Towards Detecting Knife-Carrying Individuals Using YOLO and mmWave Radar EAR Representations

1st Ogonna Okafor
Department of Computer Science
Nottingham Trent University
Nottingham, United Kingdom
ogonna.okafor@ntu.ac.uk

2nd David Ada Adama
Department of Computer Science
Nottingham Trent University
Nottingham, United Kingdom
david.adama@ntu.ac.uk

3rd Doratha Vinkemeier
Department of Computer Science
Nottingham Trent University
Nottingham, United Kingdom
doratha.vinkemeier@ntu.ac.uk

***Abstract*—Millimetre-wave (mmWave) radar offers a privacy-preserving sensing modality for indoor security applications, but effective learning from radar data remains challenging due to its high dimensionality and non-visual structure. This paper presents an exploratory study on the adaptation of standard vision-based object detectors to radar-derived representations. Specifically, the study investigates whether Elevation-Azimuth-Range (EAR) tensors obtained from a 77 GHz FMCW radar can be transformed into 2D pseudo-images suitable for training off-the-shelf YOLO detectors without architectural modification. Using a small, controlled indoor dataset with weak cross-modal supervision from synchronised RGB cameras, multiple YOLO variants are evaluated under identical training conditions, with performance analysed using both conventional detection metrics and safety-oriented measures such as Miss Rate and False Alarm Rate. The results demonstrate that modern convolutional detectors can learn discriminative patterns from EAR-based pseudo-images despite the absence of explicit geometric calibration between radar and camera coordinate frames, requiring the detector to implicitly learn a cross-modal correspondence. These findings highlight both the feasibility and limitations of representation transfer from radar to vision models. The study is intended as a methodological exploration rather than a deployment-ready system.**

***Index Terms*—mmWave radar, object detection, EAR representation, weak supervision, YOLO, security sensing**

I. Introduction

Computer vision systems [1], [2] are widely deployed in security and surveillance applications; however, their reliance on optical imagery raises concerns related to privacy, sensitivity to lighting conditions, and limited robustness under visual occlusion [3], [4]. Millimetre-wave (mmWave) radar offers an alternative sensing modality that operates independently of illumination, penetrates certain occlusions, and inherently captures less personally identifiable information [5]. These characteristics make radar attractive for privacy-aware indoor sensing. Nevertheless, semantic understanding from radar data remains challenging, as radar measurements are typically represented as high-dimensional tensors encoding spatial and spectral structure that differ fundamentally from photographic images. Consequently, many radar perception approaches [6], [7] rely on hand-crafted features or radar-specific neural architectures, limiting reusability and increasing system complexity.

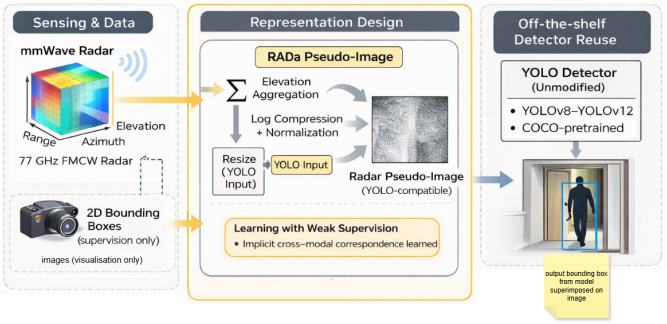

Fig. 1. End-to-end representation-transfer pipeline for radar-based knife-carrying person detection.

In contrast, the computer vision community has developed highly optimised object detection architectures, such as the YOLO family [8], which demonstrate strong performance and efficiency on natural images. An open research question is whether such mature vision-based models can be repurposed for radar data through appropriate representation design, without requiring radar-specific architectural modifications.

This paper addresses this question through a focused feasibility study. The objective is to examine whether Elevation–Azimuth–Range (EAR) radar tensors can be transformed into 2D pseudo-images that preserve sufficient structure for standard convolutional object detectors to learn from. As a case study, the detection of individuals carrying handheld threat-like objects (such as knives) in indoor environments is considered. The aim is not to propose a deployable weapon detection system, but to analyse the representational compatibility between radar data and vision-based detection models under realistic constraints, including limited data availability and weak cross-modal supervision. Unlike fully calibrated radar–camera systems, no explicit geometric alignment between modalities is assumed; instead, the extent to which standard detectors can implicitly learn this correspondence from data is examined. An overview of the proposed representation-transfer pipeline, from raw mmWave radar measurements to YOLO-based knife-

carrying person detection using EAR-derived pseudo-images, is illustrated in Fig. 1. The contributions of this work are as follows: 1) A practical pipeline for converting EAR radar tensors into 2D pseudo-images compatible with off-the-shelf YOLO detectors. 2) An empirical comparison of multiple YOLO variants trained under identical conditions on radar-derived input. 3) An evaluation using safety-oriented metrics, together with an analysis of limitations arising from weak cross-modal supervision and representation choices.

This study is intended as an exploration of the transfer of representation from the mmWave radar to vision-based detectors, rather than a deployment-ready weapon detection system. The remainder of the paper is organised as follows: Section II presents the prior art; Section III presents background and preliminaries; Section IV describes the method; Section V reports results and analysis; and Section VI concludes and highlights future directions. The anonymised dataset supporting the findings of this study is available upon reasonable request.[1] The source code for this study is publicly available.[2]

II. Related Work

Prior work has investigated mmWave radar for carried-object and weapon detection using range–azimuth–elevation (RAE) or volumetric radar representations combined with bespoke deep learning architectures. Gao *et al.* generate RAE image cubes from 77 GHz frequency-modulated continuous-wave (FMCW) radar and train a custom deep network to detect open and concealed carried objects, including knives, using a dedicated prediction and post-processing pipeline [9]. More recently, MMW-Carry expanded this approach by incorporating radar–camera fusion and multi-view temporal aggregation to improve robustness in large indoor environments [10]. In parallel, a growing body of radar perception research converts radar outputs (e.g., range–azimuth, range–Doppler, or RF heatmaps) into image-like representations for object detection, typically employing radar-specific or modified detection networks in automotive and indoor sensing applications [11], [12]. In contrast to these approaches, this work represents mmWave EAR tensors as 2D pseudo-images and trains *unmodified*, off-the-shelf YOLO object detectors directly on these representations for knife-carrying person detection. Rather than designing a radar-specific architecture, the emphasis is placed on evaluating whether representation transfer enables the reuse of mature vision-based detectors for radar-based weapon detection, thereby reducing system engineering complexity and establishing a baseline for this paradigm.

III. Background and Preliminaries

mmWave radar is widely used for indoor human monitoring due to its robustness to occlusion and lighting variations. A commonly used radar modality is the FMCW radar, operating in the 60–77 GHz band [13]. The FMCW radar transmits a periodic wideband linear frequency-modulated signal and measures reflections from targets. The received signal is mixed with the transmitted waveform to generate an intermediate frequency (IF) signal, from which the target's range and angular position can be extracted.

A. Range Estimation

For a single target, the transmitted and received signals differ in frequency by a constant value, producing an IF tone. The frequency of this tone is linearly related to the target's range r as $f_{\mathrm{IF}} = \frac{2rS}{c}$, where S is the frequency sweep slope and c is the speed of light. Multiple targets generate multiple tones in the IF signal. Applying a fast Fourier transform (FFT) along the fast-time (sample) dimension allows efficient extraction of these frequency components, yielding the target ranges.

B. Angle (Direction of Arrival) Estimation

Angle estimation is performed using an antenna array at the receiver. For a target located at angle θ in the far-field, the steering vector for a uniform linear array with N_{Rx} elements is $\mathbf{a}(\theta) = \left[1, e^{-j2\pi d \sin\theta/\lambda}, \ldots, e^{-j2\pi(N_{\mathrm{Rx}}-1)d\sin\theta/\lambda}\right]^T$, where d is the inter-element spacing and λ is the radar wavelength. The phase differences across elements encode the target direction. Direction-of-arrival (DoA) estimation can be accomplished using FFT-based beamforming, or more advanced methods such as MVDR and MUSIC [14], [15]. FFT-based methods are efficient and do not require prior knowledge of the number of targets, making them suitable for real-time applications.

C. Elevation–Azimuth–Range (EAR) Representation

By combining range and angular estimation along both transmit and receive antenna arrays, a three-dimensional tensor of target reflections is obtained, referred to as the EAR representation: $\mathbf{E} \in \mathbb{R}^{N_e \times N_a \times N_r}$, where N_e, N_a, and N_r denote elevation, azimuth, and range bins, respectively. The Doppler dimension is intentionally collapsed in this study. While Doppler information provides valuable motion cues for human activity recognition, incorporating it would require either temporal modelling or higher-dimensional representations that are incompatible with unmodified 2D object detectors. The aim is therefore to isolate the feasibility of representation transfer from spatial EAR structure to standard vision-based detectors. Starting from ADC frames $\mathbf{X} \in \mathbb{C}^{N_{\mathrm{Tx}} \times N_{\mathrm{Rx}} \times N_c \times N_s}$, spatial information is extracted through range FFT, chirp averaging, and angular FFTs. Retaining only magnitude information yields the EAR representation:

$$\mathcal{T}(\mathbf{X}) = \left| \mathcal{F}_{\mathrm{Rx}}\, \mathcal{F}_{\mathrm{Tx}}\, \frac{1}{N_c} \sum_{k=1}^{N_c} \mathcal{F}_s(\mathbf{X})(:,:,k,:) \right|, \qquad (1)$$

such that $\mathbf{E} = \mathcal{T}(\mathbf{X})$. This subsection focuses on signal processing and the physical interpretation of the EAR representation; the deterministic conversion from EAR tensors to YOLO-compatible images is described in Section IV-D.

[1] n1071552@my.ntu.ac.uk
[2] https://github.com/Ogoskino/mmw_localise

D. YOLO Object Detection

You Only Look Once (YOLO) [8] is a family of convolutional neural networks designed for real-time object detection. YOLO treats detection as a single regression problem, predicting bounding boxes and class probabilities directly from input images. Modern YOLO variants [16], [17], [18], [19], [20] incorporate multi-scale feature extraction, anchor-based detection, and optimised loss functions, achieving high accuracy and speed. By converting EAR volumes into 2D pseudo-images, radar measurements can be directly fed into YOLO models without modifying the network architecture. This approach leverages YOLO's spatial feature learning capabilities to identify subtle radar signatures corresponding to carried knives, even under concealment or occlusion.

E. Radar-Native Heatmap Baseline

To contextualise the performance of YOLO-based detectors under weak cross-modal supervision, a radar-native learned baseline is implemented. The baseline consists of a lightweight U–Net-style convolutional network [21] trained to predict a single-channel spatial likelihood heatmap from azimuth-range radar maps. Supervision is provided by filling the camera-derived bounding box corresponding to the *person-with-knife* class, using the same weak annotation procedure as employed for YOLO training. At inference time, bounding boxes are obtained via thresholding and connected-component analysis [22] on the predicted heatmap. The baseline is trained from scratch on radar input only and does not employ object-detection heads, anchor boxes, or pretrained visual features.

IV. Method

A. Testbed Setup and Radar Configuration

Our experimental setup consists of a 77-GHz mmWave radar from Texas Instruments, integrated with an Intel RealSense D455 depth camera, as illustrated in Fig. 2. The radar operates with 12 TX antennas and 16 RX antennas. By employing time-division multiplexing (TDM) across the transmitters, a 2-D virtual array with 192 elements is synthesised, achieving fine azimuth (1.35°) and elevation (19°) resolutions [13]. The radar configuration parameters are summarised in Table I. Based on these parameters, the radar achieves a range resolution of $c/(2B) = 0.06\,\text{m}$ and a maximum detectable range of $f_s c/(2S) = 15\,\text{m}$. Both devices are mounted on a custom stand to ensure an unobstructed field of view (FoV).

B. Dataset

The dataset comprises approximately 4 320 synchronised mmWave radar and RGB camera frames collected across three indoor environments to capture variability in clutter, multipath, and room geometry. Four participants executed scripted walking trajectories in left-to-right and right-to-left directions at ranges of 2.0 m and 4.0 m. Participants carried a mobile phone, a knife, both objects, or no object. Hard negatives included everyday items such as keys, lanyards, and vaseline containers. Each recording session lasted approximately 3 s. Objects were carried openly or concealed within predefined body zones and materials, yielding 24 distinct scenario configurations that span object type, carry style, concealment, walking direction, range and body location. Scenario definitions and environment–participant assignments are summarised in Tables II and III, respectively. Scenario D3 was not recorded due to collection constraints. The multimodal dataset is defined as

$$\mathcal{D} = \{(R_t, I_t, \mathcal{B}_t)\}_{t=1}^{T}, \tag{2}$$

where R_t denotes the radar measurement at time index t, I_t the corresponding RGB frame, and $\mathcal{B}_t$ the set of annotated two-dimensional bounding boxes indicating human presence and weapon status. Radar and camera streams were recorded asynchronously and aligned using hardware timestamps. The camera frames outside the radar acquisition window were removed, bounding the residual temporal misalignment to ≤ 30 ms. Then both streams were uniformly subsampled at a common length using evenly spaced indices, producing an effective frame rate of approximately 15 FPS.

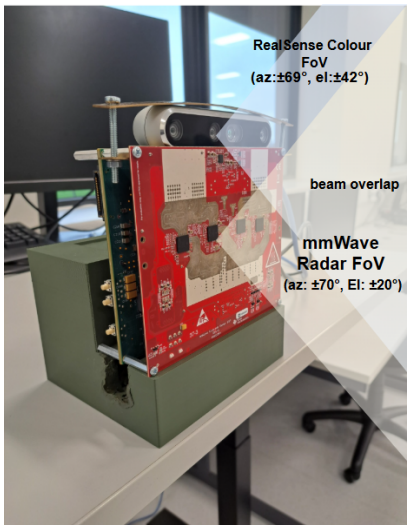

Fig. 2. Experimental testbed showing the fixed mmWave radar and RGB camera setup with overlapping fields of view (FoV) in an indoor environment. The radar coordinate frame is illustrated, where *az* denotes the azimuth angle (horizontal) and *el* denotes the elevation angle (vertical).

TABLE I
FMCW Radar Configuration Parameters

Parameter	Value
Carrier Frequency (f_c)	77 GHz
Bandwidth (B)	3.16 GHz
Frequency Slope (S)	79 MHz/μs
Sampling Frequency (f_s)	8 MS/s
Number of ADC Samples (N_s)	180
Number of Chirps per Frame (N_c)	50
Inter-frame Interval (T_f)	66.7 ms
Transmit Antennas (T_X) number	12
Chirp Interval (I_c)	45 μs
Chirp Duration (T_c)	540 μs

C. Data Splits

To avoid subject-specific leakage, a strict cross-subject split is adopted: participants A and B for training, D for validation,

TABLE II

SCENARIO CONFIGURATIONS. OBJ = OBJECT (PH = PHONE, KN = KNIFE, BT = BOTH, N = NONE); CARRY: O = OPEN, C = CONCEALED; MAT = CONCEALMENT MATERIAL (M1 = LOWER BODY/LIGHT, M2 = UPPER BODY/HEAVY); DIR = WALKING DIRECTION; ZONES: P1 = LEFT HIP/POCKET, P2 = RIGHT HIP/POCKET, P4 = CHEST OR INNER GARMENT POCKET, P5 = HAND-CARRY (OPEN ONLY).

ID	Obj	Carry	Mat	Range (m)	Dir	Zone
1	Ph	O	—	2.0	L→R	P1
2	Ph	O	—	4.0	R→L	P5
3	Ph	C	M1	4.0	L→R	P2
4	Ph	C	M2	2.0	R→L	P4
5	Ph	O	—	4.0	L→R	P5
6	Ph	C	M1	2.0	R→L	P1
7	Kn	O	—	2.0	L→R	P2
8	Kn	O	—	4.0	R→L	P5
9	Kn	C	M1	4.0	L→R	P2
10	Kn	C	M2	2.0	R→L	P4
11	Kn	O	—	4.0	L→R	P1
12	Kn	C	M2	4.0	R→L	P4
13	Bt	O	—	2.0	L→R	P1+P2
14	Bt	O	—	4.0	R→L	P5+P1
15	Bt	C	M1	4.0	L→R	P2+P1
16	Bt	C	M2	2.0	R→L	P2+P4
17	Bt	C	M1	2.0	L→R	P1+P2
18	Bt	C	M2	4.0	R→L	P1+P4
19	N	—	—	2.0	L→R	—
20	N	—	—	4.0	R→L	—
21	N	—	—	4.0	L→R	—
22	N	—	—	2.0	R→L	—
23	Ph	C	M2	4.0	L→R	P4
24	Kn	C	M1	2.0	R→L	P5

TABLE III

ENVIRONMENT × PARTICIPANT ASSIGNMENT. NUMBERS REFER TO SCENARIO IDS; MISSING SCENARIOS: D3. TRAIN: B, D; VALIDATION: A; TEST: C.

Env	A	B	C	D
1	1–8	9–16	17–24	1–8
2	17–24	1–8	9–16	17–24
3	9–16	17–24	1–8	9–16

and C held out entirely for testing. No individual appears in more than one of the training, validation, or test sets.

D. Radar-to-YOLO Processing

To enable reuse of off-the-shelf RGB detectors, each EAR tensor is converted into a deterministic three-channel pseudo-image. The elevation dimension is first collapsed into an azimuth–range representation via aggregation, trading vertical resolution for a compact format compatible with pretrained 2D backbones.

Let $\mathbf{E} \in \mathbb{R}^{N_e \times N_a \times N_r}$ denote an EAR tensor. Elevation is reduced using sum aggregation, $M(i,j) = \sum_{e=1}^{N_e} \mathbf{E}(e,i,j)$, yielding an azimuth–range map

$$M \in \mathbb{R}^{N_a \times N_r}. \tag{3}$$

As shown in Table IV, sum pooling consistently outperforms mean and max aggregation for the safety-critical *person-with-knife* class, achieving the lowest MR and FAR. Max pooling tends to amplify spurious reflections, while mean pooling suppresses weak but persistent returns. Sum aggregation is therefore adopted throughout.

TABLE IV

EFFECT OF EAR-TO-IMAGE REPRESENTATION ON VALIDATION PERFORMANCE *person-with-knife class*. THE BEST RESULTS PER METRIC ARE BOLDED.

3-channel Representation	Precision	Recall	MR ↓	FAR ↓
Mean (replicated)	0.774	0.657	0.343	0.088
Max (replicated)	0.752	0.632	0.368	0.095
Sum (replicated)	**0.803**	**0.676**	**0.324**	**0.075**
Mean–Max–Std (distinct)	0.699	**0.676**	**0.324**	0.133

For numerical stability, logarithmic compression with $\varepsilon = 10^{-6}$ is applied, yielding $L = \log(M + \varepsilon)$. The map L is resized via bilinear interpolation to a fixed exported resolution of $W \times H = 640 \times 480$. Per-frame standardisation, followed by min–max scaling to $[0, 1]$, produces

$$S = \frac{L - \mu_L}{\sigma_L}, \qquad S' = \frac{S - \min(S)}{\max(S) - \min(S)}. \tag{4}$$

The resulting pseudo-image is quantized as $I = \mathrm{uint8}\left(255 \cdot \mathrm{clip}(S', 0, 1)\right)$ and replicated across three channels to match the input dimensionality of the pretrained RGB detector. Alternative three-channel constructions, in which the elevation dimension was collapsed into distinct channels using complementary statistics (mean, maximum, and standard deviation across elevation), were also evaluated as shown in Table IV, but exhibited higher FARs without improving MR and are therefore excluded from the final system. Synchronised RGB bounding boxes $\mathcal{B}_t$ are clipped to exported image bounds and converted to normalised YOLO format (x_c, y_c, w, h), with coordinates normalised by W and H. Performance differences across representations are modest, indicating that representation choice alone does not dominate detection accuracy. As discussed in Section V-A, radar evidence supporting detection is often spatially misaligned with camera-space annotations by a distance comparable to the extent of the object, limiting the benefit of more expressive representations under weak cross-modal supervision.

E. Cross-Modal Supervision

RGB frames are manually annotated using calibrated camera imagery, yielding accurate bounding boxes in image space. These annotations are used to supervise the radar-based detector through temporal synchronisation only; no explicit geometric calibration between the radar and the camera is performed. As a result, while the annotations themselves are accurate, the correspondence between radar features and camera-aligned bounding boxes is indirect. The learning problem therefore requires the detector to implicitly infer a mapping from uncalibrated radar EAR representations to calibrated camera bounding boxes. This weak cross-modal supervision setting reflects realistic deployment constraints and is intentionally adopted in this study.

F. Implementation Details and Training Protocol

Five variants of the Ultralytics YOLO architecture (YOLOv8–YOLOv12) are evaluated, all initialised from COCO-pretrained weights. Radar pseudo-images are exported in standard YOLO format using a fixed, subject-disjoint data

split that is reused across all models to ensure fair comparison. All detectors are fine-tuned for 180 epochs with a batch size of 16 at a fixed input resolution of 480×640. Default Ultralytics optimisation settings, including the optimizer and learning-rate schedule, as well as standard data augmentations (mosaic, random scaling, flipping, and hue/saturation jitter), are used. No architectural modifications are introduced, enabling the analysis to isolate the effects of radar pseudo-image representation and model capacity. All experiments are conducted on a workstation equipped with a NVIDIA GeForce RTX 3080 Ti GPU using the Ultralytics YOLO framework.

G. Evaluation & Metrics

During evaluation, predicted and ground-truth bounding boxes are matched using the Intersection-over-Union (IoU) criterion, with detections considered correct at an IoU threshold of 0.5. Each ground-truth instance is matched to at most one prediction. Performance is measured using Precision, Recall, and F1 score. To reflect safety-critical performance, Miss Rate (MR), which measures the fraction of missed threats, and False Alarm Rate (FAR), which represents the expected number of false alerts per frame, are reported. Detection accuracy is further assessed using Average Precision at IoU 0.5 (AP_{50}) and mean Average Precision averaged over IoU thresholds from 0.5 to 0.95 (AP), following the standard YOLO evaluation protocol. Computational efficiency is reported in frames per second (FPS), measured during inference. To quantify statistical uncertainty in safety-oriented metrics, nonparametric bootstrapping with 1,000 resamples is applied to the evaluation set to estimate 95% confidence intervals (CIs) for MR and FAR. All metrics are computed for the *person-with-knife* class, treated as the safety-critical target in this work.

H. Operating Point (Threshold) Selection

The detector confidence threshold τ is selected on the validation set using Algorithm 1. The optimal threshold $\tau^\star$ maximises the score $\mathcal{S}(\tau)$ and is used for all quantitative evaluations. For qualitative analysis, detections at $\tau^\star$ are overlaid on synchronised RGB frames to assess localisation and failure modes.

Algorithm 1 Detection Threshold Selection

Require: Validation set $\mathcal{D}_{\text{val}}$, thresholds $\mathcal{T}$
Ensure: Optimal threshold $\tau^\star$
$\mathcal{S}_{\max} \leftarrow 0$
for $\tau \in \mathcal{T}$ **do**
  Evaluate on $\mathcal{D}_{\text{val}}$ at τ
  Compute $F1$, MR, FAR
  $\mathcal{S}(\tau) \leftarrow (1 - \text{MR})(1 - \text{FAR})F1$
  if $\mathcal{S}(\tau) > \mathcal{S}_{\max}$ **then**
    $\mathcal{S}_{\max} \leftarrow \mathcal{S}(\tau)$; $\tau^\star \leftarrow \tau$
  end if
end for
return $\tau^\star$

V. Results & Discussion

A. Cross-Modal Supervision Noise Analysis

Although camera annotations are accurate in image space, the absence of explicit radar–camera calibration introduces spatial inconsistency when these labels are used to supervise radar-based detectors. As a result, radar evidence supporting a detection may be spatially offset from the camera-aligned bounding box used as supervision. To quantify the severity of this effect, a proxy measure of cross-modal supervision noise is computed using the radar pseudo-images and camera-space annotations. For each radar frame, a normalised radar energy map $R(x, y)$ is extracted from the EAR-derived pseudo-image prior to quantisation. To suppress low-energy clutter and multipath reflections, the map is thresholded at a high percentile, and an energy-weighted centroid (x_r, y_r) is computed as

$$(x_r, y_r) = \frac{\sum_{x,y}(x, y)\, R(x, y)}{\sum_{x,y} R(x, y)}, \tag{5}$$

where the summation is taken over pixels that exceed the percentile threshold. The corresponding camera annotation provides a bounding box with center (x_b, y_b) and width w_b and height h_b in the same resized image coordinates. The centroid-to-box-center distance is then computed and normalised by the bounding-box diagonal,

$$d_{\text{box}} = \frac{\sqrt{(x_r - x_b)^2 + (y_r - y_b)^2}}{\sqrt{w_b^2 + h_b^2}}, \tag{6}$$

yielding a scale-invariant measure of cross-modal misalignment. Fig. 3 shows the distribution of d_{box} for *person-with-knife* instances in the evaluation set. The median misalignment corresponds to a substantial fraction of the bounding-box diagonal, with a long tail extending beyond one full diagonal. This indicates that, in many frames, the radar evidence supporting detection is spatially offset from the camera-space annotation by an amount comparable to the object extent itself. Such supervision noise fundamentally constrains learning under weak cross-modal alignment and helps explain the modest

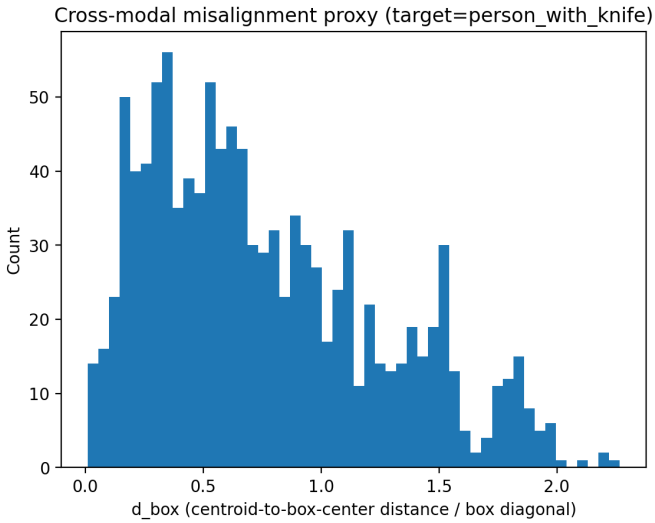

Fig. 3. Distribution of the normalised centroid-to-box distance (d_{box}) for *person-with-knife* instances. The metric measures the distance between a radar-derived energy centroid and the camera-space bounding-box center, normalized by the bounding-box diagonal. Larger values indicate greater cross-modal misalignment.

TABLE V

DETECTION PERFORMANCE ACROSS YOLO VARIANTS. **PANEL A**: OVERALL AVERAGE PERFORMANCE ACROSS ALL EVALUATION SCENARIOS. **PANEL B**: SCENARIO-WISE MISS MR AND FAR. THE BEST RESULTS PER METRIC ARE BOLDED. BOOTSTRAPPED 95% CONFIDENCE INTERVALS ARE REPORTED WHERE APPLICABLE.

Panel A: Overall Performance

Model	$\tau^\star$	$\mathbf{AP}_{50}$	**AP**	**MR (95% CI)** ↓	**FAR (95% CI)** ↓	**F1 (95% CI)**	**FPS**
YOLOv8	0.37	0.679	0.480	0.124 [0.095, 0.150]	**0.031 [0.021, 0.041]**	**0.906 [0.884, 0.924]**	53.9
YOLOv9	0.27	0.648	0.459	0.117 [0.090, 0.143]	0.046 [0.033, 0.059]	0.896 [0.873, 0.919]	77.2
YOLOv10	0.34	0.665	**0.506**	0.114 [0.087, 0.141]	0.093 [0.074, 0.113]	0.858 [0.830, 0.883]	**110.1**
YOLOv11	0.43	**0.695**	0.496	0.111 [0.086, 0.135]	0.049 [0.037, 0.062]	0.897 [0.871, 0.919]	102.0
YOLOv12	0.38	0.683	0.500	**0.093 [0.068, 0.113]**	0.062 [0.047, 0.079]	0.896 [0.872, 0.920]	85.6

Panel B: Scenario-wise MR / FAR

Model	**CK**	**LC**	**TM**	**OC**	**4 m**	**2 m**
YOLOv8	0.166 / **0.052**	0.274 / **0.055**	0.057 / 0.048	0.048 / **0.009**	0.065 / 0.031	0.179 / **0.031**
YOLOv9	0.155 / 0.072	0.257 / 0.106	0.051 / **0.037**	0.048 / 0.019	0.061 / **0.022**	0.168 / 0.071
YOLOv10	0.138 / 0.109	**0.212** / 0.146	0.062 / 0.071	0.072 / 0.075	0.083 / 0.088	**0.144** / 0.097
YOLOv11	0.146 / 0.074	0.223 / 0.109	0.068 / **0.037**	0.048 / 0.023	0.069 / 0.027	0.151 / 0.071
YOLOv12	**0.130** / 0.096	0.218 / 0.150	**0.040** / 0.041	**0.029** / 0.026	**0.036** / 0.025	0.147 / 0.100

Scenario definitions: CK = concealed knife; LC = light clothing concealment; TM = thick material concealment; OC = open carry; 4 m / 2 m = target range from radar.

TABLE VI

COMPARISON WITH RADAR-NATIVE HEATMAP BASELINE FOR THE *person-with-knife* CLASS. THE BEST RESULTS PER METRIC ARE BOLDED

Model	IoU	F1	MR ↓	FAR ↓
Heatmap Baseline	0.10	0.53	0.23	0.60
Heatmap Baseline	0.30	0.01	0.99	0.99
Heatmap Baseline	0.50	0.00	1.00	1.00
YOLOv8	0.50	**0.91**	**0.12**	**0.03**

gains observed from richer radar representations and additional cues in subsequent experiments.

B. Quantitative Results and Analysis

Table V summarises the detection performance of five YOLO variants trained under identical conditions on radar-derived pseudo-images.

1) Overall Performance: All evaluated YOLO variants achieve non-trivial performance, with AP values ranging from 0.459 to 0.506, indicating that standard vision-based detectors can learn meaningful structure from EAR radar pseudo-images without radar-specific architectural modifications. This supports the feasibility of representation transfer from mmWave radar to convolutional object detectors. YOLOv10 attains the highest AP at 0.506 but exhibits the highest FAR at 0.093, suggesting a tendency toward over-detection. In contrast, YOLOv12 achieves the lowest overall MR at 0.093, 95% CI [0.068, 0.113], reflecting improved sensitivity at the cost of moderately increased FAR. YOLOv8 and YOLOv11 provide more balanced performance, combining low FAR at 0.031 and 0.049, respectively, with stable F1 scores exceeding 0.89. YOLOv10 achieves the highest inference speed at 102 FPS but shows slightly reduced precision-related metrics, indicating a trade-off between throughput and robustness under cross-modal representation shift.

The radar-native baseline achieves reasonable coarse agreement at relaxed IoU thresholds (e.g., IoU ≥ 0.1) but fails under stricter localisation criteria, as shown in Table VI. In contrast, YOLO-based detectors maintain substantially lower miss and false alarm rates at IoU $= 0.5$, indicating more precise spatial localisation despite weak cross-modal supervision.

2) Scenario-wise Analysis: The Scenario-level results in Table V (Panel B) reveal consistent trends between models. Light clothing concealment yields the highest MR values, confirming it as the most challenging scenario. Open carry and thick material concealment are comparatively easier, particularly for YOLOv12, which achieves the lowest MR across all evaluated conditions. Detection performance is consistently lower at 2 m than at 4 m between all models. This behaviour is likely attributable to representation artefacts introduced during tensor-to-image projection at close range, indicating that performance limitations arise not only from model capacity but also from the chosen radar representation.

C. Qualitative Analysis

Fig. 4 demonstrate the behaviour of the proposed *person-with-knife* detection systems in a range of indoor environments, carrying conditions, and target ranges. The results indicate that the detector can reliably identify the presence of a human subject carrying a knife across diverse scenes, including office-like environments, cluttered laboratory spaces, and controlled settings. The consistency of detections across these environments suggests that the learned representations are not overly dependent on a single background configuration, despite the use of radar-derived pseudo-images and weak cross-modal supervision. A key strength observed is the robustness of the model to variations in body orientation and

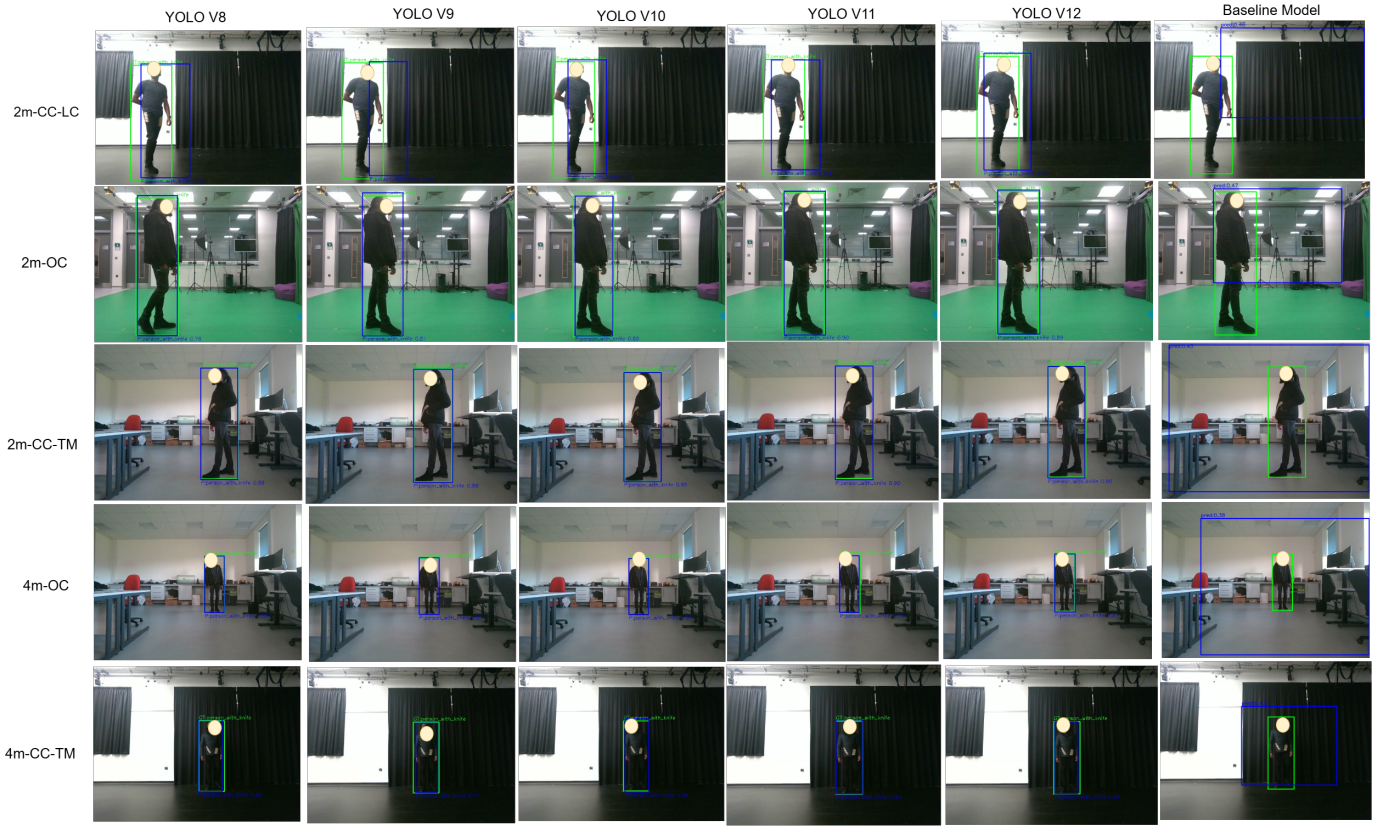

Fig. 4. Qualitative detection results for the *person-with-knife* class across YOLOv8–YOLOv12 and the radar-native heatmap baseline under different scenario conditions: OC (open carry), CK (concealed knife), LC (light clothing concealment), and TM (thick material concealment), at 2 m and 4 m ranges. Green boxes indicate ground truth; blue boxes denote model predictions.

walking direction. Subjects are detected in frontal, profile, and partially rotated poses, with bounding boxes remaining spatially stable between frames. This indicates that the detector relies on coarse, holistic radar signatures associated with human presence and carried objects, rather than pose-specific or limb-level cues. This behaviour is consistent with the EAR-to-image projection, which preserves global spatial energy patterns but suppresses fine-grained geometry. In contrast, the radar-native U-Net–style baseline performs substantially worse producing diffuse or fragmented activations that lead to unstable or missed bounding boxes. This reflects sensitivity to noise and the lack of multi-scale feature aggregation and explicit bounding-box regression under weak cross-modal supervision.

VI. Conclusion

This paper explored whether off-the-shelf vision-based object detectors can be applied to mmWave radar data by converting EAR tensors into 2D pseudo-images. Results show that standard YOLO models can learn meaningful structure from radar-derived representations under weak cross-modal supervision, without radar-specific architectural modifications, despite the absence of explicit geometric calibration between radar and camera coordinate frames. Analysis indicates that performance is limited primarily by representation design and cross-modal supervision noise rather than detector capacity, with diminishing returns from increased model complexity. These findings suggest that, in multimodal learning under realistic sensing constraints, improving representation compatibility and supervision fidelity is more impactful than adopting increasingly sophisticated detectors. Future work will focus on learning strategies that address these limitations, including temporal modelling and Doppler-aware representations.

References

[1] M. T. Bhatti, M. G. Khan, M. Aslam, and M. J. Fiaz, "Weapon detection in real-time cctv videos using deep learning," *Ieee Access*, vol. 9, pp. 34 366–34 382, 2021.

[2] S. Narejo, B. Pandey, D. Esenarro Vargas, C. Rodriguez, and M. R. Anjum, "Weapon detection using yolo v3 for smart surveillance system," *Mathematical Problems in Engineering*, vol. 2021, no. 1, p. 9 975 700, 2021.

[3] X. Gao, G. Xing, S. Roy, and H. Liu, "Ramp-cnn: A novel neural network for enhanced automotive radar object recognition," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5119–5132, 2020.

[4] O. Okafor, D. A. Adama, and D. Vinkemeier, "Machine learning-based classification of open carry and concealed dangerous object using mmwave radar data features," in *UK Workshop on Computational Intelligence*, Springer, 2025, pp. 501–514.

[5] J. Zhang et al., "A survey of mmwave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2052–2087, 2023.

[6] K. Mitchell et al., "Mmsense: Detecting concealed weapons with a miniature radar sensor," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[7] Y. Wang, J. Su, H. Murakami, and M. Tonouchi, "Pointnet++ based concealed object classification utilizing an fmcw millimeter-wave radar," *Journal of Infrared, Millimeter, and Terahertz Waves*, vol. 45, no. 11, pp. 1040–1057, 2024.

[8] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[9] X. Gao, H. Liu, S. Roy, G. Xing, A. Alansari, and Y. Luo, "Learning to detect open carry and concealed object with 77 ghz radar," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 791–803, 2022. DOI: 10.1109/JSTSP.2022.3171168.

[10] X. Gao, Y. Luo, A. Alansari, and Y. Sun, "Mmw-carry: Enhancing carry object detection through millimeter-wave radar–camera fusion," *IEEE Sensors Journal*, vol. 24, no. 9, pp. 15 091–15 100, 2024.

[11] A. Sonny, A. Kumar, and L. R. Cenkeramaddi, "Carry object detection utilizing mmwave radar sensors and ensemble-based extra tree classifiers on the edge computing systems," *IEEE Sensors Journal*, vol. 23, no. 17, pp. 20 137–20 149, 2023.

[12] 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*, 2021, pp. 504–513.

[13] Texas Instruments, *Tiduen5a: 60-ghz mmwave sensor evaluation module user's guide*, Document TIDUEN5A, Texas Instruments, 2020. [Online]. Available: https://www.ti.com/lit/ug/tiduen5a/tiduen5a.pdf.

[14] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[15] J. McWhirter and T. Shepherd, "Systolic array processor for mvdr beamforming," in *IEE Proceedings F (Radar and Signal Processing)*, IET, vol. 136, 1989, pp. 75–80.

[16] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics yolov8*, version 8.0.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[17] C.-Y. Wang and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," 2024.

[18] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al., "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.

[19] G. Jocher and J. Qiu, *Ultralytics yolo11*, version 11.0.0, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics.

[20] Y. Tian, Q. Ye, and D. Doermann, "Yolo12: Attention-centric real-time object detectors," *arXiv preprint arXiv:2502.12524*, 2025.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[22] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2018.