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

AnyThermal: Towards Learning Universal Representations for Thermal Perception

anythermal.github.io

Parv Maheshwari[1], Jay Karhade*[1], Yogesh Chawla*[2], Isaiah Adu[3], Florian Heisen[4], Andrew Porco[5], Andrew Jong[1], Yifei Liu[1], Santosh Pitla[2], Sebastian Scherer[1], Wenshan Wang[1]

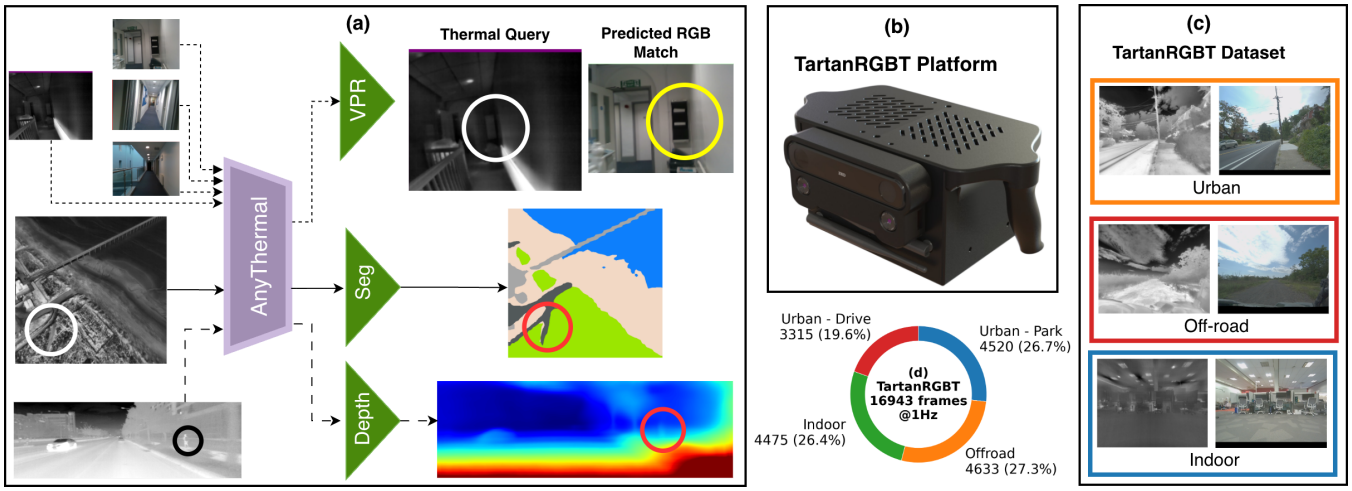

Fig. 1. **AnyThermal** is a task-agnostic thermal encoder that delivers state-of-the-art performance across diverse tasks—such as cross-modal place recognition, thermal segmentation, and monocular thermal depth estimation—and can be applied to a wide range of environments, including indoor, aerial, off-road, and urban settings. To bridge the existing data diversity gap for training AnyThermal, we build (b) an open-source data collection platform and introduce (c) **TartanRGBT**, a synchronized RGB-T dataset that spans over four types of diverse environments, as shown in (d) with a balanced distribution and a total of 16943 RGB-T pairs.

***Abstract*— We present AnyThermal, a thermal backbone that captures robust task-agnostic thermal features suitable for a variety of tasks such as cross-modal place recognition, thermal segmentation, and monocular depth estimation from thermal images. Existing thermal backbones that follow task-specific training from small-scale data result in utility limited to a specific environment and task. Unlike prior methods, AnyThermal can be used for a wide range of environments (indoor, aerial, off-road, urban) and tasks, all without task-specific training. Our key insight is to distill the feature representations from visual foundation models such as DINOv2 into a thermal encoder using thermal data from these multiple environments. To bridge the diversity gap of the existing RGB-Thermal datasets, we introduce the TartanRGBT platform, the first open-source data collection platform with synced RGB-Thermal image acquisition. We use this payload to collect the TartanRGBT dataset - a diverse and balanced dataset collected in 4 environments. We demonstrate the efficacy of AnyThermal and TartanRGBT, achieving state-of-the-art results with improvements of up to 36% across diverse environments and downstream tasks on existing datasets.**

* Equal contribution
[1] Authors are with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. {parvm, jkarhade, ajong, yifeil5, basti, wenshanw}@andrew.cmu.edu
[2] Authors are with Biological Systems Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA. {ychawla2, spitla}@nebraska.edu
[3] Authors are with Mechanical Engineering, Penn State University, University Park, PA, USA. ioa5099@psu.edu
[4] Authors are with the School of Engineering and Design, Technical University of Munich, Munich, Germany. florian.heisen@tum.com
[5] Authors are with Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. aporco@andrew.cmu.edu

I. INTRODUCTION

The utility of thermal images has been well explored in the context of robot perception in degraded environments [1]–[4]. Unlike RGB sensors that are sensitive to lighting conditions and weather changes, thermal imagery is robust to all these challenges, making it a necessary addition for resilient autonomy in scenarios like search and rescue, autonomous driving, and surveillance.

However, unlike RGB images, thermal images suffer from a scarcity of data. While RGB benefits from Internet-scale repositories that have driven major advances in deep learning [5], [6], no such large-scale resource exists for thermal data. As a result, thermal feature extractors have yet to benefit from training at scale. Consequently, many works adapt pre-trained RGB backbones with task-specific objectives [3], [7], [8]. In this work, we show that RGB-only backbones fail to capture thermal-specific cues and that training the feature-extraction backbone on thermal images for task-agnostic training yields substantially stronger representations.

Since thermal datasets are scarce, a promising approach to

improving thermal models is distilling knowledge from pretrained RGB models [9]. This leverages both the diversity of large-scale RGB data and the correspondence between RGB and thermal views of the same scene. Effective knowledge distillation, even in data-constrained domains, requires sufficient data diversity [10]. However, prior work has been limited to a single dataset from a single environment [9], restricting its generality. In this paper, we address this limitation by combining RGB-T datasets from diverse domains for distillation, and show that the resulting backbone achieves state-of-the-art performance on thermal segmentation, cross-modal place recognition, and thermal depth estimation.

While several RGB-T datasets exist, most are confined to a single type of environment (Table I). To advance knowledge distillation for thermal images, there is a clear need for RGB-T datasets spanning multiple environments. To bridge this gap, we collect a new dataset across multiple environments and demonstrate that our diverse dataset can further amplify the gains achieved from distillation.

We summarize our main contributions as follows:

- **AnyThermal**: A task-agnostic feature extractor for thermal images via cross-modal knowledge distillation. Combined with lightweight heads, it achieves state-of-the-art performance across environments on tasks like thermal segmentation and cross-modal place recognition, while outperforming similar-sized RGB backbones in thermal monocular depth estimation.
- **TartanRGBT Platform**: The first open-source hardware and software suite for synchronized stereo RGB and stereo thermal data collection.
- **TartanRGBT dataset**: Using the TartanRGBT platform, we collect a diverse, balanced data set which covers residential areas, campuses, indoor environments, off-road terrain, parks, and trails. We also show how this dataset can further boost AnyThermal's performance in various thermal downstream tasks across environments.

We open-source AnyThermal and TartanRGBT (models, code, platform, dataset) at https://anythermal.github.io

II. Related Works

A. Thermal Images for Robot Perception

Thermal perception has advanced across odometry [11]–[13], place recognition [3], segmentation [2], [14], [15], and depth estimation [1], [7], [16], [17]. However, annotated thermal data remains scarce. To mitigate this, recent strategies leverage RGB priors via domain adaptation [3], [18], weight initialization [15], or cross-modal supervision [17]. While effective, these methods are often limited to single tasks or narrow domains. In contrast, AnyThermal serves as a universal thermal encoder, demonstrating robustness across diverse tasks and environments.

B. Multi-modal Foundation Models

Foundation models [5], [6], [19] leverage large-scale pretraining for generalized representations. Unlike the task-specific methods in Section II-A, knowledge distillation targets general-purpose backbones by mimicking teacher feature spaces. This strategy has successfully produced robust backbones for depth, and LiDAR [20], [21]. While ImageBind [9] used this for thermal data, its single-environment training is brittle. AnyThermal leverages multi-environment distillation to learn a robust, task-agnostic backbone.

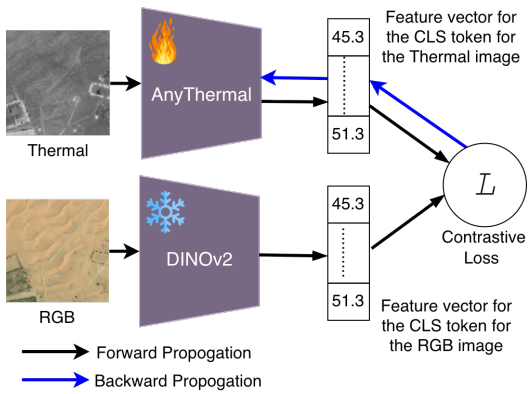

Fig. 2. We distill a frozen DINOv2 teacher into a DINOv2-initialized AnyThermal student. This initialization promotes multi-environment generalization, while distillation extracts rich thermal representations. Our choice of a task-agnostic, self-supervised objective over thermal and RGB features eliminates the need for annotated data, thereby scaling effectively with the available RGB-T data.

C. RGB-T Datasets

Existing RGB-T datasets cover urban [1], [11], [13], [22], indoor [12], aerial [2], [3], and off-road [2], [23] settings, yet most are restricted to a single environment (Table I). Furthermore, non-standardized hardware hinders scalable data collection. Given the limitations of current thermal simulation [24], progress of thermal perception relies on diverse, real-world data. To lower this barrier, we open-source the TartanRGBT platform and our diverse TartanRGBT dataset.

III. AnyThermal: Thermal Encoder

A. Overview

AnyThermal is a DINOv2-based model that has undergone knowledge distillation for thermal images. To improve generalizability across domains, the distillation is done by combining multiple datasets across domains (urban, aerial, indoor, off-road). Moreover, similar to DINOv2, we show that using AnyThermal as a task-agnostic feature extractor combined with a task-specific head can lead to state-of-the-art performance on tasks like thermal segmentation, cross-modal place recognition, and monocular depth estimation.

B. Knowledge Distillation

To perform knowledge distillation, two DINOv2 ViT-B/14 encoders are used. Both are initialized with pretrained weights. The teacher network processes RGB images and is kept frozen, while the student processes thermal images and is trainable (Fig. 2). To use DINOv2 encoders with thermal images, the images are converted from grayscale to 3-channel. After distillation, the student serves as our AnyThermal model.

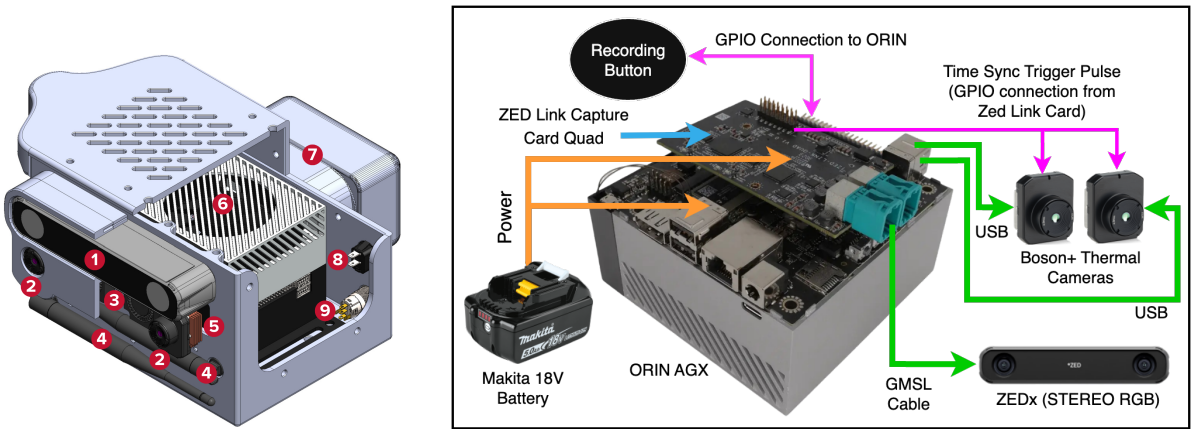

Fig. 3. **Left:** CAD of the TartanRGBT system (casing partially hidden): (1) ZED X stereo camera; (2) Teledyne FLIR Boson 640 LWIR camera; (3) 5V blower fan; (4) Wi-Fi antennae; (5) copper heat sinks; (6) NVIDIA Jetson AGX Orin (64GB); (7) 18V Makita battery; (8) power switch; (9) recording button. **Right:** System diagram, showing transfer of power (orange), sensor data (green), and signal (pink) — time synchronization and recording trigger

For RGB–thermal knowledge distillation, we apply InfoNCE loss [25] on the [CLS] tokens, leveraging the intuition that corresponding image pairs share global semantic features. Since DINOv2 [CLS] features prioritize semantic information over low-level cues like color [5], they provide a robust foundation for cross-modal alignment. Unlike patch-level losses that require pixel-perfect registration, a global contrastive objective relaxes the need for exact RGB–thermal alignment or synchronization. This is advantageous when distilling from datasets like ViVID++ and STheREo, where precise time-sync or spatial alignment is unavailable.

We train AnyThermal on five datasets across diverse domains: ViVID++ [11] (Urban), STheREo [13] (Urban), Freiburg [22] (Urban), Boson Nighttime [3] (Aerial), and TartanRGBT(urban, indoor, and off-road).

Other datasets such as MS^2 [1], CART [2], and OdomBeyondVision [12] are reserved for zero-shot evaluation on downstream tasks. M2P2, despite its large number of off-road sequences, is excluded from training AnyThermal because many sequences have poor visibility, which weakens RGB teacher features and hampers effective thermal distillation.

C. Task-Specific Head and Training

We use task-specific heads to adapt AnyThermal's generalized features to the requirements of each downstream task (segmentation, VPR, depth estimation).

D. Cross-Modal Place Recognition

A cross-modal place recognition task is to find a positive match in a database (D) of the modality A for a query (q) of modality B. Similar to [3], we use thermal queries, and a corresponding RGB database. For each training dataset, ground-truth positives are defined by a fixed, environment-specific threshold — small for indoor, larger for urban and aerial — applied as a geographical radius when GPS/odometry is available, or a temporal (frame) radius otherwise.

For VPR, methods like SALAD [26] and SGM [3] show benefits of pairing a generalized feature extractor [5], [27] with a specialized head (NetVLAD [28], SALAD). We choose SALAD due to its higher recall compared to other VPR heads [26].

Following [3], we train with a triplet margin loss [29], where each triplet (a, p, n) consists of an anchor (RGB or thermal image), a positive, and a negative. All datasets used for knowledge distillation also train the VPR head, ensuring robust clustering across environments. Unlike distillation, VPR training uses intra-dataset sampling to form harder, visually similar triplets for more effective learning.

E. Thermal Segmentation

Following DINOv2, we ablated several lightweight heads (1-layer MLP, 2-layer non-linear MLP, and DPT) for segmentation, ultimately selecting the two-layer non-linear MLP for AnyThermal. It takes patch features of size $(H/14 \times W/14 \times 768)$ from DINOv2 and outputs a mask $(H/14 \times W/14 \times C)$ for C classes, which is then upsampled and compared with the ground truth. The backbone remains frozen throughout training, with only the task head optimized using Dice loss [30], which empirically outperformed cross-entropy loss. To mitigate data scarcity, we apply augmentations including brightness, contrast, gamma, and horizontal flipping.

F. Mono-Thermal Depth Estimation

For monocular depth estimation, we use the training and evaluation code for MiDaS [31] framework from [7], which originally uses an EfficientLite3 backbone. We replace this with ViT-based backbones (frozen DINOv2 or AnyThermal), using multiscale patch features from different layers to mimic EfficientNet3's hierarchical features. The rest of the MiDaS architecture remains unchanged.

IV. TartanRGBT Platform

To collect RGB-T pairs in diverse environments, we have designed a data collection platform - TartanRGBT platform, as shown in Fig. 3, which comprises a compute module (NVIDIA Orin AGX 64GB), an 18V Makita battery, a ZEDx camera (stereo RGB + IMU), a ZEDx quad link capture card, and two FLIR Boson 640+ cameras (stereo-thermal). The

TABLE I
COMPARISON OF RGB-T DATASETS ACROSS SENSING MODALITIES, SYNCHRONISATION, AND ENVIRONMENTS.

Dataset	Plat.	# RGB-T Pairs @1Hz[a]	RGB[b]	THR[b]	Sync	Environment				
						Indoor	Offroad	Aerial	U-Drive	U-Park
MS^2 [1]	V	16215	S	S	✓	✗	✗	✗	✓	✗
ViVID++ [11]	H/V	14824	M	M	✓	✗[c]	✗	✗	✓	✗
STheReO [13]	V	8393	S	S[d]	✗	✗	✗	✗	✓	✗
CART [2]	H/D	9678	M	M	✓	✗	✓	✓	✗	✗
Boson-Nighttime [3]	D	52590/N[e]	M	M	✗	✗	✗	✓	✗	✗
OdomBeyondVision [12]	D/G/H	7129	S	M	✗	✓	✗	✗	✗	✗
M2P2 [23]	G	34362	S	M	✓	✗	✓	✗	✗	✓
Ours (TartanRGBT)	H	16943	S	S	✓	✓	✓	✗	✓	✓

[a] Number of frames is considered at 1Hz to ensure non-redundancy of data in knowledge-distillation.
[b] S: Stereo camera , M: Monocular camera
[c] While VIVID++ contains some indoor sequences, all of them are in a VICON cage and hence not diverse even for an indoor dataset
[d] The stereo thermal pair is not timesynced
[e] The frequency of thermal capture is not specified. So the N is unknown

Platform abbreviations: V = Vehicle, H = Handheld, D = Drone/UAV, G = UGV. Combinations (e.g., H/V, U/G/H) indicate multiple platforms. Sync = hardware synchronization. U-Drive and U-Park denote urban driving (campus, road, residential areas) and park environments, respectively. As shown, our dataset is the most diverse while also providing synced and stereo RGB–thermal pairs.

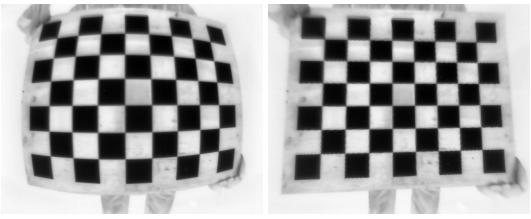

Fig. 4. Thermal calibration: original (left) and rectified (right).

sensors are hardware time-synchronized and capture images at 30Hz.

A. CAD design and 3D printing

The payload, as shown in Fig. 3, is housed in a custom 3D-printed case with ergonomic handles on the top and sides for ease of use. Each thermal camera has heatsinks and an active cooling fan to maintain stable operation. The enclosure provides access to external ports and includes air vents to ensure airflow around the onboard computer.

B. Time Syncing

The four-camera suite (2 RGB, 2 thermal) is hardware-synchronized at 30 FPS using a Master-Slave configuration. The factory-synced ZED X generates a trigger pulse via the ZED Link Capture Card; this signal drives the thermal cameras through their external sync pins, ensuring thermal frames are captured in lockstep with the RGB frames.

C. Calibration

A complete calibration of all cameras requires intrinsic, distortion, and extrinsic factors between the cameras. Factory calibration of the stereo RGB pair was used to retrieve the intrinsics and distortion coefficients of each RGB camera. To calibrate the intrinsics and distortion parameters of the thermal cameras, a custom heated checkerboard was used similar to [32]. The results after thermal rectification can be seen in Fig. 4 . The extrinsics between the RGB and thermal cameras were retrieved from the CAD design.

D. Data Collection Procedure

For ease of data collection, the payload auto-launches all software drivers (cameras, ROS2 recording, and GPIO trigger detection) via Docker at startup, eliminating manual setup. A dedicated hardware button enables one-click start/stop of recordings, and external WiFi antennas provide remote access to the ORIN.

E. Open-Source

The TartanRGBT platform is designed with off-the-shelf components to facilitate accessible RGB-T data collection. We provide the full CAD files, software stack (Docker, drivers), and assembly guides at anythermal.github.io.

V. TartanRGBT Dataset

A. Data Distribution

As shown in Table I and Fig. 1(d), TartanRGBT (collected in Pittsburgh, USA) is the first dataset to provide broad environmental diversity alongside high-quality, time-synced, and registered RGB–T imagery. While moderate in size, its emphasis on diverse environments makes it uniquely effective for knowledge distillation, outperforming larger single-domain datasets as shown in Section VI-D.

B. Modalities

The TartanRGBT platform records stereo RGB, stereo thermal, and IMU data. RGB-T frames during thermal Flat Field Correction (FFC) are filtered to account for thermal capture pauses. We register RGB-thermal pairs using stereo-estimated depth. To support VPR training (Sec. III-D), MAC-VO [33] odometry provides spatial supervision for cross-traversal pair mining. Both raw and derived modalities (depth and odometry) are included in the public release.

C. Thermal 8-bit Processing

To convert a 16-bit raw thermal image to an 8-bit image, similar to [16], we apply the following in sequence: Min-Max normalization, CLAHE, and BilateralFilter.

D. RGB-Thermal Image Registration

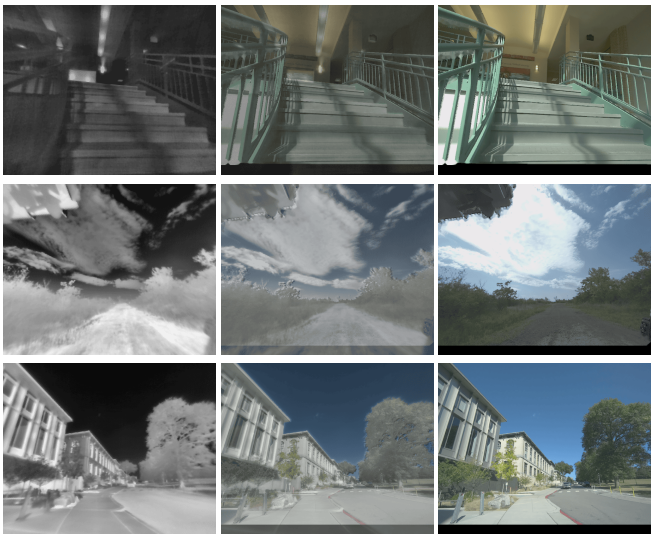

Fig. 5. **RGB–Thermal Registration** in the TartanRGBT dataset: alpha-blended overlays for indoor, off-road, and urban domains with blending factors $\alpha \in \{0.00, 0.50, 1.00\}$. The thermal camera's lower mounting position than the RGB camera causes it to capture more of the lower scene, creating a black strip of pixels at the bottom in the thermal-aligned view of the RGB image.

RGB–thermal image registration provides the spatial correspondence necessary to map teacher-side visual representations to student-side thermal representations during cross-modal knowledge distillation. Following HeatNet [22], our alignment has three stages: (1) Back-project pixels to 3D via stereo-RGB depth estimation; (2) transform 3D points into the thermal frame with pre-calibrated extrinsic; (3) project with thermal intrinsics to yield aligned RGB–thermal pairs (Fig. 5).

Similar to HeatNet [22], which employed the state-of-the-art stereo model of its time for dense depth estimation, we adopt FoundationStereo [34] to obtain dense depth. Although the estimated depth is not perfect and errors in prediction directly affect the aligned outputs, it offers a practical alternative to accurate but sparse LiDAR, as knowledge distillation requires dense supervision. Furthermore, as highlighted in Section III-B, our choice of a global contrastive loss alleviates the need for pixel-level alignment between the RGB-Thermal pair during distillation.

FoundationStereo produces a dense depth map, but during RGB–thermal alignment, black pixels arise from occlusions between the two views and from rasterizing 3D points onto discrete thermal pixels, leaving some locations unfilled. We address this with two steps. First, a z-buffer enforces visibility by retaining only the nearest depth per thermal pixel. Second, after projection to 2D, bilinear splatting improves coverage by distributing each projected sample across its four neighboring pixels with interpolation weights. As shown in Fig. 5, splatting is not applied in the lower regions of the thermal images where no RGB depth is available, as this would otherwise hallucinate content without valid 3D data.

TABLE II

RECALL@1(%) FOR CROSS-MODAL PLACE RECOGNITION ACROSS DIVERSE ENVIRONMENTS.

Model Name	Backbone	Head[a]	MS^2 r: 15	CART r: 15	OBV[b] r: 3
*DINOv2 [5]	DINOv2	X	27.21	25.98	29.49
*SALAD [26]	DINOv2	S	76.97	49.38	38.94
*ImageBind [9]	ViT-Huge	X	0.79	1.13	10.25
*SGM [3]	ResNet-18	N	20.02	45.59	21.05
AnyThermal	AnyThermal	X	75.39	45.45	45.40
AnyThermal-VPR	AnyThermal	S	**81.11**	**56.00**	**53.17**

[a] VPR Heads: **N:** NetVLAD, **S:** SALAD, **X:** No head has been used, and instead the CLS token is used as the feature vector for the images
[b] OBV: OdomBeyondVision [12]

Environment-specific radii r define positive matches:(MS^2: Urban, CART: Aerial, OBV: Indoor). We denote frozen models with * and distinguish between RGB-only and RGB–thermal training. AnyThermal belongs to the latter, using RGB-initialization followed by thermal distillation. AnyThermal consistently outperforms baselines, with its gains over *DINOv2 validating the efficacy of our RGB-to-thermal distillation.

E. Limitations

We have released dense depth and odometry to support RGB–thermal alignment and VPR training. As they are obtained from stereo-RGB algorithms, their accuracy is insufficient for benchmarking tasks such as odometry or depth estimation. Thus, we do not evaluate downstream tasks like cross-modal place recognition or depth estimation on TartanRGBT. Since VPR training does not require precise odometry, the current estimates suffice. Future work will include GPS and LiDAR for accurate odometry and depth.

VI. Results

We demonstrate the effectiveness of AnyThermal on three tasks: cross-modal place recognition, thermal segmentation, and monocular thermal depth estimation.

A. Cross-Modal Place Recognition

1) Formulation: Our cross-modal place recognition task, as described in Section III-D, is defined as: given a thermal query image, retrieve a matching RGB image from a database. To ensure proper evaluation, the paired RGB image of a query is excluded from its positive set. We report Recall@1 (R@1) in Table II, where R@1 is the probability that the top retrieved match is positive for a query.

2) Evaluation Datasets: We evaluate AnyThermal and baselines on three diverse zero-shot datasets: CART [2] (aerial), MS2 [1] (urban), and OdomBeyondVision [12] (indoor). CART and MS2 provide GPS, enabling all sequences to form a shared database, while OdomBeyondVision relies on intra-sequence odometry and is hence evaluated per sequence. For OdomBeyondVision, a weighted mean recall is reported across sequences, weighted by the number of

3) Baselines: We compare against two categories:

- *RGB Methods:* NetVLAD [28], MixVPR [35], and SALAD [26]. We report SALAD as the representative baseline due to its superior performance. Frozen RGB-DINOv2 (teacher) without a VPR head is also included

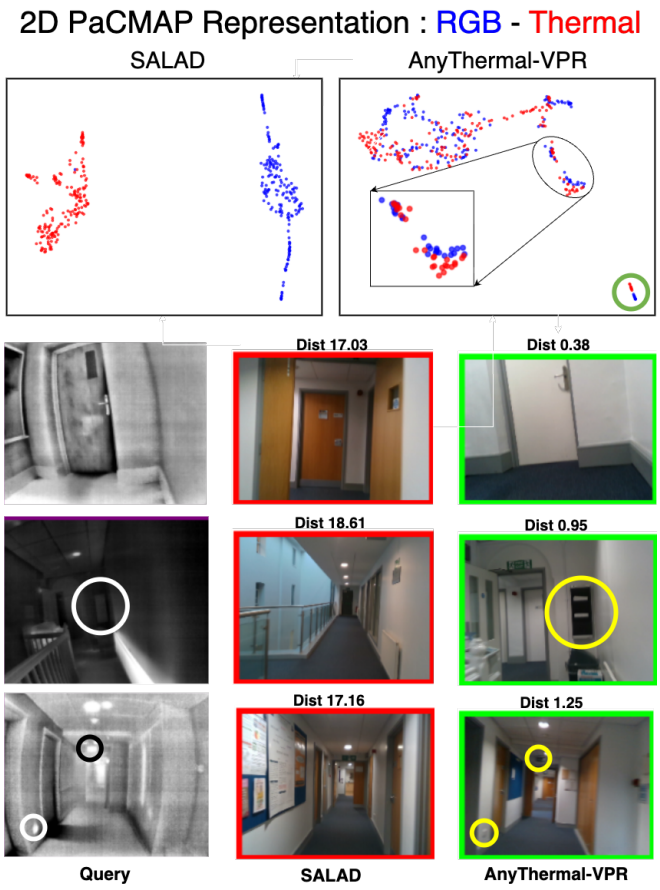

Fig. 6. **Cross-Modal VPR on OdomBeyondVision: Top**: PaCMAP [36] representations show SALAD poorly(far) aligns RGB–Thermal embeddings, while AnyThermal-VPR aligns them well in a shared representation space. **Bottom**: Example queries where SALAD fails to retrieve the correct RGB match, but AnyThermal-VPR succeeds, with key clues circled.

to isolate the impact of our knowledge distillation from the VPR head training.

- *RGB-Thermal Methods:* ImageBind [9] and SGM [3]. Although ImageBind is not trained for VPR, we include it since it is the only other method to perform knowledge distillation between RGB and thermal. SGM is trained for cross-modal place recognition, but only on Boson Nightime [3], which is an aerial-only dataset.

As shown in Table II, AnyThermal-VPR outperforms all baselines across environments. Moreover, the gap between DINOv2-X and AnyThermal-X underscores the need for knowledge distillation in thermal images and confirms that frozen RGB extractors are suboptimal. This is further evidenced by AnyThermal matching SGM's aerial performance without a VPR head, despite both being trained solely on the Boson Nighttime dataset for aerial data. Fig. 6 further illustrates that AnyThermal-VPR aligns RGB and thermal representations more effectively than the strongest baseline.

B. Thermal Segmentation

We evaluated the use of AnyThermal for thermal segmentation (Fig. 7) on the MF-Net [14] dataset using its standard

TABLE III

THERMAL SEGMENTATION ON MF-NET DATASET:

Model	# parameters(M)	mIoU (%)	FPS
RTFNET-152 [37]	196.37	47.00%	8.37
MCNET [8]	54.65	51.95	1.88
RGB_DINO-SEG	87.02	45.46%	6.79
AnyThermal-SEG	87.02	**53.47%**	6.79

The number of parameters is reported in Millions (M). The FPS is reported on ORIN AGX 64GB. We can see, AnyThermal with a 2-layer MLP head (SEG) achieves state-of-the-art performance while being 3.6x faster than the closest performing baseline

TABLE IV

MONOCULAR DEPTH ESTIMATION ON THE MS^2 DATASET

Backbone	AbsRel↓	SqRel↓	RMSE↓	RMSElog↓
efficientnet lite3	0.1015	0.3955	2.9587	0.1417
dinov2_vitb14	0.0905	0.3177	2.7493	0.1208
AnyThermal	**0.0883**	**0.3142**	**2.7432**	**0.1182**

We evaluate our proposed method with a representative MDE network (MiDaS [31]). All results are averaged over all day, night, and rainy evaluation sets of MS^2. The best performance is highlighted in **bold**.

train/val/test splits and all 9 classes (including background) for mIoU. Table III also reports FPS on an NVIDIA ORIN AGX 64GB. AnyThermal achieves state-of-the-art mIoU while delivering a 3.6× FPS boost over the closest baseline.

C. Mono-Thermal Depth Estimation

Following [7], we evaluate on the MS^2 dataset using sparse LiDAR ground truth and report multiple metrics (Table IV). We use the MIDAS [31] architecture, where we ablate the effect of replacing the EfficientLite3 backbone used in [7] with frozen DINOv2, and further with AnyThermal. The gain from EfficientLite3 to DINOv2 reflects impact of network depth, while the additional improvement with AnyThermal proves its benefits over frozen-RGB pretrained backbones.

D. Scaling Data in AnyThermal training

It is crucial to understand how multi-domain datasets in knowledge distillation affect downstream performance. Specifically, we ask whether simply adding more data improves efficacy, or if dataset diversity is essential for building robust feature extraction backbones.

We study the effect of data scaling during pre-training by training the AnyThermal backbone on progressively larger dataset combinations. For each backbone, task-specific heads are then trained separately: the VPR head is trained on the same datasets as the backbone, leveraging GPS/odometry or temporal cues for self-supervision. Segmentation and depth heads, which require labeled data absent from the pre-training datasets, are trained on the training splits of their respective task-specific benchmarks and evaluated on the test splits. This ensures a fair comparison: VPR baselines are evaluated zero-shot, while segmentation and depth baselines were trained on the same data as our task-specific heads.

As shown in Fig. 8, adding more datasets generally improves performance but not always:

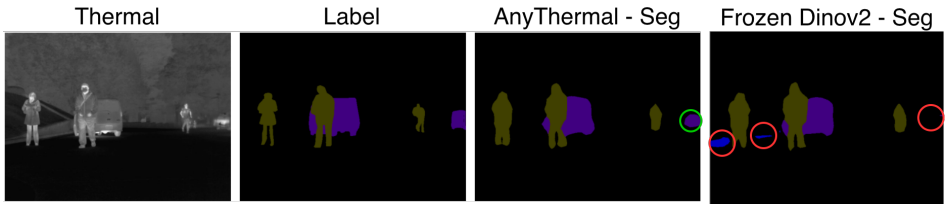

Fig. 7. **Thermal Segmentation on MF-Net [14]:** The frozen DINOv2 baseline misses objects (e.g., the car on the right) and misclassifies the background, while our AnyThermal backbone segments accurately. Red and green circles highlight incorrect and correct classification, respectively.

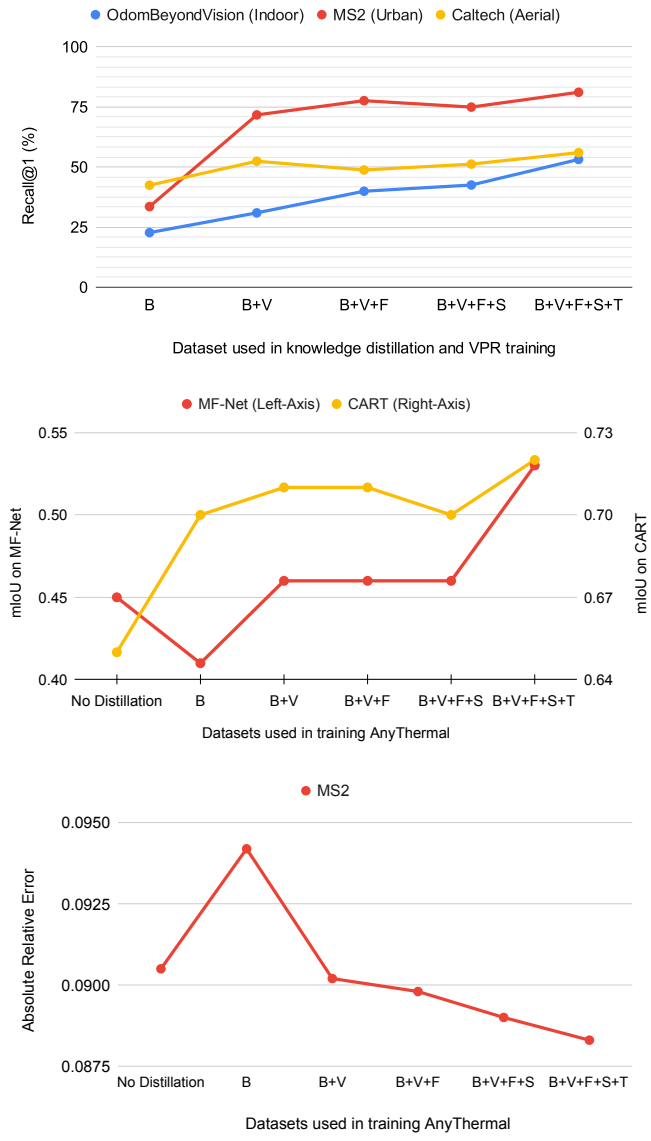

Fig. 8. Effect of scaling data in pretraining - knowledge distillation + VPR training(for top plot only)- on downstream performance. X-axis shows pretraining datasets (B: Boson Nightime, V: ViVID++, F: Freiburg, S: STheReO, T: TartanRGBT). [Top]: Recall for cross-modal place recognition (higher is better). [Middle]: mIoU for thermal segmentation on MF-Net and CART (higher is better). [Bottom]: Absolute relative (Abs_Rel) error for monocular thermal depth estimation (lower is better). Adding TartanRGBT consistently improves performance across environments and tasks, unlike Freiburg and STheReO, which add little diversity and lead to saturation.

- **Domain Gap in Single-Dataset Distillation:** In Fig. 8 (middle, bottom), a AnyThermal variant distilled only on Boson Nighttime (aerial) underperforms in urban domains (red), compared to the frozen RGB-DINOv2 (No distillation). This gap arises from its aerial-only training. Conversely, performance improves on CART (middle, yellow), as it is also aerial.
- **Performance Saturation:** In Fig. 8, adding more urban data (B+V $\rightarrow$ B+V+F $\rightarrow$ B+V+F+S) yields only marginal gains, with some aerial evaluations showing drops (e.g., thermal segmentation dip between $B+V+F$ and $B+V+F+S$).

In contrast, TartanRGBT consistently improves performance across tasks and domains. Notable gains include indoor VPR recall and improved CART segmentation, driven by our rich indoor and off-road sequences. It even boosts urban performance, suggesting TartanRGBT captures unique features absent from existing urban datasets (ViVID++, Freiburg, and STheReO).

These results show that while scaling data helps up to a point, data diversity is more critical than scale for building robust, generalizable feature extractors.

VII. Conclusion and Future Work

We present AnyThermal, a task-agnostic thermal feature extraction backbone distilled from pre-trained RGB backbones. To further advance thermal research, we introduced the TartanRGBT Platform—the first open-source RGB-T collection framework—and curated a diverse TartanRGBT dataset. Together, AnyThermal and TartanRGBT deliver up to 36% improvement across environments (urban, indoor, aerial, off-road) and tasks (cross-modal place recognition, thermal segmentation, depth estimation).

Future directions can include A) applying AnyThermal to more diverse tasks such as object detection and cross-modal matching, and B) distilling stronger backbones leveraging newer visual foundation models [6]. As shown in Fig. 8, AnyThermal's performance has not yet plateaued, suggesting further gains through scaling diverse RGB-T data. Future efforts will focus on (i) expanding TartanRGBT with additional sensors and environments (e.g., GPS, aerial); and (ii) community-driven data collection with our platform to advance generalization of thermal and cross-modal algorithms.

Acknowledgments

The authors would like to express their gratitude to the members of AirLab at Carnegie Mellon University for

their valuable insights and support throughout this work. This work was supported by Defense Science and Technology Agency (DSTA) Contract #DST000EC124000205 and DSO National Laboratories (DSO) Contract #DSOCO25020. This work used Bridges-2 at PSC through allocation cis220039p from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #213296.

REFERENCES

[1] U. Shin, J. Park, and I.-S. Kweon, "Deep depth estimation from thermal image," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1043–1053, 2023.

[2] C. T. Lee, M. Anderson, N. Raganathan, X. Zuo, K. Do, G. Gkioxari *et al.*, "Cart: Caltech aerial rgb-thermal dataset in the wild," in *European Conference on Computer Vision*, 2024.

[3] J. Xiao, D. Tortei, E. Roura, and G. Loianno, "Long-range uav thermal geo-localization with satellite imagery," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5820–5827, 2023.

[4] S. Hwang, J. Park, N. Kim, Y. Choi, and I.-S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1037–1045, 2015.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. Q. Vo, M. Szafraniec, V. Khalidov *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[6] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose *et al.*, "Dinov3," 2025.

[7] U. Shin, K. Lee, and J. Oh, "Bridging spectral-wise and multi-spectral depth estimation via geometry-guided contrastive learning," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6299–6305, 2025.

[8] H. Xiong, W. Cai, and Q. Liu, "Mcnet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene," *Infrared Physics & Technology*, vol. 113, p. 103628, 2021.

[9] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin *et al.*, "Imagebind one embedding space to bind them all," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 180–15 190, 2023.

[10] L. Frank and J. Davis, "What makes a good dataset for knowledge distillation*f*," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23 755–23 764, 2025.

[11] 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.

[12] P. Li, K. Cai, M. R. U. Saputra, Z. Dai, and C. X. Lu, "Odombeyondvision: An indoor multi-modal multi-platform odometry dataset beyond the visible spectrum," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3845–3850.

[13] S. Yun, M. J. Jung, J.-M. Kim, S. Jung, Y. Cho, M.-H. Jeon *et al.*, "Sthereo: Stereo thermal dataset for research in odometry and mapping," *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3857–3864, 2022.

[14] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115, 2017.

[15] C. T. Lee, J. G. Frennert, L. Gan, M. Anderson, and S.-J. Chung, "Online self-supervised thermal water segmentation for aerial vehicles," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7734–7741, 2023.

[16] D. Dhrafani, Y. Liu, A. Jong, U. Shin, Y. He, T. Harp *et al.*, "Firestereo: Forest infrared stereo dataset for uas depth perception in visually degraded environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3302–3309, 2025.

[17] X. Zuo, N. Ranganathan, C. T. Lee, G. Gkioxari, and S.-J. Chung, "Monother-depth: Enhancing thermal depth estimation via confidence-aware distillation," *IEEE Robotics and Automation Letters*, vol. 10, pp. 2830–2837, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276039463

[18] L. Gan, C. T. Lee, and S.-J. Chung, "Unsupervised rgb-to-thermal domain adaptation via multi-domain attention network," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6014–6020, 2022.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[20] G. Puy, S. Gidaris, A. Boulch, O. Siméoni, C. Sautier, P. Pérez *et al.*, "Three pillars improving vision foundation model distillation for lidar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 519–21 529.

[21] J. Karhade, "Towards universal place recognition," Master's thesis, Carnegie Mellon University, Pittsburgh, PA, August 2024.

[22] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8461–8468, 2020.

[23] A. Datar, A. Pokhrel, M. Nazeri, M. B. Rao, C. Pan, Y. Zhang *et al.*, "M2p2: A multi-modal passive perception dataset for off-road mobility in extreme low-light conditions," *ArXiv*, vol. abs/2410.01105, 2024.

[24] A. Upadhyay, M. Sharma, P. Mukherjee, A. Singhal, and B. Lall, "A comprehensive survey on synthetic infrared image synthesis," *ArXiv*, vol. abs/2408.06868, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271860142

[25] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49670925

[26] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 658–17 668, 2023.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[28] R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2015.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2015, p. 815–823.

[30] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, *Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations*. Springer International Publishing, 2017, p. 240–248.

[31] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.

[32] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9441–9447.

[33] Y. Qiu, Y. Chen, Z. Zhang, W. Wang, and S. A. Scherer, "Mac-vo: Metrics-aware covariance for learning-based stereo visual odometry mac-vo.github.io," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3814, 2024.

[34] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. T. Birchfield, "Foundationstereo: Zero-shot stereo matching," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260, 2025.

[35] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Mixvpr: Feature mixing for visual place recognition," *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2997–3006, 2023.

[36] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Journal of Machine Learning Research*, vol. 22, no. 201, pp. 1–73, 2021.

[37] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.