

Directional Navigability Estimation from Ultra-Low-Resolution Multi-View Thermal Sensing

Faiz Moazam*, Jana Hussien*, Abdulrahman Al-Taweel,
Gustavo Ramalho, Yaqoob Ansari, and Eduardo Feo-Flushing

Carnegie Mellon University

{fmoazam, jhussien, ataweel, gramalho, yansari, efeoflus}@andrew.cmu.edu

Abstract—Wearable navigation systems typically depend on high-resolution cameras, depth sensors, and large annotated datasets, a dependence that limits their deployability in privacy-sensitive indoor environments and raises development costs. This paper investigates whether directional navigability can be estimated solely from ultra-low-resolution thermal imagery captured by a wearable multi-camera platform. We propose a learning framework in which high-resolution RGB imagery, available only during data collection, automatically generates navigability labels through a pretrained object detector, eliminating manual annotation. Unlike conventional cross-modal approaches, the framework uses only coarse geometric alignment for automatic labeling and requires no geometric calibration at inference. Experiments with three 32×24 thermal cameras show that directional navigability is reliably inferred from thermal input alone and that feature-level fusion outperforms per-camera prediction as well as pixel- and decision-level fusion. A lightweight camera-identity classifier restores near-oracle accuracy when the input ordering changes, and every fusion architecture runs in under 3 ms per frame on CPU. These results suggest that ultra-low-resolution thermal sensing can serve as a scalable and privacy-preserving foundation for wearable indoor navigation.

I. INTRODUCTION

Wearable indoor navigation systems need reliable perception of navigable directions while staying inexpensive, privacy-preserving, and easy to deploy [1]. Most rely on RGB cameras, depth sensors, or LiDAR, which add hardware cost, raise privacy concerns, and need large annotated datasets, and multi-camera platforms often assume a sensor arrangement fixed at calibration [2].

Ultra-low-resolution thermal cameras offer an alternative that addresses many of these deployment constraints. A 32×24 thermal array is inexpensive, consumes little power, and does not capture identifiable appearance. Its low resolution, however, removes much of the information conventional perception relies on, while its limited field of view motivates using multiple cameras for coverage.

This paper investigates whether ultra-low-resolution multi-view thermal sensing provides sufficient information for reliable directional navigability estimation. As illustrated in Fig. 1, we use a high-resolution RGB camera only during data collection to automatically label synchronized thermal observations. At deployment, the system relies exclusively on thermal imagery, without requiring RGB input or geometric calibration.

*These authors contributed equally to this work.

This paper makes four contributions: a training framework that uses RGB supervision only during data collection, enabling thermal-only directional navigability estimation without manual annotation; a systematic comparison of fusion strategies at the pixel, feature, and decision levels for ultra-low-resolution thermal sensing; an evaluation of robustness to camera placement with a lightweight camera-identity correction that maintains performance under camera permutations; and a wearable multi-camera dataset of ultra-low-resolution thermal observations with their automatically generated labels, to support future research in privacy-preserving indoor navigation.

II. RELATED WORK

A. Wearable Navigation and Traversability Estimation

Existing wearable navigation systems estimate traversable space with RGB cameras [3], depth sensors, or LiDAR [4], inferring navigability from visual appearance [5] or explicit geometry. These sensors are accurate but hardware-heavy, and they raise privacy concerns for continuous wearable deployment [6].

B. Privacy-Preserving and Ultra-Low-Resolution Thermal Sensing

Thermal sensing largely avoids capturing identifiable imagery [6], [7]. DareDevil [8] uses a paired RGB camera and a pretrained YOLOv8 detector to label thermal frames automatically and then detects obstacles from thermal input alone, but it uses a single 640×480 thermal camera and binary obstacle detection, leaving directional navigability, multi-camera thermal perception, and ultra-low resolution unexplored. Ultra-low-resolution thermal arrays further improve privacy, cost, and power efficiency, but their limited spatial resolution has confined prior work to occupancy detection [6], human monitoring [9], and motion estimation, such as thermal-gyroscope odometry [10]. Whether such sensing alone can support directional navigability on a wearable multi-camera rig has not been studied.

C. Multi-View Sensor Fusion and Placement Robustness

Multi-view systems combine cameras to increase coverage and perception quality [11], [12], typically within a unified pipeline [13]. Where and how to combine the views varies across applications, and most systems assume a fixed camera configuration. That assumption is hard to maintain on wearable

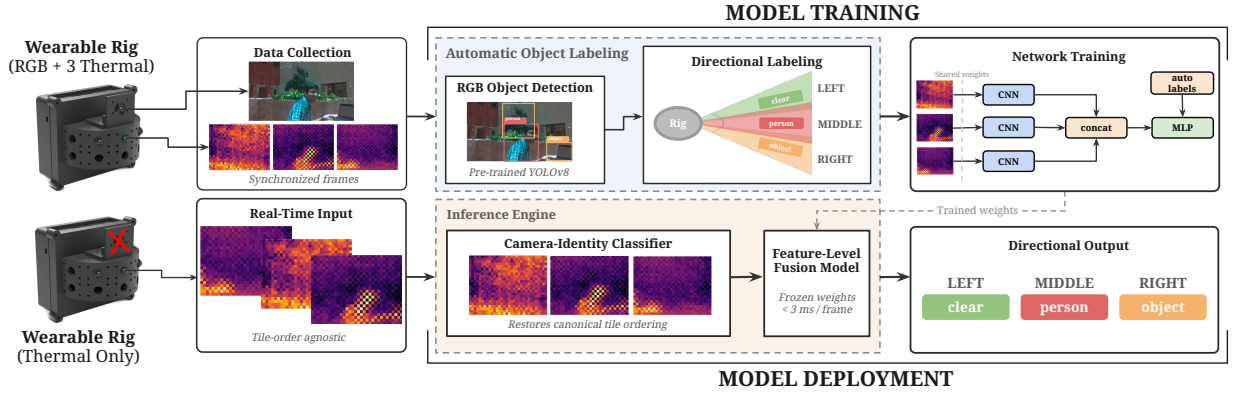


Fig. 1. End-to-end pipeline. Top: RGB/YOLOv8 auto-labels supervise the shared-weight CNN backbone with no manual annotation. Bottom: thermal-only tiles, re-sorted by the camera-identity classifier, feed the frozen Feature Fusion model for three-class directional output. Inferno colormap for display only.

platforms, where cameras are repositioned or reattached [2], and changing the mapping between cameras and network inputs can degrade fusion models that exploit camera-specific spatial relationships. We compare fusion strategies for ultra-low-resolution thermal navigability and evaluate their robustness to camera permutations.

III. METHODOLOGY

Fig. 1 provides an overview of the proposed framework, comprising two stages: RGB-assisted automatic labeling and model training, followed by thermal-only inference with camera-identity correction. The following sections describe the labeling process and the evaluated fusion architectures.

A. Problem Formulation

At each timestep the rig outputs three 32×24 thermal tiles $\{x_L, x_M, x_R\}$, one per active camera. For each direction $d \in \{\text{left}, \text{middle}, \text{right}\}$ the model predicts $y_d \in \{\text{clear}, \text{person}, \text{other object}\}$, where clear denotes an unobstructed direction. Separating person from other object lets a navigation aid respond differently, since a person may move out of the path whereas a static object will not. At inference the model observes only the thermal tiles, $f : (x_L, x_M, x_R) \mapsto (y_L, y_M, y_R)$; training labels come from the paired RGB stream (Section III-B).

B. RGB-Based Automatic Label Generation

Our labeling builds on DareDevil [8]. Unlike its single-camera formulation, detections are assigned to the three thermal cameras using coarse footprint overlap rather than pixel-accurate RGB-to-thermal alignment. This requires one RGB-to-thermal homography per camera, calibrated once during rig assembly, but no geometric calibration is needed at inference.

Each RGB frame is processed by YOLOv8x [14] at its default confidence threshold (0.25). A direction whose footprint overlaps a person detection is labeled person (even if other objects also overlap it), one overlapped only by other detections is labeled other object, and one with no overlapping detection is labeled clear. Frames without a matched RGB

frame (within 150 ms) are left unlabeled and masked during training.

C. Multi-View Fusion Architectures

We evaluate four architectures that map raw single-channel 32×24 tiles to a three-class prediction per direction and differ in how the three views are combined: pixel-, decision-, and feature-level fusion, plus feature-level fusion with camera-identity correction. Except for the stitched baseline, all models share a convolutional backbone whose depth, width multiplier, and activation are set during model selection and which returns multi-scale feature maps.

a) Model 1: Stitched Panorama (pixel-level fusion): The tiles are warped into a shared panorama with precomputed thermal-to-thermal homographies [15], [16], and a CNN predicts each direction from the panorama. Information lost in the warp cannot be recovered later.

b) Model 2: Output Fusion (decision-level fusion): The backbone processes each tile independently, a per-camera head produces a 3-class logit vector, and a small MLP refines the concatenated logits, so fusion sees only low-dimensional logits rather than spatial features.

c) Model 3: Feature Fusion: The shared backbone runs on each tile, and at every level the three feature maps are concatenated channel-wise and projected back to the original channel count by a learned 1×1 convolution, giving one fused multi-scale pyramid from which a shared head produces the per-direction logits. Model selection over depth, width, activation, and head capacity yields depth 5, width multiplier 1.5 $\times$, GELU, a 512-unit head, and dropout 0.1, used for Models 3 and 4 throughout.

d) Model 4: Feature Fusion with camera-identity correction: Feature Fusion assumes a fixed mapping between input slots and physical directions, making it sensitive to changes in camera arrangement. A lightweight camera-identity classifier, a three-layer CNN of about 26K parameters that is separate from the fusion backbone, predicts each raw tile's direction; given possibly shuffled tiles, it selects the most likely camera-to-direction assignment, and the reordered tiles are passed to

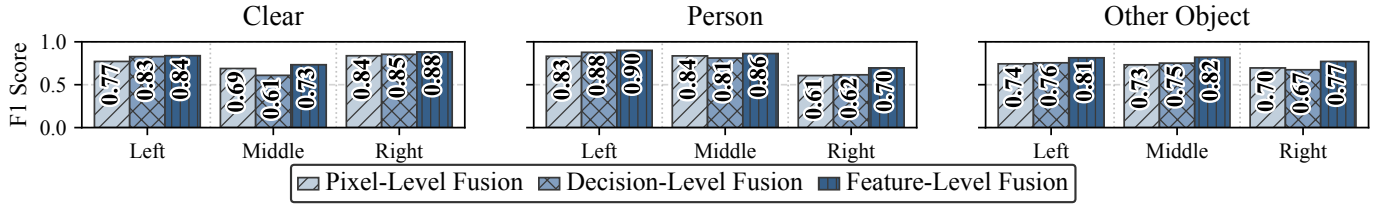


Fig. 2. Per-class F_1 by camera direction (left, middle, right) for pixel- (stitched), decision- (output fusion), and feature-level fusion, averaged over three training seeds on validation. Feature-level fusion is the strongest in every class and direction.

the unmodified Feature Fusion model. We compare against a blind condition with no correction and an oracle using ground-truth identities.

All models are trained with the same per-direction three-class cross-entropy, weighted by inverse class frequency and masked for unlabeled frames, using Adam (learning rate 10^{-3} , weight decay 10^{-4}) with a cosine schedule for 30 epochs at batch size 64 and keeping the epoch with the best validation macro F_1 (averaged over *person* and *other object*, as throughout).

IV. EXPERIMENTAL SETUP

A. Wearable Platform

The wearable platform (Fig. 1) consists of an NVIDIA Jetson Nano, three Melexis MLX90640 thermal cameras arranged in a forward-facing fan, and a co-mounted Raspberry Pi Camera v2 used only to generate labels. The thermal cameras provide left, middle, and right views, each a 32×24 image at 4 Hz. The platform is built from commodity components for about \$471, well below comparable wearable depth-based navigation systems [17], [18]. Thermal and RGB streams are timestamped during acquisition and synchronized offline, discarding poorly aligned samples before training.

B. Data Collection

The dataset¹ was collected on the Carnegie Mellon University in Qatar campus and spans a variety of indoor and outdoor scenes, planned to cover a free, a partially obstructed, and a fully obstructed path (Table I); the per-direction labels themselves come from the detector. A human operator wore the platform and walked through each environment at a natural pace. Sessions were recorded both with and without other people present, so that the automatic labeler encountered a realistic mixture of *clear*, *person*, and *other-object* events. In total, the platform recorded about 116 minutes of data. The experiments use 15,868 frames from 12 sessions, split at random into 12,697 for training and 3,171 for validation; a separate session of 988 frames, never used for training or validation, is held out as the test set for the camera-arrangement experiments. Fig. 3 shows the resulting class distribution.

¹The thermal frames, labels, and homographies are available at <https://huggingface.co/datasets/cmu-impactlab/thermal-directional-navigability>, and the code at <https://github.com/cmu-impactlab/ershad-directional-navigability>. RGB frames are withheld to protect the privacy of recorded individuals.

TABLE I
SCENE TYPES ENCOUNTERED DURING DATA COLLECTION.

Environment	Path	People
Open corridors	Free	No
Stairwells	Free	No
Covered walkways	Free	No
Doorways	Partial	No
Furniture in path	Partial	No
Corridor with bystander	Partial	Yes
Crowded lobbies	Full	Yes
Waiting areas	Full	Yes
Building entrances	Full	Yes

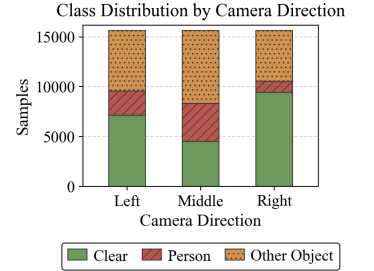


Fig. 3. Class distribution across camera directions for the combined training and validation split.

V. RESULTS

A. Navigability Estimation and Fusion Strategy

All models recover navigability above chance from raw 32×24 tiles, and Feature Fusion is the strongest in every class and camera direction (Fig. 2), with overall per-class F_1 (mean over directions) of 0.82 (*clear*), 0.82 (*person*), and 0.80 (*other object*) on validation. The stitched-panorama and output-fusion baselines reach 0.72–0.77 overall and are weakest on person detection from the right camera (0.61 and 0.62). Feature Fusion’s lead holds even when decision-level fusion is given its backbone and head settings (2.03M parameters): it stays ahead on *clear*, by 1.4 points overall and 3.4 in the middle direction, and on macro F_1 (0.811 vs. 0.808), with 8.5% fewer parameters (1.86M vs. 2.03M). Concatenating feature maps at every level preserves spatial cues that stitching discards in the warp and that output fusion never sees, consistent with the gain concentrating on free space. On the held-out test session, Feature Fusion with the correct camera order reaches 0.92, 0.92, and 0.90 (Fig. 4, oracle), so the configuration selected on validation carries over to a session never used for training or model selection; its higher scores likely reflect that session’s scene mix. On a cloud CPU, a single-sample forward pass takes at most 2.37 ms (Feature Fusion) plus 0.21 ms for the identity classifier, far below the 250 ms frame period at 4 Hz.

B. Comparison with Additional Baselines

We train three baselines with the same split, labels, and recipe. All learned models here and in Fig. 2 are averaged over three seeds, with standard deviations of at most 0.026 in F_1 (at most 0.007 for Feature Fusion and the no-fusion

TABLE II
BLOCKED-DIRECTION DETECTION (*person* OR *other object* COUNTS AS
BLOCKED) AGAINST HUMAN JUDGMENTS ON THE 241 DIRECTIONS
WHERE BOTH ANNOTATORS AGREE (155 BLOCKED).

Model	F_1	Acc.	Prec.	Rec.
RGB detector (full res.)	0.87	0.81	0.78	0.98
Feature Fusion (32×24)	0.87	0.82	0.81	0.94

model). The no-fusion model is Feature Fusion without the cross-camera connection: each direction is predicted from its own tile by a per-direction head. The permutation-equivariant model appends the mean and maximum of the pooled tile features across cameras and applies one shared head to every tile [19]. The threshold baseline labels a tile *person* when its warmest pixel lies in a body-temperature range and exceeds the tile median by a margin, and *other object* when its temperature standard deviation exceeds a threshold; thresholds are grid-searched on the training split for macro F_1 over all classes.

Removing fusion costs accuracy in every class and direction: the no-fusion model reaches 0.79, 0.81, and 0.78 overall, and for each class Feature Fusion’s weakest seed outperforms the no-fusion model’s strongest. The gain (2.9, 1.2, and 2.2 points) is largest for *clear*, the decision a navigation aid acts on, and peaks at 5.6 points in the middle direction, the only camera with neighbors on both sides. Neighboring views appear to help resolve free space, which is ambiguous in a single tile, while a warm body is easy to spot in one. Feature Fusion does this with fewer parameters at the same CPU latency. The threshold baseline reaches only 0.52 macro F_1 (0.50 on *clear*), and per-direction thresholds do not help (0.51).

C. Human Validation of Navigability

To check that the labels and the model capture what matters for navigation, two annotators labeled 100 validation frames (300 directions) from the RGB images without seeing the automatic labels. For each direction, they recorded what it contains (*clear*, *person*, or *other object*) and whether it blocks the path. Frames came from all 12 sessions, with quotas ensuring enough *person* and *other object* directions; the annotators agreed on 82% of classes ($\kappa = 0.71$) and 80% of blocking judgments ($\kappa = 0.59$); we score where they agree.

For navigation, the key question is whether a direction is blocked. Here the 32×24 thermal model matches the full-resolution RGB detector that trained it (Table II): it catches 93.5% of blocked directions, reaches the same F_1 of 0.87, and raises 21% fewer false alarms on open directions. The detector’s extra false alarms come from flagging any object in a footprint, even one that does not block the path: 28% of its *other object* directions were judged open. Feature Fusion is likely less affected because distant or thermally indistinct objects that the detector flags are hardly visible at 32×24 . The automatic labels are consistent with human judgment: they agree with the annotators’ shared label on 86.5% of directions (95% CI 80.6–91.8%, weighted to the validation class mix), and the detector missed no person both annotators saw.

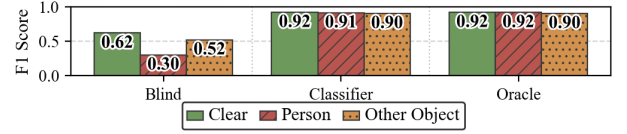


Fig. 4. Overall per-class F_1 of Model 4 under camera shuffling on the held-out test set (988 frames, not seen during training or validation), with no correction (blind), classifier re-sorting, and oracle re-sorting.

D. Camera Arrangement Robustness

Shuffling camera slots collapses uncorrected Feature Fusion, and the identity classifier restores it to near-oracle accuracy on the held-out test set (Fig. 4). Without correction, overall per-class F_1 drops to 0.62 (*clear*), 0.30 (*person*), and 0.52 (*other object*), because the fixed-order projection layers encode slot position. The classifier assigns each tile to its slot with 98.7% accuracy on the same 988 frames (chance: 33.3%), and re-sorting with it recovers 0.90–0.92 for all three classes, within 0.01 of the oracle. The tiles themselves carry direction information; what breaks under shuffling is the slot ordering, not the thermal representation.

The permutation-equivariant baseline comes within about half a point of Feature Fusion in macro F_1 (0.805 vs. 0.811), and on validation its predictions are unchanged under all five non-identity permutations without any correction. The no-fusion model offers no such protection: its direction-specific heads collapse under shuffling as well (0.61, 0.42, and 0.49 on validation). Because the equivariant model’s outputs follow the input slots, it still cannot tell which direction each tile faces, so reporting that the left path is blocked requires camera identity in either design. For Feature Fusion the classifier also restores accuracy, allowing the slightly stronger model to be used when camera inputs are reordered.

VI. CONCLUSION

This paper showed that directional navigability can be learned from three 32×24 thermal cameras using automatically generated RGB labels. Feature-level fusion outperformed per-camera prediction across all classes and directions, consistently across training seeds, with fewer parameters at comparable latency. A lightweight camera-identity classifier restored near-oracle accuracy under input permutations at negligible latency cost. A permutation-equivariant variant avoids the ordering problem but still needs camera identity to report directions. Every fusion architecture runs in under 3 ms per frame on CPU, supporting embedded wearable deployment.

This study has limitations. Navigability is defined by an RGB detector’s object labels rather than measured traversability, and the system has not been evaluated in end-to-end navigation or with users. Data come from a single campus, and camera rearrangement is simulated by permuting input slots rather than by physically re-mounting the sensors. The training/validation split is random at the frame level, which makes the held-out session the stricter test. Future work will address these with physically re-mounted sensors, recordings in new environments, and user studies.

REFERENCES

- [1] J. Yu, Y. Yang, and V. Nambodiri, "NaVIP: A Computer Vision-Based Indoor Navigation Solution for Visually Impaired Persons," *IEEE Access*, vol. 14, pp. 32 235–32 251, 2026.
- [2] J. Kuo, M. Muglikar, Z. Zhang, and D. Scaramuzza, "Redesigning SLAM for Arbitrary Multi-Camera Systems," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 2116–2122.
- [3] Z. Cao, H. Mu, Y. Sun, J. Xia, Y. Shi, Z. Ma, and M. Zhou, "Wearable Visuomotor Navigation and Mobility Assistance Device for the Partially Sighted and Visually Impaired in Unfamiliar Indoor Scenes," *IEEE Transactions on Industrial Electronics*, pp. 1–12, 2026, early access, doi: 10.1109/TIE.2026.3684167.
- [4] H. Liu, R. Liu, K. Yang, J. Zhang, K. Peng, and R. Stiefelham, "HIDA: Towards Holistic Indoor Understanding for the Visually Impaired via Semantic Instance Segmentation with a Wearable Solid-State LiDAR Sensor," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2021, pp. 1780–1790.
- [5] H. Mu, G. Zhang, Z. Ma, M. Zhou, and Z. Cao, "Dynamic Obstacle Avoidance System Based on Rapid Instance Segmentation Network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 5, pp. 4578–4592, 2024.
- [6] M. Rinta-Homi, N. H. Motlagh, A. Zuniga, H. Flores, and P. Nurmi, "How Low Can You Go?: Performance Trade-offs in Low-Resolution Thermal Sensors for Occupancy Detection: A Systematic Evaluation," *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 3, pp. 1–22, 2021.
- [7] I. Papagianopoulos, G. De Mey, A. Kos, B. Wiecek, and V. Chatziathasiou, "Obstacle Detection in Infrared Navigation for Blind People and Mobile Robots," *Sensors*, vol. 23, no. 16, p. 7198, 2023.
- [8] J. Gao, H. K. Gedawy, E. Feo-Flushing, and K. A. Harras, "A Deployable Privacy-Preserving Thermal-Based Obstacle Detection System for Indoor Navigation," in *Proc. of IEEE International Conference on Communications*. IEEE, 2025, pp. 4245–4250.
- [9] A. Rezaei, M. C. Stevens, A. Argha, A. Mascheroni, A. Puiatti, and N. H. Lovell, "An Unobtrusive Human Activity Recognition System Using Low Resolution Thermal Sensors, Machine and Deep Learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 1, pp. 115–124, 2023.
- [10] F. Mohsen and A. Safa, "Deep Fusion of Ultra-Low-Resolution Thermal Camera and Gyroscope Data for Lighting-Robust and Compute-Efficient Rotational Odometry," in *2025 IEEE/ACS 22nd International Conference on Computer Systems and Applications (AICCSA)*, 2025, pp. 1–7.
- [11] K. Zhu, W. Chen, W. Zhang, R. Song, and Y. Li, "Autonomous Robot Navigation Based on Multi-Camera Perception," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5879–5885.
- [12] A. S. Olagoke, H. Ibrahim, and S. S. Teoh, "Literature Survey on Multi-Camera System and Its Application," *IEEE Access*, vol. 8, pp. 172 892–172 922, 2020.
- [13] D. Zheng, X. Zheng, L. T. Yang, Y. Gao, C. Zhu, and Y. Ruan, "MFFN: Multi-View Feature Fusion Network for Camouflaged Object Detection," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 6221–6231.
- [14] G. Joher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [15] H. Dahmani, M. Yaqoob, M. Y. Ansari, and E. Feo-Flushing, "Thermal Homography in Photovoltaic Panels: Evaluating Deep Learning and Feature-Based Methods," in *Proc. of IEEE Texas Power and Energy Conference (TPEC)*, 2025, pp. 1–6.
- [16] E. Ansar, S. Sewil, F. Z. Maksood, and E. Feo-Flushing, "Performance Evaluation of Feature Detectors and Descriptors with Close-Range Solar Panel Images," in *Proc. of IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 2024, pp. 1–5.
- [17] Z. Li, F. Han, and K. Zheng, "An RGB-D Camera-Based Wearable Device for Visually Impaired People: Enhanced Navigation with Reduced Social Stigma," *Electronics*, vol. 14, no. 11, p. 2168, 2025.
- [18] J. Bai, Z. Liu, Y. Lin, Y. Li, S. Lian, and D. Liu, "Wearable Travel Aid for Environment Perception and Navigation of Visually Impaired People," *Electronics*, vol. 8, no. 6, p. 697, 2019.
- [19] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola, "Deep Sets," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3391–3401.