Seeing Beyond Temperature: Multimodal Self-Supervised Learning for Thermal-Only Off-Road Mobility in No-Light Conditions

Harsh Rangwala[1], Madhan Balaji Rao[1], Anuj Pokhrel[1], Andre Harrison[2], Maggie Wigness[2], and Xuesu Xiao[1]

***Abstract*— Thermal cameras offer robust perception in extreme low-light conditions where most RGB cameras fail. However, their unimodal temperature sensing is insufficient for robot mobility tasks, especially in off-road environments, where both geometric and semantic information (e.g., obstacles vs. free space, traversable vs. non-traversable areas, and rough vs. smooth terrain) are crucial for navigation. This paper utilizes multimodal Self-Supervised Learning (SSL) with LiDAR, IMU, and elevation during training to learn a versatile thermal representation space for thermal-only inference to enable four off-road mobility tasks, i.e., depth estimation, traversability estimation, roughness prediction, and navigation policy. We systematically study the efficacy of different auxiliary sensor modalities during SSL training on the downstream tasks. Our experiments on a large-scale thermal dataset for off-road mobility demonstrate our improved thermal representation across various downstream tasks and show it can navigate a physical ground robot in no-light conditions using thermal-only input.**

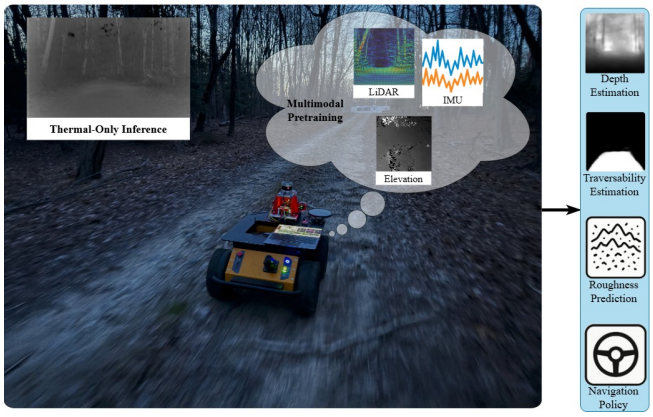

Fig. 1: Leveraging multimodal self-supervised pretraining with auxiliary modalities, SBT's thermal-only representation allows a ground robot to perform a variety of off-road mobility tasks, for which temperature alone is not sufficient.

I. Introduction

Robust autonomous navigation in unstructured off-road environments presents significant challenges for robotic systems, particularly under degraded perceptual conditions such as nighttime operation or the presence of atmospheric obscurants like fog, dust, or smoke [1]. These conditions severely impair the functionality of commonly employed perception sensors. Standard RGB cameras, while providing rich textural and color information, fail dramatically in low-light or no-light scenarios and are easily hampered by visual obscurants [2]. LiDAR sensors, widely used for their ability to provide precise 3D geometric data crucial for mapping and obstacle avoidance, suffer significant performance degradation in adverse weather like heavy rain, snow, fog, or dust due to laser beam scattering and absorption [3]. RADAR sensors offer robustness to weather and lighting conditions but typically lack the spatial resolution needed for detailed environmental modeling and object classification [4]. Furthermore, active sensors like LiDAR or RADAR emit detectable signals, which can be undesirable in applications requiring stealth [5]. Consequently, there is a critical need for perception solutions that enable reliable robot mobility around the clock, especially in challenging off-road environments, ideally leveraging passive sensing modalities.

Thermal infrared camera, specifically Long-Wave Infrared (LWIR) sensors operating in the 8-14 μm range, addresses the significant challenges posed by no-light conditions for autonomous navigation [2]. Operating passively by detecting emitted thermal radiation, they function effectively in complete darkness and varying illumination levels. Moreover, thermal imaging often exhibits superior penetration capabilities through atmospheric obscurants like fog, smoke, and dust compared to visible-light cameras and, in some cases, LiDAR. However, thermal sensors are not without limitations. Thermal images primarily represent the spatial distribution of temperature, which lacks the explicit 3D geometric information provided by LiDAR or the rich texture and color cues available from RGB cameras (in good lighting) [6]. This can impact tasks such as depth estimation, fine-grained obstacle characterization, and detailed terrain analysis. Additionally, thermal image signatures can be ambiguous due to variations in surface emissivity or environmental effects (e.g., wet surfaces appearing cooler than their surroundings).

To address this informational deficit, we introduce Seeing Beyond Temperature (SBT, Figure 1), a framework that learns to overcome the limitations of unimodal thermal sensing by adopting a learning with privileged information paradigm. We conduct an offline, multimodal Self-Supervised Learning (SSL) pretraining phase that operates under good perceptual conditions in dark environments, where auxiliary sensors—including LiDAR, IMU, and elevation—provide clean, high-fidelity data. During this phase, SBT distills the rich geometric and dynamic information from these privileged modalities into a learned thermal representation. The resulting pretrained thermal encoder, now infused with an implicit understanding of the world's structure and dynamics, can then be deployed for thermal-

[1]Department of Computer Science, George Mason University [2]US DEVCOM Army Research Laboratory

only inference. This allows the robot to perform a variety of complex off-road mobility tasks in challenging no-light conditions where the auxiliary sensors would be degraded or entirely unavailable.

The main contributions of this work can be summarized as:

1) We propose SBT, a multimodal SSL framework to learn enhanced thermal representations by leveraging auxiliary LiDAR, IMU, and elevation data during pretraining.
2) We conduct a systematic ablation study analyzing the impact of different combinations of auxiliary modalities on the quality of the learned thermal representations for thermal-only inference.
3) We demonstrate that our learned thermal representations can achieve improved performance on multiple downstream off-road mobility tasks: depth estimation, traversability estimation, roughness prediction, and navigation policy, compared to task-specific baselines.
4) We validate our approach on a large-scale real-world dataset collected in challenging nighttime off-road conditions and demonstrate that the learned thermal embeddings can navigate a physical robot platform with thermal-only input.

II. Related Work

We review related work in off-road autonomy, thermal perception for robot mobility, and self-supervised robot learning.

A. Off-Road Autonomy

Navigating unstructured off-road environments poses significant perception challenges distinct from structured settings, demanding a sophisticated understanding of complex terrain geometry and ambiguous semantics to accurately assess risks [7]–[11]. Research efforts have significantly advanced perception and navigation capabilities for off-road autonomy [12], [13]. For instance, visual self-supervised methods like V-STRONG [14] have learned terrain traversability directly from images, leveraging vision foundation models and contrastive learning to improve generalization in diverse outdoor areas. Systems such as TNES [15] have demonstrated how fusing semantic information from RGB cameras with geometric data from LiDAR enables robust real-time traversability mapping and navigation for heavy machinery like excavators in complex work sites. Addressing the specific challenge of semantic understanding, frameworks like OFFSEG [16] have improved segmentation performance on off-road datasets by pooling classes and using color segmentation for finer details within traversable regions. Furthermore, pushing beyond simple geometric or semantic considerations, competence-aware navigation systems like CAHSOR [17] have leveraged multimodal self-supervision (vision, inertia, and speed) to learn complex 6-DoF vehicle kinodynamics, enabling safer high-speed maneuvering by reasoning about the consequences of actions on varied terrain. Due to the complementary nature of sensors like LiDAR (geometry) and RGB cameras (semantics/appearance), multimodal sensor fusion, as seen in approaches like TNES [15], has become a standard technique to achieve robust perception across diverse conditions. However, this reliance on sensor fusion, particularly involving RGB cameras or active sensors such as LiDAR, introduces a critical vulnerability: These systems often fail in complete darkness, adverse weather, or scenarios requiring stealth, where passive perception becomes necessary. This limitation motivates alternative sensors, such as thermal cameras, capable of operating effectively in these demanding conditions.

B. Thermal Perception for Robot Mobility

Thermal cameras offer unique advantages for perception in challenging environmental conditions, a capability increasingly recognized through the development of dedicated multimodal datasets. For instance, ViViD++ [18] has provided diverse visual data targeting varying luminance conditions; M2P2 [19] has specifically focused on passive perception for off-road mobility in extreme low-light scenarios. GO [20] has included thermal data alongside other sensors for general unstructured environments. Building upon the potential demonstrated by such data resources, recent specific applications include ThermalVoyager [21], which has explored thermal cameras for autonomous navigation, but focused primarily on structured on-road environments with well-defined features. Shin et al. [22] have developed techniques for thermal monocular depth estimation, maximizing self-supervision from thermal images for effective depth and ego-motion learning. However, these approaches either target structured environments or address only one specific perception task, leaving a critical gap in terms of comprehensive thermal-only perception for different off-road mobility tasks in extreme conditions. SBT directly addresses this gap by enhancing thermal representations through multimodal self-supervised learning.

C. Self-Supervised Robot Learning

Self-supervised learning has emerged as a powerful paradigm for developing robust perceptual representations without relying on extensive manual annotations, particularly valuable for robotics applications in complex, unstructured environments. In the context of off-road mobility, several approaches have leveraged SSL to improve downstream task performance [17], [23]–[30]. Most existing self-supervised approaches for off-road navigation rely on sensor fusion during both training and deployment. Jeon et al. [31] used RGB and depth inputs at deployment time while leveraging robot trajectories for self-supervision. Similarly, Gasparino et al. [32] integrated RGB and depth information using convolution layers, with the robot's trajectory serving as self-generated ground truth. IRISPath [33] enhances costmaps for off-road navigation through early IR-RGB fusion for day and night traversability. Ægidius et al. [34] took a different approach, using only RGB at deployment time while employing semantic traversability estimation with pose-projected features during training. While these methods advance the state of the art in traversability estimation and navigation, they rely on RGB cameras during inference, which fail in

no-light conditions, where our thermal-only approach excels. On the other hand, recent datasets such as M2P2 [19] provide an opportunity to investigate thermal perception for off-road mobility in no-light conditions. SBT directly leverages M2P2 and develops a SSL framework that systematically leverages auxiliary modalities during pretraining to enhance thermal representations, enabling robust thermal-only inference across multiple off-road mobility tasks, a capability not demonstrated in previous research.

III. Approach

SBT is a two-stage process comprising a multimodal self-supervised pretraining phase, followed by training downstream tasks on the learned representation (Figure 2). In the pretraining stage, we leverage auxiliary sensing modalities (LiDAR, IMU, and 2.5D elevation) alongside thermal images to learn versatile thermal representations that capture geometric and semantic terrain information, which otherwise does not exist in thermal-only input. Notice that while we choose representative LiDAR, IMU, and elevation, other auxiliary modalities can be easily incorporated when necessary.

A. Learning Enriched Thermal Features with Auxiliary Sensors

The core of SBT lies in its multimodal self-supervised pretraining strategy. The objective is to train a thermal image encoder such that its output feature embeddings contain information more than temperature, e.g., geometric structure and dynamic interaction cues, guided by the auxiliary sensors during training.

LiDAR is chosen for its direct measurement of the environment's 3D structure, yielding precise point-cloud data that capture detailed spatial layout, object shapes, and distances. This explicit 3D information is crucial for grounding the learned thermal features in the physical geometry of the scene, directly compensating for the lack of inherent geometric features in thermal images. Since LiDAR points are sparse we project the 3D points onto the thermal camera frame using the LiDAR-Camera extrinsics and intrinsics. We then interpolate the sparse projected points to create dense depth maps. IMU provides high-frequency measurements of the robot's linear acceleration and angular velocity, offering a dynamic, proprioceptive view that directly reflects the physical interaction with the terrain. This proprioceptive modality captures effects such as vibrations and changes in body motion induced by the different surfaces, which thermal images cannot convey. Complementing the raw LiDAR data, we utilize 2.5D elevation maps [35] as a processed pseudo-modality specifically curated to represent the most relevant geometric information for off-road mobility. Derived from LiDAR point clouds and IMU data, these maps offer a structured 2.5D grid encoding critical ground attributes such as height, slope, and local roughness within the vicinity of the robot. Including elevation maps guides the thermal representation towards mobility-critical geometric attributes. This processed geometric view serves as a bridge between the raw spatial data from LiDAR and the dynamic interaction data from IMU, facilitating the infusion of navigable surface characteristics into the thermal representation.

By integrating these three modalities, raw geometry (LiDAR), raw dynamics (IMU), and processed mobility-focused geometry (Elevation Maps), SBT aims to produce thermal representations that not only capture spatial thermal variations, but are also implicitly aware of the underlying terrain structure, composition, and the dynamic interactions they afford.

B. Self-Supervised Representation Learning for Off-Road Navigation

SBT's SSL framework is built around a shared vision encoder that processes multiple thermal and auxiliary inputs into an augmented embedding space (Figure 2 left). At its core, a ResNet-50 [36] is used to process all modalities, including thermal images, LiDAR-derived depth images, and 2.5D elevation maps. For inertial measurements, a dedicated IMU encoder is used to process the high-frequency IMU signals over synchronized two-second windows, being processed into Power Spectral Density (PSD) of linear accelerations and angular velocities. The initial embeddings $\mathbf{y}$ from each encoder are then passed through a dedicated projector head. The projector head maps the features into a common, high-dimensional latent space, producing the final feature vectors $\mathbf{z}^t$, $\mathbf{z}^l$, $\mathbf{z}^i$, and $\mathbf{z}^e$ (for thermal, LiDAR, IMU, and elevation, respectively) upon which the self-supervised loss is computed.

We treat each modality as a different "view" of the same off-road environment, analogous to how multi-view self-supervised learning treats two augmented images [37]. Crucially, our goal is not to perform pixel-wise fusion, but to learn a semantically unified, shared embedding of scene structure across modalities. For instance, the model is trained to learn that the abstract concept of a "rock" should produce a similar feature vector, whether it is perceived as a specific thermal signature in the camera view or as a cluster of high-elevation points in the Bird's-Eye-View (BEV) elevation map.

We achieve this alignment by applying a redundancy-reduction loss between the thermal latent and each auxiliary latent representation.

Barlow Twin Loss. We adopt the Barlow Twins (BT) self-supervised objective [38] to drive the cross-modal feature alignment. Originally proposed for two augmented images, here we extend it to align four carefully selected complementary modalities: thermal images, LiDAR point clouds (represented as depth images), IMU signals, and 2.5D elevation maps. For any thermal–auxiliary pair (t, a) with $a \in \{l, i, e\}$, we first normalize each latent dimension throughout the batch to zero mean and unit variance. We then compute the BT loss for each pair:

$$\mathcal{L}_{\mathrm{BT}}^{(t,a)} = \sum_{i=1}^{d} \big(C_{ii}^{(t,a)} - 1\big)^2 \; + \; \lambda \sum_{i \neq j} \big(C_{ij}^{(t,a)}\big)^2 ,$$

where λ is a trade-off parameter and $C^{(t,a)}$ is the cross-correlation matrix computed between the outputs z^t and z^a.

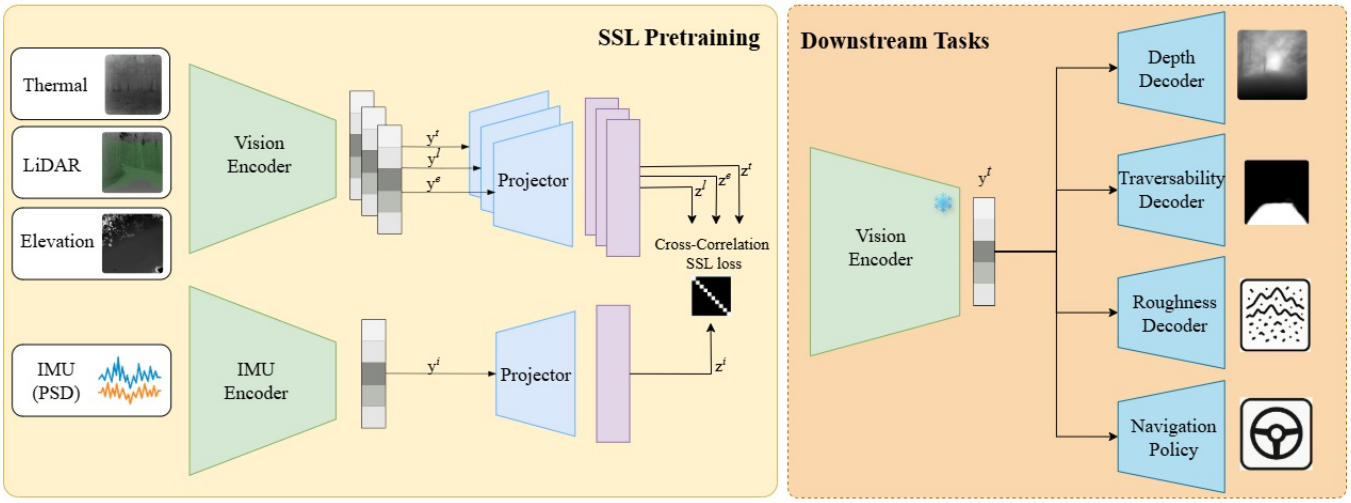

Fig. 2: Overview of our SBT Pretraining Pipeline (Left) and four Downstream Tasks (Right).

The total loss is the sum of losses for each thermal-auxiliary pair:

$$\mathcal{L}_{\text{Total\_Loss}} = \mathcal{L}_{\text{BT}}^{(t,l)} + \mathcal{L}_{\text{BT}}^{(t,i)} + \mathcal{L}_{\text{BT}}^{(t,e)} .$$

By minimizing $\mathcal{L}_{\text{Total\_Loss}}$, the model is trained to distill LiDAR geometry, IMU-derived dynamics, and terrain elevation structure directly into its temperature-based features.

C. Downstream Tasks

The goal of the SSL pretraining is to distill geometric and dynamic awareness into our SBT thermal embeddings in order to enable a diverse set of downstream off-road mobility tasks (Figure 2 right). For each downstream task, we freeze the SBT encoder and use it as a feature extractor, attaching a lightweight task-specific head to map the thermal feature embedding to the desired output. This head is trained in a supervised manner using task-specific labels. This paradigm enables us to systematically measure how effectively self-supervised thermal features support various terrain understanding and mobility objectives, without access to auxiliary sensor data at inference time.

Depth Estimation. Accurate depth estimation is fundamental for obstacle avoidance and 3D scene understanding. Monocular Depth Estimation (MDE) is inherently challenging due to scale ambiguity and reliance on image cues. Thermal MDE faces additional difficulties due to lower resolution, less texture, and potential noise compared to RGB MDE. To predict depth map, we attach a convolutional decoder to the frozen SBT encoder. The decoder uses five up-sampling stages with bilinear interpolation and skip-connections from the encoder to recover spatial detail. We train this head using a composite loss function:

$$L_{\text{depth}} = 0.15L_{\ell_1} + 0.85L_{\text{SSIM}} + 0.6L_{\text{grad}} + 0.3L_{\text{smooth}},$$

where L1 loss L_{ℓ_1} measures pixel-wise absolute differences, the Structural Similarity loss L_{SSIM} captures structural details, the gradient consistency loss L_{grad} enforces gradient consistency, and the edge-aware smoothness loss L_{smooth} promotes spatial smoothness with edge-aware weighting. The geometric knowledge implicitly encoded in the SBT thermal features, derived from LiDAR and elevation map supervision during pretraining, is expected to improve MDE accuracy compared to baseline approaches.

Traversability Estimation. Traversability estimation involves predicting a per-pixel likelihood map (range: 0-1) indicating safe navigation regions from a single thermal image. This segmentation task is critical for path planning in unstructured environments where RGB-based methods struggle with no-light or obscured conditions. To generate the ground-truth binary mask for this task, we project the robot's future footprint into each thermal image using odometry poses from Direct LiDAR-Inertial Odometry [39]. For a sequence of poses, a canonical set of points representing the robot's footprint is transformed into the camera frame and projected to the pixel coordinates. For each segment of the projected path, a perspective-aware footprint is rasterized. The width of this footprint is dynamically scaled based on its distance form camera, accurately representing the robot's constant physical width as it moves through the scene. The aggregation of these filled footprints creates a complete binary mask, labeling the projected path as traversable with pixel values of 1 and all other areas as non-traversable with pixel values of 0.

The decoder head comprises of upsampling layers with spatial attention gates, integrating multi-scale skip connections from the frozen SBT encoder. Dilated convolutions and bilinear interpolation recover fine-grained 256×256 spatial resolution while preserving mobility-critical features highlighted by the attention modules. Training uses a combination of binary cross-entropy (BCE) and Dice loss:

$$L_{\text{trav}} = L_{\text{BCE}} + L_{\text{Dice}},$$

where BCE penalizes each pixel's classification error and encourages well-calibrated confidence scores, while the Dice term directly maximizes overlap between prediction and

ground truth, making it robust to class imbalance (i.e., small traversable regions) and sharper boundary alignment.

Roughness Prediction. Terrain roughness is a critical task for autonomous ground vehicles operating in complex, off-road environments. Estimating roughness enables a robot to anticipate mechanical vibrations, maintain stability, adapt speed, and proactively avoid hazardous or high-drag terrain patches. Robust roughness prediction is essential for safety, comfort, and mission success, especially when geometric sensors like LiDAR are noisy, occluded, or unavailable.

To estimate roughness, we leverage proprioceptive measurements from the robot's onboard IMU as ground truth. For each thermal image, we synchronize and aggregate IMU data (linear accelerations and gyroscopic angular velocities) collected as the robot traverses the terrain segment appearing in the image. We compute temporal jerk (the derivative of the linear acceleration in the z-axis a_{jz}) and high-frequency power in gyroscope signals (particularly roll g_{ax} and pitch g_{ay} rates). These features are combined into a scalar weighted sum:

$$\text{Roughness} = w_{jz}|a_{jz}| + w_{gx}|g_{ax}| + w_{gy}|g_{ay}|$$

where weights for vertical shocks $w_{jz} = 0.075$, and rotational disturbances $w_{gx} = w_{gy} = 0.475$. The weighting scheme intentionally places much greater emphasis on gyroscopic terms (roll and pitch rates) than on vertical acceleration jerk. In off-road robotics, roll and pitch disturbances correspond to sustained, high-frequency shaking caused by uneven or sloped terrain. The vertical acceleration jerk, though important for detecting sharp bumps, is down-weighted to prevent isolated spikes (such as single impacts or sensor noise) from dominating the roughness score. This fusion mirrors the qualitative "felt" experience of roughness by a ground vehicle. The roughness decoder is a lightweight regression head attached to the frozen SBT encoder. Given a thermal image, the encoder produces a d-dimensional feature embedding, which is passed through three fully connected layers with BatchNorm and ReLU activations. The final output layer produces a single scalar prediction for the terrain roughness score corresponding to the input image. Performance is assessed via Mean Squared Error (MSE) loss. Optimization is performed using AdamW optimizer. This setup allows the model to robustly map thermal visual cues to experienced terrain roughness, even in the absence of explicit geometric or proprioceptive signals at inference time.

Navigation Policy. We evaluate the utility of the SBT thermal representations for navigation policy learning using Behavior Cloning (BC) [40], [41]. Given a stream of front-facing thermal images as input, the goal is to learn a policy π_θ that outputs the appropriate action $a_t = [v_t, \omega_t]$ for each timestep t, where v_t is the linear velocity and ω_t is the angular velocity command. Our navigation policy consists of a pretrained, frozen SBT vision encoder followed by a lightweight multi-layer perceptron (MLP) policy head. The policy head is trained with mean squared error loss.

IV. Experiments

We evaluate the effectiveness and demonstrate the versatility of the thermal representations learned by SBT. We detail our experimental setup, compare our method to established baseline approaches, and present quantitative results on all four downstream mobility tasks. An ablation study is conducted to analyze the contribution of each auxiliary modality during pretraining. Finally, we provide qualitative visualizations to offer further insight into our approach's performance. We also deploy the navigation policy and depth estimation downstream tasks on a physical robot to demonstrate real-world off-road navigation with thermal-only input in no-light conditions.

A. Experiment Setup

All experiments are conducted on the M2P2 dataset [19], a large-scale dataset to facilitate night-time off-road robot navigation research. This dataset provides synchronized multimodal data streams essential for our approach, including thermal camera imagery (T), LiDAR sensor point clouds (L), and IMU measurements (I).

To assess the contribution of each auxiliary modality, we conduct pretraining in three distinct ablation configurations:

- **(T+L):** Pretraining on Thermal and LiDAR only, using $\mathcal{L}_{\text{BT}}^{(t,l)}$.
- **(T+L+I):** Pretraining on Thermal, LiDAR, and IMU, with a summed objective $\mathcal{L}_{\text{BT}}^{(t,l)} + \mathcal{L}_{\text{BT}}^{(t,i)}$.
- **(T+L+I+E):** Pretraining on Thermal, LiDAR, IMU, and Elevation, minimizing the total loss $\mathcal{L}_{\text{Total\_Loss}} = \mathcal{L}_{\text{BT}}^{(t,l)} + \mathcal{L}_{\text{BT}}^{(t,i)} + \mathcal{L}_{\text{BT}}^{(t,e)}$.

All pretraining configurations are trained with AdamW [42], and a batch size of 128 for 500 epochs. The BT hyperparameter λ is set to 5×10^{-3}. Due to computational limitations, for Thermal + LiDAR + IMU + Elevation pretraining, we use a batch size of 64.

We benchmarked the computational performance of our thermal-only inference pipeline for all downstream tasks on the physical robot's hardware. The results, including latency, frame rate, and memory usage, are detailed in Table I

TABLE I: Computational Performance for all Downstream Tasks on the Husky A200 robot (NVIDIA RTX 3060, 12GB).

Task	Latency	FPS	GPU Mem.	Total Params (Encoder + Downstream Head)
Navigation Policy	20.12 ms	49.70	127 MB	23.5M+2.7M
Depth Est.	32.86 ms	30.43	1284 MB	23.5M+305.7M
Traversability Est.	27.76 ms	36.02	1305 MB	23.5M+311.3M
Roughness Pred.	22.96 ms	43.55	121 MB	23.5M+1.1M

For all four downstream tasks (Depth Estimation, Traversability Estimation, Roughness Prediction, Navigation Policy), the SBT encoder is frozen. Only a lightweight, task-specific decoder head is trained. Inference for all tasks uses only thermal image input.

B. Baselines

To demonstrate the performance of the SBT thermal representations, comparisons are made against state-of-the-

art baseline methods specifically designed for each off-road mobility task:

- **Depth Estimation:** We compare against ThermalMonoDepth [22], a self-supervised monocular depth estimation method on thermal images. Benchmarking against ThermalMonoDepth allows us to examine how well the SBT representations distill geometric priors into the thermal features in a way that directly benefits pixel-wise depth under challenging off-road and nighttime conditions.
- **Traversability Estimation:** We compare against STEPP [34]. STEPP performs semantic traversability estimation by projecting future poses and leverages pose-projected features to label traversable regions. Since STEPP was originally trained on RGB data, we train the STEPP model on M2P2.
- **Roughness Prediction:** We do not find any relevant thermal-based (or RGB-based) roughness prediction method with open-source implementation for us to (adapt based on M2P2 and) compare against. So we implement our own end-to-end approach that regresses from thermal image.
- **Navigation Policy:** We compare against the end-to-end BC approach in M2P2 [19]. M2P2 BC directly regresses from thermal images to linear and angular velocity commands without a pretraining process that leveraging auxiliary modalities like LiDAR, IMU, and elevation.

C. Quantitative Results

The performance of the SBT representations learned with three different modality combinations during pretraining is quantitatively evaluated and compared to corresponding baselines (see Table II).

The results clearly demonstrate that all variants of SBT (T+L, T+L+I, and T+L+I+E) outperform their baselines corresponding to each downstream off-road mobility task (except T+L+I for roughness prediction MSE), showcasing the effectiveness and versatility of the learned SBT thermal representations. Within SBT, the difference in task performance caused by different input modality combinations during pretraining is more subtle. The addition of IMU data (T+L+I) results in a slight degradation in depth estimation performance compared to the T+L model (Abs Rel 0.152 vs. 0.163), yet provides a notable improvement in traversability precision (0.911 vs. 0.923). This is a key finding, as it highlights a fundamental tradeoff in multimodal representation learning. While dynamics cues from the IMU are highly beneficial for traversability estimation and navigation, they introduce signals that are non-informative or act as noise for a purely static, geometric task like depth estimation. When trained with IMU data, the learning objective pulls the feature space away from a purely geometric optimum to also encode vehicle motion, resulting in the minor performance degradation on the static depth task. The further addition of elevation maps (T+L+I+E) consistently boosts segmentation metrics, particularly IoU and F1, emphasizing the value of structured terrain priors for identifying safe navigable surfaces. Notably, roughness prediction and navigation policy tasks exhibit stable performance across all configurations, indicating that SBT's thermal backbone, once seeded with geometric information, generalizes well to dynamics and control-relevant information with or without explicit IMU or elevation supervision. Overall, these findings illustrate that geometric, inertial, and structural modalities each offer synergistic benefits for representation learning, but their relative impact depends on the demands of the downstream task.

D. Qualitative Results

To complement the quantitative findings, we present qualitative examples demonstrating the impact of SBT's multimodal representation learning on downstream tasks in no-light off-road environments.

Figure 3 top shows the thermal input, ground truth depth, and reconstructed depth by ThermalMonoDepth [22] and SBT. ThermalMonoDepth does not produce depth qualitatively similar to the ground truth, while SBT captures most depth information with some blurred details. For traversability estimation, Figure 3 bottom shows that SBT is able to produce a traversable mask that correctly identifies the nontraversable area due to the tree trunk on the left, while STEPP [34] does not generate satisfactory results.

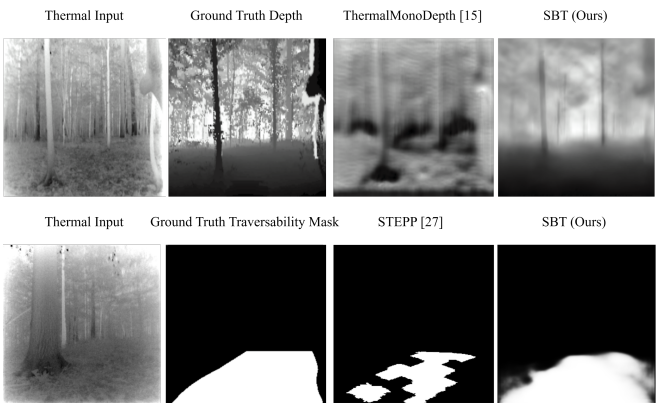

Fig. 3: Qualitative Comparison of Depth Estimation (top) and Traversability Estimation (bottom).

E. Physical Demonstration

To validate real-world performance, we deploy the learned navigation policy on a Clearpath Husky A200 robot on an unseen 50-meter, mixed-terrain off-road trail at night (Figure 4). The trail consists of a narrow paved path bordered by surrounding forest. The demonstration confirms that the knowledge distilled from multimodal pretraining enables successful generalization, allowing the robot to navigate using only thermal input. Occasional minor human interventions are required when encountering sharp turns. These interventions underscore the performance gap between short-horizon navigation and the capabilities required for extended missions, which demand more sophisticated spatial

TABLE II: Quantitative Comparison across Pretraining Configurations and Baselines.

Task	Metric	T+L	T+L+I	T+L+I+E	Baseline
					TMD [22]
Depth Estimation	Abs Rel ↓	**0.152**	0.163	0.162	1.162
	RMSE ↓	**3.424**	3.624	3.464	9.477
	δ_1 ↑	**0.789**	0.772	0.781	0.300
	δ_2 ↑	**0.953**	0.943	0.946	0.468
	δ_3 ↑	**0.986**	0.981	0.982	0.620
					STEPP [34]
Traversability Estimation	IoU ↑	0.865	0.862	**0.870**	0.1257
	F1 ↑	0.927	0.926	**0.928**	0.2233
	Precision ↑	0.911	**0.923**	0.914	0.8266
	Recall ↑	**0.945**	0.928	0.914	0.1291
					End-to-End
Roughness Prediction	MSE ↓	**0.004**	0.005	**0.004**	**0.004**
					M2P2 BC [19]
Navigation Policy	MSE ↓	**0.006**	**0.006**	0.007	0.009

and temporal reasoning. Thus, while our approach establishes foundational capability for thermal-only navigation in no-light conditions, robust long-horizon off-road autonomy at night remains an open challenge for future work.

Fig. 4: Physical robot navigation in near-darkness using SBT's thermal-only navigation policy, showing Husky robot at the beginning (left), middle (center), and end (right) of a successful, collision-free navigation sequence on a 50-meter off-road trail.

V. CONCLUSIONS

We present SBT, a self-supervised multimodal representation learning framework that leverages thermal, LiDAR, inertial, and elevation perception during pretraining to enable a variety of downstream, thermal-only, off-road mobility tasks in no-light conditions, i.e., depth estimation, traversability estimation, roughness prediction, and navigation policy. Our experiments demonstrate that SBT consistently surpasses established, task-specific baselines for each downstream task on challenging nighttime datasets. Through our systematic ablation study, we show that each auxiliary modality, geometric, inertial, and structural, provides different levels of benefits to the thermal representation, depending on the specific downstream off-road mobility task.

REFERENCES

[1] C. Min, S. Si, X. Wang, H. Xue, W. Jiang, Y. Liu, J. Wang, Q. Zhu, Q. Zhu, L. Luo *et al.*, "Autonomous driving in unstructured environments: How far have we come?" *arXiv preprint arXiv:2410.07701*, 2024.

[2] A. S. Bhadoriya, V. Vegamoor, and S. Rathinam, "Vehicle detection and tracking using thermal cameras in adverse visibility conditions," *Sensors*, vol. 22, no. 12, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/12/4567

[3] M. Dreissig, D. Scheuble, F. Piewak, and J. Boedecker, "Survey on lidar perception in adverse weather conditions," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[4] J. Vargas, S. Alsweiss, O. Toker, R. Razdan, and J. Santos, "An overview of autonomous vehicles sensors and their vulnerability to weather conditions," *Sensors*, vol. 21, no. 16, p. 5397, 2021.

[5] T. Raj, F. Hanim Hashim, A. Baseri Huddin, M. F. Ibrahim, and A. Hussain, "A survey on lidar scanning mechanisms," *Electronics*, vol. 9, no. 5, p. 741, 2020.

[6] T. X. B. Nguyen, K. Rosser, and J. Chahl, "A review of modern thermal imaging sensor technology and applications for autonomous aerial navigation," *Journal of Imaging*, vol. 7, no. 10, p. 217, 2021.

[7] A. Datar, C. Pan, and X. Xiao, "Learning to model and plan for wheeled mobility on vertically challenging terrain," *arXiv preprint arXiv:2306.11611*, 2023.

[8] A. Datar, C. Pan, M. Nazeri, and X. Xiao, "Toward wheeled mobility on vertically challenging terrain: Platforms, datasets, and algorithms," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[9] X. Xiao, J. Biswas, and P. Stone, "Learning inverse kinodynamics for accurate high-speed off-road navigation on unstructured terrain," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6054–6060, 2021.

[10] H. Karnan, K. S. Sikand, P. Atreya, S. Rabiee, X. Xiao, G. Warnell, P. Stone, and J. Biswas, "Vi-ikd: High-speed accurate off-road navigation using learned visual-inertial inverse kinodynamics," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3294–3301.

[11] T. Xu, C. Pan, M. B. Rao, A. Datar, A. Pokhrel, Y. Lu, and X. Xiao, "Verti-bench: A general and scalable off-road mobility benchmark for vertically challenging terrain," in *Robotics: Science and Systems (RSS) 2025*, 2025.

[12] X. Xiao, B. Liu, G. Warnell, and P. Stone, "Motion planning and control for mobile robot navigation using machine learning: a survey," *Autonomous Robots*, vol. 46, no. 5, pp. 569–597, 2022.

[13] K. S. Sikand, S. Rabiee, A. Uccello, X. Xiao, G. Warnell, and J. Biswas, "Visual representation learning for preference-aware path planning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 303–11 309.

[14] S. Jung, J. Lee, X. Meng, B. Boots, and A. Lambert, "V-strong: Visual self-supervised traversability learning for off-road navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 1766–1773.

[15] T. Guan, Z. He, R. Song, and L. Zhang, "Tnes: terrain traversability mapping, navigation and excavation system for autonomous excavators on worksite," *Auton. Robots*, vol. 47, no. 6, p. 695–714, Jul. 2023. [Online]. Available: https://doi.org/10.1007/s10514-023-10113-9

[16] K. Viswanath, K. Singh, P. Jiang, P. Sujit, and S. Saripalli, "Offseg: A semantic segmentation framework for off-road driving," in *2021 IEEE*

17th International Conference on Automation Science and Engineering (CASE). IEEE, 2021, pp. 354–359.

[17] A. Pokhrel, A. Datar, M. Nazeri, and X. Xiao, "CAHSOR: Competence-aware high-speed off-road ground navigation in SE (3)," *IEEE Robotics and Automation Letters*, 2024.

[18] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: Vision for visibility dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.

[19] A. Datar, A. Pokhrel, M. Nazeri, M. B. Rao, C. Pan, Y. Zhang, A. Harrison, M. Wigness, P. R. Osteen, J. Ye, and X. Xiao, "M2p2: A multi-modal passive perception dataset for off-road mobility in extreme low-light conditions," in *IEEE/RSJ International Conference on Intelligent Robots (IROS)*, 2025.

[20] P. Jiang, K. Viswanath, A. Nagariya, G. Chustz, M. Wigness, P. Osteen, T. Overbye, C. Ellis, L. Quang, and S. Saripalli, "Go: The great outdoors multimodal dataset," *arXiv preprint arXiv:2501.19274*, 2025.

[21] N. Aditya, P. Dhruval, J. Shalabi, S. Jape, X. Wang, and Z. Jacob, "Thermal voyager: A comparative study of rgb and thermal cameras for night-time autonomous navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 116–14 122.

[22] U. Shin, K. Lee, B.-U. Lee, and I. S. Kweon, "Maximizing self-supervision from thermal image for effective self-supervised learning of depth and ego-motion," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7771–7778, 2022.

[23] A. Datar, C. Pan, M. Nazeri, A. Pokhrel, and X. Xiao, "Terrain-attentive learning for efficient 6-dof kinodynamic modeling on vertically challenging terrain," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024.

[24] M. Nazeri, A. Pokhrel, A. Card, A. Datar, G. Warnell, and X. Xiao, "Vertiformer: A data-efficient multi-task transformer for off-road robot mobility," *arXiv preprint arXiv:2502.00543*, 2025.

[25] M. Nazeri, A. Datar, A. Pokhrel, C. Pan, G. Warnell, and X. Xiao, "Vertiencoder: Self-supervised kinodynamic representation learning on vertically challenging terrain," *arXiv preprint arXiv:2409.11570*, 2024.

[26] M. Nazeri, J. Wang, A. Payandeh, and X. Xiao, "Vanp: Learning where to see for navigation with self-supervised vision-action pre-training," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 2741–2746.

[27] H. Karnan, E. Yang, D. Farkash, G. Warnell, J. Biswas, and P. Stone, "Sterling: Self-supervised terrain representation learning from unconstrained robot experience," in *Conference on Robot Learning*. PMLR, 2023, pp. 2393–2413.

[28] 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.

[29] 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)*. IEEE, 2024, pp. 12 606–12 606.

[30] 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.

[31] Y. Jeon, E. I. Son, and S.-W. Seo, "Follow the footprints: Self-supervised traversability estimation for off-road vehicle navigation based on geometric and visual cues," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 1774–1780.

[32] M. V. Gasparino, A. N. Sivakumar, and G. Chowdhary, "Wayfaster: a self-supervised traversability prediction for increased navigation awareness," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 8486–8492.

[33] S. Sharma, A. Raizada, and S. Sundaram, "Irispath: Enhancing costmap for off-road navigation with robust ir-rgb fusion for improved day and night traversability," 2025. [Online]. Available: https://arxiv.org/abs/2412.03173

[34] S. Ægidius, D. Hadjivelichkov, J. Jiao, J. Embley-Riches, and D. Kanoulas, "Watch your stepp: Semantic traversability estimation using pose projected features," *arXiv preprint arXiv:2501.17594*, 2025.

[35] T. Miki, L. Wellhausen, R. Grandia, F. Jenelten, T. Homberger, and M. Hutter, "Elevation mapping for locomotion and navigation using gpu," 2022.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decoupling common and unique representations for multimodal self-supervised learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 286–303.

[38] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International conference on machine learning*. PMLR, 2021, pp. 12 310–12 320.

[39] K. Chen, R. Nemiroff, and B. T. Lopez, "Direct lidar-inertial odometry: Lightweight lio with continuous-time motion correction," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3983–3989.

[40] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed., vol. 1. Morgan-Kaufmann, 1988.

[41] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to End Learning for Self-Driving Cars," Apr. 2016.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.