# TerraEDL: Uncertainty-Aware Multi-Modal Terrain Classification via Evidential Deep Learning

Jianghuan Xu[1], Hongze Li[1] and Huijing Zhao[1]

***Abstract*— Robust terrain classification is essential for mobile robots operating in complex environments. While multi-modal fusion improves recognition accuracy, severe domain shifts and asymmetric sensor degradation often lead to unsafe overconfident decisions. Visual Foundation Models (VFMs) provide strong cross-scene generalization but are computationally prohibitive for high-frequency control. In this work, we present TerraEDL, a conflict-aware multi-modal fusion framework designed for risk-triggered robotic perception. Instead of directly aggregating heterogeneous evidences, we introduce a Consensus-Driven Dual-Branch Fusion (CRED-Fusion) mechanism that decouples shared consensus information from modality-specific residual signals. This structured fusion reduces artificial conflict under heterogeneous sensing conditions and produces uncertainty signals suitable for reliability-aware decision making. Building upon this fusion framework, we propose a deterministic integration policy that invokes a VFM only when the onboard multi-modal prediction becomes unreliable. This design enables adaptive fallback to a more generalizable model without sacrificing real-time performance. Extensive experiments demonstrate the superior performance and efficiency of our approach. The results show that TerraEDL achieves average classification accuracies of 94.47% and 91.44% on the Jackal and VP-Terrain datasets, respectively, outperforming traditional Dempster-Shafer fusion baselines. Furthermore, by using uncertainty as a dynamic trigger, our framework activates the VFM in only 10.84% of frames and achieves an efficient inference time of 23.22 ms per frame while maintaining robust cross-scene generalization.**

## I. INTRODUCTION

Reliable terrain classification is a fundamental capability for autonomous mobile robots operating in unstructured environments [1]. Accurate perception of terrain properties directly affects motion planning, traction control, and long-term system safety [2] [3]. Misclassification may lead to suboptimal navigation decisions or even catastrophic hardware damage due to loss of stability [4]. While modern deep networks have significantly improved perception accuracy, ensuring reliability under heterogeneous sensing conditions and domain shifts remains a critical challenge [3].

To enhance robustness, multi-modal fusion has become common in terrain classification. Vision provides rich semantic and long-range contextual information, whereas proprioceptive signals capture local interaction dynamics such as vibration and slip. These modalities are inherently complementary [5] [6]. However, they encode fundamentally different physical properties: vision reflects spatial semantics, while proprioception captures embodied interaction responses. Under adverse conditions—such as low illumination, sensor degradation, or cross-scene domain shifts, they often exhibit asymmetric reliability [7].

Most existing multi-modal fusion frameworks implicitly treat heterogeneous modalities as evidence residing in a shared and compatible space. Feature-level fusion assumes semantic alignment in latent representations [8], while decision-level fusion typically aggregates modality outputs without explicitly accounting for structural differences. In evidential fusion methods [9] based on Dempster–Shafer (DS) theory, belief masses[10] are directly combined under the assumption of evidence-space consistency. While mathematically elegant, this assumption fails in heterogeneous robotic perception. When modalities encode distinct physical attributes, naive aggregation may artificially inflate the conflict term, leading to unstable belief sharpening and unreliable uncertainty estimates [11].

This issue is not merely a weighting problem—it is a structural evidence alignment problem. In heterogeneous sensing, not all evidence should be directly fused. Some components represent genuine cross-modal consensus, while others correspond to modality-specific residual cues that should not be forced into aggregation. Failure to distinguish these components can result in overconfident predictions precisely when the system should express uncertainty [11].

To address this fundamental limitation, we propose TerraEDL, an uncertainty-aware multi-modal terrain classification framework built upon a **Consensus-Driven Dual-Branch Fusion (CRED-Fusion)** mechanism. As shown in Fig. 1, instead of directly combining heterogeneous evidences, we introduce a "decouple-before-combine" paradigm. Specifically, modality evidence is decomposed into two streams: a **shared branch**, which captures cross-modal consensus information suitable for evidential aggregation, and a **private branch**, which preserves modality-specific residual evidence that may be complementary rather than contradictory. By restricting Dempster–Shafer fusion to the shared branch, we mitigate artificial cross-modal conflict and prevent over-sharpened belief distributions. The private branch, in turn, ensures that discriminative modality-specific cues are retained without influencing the fusion process. This structured design transforms evidential fusion from naive belief accumulation into consensus-aware reasoning.

Beyond structured fusion, robotic systems require a principled mechanism to decide when to rely on onboard predictors

This work was supported by the National Natural Science Foundation of China under Grants U22A2061.

[1]The authors are with the State Key Laboratory of General Artificial Intelligence, Peking University, Beijing 100871, China, and also with the School of Intelligence Science and Technology, Peking University, Beijing 100871, China.

Correspondence: Jianghuan Xu, xjhcgdk@stu.pku.edu.cn and H.Zhao, zhaohj@pku.edu.cn.

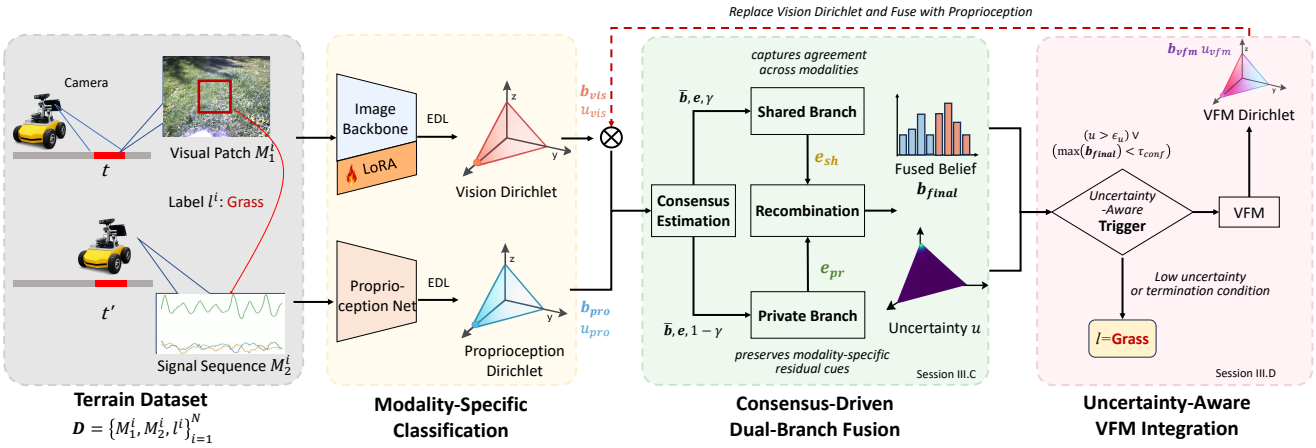



Fig. 1: Architecture overview of the proposed framework TerraEDL and Consensus-Driven Dual-Branch Fusion (CRED-Fusion).

and when to invoke more powerful models. Recent Visual Foundation Models (VFMs) offer strong cross-scene generalization but remain computationally expensive for high-frequency control loops. Rather than proposing a new uncertainty modeling approach, we treat the evidential belief and its associated reliability score as a decision interface. Building upon the conflict-aware fusion design, we introduce a deterministic integration policy that activates a VFM only when the fused prediction exhibits low confidence. This selective fallback strategy improves cross-scene robustness while preserving real-time performance.

Through extensive experiments across multiple terrain datasets and cross-scene benchmarks, we demonstrate that TerraEDL ensures high classification accuracy and robust cross-scene generalization even when the robot encounters new, unseen terrains or severe sensor degradation.

The main contributions of this work are as follows:

- We identify a structural incompatibility between conventional evidential fusion and heterogeneous robotic sensing, and propose a consensus-driven dual-branch fusion mechanism that explicitly decouples shared consensus from modality-specific residual evidence prior to aggregation.
- We introduce a deterministic uncertainty-aware integration policy that activates a highly capable Visual Foundation Model (VFM) only when the onboard multi-modal prediction becomes unreliable. This enables adaptive fallback to a more generalizable model while preserving high-frequency real-time efficiency.
- We validate the framework on public terrain benchmarks and real robotic platforms, demonstrating improved robustness, cross-scene generalization, and real-time feasibility.

The remainder of this paper is as follows: Section II reviews related work. Section III presents the TerraEDL methodology. Section IV discusses experiments and results. Section V concludes the paper and outlines future directions.

## II. RELATED WORKS

### A. Terrain Classification

Early terrain classification typically relied on a single sensing modality, such as vision [12] or proprioception [13]. Visual sensors, offering dense semantic information and ease of deployment, were used in early non-learning methods with handcrafted features like color, texture [14], or SURF descriptors [15]. Deep learning has improved visual TC using various lightweight backbones like MobileNetV2 and ResNet [6] [16]. However, purely vision-based methods are inherently constrained by their sensitivity to illumination changes and visual ambiguity (e.g., distinguishing visually similar but physically different terrains like dry and wet grass). This inherent vulnerability severely limits their robustness in real-world scenarios [17]. Conversely, proprioceptive modalities—capturing inertial measurements (IMUs) and vibration [13], or audio data [18]—remain robust to visual disturbances but provide spatially sparse and short-sighted information.

To exploit the complementary nature of these heterogeneous sensory streams, various multi-modal fusion strategies have been developed. At the feature level, Hanson et al. [16] concatenated neural feature vectors prior to multi-layer perceptron (MLP) classification, whereas Tian et al. [19] explicitly aligned visual and IMU features using a Mean Squared Error (MSE) loss. At the decision level, Li et al. [5] fused multi-modal outputs based on Bayesian principles. Despite closed-set gains, most existing fusion frameworks focus heavily on architectural complexity rather than predictive reliability [9], and typically lack explicit uncertainty modeling. Consequently, they struggle to identify when modalities conflict, fundamentally bottlenecking their cross-scene generalization capabilities.

To address this generalization bottleneck, recent studies have begun exploring Visual Foundation Models (VFMs) or Visual Language Models (VLMs) for open-world semantic understanding, such as CLIP [20]-based vision-language fusion methods [21]. While VFMs or VLMs exhibit exceptional zero-shot generalization and robustness against distri-

bution shifts, their massive parameter counts and inference latency render them computationally prohibitive for the high-frequency control loops required by mobile robots. Therefore, a critical open challenge remains: how to dynamically leverage the powerful generalization of VFMs or VLMs without sacrificing the real-time efficiency of lightweight, onboard multi-modal fusion. This highlights the necessity of equipping onboard fusion modules with robust uncertainty quantification capabilities, allowing uncertainty to act as a dynamic trigger for VFM invocation.

### B. Reliability-Aware Multi-Modal Fusion

Model uncertainty is well studied. Bayesian Neural Networks (BNNs) [22], Monte Carlo (MC) Dropout [23], and Deep Ensembles [24] provide principled probabilistic formulations and capture predictive variability. However, these approaches typically rely on sampling or multiple model instances, which may increase robotic runtime cost.

Evidential Deep Learning (EDL) [10] offers a deterministic alternative by modeling class probabilities with a Dirichlet distribution, enabling uncertainty estimation in a single forward pass. Based on this formulation, Dempster-Shafer (DS) theory has been adopted in multi-view and multi-modal settings to aggregate belief masses at the decision level [9]. These evidential fusion strategies provide a structured mechanism for incorporating reliability cues during aggregation.

Despite their advantages, most existing evidential fusion frameworks assume that modality outputs reside in a compatible evidence space and can be directly combined. In heterogeneous robotic perception, however, sensing modalities often capture fundamentally different physical properties. Under asymmetric degradation or domain shifts, naive aggregation may lead to unstable fusion behavior. This observation motivates structured aggregation strategies that explicitly account for heterogeneous sensing characteristics while maintaining decision-level reliability cues.

### C. Cross-scene Generalization

Cross-scene generalization remains a formidable challenge in terrain classification due to inevitable domain shifts across diverse environments. To tackle this, recent works have explored various robust representation learning and domain adaptation techniques. For instance, some approaches employ self-supervised contrastive learning [5] [25] to extract domain-invariant features from heterogeneous sensor streams. Other methods focus on data augmentation [6] to improve model robustness. To evaluate model robustness against such shifts, recent benchmarks like TerraBench and TerraData [21], which include various off-road datasets [6] [26] [5], have been introduced, establishing standard cross-scene evaluation protocols. Despite these efforts, most existing frameworks rely predominantly on conventional small-scale models and critically lack the explicit estimation and utilization of predictive uncertainty. Consequently, a framework capable of adaptively quantifying its uncertainty to dynamically fall back on highly generalizable Visual Foundation Models (VFMs) remains unexplored.

## III. METHODOLOGY

### A. Problem Formulation

Below we present the formulation of TerraEDL. Given a labeled multi-modal dataset $\mathcal{D} = \{M_1^i, M_2^i, \ldots, M_n^i, l^i\}_{i=1}^N$, where $M_j^i$ denotes the $j^{th}$ modality sensor data and $l^i$ represents the corresponding terrain category label, our objective is to formulate a robust uncertainty-aware fusion mapping $\mathcal{F}$. Instead of merely training isolated predictive models for each modality, we aim to leverage pre-extracted modality evidence (parameterized as Dirichlet distributions) to achieve better predictions and cross-scene generalization. Specifically, the fusion framework $\mathcal{F}$ computes a consensus belief $\bar{b}$ to decouple and recombine shared evidence $e_{sh}$ and private evidence $e_{pr}$. Ultimately, the system yields a joint terrain prediction $\hat{l}$ and an uncertainty $u$, formulated as: $(\hat{l}, u) = \mathcal{F}(M_1, \ldots, M_n)$. This explicit uncertainty $u$ then serves as a logical trigger for a Visual Foundation Model (VFM) to ensure robustness in new or unseen environments. In this paper, we focus on the two most commonly used modalities: vision and proprioception ($n = 2$).

### B. Preliminaries

*1) Evidential Deep Learning (EDL):* Unlike standard neural networks that use a softmax layer to output point estimates of class probabilities, Evidential Deep Learning (EDL) explicitly models the predictive distribution by placing a Dirichlet prior over the class probabilities. For a $K$-class classification problem, let the neural network directly output a non-negative evidence vector $\boldsymbol{e} = [e_1, e_2, \ldots, e_K]^T \geq 0$. These evidence values are typically obtained by applying a softplus activation function to the network's raw logits. The evidence characterizes the parameters of a Dirichlet distribution, denoted as $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_K]^T$, where $\alpha_k = e_k + 1$. The total Dirichlet strength, which represents the overall accumulated evidence, is defined as $S = \sum_{k=1}^K \alpha_k$. Grounded in Subjective Logic, this formulation naturally unifies the predicted class probability $p_k$ and the overall uncertainty $u$, which are computed as follows:

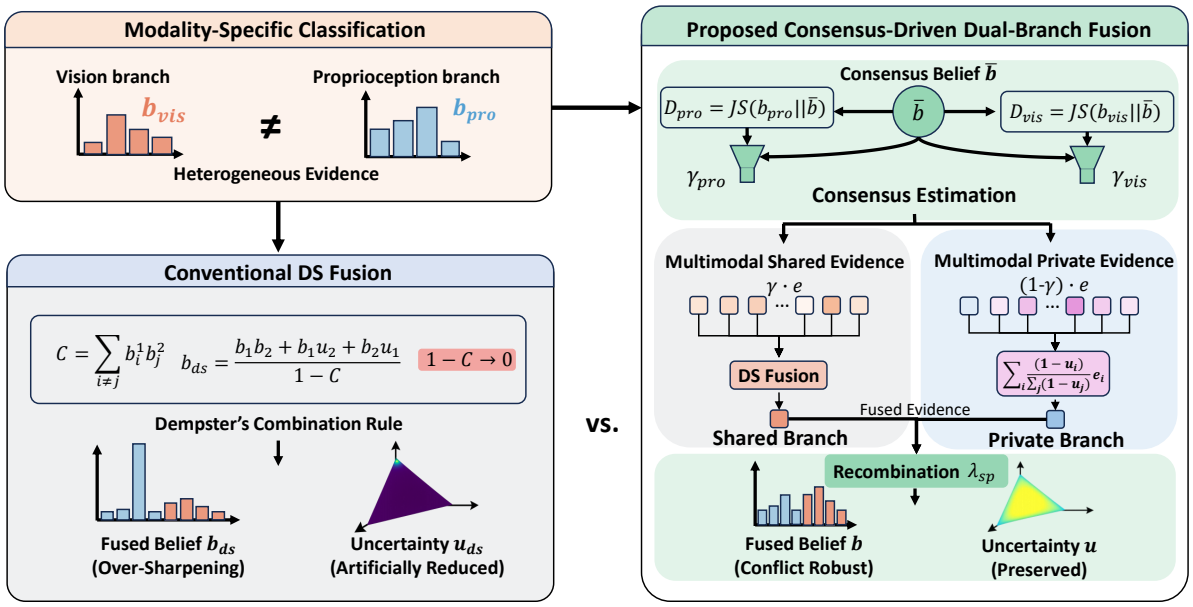
$$p_k = \frac{\alpha_k}{S}, \quad u = \frac{K}{S} \tag{1}$$

Here, the uncertainty $u$ is inversely proportional to the total evidence $S$.

*2) Dempster-Shafer (DS) Fusion:* Below we describe the Dempster-Shafer (DS) fusion framework. Given two views, denoted as 1 and 2, we can derive their respective belief masses $b_k^1 = e_k^1/S^1$ and $b_k^2 = e_k^2/S^2$, along with their overall uncertainties $u^1$ and $u^2$. The traditional DS combination rule fuses these two independent predictions to form a joint belief mass for each class $k$, formulated as:

$$b_k^{1\oplus 2} = \frac{1}{1-C}\left(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1\right) \tag{2}$$

Here, $C$ represents the degree of conflict between the two sources, calculated as $C = \sum_{i \neq j} b_i^1 b_j^2$. This conflict term mathematically measures the extent of contradictory evidence between the predictions.




*Direct aggregation under leads to conflict amplification and overconfident predictions.*

*Structured decoupling mitigates artificial conflict and preserves reliable uncertainty.*

Fig. 2: Comparison between conventional DS Fusion and proposed Consensus-Driven Dual-Branch Fusion (CRED-Fusion) under heterogeneous evidence.

### C. Consensus-Driven Dual-Branch Fusion

While the DS combination rule provides an elegant mathematical foundation for homogeneous multi-source tasks, its direct application to heterogeneous multi-modal perception—specifically vision and proprioception—reveals a critical limitation. Vision captures remote spatial semantics, such as texture and color, whereas proprioception records local physical interactions, such as vibration and vehicle kinematics. Because these modalities operate in fundamentally disjoint feature spaces, they frequently provide complementary rather than identical cues for the same terrain. However, the standard DS operator inherently assumes evidence space consistency. Consequently, it mathematically misinterprets these structurally disjoint, modality-specific features as severe contradictions, artificially inflating the conflict term $C$. This misalignment suppresses the accumulation of valid beliefs and produces unreliable uncertainty estimates, rendering the system incapable of safely triggering the VFM. To address this fundamental bottleneck, we propose a **Consensus-Driven Dual Branch Fusion** architecture, as illustrated in Fig. 2. As explicitly compared to traditional direct fusion, our approach avoids forcing immediate aggregation; instead, it adaptively decouples the shared evidence from modality-specific private evidence before the final evidential combination.

*1) Consensus Estimation:* We follow the training method in Evidential Deep Learning [10] to get Dirichlet of each modality, from which the belief mass $\boldsymbol{b}^i$ and uncertainty $u^i$ of each modality can be derived. As shown in Fig. 2, we then compute the consensus belief $\bar{\boldsymbol{b}}$:

$$\bar{\boldsymbol{b}} = \sum_{i=1}^{n} \lambda^i \boldsymbol{b}^i \tag{3}$$

where $\lambda^i$ is the weight of the $i^{th}$ modality. In our work, $\lambda^i$ can be a hyperparameter or computed in an uncertainty-aware manner:

$$\lambda^i = \frac{1-u^i}{\sum_i 1-u^i} \tag{4}$$

*2) Dual-Branch Evidence Fusion:* To fuse the shared information from each modality and preserve modality-specific information, we design a dual-branch fusion strategy. First, we compute consistency weight of each modality:

$$\gamma^i = \mathrm{e}^{-\tau \cdot D_i} \tag{5}$$

where $D_i = JS(\boldsymbol{b}^i || \bar{\boldsymbol{b}})$ describes the Jensen-Shannon Divergence between belief mass of each modality $\boldsymbol{b}^i$ and the consensus belief $\bar{\boldsymbol{b}}$. The closer $\boldsymbol{b}^i$ is to consensus $\bar{\boldsymbol{b}}$, the closer $\gamma^i$ is to 1.

We then divide the evidence from each modality into shared and private branch. In shared branch, we compute the shared evidence $\boldsymbol{e}^i_{sh} = \gamma^i \boldsymbol{e}^i$ and the corresponding Dirichlet:

$$\boldsymbol{\alpha}^i_{sh} = \boldsymbol{e}^i_{sh} + \mathbf{1}, \quad S^i_{sh} = \sum_k \alpha^{ik}_{sh}, \quad \boldsymbol{b}^i_{sh} = \frac{\boldsymbol{e}^i_{sh}}{S^i_{sh}} \tag{6}$$

where DS fusion is applied to belief masses derived from the reconstructed Dirichlet parameters after evidence splitting, yielding fused shared belief $\boldsymbol{b}_{sh}$ and uncertainty $u_{sh}$. To recombine with private evidence in the same space, we reconstruct the DS-fused shared evidence via:

$$S_{sh} = \frac{K}{u_{sh}}, \quad \boldsymbol{e}_{sh} = S_{sh} \cdot \boldsymbol{b}_{sh}, \quad \boldsymbol{\alpha}_{sh} = \boldsymbol{e}_{sh} + \mathbf{1} \tag{7}$$

In private branch, we compute the private evidence $\boldsymbol{e}^i_{pr} = (1-\gamma^i) \cdot \boldsymbol{e}^i$ and sum them through reliability weighted fusion:

$$\boldsymbol{e}_{pr} = \frac{\sum_i \lambda^i \cdot \boldsymbol{e}^i_{pr}}{\sum_i \lambda^i} \tag{8}$$

where $\lambda^i$ is the reliability weight of each modality, which can be a hyperparameter or computed like the weight in consensus estimation.

Finally, results of each branch are recombined:

$$\boldsymbol{e}_{final} = \boldsymbol{e}_{sh} + \lambda_{sp}\boldsymbol{e}_{pr} \tag{9}$$

where $\lambda_{sp}$ is the coefficient controlling the strength, and we get the fused belief $\boldsymbol{b}_{final}$, uncertainty $u$ and the prediction.

### D. Uncertainty-Aware VFM Integration

The final output of our CRED-Fusion module provides both the fused belief masses $\boldsymbol{b}_{final}$ and the overall uncertainty $u_{final}$. To balance real-time operational efficiency with robustness and generalization, we introduce a deterministic integration mechanism that selectively invokes the Visual Foundation Model (VFM) only when the onboard multi-modal perception is deemed unreliable. Specifically, the VFM is queried if the predictive uncertainty exceeds a predefined safety margin, or if the maximum belief of the predicted terrain class falls below a strictly defined threshold. This boolean trigger policy is formally expressed as:

$$\text{Trigger} = (u_{final} > \epsilon_u) \vee (\max(\boldsymbol{b}_{final}) < \tau_{conf}) \tag{10}$$

where $\epsilon_u$ and $\tau_{conf}$ are hyperparameters representing the maximum acceptable uncertainty and the minimum required confidence, respectively.

When the trigger condition is satisfied, the VFM produces classification logits $\boldsymbol{z}^{vfm}$. To ensure consistency with the evidential fusion framework, we convert these logits into non-negative evidence using a softplus activation:

$$\boldsymbol{e}^{vfm} = \text{softplus}(\boldsymbol{z}^{vfm}) \tag{11}$$

Following the same evidential formulation described in Eq. 1, we derive the VFM's Dirichlet parameters $\boldsymbol{\alpha}^{vfm}$, belief masses $\boldsymbol{b}^{vfm}$, and uncertainty $u^{vfm}$ from $\boldsymbol{e}^{vfm}$.

This conversion guarantees that the VFM prediction is represented in the same evidential space as the onboard modalities, preserving compatibility with the CRED-Fusion framework. When triggered, the original vision branch output is replaced by the VFM evidential output, while the proprioceptive branch remains unchanged. The fusion module is then re-executed using the VFM-proprioception pair.

## IV. EXPERIMENTAL EVALUATION

### A. Dataset

To validate the performance of our method, we conduct extensive experiments on public datasets and benchmarks. For the performance of multi-modal terrain classification, we choose Jackal [6] and VP-Terrain [5]. For cross-scene generalization, we choose datasets in TerraData [21], including Jackal (JA), RSCD (RS) and TerraPOSS (PO), and a comprehensive benchmark TerraBench [21].

### B. Experimental Design

*1) Baseline:* We select five vision-proprioception fusion approaches [6] [16] [19] [5]. These approaches include both decision fusion and feature fusion. Moreover, we include original DS Fusion [9] as ablation study.

*2) Implementation:* TerraEDL adopts MambaVision [27] as the vision backbone, where only LoRA parameters (rank $r = 4$) in attention and MLP layers are updated while the backbone remains frozen. The proprioception branch consists of two 1-D CNN blocks (kernel size 5, channel size 128) followed by 8 ResidualMambaBlocks (state size 32). DINOv3 [28] is used as the VFM backbone, where only the classifier head is trained. All models are optimized using the EDL loss [10] with AdamW optimizer (learning rate 0.005) for 50 epochs. The batch size is 64 for vision and 256 for proprioception. Trigger thresholds ($\epsilon_u = 0.5$, $\tau_{conf} = 0.6$) are selected on validation scenes and fixed across experiments and the VFM only triggered once for one low reliability sample as termination condition. All results are averaged over 5 runs with different random seeds. Moreover, we deploy our method on an Autolabor Pro 1 robot, whose results are included in the attached video.

### C. Analysis of Experimental Results

*1) Results of Fusion-Based Classification:* Table I compares TerraEDL with baselines using both vision and proprioception modalities across four environmental conditions in two public datasets. TerraEDL consistently achieves the highest average accuracy, showing its robustness under different conditions. Moreover, even without VFM, our proposed fusion framework surpasses the original DS Fusion, especially in dark environment of VP-Terrain. This can be attributed to the extremely dark illumination in VP-Terrain, which makes the vision modality completely fail. In such condition, DS Fusion becomes unreliable as shown in Fig. 2.

*2) Cross-Scene Generalization:* We follow the benchmark described in TerraBench [21], which includes **Zero-shot Cross-Scene Generalization** and **Few-shot Novel Class Adaption**. Table II compares TerraEDL with baseline methods in zero-shot cross-scene generalization, while Table III compares the performance of TerraEDL with baseline methods in few-shot novel class adaptation. As Table II shows, our proposed method outperforms all baselines in most settings of zero-shot cross-scene generalization, highlighting the strong generalization ability of our method. In few-shot novel class adaptation, our method still shows good performance. However, due to the limited data (1%), the uncertainties of most samples are high and VFM is triggered unduly, which cannot be fully trained with such amount of data, leading to worse results in some settings.

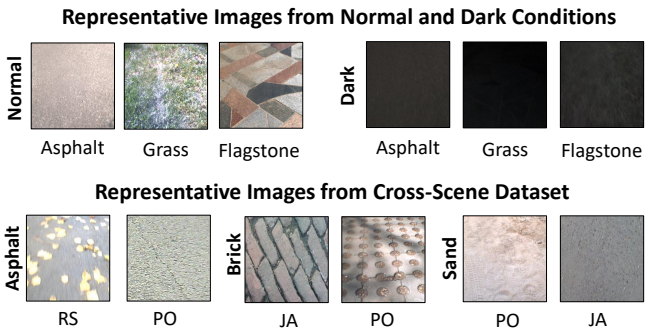


Fig. 3: Representative images of distribution shifts.

TABLE I: Accuracy (in %) of the proposed method and baseline on vision-proprioception input.

| Method | Jackal | | | | | VP-Terrain | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Normal | Dark | Fog | Sun | Avg | Normal | Dark | Fog | Sun | Avg |
| Chen et al. [6] | 99.07 | **92.70** | 93.76 | 91.41 | 94.24 | 99.71 | 82.46 | 74.18 | 93.19 | 86.40 |
| Li et al. [5] | 98.65 | 89.19 | 91.69 | 92.91 | 93.11 | 98.17 | **85.35** | 76.24 | 89.96 | 87.01 |
| Hanson et al. [16] | **99.97** | 35.51 | 65.41 | 57.05 | 64.49 | 97.65 | 56.95 | 63.28 | 82.45 | 71.46 |
| Guan et al. [19] | 99.94 | 29.08 | 58.60 | 55.59 | 60.80 | **99.95** | 26.46 | 34.57 | 79.36 | 53.37 |
| TerraEDL (DS Fusion) | 99.90 | 82.59 | 94.93 | **94.38** | 92.95 | 99.62 | 63.71 | 89.42 | 90.08 | 85.71 |
| TerraEDL (w/o VFM) | 99.91 | 83.74 | **95.17** | 94.30 | 93.28 | 99.44 | 78.15 | 92.29 | 90.83 | 90.18 |
| TerraEDL | 99.93 | 88.98 | 95.00 | 93.96 | **94.47** | 99.58 | 80.02 | **92.38** | **93.79** | **91.44** |

TABLE II: Accuracy (in %) of TerraEDL and baselines in **Zero-shot Cross-Scene Generalization**.

| Method | PO→JA | RS→JA | JA→PO | RS→PO | JA→RS | PO→RS |
|---|---|---|---|---|---|---|
| Chen et al. [6] | 48.45 | 71.26 | 35.04 | 58.41 | 9.97 | 27.35 |
| Li et al. [21] | 84.99 | **99.98** | 70.93 | 73.08 | **71.34** | 50.21 |
| Hanson et al. [16] | 39.65 | 82.38 | 37.00 | 52.78 | 19.94 | 20.35 |
| Zhao et al. [26] | 62.01 | 85.54 | 52.62 | 64.01 | 37.54 | 36.12 |
| TerraEDL (w/o VFM) | 59.81 | 99.40 | 62.02 | 71.72 | 33.96 | 26.79 |
| TerraEDL | **97.11** | 99.66 | **79.76** | **76.71** | 50.00 | **66.26** |

TABLE III: Accuracy (in %) of TerraEDL and baselines on **Few-shot Novel Class Adaptation**.

| Method | PO→JA | RS→JA | JA→PO | RS→PO | JA→RS | PO→RS |
|---|---|---|---|---|---|---|
| Chen et al. [6] | 87.49 | 90.20 | **87.62** | **89.00** | 55.84 | 52.65 |
| Li et al. [21] | **96.42** | **97.24** | 81.55 | 87.36 | 58.15 | 54.11 |
| Hanson et al. [16] | 90.57 | 75.90 | 76.55 | 77.90 | 50.18 | 53.75 |
| Zhao et al. [26] | 83.66 | 65.72 | 59.20 | 62.05 | 35.99 | 37.51 |
| TerraEDL (w/o VFM) | 72.69 | 87.94 | 71.35 | 70.84 | 41.28 | 37.30 |
| TerraEDL | 87.74 | 94.05 | 82.68 | 80.02 | **64.33** | **57.99** |

### *D. Visualization Analysis*

To reveal the reliability when the models are faced with distribution shifts, we visualize the uncertainty density distributions under two distinct challenges-**Modality Degradation (Normal→Dark)** and **Severe Domain Shift (Cross-Scene)**, as shown in Fig. 3.

*1) Modality Degradation (Normal→Dark):* Fig. 4 presents both the uncertainty density distributions and the corresponding Receiver Operating Characteristic (ROC) curves under severe illumination degradation (VP-Terrain Dark). In this scenario, the visual modality becomes highly noisy while proprioception remains reliable. When traditional Dempster-Shafer (DS) fusion is applied, the structural misalignment between the noisy visual features and the stable proprioceptive cues artificially inflates the conflict term. Consequently, traditional DS fusion suffers from belief over-sharpening, inadvertently pushing a significant portion of incorrect predictions into the low-uncertainty region (visible as the overlapping tail in the baseline density plot).

To quantitatively evaluate this fatal overconfidence, we treat the uncertainty metric as a binary classifier for identifying erroneous predictions (i.e., triggering the VFM). As shown in the ROC curves, our CRED-Fusion architecture achieves a significantly higher Area Under the Curve

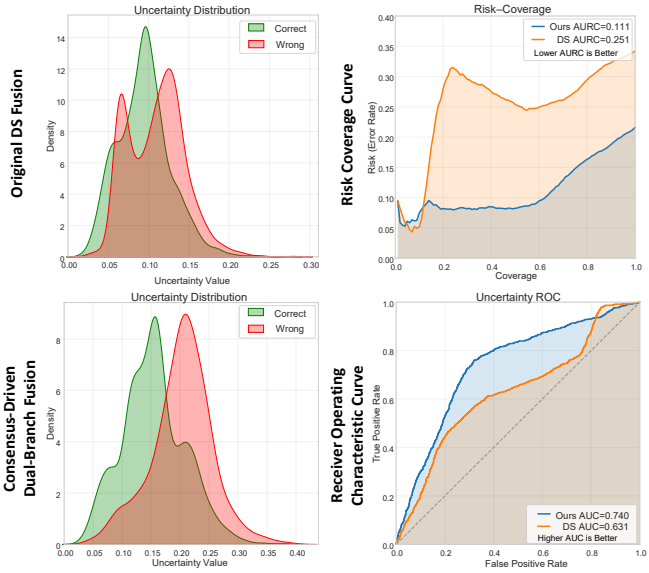



Fig. 4: Performance analysis under severe modality degradation (VP-Terrain Dark).

(AUROC) compared to the traditional DS baseline. This quantitative superiority proves that by explicitly decoupling the shared evidence from modality-specific private evidence, our method effectively minimizes hazardous high-confidence errors (false negatives in error detection), providing a strictly conservative and highly reliable trigger policy to prevent silent failures.

*2) Severe Domain Shift (Cross-Scene):* To further demonstrate the effectiveness of our uncertainty metric as a universal VFM trigger, Fig. 5 visualizes the density distribution on the challenging cross-scene subset. Under extreme shifts where local models inherently struggle, an ideal system must explicitly recognize its unfamiliarity. As shown in our plot, the distribution of correct predictions (green) peaks at a low uncertainty range, whereas the distribution for incorrect predictions (red) cleanly shifts toward the higher uncertainty spectrum. This distinct separability provides compelling empirical evidence that our estimated uncertainty dynamically and accurately reflects cross-scene generalization difficulty. Even in the absence of stable multi-modal priors, this reliable metric ensures that the computationally expensive VFM is safely and precisely invoked only when the lightweight onboard network fails.

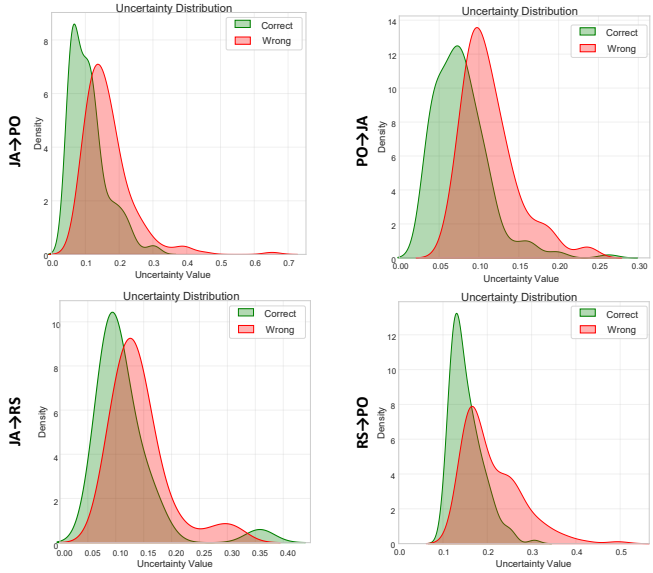



Fig. 5: Uncertainty distribution density during cross-scene generalization.

### E. System Efficiency and Real-Time Performance

A critical bottleneck in deploying Visual Foundation Models (VFMs) on mobile robots is their prohibitive inference latency. To demonstrate the practical feasibility of our proposed framework for real-time control loops, we evaluate the system's efficiency by analyzing the VFM trigger rate and the overall inference time. We measure the runtime performance on an NVIDIA Jetson Orin NX Developer Kit (8GB), accelerated by NVIDIA Tensor RT.

TABLE IV: VFM trigger rate and the overall inference time in different environmental settings.

| | Normal | Dark | Fog | Sun | Avg |
|---|---|---|---|---|---|
| VFM Trigger Rate (%) | 2.61 | 21.37 | 9.38 | 10.00 | 10.84 |
| Inference Time (ms) | 12.72 | 36.66 | 21.36 | 22.15 | 23.22 |

As presented in Table IV, relying exclusively on the VFM for semantic understanding yields high robustness but incurs an inference latency of approximately 137 ms per frame. Conversely, the standalone lightweight dual-branch network operates at a efficient 9.39 ms per frame but suffers severe performance degradation under unseen conditions.

Our uncertainty-aware VFM integration mechanism elegantly bridges this gap. Across the entire testing dataset, the VFM is invoked in only 10.84% of the total frames—specifically during highly ambiguous or structurally misaligned scenarios. By heavily amortizing the computational cost of the VFM, our adaptive pipeline achieves an average inference time of merely 23.22 ms per frame. This empirical result confirms that our framework successfully extracts the zero-shot generalization capabilities of the VFM without sacrificing the high-frequency real-time performance strictly required by mobile robotic platforms.

### F. Ablation Study

To validate the internal mechanics of our CRED-Fusion architecture, we conduct an ablation study focusing on the decoupled evidence streams. Since the superiority of our framework over traditional direct fusion was established in previous evaluations, we now isolate our network's internal branches. We compare our complete model against two structural variants: a "Shared-Only" baseline (deriving predictions purely from shared evidence) and a "Private-Only" baseline (utilizing only modality-specific private evidence). Moreover, single vision and proprioception modalities are also included.

TABLE V: Ablation results of shared-only and private-only branches, single-modality baselines and VFM trigger rate.

| method | Normal | Dark | Fog | Sun | Avg | Trigger Rate (%) |
|---|---|---|---|---|---|---|
| Vision Only | 99.11 | 30.60 | 78.60 | 87.58 | 73.97 | - |
| Proprioception Only | 88.76 | 79.41 | 88.76 | 88.76 | 86.42 | - |
| Shared Only | 99.47 | 79.69 | 89.37 | 90.03 | 89.64 | 14.25 |
| Private Only | 99.25 | 79.32 | 92.80 | 94.26 | 91.41 | 97.60 |

As reported in Table V, single modality variants show suboptimal performance. The Shared-Only configuration abandons modality-specific information, leading to a noticeable drop in accuracy. Conversely, the Private-Only configuration exhibits higher accuracy. This is attributed to its extremely high VFM Trigger rate, revealing its suffering from lacking the stabilizing foundation of multi-modal agreement, thus always triggering VFM due to high uncertainty.

### G. Summary of Experimental Results

The main results and findings are as follows:

(1) Under modality degradation, the proposed CRED-Fusion demonstrates improved robustness compared to conventional DS fusion and existing multi-modal baselines. As shown in Table I, TerraEDL achieves state-of-the-art average accuracies of 94.47% and 91.44% on the Jackal and VP-Terrain benchmarks, respectively. In contrast to DS Fusion, which amplifies artificial cross-modal conflict, the decoupling mechanism mitigates belief over-sharpening and maintains stable under asymmetric degradation.

(2) The fused belief provides a reliability signal that more effectively separates correct and incorrect predictions. As illustrated in Fig. 4, TerraEDL achieves an AUROC of 0.740 and an AURC of 0.111, outperforming conventional DS fusion in risk–coverage trade-off. This improvement indicates that the structured fusion design produces reliability scores better aligned with actual prediction correctness, which is essential for safe and deterministic downstream gating rather than for standalone uncertainty modeling.

(3) The reliability-aware VFM integration mechanism enables effective cross-scene generalization while preserving real-time efficiency. In zero-shot cross-scene experiments (Table II), TerraEDL significantly improves generalization performance when combined with selective VFM fallback. Importantly, the average trigger rate remains low at 10.84% (Table IV), leading to an average inference time of 23.22 ms per frame, which satisfies high-frequency robotic control requirements. These results demonstrate that structured evidential fusion, together with selective foundation model activation, offers a practical and computationally efficient solution for robust terrain perception.

Overall, the experimental findings substantiate the three main contributions of this work: (i) conflict-aware structured fusion for heterogeneous sensing, (ii) reliability-driven foundation model integration without sacrificing real-time performance, and (iii) improved cross-scene robustness under modality degradation.

## V. CONCLUSION AND FUTURE WORKS

We presented TerraEDL, an uncertainty-aware multimodal terrain classification framework for heterogeneous robotic sensing. By identifying structural limitations of direct evidential aggregation under heterogeneous modalities, we proposed CRED-Fusion that decouples shared consensus from modality-specific residual evidence prior to combination. This structured design improves robustness under asymmetric degradation and yields structured reliability for downstream control and model selection. Leveraging this reliability signal, we further introduced an uncertainty-aware VFM integration mechanism to enhance cross-scene generalization while maintaining real-time performance.

While the proposed framework demonstrates improved robustness and reliability, several limitations remain: (1) The current decoupling mechanism relies on divergence-based consensus estimation and hyperparameters. Although effective in our experiments, this formulation remains heuristic and may not fully capture complex inter-modal relationships. (2) As observed in few-shot novel class adaptation experiments, the system may exhibit excessive VFM triggering when modality-specific predictors are insufficiently trained. This indicates that uncertainty calibration under data-scarce regimes remains an open challenge. (3) The current framework operates at the decision level, assuming that evidential outputs from each modality are available in a shared label space, which limits the ability to address deeper feature-level misalignment. Future work will focus on addressing these limitations. In particular, we aim to explore learnable or theoretically grounded consensus estimation strategies, potentially through meta-learning or adaptive conflict modeling. We will also investigate uncertainty regularization and calibration-aware training methods to mitigate over-triggering and improve stability in data-scarce scenarios. Furthermore, extending the decoupling paradigm to representation-level alignment, or jointly modeling cross-modal correlations prior to evidential projection, may further enhance robustness under severe domain shifts.

## REFERENCES


[1] P. Papadakis, "Terrain traversability analysis methods for unmanned ground vehicles: A survey," *Engineering Applications of Artificial Intelligence*, vol. 26, pp. 1373–1385, 2013.

[2] D. LaRocque et al., "Proprioception is all you need: Terrain classification for boreal forests," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 11 686–11 693.

[3] H.-Y. Chen et al., "Terrain classification method using an nir or rgb camera with a cnn-based fusion of vision and a reduced-order proprioception model," *Computers and Electronics in Agriculture*, vol. 227, p. 109 539, 2024.

[4] L. Wellhausen et al., "Where should i walk? predicting terrain properties from images via self-supervised learning," *IEEE Robotics and Automation Letters*, vol. 4, pp. 1509–1516, 2019.

[5] H. Li et al., "Terrafusion: Semi-supervised vision-proprioception fusion for robust terrain classification," *IEEE Robotics and Automation Letters*, vol. 10, pp. 11 610–11 617, 2025.

[6] Y. Chen et al., "A cnn based vision-proprioception fusion method for robust ugv terrain classification," *IEEE Robotics and Automation Letters*, vol. 6, pp. 7965–7972, 2021.

[7] M. Bijelic et al., "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 679–11 689.

[8] D. Ramachandram et al., "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, pp. 96–108, 2017.

[9] Z. Han et al., "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 2551–2566, 2023.

[10] M. Sensoy et al., "Evidential deep learning to quantify classification uncertainty," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[11] L. A. Zadeh, "A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination," *AI Magazine*, vol. 7, no. 2, p. 85, Jun. 1986.

[12] R. Manduchi et al., "Obstacle detection and terrain classification for autonomous off-road navigation," *Autonomous robots*, vol. 18, pp. 81–102, 2005.

[13] C. Bai et al., "Three-dimensional vibration-based terrain classification for mobile robots," *IEEE Access*, pp. 63 485–63 492, 2019.

[14] A. Angelova et al., "Fast terrain classification using variable-length representation for autonomous navigation," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2007, pp. 1–8.

[15] P. Filitchkin et al., "Feature-based terrain classification for littledog," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2012, pp. 1387–1392.

[16] N. Hanson et al., "Vast: Visual and spectral terrain classification in unstructured multi-class environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2022, pp. 3956–3963.

[17] X.-Y. Zhang et al., "Towards robust pattern recognition: A review," *Proceedings of the IEEE*, vol. 108, no. 6, pp. 894–922, 2020.

[18] A. Valada et al., "Deep spatiotemporal models for robust proprioceptive terrain classification," *The International Journal of Robotics Research*, vol. 36, pp. 1521–1539, 2017.

[19] T. Guan et al., "Vinet: Visual and inertial-based terrain classification and adaptive navigation over unknown terrain," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 4106–4112.

[20] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[21] H. Li et al., "Terrax: Visual terrain classification enhanced by vision-language models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 15 383–15 390.

[22] L. V. Jospin et al., "Hands-on bayesian neural networks—a tutorial for deep learning users," *IEEE Computational Intelligence Magazine*, vol. 17, pp. 29–48, 2020.

[23] Y. Gal et al., "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48, PMLR, 20–22 Jun 2016, pp. 1050–1059.

[24] B. Lakshminarayanan et al., "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[25] J. Zürn et al., "Self-supervised visual terrain classification from unsupervised acoustic feature learning," *IEEE Transactions on Robotics*, vol. 37, pp. 466–481, 2021.

[26] T. Zhao et al., "A comprehensive implementation of road surface classification for vehicle driving assistance: Dataset, models, and deployment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8361–8370, 2023.

[27] A. Hatamizadeh et al., "Mambavision: A hybrid mamba-transformer vision backbone," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 261–25 270.

[28] O. Siméoni et al., *DINOv3*, 2025. arXiv: 2508.10104 [cs.CV].