Robust Human Detection under Visual Degradation via Thermal and mmWave Radar Fusion

Kaiwen Cai[1*], Qiyue Xia[2*], Peize Li[2], John Stankovic[3], Chris Xiaoxuan Lu[2]✉
[*]Co-primary authors, ✉ Corresponding author
[1]University of Liverpool, [2]University of Edinburgh, [3]University of Virginia
[1]k.cai@liverpool.ac.uk, [2]xqy170605@gmail.com, {peize.li, xiaoxuan.lu}@ed.ac.uk, [3]stankovic@cs.virginia.edu

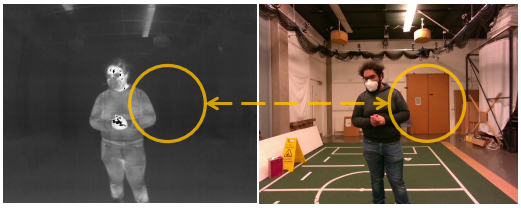
Abstract

The majority of human detection methods rely on the sensor using visible lights (e.g., RGB cameras) but such sensors are limited in scenarios with degraded vision conditions. In this paper, we present a multimodal human detection system that combines portable thermal cameras and single-chip mmWave radars. To mitigate the noisy detection features caused by the low contrast of thermal cameras and the multi-path noise of radar point clouds, we propose a Bayesian feature extractor and a novel uncertainty-guided fusion method that surpasses a variety of competing methods, either single-modal or multi-modal. We evaluate the proposed method on real-world data collection and demonstrate that our approach outperforms the state-of-the-art methods by a large margin.

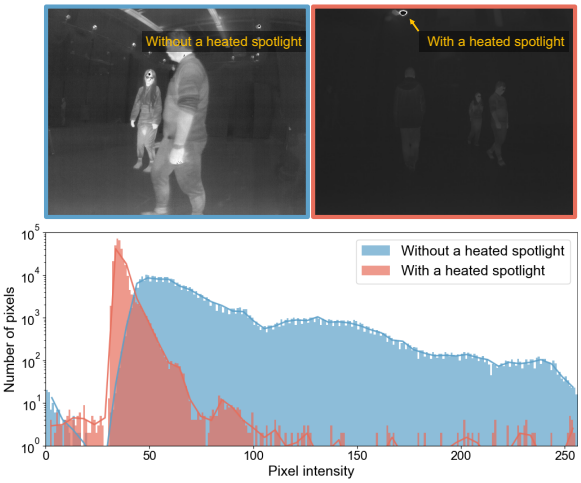
1 Introduction

Human detection is the task of locating all instances of human beings present in sensor observations. An accurate human detection module is widely regarded as an essential component in human-centred computing and cyber-physical systems. More often than not, detecting human subjects is the precursor step to enable subsequent context inference, such as pose estimation, occupancy monitoring, human tracking and crowd activity recognition etc.

Owing to their low cost and ubiquity, RGB cameras have been widely used as the *de-facto* solution to human detection and achieved great performance by leveraging recent advances in deep neural networks and computer vision. Unfortunately, as RGB cameras are a type of sensor operating with visible lights, the performance of RGB-camera-based methods is very susceptible to variable illumination (e.g., sun glare, dimness or darkness) and even fails under severe visual degradation, e.g., the smoke-filled fire areas and dust-filled construction sites. To address the intrinsic limitation of RGB cameras and visible light, a few works explore un-



(a) Thermal image (left) provides much fewer details than RGB image (right).



(b) A heated spotlight in the background decreases the contrast of the thermal image.

Figure 1: Limitation of thermal cameras.

conventional sensor modalities operating with invisible electromagnetic waves for human/object detection tasks. Typical examples include thermal cameras [46, 16] and millimeterwave (mmWave) radars [9, 23] - two emerging low-cost sensors that draw increasing attention from both academia and industry. As these sensors operate with electromagnetic waves of much larger wavelengths than visible light, they are fundamentally robust to variable illumination and airborne particles. However, these non-traditional sensors also suffer from their own challenges when it comes to human detection. Specifically, one well-known limitation of thermal cameras is their lack of distinctive image features or context when the temperature field is flat (see Figure 1a). Additionally, as shown in Figure 1b, thermal images tend to have low

contrast when the temperature of objects varies significantly. These two challenges jointly impose non-trivial detection challenges on thermal images even using cutting-edge computer vision algorithms. On the other hand, mmWave radars are known for their poor spatial resolution and high noise floor due to the multi-path effect, rendering themselves an unreliable sensor modality to detect humans in some indoor environments. For instance, Figure 2 illustrates the impact of multi-path on the mmWave radar sensing in a narrow corridor. The multi-path effect results in a cluttered point cloud with the human subjects being smeared into the background. For these reasons, the detection accuracy using thermal cameras or mmWave radars alone is far from being comparable to the one using RGB cameras in practice.

Towards a more robust human detection system under various illumination and visually degraded conditions, it is intuitive to design a multimodal fusion approach that combines the strengths of thermal cameras and mmWave radars and in the meantime, compensates for their weaknesses. While such a fusion concept is straightforward, transforming it into a useful system requires addressing multiple technical challenges. First, prior arts in this vein e.g., [44] [32] mostly adopt a late fusion strategy that individually optimises the feature extractors of each sensor modality. Such a separate feature extractor design under-exploits the cross-modal complementariness during the feature learning phase and incurs sub-optimal human detection results. Moreover, the predominant fusion operation in DNNs, e.g., the self- or cross-attention mechanisms [7, 28, 31], generates the mask from the immediate features to weaken unimportant or noisy features. However, in our settings where the input data is sparse and noisy, the learnt attention masks could be inevitably misled by data uncertainty. This challenge is particularly prominent in our context, as thermal or mmWave data tend to have many unimportant areas or noisy observations than RGB images.

In order to address these challenges, we propose UTM, a novel Uncertainty-guided Thermal and Mmwave radar fusion framework that is able to robustly detect human subjects under visual degradation by fusing the thermal and mmWave data. UTM follows an end-to-end optimization that leverages a Bayesian Neural Network (BNN) to extract the features from two modalities jointly and provide direct feature uncertainty to inform the mask generation. UTM demonstrates the feasibility of thermal-mmWave fusion for human detection tasks and provides a generic framework for different host platforms (e.g., installed as a building infrastructure or embedded on headsets). In summary, our contributions are as follows

1. This is the first-of-its-kind work that explores the usage of portable thermal cameras and single-chip mmWave radar for robust human detection.
2. We propose UTM, consisting of a novel Bayesian feature extractor (BFE) and a novel uncertainty-guided fusion (UGF) method to systematically address the detection challenges of caused by thermal images and noisy radar point clouds.
3. We evaluate our proposed UTM on real-world data collection, and the experimental results demonstrate that our method outperforms the best of competing method by 8.4% in $\text{mAP}_{50:95}$ (mean of AP_{50}, AP_{65},.., AP_{95}).
4. The collected thermal-mmWave human detection dataset and the source code of UTM are publicly released to the community at https://github.com/ramdrop/utm.

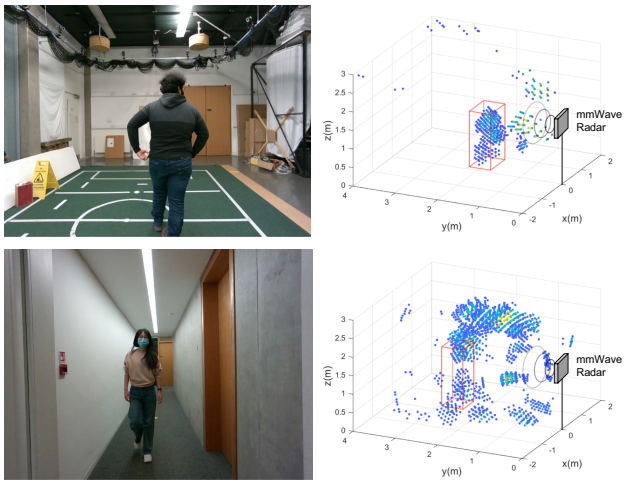


Figure 2: Top row: mmWave point cloud in an open space where the multi-path effect is negligible. Bottom row: mmWave point cloud in a narrow corridor where the multi-path effect is strong. Comparing the two, we can clearly see that the multi-path effect can significantly impact the point cloud quality and human subjects (highlighted with the red boxes) are difficult to be differentiated from the background points under such effects.

2 Preliminaries

2.1 Thermal Imaging

Thermal cameras [36] capture the infrared radiation emitted by objects and convert the detected energy into pixels' values in the thermal image. The infrared radiation that can be captured lies between visible light and microwaves within the wavelength spectrum of 0.7–1,000 μm and a common sub-division scheme is illustrated in Figure 3. Among all types of thermal cameras, the long-wavelength infrared (LWIR) camera is particularly interesting and used in this work. LWIR cameras can observe the temperature field ranging from approximately 190 $\sim$ 1,000K [13]. Importantly and similar to all other thermal cameras, LWIR cameras do not require additional sources of light or heat and can passively capture the emitted thermal energy of ambient objects. These characteristics make them a robust alternative to RGB cameras in visually degraded conditions.

2.2 Millimiter-Wave Radar

Millimiter-Wave (mmWave) radar is a transceiver device that operates with electromagnetic waves between 30GHz-300GHz. mmWave radar uses a linear 'chirp' or swept frequency transmission. When receiving the signal reflected by an obstacle, the radar front-end performs a dechirp operation by mixing the received signal with the transmitted

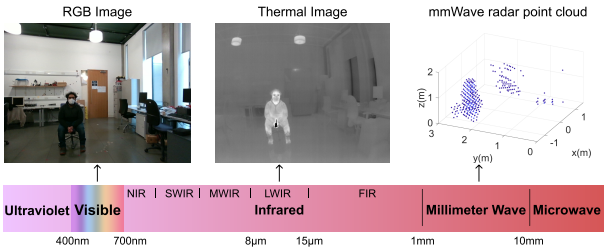


Figure 3: Electromagnetic spectrum with sub-division of infrared radiation.

signals, which produces an Intermediate Frequency (IF) signal. The distance between the object and the radar can be calculated from the IF signal. A mmWave radar can further estimate the obstacle angle by using the different virtual antennas. Signals received at different antennas might have different phases due to the distance between the receivers. Based on the distance between the receivers and the corresponding phase differences of the received signals, the angle of arrival can be estimated [40]. While mmWave radars have been widely used in human sensing, current methods are strongly susceptible to the multi-path effect common in indoor environments.

3 Method Design

The proposed cross-modal human detection pipeline, `UTM`, is shown in Figure 4. Our proposed `UTM` consists of three modules: a Bayesian Feature Extractor (BFE), an Uncertainty-Guided Feature Fusion module (UGF) and a Multiscale Detection Net (MDN). Given a raw thermal image and a raw radar point cloud, we first preprocess the data to generate a radar depth image (see Sec. 3.1), Next, we use the BFE module to extract features from both the thermal image and the radar depth image in parallel (see Section 3.2). The features from the two modalities are then fused using the UGF module (see Section 3.3). Finally, we use the MDN module to generate detection bounding boxes (see Section 3.4). We detail the design of each module in subsequent sections.

3.1 Data Preprocessing

Before feeding the data to the human detection network, mmWave radar and thermal camera data needs to be preprocessed. Particularly, we aim to have image-like array representation as the development of neural networks for image processing is more established and give us more design choice. Thermal images are naturally array data while mmWave radars give point clouds. This requires us to use point cloud projection to convert the radar data into images, leading to its final representation akin to the depth images. Concretely, the radar data in each frame is a set of points, where each point is represented by a 3-d vector. For clarity, we denote the point by $p := (x, y, z) \in \mathbb{R}^3$. The mmWave radar point clouds are projected to the 2D image plane using the extrinsic and intrinsic parameters between the thermal camera and the mmWave radar. The projection of each point follows:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} KT \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{1}$$

where K is the 3×3 thermal camera's intrinsic matrix, and T is the 3×4 extrinsic matrix between the thermal camera and the mmWave radar. (x, y, z) is the 3D location in the radar coordinate and (u, v) is the projected pixel location in the 2D image plane. For each projected pixel, we use the corresponding depth value from mmWave radar point clouds to represent the pixel value on the image, which generates an image-like depth map. Meanwhile, only the projected pixels that locate within the thermal image range are reserved so that the projected depth map has the same size and field of view as the thermal image. The same size and field of view allow for more flexible fusion strategies, including the early fusion where two types of images can be directly stacked together.

3.2 Bayesian Feature Extractor

3.2.1 Motivation behind BFE.

Convolutional Neural Networks (CNNs) have been widely used as feature extractors in many computer vision tasks [49, 43, 15] due to their expressive representation power to process image-like data. In a multi-modal detection network, two feature extractors are needed to process thermal and radar images respectively, if other fusion strategies are used other than early (input) fusion. To this end, we adopt the convolution blocks of the YOLOv5s [19] as our feature extractors due to their small footprint and a proven ability for general object detection tasks. Hereafter we refer to it as the basic Feature Extractor (FE).

While it is straightforward to use the basic FE for sensor fusion, there are significant limitations that must be addressed when using thermal and mmWave radar sensors. These sensors present unique limitations, such as the low contrast and featureless images of thermal cameras, and the sparse and noisy measurements provided by radar sensors. Intuitively, uncertainty in the extracted features is therefore inevitable as the inputs themselves can be in low fidelity. The challenge here is, however, sorting out all uncertainty sources is impossible and a general uncertainty estimation is needed. This motivates us to use a Bayesian Neural Network (BNN) [21] in the feature extraction process. BNNs are a sensible choice here because they can model the uncertainty in an agnostic manner by using predictions from weight distributions. We thus propose to upgrade the basic FE to a Bayesian Feature Extractor (BFE) which offers two advantages: 1) by using a Bayesian feature extractor, we can extract additional information, such as variance, from existing sparse and noisy data, and 2) the sensor uncertainties are mitigated by formulating the feature learning process as an optimization problem with posterior distributions.

3.2.2 BFE Design Detail

We now detail the principles of the proposed BFE. In the Bayesian framework, the feature extractor's weights $\mathcal{W}$ are treated as distributions rather than deterministic values.

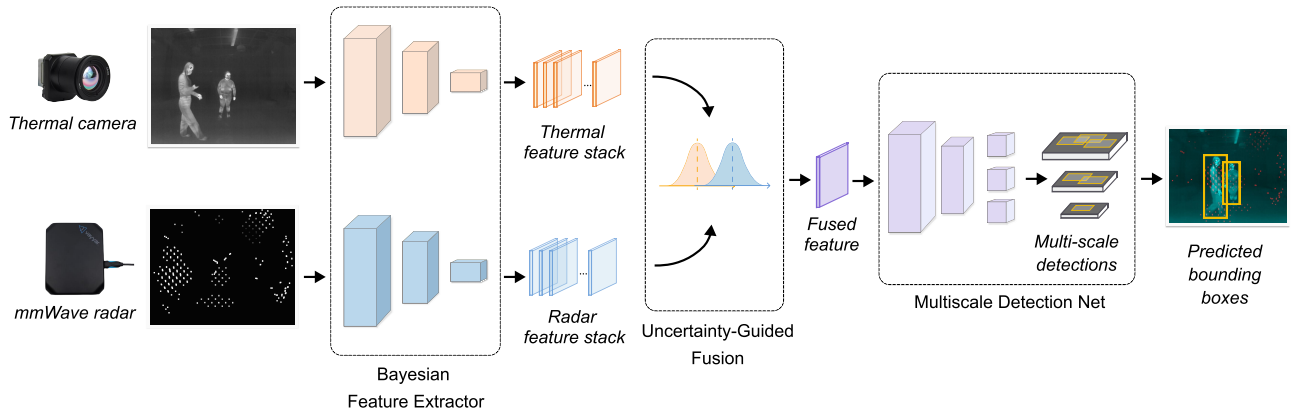


Figure 4: Overview of the proposed UTM.

Given the dataset $\mathcal{D}$, the goal is to find the posterior distribution of BFE's weights:

$$p(\mathcal{W}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{W})p(\mathcal{W})}{p(\mathcal{D})} \tag{2}$$

However, $p(\mathcal{W}|\mathcal{D})$ is intractable since the marginal probability $p(\mathcal{D})$ cannot be evaluated analytically. To solve this issue, Variational Inference (VI) is usually used to approximate the intractable distribution. VI-based methods [21] try to fit the posterior distribution $p(\mathcal{W}|\mathcal{D})$ with a tractable parametric distribution, e.g., Gaussian distribution, and then optimize over the parameters of the tractable distributions.

Based on the VI methodology, we adopt Dropout to approximate the posterior distribution. Dropout approximation assumes $p(\mathcal{W}|\mathcal{D})$ as a mixture of two Gaussian distributions with small variances and the mean of one is fixed to zero [21]. The original dropout approximation adds a dropout layer to each convolutional layer, and applies dropout during both training and inference phases. In practice, we observed that applying dropout to all convolutional layers decreases the representation ability of a feature extractor. As a result, we only include dropout layers in the final two convolutional blocks in both feature extractors. (markd in red in Figure 5).

As shown in Figure 5, there are two branches in BFE: *BFE main* and *BFE auxiliary*. BFE main is composed of three convolutional blocks outputting feature maps at $1/2$, $1/4$ and $1/8$ of input size, and BFE auxiliary shares the same architecture as BFE main. Given two cross-modal input data of size $(3,H,W)$ and $(3,H,W)$, the feature maps of the two branches have sizes of $(128,H/8,W/8)$ and $(128,H/8,W/8)$. We input thermal images to the main branch and radar depth images to the auxiliary branch. Nevertheless, the branches are interchangeable, given that they share identical structures. In the next section, we detail how we fuse these cross-modal feature maps.

3.3 Uncertainty-Guided Fusion

Formulation. Conventional fusion methods (e.g., [7, 4] ignore the importance of feature map uncertainties in guiding multimodal fusion, rendering sub-optimal multimodal detection results. To this end, we propose an uncertainty guided fusion module, called UGF, which fuses two feature maps with the guidance of their variance maps provided by the BFE. The principle of the UGF is as follows. To capture the model uncertainty during training, the BFE forward propagates the input data N times and thus produces a stack of N feature maps in each branch, which we refer to as the main feature stack $\mathcal{F}_m$ and the auxiliary feature stack $\mathcal{F}_a$, which are written as follows:

$$\begin{aligned} \mathcal{F}_m &= \{F_{m,i}|F_{m,i} \in \mathbb{R}^{H\times W\times C}, i = 1,2,..,N\}, \\ \mathcal{F}_a &= \{F_{a,i}|F_{a,i} \in \mathbb{R}^{H\times W\times C}, i = 1,2,..,N\}. \end{aligned} \tag{3}$$

Figure 6 shows the overview of the UGF module. We first derive the mean and the variance map from $\mathcal{F}_m$. As the variance indicates where the BFE is uncertain, we apply a sigmoid operation and spatial softmax operation to the variance map, such that the value at each pixel location is converted to the weight of the feature (i.e., the sum of weight is a unit). We apply the same operations to $\mathcal{F}_a$. The fused feature map, F_{fused}, is the weighted sum of the feature maps of two branches. We put it more formally as

$$F_{fused} = \phi_\mu(\mathcal{F}_m) \odot \phi_{ss}\phi_{sg}(\phi_\sigma(\mathcal{F}_m)) + \phi_\mu(\mathcal{F}_a) \odot \phi_{ss}\phi_{sg}(\phi_\sigma(\mathcal{F}_a)),$$

where $\odot$ means the Hadamard product; ϕ_μ and ϕ_σ means obtaining the mean and variance of samples, respectively; ϕ_{ss} and ϕ_{sg} means applying a sigmoid operation and spatial softmax operation, respectively.

Fusion Explanation. The intuition behind the UGF is that features are not equal to the fused feature map and should be differently treated in the *training phase*. As high variance indicates high uncertainty, an effective training phase should let the model focus more on those features of high uncertainty (i.e., hard examples) so that the model can gradually familiarise itself to handle difficult features coming from both sensors, regardless of what the uncertainty causes are. Through this 'hard training', the trained model is expected to predict well even when being fed with high-uncertainty feature maps when it comes to the *inference stage*.

Figure 7 illustrates the difference between the conventional Attention Mechanism (AM) [24] and the UGF. It can be observed that for the same input thermal image, the

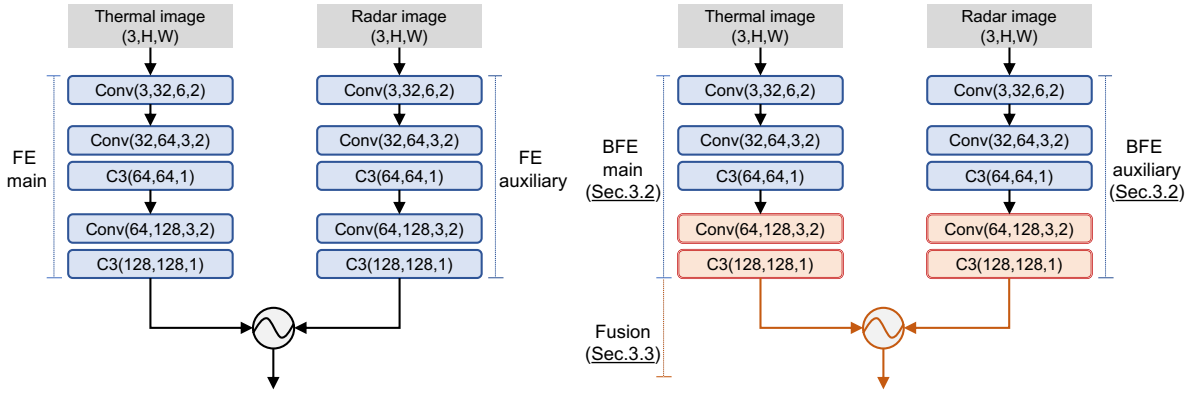


Figure 5: The basic Feature Extractor (FE) and the proposed Bayesian Feature Extractor (BFE): Conv(in,out,kernel_size,stride): The weights of FE are deterministic, while those of BFE are distributional and approximated with a mixture of two Gaussian distributions (see Sec.4.2 for details).

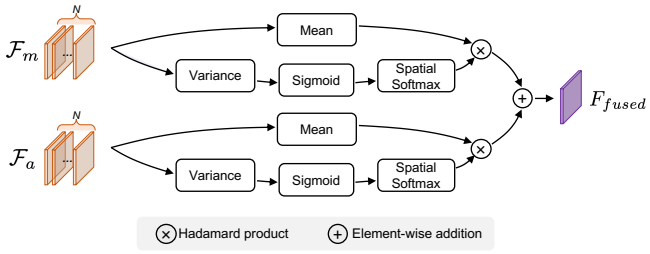


Figure 6: The proposed UGF: $\mathcal{F}_m$ denotes the main feature map stack, and $\mathcal{F}_a$ the auxiliary feature map stack.

weighted thermal feature map by AM shows that the model is focusing more on regions of human heads, which are typically considered as the characteristics of humans. In contrast, the weighted thermal feature map generated by the UGF emphasizes the overlapping regions of people, where the model typically has difficulty in differentiating between humans and objects. This demonstrates that the UGF forces the model to concentrate on challenging regions during the training in order to improve the inference robustness of the trained model.

3.4 Multiscale Detection Net

Once the fused feature map is obtained, it will be passed to the Multiscale Detection Net (MDN), which is based on the YOLOv5s [19] network. The YOLOv5s network is explained in detail in reference [19], but the principle of MDN is also provided in this section for completeness. Figure 8 shows the network architecture of MDN, which is composed of Convolutional layers or blocks, including the single convolutional layer, the C3 block and the Spatial Pyramid Pooling (SPP) block. The convolutional layers and blocks progressively extract feature maps at $1/8$, $1/16$ and $1/32$ of the input image size. The detection is made on feature maps at three scales, which accounts for small, medium, and large objects. Finally, the Non-Maximum Suppression (NMS) method is used to generate the final detection results.

4 Evaluation

4.1 Hardware and Data Collection

As there is currently a lack of publicly available datasets that incorporate both thermal camera and mmWave radar sensors, we designed a sensing platform and gathered data from a laboratory testbed for analysis. We are committed to making this dataset available to the research community and plan to release it publicly upon acceptance of our paper.

Sensing Platform Setup. Figure 9 presents the sensing platform comprising a mmWave radar, a thermal camera, and an RGB camera, affixed to a 3D printed carrier board.

The mmWave radar, Vayyar vTrigB [3], is the current best-in-class 4D radar operating between 62-69GHz. It has 20 TX and 20 RX antennas, capable of producing IQ signals with point cloud data at a density of 1000 to 3000 points per frame. The weight of this probe is 110g and the probe size is 105mm × 85mm. The mmWave radar is capturing at a frequence of 6 $\sim$ 13 Hz.

The thermal camera, the FLIR Boson 640 [2] thermal camera, has a 640 × 512 resolution, 9 Hz frame rate, and a 95° HFov that can detect temperatures up to 140 °C with high gain and 500 °C with low gain. This 79g probe has a small size of 21mm × 21mm × 11mm.

The Intel RealSense D455 [1] RGB camera has a working frequency of up to 90 Hz, a resolution of 640 × 480, and a 95 × 65° Fov. It measures 124mm × 26mm × 29mm and weighs 389g. We employ the RGB camera solely for labeling purposes.

Data Collection. To ensure the dataset's versatility, we collected human detection data[1] from various indoor environments, including offices, corridors, atriums, and kitchens. These settings possess varying lighting, spatial arrangements, and temperature conditions, as illustrated in Figure 10. We recruited volunteers to move around in the scene with different poses, while our platform detected and recorded the activity. Ultimately, we collected a total of 24,241 frames with each frame featuring 1 to 4 individuals

[1]This work has received the ethical approval XXX (no name for now due to double-blind review).

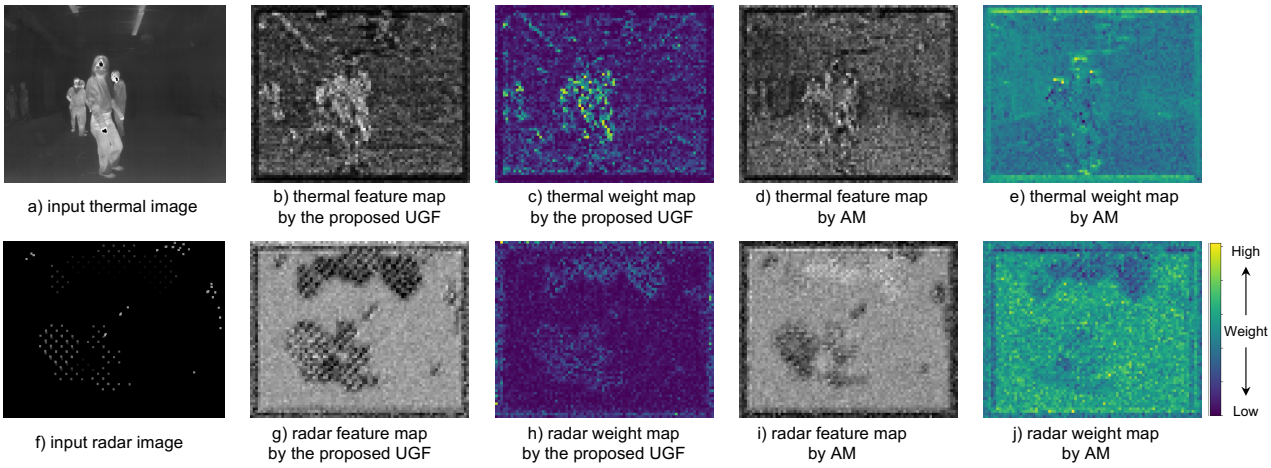


Figure 7: The inputs, feature maps and weight maps of the AM [24] and the proposed UGF module: AM is focusing more on easily-detected parts, e.g., heads, while UGF is emphasizing on noisy regions, e.g., overlapping human bodies (see highlights in their thermal weight maps). The radar weight maps show that the proposed UGF effectively concentrates on noisy measurements, while AM does not demonstrate a clear concentration.

and 12 volunteers participating in our data collection to ensure diversity. We randomly separated the frames into three distinct sets: 15,641 for training, 4,300 for validation, and 4,300 for testing.

To speed up the labelling process, we utilized the advanced vision-based detection tool detectron2 [50]. We initially obtained precise ground truth boxes on the RGB images. By leveraging the known intrinsic and extrinsic parameters of the RGB camera, thermal camera, and mmWave radar, we are able to effortlessly derive the ground truth boxes on both thermal and radar depth images through pose transformation.

4.2 Implementation Details

The previously mentioned components - BFE, UGF, and MDN - make up the complete pipeline of `UTM`, which we train in an end-to-end manner. To begin with, the thermal images and radar depth images are synchronized using the Robot Operating System [42] and then resized to a resolution of 640×512. In the UGF module, we set the dropout rate to $p = 0.2$ and the number of forward passes to $N = 5$. To optimize our model, we utilize a stochastic gradient descent optimizer with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 5×10^{-4}. We decrease the learning rate linearly over time until it reaches zero at the final epoch. Our training process spans 100 epochs, as we observe that the model's performance plateaus after this point. The selection of the best model is based on its performance on the validation set.

4.3 Evaluation Metrics

We evaluate the performance of `UTM` using predominant object detection metrics, including **precision**, **recall**, **max F1 score (mF1)** and **Average Precision (AP)** [12] at different Intersection-over-Union (IoU) thresholds. The IoU is defined as the ratio of the intersection between predicted and ground truth bounding boxes to the union of the same:

$$IoU = \frac{PredictedBox \cap GTBox}{PredictedBox \cup GTBox}, \tag{4}$$

and a detection is considered to be a true positive if the IoU is greater than a predefined IoU threshold. When assessing all detections, the precision and recall values are calculated as:

$$\begin{aligned} Precision &= \frac{TruePositives}{TruePositives + FalsePositives}, \\ Recall &= \frac{TruePositives}{TruePositives + FalseNegatives}, \end{aligned} \tag{5}$$

AP is the average detection precision under different recall values, i.e., the area under the PR curve. We follow [27] and evaluate AP with IoU thresholds from 0.50 to 0.95, denoted as AP_{50}, AP_{65}, .., AP_{95}, and calculate the mean AP over all IoU thresholds, denoted as $AP_{50:95}$.

F1 score is calculated as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

and the mF1 is the maximum F1 overall precision and recall values. Similar to AP and for fairness reasons, we evalaute the mF1 with different IoU thresholds from 0.50 to 0.95, denoted as $mF1_{50}$, $mF1_{65}$, ..,$mF1_{95}$, and calculate the mean max F1 score over all IoU thresholds, denoted as $mF1_{50:95}$.

4.4 Experimental Results

The following sections give a numerical analysis and a video demo shot in the dark environment can be found at https://github.com/ramdrop/utm.

4.4.1 Competing methods

We compare the proposed `UTM` against the following methods:

SOD [29] fuses the RGB channel and the depth channel by composing a four-channel image. We follow this idea to

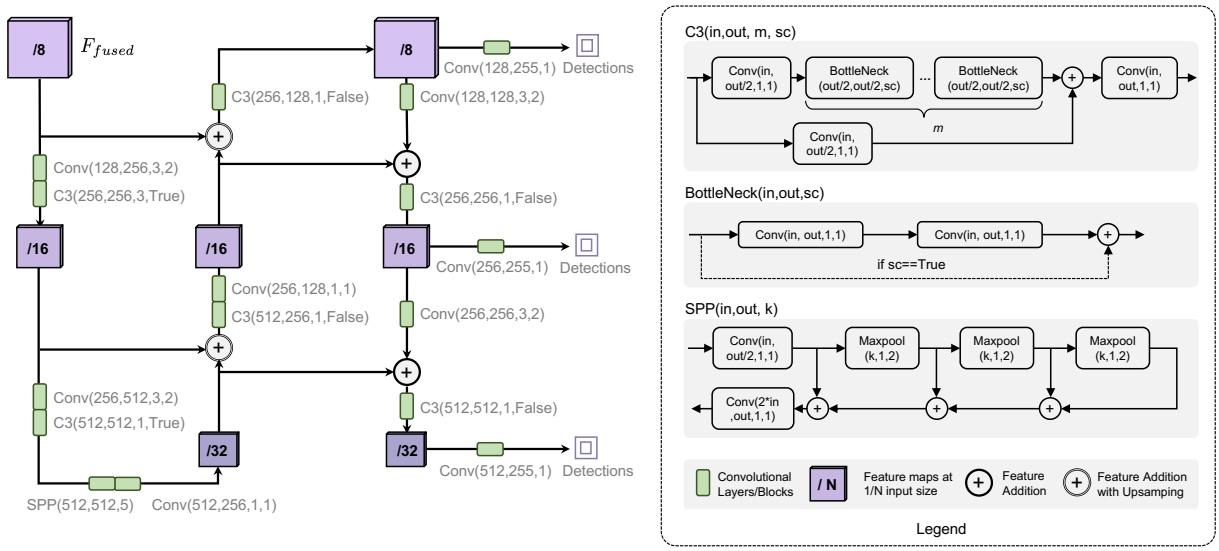


Figure 8: The Multiscale Detection Net: Feature maps are extracted and processed at three scales, $1/8$, $1/16$ and $1/32$ of the input image size. Detections are generated at each scale and then summarized.

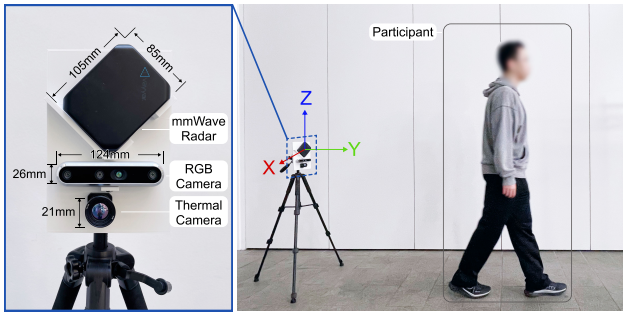


Figure 9: The sensing platform setup. Note that the RGB camera is only used for the pseudo-labelling in the training stage and is NOT used in the inference stage.

fuse the single-channel thermal image and radar depth image. Specifically, we concatenate two thermal images and one radar depth map to form a pseudo RGB image. To ensure fair comparison, we retain the other architectures intact while replacing our dual-branch BFE with a single-branch FE.

TOD [22] is a thermal camera-based object detection method, which we evalaute by replacing our dual-branch BFE with a single-branch FE while maintaining the same architectures. Furthermore, in order to examine the impact of mmWave radar sensor, we replace the input thermal image with the radar depth image, which we term as **TOD-Radar**.

MilliEye [44] uses mmWave radar and RGB cameras for human detection, where the fusion is performed at the decision level. It consists of an RGB image-based object detector trained on the public datasets COCO [27] and ExDark [30], a radar-based object tracker that produces box proposals to combine with image-based proposals, and an ROI-wise refinement head. We fine-tune MilliEye on our dataset.

Vanilla Addition (VA) [7] is a fusion technique that operates at the feature level. To ensure fairness, we evaluate it by solely substituting the proposed UGF module of `UTM` with VA. Specifically, the BFE propagates the input data N times, resulting a stack of N feature maps at each branch during the training. We obtain a single average feature map by computing the mean of the stack of feature maps. The two mean feature maps are added together and passed to the subsequent layers.

Attention Mechanism (AM) [24] is another feature-level fusion method. We evaluate it by solely substituting the proposed UGF module of `UTM` with AM. we first obtained mean feature maps from each branch, then convolved and applied a softmax operation on the radar branch to generate an attention mask. In parallel, we applied convolutional operations on the thermal branch to generate a query map, which was multiplied by the attention mask. The resulting feature map was then passed to the subsequent layers.

4.4.2 Overall Performance

Table 1 shows the AP and mF1 of various methods under different IoU thresholds when evaluated on our dataset with an NMS IoU threshold of 0.6 and Figure 11 illustrates the AP, mF1 and precision-recall curve of different methods on our dataset. We note that our evaluation involves three levels of fusion: input fusion, feature fusion, and decision fusion. SOD represents the input (early) fusion technique. Regarding feature fusion, we tested three methods, namely AM, VA, and `UTM`. And MilliEye is a decision (late) fusion-based approach.

Among sensor fusion methods, MilliEye performs the worst with the lowest $\text{mAP}_{50:95}$ of 0.343 and $\text{mmF1}_{50:95}$ of 0.375 and its curve is clearly surpassed by those of other methods in Figure 11. The under performance of MilliEye can primarily be attributed to the fact that its image-based object detector is trained on public RGB object detection

(a) Office 1 (b) Office 2 (c) Office 3 (d) Corridor 1 (e) Corridor 2

(f) Atrium 1 (g) Atrium 2 (h) Kitchen (i) Corridor 3 (j) Corridor 4

Figure 10: Ten scenes with different illuminations, spatial configurations, and temperature conditions.

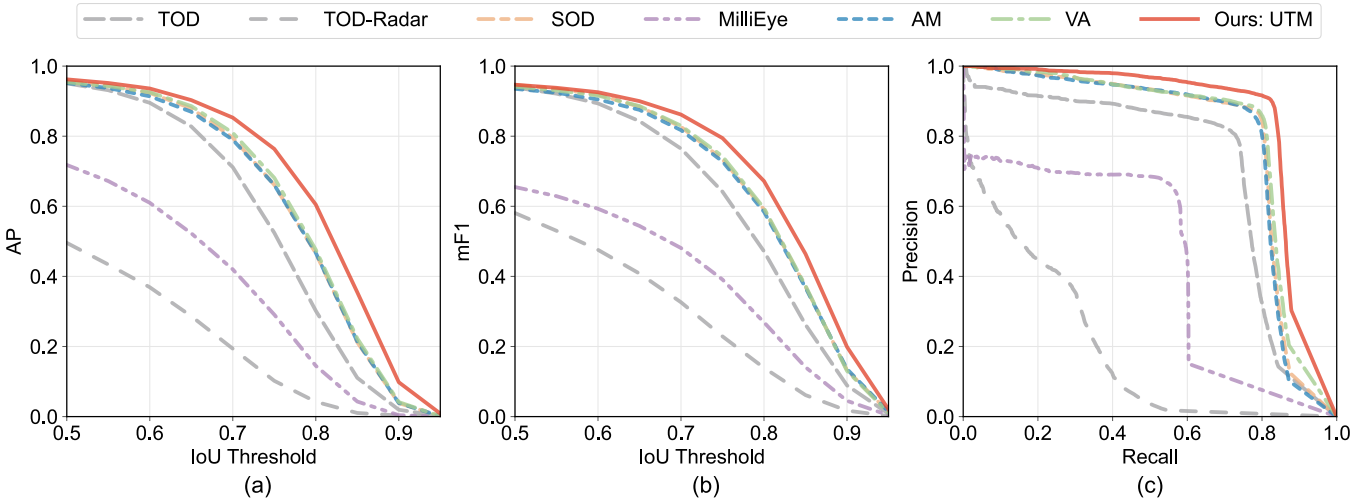


Figure 11: The mAP, mF1 and precision-recall curve of different methods under different IoU thresholds with a NMS IoU threshold of 0.6 and precision and recall are with a detection IoU threshold of 0.7. Our UTM outperforms the competing methods by a large margin in the three metrics.

datasets while our dataset comprises thermal and radar-based imagery, leading to a domain gap. Consequently, the performance of the MilliEye system is limited even after fine-tuning its image-based object detector using thermal images and mmWave radar point clouds.

SOD, AM, and VA exhibit similar values for AP and mF1 across various IoU thresholds in Table 1, as also demonstrated by their overlapping curves in Figure 11. This suggests that input fusion and feature fusion techniques perform comparably. Nevertheless, UTM surpasses them by a substantial margin, with a $\text{mAP}_{50:95}$ of 0.644 and a $\text{mmF1}_{50:95}$ of 0.672, outperforming SOD, AM and VA with $\text{mAP}_{50:95}$ boosts of 9.5%, 10.2% and 8.4%, and $\text{mmF1}_{50:95}$ boosts of 6.8%, 5.8% and 5.8%, respectively.

Although both AM and UTM are attempting to shift the model's focus to particular regions during training, the findings demonstrate that UTM, leveraging the proposed UGF, is superior to AM in integrating thermal images and radar depth images. Compared with VA, the UGF of UTM is able to learn the optimal fusion weights for each sensor, which is more flexible than the fixed weights in VA. Note that as illustrated in Figure 11, the performance gaps between UTM and other competing methods are more significant for high IoU thresholds (IoU = $0.7 \sim 0.9$). Successful detection under a high threshold means the model producing more precise bounding boxes that accurately match the ground truth box. Detections under high IoU thresholds are particularly desired in scenarios where safety is crucial. Therefore, UTM is superior in safety-related scenarios.

Figure 14 presents the qualitative results of different methods on our dataset. It is shown that detecting all human bodies when they are overlapping can be difficult. However, UTM was able to successfully detect all human bodies in the scene, while other methods failed.

Table 1: The AP and mF1 of different methods under different IoU thresholds with an NMS IoU threshold of 0.6.

Method	Input		AP_{50}	AP_{65}	AP_{75}	AP_{95}	$mAP_{50:95}$ ↑	$mF1_{50}$	$mF1_{65}$	$mF1_{75}$	$mF1_{95}$	$mmF1_{50:95}$ ↑
	R[1]	T[2]										
TOD [22]		✓	0.951	0.828	0.524	0.001	0.527	0.939	0.843	0.642	0.008	0.583
TOD-Radar [22]	✓		0.496	0.286	0.102	0.000	0.194	0.581	0.408	0.229	0.002	0.277
SOD [29]	✓	✓	0.957	0.879	0.661	0.001	0.588	0.947	0.884	0.734	0.013	0.635
MilliEye [44]	✓	✓	0.718	0.523	0.291	0.000	0.343	0.655	0.544	0.391	0.004	0.375
AM [24]	✓	✓	0.952	0.870	0.661	0.001	0.584	0.935	0.875	0.728	0.014	0.629
VA [7]	✓	✓	0.954	0.886	0.681	0.001	0.594	0.941	0.886	0.743	0.010	0.635
Ours: UTM	✓	✓	**0.962**[3]	**0.903**	**0.764**	**0.009**	**0.644**	**0.947**	**0.900**	**0.795**	**0.021**	**0.672**

[1] denotes Radar depth image.
[2] denotes Thermal image.
[3] **bold** denotes the best performance among all methods.

4.4.3 *Impact of Sensors*

The superiority of SOD over TOD and TOD-Radar is expected, as evidenced by the $mAP_{50:95}$ values of 0.588, 0.527, and 0.194, respectively. This is due to the fact that more sensor information leads to better detection performance. Thermal images significantly improve the performance of the SOD, more than radar images. This is due to the fact that radar point clouds tend to be sparser and noisier than thermal images, providing less useful information for object detection. This trend is also evident in Figure 11, where the SOD method betters the TOD and TOD-Radar by large gaps.

4.4.4 *Impact of IoU threshold for NMS*

NMS is a widely-used technique in object detection that helps to eliminate overlapping bounding boxes and is also used in UTM. The IoU threshold used in NMS can significantly impact the precision and recall values of an object detection model. For this consideration, we evaluate the sensitivity of our UTM to the IoU threshold for NMS. Figure 12 presents the $mAP_{50:95}$ and mean max $F1_{50:95}$. It can be seen that our UTM shows a negligible performance change as the IoU threshold for NMS increases from 0.30 to 0.90 while maintaining the best $mAP_{50:95}$ and mean max $F1_{50:95}$ among all methods. This implies satisfying model robustness against this important hyperparameter.

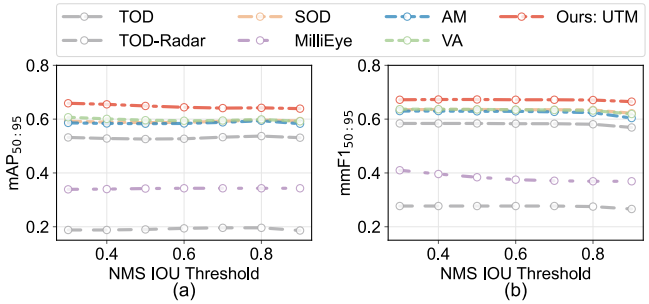


Figure 12: The mAP and mean Max F1 of different methods under different NMS IoU thresholds. Our UTM consistently performs well across varying NMS IoU thresholds.

4.4.5 *Impact of BFE Parameters*

Table 2 and Figure 13.a) display the AP of BFE with dropout enabled at various layers when p is fixed at 0.20.

Table 2: The mAP of BFE with dropout enabled at various layers when p=0.20.

	Dropout @ i^{th} layer					$mAP_{50:95}$ ↑
	1	2	3	4	5	
$i=5$					✓	0.619
$i=4,5$				✓	✓	**0.644**
$i=3,4,5$			✓	✓	✓	0.595
$i=2,3,4,5$		✓	✓	✓	✓	0.611
$i=1,2,3,4,5$	✓	✓	✓	✓	✓	0.604

Table 3: The AP of BFE with varying values of p while dropout is enabled in $i=4,5$ layers.

	p=0.05	p=0.10	p=0.15	p=0.20	p=0.25
$mAP_{50:95}$ ↑	0.634	0.642	0.619	**0.644**	0.637

It can be seen that the highest $mAP_{50:95}$ was attained by enabling dropout at the last two layers. We hypothesize that this is because insufficient dropout layers fail to adequately approximate a BNN, while excessive dropout layers may compromise the feature extractor's representational capacity.

Table 3 and Figure 13.b) present the AP of BFE with varying values of p while dropout is enabled in $i=4,5$ layers. It can be observed that UTM achieves the highest highest $mAP_{50:95}$ when $p=0.20$. However, it is worth noting that the AP gaps among the different BFEs are relatively small, suggesting that UTM's performance is not significantly affected by changes in the dropout rate.

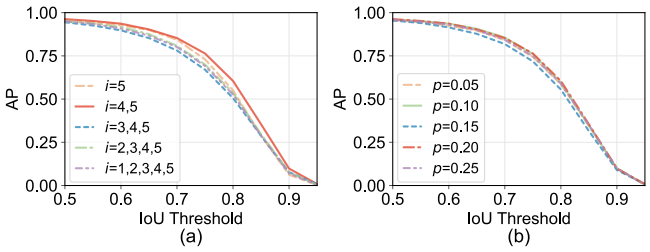
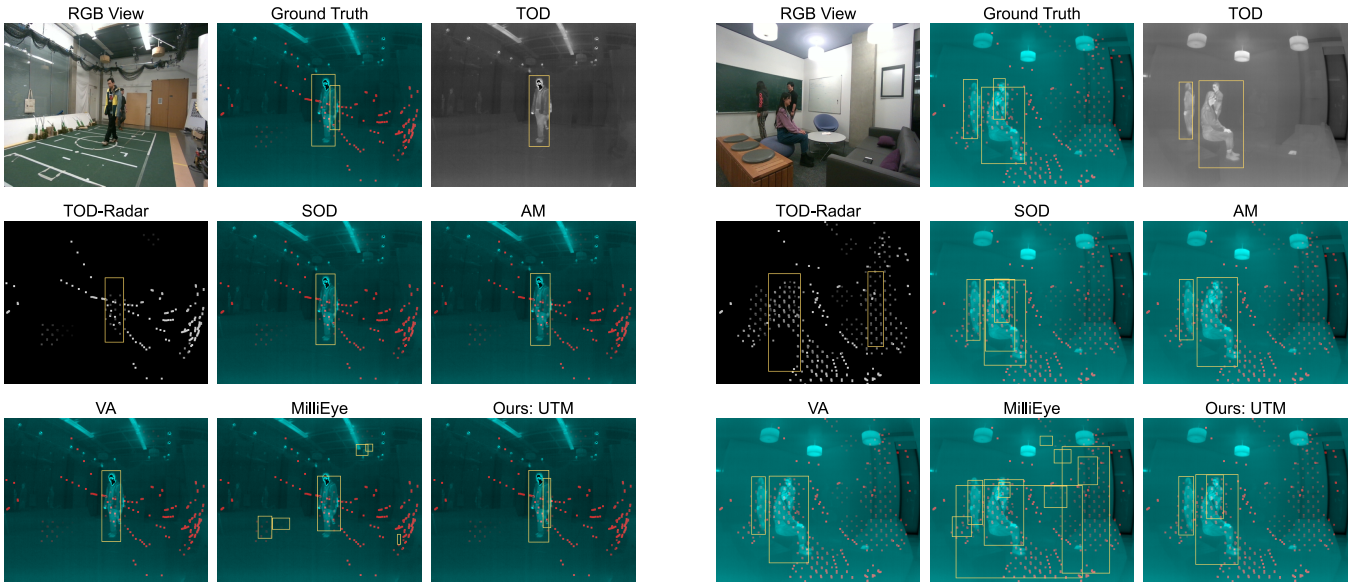


Figure 13: The mAP of BFE with varying configurations.



(a) Scene 1

(b) Scene 2

Figure 14: The qualitative human detection results on our dataset: it is shown that detecting all human bodies when they are overlapping can be difficult. UTM was able to successfully detect all human bodies in the scene, while other methods failed. (To visualize the thermal image and radar point cloud together in the same image, we create a pseudo RGB image by assigning the red and green channels to the thermal image and the blue channel to the radar point cloud.)

4.4.6 Runtime Efficiency

Table 4 presents the model size and the inference speed of UTM tested across various embedded devices, i.e., NVIDIA Jetson Xavier NX [39], and NVIDIA Jetson AGX Xavier [38]. The results show that UTM has a similar model size and computation overhead when compared to the competing methods. Additionally, UTM attains an inference speed of approximately 5 frames per second on Jetson Xavier NX and about 7 frames per second on Jetson AGX Xavier. Such latency is already able to support real-time inference and we believe the efficiency can be further improved by leveraging cutting-edge model compression methods (out of the scope of this work).

Table 4: The model size and the inference latency of UTM tested across various embedded devices.

Method	Paras	FLOPs	Frame Per Second on	
			Jetson Xavier NX	Jetson AGX Xavier
VA [7]	7.24M	61.4G	4.94	7.29
AM [24]	7.28M	61.8G	4.88	7.44
Ours: UTM	7.24M	61.4G	4.94	7.21

5 Related work

Object Detection with Thermal Cameras The majority of thermal object detection works are dedicated to human detection and start with the adaptation from the established RGB-based object detection models. The authors in [17] adopt the neural network models trained for RGB images and use transfer learning or cross-domain adaption to detect human subjects captured by thermal cameras. Wager et al. [46] extend the amount of training data by utilizing the RGB red channel of the pure visual Caltech dataset to simulate additional thermal human detection data for more robust human detection. Guo et al. [16] achieved further improvement by combining real and synthetic training data. The authors in [14] tackle the situation that thermal images are less distinguishable due to the insufficient thermal contrast between people and their surroundings in the daytime. They propose to augment thermal images with their saliency maps to serve as an attention mechanism for pedestrian detection. In addition to human subject detection, the authors in [25] utilise the YOLOv5 model architecture [19] for thermal detection on the FLIR dataset and the KAIST dataset. They modified the existing model architecture of the YOLOv5s by taking the CSP Bottleneck and applying an SK attention module. However, due to the limited contrast and textures of thermal images, the robustness of thermal-only solutions remains a question.

Object Detection with mmWave Radars mmWave radars have been increasingly adopted in object detection tasks. These works can be categorised into radar-only solutions and radar-assisted multi-sensor fusion solutions. For radar-only solutions [48, 51], the radio frequency (RF) data is input into neural networks to classify different types of objects on the road. [34] leverages graphical representations of raw radar tensor data to gain a significant improvement in detection accuracy. [10, 45] use radar point cloud for vehicle detection. While the resolution of radar has been improved, the semantic information from the sparse and noisy radar data is noticeably challenging to extract and unsuitable for effec-

tive single-modality object detection. Radar-assisted multi-sensor fusion thus emerges to improve the detection robustness recently. For example, [33, 35, 18] present the camera-radar fusion architecture for accurate 3D object detection in safe autonomous driving. As mmWave radars are impervious to extreme weather conditions in which LiDAR falls short, [26, 41] use radar-LiDAR fusion for vehicle detection. The radar-assisted multi-sensor fusion is also increasingly adopted in pedestrian or human detection. Authors in [24] propose a camera-radar fusion system for pedestrian liveness detection by exploiting the distinct reflection features of real pedestrians and distracting objects. Taking advantage of the penetrability of radar signals and the ability to be unaffected by light, camera-radar fusion object detection system [44, 11] is developed to detect humans in dark environments or through the fog. However, unlike UTM, none of the above works systematically investigates the fusion of thermal cameras and mmWave radars for human detection.

Neural Sensor Fusion Based on the position of the fusion operation in a neural network, existing neural sensor fusion strategies can be categorized into three types: input fusion, feature fusion and decision fusion. Feature fusion is the most widely used in many areas: For disease diagnosis tasks, UncertaintyFuseNet [5] and [4] fuse cross-modal medical images, e.g., CT scans and X-ray images, by concatenating feature maps of two or more branches. For human activity recognition, RFCam [8] fuses Angle-of-Arrival, distance and activity features by proposed similarity scores. For object detection, the fusion between RGB cameras and other sensors, including radar sensors and LiDAR sensors, has been explored to improve detection performance: DeepFusion [23] and PointAugmenting [47] augments points with their corresponding RGB features, [35] enriches RGB images with radar range measurement, CameraRadarFusionNet [37], RVNet [20] and [6] fuses features of RGB branch and radar branch by channel concatenation. [24] and [7] apply attention mechanism to fuse salient features. However, the above sensor fusion strategies usually cover only common sensors, e.g., RGB cameras, radar sensors and LiDAR sensors, while the novel thermal sensors are less explored. Besides, unlike RGB camera that outputs rich RGB information about the environment, both thermal images and radar images are weak in providing multi-channel measurements of objects (1-channel temperature information from thermal images and extremely sparse spatial location from radar images), it remains unknown whether and how thermal cameras and radar sensors can work together for optimum output.

6 Conclusion

By using thermal cameras and mmWave radars, this work presents a novel human detection approach UTM as a robust alternative to RGB camera-based methods under visual degradation. UTM utilizes a Bayesian feature extractor and an uncertainty-guided fusion method to systematically overcome the detection challenges caused by the low-resolution thermal images and the noisy radar point clouds. Our experimental results demonstrate that UTM achieves superior performance compared to the state-of-the-art single-modal and neural sensor fusion methods. For future work, we plan to further improve the runtime efficiency of UTM by using model compression techniques and generalize the uncertainty-guided fusion concept to other sensor combinations.

7 References

[1] Introducing the intel® realsense™ depth camera d455, 2023.

[2] Uncooled, longwave infrared (lwir) oem thermal camera module boson®, 2023.

[3] Vayyar imaging - home, 2023.

[4] M. Abdar, M. A. Fahami, S. Chakrabarti, A. Khosravi, P. Pławiak, U. R. Acharya, R. Tadeusiewicz, and S. Nahavandi. Barf: A new direct and cross-based binary residual feature fusion with uncertainty-aware module for medical image classification. *Information Sciences*, 577:353–378, 2021.

[5] M. Abdar, S. Salari, S. Qahremani, H.-K. Lam, F. Karray, S. Hussain, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. Uncertaintyfusenet: robust uncertainty-aware hierarchical feature fusion model with ensemble monte carlo dropout for covid-19 detection. *Information Fusion*, 90:364–381, 2023.

[6] S. Chadwick, W. Maddern, and P. Newman. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8311–8317. IEEE, 2019.

[7] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng, and Z. Wei. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors*, 20(4):956, 2020.

[8] H. Chen, S. Munir, and S. Lin. Rfcam: Uncertainty-aware fusion of camera and wi-fi for real-time human identification with mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(2):1–29, 2022.

[9] K. Cui, Q. Yang, L. Shen, Y. Zheng, and J. Han. Integrated sensing and communication between daily devices and mmwave radars. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2022.

[10] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer. 2d car detection in radar data with pointnets. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 61–66. IEEE, 2019.

[11] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, and H. Ma. Global-local feature enhancement network for robust object detection using mmwave radar and camera. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4708–4712. IEEE, 2022.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[13] R. Gade and T. B. Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25:245–262, 2014.

[14] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman. Pedestrian detection in thermal images using saliency maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 988–997, 2019.

[15] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019.

[16] T. Guo, C. P. Huynh, and M. Solh. Domain-adaptive pedestrian detection in thermal images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1660–1664. IEEE, 2019.

[17] C. Herrmann, M. Ruf, and J. Beyerer. Cnn-based thermal infrared person detection by domain adaptation. In *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643, pages 38–43. SPIE, 2018.

[18] J.-J. Hwang, H. Kretzschmar, J. Manela, S. Rafferty, N. Armstrong-Crews, T. Chen, and D. Anguelov. Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 388–405. Springer, 2022.

[19] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy, Lorna, Z. Yifu, C. Wong, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, V. Sonck, tkianai, yxNONG, P. Skalski, A. Hogan, D. Nair, M. Strobel, and M. Jain. ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation, Nov. 2022.

[20] V. John and S. Mita. Rvnet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments. In *Pacific-Rim Symposium on Image and Video Technology*, pages 351–364. Springer, 2019.

[21] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[22] M. Krišto, M. Ivasic-Kos, and M. Pobar. Thermal object detection in difficult weather conditions using yolo. *IEEE Access*, 8:125459–125476, 2020.

[23] S. Lee. Deep learning on radar centric 3d object detection. *arXiv preprint arXiv:2003.00851*, 2020.

[24] H. Li, R. Liu, S. Wang, W. Jiang, and C. X. Lu. Pedestrian liveness detection based on mmwave radar and camera fusion. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 262–270. IEEE, 2022.

[25] S. Li, Y. Li, Y. Li, M. Li, and X. Xu. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE Access*, 9:141861–141875, 2021.

[26] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 908–917, 2022.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] T. Liu, M. Gao, F. Lin, C. Wang, Z. Ba, J. Han, W. Xu, and K. Ren. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 2021.

[29] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao. Salient object detection for rgb-d image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019.

[30] Y. P. Loh and C. S. Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019.

[31] C. X. Lu, M. R. U. Saputra, P. Zhao, Y. Almalioglu, P. P. De Gusmao, C. Chen, K. Sun, N. Trigoni, and A. Markham. milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020.

[32] J. X. Lu, J. C. Lin, M. Vinay, P.-Y. Chen, and J.-I. Guo. Fusion technology of radar and rgb camera sensors for object detection and tracking and its embedded system implementation. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1234–1242. IEEE, 2020.

[33] M. Meyer and G. Kuschk. Deep learning based 3d object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*, pages 133–136, 2019.

[34] M. Meyer, G. Kuschk, and S. Tomforde. Graph convolutional networks for 3d object detection on radar data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.

[35] R. Nabati and H. Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2021.

[36] S. Narayana, V. Rao, R. V. Prasad, A. K. Kanthila, K. Managundi, L. Mottola, and T. V. Prabhakar. Loci: privacy-aware, device-free, low-power localization of multiple persons using ir sensors. In *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 121–132. IEEE, 2020.

[37] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–7. IEEE, 2019.

[38] NVIDIA. Nvidia jetson agx xavier developer kit, 2023.

[39] NVIDIA. Nvidia jetson xavier nx developer kit, 2023.

[40] R. Peng and M. L. Sichitiu. Angle of arrival localization for wireless sensor networks. In *2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks*, volume 1, pages 374–382, 2006.

[41] K. Qian, S. Zhu, X. Zhang, and L. E. Li. Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 444–453, 2021.

[42] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A. Y. Ng, et al. Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[43] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[44] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji, and G. Xing. Millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, IoTDI '21, page 145–157, New York, NY, USA, 2021. Association for Computing Machinery.

[45] M. Ulrich, S. Braun, D. Köhler, D. Niederlöhner, F. Faion, C. Gläser, and H. Blume. Improved orientation estimation and detection with hybrid object detection networks for automotive radar. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 111–117, 2022.

[46] J. Wagner, V. Fischer, M. Herman, S. Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016.

[47] C. Wang, C. Ma, M. Zhu, and X. Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11794–11803, 2021.

[48] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 504–513, 2021.

[49] F. Warburg, M. Jørgensen, J. Civera, and S. Hauberg. Bayesian triplet loss: Uncertainty quantification in image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12158–12168, 2021.

[50] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2, 2019.

[51] Z. Zheng, X. Yue, K. Keutzer, and A. Sangiovanni Vincentelli. Scene-aware learning network for radar object detection. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 573–579, New York, NY, USA, 2021. Association for Computing Machinery.