technologies

MDPI

Article

An Investigation of Infrared Small Target Detection by Using the SPT–YOLO Technique

Yongjun Qi [1], Shaohua Yang [2], Zhengzheng Jia [1], Yuanmeng Song [1], Jie Zhu [1,*], Xin Liu [3] and Hongxing Zheng [1,*]

[1] School of Computer Science and Engineering of North China Institute of Aerospace Engineering, Langfang 065000, China; qyjun@189.cn (Y.Q.); jzz186@stumail.nciae.edu.cn (Z.J.); sym796@stumail.nciae.edu.cn (Y.S.)
[2] School of Aeronautics and Astronautics of North China Institute of Aerospace Engineering, Langfang 065000, China; ysh0130@stumail.nciae.edu.cn
[3] School of Optoelectronic Engineering, Xidian University, Xi'an 710071, China; xinliu@mail.xidian.edu.cn
* Correspondence: zhujie0424@126.com (J.Z.); hxzheng@hebut.edu.cn (H.Z.)

Abstract: To detect and recognize small-size and submerged complex background targets in infrared images, we combine a dynamic receptive field fusion strategy and a multi-scale feature fusion mechanism to improve the detection performance of small targets significantly. The space-to-depth convolution module is introduced as a downsampling layer in the backbone first and achieves the same sampling effect. More detailed information is retained at the same time. Thus, the model's detection capability for small targets has been enhanced. Then, the pyramid level 2 feature map with minimum receptive field and maximum resolution is added to the neck, which reduces the loss of positional information during feature sampling. Furthermore, x-small detection heads are added, the understanding of the overall characteristics and structure of the target is enhanced much more, and the representation and localization of small targets have been improved. Finally, the cross-entropy loss function in the original network model is replaced by an adaptive threshold focal loss function, forcing the model to allocate more attention to target features. The above methods are based on a public tool, the eighth version of You Only Look Once (YOLO) improved, it is named SPT–YOLO (SPDConv + P2 + Adaptive Threshold + YOLOV8s) in this paper. Some experiments on datasets such as infrared small object detection (IR-SOD) and infrared small target detection 1K(IRSTD-1K), etc. have been executed to verify the proposed algorithm; and the mean average precision of 94.0% and 69% under the condition of threshold at 0.5 and over a range from 0.5 to 0.95 is obtained, respectively. The results show that the proposed method achieves the best performance of infrared small target detection compared to existing methods.

Keywords: infrared small target detection; dynamic receptive field fusion; multi-scale feature fusion; space-to-depth convolution; YOLO

check for updates

Received: 21 October 2024
Revised: 23 December 2024
Accepted: 14 January 2025
Published: 17 January 2025

Citation: Qi, Y.; Yang, S.; Jia, Z.; Song, Y.; Zhu, J.; Liu, X.; Zheng, H. An Investigation of Infrared Small Target Detection by Using the SPT–YOLO Technique. *Technologies* **2025**, *13*, 40. https://doi.org/10.3390/technologies13010040

Copyright: © 2025 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

Infrared small target detection is a key technology for recognizing and detecting tiny targets in infrared images. Compared with visible light, infrared sensors have strong penetrating power, can resist external interference, and can work continuously under both day and night conditions. This technology is widely used in the military field, including nighttime ship inspection [1] and military early warning [2]. In the civilian field, it can also be used in application scenarios such as disaster monitoring [3], crop detection [4], and fault detection [5].

Currently, infrared small targets typically have very low contrast (less than 0.15), a low signal-to-noise ratio (less than 1.5), and occupy less than 0.15% of the image's total pixels [6]. These characteristics make the detection of infrared small targets a challenging problem. Currently, two primary detection methods are employed: detection-based tracking, which uses single-frame images followed by tracking, and tracking-based detection, which employs continuous frame images followed by detection [7]. However, due to rapid changes in both the target and background within real environments, especially during the rapid motion of the infrared sensor, the trajectory captured by tracking may deviate from the target's actual path. This deviation increases computational complexity and limits the practical application of this method [8]. Therefore, studying single-frame detection methods is particularly important.

Depending on the principles of the detection methods, single-frame infrared target detection approaches are primarily divided into three categories [9], filtering-based methods, local contrast-based methods, and data structure-based methods. Although filtering-based methods are widely applied [10–12], they are constrained by smooth and slowly changing backgrounds, lack robustness to variations in target size, and struggle with targets exhibiting low signal-to-noise ratios. Local contrast-based methods [13] depend on the significant difference between targets and the background but are prone to generating a False Positive when there is substantial interference in the background. Data structure-based methods, such as low-rank matrix recovery [14], dictionary learning [15], and tensor modeling [16], enhance detection performance but perform poorly in handling heterogeneous and dynamic backgrounds and are sensitive to noise.

Unlike traditional machine learning methods that require expert knowledge and experience to manually design feature extractors, deep-learning algorithms learn feature extraction automatically from datasets through end-to-end training [13]. Current deep-learning algorithms for target detection can be categorized into two-stage and single-stage algorithms. Two-stage algorithms, such as region-based convolutional neural network [17] and its faster [18], generate candidate regions before performing target classification and localization. Single-stage algorithms, including single-shot multibox detector [19] and You Only Look Once (YOLO) [20], predict the target class and the corresponding bounding box location directly from the image. Compared to two-stage algorithms, single-stage algorithms offer the advantages of simplicity, speed, and efficiency. However, one directly applying deep-learning methods to infrared small target detection still faces several challenges.

(i) The low resolution of infrared images and the small size of targets make it difficult to clearly distinguish and locate the targets within the images.

(ii) Repeated downsampling leads to the loss of small target information in deeper feature maps, significantly reducing the detection capability of the model.

(iii) There is a significant imbalance between targets and background in the images, causing the model to focus more on background features rather than target features.

To address the above problems, this paper is based on the eighth version of YOLO [21] improvement. Based on the characteristics of infrared small targets, we propose a new network model. The main works are as follows. First, the space-to-depth convolution (SPDConv) module replaces the maximum pooling and sampling layers [22], which improves the feature representation capability by capturing the global feature information, thus enhancing the detection of small targets. Then, the pyramid level 2 (P2) feature map increases with the minimum receptive field in the neck of the model to reduce the loss of position information during feature map downsampling; meanwhile, the x-small detection head [23] is added to further enhance the detection accuracy of small targets. Thirdly, the adaptive threshold focal loss function [24] is proposed to replace the cross-entropy loss function in the original network model. By dynamically adjusting the loss weights, it

enhances the model's focus on difficult-to-detect targets and significantly improves the overall detection performance. This proposed network model is named SPT–YOLO, where S refers to the SPDConv structure, P means the P2 feature map added to the feature pyramid network, and T is the adaptive threshold focal loss function. Finally, the self-collected infrared small target dataset, infrared small object detection (IR-SOD) covers diverse scenes and target classes, including cars and ships.

In summary, the contributions of this work can be outlined as follows.

(1) A non-strided convolution replaces the traditional max-pooling and strided convolution modules, enhancing the model's ability to capture complex details.

(2) The P2 feature map, which has the smallest receptive field, is integrated into the feature pyramid network, and an additional x-small detection head is introduced, effectively improving the detection performance for small targets.

(3) An adaptive threshold focal loss function replaces the cross-entropy loss function in the original model, dynamically adjusting weights to enhance focus on hard-to-detect targets, significantly improving overall performance. Furthermore, a self-collected infrared small object dataset, IR-SOD, is constructed, covering diverse scenes and target types, providing crucial support for model training and validation.

2. Fundamentals of IR Small Target Detection

Current infrared target detection methods can be broadly categorized into two groups. One category is traditional detection methods based on manually designed features, including background feature-based methods [25,26], which exploit the differences between the target and the background for detection, e.g., by modeling and analyzing the statistical properties of the background to identify the target and target feature-based methods [27,28], which focus on extracting the features of the target itself, such as shape, texture, and color. In addition, low-rank sparse decomposition-based methods [29,30] are also commonly used for infrared target detection, exploiting the low rank and sparsity of targets in infrared images. Another class of approaches is deep-learning-based methods that improve upon existing good target detection models such as the faster region-based convolutional neural network [18], single shot multibox detector [19], and YOLO [20]. These methods have achieved good detection performance in visible images and can be extended to infrared target detection tasks with proper tuning and optimization. In addition, some researchers have also utilized custom-designed models for infrared small target detection [31,32], which usually consider the special characteristics and requirements of infrared images to improve detection performance. In these methods, since the eighth version occurs, we think that the YOLO has greater potential for application for the following reasons.

In 2016, Redmon et al. [20] proposed an end-to-end target detection network called YOLO, an approach that has attracted a lot of attention. It significantly reduces the time required for target detection by integrating the target detection network into a single neural network and excels in accuracy. With the introduction of version one, the single-stage target detection method gained traction. Subsequently, after several upgrades, the network versions were released from 1.0 to 10.0 [33–36]. The continuous evolution of these versions combined several factors such as detection accuracy, detection speed, and network size, with the eighth version being more representative and widely popular. It is a lightweight deep-learning model consisting of the backbone, neck, and head networks, as shown in Figure 1.

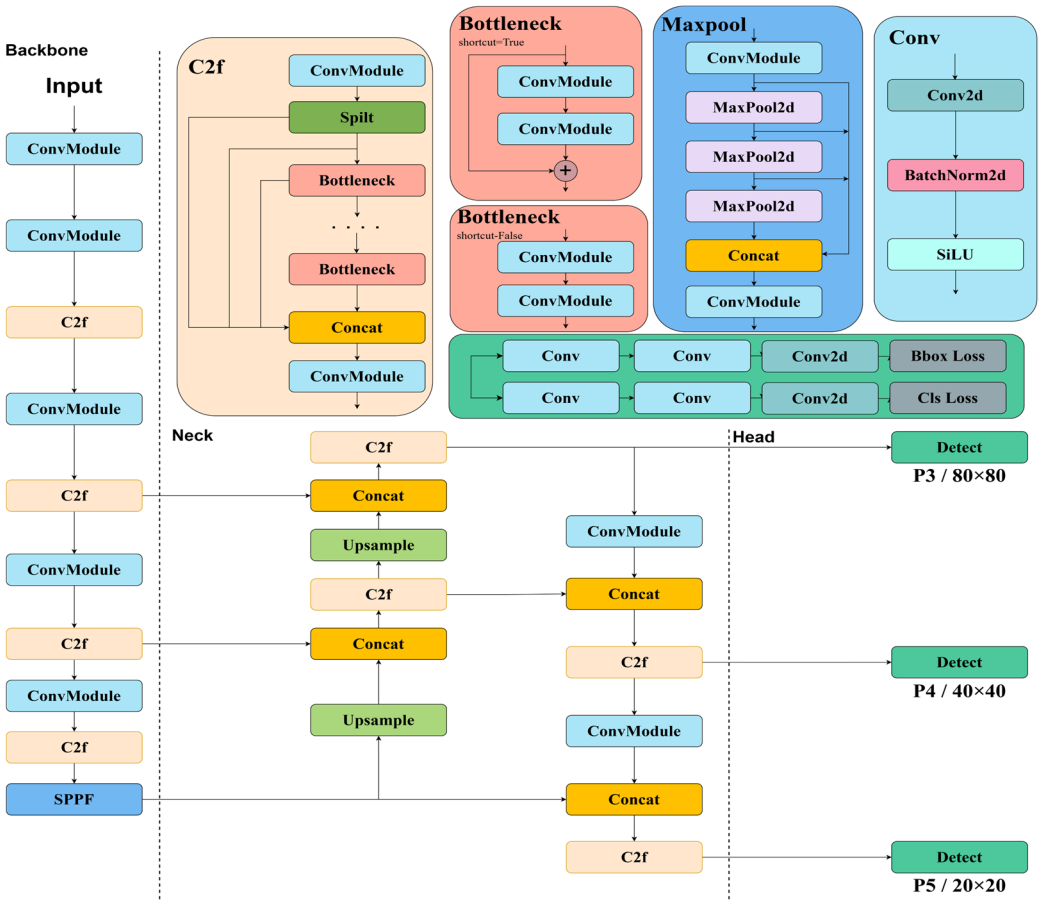

Figure 1. A lightweight deep-learning model consisting of the backbone, neck, and head networks, and the YOLOv8 network architecture depicted in detail.

The backbone network extracts features through Conv, C2f, and spatial pyramid pooling—fast (SPPF) modules, and enhances the model convergence through normalization operations to improve the target detection accuracy. The C2f module consists of two standard convolutional layers and multiple Bottleneck modules, which can efficiently fuse the features of different layers. The SPPF module, on the other hand, is used in YOLOv8 to quickly expand the sensory field by introducing the same size of the Maximum Pooling layer for parallel downsampling of the input feature map to quickly expand the sensory field, and unlike the previous version of the spatial pyramid pooling (SPP) module, SPPF further optimizes the computational efficiency while maintaining the contextual information and edge features. As a feature fusion network, the neck network completes the multi-scale fusion of images by combining the feature pyramid network and the path aggregation network. The complete intersection loss function and the non-maximal suppression algorithm are used in the prediction network for detection. The YOLOv8s' loss function consists of three parts, namely, the confidence loss, the classification loss, and the position loss, also as shown in Figure 1.

$$L = L_{conf} + L_{cls} + L_{box} \tag{1}$$

Loss of confidence,

$$L_{conf} = -\sum(\text{ylog}(\hat{y})) + (1-y)\text{log}(1-\hat{y}) \tag{2}$$

where y is the labeled value and $\hat{y}$ is the model predicted value. Classification loss,

$$L_{cls} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{3}$$

where y_i is the true label of class i and $\hat{y}_i$ is the class probability predicted by the model.

$$L_{box} = 1 - IoU + \frac{\rho^2\left(b, b^{g^t}\right)}{c^2} + \alpha v \tag{4}$$

where intersection over Union is the intersection and concurrency ratio of the predicted and real frames, ρ is the Euclidean distance between the centroids of the two, c is the diagonal length of the minimum envelope frame, and α and v are aspect ratio-related parameters.

$$\alpha = \frac{v}{1 - IoU + v} \tag{5}$$

$$v = \frac{4}{\pi^2}\left(arctan\frac{\omega_{g^t}}{h_{g^t}} - arctan\frac{\omega}{h}\right) \tag{6}$$

where ω_{g^t} and h_{g^t} are the width and height of the true box; ω and h are the width and height of the predicted bounding box, respectively.

3. Method of SPDConv, P2 and Adaptive Threshold

3.1. Architecture

Considering the accuracy requirements of the infrared small target detection task, the state-of-the-art one-stage deep-learning algorithm YOLOv8s is adopted, improved, and optimized. The basic framework of this network mainly includes the backbone network, the neck structure, and the multi-scale detection head. Aiming at the problem that traditional convolution is prone to lose detail information during multiple sampling, we introduce the SPDConv module in the backbone network to replace the traditional maximum pooling and downsampling layers, which combines a space-to-depth conversion layer and a non-strided convolution, which can reduce the spatial resolution while retaining more details and improve the feature expression ability. In the neck part of the model, based on the feature pyramid network (FPN) and path aggregation network (PAN) frameworks, the P2 feature map with minimal receptive fields is added and extended to four layers of feature output. The added x-small detection head, combined with the 160 × 160 high-resolution feature maps, effectively improves the detection accuracy of the model on small-sized targets by integrating shallow feature information. In addition, the traditional cross-entropy loss function is replaced with an adaptive threshold focal loss, which can dynamically adjust the loss weights according to the sample distribution, especially to give more attention to the difficult-to-detect small targets, thereby improving the detection performance of the model in complex scenes, and the SPT–YOLO is shown in Figure 2.

3.2. SPDConv Module for Downsampling

Detecting small targets is a challenging task because of their inherently low resolution and limited contextual information. In addition, infrared images suffer from low resolution and lack of color features. In the same infrared image, large targets tend to dominate, making it difficult for small targets to be effectively detected. A root cause of this problem lies in the design of a traditional strided convolution and pooling layer. In most research scenarios, images usually have good resolution and moderate target size, and the strided convolution and pooling layer can easily skip a large amount of redundant pixel infor-

mation, while still allowing the model to learn effective features. However, in infrared small target detection, where the target is small and the image is blurred, this assumption of skipping redundant pixel information is no longer valid, leading to loss of detailed information and poor feature learning.

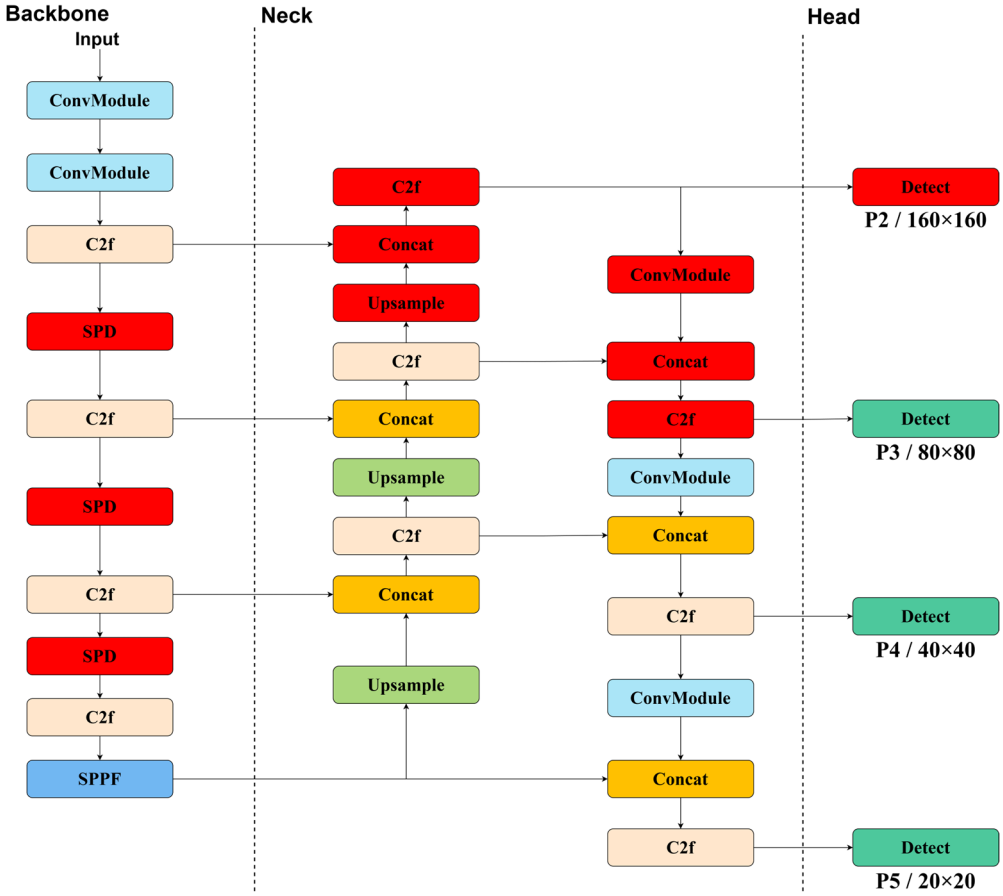

Figure 2. The architecture of the SPT–YOLO detection model, highlights the improvements made to the original YOLOv8s. The red boxes in the backbone network represent the integration of SPDConv (labeled as SPD). In the neck network, the red-highlighted section corresponds to the addition of the P2 feature layer. Similarly, in the detection head, the red box denotes the inclusion of an x-small detection head.

In order to solve the above problems, our proposal is to apply SPDConv as an alternative to the strided convolution and maximum pooling operations responsible for the downsampling layer in the YOLOv8 network structure, SPDConv consists of a space-to-depth (SPD) layer and a non-strided convolution layer. Specifically, the input feature map is initially transformed using the SPD layer, followed by a convolution operation using the non-strided convolution layer. This combination effectively reduces the spatial dimensions while preserving channel information without sacrificing detail. The conventional maximum pooling and strided convolution schematic is shown in Figure 3a, in which the feature map is divided into several equal-sized rectangular regions called pooling windows, and then the size and step size of the pooling windows are specified in the implementation process. If the downsampling multiplicity of the feature map is 2 and the pooling window is set to 2×2. Then, the shape of the feature map after downsampling is [H/2, W/2, C2]. In contrast, SPDConv adopts a more refined approach as shown in Figure 3b, with the same downsampling multiplicity, the SPDConv first divides the input feature map (X) into several smaller subregions. The splitting process can be expressed as follows.

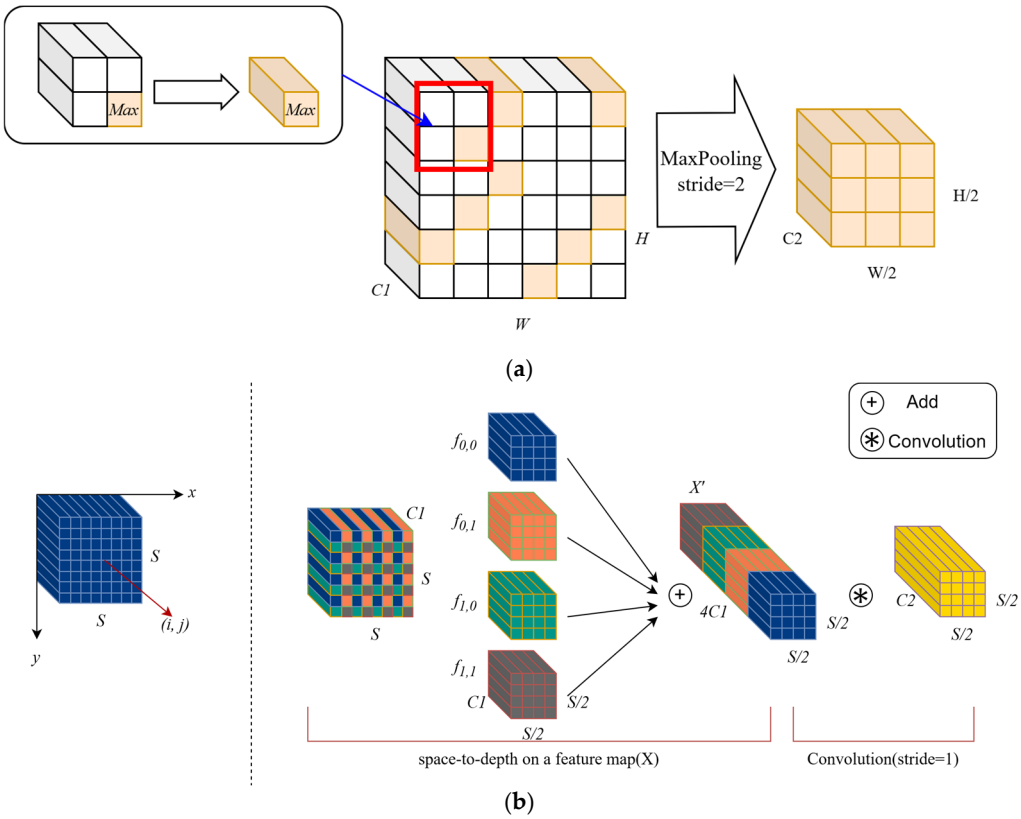

Figure 3. Conventional maximum pooling and strided convolution schematic; (**a**) several equal-sized rectangular regions called pooling windows; (**b**) SPDConv Module.

$$f_{0,0} = X[0:S:scale, 0:S:scale], f_{1,0} = X[1:S:scale, 0:S:scale], \ldots\ldots,$$

$$f_{scale-1,0} = X[scale-1:S:scale, 0:S:scale];$$

$$f_{0,1} = X[0:S:scale, 1:S:scale], f_{1,1} = X[1:S:scale, 1:S:scale], \ldots\ldots,$$

$$f_{scale-1,1} = X[scale-1:S:scale, 1:S:scale] \ldots$$

$$f_{0,scale-1} = X[0:S:scale, scale-1:S:scale], f_{1,scale-1} = X[1:S:scale, scale-1:S:scale], \ldots\ldots,$$

$$f_{scale-1,scale-1} = X[scale-1:S:scale, scale-1:S:scale] \quad (7)$$

Here, *scale* is a scaling factor that determines the downsampling rate applied to the input feature map X. The notation $f_{(i,j)}$ represents the sub-feature maps obtained through iterative operations, where i and j range from 0 to (*scale*-1). Each sub-feature map $f_{(i,j)}$ has a size of $S/scale \times S/scale \times C_1$, where S denotes the spatial dimensions of the input feature map, and C_1 represents the number of channels. For example, when *scale* = 2 (Figure 3b), the input feature map is divided into four sub-feature maps, each with a size of $S/2 \times S/2 \times C_1$. These sub-feature maps are then concatenated along the channel dimension to form a new feature map X', with dimensions $S/2 \times S/2 \times 4C_1$. Finally, a 1×1 convolution kernel of size C_2 is applied to the concatenated feature map X' to convert it into the desired feature map size.

SPDConv can retain more detailed information while achieving the same downsampling effect compared to traditional strided convolution and maximum pooling, because SPDConv splits the feature map into multiple sub-feature maps and splices between these sub-feature maps during the spatial-to-channel conversion process. Secondly, while traditional methods only focus on the information in the local area, SPDConv can consider the

information of the whole feature map at the same time, and therefore can better capture the contextual relationship, so that the model can more accurately understand the relationship between the target and the background.

3.3. P2 Feature Map for Enhanced Location Information

The task of detecting small targets is often challenging due to their low resolution and limited contextual information. To cope with this problem, the network architecture of YOLOv8 combines two different sampling approaches in its Neck module: FPN [37] and PAN [38] for effective fusion of multi-scale features. In this architecture, the Conv module is used for feature extraction, the Concat module and the upsampling operation are used to generate feature maps at different scales, while the C2f module further enhances the feature fusion capability.

In the original YOLOv8, the feature maps P3, P4, and P5 from the backbone network were input to the neck network for feature fusion, while the shallower feature map P2 (which has less semantic information) was not used. However, due to the small amount of information contained in the small targets, it is often difficult to retain the location information of the underlying feature maps after multiple downsampling. The P2 feature maps have the smallest sensory field and the highest resolution in the original image and thus are well suited for classifying weak and small infrared targets that lack feature information and are difficult to locate accurately.

Therefore, it is necessary to introduce the P2 feature map into the FPN and PAN. This introduction not only helps the PAN to transfer the underlying features and spatial location information from the T2 to the T5 feature map, which improves the richness of feature fusion but also adds a smaller detector head, so that the whole network forms a four-detector head structure, which is different from the three-detector head structure of the original YOLOv8. Figure 4 shows the feature fusion process after the introduction of P2. In the FPN, the feature map with the same resolution as P2 is called F2. In Figure 4, the Concat operation in the dashed box is to connect P2 with the input nodes of the FPN structure, and this operation plays a key role in the multi-scale feature fusion. The formula is as follows.

$$F2 = Concat(P2, Upsample(F3)) \tag{8}$$

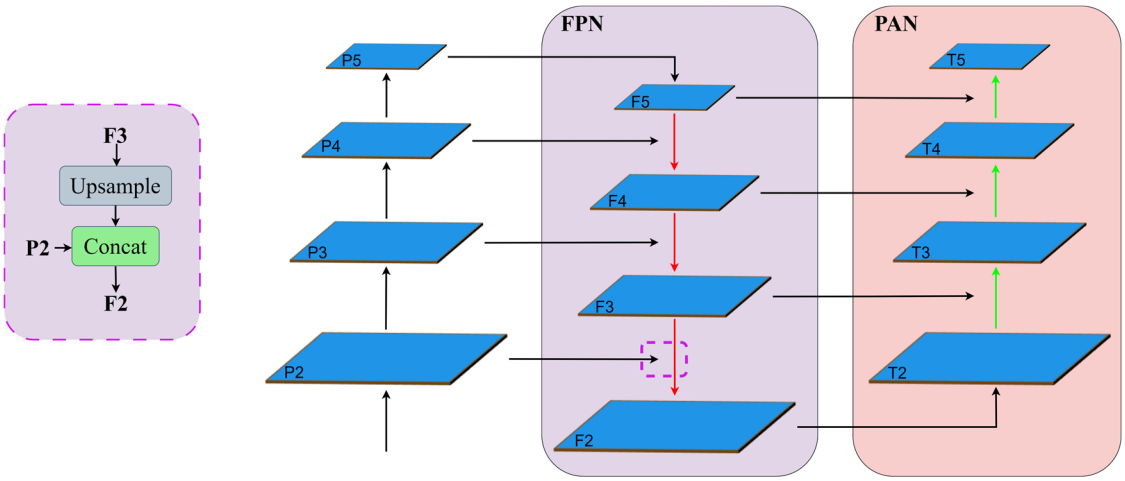

Figure 4. Feature fusion process of the network after adding P2. The process from F3 to F2 is detailed in the purple box on the left, which includes the upsampling and fusion modules. F3 is the first upsampling to match the size of P2, after which it is fused with the P2 feature map.

3.4. Adaptive Threshold Focal Loss for Enhanced Target Features

In infrared images, the background occupies most pixels, while the target comprises only a small portion, as shown in Figure 5. During training, the model finds it easier to learn background features than target features. Background samples can be regarded as "easy samples", while target samples are classified as "hard samples". However, even when background features are well learned, their numerical dominance still significantly impacts the loss function during training. As background samples dominate the gradient update direction, the importance of target information is diminished, severely limiting the model's performance in small target detection tasks.

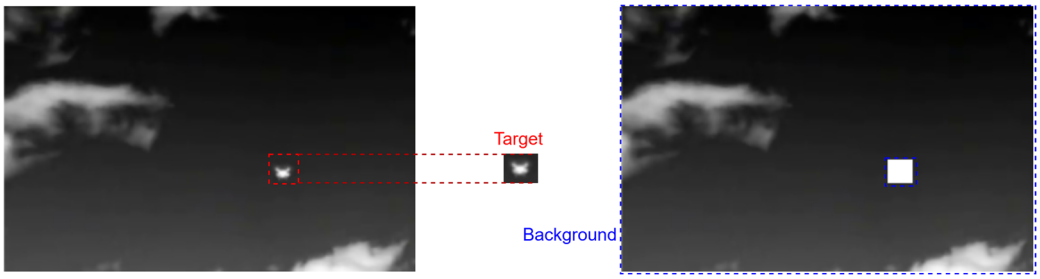

Figure 5. Imbalance between target and background. The left panel shows the original image with the target of interest highlighted within the red dashed border. The right panel shows the background of the image after target removal, with the green dashed border marking where the target was previously located.

The traditional binary cross entropy (BCE) loss function [39] assigns equal weights to all samples, making it ineffective in addressing the significant imbalance between targets and background. In infrared small target detection tasks, this design causes the loss computation to be dominated by the numerous background samples, leading the model to focus more on learning background features while neglecting target samples. This limitation is particularly pronounced in scenarios with small target sizes and low contrast.

To address this issue, we propose the use of an adaptive threshold focal loss (ATFL) to replace the traditional BCE loss function. First, this method sets dynamic thresholds to effectively distinguish between easily recognizable backgrounds and difficult-to-identify targets. Second, it increases the weight of losses associated with targets while reducing the influence of background-related losses, thereby encouraging the model to focus more on target features and mitigating the imbalance between targets and background. Finally, an adaptive hyperparameter design is introduced, enabling the loss function to dynamically adjust parameters, reducing the time cost associated with manual hyperparameter tuning.

The ATFL decouples targets and backgrounds based on a predefined threshold. By adaptively adjusting the loss value according to predicted probability, it aims to enhance the detection performance of infrared small targets.

The classical cross-entropy loss function can be expressed as

$$L_{BCE} = -(y\log(p) + (1-y)\log(1-p)) \tag{9}$$

where p denotes the predicted probability and y denotes the true label, which is succinctly represented as

$$L_{BCE} = -log(p_t) \tag{10}$$

where

$$p_t = \begin{cases} p, \; if \; y = 1 \\ 1-p, \; other \end{cases} \tag{11}$$

Cross-entropy loss cannot solve the imbalance between samples, so the focus loss function introduces a modulation factor $(1-p_t)^{\gamma}$, which reduces the loss of easy-to-classify samples by adjusting the focusing parameter γ. However, the modulation factor reduces the value of the loss of difficult samples while reducing the loss of easy samples, which is not conducive to the learning of difficult samples. Thus, the threshold focus loss function effectively mitigates the impact of easy samples by reducing the loss weight of easy samples while increasing the loss weight assigned to difficult samples. Specifically, samples with $p_t > 0.5$ are treated as "easy", while those with $p_t \leq 0.5$ are classified as "difficult". The expression is as follows.

$$TFL = \begin{cases} -(\lambda - p_t)^{\eta}\log(p_t), & p_t \leq 0.5 \\ -(1-p_t)^{\gamma}\log(p_t), & p_t > 0.5 \end{cases} \tag{12}$$

where $\eta, \gamma, \lambda(>1)$ are hyperparameters.

For different datasets and models, the hyperparameters need to be adjusted several times to achieve the best performance. Each training takes a lot of time, resulting in expensive time costs. Therefore, the hyperparameters are adapted so that for easy sampling, the loss value decreases with increasing p_t, which further reduces the loss from easy sampling. At the beginning of training, even simple samples will have relatively low predictive probabilities and gradually rise as the training progresses and γ should gradually approach 0. The predictive probability value of the true target can be used to mathematically model the progress of model training. The formula is as follows.

$$\hat{p}_c = 0.05 \times \frac{1}{t-1}\sum_{i=0}^{t-1}\overline{p_i} + 0.95 \times p_t \tag{13}$$

where $\hat{p}_c$ denotes the predicted value of the next iteration, and $\overline{p_i}$ denotes the current average predicted probability value. According to Shannon's information theory, the larger the probability value of an event, the smaller the amount of information it brings; and conversely, the larger the amount of information it brings. Therefore, the adaptive conditioning factor γ can be expressed as

$$\gamma = -\ln(\hat{p}_c) \tag{14}$$

However, in the later stages of training, too large a value of the expected probability reduces the proportion of difficult samples, so η is denoted as

$$\eta = -\ln(p_t) \tag{15}$$

By transforming the above Equations (14) and (15) into (12), the expression for the adaptive threshold focal loss function can be expressed as

$$ATFL = \begin{cases} -(\lambda - p_t)^{-\ln(p_t)}\log(p_t), & p_t \leq 0.5 \\ -(1-p_t)^{-\ln(\hat{p}_c)}\log(p_t), & p_t > 0.5 \end{cases} \tag{16}$$

4. Numerical Experiments

4.1. Datasets

(a) IR-SOD. For the small target detection task, we utilize the IR-SOD dataset, an infrared thermal imaging dataset collected using an unmanned aerial vehicle (UAV). This dataset comprises 4000 high-resolution images with dynamic and diverse scenarios, including urban, rural, maritime, and low-light conditions. The main object categories are

cars and ships, which are representative targets for real-world surveillance applications. To facilitate comprehensive model training and evaluation, the dataset is split into training (60%), validation (20%), and test (20%) sets. Unlike many existing datasets, IR-SOD is characterized by dynamic viewpoints due to UAV movement, along with significant background noise and complex scene variations caused by weather, lighting, and environmental factors. These challenges, including low contrast, cluttered backgrounds, and small target sizes, make IR-SOD a suitable benchmark for evaluating the robustness and adaptability of small target detection algorithms in real-world applications.

(b) IRSTD-1K. It focuses on infrared small target detection. The dataset contains 1000 infrared images of different scenes with a variety of image resolutions and is mainly used for testing and evaluating the performance of detection algorithms in dealing with infrared small targets with low contrast, low signal-to-noise ratio, and different background complexities [40].

(c) NUDT-SIRST (sea). It focuses on the detection of small infrared targets in the marine environment. The images in the dataset are derived from actual marine scenes, which contain small targets such as ships and marine outposts, and are complex and noisy. The dataset contains 48 spatial infrared images and 17,598 pixel-level annotations of small ships. Each image covers an area of about 10,000 square kilometers and has a resolution of 10,000 $\times$ 10,000 pixels [41].

4.2. Experimental Configuration

Ubuntu 20.04 was used as the operating system and Python 3.8, PyTorch 2.3.1, and CUDA 12.2 as the desktop computing software environment. The experiments were conducted using an NVIDIA GeForce RTX 3090 graphics card as hardware. The neural network implementation code was modified from Ultralytics version 8.0.105. For data augmentation, only the built-in methods provided by YOLO, such as random flipping, cropping, and scaling, were utilized. While these methods introduced variability in the training data, advanced augmentation strategies, including brightness adjustment, noise injection, and perspective transformations, were not explored, which could further enhance the model's robustness across diverse scenarios and mitigate overfitting, especially given the dataset's specific and limited nature. The hyperparameters used during training, testing, and validation of the experiments were kept the same. The number of training iterations was 200 epochs and the image input to the network was rescaled to 640 $\times$ 640. An SGD optimizer was used with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001.

4.3. Datasets Training

The original dataset and YOLOv8 were used to develop the infrared small target detection model, with Figure 6 illustrating its performance on the training and validation sets by tracking the changes in three types of loss functions: bounding box regression loss, classification loss, and distribution focal loss. The bounding box regression loss quantifies the discrepancy between predicted and ground truth bounding boxes, where lower values indicate more precise predictions. Classification loss reflects the accuracy of category predictions, with lower values signifying better performance. Distribution focal loss measures prediction confidence, helping the model optimize the detection process. During the initial iterations from 0 to 100, the loss functions exhibited significant variations, while from 100 to 200 iterations, the model demonstrated continuous performance improvements and eventual stabilization. Importantly, the trends in loss variations for the training and validation sets remained consistent, with no notable performance gaps or signs of

overfitting, demonstrating that the proposed model effectively generalizes despite the limited dataset.

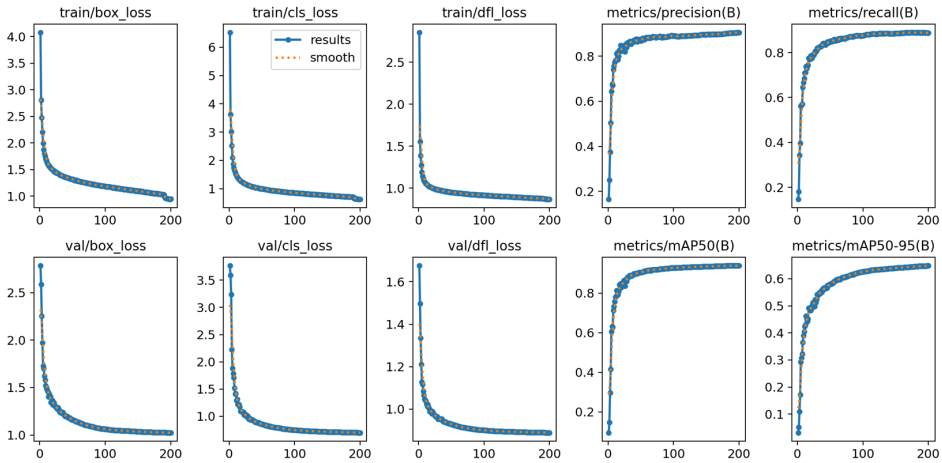

Figure 6. Visualization of model evaluation metrics during training (precision, recall, and mAP@0.5).

4.4. Evaluation Metrics

To evaluate the performance of SPT–YOLO, its detection accuracy is measured based on the mean average precision (mAP) across all categories. The mAP combines precision and recall functions as shown in the following:

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

where TP, FP, and FN denote true positive, false positive, and false negative, respectively. The precision–recall curve (PRC) can be plotted based on recall and precision, and the average precision is the area under the PRC. The average precision of a single category and the average precision of all categories are shown as follows:

$$AP = \int_0^1 P(R)dR \tag{19}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP \tag{20}$$

where the P denotes precision, and R denotes recall, and N is the number of categories. F1 is the harmonic mean of precision and recall, providing a balanced measure of model performance. The calculation formula is as follows:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{21}$$

4.5. Comparison with Other Methods

In order to confirm the effectiveness of SPT–YOLO, the proposed method is compared with mainstream methods such as YOLOv5, YOLOv8, YOLOv9, YOLOv10, YOLO11 [42], DNANet, AGPCNet [43] on self-researched datasets and two publicly available datasets, respectively. SPT–YOLO achieved mAP0.5 of 94.3%, 91.6%, and 88.2% on the IR-SOD, IRSTD-1K dataset, and the NUDT-SIRST (sea) dataset, respectively. mAP0.5:0.95 reached 69%, 65%, and 60%, respectively. Tables 1–3 illustrate the comparison of the test results on the three datasets.

Table 1. Comparison of test results on the IR-SOD dataset.

Detection Method	mAP0.5	mAP0.5:0.95	Precision	Recall	F1
YOLOv5	0.80	0.51	0.85	0.68	0.76
YOLOv8	0.82	0.57	0.86	0.81	0.84
YOLOv9	0.80	0.52	0.85	0.80	0.83
YOLOv10	0.81	0.54	0.82	0.88	0.83
YOLO11	0.81	0.53	0.83	0.84	0.76
DNANet	0.84	0. 59	0.88	0.84	0.84
AGPCNet	0.82	0. 42	0.82	0.88	0.83
SPT–YOLO	**0.94**	**0.69**	**0.96**	**0.93**	**0.94**

Table 2. Comparison of test results on the IRSTD-1K dataset.

Detection Method	mAP0.5	mAP0.5:0.95	Precision	Recall	F1
YOLOv5	0.79	0.51	0.86	0.72	0.79
YOLOv8	0.80	0.39	0.82	0.75	0.79
YOLOv9	0.82	0.39	0.84	0.75	0.49
YOLOv10	0.81	0.39	0.81	0.74	0.77
YOLO11	0.81	0.39	0.85	0.74	0.80
DNANet	0.87	0.46	0.91	0.89	0.90
AGPCNet	0.37	0.48	0.37	0.68	0.48
SPT–YOLO	**0.91**	**0.65**	**0.88**	**0.86**	**0.87**

Table 3. Comparison of test results on the NUDT-ISRST (sea) dataset.

Detection Method	mAP0.5	mAP0.5:0.95	Precision	Recall	F1
YOLOv5	0.34	0.18	0.64	0.27	0.39
YOLOv8	0.34	0.19	0.65	0.27	0.39
YOLOv9	0.35	0.19	0.67	0.28	0.40
YOLOv10	0.33	0.19	0.63	0.27	0.38
YOLO11	0.31	0.17	0.63	0.26	0.39
DNANet	0.76	0. 57	0.76	0.72	0.74
AGPCNet	0.41	0. 24	0.51	0.47	0.44
SPT–YOLO	**0.88**	**0.60**	**0.87**	**0.82**	**0.84**

The methods used for comparison include both general target detection methods for deep-learning and the new methods proposed in the last two years for infrared small target detection tasks. The experimental results show that on all three datasets, SPT–YOLO has better infrared detection performance in both the detection paradigm-based and segmentation paradigm-based approaches. Specifically, on the three datasets, SPT–YOLO achieves mAP0.5 of 0.94, 0.91, and 0.88 in the metrics that best reflect the model's integrated detection capability, compared to the methods used for comparison. As the IoU threshold increases, SPT–YOLO still achieves better performance in the methods used for comparison. This means that SPT–YOLO can still maintain high accuracy and robustness when dealing with more stringent detection criteria.

In order to further validate the performance enhancement of the proposed network model in infrared small target detection, comparative experiments are conducted in this study on Precision-Recall (PR) curves as shown in Figure 7. The PR curves can demonstrate the trade-off between target detection accuracy and the recall of the model under different decision thresholds. The experimental results show that the improved network model SPT–YOLO exhibits better performance on the PR curve compared with the current mainstream methods, which further confirms the effectiveness of the present method. Through meticulous comparative experiments and in-depth analysis of PR curves, this

study demonstrates that the proposed SPT–YOLO algorithm has significant advantages in the task of infrared small target detection.

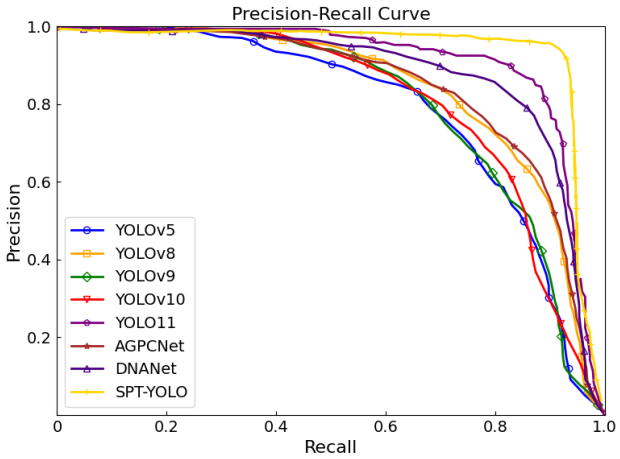

Figure 7. Comparison of PR of different methods on IR-SOD dataset.

4.6. Visualization

In contrast, Figures 8–10 show the YOLOv5, YOLOv8, YOLOv9, YOLOv10, and SPT–YOLO. Images were selected from the IR-SOD dataset, and in this experiment, target categories and objective confidence levels were labeled with detection boxes.

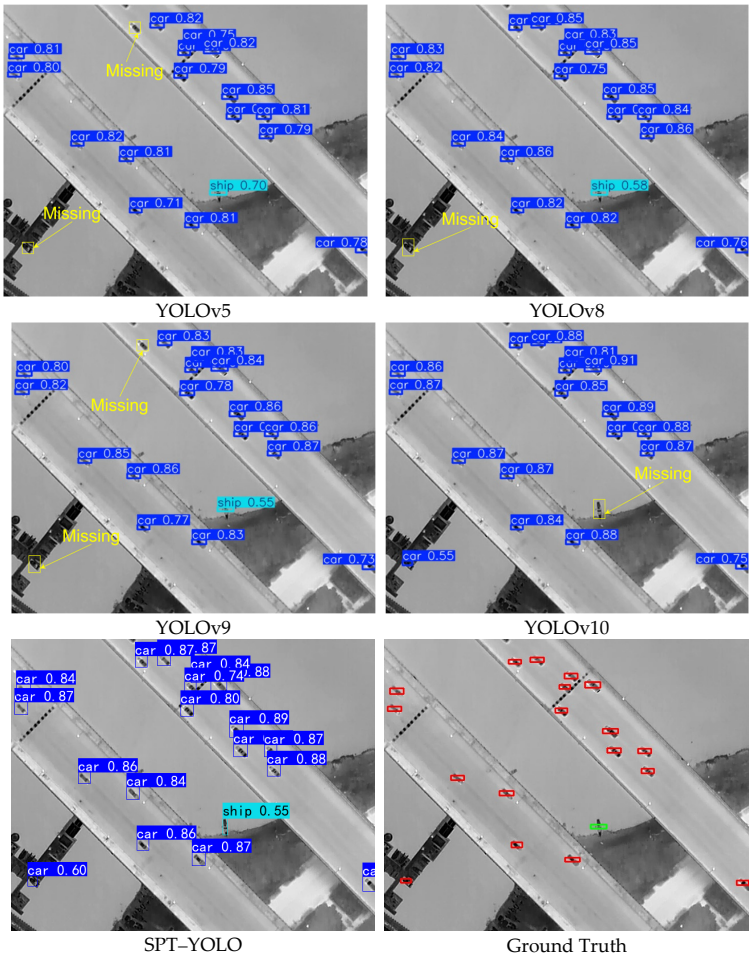

Figure 8. Comparison of multiple target detection in a scene, where the blue bounding boxes represent cars, and the indigo boxes represent ships. In the ground truth annotations, green boxes denote cars and blue boxes denote ships, while the yellow bounding boxes highlight the missed detections.

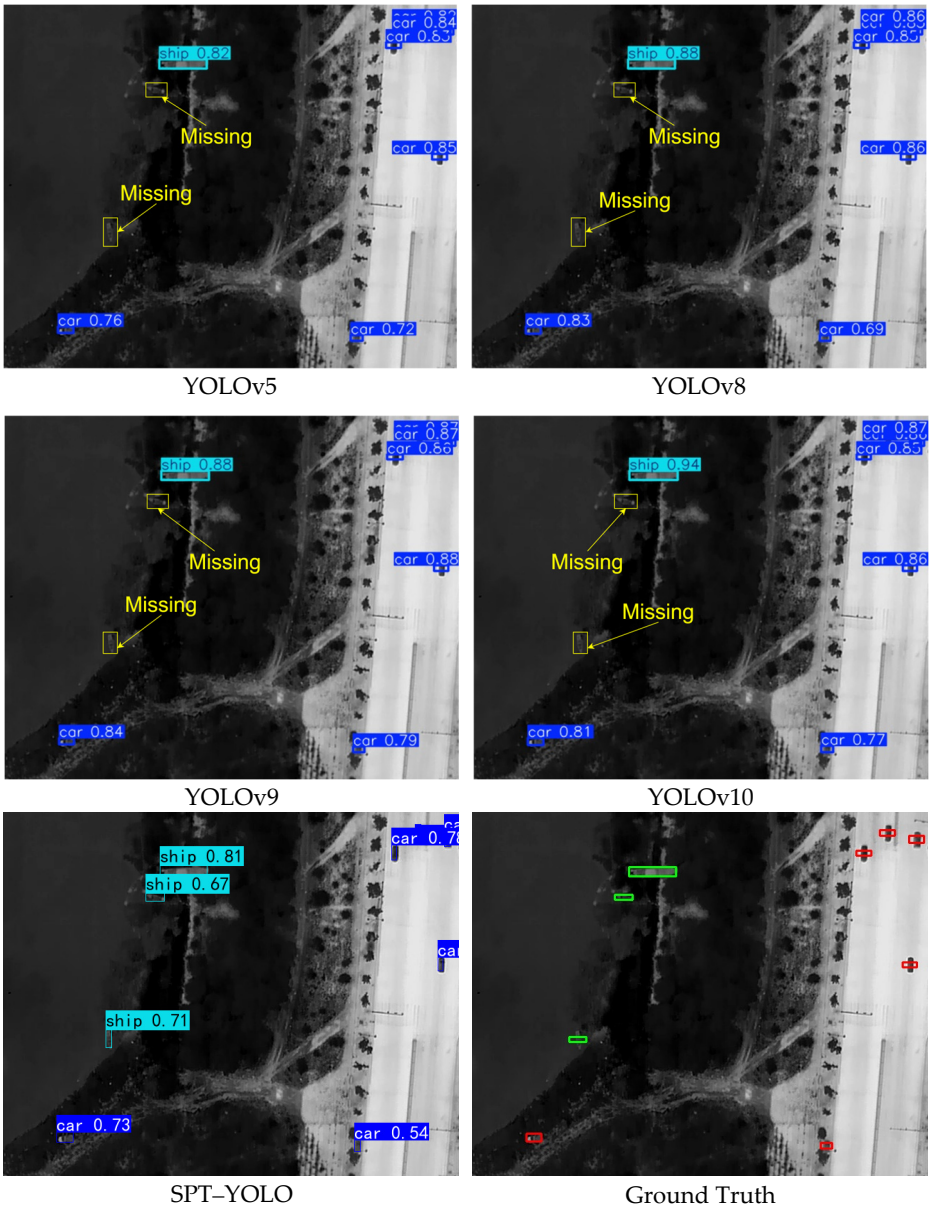

Figure 9. Presentation of a comparative analysis of various models in complex scenes.

We evaluated the performance of SPT–YOLO on the IR-SOD dataset in three challenging scenarios: high-density target crowding, complex background, and sparse target scenes. The results showed that SPT–YOLO exhibited consistent detection performance in all scenarios. As shown in Figure 8, in the high-density target crowding scene, YOLOv5, YOLOv8, YOLOv9, and YOLOv10 all exhibited varying degrees of missed detections, with some models misidentifying targets as background. In contrast, SPT–YOLO successfully detected all targets, with significantly higher IoU values for the detection results. In the complex background scene (Figure 9), YOLOv5, YOLOv8, YOLOv9, and YOLOv10 still faced significant challenges in accurately identifying ships, especially for those ships with complex backgrounds or unclear targets, often leading to false detections. SPT–YOLO demonstrated its superior performance in complex scenes by successfully detecting these ambiguous targets. Furthermore, in the sparse target scene (Figure 10), SPT–YOLO not only accurately detected all targets but also provided higher IoU values, demonstrating its precision in target localization.

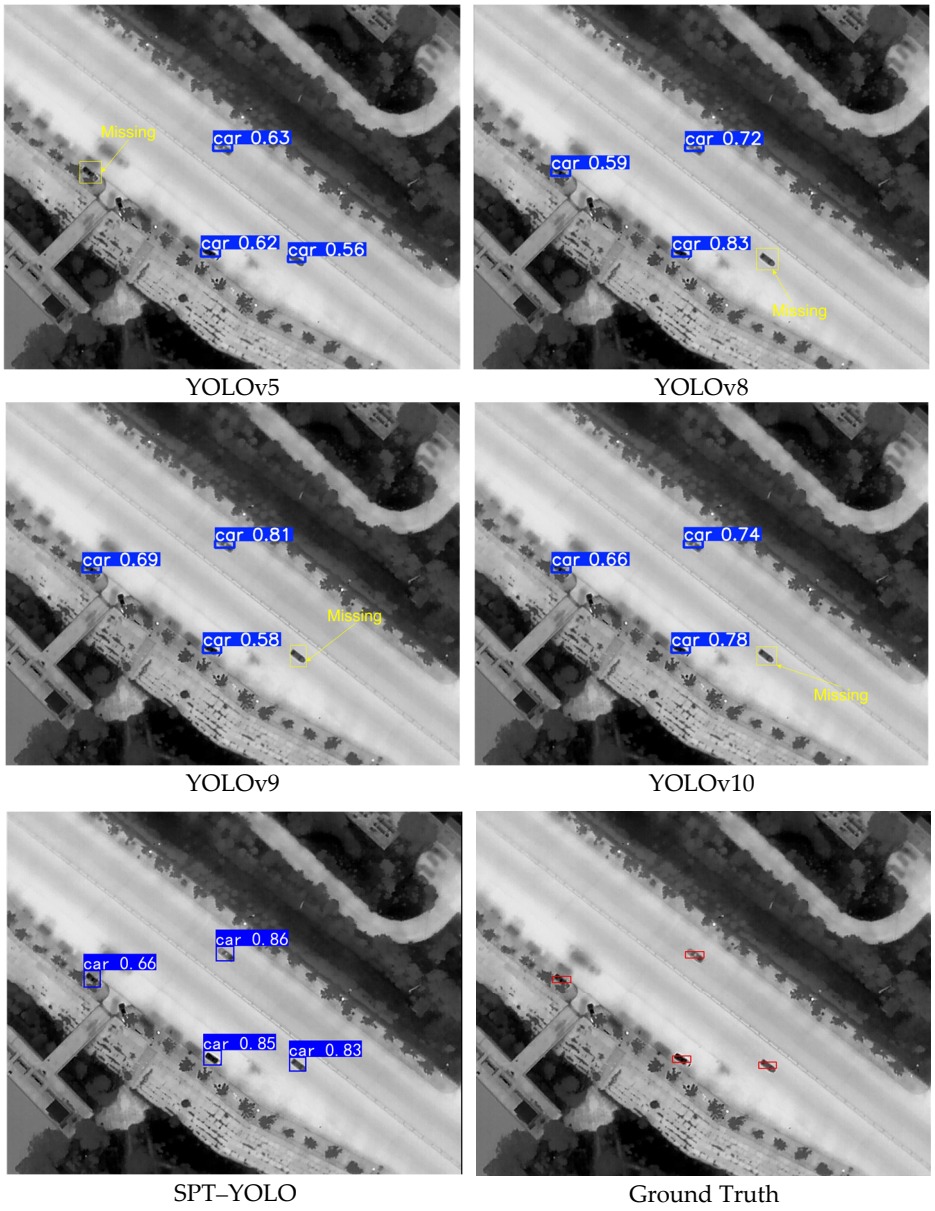

Figure 10. Illustration of a comparative analysis of target detection in scenes with few targets, where the purple bounding boxes indicate missed detections.

4.7. Ablation Study

To validate the effectiveness of the proposed SPT–YOLO model, we conducted a series of ablation experiments on the IR-SOD dataset to evaluate the impact of each module on SPT–YOLO. Table 4 shows the results of the SPT–YOLO ablation experiments, where SPD represents SPDConv, P2 + x-small represents the P2 feature layer and the x-small detection head, and ATFL represents the adaptive threshold focus loss function. As shown in Table 4, compared to the baseline model, the other three improved models showed improved detection performance, with the hybrid model showing the largest gain. Compared to the original YOLOv8s, the mAP0.5 increased by 4.88%, and mAP0.5:0.95 increased by 5.26%. Precision, Recall, and F1 scores improved by 11.63%, 14.81%, and 11.90%, respectively. The results indicate that these improvements complement each other in the detection task, enabling the model to achieve the best detection performance. Each added improvement module increases the model's mAP, Precision, and Recall. The individual ablation experiments for each improvement step will be analyzed in the following sections.

Table 4. Results of the SPT–YOLO ablation experiment performed on the IR-SOD dataset.

YOLOv8s	SPD	P2 + x-Small	ATFL	mAP0.5	mAP0.5:0.95	Precision	Recall	F1
√	×	×	×	0.82	0.57	0.86	0.81	0.84
√	√	×	×	0.85(3.66%↑)	0.60(5.26%↑)	0.88(2.32%↑)	0.82(1.23%↑)	0.87(3.57%↑)
√	√	√	×	0.90(6.10%↑)	0.66(10.52%↑)	0.91(5.81%↑)	0.81	0.88(4.76%↑)
√	√	√	√	0.94(4.88%↑)	0.69(5.26%↑)	0.96(11.63%↑)	0.93(14.81%↑)	0.94(11.90%↑)

(a) SPDConv. To validate the effectiveness of the SPDConv module in retaining details during the downsampling process of infrared small target detection, we conducted ablation experiments on the IR-SOD dataset using the SPT–YOLO model as the baseline. The verification results are shown in Table 5. When using the SPDConv module, mAP0.5 and mAP0.5:0.95 improved by 3% and 4%, respectively.

Table 5. Performance of different loss functions compared between Conv and SPDConv.

Conv Module	mAP0.5	mAP0.5:0.95	Precision	Recall
Conv	0.91	0.65	0.92	0.90
SPDConv	**0.94**	**0.69**	**0.96**	**0.93**

SPT–YOLO outputs the shallowest feature map from the third convolution module for subsequent feature extraction, which contains relatively complete image features suitable for visual observation and comparison. To visually analyze the improvement of SPDConv over the original convolution module in feature extraction, we performed detection on the same image before and after replacing the convolution modules in the backbone network with SPDConv and directly compared the third convolution module output feature maps of both networks. Six adjacent, identical channels from the third convolution output feature maps of both networks were selected for observation and comparison, as shown in Figure 11.

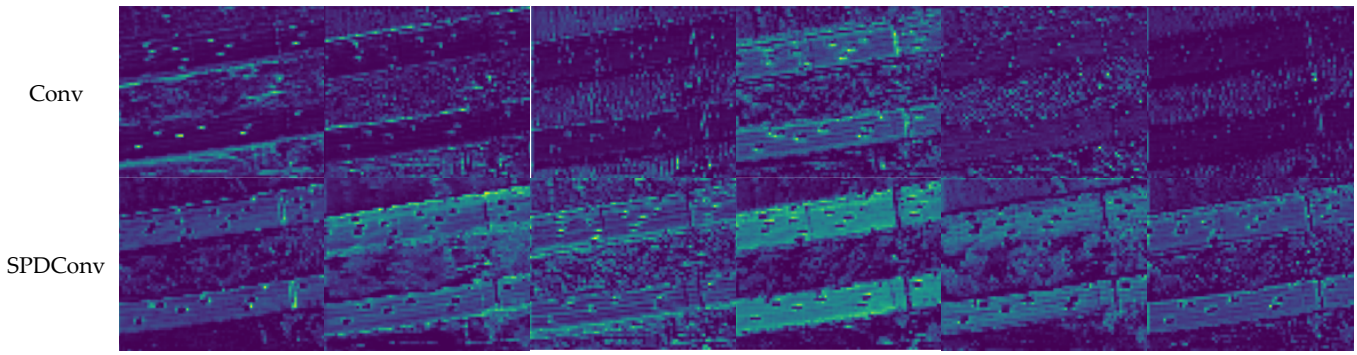

Figure 11. Comparison of feature maps for the output by the third convolutional module of the network.

From Figure 11, we can clearly see that when using the Conv module, the extracted infrared small target information is mixed with background information, appearing darker, which makes it easier to misidentify background information as targets. However, after using the SPDConv module, the extracted small target features are clearer and brighter, almost completely separated from the background. This reflects that replacing the Conv module in the backbone network with the SPDConv module is beneficial for infrared small target detection.

(b) P2 + 4Head. To evaluate the effectiveness of the P2 feature layer and x-small detection head in infrared small target detection, we conducted ablation experiments on

the IR-SOD dataset using the SPT–YOLO model as the baseline. The verification results are shown in Table 6. To visually analyze the improvement of the P2 feature layer and x-small detection head over the original model in infrared small target detection, we added the P2 feature layer and x-small detection head to the original network model and performed detection on the same image before and after the addition. The output results were compared using heatmaps. As shown in Figure 12, comparing the results before and after adding the P2 feature layer and x-small detection head, we can clearly see that after adding these components, every target was accurately detected, with no missed detections as seen before the improvements. This enhancement enables the model to focus better on infrared small targets, improving its accuracy and making it more suitable for infrared small target detection.

Table 6. Performance of different loss functions compared between 3 head and P2 + x-small head.

Number of Detection Heads	mAP0.5	mAP0.5:0.95	Precision	Recall
3Head	0.88	0.62	0.88	0.90
P2 + x-small Detection Head	**0.94**	**0.69**	**0.96**	**0.93**

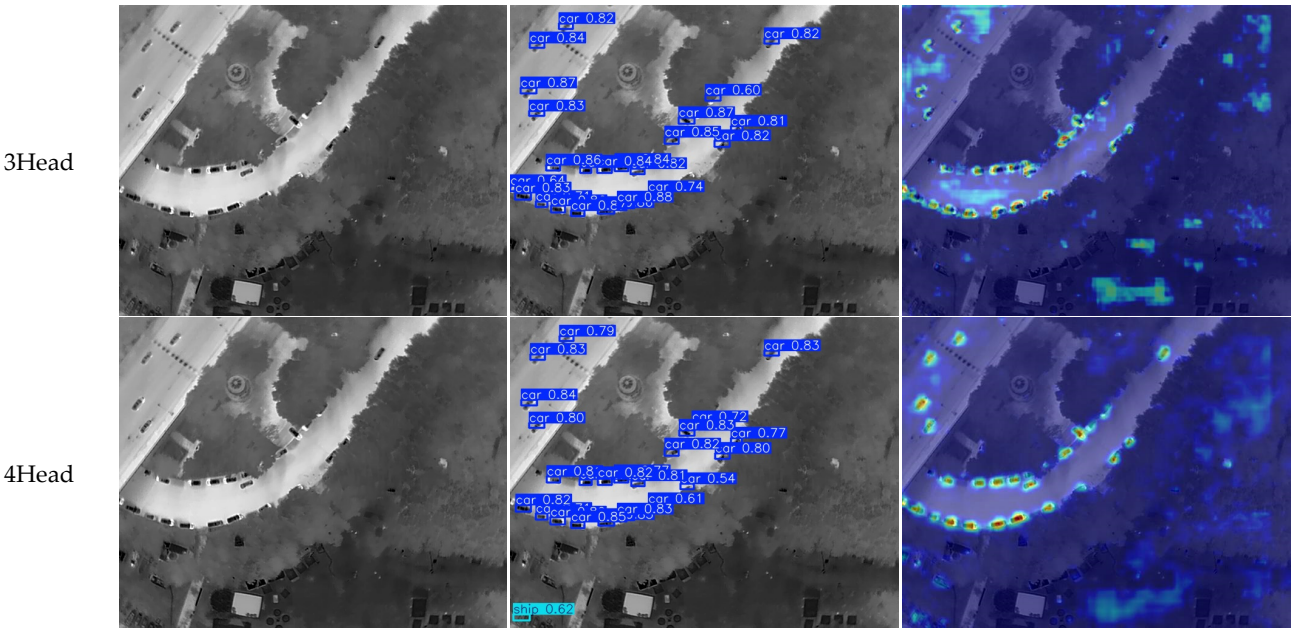

Figure 12. Comparison of detection performance with 3 and 4 detection heads.

(c) Adaptive Threshold Focus Loss. To explore the impact of different loss functions on detection performance, we conducted a comparative experiment on the IR-SOD dataset. The verification results are shown in Table 7. The adaptive threshold focus loss significantly outperformed the other four loss functions and effectively addressed the class imbalance issue in infrared images. mAP0.5 and mAP0.5:0.95 reached 0.94 and 0.69, respectively, which is 1% and 5% higher than the performance of the well-performing VariFocal Loss. This improvement effectively solves the class imbalance problem in infrared images, enabling the model to allocate more weight to difficult samples, thus improving the model's accuracy and making it more suitable for infrared small target detection.

Table 7. Performance of different loss functions for comparison of 5 cases.

Loss Function	mAP0.5	mAP0.5:0.95	Precision	Recall	F1
Slide Loss	0.91	0.60	0.87	0.74	0.80
Focal Loss	0.92	0.63	0.90	0.85	0.87
VariFocal Loss	0.93	0.64	0.90	0.86	0.88
Quality Focal Loss	0.92	0.62	0.88	0.86	0.87
ATFL	**0.94**	**0.69**	**0.96**	**0.93**	**0.94**

5. Conclusions

An improved model is proposed for infrared small target detection. The model introduces SPDConv instead of a partial downsampling layer in the backbone network, which can retain more detailed information while achieving the same downsampling effect. In order to enhance the detection of small targets, a P2 feature layer is added to the neck network of the model, and an additional x-small detection head is added to the detection head part. This series of improvements significantly enhance the model's detection performance for small-sized targets. In addition, the model introduces an adaptive threshold focal loss function to replace the traditional cross-entropy loss, which effectively solves the problem of category imbalance in infrared images. Compared with the baseline method and other mainstream detection algorithms, this model achieves a comprehensive improvement in detection accuracy.

Despite the significant progress achieved by the proposed model in infrared small target detection, there are certain limitations that need to be addressed. First, the model's performance is highly dependent on the quality and diversity of the training data. In scenarios where the training data fails to cover a wide range of conditions, the model's generalization ability may degrade in unseen environments. Second, the computational complexity of the model poses a challenge for deployment in resource-constrained settings, such as edge devices or systems with limited hardware capabilities.

Future research will focus on addressing the limitations identified in this study through the following key areas. First, more efficient network architectures will be explored to reduce the computational burden of the model, making it suitable for deployment on resource-constrained devices without compromising detection performance. Second, efforts will be directed towards expanding the dataset to include more diverse scenarios, such as varying environmental conditions, complex backgrounds, and a wider range of target types, to enhance the model's generalization ability. Finally, the model's practical applicability will be evaluated through real-world implementations in systems such as autonomous surveillance and drone-based monitoring, ensuring its robustness and effectiveness in operational settings.

Author Contributions: Conceptualization, Y.Q. and S.Y.; methodology, Y.Q. and S.Y.; validation, Z.J., Y.S. and J.Z.; formal analysis, S.Y.; investigation, S.Y.; resources, Y.Q.; data curation, Z.J. and Y.S.; writing—original draft preparation, Y.Q. and S.Y.; writing—review and editing, H.Z. and X.L.; visualization, S.Y.; supervision, H.Z. and X.L.; project administration, Y.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Hebei Province Central Guiding Local Science and Technology Development Fund Project under grant number 236Z4901G and the Science and Technology Innovation Program of Hebei, China (Grant No. SJMYF2022X18). Additionally, this research was funded by three projects from the North China Institute of Aerospace Engineering: Grant Nos. YKY-2024-70.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

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

References

1. Cascelli, G.; Guaragnella, C.; Nutricato, R.; Tijani, K.; Morea, A.; Ricciardi, N.; Nitti, D.O. Use of a residual neural network to demonstrate feasibility of ship detection based on synthetic aperture radar raw data. *Technologies* **2023**, *11*, 178. [CrossRef]
2. Bertrand, N.P.; Lee, J.; Prussing, K.F.; Shapero, S.; Rozell, C.J. Infrared search and track with unbalanced optimal transport dynamics regularization. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 2072–2076. [CrossRef]
3. Hewarathna, A.I.; Hamlin, L.; Charles, J.; Vigneshwaran, P.; George, R.; Thuseethan, S.; Wimalasooriya, C.; Shanmugam, B. Change detection for forest ecosystems using remote sensing images with Siamese attention U-Net. *Technologies* **2024**, *12*, 160. [CrossRef]
4. Sanida, M.V.; Sanida, T.; Sideris, A.; Dasygenis, M. An efficient hybrid CNN classification model for tomato crop disease. *Technologies* **2023**, *11*, 10. [CrossRef]
5. Polymeropoulos, I.; Bezyrgiannidis, S.; Vrochidou, E.; Papakostas, G.A. Enhancing solar plant efficiency: A review of vision-based monitoring and fault detection techniques. *Technologies* **2024**, *12*, 175. [CrossRef]
6. Chapple, P.B.; Bertilone, D.C.; Caprari, R.S.; Angeli, S.; Newsam, G.N. Target detection in infrared and SAR terrain images using a non-Gaussian stochastic model. In *Targets and Backgrounds: Characterization and Representation V*; SPIE: Bellingham, WA, USA, 1999.
7. Marvasti, F.S.; Mosavi, M.R.; Nasiri, M. Flying small target detection in IR images based on adaptive toggle operator. *IET Comput. Vis.* **2018**, *12*, 527–534. [CrossRef]
8. Bi, Y.; Bai, X.; Jin, T.; Guo, S. Multiple feature analysis for infrared small target detection. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1333–1337. [CrossRef]
9. Zhao, M.; Li, W.; Lu, L.; Hu, J.; Ma, P.; Tao, R. Single-frame infrared small-target detection: A survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 87–119. [CrossRef]
10. Lu, D.; Wang, M.; Yang, X.; Teng, L.; Tan, J.; Tian, Z.; Gu, G. A small target detection method for sea surface based on guided filtering and local mean gray difference. *J. Comput. Commun.* **2023**, *11*, 49–63. [CrossRef]
11. Deshpande, S.D.; Er, M.H.; Venkateswarlu, R.; Chan, P. Max-mean and max-median filters for detection of small targets. In *Signal and Data Processing of Small Targets*; SPIE: Bellingham, WA, USA, 1999; Volume 3809, pp. 74–83.
12. Bae, T.W.; Sohng, K.I. Small target detection using bilateral filter based on edge component. *J. Infrared Millim. Terahertz Waves* **2010**, *31*, 735–743. [CrossRef]
13. Chen, C.; Li, H.; Wei, Y.; Xia, T.; Tang, Y.Y. A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 574–581. [CrossRef]
14. Peng, Y.; Suo, J.; Dai, Q.; Wang, X. Reweighted low-rank matrix recovery and its application in image restoration. *IEEE Trans. Cybern.* **2014**, *44*, 2418–2430. [CrossRef] [PubMed]
15. Li, Z.-Z.; Chen, J.; Hou, Q.; Fu, H.-X.; Dai, Z.; Jin, G.; Li, R.-Z.; Liu, C.-J. Sparse representation for infrared dim target detection via a discriminative over-complete dictionary learned online. *Sensors* **2014**, *14*, 9451–9470. [CrossRef] [PubMed]
16. Sun, Y.K.; Yang, J.; Long, Y.; Shang, Z.; An, W. Infrared patch-tensor model with weighted tensor nuclear norm for small target detection in a single frame. *IEEE Access* **2018**, *6*, 76140–76152. [CrossRef]
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
18. 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]
19. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016.
20. Redmon, J. You Only Look Once: Unified, Real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Hussain, M. YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature Toward Digital Manufacturing and Industrial Defect Detection. *Machines* **2023**, *7*, 677. [CrossRef]
22. Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A new CNN building block for low-resolution images and small objects. In Proceedings of the ECML PKDD 2022, Grenoble, France, 19–23 September 2022.
23. Li, H.; Qu, H. DASSF: Dynamic-Attention Scale-Sequence Fusion for Aerial Object Detection. *arXiv* **2024**, arXiv:2406.12285.
24. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing feature learning network for infrared small target detection. *IEEE Trans. Geosci Remote Sens.* **2024**, *62*, 5906511. [CrossRef]

25. Bai, X.; Zhou, F. Analysis of new top-hat transformation and the application for infrared dim small target detection. *Pattern Recognit.* **2010**, *43*, 2145–2156. [CrossRef]
26. Bai, X.; Zhou, F. Hit-Or-Miss Transform based infrared dim small target enhancement. *Opt. Laser Technol.* **2011**, *43*, 1084–1090. [CrossRef]
27. Han, J.; Moradi, S.; Iman, F.; Liu, C.; Zhang, H.; Zhao, Q. A local contrast method for infrared small-target detection utilizing a tri-layer window. *IEEE Trans. Geosci. Remote Sens.* **2020**, *17*, 1822–1826. [CrossRef]
28. Han, J.; Moradi, S.; Iman, F.; Zhang, H.; Zhao, Q.; Zhang, X.; Li, N. Infrared small target detection based on the weighted strengthened local contrast measure. *IEEE Trans. Geosci. Remote Sens.* **2021**, *18*, 1670–1674. [CrossRef]
29. Kong, X.; Yang, C.; Cao, S.; Li, C.; Peng, Z. Infrared small target detection via nonconvex tensor fibered rank approximation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5000321. [CrossRef]
30. Zhang, P.; Zhang, L.; Wang, X.; Shen, F.; Pu, T.; Fei, C. Edge and corner awareness-based spatial–temporal tensor model for infrared small-target detection. *IEEE Trans Geosci. Remote Sens.* **2021**, *59*, 10708–10724. [CrossRef]
31. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Trans. Image Process.* **2023**, *32*, 1745–1758. [CrossRef]
32. Wang, K.; Du, S.; Liu, C.; Cao, Z. Interior attention-aware network for infrared small target detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5002013. [CrossRef]
33. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
34. Jocher, G. YOLOv5 by Ultralytics. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 28 February 2023).
35. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision (ECCV) 2024, Milan, Italy, 29 September–4 October 2024.
36. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
37. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
38. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
39. Mao, A.; Mohri, M.; Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
40. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Pro-ceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA, 22 June 2022.
41. Wu, T.; Li, B.; Luo, Y.; Wang, Y.; Xiao, C.; Liu, T.; Yang, J.; An, W.; Guo, Y. MTU-Net: Multilevel transUNet for space-based infrared tiny ship detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5601015. [CrossRef]
42. Jocher, G.; Qiu, J. Ultralytics YOLO11. 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 1 October 2024).
43. Zhang, T.; Cao, S.; Pu, T.; Peng, Z. AGPCNet: Attention-guided pyramid context networks for infrared small target detection. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 4250–4261. [CrossRef]

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