

# Learning-on-the-Drive: Self-supervised Adaptive Long-range Perception for High-speed Offroad Driving

Eric Chen<sup>\*1</sup>, Cherie Ho<sup>\*2</sup>, Mukhtar Maulimov<sup>2</sup>, Chen Wang<sup>3</sup>, and Sebastian Scherer<sup>2</sup>

**Abstract**— Autonomous offroad driving is essential for applications like emergency rescue, military operations, and agriculture. Despite progress, systems struggle with high-speed vehicles exceeding 10m/s due to the need for accurate long-range (> 50m) perception for safe navigation. Current approaches are limited by sensor constraints; LiDAR-based methods offer precise short-range data but are noisy beyond 30m, while visual models provide dense long-range measurements but falter with unseen scenarios. To overcome these issues, we introduce *ALTER*, a *learning-on-the-drive* perception framework that leverages both sensor types. *ALTER* uses a *self-supervised* visual model to learn and adapt from near-range LiDAR measurements, improving long-range prediction in new environments without manual labeling. It also includes a model selection module for better sensor failure response and adaptability to known environments. Testing in two real-world settings showed on average 43.4% better traversability prediction than LiDAR-only and 164% over non-adaptive state-of-the-art (SOTA) visual semantic methods after 45 seconds of online learning.

## I. INTRODUCTION

Autonomous off-road driving is pivotal in diverse applications including emergency rescue operations [1], military maneuvers [2], and agriculture [3]. These fields demand high-precision and reliable navigation systems capable of operating under varied and challenging conditions. The essence of this requirement lies in the ability to traverse off-road terrains at high velocities, often exceeding 10m/s, which necessitates long-horizon perception and decision-making capabilities.

Despite current autonomous offroad driving systems having made significant strides, they continue to grapple with the complexities presented by high-speed vehicular movement. As shown in Fig. 1, the crux of this challenge is the need for accurate long-range perception, specifically beyond 50m. Such capabilities are critical for early decision-making, which is essential for safe and effective navigation in dynamic off-road environments. For example, a robot can operate faster if it distinguishes long-range features like impassable bushes, risky tall grass, and efficient trails [4], [5].

Most autonomous off-road driving systems are designed for slow speeds below 5m/s and tailored for specific scenarios, as evidenced by various studies [6]–[9]. A key challenge in developing high-speed vehicles is the limitations of sensor technologies. For example, LiDAR-based perceptual models,

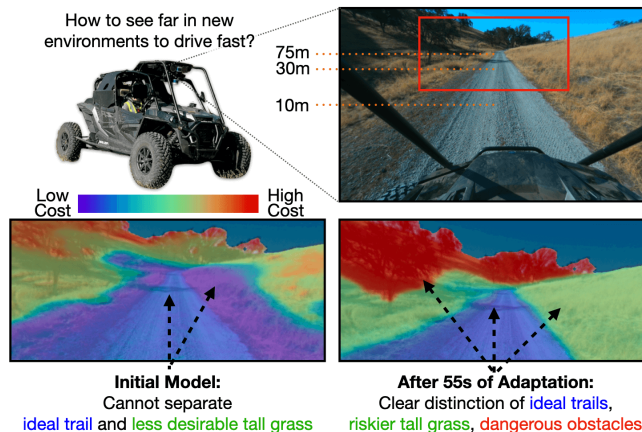


Fig. 1: We present *ALTER*, a long-range *learning-on-the-drive* perception framework for high-speed offroad driving. We validate it using All-Terrain Vehicle data from two real-world offroad scenarios where key terrains (trails, grass and obstacles) can be differentiated at long range (> 50m) with less than 1 min of online learning.

though accurate and reliable against visual variations, see reduced effectiveness beyond 30m due to sparse, noisy data, complicating detailed analysis [7], [10]. In contrast, visual perceptual models offer denser and longer-range observations but rely on extensive human-annotated datasets and falter in unseen scenarios and environments [9], [11].

To effectively utilize the strengths of LiDARs and cameras, an effective way is to self-supervise long-range visual models by near-range LiDAR measurements in an online adaptive setting. For example, traversability can be learned in colored point cloud space [12], however restricting the prediction to within reliable LiDAR range may limit strategic planning. Some methods [4], [5], [13] learn from First-Person-View (FPV) images but only yield binary segmentation masks. Such methods are insufficient for high-speed driving which requires *finer-grained terrain reasoning* for early, precise decisions.

To produce fine-grained terrain traversability estimation and achieve online adaptation (< 1 min) simultaneously, we present *ALTER* (Adaptive Long-range Traversability Estimator), a novel learning-on-the-drive perception framework that is engineered to be both simple and highly efficient. It features a straightforward visual model capable of learning from near-range LiDAR measurements in real-time, delivering accurate, pixel-wise estimates. This involves accumulating near-range LiDAR data to generate a comprehensive long-range map and then projecting these 3D labels onto images for self-supervised learning, leading to a precise, fine-grained, long-range, and online-adaptable perception model.

Additionally, to address potential forgetting issues and

\*Both authors contributed equally.

<sup>1</sup>Eric Chen is with the Department of Computer Science and Mathematics, Harvey Mudd College, USA erchen@hmc.edu

<sup>2</sup>Cherie Ho, Mukhtar Maulimov, and Sebastian Scherer are with the Robotics Institute, Carnegie Mellon University, USA {cherieh, mmaulimo, basti}@andrew.cmu.edu

<sup>3</sup>Chen Wang is with the Spatial AI & Robotics Lab, University at Buffalo, NY 14260, USA chenw@sairlab.org

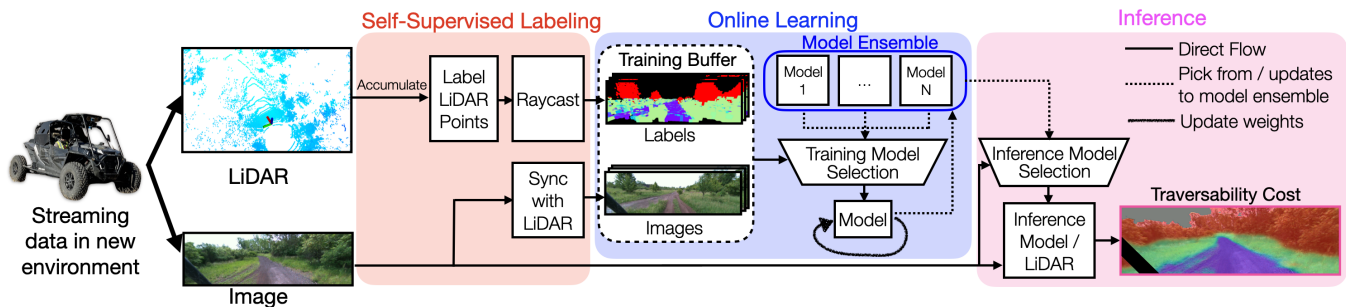


Fig. 2: **ALTER Framework:** Our adaptive approach estimates traversability from images using models trained online with new LiDAR data. **Online Labeling:** We accumulate LiDAR features to label their costs in 3D. These labels are projected onto the image plane to form training labels. **Online Learning:** *ALTER* picks the most similar model or makes a new model for training to adapt to new appearances. **Online Inference:** We pick the best inference model given image similarity or LiDAR to predict pixel-wise fine-grained traversability scores.

manage sensor failures, we introduce a model selection mechanism that dynamically chooses the segmentation head to finetune, creating an ensemble of local experts. This approach provides consistent performance upon re-visiting previously learned environments and yields robust predictions in case of sensor impairments, like cameras obscured by mud, a common and previously overlooked issue in off-road driving. Our contributions can be summarized as follows:

- **Online Learning Framework:** We present a simple yet effective online adaptable traversability estimation framework, *ALTER*, based on self-supervised learning. To the best of our knowledge, it is one of the first methods to generate precise, fine-grained, and pixel-wise traversability scores by combining the strengths of LiDAR and camera. Online adaptation is made possible using accumulated LiDAR labels during operation, which also helps to extend the fine-grained perception range from 30m to 100m, enabling high-speed off-road driving. In addition, we introduce a model selection module employing image-based novelty detection and online validation loss, alleviating the forgetting issue and enhancing robustness against visual sensor failures.
- **Real-World Adaptive Performance:** To demonstrate the performance for online adaptation, we evaluate *ALTER* on real-world visually distinct off-road data, which is collected from forests and hills in Pittsburgh, PA, and San Luis Obispo, CA, respectively. *ALTER* outperforms LiDAR-only methods on average by 43.4% and non-adaptive methods by 164% with under 1 min of online learning. Additionally, we show that the model selection module effectively counters catastrophic forgetting and boosts robustness to visual sensor failures.

## II. RELATED WORKS

**Traversability estimation:** Traversability estimation involves determining how suitable a specific terrain is for the vehicle to drive over. The two most common sensors for offroad perception are cameras and LiDARs. We find works that use LiDARs to quantify traversability by building a grid map whose cells contain geometric statistics [6], [10], [14]. However, LiDAR measurements are sparse, and therefore noisy, at the distance. Many works incorporate visual predictions to increase the range [9], [11], [15]. For example,

Maturana et al. [9] classifies an image into semantic classes, such as trail, grass, etc. However, these visual models are trained with copious hand-labeled data and fail when deployed in environments different from the training distribution.

**Cross-modal Self-supervised Learning:** To alleviate the need for hard-to-acquire human annotation, self-supervised learning (SSL) is often employed to generate labeled data automatically, with applications in healthcare [16] and robotics [17]. This paradigm has been used to combine complementary information across multiple sensors. For example, visual models have successfully learned from IMU [18], [19], force-torque signals [20], acoustic waveforms [21] and odometry [22]–[24]. Recently, SSL has leveraged exciting progress in visual foundation models like Segment Anything [25] for improved perception. However, the above works train on static datasets and may still fail when out of distribution.

**Online Learning:** Online learning is a way to address data drift by adapting to new data distributions. Common techniques include repeatedly training with the latest training buffer [26] or discarding outdated clusters of data [4]. With larger image datasets, finetuning from pretrained model enables more sample-efficient online learning [27]. However, fine-tuning a model on new tasks can lead to "forgetting" of previously learned information, known as catastrophic forgetting [28]. Continual learning approaches prevent forgetting by memory-based [29] and regularization-based methods [30].

**Online Self-Supervised Learning for Traversability Estimation:** For high-speed off-road driving, it is important to have accurate sensing at range, especially in new settings. We find works that leverage cross-modal self-supervised learning and online learning to provide dense visual traversability estimation in unseen environments. Several works train visual traversability models using stereo measurements as supervision [26], [31]. Hadsell et al. [26] segments images into multiple classes based on stereo-estimated object height. However, the authors also state the unreliability of stereo range sensing to appearance and lighting changes as a key limitation. Recent works leverage odometry error to self-supervise visual models online [32], [33]. However, they are constrained by the need for direct terrain traversal to learn terrain costs. Frey et al. [33] address this by predicting a confidence score to capture uncertain or untraversable regions, but this may lead to initial conservative estimates.

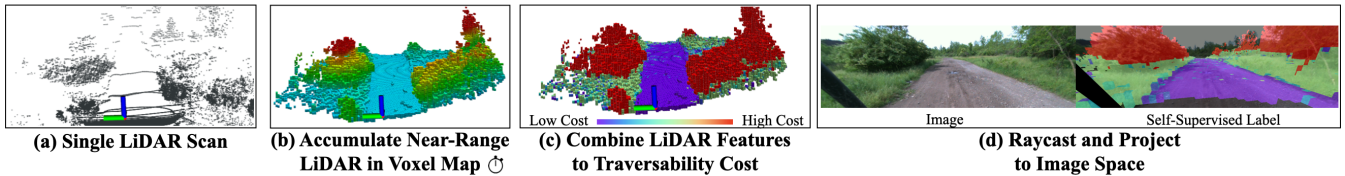


Fig. 3: **Self-Supervised Traversability Labeling Pipeline:** Our self-supervised approach generates pixel-wise labels without manual labeling. First, LiDAR scans are registered and accumulated into a voxel map. Then, LiDAR features (object height, terrain slope, planarity) are extracted from the voxel map to calculate traversability costs. Finally, the 3D labels are projected into image space to generate pixel-wise labels. Our labels are able to distinguish between regions of different traversability, such as trail, grass and tall obstacles.

Alternatively, LiDAR offers both long-range and untraversable terrain labeling, largely addressing these limitations.

Our work builds on successful approaches that use online self-supervised learning to combine LiDAR and camera [4], [8], [12]. Sofman et al. [12] learns a mapping between LiDAR-based traversability to a colored point cloud space. However, the estimation is limited to within reliable LiDAR range while our method goes beyond. We also find works that infer on images [4], [5], [13]. However, in these works, the prediction is limited to *binary* traversable or non-traversable classes, which is overly conservative for complex off-road scenes. In contrast, *ALTER* provides *fine-grained* traversability estimation. To address the limitations of past work, we seek to build an adaptive visual model that learns from LiDAR measurements to provide fine-grained, pixel-wise predictions.

### III. APPROACH

To estimate fine-grained traversability at long-range while being robust to previously unseen environments, we adapt our model online for long-range visual observations by accumulating near-range LiDAR measurements. The framework of *ALTER* is shown in Fig. 2, consisting of three key steps:

- 1) Extracting image labels from accumulated LiDAR scans;
- 2) Adapting a visual model with the self-labeled data;
- 3) Infer on incoming images with online-learned models.

#### A. Online Self-Supervised Labeling

LiDAR can produce accurate near-range traversability labels, hence with the vehicle moving forward, we can obtain the long-range visual labels for the current time step by simply *accumulating* the future near-range LiDAR measurements. This involves (1) pixel-wise labeling in LiDAR space and (2) projecting LiDAR labels to image space.

1) *Pixel-wise Labeling in LiDAR Space:* To better label LiDAR scans for traversability estimation, we first a) accumulate LiDAR scans into voxel maps; then b) extract the object height, slope, and planarity features; and last c) combine features to inform traversability for off-road driving.

a) *Accumulating LiDAR Scans into Voxel Map:* LiDAR scan accumulation is sensitive to the transform accuracy, hence we first register LiDAR scans with a LiDAR odometry method SuperOdometry [34] as shown in Fig. 3(a). SuperOdometry is a real-time and robust odometry method, which can meet our expectation of online adaptation. The resulting registered scans are then accumulated in a 120m  $\times$  120m robot-centered 3D voxel map at a 0.25m resolution (Fig. 3(b)).

b) *Extracting LiDAR Features:* After getting an online voxel map, we extract object height, gravity-aligned ground slope, and planarity features from the accumulated map to inform traversability. Leveraging these features together is valuable for off-road driving [10], [35]. For example, object height indicates tall obstacles, slope helps identify hard-to-traverse terrain, and planarity can differentiate between free space areas such as grass and trail. For object height  $f_h$ , we estimate the ground plane with a Markov Random Field-based Filter that interpolates between the heights of the neighboring lowest dense cells. For each voxel in a voxel pillar, we label its object height by the difference between the height of the tallest voxel and ground. The ground plane is also used to derive terrain slope  $f_s$ . For planarity  $f_p$ , we find the eigenvalues for each cell by applying singular value decomposition on the cell's LiDAR points. Using the eigenvalues, we calculate planarity by  $\frac{(\lambda_2 - \lambda_3) \times 2}{\lambda_1}$ . Planarity calculation benefits from accumulation as it requires dense and accurately registered LiDAR data, which is less likely at a distance.

c) *Combining Features to Traversability Costs:* We generate a continuous traversability score by combining LiDAR features into a cost function. This function scores terrain by increasing severity: a trail with low height and high planarity will be low cost, grass with low height and low planarity will be medium cost, regions with high terrain slope is also medium cost, and untraversable obstacles above a height threshold have high costs. For efficiency, we simply use the following cost function  $J$ , where  $h_{\text{thresh}} = 1\text{m}$ :

$$J(f_h, f_p, f_s) = \begin{cases} 10 & \text{if } f_h \geq h_{\text{thresh}} \\ 6 - f_p + f_s \in [0, 6] & \text{if } f_h < h_{\text{thresh}} \end{cases} \quad (1)$$

An example labeled 3D voxel map is shown in Fig. 3(c).

2) *Projecting LiDAR Labels to Image:* Given labeled voxels in 3D space, we project them to image space to generate pixel-wise label masks for visual traversability learning. Fig. 3(d) shows an example image/label pair.

a) *Near-to-far Accumulation:* To mitigate long-range LiDAR noise, we perform a *near-to-far learning* similar to [36], which accumulates data from a short-range sensor to supervise long-range models. Specifically, we associate an image taken at time  $t$  with a voxel map after a delay at  $t + d$ . The resulting labels are more accurate at farther range and cover a larger distance. To capture accurate labels, we trim the admissible voxels to be within 30m of past and future robot odometries, within which LiDAR provides good estimates, and within 100m of the odometry from the image frame.

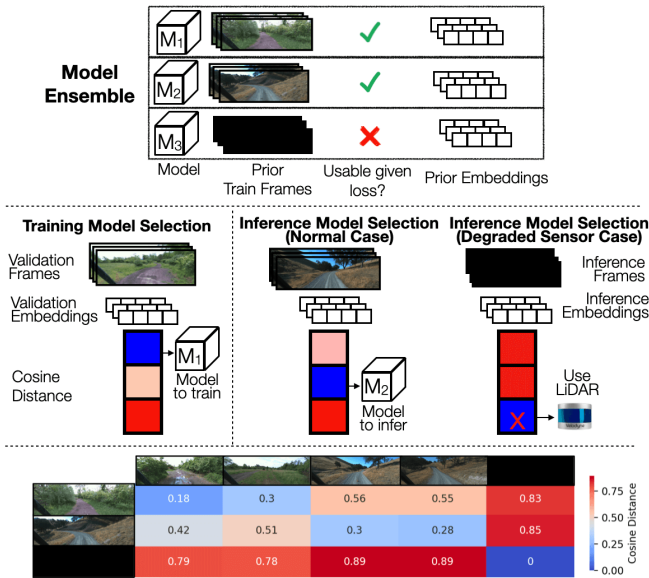


Fig. 4: **Training and Inference Model Selection:** For each training cycle, the model with the most similar embeddings to the validation frames is trained. If all models are dissimilar, a new model head is created. The inference model is selected using cosine distance. If this model has not been marked as *usable*, ALTER chooses LiDAR. The bottom shows cosine distances between different frames.

b) *Raycasting from Camera View:* While projecting a future accumulated map improves the labeling accuracy, a naive projection of LiDAR labels will also include labels that would have been occluded from the initial viewpoint. We use Open3D [37] to render the raycasted view of the voxel labels from the pose where the image is taken. By labeling *accumulated* LiDAR measurements in a near-to-far manner, we can accurately label pixels up to 100m along the driven-over path, instead of just 30m. This results in more accurate and filled-out pixel-wise labels in image space.

## B. Online Learning

As new image/label pairs are generated from self-supervision, we continuously train models to *adapt-on-the-drive* and predict fine-grained pixel-wise traversability scores.

Training and inference are done in the top-half of the image as the goal is long-range perception. We use a U-Net model [38] as the base architecture with a MobileNet backbone [39] pre-trained on ImageNet [40] for fast training. The encoder is frozen to keep embeddings aligned during continued training.

1) *Training Model Selection:* During operation, the vehicle may visit new and old environments with appearance variations. We therefore keep an ensemble of up to  $N$  specialized models, each fine-tuned for their settings, allowing for fast reactions when revisiting environments. Our model selection process is similar in spirit to [4] with three main differences, (1) their similarity measure is color-based while ours is image-based to include more global context, (2) theirs is purely similarity-based which lacks robustness, and (3) cannot degrade to LiDAR. Fig. 4 illustrates our *joint loss-similarity model selection* process. At a high level, we combine (1) image feature similarity-based selection which is generally a real-time signal, and (2) loss-based selection which provides

an absolute performance metric but lags behind the actual image as labels rely on accumulated LiDAR. The training model selection starts upon receiving an incoming training data buffer. From the ensemble, we select one candidate to train further on the incoming frames by using the shared encoder to generate embeddings and choosing the model whose prior training data embeddings have the lowest cosine difference to the embeddings of the new frames. If all models' cosine difference exceeds a threshold  $CD_{new}$ , a new model is made by copying the weights of the most similar model. If the model ensemble has reached capacity, the least recently used model is discarded. This strategy is vital for maintaining model relevance while adapting to new environments.

2) *Model Training:* To enable fast adaptation, we train the selected model for 10 seconds with a buffer of 10 frames from the most recent experience. Then, the training process restarts using the next selected model and buffer. The training duration is aligned to the buffer creation time so that the next cycle of model selection and training begins as soon as a new buffer of training data is available, improving model reactivity. To mitigate the impact of unknown pixels, the visual model is trained with an adapted Mean Squared Error using only pixels with self-assigned labels. To counter data imbalance, we weight losses on pixels of different traversability ranges using median frequency balancing [41].

3) *Training Model Usability Evaluation:* New model heads are not always preferred due to reasons such as sensor failures, e.g., a simple occlusion of the field of view can corrupt a visual model with fine-tuning. Therefore, we evaluate if the newly trained model head is usable instead of LiDAR after each training cycle. Specifically, if the model's loss on the last 20s of data is below a threshold  $L_{usable}$ , then the model will be marked as *usable*. Otherwise, we conservatively opt for LiDAR. We find adding this loss-based component is critical in two settings – when a model just started learning in a new setting and when a model is training with visually degraded data. LiDAR is the better option in both scenarios.

## C. Inference Method Selection and Traversability Inference

For efficiency, we select one model head for inference on a per-frame basis, which might not be the model currently training due to training data collection delays. To be robust to new environments, we compare the last 10 incoming frames' embeddings with those from each model in the ensemble. If the nearest model is similar enough given a cosine distance threshold  $CD_{LiDAR}$  and has been marked as *usable* given previous loss, the model will be used, otherwise LiDAR will be used. Fig. 4 shows an example of sensor failure (camera occluded) where LiDAR is used for inference.

## IV. RESULTS

Through our experiments, we aim to answer the following research questions: **Q1:** Does ALTER enable more robust performance in new environments than baseline visual methods? **Q2:** Does our visual model provide more accurate long-range perception than LiDAR-only estimates? **Q3:** How does ALTER perform in previously-learned environments? **Q4:** How does

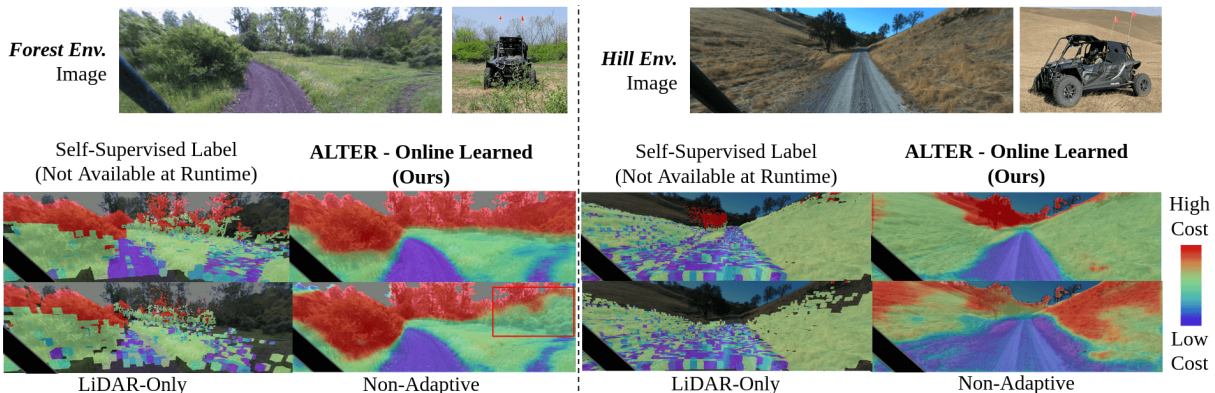


Fig. 5: **Qualitative Continuous Cost Prediction Results after 45s of Adaptation:** Our online-learned method is able to differentiate nuanced terrain (trail, grass, obstacle) at long-range with less than one minute of combined data collection and training. Compared to LiDAR-only baseline, our method can better differentiate grass and trail at a distance, even beyond voxel range. Compared to a non-adaptive model, our method provides more robust predictions when deployed in a new environment.

model selection result in a long-range perception system more robust to visual sensor failures?

#### A. Experimental Setup

1) *Platform and Dataset:* We use a Polaris RZR All-terrain Vehicle (ATV) platform equipped with RGB cameras (Multisense S27) and LiDARs (Velodyne VLP-32C) for data collection. To demonstrate the generalizability of online adaptation, we select visually distinct environments including *forests* featuring dark-green vegetation, green grass, and dark-brown colored trails in Pittsburgh, PA, and *hills* featuring dry yellow grass with sparse trees in San Luis Obispo, CA. We conduct experiments with two sequences from each environment and confine the duration of each sequence to 75s to showcase the capability for online adaptation.

2) *Implementation:* The LiDAR voxel map accumulates data for 5s for each image frame. Training data are gathered in 10s buffers, each containing 10 training and 4 validation frames. Labels are generated at an average of 0.65s per frame. After a 10s buffer is collected, the model trains for 10s at a learning rate of  $3e^{-3}$ . Images are captured at  $1024 \times 750$  pixels and resized to  $512 \times 384$  pixels. We use  $L_{\text{usable}} = 6$ ,  $CD_{\text{new}} = 0.45$  and  $CD_{\text{LiDAR}} = 0.75$ .

3) *Metrics:* We provide intersection-over-union (IOU) metrics using two comparisons, (1) *mIoU-10*: Pixels beyond 10 meters where all methods have labels, and (2) *mIoU-H*: full top-half image (removing sky and robot cage) to include areas beyond LiDAR, using bilinear interpolation to fill LiDAR gaps. Metrics are calculated against images manually segmented into three classes (low, medium, and high cost/untraversable) based on an expert operator’s sense of terrain desirability. In general, low cost reflects smooth trails, medium cost reflects grassy or steep areas, and high-cost reflects untraversable obstacles or dangerous cliffs. For comparison, we convert continuous cost predictions to discrete classes and map semantic class predictions from GANAV [11] to these three classes, as common in semantic navigation [9]. For *ALTER*, we threshold 0-2.5 to low, 2.5-7.5 to medium and 7.5 to high-cost. For GANAV, we map *smooth*, *bumpy* to low, *rough* to medium and *obstacle*, *forbidden* to high-cost.

4) *Baselines:* We select four baselines for comparison at long range and in out-of-distribution environments.

- 1) **LiDAR-only:** The geometric-based traversability map built from accumulated LiDAR points up to current time is commonly used for offroad driving [7], [10]. For comparison, the 3D scores are raycasted into image space. This differs from self-supervised labels, which incorporate future data and are not available in real-time.
- 2) **GANAV [11] (Trained on RELIS dataset [42]):** State-of-the-art visual semantic model for offroad driving, with author-provided weights pretrained on RELIS dataset.
- 3) **GANAV [11] (Trained on RUGD dataset [43]):** As above, with author-provided weights for RUGD dataset.
- 4) **Adaptive SVM, based on [5]:** This classic method incorporated a self-supervised adaptive visual model, the closest method to ours. It adapts a Support Vector Machine (SVM) online given LiDAR labels. For efficiency, we implement a version with Linear SVM and Stochastic Gradient Descent. As it predicts binary labels, we modify the approach to output three classes for fair comparison.

While there has been recent work on adaptive self-supervised visual traversability estimation using odometry [32], [33], the unavailability of source code and use of a different supervision source render a direct comparison inappropriate.

#### B. Does ALTER enable robust prediction in new scenarios?

First, we test *ALTER*’s robustness in new environments. We test on four sequences total across two environments, with the first 45s for training and the next 30s for evaluation. *ALTER* starts with a model trained on both sequences of the opposing environment.

Table I quantitatively compares *ALTER* with the baselines, showing *ALTER* provides more accurate long-range results across all sequences, with less than 1 minute of adaptation each. Adaptive SVM struggles to distinguish grass and trees of similar colors due to its reliance on color and texture features. Our U-Net-based approach leverages larger contextual information and attain more precise estimates. To isolate the impact of online learning, we compare to *ALTER without adaptation*, and show an average of 61%

TABLE I: **IOU Comparison After 45s of Adaptation in New Environments:** We compare our adaptive method to baselines on hand-labeled ground truths. *ALTER* is more robust in new environments than the baseline visual model and outperforms all baselines with less than a minute of data collection and adaptation time. (\*) Self-supervised labels are not available at runtime as they accumulate future data. They are included for comparison to show that *ALTER* outperforms even its own training data which while informative can be noisy.

	Forest Seq. 1		Forest Seq. 2		Hill Seq. 1		Hill Seq. 2		Over All Seqs.	
	mIoU-10	mIoU-H	mIoU-10	mIoU-H	mIoU-10	mIoU-H	mIoU-10	mIoU-H	mIoU-10	mIoU-H
<b>ALTER (Ours)</b>	<b>0.68</b>	<b>0.75</b>	<b>0.62</b>	<b>0.64</b>	<b>0.71</b>	<b>0.71</b>	<b>0.63</b>	<b>0.68</b>	<b>0.66</b>	<b>0.70</b>
Adaptive SVM	0.35	0.42	0.28	0.34	0.55	0.59	0.45	0.52	0.41	0.47
ALTER w/o Adapt.	0.63	0.69	0.48	0.57	0.26	0.33	0.26	0.31	0.41	0.48
GANAV (RELLIS)	0.32	0.35	0.29	0.26	0.11	0.13	0.11	0.13	0.21	0.22
GANAV (RUGD)	0.48	0.45	0.46	0.45	0.12	0.13	0.13	0.18	0.30	0.30
Lidar-Only	0.36	0.44	0.38	0.50	0.52	0.54	0.56	0.58	0.46	0.52
Self-supervised Label*	0.54	0.59	0.55	0.61	0.59	0.63	0.61	0.63	0.57	0.62

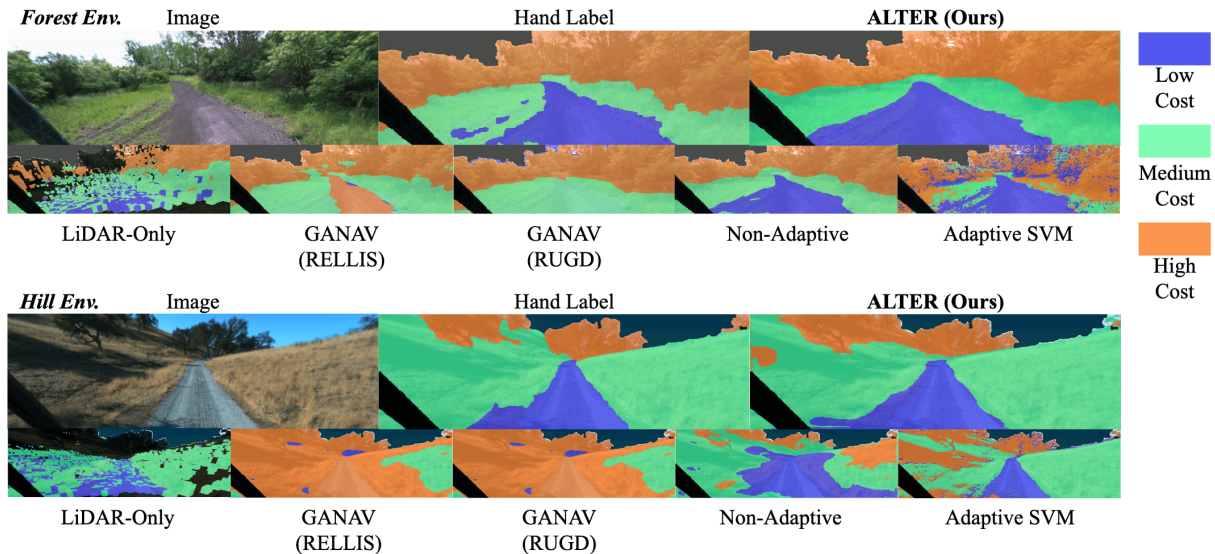


Fig. 6: **Qualitative Comparison to Hand-labels after 45s of Adaptation:** *ALTER* enables more accurate differentiation of the classes at long-range. Compared to non-adaptive methods, our adaptive method provides more robust predictions when deployed in a new environment.

TABLE II: **Comparison when Adapting to Multiple Domains:** Adaptive methods (top rows) outperform non-adaptive methods (mid rows) on sequences with multiple environments. *ALTER*'s model selection helps it to quickly adapt, leading to improved performance within the first 15 seconds (F15) of revisiting a similar environment.

mIoU	Pretrained → Forest 2 → Hill 2				Pretrained → Hill 2 → Forest 2			
	10	H	10-F15	H-F15	10	H	10-F15	H-F15
<b>ALTER (Ours)</b>	<b>0.69</b>	<b>0.67</b>	<b>0.64</b>	<b>0.66</b>	<b>0.70</b>	<b>0.68</b>	<b>0.64</b>	<b>0.66</b>
ALTER w/o MS	0.60	0.63	0.39	0.52	0.65	0.65	0.51	0.58
Adaptive SVM	0.35	0.39	0.16	0.18	0.44	0.46	0.39	0.36
GANAV (RELLIS)	0.21	0.22	0.23	0.26	0.20	0.22	0.23	0.26
GANAV (RUGD)	0.25	0.27	0.18	0.24	0.24	0.26	0.18	0.24
Lidar-Only	0.48	0.52	0.44	0.49	0.48	0.52	0.44	0.49

improvement. Importantly, we found that GANAV [11], a SOTA visual semantic model trained on open datasets, fails in our data, suggesting the need for online adaptation. Fig. 5 shows representative qualitative outputs and Fig. 6 shows comparisons to hand-labeled classes. *ALTER* clearly distinguishes low-cost planar trails, medium-cost grassy and sloped regions, and high-cost obstacles in both environments. In contrast, the non-adaptive methods fail to predict the trail accurately, as shown in the Hill Env. This result shows that *ALTER* improves robustness in new settings.

TABLE III: **Comparison over a Mission with Simulated Sensor Failure:** *ALTER* outperforms all baselines while *ALTER w/o MS* and other vision-based methods show poor resilience.

mIoU	Forest 2 w/ Failure				Hill 2 w/ Failure			
	10	H	10-F15	H-F15	10	H	10-F15	H-F15
<b>ALTER (Ours)</b>	<b>0.52</b>	<b>0.59</b>	<b>0.52</b>	<b>0.60</b>	<b>0.65</b>	<b>0.61</b>	<b>0.68</b>	<b>0.60</b>
ALTER w/o MS	0.30	0.34	0.26	0.32	0.38	0.35	0.36	0.34
Adaptive SVM	0.27	0.31	0.30	0.35	0.27	0.30	0.34	0.38
GANAV (RELLIS)	0.26	0.29	0.25	0.27	0.16	0.14	0.15	0.16
GANAV (RUGD)	0.42	0.29	0.43	0.32	0.13	0.20	0.12	0.16
Lidar-Only	0.40	0.50	0.38	0.48	0.55	0.53	0.52	0.48

### C. Does *ALTER* provide longer-range estimates?

This section examines *ALTER*'s ability to determine traversability at long-range compared to LiDAR. We observe that *ALTER* can provide 43% improvement in traversability estimates over LiDAR at long range. In the first row of Fig. 6, we show a scenario where our online-learned prediction differentiates the trail from the grass region at long range. In contrast, the LiDAR-only map does not distinguish the two areas well because the planarity feature of the trail requires a high density of points to be accurate. Moreover, in the second row of Fig. 6, the LiDAR-only baseline is unable to provide estimates at long range due to the small voxel map range

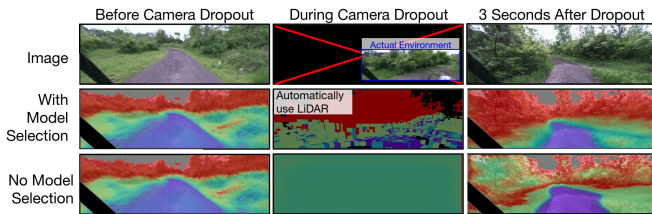


Fig. 7: **Camera Dropout Experiment:** Without model selection, the system predicts erroneously using the visual model during dropout. In contrast, *ALTER* automatically uses LiDAR during camera dropout and quickly recovers when the sensor is back.

(60m radius) needed for fast runtime. In contrast, *ALTER* can predict the trail and grass much further than even the voxel map size. As observed, *ALTER* is able to provide better long-range traversability estimates than using only LiDAR.

#### D. How does *ALTER* perform in learned environments?

We also validate *ALTER*'s ability to overcome catastrophic forgetting [28] with a multiple-scenario test setup. We provide two starting models for our system, one model trained on Seq. 1 of each environment. We test on two scenarios, Forest→Hill, Hill→Forest, where the vehicle transitions from one starting environment to another. To isolate the benefit of model selection, we also compare to *ALTER without model selection (ALTER w/o MS)*. For adaptive methods without model selection mechanism, we average the results from runs beginning from each starting model. In addition, we report the mIoU of the first 15s in each environment, denoted *F15*, to quantify a method's reactivity to a new environment.

Table II compares the different methods, with *ALTER* consistently outperforming baselines. With model selection, *ALTER* significantly improves its reactivity to new environment with much higher accuracy in the first 15 seconds than *ALTER without model selection*. While slower to react, we find even without model selection, *ALTER* also outperforms all other baselines, giving credit to the adaptive nature.

#### E. How does model selection operate for visual degradation?

We next examine how *ALTER*'s model selection module leads to more robustness against visual sensing failures. We follow the same experimental setup as the previous subsection. We test with sudden simulated *View Drop*, a standard camera failure tested in Computer Vision communities [44]. We simulate camera dropout of various lengths (between 20s and 50s) in each environment, to test each method's ability in times without camera images and when camera images come back. Table III and Fig. 7 display the comparison, where *ALTER* outperforms all methods. *ALTER without model selection* performs worse than LiDAR-only method, which is expected as LiDAR provides consistent, albeit shorter-range measurements during visual sensing failure. In addition, without model selection, the method continues to fine-tune on the faulty sensor data and cannot rapidly recover when the camera is normal. In contrast, *ALTER* can switch immediately to LiDAR when it senses it is out-of-distribution and then immediately uses the visual model when the sensor gives in-distribution inputs. This experiment validates that the model selection module can robustify against visual impairment.

## V. CONCLUSIONS

In this paper, we present *ALTER*, an offroad perception framework that *adapts on-the-drive* to predict dense traversability map in novel environments. *ALTER* is a self-supervised framework that *adapts* a visual model online using near-range LiDAR measurements. We incorporate a model selection module to pick the best model to train and, at inference, pick the best model for predictions or use LiDAR instead. We validate with real-world offroad driving datasets that *ALTER* significantly increases robustness when operating in new environments. We showed that the adapted visual model can provide more accurate long-range measurements than LiDAR, especially on terrains that require dense points to differentiate. Finally, we show that *ALTER* provides more reactivity when revisiting previously learned environments and improved robustness against visual sensing degradation. For future work, we are interested in incorporating visual foundation models [25], [45] for faster adaptation in harder-to-segment images, pretraining on large multi-modal datasets [46], [47], incorporating more data sources such as proprioception [18] for more robustness against LiDAR noise, and more intelligent data sampling [48] for improved performance.

## ACKNOWLEDGMENT

Distribution Statement 'A' (Approved for Public Release, Distribution Unlimited). This research was sponsored by DARPA (#HR001121C0189). The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. We thank JSV and EF for their assistance.

## REFERENCES

- [1] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, "Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision," *Ieee Access*, vol. 8, pp. 191 617–191 643, 2020.
- [2] J. E. Naranjo, M. Clavijo, F. Jiménez, O. Gomez, J. L. Rivera, and M. Anguita, "Autonomous vehicle for surveillance missions in off-road environment," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2016, pp. 98–103.
- [3] A. Ghobadpour, G. Monsalve, A. Cardenas, and H. Mousazadeh, "Off-road electric vehicles and autonomous robots in agricultural sector: trends, challenges, and opportunities," *Vehicles*, vol. 4, no. 3, pp. 843–864, 2022.
- [4] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski, "Self-supervised monocular road detection in desert terrain." in *Robotics: science and systems*, vol. 38. Philadelphia, 2006.
- [5] S. Zhou, J. Xi, M. W. McDaniel, T. Nishihata, P. Salesses, and K. Iagnemma, "Self-supervised learning to visually detect terrain surfaces for autonomous robots operating in forested terrain," *Journal of Field Robotics*, vol. 29, no. 2, pp. 277–297, 2012.
- [6] D. Langer, J. Rosenblatt, and M. Hebert, "A behavior-based system for off-road navigation," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 776 – 783, December 1994.
- [7] A. Kelly, O. Amidi, M. Bode, M. Happold, H. Herman, T. Pilarski, P. Rander, A. Stentz, N. Vallidis, and R. Warner, "Toward reliable off road autonomous vehicles operating in challenging environments," in *Proceedings of 9th International Symposium on Experimental Robotics (ISER '04)*, June 2004, pp. 599 – 608.
- [8] J. A. Bagnell, D. Bradley, D. Silver, B. Sofman, and A. Stentz, "Learning for autonomous navigation," *IEEE Robotics and Automation Magazine*, vol. 17, no. 2, pp. 74–84, 2010.

- [9] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time semantic mapping for autonomous off-road navigation," in *Proceedings of 11th International Conference on Field and Service Robotics (FSR '17)*, September 2017, pp. 335 – 350.
- [10] S. Triest, M. G. Castro, P. Maheshwari, M. Sivaprakasam, W. Wang, and S. Scherer, "Learning risk-aware costmaps via inverse reinforcement learning for off-road navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 924–930.
- [11] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, "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.
- [12] B. Sofman, E. Lin, J. A. Bagnell, N. Vandapel, and A. Stentz, "Improving robot navigation through self-supervised online learning," in *Robotics: Science and Systems*, 2006.
- [13] D. Maier, M. Bennewitz, and C. Stachniss, "Self-supervised obstacle detection for humanoid navigation using monocular vision and sparse laser data," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1263–1269.
- [14] T. Overbye and S. Saripalli, "G-vom: A gpu accelerated voxel off-road mapping system," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 1480–1486.
- [15] J. Mei, Y. Yu, H. Zhao, and H. Zha, "Scene-adaptive off-road detection using a monocular camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 242–253, 2017.
- [16] E. Tiu, E. Talius, P. Patel, C. Langlotz, A. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning," *Nature Biomedical Engineering*, vol. 6, pp. 1–8, 09 2022.
- [17] M. Sivaprakasam, S. Triest, W. Wang, P. Yin, and S. Scherer, "Improving off-road planning techniques with learned costs from physical interactions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2021, p. 4844–4850.
- [18] M. G. Castro, S. Triest, W. Wang, J. M. Gregory, F. Sanchez, J. G. Rogers, and S. Scherer, "How does it feel? self-supervised costmap learning for off-road vehicle traversability," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 931–938.
- [19] O. Mayuku, B. W. Surgenor, and J. A. Marshall, "A self-supervised near-to-far approach for terrain-adaptive off-road autonomous driving," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 14 054–14 060.
- [20] L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hutter, "Where should i walk? predicting terrain properties from images via self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1509–1516, 2019.
- [21] J. Zürn, W. Burgard, and A. Valada, "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 466–481, 2021.
- [22] R. Schmid, D. Atha, F. Schöller, S. Dey, S. Fakoorian, K. Otsu, B. Ridge, M. Bjelonic, L. Wellhausen, M. Hutter, and A.-a. Agha-mohammadi, "Self-supervised traversability prediction by learning to reconstruct safe terrain," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 12 419–12 425.
- [23] X. Yao, J. Zhang, and J. Oh, "Rca: Ride comfort-aware visual navigation via self-supervised learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7847–7852.
- [24] J. Seo, S. Sim, and I. Shim, "Learning off-road terrain traversability with self-supervisions only," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4617–4624, 2023.
- [25] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.
- [26] R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller, and Y. LeCun, "Deep belief net learning in a long-range vision system for autonomous off-road driving," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 628–633.
- [27] K.-Y. Lee, Y. Zhong, and Y.-X. Wang, "Do pre-trained models benefit equally in continual learning?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2023, pp. 6485–6493.
- [28] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, 1999.
- [29] C. de Masson d'Autume, S. Ruder, L. Kong, and D. Yogatama, "Episodic memory in lifelong language learning," *CoRR*, vol. abs/1906.01076, 2019.
- [30] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [31] M. Happold, M. Ollis, and N. Johnson, "Enhancing supervised terrain classification with predictive unsupervised learning," in *Proceedings of Robotics: Science and Systems*, Philadelphia, USA, August 2006.
- [32] A. J. Sathyamoorthy, K. Weerakoon, T. Guan, J. Liang, and D. Manocha, "Terrapn: Unstructured terrain navigation using online self-supervised learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 7197–7204.
- [33] J. Frey, M. Mattamala, N. Chebroul, C. Cadena, M. Fallon, and M. Hutter, "Fast Traversability Estimation for Wild Visual Navigation," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.
- [34] S. Zhao, H. Zhang, P. Wang, L. Nogueira, and S. Scherer, "Super odometry: Imu-centric lidar-visual-inertial estimator for challenging environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8729–8736.
- [35] J.-F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert, "Natural terrain classification using three-dimensional lidar data for ground robot mobility," *Journal of Field Robotics*, vol. 23, no. 10, pp. 839–861, 2006.
- [36] C. Wellington, A. Courville, and A. T. Stentz, "A generative model of terrain for autonomous navigation in vegetation," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1287–1304, 2006.
- [37] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.
- [38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [39] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.
- [40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [42] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.
- [43] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.
- [44] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for bev detection and map segmentation," *arXiv preprint arXiv:2304.09801*, 2023.
- [45] S. Jung, J. Lee, X. Meng, B. Boots, and A. Lambert, "V-strong: Visual self-supervised traversability learning for off-road navigation," *arXiv preprint arXiv:2312.16016*, 2023.
- [46] M. Sivaprakasam, P. Maheshwari, M. G. Castro, S. Triest, M. Nye, S. Willits, A. Saba, W. Wang, and S. Scherer, "Tartandrive 2.0: More modalities and better infrastructure to further self-supervised learning research in off-road driving tasks," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12 606–12 606.
- [47] D. Cheng, F. C. Ojeda, A. Prabhu, X. Liu, A. Zhu, P. C. Green, R. Ehsani, P. Chaudhari, and V. Kumar, "Treescope: An agricultural robotics dataset for lidar-based mapping of trees in forests and orchards," 2023.
- [48] C. Wang, Y. Qiu, W. Wang, Y. Hu, S. Kim, and S. Scherer, "Unsuper-vised online learning for robotic interestingness with visual memory," *IEEE Transactions on Robotics (T-RO)*, 2021.