applied sciences

MDPI

Article

Swin-YOLO for Concealed Object Detection in Millimeter Wave Images

Pingping Huang [1,2], Ran Wei [1,2], Yun Su [1,2,*] and Weixian Tan [1,2]

[1] College of Information Engineering, Inner Mongolia University of Technology, Hohhot 010051, China
[2] Inner Mongolia Key Laboratory of Radar Technology and Application, Hohhot 010051, China
* Correspondence: suyun@imut.edu.cn

Abstract: Concealed object detection in millimeter wave (MMW) images has gained significant attention in the realm of public safety, primarily due to its distinctive advantages of non-hazardous and non-contact operation. However, this undertaking confronts substantial challenges in practical applications, owing to the inherent limitations of low imaging resolution, small concealed object size, intricate environmental noise, and the need for real-time performance. In this study, we propose Swin-YOLO, an innovative single-stage detection model built upon transformer layers. Our approach encompasses several key contributions. Firstly, the integration of Local Perception Swin Transform Layers (LPST Layers) enhanced the network's capability to acquire contextual information and local awareness. Secondly, we introduced a novel feature fusion layer and a specialized prediction head for detecting small targets, effectively leveraging the network's shallow feature information. Lastly, a coordinate attention (CA) module was seamlessly incorporated between the neck network and the detection head, augmenting the network's sensitivity towards critical regions of small objects. To validate the efficacy and feasibility of our proposed method, we created a new MMW dataset containing a large number of small concealed objects and conducted comprehensive experiments to evaluate the effectiveness of overall and partial improvements, as well as computational efficiency. The results demonstrated a remarkable 4.7% improvement in the mean Average Precision (*mAP*) for Swin-YOLO compared with the YOLOv5 baseline. Moreover, when compared with other enhanced transformer-based models, Swin-YOLO exhibited a superior accuracy and the fastest inference speed. The proposed model showcases enhanced performance and holds promise for advancing the capabilities of real-world applications in public safety domains.

Keywords: millimeter wave images; concealed object detection; Swin Transformer; attention mechanism

check for updates

Citation: Huang, P.; Wei, R.; Su, Y.; Tan, W. Swin-YOLO for Concealed Object Detection in Millimeter Wave Images. *Appl. Sci.* **2023**, *13*, 9793. https://doi.org/10.3390/app13179793

Academic Editor: Amit Kumar Mishra

Received: 25 July 2023
Revised: 26 August 2023
Accepted: 28 August 2023
Published: 30 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the increasing emphasis on public safety in modern society, security inspection has become a necessary part of social development. It requires the fast and accurate detection of hazards from the source to minimize unnecessary losses. The traditional methods of security inspection are widely used, but they generally suffer from different defects: optical methods, as an early detection method, have a short wavelength and are unable to penetrate clothing, which makes it difficult to detect hidden objects; metal detection instruments can only identify metal targets of certain sizes and are unable to identify non-metallic items [1], such as fire sources; and X-ray security inspections, although penetrating, will generate harmful ionizing radiation to the human body, affecting human health [2]. These negative factors have greatly affected the effectiveness and reliability of security checks, resulting in a hidden danger for safety.

In recent years, millimeter-wave (MMW) imaging technology has received widespread attention in the field of security inspection, owing to the excellent penetration characteristics and harmlessness of MMW radar [3,4]. MMW radar has a wide-ranging detection capability

for both metallic and non-metallic objects [5]. The MMW imaging system emits non-ionizing electromagnetic waves [6] and receives reflected energy from various parts of the human body and their surface attachments. MMW images are generated based on the changes in these reflected energies. MMW imaging technology can accurately reflect the types and distribution of human concealed objects, so the object detection of concealed object images naturally becomes the major focus of current research. Figure 1 illustrates the MMW imaging principle.

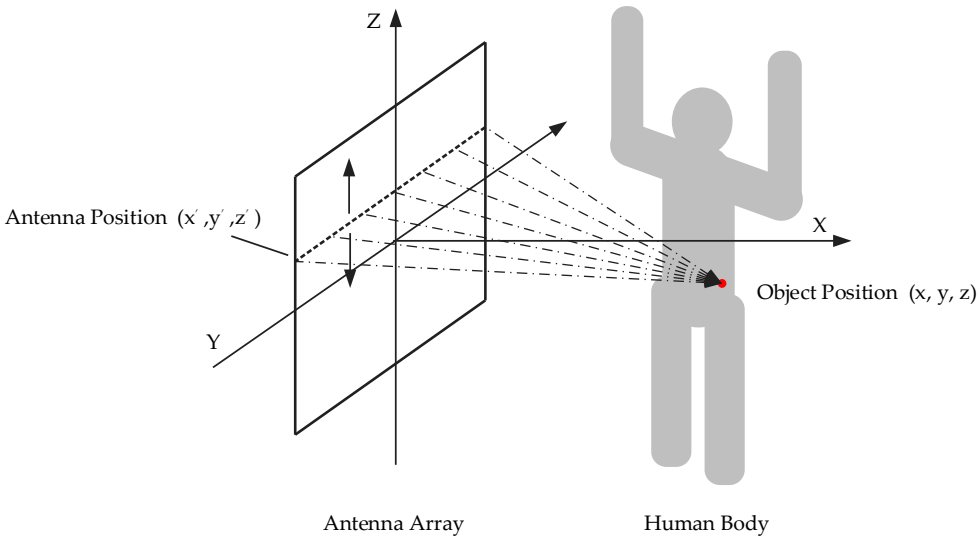


Figure 1. The model of millimeter wave imaging principle.

Traditional methods for detecting concealed objects in images rely on geometric feature matching to segment the object from the body region [7], followed by an analysis of the target area for detection and classification. However, these methods are limited to analyzing single images and have poor real-time performance, which cannot meet the demands of modern security inspection. With the rapid development of deep learning, deep learning-based object detection methods have been widely applied in the concealed object detection for the MMW image. For instance, Yang et al. [8] employed a deep learning-based deformable body partition model to combine suspicious target information with body part information and detect concealed objects, while Meng [9] proposed a convolutional neural network (CNN) with pose segmentation functionality to divide human images into body part images, and then detected clutter anomalies in body parts to identify objects, both of which achieved remarkable performance and effectively addressed the challenges posed by traditional methods. He et al. [10] proposed a detection framework using multi-level features information from cross-section sequences for detecting MMW concealed objects. The framework consists of a CNN and a Long Short-Term Memory (LSTM) network, achieving more effective feature fusion.

Although achievements have been made in MMW image concealed object detection based on deep learning, it still faces a series of problems. Firstly, in terms of image quality, the difference in the intensity of MMW reflection between the human body area and the concealed objects is not obvious, and there is insignificant difference in image grayscale [11]. This makes it difficult to distinguish the texture of human body parts and concealed objects. Moreover, MMW radar imaging is affected by environmental and random noise, resulting in low imaging signal-to-noise ratio. The reduction in available information in the concealed object image area, as well as the complex image information and noise impact, all have an impact on the detection accuracy of the network. Secondly, with regards to detecting targets, the concealed objects in MMW images are small, their size is much smaller than the human body, and the resolution of concealed objects is low, resulting in insufficient feature information [10]. In deep learning networks, the features of small objects gradually decrease as the network deepens, bringing great difficulties to detection.

For the problems in the MMW image of detecting hidden objects, a novel concealed object detection method based on the YOLOv5 [12] model is proposed. YOLOv5 was selected as the framework for its performance in harmonizing accuracy and speed, and it concurrently ensured engineering forward compatibility. Moreover, a series of enhancements were introduced. First, to make the YOLOv5 network have the ability of global context modeling and have better performance in detecting complex concealed object features, we have made improvements to the neck network and added Local Perception Swin Transformer Layers, which can enhance the context information connection of the feature map through a self-attention mechanism and make better use of local information through a Local Perception Module. Second, we added a detection branch for small objects in the shallow layers of the network and introduced a bidirectional feature pyramid network (BiFPN) in the path aggregation neck. Through the bidirectional cross-scale connections and fast normalization fusion, the feature fusion capability of the network was enhanced. The small object detection branch can retain the feature information of small concealed objects to the greatest extent, thus solving the problem of feature information loss caused by neural network down-sampling. Finally, we added the Coordinate Attention Module at the connection between each detection head and the neck network, which makes the network more sensitive to positional information and focuses on critical image information. Experimental results showed that, in the detection of concealed objects in MMW concealed images, the performance of our proposed object detector was superior to the original YOLOv5 network.

The main contributions of this paper can be summarized as follows:

1. More comprehensive contextual information is utilized through improved Swin Transformer. The Local Perception Swin Transformer Layers are added to the neck network of YOLOv5, which enhances the network's ability to model feature context through the introduction of a window self-attention mechanism and Local Perception Module.
2. The features of small targets are sufficiently extracted through more refined feature extraction and fusion networks. Additional feature fusion layers and prediction heads for small concealed objects are added to YOLOv5, enhancing the network's detection performance for MMW targets. BiFPN is introduced to strengthen the network's feature fusion ability.
3. The key features of the concealed object are fully used. The lightweight Coordinate Attention Modules are added to the connection between the neck and detection heads of YOLOv5, which improves the extraction and utilization of critical features.
4. We collected an MMW concealed object image dataset to support the training of different network models and test the detection performance of concealed objects. The dataset includes five classes of suspicious items and a total of 82,014 MMW images.

2. Related Work

2.1. Object Detection

The object detection framework based on convolutional neural networks (CNN) is currently a hot topic in computer vision research. Researchers have widely classified it into two-stage detection methods and single-stage detection methods. The two-stage detection, also known as region proposal-based detection, achieves a high detection accuracy and mainly includes the Region-CNN (RCNN) series. In the first stage of the RCNN [13] network, a selective search algorithm is used to generate candidate regions on the input image, and convolutional neural networks are used to extract features from these candidate regions, ensuring sufficient precision and recall. In the second stage, the objects within the candidate regions are classified and regressed. The Fast-RCNN [14] algorithm integrates RCNN and a Spatial Pyramid Pooling Network (SPP-Net) [15] and uses Region of Interest-Pooling (ROI-Pooling) to map candidate region feature windows of different sizes onto feature vectors of the same size, thus unifying the input size. Faster-RCNN [16], on the other hand, employs the Region Proposal Network (RPN) in the first stage and the anchor mechanism to generate object candidate regions, and classifies the objects within the anchor

in the second stage, further improving the detection speed and accuracy of the network. Although the RCNN network has advantages in accuracy, the two-stage object detection network still cannot achieve real-time detection.

Single-stage detection methods consider object detection as a one-time problem and use regression methods to obtain the coordinates of object bounding boxes and class confidence. This enables the high-speed positioning and classification of objects in an end-to-end manner. Single-stage methods mainly include the Single Shot Detector (SSD) [17] and the YOLO series algorithms, where SSD introduces the RPN structure and uses multi-scale feature maps to detect objects. YOLO [18] is the landmark algorithm for single-stage object detection, establishing the main idea of the subsequent framework, using GoogLeNet as the feature extraction network. YOLOv2 [19] improves the backbone network by using Darknet-19 to extract object features and adding batch normalization layers to each convolutional layer to reduce the sensitivity of the model to parameters. YOLOv3 [20] introduces a residual network module and uses Darknet-53 as the backbone network, predicting objects separately with multi-scale feature maps, which improves the detection performance for objects with large scale differences. YOLOv4 [21] introduces Mosaic data augmentation at the input end, the backbone network uses the Cross Stage Partial Network Darknet-53 (CSPDarknet-53) to avoid the problem of repeated gradient information, and it uses Dropblock to randomly reduce the number of neurons to simplify the network operation. Also, YOLOv4 uses a Feature Pyramid Network (FPN) [22] combined with a Path Aggregation Network (PANet) [23] structure for multi-scale feature extraction to fully integrate multi-scale features, and achieves higher accuracy than YOLOv3 at the same speed. The strategy of not relying on region detection greatly improves the detection speed of the single-stage methods, balancing accuracy and detection speed well. YOLOv5 [12], as the culmination of the YOLO series, has all the advantages of this series of networks, achieving further improvements in detection accuracy and speed. It includes four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. With increasing network depth and width, the extraction and fusion ability of features are continuously enhanced, which is suitable for different scenarios and meets the requirements of modern object detectors. Although there are many new versions of the YOLO series [24–26], the specific architecture tends to be consistent with the design philosophy of YOLOv5. Therefore, this article adopts YOLOv5l as the basic framework for object detection.

2.2. The Attention Mechanism

The attention mechanism is a structure that focuses on key information and has profound implications on computer vision image processing. By focusing on specific regions of an image with high weights, the attention mechanism simulates human attention and enables neural networks to have perceptual capabilities. Specifically, it reduces the amount of sample data that needs to be processed and increases the feature matching of samples. Earlier, researchers combined Recurrent Neural Network (RNN) models with attention mechanisms [27] for image classification, which achieved correlation expression between features based on solving the problem of gradient explosion in RNN networks. In CNN, a Squeeze-and-Excitation Network (SE-Net) [28] focuses on the relationship between image channels. It weights the features between different channels to emphasize the effective information and suppress the invalid information, and to improve the detection performance of neural networks. The Convolutional Block Attention Module (CBAM) [29], as a lightweight attention module that is easy to deploy, combines channel and spatial attention. It adaptively adjusts the original input feature maps based on the attention feature maps in both channel and spatial dimensions.

The self-attention mechanism is another form of attention mechanism, which, unlike spatial and channel attention, shows good performance in characterizing context relationships and is widely used in natural language processing (NLP) tasks. In computer vision tasks, a Non-Local network [30] was the first to apply the self-attention mechanism to the field of object detection to address the limitation of convolutional units in local region

operations. Self-attention can enhance the acquisition of global information and guide the feature extraction of convolutional neural networks. Moreover, this mechanism can also enhance the global semantic characteristics of the deep network features, strengthening the contextual connection between foreground and background. Visual Transformer (Vit) [31], which is based on pure self-attention mechanisms, has demonstrated that a pure attention network can achieve a classification performance that is comparable to that of CNNs. The network uses a standard Transformer Encoder to globally model image patches. A Swin Transformer [32], on the other hand, combines the advantages of Transformers and CNN, uses a hierarchical structure to obtain information at different scales, and improves the computational efficiency of the Transformer [33] using shifted windows. Swin Transformers perform well in multiple tasks such as detection, classification, and segmentation, filling the gap in the downstream tasks of Vit. Since the concealed object within MMW images are generally small and the MMW image features are more complex, the network requires a better detector to accomplish the detection of MMW concealments. The local information modeling of CNN has been proven effective, while the Swin Transformer can fully extract high-level semantic features of the image and establish context information connections in the image. Therefore, we explore the combination of CNN and Swin Transformers to improve the performance of MMW object detection.

2.3. Concealed Object Detection on Millimeter Wave Images

Earlier, researchers used hand-designed features to detect concealed objects in MMW images and segment the concealed object part. Shen et al. [34] modeled the hidden object by analyzing the denoised image histogram and the internal temperature distribution, then evolved curves along the isocontours of the image to separate objects from the background. Yeom et al. [35] extracted geometric feature vectors from binary segmented images using preprocessing, Principal Component Analysis (PCA), and size normalization to separate concealed objects from body regions. Lee et al. [36] used a multi-level segmentation approach containing the Expectation-Maximization (EM) algorithm to detect and segment concealed object regions. However, due to the small size of the feature regions and the limited number of feature parameters, hand-designed features were unable to fully and effectively express the characteristics of concealed objects.

With the development of deep learning, researchers have adopted CNN for MMW concealed object image detection. This is because CNNs can encode features and implicitly learn low-order to high-order features from training data, making the extracted features more discriminative. Liu et al. [37] proposed a context-embedded object detection network and applied dilated convolution to expand the resolution of feature maps. By extracting detailed features and contextual information, this network improves the detection performance of small objects. Yang et al. [8] proposed a deformable body partition model that divides MMW human body images into multiple regions and detects hidden objects for different body regions, dynamically tracking objects from different angles using spatiotemporal context information. Meng et al. [9] used the clutter features of MMW human body segmentation images for deep learning training, and the clutter anomaly detector had a stronger generalization performance. Li et al. [38] proposed a two-stage MMW human body concealed object image detector. The network improved the fusion relationship between adjacent feature maps in the FPN and used multi-angle images to optimize the detection results. Zhang [39] proposed a domain adaptive detector for MMW image concealed objects. The network is composed of a detection module with hierarchical attention units and an unsupervised domain adaptive module. The detection module fused low-level and high-level features, and the domain adaptive module extracted the domain differences of features. Yuan et al. [40] proposed a multi-path feature pyramid (MPFP) model and an improved residual block distribution to enhance the recognition accuracy. Wang et al. [41] proposed a concealed object detection method based on normalized accumulation maps (NAM), which can reveal frequently occurring concealed object locations. Most of the methods in the above-mentioned research are based on two-stage detectors. However,

MMW human body concealed object detection requires high real-time performance. Thus, it is necessary to explore object detectors that balance real-time performance and accuracy. Table 1 shows the comparison of the methods above. Detailed information on the key methods, models, average precision, and inference time used in each of the studies are provided.

Table 1. The comparison of the concealed object detection on millimeter wave images.

Reference	Model/Method	AP (%)	Inference Time (FPS)
Shen et al. [34]	Denoised Image Histogram + Internal Temperature Distribution + Isocontours	/	/
Yeom et al. [35]	Preprocessing + PCA + Size Normalization	/	/
Lee et al. [36]	EM	/	/
Liu et al. [37]	ZF-Net + High-Resolution Features + Context Embedded Module	84.65	/
Yang et al. [8]	Darknet-53 + Deformable Body Partition Model	80.1	25
Meng et al. [9]	Convolution Pose Machine	/	/
Li et al. [38]	ResNet50 + FPN	82	/
Zhang et al. [39]	SSD + Bottom-Up Manner + Top-Down Manner + K-means	88.26	15
Yuan et al. [40]	Residual Block + MPFP	82.39	24.2
Wang et al. [41]	YOLOv2 + NAM	69.67	/

3. Theoretical Model

3.1. Overview of the YOLOv5 Method

YOLOv5 is a single-stage object detection network consisting of three components: the backbone network, the neck network, and the detection head. The backbone network is responsible for image feature extraction and generates multi-scale feature maps through layer-by-layer convolution and pooling operations, The backbone network consists of the CSPDarknet53 network [42] and the Spatial Pyramid Pooling-Fast (SPPF) layer. Among them, the design of the CSP structure reduces the size of the network through cross-layer connections and maintains accuracy while reducing computational loss. The SPPF module improves the structure of the SPP layer [15] by fusing local and global features to provide a more comprehensive feature representation. The neck network processes the multi-scale feature maps and employs a pyramid structure of PANet [23] to fuse information from low-level spatial feature maps and high-level semantic feature maps. Regarding the detection head part, YOLOv5 transforms the object detection task into a regression task [43,44], using the feature maps from the neck network as the basis for classification and localization. YOLOv5 performs accurate classification, localization, and confidence estimation by employing grid-based anchor box regression. For predicting anchor boxes, it utilizes the CIoU loss [45] as the bounding box loss function.

3.2. The Proposed Object Detection Network

Based on the characteristics of abundant small objects and special distribution in MMW images, we improved the YOLOv5 network architecture. Figure 2 illustrates the architecture of Swin-YOLO proposed for object detection on MMW human concealed object images.

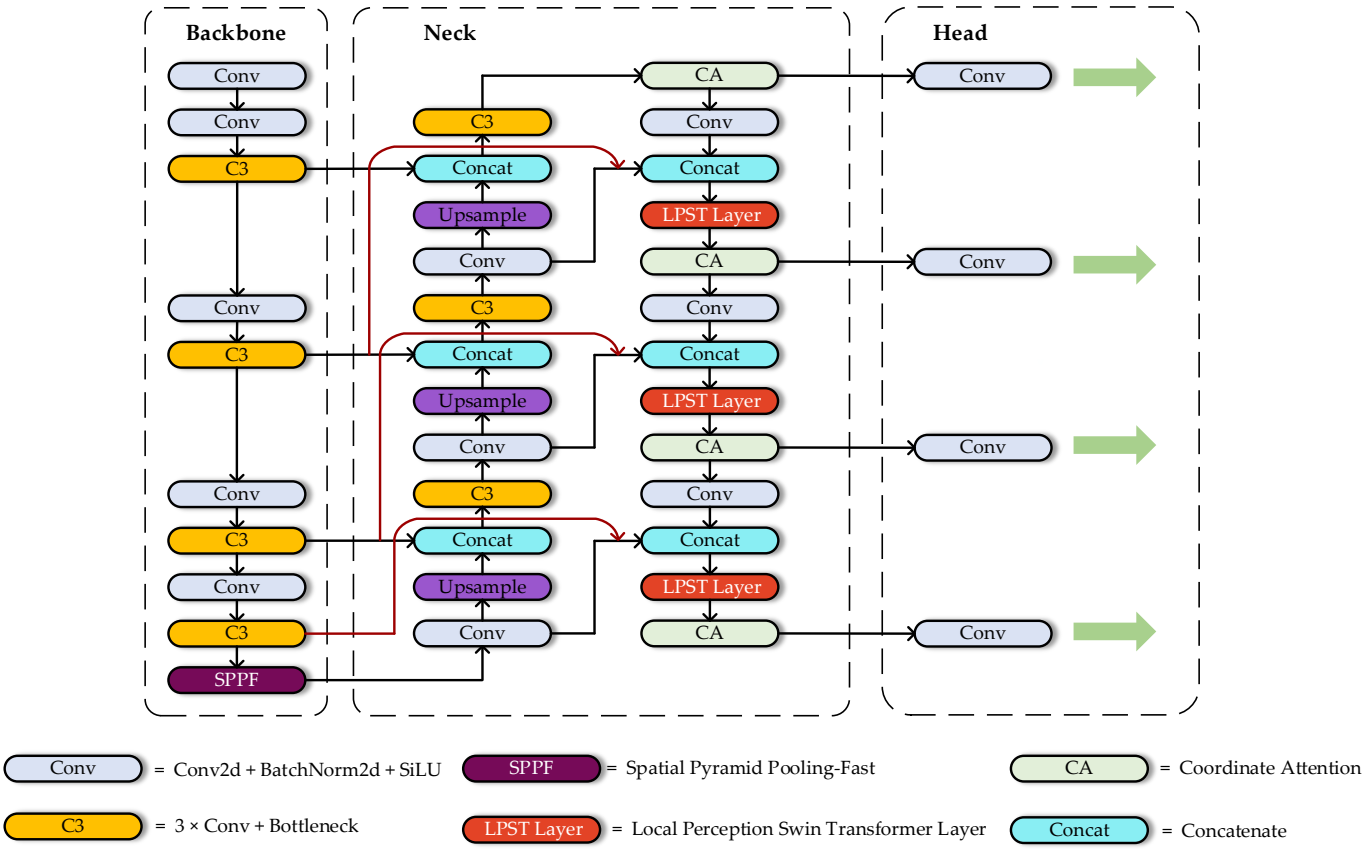


Figure 2. The architecture of Swin-YOLO. We added a small object feature extraction detection layer and additional detection heads, and used BiFPN to fuse features. The red lines represent additional feature fusion; Local Perception Swin Transformer Layers were introduced to the network neck; Coordinated Attention Modules were added at the connection between each detection head and neck network.

3.2.1. Feature Representation Layers and Prediction Head for Small Objects

The YOLOv5 network extracts features from three different-sized feature maps. The high-level feature map contains rich semantic information but has a lower resolution, resulting in imprecise object localization and a significant loss of information for small objects. In contrast, the low-level feature map has higher amounts of localization information but less semantic information, which is beneficial for detecting small objects.

In MMW human body images, concealed objects are small in size and have complex features. To make better use of the shallow feature information extracted from the backbone, we added a feature representation layer and a detection head for small-scale objects on top of the original YOLOv5 neck network. The new added layers will compensate for the loss of information caused by multiple down-sampling. Furthermore, it can enhance the feature fusion effect of small objects to adapt to small object datasets in MMW images.

The small-scale feature representation layer utilizes the low-level information. A horizontal residual connection path was established with the shallow layers of the backbone, obtaining high-resolution mapping. The small-scale feature representation layer connects the original feature extraction layer through a top-down pathway, which fuses low-resolution, high-semantic features with high-resolution, and low-semantic features. The small-object detection head is set at the newly added feature representation layer, making the object detector more sensitive to small objects. With the added feature fusion layer, the neck network now has a four-layer detection head structure, which is more suitable for detecting small concealed objects in MMW images.

3.2.2. Coordinate Attention Module

The Coordinate Attention (CA) Module [46] is a lightweight attention module that embeds the object positional information in the channel attention. It can generate attention feature maps in two coordinate directions and resize the weights of the two-dimensional feature maps. By multiplying the input feature map with the attention feature, the final attention feature is obtained, as shown in Figure 3.

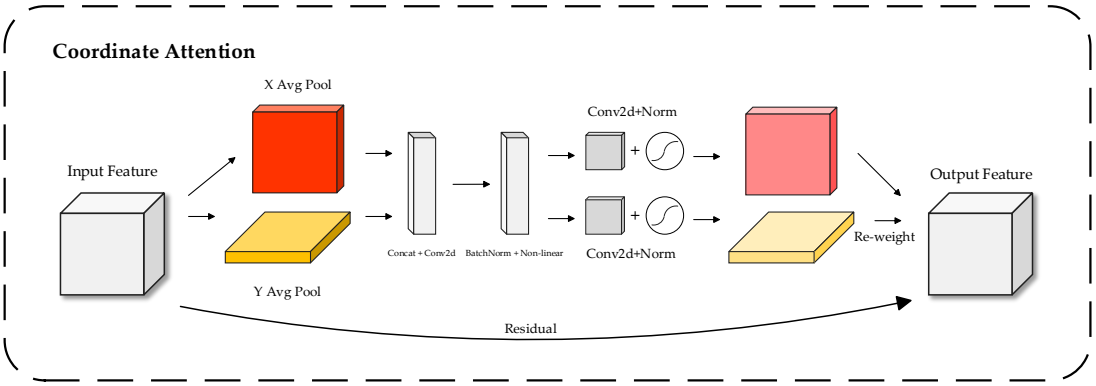


Figure 3. The architecture of Coordinate Attention. The Coordinate Attention sub-channel generates attention simultaneously in both X and Y directions. X Avg Pool and Y Avg Pool represent 1D horizontal global pooling and 1D vertical global pooling, respectively.

The feature map of the input image is represented as F. The CA module first performs average pooling operations along the horizontal and vertical coordinates of each channel of the input feature vector, obtaining a pair of direction-sensitive feature maps z^h and z^w. These two 1D feature encoding processes allow the model to capture long-range dependencies between channels in one direction while preserving the object's positional information in the other direction.

The intermediate feature map encodes the spatial feature maps along the horizontal and vertical directions. We first concatenate the z^h and z^w feature maps along the spatial dimension and encode them using a shared 1×1 convolutional function F_1, followed by a nonlinear activation function δ to output the intermediate feature map f. This is shown in Equation (1):

$$f = \delta\left(F_1\left(\left[\left(z^h\right) + \left(z^w\right)\right]\right)\right), \tag{1}$$

Then, we split f along the spatial dimension and use 1×1 convolutional functions F_h and F_w along different spatial directions to transform the channel numbers of f^h and f^w. We then obtain attention weights g^h and g^w with the same number of channels as the input F, as shown in Equation (2):

$$g^{(h,w)} = \sigma\left(F_{(h,w)}\left(f^{(h,w)}\right)\right), \tag{2}$$

where (h, w) represents either h or w, and σ represents the sigmoid activation function.

Finally, g^h and g^w are applied to the input feature F to obtain the final CA feature map M_c, as shown in Equation (3):

$$M_c = F_i \times g^h \times g^w, \tag{3}$$

For MMW human body images, the features of the concealed object are similar to those of the human body region. The concealed objects are small with complex textures, which can cause confusion between the detected object and the background, making it difficult to accurately identify the concealed object. Therefore, it is especially important to focus on the key areas of the feature map. After extracting the concealed object features with the backbone network, the multi-scale features are further processed by the neck network. As the key to feature fusion, the neck network can utilize the feature map better.

Thus, in order to refine the position information in the neck network and fuse the more important information from different depth feature maps, we combine the lightweight Coordinate Attention Module with the neck network of Swin-YOLO to improve the detection performance of the detection head.

3.2.3. Local Perception Swin Transformer (LPST) Layer

The proposed Local Perception Swin Transformer (LPST) Layer consists of two consecutive Local Perception Swin Transformer Blocks, which, combined with the Local Perception Module (LPM), enable the Swin Transformer to obtain local information. The structure of the LPST Layer is shown in Figure 4a.

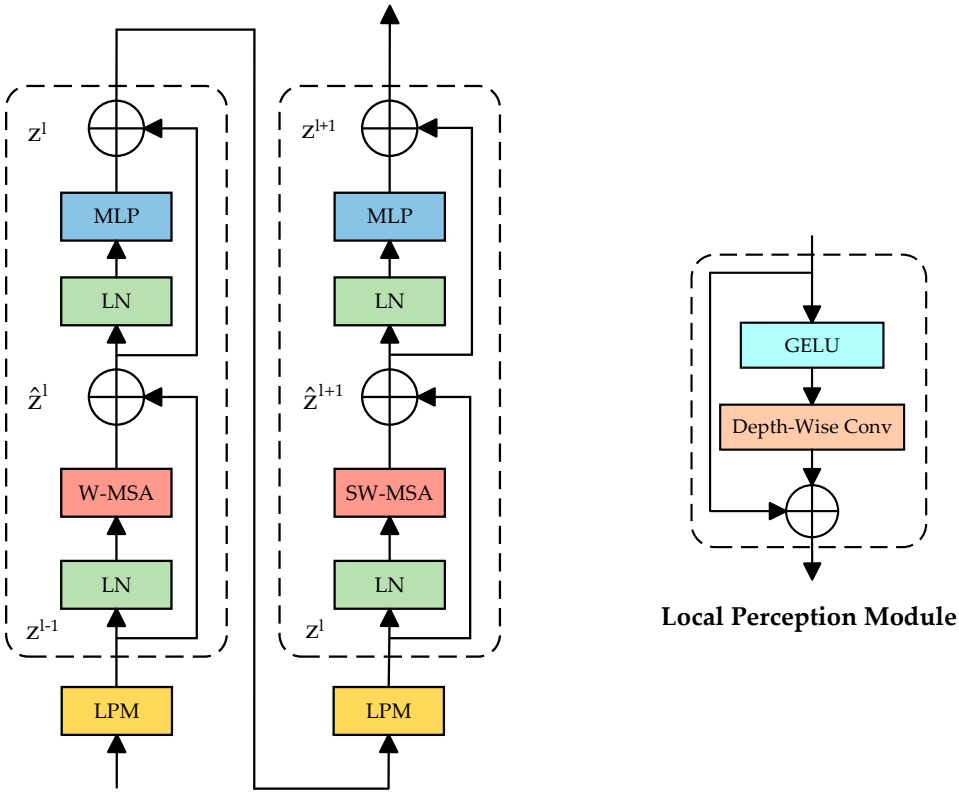


Figure 4. The architecture of the Local Perception Swin Transformer (LPST). (**a**) The detailed structure of the Local Perception Swin Transformer Layer; (**b**) The detailed structure of the Local Perception Module.

The key component of the LPST Layer is the Swin Transformer Layer. The Swin Transformer Layers consist of two successive Swin Transformer Blocks in series. The two Swin Transformer Blocks are alternatively equipped with Window Multi-head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) modules. The first Swin Transformer Block computes multi-headed attention (MSA) in small windows within an image, and establishes a self-attention interaction between windows in the second Block.

The Swin Transformer Block has a similar structure to the Transformer Block, with the difference that the MSA is replaced by the W-MSA and SW-MSA, while the rest of the structure remains the same. Specifically, the Swin Transformer Block consists of a window-based MSA module and an MLP layer. The W-MSA module, the SW-MSA module, and the MLP module are connected with LayerNorm layers and use the residual structure.

MSA is the core of the Transformer, which integrates the self-attention matrices of multiple independent subspaces through parallel self-attention calculations. For a feature maps $X \in R^{H\times W\times C}$, after being transformed by linear projection and reshaping operations, it becomes Q, K, $V \in R^{N\times C'}$, where $N = H \times W$. The self-attention can be expressed as follows:

$$Z = AV, \tag{4}$$

$$A = softmax\left(QK^T\right), \tag{5}$$

where $A \in R^{N\times N}$ is the attention matrix that represents the relationship between all elements and other elements in the feature map. Output Z aggregates global information.

The W-MSA module is an improvement of MSA. Unlike the Transformer, it calculates self-attention within a window and introduces cross-window connections through the shifted windows. The design of shifted windows reduces the computational complexity of the Transformer and enables information transmission between different windows. In particular, for the feature map $X \in R^{H\times W\times C}$, the window size is m × m and the computational complexity Ω is as follows:

$$\Omega(\text{MSA}) = 4HWC^2 + 2(HW)^2C, \tag{6}$$

$$\Omega(\text{W-MSA}) = 4HWC^2 + 2(HW)^2C, \tag{7}$$

The SW-MSA module solves the problem of a lack of information interaction of W-MSA. It achieves cross-window communication through the shifted windows, as shown in Figure 5. First, the Swin Transformer uses W-MSA in Figure 5a to performs window partitioning of the input feature map and calculates the self-attention in each window. Subsequently, as shown in Figure 5b, each window is moved upwards and to the left by half of the window size, which generates new windows that provide connections between windows of Figure 5a. For efficient batch computation, the green, yellow, and red regions in Figure 5c are moved to the bottom, right, and bottom-right of the image, respectively, as illustrated in Figure 5d. In comparison with the YOLOv5 improved network that utilizes the Swin Transformer Encoder [47], SW-MSA effectively facilitates cross-window information interaction.

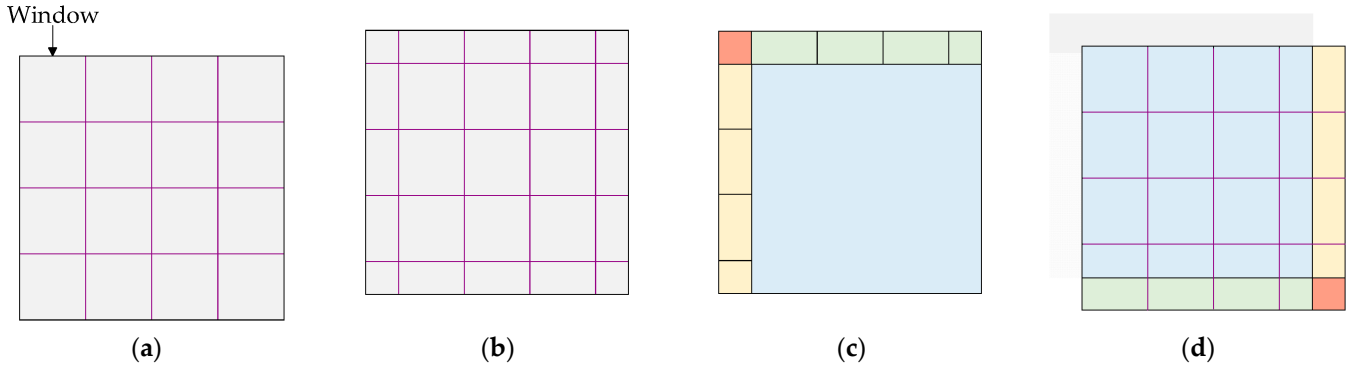


Figure 5. Schematic diagram of SW-MSA. (**a**) Window partitioning scheme of the regular W-MSA; (**b**) Shifted window partitioning of the SW-MSA; (**c**) Action of the shifted windows; (**d**) Effective batch calculation process used for self-attention in the shift window partitioning.

Although the Swin Transformer alleviates the problem of lacking local correlation and structural information caused by positional encoding through shifted windows, large-scale spatial contextual information still cannot be well encoded. Therefore, we added the key component of extracting local information, the Local Perception Module, before each Swin Transformer Block. The specific structure of LPM is shown in Figure 4b, which consists of depth-wise convolution and GELU activation function, and the local feature is extracted

through residual structure. Depth-wise convolution can increase the receptive field of spatial images, so as to better encode a wide range of contextual information.

LPM can be defined as the Equation (8):

$$\mathrm{LPM}(X) = \mathrm{GELU}(\mathrm{DWConv}(X)) + X, \tag{8}$$

where $X \in R^{H \times W \times d}$, $H \times W$ is the resolution input, d represents the dimensionality of the feature, and DWConv $(\cdot)$ represents depth-wise convolution.

The Swin Transformer possesses a powerful context modeling capability to capture long-range dependencies, but it still faces challenges. Recently, in the field of optical image object detection, researchers have incorporated the Transformer into the basic YOLOv5 module [48], using the Transformer Block combined with the C3 module to replace the CSP Bottleneck in the original version of YOLOv5, which is called C3TR, and its specific structure is shown in Figure 6.

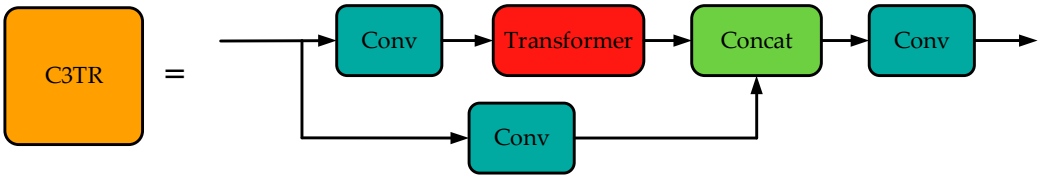


Figure 6. The architecture of C3TR.

However, according to the experimental results in Section 4.3, the Transformer brings a higher computational cost and the hybrid structure of C3TR performs poorly. Therefore, we use the LPST Layer, which alleviates the impact of the Transformer's quadratic computational complexity. Moreover, LPST restricts self-attention calculation within local windows through a shifting window mechanism, which brings higher efficiency in self-attention computation. LPST also combines the advantages of CNNs to optimize local information acquisition. Since MMW human concealment security inspection requires real-time performance, we add LPST Layer in the neck network part, which can not only enable the YOLOv5 network to obtain global contextual information for efficient detection of MMW small objects but also effectively save computational costs.

3.2.4. Improvement of Multi-Scale Feature Fusion Network

The fusion of multi-scale features is a key design of YOLOv5 object detection. In the neck network, given the input feature P_i^{in} of level i, YOLOv5 converts the intermediate layer feature P^{td} from different network levels to suit the resolution of the current input feature level i. Then, the feature fusion is completed using channel addition, and output feature P_i^{out} from level i. The intermediate feature P^{td} and output feature P^{out} in the neck network are shown in Equations (9) and (10).

$$P_i^{td} = Conv\left(P_i^{in} + Resize\left(P_{i+1}^{td}\right)\right), \tag{9}$$

$$P_i^{out} = Conv\left(P_i^{out} + Resize\left(P_{i-1}^{out}\right)\right), \tag{10}$$

where *Conv* represents convolutional operations related to feature processing, and *Resize* represents operations used for resolution matching.

However, this simple fusion method fails to distinguish between feature maps and neglects the differences between them. Since feature maps of different resolutions contain different information, their importance for fusing input feature maps varies. Moreover, deep feature maps contain stronger semantic information while shallow feature maps contain stronger positional information. Therefore, building upon the framework discussed in Section 3.2.1, we incorporate the concept of a Bi-Directional Feature Pyramid Network (BiFPN) [49] and reassess the contribution of different scale features to output features. The specific structure of the feature fusion layer is illustrated in Figure 7.

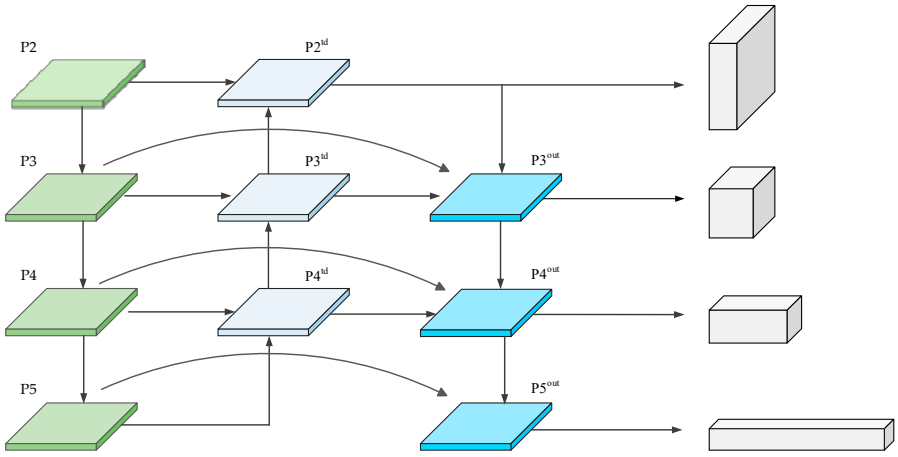


Figure 7. Improved multi-scale feature fusion network.

The improved neck network first introduces bi-directional cross-scale connections to aggregate the input features of levels P3, P4, and P5, ensuring the integrity of the input features. Secondly, a fast normalization fusion method is designed, using a trainable parameter w_i as the weight for different resolution features. Through the optimization of trainable weights, the feature representation capability of the network can be further improved. The improved feature fusion formula is illustrated in Equations (11) and (12):

$$P_i^{td} = Conv\left(\frac{w_1 P_i^{in} + w_2 Resize\left(P_{i+1}^{td}\right)}{w_1 + w_2 + \varepsilon}\right), \tag{11}$$

$$P_i^{out} = Conv\left(\frac{w_1' P_i^{in} + w_2' Resize\left(P_i^{td}\right) + w_3' Resize\left(P_{i-1}^{out}\right)}{w_1 + w_2 + w_3 + \varepsilon}\right), \tag{12}$$

Particularly in MMW images, where small objects may have low resolution and fewer features, weight-based feature fusion can more effectively represent small object features from different resolutions. Furthermore, the cross-scale design can address the issue of feature loss that arises from the deepening of the network, which is highly advantageous for detecting small concealed objects.

4. Experiments

To validate the detection effectiveness of the proposed model for MMW images, we collected the MMW concealed objects dataset, which is shown in Section 4.1, and conducted relevant experiments on this dataset. The experimental results demonstrate that the proposed improvements to the YOLOv5 in this paper lead to enhanced accuracy for MMW concealed object detection, and the algorithm is particularly well-suited for the detection of small concealed objects in MMW images.

4.1. Dataset and Evaluation Metrics

The MMW dataset used in our study was acquired by an experimental MMW radar system, which is illustrated in Figure 8. The experimental radar system generates multi-angle human images by scanning the entire 360-degree view.

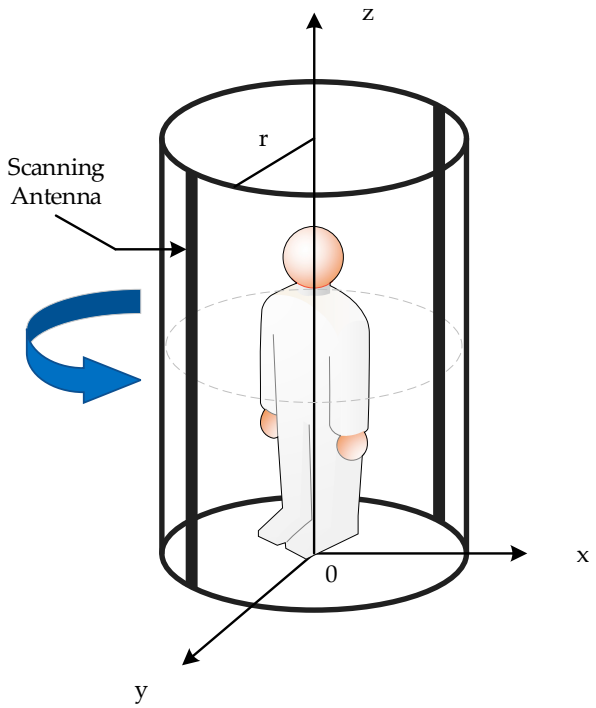


Figure 8. Experimental MMW Radar System Model.

This dataset mainly simulates the conditions for production units to inspect fire sources, electronic devices, and other suspicious items. For data collection, volunteers were requested to conceal and carry pre-arranged items, including cigarettes, lighters, matches, telephones, and other suspicious objects, on their limbs and anterior and posterior torso positions, after which the volunteers entered the MMW radar system for body scanning. The MMW scanner rotated a full circle to perform a body scan of the volunteer and generated multi-angle MMW images of the human body concealed objects, including 7 angles of the front and 7 angles of the back surface of the body. All objects in the images were independently labeled at different angles and all had different patterns, and the multi-angle images enabled the dataset to possess concealed object features that were visible at specific angles. Due to the relatively stable collection environment of this dataset, it was less affected by environmental noise. In terms of dataset size, the MMW dataset included a total of 82,014 MMW radar imaging pictures, with 266,109 manually labeled target labels among the 5 classes. Figure 9 presents the label counts for each class. The dataset is divided into training set, test set, and validation set in an 8:1:1 ratio.

Table 2. The distributions of the instance size for the MMW dataset.

Dataset	0~32 × 32 Pixels	32 × 32~64 × 64 Pixels	>64 × 64 Pixels
MMW Dataset	66%	33%	1%

For image detection evaluation, we chose precision (*P*), recall (*R*), average precision (*AP*), and mean average precision (*mAP*) as evaluation metrics. True positive (*TP*), false positive (*FP*), true negative (*TN*), and false negative (*FN*) were used to define *P* and *R*, as shown in Equations (13) and (14) below. If the Intersection over Union (IoU) threshold of the between the detected bounding box and the ground truth is greater than 0.5, the detection box is marked as *TP*; Otherwise, it is marked as *FP*. If a concealed object is not matched to any detection box, it is marked as *FN*. *TN* is not considered in the binary classification problem.

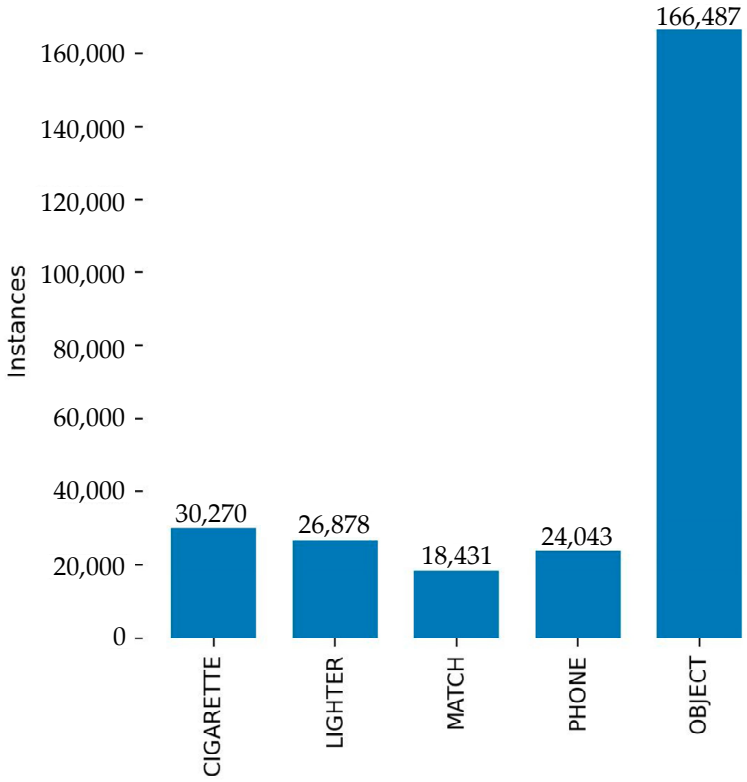


Figure 9. The distribution of the classes of the labels for the MMW dataset.

In addition, we conducted a statistical analysis of the object size of the MMW dataset images, and objects with a pixel size smaller than 32 × 32 were defined as small objects. As shown in Table 2, the number of small objects accounted for 66% of the dataset, while targets with pixel sizes ranging from 32 × 32 to 64 × 64 accounted for 33%. It was evident that the majority of concealed objects in the MMW dataset were small objects, which posed a certain challenge to the training of the YOLOv5 network.

$$P = \frac{TP}{(TP + FP)}, \tag{13}$$

$$R = \frac{TP}{(TP + TN)}, \tag{14}$$

AP is the integral of the precision–recall curve (*P*–*R* curve) in the range of 0 to 1. A higher *AP* value indicates higher detection accuracy, as defined in Equation (15).

$$AP = \int_0^1 P(R)dR, \tag{15}$$

In multi-class object detection tasks, the detection accuracy of a model is evaluated by calculating *mAP* (the average of *APs* of all classes) as shown in Equation (16) below, where *N* represents the number of classes:

$$mAP = \frac{1}{N}\sum AP, \tag{16}$$

*F*1 score is the harmonic mean of *P* and *R*, as expressed in Equation (17):

$$F1 = \frac{2 \times P \times R}{P + R}, \tag{17}$$

4.2. Implementation Details

Our proposed Swin-YOLO is implemented in Python 3.8 and PyTorch 1.8.1 compilation environment. The training, validation, and testing were performed on an NVIDIA Quadro RTX4000 GPU, and the operating system used was Ubuntu 18.04 LTS. For training part, we used the pre-trained model of the YOLOv5 baseline network on the COCO dataset [50]. Due to changes in the original structure, our proposed network used some of the pre-trained models of YOLOv5 and shared weights in the backbone network to improve the model's generalization ability. The pre-trained model also helped to save training time. In terms of model size, we chose YOLOv5l as the baseline. We trained the model on the laboratory's MMW concealed object dataset for 50 iterations, with a warm-up training of 3 iterations to stabilize the gradient change. We set the initial learning rate to 0.001, weight decay coefficient to 0.0005, and used the One Cycle strategy to gradually reduce the learning rate. The batch size was 4. The detector updated parameters and reduced losses using the SGD optimizer, with a momentum of 0.937. In terms of Anchor setting, we selected the anchor boxes from the initial settings of YOLOv5 and calculated through the Autoanchor, obtaining an anchor above threshold (AAT) of 6.88 and a best recall rate (BRP) of 1.000, which is well adapted to the dataset.

4.3. Experimental Results

To validate the effectiveness of the proposed method, we conducted experiments with Swin-YOLO on the MMW dataset proposed in Section 4.1 and compared the results with other YOLO series models such as YOLOv8 [26], TPH-YOLOv5 [48], and STPH-YOLOv5. STPH-YOLOv5 combined the Swin Transformer with the C3 module based on the idea of C3TR in TPH-YOLOv5. As shown in Table 3, our proposed method achieved a *mAP* of 92% on the MMW dataset, which is 4.7% higher than the YOLOv5 baseline method, 1.3% higher than that of the multi-head Transformer network, and 1.5% higher than that of the C3STR network improved by the Swin Transformer. Moreover, the *mAP* of Swin-YOLO is 3.6% higher than most advanced single-stage detector YOLOv8 and 2.9% higher than two-stage detector Faster-RCNN. These results demonstrate the effectiveness of Swin-YOLO in detecting small concealed objects with complex features in MMW images.

Table 3. Test results on the MMW dataset for different detection models.

Method	*P*	*R*	*mAP*
YOLOv5	88.7	82	87.3
YOLOv8	91.5	82.3	88.4
Faster-RCNN	90.4	83.8	89.1
TPH-YOLOv5	90.6	86.1	90.7
STPH-YOLOv5	90.3	85.8	90.5
Swin-YOLO	91.8	87.3	92.0

Figure 10 shows the *P*–*R* curves for each class of Swin-YOLO, which can reflect the relationship between precision and recall. For each class, the *AP* value is represented in the figure as the area of the *P*–*R* curve, and the results are shown in the legend of the figure. A higher *AP* value indicates the better detection performance of the model. From Figure 10, it can be seen that the improved model achieved good detection results for each class of objects. The class Cigarettes performed best with an accuracy of 96.1%, Lighters with an accuracy of 91.9%, Matches with an accuracy of 88.8%, Telephones with an accuracy of 95.1%, and Objects with an accuracy of 88%.

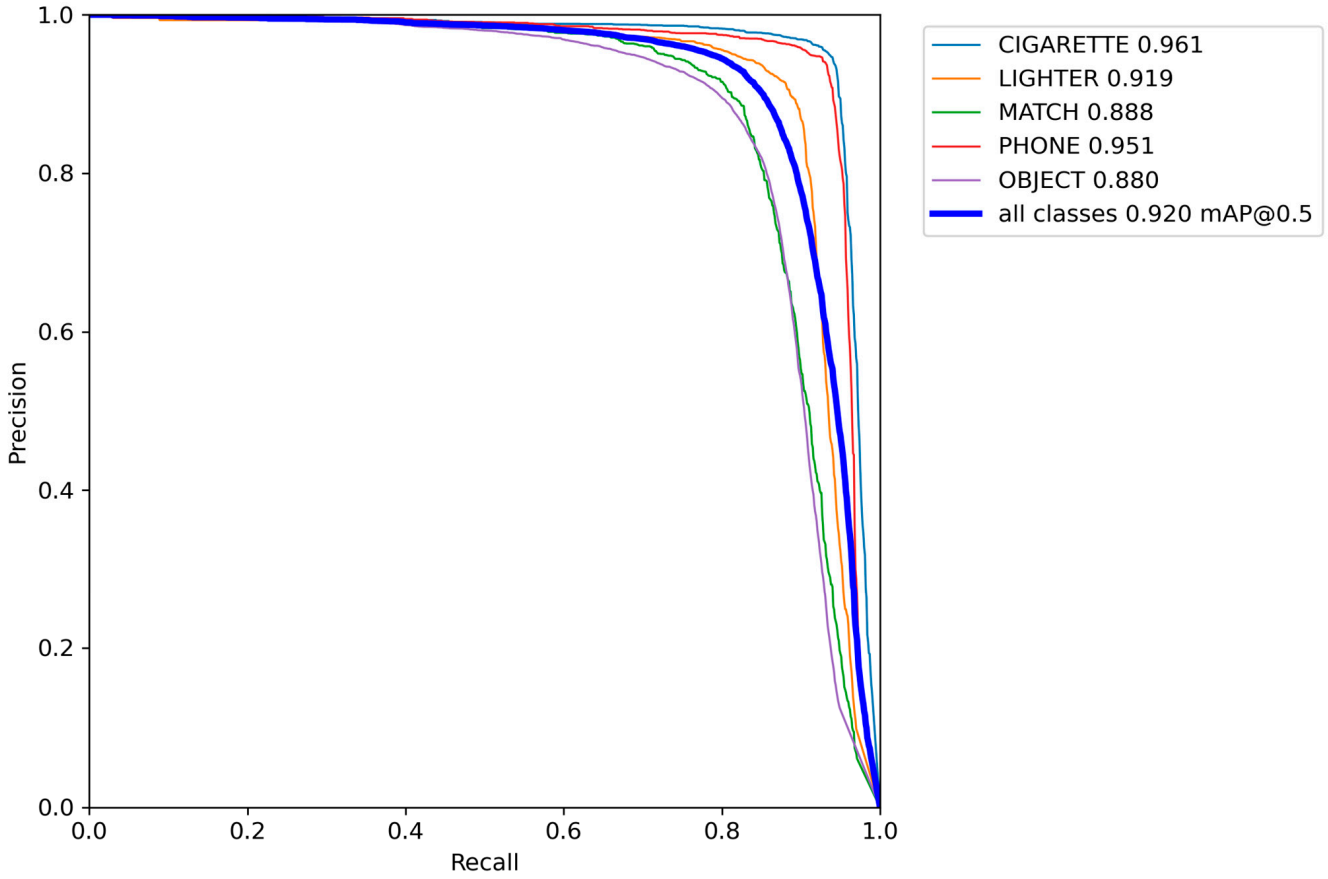


Figure 10. P–R curve results of Swin-YOLO on MMW dataset.

The intersection-over-union (IoU) threshold and confidence threshold are two fundamental measures of deep learning models. The confusion matrix of Swin-YOLO on the MMW dataset was calculated, as shown in Figure 11. We used an IoU threshold of 0.5 and a confidence threshold of 0.25, and visualized the classification of each class. In the confusion matrix plot, each row represents the predicted class, each column represents the actual class, and the data on the diagonal represents the proportion of correctly classified classes. As shown in Figure 11, most objects in the five classes were correctly predicted, indicating that the model has good performance. However, there is a high probability that the Object and Match classes are identified as background. For the class Object, the large number of objects and numerous object features can easily lead to misjudgments by the model. As for the class Match, the limited feature extraction due to the relatively small number of training samples leads to a higher false negative rate and a lower *AP* value compared with other classes. For the class Lighter, some targets were also mistakenly identified as background due to their extremely small size.

Figure 12 shows the variation curves of loss values, including localization loss (Figure 12a), confidence loss (Figure 12b), and classification loss (Figure 12c). As shown from the figure, with the increase in the number of iterations, the three types of losses decrease steadily. The model converges after 50 iterations.

The *F*1 score curve shows the classification performance of the proposed model. As shown in Figure 13, Swin-YOLO has an effective classification effect on all classes, especially for class cigarette and class phone, followed by class lighter, class object and class match.

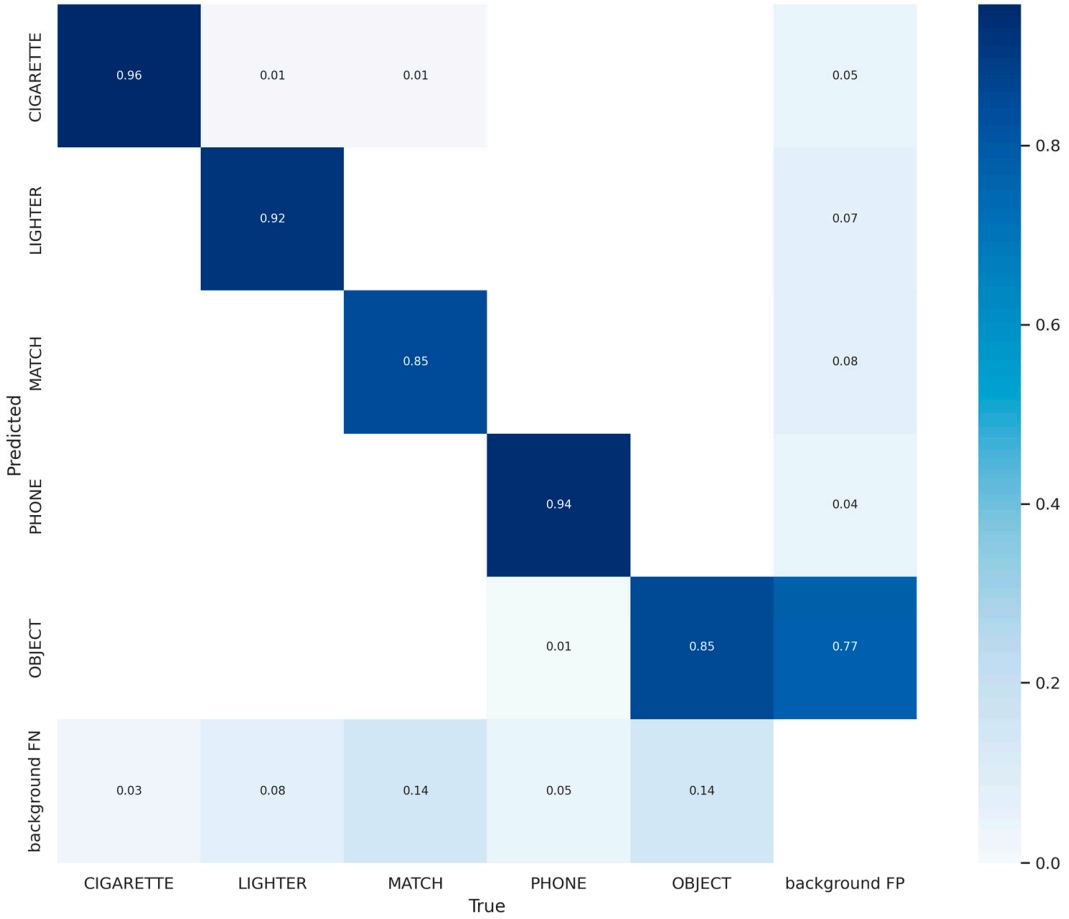
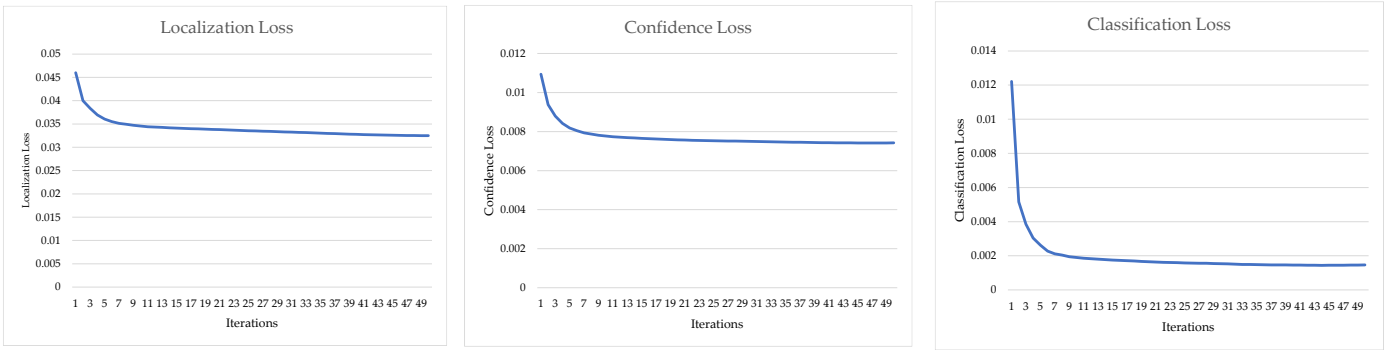


Figure 11. The confusion matrix from the Swin-YOLO results on the MMW dataset.



(**a**) (**b**) (**c**)

Figure 12. The Loss variation curves from the Swin-YOLO results: (**a**) localization loss; (**b**) classification loss; (**c**) confidence loss.

To evaluate the computational efficiency of the proposed method, we compared the inference time model size of the proposed detection model with baseline and the Transformer-based TPH-YOLOv5, as shown in Table 4. Swin-YOLO demonstrated a speed advantage over TPH-YOLOv5, which verifies the advantage of the Swin Transformer in saving more computational resources than the Transformer. However, despite the improved detection performance, compared with the baseline, Swin-YOLO is slower than YOLOv5, indicating that simplifying the model to improve inference speed is a future research direction. In addition, the model size of Swin-YOLO has increased, compared with the original YOLOv5, due to these improvement measures, but is smaller than TPH-YOLOv5.

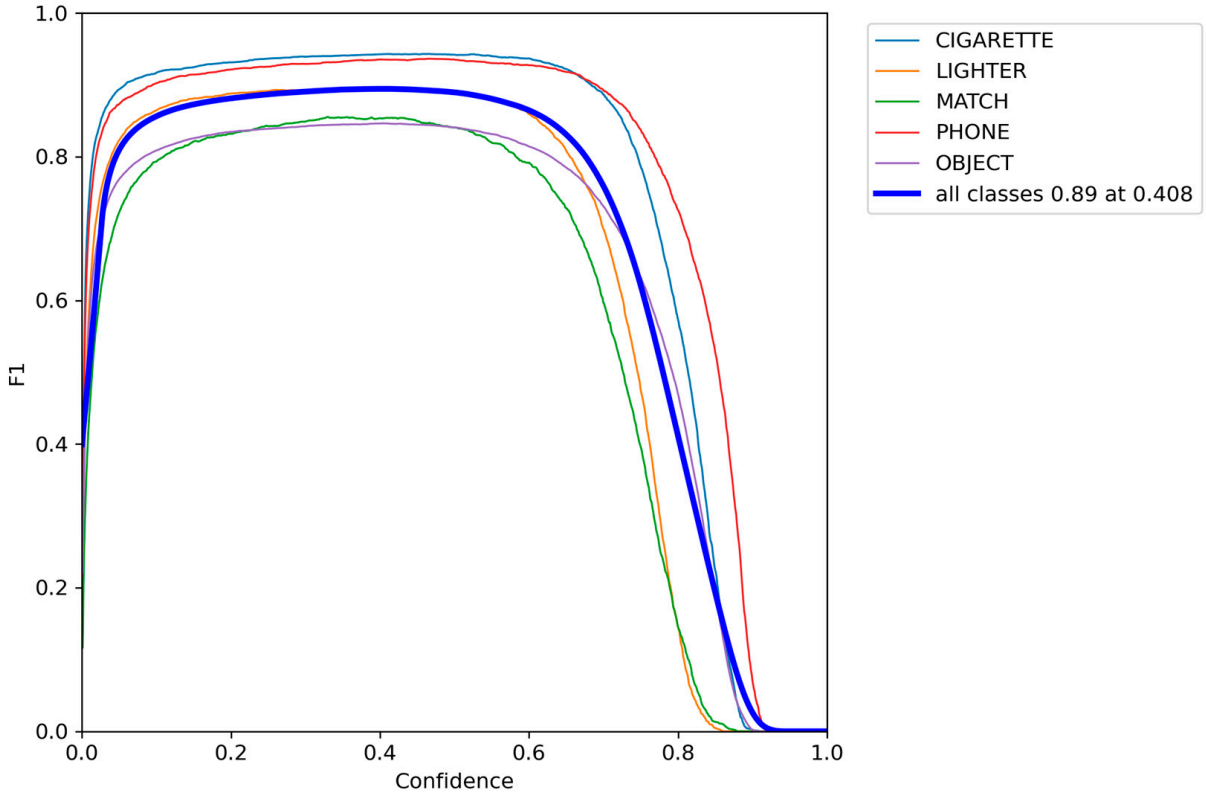


Figure 13. *F*1 curve of Swin-YOLO on MMW dataset.

Table 4. Comparison of inference time of different models on the MMW dataset.

Method	Model Size (MB)	Inference Time (Per Picture)
YOLOv5	92.8	10.6 ms
TPH-YOLOv5	121.8	20.6 ms
Swin-YOLO	110.5	16.7 ms

In addition to inference time and model size, we also calculated the computational complexity of the proposed model, represented by giga floating-point operations per second (GFLOPs), which can be used to measure the computational requirements of the network. Swin-YOLO requires 231.5 Gflops. To better address the limitations of computing resources, a lightweight version of the model will be one of our focuses in the future.

We present the detection results of the proposed model on a laboratory MMW dataset in Figure 14. Swin-YOLO can detect small objects with complex features and is more suitable for the detection of small objects, especially for the small class Cigarette. The addition of the Swin Transformer enables the model to capture rich global contextual information and enhance feature representation through self-attention mechanisms, which demonstrates the assisting value of global contextual information in object detection. In addition, the attention mechanism can effectively focus on key object regions for objects with indistinct features and biased angles. Moreover, the improvement of the layer structure, connections, and feature fusion methods of the neck network can enrich feature information and efficiently fuse features. Therefore, Swin-YOLO can better utilize the complex low-level and high-level features, which facilitates the detection of small objects.

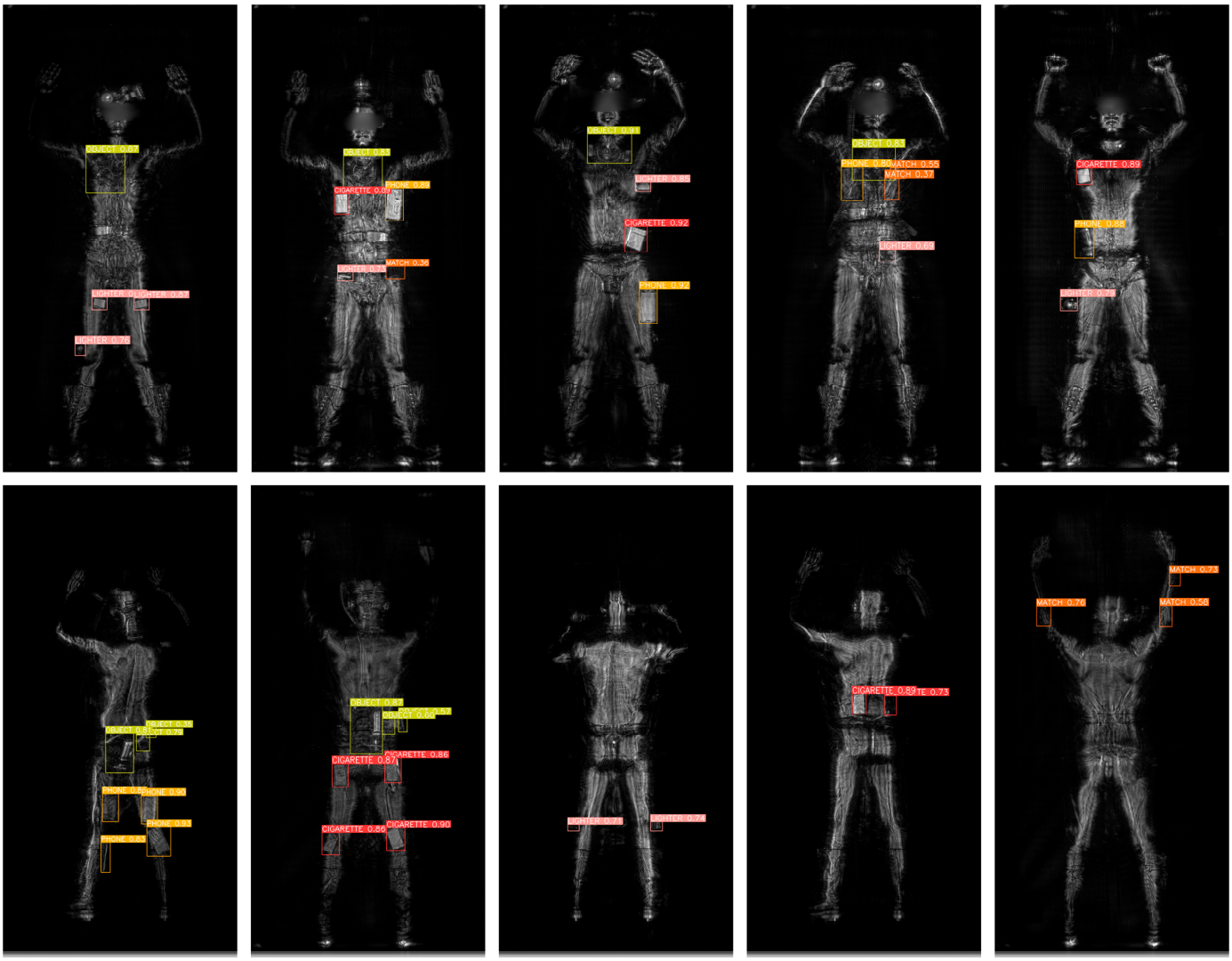


Figure 14. Some detection results of Swin-YOLO on MMW dataset.

However, from the detection results, although most items have been well detected, there are slight differences in the performance of the model for items displayed in different directions of the same class, especially in asymmetric images. When carrying items on the side of the human body, the hidden object features are easily confused with the background, posing a challenge for model detection. Therefore, in future work, the detection of side-carried objects will be a key research topic. Moreover, data detection in more complex detection environments, such as areas affected by noise and densely populated areas, is also one of the key issues to consider in the future.

4.4. Ablation Experiments

In this section, we conducted ablation experiments on the improvements proposed in this paper to better understand the performance of the neural network. The experimental results are shown in Table 5. The detection *mAP* of YOLOv5 is 87.3%. After the improvement of using the added small object detection head (termed as P2), the model's *mAP* was increased by 0.7% to reach 88%, which demonstrates the effectiveness of the small object detection branch. Subsequently, we evaluated the Local Perception Swin Transformer layer and Coordinate Attention. The addition of the LPST Layer has brought a further 3.5% improvement, indicating that the LPST improves the detection performance of small objects. It provides the network with an enhancement in self-attention and global contextual information, which is critical. Due to the large number of small objects in MMW images, such

improvements are necessary. Building upon the improvement of the small object detection head, the addition of Coordinate Attention resulted in a 2.4% increase in performance, demonstrating that coordinate-based attention can optimize the feature information of small objects and improve the performance of the small object detection head. The addition of BiFPN improved the network's *mAP* by 0.2% compared with the modification of P2, CA, and LPST, proving the effectiveness of cross-scale fusion and normalized feature map weights. In comparison with the Swin Transformer, the *mAP* of LPST was 0.6% higher than the Swin Transformer, which proved the effectiveness of LPM in providing the Swin Transformer with local perception.

Table 5. Ablation results on the MMW dataset.

Method	*P*	*R*	*mAP*
YOLOv5	88.7	82	87.3
YOLOv5 + P2	91.3	82.5	88
YOLOv5 + P2 + CA	89.5	84.5	89.7
YOLOv5 + P2 + LPST	90.9	85.9	90.8
YOLOv5 + P2 + CA + Swin Transformer	91.7	85.6	91.2
YOLOv5 + P2 + CA + LPST	91.4	87.1	91.8
Swin-YOLO	91.8	87.3	92

When all these improvements were combined, the object detection network achieved a 4.7% improvement over the YOLOv5 baseline. Therefore, Swin-YOLO achieves competitive detection performance, demonstrating its potential for MMW object detection.

4.5. Discussion

Despite the performance improvement of the proposed Swin YOLO, there is still some room for improvement in the future. Firstly, considering the stability in complex environments and the particularity of only allowing one person to pass through the security check environment at a time, if there is large-scale congestion or high-speed passing through the security check, the imaging quality will be affected to some extent, which will bring about changes in detection performance. Therefore, in future work, improving model robustness is a key step; secondly, there are limitations of computing resources. The security environment has strict requirements for computing resources and modern security inspections have stricter requirements for computing resource conditions. The different computational resource requirements of detectors may affect their actual application scenarios. As a consequence, future lightweight models are also a key direction of work.

5. Conclusions

This paper presented Swin-YOLO, a novel MMW object detection algorithm that combines Local Perception Swin Transformer and YOLOv5 to address the characteristics of small target size and complex features in MMW concealed object images. Firstly, to capture shallow features, extra feature fusion layers were added to obtain more low-level feature information. An additional prediction head was also added for the detection of tiny objects. Secondly, Local Perception Swin Transformer Layers were introduced to the neck network to enhance the global contextual information acquisition ability and increase the contextual connections of feature maps, effectively improving the detection performance of MMW small concealed objects. Moreover, Coordinate Attention modules were added to each detection head, and the coordinate-based attention made the network focus on the key areas of the MMW images and improved the localization accuracy of small object detection. Finally, a bidirectional feature pyramid network was employed to improve the feature fusion network, better integrating the extracted shallow and deep semantic features. To validate the detection performance of the proposed detector, an MMW concealed object image dataset collected by experimental radar was created. The experimental results show that the *mAP* value of the proposed method reached 92% on the above dataset, which

outperformed the YOLOv5 baseline network. Moreover, comparison experiments with other Transformer multi-head networks showed that the proposed detector achieves a balance between real-time performance and detection accuracy and is more suitable for small object detection in MMW human concealed object images. Still, the data in this article was not collected under complex and dynamic conditions. Therefore, we are not confident that our proposed model is valid on this data. In future work, improving the detection performance of the model in complex environments is a key task. However, in order to address the limitations of computing resources in security inspection scenarios, it can be considered to design a more lightweight model to improve detection efficiency.

Author Contributions: Conceptualization, P.H., R.W. and Y.S.; methodology, R.W.; software, P.H. and R.W.; validation, R.W., Y.S. and W.T.; formal analysis, Y.S.; investigation, R.W.; resources, P.H.; data curation, P.H.; writing—original draft preparation, P.H., R.W. and Y.S.; writing—review and editing, R.W., Y.S. and W.T.; visualization, R.W.; supervision, Y.S. and P.H.; project administration, W.T.; funding acquisition, P.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No. 61971246), Inner Mongolia Autonomous Region 2022 Science and Technology Leading Talent Team Project Task Letter (Grant No. 2022LJRC0002), Project Plan and Task Letter for Basic Research Business Expenses of Universities Directly under the Autonomous Region (Grant No. JY20220077).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Jianxin Zhang and Diankun Zhang from OBE Terahertz Science and Technology (Beijing) Co., Ltd. (Beijing, China) for the kind support with data collecting and labeling. The authors appreciate that the anonymous reviewers put forward their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gao, J.; Qin, Y.; Deng, B.; Wang, H.; Li, X. A novel method for 3-D millimeter-wave holographic reconstruction based on frequency interferometry techniques. *IEEE Trans. Microw. Theory Tech.* **2017**, *66*, 1579–1596. [CrossRef]
2. Haraz, O.M.; Ashraf, M.A.; Sebak, A.R.; Alshebeili, S.A. Detection of metallic and nonmetallic concealed targets based on millimeter-wave inverse scattering approach. *Int. J. RF Microw. Comput.-Aided Eng.* **2020**, *30*, e22290. [CrossRef]
3. Yeom, S.; Lee, D.S.; Son, J.Y.; Kim, S.H. Concealed object detection using passive millimeter wave imaging. In Proceedings of the IEEE 2010 4th International Universal Communication Symposium, Beijing, China, 18–19 October 2010; pp. 383–386.
4. Lee, D.S.; Yeom, S.; Son, J.Y.; Kim, S.H. Automatic image segmentation for concealed object detection using the expectation-maximization algorithm. *Opt. Express* **2010**, *18*, 10659–10667. [CrossRef] [PubMed]
5. Pang, L.; Liu, H.; Chen, Y.; Miao, J. Real-time concealed object detection from passive millimeter wave images based on the YOLOv3 algorithm. *Sensors* **2020**, *20*, 1678. [CrossRef]
6. Zhang, Y.; Wu, C.; Liu, X.; Wang, L.; Dai, C.; Cui, J.; Li, Y.; Kinar, N. The Development of Frequency Multipliers for Terahertz Remote Sensing System. *Remote Sens.* **2022**, *14*, 2486. [CrossRef]
7. Yeom, S.; Lee, D.S. Multi-level segmentation for concealed object detection with multi-channel passive millimeter wave imaging. In Proceedings of the 2013 International Conference on IT Convergence and Security (ICITCS), Macau, China, 16–18 December 2013; pp. 1–4.
8. Yang, X.; Wei, Z.; Wang, N.; Song, B.; Gao, X. A novel deformable body partition model for MMW suspicious object detection and dynamic tracking. *Signal Process.* **2020**, *174*, 107627. [CrossRef]
9. Meng, Z.; Zhang, M.; Wang, H. CNN with pose segmentation for suspicious object detection in MMW security images. *Sensors* **2020**, *20*, 4974. [CrossRef] [PubMed]
10. He, W.; Zhang, B.; Wang, B.; Sun, X.; Yang, M.; Wu, X. Concealed Object Detection in Millimeter Wave Image Based on Global Correlation of Multi-level Features in Cross-section Sequence. *J. Infrared Millim. Waves* **2021**, *40*, 738–748.
11. Zhang, K.S. Detection of Contraband Based on Millimeter Wave Image. Master's Thesis, Hangzhou Dianzi University, Hangzhou, China, May 2022.
12. Ultralytics. Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 9 June 2020).

13. 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.
14. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Berlin, Germany, 11–14 March 2015; pp. 1440–1448.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
17. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016, Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2022*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
18. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
19. Redmon, J.; Farhadi, A. Yolo9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–27 July 2017; pp. 6517–6525.
20. Joseph, R.; Ali, F. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
21. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
22. 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.
23. 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.
24. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
25. 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. *arXiv* **2022**, arXiv:2207.02696.
26. Ultralytics. Yolov8. Available online: https://github.com/ultralytics/ultralytics (accessed on 10 January 2023).
27. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2204–2212.
28. 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.
29. 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.
30. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
33. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2021.
34. 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]
35. Yeom, S.; Lee, D.S.; Chang, Y.S.; Lee, M.K.; Jung, S.W. Concealed object recognition based on geometric feature descriptors. *Passiv. Act. Millim.-Wave Imaging XV* **2012**, *8362*, 135–140.
36. Lee, D.S.; Yeom, S.; Chang, Y.S.; Lee, M.K.; Sang-Won Jung, S.W. Real-time computational processing and implementation for concealed object detection. *Opt. Eng.* **2012**, *51*, 071405. [CrossRef]
37. 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]
38. Li, X.; Yang, K.; Fan, X.; Hu, L.; Li, J. Fast and accurate concealed dangerous object detection. *J. Electron. Imaging* **2022**, *31*, 023021.
39. Zhang, B.; Wang, B.; Wu, X.; Zhang, L.; Yang, M.; Sun, X. Domain adaptive detection system for concealed objects using millimeter wave images. *Neural Comput. Appl.* **2021**, *33*, 11573–11588. [CrossRef]
40. Yuan, M.; Zhang, Q.; Li, Y.; Yan, Y.; Zhu, Y. A Suspicious Multi-Object Detection and Recognition Method for Millimeter Wave SAR Security Inspection Images Based on Multi-Path Extraction Network. *Remote Sens.* **2021**, *13*, 4978.
41. 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]
42. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1571–1580.

43. Sun, Z.; Li, P.; Meng, Q.; Sun, Y.; Bi, Y. An Improved YOLOv5 Method to Detect Tailings Ponds from High-Resolution Remote Sensing Images. *Remote Sensing* **2023**, *15*, 1796. [CrossRef]
44. Bao, W.; Du, X.; Wang, N.; Yuan, M.; Yang, X. A Defect Detection Method Based on BC-YOLO for Transmission Line Components in UAV Remote Sensing Images. *Remote Sensing* **2022**, *14*, 5176. [CrossRef]
45. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-iouloss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12993–13000.
46. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
47. Gong, H.; Mu, T.; Li, Q.; Dai, H.; Li, C.; He, Z.; Wang, W.; Han, F.; Tuniyazi, A.; Li, H.; et al. Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sens.* **2022**, *14*, 2861. [CrossRef]
48. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone- captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 19–25 June 2021; pp. 2778–2788.
49. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.

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.