Real-time Concealed Weapon Detection on 3D Radar Images for Walk-through Screening System

Nagma S. Khan[1] khan_nagma@nec.com
Kazumine Ogura[1] k-oguraay@nec.com
Eric Cosatto[2] cosatto@nec-labs.com
Masayuki Ariyoshi[1] m.ariyoshi@nec.com

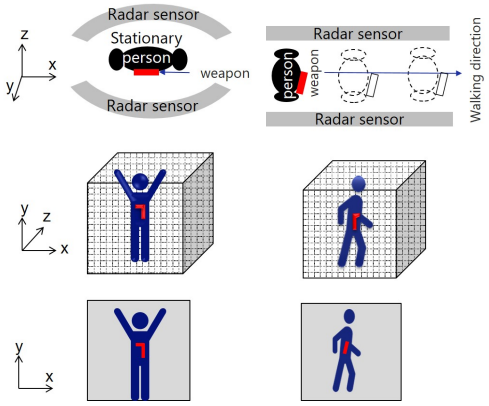
[1]R&D Division, NEC Corporation, Japan
[2]NEC Laboratories America, Inc.

Abstract

This paper presents a framework for real-time concealed weapon detection (CWD) on 3D radar images for walk-through screening systems. The walk-through screening system aims to ensure security in crowded areas by performing CWD on walking persons, hence it requires an accurate and real-time detection approach. To ensure accuracy, a weapon needs to be detected irrespective of its 3D orientation, thus we use the 3D radar images as detection input. For achieving real-time, we reformulate classic U-Net based segmentation networks to perform 3D detection tasks. Our 3D segmentation network predicts peak-shaped probability map, instead of voxel-wise masks, to enable position inference by elementary peak detection operation on the predicted map. In the peak-shaped probability map, the peak marks the weapon's position. So, weapon detection task translates to peak detection on the probability map. A Gaussian function is used to model weapons in the probability map. We experimentally validate our approach on realistic 3D radar images obtained from a walk-through weapon screening system prototype. Extensive ablation studies verify the effectiveness of our proposed approach over existing conventional approaches. The experimental results demonstrate that our proposed approach can perform accurate and real-time CWD, thus making it suitable for practical applications of walk-through screening.

1. Introduction

Security concerns in public places have increased the demand for systems that can detect concealed hand-held weapons to inhibit terrorist activities. Concealed weapon detection (CWD) is typically performed by body scanners which employ radar imaging technology to visualize concealed items on a person. Conventional body scanners, such as the ones widely used at airports, require the person to be stationary and adopt a pre-defined pose during the scan, as shown in Fig. 1a [16, 17]. This makes the scanning procedure time-consuming, leading to low throughput. Therefore, they are not suitable to be deployed in crowded public places like railway stations, malls, etc, which require fast scanning. To overcome these problems, we developed a walk-through screening system prototype (Fig. 1b) capable of performing CWD on walking persons.



(a) Stationary (b) Walk-through

Fig. 1: Stationary vs walk-through scaninng: *Top row*: Top-view of the scan process. *Middle row*: 3D radar image of the scan, which is composed of ordered voxels. *Bottom row*: Corresponding z-axis projected 2D image.

The CWD requirements of walk-through screening are two-folds: real-time processing and accuracy. Since the target is a walking person, the occlusion of weapon can happen by arms or legs leading to reduced accuracy of detection. This problem can be tackled if the system can scan at a high frame rate so as to capture the changing poses of a walking person. Based on the transmission time of the radar sensors, our system has a scan rate of 20fps (details in Section 2). Each scan generates a radar image frame and so, the detection approach needs to have a run-time of 20fps to inspect all the frames. Otherwise, any dropped frame can risk missing the detection of weapon. Therefore, a real-time (20fps) CWD approach is required for our system. Some existing approaches [8, 18] have shown that convolutional neural networks (CNN) are well-suited to the task of CWD from radar images. However, these approaches project the originally 3-dimensional (3D) radar image to 2D prior to detection. The 2D projection based-approach is suitable for

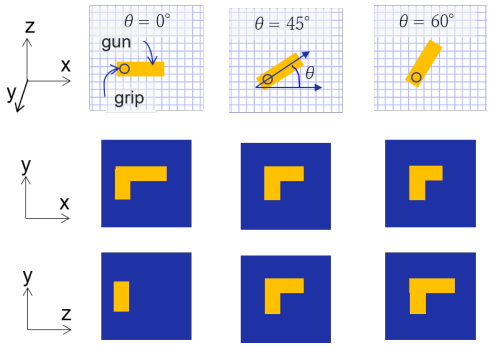
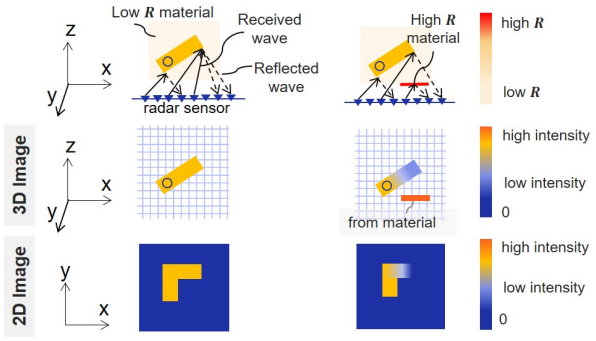


Fig. 2: Top view of different orientations of the weapon (top row), corresponding z-axis projected image (middle row), and x-axis projected image (bottom row).



(a) Easy case (b) Difficult case

Fig. 3: Received reflections from the gun (top row), Intensity distribution of gun in 3D image and in 2D projected image respectively (middle and bottom row).

conventional stationary scanning, as shown in Fig. 1a. In this case, the scan surface area is large as it is frontal scanning. Thus, the projection of 3D image to 2D conserves the shape of the target to be detected. Whereas in walk-through scanning (Fig. 1b), the scan surface area is smaller due to lateral scanning. Therefore, 2D projection leads to the loss of signature shape of the gun, as shown in the bottom row of Fig. 1b. Fig. 2 further illustrates how changing orientation of the gun, carried by the human, impacts the loss of shape information in the 2D projected image.

This shape loss is worsened when intensity of weapon's concealing material becomes higher compared to the weapon. This can happen if whole or part of material has stronger reflectance due to the presence of metal tags, zips, etc. Consequently, the 2D projection operation ignores contribution from the weapon as, by design, it favors the selection of higher intensity regions i.e. material. The reflectance R refers to the effectiveness of a surface to reflect the received radar waves. As shown in Fig. 3b, the right side of the gun has weaker reflection due to presence of high R material, thus comparatively lowering its intensity in the resulting 3D radar image. Hence, the gun's right side shape is lost in 2D projection along z-axis as its intensity was lower compared to that of the material. Therefore, considering weapon's shape information loss during 2D projection, an accurate detection approach should process the 3D radar image directly. Thus, a real-time (run-time$\leq$ 50ms) 3D detection approach is required for our walk-through weapon screening system.

While research in 3D object detection has progressed lately, the majority of the studies are designed for orderless point clouds [6]. Such approaches are not appropriate for the ordered 3D radar image as they do not consider the spatial relationship between the neighboring voxels. On the other hand, the 3D CNN based object detection networks capture such spatial relationship and are often utilized in the medical domain [20]. In many instances, they are based on computationally-intensive Region Proposal Network (RPN) and are therefore not well suited for our severe real-time constraints. Thus, to the best of our knowledge, there is no existing approach for 3D object detection which fulfils our requirements of real-time and accuracy.

This paper proposes a framework for real-time and accurate CWD on 3D radar images for walk-through weapon screening system. More specifically, we reformulate a 2D segmentation network U-Net [15] for a 3D detection task. U-Net has a fully-convolutional architecture which can be readily extended to 3D. Additionally, U-Net has a proven track record in challenging domains such as the biomedical field [5]. To facilitate positional inference, our 3D U-Net implementation is trained to predict a 3D peak-shaped probability map instead of the usual voxel-wise label map. In the probability map, a peak signifies the presence of a weapon, so the weapon detection task translates to peak detection on the map. The peak detection operation can be efficiently implemented in 3D, thus providing the fast detection output. To prove the effectiveness of the proposed framework, we evaluate it on a realistic walk-through 3D radar image dataset. Extensive experimental evaluations confirm that the proposed real-time CWD approach fulfils the real-time (20fps) and accuracy requirements for use in a walk-through weapon screening system.

2. Walk-through weapon screening system

Our walk-through weapon screening system is a body scanner capable of performing CWD on a walking person. Apart from body scanners, closed-circuit television (CCTV) [10] and X-Ray scanners [3] may be used for security applications but they are outside the scope of CWD. CCTV based technologies generally employ optical camera sensors thus making them unsuitable to detect concealed weapons. And, X-Ray based technologies employ high power waves which are harmful for humans. Hence, both CCTV and X-Ray are unsuitable for CWD.

2.1. Basics of CWD

Body scanners, which perform CWD, employ low power safe waves like radio and hence are suitable to scan the hu-

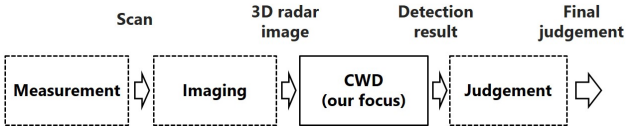


Fig. 4: Processing flow of a body scanner performing CWD. Our proposed framework focuses on the CWD block, whose input is 3D radar image and output is detection result.

Table 1: Main difference points of our walk-through screening vs conventional stationary body scanner

CWD system	Throughput (person/H)	Pose restriction
Stationary [8, 18]	300	Exists
Walk-through	2400	None

man for concealed dangerous objects like weapons. Generally, systems performing CWD have a processing flow as shown in Fig. 4. First, in measurement or scanning step, the radar antennas transmits radio waves and receives their reflections from a subject. Second, in imaging step, the acquired scan is synthesized into a 3D radar image frame. The radar image is in the form of a 3D voxelized cuboid, as shown in middle row of Fig. 1. Third, in detection (CWD) step, each frame of the radar image is inspected to detect concealed weapons by either a human operator or by automated CWD e.g., artificial intelligence based detection. Our proposed framework falls under automated CWD, as automation is required to support the high throughput of our walk-through screening system. Hence, references to CWD in the following sections of this paper implies automated CWD. Finally, in judgement step, the frame-based detection result is integrated to perform a final judgement, as to whether the weapon is detected on the person or not.

2.2. System overview

Our walk-through weapon screening system is a walk-through version of the body scanner, as opposed to the conventional stationary version (Fig. 1). Major differentiating points are summarized in Table 1. Our system consists of two parallel radar sensor panels, forming a gate, between which the person walks during scanning, as shown in the top row of Fig. 1b. Each of the panel contain multiple antennas spread across it, which transmit and receive radio waves at different frequencies to scan the 3D space between them. All these antennas must transmit sequentially in order to prevent any signal interference, and then receive the reflected wave to acquire a single scan. The sequential transmission constrains the scanning speed of the system to 20fps i.e., 50ms per scan. The acquired scan is synthesized into a 3D radar image $I \in \mathbb{R}^3$ using beam-forming algorithm [2], as per Eq. 1.

$$I(P) = |\sum_{\forall f}\sum_{\forall R}\sum_{\forall T} s(T,R,f)e^{j\frac{2\pi f}{c}(r_{PT}+r_{PR})}| \quad (1)$$

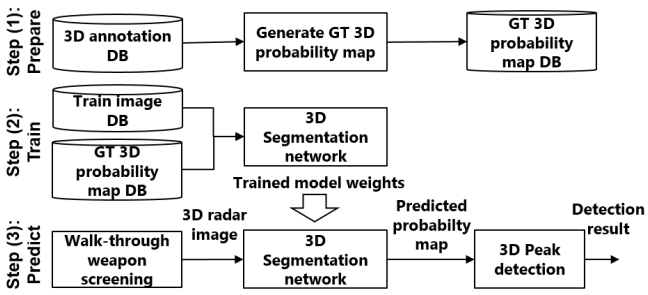


Fig. 5: Overview of the real-time 3D object detection framework.

Here $P \in \mathbb{R}^3$ defines a point in 3D space i.e., voxel, $I(P)$ is the absolute value of the radar image at Pth voxel, $T \in \mathbb{R}^3$ is the transmitter's position, $R \in \mathbb{R}^3$ is the receiver's position, f is the frequency of radar wave, c is the speed of light, $r_{PT} = ||P - T||_2$, $r_{PR} = ||P - R||_2$ and $s(.)$ is the acquired radar signal. The $|.|$ operation takes the absolute value and $||.||_2$ calculates the L2 norm. The synthesized 3D radar image I is given as input for CWD which outputs the location of detected weapon(s), if any.

3. Related Work

Concealed Weapon Detection (CWD). There is only a handful of available research on CWD. In one of the work [4], the radar image is divided into 2D patches. Patch-wise SIFT features are extracted and fed to a support vector machine (SVM) for classification into weapon/no-weapon. Due to the patch based nature of this approach, it cannot meet the real-time constraints of walk-through screening. Recent works in CWD [8, 18] use deep learning based methods and can achieve good performance with real-time processing speed. However, these approaches project the 3D image to 2D prior to detection, thus losing weapon shape information in walk-through screening, as shown in Fig. 1b. Thus, existing work in CWD are unsuitable to meet the requirements of walk through weapon screening. Hence, the motivation for this work.

3D Deep Learning. Almost all of the existing research in 3D deep learning is either in point cloud domain [6] or medical domain [1, 20]. The point cloud does not have a regular grid-like structure as the radar image, thus the approaches designed for point cloud mostly employ multi-layer perceptrons (MLP) [7] which do not consider the spatial order of the input. A few approaches [19, 9, 11] which do consider the spatial order by converting the point cloud to ordered 3D voxelized format are designed for classification tasks as opposed to detection. Zhao and Tuzel [21] attempted point cloud detection but the feature extraction backbone was partially designed with multi-layer perceptron so it ignored the 3D spatial relationship.Thus, point cloud 3D object detection approaches are not appropriate for our task.

The existing 3D deep learning approaches designed for medical data take 3D ordered data format as input, similar to ours, but they are not suitable for CWD application as explained next. The system proposed in [1] is designed for pixel-wise segmentation as opposed to detection, while the system proposed in [20] has large processing time due to the use of 3D RPN.

4. Method

In this section, we introduce our proposed approach to perform real-time and accurate CWD on 3D radar images for walk-through screening. Initially, the overview of the proposed method is provided. That is followed by details of the probability map's design, network architecture and loss function.

4.1. Overview of Proposed Method

In order to adapt the 3D segmentation network for detection, it needs to be trained to predict probability maps. The ground truth (GT) 3D probability maps are prepared first as shown in Step (1) of Fig. 5. This step utilizes the 3D annotation which contains the GT position and size information of the weapon; more details will be presented in Section 4.2. Next, in Step (2), the generated GT 3D probability maps and the corresponding radar images are utilized to train our real-time 3D segmentation network, details of which are introduced in 4.3 and 4.4. In Step (3), the trained network first generates a predicted probability map on a radar image, and then a peak detection operation is used to detect and localize the peaks, i.e., weapons, on the map. The peak detection is a simple mathematical operation which analyzes the gradient in the probability map to localize maxima if present. The peak detection operation finally provides the detected peak as prediction result after ensuring that the peak's value is higher than a pre-decided detection threshold. In a multi-class setup, the peak detection operation is performed for each of the output class maps

Although the proposed detection approach is capable of localizing multiple weapons per image by virtue of a multi peak shaped probability map, the rest of the paper will focus on single weapon detection per image.

4.2. Design of Probability Map

We use a peak-shaped Gaussian function to model the weapon in the probability map. The 3D annotations for the weapon, which are available as 3D bounding boxes, are used to generate the Gaussian function on the GT probability map. The bounding box center $X_i \in \mathbb{R}^3$ denotes its mean $\mu_i \in \mathbb{R}^3$, whereas the bounding box's dimension $l_i \in \mathbb{R}^3$ determines its standard deviation $\sigma_i \in \mathbb{R}^3$ for the ith image as

$$\sigma_i = kl_i, \forall i \in 1, 2, \ldots, N \tag{2}$$

Here, N is the total number of images and $k \in \mathbb{R}$ is a proportionality factor which controls the size of the Gaussian function by adjusting it's σ_i along each dimension in order to capture the weapon shape appropriately. In mathematical terms, k controls the extent to which the Gaussian function overlaps with the bounding box; e.g., for $k = 0.25$, the $[-2\sigma, 2\sigma]$ range of the Gaussian function overlaps with the bounding box. The values in the probability map lie between $[0, 1]$, where 1 indicates the peak of the Gaussian function. In the case when a weapon is not present in a radar image, the corresponding probability map consists of all zeros. In our implementation, the 3D bounding box dimensions are tightly adjusted around the weapon and aligned with the x, y, and z axis.

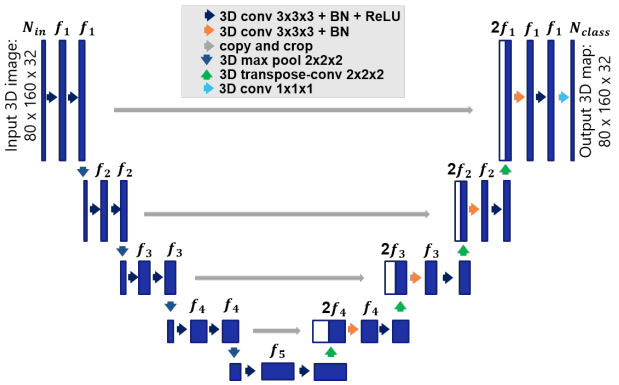


Fig. 6: Proposed 3D network's architecture. Here, $f_1 = 32$, $f_2 = 64$, $f_3 = 128$, $f_4 = 256$, and $f_5 = 512$.

4.3. 3D Segmentation Network

To achieve real-time CWD on 3D radar images, the 2D U-Net segmentation network [15] is adapted to provide detection output. The U-Net is first extended to 3D by replacing all 2D operations (convolution, transposed convolution, and max-pooling) with their 3D versions. To meet the real-time requirement, we further reduce the number of trainable weights to one-third by halving the feature map count. Thus, we obtain our real-time 3D segmentation network as shown in Fig. 6, where N_{in} is the number of input channels and N_{class} is the number of classes. Lastly, this network is trained to generate a peak-shaped 3D probability map to facilitate weapon detection through peak detection.

4.4. Loss Function

Our 3D segmentation network is trained using the loss,

$$L = -\sum_{v \in \Omega} \sum_{c \in N_{class}} w_v y_v^c log(p_v^c) \tag{3}$$

Here p_v^c is the predicted probability map's value and y_v^c is the GT probability map's value for the cth class channel's vth voxel in $\Omega \in \mathbb{R}^3$. Here, w_v is the weight for the vth voxel, $w_v > 1$ for the voxels where $y_v^c > 0$, otherwise $w_v = 1$. The weight helps deal with severe class imbalance as the weapon occupies only 2–3% of the total volume in a 3D image.

(a) Model guns (b) With-weapon scenarios

Fig. 7: (a) Model guns: (top) Pistol and (bottom) Revolver; (b) Experiment scenarios: (left) gun concealed in bag, and (right) gun concealed on waist.

Table 2: Data distribution of the radar image dataset

Data split	With-weapon images	No-weapon images
Train	9800	10000
Test	2400	2500
Validation	2500	-

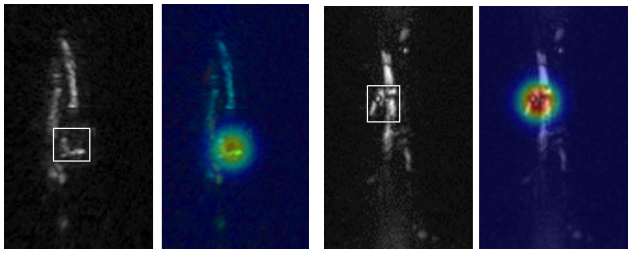
5. Experimental Evaluation

In this section, we first introduce our dataset and evaluation method. Then, we validate our proposed 3D approach's suitability for walk-through CWD by comparing it with other candidate architectures, existing 3D-RCNN and 2D approaches. Finally, we present extensive ablation studies to validate the effectiveness of the proposed approach.

Dataset. Our proposed method is evaluated on a radar image dataset collected using our walk-through weapon screening system prototype. Multiple subjects walked through the screening system, either carrying a concealed weapon (with-weapon) or without weapon (no-weapon). In the with-weapon scenarios, the weapon is concealed in realistic positions e.g., inside bags, in holsters at the side of the waist, etc. Two types of model guns as shown in Fig. 7a - a pistol and a revolver, are used as target weapons. In the no-weapon scenarios, the subjects carried daily-use items like cell-phone, laptop, water bottle, etc, to make the evaluation more realistic. The experimental scenarios are shown in Fig. 7b and the data distribution is given in Table 2. The validation set was used to tune the class-weight w_v and to choose the best network in terms of validation loss for evaluation. 2D visualizations of the probability map and their corresponding bounding box labels are shown in Fig. 8.

Evaluation method. The network parameters are set as single input channel i.e. $N_{in} = 1$ which is radar image's intensity, and $N_{class} = 1$ as there in only a gun class for this experiment. To generate the Gaussian function, $k = 0.25$ is used in Eq. 2 as it was experimentally determined to be optimal. The weapon detection output is obtained as shown in Step (3) of Fig. 5.

During the prediction phase, the trade-off between the true positive rate (TPR) and false positive rate (FPR) is controlled by means of a detection threshold. To compare the performances of different approaches, we use the Area-under-Curve (AUC) metric of the receiver operating characteristics (ROC) curve. The ROC curve is generated by evaluating the TPR and FPR at different detection thresholds. TPR, or more commonly known as recall measures an approach's capability to detect the weapon when it is present in the image. Thus, recall is evaluated on the with-weapon images as mentioned next - if peak i.e. weapon is detected and predicted peak's position is inside the GT bounding box then we consider it to be true positive (TP) else a false negative (FN). Whereas in CWD, the aim of FAR is to evaluate the proportion of false alarms i.e., instances where weapon is wrongly detected in a no-weapon image. Hence, FAR is evaluated on only no-weapon images as follows: if a peak is detected in a no-weapon image then its considered a false positive (FP) else a true negative (TN). Since high FAR in CWD is detrimental to its practical applicability, we use the partial AUC [12] i.e., AUC_{eff} as our accuracy metric by limiting the range of FAR between 0% and 10%. The AUC_{eff} is an appropriate overall metric of CWD accuracy as it captures the balance between the recall and FAR for various detection thresholds. We also present the more comprehensive ROC curves for finer comparisons.



(a) Gun in bag (b) Gun on waist

Fig. 8: Sample radar images (2D visualization): Left image shows the bounding box label and the right image shows the Gaussian function of the probability map.

5.1. Comparison of accuracy and run-time

We experimentally show that our proposed approach is the best choice for our walk-through screening system in terms of run-time and accuracy requirements. Run-time is defined as the time taken to output the detection result after the input 3D image is provided and AUC_{eff} is used as a measure of accuracy. We compare our proposed 3D architecture with other candidate architectures. The network architectures used in this comparison are based on the U-Net of Fig. 6. The feature map count f_i for $i \in \{1, 2, ..., 5\}$ differs for each candidate architecture, as summarized in Table 3. The proposed 3D network's accuracy and run-time is compared with the default U-Net [15] (3D-UNet-default) and its variations where default feature map count is reduced by 75% (3D-UNet-3-by-4), and 25% (3D-UNet-1-by-4), as well as the 3D-RPN [20] and 2D approaches (2D-z_int and 2D-multi-view) for completeness. The comparative evaluation results are as shown in Fig. 9. The details

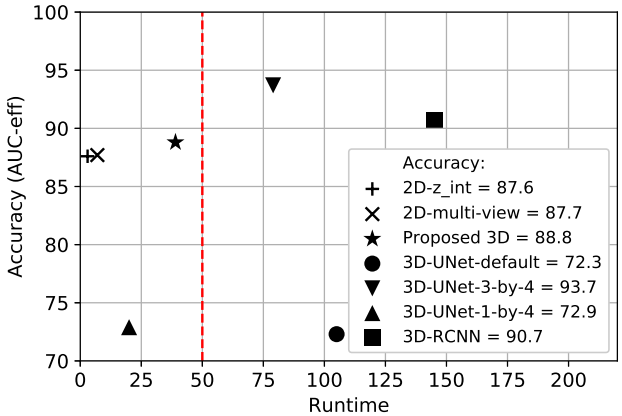


Fig. 9: Comparison of accuracy vs run-time of candidate architectures, the red dotted line signifies the walk-through system's run-time requirement of 50ms (20fps).

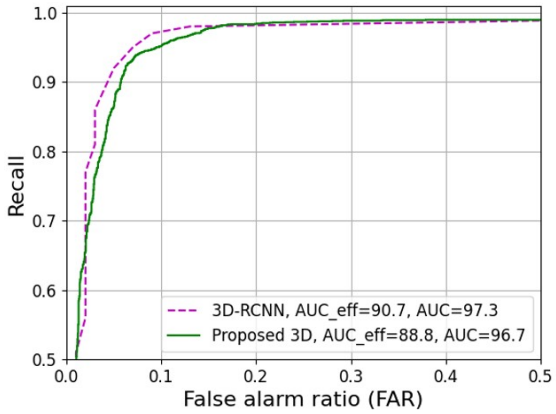


Fig. 10: Performance of proposed 3D approach (run-time = 20fps) compared with conventional 3D RPN-based object detection (run-time = 7fps) [20].

Table 3: Feature map and parameter count (param.) for candidate architectures for CWD in walk-through screening, proposed 3D architecture is shown in bold

Architecture	f_1	f_2	f_3	f_4	f_5	param.
Proposed 3D	**32**	**64**	**128**	**256**	**512**	**18M**
3D-UNet-default	64	128	256	512	1024	72M
3D-UNet-3-by-4	48	96	192	384	768	42M
3D-UNet-1-by-4	16	32	64	128	256	4M

Table 4: Performance of proposed 3D vs 2D projection based approaches, using the AUC_{eff} metric. Here overall performance is evaluated on the whole test set.

Approach	overall	easy case	difficult case
Proposed 3D	**88.8**	**88.2**	**74.9**
2D multi-view	87.7	87.0	67.4
2D z_int	87.6	87.1	68.3
2D x_int	83.8	82.7	65.1

of 3D RPN and 2D approaches will be introduced in subsequent subsections. All the frameworks are implemented in PyTorch [13] and a single Quadro RTX 5000 GPU is used for the evaluation.

As it can be seen in the Fig. 9, most of the candidate architectures are to the right of the red dashed line. This implies they exceed the walk-through run-time requirement of $50ms$. Accuracy-wise, interestingly, the 3D-UNet-3-by-4 achieves high value and a significant improvement over 3D-UNet-default. We suspect that the reduced parameter count (30M less than the default), led to a better generalization capability for 3D-UNet-3-by-4. However, its run-time is significantly higher than our requirement. Amongst the architectures to the left of the red dashed line, the proposed 3D has the highest accuracy. It is because the proposed 3D architecture has enough representational power, without being overly complicated, to match the required complexity of the detection task. Thus, the proposed 3D best fulfils the accuracy and run-time requirements for CWD in walk-through weapon screening systems.

5.2. Comparison with existing work

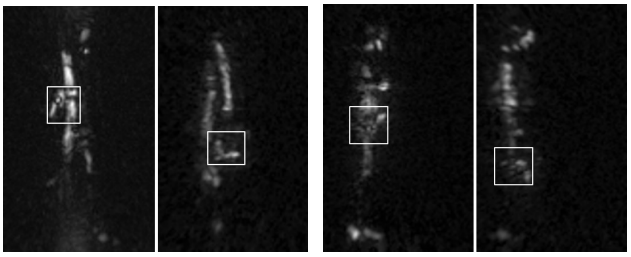
We compare the run-time and accuracy of our proposed 3D approach with existing RPN-based method [20] to validate that the existing work does not satisfy our system requirements. We tune the hyper-parameters of the 3D-RCNN network, mainly the RPN, on our validation dataset. Namely, we adjust the number of anchors, their sizes and the bounding box overlap thresholds which control the count of region proposals. The run-time of the 3D-RCNN network is reduced compared to [20] due to these optimizations, but it is still much higher than our system's requirement. Post-optimization, the run-time of 3D-RCNN is $145ms$ (7fps), where 3D RPN alone has a run-time of 100ms. This confirms our understanding that RPN contributes to increased run-time thus, making it unsuitable for our requirement of real-time processing (20fps). Next we perform accuracy comparison. For fairness, we use the same method of accuracy evaluation as our proposed approach instead of conventional Intersection-over-Union (IoU) i.e. if center of predicted 3D bounding box is inside GT box then we say its a TP else not. As shown in Fig. 10, the RPN-based approach has comparative performance as our approach.

5.3. Comparison with 2D Approach

We experimentally validate the performance superiority of using 3D approach over 2D for our application, by comparing proposed 3D approach with 2D-projection based approaches, as in [8, 18]. A 2D projected image I_p is obtained from a 3D absolute valued image I by using the max-projection strategy along projection axis p as,

$$I_p = max(I, axis = p) \tag{4}$$

For the 2D approach, two orthogonal projection axes, the z-axis and x-axis, are considered to obtain the 2D images



(a) Easy case (b) Difficult case

Fig. 11: Easy vs difficult images: Gun's position is shown by white bounding box. The left and right images show gun concealed on the waist and in the bag respectively.

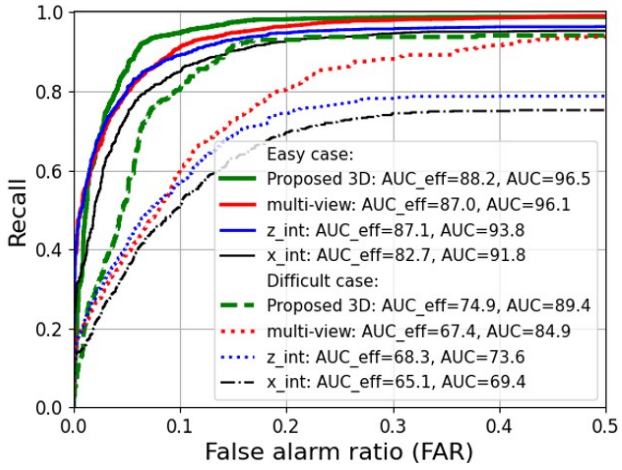
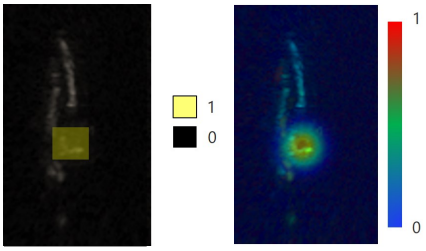


Fig. 12: ROC curves of proposed 3D approach vs 2D projection-based approaches. Proposed 3D has higher performance in both cases, but for the difficult case it surpasses 2D approaches by a large margin.

I_z and I_x, respectively. Two separate 2D networks, z_int and x_int, are trained with I_z and I_x as detection input, respectively. The 2D detection method is implemented with a 2D U-Net [15] and 2D Gaussian function in a probability map. We also compare the proposed 3D approach with a multi-view approach as multi-view 2D network is frequently employed in other domains [14] to learn 3D shape. The multi-view is an ensemble approach which combines the prediction outputs of z_int and x_int using the OR rule. If a weapon is detected by either the z_int or the x_int approach, it is counted as detected in the multi-view. The results are presented in Table 4 (overall).

For a more detailed analysis, we further split the with-weapon test set of 2400 count into easy (2159 counts) and difficult case (249 counts). As explained in Fig. 3, the 3D radar images which retain the shape of the gun after 2D projection are tagged as easy case, whereas the images which suffer from shape loss are tagged as difficult case. Since such a shape loss is challenging to quantify, we instead use a qualitative visual check to identify such images. Even



(a) Binary (b) Gaussian

Fig. 13: Types of mask (2D visualization): (a) Voxel-wise binary mask vs (b) Peak-shaped Gaussian probability mask.

though the 2D projected image is used for the qualitative check, understandably the images are difficult in 3D as well due to lower weapon intensity values (middle row of Fig. 3b). Some sample images for both the cases are shown in Fig. 11. We present the ROC curves for the same in Fig. 12, and the accuracy values in Table 4 for completeness.

Amongst the 2D-projection based approaches, the multi-view approach has better accuracy than the "single-view" (z_int and x_int) approaches. This is understandable because the multi-view is an ensemble approach utilizing the shape information from two orthogonal projection axes, z and x, for detection. But the proposed 3D approach performed most effectively, compared to the 2D approaches, in particular for the difficult cases. As expected, all approaches have lower performance for difficult cases when compared with easy cases. But the 3D approach proves to be most robust, as the performance gap between easy and difficult cases is the least. Hence, it is confirmed that our proposed 3D method is superior compared to 2D.

5.4. Ablation Study

Peak shaped probability mask vs voxel-wise mask. We compare our peak shaped probability mask with the conventionally used voxel-wise mask to show that the former is more meaningful. Since our original GT annotations are available as bounding boxes, we suggest generating GT voxel-wise masks as cuboid-shaped binary masks, as shown in Fig. 13. Such a binary mask resembles a solid 3D bounding box, the inside of the mask containing the weapon has value of 1 whereas its outsides has value of 0. So it can be said that binary masks are hard labels whereas our peak-shaped probability masks are soft labels. During prediction with voxel-wise binary masks, we calculate the center-of-mass of the thresholded output map to infer the weapon's position. We compare the accuracy of the peak-shaped probability map approach with the voxel-wise binary mask approach, as shown in Fig. 14, and observe that our proposed approach outperforms significantly.

Thus, we can conclude that our proposed approach of using peak-shaped probability maps is more meaningful than the conventional approach of using binary mask. We would

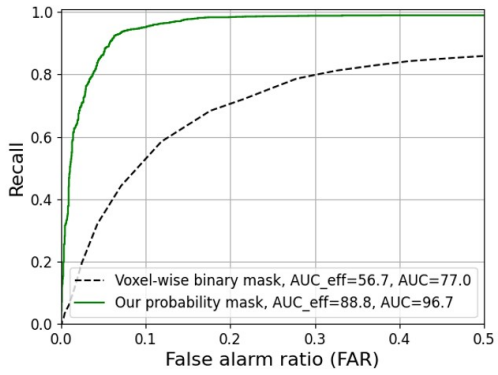
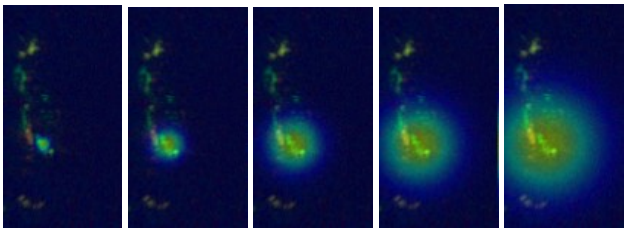


Fig. 14: Performance comparison between peak-shaped probability mask vs voxel-wise binary mask.



(a) 0.125 (b) 0.25 (c) 0.5 (d) 0.75 (e) 1.0

Fig. 15: Impact of k's value on Gaussian function's size (Eq. 2), shown for $k \in \{0.125, 0.25, 0.5, 0.75, 1.0\}$. The Gaussian map is superimposed on the radar image.

like to add here that the suggested approach to find the center of a weapon in binary mask does not extend to the case of multiple weapons per image, since there is no straightforward way to find multiple centers.

Impact of probability map size on accuracy. The impact of the Gaussian function size σ_i on the accuracy is studied by varying the value of the proportionality factor k in Eq. 2. We actually wish to find the the optimal ratio k_{opt} between σ_i and the ground truth bounding box's size l_i, thus we vary $k \in [0.125, 0.25, 0.5, 0.75, 1.0]$. If k is too small as shown in Fig. 15a, the Gaussian function does not cover the entire weapon shape, leading to loss of shape information. Meanwhile, if k is too large, as shown in Fig.15e, the Gaussian function treats the surrounding context, e.g., the person's body, as weapon. The ROC curves for various k are shown in 16, where the maximum AUC score is obtained for $k = 0.25$, so we set $k_{opt} = 0.25$. We would like to add here that k_{opt} is independent of a gun's size as l_i already captures such size variation.

6. Discussion

From the experimental evaluation, we can see that our proposed 3D approach achieves good accuracy, while fulfilling the run-time requirements of the walk-through CWD system (Fig. 9). The 3D-UNet-3-by-4 as well as 3D-RCNN have higher accuracy but they significantly exceed the walk-through run-time requirements, and hence are deemed unsuitable. On the other hand, the 2D approaches are faster and seem to have good accuracy in overall metric but detailed analysis reveal their shortcomings (Table 4). If one examines the curves for difficult cases in Fig. 12, then the large gap in performance between 3D and 2D is revealed. This confirms our understanding that 2D approaches are not a suitable choice for our system as: (1) the detection performance suffers immensely due to shape loss during 2D projection, and (2) multiple 2D projections (multi-view) are unable to represent the 3D shape and perform worse than 3D approach as projection itself is a lossy operation. The ablation studies justify our choice of probability map as the segmentation mask, instead of the conventional binary mask. And we inform the optimal ratio between the probability map's size and the ground truth bounding box's size i.e. k to aid in the probability map's design.

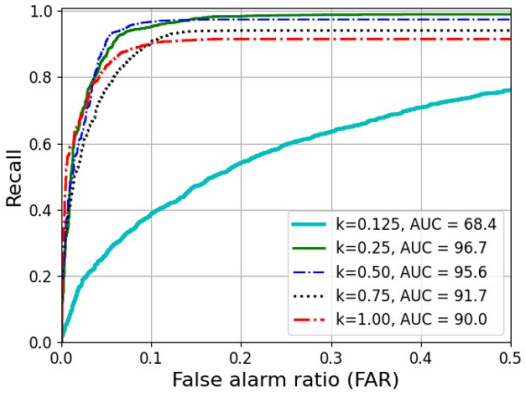


Fig. 16: Impact of probability map size on accuracy, σ_i of Eq. 2 is varied by varying the proportionality factor k.

To sum up, we verify the effectiveness of our proposed approach through extensive experimental studies and confirm that it is indeed the most suitable choice to perform accurate and real-time walk-through screening.

7. Conclusion

We developed a framework for real-time and accurate CWD on 3D radar images for walk-through weapon screening systems by reformulating a 2D segmentation network. To facilitate positional inference, the segmentation network is trained to predict peak-shaped probability maps where a peak marks a weapon's position. The predicted weapon probability maps are given to an elementary peak detector to obtain the weapon detection output. The effectiveness of the proposed approach is validated by extensive experimental studies on a realistic dataset of walk-through 3D radar images. The proposed framework performed accurately and in real-time, thus making it suitable for use in walk-through weapon screening.

References

[1] Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.

[2] Sherif Sayed Ahmed, Andreas Schiessl, Frank Gumbmann, Marc Tiebout, Sebastian Methfessel, and Lorenz-Peter Schmidt. Advanced microwave imaging. *IEEE Microwave Magazine*, 13(6):26–43, 2012.

[3] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022.

[4] Leonardo Carrer and Alexander G. Yarovoy. Concealed weapon detection using uwb 3-d radar imaging and automatic target recognition. In *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, pages 2786–2790, 2014.

[5] Eric Cosatto, Kyle Gerard, Hans-Peter Graf, Maki Ogura, Tomoharu Kiyuna, Kanako Hatanaka, Yoshihiro Matsuno, and Yutaka Hatanaka. A multi-scale conditional deep model for tumor cell ratio counting. In *Digital and Computational Pathology*, page 5, 02 2021.

[6] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43, 2021.

[7] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[8] J.Zhang, W.Xing, M.Xing, and G.Sun. Terahertz image detection with the improved faster region-based convolutional neural network. *Sensors*, 18:2387, 2018.

[9] Truc Le and Ye Duan. Pointgrid: A deep network for 3d shape understanding. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9204–9214, 2018.

[10] Jun Yi Lim, Md Istiaque Al Jobayer, Vishnu Monn Baskaran, Joanne Mun Yee Lim, John See, and Kok Sheik Wong. Deep multi-level feature pyramids: application for non-canonical firearm detection in video. *Engineering Applications of Artificial Intelligence*, 97, 2021.

[11] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015.

[12] Donna McClish. Analyzing a portion of the roc curve. *Medical decision making : an international journal of the Society for Medical Decision Making*, 9:190–5, 08 1989.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[14] Shaohua Qi, Xin Ning, Guowei Yang, Liping Zhang, Peng Long, Weiwei Cai, and Weijun Li. Review of multi-view 3d object recognition methods based on deep learning. *Displays*, 69, 2021.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*, pages 234–241. Springer International Publishing, 2015.

[16] D. M. Sheen, D. L. McMakin, and T. E. Hall. Three-dimensional millimeter-wave imaging for concealed weapon detection. *IEEE Transactions on Microwave Theory and Techniques*, 49:1581–1592, 2001.

[17] S.S.Ahmed, A.Schiessl, F.Gumbmann, M.Tiebout, S.Methfessel, and L.Schmidt. Advanced microwave imaging. *IEEE Microwave Magazine*, 13:26–43, September 2012 September 2012.

[18] T.Liu and Y.Zhao. Concealed object detection for activate millimeter wave image. *IEEE Transactions on Industrial Electronics*, 66:9909–9917, 2019.

[19] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *2015 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[20] Xuanang Xu, Fugen Zhou, Bo Liu, Dongshan Fu, and Xiangzhi Bai. Efficient multiple organ localization in ct image using 3d region proposal network. *IEEE Transactions on Medical Imaging*, 38, 2019.

[21] Y. Zhao and O. Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.