*electronics*

MDPI

*Article*

# Multi-Dimensional Information Fusion You Only Look Once Network for Suspicious Object Detection in Millimeter Wave Images

**Zhenhong Chen *, Ruijiao Tian, Di Xiong, Chenchen Yuan, Tang Li and Yiran Shi**

Beijing Institute of Radio Metrology and Measurement, Beijing 100854, China; tiansrx123@163.com (R.T.); maodiy@126.com (D.X.); yuanchen@buaa.edu.cn (C.Y.); litang0216@163.com (T.L.); tnt957359095@163.com (Y.S.)
* Correspondence: chenzhenhong_czh@163.com

**Abstract:** Millimeter wave (MMW) imaging systems have been widely used for security screening in public places due to their advantages of being able to detect a variety of suspicious objects, non-contact operation, and harmlessness to the human body. In this study, we propose an innovative, multi-dimensional information fusion YOLO network that can aggregate and capture multimodal information to cope with the challenges of low resolution and susceptibility to noise in MMW images. In particular, an MMW data information aggregation module is developed to adaptively synthesize a novel type of MMW image, which simultaneously contains pixel, depth, phase, and diverse signal-to-noise information to overcome the limitations of current MMW images containing consistent pixel information in all three channels. Furthermore, this module is capable of differentiable data enhancements to take into account adverse noise conditions in real application scenarios. In order to fully acquire the augmented contextual information mentioned above, we propose an asymptotic path aggregation network and combine it with YOLOv8. The proposed method is able to adaptively and bidirectionally fuse deep and shallow features while avoiding semantic gaps. In addition, a multi-view, multi-parameter mapping technique is designed to enhance the detection ability. The experiments on the measured MMW datasets validate the improvement in object detection using the proposed model.

**Keywords:** MMW images; object detection; YOLOv8; multimodal enhancement; information fusion

check for updates

**Citation:** Chen, Z.; Tian, R.; Xiong, D.; Yuan, C.; Li, T.; Shi, Y. Multi-Dimensional Information Fusion You Only Look Once Network for Suspicious Object Detection in Millimeter Wave Images. *Electronics* **2024**, *13*, 773. https://doi.org/10.3390/electronics13040773

Academic Editor: Adão Silva

Received: 17 January 2024
Revised: 12 February 2024
Accepted: 13 February 2024
Published: 16 February 2024

## 1. Introduction

In recent years, due to the increasing emphasis on public safety, millimeter wave (MMW) imaging systems [1–3] have gradually become a necessary screening technique. Traditional security screening techniques, which mainly include X-ray machines and metal detectors, have different limitations. X-ray machines are not suitable for human screening due to the use of ionizing radiation, and metal detectors cannot detect non-metallic objects. In contrast, MMW scanners can not only penetrate clothing to detect concealed suspicious items, but they are also harmless to humans due to their utilization of nonionizing electromagnetic waves. Furthermore, MMW scanners are favored for their noncontact and privacy-protecting advantages. However, as illustrated in Figure 1, achieving the high-precision detection of suspicious objects in MMW images faces the following challenges [4–6]: (1) Compared with optical images, MMW images suffer from low-resolution problems. (2) Current MMW images are actually three-channel grayscale images, so they cannot provide enough information to find statistically significant differences between suspicious items and the human body. (3) The size gaps among suspicious items are large, and there are a variety of small targets in MMW images. (4) There is inherent system noise and multipath-reflection noise. Early MMW detection algorithms [7–10] are mainly based on statistical theory. Because they rely on hand-designed features and lack analyses utilizing big datasets, these algorithms have difficulty solving the above real-world application issues.

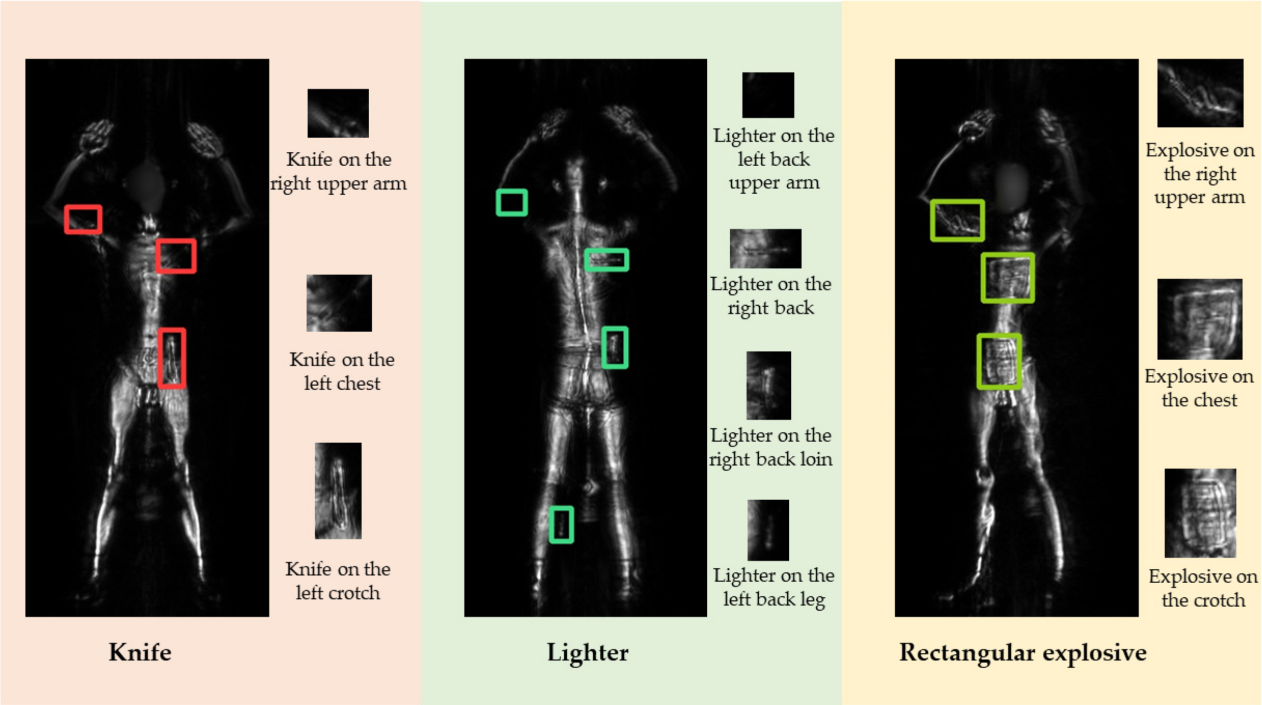

Knife on the right upper arm
Knife on the left chest
Knife on the left crotch
Knife
Lighter on the left back upper arm
Lighter on the right back
Lighter on the right back loin
Lighter on the left back leg
Lighter
Explosive on the right upper arm
Explosive on the chest
Explosive on the crotch
Rectangular explosive

**Figure 1.** MMW images of a human carrying knives (red boxes), lighters (green boxes), and rectangular explosives (yellow boxes), which have obvious size and shape differences.

Recently, deep learning has achieved great success in the field of optical image detection, which benefits from the powerful feature extraction capabilities of convolutional neural networks (CNNs) [11] and attention mechanisms [12–15]. Based on whether candidate regions are generated or not, these algorithms can be mainly categorized into two-stage methods and one-stage methods. The representative two-stage algorithm is the region-based convolution neural network (RCNN) series, including R-CNN [16], Fast-RCNN [17], Faster-RCNN [18], Cascade-RCNN [19], Dynamic-RCNN [20], etc. The RCNN series generates region proposals in the first stage and refines the localization and classification in the second stage. Meanwhile, many remarkable improvements, such as feature pyramid networks (FPNs) [21], region-based fully convolutional networks (R-FCNs) [22], and regions of interest (ROIs) [23] have been successively applied to enhance the semantic interaction capability and estimation accuracy. Because they combine coarse-grained and fine-grained operations, two-stage algorithms can obtain a high accuracy. Compared with two-stage algorithms, one-stage methods are free of candidate-region generation and directly perform predictions. As a result, one-stage methods have a higher efficiency and are more suitable for environments with high foot traffic. The single-shot, multibox detector (SSD) [24] and You Only Look Once (YOLO) [25] algorithms are two well-known early one-stage algorithms. Although they can detect in real time, they cannot match the accuracy of two-stage algorithms. Over time, more and more YOLO versions [26–31] were developed to reduce the accuracy gap between one-stage and two-stage methods. Among them, YOLOv5 [32] has become the current main solution in industrial inspection areas due to its excellent performance. Through jointly utilizing the cross-stage partial (CSP) theory [33], mosaic processing, and a path aggregation network (PANet) [34], the accuracy of YOLOv5 has been as good as those of two-stage methods. Building upon the success of YOLOv5, the latest state-of-the-art (SOTA) model, YOLOv8, was proposed in [35]. To promote convergence, YOLOv8 designs the C2f module to control the shortest longest gradient path [36]. Decoupled-head [37] and task-aligned assigners [38] have also been introduced to enhance its capabilities. However, directly applying the above methods to the detection of suspicious objects in MMW images does not fully exploit their advantages. This is because MMW images and optical images have many differences, as stated earlier.

Thanks to the development of deep learning, scholars have proposed some CNN-based detection algorithms for MMW images. In [39], a self-paced, feature attention fusion network was proposed to fuse MMW features with different scales in a top–down manner. A multi-source aggregation transformer was presented in [40] to model the self-correlation and cross-attention of multi-view MMW images. To address the issues in detecting small targets in MMW images, Huang et al. [5] combined YOLOv5 and a local perception swin transformer to increase the algorithm's global information acquisition ability. Liu et al. [41] adopted dilated convolution to construct a network with an enhanced multi-size target detection ability. Wang et al. [42] used normalized accumulation maps to locate targets. A Siamese network was utilized in [43] to change the MMW detection task into a similarity comparison task with a low complexity. However, the susceptibility to noise and insufficient channel information are still challenges that MMW detectors need to solve.

In this paper, we propose an improved YOLO detection algorithm that aims to increase the available information and extend its feature aggregation ability. We named this innovative method the multi-dimensional information fusion YOLO (MDIF-YOLO) network. Current MMW images are actually grayscale images containing three channels, from which the data information cannot be fully mined by existing networks. In view of this issue, we designed a data information aggregation (DIA) module, which can jointly use pixel, depth, phase, and different signal-to-noise ratio (SNR) information to generate a novel type of multi-channel MMW images. Moreover, the DIA module realizes differentiable image enhancement in the generation procedure. The corresponding enhancement parameters can be learned end-to-end during the training of the model. It is worth noting that one can arbitrarily select the information type (pixel, depth, phase, or SNR) for fusion. Therefore, the DIA module has a wide range of application scenarios and can significantly increase the available information and robustness to various types of noise. After the DIA module, the latest YOLOv8 is applied as the framework to construct a detection model for multi-channel MMW images. In order to better mine the multimodal features, we designed an asymptotic path aggregation network (APAN) and utilized it to replace the original PANet neck of YOLOv8. Unlike the PANet neck that uses simple concat fusion, the APAN adopts an adaptive spatial weighted fusion strategy, which can address the inconsistencies of different layers. Furthermore, the APAN extends the asymptotic incorporation theory [44,45] to realize bidirectional asymptotic aggregation in both top–down and bottom–up paths, which contributes to avoiding feature gaps during network transmission. At the output of the detection head, a multi-view, multi-parameter mapping technique is utilized to refine the detection results. This technique performs a cross-correlation mapping refinement among the three sets of multi-view images generated by using three sets of different DIA parameters for the same scan group, and applies a padding strategy to enhance the detection of unclear targets. As a result, an improved detection precision and a reduction in false positives can be achieved.

The proposed MDIF-YOLO network possesses the following advantages: (1) The DIA module increases the information richness of MMW images and enhances the network's robustness to noise. (2) To our knowledge, we are the first to propose an end-to-end, online, multi-dimensional data synthesis method in the MMW security field, and we improved the mainstream MMW grayscale image detection method. (3) The constructed APAN addresses feature inconsistencies and avoids semantic gaps during the fusion of deep and shallow layers. (4) A higher detection precision and fewer false positives can be obtained through the proposed multi-view, multi-parameter mapping technique. Extensive experiments verified the favorable performance of the MDIF-YOLO. The rest of this paper is organized as follows: Section 2 details the overall framework and design of the proposed MDIF-YOLO network. The experiments and analysis are provided in Section 3. Section 4 concludes this article.

## 2. Materials and Methods

### *2.1. Overview of the System*

In this study, we propose an innovative MDIF-YOLO network for suspicious object detection in MMW images. The MMW dataset [46,47] used in this research was obtained using our BM4203 series MMW security scanners [48,49] that adopt cylindrical scanning and active imaging, as shown in Figure 2. In one scan, ten MMW images from different angles can be obtained, where the first five images are generated by imaging the front of the person and the last five images are of the back. In order to protect the privacy of the person being scanned, BM4203 series MMW security scanners will automatically blur their face. The signal wideband of the system ranges from 24 GHz to 30 GHz, and the resolution is about 5 mm. As representative security products that have been widely applied in airports, stations, and stadiums, BM4203 series MMW security scanners have a high reliability and good image resolution. Through our scanners, the corresponding phase, depth, and diverse SNR information are simultaneously stored together with the MMW pixel images.

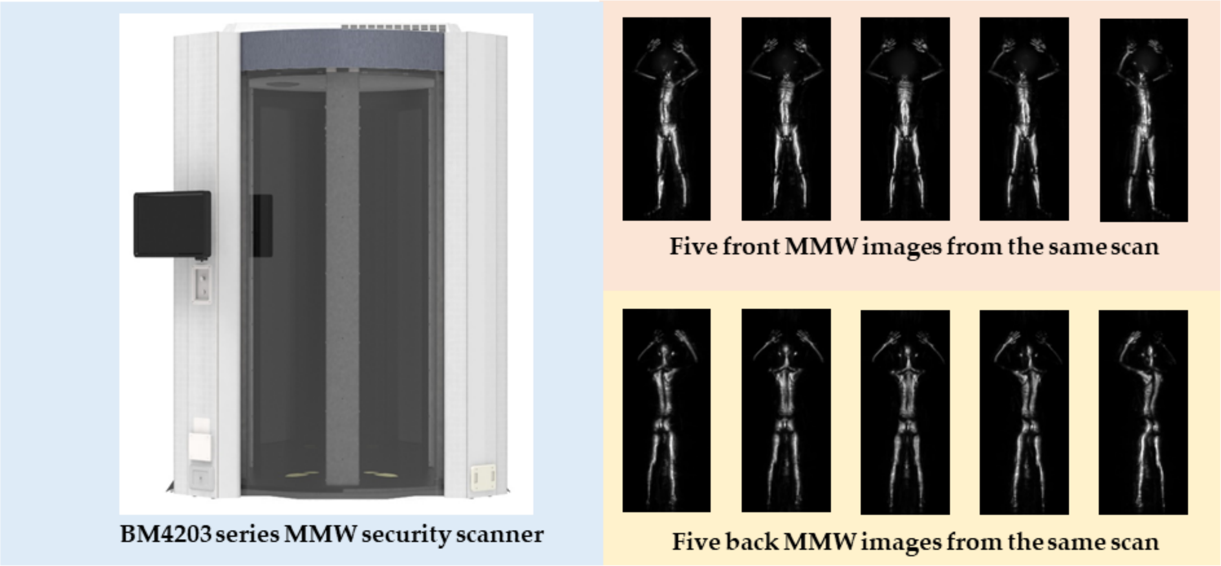

BM4203 series MMW security scanner
Five front MMW images from the same scan
Five back MMW images from the same scan

**Figure 2.** BM4203 series MMW security scanner and ten MMW images generated in one scan.

Figure 3 depicts the entire architecture of the proposed MDIF-YOLO network. The MDIF-YOLO consists of three main modules: the DIA module, YOLOv8–APAN module, and multi-view, multi-parameter mapping module. At the beginning, the DIA module receives original MMW pixel images together with one or more types of information (depth, phase, or SNR). In the DIA module, these multimodal data can be successively aggregated in both coarse-grained and fine-grained manners, where coarse-grained fusion focuses on estimating enhancement parameters that can be trained by backpropagation, and fine-grained fusion concentrates on further integrating the enhanced data. Then, the generated, novel, robust multi-channel MMW images are sent to the YOLOv8–APAN module. The roles of the YOLOv8 backbone, APAN neck, and detection head in the YOLO module are, respectively, multi-scale feature extraction, deep–shallow layer aggregation, and prediction. In particular, the proposed APAN in this paper combines bidirectional, asymptotic aggregation and an adaptive spatial weighted technique to avoid multi-scale gaps. Finally, the prediction results are sent to the mapping module. By jointly mapping the results of three sets of images with different DIA parameters from the same scan, the mapping module provides sufficient multi-view and multi-dimensional information to fine-tune the results. For unclear targets, a padding strategy is adopted to increase the detection performance.

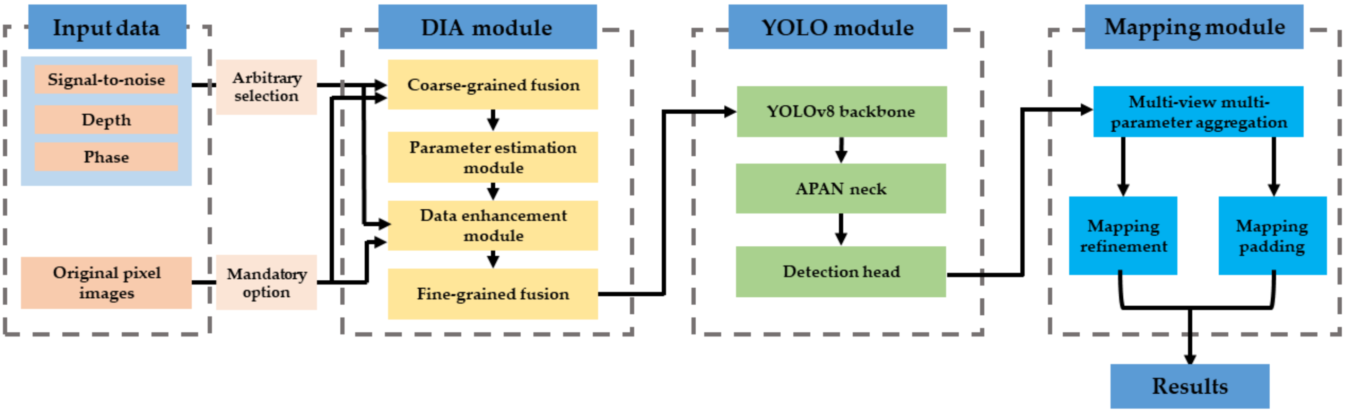


**Figure 3.** MDIF-YOLO network architecture.

### *2.2. Data Information Aggregation Module*

Although the current grayscale MMW images have three channels, the elements at the same spatial location in the three channels are identical pixels. The aim of the proposed DIA module is to synthesize a novel type of MMW image containing multi-dimensional information. Figure 4 shows the module's overall structure, which comprises coarse-grained fusion, parameter estimation, data enhancement, and fine-grained fusion submodules.

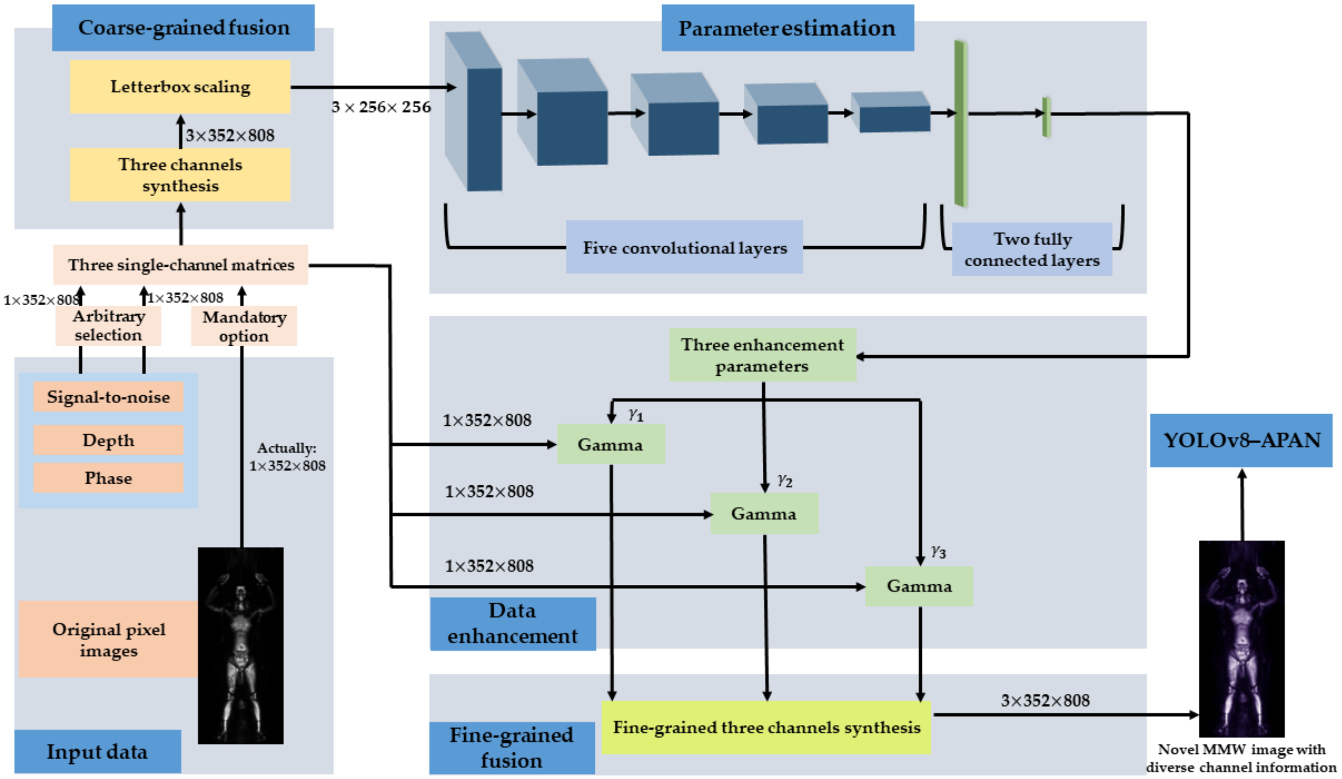


**Figure 4.** DIA module architecture.

As previously mentioned, ones can arbitrarily choose from available pixel, depth, phase, or SNR data for fusion. In this study, the pixel, depth, and SNR information were chosen to generate a new type of image. Their three single-channel matrices with dimensions of 1 × 352 × 808 were sent to the coarse-grained fusion submodule for the preliminary construction of a 3 × 352 × 808 matrix. To improve the computational efficiency, a letterbox scaling operation [35] was adopted to achieve downsampling so that the dimensions can be reduced to 3 × 256 × 256. Note that pixel, depth, and phase data can be easily obtained by using a wavenumber domain reconstruction algorithm [8] to process the MMW 3D echo data, which is expressed as

$$s(x_r, y_r, \omega) = \int\int\int f(x_t, y_t, z_t) \times e^{-j2k\sqrt{(x_t - x_r)^2 + (y_t - y_r)^2 + (z_t - z_r)^2}} dx_t dy_t dz_t, \tag{1}$$

where $(x_t, y_t, z_t)$ indicates the coordinates of the target, $(x_r, y_r, z_r)$ are the coordinates of the radar element, and $f(x_t, y_t, z_t)$ is the reflectivity. Different SNR data can be obtained by changing the logarithmic imaging threshold or by using simple image processing methods, e.g., the CLAHE method [50].

Based on the differentiable training theory presented in [51,52], after the coarse-grained fusion submodule, the parameter estimation submodule is used to extract three enhancement parameters that can be trained during backward propagation. Similar to [51], the parameter estimation submodule is a tiny CNN with five convolutional layers (output channel numbers are 16, 32, 32, 32, and 32) and two fully connected layers. Every convolutional layer contains 3 × 3 convolutions with stride 2, batch normalization, and a swish activation function. The fully connected layers output three trainable enhancement parameters.

The data enhancement submodule receives three chosen single-channel matrices from the input. Each matrix will be online gamma enhanced through setting the output parameters $\gamma_1$, $\gamma_2$, and $\gamma_3$ of the parameter estimation submodule as the enhancement parameters. Gamma enhancement can be expressed as $y = ls^{\gamma}$, where $l$ represents the luminance coefficient that is generally set to 1, $s$ is any matrix element, and $\gamma$ is the gamma enhancement parameter. Obviously, the power exponential operation is differentiable.

At the end of the DIA module, the fine-grained fusion submodule will synthesize the above three enhanced single-channel matrices. As a result, a novel type of real three-channel MMW image with multimodal information can be generated. Figure 5 shows a comparison between a traditional MMW pixel image and the corresponding novel multimodal MMW image. Even to the human eyes, the proposed new type of image has richer textures, particularly for poorly imaged areas of the human body, such as the arms. Additionally, from the computer vision perspective, the new image has three different channels of information. In contrast, traditional MMW images only have one channel of information. Therefore, from the perspective of both image quality and feature extraction, the new image has more advantages. In addition, the online trainable data enhancement increases the network's robustness to various types of noise. In particular, through learning how to use the new kind of image, even the detection ability for the original pixel image can be greatly improved. The reason is that the learning of multi-dimensional information also strengthens the ability to analyze pixels.

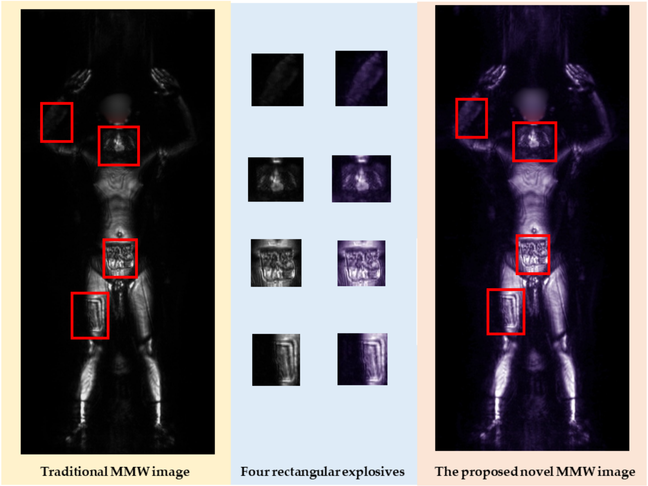


**Figure 5.** Comparison between traditional MMW image and the proposed novel MMW image. The person being scanned is carrying four rectangular explosives.

### 2.3. YOLOv8–Asymptotic Path Aggregation Network

In order to adequately extract and analyze the augmented information of the novel MMW image, an APAN was designed and embedded into YOLOv8 as its neck architecture. This novel network is named YOLOv8–APAN; its structure is depicted in Figure 6. YOLOv8–APAN comprises a YOLOv8 backbone, APAN neck, and detection head, where the backbone extracts features, the neck aggregates the extracted features, and the head produces the detection results. Like YOLOv8, in the backbone and neck, the proposed method applies C2f instead of C3 in YOLOv5. Compared with C3, C2f can control the shortest longest gradient path while keeping it lightweight, so that a higher capability boundary can be achieved. The detection strategy of the head uses the anchor-free method rather than the anchor-based method. In this way, the limitations of the anchor box design and the computational complexity of NMS will be alleviated. The structural components of C2f and the detection head can be found in [35].

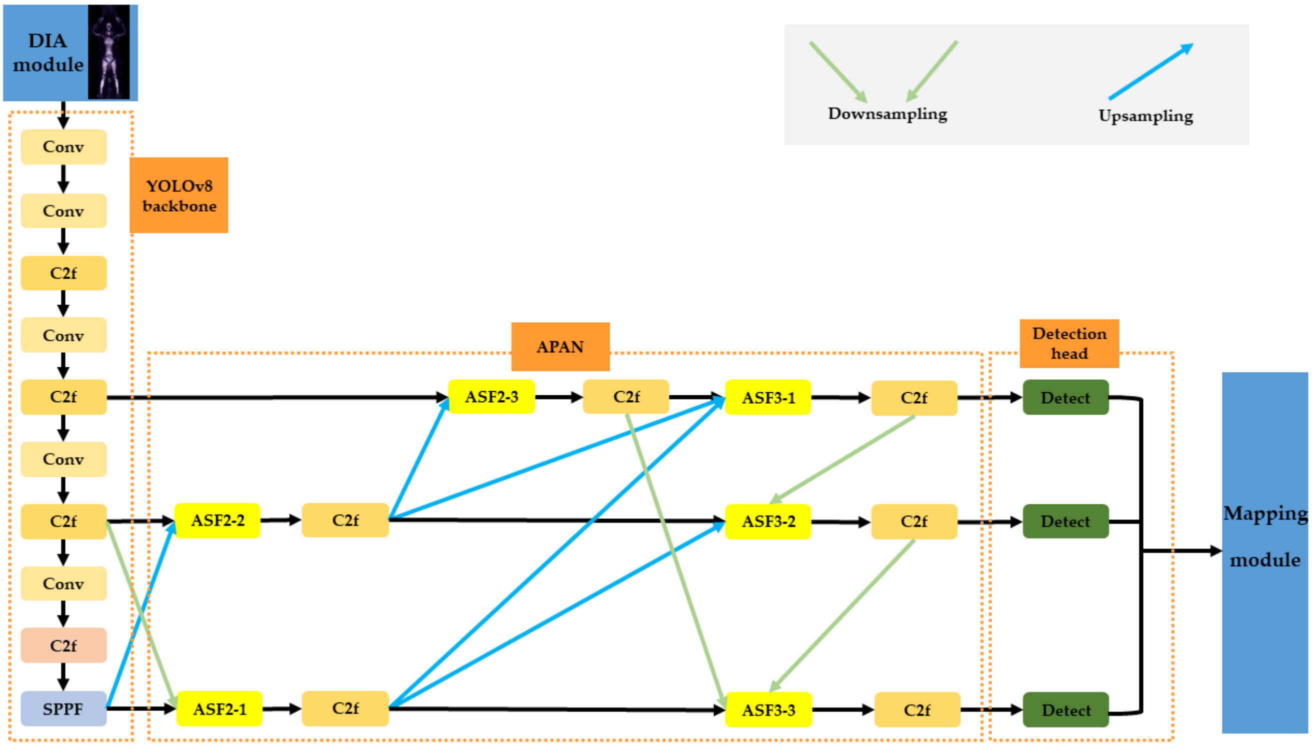


**Figure 6.** The framework of YOLOv8–APAN.

It is well known that current neck networks generally choose FPN and its optimized versions, e.g., PANet, because of their strong deep and shallow feature fusion capabilities. Recently, an asymptotic feature pyramid network (AFPN) was proposed in [44] as a new and improved version of the FPN. Through integrating adjacent shallow-layer features and asymptotically uniting deep-layer features, the AFPN is able to eliminate semantic gaps between non-adjacent levels. This paper extends the asymptotic incorporation theory to PANet and proposes APAN. As the improved version of AFPN, the proposed APAN realizes asymptotic integration not only in top–down paths but also in bottom–up paths, which is shown in Figure 7. Unlike AFPN, which first considers shallow-level features and then expands to deeper layers, our APAN involves bidirectional expansion union, i.e., ASF2-1, ASF2-2, and ASF2-3 first follow a deeper-to-lower direction, and then ASF3-1, ASF3-2, and ASF3-3 take a lower-to-deeper path. According to [44], different layers that are far apart have semantic gaps that cannot be ignored. Thus, it is not suitable to fuse them directly. In APAN, feature differences between two distant layers are greatly reduced by first merging adjacent layers in pairs, and then further considering the fusion of distant layers. For example, input1 and input2 are the first to be aggregated through ASF2-1 and

ASF2-2, and then the fusion of input3 and the output of ASF2-2 is carried out through ASF2-3. Although input1 and input3 are non-adjacent layers, their indirect information fusion in ASF2-3 can avoid the gap issue. This is because input1 and input2 as well as input2 and input3 are adjacent layers, and the information of input2 in ASF2-2 plays an important role in regulating the conflict between input1 and input3. The same theory applies to ASF3-1, ASF3-2, and ASF3-3, except that they aim at three layers. Following each ASF submodule, C2F is deployed for processing and learning the fused feature. Moreover, upsampling and downsampling operations are introduced to align the dimensions during fusion. Upsampling consists of convolution and an interpolation technique, and downsampling is achieved by convolution.

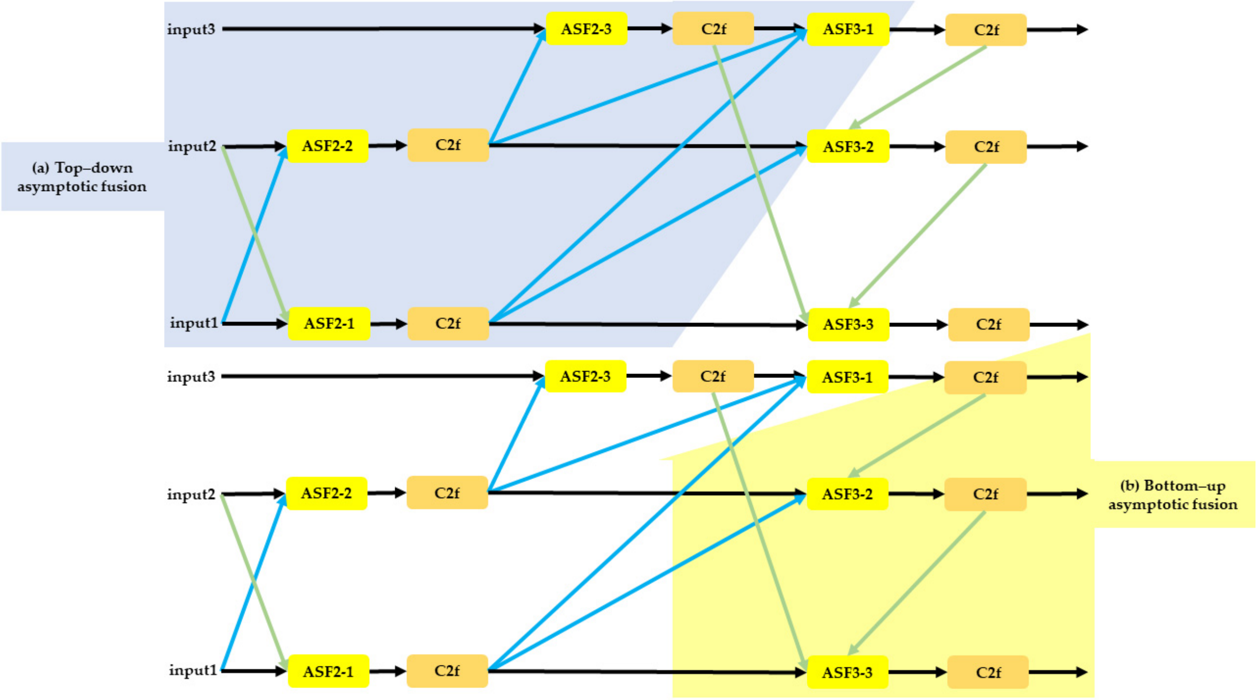

(a) Top–down asymptotic fusion
input3
input2
input1
ASF2-1
ASF2-2
ASF2-3
C2f
ASF3-1
ASF3-2
ASF3-3
(b) Bottom–up asymptotic fusion

**Figure 7.** The bidirectional asymptotic aggregation paths of the APAN.

As discussed above, ASF2 and ASF3 are the key components of the APAN. They are mainly used for adaptive spatial weighted fusion. ASF2 is used for the weighted union of two layers and ASF3 is applied for the weighted union of three different-level layers. Figure 8 illustrates their structure. The computing procedure of ASF2 can be defined as

$$ASF2(\mathbf{X}_1, \mathbf{X}_2) = C2f(SiLU(BN(Conv2d(\mathbf{A}_1 \odot \mathbf{X}_1 + \mathbf{A}_2 \odot \mathbf{X}_2)))), \tag{2}$$

$$\mathbf{A}_1, \mathbf{A}_2 = softmax(Conv2d(concat(SiLU(BN(Conv2d(\mathbf{X}_1))), SiLU(BN(Conv2d(\mathbf{X}_2)))))), \tag{3}$$

where *Conv2d*, *BN*, *concat*, and *softmax* represent the convolution, batch normalization, and softmax operations, respectively. SiLU is the swish activation function. $\odot$ represents element-by-element multiplication. In Equation (3), both the kernel size and stride of the convolution operations for $\mathbf{X}_1$ and $\mathbf{X}_2$ are 1, and the output channel number is 8. The kernel size and stride of the last convolution operation in Equation (3) are also 1, and the output channel number is 2. For Equation (2), the kernel size and stride of its convolution operation are 3 and 1, respectively, and the corresponding output channel number is equal to the channel number of the input feature without upsampling and downsampling operations. Similarly, ASF3 can be expressed as

$$ASF3(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3) = C2f(SiLU(BN(Conv2d(\mathbf{A}_1 \odot \mathbf{X}_1 + \mathbf{A}_2 \odot \mathbf{X}_2 + \mathbf{A}_3 \odot \mathbf{X}_3)))), \tag{4}$$

$$\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 = softmax(Conv2d(concat(SiLU(BN(Conv2d(\mathbf{X}_1))), SiLU(BN(Conv2d(\mathbf{X}_2))), SiLU(BN(Conv2d(\mathbf{X}_3)))))). \tag{5}$$

In Equation (5), the convolution operations for $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ have the same parameters as those in Equation (3), i.e., both the kernel size and stride are 1, and the output channel number is 8. The kernel size and stride of the last convolution operation are 1, and the output channel number becomes 3 due to there being three inputs. The parameters of the convolution operation in Equation (4) are the same as those of the convolution operation in Equation (2), i.e., the kernel size is 3, the stride is 1, and the output channel number is equal to the channel number of the input feature, which does not need to perform upsampling and downsampling operations. It is well known that for different-level layer fusion, even if upsampling and downsampling operations are performed, the target characteristics inevitably present distortions. Fortunately, ASF2 and ASF3 can generate adaptive coefficients based on the evaluation of the importance of the characteristic, thus strengthening the desired information and weakening the jamming information. Through this way, inconsistencies of different layers in the aggregation procedure can be relieved. From Figure 8, ASF3 is a more complex version of ASF2, because the input number has been changed from two layers to three layers. In future work, we can consider designing an ASF4 submodule with four inputs to further optimize the network model.

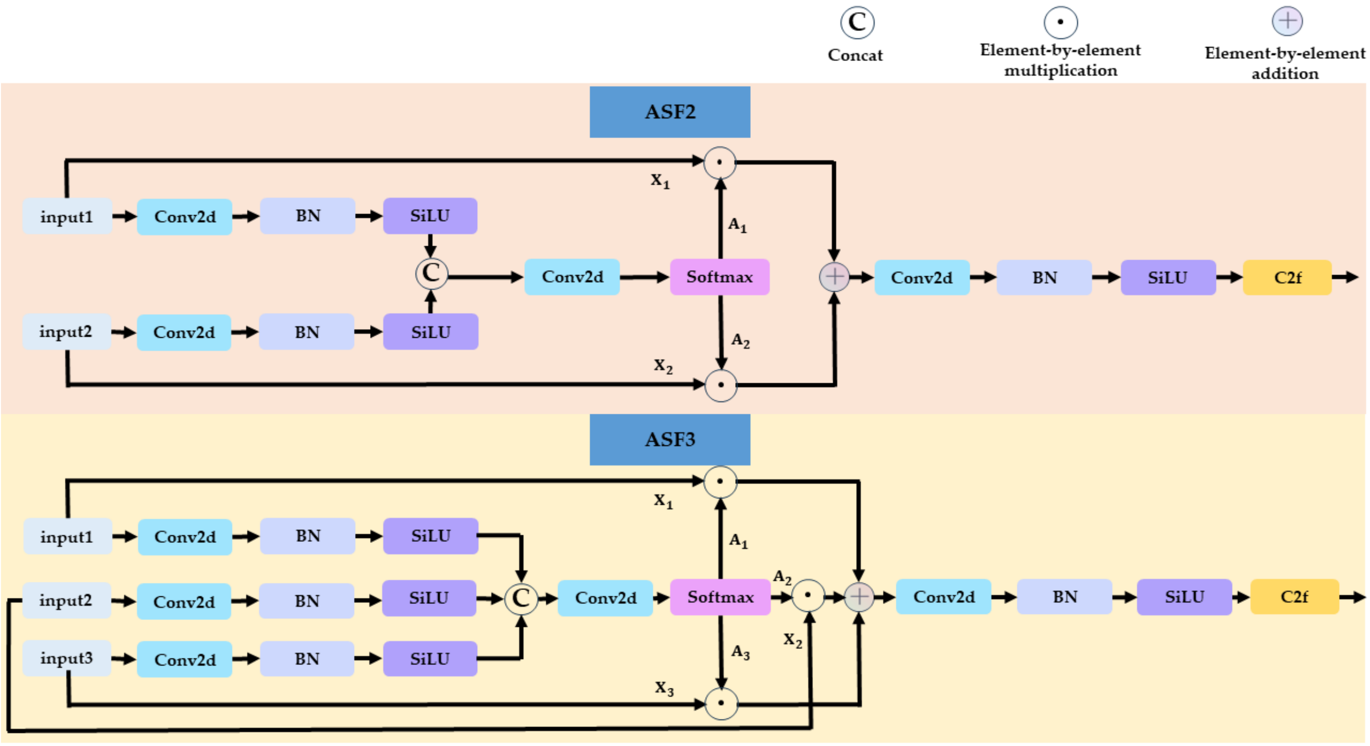


**Figure 8.** The structure component of ASF submodule in an APAN.

The APAN has three outputs with different receptive fields, which are sent separately to three detection heads. The three detection heads are decoupled, which means they can be divided into two branches, i.e., regression and classification. None of them need to design anchors in advance and can choose to use anchor-free detection.

### 2.4. *Multi-View, Multi-Parameter Mapping Module*

Although the above DIA module and YOLOv8–APAN can significantly improve the detection capability, fusing multi-view and different DIA pattern information into the whole network can help to further refine the detection results and improve the accuracy. In this section, a multi-view, multi-parameter mapping technique is introduced. This technique can simultaneously utilize the spatial–temporal and wide domain information of three sets of images with different DIA parameters. This way, a higher performance and better robustness against noise can be obtained.

As shown in Figure 9. this technique consists of two parts: (1) multi-view, multi-parameter aggregation and (2) mapping refinement and padding. The multi-view, multi-parameter aggregation submodule is constructed to fuse multi-angle and multiple DIA pattern information. The mapping refinement and padding submodule is designed for screening and modifying the results.

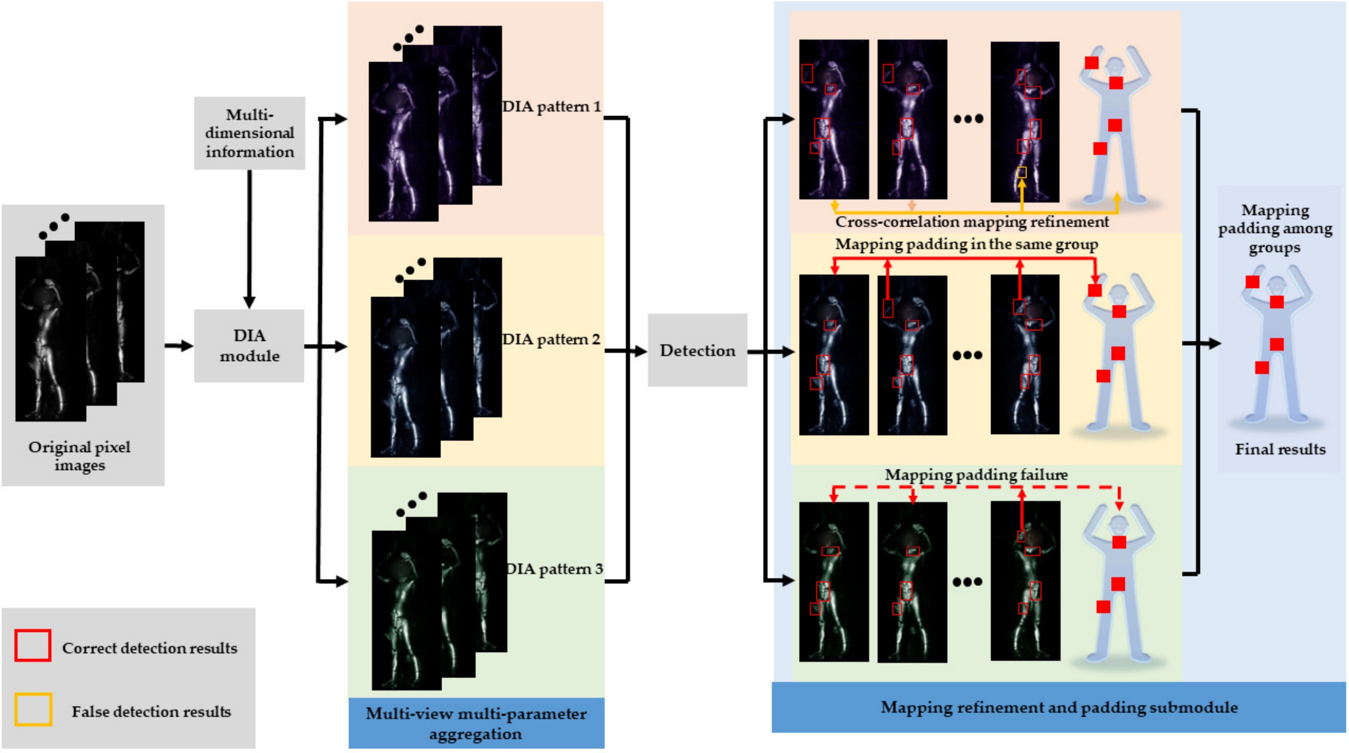

Multi-dimensional information
DIA module
Original pixel images
DIA pattern 1
DIA pattern 2
DIA pattern 3
Multi-view multi-parameter aggregation
Detection
Cross-correlation mapping refinement
Mapping padding in the same group
Mapping padding failure
Mapping padding among groups
Final results
Mapping refinement and padding submodule
Correct detection results
False detection results

**Figure 9.** The overall process of multi-view, multi-parameter mapping technology.

The multi-view, multi-parameter aggregation submodule is used at the beginning of the model detection process. In fact, this submodule can be seen as creating three parallel branches that feed three sets of DIA outputs into the YOLOv8–APAN. As stated before, the DIA module can realize online differentiable image enhancement, during which three DIA enhancement parameters for three channels are obtained. In the training process, some optimal DIA parameters can be concluded. In the multi-view, multi-parameter aggregation submodule, we choose three sets of optimal DIA parameters to construct novel, multi-dimensional MMW images in three different DIA patterns, i.e., DIA pattern 1, DIA pattern 2, and DIA pattern 3. It should be noted that the three patterns have the same input data from the same scan, but with different DIA parameters. This means that the three patterns have the same multi-dimensional, wide domain information type, but are in different SNR cases, because the gamma enhancement method could increase or suppress noise by varying its parameters. Their comparison is depicted in Figure 10. Obviously, different DIA parameters result in different representations because of varying channel weightings. In other words, the emphasis on different kinds of information is more diverse. Therefore, jointly using three DIA patterns could improve the applicability and stability of the method. DIA pattern 1, DIA pattern 2, and DIA pattern 3 each have 10 images (equal to the original pixel image number in the same scan). The 10 images provide multi-angle information of the detected human body. In addition, due to the powerful parallel computing capabilities of YOLO and GPU-CUDA technology, combining the three patterns would not affect the high efficiency of the detection product. To summarize, multi-angle 3D information, multi-type input information, and multi-SNR information can be aggregated through the multi-view, multi-parameter aggregation submodule to achieving a better performance.

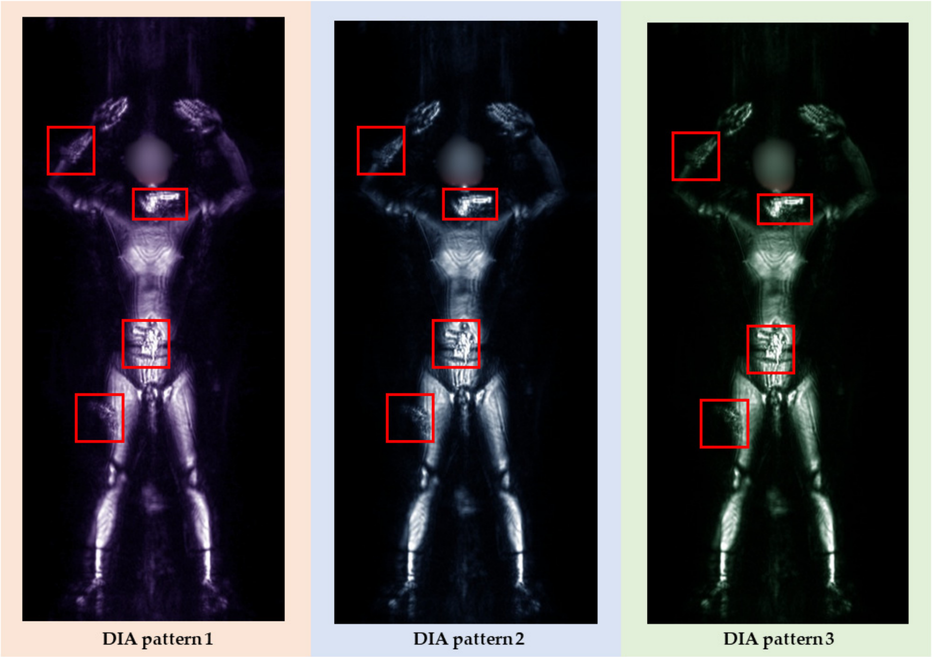


**Figure 10.** Comparison of the constructed multi-dimensional images with different DIA parameters. The person being scanned is carrying four guns, which have been marked by rectangular boxes in the figure.

The output data of the multi-view, multi-parameter aggregation submodule are delivered to YOLOV8–APAN for detection. Then, the detection results of the 30 images from the same scan are fed into the mapping refinement and padding submodule for the final adjustment. Figure 11 shows its detailed structure, which mainly comprises three types of components, i.e., a filtering and mapping component, refinement component, and padding component.

The filtering and mapping component is used for screening and mapping the detection results from each DIA pattern image group. Every detection result contains target coordinates and confidence. In order to reduce false positives and redundancy, the filtering and mapping component first filters the results by setting confidence thresholds according to the type of targets. Easily distinguishable target categories, such as guns, have high thresholds, while unclear categories, such as powder, have low thresholds. Assume that there are $M$ kinds of targets, and the given threshold of each kind is $T_m, m = 1, 2, \ldots, M$. Then, whether the $i$th ($I > 0$) result in the $j$th ($j = 1, 2, \ldots, 10$) image is filtered by the category threshold can be determined by the following mask function:

$$mask_{i,j,m} = \begin{cases} 1, & if \quad conf_{i,j,m} \geq T_m \\ 0, & if \quad conf_{i,j,m} < T_m \end{cases}. \tag{6}$$

If the confidence is no less than the category threshold, the mask is set as 1, and the result will be saved. Otherwise, it will be deleted. Furthermore, filtering can be further adjusted based on whether the target coordinates are located in a faint body part area. For example, arms are the dimmest body part due to their small scattering area. As for the targets located in arms, the mask function can be rewritten as

$$mask_{i,j,m} = \begin{cases} 1, & if \quad conf_{i,j,m} \geq \alpha_m T_m \\ 0, & if \quad conf_{i,j,m} < \alpha_m T_m \end{cases}, \tag{7}$$

where $0 < \alpha_m < 1$. The arm region can be obtained using a human posture recognition method or be roughly set using simple image zoning. This region-based threshold setting facilitates accurate screening. In this study, for convenience, we only chose basic threshold filtering in Equation (6). After filtering, the detection result coordinates of five front images and five back images in each DIA pattern group are mapped to the same front and back image in the group, respectively. The mapping can be easily achieved by 3D geometric transformation [46] because the depth, rotation angle, and 2D plane coordinates can be acquired by an arbitrary MMW security system. Moreover, there are other available mapping methods, such as the feature point matching method [53] or the tracking technique [47].

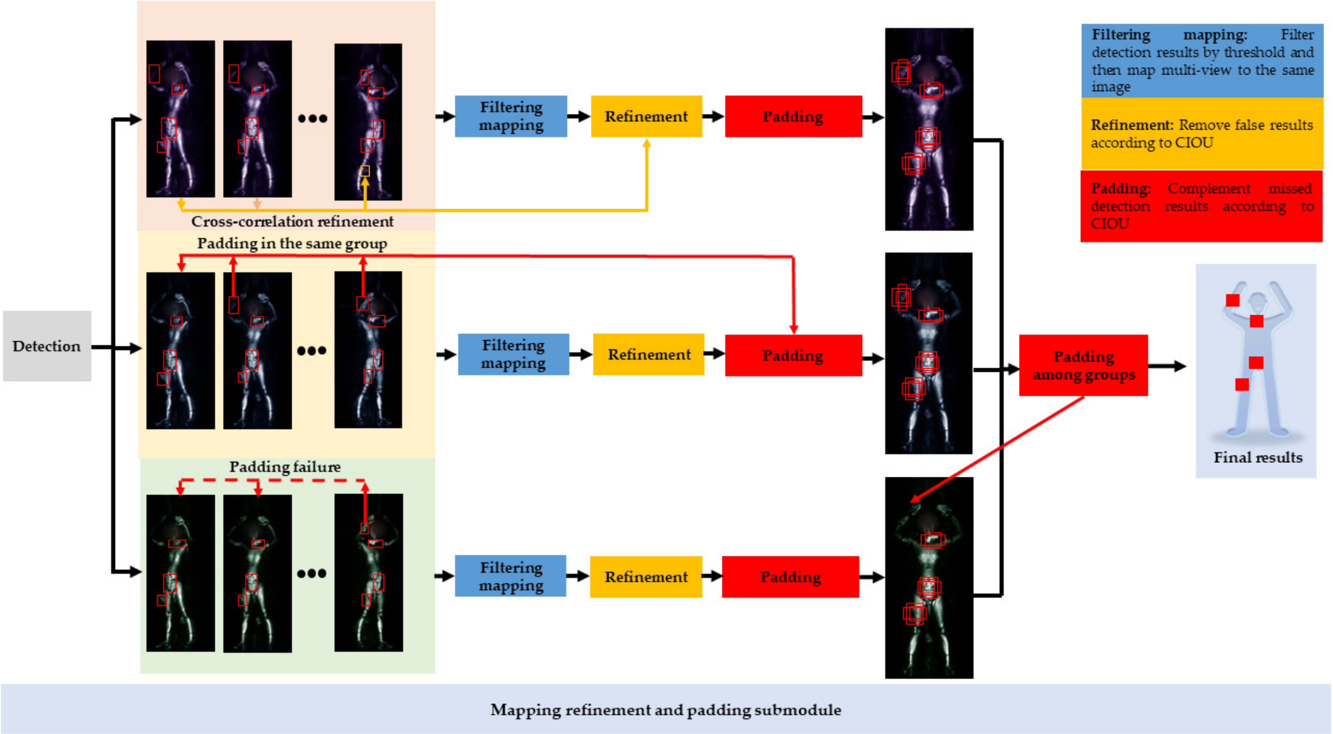

Detection
Cross-correlation refinement
Padding in the same group
Padding failure
Filtering mapping
Refinement
Padding
Padding among groups
Final results
Filtering mapping: Filter detection results by threshold and then map multi-view to the same image
Refinement: Remove false results according to CIOU
Padding: Complement missed detection results according to CIOU
Mapping refinement and padding submodule

**Figure 11.** The structure of the mapping, refinement, and padding submodule.

The refinement component applies CIOU to calculate the correlations among the filtering and mapping results. For each mapping detection result, the CIOU between it and the other results are saved in a list. If the list of a mapping detection result only contains zeros, the corresponding mapping detection results are deleted. The reason is that all CIOUs being zero means that there are no existing intersecting detection boxes, so it is highly likely that the detection result is a false positive. For instance, the first group of images (DIA pattern 1 images) in Figure 11 has one false positive on the leg in the final front image. Since there are no false positives in close positions in the other front images, this false positive can be deleted by the refinement component. Please note that since all the suspicious objects in the example are placed on the front of the human body, only the front images are shown.

The role of the padding component is complementary to that of the refining component. For one position, if more than one image has a detection result, it is determined that there is a target at all the corresponding positions in the same-side images, thus correcting some missing detections. In the first image of the second image group (DIA pattern 2 images) in Figure 11, the network failed to detect the target. Fortunately, this absence can be complemented by more than one correct detection result at close positions in other images. If only one image has a result, the padding process does not work, such as the example in the third image group (DIA pattern 3 images) of Figure 11.

Once the three DIA pattern groups obtain their refined and padded results, padding among groups is carried out. Different from in-group padding requiring no less than two correct results at close locations, padding among groups needs to complement an absence, even if only one group detects the target at a certain location. This helps to break down the limitations of some DIA pattern images. The final results of all three groups are padded and aggregated onto the same puppet image, as shown in Figure 11. Clearly, a higher detection precision and fewer false positives can be achieved through the proposed multi-view, multi-parameter mapping technique.

## 3. Experiment Results and Discussion

In this section, the performance of the proposed MDIF-YOLO algorithm is evaluated. Many experiments were conducted using the real-world MMW dataset obtained from our BM4203 series MMW security scanners. All experiments were implemented in the WIN 10 system using a 24GB NVIDIA TITAN RTX GPU. In the following, we will expatiate the dataset, experiment details, comparisons between the proposed method and other SOTA methods, and the corresponding discussions.

### *3.1. Dataset*

For reasons of commercial confidentiality, product copyright, and privacy protection, there are few large, available MMW datasets collected from practical application scenarios, let alone other non-pixel signal data such as depth and phase, even though non-pixel raw signal data are as readily available as pixels in every MMW screening system. Another reason for not providing non-pixel signal data is that current MMW detection algorithms do not take into account the use of multi-dimensional information. How to obtain non-pixel signal data has never been an obstacle for arbitrary MMW security products.

The MMW image dataset [46,47] and corresponding multi-dimensional data used in this study were collected using our BM4203 series MMW security scanners [48,49], which have been widely used in many airports, stations, and stadiums. BM4203 series MMW security scanners can perform almost 360-degree imaging of the person being scanned. Each scan can generate 10 images, where the first 5 images display the front of the body, and the last 5 images display the back. The resulting images preserve the person's privacy by blurring the face, and a good resolution of 5 mm can be achieved. Based on the application of BM4203 series MMW security scanners in multiple practical scenarios, the large dataset created contains 186610 real MMW security images and the corresponding depth, phase, and SNR information data. This study selected pixel, depth, and SNR information to use in creating novel multi-dimensional MMW images. Note that the novel image type would not affect the labels, so the labels do not need to be modified. The number of men and women in the dataset are balanced, and the age distribution is 18 to 70 years old. The testers wore a variety of clothes that are common in the four seasons, and carried guns, knives, rectangular explosives, lighters, powders, liquids, bullets, and phones. The dataset acquisition details can be summarized as follows:

(1) The collection of the dataset was accomplished through four BM4203 series MMW security scanners applied in four different practical scenarios, including an airport, station, stadium, and office building.

(2) There were about 200 testers, whose ages ranged from 18 to 70 years old. The number of men and women was nearly equal, and their body mass indexes covered the thin, normal, and obese ranges.

(3) Since spring and autumn clothing are almost identical, the clothing worn by the testers can be divided into three types: winter type, spring and autumn type, and summer type. The frequencies of these three types of clothing in the dataset are similar.

(4) There are eight kinds of common suspicious objects, which include guns, knives, rectangular explosives, lighters, powders, liquids, bullets, and phones. The testers were asked to hide suspicious objects on various body parts, including the upper and

lower arms, shoulders, neck, chest, abdomen, crotch, back, waist, buttocks, and legs. For each body part, the suspicious items were placed randomly.

(5) During the scan, the testers maintained a fixed posture with their hands raised upward. Each scan generated 10 images at 10 different angles. When one scan was complete, the system prompted the testers to leave. If the testers moved or had the wrong posture during the scan, the system prompted them to rescan, thus avoiding image blurring or image occlusion.

(6) The systems adopted a wave-number domain imaging algorithm [8]. The depth and phase information were stored together with the pixel information (i.e., traditional images) during its maximum value projection procedure. Different SNR information can be obtained by changing the logarithmic threshold of the imaging method or using simple image processing methods, e.g., the CLAHE method.

(7) The dataset was labeled using labeling software.

As shown in Figure 12, the division strategy of the dataset can be summarized as follows:

(1) After finishing the data collection and data labeling, a test set was first constructed by selecting one-tenth of the data from the full dataset. In particular, we ensured that the testers associated with the test set were not correlated with the remaining nine-tenths of the dataset.

(2) The remaining nine-tenths of the dataset were stored as eight subgroups according to the eight types of suspicious objects, i.e., guns, knives, rectangular explosives, lighters, powders, liquids, bullets, and phones.

(3) In each subgroup, the data produced by the same scan were treated as one unit, which was named as a scan unit in this paper. Then, each subgroup was shuffled by the scan unit, which means that the data produced by the same scan were not shuffled, while different scan units were shuffled.

(4) After the shuffle operation, each subgroup was divided into a training subset and validation subset. The partition ratio was 9:1.

(5) Finally, the training subset and validation subset from the eight subgroups were combined into the final training set and validation set.

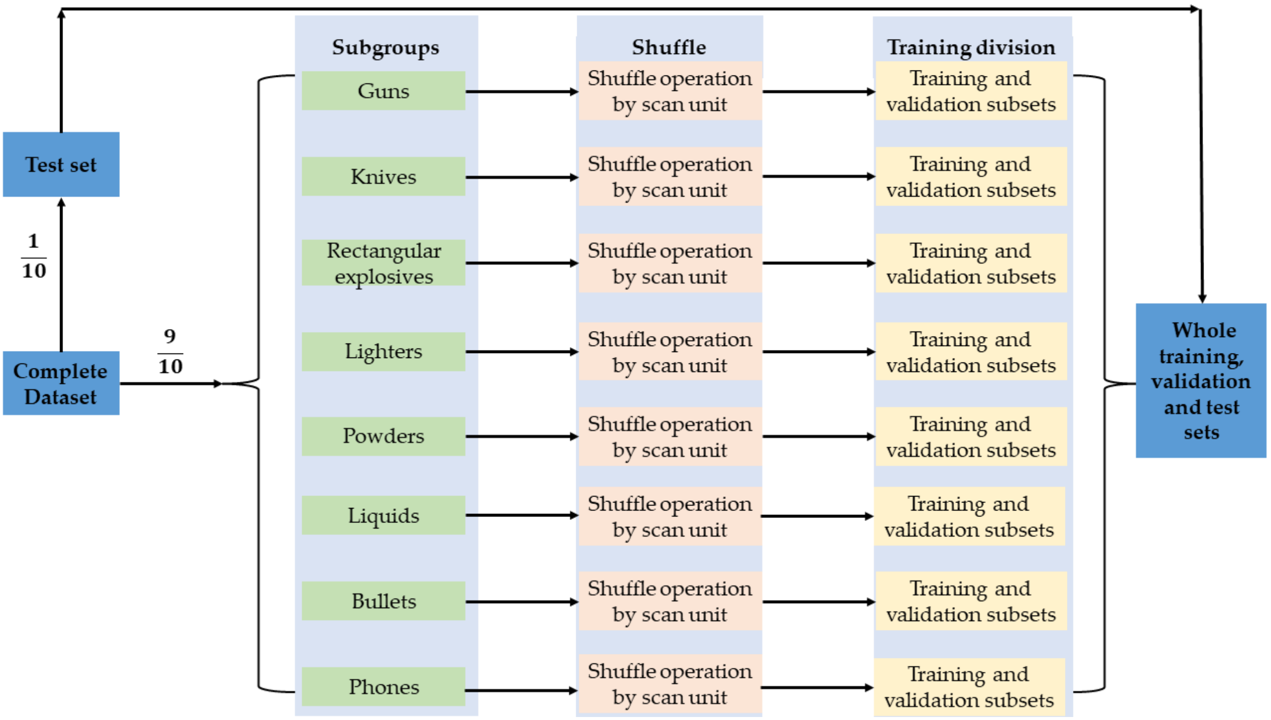


**Figure 12.** Dataset division process.

In conclusion, the partition ratio of the training, validation, and test sets was $\frac{81}{100} : \frac{9}{100} : \frac{10}{100}$.

### 3.2. Evaluation Metrics

Following the main detection and evaluation indicators used in the security inspection field, this study utilized Precision, Recall, and mean average precision (mAP) to evaluate the method's performance. In addition, a metric named the fake alert rate (FAR) was defined for a more comprehensive evaluation. These metrics can be expressed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

$$\text{FAR} = \frac{\text{FP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{10}$$

$$\text{mAP} = \frac{1}{C}\sum_{C}\int_0^1 \text{Precesion}(\text{Recall})d(\text{Recall}), \tag{11}$$

where TP, FP, TN, and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively, and $C$ is the number of target classes. Precision indicates the proportion of detected real targets among all detected results, and Recall measures what percent of the real targets were detected. Here, the definition of FAR is different from common false alarm rate concepts in the optical detection field. FAR is used to evaluate the rate of false positives in the entire dataset. mAP is a comprehensive evaluation and is the average of the integral of the Precision–Recall curve.

### 3.3. Implementation Details

The original dimensions of the pixel images were 3 × 352 × 808, where the three channels were actually the same. Thus, the images' real dimensions are seen as 1 × 352 × 808. During the parameter estimation of the DIA module, the data size was reshaped to 3 × 256 × 256 to reduce the number of calculations. Through DIA, a novel type of multi-dimensional MMW image, which is actually multi-channel and has dimensions of 3 × 352 × 808, is generated. When the data entered the YOLOv8–APAN, a mosaic enhancement and the letterbox process were applied. The novel image size became 3 × 640 × 640 at the input of the YOLO network until the end. During the training process, the proposed MDIF-YOLO algorithm end-to-end updates the parameters using an SGD optimizer with a momentum of 0.937. To ensure the stability of convergence, a warm-up policy was first applied for the first three epochs; then, the one-cycle strategy was used to gradually and nonlinearly decrease the learning rate. We set the initial learning rate and optimizer weight decay as 0.01 and 0.0005, respectively. The mosaic enhancement kept running during the training. The total number of training epochs was 100, and the batch size was set as 16.

### 3.4. Performance Comparison

In this subsection, the performance of the proposed MDIF-YOLO was compared to several other networks that are the most effective and widely used in the field of MMW security. In the practical application of MMW security products, most manufacturers still prefer to use the SOTA deep learning detectors in each period because of their good end-to-end training capabilities and their powerful deployment. As mentioned previously, SOTA deep learning detectors in different periods mainly include the RCNN series and YOLO versions. Here, we select Faster-RCNN, Cascade-RCNN, Dynamic-RCNN, YOLOv3, YOLOv5, and YOLOv8 as the comparison items. For security products, the detection ability and fake alert rate are the most important performance indicators. The results are shown in Table 1 and Figure 13. It can be seen that the performance of early YOLO methods (i.e., YOLOv3 and earlier versions) are not comparable to those of the RCNN algorithms. However, the recently proposed YOLOv5 and YOLOv8 algorithms showed comparable performances to those of the two-stage algorithms. By overcoming the limitations of traditional MMW images and aggregating multi-dimensional information, the MDIF-YOLO proposed in this paper outperformed the other six SOTA detectors. Thanks to the synthesis of much more abundant image information and its powerful feature aggregation ability, MDIF-YOLO obtained better Precision and mAP values. Furthermore, the combination of multi-dimensional data, including pixel, depth, multi-SNR, and multi-view information, can also improve the robustness and applicability. As a result, the proposed method had a significantly lower fake alert rate compared to the other methods.

**Table 1.** Detection abilities and fake alert rates of Faster-RCNN, Cascade-RCNN, Dynamic-RCNN, YOLOv3, YOLOv5, YOLOv8, and the proposed MDIF-YOLO.

| Algorithm | Precision | FAR | mAP |
|---|---|---|---|
| Faster-RCNN | 83.3 | 17.5 | 80.2 |
| Cascade-RCNN | 86.7 | 14.5 | 85.1 |
| Dynamic-RCNN | 87.2 | 14.2 | 86.8 |
| YOLOv3 | 82.6 | 17.7 | 79.8 |
| YOLOv5 | 87.1 | 12.8 | 86.6 |
| YOLOv8 | 89.3 | 11.5 | 88.7 |
| MDIF-YOLO | 92.2 | 6.4 | 91.8 |

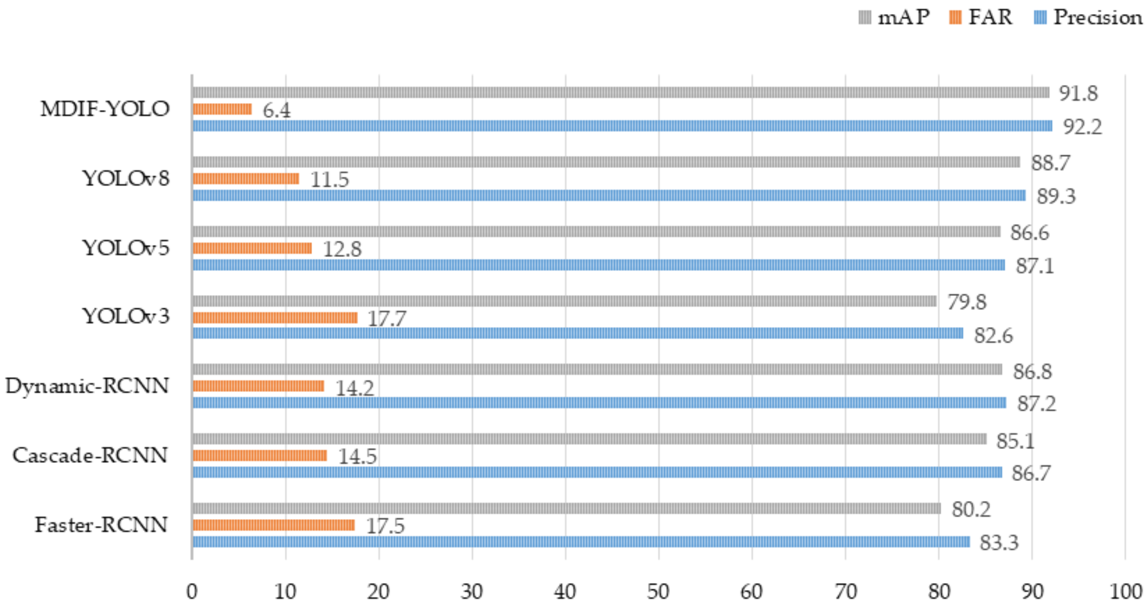


**Figure 13.** Performance comparison of Faster-RCNN, Cascade-RCNN, Dynamic-RCNN, YOLOv3, YOLOv5, YOLOv8, and the proposed MDIF-YOLO.

In order to give the reader a clearer understanding of the performance improvement of the proposed method, Figure 14 visualizes the comparison between MDIF-YOLO and YOLOv8. As discussed in the DIA module subsection, learning the DIA pattern images can greatly improve the detection ability of the original pixel images. To prove this more succinctly, in the next experiment, we omitted the multi-view, multi-parameter mapping process from MDIF-YOLO and directly predicted pixel images and the corresponding group of DIA pattern images. Note that SOTA algorithms, such as YOLOv8, do not have the structure to generate and learn multi-dimensional MMW images. Meanwhile, the existing SOTA methods are not robust enough to deal with the various types of SNR information contained in DIA pattern images, which has been tested in the actual operation of our MMW security products. Therefore, applying them to DIA pattern images will cause a significant degradation in the performance, and it is pointless to list their test results for DIA pattern images. This experiment chose bullets as the detection objects, which are the most difficult small targets to detect in MMW images. In Figure 14, the bullets are very small relative to the human body, and are indistinguishable from the body texture, especially when the bullets are hidden on the unclear body parts, such as the back of the head, legs, and arms. YOLOv8 failed in identifying the bullets at the back of the head and leg. In contrast, the proposed MDIF-YOLO network identified almost all the targets. Applying MDIF-YOLO to both the pixel images and the DIA pattern images had almost the same effectiveness, which verifies that aggregating the multi-dimensional information can improve the robustness and ability to analyze various types of MMW data. Another representation for robustness is the FAR, which has not been used as much as Precision and mAP in the relevant research. In effect, FAR can fully reflect the stability when there are signal oscillations and changes in the SNR. From Figures 13 and 14, we can observe that the proposed method had a significantly lower FAR for different types of images, verifying that the proposed method is more robust.

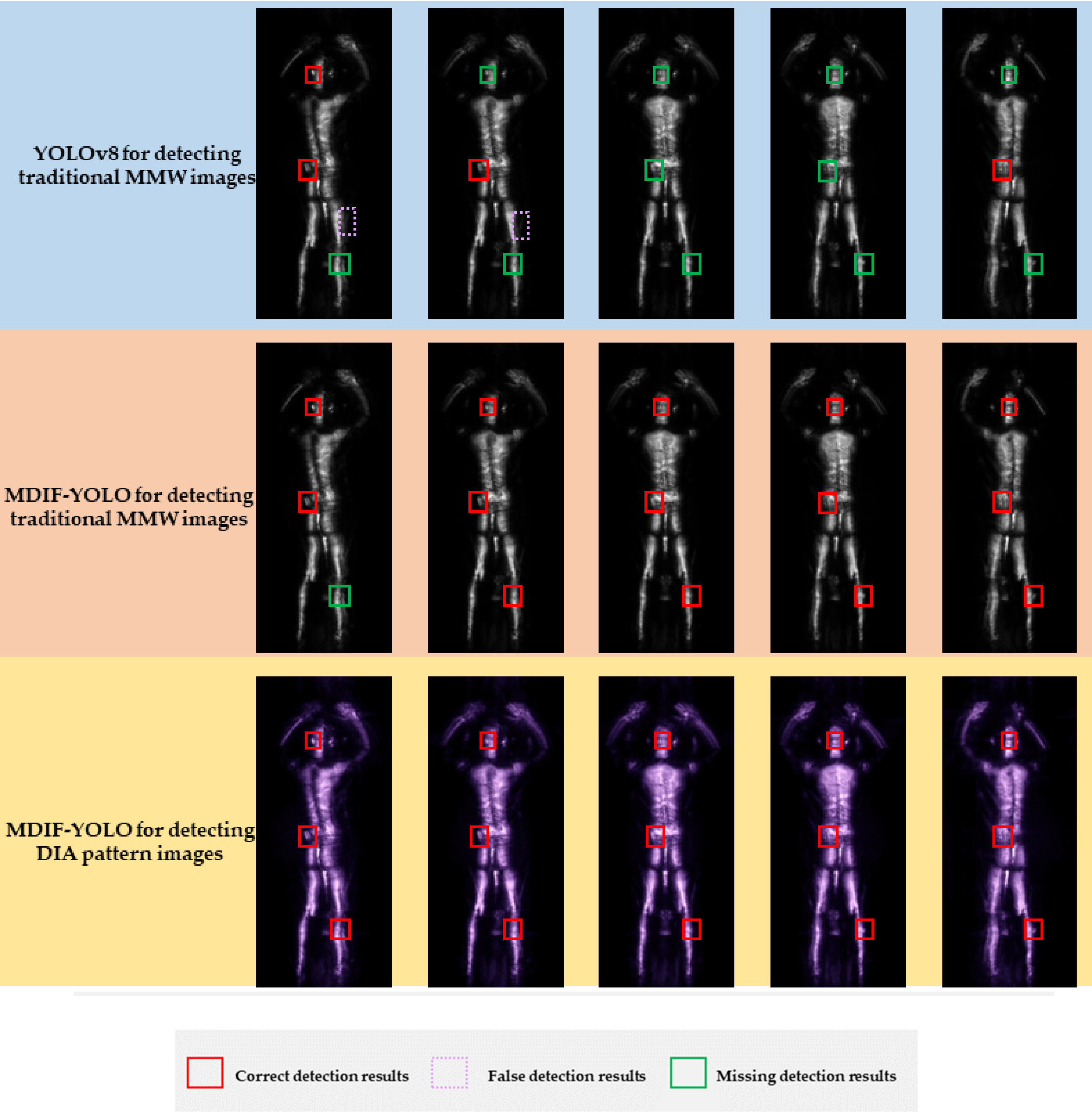

YOLOv8 for detecting traditional MMW images
MDIF-YOLO for detecting traditional MMW images
MDIF-YOLO for detecting DIA pattern images
Correct detection results
False detection results
Missing detection results

**Figure 14.** Visualized detection comparison between YOLOv8 and MDIF-YOLO. YOLOv8 uses pixel MMW images, while MDIF-YOLO uses both pixel MMW images and a new type of MMW images. The person being scanned is hiding bullets in the back of the head, lower back, and right leg.

Another comparison of MDIF-YOLO and the detectors specifically designed for MMW security is provided in Table 2 and Figure 15. Since the relevant research in the MMW security field is not sufficient and many related methods are not as good as the SOTA deep learning methods, we chose the recently proposed multi-source aggregation transformer (MSAT) [40] and Swin–YOLO [5] algorithms as comparisons. The comparison results showed that although these two latest specialized MMW detectors have a certain improvement over the general purpose SOTA methods, the proposed MDIF-YOLO was still superior, because it revolutionizes how the MMW data are used and aggregated, which proves its effectiveness in the MMW detection field. A lower FAR shows the stability of the proposed method and makes its application scenario more comprehensive, thus increasing its competitiveness.

**Table 2.** Detection abilities and fake alert rates of MSAT, Swin–YOLO, and the proposed MDIF-YOLO.

| Algorithm | Precision | FAR | mAP |
|---|---|---|---|
| MSAT | 90.2 | 11.0 | 89.9 |
| Swin–YOLO | 89.8 | 11.4 | 89.7 |
| MDIF-YOLO | 92.2 | 6.4 | 91.8 |

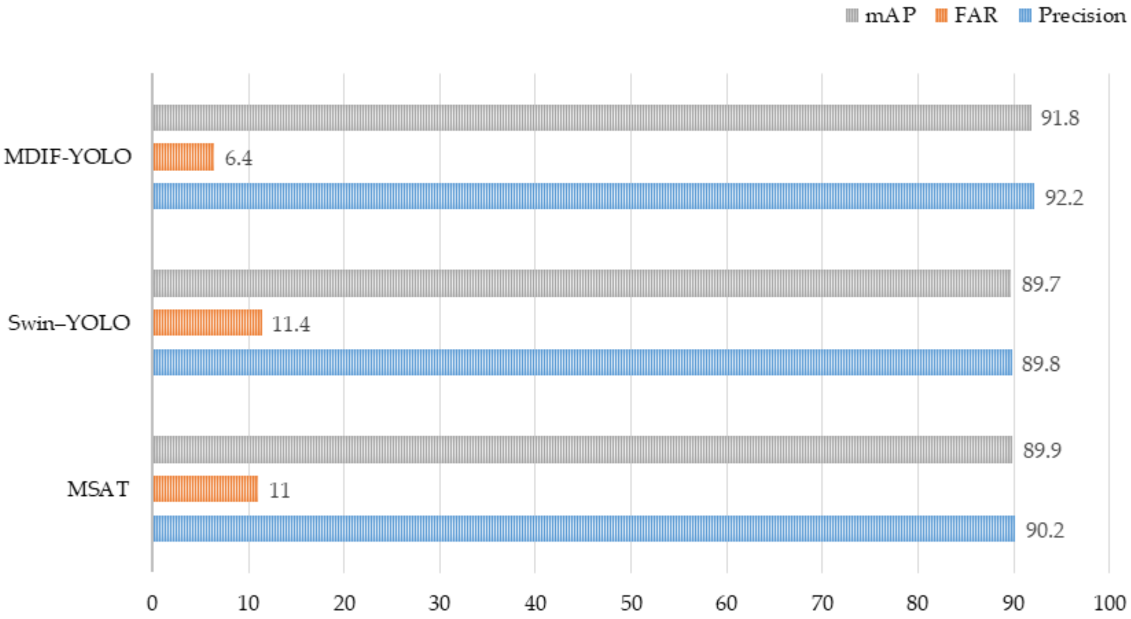


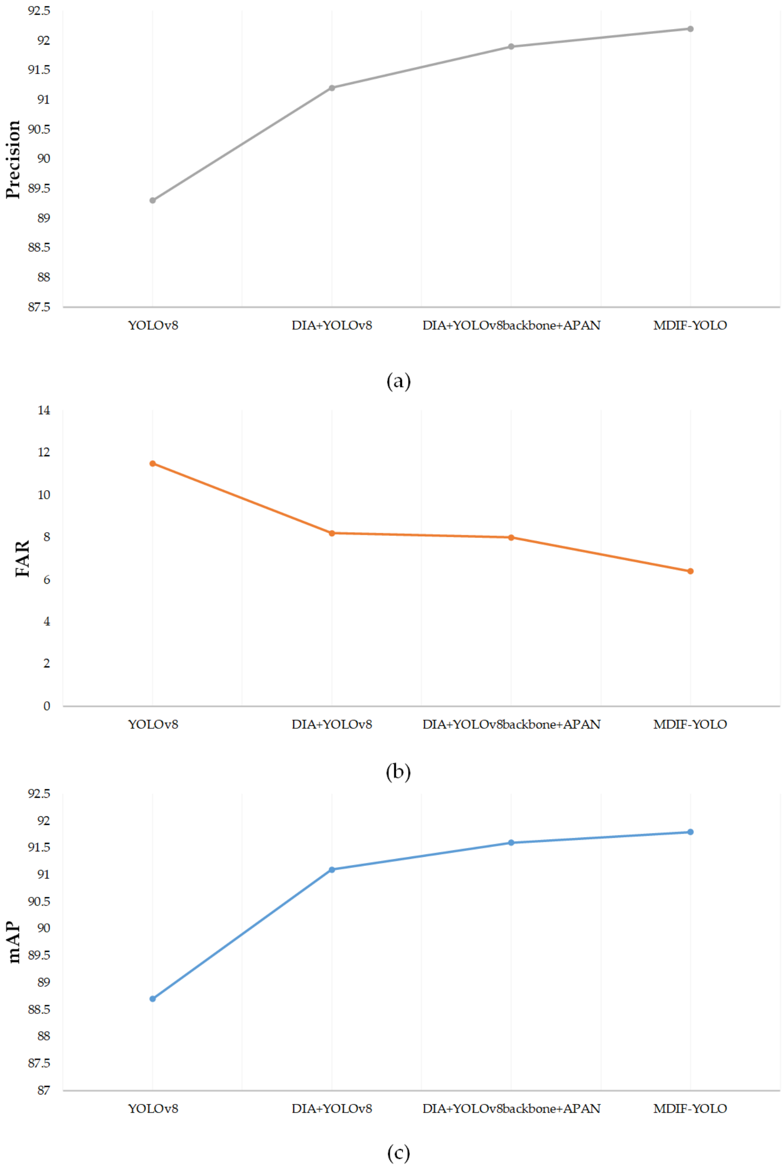
**Figure 15.** Performance comparison of MSAT, Swin–YOLO, and the proposed MDIF-YOLO.

### *3.5. Ablation Experiments*

Ablation experiments were conducted to evaluate the effectiveness of each proposed module in MDIF-YOLO. The results are given in Table 3 and Figure 16. The Precision and mAP results of YOLOV8 were 89.3% and 88.7%. When adding the proposed DIA module to YOLOV8, the two metrics reached 91.2% and 91.1%, an increase of about 2%. The FAR dropped from 11.5% to 8.2%, a decrease of 3.3%. Evidently, the multi-dimensional information fusion and differentiable data enhancement accomplished by the DIA module raised the upper limit of the available information and enhanced the overall performance. Subsequently, the APAN was used to replace the YOLOv8 neck, which led to a 2.6% and 2.9% growth for Precision and mAP compared with YOLOv8. The FAR decreased to 8%. The reason for this improvement is that the APAN relieves feature inconsistencies and avoids semantic gaps. Finally, the multi-view, multi-parameter mapping module was added to form the complete MDIF-YOLO method. The Precision and mAP achieved, respectively, a 2.9% and 3.1% improvement to 92.2% and 91.8%. The FAR had a significant 5.1% reduction to 6.4%. Thus, it was proven that the multi-view, multi-parameter mapping module plays a role in fine-tuning the detection results, supplementing missed detections, and suppressing error detections. From Figure 16, the use of multi-dimensional information fusion resulted in obvious performance gaps between MDIF-YOLO and YOLOv8. These ablation experiments verified that the DIA module, APAN, and multi-view, multi-parameter mapping module cooperate with each other to improve the performance and practicability of the algorithm from different aspects.

**Table 3.** Ablation studies of the proposed MDIF-YOLO.

| Algorithm | Precision | FAR | mAP |
|---|---|---|---|
| YOLOv8 | 89.3 | 11.5 | 88.7 |
| DIA + YOLOv8 | 91.2 | 8.2 | 91.1 |
| DIA + YOLOv8backbone + APAN | 91.9 | 8.0 | 91.6 |
| MDIF-YOLO | 92.2 | 6.4 | 91.8 |



(a)

(b)

(c)

**Figure 16.** Ablation experiment comparisons. (**a**) Precision comparisons; (**b**) FAR comparisons; (**c**) mAP comparisons.

## 4. Conclusions

We proposed the MDIF-YOLO algorithm, which pioneers fusing multimodal data in the MMW detection field. Multiple types of MMW data, such as the pixel, depth, phase, SNR, and multi-view information, are jointly used in the MDIF-YOLO to break the limitations of traditional MMW detectors that only using pixel information. Using online, differentiable, multimodal data enhancement, the DIA module in the proposed method increased the algorithm's robustness and generalization. Moreover, the extended multi-dimensional information is fully mined, extracted, and aggregated using the YOLOv8–APAN module. The bidirectional asymptotic aggregation and adaptive spatial weighted fusion techniques are combined in APAN to relieve feature inconsistencies and prevent semantic gaps. Finally, the multi-view and different DIA pattern detection results are refined in the multi-view, multi-parameter mapping module. A higher performance and better robustness can be achieved through mapping, fine-tuning, and padding. The experimental results demonstrate the superiority of the proposed MDIF-YOLO algorithm.

**Author Contributions:** Conceptualization, Z.C.; methodology, Z.C.; software, Z.C., R.T. and C.Y.; validation, Z.C., R.T. and D.X.; investigation, Z.C., C.Y., T.L. and Y.S.; writing—original draft preparation, Z.C.; writing—review and editing, Z.C.; visualization, Z.C., D.X. and C.Y.; supervision, Z.C. and R.T.; project administration, Z.C., D.X. and R.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Although author's organization Beijing Institute of Radio Metrology and Measurement does not have an ethics committee, author's BM4203 series MMW security scanners and the corresponding data utilization are in compliance with Chinese laws and regulations and ethics, and we have obtained relevant Chinese government utilization certificates.

**Informed Consent Statement:** The human data used in this paper were obtained from several of authors, and authors agree that this paper uses these data.

**Data Availability Statement:** The data in this research are available upon reasonable request due to restrictions involving privacy protection and product copyrights.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Liu, H.; Wang, S.; Jing, H.; Li, S.; Zhao, G.; Sun, H. Millimeter-Wave Image Deblurring via Cycle-Consistent Adversarial Network. *Electronics* **2023**, *12*, 741. [CrossRef]
2. Chen, X.; Yang, Q.; Wang, H.; Zeng, Y.; Deng, B. Adaptive ADMM-Based High-Quality Fast Imaging Algorithm for Short-Range MMW MIMO-SAR System. *IEEE Trans. Antennas Propag.* **2023**, *71*, 8925–8935. [CrossRef]
3. Li, S.; Wu, S. Low-Cost Millimeter Wave Frequency Scanning Based Synthesis Aperture Imaging System for Concealed Weapon Detection. *IEEE Trans. Microw. Theory Tech.* **2022**, *70*, 3688–3699. [CrossRef]
4. Meng, Z.; Zhang, M.; Wang, H. CNN with Pose Segmentation for Suspicious Object Detection in MMW Security Images. *Sensors* **2020**, *20*, 4974. [CrossRef]
5. Huang, P.; Wei, R.; Su, Y.; Tan, W. Swin-YOLO for Concealed Object Detection in Millimeter Wave Images. *Appl. Sci.* **2023**, *13*, 9793. [CrossRef]
6. Su, B.; Yuan, M. Object Recognition for Millimeter Wave MIMO-SAR Images Based on High-resolution Feature Recursive Alignment Fusion Network. *IEEE Sens. J.* **2023**, *23*, 16413–16427. [CrossRef]
7. Haworth, C.D.; De Saint-Pern, Y.; Clark, D.; Trucco, E.; Petillot, Y.R. Detection and Tracking of Multiple Metallic Objects in Millimetre-Wave Images. *Int. J. Comput. Vis.* **2007**, *71*, 183–196. [CrossRef]
8. Sheen, D.M.; McMakin, D.L.; Hall, T.E. Three-Dimensional Millimeter-Wave Imaging for Concealed Weapon Detection. *IEEE Trans. Microw. Theory Tech.* **2001**, *49*, 1581–1592. [CrossRef]
9. Grossman, E.N.; Miller, A.J. Active Millimeter-Wave Imaging for Concealed Weapons Detection. In *Passive Millimeter-Wave Imaging Technology VI and Radar Sensor Technology VII*; SPIE: Bellingham, WA, USA, 2003; Volume 5077, pp. 62–70.
10. Shen, X.; Dietlein, C.R.; Grossman, E.; Popovic, Z.; Meyer, F.G. Detection and Segmentation of Concealed Objects in Terahertz Images. *IEEE Trans. Image Process.* **2008**, *17*, 2465–2475. [CrossRef] [PubMed]
11. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Chen, T. Recent Advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]

12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
13. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
14. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
15. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic Detr: End-to-End Object Detection with Dynamic Attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 2988–2997.
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
19. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
20. Zhang, H.; Chang, H.; Ma, B.; Wang, N.; Chen, X. Dynamic R-CNN: Towards High Quality Object Detection via Dynamic Training. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 260–275.
21. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
22. Dai, J.; Li, Y.; He, K.; Sun, J. R-Fcn: Object Detection via Region-Based Fully Convolutional Networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1–9.
23. Gong, T.; Chen, K.; Wang, X.; Chu, Q.; Zhu, F.; Lin, D.; Feng, H. Temporal ROI Align for Video Object Recognition. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 1442–1450. [CrossRef]
24. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single Shot Multibox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14. Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
25. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
26. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo Algorithm Developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
27. Tian, Y.; Yang, G.; Wang, Z.; Wang, H.; Li, E.; Liang, Z. Apple Detection during Different Growth Stages in Orchards Using the Improved YOLO-V3 Model. *Comput. Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
28. Diwan, T.; Anirudh, G.; Tembhurne, J.V. Object Detection Using YOLO: Challenges, Architectural Successors, Datasets and Applications. *Multimed. Tools Appl.* **2023**, *82*, 9243–9275. [CrossRef] [PubMed]
29. Han, L.; Ma, C.; Liu, Y.; Jia, J.; Sun, J. SC-YOLOv8: A Security Check Model for the Inspection of Prohibited Items in X-ray Images. *Electronics* **2023**, *12*, 4208. [CrossRef]
30. Wang, J.; Wang, J.; Zhang, X.; Yu, N. A Mask-Wearing Detection Model in Complex Scenarios Based on YOLOv7-CPCSDSA. *Electronics* **2023**, *12*, 3128. [CrossRef]
31. Casas, E.; Ramos, L.; Bendek, E.; Rivas-Echeverría, F. Assessing the Effectiveness of YOLO Architectures for Smoke and Wildfire Detection. *IEEE Access* **2023**, *11*, 96554–96583. [CrossRef]
32. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Jain, M. Ultralytics/Yolov5: V7. 0-Yolov5 Sota Realtime Instance Segmentation. *Zenodo* **2022**.
33. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
34. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
35. Jocher, G.; Chaurasia, A.; Qiu, J. *YOLO by Ultralytics*; Ultralytics: Los Angeles, CA, USA, 2023.
36. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-The-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
37. Gupta, C.; Gill, N.S.; Gulia, P.; Chatterjee, J.M. A Novel Finetuned YOLOv6 Transfer Learning Model for Real-Time Object Detection. *J. Real-Time Image Process.* **2023**, *20*, 42. [CrossRef]
38. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-Aligned One-Stage Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 3490–3499.

39. Wang, X.; Gou, S.; Li, J.; Zhao, Y.; Liu, Z.; Jiao, C.; Mao, S. Self-Paced Feature Attention Fusion Network for Concealed Object Detection in Millimeter-Wave Image. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 224–239. [CrossRef]
40. Sun, P.; Liu, T.; Chen, X.; Zhang, S.; Zhao, Y.; Wei, S. Multi-Source Aggregation Transformer for Concealed Object Detection in Millimeter-Wave Images. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6148–6159. [CrossRef]
41. Liu, T.; Zhao, Y.; Wei, Y.; Zhao, Y.; Wei, S. Concealed Object Detection for Activate Millimeter Wave Image. *IEEE Trans. Ind. Electron.* **2019**, *66*, 9909–9917. [CrossRef]
42. Wang, C.; Shi, J.; Zhou, Z.; Li, L.; Zhou, Y.; Yang, X. Concealed Object Detection for Millimeter-Wave Images with Normalized Accumulation Map. *IEEE Sens. J.* **2021**, *21*, 6468–6475. [CrossRef]
43. Guo, D.; Tian, L.; Du, C.; Xie, P.; Chen, B.; Zhang, L. Suspicious Object Detection for Millimeter-Wave Images with Multi-View Fusion Siamese Network. *IEEE Trans. Image Process.* **2023**, *32*, 4088–4102. [CrossRef]
44. Yang, G.; Lei, J.; Zhu, Z.; Cheng, S.; Feng, Z.; Liang, R. AFPN: Asymptotic Feature Pyramid Network for Object Detection. *arXiv* **2023**, arXiv:2306.15988.
45. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
46. Jiang, Y.; Cui, J.; Chen, Z.; Wen, X. Concealed Threat Detection Based on Multi-View Millimeter Wave Imaging for Human Body. In Proceedings of the 2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP), Chongqing, China, 11–13 December 2019; pp. 1–4.
47. Jiang, Y.; Chen, Z.; Xiong, D.; Shi, J. Video-Rate Suspicious Object Detection for MMW Walk-Through Imaging System. In Proceedings of the 2021 46th International Conference on Infrared, Millimeter and Terahertz Waves (IRMMW-THz), Chengdu, China, 29 August–3 September 2021; pp. 1–2.
48. Wen, X.; Zhang, L.; Guo, W.; Fei, P. Active Millimeter-Wave Near-Field Cylindrical Scanning Three-Dimensional Imaging System. In Proceedings of the 2018 International Conference on Microwave and Millimeter Wave Technology (ICMMT), Chengdu, China, 7–11 May 2018; pp. 1–3.
49. Xiong, D.; Chen, Z.; Guo, W.; Wen, X. Near-field Millimeter Wave 3D Imaging Method for On-the-move Personnel. In Proceedings of the 2021 4th International Conference on Information Communication and Signal Processing (ICICSP), Shanghai, China, 24–26 September 2021; pp. 432–437.
50. Sahu, S.; Singh, A.K.; Ghrera, S.P.; Elhoseny, M. An Approach for De-Noising and Contrast Enhancement of Retinal Fundus Image Using CLAHE. *Opt. Laser Technol.* **2019**, *110*, 87–98.
51. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1792–1800. [CrossRef]
52. Hu, Y.; He, H.; Xu, C.; Wang, B.; Lin, S. Exposure: A White-Box Photo Post-Processing Framework. *ACM Trans. Graph. TOG* **2018**, *37*, 1–17. [CrossRef]
53. Li, X.; Yang, K.; Fan, X.; Hu, L.; Li, J. Fast and Accurate Concealed Dangerous Object Detection for Millimeter-Wave Images. *J. Electron. Imaging* **2022**, *31*, 023021. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.