remote sensing

MDPI

Article

Low Contrast Infrared Target Detection Method Based on Residual Thermal Backbone Network and Weighting Loss Function

Chunhui Zhao [1,2], Jinpeng Wang [1,2], Nan Su [1,2,*], Yiming Yan [1,2] and Xiangwei Xing [3]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China; zhaochunhui@hrbeu.edu.cn (C.Z.); wangjinpeng9521@hrbeu.edu.cn (J.W.); yanyiming@hrbeu.edu.cn (Y.Y.)
[2] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin 150001, China
[3] Beijing Remote Sensing Information Institute, Beijing 100094, China; xingxiangwei@nudt.edu.cn
* Correspondence: sunan08@hrbeu.edu.cn

Abstract: Infrared (IR) target detection is an important technology in the field of remote sensing image application. The methods for IR image target detection are affected by many characteristics, such as poor texture information and low contrast. These characteristics bring great challenges to infrared target detection. To address the above problem, we propose a novel target detection method for IR images target detection in this paper. Our method is improved from two aspects: Firstly, we propose a novel residual thermal infrared network (ResTNet) as the backbone in our method, which is designed to improve the feature extraction ability for low contrast targets by Transformer structure. Secondly, we propose a contrast enhancement loss function (CTEL) that optimizes the weights about the loss value of the low contrast targets' prediction results to improve the effect of learning low contrast targets and compensate for the gradient of the low-contrast targets in training back propagation. Experiments on FLIR-ADAS dataset and our remote sensing dataset show that our method is far superior to the state-of-the-art ones in detecting low-contrast targets of IR images. The mAP of the proposed method reaches 84% on the FLIR public dataset. This is the best precision in published papers. Compared with the baseline, the performance on low-contrast targets is improved by about 20%. In addition, the proposed method is state-of-the-art on the FLIR dataset and our dataset. The comparative experiments demonstrate that our method has strong robustness and competitiveness.

Keywords: residual thermal infrared network; contrast enhancement loss function; low contrast target detection; infrared (IR) target detection

Citation: Zhao, C.; Wang, J.; Su, N.; Yan, Y.; Xing, X. Low Contrast Infrared Target Detection Method Based on Residual Thermal Backbone Network and Weighting Loss Function. *Remote Sens.* **2022**, *14*, 177. https://doi.org/10.3390/rs14010177

Academic Editor: Józef Lisowski

Received: 7 December 2021
Accepted: 28 December 2021
Published: 1 January 2022

1. Introduction

Infrared image shows an outstanding advantage in the bad illuminance environment compared to optical images (the RGB images). Infrared target detection plays a very important role in many applications, such as early warning system and marine monitoring system [1,2]. At present, most methods rely on the information about the inherent feature of the image to complete the target detection task. Inherent features refer to the features that can distinguish the target from the background in the image. For example, if the color information in the optical image is used to detect the red football on green grass, the task is simple, while this color information is lacking in the infrared image. Inherent features also include texture features, contour features, and so on. However, compared with optical images, infrared images obviously lack information of inherent features. In addition, the infrared image often suffers from strong background clutter, noise and low contrast. The above problems make infrared target detection an arduous task.

Target detection methods for infrared can generally be divided into three categories [3]: methods based on local prior, methods based on nonlocal prior and methods based on learning. In local prior methods [4–6], scholars rely on local information to search targets from images. In addition, to optimize the algorithm performance, various improved filters are used to enhance the target information before detection. However, these methods are sensitive to complex backgrounds and perform poorly on low-contrast targets. Excellent detection performance can only be obtained in a simple background but often fail in real scenes. The detection method based on nonlocal prior has shown competitive detection performance in recent years [7]. By using the nonlocal similarity of background and the sparsity of target, a matrix-level pre-background separation method is proposed. However, this method still cannot have good performance in low contrast targets.

Recently, deep-learning based methods have also been used for infrared target detection. In the deep-learning based method, the overall model can be mainly divided into feature extraction part and result prediction parts. RISTDnet [8] improves the feature extraction ability of infrared image targets by integrating multi-scale convolution results. A backbone network using multi-residual structure is proposed in [3], and the loss of target information in deep networks is suppressed by establishing each layer of the information path. These methods are focused on the information loss of IR images in the process of feature extraction, without considering the loss of inherent feature in IR images. ThermalDet [9] was used for infrared target detection and achieved competitive performance. This structure can adaptively reweight each channel of the feature after fusing different levels of features. Deep-IRTarget [10] establishes the information path between each feature channel through the self-attention networks to further extract the spatial information of the image. However, these methods do not take into account the context information between the target area and the surrounding background area in the image. MMTOD [11] uses optical image information to compensate IR image information and improves the detection performance of the algorithm. However, when the infrared image view information is more reliable, such as low illumination, smoke, and so on, optical image information is an interference to the target detection network. In [12], a target detection method based on infrared image feature contrast is proposed. In other words, by optimizing the calculation method of local contrast, the target position is more accurately searched by contrast in the feature map. This method can only achieve good performance in a simple background and cannot be applied to scenes with complex ones.

In the past few years, due to the wide application scenarios of optical images, target detection algorithms have developed faster in the field of optical images than in the field of infrared images. The performance of Cascade RCNN [13] with multi-level connection structure on optical images public dataset is amazing. It optimizes the performance of the method by multistage screening of the back propagation samples. However, the low-contrast difficult samples do not occupy the majority of the data set, and the learning effect of the algorithm on the low-contrast samples may be interfered with by other samples. The widely applied YOLO [14,15] series and the most advanced YOLOX [16] algorithm in the field of optical images had designed a large number of residuals and branch structures in the backbone network, which enhances the feature extraction ability of the network. However, the loss of inherent feature for the target under infrared imaging conditions is not taken into account.

The low contrast targets in infrared images are seriously submerged by background, which is the difficult target for target detection tasks. These difficult targets have a great influence on the performance of the detection method. The above current detection methods focus on how to obtain more feature information, without using context information to enhance the key features in the image. In addition, there is a lack of network training objectives, namely loss function targeted optimization. There is a loss of inherent features in infrared images, so it is necessary to extract the feature information fully. In addition, due to the imbalance between low contrast samples and other samples in dataset, the training effect of low contrast targets is not ideal. To solve the problems above, we propose a novel

infrared target detection method. The method improves a deep learning network with physical characteristics of infrared images for target detection. Specifically, a backbone network is proposed to enhance the target feature by integrating the context information to improve the feature extraction ability of low-contrast targets. In addition, we propose a new loss function to improve the learning effect of low-contrast targets.

In summary, the main contributions of this article are listed as follows:

1. We propose a novel residual thermal infrared network (ResTNet) based on an attention mechanism to alleviate the inherent feature loss problem of infrared image targets. A novel multi-spatial attention network (MSAN) is designed in ResTNet, which uses a Transformer structure for attention operation. The network can establish the information path between local regions of different scales in each position of the image in the feature extraction process, so as to enhance the extraction of target feature in the image by integrating the context information;
2. A contrast enhancement loss function (CTEL) is proposed related to target contrast to suppressing the imbalance between low contrast targets and other targets. Specifically, the weight of low contrast targets loss values is optimized by CTEL, which effectively improved the detection effect of low-contrast difficult targets and compensate for the gradient of the low-contrast targets in training back propagation;
3. We produce a new infrared dataset about remote sensing and verify our method on it. The experimental results on the FLIR dataset and our dataset show that our method is far superior to the current most advanced algorithm.

The rest of this article is structured as follows: in Section 2, we describe the network structure and methods in detail. Section 3 gives the details of our work and experimental results and related comparison to verify the effectiveness of our method. Discussions of proposed improvements are presented in Section 4. Finally, we summarize the research content in the fifth part.

2. Method

2.1. Overall Network Architecture

This article presents a two-stage method for IR images' target detection as shown in Figure 1. The overall structure is divided into three parts: backbone, neck, and detection head network. In our method, the proposed ResTNet is used as the backbone network, and the feature pyramid network (FPN) is used as the neck in our method. The detection head is composed of a regional proposal network (RPN) [17] and result prediction network.

Firstly, the IR images as the input of backbone network to extract global features, and enhance target features by the designed MSAN. Secondly, the FPN fuses the feature information extracted from the image and enters it into the RPN to get region of interest (ROI). Finally, the detection head combines the ROI and image features to predict the position and category of targets. In particular, the proposed CTEL is used in the result prediction network to enhance the detection performance of usually difficult low-contrast targets by optimizing the weights of low contrast targets and other target loss function values.

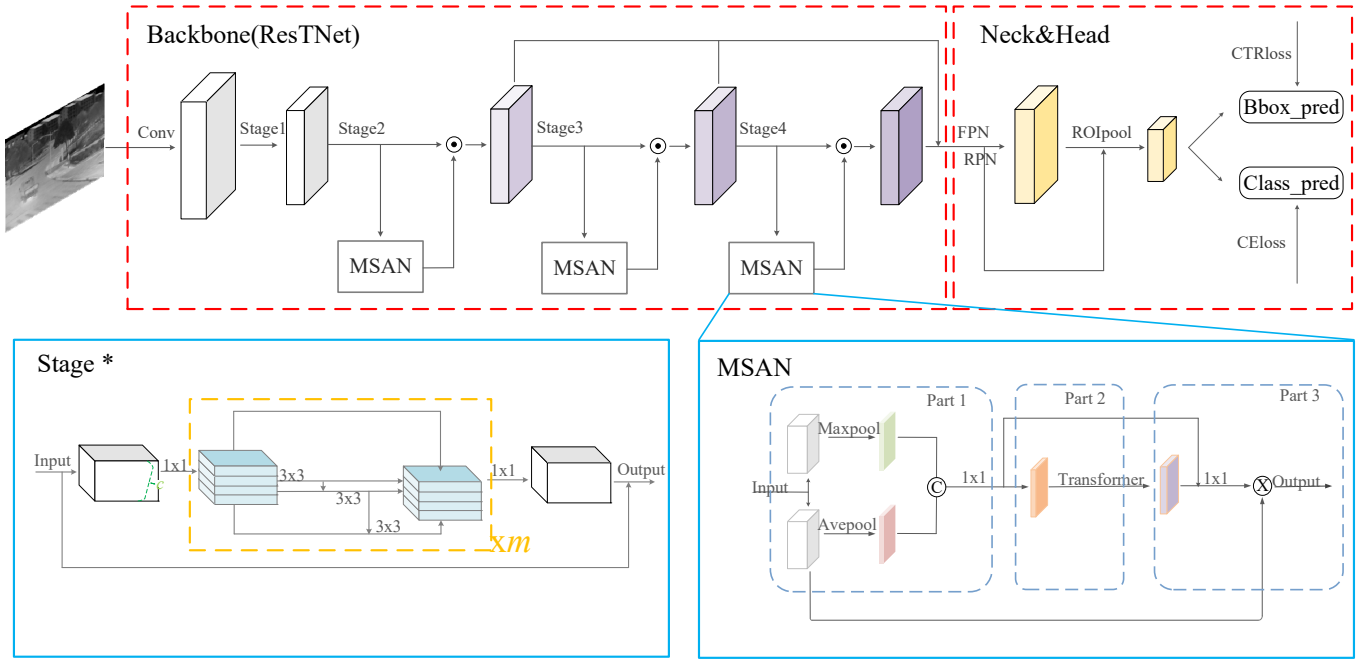

Figure 1. The overall scheme of the proposed method. The part marked with 1 by 1 (3 by 3) denotes convolution as the same size before and after input using kernel size 1 (3). The intersection part of the arrow represents the element addition operation between tensors.

2.2. ResTNet

Compared with optical images, infrared images have inherent defects such as loss of features and low contrast. Many popular target detections methods in optical images based on deep learning cannot be well applied to IR images. To solve the above problems, a more powerful backbone network, ResTNet, is designed. That backbone network can further extract the spatial feature response from the infrared image, and search the internal relations to each local area of the feature in a higher dimension. Meanwhile, the context feature information is integrated to highlight the target features in the image.

The ResTNet and internal structure are shown in the blue boxes in Figure 1. The proposed ResTNet is improved based on ResNet. The ResNet consists of five parts: the shallowest convolution layer and four stages. The Stage part of the feature extraction network is composed of many convolution blocks. In addition, the number of convolution blocks at each stage is not necessarily the same. The m in the figure represents the number of convolution blocks stacked in a certain stage, namely the depth of the network. In our method, the convolution block is structured using the structure used in [18]. The feature tensor obtained by 1×1 convolution operation of input features is divided into four groups along the channel, and then the addition and convolution operations are carried out in turn. To supplement the feature information of targets in IR images by integrating image context features, MSAN is used at the end of the last three convolution blocks of ResTNet. The MSAN is mainly composed of three parts. In the first part, the MSAN convolutes the results of the maximum pooling and average pooling of the input feature tensor along the channel direction, and takes the convolution result as the input of the Transformer structure. The weight matrix is obtained by combining the input and output of the Transformer network. In the third part, the element product of the input feature and the weight matrix are used as the output. In general, the correlation of local regions in the image is inversely proportional to the distance. In other words, the target's feature information is context-related only to a limited range of local areas.

Inspired by the Transformer [19] structure, we design a novel Transformer structure in the ResTNet as shown in Figure 2, which computes the cross-attention of feature responses at different scales for each local region in the feature space response map. At the same time,

information paths are established between local regions of different scales at each location of the image.

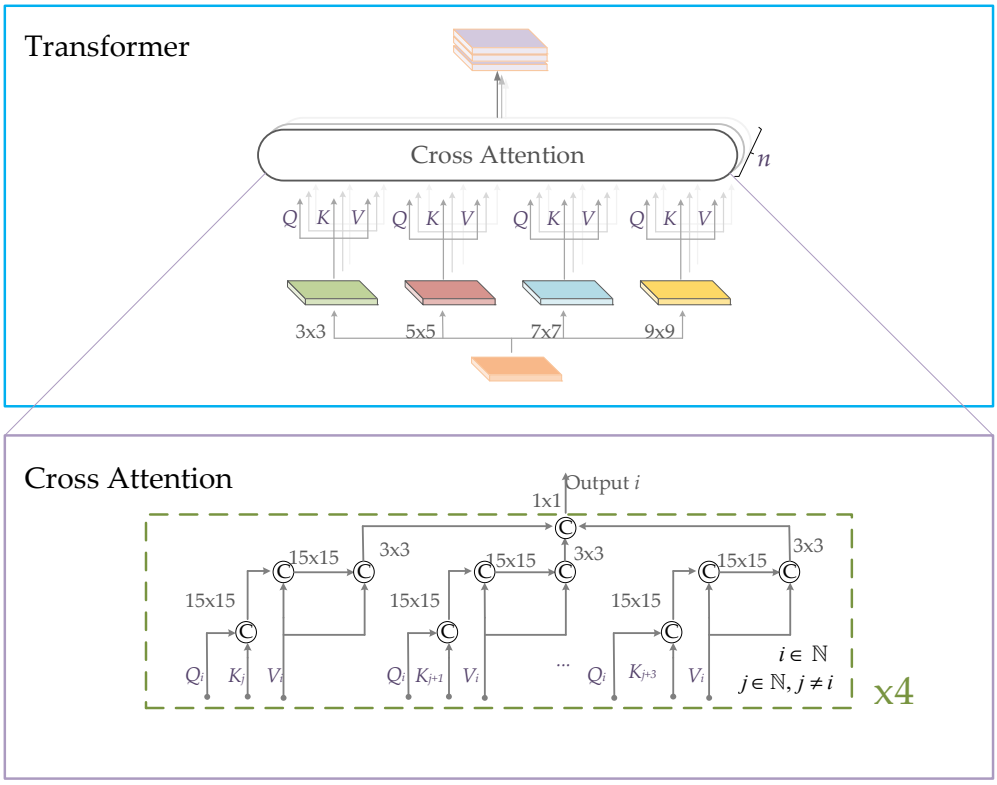

Figure 2. Our transformer structure and its internal structure diagram.

As shown in Figure 2, the calculation formula of cross-attention module is as follows:

$$CrA_i(s_1, s_2, \ldots, s_n) = C_3(s_i, \sum_{j \in \mathbb{N}^n} C_{15}(v_i, C_{15}(q_i, k_j))) \tag{1}$$

where q_i, k_i, and v_i are the query, key, and value response of the ith input tensor, which are further abstracted by the input tensor through convolution operation. In addition, $C_k(x, y)$ refers to the convolution operation with kernel size $k \times k$, step size 1 and padding value $(k-1)/2$ after splicing the two variables x and y. In other words, width and height of input and output of convolution are the same. In order to notice the information in different subspaces, the attention calculation for each input feature is the fusion of the calculation results from multiple cross-attention modules with different parameters. The function expression of the transformer structure is as follows:

$$f(X) = C_1(CrA_1(S), CrA_2(S), CrA_3(S), \ldots, CrA_n(S)) \tag{2}$$

To make the network focus on the spatial information of features, the joint expression of two different pooling operations for input features is used as the input of the proposed converter structure. Finally, the overall spatial response map is mapped between 0 and 1 as the spatial feature enhancement coefficient matrix of input features. The calculation process of MSAN can be seen in the equation:

$$Y = C_1(f(Y_1), Y_1) \cdot X \tag{3}$$

$$Y_1 = C_1(\frac{1}{k}\sum_k X_{ijk}, MAX_k(X_{ijk})), X \in \mathbb{R}^{i \times j \times k} \tag{4}$$

where X is the input feature tensor, and MAX_k refers to the max pooling operation of input features in the direction of the k dimension. In addition, the Y_1 is the joint expression of the average pooling response map and the maximum pooling response map of the input feature.

2.3. Contrast Enhancement Loss Function

There are usually difficult low-contrast targets in the infrared dataset that is difficult to learn. Aiming at the imbalance between low contrast targets and other targets, a new loss function is proposed. Our loss function includes two parts: classification and regression. The CIoU loss function [20] is used as the basic regression loss function for this article. At the same time, the cross entropy loss function is used for classification training. Compared with the traditional IoU [21] loss function, CIoU more accurately reflects the regression effect of the anchor box by taking into account the overlap area, relative distance, and aspect ratio of the prediction box and the real box. On the basis of CIoU, CTEL optimizes the loss function weight value of the target according to the contrast to balance the learning effect of low contrast samples and other samples.

To evaluate the difficulty of the sample, we design a measure to represent the difference between the target area and the background area. Firstly, for the target with scale (h, w), we obtain the region containing the target and the surrounding background with scale $((1+\alpha)\cdot h, (1+\beta)\cdot w)$ to calculate the contrast. In this paper, the values of α and β are both 1.

Secondly, the selected region is divided into five parts shown in Figure 3: one Target area on the left side of the figure and the four Background areas on the right side of the figure. Finally, the gray value distribution difference between the target and the background in each direction is calculated and the information entropy [22] as the mathematical model of calculating the discrete degree of gray value distribution in a local area. At the same time, the gray value of the region is multiplied by its corresponding information entropy, so that our measurement can take into account the center of the distribution.

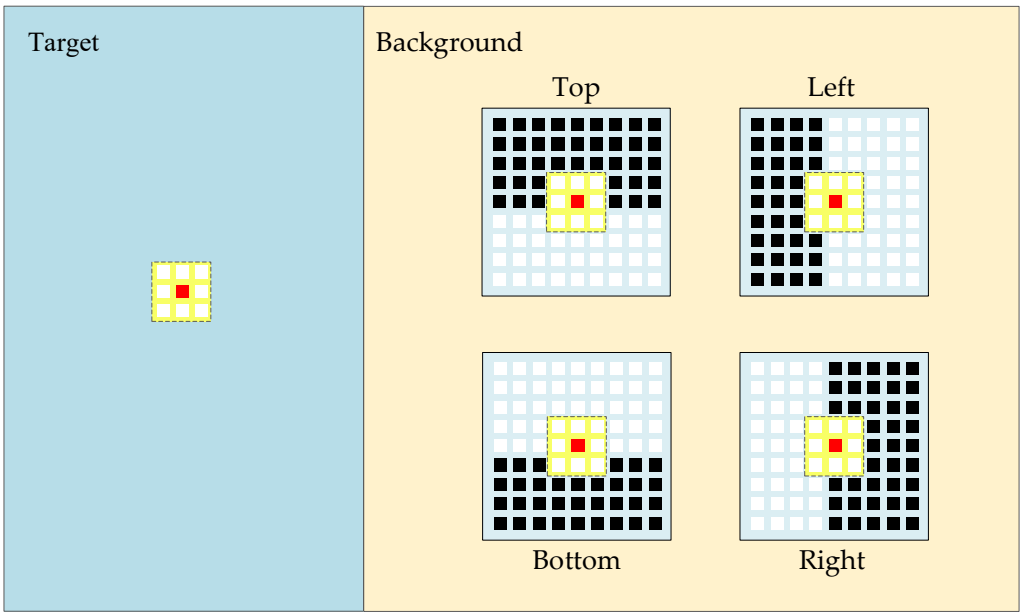

Figure 3. Target and surround background area division diagram. In the process of background division, black squares represent the selected area.

If the background areas in all directions have different degrees of submergence to the target, the difference between the target area and the background gray value distribution in all directions will be small. Based on this theory, the maximum difference between the

target area and the background distribution in each direction as the contrast value of the target, which can evaluate its difficulty.

$C_{(t)}$, that is, the contrast of target t is defined as the following equation:

$$C_{(t)} = \underset{d \in Dir}{Maximum} \left| \sum_{i=0}^{255} (1+i) P_{i,t} \log P_{i,t} - (1+i) P_{i,d} \log P_{i,d} \right|, Dir = top, bottom, left, right\} \quad (5)$$

where t and d are the target area and the corresponding background area, respectively. The t is the target area, and Dir represents the different background areas in the area intercepted on the original image centered on the center point of the target area. The P_{it} denotes the probability that the i pixel value will appear in the t region. In addition, our logarithmic operation is based on ten.

The specific contrast calculation method is as Algorithm 1:

Algorithm 1. Contrast Calculation Method.

Require: *ImgData*: the matrix of the image; *TargetRegion*: location of the target; α: visual field coefficient; β: visual field coefficient;
Ensure: contrast x
get the matrix of the target region with scale (h,w) as t;
get the matrix of surrounding the target region with scale ((1+α)h,(1+β)w);
split four background regions as [b_1,b_2,b_3,b_4]
for k = 1; k < 4; k++ **do**
m_k =0
for i = 1; i <= 255; i++ **do**
compute the number of pixel value i
compute the frequency of pixel value i in the target area and background area as p_t and p_d
m_k += | |$(i+1)p_t \log(p_t) - (i+1)p_d \log(p_d)$| |
end for
end for
x = max(m_1,m_2,m_3,m_4)

Gaussian function has the characteristic of being farther away from the center, the smaller the weight. In addition, there is a σ parameter that can easily adjust the interval. If our mapping function shows a trend of Gaussian function in the main distribution range of contrast probability density, it can achieve the purpose of enhancing difficult samples and ensuring network stability.

Thus, the Gaussian function is applied as the basic math model of the contrast enhancement coefficient mapping function. In order to enhance the learning effect of difficult samples and accelerate the convergence rate of simple samples, the sum of the mapping coefficient and the super parameter b are used for the final additional enhancement coefficient. The mapping function is as follows:

$$\lambda = 1 + \frac{k}{\sqrt{2\pi}\sigma} e^{-\frac{C_{(t)}^2}{2\sigma^2}} + b \quad (6)$$

where both k, σ, and b are our preset hyper-parameters. The square of the target contrast is utilized for the index of the natural base number e and multiply it with a mapping coefficient. Then, the results are added with a super parameter b as the additional enhancement coefficient of the loss function.

The function of CTEL is as follows:

$$L_{CTEL} = (1 - IoU + \frac{40}{\pi}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2) \cdot \lambda \quad (7)$$

where w and w^{gt} refer to the width of the prediction box and the width of the real box, respectively, so does h.

The process of CTEL calculation is as Algorithm 2:

Algorithm 2. CTEL Calculation Method.

Require: *PreResult*: a prediction result; *GT*: a annotation box; F_{ciou}: calculation function of CIoU; δ: hyper-parameter of mapping function; *k:* hyper-parameter of mapping function; *b:* hyper-parameter of mapping function;
Ensure: regression loss l
compute CIoU loss $L_{ciou} = F_{ciou}(GT, PreResult)$
compute the contrast of the target as x
$Index = \frac{x^2}{2\delta^2}$
$c = \frac{k}{\sqrt{2\pi}\delta}$
$b = b + 1$
$w = ce^{-Index} + b$
compute value of regression loss $l = wL_{ciou}$

In summary, the overall network loss function and classification loss is as follows:

$$L_{cls} = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{8}$$

$$Loss = L_{cls} + L_{CTEL} \tag{9}$$

where i means a sample, and c denotes a class. y_{ic} is a symbolic function. If the real category of sample i is equal to c, it takes 1; otherwise, it takes 0. p_{ic} is the probability that sample i belongs to category c.

3. Experiment and Analysis

This section introduces the experimental dataset, implementation details, and related evaluation indexes. In addition, a large number of comparative experiments were designed to verify the advancement and robustness of the proposed method.

3.1. Dataset Introduction

A series of experiments are designed on two datasets. The proposed method has been compared with other SOTA methods on two datasets.

FLIR-ADAS dataset [23] takes a vehicle camera view from November to May on the streets and roads of Santa Barbara, California, from sunny day (60%) and night (40%) to cloudy weather. The dataset contains pedestrians, dogs, bicycles, small cars, and other cars. Since the number of samples about dogs and other cars in the dataset are much smaller than that of other categories, it will bring certain contingency in the experimental evaluation stage. In order to be consistent with other researchers, the labeling information of dogs and other vehicles is ignored in the experimental stage. In addition, basic data enhancement methods such as offset and rotation are used for bicycle samples with less training samples.

Our infrared dataset for remote sensing was collected on the road by a drone equipped with an infrared sensor camera. Due to the different imaging bands of infrared images and the great influence of ambient temperature, in order to ensure the integrity of the dataset, and the dataset spans the temperature difference of 25 degrees Celsius. During the period from May to November, the day (50%) and night (50%) data were captured on the roads and streets of Harbin.

Low contrast in infrared images has a great influence on the performance of the detection method. We aim to optimize improving the detection performance of low-contrast targets. To verify the improvement effect of the proposed method on low contrast samples, the contrast distribution of samples in each dataset is calculated according to Formula (5).

Specifically, it is the maximum value of the difference between the distribution measurement of the target and the surrounding four backgrounds. The statistical distribution is shown in Figure 4. Although the overall distribution of the two datasets is similar, our dataset contains more low-contrast samples. That is, the proportion of difficult targets in remote sensing dataset is larger, the number of simple targets is smaller, and the target detection algorithm has more difficulty with achieving success. Experiments show that our method has achieved the best detection effect in the two datasets.

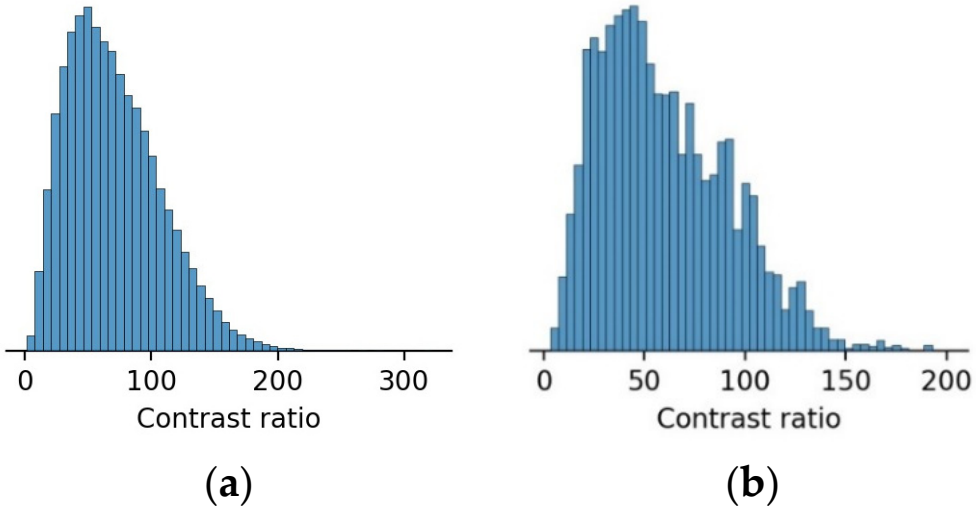

Figure 4. (**a**) The distribution map of contrast and quantity of targets in FLIR dataset; (**b**) the distribution map of contrast and quantity of targets in our dataset.

Some samples in the above dataset are shown in Figure 5.

Figure 5. Display of some images in the dataset mentioned in this article. (**a**) some images of FLIR dataset; (**b**) some images of our dataset.

3.2. Implementation Details

In this article, all experiments were implemented in Pytorch on a PC with Intel(R) Xeon(R) Silver 4210R CPU, NVIDIA RTX 3090 GPU. The PC operating system was Ubuntu 18.04.

The Stochastic Gradient Descent (SGD) algorithm is used for the optimizer. The initial learning rate is 0.002, the attenuation weight is 0.0001, and the momentum is 0.9. In addition, there are many other network parameters. In addition, other parameters and contrast threshold in this experiment are shown in Table 1. The m parameter of ResTNet refers to the number of convolution blocks in each stage, and the n parameter refers to the number of cross-attention modules. The parameters of CTEL in the Table 1 are used in

Formula 6, which are the calculation parameters of the mapping function of contrast and loss weight.

Table 1. Other parameters.

Parameter	Value
ResTNet.m_{1-4}	3, 4, 23, 3
ResTNet.n	3
CTEL.delta	100
CTEL.k,b	0.75, 0.1

3.3. Evaluation Metrics

Precision, Recall, and Mean Average Precision (mAP) were used to evaluate the detection performance of different methods. In addition, in order to further verify the effectiveness of the improved method, the detection performance of samples in each contrast range is evaluated. The precision and recall calculation methods are as follows:

$$\mathrm{P_{r\ }}_{ecision} = \frac{TP}{TP + FP} \tag{10}$$

$$\mathrm{R_{e\ }}_{call} = \frac{TP}{TP + FN} \tag{11}$$

where TP (True Positive) refers to the positive sample of correct detection, and FP (False Positive) refers to the positive sample of error detection. Similarly, TN and FN refer to negative samples detected correctly and false samples detected incorrectly.

In the field of target detection, the IoU as a criterion to predict whether it is a positive sample. The IoU formula is as follows:

$$IoU = \frac{P \cap T}{P \cup T} \tag{12}$$

where P and T refer to target area and real box area. All experiments in this paper use 0.5 as the IOU threshold in the evaluation process, that is, samples with IOU greater than 0.5 are considered positive samples.

AP refers to the area value of $P - R$ curve and the area surrounded by the coordinate axis. The closer the AP value is to 1, the better the detection of the algorithm. The calculation process can be summarized as follows:

$$AP = \int P(R)dR \tag{13}$$

The mAP represents the average value of various APs, which is used to fairly measure the detection performance of multi-class target detection tasks.

3.4. Analysis of Results

Firstly, ablation and contrast experiments are set on the FLIR datasets. Secondly, in order to verify the competitive of the proposed method, the performance comparison experiments of each method in different contrast intervals are designed. In addition, experimental results on our dataset show that our algorithm is also still competitive and robust for remote sensing datasets.

3.4.1. Experiments on the FLIR-ADAS Dataset

(1) Comparison with other state-of-the-art methods

On the FLIR-ADAS dataset, a series of comparative experiments and ablation experiments are designed to verify the advantage. Comparative methods include popular ones, the latest ones, and ones designed for detection task in infrared images. Popular

methods such as SSD, YOLOv5s, and Faster R-CNN are widely used in various target detection tasks because of their excellent performance. YOLOX and YOLOF are the most advanced methods in the field of target detection. In addition, MMTOD and ThermalDet have been designed by researchers for target detection in infrared images in recent years. The performance of the target detection methods above on the FLIR dataset is shown in Table 2.

Table 2. This is the performance of various advanced methods on FLIR dataset. Top2 is highlighted using red and green, respectively.

Method	Person	Car	Bicycle	mAP
SSD [24]	40.9	61.6	43.6	48.7
Faster R-CNN [17]	39.6	67.5	54.6	53.9
Retinanet [25]	52.3	71.5	61.3	61.7
FCOS [26]	69.7	79.7	67.4	72.3
MMTOD-UNIT [11]	49.4	70.7	64.4	61.5
MMTOD-CG [11]	50.2	70.6	63.3	61.4
RefineDet [27]	77.1	84.5	57.2	72.9
ThermalDet [9]	**78.2**	85.5	60.0	74.6
Cascade R-CNN [13]	77.3	79.8	84.3	80.5
YOLOv5s [14]	68.3	80.0	67.1	71.8
YOLOF [15]	67.8	79.4	68.1	71.8
YOLOX [16]	**78.2**	80.2	**85.4**	81.2
baseline	69.7	79.9	61.5	70.4
baseline + CTEL	74.9	84.0	75.7	79.2
baseline +ResTNet	77.8	**87.5**	83.6	**82.9**
ResTNet+ CTEL	78.0	**87.4**	**87.4**	**84.3**

The results above the dotted line in Table 2 are the comparative experiment of other methods, and the results below the dotted line are the ablation experiment of the proposed method. The two-stage structure in Figure 1 is utilized as the prediction framework. In addition, the baseline in the experimental part of this paper takes the ResNet50 network as the backbone and CIoU as regression loss function.

The ResTNet improves the ability of feature extraction for low contrast targets by integrating the context information and enhancing the target features. Compared with the baseline, ResTNet increased the mAP in the dataset by 13%. Because the ResTNet can not only act on low-contrast targets, but also enhances all targets, this has brought more obvious improvement effects. The CTEL improves the learning effect of low-contrast targets by balancing the reverse propagation of samples on the dataset, thus increasing its mAP by 9%. The proposed method contains ResTNet and CTEL, and the mAP reaches the best 84% on the FLIR dataset. The ThermalDet is the best detection method on the FLIR dataset for infrared design that can be found in public papers, but the mAP of our method is about 10% higher than it. The YOLOX is the most advanced method in the field of target detection, and the mAP of our method is about 3% higher than it.

For better visibility, drawing the yellow box is placed after drawing the green box to highlight the difference between the two. As shown in Figure 6, the subjective results of ablation experiments were given. The targets in the circle are that our method detects correctly and the basic method misses. Figure 6 shows that our method is much more capable of detecting infrared targets than the baseline.

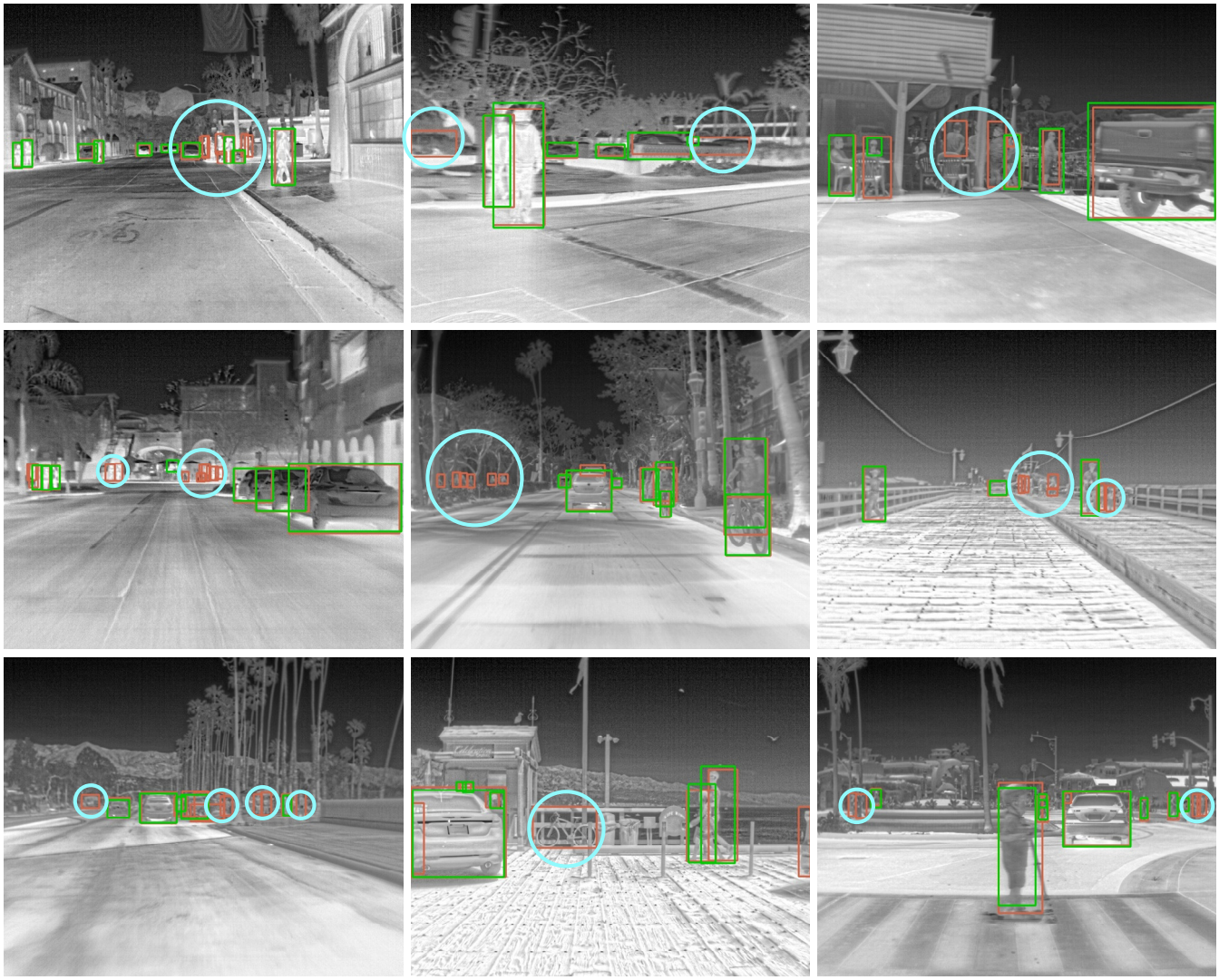

Figure 6. Our method detects results on some typical images in the FLIR dataset. The prediction box of the baseline is marked as green, and the prediction box of the proposed method is marked as orange-red.

The difference target in the images in the first row and the second column is the car. The difference targets in the images of the first and second columns of the third line contain cars and bicycles. The different targets in other images are pedestrians. The circled targets of the image in the middle of Figure 5 are low-contrast targets severely overlapped by the background, which is the difficult target to detect in the target detection task. Figure 5 shows that our method improves the feature extraction ability of usually difficult low-contrast targets by establishing the information path of image context.

To intuitively compare the detection performances of the proposed method and popular ones, we visualize the detection results of some methods in Figure 7. The Figure 7 shows some of the detection results from FCOS and YOLOv5s, as well as the methods proposed in this paper. The detection results of different categories were annotated with different colors. The person is labeled as red boxes, bicycles are labeled as green boxes, and cars are labeled as orange boxes.

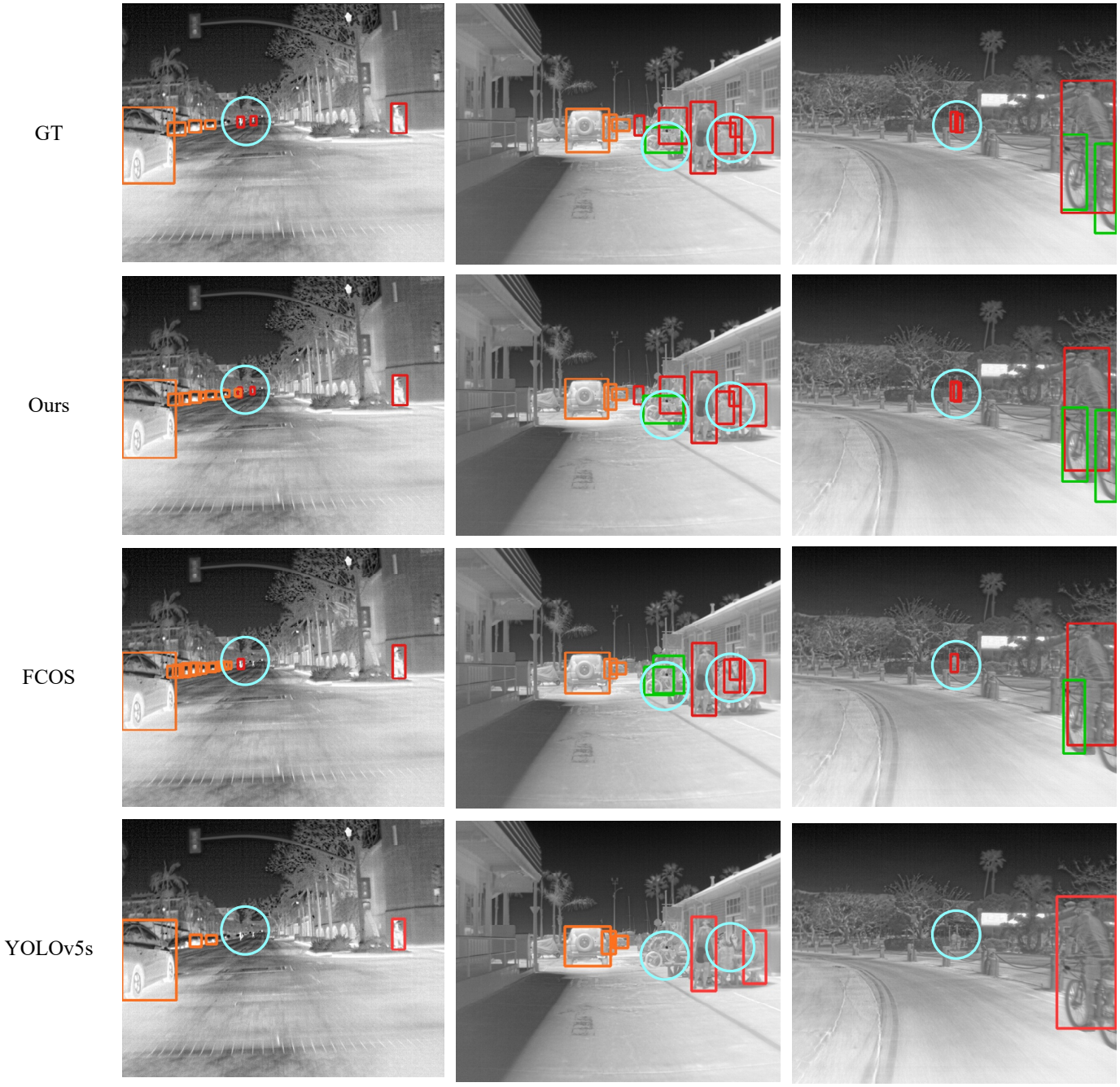

Figure 7. Detection effects diagram of each popular method. From top to bottom is the real label, the proposed method results, FCOS results, and YOLOv5s results.

The targets in the blue circle are low contrast samples, which are severely cluttered by the background and difficult to recognize even for humans. Figure 5 shows that the detection performance of our method on low-contrast targets is amazing. The proposed CTEL improves the detection performance of low-contrast samples by balancing the learning of difficult samples and simple ones. In addition, our ResTNet extracts target feature information better from a complex background.

(2) Comparison of different contrast intervals

To clearly compare the performance of each method on different contrast samples, three contrast intervals are divided. The target contrast less than 20 is defined as low contrast, between 20 and 120 is medium contrast, and more than 120 is high. The performance of some methods in different contrast ranges is compared, aiming to verify the advantage

of the proposed method on low-contrast samples. The mAP of samples with different contrast intervals is shown in Table 3.

Table 3. The mAP of different comparative samples in some methods are shown in the table. The Top2 in different contrast intervals in the table are highlighted with red and green.

Method	Low	Mid	High
SSD [24]	32.5	47.9	63.4
Retinanet [25]	44.8	64.4	79.5
FCOS [26]	57.4	75.1	86.6
Cascade R-CNN [13]	64.3	79.8	90.3
YOLOv5s [14]	61.2	77.8	86.1
YOLOF [15]	61.2	77.7	86.2
YOLOX [16]	68.2	79.0	86.3
baseline	61.3	76.7	87.2
baseline + CTEL	63.9	81.5	88.9
baseline + ResTNet	**71.1**	**84.1**	**91.7**
ResTNet + CTEL	**73.3**	**84.2**	**91.6**

The experimental results show that our method performs best on samples in all intervals compared with other comparison methods. Our method can effectively improve the detection rate of low-contrast targets.

Almost all methods can get ideal performance in a high contrast range, but poor performance in a low contrast range. Our method obviously performs better in low contrast targets, which indicates that the two proposed ResTNet and CTEL are effective.

The CTEL mainly improves the detection performance of low-contrast targets. Meanwhile, ResTNet comprehensively improves the detection performance of the contrast of each interval. At present, the most advanced target detection algorithm YOLOX has achieved great success in the field of optical images, and the overall performance exceeds other popular detection algorithms in almost all datasets. Table 3 shows that, compared with YOLOX, our method has stronger performance on low-contrast targets. In the low contrast range, the mAP of our method is about 5% higher than that of the YOLOX method, and has achieved grand achievements. In addition, compared with the base, the mAP in the low contrast range has achieved huge improvement of more than 10%.

The comparison results between the baseline and the proposed detection architecture are shown in Figure 8. The first line is the example with lower contrast targets, then the second and third lines are the examples with more moderate contrast targets and more high contrast targets, respectively. The green boxes in the graph are the true value. In addition, low-contrast targets are labeled by L, medium-contrast targets by M, and high-contrast targets by H. The blue box in Figure 8 is the detection result of the basic network, and the red box is the detection result of the architecture proposed in this paper.

For ease of observation, the differences between the proposed method and the baseline are labeled using a blue circle. As shown in Figure 8, most of the error targets in the baseline have small entropy difference with the background, which is difficult to distinguish and fall into the range of low contrast and medium contrast. After our improvement, the missing targets by the baseline are re-detected. This result is consistent with the data in Table 3.

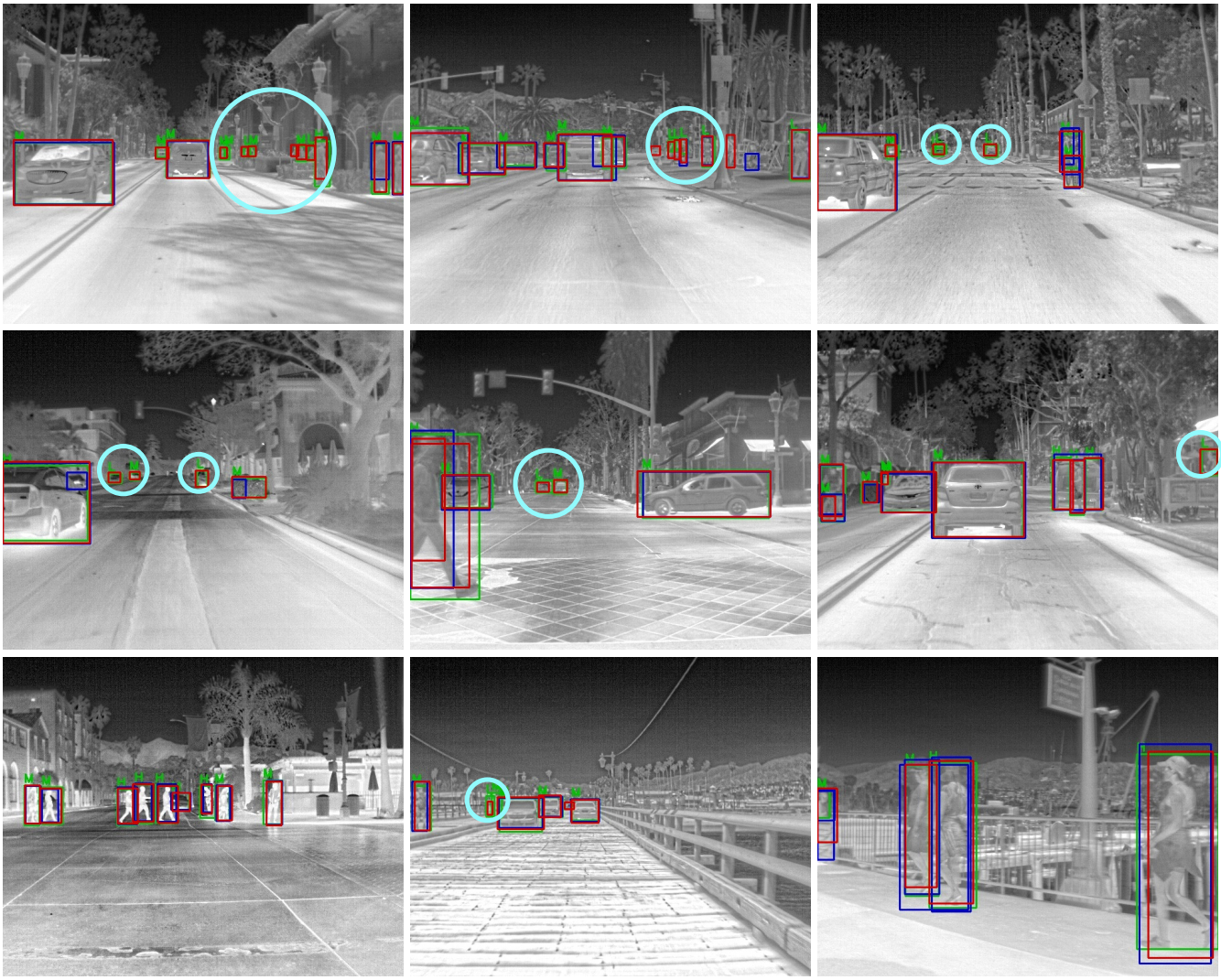

Figure 8. Schema checking results' comparison diagram between basic network and architecture proposed in this paper.

3.4.2. Experiments on the Remote Sensing Infrared Dataset

The loss of inherent features from some targets in remote sensing infrared images is more serious than that of general infrared image data. Our dataset has more challenges such as dense targets and wide-scale distribution. In this more difficult infrared dataset for remote sensing, our ResTNet and CTEL can still effective.

A series of comparative experiments on our infrared dataset for remote sensing is designed to demonstrate the robustness and advantage of the proposed method. The comparative experimental results are shown in Table 4. The results above the dotted line in Table 2 are the comparative experiment of other methods, and the results below the dotted line are the ablation experiment of the proposed method.

Table 4. The experimental results of various popular methods on our datasets are shown in tables. Top2's mAP metrics is highlighted in red and green.

Method	mAP
SSD [24]	13.9
Retinanet [25]	41.1
FCOS [26]	51.0
Cascade R-CNN [13]	58.4
YOLOv5s [14]	62.4
YOLOF [15]	52.5
YOLOX [16]	60.5
baseline	49.7
baseline + CTEL	59.8
baseline + ResTNet	**66.6**
ResTNet + CTEL	**67.2**

As shown in Table 4, our improvement brings great performance improvement for the baseline on the remote sensing dataset. By improving the ability of feature extraction and learning low-contrast targets, the mAP of this architecture on our remote sensing infrared data set is increased by about 18%, which is better than the most advanced method at present.

The detection results of our method and the baseline are shown in Figure 7.

In Figure 9, the blue circles are the difficult samples, such as the samples in the first row. These samples lack inherent characteristic information, which has a great impact on the performance of the detection method. Using our method, the context information path of the image is established, and the feature information of the target is better utilized. The proposed method detects many usually difficult low-contrast targets.

Experiments show that the proposed method improves the performance of detection on low contrast targets. Our method obtains the best performance on the two datasets after combining ResTNet and CTEL. The proposed target detection method for infrared images task is competitive and robust.

Figure 9. Our method detects results on some typical images in remote sensing datasets. The prediction box of the baseline is marked as green, and the prediction box of the proposed method is marked as yellow.

4. Discussion

As shown in Section 2, a novel target detection method is proposed, which improves the feature extraction ability of infrared image targets through attention modules about integration of context information. The feature thermal diagram of the backbone network is shown in Figure 10. The high temperature region is red, and the low is blue. The color in the heat map corresponds to the dependence of the prediction results on the characteristics of each region. The color red has the highest dependence, followed by yellow, and the lowest blue. It can be seen from Figure 10 that our method focuses its attention on the area of the target.

It can be seen from Figure 10 that, compared with the baseline, the hotspot distribution in the feature map constructed by the proposed method is more concentrated on the target. This proves that the feature extraction ability of this method for infrared images task is much stronger than that of the baseline.

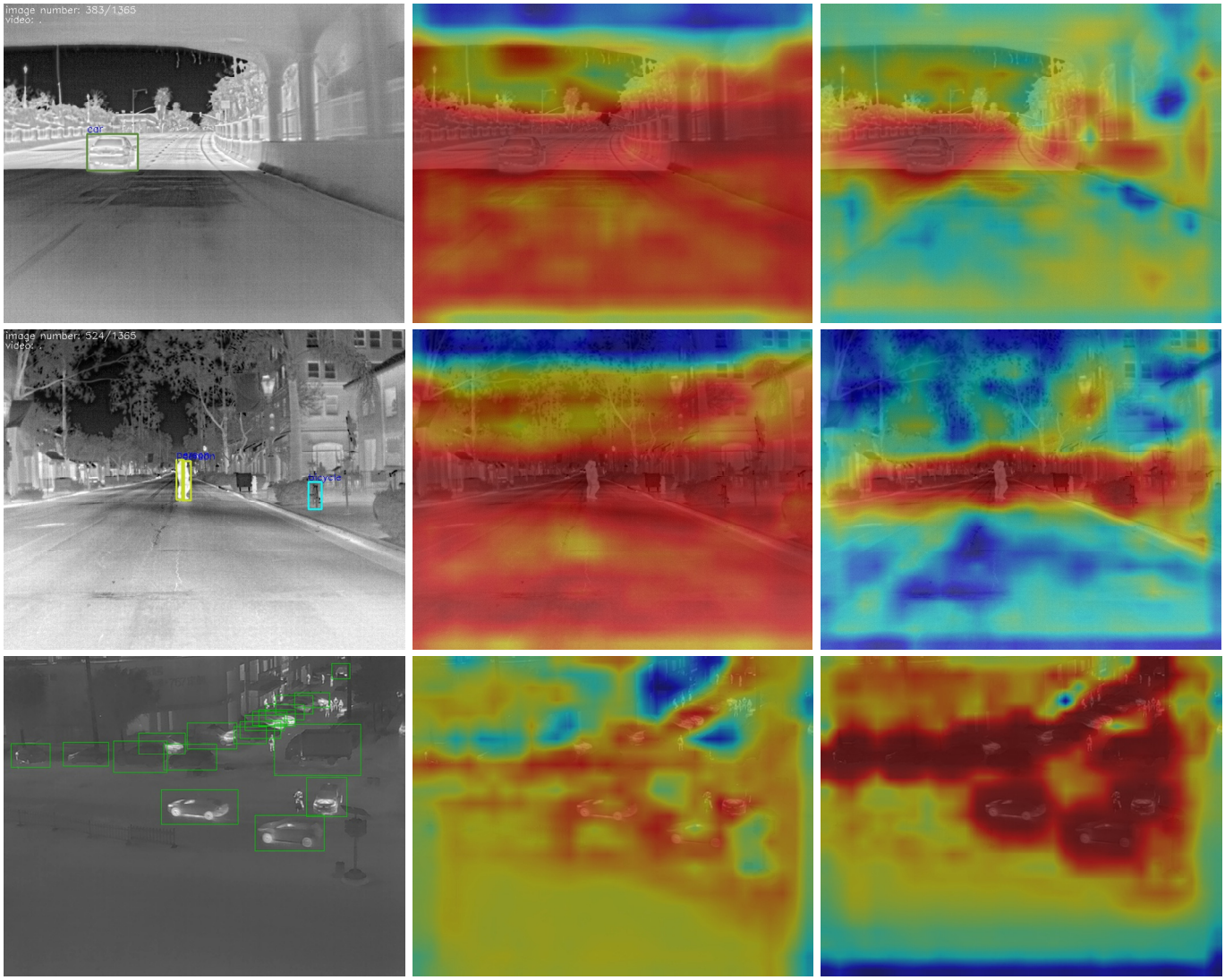

Figure 10. These are some images and feature maps extracted from them. The first of each line is the labeled infrared image. The second and third are the feature thermal images output by baseline and our method, respectively.

5. Conclusions

In this paper, aiming at the problems of low contrast and loss of inherent characteristics in IR images, we propose a novel detection method for IR images.

Firstly, to improve the feature extraction ability in IR images, a novel backbone network ResTNet is proposed. The ResTNet integrates context information by the Transformer structure in MSAN to improve the feature extraction ability of low-contrast targets in IR images.

Secondly, we design a contrast enhancement loss function based on the physical properties of IR images. The proposed loss function weights the loss values of low-contrast targets and other targets respectively, aiming to balance the backward propagation gradient of low-contrast samples and other samples in the network.

Extensive experiments on the FLIR dataset and our remote sensing dataset show that the proposed improvements can obtain encouraging results, especially for low contrast targets. The proposed method is SOTA, competitive, and robust. The proposed method can optimize the performance of different difficulty targets as a whole, but we have more optimization on the performance of low contrast targets. Compared with the baseline,

our method improves the performance of low contrast targets by about 20% on the FLIR dataset, but not so much on the simple targets.

Author Contributions: Conceptualization, C.Z. and N.S.; methodology, N.S. and J.W.; software, J.W.; validation, Y.Y. and C.Z.; formal analysis, C.Z.; data curation, Y.Y. and X.X.; writing—original draft preparation, J.W. and N.S.; writing—review and editing, N.S., X.X. and J.W.; funding acquisition, C.Z., Y.Y. and N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 61801142, No. 62071136, No. 61971153, No. 62002083) Heilongjiang Postdoctoral Foundation LBH-Q20085, LBH-Z20051, and the Fundamental Research Funds for the Central Universities Grant 3072021CF0814, 3072021CF0807, and 3072021CF0808.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The FLIR dataset is obtained from https://www.flir.cn/oem/adas/, accessed on 10 August 2021. The rest of the data was produced by ourselves, which is not publicly available.

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

References

1. 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]
2. Chen, Y.; Shin, H. Pedestrian Detection at Night in Infrared Images Using an Attention-Guided Encoder-Decoder Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 809. [CrossRef]
3. 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**. [CrossRef]
4. Deshpande, S.D.; Er, M.; Ronda, V.; Chan, P. Max-mean and max-median filters for detection of small-targets. *Proc. SPIE* **1999**, *3809*, 74–83.
5. He, Y.; Zhang, C.; Mu, T.; Yan, T.; Wang, Y.; Chen, Z. Multiscale Local Gray Dynamic Range Method for Infrared Small-Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1846–1850. [CrossRef]
6. Han, J.; Liu, S.; Qin, G.; Zhao, Q.; Zhang, H.; Li, N. A Local Contrast Method Combined with Adaptive Background Estimation for Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1442–1446. [CrossRef]
7. Zhang, L.; Peng, Z. Infrared small target detection based on partial sum of the tensor nuclear norm. *Remote Sens.* **2019**, *11*, 382. [CrossRef]
8. 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**. [CrossRef]
9. Cao, Y.; Zhou, T.; Zhu, X.; Su, Y. Every Feature Counts: An Improved One-Stage Detector in Thermal Imagery. In Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 6–9 December 2019; pp. 1965–1969. [CrossRef]
10. Zhang, R.; Xu, M.; Shi, Y.; Fan, J.; Mu, C.; Xu, L. Infrared Target Detection Using Intensity Saliency and Self-Attention. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 1991–1995. [CrossRef]
11. Devaguptapu, C.; Akolekar, N.; Sharma, M.M.; Balasubramanian, V.N. Borrow from Anywhere: Pseudo Multi-Modal Object Detection in Thermal Imagery. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1029–1038. [CrossRef]
12. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Attentional Local Contrast Networks for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 9813–9824. [CrossRef]
13. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
14. Glenn, J.; Alex, S.; Jirka, B.; NanoCode012; Christopher, S.; Liu, C.; Prashant, R. Yolov5. 2021. Available online: https://github.com/ultralytics/yolov5 (accessed on 29 August 2021).
15. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.I.; Cheng, J.; Sun, J. You Only Look One-level Feature. *arXiv* **2021**, arXiv:2103.09460.
16. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
17. 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] [PubMed]
18. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef] [PubMed]

19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
20. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *arXiv* **2020**, arXiv:1911.08287. [CrossRef]
21. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM international conference on Multimedia (MM'16), Amsterdam The Netherlands, 15–19 October 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 516–520. [CrossRef]
22. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
23. FA Group. Flir Thermal Dataset for Algorithm Training. 2018. Available online: https://www.flir.in/oem/adas/adas-dataset-form/ (accessed on 29 August 2021).
24. 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, Graz, Austria, 8 December 2015.
25. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007. [CrossRef]
26. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**, arXiv:1904.01355.
27. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212.