M2P2: A Multi-Modal Passive Perception Dataset for Off-Road Mobility in Extreme Low-Light Conditions

Aniket Datar*[1], Anuj Pokhrel*[1], Mohammad Nazeri*[1], Madhan B. Rao*[1], Harsh Rangwala[1], Chenhui Pan[1], Yufan Zhang[1], André Harrison[2], Maggie Wigness[2], Philip R. Osteen[2], Jinwei Ye[1], and Xuesu Xiao[1]

***Abstract*—Long-duration, off-road, autonomous missions require robots to continuously perceive their surroundings regardless of the ambient lighting conditions. Most existing autonomy systems heavily rely on active sensing, e.g., LiDAR, RADAR, and Time-of-Flight sensors, or use (stereo) visible light imaging sensors, e.g., color cameras, to perceive environment geometry and semantics. In scenarios where fully passive perception is required and lighting conditions are degraded to an extent that visible light cameras fail to perceive, most downstream mobility tasks such as obstacle avoidance become impossible. To address such a challenge, this paper presents a Multi-Modal Passive Perception dataset, M2P2, to enable off-road mobility in low-light to no-light conditions. We design a multi-modal sensor suite including thermal, event, and stereo RGB cameras, GPS, two Inertia Measurement Units (IMUs), as well as a high-resolution LiDAR for ground truth, with a multi-sensor calibration procedure that can efficiently transform multi-modal perceptual streams into a common coordinate system. Our 10-hour, 32 km dataset also includes mobility data such as robot odometry and actions and covers well-lit, low-light, and no-light conditions, along with paved, on-trail, and off-trail terrain. Our results demonstrate that off-road mobility and scene understanding under degraded visual environments is possible through only passive perception in extreme low-light conditions. The project website can be found at https://cs.gmu.edu/~xiao/Research/M2P2/.**

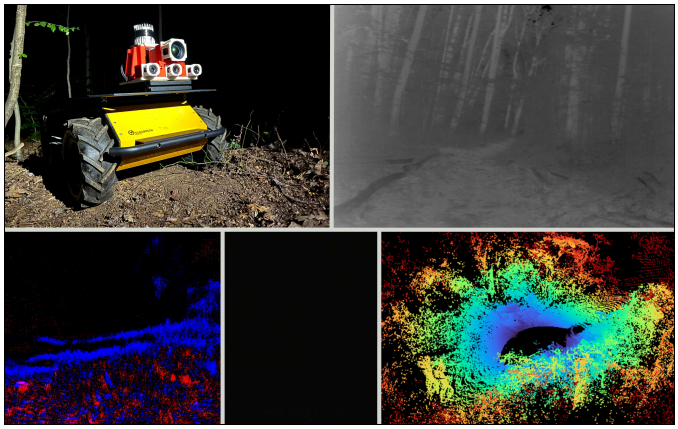

Fig. 1: Multi-Modal Passive Perception Data Collection in an Off-Road Forest Environment in Complete Darkness. Top Left: Clearpath Husky with the Sensor Suite (flashlight for visualization only); Top Right: Thermal Image; Bottom Left: Event Stream; Bottom Middle: RGB Image (fail to perceive); Bottom Right: LiDAR Point Cloud (for ground truth).

I. Introduction

Autonomous mobile robots have found their way out of controlled lab, factory, and warehouse environments into the wild [1]. On their way to deliver packages [2], inspect infrastructure [3], maintain agricultural fields [4], and conduct search and rescue missions [5], those robots constantly perceive their surroundings with their onboard sensors. The perceived geometric and semantic world representations allow them to move to their goals while avoiding collisions. Such an extension in Operational Design Domain requires robot perception systems to address challenges around the clock, ranging from well-lit to no-light conditions, as well as from paved to completely off-road terrain in the wild.

Existing perception systems for mobile robots rely heavily on active sensing. For example, LiDAR range finders [6] use pulsed laser beams to detect distance and perceive environmental geometry, while Time-of-Flight sensors [7] use infrared light and measure the time it takes for the light signal to travel to the target and back. Despite working well in all lighting conditions, many active sensors suffer from significant noise in heavy rain, snow, and fog. Furthermore, the reliance on the emission of active light signals will expose the presence of the robot, making those active sensors less ideal for covert operations, e.g., in military settings.

Non-active, visible light imaging sensors, e.g., RGB cameras, are also widely used in robot perception systems, relying on reflected light to form images for non-light emitting objects. Stereo camera pairs can triangulate to determine distance and use different RGB color channels to reason about semantics. Those sensors work well in well-lit indoor and outdoor environments and provide similar sensing as human perception. However, visible light imaging sensors require good lighting conditions to perceive reflected light and form visible pixels, and therefore suffer from degraded perception quality in low-light to no-light conditions.

These aforementioned limitations of existing active and visible light imaging sensors present challenges for long-duration, off-road, autonomous missions, since robots need to perceive their surroundings around the clock regardless of the ambient lighting conditions and are also oftentimes required to be fully passive to maintain stealth. To operate in low-light to no-light conditions without emitting any active light signatures, novel sensing modalities, including thermal and event cameras, show promise by passively sensing infrared radiation from all objects with a temperature above absolute zero or per-pixel brightness changes (also called "events") asynchronously with low latency, high dynamic range, and low power consumption, respectively.

In this paper, we propose to use multi-modal passive percep-

tion modalities to enable robot perception in extreme low-light conditions so as to facilitate downstream off-road mobility tasks (Fig. 1). To be specific, our contributions include:

- a multi-modal sensor suite including thermal, event, and stereo RGB cameras, GPS, two IMUs, and a high-resolution LiDAR for ground truth;
- a precise multi-sensor calibration procedure for multi-modal perceptual streams;
- a Multi-Modal Passive Perception dataset, M2P2, with data ranging from different lighting conditions (well-lit to no-light) and various off-road terrain conditions (paved to off-trail), along with mobility data like robot odometry and actions; and
- experimental results demonstrating off-road mobility, depth reconstruction, and vehicle odometry through only passive perception in extreme low-light conditions.

II. Related Work

In this section, we review related work in off-road perception systems and passive perception sensors.

A. Off-Road Perception

Perception in off-road environments requires both exteroceptive and interoceptive sensing to understand the environment and the robot's interaction with it. The availability of a wide array of sensors makes safe traversal through off-road environments possible. While a single modality may suffice for navigation in structured environments, the inclusion of multiple modalities in challenging environments adds robustness and redundancy,

ensuring that navigation can continue even if one or more sensors are unable to work at full capacity because of adverse environmental conditions. By combining complementary data from multiple sensors, robots can also better perceive and interpret complex environmental features for comprehensive understanding in a variety of off-road unstructured scenarios.

Active sensing modalities like LiDAR and RADAR detect and perceive environmental geometry, enabling the creation of 2D, 3D, or 2.5D elevation maps [11, 12, 13, 14, 15] of the environment. Although LiDAR-based systems are highly popular for their robustness and precision, they can suffer in heavy rain, snow, and fog, and may struggle to map terrain at greater distances [16]. Additionally, the use of pulsated beams can expose the presence of the robot. On the other hand, vision-based navigation systems utilize visible light imaging sensors, e.g., RGB or RGB-D cameras, to understand the terrain semantics [17, 18, 19, 20], create elevation maps [16, 18], and map off-road terrain [21, 22]. Although vision-based navigation systems are advantageous due to their passive sensing capabilities and ability to provide rich environmental information, their reliance on visible light causes poor performance in low-light conditions.

While also being passive, interoceptive sensors like IMUs and force sensors measure robot internal states during environment interactions, which can be used to generate traversability maps [17, 23] and model terrain response [24, 25] when combined with exteroception.

Combining the advantages of the aforementioned perception modalities expands robots' Operational Design Domain in varying environmental conditions around the clock, such as low visibility or extreme weather, with the possibility of staying passive. With the recent advancement in data-driven approaches [1], multi-modal off-road datasets [8, 26, 27] are essential for developing and refining perception and mobility algorithms, providing a foundation for training, testing, and benchmarking. Our multi-modal sensor suite offers passive sensing capabilities with precise ground truth from active perception, enabling navigation in extremely low-light off-road environments. The sensor suite is resilient to environmental degradation like dust, smoke, fog, snow, and rain, and can be calibrated in a single step for effective off-road navigation.

B. Related Datasets

A few existing datasets provide a variety of sensor modalities and ground truth data, enabling the development and benchmarking of algorithms in areas such as SLAM, object recognition, and autonomous navigation (Table I): MVSEC [28] is the first dataset that synchronizes stereo event cameras and provides accurate ground truth depth from LiDAR and SLAM and ground truth pose using a motion capture system and GPS; UZH-FPV [29] dataset utilized fast, aggressive, and agile drones to capture event camera data for extreme motion scenarios, but does not contain depth information; For night and day place recognition tasks, Maddern and Vidas [30] built a capture platform consisting of GPS, RGB camera, and thermal camera to capture data from before dawn to after dusk; The KAIST Multi-Spectral Day/Night Dataset [31] introduced a sensor system designed for SLAM, comprising stereo RGB cameras, LiDAR, and thermal camera; Aiming at off-road environments such as forests and urban areas, M3ED [32] used high resolution stereo event cameras, grayscale and RGB cameras, IMU, LiDAR, and RTK localization to collect a high-speed dynamic motion dataset; ViViD++ [8] is the first dataset to feature aligned information from multiple types of alternative vision sensors, including RGB, thermal, event, depth, and inertial measurements. Compared to existing datasets, our M2P2 dataset is the first dataset that focuses on off-road mobility in extremely low-light environments with the most perception modalities and highest sensor quality, as well as a precise multi-modal calibration procedure with accurate synchronization (see Table I for comparison).

III. Multi-Modal Sensor Suite

Our multi-modal sensor suite comprises a thermal and an event camera, stereo RGB cameras, two IMUs, GPS, and LiDAR for ground truth. All sensors are assembled in a custom-designed 3D-printed structure, which can be easily mounted on most mobile robot platforms (Fig. 2). The total dimensions of the sensor suite are $0.31 \times 0.26 \times 0.24$ m, with a total weight of 2 kg.

TABLE I: Comparison with alternative vision datasets.

Dataset	Sensor Modality							Hardware	Environments	Lighting
	RGB	Depth	Thermal	Event	LiDAR	IMU	GPS			
ViViD++ [8]	✓	✓	✓	✓	✓	✓	✓	Vehicle	Indoor/Urban	Day/Night
DiTer++ [9]	✓	✓	✓	✗	✓	✓	✓	Legged	Diverse Terrain	Day/Night
TartanDrive 2.0 [10]	✓	✓	✗	✗	✓	✓	✓	Wheeled	Off-road	Day
M2P2	✓	✓	✓	✓	✓	✓	✓	Wheeled	Off-road	Day/Night

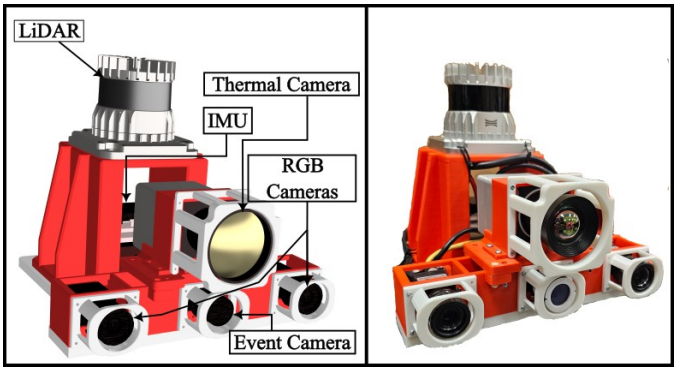

Fig. 2: Sensor Suite CAD (Left) and Hardware (Right).

A. Thermal Camera

Our sensor suite includes a Xenics Ceres T 1280 thermal camera, which features Long Wave Infrared (LWIR) imaging at a high resolution of 1280×1024. The camera can capture images at a maximum of 45 FPS via the GigE Vision interface. The thermal camera is paired with a wide-angle lens of 11 mm with 71.7° Horizontal Field of View (HFoV), 58.9° Vertical FoV (VFoV), and an aperture of f/1.2. Notice that our wide-angle LWIR camera provides the highest quality thermal images compared to any existing open-source datasets.

B. Event Camera

We use a Prophesee Metavision EVK4 as our event camera. The camera has a latency of 220 μs within a compact size with a sensor resolution of 1280×720. We use a lens with 46.8° HFoV and 36° VFoV with an aperture range from f/2-11 (fixed at f/4.0). The camera has a time resolution equivalent to 10K FPS and a low-light cutoff of 0.08 lx. To prevent LiDAR pulses from introducing noisy events, we apply an IR filter in front of the event camera lens.

C. Stereo RGB Cameras

We use two FLIR Blackfly S cameras for capturing images in the RGB spectrum. The cameras have a resolution of 1616×1240, which can be captured at a maximum of 175 FPS (fixed at 10 FPS).

While our stereo RGB cameras fail to perceive in no-light conditions, they can still perceive in environments featuring only partial degradation or with some ambient lighting.

D. IMUs

We use a Yahboom 10-DoF IMU featuring a 3-axis accelerometer, 3-axis gyroscope, 3-axis magnetometer, and a barometer. The sample rate of the IMU is 200 Hz. It features built-in data fusion and gyro stabilization.

We also include the IMU embedded in the LiDAR (see details below).

E. LiDAR for Ground Truth

A 3D Ouster OS1-128 LiDAR is used to provide ground truth with 128 lines of vertical divisions in 45° VFoV and selectable 512, 1024, and 2048 angle divisions in 360° HFoV at 10/20 Hz. For best data efficiency, LiDAR point clouds are recorded with 1024 angle divisions at 10 Hz. The LiDAR also features a built-in 6-DoF IMU with a 125 Hz sample rate for LiDAR frame calibration.

IV. Sensor Suite Calibration

To understand how the multi-modal perception streams from the sensor suite transform real-world features in world coordinates into their corresponding sensor readings, as well as how they correlate with each other in terms of a common coordinate system, we develop a streamlined multi-modal calibration procedure to calibrate all the sensors with different modalities in the sensor suite.

Traditional calibration methods use distances measured by geometric features, such as a printed black and white checkerboard with squares of known sizes for camera intrinsics and camera-to-camera extrinsics calibration, or a flat surface for LiDAR-to-camera extrinsics calibration. However, for our multi-modal sensor suite, those methods pose a limitation as conventional calibration targets are not visible in the infrared range of a thermal camera. Furthermore, static calibration targets are not visible by an event camera, which needs motion to detect the changes in intensity.

Therefore, our multi-modal sensor suite requires a common calibration target that can be perceived by all sensors as to calibrate both intrinsic and extrinsic parameters.

A. Thermal Checkerboard

The first challenge of calibrating our sensor suite comes from the thermal camera, which requires different thermal signatures to reflect distances of geometric features. To introduce a contrast thermal signature, we create a calibration target using an aluminum sheet of 3 mm thickness and carbon fiber squares of 35 mm. The sheet and the carbon fiber squares are precision milled with CNC achieving an accuracy of 0.05 mm. Since the aluminum sheet is highly reflective in the long wave infrared (IR) spectrum (similar to a mirror in the visible spectrum), we anodize the aluminum sheet to eliminate

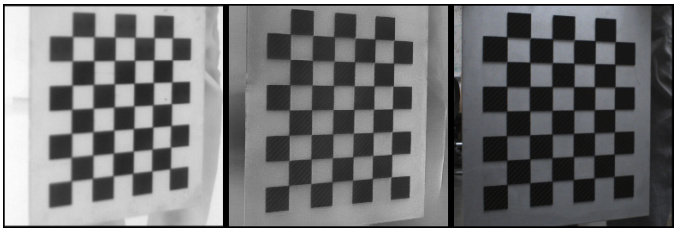

Fig. 3: Calibration Target (Thermal, Event, and RGB Image).

unwanted reflection in the IR spectrum. After heating the calibration target to roughly 45°C, due to a large difference in emissivity of aluminum and carbon fiber, the checkerboard pattern appears in the thermal image (Fig. 3 left). Due to the contrast in color of aluminum and carbon fiber, the same pattern is visible in both RGB cameras (Fig. 3 right).

B. Event Reconstruction

To address the second calibration challenge of correlating asynchronous event data with other synchronous data streams, such as thermal and RGB images, we employ a two-step approach. First, we reconstruct a grayscale image from the raw event stream using E2Calib [33] (Fig. 3 middle). Additionally, we utilize the trigger input functionality of the event camera to precisely mark timestamps for frame reconstruction, enabling accurate temporal alignment between the reconstructed event frames and corresponding frames from other sensors. This method allows us to overcome the inherent asynchronous nature of event data and establish reliable temporal relationship with synchronous data streams, facilitating multi-modal sensor fusion and calibration.

C. Multi-Modal Synchronization

With a common calibration target visible in all four cameras in the sensor suite, with another RGB camera in the stereo pair, the last calibration challenge is the precise synchronization among multiple asynchronous and un-synchronized data streams to achieve calibration convergence. To address this, we implement a synchronization scheme as illustrated in Fig. 4.

We synchronize all four cameras to the LiDAR, which generates a 10 Hz sync pulse aligned to its encoder angle at 360°. This pulse triggers frame acquisition in the RGB and thermal cameras, with its edges marking temporal points in the event camera stream. The pulse width matches the RGB camera's exposure time, and its falling edge is used for event camera frame reconstruction. This approach aligns the reconstructed frame with the RGB camera's exposure completion, ensuring precise temporal correlation across all sensors.

D. All-in-One Calibration Procedure

Finally, we splice all the synchronized frames and create a ROS-bag that can be used with any calibration toolkit. In our implementation, we use Kalibr [34] calibration toolkit to generate camera intrinsic and extrinsic parameters. Furthermore, we need to calibrate the camera and IMUs to complete the transformation tree for the entire sensor suite. As the

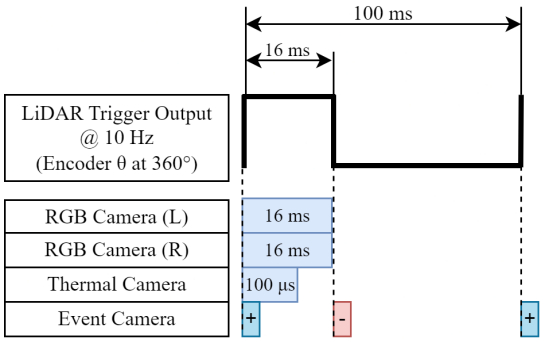

Fig. 4: Multi-Modal Synchronization: LiDAR trigger synchronized to internal encoder angle ($\theta = 360°$) initiates frame acquisition at a rate of 10 Hz for RGB and thermal cameras, with event camera recording trigger edges for frame reconstruction.

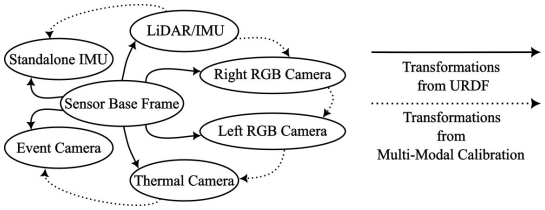

Fig. 5: Transformation Tree of the Sensor Suite: Solid arrows indicate direct hardware transformations, while dotted arrows represent transformations from our multi-modal calibration.

Ouster IMU features a 6-DoF IMU with factory-calibrated transformation from the LiDAR base to the IMU frame, we use the Ouster base as a reference frame to bind everything into a single tree. The entire transformation tree of the sensor suite from our multi-modal calibration, as well as from our hardware design, is shown in Fig. 5.

Fig. 6 shows the LiDAR point cloud overlaid on the corresponding RGB image, along with the reconstructed event frame and thermal image, demonstrating the spatial and temporal alignment of the multi-modal data.

V. Multi-Modal Passive Perception Dataset

M2P2 dataset encompasses over 10 hours of data collected across various challenging terrain conditions (Fig. 7). The data are gathered with the sensor suite mounted on a Clearpath Husky A200 robot. The dataset includes sequences from a diverse range of environments, progressing from fully prepared

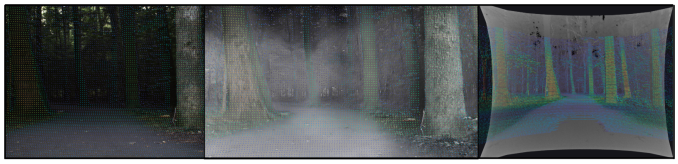

Fig. 6: Multi-modal data from the M2P2 dataset, showcasing spatial and temporal alignment in a low-light, off-road forest environment. LiDAR point cloud overlaid on RGB image (left), reconstructed event frame at the trigger's falling edge (middle), and thermal image (right).

Fig. 7: >32 km, 10.15-Hour M2P2 Data Collection across Different Locations: Maps show diverse environments including lakeside trails, urban parks, and dense forests, highlighting the variety of terrain and conditions captured in the dataset.

paved trails to non-paved off-road paths, and ultimately to unprepared off-trail environments within densely forested areas featuring thick vegetation and narrow passages. To capture a comprehensive range of lighting conditions, data collection is conducted at dusk, with luminosity levels varying from 20 lx to complete darkness (0 lx). This approach ensures the dataset's applicability to both well-lit and no-light scenarios, addressing the challenges of navigation in varying environmental conditions.

The dataset is structured as ROS-bag files, consisting of compressed RGB and thermal images at 10 FPS, asynchronous raw event stream, 3D point cloud data from LiDAR, IMU data, GPS coordinates, robot odometry and status messages, and human-commanded joystick inputs. All camera data are synchronized using the trigger pulse from the LiDAR, ensuring temporal alignment across multi-modal sensor inputs. Due to the dense canopy of the trees the GPS data is only available for 87.97% of the total dataset. However, it is possible to fuse LiDAR, IMU, and GPS, when available, with LIO-SAM [35], relying primarily on lidar-inertial odometry. Fig. 8 shows a LIO-SAM-generated map overlaid on a satellite image. The LiDAR point cloud aligns well with visible features (e.g., trail edges and vegetation), demonstrating mapping accuracy. The inset compares the estimated trajectory (blue) to raw GPS (green); the latter deviates significantly under dense tree cover, reflecting degraded signal quality, while the LIO-SAM trajectory remains consistently accurate.

To facilitate accurate sensor placement replication, we provide the URDFs (Unified Robotics Description Format) for the sensor suite configuration on the Husky platform, along

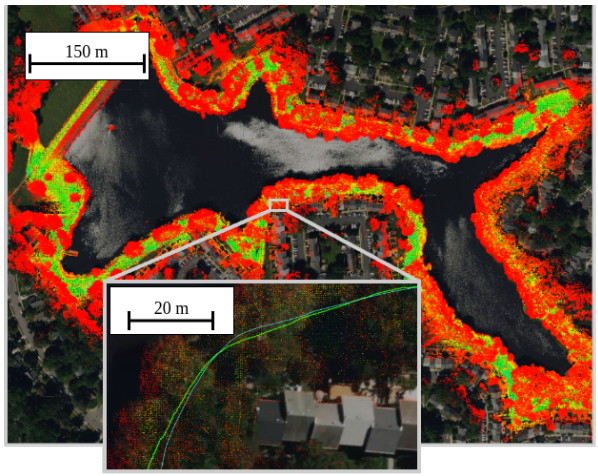

Fig. 8: LIO-SAM Mapping Results on Lake Braddock Trail: The LiDAR point cloud (colored points) is overlaid on a satellite image. Inset: Comparison of LIO-SAM estimated trajectory (blue) and raw GPS trajectory (green).

with the calibrated transformations. Table II shows the main statistics of the M2P2 dataset. The multi-modal synchronization scheme achieves near-perfect alignment between RGB images and LiDAR point clouds, with only six instances of mis-synchronization. The slightly reduced number of thermal images compared to RGB images is due to the thermal camera's automatic shutter calibration, which interrupts the image stream for approximately 0.4 seconds to correct for non-uniformities.

TABLE II: M2P2 Statistics

Attribute	Quantity
Total Size	≈2 TB
Total Distance	>32 km
Total Time	10.15 h
Total GPS Lock Time	8.93 h
Average Speed	0.95 m/s
Number of RGB Images	730606
Number of Thermal Images	361685
Number of Events	1.15×10^{11}
Number of Point Clouds	365297

TABLE III: Quantitative Depth Prediction Comparison on Unseen Data.

Model	#Params (M) ↓	Abs Rel ↓	RMSE ↓	δ_1 ↑
DepthAnythingV2	335.3	0.66	8.43	0.03
U-Net + M2P2	31	0.13	2.12	0.82

VI. Experiment Results

We conduct three experiments using our M2P2 dataset to demonstrate its usefulness in off-road navigation and perception under degraded lighting conditions.

A. End-to-End Navigation Learning

To demonstrate the effectiveness of the dataset to enable end-to-end learning for autonomous navigation, we train an end-to-end behavior cloning (BC) model that outputs linear and angular velocities [36, 37] based on thermal camera input into a ResNet-18. Considering the difference in absolute temperature, we normalize each pixel value based on the max and min values of the current thermal image to get the relative temperature readings. We deploy this BC model on the Husky robot for a 3.6 km autonomous navigation task on a paved hiking trail, as illustrated in Fig. 9. The luminosity during the experiment ranges from 235 lx to 0 lx (indicated by the color of the path), with the robot completing most of the navigation in complete darkness (0 lx). The robot successfully completes the navigation, requiring only 11 human interventions when it goes off-course. Most interventions are because the pavement and the gravel on the side show similar temperature in the thermal input and therefore confuse the robot. More sophisticated techniques that leverage other sensor modalities, e.g., event camera, are necessary to enable more robust navigation.

B. Perception in Degraded Visual Environments

To evaluate the efficacy of M2P2 in enabling scene perception in degraded visual environments, we conduct a comparative analysis of metric depth estimation. Specifically, we train a U-Net [38], 31M parameters, to learn a mapping between thermal infrared imagery and corresponding depth information derived from the LiDAR point clouds. We compare the performance of this U-Net, trained on the M2P2 dataset, against DepthAnythingV2-Large [39], a monocular metric depth estimation model with approximately 335.3 million parameters. Quantitative results, detailed in Table. III, reveal a substantial performance superiority of U-Net despite its significantly lower parameter count. Notably, DepthAnythingV2-Large demonstrates limited generalization to the infrared domain.

Fig. 9: Autonomous Navigation around a 3.6 km Trail with a BC model and Thermal Input: Lighting conditions drops from 255 lx at the beginning (light gray on the path, lower right) to 0 lx (black, upper left). 11 interventions (red crosses) are necessary to correct the robot when going off-course.

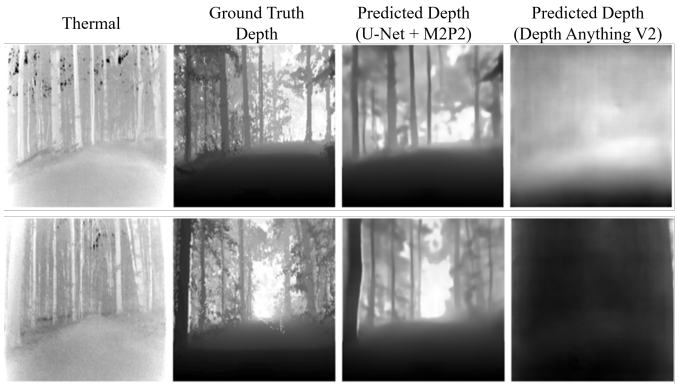

Fig. 10: Qualitative Depth Prediction Comparison on Unseen Data.

Qualitative evaluations, illustrated in Fig. 10, further reinforce these findings. Qualitative inspection confirms that the U-Net trained on M2P2 generates depth maps of considerably higher fidelity compared to those produced by DepthAnythingV2-Large. This observation highlights the pivotal role of domain-specific datasets like M2P2 in enabling the development of robust perception models for degraded visual environments, where traditional RGB-based methods are inherently challenged. Our results suggest that such datasets are indispensable for bridging the gap between standard visual perception and the complexities introduced by atypical sensory inputs.

C. Passive Visual Odometry with Thermal and Event Data

A unique characteristic of M2P2 is the inclusion of calibrated, synchronized thermal and event camera data, en-

TABLE IV: Translational ATE with Thermal-Event Fusion

Event Percentage	Translational ATE (m) ↓
100% (Full Event Data)	8.79
80%	11.60
50%	12.79
25%	12.49

abling exploration of passive perception in extremely low-light conditions. While prior work has investigated visual-inertial odometry using RGB and event cameras [40], the fusion of thermal and event data for odometry remains relatively underexplored. This combination holds significant promise for applications where visible light is scarce or unavailable, such as nighttime off-road navigation or covert operations. The closest existing work is RAMP-VO [40], and M2P2 helps advance this area of research.

To demonstrate the potential of this multi-modal fusion, we adapt the RAMP-VO framework, originally designed for RGB and event data, to process thermal and event data from M2P2. We focus on a challenging 157.5 m segment of the Burke Lake trail to evaluate the robustness of the approach under varying light levels.

Crucially, we simulate reduced lighting conditions by systematically subsampling the event stream. This allows us to assess the performance of the thermal-event odometry system as the available information from the event camera decreases. We experiment with retaining 80%, 50%, and 25% of the original events, representing progressively darker scenarios, in addition to using the full event data (100%).

Table IV presents the translational Absolute Trajectory Error (ATE) for each event subsampling level. As expected, the error generally increases as the event data becomes sparser.

VII. Conclusions and Future Work

This paper introduces M2P2, a novel multi-modal passive perception dataset specifically designed to address the challenges of off-road robot mobility in extreme low-light conditions. Unlike existing datasets, M2P2 uniquely combines thermal, event, and stereo RGB cameras, along with IMUs, GPS, and LiDAR for ground truth, providing a comprehensive representation of challenging off-road, low-light environments. We make the M2P2 dataset, along with our sensor suite design, publicly available to facilitate further research. We also present a robust multi-sensor calibration procedure, ensuring accurate data alignment across all modalities. Our initial experiments demonstrate that, even in complete darkness, off-road navigation, scene understanding, and vehicle state estimation are achievable using purely passive sensing.

While these initial experiments showcase the promise of individual modalities and limited fusion, the full realization of M2P2's potential requires deeper exploration of advanced sensor fusion techniques and their application to a wider range of mobility tasks. As the first step toward fully passive perception for off-road mobility in extreme low-light conditions, this work opens up a new avenue of future research. Some of the areas that could benefit from M2P2 include Visual Inertial Odometry [41, 42, 43], SLAM [44, 45, 46], and off-road kinodynamics modeling [47, 48, 49, 50, 51, 52], all with the purely passive modalities available from our multi-modal sensor suite and dataset.

References

[1] 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.

[2] J. Hooks, M. S. Ahn, J. Yu, X. Zhang, T. Zhu, H. Chae, and D. Hong, "Alphred: A multi-modal operations quadruped robot for package delivery applications," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5409–5416, 2020.

[3] L. Van Nguyen, S. Gibb, H. X. Pham, and H. M. La, "A mobile robot for automated civil infrastructure inspection and evaluation," in *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2018, pp. 1–6.

[4] L. F. Oliveira, A. P. Moreira, and M. F. Silva, "Advances in agriculture robotics: A state-of-the-art review and challenges ahead," *Robotics*, vol. 10, no. 2, p. 52, 2021.

[5] R. R. Murphy, *Disaster robotics*. MIT press, 2014.

[6] U. Wandinger, "Introduction to lidar," in *Lidar: range-resolved optical remote sensing of the atmosphere*. Springer, 2005, pp. 1–18.

[7] L. Li *et al.*, "Time-of-flight camera—an introduction," *Technical white paper*, no. SLOA190B, 2014.

[8] 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.

[9] J. Kim, H. Kim, S. Jeong, Y. Shin, and Y. Cho, "Diter++: Diverse terrain and multi-modal dataset for multi-robot slam in multi-session environments," *arXiv preprint arXiv:2412.05839*, 2024.

[10] M. Sivaprakasam, P. Maheshwari, M. G. Castro, S. Triest, M. Nye, S. Willits, A. Saba, W. Wang, and S. Scherer, "Tartandrive 2.0: More modalities and better infrastructure to further self-supervised learning research in off-road driving tasks," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 12 606–12 606.

[11] K. Ebadi, Y. Chang, M. Palieri, A. Stephens, A. Hatteland, E. Heiden, A. Thakur, N. Funabiki, B. Morrell, S. Wood *et al.*, "LAMP: Large-scale autonomous mapping and positioning for exploration of perceptually-degraded subterranean environments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 80–86.

[12] R. Thakker, N. Alatur, D. D. Fan, J. Tordesillas, M. Paton, K. Otsu, O. Toupet, and A.-a. Agha-mohammadi, "Autonomous off-road navigation over extreme terrains with perceptually-challenging conditions," in *Experi-*

mental Robotics: The 17th International Symposium. Springer, 2021, pp. 161–173.

[13] Y. Chang, K. Ebadi, C. E. Denniston, M. F. Ginting, A. Rosinol, A. Reinke, M. Palieri, J. Shi, A. Chatterjee, B. Morrell, A.-a. Agha-mohammadi, and L. Carlone, "LAMP 2.0: A robust multi-robot slam system for operation in challenging large-scale underground environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9175–9182, 2022.

[14] M. Wermelinger, P. Fankhauser, R. Diethelm, P. Krüsi, R. Siegwart, and M. Hutter, "Navigation planning for legged robots in challenging terrain," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 1184–1189.

[15] L. Sharma, M. Everett, D. Lee, X. Cai, P. Osteen, and J. P. How, "RAMP: A risk-aware mapping and planning pipeline for fast off-road ground robot navigation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5730–5736.

[16] C. Chung, G. Georgakis, P. Spieler, C. Padgett, A. Agha, and S. Khattak, "Pixel to elevation: Learning to predict elevation maps at long range using images for autonomous offroad navigation," *IEEE Robotics and Automation Letters*, 2024.

[17] 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.

[18] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3019–3026, 2018.

[19] 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)*. IEEE, 2019, pp. 5000–5007.

[20] X. Meng, N. Hatch, A. Lambert, A. Li, N. Wagener, M. Schmittle, J. Lee, W. Yuan, Z. Chen, S. Deng *et al.*, "Terrainnet: Visual modeling of complex terrain for high-speed, off-road navigation," *arXiv preprint arXiv:2303.15771*, 2023.

[21] P. Sermanet, R. Hadsell, M. Scoffier, U. Muller, and Y. LeCun, "Mapping and planning under uncertainty in mobile robots with long-range perception," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 2525–2530.

[22] M. Bajracharya, J. Ma, M. Malchano, A. Perkins, A. A. Rizzi, and L. Matthies, "High fidelity day/night stereo mapping with vegetation and negative obstacle detection for vision-in-the-loop walking," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 3663–3670.

[23] 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.

[24] 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)*. IEEE, 2022, pp. 2931–2937.

[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] S. Jeong, H. Kim, and Y. Cho, "Diter: Diverse terrain and multi-modal dataset for field robot navigation in outdoor environments," *IEEE Sensors Letters*, vol. PP, pp. 1–4, 03 2024.

[27] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "Rellis-3d dataset: Data, benchmarks and analysis," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 1110–1116.

[28] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, "The multivehicle stereo event camera dataset: An event camera dataset for 3d perception," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018.

[29] J. Delmerico, T. Cieslewski, H. Rebecq, M. Faessler, and D. Scaramuzza, "Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6713–6719.

[30] W. Maddern and S. Vidas, "Towards robust night and day place recognition using visible and thermal imaging," in *Proceedings of the RSS 2012 Workshop: Beyond laser and vision: Alternative sensing techniques for robotic perception*. University of Sydney, 2012, pp. 1–6.

[31] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[32] K. Chaney, F. Cladera, Z. Wang, A. Bisulco, M. A. Hsieh, C. Korpela, V. Kumar, C. J. Taylor, and K. Daniilidis, "M3ed: Multi-robot, multi-sensor, multi-environment event dataset," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 4016–4023.

[33] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, "How to calibrate your event camera," in *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, June 2021.

[34] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 1280–1286.

[35] 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.

[36] 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.

[37] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[39] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[40] R. Pellerito, M. Cannici, D. Gehrig, J. Belhadj, O. Dubois-Matra, M. Casasco, and D. Scaramuzza, "Deep visual odometry with events and frames," in *IEEE/RSJ International Conference on Intelligent Robots (IROS)*, June 2024.

[41] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7244–7251.

[42] Z. Huai and G. Huang, "Robocentric visual–inertial odometry," *The International Journal of Robotics Research*, vol. 41, no. 7, pp. 667–689, 2022.

[43] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct ekf-based approach," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 298–304.

[44] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[45] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[46] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual slam algorithms: A survey from 2010 to 2016," *IPSJ transactions on computer vision and applications*, vol. 9, pp. 1–11, 2017.

[47] 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.

[48] 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.

[49] P. Atreya, H. Karnan, K. S. Sikand, X. Xiao, S. Rabiee, and J. Biswas, "High-speed accurate robot control using learned forward kinodynamics and non-linear least squares optimization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 789–11 795.

[50] 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.

[51] 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.

[52] 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.