this work we discuss the improvements to our previous dataset, TartanDrive, that we are currently working on to fulfill these needs in the context of off-road driving. We address the challenges of copious data collection in order to provide an expansive dataset containing several modalities collected in an outdoor area with approximately 225 acres of diverse terrain. Moreover, we will provide scripts capable of re-configuring this data (such as by filtering by location or formatting to fit specific use-cases/conventions) and release a framework that will allow others to not only use our data but collect their own in a way that enables them to use our scripts. By leveraging this dataset, we hope to facilitate the advancement of robotics and reduce the barrier to entry that is often associated with data at this scale.**

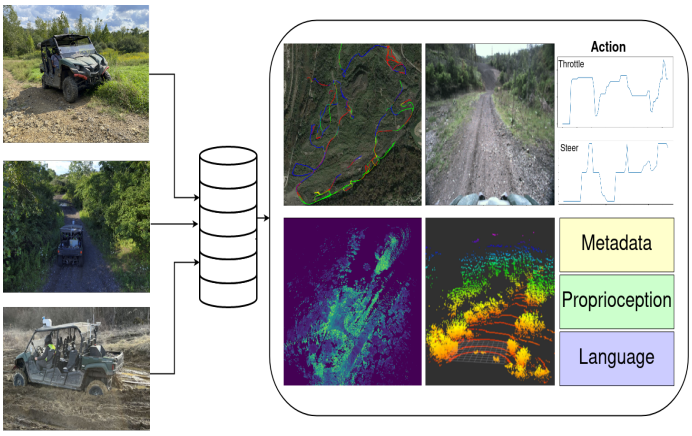

Fig. 1: We present our goals for the next generation of TartanDrive, including more modalities, better structure and data pipelines, and a framework for others to process their own data through our pipelines.

I. Introduction

In many robotics applications, systems are required to perform complicated tasks in diverse environments. It is often too difficult to hand-tune parameters in traditional algorithms for these tasks, which is why more recent works tend to incorporate learned models into their software as submodules or in an end-to-end manner. Recently, there has been a trend in exploring how to re-purpose large models developed for language and vision, such as GPT or CLIP [1,2] for robotics. These models have been shown to generalize effectively, and some works have already shown their potential improvement in certain robotics applications [3–6]. However, their impact in robotics as a whole is constrained by the data that they are pre-trained on, which is limited with respect to key robotics concepts such as multimodality and physical interactions. By using large amounts of robotics data when initially training these large models, we can potentially obtain richer robotics-specific representations that are powerful enough to handle even more complicated downstream tasks. One explanation for the lack of progress in this regard is the fact that accumulating enough data presents a challenge in several domains, especially in field robotics applications such as autonomous off-road driving, where there are tighter requirements with respect to performance and safety.

There exist very large autonomous driving datasets, but the largest ones are from urban environments such as cities [7–10]. These bigger datasets have been feasible partially due to the resources and infrastructure available to the institutions that collected them. The additional difficulties present in collecting off-road data has caused the existing datasets to be more limited in comparison [11–17]. With a few exceptions, most have smaller amounts of data due to various constraints. For example, many focus on tasks that require explicit labels such as semantic segmentation. With these types of datasets, even if more data was available, it is difficult to scale them up due to the barrier of obtaining high-quality labels. The ambiguity present in off-road environments has also caused these datasets to be somewhat inconsistent with respect to each other [18]. For example, some might focus on distinguishing between traversable and non-traversable terrain [14], while others might be more concerned with distinguishing different object classes such as trees and bushes [12,13]. This makes it hard to merge datasets without significant engineering effort [19], and the way they are merged is task- and system-specific (bushes might get mapped to traversable terrain for large robots but not for small robots).

We claim that scaling off-road datasets up to the needs of large models requires a few specific key qualities. Regardless

* This work was supported by ARL awards #W911NF1820218 and #W911NF20S0005.

[1] Robotics Institute, Carnegie Mellon University, msivapra,striest,mguamanc,mmaulimo,cherieh,wenshanw,basti@andrew.cmu.edu

[2] Department of Mechanical Engineering, University of Pittsburgh. man172@pitt.edu

[3] Department of Mathematics, Indian Institute of Technology Kharagpur. parvmaheshwari2002@iitkgp.ac.in

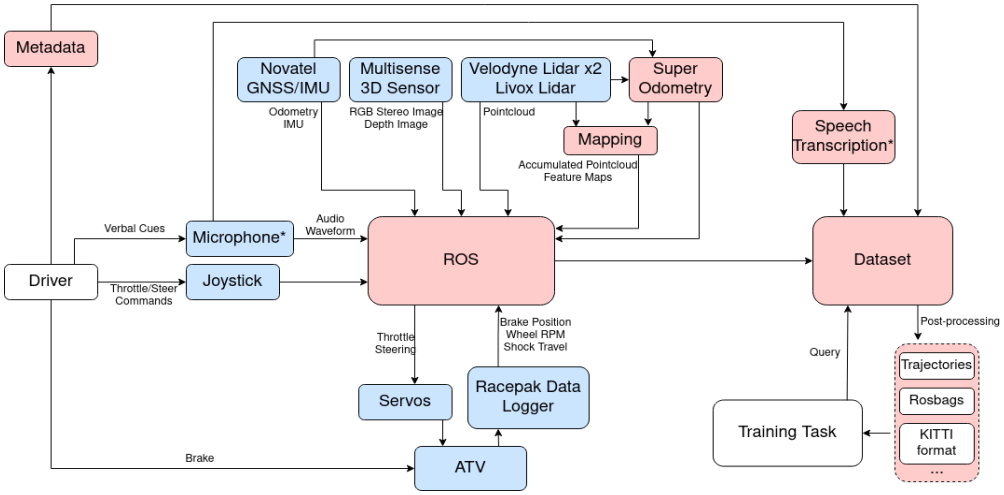

Fig. 2: The high-level flow of information in our data collection process, with software in red and hardware in blue.

of whether or not the dataset contains manually-labeled samples, it should provide information that allows models to be trained in a self-supervised manner. This will help to alleviate the scaling issues present with hand-labeled data. It should also be highly accessible, meaning that it should be useful to as many works as possible, regardless of the downstream task. Finally, it should be consistent with other off-road datasets so that domain overlap can be exploited when present.

Last year, we released a large multimodal off-road driving dataset, and now we are working on an improved version that we will release at the end of this summer. Taking inspiration from highly-organized and accessible examples in other works [7,9,12,20], we present our current progress and ideas towards this dataset that fulfills the previously-mentioned requirements in the following ways:

1) Rather than dedicating effort into manually labeling existing data, we are adding more modalities, such as lidar as shown in Fig. 1. We continue to provide other signals as well, such as teleoperation inputs and GPS, that have the potential to facilitate self-supervision.
2) We are implementing better infrastructure and pipelines in order to improve ease-of-use and accessibility. This includes processed forms of the raw data (e.g. birds-eye-view maps from lidar or stereo camera) that can be directly used to test non-perception tasks such as planning. Additionally, we will provide the capability to re-configure this processed data based on the needs of the user.
3) We will release a framework for structuring datasets the same way we do. This will enable others to utilize our pipelines with their own data and will encourage consistent data conventions.

II. Related Work

The largest datasets for autonomous driving come from urban scenarios, where cars are generally driving on roads and in relatively well-structured environments. For example, the Waymo Open Dataset [7,8] includes data from urban vehicles equipped with lidars and cameras in order to provide 100,000 labeled images and 1,200 segments of labeled lidar data. Nuscenes [9,10] is similar, including labeled lidar and camera data, but also exposes other information such as IMU data and vehicle information such as wheel speed and steering angle. Argoverse provides processed maps alongside their raw data in order to streamline certain tasks. For example, one of the maps they incldue allows the user to filter out ground heigh in order to detect obstacles more efficiently.

There is also a growing number or off-road driving datasets. Rellis-3D and RUGD [12,13] are two popular datasets, with RUGD providing over 7,000 annotated images with 24 object classes, and Rellis-3D containing over 6,000 annotated images and 13,000 labeled LiDAR scans. Rellis-3D also provides the rosbag files they collected, which include stereo camera and IMU data as well. Rather than focus on object segmentation, the CaT: CAVS Traversability Dataset classifies different types of terrain as traversable or non-traversable [14], providing different labels for different types of vehicles such as sedan, pickup truck, and off-road vehicle. While these works are a good source of training data, their size isn't enough to train larger models. Moreover, they lack the infrastructure that the larger urban driving datasets provide.

III. The Dataset

The first version of TartanDrive was designed with dynamics modeling in mind. Our new version will have the same capabilities as before but, thanks to better organization and additional sensor modalities, it will also be suitable for a number of other tasks. The overall flow of data is outlined in Fig. 2.

A. The Platform

As in TartanDrive, we use a Yamaha Viking All-Terrain Vehicle (ATV) as our data collection platform pictured in Fig. 3. The previous hardware still exists on the platform, but we will now also be including lidar information from three lidar sensors. Two Velodyne VLP-32 lidar sensors are mounted on top of the front of the ATV, one of them being

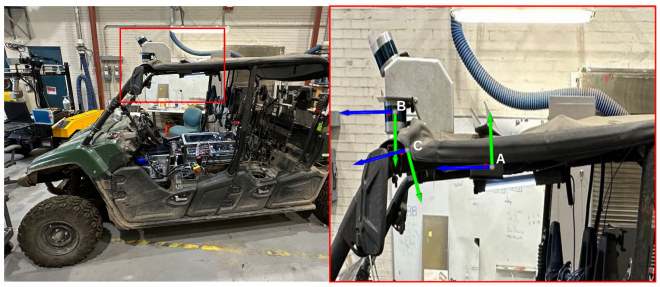

Fig. 3: The Yamaha Viking ATV that we use for data collection and testing. Relevant frames include the GPS/INSS system (A), Velodyne (B), and Multisense (C)

tilted down at an angle as shown in Fig. 3. Before collecting the full dataset, we will be adjusting the tilt angle of the Velodyne and also add a Livox Mid-70 to maximize sensor coverage. We also will provide a 3D pointcloud model of the car generated using a Faro Focus Scanner so that accurate measurements between different components of the car can be taken.

B. Data Collection

The main location for collecting data is the same as before, as it covers a wide set of terrains such as dirt paths, grassy hills, and narrow trails surrounded by dense foliage. Again, the vehicle will be tele-operated by a human during collection, but each session will now have an associated metadata file containing information such as:

- Name of driver and robot
- Date/time
- Context (e.g. data collection)
- Weather conditions (e.g. sunny, damp, snow)
- Active sensors and algorithms/modules
- Algorithm parameters

as well as a file containing time-stamped notes taken while driving. This added structure combined with our post-processing pipelines will allow the querying of data in several useful manners (e.g. evaluating a perception algorithm in dim vs bright conditions). We emphasize that taking time to populate these metadata files can significantly improve the utility, especially when it comes to large dataset management and usage.

C. Raw Data

Our dataset will provide the raw data from a number of sensors:

1) Pointclouds: We provide pointcloud data from 2 Velodyne VLP-32 lidar sensors, as well as a Livox Mid-70. We include extrinsics so that they can be merged into one cloud.

2) Images: A Carnegie Robotics Multisense S21 provides image data, specifically monocular RGB, depth, and the grayscale images from the stereo cameras used to generate the depth.

3) IMU and Pose: A NovAtel PROPAK-V3-RT2i GNSS provides an estimate of global pose, and also gives IMU data. We also have an additional Xsens MTi-30 AHRS IMU, and we are exploring the possibility of setting up an RTK system for higher GPS accuracy.

4) Teleoperation: Joystick controls are used to drive the ATV, and these inputs are recorded in the form of a throttle and steer angle command.

5) Proprioceptive Information: In addition to the commanded steer angle and throttle, we include the actual steer angle and throttle at each timestep. A Racepak G2X Pro Data Logger is used to record suspension shock travel and wheel RPM (for each wheel), as well as brake pedal position.

D. Post-Processed Data

We also provide data in a post-processed form similar to what we currently use in our own autonomy algorithms:

1) Odometry and Registered Pointcloud: In addition to the odometry estimate provided by the Novatel GNSS system, we record an estimate provided by running Super Odometry [21]. We also use the Super Odometry estimate to register pointclouds from N timesteps together to form a global pointcloud. We will also provide a global pointcloud accumulated across all scans of a given run, similar to what is provided in [17].

2) Depth Estimation and Semantic Segmentation: We use the raw stereo images from the MultiSense as input into TartanVO [22] in order to predict a depth image at each timestep. We use a pretrained GANav model [23] to provide a semantic segmentation prediction as well.

3) Local Maps: We process the registered pointcloud into a local birds-eye-view map that is 200x200m at .5m resolution. For each voxel, we calculate several features/statistics and represent each as a channel in the map, some of which are shown in Fig. 4. The local lidar map calculates for each cell the following features:

1) Min/Max/Mean Height of Points
2) Roughness
3) SVD Features
4) Estimated Ground Height
5) Estimated Ground Slope (X,Y,Magnitude)
6) Semantic Class

We also provide the image maps from the original TartanDrive dataset in the form of height and RGB maps at a resolution of .02m extending out 10 meters in front of the vehicle and 5 meters on either side.

E. Data Pipelines

1) Formatting: All our data is collected as rosbags, and replaying the rosbags is a convenient way to test real-time algorithms. However, they are less convenient for training models offline. Alongside the rosbags, we provide utilities that convert them into a dataset of trajectories where each trajectory contains N seconds of all the above modalities time-synced together and saved in a dictionary. We also will provide similar utilities that instead convert the data into other common dataset formats such as KITTI [24].

2) Reconfiguration: We chose the default map size/resolution parameters to fit the dimensions and capabilities of our ATV. However, one goal of this dataset is to aid others in training models to deploy on their own systems that might be physically very different to ours

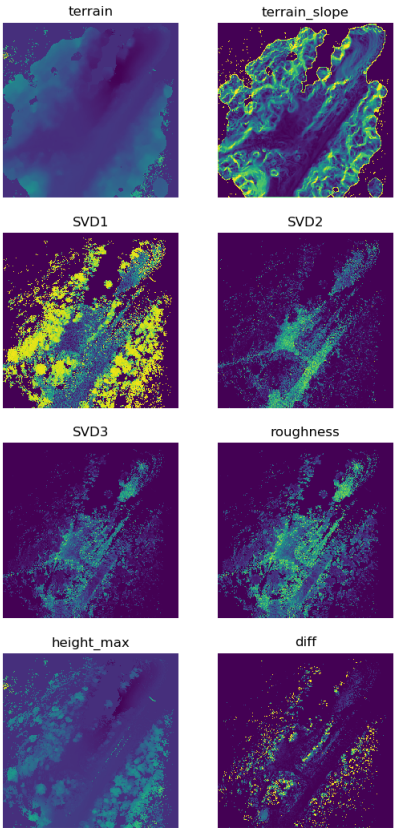

Fig. 4: Some of the feature maps calculated using the registered pointcloud that we provide.

(e.g. a .5m resolution might not be fine enough for a smaller and more dexterous vehicle, and a 200m range at a finer resolution might not be feasible given computational constraints). To that end, we aim to provide scripts that can regenerate our post-processed data in user-specified configurations.

3) Utilites: One challenge that arises when working with large datasets is extracting only the information that is useful for a given task. We will provide utilities that take advantage of our metadata system and other signals to easily filter and group data by various queries. For example, a user would be able to split the data by time of day, by location, or by who was driving the ATV. This makes it easy to extract only the information that is necessary for an experiment, and for common tasks such as creating a validation test set.

F. Data Framework

Collecting our new dataset with the features and infrastructure described above gives it the potential of much higher quality than the original TartanDrive dataset. We will release a framework for others to follow so that they can take advantage of our tools in their datasets as well. If they follow our framework, they will be able to easily use our post-processing tools for easy-to-use data. Using the same conventions as that in our dataset will also make it easier to merge them together (for example in order to create a multi-robot dataset).

G. Additional Goals

While the majority of the new dataset will come from the same testing site, we are exploring the possibility of collecting more data at other locations. While our existing site has a wide range of terrain, even more diversity will make our dataset more compelling. The logistics of transporting our ATV to several other locations is non-trivial however. Another key advantage of detailing our data collection process as a framework is that it would make it easier to collaborate with other groups and accumulate data that would be to difficult to obtain separately.

We are also exploring the idea of including audio and language in our dataset. We believe that there is something to be learned about driving with respect to the words a driver uses to describe where their goal is and how they wish to reach it (for example avoiding obstacles that might be difficult to detect with perception alone). There is also possible information to be gained from the sounds generated by the ATV as it drives through different environments, such as change in engine volume or the sound of wheels driving across various terrains.

IV. Impact

Though we are still working on improving our overall setup as described in this work, our current progress with our data has already enabled us to train two models without human labels that help the ATV autonomously navigate complex terrain. In Triest et al., inverse reinforcement learning is used to predict costmaps based on the lidar feature maps, using our trajectories as expert demonstrations [25]. Guaman Castro et al. use a birds-eye-view representation using the stereo camera images to predict a costmap that is supervised by a bumpiness cost calculated using the IMU [26]. Many existing off-road datasets either don't contain enough modalities or don't contain enough data to enable self-supervised approaches such as these.

Our setup will also allow for a number of task-specific benchmarks. For example, GPS is not only a potential self-supervision signal, but it also provides an effective ground-truth for testing SLAM algorithms run using various combinations of the different modalities we provide.

V. Conclusion

We present our current work and goals for the next generation of our TartanDrive multimodal off-road driving dataset, specifically how better infrastructure and an emphasis on self-supervision can allow it to scale up to the needs of large models. We highlight the new hardware and modalities we are adding, as well as the tools we provide to process them into easily-manipulable datasets. By setting up a framework for others to feed their data into, we hope to facilitate the creation of more task-independent datasets that are unified enough to be merged together as needed into even bigger datasets without a loss of quality.

REFERENCES

[1] OpenAI, "Gpt-4 technical report," 2023.

[2] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[3] A. Brohan, N. Brown, J. Carbajal *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *arXiv preprint arXiv:2212.06817*, 2022.

[4] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," Microsoft, Tech. Rep. MSR-TR-2023-8, February 2023. [Online]. Available: https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilities/

[5] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*. PMLR, 2022, pp. 894–906.

[6] D. Shah, B. Osiński, S. Levine *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on Robot Learning*. PMLR, 2023, pp. 492–504.

[7] P. Sun, H. Kretzschmar, X. Dotiwalla *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] S. Ettinger, S. Cheng, B. Caine *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset."

[9] H. Caesar, V. Bankiti, A. H. Lang *et al.*, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[10] J. H. L. Z. H. C. O. B. A. V. W. Fong, R. Mohan, "Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking," in *ICRA*, 2022.

[11] S. Triest, M. Sivaprakasam, S. J. Wang *et al.*, "Tartandrive: A large-scale dataset for learning off-road dynamics models," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2546–2552.

[12] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," 2020.

[13] M. Wigness, S. Eum, J. G. Rogers *et al.*, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[14] S. Sharma, L. Dabbiru, T. Hannis *et al.*, "Cat: Cavs traversability dataset for off-road autonomous driving," *IEEE Access*, vol. 10, pp. 24 759–24 768, 2022.

[15] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Field and Service Robotics*, M. Hutter and R. Siegwart, Eds. Cham: Springer International Publishing, 2018, pp. 335–350.

[16] A. Datar, C. Pan, M. Nazeri, and X. Xiao, "Toward wheeled mobility on vertically challenging terrain: Platforms, datasets, and algorithms," 2023.

[17] J. Knights, K. Vidanapathirana, M. Ramezani *et al.*, "Wild-places: A large-scale dataset for lidar place recognition in unstructured natural environments," 2023.

[18] Z. Yang, Y. Tan, S. Sen *et al.*, "Uncertainty-aware perception models for off-road autonomous unmanned ground vehicles," *CoRR*, vol. abs/2209.11115, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2209.11115

[19] A. Shaban, X. Meng, J. Lee *et al.*, "Semantic terrain classification for off-road autonomous driving," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 619–629. [Online]. Available: https://proceedings.mlr.press/v164/shaban22a.html

[20] W. Wang, D. Zhu, X. Wang *et al.*, "Tartanair: A dataset to push the limits of visual slam," 2020.

[21] S. Zhao, H. Zhang, P. Wang *et al.*, "Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments," 09 2021, pp. 8729–8736.

[22] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," 2020.

[23] T. Guan, D. Kothandaraman, R. Chandra *et al.*, "Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8138–8145, 2022.

[24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[25] S. Triest, M. G. Castro, P. Maheshwari *et al.*, "Learning risk-aware costmaps via inverse reinforcement learning for off-road navigation," 2023.

[26] M. G. Castro, S. Triest, W. Wang *et al.*, "How does it feel? self-supervised costmap learning for off-road vehicle traversability," 2023.