Intelligent Automation
& Soft Computing

Tech Science Press

DOI: 10.32604/iasc.2023.038257
Article

PF-YOLOv4-Tiny: Towards Infrared Target Detection on Embedded Platform

Wenbo Li, Qi Wang* and Shang Gao

School of Computer, Jiangsu University of Science and Technology, Zhenjiang, 212003, China
*Corresponding Author: Qi Wang. Email: wangqi@just.edu.cn

Received: 05 December 2022; Accepted: 15 February 2023

Abstract: Infrared target detection models are more required than ever before to be deployed on embedded platforms, which requires models with less memory consumption and better real-time performance while considering accuracy. To address the above challenges, we propose a modified You Only Look Once (YOLO) algorithm PF-YOLOv4-Tiny. The algorithm incorporates spatial pyramidal pooling (SPP) and squeeze-and-excitation (SE) visual attention modules to enhance the target localization capability. The PANet-based-feature pyramid networks (P-FPN) are proposed to transfer semantic information and location information simultaneously to ameliorate detection accuracy. To lighten the network, the standard convolutions other than the backbone network are replaced with depthwise separable convolutions. In post-processing the images, the soft-non-maximum suppression (soft-NMS) algorithm is employed to subside the missed and false detection problems caused by the occlusion between targets. The accuracy of our model can finally reach 61.75%, while the total Params is only 9.3 M and GFLOPs is 11. At the same time, the inference speed reaches 87 FPS on NVIDIA GeForce GTX 1650 Ti, which can meet the requirements of the infrared target detection algorithm for the embedded deployments.

Keywords: Infrared target detection; visual attention module; spatial pyramid pooling; dual-path feature fusion; depthwise separable convolution; soft-NMS

1 Introduction

Objects in natural environments emit infrared radiation continuously if their temperature is above absolute zero. Therefore, corresponding to the temperature distribution of the scene, a thermal image can be generated by collecting and detecting these infrared radiant energies. The thermal image is capable of recreating the difference in temperature and radiation emissivity of the scene's various parts, thus showing the features of the object and forming an infrared image. In some particular weather conditions, such as rain, fog, nighttime, and lack of visible light sources, pictures taken based on visible light cannot be exploited for target detection due to the poor visibility distance. Infrared imaging technology has strong penetration ability, long working distance, strong anti-interference power, high measurement accuracy, works day and night, etc. Therefore, target detection based on images obtained by infrared imaging technology has attracted widespread attention in research, and the market demand for it has also increased.

Recently, infrared target detection has been extensively employed in various spheres, such as vessel detection in the horizontal plane [1], intelligent inspection of power equipment [2], and traffic sign recognition [3]. Based on feature distillation, Beiming Li et al. put forward an improved Ghost-YOLOv5 (the version 5 of You Only Look Once algorithm) algorithm [4] for the problems of poor real-time and high computational complexity of infrared target detection model YOLOv5s [5]; Lang Shu et al. proposed a network structure that incorporates the features of DenseNet network and YOLOv5. Because of the lack of open-source infrared data and the problems of the low signal-to-noise ratio, the low resolution of infrared images seriously affect the performance of target detection. Based on the characteristics of infrared images, Ruzhen Zhang et al. analyzed the current mainstream image augmentation methods. They proposed a channel expansion-based infrared image data augmentation algorithm, which powerfully increases the amount of information in the original image to raise the detection accuracy of the network [6]. Although the above methods effectively ameliorate the accuracy, they are all proposed based on common platforms with low requirements for real-time detection. Nevertheless, in other application areas, they can only be deployed on embedded platforms and have high requirements for real-time performance in many cases. Therefore, the study of infrared target detection algorithms for embedded systems has become one of the popular research topics in academia and industry in recent years.

The main goal of this work is to design an infrared target detection algorithm PF-YOLOv4-Tiny with high recognition accuracy and robustness for embedded platforms to address the problems of blurred edges of infrared images, severe occlusions, difficulties in small target recognition, and high real-time requirements. Our contributions are summarized as follows:

1) Select YOLOv4-Tiny [7] with fewer parameters as the detection framework to accommodate embedded deployment. Mosaic [7] is applied to pre-process the images, and modules such as spatial pyramidal pooling (SPP) [8] and attention module squeeze-and-excitation (SE) are fused to mitigate the impact of the low contrast of infrared images and the problem of blurred target contours. At the same time, the ameliorated soft-non-maximum suppression (soft-NMS) [9] algorithm is employed as post-processing to lower the occurrence of misdetection and false detection problems and improve the detection accuracy.
2) Enhance the original feature pyramid networks (FPN) based on the viewpoint of path aggregation network (PANet) [10] dual-path feature fusion to transfer semantic information along with location information to heighten the detection accuracy.
3) For the purpose of lightweighting the network, depthwise separable convolutions are used to replace conventional convolutional operations except for the backbone network.
4) Regenerate anchor boxes via the K-means clustering method to cause them to be more compatible with the sizes of the actual objects.

The remaining sections of this article are organized as follows: Section 2 describes the relevant background knowledge related to this study, such as the introduction of the infrastructure, commonly employed enhanced theories, and so on; Section 3 introduces the methods adopted in the proposed algorithm PF-YOLOv4-Tiny and related details; Section 4 gives the experimental operating environment, relevant parameters settings, experimental results, and specific analysis; Section 5 concludes the work.

2 Related Works

2.1 Target Detection Algorithms

Mainstream target detection algorithms are divided into two categories: one-stage-based methods and two-stage-based methods. The one-stage target detection algorithms derive from the regression paradigm, which does not bring candidate frames but directly categorizes objects and predicts the candidate frames. The algorithms simplify the network structure, and the accuracy is lower than the two-stage target detection framework, but the detection speed is faster. The iconic algorithms are You Only Look Once (YOLO) [11] series, Single Shot MultiBox Detector (SSD) series [12], Anchor-Free [13–15] series, etc. The two-stage target detection algorithms split the detection process into two steps, first generating candidate frames by region selection methods, then combining with convolutional neural network (CNN) to classify and regress the positions of candidate frames. The algorithms are slower than the one-stage-based methods but have higher accuracy, and the representative algorithms are R-CNN [16], Fast-RCNNN [17], Faster-RCNN [18], etc.

In this study, we opt for YOLOv4-Tiny as the detection framework. YOLOv4-Tiny is a lightweight version of YOLOv4 [7] with multi-tasking, end-to-end, multi-scale, and attention mechanism features. Multi-tasking signifies that YOLOv4-Tiny is capable of completing the classification and regression of targets simultaneously. End-to-end implies that the model can give classification and regression prediction information directly after receiving image data. Multi-scale represents that the output feature maps of different convolutional layers are scaled to a uniform size to contain general global and local detail information. The attention module denotes that the model will focus on the target region features and process them in detail to obtain more detailed information and suppress other useless information, thus accelerating the processing speed.

Since YOLOv4-Tiny performs better than YOLOv3 [19] at the detection accuracy, and with fewer parameters, only 5918006, which is one-tenth of YOLOv4, YOLOv4-Tiny is faster and more suitable for embedded platforms. Its network architecture is shown in Fig. 1. The overall network structure has 38 layers, and the backbone network is Cross Stage Partial DarkNet53_Tiny (CSPDarkNet53_Tiny), which consists of CBL (Conv2D-BN-LeakyReLU) and Res_Block. CBL is a stack of three basic operations: the Conv2D convolutional layer, the BN normalization layer, and the LeakyReLU activation layer. Res_Block comprises the CBL module, the residual network, and the MaxPool2d maximum pooling. In the Res_Block, the channel segmentation of the feature extraction network is performed first, and the channel of the feature layer output after 3×3 convolution is divided into two parts, the first part is kept, and the second part is taken for the residual network. Finally, the result of the second part is fused with the first part again, as shown in Fig. 2. The feature pyramid FPN is applied when merging the two valid feature layers of the backbone network output. Then, the network output is fed into the detection head.

In convolutional networks, semantic features are lightly responded by the deep network, and the shallow network rashly reflects image features. However, the deep network possesses little geometric information, which is unfavorable for target detection; the shallow network contains more geometric information but few semantic features, which prejudices the classification. YOLOv4-Tiny exploits FPN, which combines multi-level scale features to solve the multi-scale problem and enhance the detection accuracy of small targets. Yet the location information is not conveyed, although its top-down structure enhances the semantic information. In order to work out the above matters, the PANet structure [20] was first proposed, which puts forward a bottom-up secondary fusion approach. The structure is to simply add the bottom-up fusion path on top of the FPN to pass up the strong localization features of the bottom layer, making the information path between the bottom and top features shorter. Subsequently, more complex FPN bidirectional fusion improvements emerged, such as ASFF [21], NAS-FPN [22], and BiFPN [23].

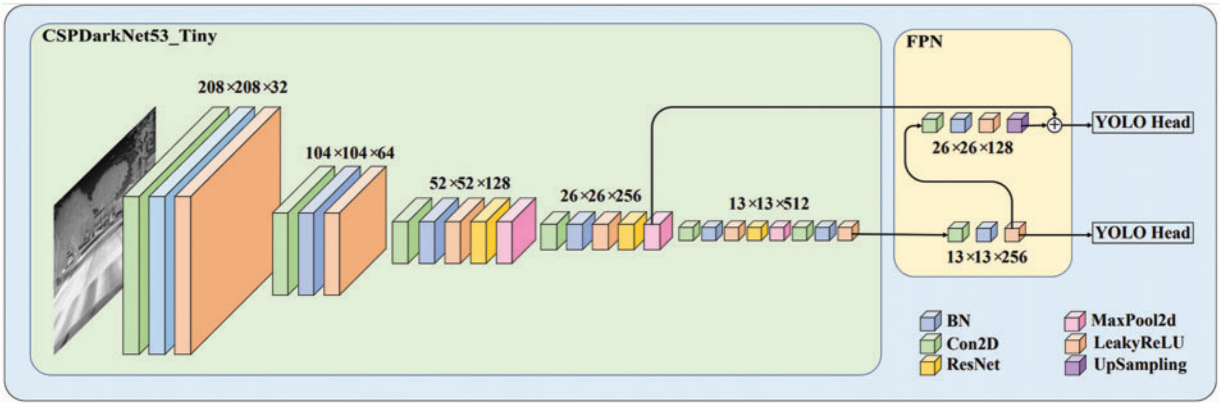

Figure 1: YOLOv4-Tiny network architecture. Extracted from the input image by CSPDarkNet53_Tiny, features will be fused in the FPN module and passed into YOLO head

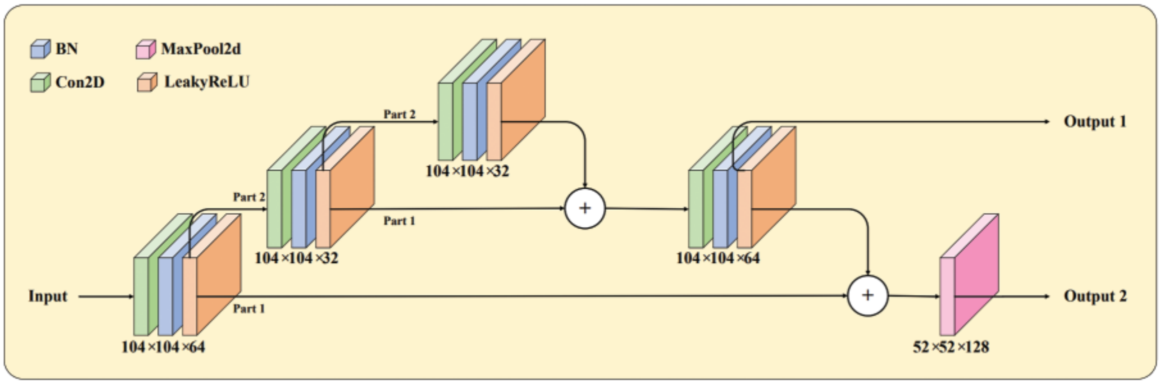

Figure 2: Res_Block module structure. In this module, residual feature fusion is performed twice

2.2 *Infrared Datasets*

Dataset is one of the essential factors affecting the merit of deep learning models. The clearer the features of images in the dataset, the better the acquired features, and the better the detection results will usually be. Unlike visible image data, which can be directly obtained by taking multiple images and averaging them, infrared datasets full of noise-reduction images are more challenging to produce. The number of publicly available infrared image datasets nowadays is relatively tiny in comparison, and the main ones are as follows.

1) SCUT FIR Pedestrian Dataset [24]

SCUT FIR pedestrian dataset is a large far-infrared pedestrian detection dataset gathering from 11 kinds of road scenarios of the under downtown, suburbs, expressway, and the campus in Guangzhou, China. The images are separated into walk person, ride person, squat person, and people, etc.

2) OTCBVS Dataset [25]

The OTCBVS dataset is a public dataset designed to test and evaluate original and advanced computer vision algorithms or non-visual (e.g., infrared) computer vision algorithms. It includes 13 sub-datasets, such as OSU Thermal Pedestrian Database, IRIS Thermal, Visible Face Database, OSU Color-Thermal Database, and OSU Color-Thermal Database.

3) KAIST Dataset [26]

The KAIST pedestrian dataset consists of 95,328 images covering a variety of regular traffic scenes on campuses, streets, and in the countryside during day and night. Each image contains both an RGB color image and an infrared image version, labeled into three categories: person, people, and cyclist.

4) Teledyne FLIR Thermal Dataset [27]

Providing fully annotated thermal and visible spectral frames, the Teledyne FLIR ADAS dataset helps developers exploit and train CNN so as to enable the auto industry to create safer and more efficient ADAS and driverless vehicle systems via FLIR's cost-effective thermal imager. The Teledyne FLIR thermal sensors can realize the detection and classification even though the conditions are challenging, such as total darkness, smoke, and glare. The detection range is four times that of an ordinary headlight. Detection objects are divided into 15 categories: People, Bicycles, Cars, Trains, and so on.

2.3 Image Data Augmentation Methods

As a part of pre-processing, data augmentation uses the existing data to produce values higher than itself without adding additional data. It can be split into two categories: supervised data augmentation and unsupervised data augmentation.

Supervised data augmentation refers to the augmentation of existing data using established data transformation rules and involves single-sample data augmentation and multiple-sample data augmentation. Single-sample data augmentation manipulates the sample, such as common geometric transformations (flip, crop, stretch, etc.) and color transformations (noise, erase, fill, etc.). Multi-sample data augmentation applies to the formation of new samples using multiple samples, such as Mixup [28], SMOTE [29], and SamplePairing [30].

Unsupervised data augmentation includes two types. The first one is represented by Generative Adversarial Network (GAN) [31], which randomly generates images consistent with the distribution of the training dataset by learning the distribution of the data; the second augmentation method is supported by AutoAugment [32], which derives the data augmentation method befitting for the task by model learning.

2.4 Visual Attention Module

Due to the disadvantages of infrared images, such as low pixel resolution, poor contrast, and blurred edges, if attention is equally allocated to the complete input image, the model will be challenging to focus on the target areas quickly and learn more practical feature information, thus having a more significant impact on the classification and regression results. Especially for small infrared targets, which may seriously lower the recognition results or even be impossible to recognize because they are not easily noticed in the first place.

The visual attention module is the brain signal processing module unique to human vision, and the attention module in deep learning borrows from the attention mindset of human vision.

It is essentially a resource allocation mechanism that assists the model in screening out the target regions with important value information from many irrelevant background regions and allocates more attention to that region, thus ameliorating the recognition efficiency. Commonly used optical attention modules include the channel attention module Squeeze-and-Excitation Networks (SENet) [33], and the mixed spatial and channel attention module Convolutional Block Attention Module (CBAM) [34].

2.5 Depthwise Separable Convolution

Depthwise separable convolution is a process of splitting the original convolution operation into two: depthwise convolution and pointwise convolution. In depthwise convolution process, one convolution kernel is responsible for one channel, one channel is convolved by only one convolution kernel, and the number of channels of the feature maps produced by this process is exactly the same as the number of channels of the input. Taking the feature maps output after depthwise convolution process in depthwise convolution, pointwise convolution is the process of weighing the feature maps in the depthwise direction. In this process, the number of convolution kernels is equal to the number of the final feature maps.

Assume that the size of the convolution kernel is W × H, the size of the input feature maps is $W_i \times H_i \times C_{in}$ (C_{in} is the number of channels), and the size of the output feature maps is $W_o \times H_o \times C_{out}$, then in the depthwise convolution process, the Params P_1:

$$P_1 = W \times H \times C_{in} \tag{1}$$

and GFLOPs G_1:

$$G_1 = W \times H \times W_o \times H_o \times C_{in} \tag{2}$$

In the pointwise convolution process, the Params P_2:

$$P_2 = 1 \times 1 \times C_{in} \times C_{out} \tag{3}$$

and GFLOPs G_2:

$$G_2 = 1 \times 1 \times W_o \times H_o \times C_{in} \times C_{out} \tag{4}$$

While for the standard convolution, the Params P_3:

$$P_3 = W \times H \times C_{in} \times C_{out} \tag{5}$$

and GFLOPs G_3:

$$G_3 = W \times H \times W_o \times H_o \times C_{in} \times C_{out} \tag{6}$$

It is calculated that the Params P ($P = P_1 + P_2$) and the GFLOPs G ($G = G_1 + G_2$) of the depthwise separable convolution is less than the standard convolution.

$$\frac{P}{P_3} = \frac{G}{G_3} = \frac{1}{C_{out}} + \frac{1}{W \times H} \tag{7}$$

In summary, the depthwise separable convolution can effectively reduce the number of Params and improve the efficiency of convolutional kernels' parameters, which is more suitable for the lightweight networks.

2.6 Bounding Box Filtering Algorithms

As a universal post-processing method, the non-maximum suppression (NMS) algorithm applied by the YOLOv4-Tiny aims at keeping the highest-scoring Bounding Box (BBox) among the many candidate BBoxes generated when detecting the same target and setting the scores of other candidate BBoxes with which the IOU value exceeds the threshold value to 0. Despite this algorithm ensuring that each target corresponds to only one BBox, it will directly affect the detection of overlapping targets, resulting in false and missed detection, and it is too absolute to let the BBoxes that the scores are below the threshold to 0. Therefore, the improved algorithms such as DIOU-NMS [35], and soft-NMS have been developed to solve the above problems.

In brief, considering the model parameters, accuracy, and deployment requirements of embedded platforms, we select YOLOv4-Tiny as the basis for improving the infrared target detection model. Meanwhile, taking the dataset's application scenarios and label categories into account, the Teledyne FLIR Thermal Dataset is finally chosen as the experimental dataset, and suitable data augmentation methods and BBox filtering algorithms are utilized to enhance the detection accuracy. In addition, via deploying the attention module, a more reasonable feature fusion structure is proposed. Through the analysis of the detection results of its labeled infrared targets, the experimental base model YOLOv4-Tiny is refined and perfected so that the final infrared target detection model gets better recognition effects.

3 PF-YOLOv4-Tiny Algorithm

The objective of this work is to design an infrared target detection algorithm with high accuracy, high recall, and high real-time performance based on YOLOv4-Tiny for the puzzles such as blurred edges of infrared images, severe occlusion, difficulty in small target recognition and high real-time requirements for embedded platforms. To this end, we make the following retrofits on the basis of the YOLOv4-Tiny algorithm.

1) Exploit Mosaic for image pre-processing

Mosaic is a higher-order version of the multi-sample data augmentation method, which utilizes four images, stitches them together, and then passes the stitched new image into the model for learning. It is equivalent to importing four images at a time, greatly enhancing the learning efficiency, as demonstrated in Fig. 3.

2) Incorporate SE Block to locate the target location

SE Block, a substructure of SENet, is a kind of channel attention module. SE Block has three layers: a global average pooling layer and two fully connected layers applying the ReLU activation function and Sigmoid activation function, respectively, with the structure printed in Fig. 4. It models the interdependencies between feature map channels to enhance feature representation.

3) Combine SPP module

In this work, due to the blurred infrared target contours and low contrast, the SPP module is set to fuse features of different scales, enrich the feature map information, perfect the model localization ability, and enhance its robustness. The framework for integrating SPP is rough: image input, convolutional layer to extract features, SPP to extract fixed size features, and Concatenate layer. Like YOLOv3, we choose windows of sizes 1, 5, 9, 13 to pool the feature maps, as shown in Fig. 4.

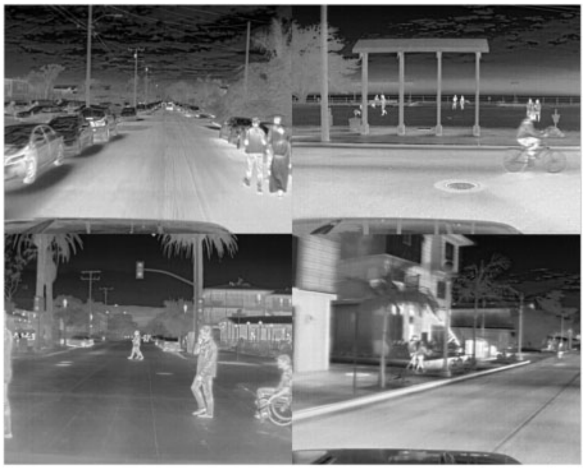

Figure 3: Mosaic augmentation method. Four images are spliced, and the generated new images are transmitted for learning

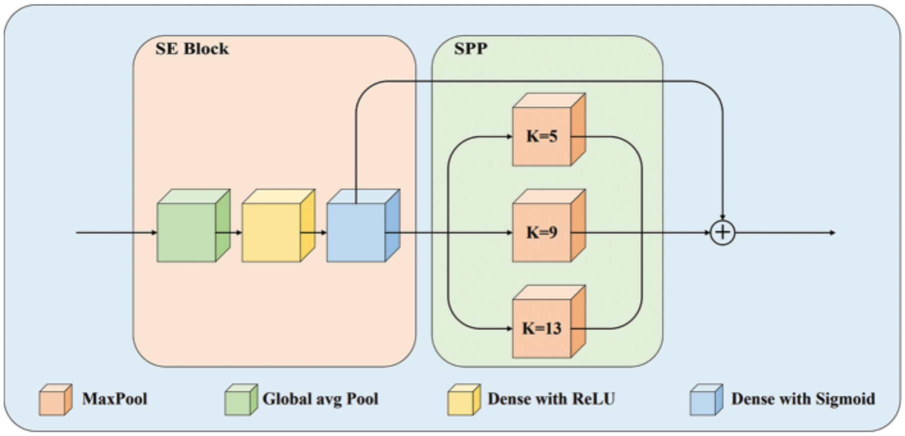

Figure 4: Structure of the module that incorporates the SE and the SPP. It is responsible for enhancing the receptive field and the target localization capability

4) Propose dual-path feature fusion structure P-FPN

We refactor the FPN in YOLOv4-Tiny based on the idea of dual-path feature fusion of PANet and blend in a bottom-up pyramid structure in front of the original FPN structure. The structure of PANet-based-feature pyramid networks (P-FPN) is exhibited in Fig. 5.

5) Switch to depthwise separable convolution

Depthwise separable convolution is an excellent choice for embedded deployments because it can significantly reduce the number of parameters, significantly reducing the model size. However, it needs to be used with caution because it decouples the spatial information from the depth information, resulting in incoherence between the information, which may affect the model's capability. In this paper, we want to build a lightweight model so that they can be ported to embedded devices. In order to minimize the adverse effects of this convolution, we decided not to change the backbone network but only to change the remaining part of the regular convolution to depthwise separable convolution, trying to reduce the number of model parameters without changing the detection capability of the model.

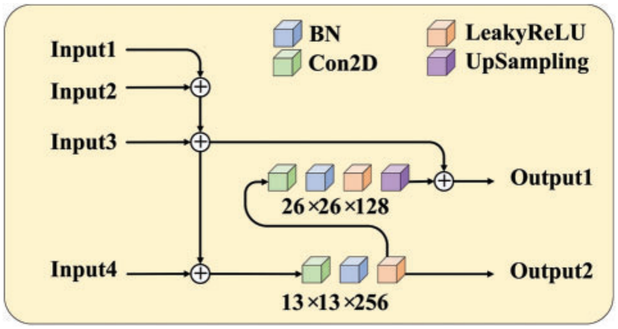

Figure 5: P-FPN structure. A bottom-up aggregation path is applied to the FPN

6) Utilize soft-NMS algorithm for post-processing

As described in Section 2.6, the NMS algorithm performs post-processing by sorting the BBoxes by score, keeping the one with the highest score while deleting other candidate BBoxes with which the IOU value exceeds the threshold value. This algorithm is subject to mistaken deletion of neighboring object BBoxes and is not suitable for target detection in relatively densely placed parts of objects. In the dataset, the infrared targets in some images are occluded and overlapped. In order to alleviate the impact of the above problem, the soft-NMS algorithm is a substitute. The mentality of this algorithm is to first sort the BBoxes according to their scores, retaining the BBox with the highest score, and reduce the scores of candidate BBoxes with an overlapping area more significant than a certain percentage of that BBox, as shown in Fig. 6. In other words, a decay function is set for the neighboring BBoxes based on the overlap size instead of ultimately setting their scores to zero. If a candidate BBox has considerable overlap with the highest scoring BBox, it will have a meager score. In contrast, if it has only a small overlap with the highest scoring BBox, then its original score will not be affected much. Therefore, soft-NMS can alleviate the occlusion problem to some extent.

7) Generate anchor size by K-means clustering method

The anchor size set in the original YOLOv4-Tiny is a generic one, which is designed by taking into account various considerations, and therefore may not be the most suitable anchor size for a particular dataset. K-means is a clustering algorithm that clusters the already labeled target box sizes into a specific number of categories, making the generated anchor more consistent with the size of the labeled targets in a specific data set species, thus making the detection better. For the YOLO network, we need to cluster them into 9 categories.

The final infrared target detection network PF-YOLOv4-Tiny structure is illustrated in Fig. 7. PF-YOLOv4-Tiny structure carries the relatively well-developed P-FPN pyramid structure for bi-directional transmission of feature information to synchronously transfer semantic information and location information to improve the detection accuracy. The SPP is combined with the channel attention module SE to quickly locate target information while fusing features of different scales to enhance the detection capability of the algorithm. All convolutions except the backbone network are replaced with depthwise separable convolutions to reduce the parameters and size of the model. The more suitable soft-NMS algorithm is applied to mitigate the problem of missed detection rate and false detection caused by inter-target occlusion.

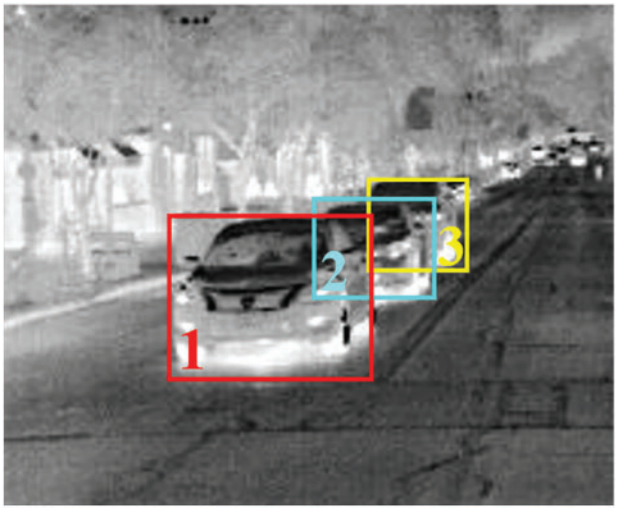

Figure 6: Bounding box filtering. Since object 2 is partially occluded by object 1, it may lead to the false deletion due to the large IOU between the candidate BBox of object 2 and the candidate BBox with the highest score of object 1. Soft-NMS algorithm attenuates the candidate BBox score with a large IOU instead of setting to 0, thus alleviating the problem of false deletion and missed detection caused by occlusion. The detection process of object 3 is the same

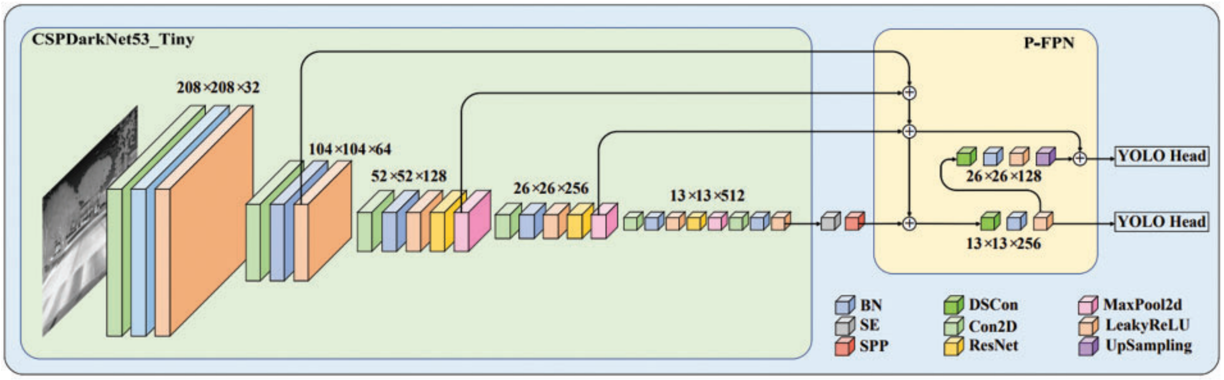

Figure 7: PF-YOLOv4-Tiny infrared target detection network structure. The SE channel attention module, SPP module, and P-FPN pyramid structure are contained here and DSCon stands for depthwise separable convolution

In addition, to make the recognition results more desirable, we bring the K-means clustering method into PF-YOLOv4-Tiny to regenerate the anchor boxes, which makes them more compatible with the sizes of the actual objects.

4 Experiments

Given the quality of dataset images, application feasibility, and the number of dataset images, the Teledyne FLIR Thermal Dataset is used in this experiment. To avoid the unnecessary effects of the unbalanced number of samples in other categories, we focus on recognizing the three types of objects: People, Bicycles, and Cars. 7177 images out of 8862 images in this dataset are settled as the training set, 798 images are employed as the validation set, and the remaining 887 images are taken as the training set. In addition, a large number of small targets and mutual occlusion problems in this dataset need to be taken into consideration.

To test the effect of different improvements on model accuracy, we conduct ablation experiments of the attention module SE, soft-NMS algorithm, SPP module, P-FPN structure, and depthwise separable convolution on the above dataset.

4.1 Experimental Setup

The experiments are based on the deep learning framework TensorFlow, implemented via Python. The training platform is the desktop computer with windows 10, i7-10750H CPU, NVIDIA GeForce GTX 1650 Ti GPU, and CUDA version 10.1. And tensorflow_gpu 2.2.0 is used in the experimental environment.

During training, the initial learning rate is set to 0.01, the Stochastic Gradient Descent (SGD) algorithm is utilized, and the cosine annealing algorithm is employed as the learning rate optimization strategy. In the pre-processing of the input images, the Mosaic data augmentation method is covered in addition to the traditional data augmentation methods such as random flip, color gamut transformation, and distortion. Meanwhile, the anchor boxes are regenerated by the K-means clustering method, which results in a higher overlap between the anchor boxes sizes and the sizes of the actual objects in the final prediction map.

The model is evaluated by the mean average accuracy (mAP) and the recall, which aims to compare and analyze the experimental results and thus determine whether a particular part of the improvement is necessary. The recall represents the proportion of predicted positive samples among all positive samples.

4.2 Ablation Experiments

4.2.1 Ablation Experiments of Attentional Module

The most frequently employed visual attention modules are SE and CBAM, which are availed by the original network YOLOv4-Tiny. After training and validation on the test set, the experimental results are shown in Table 1.

Table 1: Ablation experiments of attentional module

Model	AP			Recall	mAP
	Car	People	Bicycle		
YOLOv4-Tiny	62.89	48.22	36.53	47.31	49.21
YOLOv4-Tiny + CBAM	65.99	50.38	38.58	50.12	51.65
YOLOv4-Tiny + SE	67.02	49.95	44.30	51.22	53.76

Learning from the data in Table 1, the addition of the CBAM module enhances the recall by 2.81% and the mAP by 2.44% in contrast to the original network YOLOv4-Tiny, and the addition of the SE module raises the recall to 3.91% and the mAP by 4.55%. From the accuracy rate of three kinds of targets, the following conclusion can be reached that the SE visual attention module is more effective in bettering the accuracy of both Cars and Bicycles. In other words, the SE attention module based on channel selection is more efficient in ameliorating the target detection results of the FLIR infrared dataset.

4.2.2 Ablation Experiments of Soft-NMS

In the FLIR dataset, the issue of mutual target occlusion is serious. The application of the NMS algorithm for recognition may readily cause the problem of error deletion between BBoxes. In this case, soft-NMS is particularly necessary. It significantly reduces the false censoring rate of BBoxes between occluded objects and enhances the recognition results as much as possible. The specific experimental results are shown in Table 2.

Table 2: Ablation experiments of soft-NMS

Model	AP			Recall	mAP
	Car	People	Bicycle		
YOLOv4-Tiny	62.89	48.22	36.53	47.31	49.21
YOLOv4-Tiny + soft-NMS	70.07	53.83	47.31	55.45	57.07
YOLOv4-Tiny + CBAM	65.99	50.38	38.58	50.12	51.65
YOLOv4-Tiny + CBAM + soft-NMS	68.03	49.52	38.97	49.80	52.17
YOLOv4-Tiny + SE	67.02	49.95	44.30	51.22	53.76
YOLOv4-Tiny + SE + soft-NMS	69.58	54.64	48.28	55.53	57.50

The results indicate that the soft-NMS algorithm can perfect the detection effect, especially for recognizing Car and Bicycle target categories with a dense distribution. In contrast with the YOLOv4-Tiny model, the recognition accuracy of the YOLOv4-Tiny + soft-NMS model for Cars, People, and Bicycles is improved by 7.18%, 5.61%, and 10.78%, respectively. Compared with the YOLOv4-Tiny + SE model, the recall and mAP of YOLOv4-Tiny + soft-NMS + SE are also enhanced by 4.31% and 3.47%.

Meanwhile, we compare the recognition effects of the YOLOv4-Tiny model before and after applying the soft-NMS algorithm, as shown in Fig. 8. Fig. 8a exhibits the results of the YOLOv4-Tiny model before using the soft-NMS algorithm, and Fig. 8b displays the results of the YOLOv4-Tiny model after using the soft-NMS algorithm. From Fig. 8a, we can see that 2, 3, and 3 targets are detected from top to bottom separately, while Fig. 8b with the soft-NMS algorithm has 3, 4, and 5 targets detected, and the confidence of the boxes are all greater than or equal to the Fig. 8a.

Pondering the above data, replacing the soft-NMS algorithm is essential for infrared target detection in the FLIR dataset.

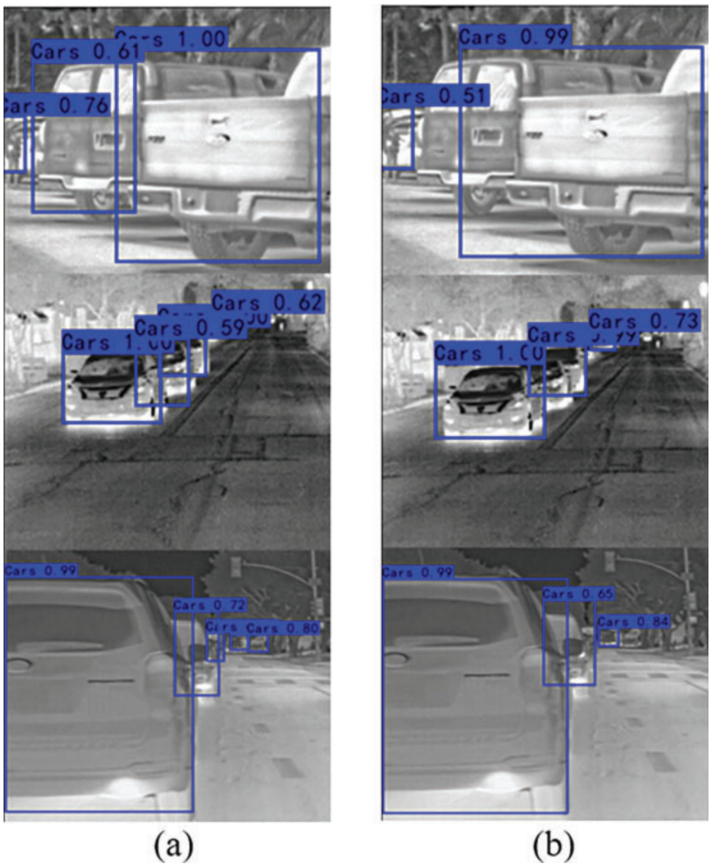

Figure 8: Comparison of before and after use of soft-NMS (a is before use, b is after use). The confidences and number of Fig. 8a both perform better than Fig. 8b by contrast

4.2.3 Ablation Experiments of SPP

Based on blending in the attention module and the soft-NMS algorithm, the YOLOv4-Tiny network is consummated by integrating the SPP module to the last layer of the backbone network output to expand the receptive field and fuse the high-level, bottom-level features for better target localization. The experimental results are shown in Table 3.

It implies that the addition of the SPP module indeed improves the recall and mAP of the YOLOv4-Tiny + SE + soft-NMS network, and the best improvement is achieved for certain targets with less obvious features in the images, such as the accuracy improvement of Bicycle by 5.64%. Compared with YOLOv4-Tiny, the recall and the mAP of the YOLOv4-Tiny + SPP model are raised to 10.5% and 8.35%, respectively. On the side of YOLOv4-Tiny + SE + soft-NMS, the overall recall of the YOLOv4-Tiny + SE + soft-NMS + SPP model is enhanced by 4.27%, and 2.59% improves mAP. The experimental results prove that it is helpful to introduce the SPP module to enhance detection accuracy.

Table 3: Ablation experiments of SPP

Model	AP			Recall	mAP
	Car	People	Bicycle		
YOLOv4-Tiny	62.89	48.22	36.53	47.31	49.21
YOLOv4-Tiny + SPP	69.10	53.65	49.95	57.81	57.56
YOLOv4-Tiny + SE + soft-NMS	69.58	54.64	48.28	55.53	57.50
YOLOv4-Tiny + SE + soft-NMS + SPP	71.43	54.92	53.92	59.80	60.09

4.2.4 Ablation Experiments of P-FPN

Enlightened by the thought of dual-path feature fusion of PANet, the FPN structure of YOLOv4-Tiny is established by joining a bottom-up pyramid structure to generate the new P-FPN structure. It transfers location information on top of the semantic information conveyed by the FPN, thus improving the detection accuracy, and the specific experimental results are shown in Table 4.

Table 4: Ablation experiments of P-FPN

Model	AP			Recall	mAP
	Car	People	Bicycle		
YOLOv4-Tiny	62.89	48.22	36.53	47.31	49.21
YOLOv4-Tiny + P-FPN	69.59	54.10	44.59	54.43	56.09
YOLOv4-Tiny + SE + soft-NMS + SPP	71.43	54.92	53.92	59.80	60.09
YOLOv4-Tiny + SE + soft-NMS + SPP + P-FPN	74.18	60.94	58.61	63.36	64.58

The recall and mAP of the YOLOv4-Tiny + P-FPN model are raised to 7.12% and 6.88% compared to the YOLOv4-Tiny model. At the same time, compared with the model YOLOv4-Tiny + soft-NMS + SE + SPP, the detection accuracy of the model with P-FPN for Cars, People, and Bicycles also improved by 2.75%, 6.02%, and 4.69%, respectively, and the overall recall and detection accuracy raised to 3.56% and 4.49%, which can be better applied to the detection of infrared targets in FLIR dataset.

4.2.5 Ablation Experiments of Depthwise Separable Convolution

At the end of all the improvements, we plan to replace the regular convolution with depthwise separable convolution to reduce the number of parameters and computational complexity. Considering the performance issues in model accuracy, we keep the convolution of the backbone network part used for feature extraction and replace the remaining part of the regular convolution with depthwise separable convolution, such as SPP, P-FPN, etc. Furthermore, the effect is shown in Table 5. FPS indicates the number of image frames per second that the model can process, reflecting its real-time nature. Params are the total number of parameters to be trained in the network model, and GFLOPs represent the amount of computation performed in the network model.

Table 5: Ablation experiments of depthwise separable convolution

Model	Recall	mAP	FPS	Params (M)	GFLOPs
YOLOv4-Tiny + SE + soft-NMS + SPP + P-FPN	63.36	64.58	41	26.350	30.68
PF-YOLOv4-Tiny	59.80	61.75	86.54	9.343	11.14

As seen from the data in the table above, the employment of depthwise separable convolution does cause the undesirable consequence of a decrease in model accuracy, with a 2.83% decrease in mAP for the model after replacement. However, after the replacement, the overall capability of the model has improved significantly, with FPS more than twice as fast as before, Params reduced to 35%, and GFLOPs even reduced by 64%. Combining the above data, we know that the use of depthwise separable convolution is correct, making the model lighter and more suitable for embedded devices.

4.2.6 Comparison Experiments

In this section, we compare the improved PF-YOLOv4-Tiny with the mainstream target detection algorithms YOLOv3 and YOLOv5, as shown in Table 6.

Table 6: Comparison experiments

Model	mAP	FPS	Params (M)	GFLOPs
YOLOv3	72.92	24.58	61.533	65.61
YOLOv5	68.94	70.42	7.018	15.8
YOLOv4-Tiny	49.21	131.93	5.879	6.837
PF-YOLOv4-Tiny	61.75	86.54	9.343	11.14

The data in the table show that although our PF-YOLOv4-Tiny model does slightly worse than YOLOv3 and YOLOv5 in terms of mAP, it is much better than both in terms of FPS, Params, and GFLOPs are much higher than YOLOv3. The experiments illustrate that our improvement for YOLOv4-Tiny is successful. The improved model PF-YOLOv4-Tiny has a balanced performance and is suitable to be deployed on embedded devices.

4.3 Experimental Results Analysis

From the above experiments, the conclusion that the soft-NMS algorithm is good at alleviating the accuracy degradation caused by inter-target occlusion like Cars can be drawn. The combination of SPP and SE modules greatly enhances the localization ability of the model and the recognition of Bicycles and People. The P-FPN structure shortens the transfer path of shallow features and further improves the accuracy. Also, Depth separable convolution has dramatically lightened the model, significantly improving the performance of the model's Params, FPS, and GFLOPs. Besides, the comparison experiments of PF-YOLOv4-Tiny with YOLOv3, YOLOv4-Tiny, and YOLOv5 illustrate that our model can implement infrared target detection on embedded devices.

5 Conclusion

To meet the performance requirements of embedded platforms in terms of real-time algorithms and accuracy, this paper selects YOLOv4-Tiny as the base network structure and Mosaic as the data augmentation method to amplify the dataset. SPP module is introduced to fuse features at different scales, which enriches the feature map information and enhances the localization capability. The SE serves as a channel attention module to allocate more attention to the target region, thus refining the recognition efficiency. The P-FPN structure is raised by enrolling a bottom-up pyramid structure in the FPN to convey both semantic and location information, which perfects the detection accuracy. To lighten the model, we replaced part of the regular convolutions in the model with depth-separable convolutions. At the same time, we apply the soft-NMS algorithm to replace the NMS algorithm of YOLOv4, which alleviates the occlusion issue in the process of dataset target recognition and further optimizes the infrared target detection accuracy effectively. In addition, the introduction of the K-means algorithm generates more suitable anchor boxes sizes for the dataset so as to make the detection results more desirable. As opposed to the YOLOv4-Tiny, the mAP of the proposed PF-YOLOv4-Tiny algorithm is improved to 61.75%, which is a 12.54% increase. It also has better real-time performance compared to mainstream one-stage networks. The inference speed reaches 87 FPS on NVIDIA GeForce GTX 1650 Ti, which can better meet the deployment requirements of IR target detection for embedded platforms.

Funding Statement: The article is supported by The Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grants No. 19JKB520031).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. W. Liu, "Transmission and identification of infrared images of ships," *Computer Simulation*, vol. 28, no. 2, pp. 316–319, 2011.

[2] P. Chen and L. M. Qin, "Deep learning based infrared image recognition of power equipment," *Journal of Shanghai Electric Power University*, vol. 37, no. 3, pp. 217–230, 2011.

[3] M. M. Lopez, R. U. Orozco, A. M. Sánchez, K. Picos and O. Ross, "Evaluation method of deep learning-based embedded systems for traffic sign detection," *IEEE Access*, vol. 9, pp. 101217–101238, 2021.

[4] B. M. Li, R. L. Jin and Z. F. Xu, "Improved ghost-YOLOv5 infrared target detection algorithm based on feature distillation," *Journal of Zhengzhou University*, vol. 43, no. 1, pp. 20–26, 2022.

[5] L. Shu, Z. J. Zhang and B. Lei, "Research on dense-yolov5 algorithm for infrared target detection," *Optical and Optoelectronic Technology*, vol. 19, no. 1, pp. 69–75, 2021.

[6] R. Z. Zhang, J. L. Zhang, X. P. Qi, H. R. Zuo and Z. Y. Xu, “Infrared target detection and recognition in complex scene,” *Opto-Electronic Engineering*, vol. 47, no. 10, pp. 200314-1–200314-10, 2020.

[7] A. Bochkovskiy, C. Y. Wang and H. Y. Liao, “YOLOv4: Optimal speed and accuracy of object detection,” arXiv e-prints, pp. 2004.10934, 2020.

[8] K. M. He, X. Y. Zhang, S. Q. Ren and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[9] N. Bodla, B. Singh, R. Chellappa and L. Davis, “Soft-NMS–improving object detection with one line of code,” in *Proc. ICCV*, Venice, Italy, pp. 5561–5569, 2017.

[10] S. Liu, L. Qi, H. F. Qin, J. P. Shi and J. Y. Jia, “Path aggregation network for instance segmentation,” in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 8759–8768, 2018.

[11] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. CVPR*, Las Vegas, NV, USA, pp. 779–788, 2016.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* “Ssd: Single shot multibox detector,” in *Proc. ECCV*, Amsterdam, Netherlands, pp. 21–37, 2016.

[13] H. Law and J. Deng, “Cornernet: Detecting objects as paired keypoints,” in *Proc. ECCV*, Munich, Germany, pp. 734–750, 2018.

[14] X. Y. Zhou, J. C. Zhou and P. Krahenbuhl, “Bottom-up object detection by grouping extreme and center points,” in *Proc. CVPR*, Long Beach, CA, USA, pp. 850–859, 2019.

[15] Z. Yang, S. H. Liu, H. Hu, L. W. Wang and S. Lin, “Reppoints: Point set representation for object detection,” in *Proc. ICCV*, Seoul, Korea, pp. 9657–9666, 2019.

[16] R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. CVPR*, Columbus, USA, pp. 580–587, 2014.

[17] R. Girshick, “Fast r-cnn,” in *Proc. ICCV*, Santiago, Chile, pp. 1440–1448, 2015.

[18] S. Q. Ren, K. M. He, R. Girshick and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[19] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” arXiv e-prints, pp. 1804.02767, 2018.

[20] K. X. Wang, J. H. Liew, Y. T. Zou, D. Q. Zhou and J. S. Feng, “Panet: Few-shot image semantic segmentation with prototype alignment,” in *Proc. ICCV*, Seoul, Korea, pp. 9197–9206, 2019.

[21] S. T. Liu, D. Huang and Y. H. Wang, “Learning spatial fusion for single-shot object detection,” arXiv e-prints, pp. 1911.09516, 2019.

[22] G. Ghiasi, T. Y. Lin and Q. V. Le, “Nas-fpn: Learning scalable feature pyramid architecture for object detection,” in *Proc. CVPR*, Long Beach, CA, USA, pp. 7036–7045, 2019.

[23] M. X. Tan, R. M. Pang and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *Proc. CVPR*, Seattle, WA, USA, pp. 10781–10790, 2020.

[24] Z. W. Xu, J. J. Zhuang, Q. Liu, J. K. Zhou and S. W. Peng, “Benchmarking a large-scale FIR dataset for on-road pedestrian detection,” *Infrared Physics & Technology*, vol. 96, pp. 199–208, 2019.

[25] B. Klare and S. Sarkar, “Background subtraction in varying illuminations using an ensemble based on an enlarged feature set,” in *Proc. CVPR*, Miami Beach, Florida, pp. 66–73, 2009.

[26] S. Hwang, J. Park, N. Kin, Y. Y. Choi and I. So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proc. CVPR*, Boston, MA, USA, pp. 1037–1045, 2015.

[27] F. A. Group, “Flir thermal dataset for algorithm training,” [Online]. Available: https://www.flir.com/oem/adas/adasdataset-form/

[28] H. Y. Zhang, M. Cisse, Y. N. Dauphin and P. D. Lopez, “Mixup: Beyond empirical risk minimization,” arXiv e-prints, pp. 1710.09412, 2017.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[30] H. Inoue, “Data augmentation by pairing samples for images classification,” arXiv e-prints, pp. 1801.02929, 2018.

[31] I. Goodfellow, A. J. Pouget, M. Mirza, B. Xu, F. D. Warde *et al.,* "Generative adversarial nets," arXiv e-prints, pp. 1406.2661, 2014.

[32] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan and Q. V. Le, "Autoaugment: Learning augmentation policies from data," arXiv e-prints, pp. 1805.09501, 2018.

[33] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. CVPR*, Salt Lake City, UT, USA, pp. 7132–7141, 2018.

[34] S. Woo, J. Park, J. Y. Lee and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, pp. 3–19, 2018.

[35] Z. H. Zheng, P. Wang, W. Liu, J. Li, R. G. Ye *et al.,* "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, New York, NY, USA, pp. 12993–13000, 2020.