


Article

IRSDet: Infrared Small-Object Detection Network Based on Sparse-Skip Connection and Guide Maps

Xiaoli Xi ^{1,2} , Jinxin Wang ^{1,2}, Fang Li ^{1,2} and Dongmei Li ^{1,2,*}

¹ Laboratory of Optoelectronic System, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China; xixiaoli@semi.ac.cn (X.X.); wangjx@semi.ac.cn (J.W.); lifang@semi.ac.cn (F.L.)

² College of Materials Science and Opto-Electronic Technology, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: lidongmei@semi.ac.cn

Abstract: Detecting small objects in infrared images remains a challenge because most of them lack shape and texture. In this study, we proposed an infrared small-object detection method to improve the capacity for detecting thermal objects in complex scenarios. First, a sparse-skip connection block is proposed to enhance the response of small infrared objects and suppress the background response. This block is used to construct the detection model backbone. Second, a region attention module is designed to emphasize the features of infrared small objects and suppress background regions. Finally, a batch-averaged biased classification loss function is designed to improve the accuracy of the detection model. The experimental results show that the proposed small-object detection framework significantly increases precision, recall, and F1-score, showing that, compared with the current advanced detection models for small-object detection, the proposed detection framework has better performance in infrared small-object detection under complex backgrounds. The insights gained from this study may provide new ideas for infrared small object detection and tracking.

Keywords: infrared image; small object; object detection; SSD



Citation: Xi, X.; Wang, J.; Li, F.; Li, D.

IRSDet: Infrared Small-Object Detection Network Based on Sparse-Skip Connection and Guide Maps. *Electronics* **2022**, *11*, 2154. <https://doi.org/10.3390/electronics11142154>

Academic Editor: Dah-Jye Lee

Received: 9 June 2022

Accepted: 8 July 2022

Published: 9 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the development of infrared image-sensor technology, infrared spectral imaging technology has provided new information for object-detection tasks [1,2]. Currently, the object detection method based on infrared images is one of the best methods for detecting remote thermal objects because the infrared features of objects are more noticeable than their visible features [3]. In remote detection tasks, most infrared objects are considered small objects because of fewer pixels, a lower signal-to-clutter ratio (SCR), unclear contours, and sparse texture features. Because of these characteristics, infrared small-object detection remains a significant challenge.

Convolutional neural networks (CNNs) provide a broader perspective on object detection. Compared to traditional methods, CNN-based object detection methods can adaptively learn object locations and semantic information in sample images, resulting in higher accuracy and robustness. Object detection models based on CNN include two- and one-stage models. The former are not suitable for high real-time detection because of the slow inference speed that divides positioning and classification into two steps, such as RCNN [4]. The latter, such as YOLO [5] and SSD [6], have a fast inference speed and good accuracy.

Some optimized CNN-based models have a good detection capacity for small objects. ResNet [7], DenseNet [8], and ResNext [9] propose shortcut connections that can transfer information by skipping one or more layers to address the degradation problem. This is helpful in reducing the feature loss of small objects during information transmission. In DetNet [10], downsampling blocks in deep layers are eliminated to preserve the resolution of high-level feature maps, which can improve the positioning accuracy of small objects. DSSD [11], RSSD [12], and FSSD [13] propose specific multiscale feature fusion methods to

suppress the static noise in low-level feature maps. In RFBNet [14], multiple branches with different kernels and dilated convolution layers are concatenated to expand the receptive field and enhance the deep features of lightweight CNNs. Extensive studies have shown that the above methods can improve the accuracy of small-object detection but still do not achieve satisfactory results. One of the most important factors is that the methods do not optimize the model structure specifically for the characteristics of small objects, such as size and texture.

Many researchers have been inspired by small-object detection methods and have proposed detection models suitable for small objects. The optimized methods of these models can be categorized into spatial-temporal information fusion [15–17], residual/background information prediction [18,19], optimized region proposal [20,21], and multiscale information fusion [22–25]. The spatial-temporal information fusion method reduces static noise by combining adjacent frames in an infrared image sequence. The residual/background information prediction method is an indirect method that first predicts the background information and then subtracts it from the original image to obtain the object's position. Traditional methods or CNN-based methods are used in optimized region proposal methods to filter the potential region of the object. Subsequently, a classifier is designed to process the potential region image.

In summary, these studies support the notion that there are many essential differences between visible and infrared objects, such as the number of image channels, image signal-to-noise ratio, and the number of hard negative samples. Therefore, infrared small-object detection methods are different from visible small-object detection algorithms; the former focuses on reducing false alarms, while the latter aims to reduce misdetections.

An infrared small detection framework called IRSDet is proposed to address these issues. The main contributions of this study are as follows.

1. A sparse-skip connection module is proposed to construct the backbone that can reduce the feature loss of infrared small objects in information transmission.
2. A feature map enhancement method based on the region attention mechanism is proposed to reduce background noise interference and emphasize the objects' potential region.
3. A batch-averaged biased classification loss method under limited memory usage is proposed to alleviate the drastic fluctuation of classification loss under the small-batch configuration and avoid the gradient explosion of the focal loss function in the initial training process.

Experimental results showed that the proposed method has high precision and recall. The insights gained from this study may provide new ideas for infrared small object detection and tracking.

2. Related Works

2.1. Small-Object Detection Methods Based on CNN

In recent years, the optimization methods of CNN-based small-object detection models have been divided into the following aspects:

Receptive field and attention mechanism: Sun et al. [26] proposed a mask-guided SSD. The method enhances features with contextual information and introduces segmentation masks to eliminate the background regions. However, segmentation masks, including the object region, require pre-labeling. FD-SSD [27] adopts deformable convolutional layers that can optimize the position of the receptive field to better adapt to the geometric and shape changes of small objects, but they increase the computational cost. Lim et al. [28] proposed FA-SSD, which uses a residual attention module and context information to enhance the feature representation of low-level feature maps, thus improving the accuracy of small-object detection.

Multiscale information fusion: Cui et al. [29] proposed a multiscale deconvolutional SSD (MDSSD). The method can simultaneously upsample high-level feature maps of different layers and fuse them with non-adjacent low-level feature maps to form a

clearer feature for small objects. Zhai et al. [30] proposed a DF-SSD that designed a backbone network based on dense connections. To enhance the representation of features in the output image, adjacent feature maps are fused to supplement semantic information and details. Although dense connections can suppress feature loss, they retain a large amount of static background noise, which is a severe problem for infrared small-object detection. Pan et al. [31] proposed a top-down feature fusion module that iteratively fuses high-level features containing semantic information with low-level features containing boundary information.

Additionally, data augmentation methods were considered to preprocess small-object samples to improve the training effect of the detection models. Kisantal M et al. [32] proposed a sample replication method to increase the number of small objects in each image to address the issue of a small number of positively matched anchors. Bai Y et al. [33] proposed a super-resolution small-object generation method, SOD-MTgan, which can upsample small, blurred objects to recover more details.

2.2. Infrared Small-Object Detection Methods Based on CNN

Spatial-temporal information fusion: Park et al. [15] proposed an infrared small-object detection method for pedestrian image sequences that manually introduces spatial-temporal information and potential object regions. To avoid position errors caused by residual and mask images, adjacent similar pixels are merged into a single object using the connecting component algorithm. To eliminate the influence of static noise in an infrared image sequence on the detection of small infrared objects, Yao et al. [16] proposed an optimized FCOS network model that uses traditional filtering methods and spatial-temporal feature fusion to preprocess sample images. Du et al. [17] proposed an interframe energy accumulation (IFEA) enhancement mechanism to effectively extract spatial-temporal information in the infrared sequence. The method is specially designed to suppress strong spatially nonstationary clutter, enhance the object, and improve accuracy.

Residual/Background information: Shi et al. [18] proposed a convolutional and denoising autoencoder network (CDAE) that uses residual images as output images. Additionally, perceptual loss is employed to solve the problem of background texture feature loss in the encoding process, and structural loss is proposed to compensate for the perceptual loss defect in which small objects appear. This method was supported by Fang et al. [19], who stated that too many details are lost during the pooling operation in the downsampling of the encoding process; thus, it is difficult to reconstruct the high-frequency details well in the decoding stage. To address this issue, they proposed a multiscale U-Net. The constructed image-to-image network integrates the global and local dilated residual convolution blocks into the U-Net, predicting the residual information between the input and output images for small infrared UAV object detection.

Optimized region proposal: Fan et al. [20] proposed an infrared small-object detection method based on region proposal and a CNN module to separate real objects from the background and significantly reduce the false alarm rate caused by complex background clutter. First, the small-object intensity is enhanced based on the local intensity characteristics. Potential object regions are then proposed by corner detection to ensure a high detection rate of the method. The approach used in the research by Ren et al. [21] is similar to that described above. They designed a simple structured region context network (RCN) to extract possible regions. Then, an optimized GAN network is used to process region images to generate super-resolution results with more detailed features.

Multiscale information fusion: The downsampling of CNN-based models may cause information loss, decreasing the accuracy of detecting small infrared objects. Ding et al. [22] and Du et al. [23] used high-resolution low-level images as feature maps to address this issue. Moreover, multilevel feature-fusion methods are used to suppress false alarms in low-level feature maps. Ju et al. [24] went one step further. They used an hourglass image-filtering module to obtain a fusion image to substitute the original input image, aiming to enhance the response of small infrared objects and suppress the

background response. However, this module directly processes the original image without distinguishing background noise from objects. Hou et al. [25] adopted a more efficient method for replacing the input image with a fusion image. The framework in their research used parallel convolutional layers to extract the contrast information of small objects and neighborhoods. The kernels of the parallel convolutional layers have different sizes for extracting different-scale spatial information.

Training strategies: Some studies have optimized training strategies for small-object detection. For instance, Du et al. [23] specially designed an IOU threshold and anchor size for small objects. Bai et al. [34] proposed a regular constraint loss (RCL) to restrict multiscale feature fusion learning and obtain more accurate object location information.

3. Proposed Method

Small infrared objects exhibit three characteristics. First, the feature of small objects is apt to lose in information transfer. Second, limited by the performance of current infrared sensors, the signal-to-noise ratio of infrared images is low, and there are numerous false alarms. Finally, most infrared objects lack texture and detailed features.

This study, therefore, proposes an infrared small-object detection model (IRSDet) to address these issues. The structure of IRSDet is shown in Figure 1. The SSD is used as the detection head in the proposed method. In this section, we first discuss the structure and function of the backbone and then describe a feature enhancement method based on the proposed guide block. Finally, we describe the batch-averaged biased classification loss function.

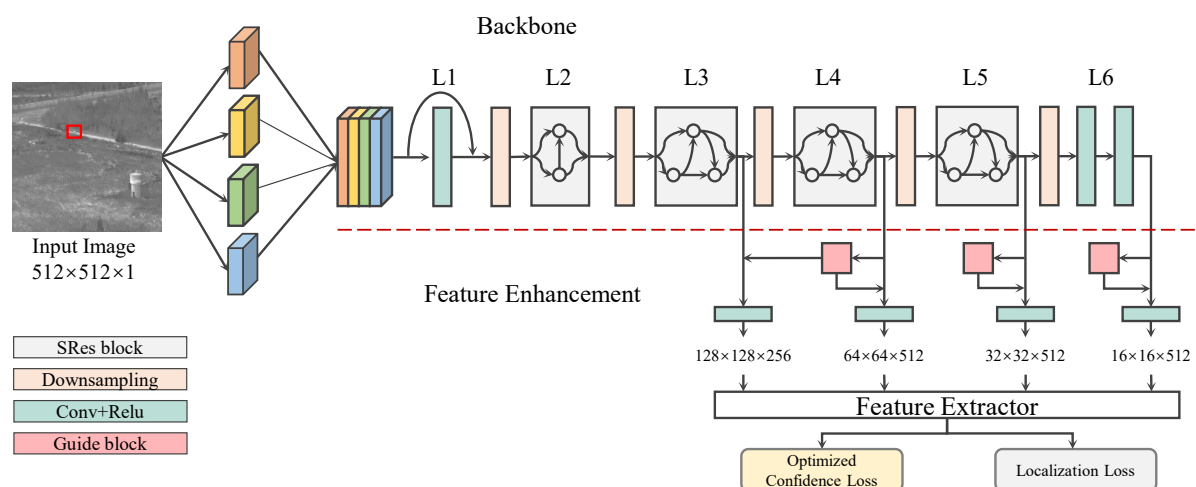


Figure 1. Proposed infrared small-object detection framework. The backbone has 16 weight layers, and L2–L5 layers have 2, 3, 3, 3 SRes blocks. The first layer of the backbone is a parallel convolutional layer with different atrous rates.

3.1. IRS16: Backbone of IRSDet

3.1.1. Sparse Residual Block

Small infrared objects are more challenging to spot than small visible objects. First, there is only one information channel in infrared images, and many false alarms have features similar to real objects. Second, serial convolutional layers can enhance the information extraction of the surrounding receptive field. The output signal of one of the serial convolutional layers cannot sufficiently extract the features of the real objects. Consequently, the subsequent convolutional layers cannot obtain accurate information. Bias accumulates over multiple convolutional layers, resulting in a severe loss of object features in deep layers.

A sparse residual block (SRes block) is proposed to address these issues. The SRes block is an alternative parallel structure that can transmit signals on parallel branches. The

signals of the two branches are combined after each convolutional layer. The structure of an SRes block is shown in Figure 2.

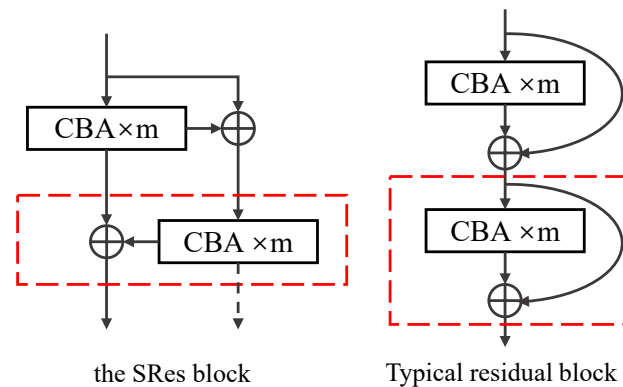


Figure 2. Proposed SRes block and typical residual block. CBA = convolutional layer + batch normalization + activation function.

The n -th $CBA \times m$ can be defined as function $F_n()$, and the output is X_n . Therefore, the output of the SRes block is:

$$X_n = F_n(X_{n-1}) + F_{n-1}(X_{n-2}) \quad (1)$$

The output of the typical residual block is:

$$X_n = F_n(X_{n-1}) + F_{n-1}(X_{n-2}) + X_{n-2} = \sum_{i=1}^n F_i(X_{n-i}) + X_0 \quad (2)$$

The output of the SRes block is related only to the outputs of the two adjacent convolutional layers. However, the output of typical residual blocks is related to all the previous layers. We consider that the excessive use of low-level information brings much background noise and, therefore, reduces the detection accuracy.

3.1.2. Adaptive Receptive Field Block

An adaptive receptive field (ARF) block is designed as the first convolutional layer to better adapt to the size change in the objects. The ARF block adopts parallel convolution kernels with different atrous rates (Figure 3a). These convolution kernels can collect features from regions of different sizes in an input image. They are then concatenated to create a large kernel with sparse receptive fields (Figure 3b). Thus, more features can be extracted to adapt to the geometric and shape changes of an object. We added a 1×1 kernel convolutional layer in the parallel stage to counteract the overlap in the center of the large convolution kernels. Convolutional layers with different atrous rates can adjust the weights of different regions.

ReLU can suppress negative signal transmission, resulting in the feature loss of small objects. Therefore, an extra branch is appended to transmit negative signals. The positive and negative signals are concatenated and then transmitted to a 1×1 convolutional layer with two functions: (1) to optimize the internal weight allocation of the sparse large convolution kernel to increase its sensitivity to object features; and (2) to reduce the dimensions of the output maps and suppress less-valuable channels.

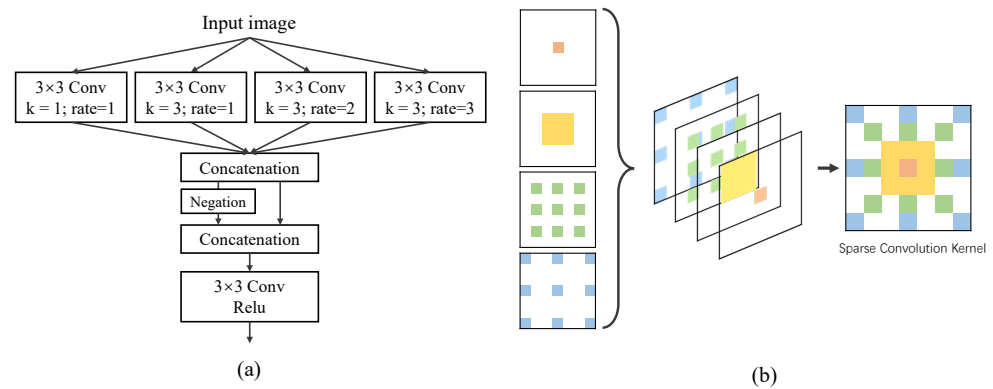


Figure 3. (a) shows the ARF block. (b) shows the sparse convolutional kernel. The convolutional layers in the parallel stage have different atrous rates.

3.1.3. Extremum Pooling

Strided convolution and pooling methods are currently the most popular downsampling methods. However, with the strided convolution and Avg-pooling, it is hard to avoid decreasing the contrast between small objects and neighborhoods. Conversely, Max-pooling can adaptively select the maximum grayscale from the region and exhibits outstanding performance in transferring semantic information. However, a specific drawback associated with Max-pooling is that it might ignore the critical details of small objects in shallow convolutional layers. The results of the downsampling are shown in Figure 4a.

We propose an optimized downsampling method called extreme pooling (Ext-pooling) to address this issue. Ext-pooling (Figure 4b) has two branches that can simultaneously transfer the local maximum and minimum to the next layer. The output $y_{\frac{m}{s} \times \frac{n}{s}}^d$ of Ext-pooling can be expressed as

$$y_{\frac{m}{s} \times \frac{n}{s}}^d = \text{Ext-pooling}(x_{m \times n}^d) \quad (3)$$

where $x_{m \times n}^d$ is an input feature map; s is the stride of the Ext-pooling layer.

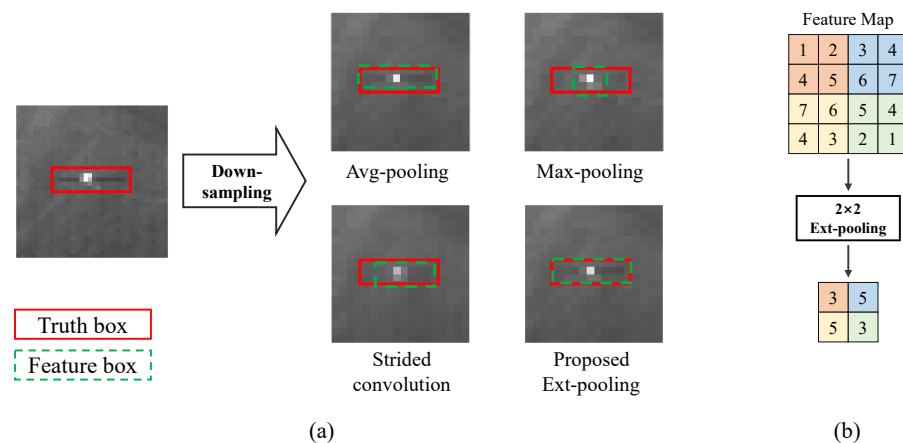


Figure 4. (a) shows the comparison of different downsampling methods. (b) shows an example of Ext-pooling, $\text{Ext-pooling}(1, 2, 4, 5) = \frac{1}{2}[(\max(1, 2, 4, 5) + \min(1, 2, 4, 5))]$.

3.2. Feature Enhancement Based on Attention Mechanism

We utilized low-level images as feature maps to improve the recall of small objects. However, low-level feature maps have undesirable noise because of the complex clutter background. Multilevel feature fusion methods were used in [22,23] to suppress background noise; however, computational costs were proportionally increased. To reduce the

interference of false alarms and noise, this study proposes a region attention mechanism block, namely, the guide block.

The structure of the guide block is shown in Figure 5. Max-pooling and Avg-pooling were used to process the branch feature maps. The former was used to recover the contour of an object's potential region, which may lose information transmission, and the latter was used to suppress noise and smoothen the image. Two processed images were then combined by multiplication. Finally, a CBA module was appended to eliminate redundant information from the image, thereby generating a guide map. Potential object regions have high weights in the guide map, whereas the background region has weak weights.

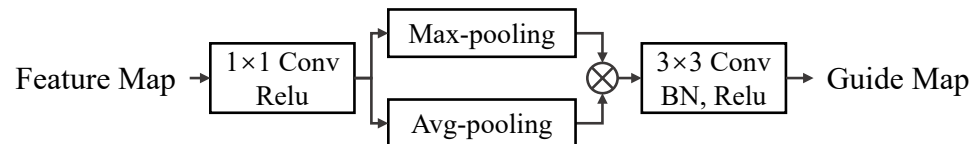


Figure 5. Structure of the guide block.

The guide map is used to activate the corresponding original feature map by element-wise multiplication, aiming to enhance the response of infrared small objects and suppress the response of the background. The output image was processed through an additional 3×3 convolutional layer to adjust the grayscale information distribution. Note that the feature map of L3 has more background noise, making it difficult to generate an accurate guide map. To address this problem, we processed the feature map of L3 layer using the guide map of its adjacent L4 layer. We adopted a bilinear interpolation algorithm and 1×1 convolutional layer to adjust the feature map resolution and number of channels, respectively.

3.3. Batch-Averaged Biased Classification Loss

The confidence loss of SSD, L_{conf} , is the softmax loss over multiple classes confidences (c). It is defined as

$$L_{conf}(x, c) = - \sum_{i \in Pos} x_{ij}^p \log(\hat{c}_i^p) - \sum_{i \in Neg} \log(\hat{c}_i^0) \quad \text{where} \quad \hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \quad (4)$$

$x_{ij}^p = \{1, 0\}$ is the indicator for matching the i th default box to the j th ground-truth box of category p ; N is the number of prior boxes matching ground-truth boxes. Ground truth is the category of each object in the image and its real bounding box.

There are many hard samples in infrared small-object images, and how to distinguish them is a critical issue. To improve the capacity to detect hard samples, Lin et al. [35] proposed an adaptive weight classification loss called focal loss. The focal loss is defined as

$$L_{fl} = -(1 - p_t)^\gamma \log(p_t) \quad \text{where} \quad p_t = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise} \end{cases} \quad (5)$$

$p_t \in [0, 1]$ is the model's estimated probability for the class with label $y = 1$; $\gamma \geq 0$ is a tunable focusing parameter.

However, the critical issue is that the characteristics of infrared and visible images are different, which makes the typical focal loss perform poorly in the training of infrared small-object detection models. To solve this problem, we propose a batch-averaged biased classification loss (Ba loss) based on focal loss.

First, the extreme class imbalance of positive and hard-negative signals encountered during the training of detectors is a central issue. In response, we set a scale factor β to adjust the proportion of positive and negative samples involved in calculating the final classification loss function, thereby suppressing the excessive interference of hard-negative

samples in the model training process. For instance, if there are N positive examples after classification, we sort negative examples using the highest confidence loss for each anchor box and pick the top $\beta \cdot N$ examples. These positives and negatives are used to compute the final classification loss.

Second, at the beginning of the training process—limited by the performance of the initial detection model and characteristics of the sample images—it is difficult to avoid several classification errors. These classification errors enormously increase the classification loss value and significantly affect or even terminate the training of the detection model. To address this issue, we added a small bias factor to L_{fl} , to avoid gradient explosion. In this study, the bias was set to 1×10^{-3} . The optimized L_{fl} is:

$$L'_{fl} = -(1 - p_t)^\gamma \log(p_t + \text{bias}) \quad (6)$$

Finally, the batch size per iteration was not sufficiently large because of the model and hardware memory size limitations. Thus, the classification loss in successive iterations is volatile, particularly in infrared datasets with complex scenes. It is unreliable to evaluate the detection accuracy of the model using a single-batched classification loss in the later training period. To solve this problem, a smoothing method was adopted in this study to adjust the weights of the classification loss of multiple batch samples. The latest confidence loss has a large weight because it reflects the current situation of the model; early confidence losses have low weights. The modified confidence loss is

$$\begin{aligned} L'_{conf_n} &= \frac{1}{2}L_{conf_n} + \frac{1}{2}L'_{conf_{n-1}} = \frac{1}{2}L_{conf_n} + \frac{1}{2^2}L_{conf_{n-1}} + \cdots = \sum_{j=1}^n \frac{1}{2^j}L_{conf_{n-j+1}} \\ &= \sum_{j=1}^n \frac{1}{2^j}L'_{fl_{n-j+1}} \end{aligned} \quad (7)$$

4. Experiments and Results

4.1. Dataset

According to the definition of SPIE, an object with less than 80 pixels in an image of 256×256 pixels is a small object. The dataset [36] selected in this study contained 15,546 images in which the objects were small fixed-wing UAVs. The dataset acquisition scene covered the sky, ground, and a variety of complex scenes. Some of the images in the dataset are shown in Figure 6.

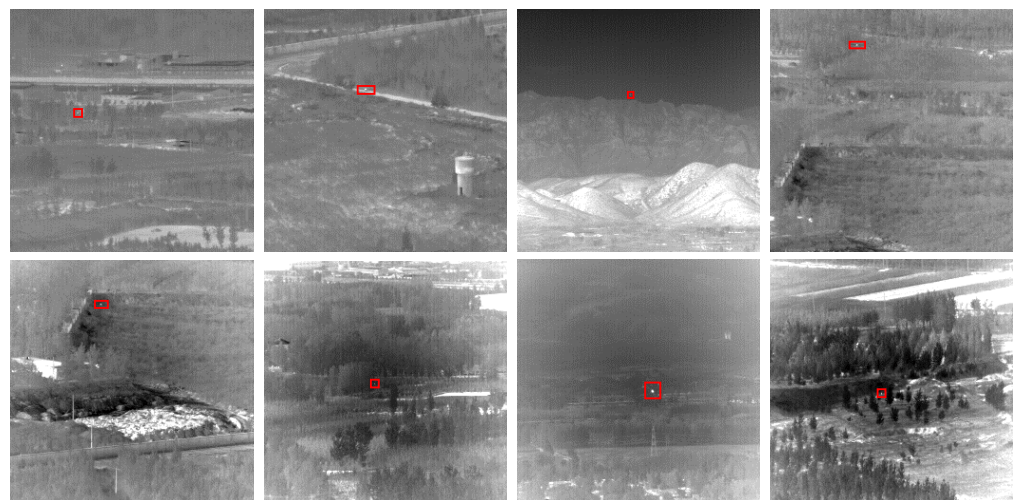


Figure 6. Some images in the experimental dataset. Red boxes mark the real locations of objects. The resolution of the images is 256×256 .

The size distribution of objects in the experimental dataset is shown in Figure 7. A total of 82.2% are below 20 pixels, 10.8% are 20~40 pixels, 4.3% are 40~60 pixels, and 2.7% are 60~80 pixels.

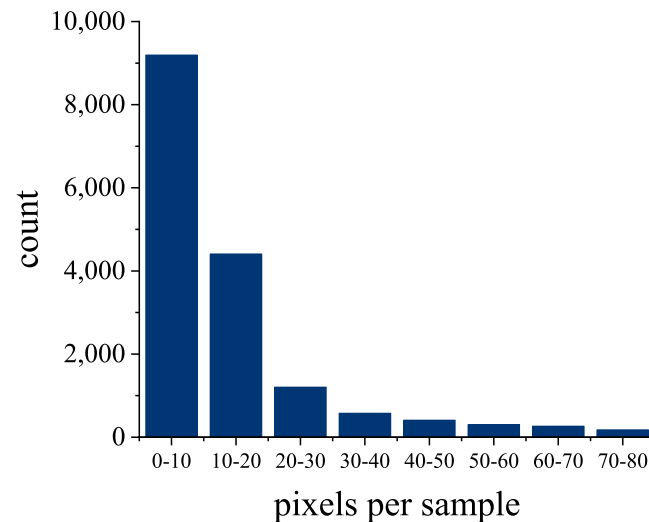


Figure 7. Sizes of objects in the dataset.

The dataset contained 21 scenes, and the training and test sets were divided based on the serial number of scenes to ensure that the sample ratio was 4:1. The training and test sets contained data from different backgrounds, and the details are listed in Table 1.

Table 1. Division of the dataset.

Class	Training Set	Test Set
Data Serial Number	2; 3; 5; 6; 7; 9; 10; 11; 13; 14; 15; 17; 18; 19; 21; 22	4; 8; 12; 16; 20
Number of Images	12,355	3191
Number of Samples	12,954	3590
Average SCR	3.6	4.5

4.2. Experiments Settings

The experiments in this study were run on Ubuntu 20.04, and the deep learning framework was PyTorch 1.8.1. The GPU was 11 GB RTX3080Ti. We used the cosine decay method to adjust the learning rate in the training process. The initial learning rate is 1×10^{-3} , which finally decreases to 1×10^{-7} . The number of training iterations was set as 160,000. The batch size was set as 8. β in the loss function was set as 14. We used the k-means method to cluster the size of ground-truth boxes of the dataset and then preset anchor box parameters to accelerate the reduction of regression loss.

4.3. Evaluation Criteria

Infrared images tend to have more false alarms compared to visible images. Thus, visible small-object detection tasks focus on *FN*, whereas infrared small-object detection tasks should consider *FP*. To address this issue, precision, recall, and *F₁ Score* were used as the evaluation criteria in our experiments for infrared small-object detection.

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F_1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (10)$$

TP is true positive, FN is false negative, and FP is false positive.

The loss curve is a time series curve, and we use Moving Standard Deviation (MSD) and Moving Average (MA) as the evaluation criteria for Ba loss and typical focal loss curves. MSD is used to evaluate curve smoothness, and MA is used to evaluate curve trends. The formulas for MSD and MA are represented as:

$$MSD = \sqrt{\frac{1}{N-1} \sum_{i \in L} |A_i - MA|^2} \quad (11)$$

$$MA = \frac{1}{N} \sum_{i \in L} A_i \quad (12)$$

where L is the moving window, N is the length of L , and A_i is the point i on curve A .

4.4. Results and Analysis

4.4.1. Ablation Studies

This section assesses the functions of the proposed blocks. We used the SSD as the baseline and modified the backbone, feature enhancement method, and loss function according to the method proposed in this study. A comparison of the detection methods with different configurations is presented in Table 2.

Table 2. Experiment results of detection methods with different configurations.

No.	SSD Detector			TP	FN	FP	Precision (%)	Recall (%)	mAP (%)	F_1Score (%)
	Backbone	Feature Enhancement	Optimized Loss							
1	VGG16			3121	469	137	95.8	86.9	86.3	91.2
2		✓		3182	408	124	96.2	88.6	87.8	92.3
3			✓	3262	328	206	94.1	90.9	90.2	92.4
4		✓	✓	3233	357	109	96.7	90.1	89.4	93.3
5				3206	384	159	95.3	89.3	88.3	92.2
6	IRS16(ours)	✓		3287	303	46	98.6	91.6	90.8	95.0
7			✓	3372	218	275	92.5	93.9	92.7	93.2
8		✓	✓	3396	194	43	98.8	94.6	94.5	96.6

bold number: Optimal result.

First, to assess whether the proposed evaluation criteria were rational, we plotted the precision, recall, mAP, and F_1Score in Table 2, as shown in Figure 8. Further analysis showed that the trend of mAP was consistent with the recall trend but was not sensitive to changes in precision. F_1Score was sensitive to changes in both recall and precision. Therefore, this study used F_1Score instead of mAP as the evaluation criterion.

It is apparent from Table 2 that IRS16 can transmit more details about small objects. Compared to No. 1, No. 5 showed an 18.1% decrease in FN , thus resulting in a 2.4% increase in recall. The result is significant that the guide maps have improved the precision and recall of the detection model. FN and FP significantly decreased when using guide blocks as feature enhancement methods, regardless of VGG16 or IRS16. Some of the detection results and corresponding guide maps are shown in Figure 9. Moreover, the batch-average-biased classification loss function was more conducive to the detection of small infrared objects. This effectively improved the recall rate of the detection model. The results of Nos. 3 and 7 show that the recall rates of VGG16 and IRS16 increased by 4.0% and 4.6%, respectively.

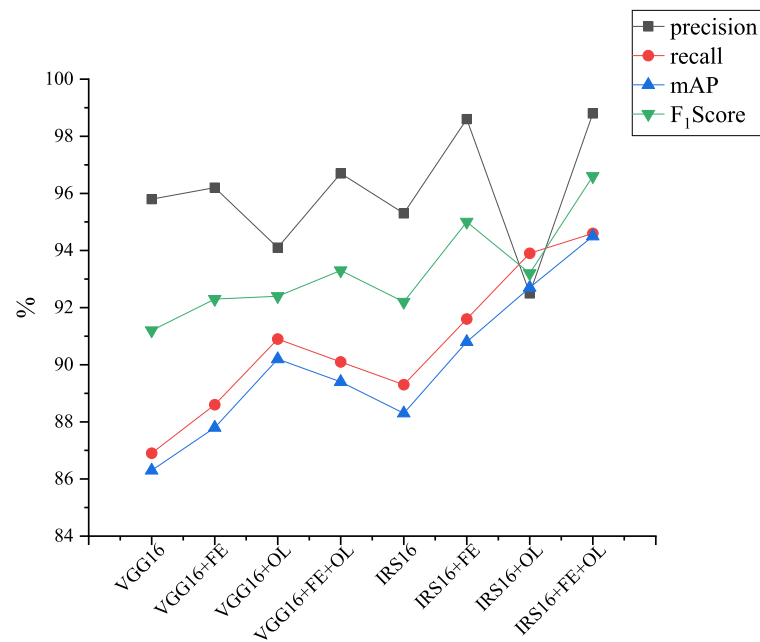


Figure 8. Experiment results of different methods in Table 2. FE means proposed feature enhancement method, and OL means proposed classification loss.

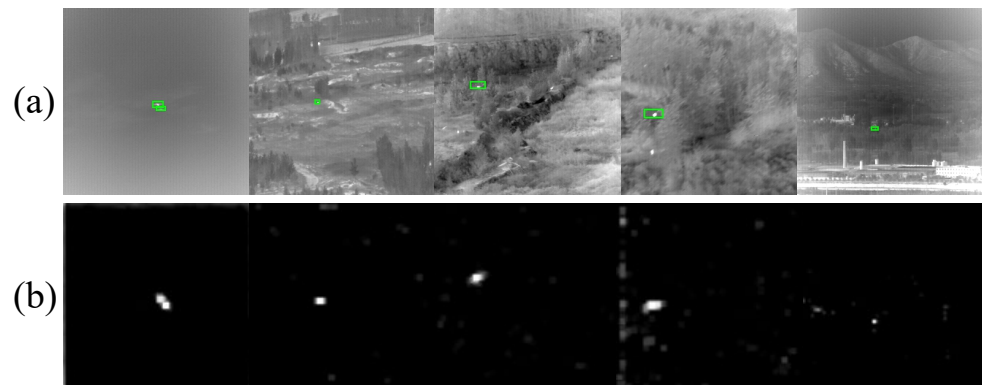


Figure 9. (a) Detection results and (b) corresponding guide maps. The guide maps highlight the object regions.

Overall, these results indicate that each block proposed in this study can improve the accuracy of the detection model. A comparison of SSD512 (No.1 in Table 1) and IRSDet (No. 8 in Table 1) shows that the latter is more suitable for detecting small infrared objects. Remarkably, the F_1 Score of the IRSDet significantly increased by 5.4% compared with that of SSD512, reaching 96.6%.

4.4.2. Different Configurations of the Proposed Model

In this section, we changed the IRS16 configuration and feature enhancement method. The experimental results are listed in Table 3. The models in Table 3 adopted the classification loss proposed in this study.

Table 3. Different configurations of the proposed model.

No.	Backbone		Feature Enhancement	TP	FN	FP	Precision (%)	Recall (%)	mAP (%)	F ₁ Score (%)
	Down-Sampling	Feature Extraction								
1	Ext-pooling	None	Guide block	3411	179	203	94.4	95.0	94.1	94.7
2		Residual block		3357	233	117	96.6	93.5	93.1	95.0
3		SRes block		3396	194	43	98.8	94.6	94.5	96.6
4	Max-pooling	SRes block		3399	191	78	97.8	94.7	94.4	96.2
5	Strided Convolution			3083	507	789	79.6	85.9	81.3	82.6
6	Ext-pooling	SRes block	FPN	3446	144	191	94.8	96.0	95.3	95.4

bold number: Optimal result.

Feature extraction: Comparison of the results for 1, 2, and 3. The detection model using the serial block has many *FPs*, which means that the serial block will lose the texture of the objects, weakening the difference between the noise and objects. In contrast, the residual and SRes blocks have lower *FPs* and can improve the accuracy of the detection model. Remarkably, the excessive use of low-level information of the residual blocks introduced background noise and, therefore, did not substantially decrease the *FPs* and *FNs*.

Down-sampling: Comparison of the detection results of 3, 4, and 5. Max-pooling and Ext-pooling could improve the precision of the detection models. However, the number of *FPs* in the latter was 45% lower than that in the former. The detection model using convolutional downsampling is inferior to the other detection models. This indicates that convolutional downsampling is inappropriate for infrared small-object detection.

Feature enhancement method: It is apparent from Table 3 that the detection model using FPN has the least number of *FNs* compared to the other detection models, which reveals that multiscale feature fusion can combine the object information in multiple feature maps to enhance the features of real objects. However, it also stresses the characteristics of static noise, resulting in an undesired increase in the *FPs*.

4.4.3. Convergence Analysis of Gradient Descent

This study compared the focal loss with the proposed classification loss function. We used the default classification loss function of SSD to pretrain the initial detection model to avoid gradient explosion owing to focal loss at the beginning of the training process. The number of iterations of pretraining was 40,000, and the batch size was set to 8. Subsequently, the focal and Ba loss functions were employed in the model. The number of iterations was 40,000, and the batch size was set to 8. The learning rate was 1×10^{-3} . The results are shown in Figure 10. The curve of the classification loss function proposed in this study is smoother than that of the focal loss function curve and has a faster convergence rate.

The experimental results show that in the training process of the infrared small-object detection model, the batch-averaged method can effectively solve the loss fluctuation problem caused by the limitation of GPU memory. Moreover, the scale factor of positive to negative helps the detection model eliminate the learning dilemma owing to the extreme class imbalance between positive and negative samples. Using these methods, the model can focus on distinguishing between hard samples, thereby reducing the loss value for the detection model.

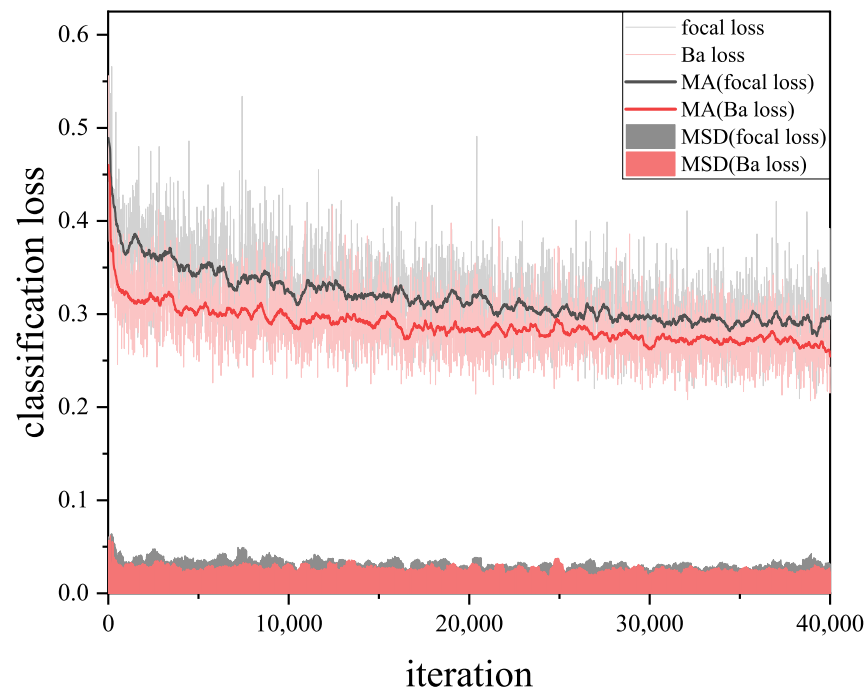


Figure 10. Comparison of different classification losses. MA = Moving Average; MSD = Moving Standard Deviation.

4.4.4. Comparison of Advanced Detection Models

Clearly, from Table 4, the method in this study performed well at infrared small-object detection. The *TPs*, *FNs*, *FPs*, and precision of the proposed model reached a suboptimal level, and the recall, mAP, and *F₁ Score* reached an optimal level. Using high-resolution feature maps inevitably reduces the inference speed of the model; however, it decreases the *FNs* and *FPs*. Some of the detection results are presented below.

Table 4. Comparison of different detection methods.

Method	<i>TP</i>	<i>FN</i>	<i>FP</i>	Precision (%)	Recall (%)	mAP (%)	FPS	<i>F₁ Score</i>
SSD512	3121	469	137	95.8	86.9	86.3	117	91.2
IRSDet(ours)	<u>3396</u>	<u>194</u>	<u>43</u>	<u>98.8</u>	94.6	94.5	72	96.6
DSSD	3284	306	255	92.8	91.5	88.5	<u>105</u>	92.1
FSSD	3286	304	124	96.4	91.5	91.1	106	93.9
SSD-ST	3207	383	99	97.0	89.3	88.3	83	93.0
FA-SSD	3211	379	271	92.2	89.4	83.8	62	90.8
FD-SSD	3311	279	31	99.1	92.2	92.2	44	<u>95.5</u>
DF-SSD	2920	670	136	95.6	81.3	79.0	81	87.9
YOLOv3	3308	282	435	88.4	92.2	86.4	66	90.2
YOLOv4	3397	193	446	88.4	<u>94.6</u>	<u>93.1</u>	67	91.4

bold number: Optimal result, underline number: Suboptimal result.

It can be seen from the above table that the method in this paper has a good performance in infrared small object detection. The proposed model's *TPs*, *FNs*, *FPs*, and precision reached a suboptimal level, and the recall, mAP, and *F₁ Score* reached an optimal level. The use of high-resolution feature maps inevitably reduces the inference speed of the model but also decreases *FNs* and *FPs*. Some detection results are shown in Figure 11.

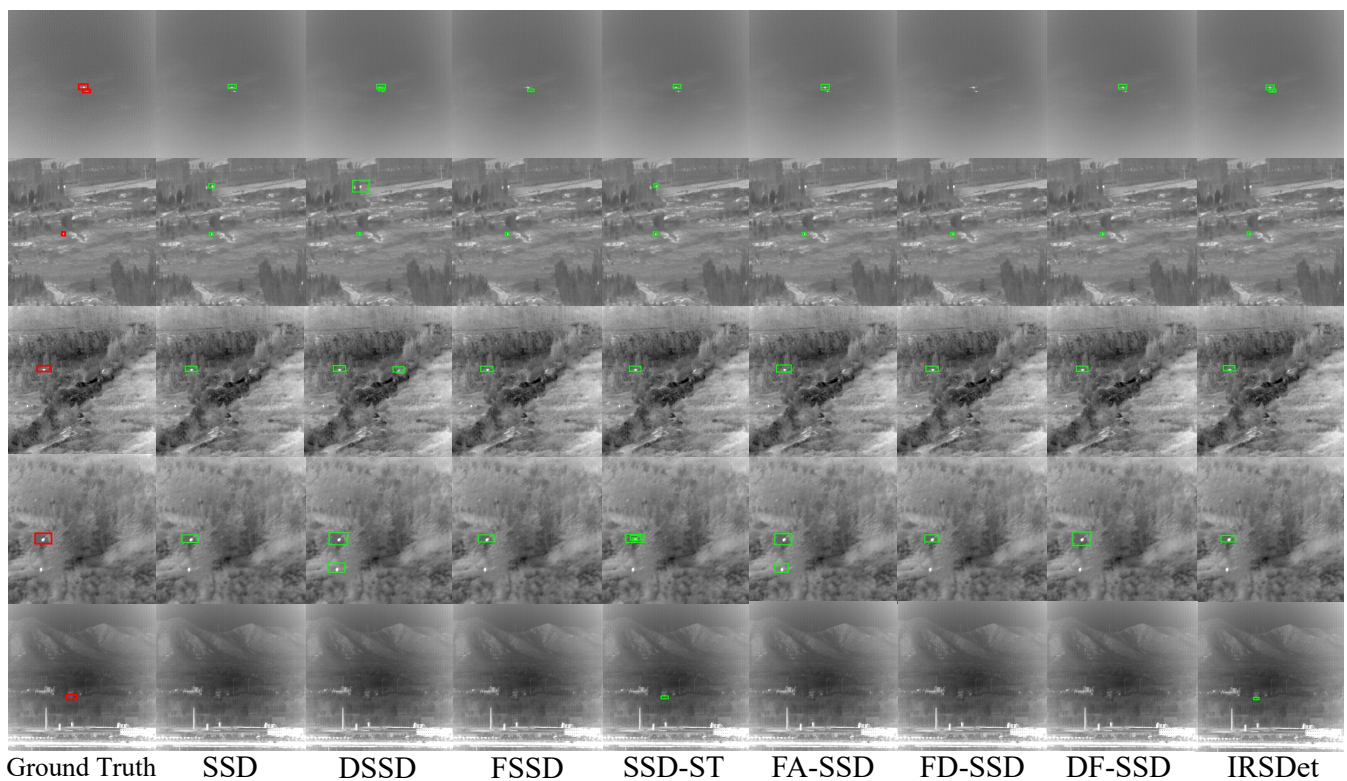


Figure 11. Detection results of advanced detection models.

5. Conclusions

This study proposed an infrared small-object detection framework based on deep learning to improve the detection capacity for small objects such as drones and vehicles in complex backgrounds. First, we proposed a backbone that uses sparse skip connection and the optimized downsampling method to enhance the feature representation of small objects. Then, we proposed a feature enhancement module based on the attention mechanism to filter potential object regions. Finally, the classification loss function was modified to improve the detection accuracy for infrared hard samples. A small public infrared dataset was used to evaluate the detection model. The experimental results show that the IRSDet proposed in this study performed better than the other advanced small-object detection methods. The precision and recall rates were 98.8% and 94.6%, respectively, and the F_1 Score reached 96.6%.

This paper provides deeper insight into research in the field of infrared object detection and tracking. The limitations of this study are that we did not optimize the location loss function, and the inference speed of the current detection models was not sufficiently fast. Therefore, our future research direction is to explore the position loss function suitable for small infrared objects and determine an efficient combination of traditional and deep learning methods.

Author Contributions: Methodology, X.X.; Software, X.X. and J.W.; Investigation, X.X. and J.W.; Data curation, X.X. and J.W.; Writing—original draft preparation, X.X.; Writing—review and editing, X.X., J.W., F.L. and D.L.; Supervision, D.L.; Project administration, F.L.; Funding acquisition, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

References

1. Liu, T.; Li, Y.F.; Liu, H.; Zhang, Z.; Liu, S. RISIR: Rapid infrared spectral imaging restoration model for industrial material detection in intelligent video systems. *IEEE Trans. Ind. Inform.* **2019**. [\[CrossRef\]](#)
2. Liu, T.; Liu, H.; Li, Y.F.; Chen, Z.; Zhang, Z.; Liu, S. Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing. *IEEE Trans. Ind. Inform.* **2019**, *16*, 544–554. [\[CrossRef\]](#)
3. Yavariabdi, A.; Kusetogullari, H.; Celik, T.; Cicek, H. FastUAV-net: A multi-UAV detection algorithm for embedded platforms. *Electronics* **2021**, *10*, 724. [\[CrossRef\]](#)
4. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
6. 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, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
8. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
9. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
10. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Detnet: Design backbone for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 334–350.
11. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. Dssd: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
12. Jeong, H.P.J.; Kwak, N. Enhancement of SSD by concatenating feature maps for object detection. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017; Kim, T.-K., Zafeiriou, G.B.S., Mikolajczyk, K., Eds.; BMVA Press: London, UK, 2017; pp. 76.1–76.12. [\[CrossRef\]](#)
13. Li, Z.; Zhou, F. FSSD: feature fusion single shot multibox detector. *arXiv* **2017**, arXiv:1712.00960.
14. Liu, S.; Huang, D. Receptive field block net for accurate and fast object detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 385–400.
15. Park, J.; Chen, J.; Cho, Y.K.; Kang, D.Y.; Son, B.J. CNN-Based Person Detection Using Infrared Images for Night-Time Intrusion Warning Systems. *Sensors* **2019**, *20*, 34. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Yao, S.; Zhu, Q.; Zhang, T.; Cui, W.; Yan, P. Infrared Image Small-Target Detection Based on Improved FCOS and Spatio-Temporal Features. *Electronics* **2022**, *11*, 933. [\[CrossRef\]](#)
17. Du, J.; Lu, H.; Zhang, L.; Hu, M.; Chen, S.; Deng, Y.; Shen, X.; Zhang, Y. A Spatial-Temporal Feature-Based Detection Framework for Infrared Dim Small Target. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–12. [\[CrossRef\]](#)
18. Shi, M.; Wang, H. Infrared Dim and Small Target Detection Based on Denoising Autoencoder Network. *Mob. Netw. Appl.* **2020**, *25*, 1469–1483. [\[CrossRef\]](#)
19. Fang, H.; Xia, M.; Zhou, G.; Chang, Y.; Yan, L. Infrared Small UAV Target Detection Based on Residual Image Prediction via Global and Local Dilated Residual Networks. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [\[CrossRef\]](#)
20. Fan, M.; Tian, S.; Liu, K.; Zhao, J.; Li, Y. Infrared Small Target Detection Based on Region Proposal and CNN Classifier. *Signal Image Video Process.* **2021**, *15*, 1927–1936. [\[CrossRef\]](#)
21. Ren, K.; Gao, Y.; Wan, M.; Gu, G.; Chen, Q. Infrared Small Target Detection via Region Super Resolution Generative Adversarial Network. *Appl. Intell.* **2022**, *52*, 11725–11737. [\[CrossRef\]](#)
22. Ding, L.; Xu, X.; Cao, Y.; Zhai, G.; Yang, F.; Qian, L. Detection and tracking of infrared small target by jointly using SSD and pipeline filter. *Digit. Signal Process.* **2021**, *110*, 102949. [\[CrossRef\]](#)
23. Du, J.; Lu, H.; Hu, M.; Zhang, L.; Shen, X. CNN-based Infrared Dim Small Target Detection Algorithm Using Target-Oriented Shallow-Deep Features and Effective Small Anchor. *IET Image Process.* **2021**, *15*, 1–15. [\[CrossRef\]](#)
24. Ju, M.; Luo, J.; Liu, G.; Luo, H. ISTDet: An efficient end-to-end neural network for infrared small target detection. *Infrared Phys. Technol.* **2021**, *114*, 103659. [\[CrossRef\]](#)
25. Hou, Q.; Wang, Z.; Tan, F.; Zhao, Y.; Zheng, H.; Zhang, W. RISTDnet: Robust infrared small target detection network. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [\[CrossRef\]](#)
26. Sun, C.; Ai, Y.; Wang, S.; Zhang, W. Mask-guided SSD for small-object detection. *Appl. Intell.* **2021**, *51*, 3311–3322. [\[CrossRef\]](#)
27. Yin, Q.; Yang, W.; Ran, M.; Wang, S. FD-SSD: An improved SSD object detection algorithm based on feature fusion and dilated convolution. *Signal Process. Image Commun.* **2021**, *98*, 116402. [\[CrossRef\]](#)
28. Lim, J.S.; Astrid, M.; Yoon, H.J.; Lee, S.I. Small object detection using context and attention. In Proceedings of the 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIC), Jeju Island, Korea, 13–16 April 2021; pp. 181–186.
29. Cui, L.; Ma, R.; Lv, P.; Jiang, X.; Xu, M. MDSSD: Multi-scale deconvolutional single shot detector for small objects. *Sci. China Inf. Sci.* **2020**, *63*, 120113. [\[CrossRef\]](#)

30. Zhai, S.; Shang, D.; Wang, S.; Dong, S. DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion. *IEEE Access* **2020**, *8*, 24344–24357. [[CrossRef](#)]
31. Pan, H.; Jiang, J.; Chen, G. TDFSSD: Top-down feature fusion single shot MultiBox detector. *Signal Process. Image Commun.* **2020**, *89*, 115987. [[CrossRef](#)]
32. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296.
33. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 206–221.
34. Bai, Y.; Li, R.; Gou, S.; Zhang, C.; Chen, Y.; Zheng, Z. Cross-Connected Bidirectional Pyramid Network for Infrared Small-Dim Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 7506405. [[CrossRef](#)]
35. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
36. Hui, B.; Song, Z.; Fan, H. A dataset for infrared detection and tracking of dim-small aircraft targets under ground/air background. *China Sci. Data* **2020**, *5*, 291–302.