2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

How Bumpy Is It? Incremental Online Learning of Terrain-Induced Bumpiness Costs for Off-Road Vehicles

Haoyu Yuan[1†], Tianwei Niu[1†], Shengshan Ma[1], Runjiao Bao[1] and Shoukun Wang[1]

***Abstract*— Stable autonomous driving in unstructured off-road environments remains a longstanding challenge. In the absence of structured roads and in the presence of uneven terrain, vegetation, and soil slopes, vehicles must rely on LiDAR–Camera fusion to identify stable and traversable roads. However, existing terrain perception methods largely remain at the level of semantic segmentation and struggle to capture physical attributes such as surface roughness and load-bearing capacity. Meanwhile, constructing datasets annotated with accurate physical properties is prohibitively costly and inherently limited in class diversity, making it difficult to cover unseen terrains. To address these limitations, we propose an online ground bumpiness cost learning framework for off-road vehicles, which enables continuous and direct learning of terrain-specific bumpiness costs during operation without the need for manual annotation. The framework consists of four key components: (i) ground bumpiness cost computation, (ii) a lightweight multimodal terrain segmentation model, (iii) an instance-level incremental update strategy, and (iv) a bumpiness cost mapping module. Extensive experiments on the EV-56 vibroseis truck demonstrate that the proposed framework can finely discriminate terrains with varying bumpiness costs and incrementally estimate costs for previously unseen terrains, thereby providing strong support for safe and reliable off-road autonomous driving.**

I. INTRODUCTION

Autonomous driving technologies for structured road have reached a high level of maturity. However, when vehicles operate in unstructured off-road environments, the applicability of existing methods degrades markedly [1], [2], [3], [4]. Vibroseis trucks—large, purpose-built vehicles for oil and gas exploration—must frequently traverse deserts, grasslands, and Gobi terrain. As shown in Fig. 1, these scenarios not only lack clear road markings and prior maps but also feature diverse and complex terrain conditions, including soft sandy soil, gravel surfaces, and hard rock formations. For such large platforms with stringent operational requirements, conventional road-detection pipelines or generic semantic-segmentation methods [5], [6] are insufficient to meet task needs.

In such environments, the core challenge of terrain perception lies not merely in "seeing the surface" but in "understanding it". Whether a vehicle can drive stably and safely depends on the physical properties of the ground, including surface roughness, slope, and load-bearing capacity [7].

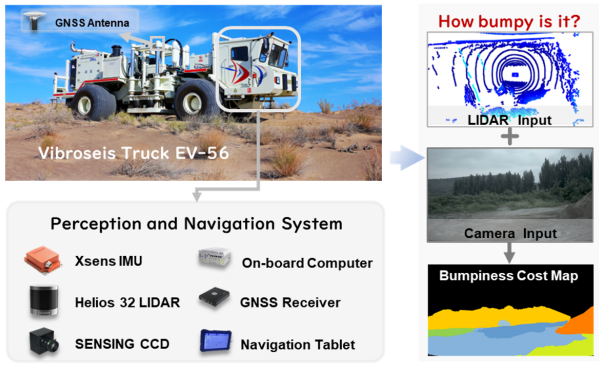

Fig. 1. A typical scenario in field operation environments. Equipped with an onboard perception system, the EV-56 truck can successfully predict the cost of terrain bumps in its surroundings using our proposed online learning framework.

These properties directly affect vehicle vibration, energy consumption, and operational efficiency. However, most vision- or LiDAR-based learning approaches remain focused at the semantic level [8], [9] and lack explicit modeling of these physical effects. They typically assign uniform traversability to all terrain within the same class, ignoring intra-class variations. In reality, terrains of the same semantic class may exhibit markedly different drivability—for example, firm grassland may be easily traversable, whereas soft grassland can be difficult to cross. This discrepancy creates a gap between perception results and vehicle control requirements, necessitating additional post-processing to convert perception outputs into cost information usable for planning and control, thereby increasing system complexity and potential risk.

On the other hand, the diversity and complexity of off-road environments make it challenging to train perception models via supervised learning. The primary obstacle lies in the prohibitive cost of constructing image datasets with accurate physical-property annotations. Several studies [8], [10] have attempted to leverage self-supervised methods by associating ground-truth terrain attributes with color images. However, these approaches require retraining on data collected online and are restricted to a predefined, limited set of terrain classes, which limits their ability to handle previously unseen terrain and reduces the efficiency of terrain assessment in unknown environments. Furthermore, since the robot typically traverses only a small portion of the imaged surface, the labels obtained through slice-based methods [10] are sparse, resulting in slow training convergence and elevated prediction noise.

*This study was supported by the National Natural Science Foundation of China (62473044).

†Haoyu Yuan and Tianwei Niu contributed equally to this work.

[1]Shoukun Wang (Corresponding author), Haoyu Yuan, Tianwei Niu, Shengshan Ma, Runjiao Bao are with School of Automation, Beijing Institute of Technology, 100081, China. Email: bitwsk@bit.edu.cn

To address the above challenges, this work proposes an online learning framework for terrain-induced bumpiness cost, tailored for off-road vehicles. The framework comprises four key components: (i) a ground bumpiness cost computation module for generating pseudo-labels of the current terrain; (ii) a lightweight multimodal terrain segmentation model for recognizing different terrain classes; (iii) an instance-level incremental update mechanism for maintaining label consistency of the same terrain type across space and time; and (iv) a bumpiness cost mapping module for constructing a cost map that encodes terrain-specific bumpiness characteristics. The proposed framework is deployed on an EV-56 vibroseis truck and validated through extensive experiments, demonstrating its effectiveness.

The main contributions of this work are summarized as follows:

- We propose an online ground bumpiness cost learning framework tailored for off-road vehicles, capable of real-time prediction without additional pretraining.
- We design a bumpiness cost computation method derived from IMU signals, enabling learning without manual annotations.
- We introduce an incremental spatio-temporal update strategy for terrain instances, allowing classification of unseen terrains without relying on predefined classes.
- We develop a bumpiness cost mapping module that continuously refines the relationship between terrain classes and bumpiness costs.

II. RELATED WORK

A. Off-Road Terrain Perception

In autonomous navigation, the ability to perceive surrounding terrain in real time under off-road conditions is a critical prerequisite for reliable path planning and safe driving. Traditional approaches typically model the environment using occupancy grids [11], geometric features [12], or semantic labels [13], and project the results into a bird's-eye view (BEV) representation for integration with the planner [14], [15]. For instance, SimpleBEV [16] adopts a forward-sampling strategy to obtain BEV features, while TerrainNet [7] combines stereo depth completion with soft quantization to improve both accuracy and runtime efficiency in off-road environments.

In recent years, semantic scene completion–based terrain perception methods [7], [17] have attracted significant attention. These approaches generate high-precision terrain segmentation by predicting dense 3D maps [18]. However, they typically rely on complete point clouds from continuous LiDAR scans and struggle to maintain stable predictions in challenging environments with only sparse LiDAR input. Other studies have explored image–point cloud fusion methods [19], [20], which extract complementary features in a unified BEV space and leverage depth completion and uncertainty modeling to mitigate the effects of point cloud sparsity and occlusion. Although such methods have achieved excellent performance in structured road scenarios, they still face challenges of generalization and real-time operation in unmarked, undulating off-road environments.

B. Terrain Traversability Analysis

Terrain traversability aims to quantify the feasibility of robot motion over a given terrain and is a key enabler for off-road navigation. Early studies primarily relied on vision-based or geometric features, employing classifiers to categorize terrain into traversable and non-traversable regions [21], [22], or estimating surface roughness scores using indicators such as planarity [23] or eigenvalue-based metrics [24]. With the advancement of semantic segmentation, several works [9], [18] have attempted to first partition terrain into discrete semantic classes and then map them to cost layers for planning. However, such hard-coded mappings lack flexibility and struggle to adapt to varying terrain conditions.

Recent research has shifted toward directly learning continuous cost maps from multimodal inputs. For example, the BADGR system [25], [26] employs a binary predictor to determine whether a candidate action sequence will induce excessive bumpiness. Other studies leverage neural networks to generate continuous cost maps from LiDAR or RGB data [27], [28], and incorporate conditional risk modeling to account for uncertainty [28]. 3DTTNet [29] further integrates image semantics with LiDAR geometry to construct multi-layer cost maps and enriches them with attributes such as slope and roughness to characterize terrain cost more comprehensively. Nevertheless, these methods still struggle with inference in occluded regions and maintaining global 3D consistency, making it challenging to deliver stable cost estimates in highly complex terrains.

III. METHOD

This section presents our terrain-induced bumpiness cost aware incremental online learning framework, as illustrated in Fig. 2. The framework consists of four key modules: (i) computation of the bumpiness cost for the current terrain surface; (ii) lightweight LiDAR-camera terrain segmentation; (iii) terrain instance incremental update and aggregation; and (iv) mapping from terrain classes to bumpiness costs.

A. Pseudo-Labeling of Ground Bumpiness Cost

Our goal is to learn a continuous and normalized ground bumpiness cost function that characterizes the interaction between the robot and different terrain regions, and can be directly exploited by the downstream path planning module. Previous studies [8], [30] have shown that the frequency response of IMU accelerations serves as a reliable indicator of terrain roughness, undulation, and deformability. Unlike prior approaches that consider only the IMU's vertical acceleration, we additionally incorporate the robot's angular accelerations about the X- and Y-axes to capture pitch- and roll-induced disturbances caused by uneven terrain. The joint use of these three signals enables a comprehensive characterization of bumpiness across varying velocities and attitudes. To derive a scalar quantity that reflects the overall

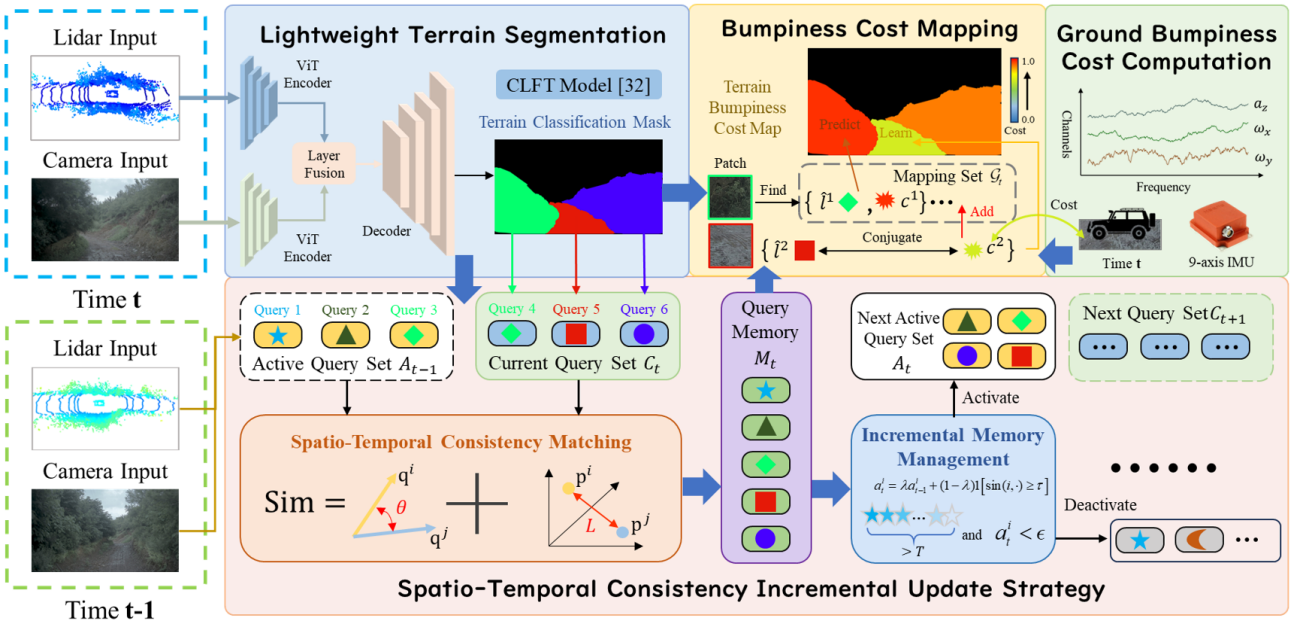

Fig. 2. Overview of the Proposed Online Terrain Bump Cost Learning Framework. The lightweight terrain segmentation module leverages RGB images and LiDAR point clouds to categorize the current terrain, and generates a query pair E_t for each category (illustrated as color-coded primitive shapes). The spatio-temporal consistency incremental update module performs similarity matching between the newly generated query set and the historical active query set: newly emerging queries are registered (red squares, blue circles), identical queries are merged (green diamonds), and queries that do not appear for multiple consecutive frames are marked as inactive (light-blue pentagrams). The bumpiness cost mapping module is used to predict or learn the bumpiness level of the current terrain. For newly emerging queries, they are associated with the bumpiness costs computed by the bumpiness cost computation module and added to the mapping set ($\{\hat{l}^2, C^2\}$); for existing queries, the corresponding bumpiness costs are directly predicted ($\{\hat{l}^1, C^1\}$).

terrain roughness, bumpiness, and deformability, we compute the band power of the vertical acceleration a_z, pitch angular acceleration ω_x, and roll angular acceleration ω_y, after normalizing each signal to zero mean and unit variance to eliminate scale and magnitude discrepancies. The ground bumpiness cost C is then defined as:

$$C = \int_{f_{\min}}^{f_{\max}} \left[\tilde{B}_{a_z}(f) + \tilde{B}_{\omega_x}(f) + \tilde{B}_{\omega_y}(f)\right] df, \quad (1)$$

where $\tilde{B}_{a_z}$, $\tilde{B}_{\omega_x}$, and $\tilde{B}_{\omega_y}$ denote the power spectral densities (PSDs) of the standardized a_z, ω_x, and ω_y signals, respectively, computed using the Welch method [31]. The parameters $f_{\min}$ and $f_{\max}$ specify the frequency range used for the band-power calculation. Based on empirical annotation experience, a frequency band of 5–35 Hz is selected, and the resulting band powers are further normalized with respect to the statistics of the measured trajectory data.

B. Lightweight Lidar-Camera Segmentation Model

Given the ground RGB image $\mathbf{I}_t$ and the synchronized LiDAR point cloud $\mathbf{L}_t$ at time t, we employ the pretrained CLFT model [32] to jointly parse the two modalities and generate a pixel-wise terrain mask $\mathbf{S}_t \in \{0,1\}^{H\times W\times c}$, where H and W denote the image height and width, respectively, and c is the number of terrain classes. To satisfy the real-time and computational constraints of the embedded platform, we replace the default ViT-Base encoder in the CLFT framework with a lighter ViT-Tiny variant, reducing the embedding dimension to 192 while keeping the patch size fixed at 16 to ensure feature alignment with the decoder. In the decoding stage, the projection dimension D is proportionally reduced to 128, and sparse regularization is applied to the convolutional layers of the decoder. Finally, channel pruning and knowledge distillation are employed to further compress the computational load without causing a significant loss of segmentation accuracy.

Considering the typical error patterns in unsupervised settings, we explicitly distinguish two common types of mismatch in the pixel-level post-processing of CLFT outputs: **Inter-class Merging** (different terrain types erroneously classified as the same class) and **Intra-class Over-segmentation** (a single terrain region fragmented into multiple segments). The former directly compromises geometric and semantic consistency, degrading subsequent cost-map construction, whereas the latter can be mitigated by the subsequent incremental aggregation module. To suppress inter-class merging while tolerating moderate over-segmentation, we apply a dual-threshold strategy to the class posterior $\mathbf{P}_t \in \mathbb{R}^{H\times W\times c}$: a pixel is accepted only if $\max_k \mathbf{P}_t(x,y,k) \geq \tau_{\text{conf}}$, where the confidence threshold is set to $\tau_{\text{conf}} = 0.5$; region merging is performed only when the mask intersection-over-union (IoU) satisfies IoU $\geq \tau_{\text{IoU}} = 0.9$. This high

IoU threshold and moderate confidence constraint encourage the model to push the operating point toward the finer boundary and minimize merging, meaning it would rather produce moderate oversegmentation than incorrectly merge different terrains into the same category. Ultimately, the lightweight CLFT model achieves high-precision, real-time terrain semantic segmentation under limited computational resources. It is important to note that we employed the pre-trained CLFT model to accurately segment different terrain categories, not to focus on the semantic classification of the segmented terrain. Therefore, we do not need to be concerned about the impact on semantic accuracy.

C. Incremental Update Strategy for Terrain Instance Spatio-Temporal Consistency

Since we only perform terrain category segmentation, the terrain labels generated by the CLFT model may be inconsistent across different times and spaces. In images and LiDAR point clouds captured from different perspectives and at different times, terrain belonging to the same category may not share the same label. To enable online discovery and incremental updating of newly observed terrain classes, we propose a query-based framework for maintaining spatio-temporal consistency of terrain instances. Leveraging the temporal modeling capability of queries, each segmented region is represented as an entity that evolves across frames. Through spatio-temporal consistency matching, query updates, and incremental memory management, the framework achieves long-term label consistency and facilitates semantic expansion.

1) Query representation: At time t, the input data is defined as $\mathcal{X}_t = \{\mathbf{I}_t, \mathbf{L}_t\}$, where $\mathbf{I}_t$ is the RGB image captured by the onboard camera and $\mathbf{L}_t$ is the synchronized 3D point cloud. The CLFT segmentation model produces a pixel-wise semantic label map $\mathbf{Y}_t \in \{1, \dots, c\}^{H \times W}$, where each pixel is assigned to one of the c terrain classes. To convert the pixel-level output into a manageable set of instances, for each class c we extract connected components $R_t^{(c,k)}$ and compute the 3D geometric center on the corresponding projected point set $\mathcal{P}_t^{(c,k)}$ as:

$$\mathbf{p}_t^{(c,k)} = \frac{1}{\left|\mathcal{P}_t^{(c,k)}\right|} \sum_{\mathbf{x} \in \mathcal{P}_t^{(c,k)}} \mathbf{x}, \tag{2}$$

where (c, k) denotes the k-th connected component of class c, and $\mathbf{x}$ represents the 3D coordinates of points in $\mathcal{P}_t$. The query vector for each region, $\mathbf{q}_t^{(c,k)}$, is obtained by jointly embedding its appearance and geometric features:

$$\mathbf{q}_t^{(c,k)} = \phi\big(f_{\text{rgb}}(R_t^{(c,k)}),\ f_{\text{geom}}(\mathcal{P}_t^{(c,k)})\big), \tag{3}$$

where $f_{\text{rgb}}(\cdot)$ extracts image texture features, $f_{\text{geom}}(\cdot)$ extracts 3D shape information, and $\phi(\cdot)$ denotes a multilayer perceptron for joint embedding.

Each candidate region is represented as a query pair $E_t^{(c,k)} \triangleq (\mathbf{q}_t^{(c,k)}, \mathbf{p}_t^{(c,k)})$, and the query set of the current frame is defined as $\mathcal{C}_t = \{E_t^{(c,k)}\}$, which consists of multiple query pairs. Meanwhile, we maintain a global query memory $\mathcal{M}_t = \{(\bar{\mathbf{q}}^i, \mathbf{p}^i), (\ell^i, a^i, \mathcal{H}^i)\}$ that evolves over time. This memory maintains the states of all instances, where $\bar{\mathbf{q}}^i$ is the temporally smoothed query vector, $\mathbf{p}^i$ is the 3D reference point, ℓ^i is the terrain class label, a^i represents the instance activity score, and $\mathcal{H}^i$ records the historical observations. If a candidate region cannot be matched to any existing instance in the memory, it is marked as a new instance, a new query pair E_{new} is initialized and registered into $\mathcal{M}_t$, and a new label ℓ^{new} is assigned to ensure correct temporal association in subsequent frames.

2) Spatio-temporal consistency matching: To align the labels between the current and previous frames, the system performs optimal matching between the active query set from the previous frame $\mathcal{A}_{t-1} \subset \mathcal{M}_{t-1}$ and the candidate set $\mathcal{C}_t$ at each time step. We first predict the reference point of each historical query in the current coordinate frame:

$$\widetilde{\mathbf{p}}_t^i = T_{t-1 \to t}\, \mathbf{p}_{t-1}^i, \tag{4}$$

where $T_{t-1 \to t}$ is the rigid-body transformation matrix derived from the vehicle odometry, extrinsic parameters, and the elapsed time. The appearance–geometry similarity between a candidate region j and a historical query i is then computed as:

$$\text{Sim}(i,j) = \alpha \frac{\bar{\mathbf{q}}_{t-1}^i \cdot \mathbf{q}_t^j}{\|\bar{\mathbf{q}}_{t-1}^i\| \|\mathbf{q}_t^j\|} + \beta \exp\left(-\frac{\|\widetilde{\mathbf{p}}_t^i - \mathbf{p}_t^j\|^2}{2\sigma^2}\right), \tag{5}$$

where the first term is the cosine similarity of query features, the second term measures the spatial proximity after motion compensation, and α, β balance the semantic and geometric terms. The optimal matching π^* is obtained by maximizing the total similarity using the Hungarian algorithm:

$$\pi^* = \arg\max_{\pi \in \Pi} \sum_{(i,j) \in \pi} \text{Sim}(i,j), \tag{6}$$

where Π represents the set of all feasible matchings satisfying the one-to-one constraint. A candidate inherits the label of its matched query if its optimal matching similarity exceeds a threshold τ; otherwise, it is treated as a new instance with a newly assigned label:

$$\ell_t^j = \begin{cases} \ell_{t-1}^{i^*}, & \text{if } \text{Sim}(i^*, j) \geq \tau, \\ \ell_{\text{new}}, & \text{otherwise.} \end{cases} \tag{7}$$

After matching, the query state is updated using the latest observation to absorb the effects of vehicle motion and local perception noise. Let $(\mathbf{q}_t^{j^*}, \mathbf{p}_t^{j^*})$ denote a successfully matched query pair, the reference point is updated in a residual form:

$$\mathbf{p}_t^i = \widetilde{\mathbf{p}}_t^i + \gamma\big(\mathbf{p}_t^{j^*} - \widetilde{\mathbf{p}}_t^i\big), \tag{8}$$

where $\gamma \in (0, 1]$ controls the correction strength. The query vector is updated using an exponential moving average (EMA):

$$\bar{\mathbf{q}}_t^i = (1 - \eta)\, \bar{\mathbf{q}}_{t-1}^i + \eta\, \mathbf{q}_t^{j^*}, \tag{9}$$

where $\eta \in (0, 1]$ is the EMA coefficient.

For cross-class information management, the system maintains a prototype dictionary $\mathcal{D}_t = \{(\mathbf{m}_c, \Sigma_c)\}$, where $\mathbf{m}_c$ and Σ_c represent the mean feature and covariance of class c, respectively. When a query is associated with class c, the prototype statistics are updated as:

$$\begin{cases} \mathbf{m}_c \leftarrow (1-\rho)\,\mathbf{m}_c + \rho\,\overline{\mathbf{q}}_t^i, \\ \Sigma_c \leftarrow (1-\rho)\,\Sigma_c + \rho\,(\overline{\mathbf{q}}_t^i - \mathbf{m}_c)(\overline{\mathbf{q}}_t^i - \mathbf{m}_c)^\top. \end{cases} \tag{10}$$

When the Mahalanobis distances d_c between a candidate region and all existing class prototypes exceed a threshold δ, a new class is created, and its entry is registered into both $\mathcal{D}_t$ and $\mathcal{M}_t$, thus enabling online discovery and incremental storage of previously unseen terrain types.

3) Incremental memory management: To ensure long-term stability and storage efficiency during continuous operation, each query maintains an activity score a_t^i, with an initial value of 1. This value is updated as follows:

$$a_t^i = \lambda a_{t-1}^i + (1-\lambda)\mathbf{1}\left[\mathrm{Sim}(i,\cdot) \geqslant \tau\right], \tag{11}$$

where $\lambda \in [0,1]$ is the activity decay factor. If a query remains unmatched for T consecutive frames and $a_t^i < \epsilon$, it is removed from the active set $\mathcal{A}_t$ but retained in the global memory $\mathcal{M}_t$. If a candidate instance j cannot be matched to any element in $\mathcal{A}_t$, but there exists a class c in the prototype dictionary $\mathcal{D}_t$ such that:

$$d_c(j) = \sqrt{(\mathbf{q}_t^j - \mathbf{m}_c)^\top \Sigma_c^{-1} (\mathbf{q}_t^j - \mathbf{m}_c)} < \delta. \tag{12}$$

we regard this candidate as belonging to a previously known but currently inactive class c. Accordingly, we reinsert it into the active set $\mathcal{A}_t$ and reset its activity score to 1. This design prevents inactive instances from interfering with matching while preserving their semantic and historical features, thus avoiding redundant label creation and preventing the loss of rare terrain types.

Moreover, to mitigate semantic drift and catastrophic forgetting, a historical buffer $\mathcal{H}_i = \{\mathbf{q}_\tau^i \mid \tau \leq t\}$ is maintained for each query $\overline{\mathbf{q}}_t^i$. On demand, the buffer can be aggregated using a sliding average or attention-weighted fusion to obtain a temporally stable representation:

$$\widetilde{\mathbf{q}}_t^i = \frac{1}{|\mathcal{H}^i|} \sum_{\tau \in \mathcal{H}^i} \mathbf{q}_\tau^i. \tag{13}$$

This mechanism allows the system to leverage the appearance and geometric priors stored in $\mathcal{H}_i$ to recover from short-term occlusion, partial missing observations, or illumination changes, while providing stable features for incremental updates of rare terrain samples.

With the integration of the above three components, terrain instances form a persistent and incrementally extensible query set that maintains consistent labeling of the same terrain type over time and continuously absorbs new classes with optimized representations, providing stable classification support for the subsequent construction of the terrain bumpiness cost map. The empirically selected hyperparameters for this module are summarized in Table I.

TABLE I: Parameter definitions and ranges.

Symbol	Meaning	Range
α	Semantic weight	$0.6 \sim 0.7$
β	Geometric weight	$0.3 \sim 0.4$
τ	Matching threshold	$0.55 \sim 0.7$
γ	Correction strength coefficient	$0.3 \sim 0.5$
η	EMA update coefficient	$0.05 \sim 0.2$
ρ	Prototype momentum coefficient	$0.02 \sim 0.1$
δ	Mahalanobis distance threshold	$2.5 \sim 3.0$
λ	Activity decay factor	$0.85 \sim 0.95$
T	Continuity window size	$5 \sim 10$
ϵ	Activeness threshold	$0.1 \sim 0.2$

D. Bumpiness cost mapping

Building upon Section III-A and Section III-C, we design a bumpiness cost mapping method that incrementally establishes a statistical relationship between terrain class labels and bumpiness costs, thereby constructing a real-time traversability-aware bumpiness cost map around the robot.

We first utilize the LiDAR odometry from lio-sam [33] to aggregate all local submaps into a global map and determine the robot's pose within the map. Based on the actual chassis dimensions, we extract a local patch B_t of approximately 2×2 m centered at the robot's current position. Using the method described in Section III-C, we obtain the semantic label of each grid cell within the patch $\ell_t(u,v)$, leading to the label distribution $L_t = \{\ell_t(u,v) \,|\, (u,v) \in B_t\}$. We then associate the bumpiness cost with the dominant terrain label by selecting the most frequent class:

$$\hat{l}_t = \arg\max_{l \in L_t} \mathrm{freq}(l), \tag{14}$$

where $\mathrm{freq}(l)$ denotes the occurrence frequency of label l within the patch. The instantaneous bumpiness cost C_t is paired with the selected label to form an association tuple $U_t = (\hat{l}_t, C_t)$. To accumulate the cost statistics over time, we maintain a global set $\mathcal{G}_t$ storing the label–cost mapping:

$$\mathcal{G}_t = \mathcal{G}_{t-1} \cup U_t. \tag{15}$$

For an existing label $\hat{l}_t$, its cost estimate is updated using an exponential moving average:

$$\bar{C}_t(\hat{l}_t) = (1-\rho)\,\bar{C}_{t-1}(\hat{l}_t) + \rho\, C_t, \tag{16}$$

where $\bar{C}_t(\hat{l}_t)$ is the cumulative estimate of the bumpiness cost for label $\hat{l}_t$, and ρ is the momentum coefficient. If a new semantic label $\hat{l}_{\mathrm{new}}$ appears within the patch, a new entry is created in $\mathcal{G}_t$ and its corresponding C_t is measured in subsequent observations.

Through the above process, a progressively refined mapping is established between different terrain types and their associated bumpiness costs. As more samples are observed, the robot gradually improves its online prediction capability for bumpiness cost in complex outdoor terrains, rather than being limited to a fixed set of predefined semantic classes.

IV. EXPERIMENTS

In this section, we conduct extensive experiments to validate the proposed framework and each of its components.

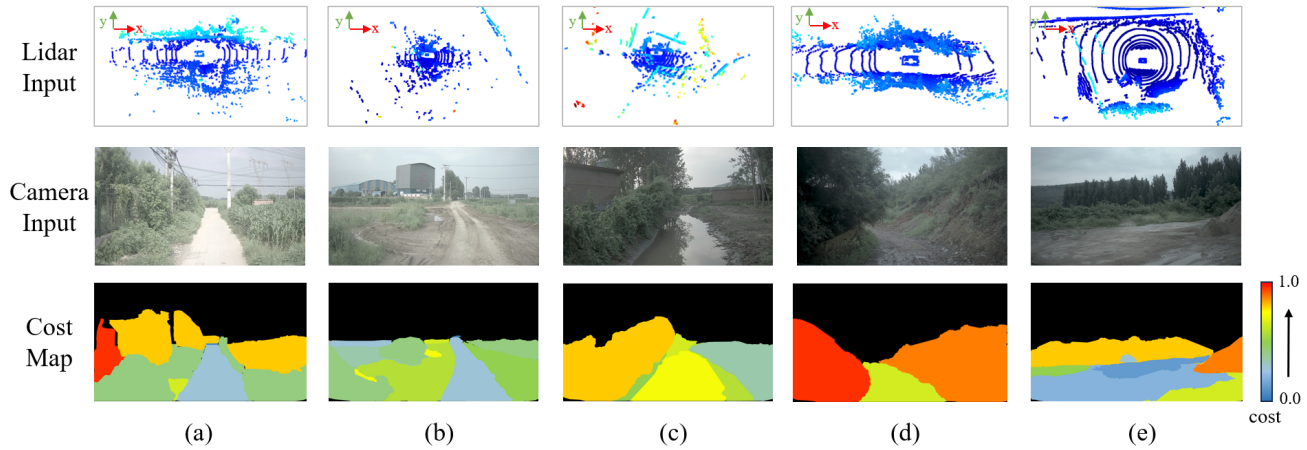

Fig. 3. Qualitative results of the bumpiness cost learning framework. Different bumpiness costs are visualized using distinct colors. Subfigures (a)-(e) present representative examples under varying semantic and geometric conditions.

All experiments are carried out on an EV-56 vibroseis truck, as illustrated in Fig. 1. The platform is equipped with a SENSING RGB camera, a Helios 32 LiDAR, an Xsens IMU, and a GPS module. All models are trained on a server with four NVIDIA L40 GPUs and deployed for real-time operation on an industrial computer with an NVIDIA Jetson AGX Orin 64GB module.

A. Collected Dataset

We conducted data collection using the EV-56 vibroseis truck across multiple open-field sites in Xushui County, Hebei Province, China. The vehicle was driven at a speed of 25 km/h. RGB images were recorded by a SENSING camera with a resolution of 1920×1080 pixels at 12 Hz, while LiDAR point clouds were captured at 20 Hz, and IMU/GNSS data were logged at 50 Hz. The resulting dataset comprises 34 manually driven sequences. After discarding invalid data, we obtained 58,779 image frames, 193,123 LiDAR scans, and 33 IMU/GNSS trajectories. The dataset covers a wide variety of representative off-road terrains, including dirt roads, grassland, puddles, and steep slopes, as illustrated in Fig. 3.

B. Overall Effectiveness of the Bumpiness Cost Learning Framework

In this section, the proposed framework is deployed on the EV-56 vibroseis truck and validated through field tests across multiple off-road environments. Representative visualization results from the testing process are shown in Fig. 3.

As illustrated in the figure, the learned cost map effectively distinguishes terrains with varying traversability and their associated costs. In the countryside path scenario (Fig. 3.a), grass-covered surfaces incur higher costs than flat dirt roads. In the waterlogged road scenario with prominent semantic features (Fig. 3.c), puddles are assigned higher costs than the surrounding mud. In the mountainous road scenario with dominant geometric features (Fig. 3.d), steep slopes exhibit significantly higher costs than gentle dirt surfaces. In contrast, in the gravel road scenario where neither semantic nor geometric cues are salient (Fig. 3.e), sandy and dirt roads still show clearly different bumpiness costs. These results demonstrate that the proposed framework not only integrates semantic and geometric terrain information but also, through the IMU-based bumpiness quantification, effectively differentiates terrain costs even in the absence of explicit semantic or geometric distinctions.

In addition, Fig. 3.b shows that even within the same mud road, the cost of the section adjacent to the puddle on the left is higher than that of the central portion. This further indicates that the proposed bumpiness cost learning framework can finely discriminate cost variations within the same terrain class, thereby avoiding estimation errors that arise from naively assigning identical costs to all instances of a given class.

C. Effectiveness of Segmentation and Spatio-Temporal Aggregation

1) Qualitative Experiments: To assess the effectiveness of the proposed spatio-temporal aggregation method for terrain instances, we evaluated it on perception data collected across different time steps. Representative terrain slices from the test set are shown in Fig. 4. As illustrated, even for terrain regions observed under entirely different viewpoints and lighting conditions, our method consistently assigns identical bumpiness costs. This demonstrates its robustness in maintaining spatio-temporal alignment.

TABLE II: Ablation Study on Spatio-Temporal Aggregation of Terrain Instances.

Method	Acc (%)	Ter (%)	Ier (%)
w/o SCM & w/o IMM	36.4	33.7	29.9
w/o SCM	52.5	29.3	18.2
w/o IMM	79.3	20.7	0.0
Ours	**90.7**	**9.3**	**0.0**

2) Quantitative Ablation: To evaluate the contribution of different components to the performance of spatio-temporal

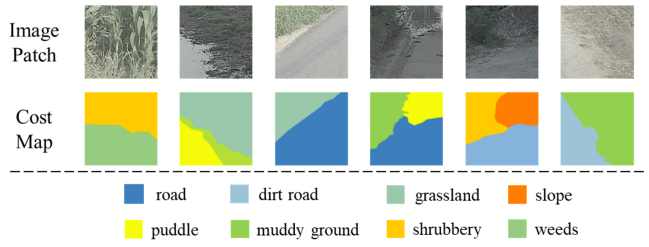

Fig. 4. Qualitative results of spatio-temporal aggregation of terrain instances. Legends in different colors correspond to the same terrain regions with consistent bumpiness costs, preserved across varying viewpoints, times, and illumination conditions.

aggregation, we conducted ablation experiments. Using 500 randomly selected test samples, we measured classification accuracy (Acc), tolerable error rate (Ter), and intolerable error rate (Ier) as performance metrics. Here, Ter is defined as the proportion of cases where identical terrain regions are split into different classes. While such errors reduce perception accuracy, they do not mislead the estimation of bumpiness cost. In contrast, Ier is defined as the proportion of cases where distinct terrain regions are erroneously merged into the same class. This type of error directly causes incorrect bumpiness cost estimation, thereby posing risks to safe motion control.

The results of the ablation study are summarized in Table II, where w/o SCM denotes replacing spatio-temporal consistency matching with conventional feature clustering, and w/o IMM indicates disabling incremental memory management. As shown in Table II, SCM effectively prevents different terrains from being mistakenly merged into the same category, thereby enabling accurate matching of related terrains across time and space. IMM further enhances the discriminability within the same terrain class by leveraging prior information to provide feature recovery from memory. When combined, the two modules yield a 54.3% improvement in classification accuracy over the baseline approach.

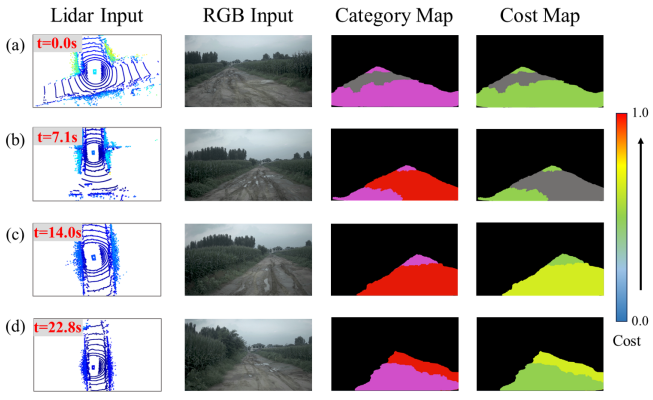

Fig. 5. Detailed process of incremental updates for unknown terrain. (a) Robot encounters a new terrain. (b) Establishes new terrain category labels. (c) Acquires new terrain bumpiness costs. (d) Predicts remembered terrain bumpiness costs.

3) Effectiveness of Incremental Updating: To further validate the effectiveness of the proposed incremental updating mechanism, we selected several timestamps from a representative sequence and illustrate the process of incrementally learning bumpiness costs for previously unseen terrains, as shown in Fig. 5. The mask colors in the Category map represent terrain category labels. To highlight changes in the central road prediction, we removed terrain category labels and bump cost masks outside the road.

When the robot first encounters an unknown terrain (the pothole region in Fig. 5.a), the lack of sufficient feature samples results in this terrain being labeled as unknown, with its bumpiness cost assigned as NaN; both are displayed as gray regions in Fig. 5.a. As the robot moves forward and establishes a new terrain class through feature queries, the region is assigned a new semantic label (the red region in Fig. 5.b). However, because the robot has not yet traversed this terrain, its bumpiness cost remains NaN. Once the robot enters the previously unseen terrain, the bumpiness cost mapping module assigns an estimated cost to this class, which is written into the corresponding entry (Fig. 5.c, yellow region). Thereafter, when the robot encounters the same terrain type again, the system can proactively predict its bumpiness cost, thereby enabling incremental updates of terrain-specific cost estimates.

V. CONCLUSIONS

This paper presents an online terrain bumpiness cost learning framework tailored for off-road vehicles. The framework comprises four key components—ground bumpiness cost computation, a lightweight multimodal terrain segmentation model, an incremental instance-level updating method, and a bumpiness cost mapping module—enabling off-road autonomous vehicles to incrementally learn and predict the bumpiness costs of surrounding terrains in real time. Extensive experiments conducted on a vibroseis truck validate the framework's effectiveness in predicting and incrementally updating bumpiness costs for previously unseen terrains.

Nevertheless, the current design does not explicitly account for vehicle dynamics, such as velocity and suspension performance. A promising direction for future research is to incorporate vehicle dynamics into the framework to improve adaptability across different vehicle platforms.

REFERENCES

[1] P. U. Lima, "Search and rescue robots: The civil protection teams of the future," in *Proceedings of the 2012 Third International Conference on Emerging Security Technologies*, ser. EST '12. USA: IEEE Computer Society, 2012, p. 12–19.

[2] T. Duckett, S. Pearson, S. Blackmore, B. Grieve, and M. Smith, "White paper - agricultural robotics: The future of robotic agriculture," 2018. [Online]. Available: https://uwe-repository.worktribe.com/output/866226

[3] E. Krotkov and J. Blitch, "The defense advanced research projects agency (darpa) tactical mobile robotics program," *The International Journal of Robotics Research*, vol. 18, no. 7, pp. 769–776, 1999. [Online]. Available: https://doi.org/10.1177/02783649922066457

[4] A. Thoesen and H. Marvi, "Planetary surface mobility and exploration: A review," *Current Robotics Reports*, vol. 2, no. 3, pp. 239–249, 2021.

[5] O. Kim, J. Seo, S. Ahn, and C. H. Kim, "Ufo: Uncertainty-aware lidar-image fusion for off-road semantic terrain map estimation," in *2024 IEEE Intelligent Vehicles Symposium (IV)*, 2024, pp. 192–199.

[6] H. Lin, H. Li, and Y. Gao, "See-touch-predict: Active exploration and online perception of terrain physics with legged robots," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3470–3477, 2025.

[7] X. Meng, N. Hatch, A. Lambert, A. Li, N. Wagener, M. Schmittle, J. Lee, W. Yuan, Z. Chen, S. Deng, G. Okopal, D. Fox, B. Boots, and A. Shaban, "Terrainnet: Visual modeling of complex terrain for high-speed, off-road navigation," 2023. [Online]. Available: https://arxiv.org/abs/2303.15771

[8] 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.

[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] 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.

[11] K. Doherty, T. Shan, J. Wang, and B. Englot, "Learning-aided 3-d occupancy mapping with bayesian generalized kernel inference," *Trans. Rob.*, vol. 35, no. 4, p. 953–966, Aug. 2019. [Online]. Available: https://doi.org/10.1109/TRO.2019.2912487

[12] J. Seo, T. Kim, S. Ahn, and K. Kwak, "Metaverse: Meta-learning traversability cost map for off-road navigation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 13 190–13 197.

[13] J. Wilson, Y. Fu, A. Zhang, J. Song, A. Capodieci, P. Jayakumar, K. Barton, and M. Ghaffari, "Convolutional bayesian kernel inference for 3d semantic mapping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8364–8370.

[14] M. Stolzle, T. Miki, L. Gerdes, M. Azkarate, and M. Hutter, "Reconstructing occluded elevation information in terrain maps with self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1697–1704, 2022, green Open Access added to TU Delft Institutional Repository 'You share, we take care!' - Taverne project https://www.openaccess.nl/en/you-share-we-take-care Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

[15] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 2273–2280.

[16] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "A simple baseline for bev perception without lidar," in *arXiv:2206.07959*, 2022.

[17] J. Fei, K. Peng, P. Heidenreich, F. Bieder, and C. Stiller, "Pillarsegnet: Pillar-based semantic grid map estimation using sparse lidar data," in *2021 IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 838–844.

[18] A. Shaban, X. Meng, J. Lee, B. Boots, and D. Fox, "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

[19] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[20] 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 *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 5000–5007.

[21] A. Howard, M. Turmon, L. Matthies, B. Tang, A. Angelova, and E. Mjolsness, "Towards learned traversability for robot navigation: From underfoot to the far field," *Journal of Field Robotics*, vol. 23, no. 11-12, pp. 1005–1017, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20168

[22] K. Konolige, M. Agrawal, M. R. Blas, R. C. Bolles, B. Gerkey, J. Solà, and A. Sundaresan, "Mapping, navigation, and learning for off-road traversal," *Journal of Field Robotics*, vol. 26, no. 1, pp. 88–113, 2009. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20271

[23] J.-F. Lalonde, N. Vandapel, D. F. Huber, and M. Hebert, "Natural terrain classification using three-dimensional ladar data for ground robot mobility," *Journal of Field Robotics*, vol. 23, no. 10, pp. 839–861, 2006. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.20134

[24] P. Krüsi, P. Furgale, M. Bosse, and R. Siegwart, "Driving on point clouds: Motion planning, trajectory optimization, and terrain assessment in generic nonplanar environments," *Journal of Field Robotics*, vol. 34, no. 5, pp. 940–984, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21700

[25] G. Kahn, P. Abbeel, and S. Levine, "Badgr: An autonomous self-supervised learning-based navigation system," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1312–1319, 2021.

[26] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine, "Rapid exploration for open-world navigation with latent goal models," 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. 674–684. [Online]. Available: https://proceedings.mlr.press/v164/shah22a.html

[27] D. D. Fan, A.-a. Agha-mohammadi, and E. A. Theodorou, "Learning risk-aware costmaps for traversability in challenging environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 279–286, 2022.

[28] X. Cai, M. Everett, J. Fink, and J. P. How, "Risk-aware off-road navigation via a learned speed distribution map," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 2931–2937.

[29] Z. Chen, C. Sun, S. Nie, C. Min, C. Ning, H. Li, and B. Wang, "3dttnet: Multimodal fusion-based 3d traversable terrain modeling for off-road environments," 2025. [Online]. Available: https://arxiv.org/abs/2412.08195

[30] D. Stavens and S. Thrun, "A self-supervised terrain roughness estimator for off-road autonomous driving," 2012. [Online]. Available: https://arxiv.org/abs/1206.6872

[31] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[32] J. Gu, M. Bellone, T. Pivoňka, and R. Sell, "Clft: Camera-lidar fusion transformer for semantic segmentation in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, p. 1–12, 2024. [Online]. Available: http://dx.doi.org/10.1109/TIV.2024.3454971

[33] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and R. Daniela, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5135–5142.