Hidden Dangerous Object Recognition in Terahertz Images Using Deep Learning Methods

Samuel Akwasi Danso 1,2,*; Liping Shang 1; Deng Hu 1; Justice Odoom 1; Quancheng Liu 1; and Benedicta Nana Esi Nyarko 1

1 School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China; shangliping@swust.edu.cn (L.S.); denghu@swust.edu.cn (D.H.); justnetonline@outlook.com (J.O.); liuqc@swust.edu.cn (Q.L.); benedictanyarko41@gmail.com (B.N.E.N.)
2 Faculty of Engineering, Ghana Communication Technology University, Accra PMB 100, Ghana
* Correspondence: sdanso@gctu.edu.gh

Abstract: As a harmless detection method, terahertz has become a new trend in security detection. However, there are inherent problems such as the low quality of the images collected by terahertz equipment and the insufficient detection accuracy of dangerous goods. This work advances BiFPN at the neck of YOLOv5 of the deep learning model as a mechanism to improve low resolution. We also perform transfer learning, thereby fine-tuning the pre-training weight of the backbone for migration learning in our model. Results from experimental analysis reveal that mAP@0.5 and mAP@0.5:0.95 values witness a percentage increase of 0.2% and 1.7%, respectively, attesting to the superiority of the proposed model to YOLOv5, which is the state-of-the-art model in object detection.

Keywords: terahertz image; object detection; hidden object; airport scanned object

1. Introduction

The frequencies from 0.1 to 30 THz form the Terahertz domain of the electromagnetic (EM) band region and act as a gap of convergence between the microwave and the infrared band in the EM spectrum, as shown in Figure 1.

Figure 1. EM spectrum band.

Terahertz is a non-ionizing frequency which has high penetration in non-dielectric materials. Several non-crystalline materials are transparent for THz rays, for example, clothes, plastic, material, paper, etc. Its impact on organic tissues implies that it cannot damage DNA, making it safe for human applications such as medical imaging. Terahertz rays suffer from the reflection and absorption of metal surfaces and polar liquids such as water, respectively [1]. Recently, terahertz technology and deep learning have been a focus of research. The applied areas in the domain of object detection include the detection of agricultural products [2–5], breast cancer and other medical conditions [6–8], and hidden object detection [9–12]. Nevertheless, terahertz suffers from low image resolution...
due to blurriness and noisy dark-spotted images, which stem from low-energy power sources [13–15] consequently affecting the detection accuracy and rate. Therefore, it stands to reason that any attempt to increase the detection rate and accuracy must first address the challenge of low resolution while at the same time revamping the deep learning model.

2. Related Work

In this section, we review previous work focusing on terahertz image processing, especially on terahertz image recognition.

2.1. Terahertz Image Acquisition & Image Processing

Due to a capture rate of up to 5000 lines per second, the term fast-256 device is capable of a scan speed up to 15 m/s. A single sensitivity band at 100 ± 10 GHz is characteristic of the sensor, but the experimental power source is around 100 GHz. Meanwhile, image capturing is made possible by the conveyor belt, which speeds at 10.1 m/s. In Figure 2, we show the active THz image device used in this work, which uses a coherent source. The THz detectors and source utilize transmission or reflection geometries.

![Figure 2. THz-scanned imaging system for security screening application.](image)

Dataset Description

This section presents the acquisition steps of the terahertz image and the expansion methods for the image dataset. Primarily, these encompass affine transformation, rotation, transmission transformation, and translation, among others. Subsequently, we perform statistical analysis of the expanded dataset. The size of the image data collected by the device is $512 \times 256$ px. In all, a total of eight different kinds of terahertz images of objects were collected. They included four types of weapon images as well as four types of non-weapon images (329 instances as a whole, since there might be more than one instance of a single image). The raw data information is shown in Table 1 and Figure 3.

![Figure 3. Terahertz acquisition image system and samples of hidden images.](image)
Table 1. Original terahertz image data.

| Class        | Screwdriver | Blade | Knife | Scissors | Boardmarker | Mobile Phone | Wireless Mouse | Water Bottle |
|--------------|-------------|-------|-------|----------|-------------|---------------|----------------|--------------|
| No.          | 65          | 21    | 66    | 59       | 40          | 40            | 40             | 40           |
| Avg. bounding box   | 108 px × 84 px | 36 px × 35 px | 89 px × 75 px | 104 px × 91 px | 78 px × 68 px | 110 px × 87 px | 70 px × 75 px | 118 px × 91 px |

In order to gain insights and better understand the characteristics of the data, analysis from the statistical perspective is pivotal, consequently aiding in model optimization. In the first place, statistics on the number of instances and the average bounding box size of eight (8) categories are shown in Table 1. As can be seen from Table 1, with weapons as our target discussion, the cardinarity of blade categories is the fewest with the average bounding box being the smallest. The largest in number is the knife followed by Screwdriver with a relatively large bounding box. From the results of Figure 3, it can be seen that about 3% to 10% constitutes a major portion of the box area ratio. Meanwhile, a little section is concentrated in the 1% area ratio with the maximum proportion no more than 25%. Furthermore, in Figure 4, size distribution analysis on different types of bounding boxes is shown. Anchor as used in Figure 4 denotes a set of predefined bounding boxes of specific height and width and essentially captures the scale and aspect ratio of specific object classes of interest to us. In addition, the THz image pixels are represented by 0–255 values as RGB (Red Green Blue). Note that in RGB, a color is defined as a mixture of pure red, green, and blue lights of varying strengths with red, green and blue light levels encoded as a value in the range 0–255 where 0 and 255 denote zero light and maximum light, respectively.

2.2. Terahertz Image Detection

Image object detection is a fundamental task in image processing. The task needs to judge not only what the object is but also the position of the object. In recent years, with the outbreak of deep learning, target detection technology in the field of machine vision has developed rapidly. Traditional manual feature design methods, such as HOG, SIFT and LBP [16–19] can only achieve good results in specific scenes and cannot adapt to complex and diverse large-scale image data. The abstract features that can adequately describe objects are often ignored when designing features manually [20], and the designed network usually needs to be trained separately to conduct multi-level positioning. In recent years, the boom of deep learning based on a convolutional neural network (CNN) provides another idea for object recognition [21]. The target detection algorithm based on a convolution neural
network automatically extracts image features through the convolution layer, which has greatly improved the efficiency and accuracy. The earliest CNN-based target detection algorithm is the RCNN algorithm proposed by Ross Girshick in 2015, which is followed by classic two-stage detection networks such as Fast RCNN and Mask RCNN [22–26]. Because the two-stage detection network consumes additional computing resources in the region proposal stage and the amount of model parameters is large, the overall detection speed of the two-stage detection network is slow. Researchers also proposed the one-stage detection network, which is represented by YOLO series [16,27–30]. The one-stage detection network directly uses the output of the convolution feature map for classification prediction and position fitting to reduce the amount of calculation. In addition, there are some anchor-free algorithms such as FCOS, CenterNet, etc.

Although the above algorithms have achieved good index results in public datasets such as PASCAL VOC [31,32] and COCO [33] datasets, there are still some problems in the application of terahertz image object detection. These problems are mainly caused by the characteristics of the terahertz dataset, namely: (1) image blur and (2) uneven distribution of the image histogram (as shown in Figure 4). The characteristics of these datasets will cause certain detection errors in the existing detection framework model for terahertz. Based on the YOLOv5s model, this paper redesigns the head structure of the detection model for terahertz datasets, adopts the BiFPN structure [34,35] and realizes skip connection in convolution feature fusion, which can fuse richer image features than the original PANet [36,37].

The main contribution of this work is summarized as follows:
1. Improving low resolution using BiFPN at the neck of YOLOv5 of the deep learning model.
2. Transfer learning is done using the fine-tuning process to the pre-training weight of the backbone for migration learning in our model.

The remainder of this work is as follows. In Section 3, we advance our model encompassing the backbone and neck, whereas in Section 4, we present experimental work involving image processing, model comparison and model transfer learning. We conclude the paper in Section 5.

3. Proposed Model

3.1. Model Backbone

A key design section of a detection model is the backbone, which determines the quality of image feature extraction. It also affects the subsequent object classification, recognition and object location. ResNet series is a widely used backbone network. It uses the residual structure to solve the problem of gradient disappearance or gradient explosion in the training process of a deep convolution network. The classical fast RCNN model and RetinaNet model use the ResNet network as the backbone to extract rich image features. The detection model in this paper uses a cross-stage partial structure with less computation [38]. This structure optimizes the characteristic diagrams of different stages of different ResNet networks, as shown in Figure 5. The input feature split its channel into two signal flows, which finally concatenate together. This way, the variability of the gradients through the network is considered.

It is noteworthy in Figure 5 that the computational complexity of the ResNet network is O(CHW2) with the complexity of the cross-stage partial basically computed by the product of the value and key branch. It is imperative to note that the dimensions of the input feature maps are $C \times H \times WC \times H \times W$. Note that $\otimes$ denotes matrix multiplication, whereas $\oplus$ is the broadcast element-wise addition. In the ResNet network, the multiplication operation plays a pivotal role in capturing the attentional feature map. It is possible to have an operation that satisfies both the acquisition of an attentional feature map and the cross-channel communication of information given that the multiplication operation is very similar to the multiplication operation in positional attention. The calculation amount and memory occupation of channel attention are significantly reduced compared to positional
attention with respect to the attention mechanism. As a consequence, the channel attention mechanisms are utilized instead of positional attention mechanisms. This way, the memory occupation and time complexity can be greatly decreased with performance not sacrificed in any way.

![Figure 5. Cross-stage partial connection bottleneck.](image)

### 3.2. Model Neck

The neck part of the whole detection network served in the role of convolution feature fusion. In the original YOLOv5s implementation, PANet is used as the neck, which adds a bottom–up pathway on top of the FPN. For the singularity of the terahertz image dataset, it is hard to obtain and fuse the significant features. As an extension of PANet, a bi-directional feature pyramid network (BiFPN) is adopted as our model’s neck, as shown in Figure 6. It takes level 3–5 input features $P_i = P_3, P_4, P_5$, where $P_i$ represents a feature level with a resolution of $1/2^i$ of the input image. For instance, our input terahertz image is transformed into 640 px by 640 px; $P_3$ then represents level 3 with resolution 80 px by 80 px $(640/2^3 = 80)$. A skip connection is applied after input $P_4$ and $P_5$ to enhance the feature representation. The different features in BiFPN are concatenated with the same size after upsampling or downsampling. The output features of BiFPN can be calculated as:

\[
\begin{align*}
O_3 &= \text{Conv}(P_3 \oplus \text{Down}(\text{Conv}(P_4))) \\
O_4 &= \text{Conv}(\text{Up}(O_3) \oplus P_4 \oplus \text{Conv}(P_4) \oplus \text{Down}(\text{Conv}(P_5 \oplus \text{Conv}(P_4)))) \\
O_5 &= \text{Conv}(\text{Up}(O_4) \oplus P_5 \oplus \text{Conv}(P_5) \oplus \text{Up}(\text{Conv}(P_4)))
\end{align*}
\]  

(1)

### 3.3. Classification and Regression Loss

The loss of this model consists of classification loss and regression loss. The classification loss adopts binary cross-entropy loss, which is defined as:

\[
\text{loss}_{clf}(p, y) = \frac{1}{N} \sum_i [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]
\]  

(2)
The return loss takes into account the GIoU loss of the bounding box:

\[ \text{loss}_{\text{reg}} = 1 - (\text{IoU} - \frac{|C \setminus (A \cup B)|}{|C|}) \]  

(3)

where IoU is expressed as:

\[ \text{IoU} = \frac{|A \cap B|}{|A \cup B|} \]  

(4)

The calculation of IoU and C under bounding box A and B can be seen in Figure 7 where C denotes the smallest enclosing convex object. IoU fulfills all properties of a metric such as non-negativity [39].

3.4. Models

In this subsection, we elucidate on the various models used in this paper (for the purpose of comparison).

Figure 6. Scheme diagram of the proposed model.

Figure 7. Diagram of IoU and C.

Note, however, that IoU loss only works when the bounding boxes have overlap, and it would not provide any moving gradient for non-overlapping instances. In other words, IoU does not accurately reflect if two shapes are in close proximity to each other or very far from each other. To address such shortcomings, we adopt the GIoU.
3.5. YOLOv5 and Variants

The framework architecture of YOLOv5 is composed of three main parts: backbone, neck, and predict head. Primarily, the backbone is a convolutional neural network that aggregates and forms image features at different granularities. It extracts feature information from input pictures. On the other hand, the neck is a series of layers to mix and combine image features to pass them forward to prediction. Typically, the neck combines the gathered feature information and creates three different scales of feature maps. The prediction head consumes features from the neck and takes box and class prediction steps. This is completed by detecting objects based on the created feature maps. In fact, the YOLO model was the first object detector to connect the procedure of predicting bounding boxes with class labels in an end-to-end differentiable network.

It is worth mentioning that YOLOv5 utilizes the CSPDarknet53 framework with an SPP layer as the backbone, the PANet as the neck, and the YOLO detection head. The best anchor frame value is calculated in YOLOv5 by adapting the clustering algorithm in different training datasets. The several activation functions tried by YOLOv5 include sigmoid, leakyReLU, and SiLU.

The five derived models for YOLOv5 include YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Although they share the same model architecture, inherent in each model is different model widths and depths. Noteworthy is the fact that smaller models are faster; hence, they are usually designed for mobile deployment, whereas larger models, although more computationally intensive, have a better performance.

In other variants of YOLOv5, CSP-Darknet is used (for instance, in YOLOv5-P7). Such an architecture often has seven stages (or block groups) commonly referred to as \([P_1, P_2, P_3, P_4, P_5, P_6, P_7]\) with strides \([2, 4, 8, 16, 32, 64, 128]\) relative to the input image, respectively. Stacks \([P_1, P_2, P_3, P_4, P_5, P_6, P_7]\) by design consist of multiple CSPDark blocks with Cross Stage Partial (CSP) connections (example: CSP-Darknet in YOLOv5-P7 has \([1, 3, 15, 15, 7, 7, 7]\) CSPDark blocks).

In the case of YOLOv5-P6, an additional large object output layer P6, following the EfficientDet example of increasing output layers for larger models is added. However, unique to this case is the fact that it is applied to all models. Note that current models have P3 (stride 8, small) to P5 (stride 32, large) outputs, but the P6 output layer is stride 64 and designed for extra-large objects. Notice that the architecture changes made to add a P6 layer here are key. The backbone is extended down to P6, and the PANet head goes down to P3 (consistent with state-of-the-art) and back up to P6 now instead of stopping at P5. However, new anchors are also added, which are evolved at img 1280. For brevity, we show in Figure 8 [40] the generalized architecture for Yolov5.

![Figure 8. Architecture of Yolov5.](image-url)
3.6. YOLOv5 Ghost

In this model, the focus is on a reduction in redundancy in intermediate feature maps calculated by mainstream CNNs. Toward this end, the model reduces the required resources (convolution filters used for generating them). In practice, we are given the input data \( x \in \mathbb{R}^{c \times h \times w} \) where \( c \) is the number of input channels whereas \( h \) and \( w \) denote the height and width of the input data, respectively. It must be noted that the operation of an arbitrary convolutional layer for producing \( n \) feature maps can be formulated as:

\[
Y = X \ast f + b
\]

where \( \ast \) is the convolution operation, \( b \) is the bias term, and \( Y \in \mathbb{R}^{h' \times w' \times n} \) is the output feature map with \( n \) channels. Meanwhile, \( f \in \mathbb{R}^{c \times k \times k \times n} \). In addition, \( h' \) and \( w' \) represent the height and width of the output data, whereas \( k \times k \) is the kernel size of convolution filters \( f \). As part of the convolution process, the required number of FLOPs is calculated as \( n.h'.w'.c.k.k \). This is often as large as hundreds of thousands, since the number of filters \( n \) and the channel number \( c \) are generally very large (for instance, 256 or 512).

From Equation (5), optimization of the number of parameters (in \( f \) and \( b \)) is explicitly determined by the dimensions of input and output feature maps. The architecture of the model is shown in Figure 9 [41].

![Figure 9. Architecture of Ghost.](image)

3.7. YOLOv5-Transformer

As shown in Figure 10 [42], in YOLOv5 Transformers (TRANS), a merger of MixUp, Mosaic and traditional methods in data augmentation are employed. Integrated into the YOLOv5 is the Transformer Prediction Heads (TPH). Essentially, these accurately localize objects in high-density scenes. In addition, the original prediction heads are replaced with Transformer Prediction Heads (TPH) with the self-attention mechanism. As a consequence, prediction is enhanced.

![Figure 10. Architecture of YOLOv5-transformer.](image)

The architecture of the transformer adopts stacked self-attention and point-wise, fully connected layers for both the encoder and decoder (see the left and right halves of Figure 11 [43]).
3.8. YOLOv5-Transformer-BiFPN

In this model, exploring the prediction potential of self-attention based on the YOLOv5, the TRANS module is integrated into the prediction heads instead of the original prediction heads. This accurately localizes objects in high-density scenes and can handle the large-scale variance of objects. Moreover, at the neck of the network, PANet is replaced with a simple but effective BiFPN structure to weight the combination of multi-level features of backbone. The specific details of TRANS together with the BiFPN are depicted in Figure 12 [44].

Figure 11. Model architecture of transformer.

Figure 12. Architecture of YOLOv5-Transformer-BiFPN.
3.9. YOLOv5-FPN

YOLOv5-FPN uses PANet to aggregate image features. As demonstrated in Figure 13 [45], PANet builds on FPN’s deep-to-shallow unidirectional fusion by incorporating secondary fusion from the bottom up and employing precise low-level localization signals to improve the overall feature hierarchy and encourage information flow.

Figure 13. YOLOv5-FPN Structure [45].

4. Experimental and Discussion

4.1. Terahertz Image Processing

There are 329 images in the original dataset. Each image size is 512 px by 256 px. After image augmentation (such as flipping, warping, rotating and blending), there are 1884 images in total. The average size of the bounding box is 89.52 px by 74.45 px. These datasets are divided into a training set and test set according to the ratio of 8:2, which yields 1507 and 377 as the training set and test set, respectively. During the training, we also enabled the mosaic online amplification method, as shown in Figure 14. That is, each input graph is randomly fused with four sub-graphs. During training and testing, the input image size is set to 640 px by 640 px.

Figure 14. Mosaic augmentation for training process.
To avoid the influence of pre-training weight, all comparison models adopt from scratch the training method, the batch size is set to 16, takes turns on a 2080 (8G) graphics card, and the number of training rounds as epochs is set to 200. In addition, the reasons for the improvement of the model’s effect are also analyzed.

4.2. Model Comparison

In this section, we compare the performance between the proposed model and the existing general detection models. In this work, the detection metrics introduced in [33] are adopted, which includes average precision (AP) and average recall (AR) over multiple Intersection over Union (IoU) values. The detection metrics are listed in Table 2, and the true positive and false positive matrices for calculating the precision and recall are shown in Figure 15.

![Confusion Matrix](image)

**Figure 15.** Calculation of precision and recall.

The performance indicators are precision, recall mAP@0.5 as well as mAP@0.5:0.95. The average of AP is mAP (mean average precision). The AP is computed for each class and combined in certain situations. However, in some situations, they are interchangeable. There is no distinction between AP and mAP in the COCO sense, for instance [33].

4.2.1. Experiment Results

Because the research in this paper is based on YOLOv5, different improvement methods are tried in the experimental process, such as changing the transformer-based backbone, using the FPN network as neck, adding an additional prediction head, etc. The relevant experimental comparison results are shown in Table 2 and Figure 16.

It can be seen from the results that the best effect is achieved on the test set by using the BiFPN network as the neck. The test results in each category are shown in Table 3.
Figure 16. (a) Training-Loss and (b) Accuracy Graph.

Table 2. Model evaluation results on the test dataset.

| Model                        | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|------------------------------|-----------|--------|---------|--------------|
| YOLOv5-BiFPN (ours)          | 0.991     | 0.991  | 0.993   | 0.857        |
| YOLOv5                       | 0.99      | 0.96   | 0.995   | 0.862        |
| YOLOv5-fpn                   | 0.994     | 0.96   | 0.995   | 0.845        |
| YOLOv5-ghost                 | 0.987     | 0.983  | 0.992   | 0.855        |
| YOLOv5-p2                    | 0.98      | 0.974  | 0.981   | 0.835        |
| YOLOv5-p7                    | 0.99      | 0.988  | 0.993   | 0.847        |
| YOLOv5-p6                    | 0.991     | 0.98   | 0.99    | 0.85         |
| YOLOv5-Transformer           | 0.989     | 0.994  | 0.994   | 0.853        |
| YOLOv5-Transformer-BiFPN     | 0.993     | 0.987  | 0.994   | 0.854        |
| CSPDarknet53-PANet-SPP [46]  |           |        |         | 0.804        |
Table 3. Performance on each class.

| Model        | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|--------------|-----------|--------|---------|--------------|
| all          | 0.991     | 0.991  | 0.993   | 0.857        |
| screw_drive  | 0.975     | 0.987  | 0.992   | 0.705        |
| blade        | 0.992     | 1      | 0.995   | 0.793        |
| knife        | 0.989     | 0.988  | 0.995   | 0.782        |
| scissors     | 0.986     | 0.99   | 0.995   | 0.832        |
| board_marker | 0.995     | 1      | 0.995   | 0.914        |
| mobile_phone | 0.995     | 1      | 0.995   | 0.966        |
| wireless_mouse| 0.994    | 1      | 0.995   | 0.941        |
| water_bottle | 0.995     | 1      | 0.995   | 0.963        |

4.2.2. Model Analysis

To analyze the detection difference of each model, we analyze the convolution characteristic graphs of different models. Let the input image size be \((C, H, W)\) and the convolution layer feature image be \((c, h, w)\). First, we reduce the dimension in the channel \(c\) dimension; then, we take the average value of the \((h, w)\) dimension, scale the feature image size to the original image size, and finally overlay to output the final effect. Figure 17 shows the different feature maps of different models. The (a)–(i) values are the same as shown in Table 3. It can be seen from the feature map that the BiFPN network structure suppresses the non-target features and reduces the feature noise, while the original YOLOv5 model still has more feature representations at the edge of the object. Other models still have large errors in terahertz image feature extraction, which reduces the accuracy of the model.

4.3. Model Transfer Learning

Transfer learning is a common skill to accelerate model convergence in the field of deep learning. Previous studies have adopted ab initio training to ensure consistency. This section will discuss the acceleration effect of transfer learning on the model. Since the backbone of the proposed network is consistent with the original YOLOv5 network, we can use the pre-training weight of the backbone for migration learning. The changes of various indicators in the training process are shown in Figure 18, and the evaluation results in the test set are shown in Table 4.

It can be seen from Figure 18 that the method of transfer learning can accelerate the convergence of network training and shorten the model training time under the condition of ensuring the same model accuracy. The fine-tuned network has achieved better results in the test set, especially in the detection of some dangerous goods as observed from Tables 3 and 4, where mAP@0.5 and mAP@0.5:0.95 values witness a percentage increase of 0.2% and 1.7%, respectively. It is also obvious from Table 2 that our model compared to [46] enjoys a 0.5% and 7% percentage increase concerning detection accuracy using the same THz dataset at COCO’s evaluation metric.
Figure 17. Convolutional feature map visualization. sub graph (a–i): (x-axis = Epoch and y-axis = Colour intensity).

Table 4. Performance on each class with fine-turning.

| Model          | Precision | Recall | mAP@0.5 | mAP@0.5:0.95 |
|----------------|-----------|--------|---------|--------------|
| all            | 0.992     | 0.998  | 0.995   | 0.874        |
| screw_drive    | 0.987     | 0.987  | 0.994   | 0.739        |
| blade          | 0.985     | 1      | 0.995   | 0.792        |
| knife          | 0.982     | 1      | 0.994   | 0.786        |
| scissors       | 1         | 1      | 0.995   | 0.861        |
| board_marker   | 0.996     | 1      | 0.995   | 0.933        |
| mobile_phone   | 0.995     | 1      | 0.995   | 0.967        |
| wireless_mouse | 0.994     | 1      | 0.995   | 0.931        |
| water_bottle   | 0.996     | 1      | 0.995   | 0.982        |
5. Conclusions

Terahertz technology is a harmless security detection method, which is of great significance to the rapid and correct recognition of terahertz images. In this paper, a terahertz image target detection method based on BiFPN network feature fusion is proposed. The research results show that when using user-defined datasets, the proposed method is better than other improved models in terahertz feature extraction and classification. In subsequent research, we will focus on how to improve the terahertz image dataset and make it suitable for general target detection algorithms in the field of machine vision.

Author Contributions: Idea conceptualization, Methodology, Writing, S.A.D.; Formal analysis, Supervision and Funding acquisition, L.S.; Formal analysis and Supervision, D.H.; Writing—Editing and Review, J.O.; Formal analysis, Q.L.; Writing—Editing and Review, B.N.E.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 11872058) and the Sichuan Science and Technology Program of China (Grant No. 2019YFG0114).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors have no conflict of interest to declare that are relevant to the content of this article.

References
1. Danso, S.; Liping, S.; Deng, H.; Odoom, J.; Appiah, E.; Etse, B.; Liu, Q. Denoising Terahertz Image Using Non-Linear Filters. *Comput. Eng. Intell. Syst.* 2021, 12. [CrossRef]
2. Penkov, N.V.; Goltyaev, M.V.; Astashev, M.E.; Serov, D.A.; Moskovskiy, M.N.; Khort, D.O.; Gudkov, S.V. The Application of Terahertz Time-Domain Spectroscopy to Identification of Potato Late Blight and Fusariosis. *Pathogens* 2021, 10, 1336. [CrossRef] [PubMed]
3. Hu, J.; Xu, Z.; Li, M.; He, Y.; Sun, X.; Liu, Y. Detection of Foreign-Body in Milk Powder Processing Based on Terahertz Imaging and Spectrum. *J. Infrared Millimeter Terahertz Waves* 2021, 42, 878–892. [CrossRef]
4. Pan, S.; Qin, B.; Bi, L.; Zheng, J.; Yang, R.; Yang, X.; Li, Y.; Li, Z. An Unsupervised Learning Method for the Detection of Genetically Modified Crops Based on Terahertz Spectral Data Analysis. *Secur. Commun. Netw.* 2021, 2021, 5516253. [CrossRef]

5. Ge, H.; Lv, M.; Lu, X.; Jiang, Y.; Wu, G.; Li, G.; Li, L.; Li, Z.; Zhang, Y. Applications of THz Spectral Imaging in the Detection of Agricultural Products. *Photonics* 2021, 8, 518. [CrossRef]

6. Wang, L. Terahertz Imaging for Breast Cancer Detection. *Sensors* 2021, 21, 6465. [CrossRef]

7. Yin, X.X.; Hadjiiloucas, S.; Zhang, Y.; Tian, Z. MRI radiogenomics for intelligent diagnosis of breast tumors and accurate prediction of neoadjuvant chemotherapy responses—A review. *Comput. Methods Programs Biomed.* 2021, 214, 106510. [CrossRef]

8. Kansal, P.; Gangadharapappa, M.; Kumar, A. Terahertz E-Healthcare System and Intelligent Spectrum Sensing Based on Deep Learning. In *Advances in Terahertz Technology and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 307–335.

9. Liang, D.; Xue, F.; Li, L. Active Terahertz Imaging Dataset for Concealed Object Detection. *arXiv* 2021, arXiv:2105.03677.

10. Owda, A.Y.; Salmon, N.; Owda, M. Indoor passive sensing for detecting hidden objects under clothing. In Proceedings of the Emerging Imaging and Sensing Technologies for Security and Defence VI, Online, 13–18 September 2021; Volume 11868, pp. 87–93.

11. Dixit, N.; Mishra, A. Standoff Detection of Metallic Objects Using THz Waves. In *ICOL-2019*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 911–914.

12. Xu, F.; Huang, X.; Wu, Q.; Zhang, X.; Shang, Z.; Zhang, Y. YOLO-MSFG: Toward Real-Time Detection of Concealed Objects in Passive Terahertz Images. *IEEE Sens. J.* 2021, 22, 520–534. [CrossRef]

13. Xie, X.; Lin, R.; Wang, J.; Qiu, H.; Xu, H. Target Detection of Terahertz Images Based on Improved Fuzzy C-Means Algorithm. In Proceedings of the 2021 Chinese Intelligent Systems Conference, Fuzhou, China, 16–17 October 2022; pp. 761–772.

14. Wang, T.; Wang, K.; Zou, K.; Shen, S.; Yang, Y.; Zhang, M.; Yang, Z.; Liu, J. Virtual unrolling technology based on terahertz computed tomography. *Opt. Lasers Eng.* 2022, 151, 106924. [CrossRef]

15. Mao, Q.; Liu, J.; Zhu, Y.; Lv, C.; Lu, Y.; Wei, D.; Yan, S.; Ding, S.; Ling, D. Developing industry-level terahertz imaging resolution using mathematical model. *IEEE Trans. Terahertz Sci. Technol.* 2021, 11, 583–590. [CrossRef]

16. Widyastuti, R.; Yang, C.K. Cat’s nose recognition using you only look once (YOLO) and scale-invariant feature transform (SIFT). In Proceedings of the 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 9–12 October 2018; pp. 55–56.

17. Thu, M.; Suvonvorn, N. Pyramidal Part-Based Model for Partial Occlusion Handling in Pedestrian Classification. *Adv. Multimed.* 2020, 2020, 6153580. [CrossRef]

18. Huang, B.; Chen, R.; Xu, W.; Zhou, Q.; Wang, X. Improved Fatigue Detection Using Eye State Recognition with HOG-LBP. In Proceedings of the 9th International Conference on Computer Engineering and Networks, Dubai, United Arab Emirates, 19–20 February 2022; pp. 365–374.

19. Hazgui, M.; Ghazouani, H.; Barhoumi, W. Genetic programming-based fusion of HOG and LBP features for fully automated texture classification. *Vis. Comput.* 2021, 38, 457–476. [CrossRef]

20. Pu, Y.; Apel, D.B.; Szmigiel, A.; Chen, J. Image recognition of coal and coal gangue using a convolutional neural network and transfer learning. *Energies* 2019, 12, 1735. [CrossRef]

21. Zhou, Z.; Lu, Q.; Wang, Z.; Huang, H. Detection of Micro-Defects on Irregular Reflective Surfaces Based on Improved Faster R-CNN. *Sensors* 2019, 19, 5000. [CrossRef] [PubMed]

22. Zhang, M.; Li, H.; Xia, G.; Zhao, W.; Ren, S.; Wang, C. Research on the application of deep learning target detection of engineering vehicles in the patrol and inspection for military optical cable lines by UAV. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; Volume 1, pp. 97–101.

23. Li, W.; Feng, X.S.; Zha, K.; Li, S.; Zhu, H.S. Summary of Target Detection Algorithms. *J. Phys. Conf. Ser.* 2021, 1757, 012003. [CrossRef]

24. Liang, F.; Zhou, Y.; Chen, X.; Liu, F.; Zhang, C.; Wu, X. Review of Target Detection Technology based on Deep Learning. In Proceedings of the 5th International Conference on Control Engineering and Artificial Intelligence, Online, 15 January 2021; pp. 132–135.

25. Dai, Y.; Liu, Y.; Zhang, S. Mask R-CNN-based Cat Class Recognition and Segmentation. *J. Phys. Conf. Ser.* 2021, 1966, 012010. [CrossRef]

26. Shi, J.; Zhou, Y.; Zhang, W.X.Q. Target detection based on improved mask rccnn in service robot. In Proceedings of the 2019 Chinese Control Conference (CCC), Guangzhou, China, 27–30 July 2019; pp. 8519–8524.

27. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

28. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustain. Cities Soc.* 2021, 65, 102600. [CrossRef]

29. Kumar, A.; Kumar, A.; Bashir, A.K.; Rashid, M.; Kumar, V.A.; Kharel, R. Distance based pattern driven mining for outlier detection in high dimensional big dataset. *ACM Trans. Manag. Inf. Syst.* 2021, 13, 1–17. [CrossRef]

30. Chien, S.; Chen, Y.; Yi, Q.; Ding, Z. Development of Automated Incident Detection System Using Existing ATMS CCTV; Purdue University: West Lafayette, IN, USA, 2019.
31. Jaszewski, M.; Parameswaran, S.; Hallenborg, E.; Bagnall, B. Evaluation of maritime object detection methods for full motion video applications using the pascal voc challenge framework. In Proceedings of the Video Surveillance and Transportation Imaging Applications, San Francisco, CA, USA, 8–12 February 2015; Volume 9407, p. 94070Y.

32. Zhou, P.; Ni, B.; Geng, C.; Hu, J.; Xu, Y. Scale-transferrable object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 528–537.

33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.

34. Ping-Yang, C.; Hsieh, J.W.; Chen, Y.S. Light-Weight Mixed Stage Partial Network for Surveillance Object Detection with Background Data Augmentation. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3333–3337.

35. Liao, J.; Zou, J.; Shen, A.; Liu, J.; Du, X. Cigarette end detection based on EfficientDet. J. Phys. Conf. Ser. 2021, 1748, 062015. [CrossRef]

36. Wang, K.; Liew, J.H.; Zhou, Y.; Zhou, D.; Feng, J. Panet: Few-shot image semantic segmentation with prototype alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9197–9206.

37. Chen, Z.; Cong, R.; Xu, Q.; Huang, Q. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. IEEE Trans. Image Process. 2020, 30, 7012–7024. [CrossRef] [PubMed]

38. 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.

39. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 658–666.

40. Xu, R.; Lin, H.; Lu, K.; Cao, L.; Liu, Y. A forest fire detection system based on ensemble learning. Forests 2021, 12, 217. [CrossRef]

41. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 1580–1589.

42. 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.

43. Zolotareva, E.; Tashu, T.M.; Horváth, T. Abstractive Text Summarization using Transfer Learning. In Proceedings of the ITAT, Oravská Lesná, Slovakia, 18–22 September 2020; pp. 75–80.

44. Guo, Z.; Wang, C.; Yang, G.; Huang, Z.; Li, G. MSFT-YOLO: Improved YOLOv5 Based on Transformer for Detecting Defects of Steel Surface. Sensors 2022, 22, 3467. [CrossRef] [PubMed]

45. Qiu, Z.; Zhao, Z.; Chen, S.; Zeng, J.; Huang, Y.; Xiang, B. Application of an Improved YOLOv5 Algorithm in Real-Time Detection of Foreign Objects by Ground Penetrating Radar. Remote Sens. 2022, 14, 1895. [CrossRef]

46. Danso, S.A.; Liping, S.; Deng, H.; Odoom, J.; Chen, L.; Xiong, Z.G. Optimizing Yolov3 detection model using terahertz active security scanned low-resolution images. Theor. Appl. Sci. 2021, 3, 235–253. [CrossRef]