AnyThermal: Towards Learning Universal Representations for Thermal Perception

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

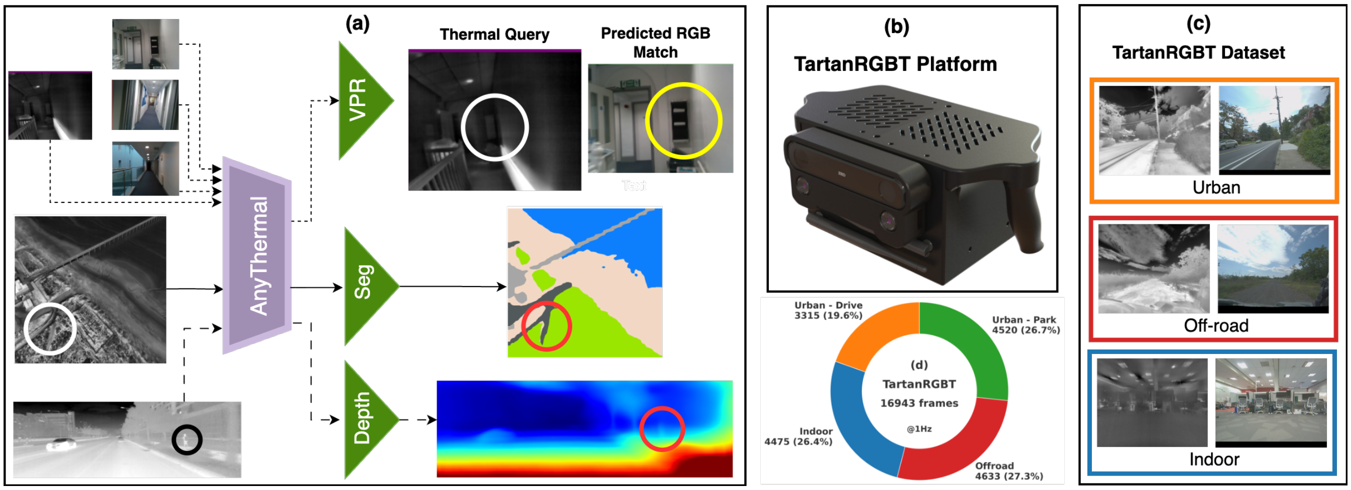

Fig. 1. **AnyThermal** is a task-agnostic thermal encoder that delivers state-of-the-art performance across diverse tasks—such as cross-modal place recognition, thermal segmentation, and monocular thermal depth estimation—and can be applied to a wide range of environments, including indoor, aerial, off-road, 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 using 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**

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]–[7], 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], [8], [9]. In this work, we show that such RGB-only backbones fail to capture thermal-specific cues, and that using thermal images for task-agnostic training of the feature extraction backbone yields substantially stronger representations.

Since thermal datasets are scarce, a promising approach to improving thermal models is distilling knowledge from pre-trained RGB models [10]. This leverages both the diversity of large-scale RGB data and the correspondence between

* Equal contribution

[1] Authors are with Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. {parvm, jkarhade, fheisen, 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 Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. aporco@andrew.cmu.edu

RGB and thermal views of the same scene. Effective knowledge distillation, even in data-constrained domains, requires sufficient data diversity [11]. However, prior work has been limited to a single dataset from a single environment [10], 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 obtained through knowledge distillation between RGB and thermal images. We show that AnyThermal when combined with task-specific heads, achieves state-of-the-art performance across environments on downstream tasks like thermal segmentation and cross-modal place recognition, while outperforming RGB-based backbones of comparable size for tasks like monocular depth estimation using thermal images.
- TartanRGBT platform: an open source data collection platform for collecting simultaneously captured stereo RGB and stereo thermal images. To the authors' best knowledge, this is the first open-source data collection platform for thermal images.
- TartanRGBT dataset: we collect a diverse, balanced data set using the TartanRGBT platform. The dataset 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 will release the models and code for AnyThermal, and open-source the TartanRGBT platform along with the collected TartanRGBT dataset upon acceptance.

II. Related Works

A. Thermal Images for Robot Perception

Thermal images have been applied to odometry [12]–[14], cross-modal place recognition [3], segmentation [2], [15], detection [4], and depth estimation [1], [16] across environments including indoor [12], [13], aerial [2], [3], off-road [17], and urban [1], [4]. Although thermal algorithms have diverse applications, they are often studied in narrow tasks or domains, limiting their utility. In contrast, we evaluate AnyThermal across a wide sets of tasks and environments to showcase its robustness and utility as a thermal encoder.

B. Multi-modal Foundation Models

Foundation models [5], [6], [18] have shown that large-scale pretraining enables generalized vision and language backbones. This has motivated robotics to adopt them for other modalities, with works like [10] demonstrating distillation of visual models into non-visual domains and building multimodal representations. Distillation has proven effective for depth/lidar [19]–[22], improving tasks such as segmentation, classification, and place recognition. Success in transferring foundation model priors to the thermal domain [10] has been limited by the use of scarce and non-diverse datasets. With AnyThermal, we show that training on multiple datasets enables effective distillation of foundation model priors, given that the datasets are collectively diverse.

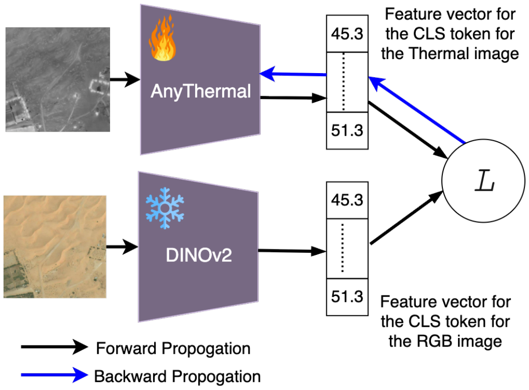
AnyThermal
45.3
51.3
Feature vector for the CLS token for the Thermal image
L
DINOv2
45.3
51.3
Feature vector for the CLS token for the RGB image
Forward Propogation
Backward Propogation

Fig. 2. We perform knowledge distillation between a frozen DINOv2 and a trainable DINOv2 network (AnyThermal), both initialized with pre-trained DINOv2 weights. The frozen network serves as the teacher, while the trainable AnyThermal backbone learns from it. Pre-trained initialization enables AnyThermal to generalize across environments, and distillation on thermal images allows it to extract meaningful thermal features. Training is task-agnostic, using self-supervised losses between thermal features from AnyThermal and RGB features from the frozen teacher. This approach requires no labels and scales naturally with increasing RGB-T datasets.

C. RGB-T Datasets

Recent RGB–T datasets span urban [1], [12], [14], [23], indoor [13], aerial [2], [3] and off-road [2], [17], yet most cover only a single environment (Table I). Moreover, each uses a distinct acquisition platform, and this non-standardization limits scalable, diverse collection. As realistic thermal simulation is not yet feasible, research progress with thermal images relies on real-world data collection, highlighting the need for community-driven efforts to collect data across environments and embodiments. To lower this barrier, we will be open-sourcing our TartanRGBT platform, whose efficacy is demonstrated through TartanRGBT dataset (Section VI-D).

III. AnyThermal: Thermal Feature-Extraction Backbone

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 feature extraction backbone 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.

For RGB–thermal knowledge distillation, we apply a contrastive loss on CLS token features, leveraging the intuition that corresponding RGB–thermal pairs should share similar global semantics. CLS token features from the final layer of DINOv2 capture semantic information [5], rather than low-level cues like color, making them a strong basis for alignment. Moreover, using contrastive loss on the CLS token, as compared to any form of patch loss (losses calculated on corresponding patches from the RGB–thermal pair), also relaxes constraints on RGB-thermal image alignment or exact synchronization. This is particularly advantageous when distilling using datasets like VIVID++ and STheReO, where perfectly aligned RGB–thermal pairs or precise time-sync are not available.

We used five datasets to train AnyThermal, distributed as:

- Urban: ViVID++ (Outdoor Driving Sequences) [12], STheREo [14], Freiburg [23] and TartanRGBT (ours)
- Aerial: Boson Nighttime Dataset [3]
- Indoor: TartanRGBT (ours)
- Offroad: TartanRGBT (ours)

Other datasets such as MS^2 [1], CART [2], and OdomBeyondVision [13] are reserved for zero-shot evaluation on downstream tasks. M2P2, despite its large size 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

As the feature descriptors from a ViT-based model can be quite large, they are combined with task-specific heads, which can be trained for a particular task like segmentation, visual place recognition (VPR), depth estimation, etc. In Section VI, we showcase how AnyThermal, when combined with task-specific heads, can lead to state-of-the-art performance on downstream tasks.

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. Moreover, for each training dataset, an environment-specific radius defines ground-truth positives, chosen as a geographical radius when odometry/GPS is available or a temporal(frame) radius otherwise.

For VPR, methods like SALAD [24] and SGM [3] show benefits of pairing a feature extractor [5], [25] with a specialized head (NetVLAD [26], SALAD). We choose SALAD due to its higher recall compared to other VPR heads [24].

Following [3], we train with a triplet margin loss [27], 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

As suggested in DINOv2 [5], ViT feature extractors (e.g., DINOv2, AnyThermal) can pair with lightweight heads for segmentation. After ablations with a single-layer MLP, two-layer non-linear MLP, and a DPT head, we select 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 upsampled and compared with the ground truth. The backbone remains frozen, training only the head with Dice loss [28], which outperformed cross-entropy. 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 [29] framework from [8], 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 sensors are hardware timesynced and capture data 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

In our platform, all four cameras (2 RGB, 2 thermal) are hardware-synchronized. The stereo RGB pair from the ZEDx is factory-synced, while a trigger pulse from the ZED Link Capture Card Quad synchronizes the thermal cameras. The pulse, aligned with RGB frame capture, is fed to both thermal

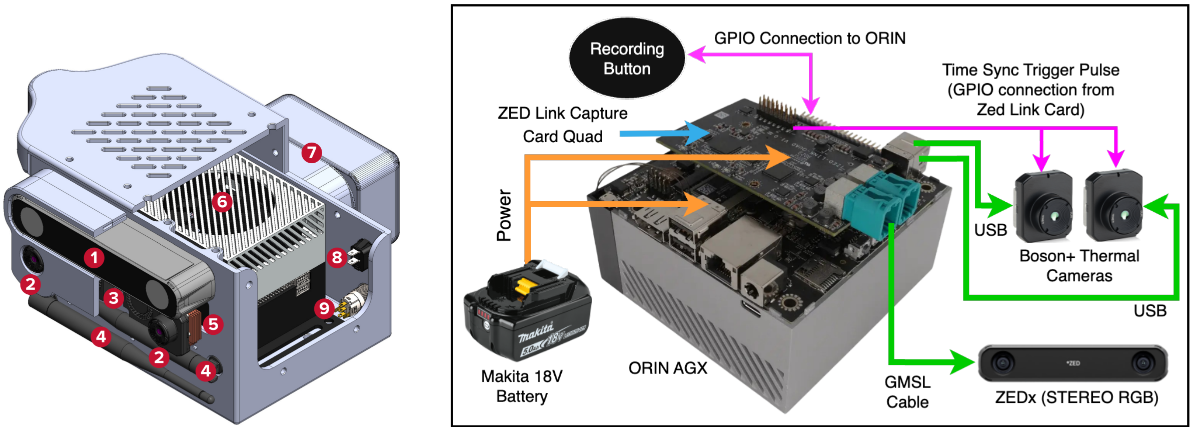

Fig. 3. **Left:** CAD model of the TartanRGBT system with half of the camera's and payload's casing hidden. Numbered components: (1) ZED X stereo camera; (2) Teledyne FLIR Boson 640 × 512, 4.9 mm, 95° HFoV, short-lens Shutterless LWIR thermal camera; (3) 5 V, 30 mm blower fan; (4) Wi-Fi antennae; (5) copper heat sinks (surrounding the thermal camera body); (6) NVIDIA Jetson AGX Orin Developer Kit, 64 GB; (7) Makita 18 V LXT® lithium-ion 4.0 Ah battery with adapter; (8) power switch; (9) recording button. **Right:** Overview of the connections between components, showing power (orange), sensor data transfer (green), and signal transfer(pink) —time synchronization and recording button trigger.

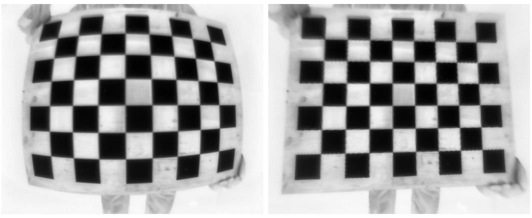

Fig. 4. Thermal checkerboard calibration image before (left) and after (right) fisheye rectification

cameras via their external sync pins. Configured in SLAVE mode with the BOSON SDK, the thermal cameras capture at 30 FPS in sync with the RGB cameras.

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 [30]. 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 sensor drivers (cameras, ROS2 recording, GPIO) via Docker at startup, eliminating manual setup. A hardware button enables one-click start/stop of recordings, and external WiFi antennas provide remote access to the ORIN.

E. Open-Source

In order to open-source the TartanRGBT platform, care has been taken to use easily available parts for assembly. Upon acceptance, we will release the CAD files, software stack (Docker, sensor drivers), component list, and assembly instructions. With this, our hope is to lower the entry barrier for the research community to collect RGB-T data.

V. TartanRGBT Dataset

A. Data Distribution

As shown in Table I, the TartanRGBT dataset is the first of its kind in offering broad environmental diversity alongside high-quality time-synced and registered RGB–thermal images. Although its size is moderate compared to other datasets, the emphasis on diversity during collection makes it impactful in knowledge distillation compare to existing datasets, as demonstrated in Section VI-D.

B. Modalities

Using the TartanRGBT platform, we record stereo RGB, stereo thermal, IMU, and thermal FFC status (manually triggered and timesynced). FFC frames are filtered since thermal capture pauses during calibration. To generate registered RGB-thermal pairs, we use FoundationStereo [31] for dense depth from stereo RGB, which will also be released. For training applications such as visual place recognition (Section III-D), we generate odometry using MAC-VO [32].

C. Thermal 8-bit Processing

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

D. RGB-Thermal Image Registration

Pixel-level RGB–thermal registration ensures spatial correspondence, improving distillation supervision and enabling tasks such as RGB–thermal translation, label transfer, and cross-modal learning. Existing aligned datasets ([2], [23], [15]) are limited and in singular environments, making our diverse dataset valuable for training and benchmarking.

Following [23], alignment has three stages: (1) estimate depth from rectified stereo RGB using FoundationStereo [31]

TABLE I

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

Dataset	Plat.	# RGB-T Pairs @1Hz[a]	RGB	THR	Sync	Reg.	Environment				
							Indoor	Offroad	Aerial	U-Drive	U-Park
MS^2 [1]	V	16215	S	S	✓	✗	✗	✗	✗	✓	✗
ViVID++ [12]	H/V	14824	M	M	✓	✗	✗[b]	✗	✗	✓	✗
STheReO [14]	V	8393	S	S[c]	✗	✗	✗	✗	✗	✓	✗
CART [2]	H/D	9678	M	M	✓	✓	✗	✓	✓	✗	✗
Boson-Nighttime [3]	D	52590/N[d]	M	M	✗	✓	✗	✗	✓	✗	✗
OdomBeyondVision [13]	D/G/H	7129	S	M	✗	✗	✓	✗	✗	✗	✗
M2P2 [17]	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] While VIVID++ contains some indoor sequences, all of them are in a VICON cage and hence not diverse even for an indoor dataset
[c] The stereo thermal pair is not timesynced
[d] 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. Reg. = registered (aligned) RGB–T pairs; 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 registered RGB–thermal pairs.

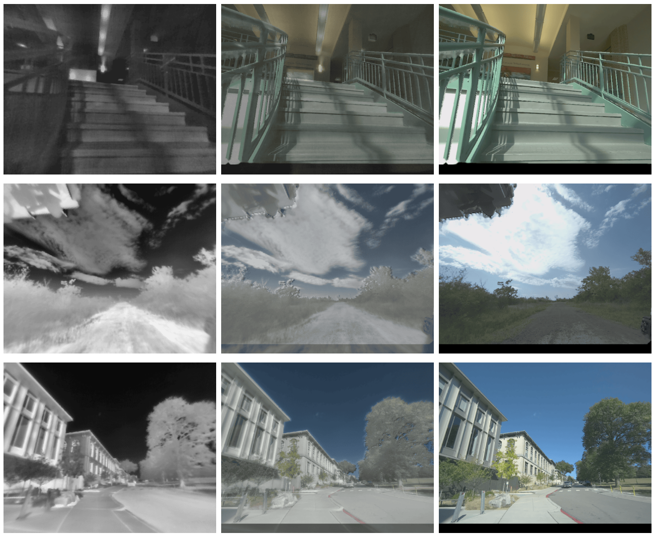

Fig. 5. **RGB–Thermal Registration** in the TartanRGBT dataset: alpha-blended overlays for indoor, off-road, and urban domains with blending factors $\alpha \in \{0.00, 0.50, 1.00\}$. Due to sensor geometry (thermal mounted below RGB), the thermal view includes more of the lower scene, resulting in additional pixels at the bottom of the thermal images that are not present in the RGB images, producing black regions where RGB pixels are absent.

to back-project pixels into 3D; (2) transform 3D points into the thermal frame with pre-calibrated extrinsics; (3) project with thermal intrinsics to yield aligned RGB–thermal pairs (Fig. 5).

Similar to [23], which employed the state-of-the-art stereo model of its time for dense depth estimation, we adopt FoundationStereo to obtain dense pixel-level alignment. 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, continued advances in depth estimation models are expected to further improve alignment quality.

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.

E. Limitations

We will release 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 also 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 [13] (indoor). CART and MS2 provide GPS, enabling all sequences to form a shared database, while OdomBeyondVision relies on intra-sequence odometry and is evaluated per sequence. For OdomBeyondVision, a weighted mean recall is reported across sequences, weighted by the number of queries in each.

3) Baselines: We compare against two categories:

- *RGB Methods:* R2former [33], NetVLAD [26], MixVPR [34], and SALAD [24]. Since SALAD consistently outperforms the others, we report it as the representative RGB baseline. We also include frozen RGB-DINOv2 (teacher) without a VPR head.
- *RGB-Thermal Methods:* ImageBind [10] 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.

TABLE II

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 [24]	DINOv2	S	76.97	49.38	38.94
*ImageBind [10]	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 [13]

The positive radius (r, in meters) used for determining positive matches, is chosen per environment (MS^2: Urban, CART: Aerial, OBV: Indoor). * denotes frozen models (backbone + head from original papers). Red indicates RGB-only training, while blue indicates RGB–thermal training. AnyThermal belongs to the blue category, as it is initialized with RGB-pretrained weights and distilled on thermal images. The upper section lists RGB-only methods, and the lower section lists RGB–thermal methods. AnyThermal, especially with a VPR head, outperforms baselines; The gap between *DINOv2 and AnyThermal shows the benefit of distilling RGB-pretrained backbones on thermal data.

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 [15] dataset using its standard 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.

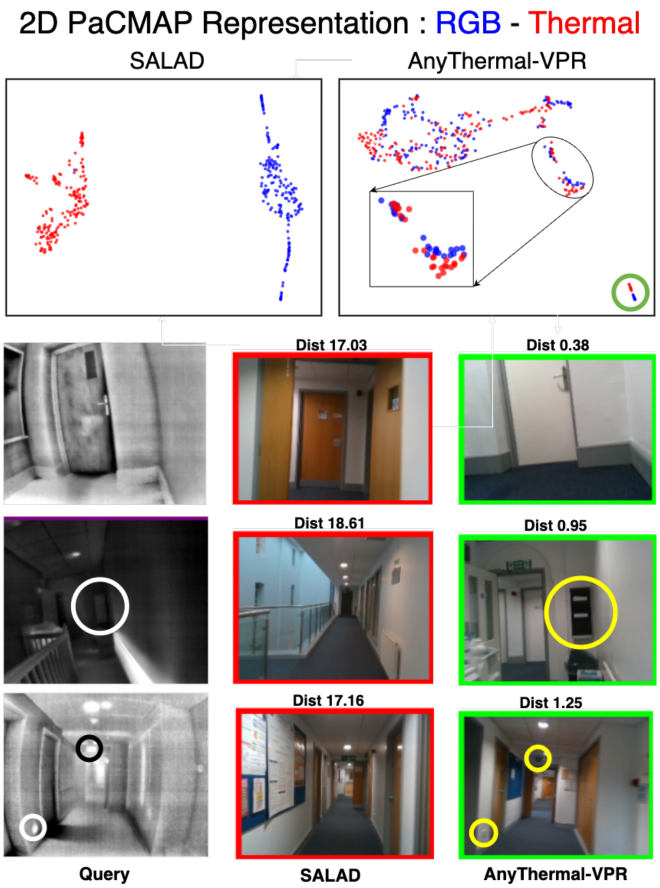

Fig. 6. **Cross-Modal VPR on OdomBeyondVision: Top**: PaCMAP [35] representations show SALAD poorly(far) aligns RGB–Thermal embeddings, while AnyThermal-VPR aligns them well in a shared representation space. **Bottom**: Example queries where SALAD fails to retrieve the correct RGB match, but AnyThermal-VPR succeeds, with key clues circled.

TABLE III

THERMAL SEGMENTATION ON MF-NET DATASET:

Model	# parameters(M)	mIoU (%)	FPS
RTFNET-152 [36]	196.37	47.00%	8.37
MCNET [9]	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

C. Mono-Thermal Depth Estimation

Following [8], we evaluate on the MS^2 dataset using sparse LiDAR ground truth and report multiple metrics (Table IV). We use the MIDAS [29] architecture, where we ablate the effect of replacing the EfficientLite3 backbone used in [8] with frozen DINOv2, and further with AnyThermal. The

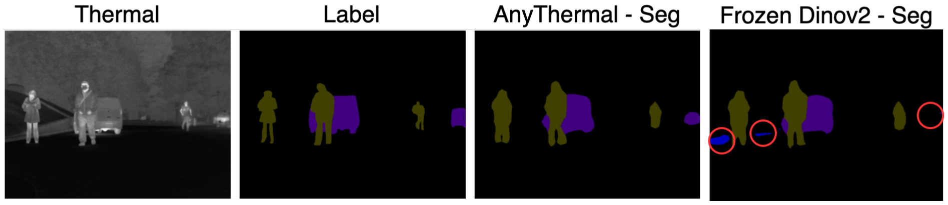

Fig. 7. **Thermal Segmentation on MF-Net [15]:** The frozen DINOv2 baseline misses objects (e.g., the car on the right) and misclassifies the background, while our AnyThermal backbone segments accurately.

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 [29]). All results are averaged over all day, night, and rainy evaluation sets of MS^2. The best performance is highlighted in **bold**.

gain from EfficientLite3 to DINOv2 reflects 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 distilling knowledge into the AnyThermal backbone and training the VPR head. Among task-specific heads, only VPR is included in pre-training, since it can leverage GPS/odometry or temporal cues and be evaluated zero-shot. In contrast, segmentation and depth require labeled data, so these tasks are trained and evaluated on the respective splits of their evaluation datasets. This setup ensures fairness: VPR baselines are evaluated zero-shot, while segmentation and depth baselines are trained on the evaluation datasets.

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

- **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 → B+V+F → B+V+F+S) yields only marginal gains, with some aerial evaluations showing drops (e.g., thermal segmentation dip between $B+V+F$ and $B+V+F+S$).

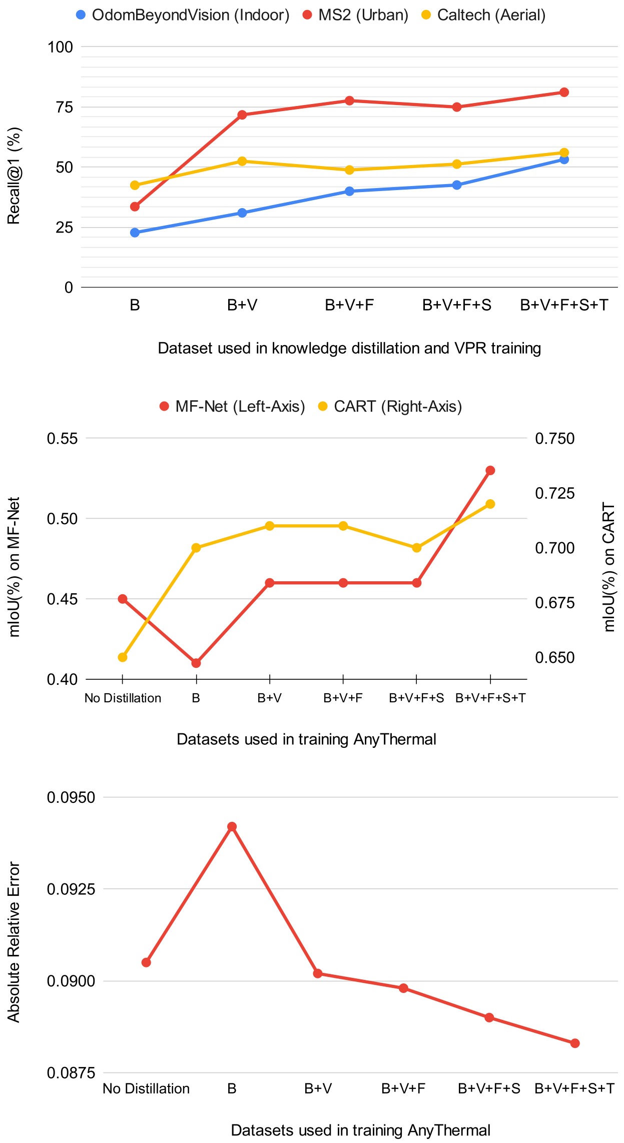

Fig. 8. Effect of scaling data in pretraining - knowledge distillation + VPR training(for top plot only)- on downstream performance. X-axis shows pretraining datasets (B: Boson Nightime, V: ViVID++, F: Freiburg, S: STheReO, T: TartanRGBT). [Top]: Recall for cross-modal place recognition (higher is better). [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.

In contrast, adding our TartanRGBT consistently improves performance across tasks and domains, with notable gains in indoor VPR recall (from rich indoor sequences), improved segmentation on CART (from off-road coverage since CART segmentation includes off-road data), and even boosts in urban domains despite existing urban datasets.

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.

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] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[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] 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.

[18] 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.

[19] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan *et al.*, "Segment any point cloud sequences by distilling vision foundation models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37 193–37 229, 2023.

[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] Y. Xia, Z. Li, Y.-J. Li, L. Shi, H. Cao, J. F. Henriques *et al.*, "Uniloc: Towards universal place recognition using any single modality," *arXiv preprint arXiv:2412.12079*, 2024.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "foundationstereo," in *CVPR*, 2025, pp. 5249–5260.

[32] 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.

[33] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 370–19 380.

[34] 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.

[35] 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.

[36] 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.