文章编号：1001-9014(2023)05-0701-10 DOI：10. 11972/j. issn. 1001-9014. 2023. 05. 015

ACE-STDN: An infrared small target detection network with adaptive contrast enhancement

YE Xin-Yi[1,2]， GAO Si-Li[2]， Li Fan-Ming[2*]

（1. School of Information Science and Technology，ShanghaiTech University，Shanghai 201210，China；
2. Key Laboratory of Infrared System Detection and Imaging Technology，Shanghai Institute of Technical Physics，
Chinese Academy of Sciences，Shanghai 200083，China）

Abstract：Due to the long distance and complex background，it is hard for the infrared detecting and tracking system to find and locate the dim-small targets in time. The proposed method，ACE-STDN，aims to tackle this difficult task and improve the detection accuracy. First of all，an adaptive contrast enhancement subnetwork preprocesses the input infrared image，which is conducive for the low-contrast dim targets. Next，a detection subnetwork with a hybrid backbone takes advantage of both convolution and self-attention mechanisms. Besides，the regression loss is designed based on 2D Gaussian distribution representation instead of Intersection over Union measurement. To verify the effectiveness and efficiency of our method，we conduct extensive experiments on two public infrared small target datasets. The experimental results show that the model trained by our method has a significant improvement in detection accuracy compared with other traditional and data-based algorithms，with the average precision reaching 93. 76%. In addition，ACE-STDN achieves outstanding detection performance in a multiclass object dataset and a general small object dataset，verifying the effectiveness and robustness.

Key words：infrared image，small target detection，deep learning

基于自适应对比度增强的红外小目标检测网络

叶昕怡[1,2]， 高思莉[2]， 李范鸣[2*]

（1. 上海科技大学 信息科学与技术学院，上海 201210；
2. 中国科学院上海技术物理研究所 红外探测与成像技术重点实验室，上海 200083）

摘要：红外探测系统需要尽早发现目标以便及时拦截，但是红外图像上的小目标检测是一个挑战十足的任务。为了提高检测准确率，提出一种基于自适应对比度增强的红外小目标检测方法。为了利用自注意力机制和卷积各自的优势，设计了一个高效的特征提取网络和一个面向小目标的检测头。同时为了解决实际应用中出现的弱目标，在检测子网络前添加了一个图像预处理子网络，该模块可以自适应地调节图像对比度。在红外空中小目标数据集上的实验表明，提出的方法能达到93.76%的检测精度，与经典的检测方法相比，能够更好地平衡检测精度和召回率，证明了方法的巨大应用潜力。

关 键 词：红外图像；小目标检测；深度学习

中图分类号：TP391.4 **文献标识码**：A

Introduction

Single-frame infrared small target detection（SIRST）aims at locating small targets from complex backgrounds using the infrared radiation difference，which is one of the research hot spots in many applications. Infrared acquisition technology surpasses visible and radar detecting systems in several aspects，i. e.，the

Received date：2022- 12- 19，**revised date**：2023- 07- 21 **收稿日期**：2022- 12- 19，**修回日期**：2023- 07- 21

Foundation items：Supported by the Youth Innovation Promotion Association CAS（2014216）；Supported by the National Pre-research Program during the 14th Five-Year Plan（514010405）.

Biography：YE Xin-Yi（1998-），female，Gansu，China. M. A. The research area involves small object detection and image processing. E-mail：xinyiye15@163. com

[*] **Corresponding author**：E-mail：lfmjws@163. com

strong shielding property of passive detecting, weather-protected long imaging range, and high sensitivity to boosting missiles plumes. Therefore, it is widely used in the military field, such as long-range precision strikes, aerospace defense confrontation, battlefield intelligence, and reconnaissance. Besides, it also makes remarkable achievements in remote sensing, medical imaging, and industrial flaw detection.

The bottleneck of the infrared acquisition system and its application lies in the capability to detect small-dim targets. It is a very challenging problem since an infrared small-dim target only contains less than 9 ×9 pixels (of a 256 ×256 image) and the images have low contrast and low signal-to-noise ratio (SNR). Therefore, the performance and efficiency of detecting small objects are far from satisfactory, and the technical roadblocks are as follows. (1) Targets lack structural features like fine texture and shape. (2) Small objects have similar characteristics to unpredictable background clutter. (3) Infrared imagery is frequently accompanied by smog and waves. (4) There are a few public infrared small target datasets.

Up to now, the research in the single-frame detection domain can be categorized into four groups. Filter-based algorithms [1-2] concentrate on assuming and suppressing the background, or using the frequency difference between target and clutter to approach saliency detection. Methods based on human visual system (HVS) [3-4] imitates that human eyes can be attracted by the target in an image. LCM [3] refers to the contrast mechanism of HVS and measures the dissimilarity between the current location and its neighborhoods. Algorithms based on low-rank assumption [5-6] exploit the property of non-local self-correlation of the global infrared patch image. IPI [5] assumes that the target patch-image is a sparse matrix and the background patch-image would be a low-rank matrix. The small target detection task is then transformed into an optimization problem of recovering the low-rank and sparse matrices. Although differences exist in imaging using the standard scopes and the infrared scope, deep learning is driven by data and concentrates on data distribution. Therefore, when facing the infrared small target detection task, we can draw on the state-of-the-art networks in the visible field.

Small object detection has been a hot topic in computer vision for years. Multiple methods are proposed to solve this difficult but important task. The first kind of method goes deep into multi-scale feature learning. Feature pyramid network (FPN) [7] assembles the spatial information from low levels and semantic information from high levels to strengthen the features of targets. The receptive fields of small objects are smaller than big objects, so TridentNet [8] focuses on designing three dilated convolutions with different dilation rates to construct multi-size detection branches. The second type uses generative adversarial networks (GAN) to generate high-resolution images [8] or high-resolution features [10-11]. The generator generates fake high-resolution images or feature maps from low-resolution ones, and the discriminator takes charge to discriminate between the fake and the real. Context-based methods [12-14] dig into the contextual information and the relationship between the small object and its neighbors which are easier to detect. Another way to promote the performance of small object detection is by adding special designs to generic object detection architectures. S^{3}FD [15] develops a scale compensation anchor matching strategy and a max-out background label to improve the recall rate and reduce the false positive rate. FaceBoxes [16] proposes a new anchor densification strategy ensuring different types of anchors have the same density on the image, which significantly improves the recall rate of small faces. Meanwhile, some researchers observe that Intersection over Union (IoU), the most widely used metric in object detection, is sensitive to slight offsets between predicted bounding boxes and ground truths when detecting tiny objects. Thus, new metrics are proposed to improve the performance of anchor-based detectors, e. g. , GIoU [17], DIoU [18], CIoU [18], DotD [19], and NWD [20].

In recent years, some awesome learning-based methods are proposed for infrared small target detection. MDvsFA-cGAN [21] proposes a conditional GAN comprising two generators that utilize the different sizes of receptive fields. The dense nested attention network (DNA-Net) [22] contains a tri-directional dense nested interactive module (DNIM) to achieve progressive feature interaction and adaptive feature enhancement. Dai [23] designs a plug-in module named asymmetric contextual modulation (ACM), which can encode the smaller scale visual details into deeper layers. In EAAU-Net [24], an enhanced asymmetric attention (EAA) module is designed to improve performance using cross-layer feature fusion and spatial-channel information exchange between the same layers. Liu [25] adopts the self-attention mechanism of the transformer to learn the interaction information in a larger range, which is the first work to explore the transformer to detect the infrared small-dim target.

This work is motivated by the following thought that we gain from SIRST tasks. Due to the limited pixels of the targets, producing high-quality detection performance requires an efficient method to extract features and make full use of the information from the infrared image. However, such satisfactory performance may not be sufficiently achieved by a single detection network, especially when the small target is dim as well. Inspired by the traditional background suppression-based methods, we find that a better approach may be adding an adjunctive preprocessing module to reduce noise and improve contrast.

Following the above idea, we propose an end-to-end deep learning framework to improve the performance of SIRST. In this framework, the pipeline can be divided into two phrases, i. e. , enhancing the contrast and detecting small targets. Two deep neural networks are constructed to focus on the two stages respectively. The contrast enhancement subnetwork works to improve the SNR and the relative local thermal contrast adaptively, then its output passes to the small target detection subnetwork. Existing CNN-based detection networks can extract features by convolution within a small neighborhood, but the limited receptive field makes it hard to capture global dependency. Another framework, Transformer, becomes more popular and dominates among various vision tasks for the capacity to capture long-range dependency and learn global contextual information via self-attention. However, it induces costly computation at the same time, which means a pure transformer structure is not a wise choice. In addition, these state-of-the-art networks are designed for generic image datasets. Directly using them for infrared small target detection can fail catastrophically due to the large difference in the data distribution. Thus, it is of great importance to re-design the structure to handle infrared small target detection tasks. To make the best use of these two dominant frameworks and alleviate their respective deficiencies, we design a hybrid detection subnetwork that integrates the self-attention mechanism into deep layers and applies a novel transformer-styled convolution block. In this way, the network can extract long-range information and overcome redundancy and dependency simultaneously.

The contributions of this paper can be summarized as follows. **Firstly**, we propose a novel framework for infrared small object detection, using a two-stage learning paradigm. Compared with the existing learning-based methods that use a single network for detection, our approach operates in favor of the dim targets. **Secondly**, an adaptive contrast enhancement subnetwork is proposed to preprocess the input infrared image by suppressing the complex background and highlighting the target. **Thirdly**, a hybrid backbone is designed in the small target detection subnetwork, which is beneficial to find and locate small targets under difficult circumstances of long distances and complex backgrounds. Besides, this backbone proves the availability and superiority of the mix of both self-attention and convolution. **Last**, we use a loss function that measures the similarity between bounding boxes of tiny objects by the distance of their corresponding Gaussian distributions instead of IoU-based measurement series.

1　Method

In this section, we introduce the proposed infrared dim-small target detection pipeline at length. Ordered by the workflow, the contrast enhancement subnetwork adjusts the thermal contrast adaptively to suppress the complex background and highlight the target. Then the hybrid backbone extracts features combining the advantages of both convolution and self-attention mechanisms. At last, the small-target-oriented detector predicts infrared dim-small targets with two detection heads based on feature maps from different layers. Besides, we apply a novel regression loss to elevate accuracy and speed up convergence. Figure 1 illustrates the framework of ACE-STDN.

1.1　Contrast Enhancement Preprocessing

The image captured by the infrared imaging system usually has a low signal-to-noise ratio, and lacks in the relative local thermal contrast, causing some small targets with indistinctive characteristic. These infrared dim-small targets bring difficulties in object detection. To tackle this challenge, we design an adaptive contrast enhancement subnetwork (ACESN) to preprocess the input infrared image before detection. Given an infrared image $I \in \mathbb{R}^{W \times H}$, the preprocessing procedure can be modeled as:

$$\hat{I} = \mathcal{F}(I, \theta), \tag{1}$$

where $\hat{I} \in \mathbb{R}^{W \times H}$ represents the enhanced infrared image. $\mathcal{F}$ is the enhancement network with trainable parameter θ, which is illustrated in Fig. 2.

The ACESN can be divided into three modules. First of all, the multi-level feature extraction module (MLFEM) is a simple 4-layer CNN, while in each convolution layer, the kernel is 3 × 3 in size and 1 in stride. Besides, it applies ReLU as the activation function. The input of MLFEM is the low-contrast infrared image, and each feature map is the input to its corresponding feature enhancement sub-module as well as the input of the next layer. Secondly, the branch-independent enhancement module (BIEM) is composed of 4 feature enhancement sub-modules. The output of each branch is an enhanced image $I_i \in \mathbb{R}^{W \times H}$. Each sub-module has an identical symmetric architecture, operating downsampling and upsampling. Except for the kernel size, these layers have the same settings in MLFEM, stride 1, and ReLU nonlinearity. The last one is the fusion module (FM), which concatenates the 4 output images from BIEM to produce the final enhanced image $\hat{I}$ using a 1 × 1 convolution kernel. This merging equals a weighted sum with learnable weights.

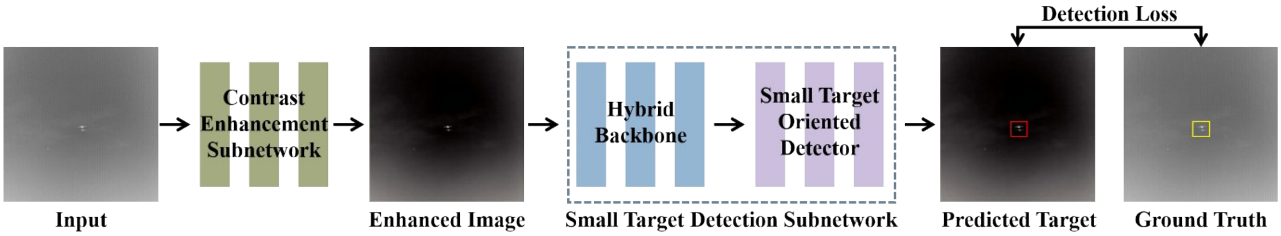
Detection Loss
Contrast Enhancement Subnetwork
Hybrid Backbone
Small Target Oriented Detector
Input
Enhanced Image
Small Target Detection Subnetwork
Predicted Target
Ground Truth

Fig. 1　The training pipeline of the proposed ACE-STDN framework. Our method consists of two subnetworks to preprocess the infrared image and detect small targets respectively. The contrast enhancement subnetwork aids the small target detection subnetwork to achieve better performance, especially for dim targets.
图 1　本文提出的ACE-STDN的网络框架与训练流程

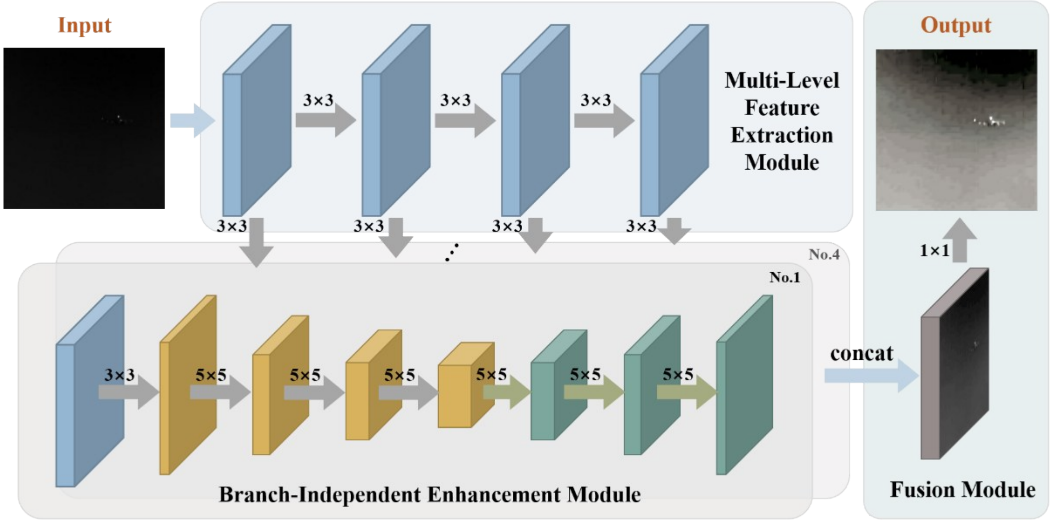

Fig. 2 The adaptive contrast enhancement subnetwork for infrared images. This network consists of three main modules, where gray arrows denote convolution layers, and the green ones are deconvolution layers
图2 红外图像自适应对比度增强网络

1.2 Hybrid Feature Extraction Network

The feature extraction network, known as the backbone, is the bedrock of detection tasks. It conducts convolution on input images and provides concise semantic information for the subsequent detectors. However, in most classic networks, the convolution block has a relatively small receptive field, which leads to poor performance and needs a proper solution. Impressed by the effectiveness of vision transformer, we propose a novel hybrid backbone (HB) with transformer-styled convolution (TSConv) blocks and a transformer encoder block.

Instead of simply connecting transformer encoders with convolution blocks, we design the TSConv block to uniform them. Following the instruction in ConvNeXt[26], we modify the structure of the classic ResNet block at first. In the transformer encoder block, the hidden dimension of the MLP block is wider than the input dimension, which forms an inverted bottleneck. Therefore, we alter the ResNet from a bottleneck structure to an inverted bottleneck structure by rearranging the convolutions. Moreover, we use the depthwise convolution to imitate the weighted sum operation in the self-attention mechanism, which only mixes the information in the spatial dimension. The proposed TSConv block contains a 3 × 3 depthwise convolution followed by two 1 × 1 convolutions, and each operation merely mixes the information across one dimension, spatial-wise or channel-wise. In Fig. 3., (a) and (b) illustrate the differences between TSConv Block and ConvNeXt Block, which make the module more suitable in infrared dim-small target detection, e. g., the 7×7 convolution is too large to maintain and transmit the information of a tiny target. Additionally, the proposed TSC3 module (in Fig. 3 (c)) imitates the structure of the C3 module in YOLOv5, which con-

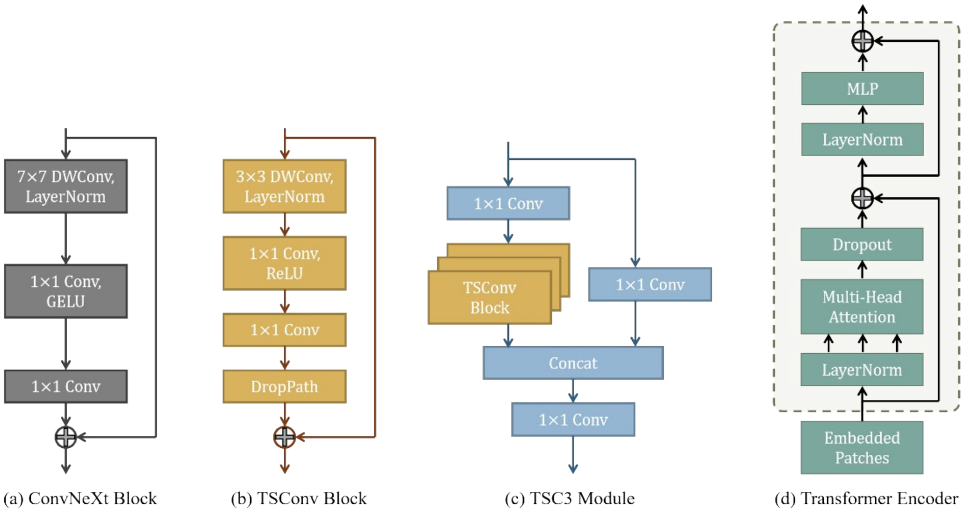

Fig. 3 The structure of the transformer encoder block and TSConv block.
图3 特征提取网络中组件的结构

catenates three stacking TSConv blocks with a standard convolution.

We apply a transformer encoder block to capture long-range dependency and learn global contextual information. As shown in Fig. 3（d）, each transformer encoder block contains two sub-units, including the multi-headed self-attention（MSA）, and two fully-connected layers with a GELU non-linearity（MLP）. LayerNorm（LN）is applied before each sub-unit, and residual connections are used after each sub-unit. Transformer encoder block increases the ability to capture global dependency. It also explores the feature representation potential via the self-attention mechanism.

Because the self-attention mechanism is inefficient to encode local features in the shallow layers. It simply captures detailed visual features, resembling the feature extraction result of convolution. Moreover, the self-attention applied on high-resolution shallow feature maps brings a large and unnecessary computation burden. In contrast, the convolution operation focuses on local dependency within a small neighborhood. It can obtain similar features in the shallow layers while reducing local redundancy and avoiding unnecessary computation. Therefore, as shown in Fig. 4. , the proposed TSC3 modules are used in the inception phase of the feature extraction network, and the transformer encoder block is located at the end.

1.3 Small Target-Oriented Detector

In the definition by SPIE, an infrared dim-small target only occupies less than 9 × 9 pixels of a 256 × 256 image. After multiple feature extraction layers, the feature map fed into the prediction head lacks enough information of small objects. Therefore, we need to redesign the prediction head using low-level, high-resolution feature maps.

First of all, we use a weighted bi-directional feature pyramid network（BiFPN）[27] as a substitute for FPN and PAN in YOLOv5 to make the best use of the feature maps extracted by the HB at different stages. As shown in Fig. 4, BiFPN deletes those nodes with only one input edge, because these kind of nodes have less contribution to features fusion. Besides, BiFPN adds an extra edge from the original input to the output node if they are at the same level, which aims to fuse more features without adding much cost.

Secondly, in order to improve the detection performance for infrared dim-small targets, we apply the coordinate attention mechanism to adaptively enhance multi-level features, which embeds smaller-scale details into high-level coarse feature maps. The coordinate attention embeds positional information into channel attention, aggregating features along the two spatial directions. One spatial direction captures long-range dependencies and the other preserves precise position information. In this way, the resulting feature maps can be applied to the input feature map to augment the representations of the infrared dim-small targets.

At the end of the detection pipeline, we remove the detection head for large-scale and medium-scale objects in YOLOv5. At the same time, we conduct one more upsampling and add a new prediction head for tiny infrared targets, illustrated in Fig. 5. Compared with YOLOv5, our framework adds a detection head for tiny objects, and removes two heads for medium and large size objects. The blue arrows and orange ones stand for convolution and de-convolution respectively. It is worth noting that we omit some connections and concatenation in this figure to explain and highlight our main modification. The new prediction head is generated from a low-level, high-resolution feature map, which is more sensitive to dim-small targets in an infrared image.

1.4 2D Gaussian Distribution Regression Loss

Interfered with atmospheric scattering, atmospheric refraction, optics compensation, etc. , the signals of small objects received by the infrared imaging system are extremely weak. As illustrated in Fig. 6, the 3D intensity distribution of a dim-small infrared target shows that the shape is centrosymmetric and the intensity decreases in concentric circles, which is similar to the 2D Gaussian distribution function. Kim [28] models the dim-small infrared targets as 2D Gaussian distributions:

$$f_T(r) = f_{T0}(x,y) = \lambda exp\left(-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right]\right), \quad (2)$$

where σ_x and σ_y are the scale parameters of horizontal and vertical respectively. λ is the gray value of the object. And $f_T(r)$ denotes the gray-level spatial cumulative distribution function of this dim-small target.

Inspired by the formula above, we investigate the infrared dim-small target dataset and find that these small instances do not correspond to the rectangle shape perfectly. A bounding box contains both target and back-

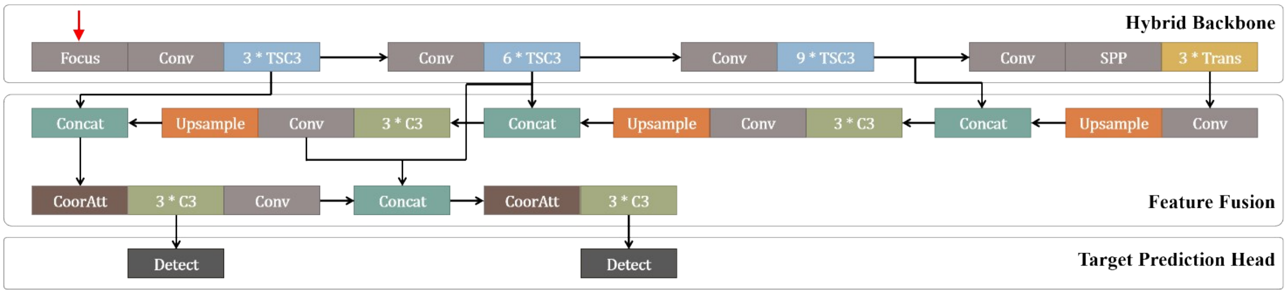

Fig. 4 The architecture of the detection subnetwork
图 4 目标检测子网络的网络架构

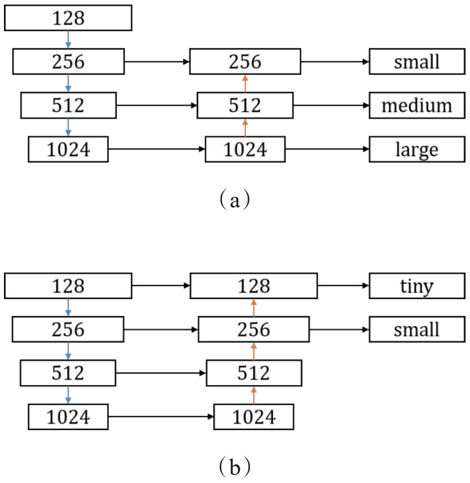

Fig. 5 Two different frameworks: (a) the framework of YOLOv5; (b) our improved framework
图 5 两种不同的检测框架:(a) YOLOv5 的检测框架;(b) 本文改进的检测框架

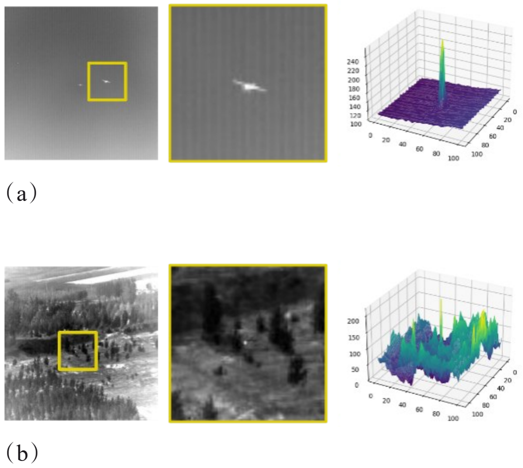

Fig. 6 Infrared small-dim targets in the real world and their local intensity distribution: (a) simple background; (b) complex background
图 6 真实红外弱小目标的领域强度分布:(a) 简单背景;(b) 复杂背景

ground information, while the background information is distributed near the boundaries. Therefore, we can use a 2D Gaussian distribution to describe the bounding box, modeling the importance of pixels inside the bounding box by weights. Use a random bounding box $R = (cx, cy, w, h)$ as an example, where (cx, cy), w and h denote the center coordinates, width, and height respectively. Its inscribed ellipse can be represented as:

$$\frac{\left(x-\mu_x\right)^2}{\sigma_x^2}+\frac{\left(y-\mu_y\right)^2}{\sigma_y^2}=1, \tag{3}$$

where $\left(\mu_x,\mu_y\right)$ is the center coordinates of the ellipse, σ_x and σ_y are the lengths of semi-axes along x and y axes. Accordingly, $\mu_x = cx$, $\mu_y = cy$, $\sigma_x = \frac{w}{2}$, $\sigma_y = \frac{h}{2}$. This ellipse will be a density contour of the 2D Gaussian distribution. Therefore, the mentioned bounding box R can be modeled as:

$$\mathcal{N}\left(\mu,\Sigma\right) \qquad \mu=\begin{bmatrix} cx \\ cy \end{bmatrix} \quad \Sigma=\begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix}, \tag{4}$$

Therefore, we can measure the similarity between two bounding boxes by the distribution distances of their corresponding Gaussian distributions, replacing the common measurement, IoU. Both measurements are illustrated in Fig. 7.

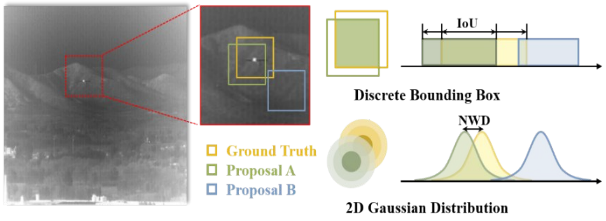

Fig. 7 The schematic diagram of measurements using a discrete bounding box and 2D Gaussian Distribution
图 7 使用分布距离衡量包围框相似性

Following the Ref. [20], we use the Wasserstein distance as our regression loss function, which comes from optimal transport theory. In the first place, we represent the predicted bounding box and the ground truth in the form of 2D Gaussian distribution, $\mathcal{N}_p\left(\mu_p,\Sigma_p\right)$ and $\mathcal{N}_{gt}\left(\mu_{gt},\Sigma_{gt}\right)$. Then we can put forward the 2^{nd} order Wasserstein distance between $\mathcal{N}_p$ and $\mathcal{N}_{gt}$:

$$W_2^2\left(\mathcal{N}_p,\mathcal{N}_{gt}\right) = \left\|\left(\left[cx_p,cy_p,\frac{w_p}{2},\frac{h_p}{2}\right]^{\mathrm{T}},\left[cx_{gt},cy_{gt},\frac{w_{gt}}{2},\frac{h_{gt}}{2}\right]^{\mathrm{T}}\right)\right\|_2^2, \tag{5}$$

Next, the distance is changed into its exponential form in which the value is constrained within [0, 1]:

$$NWD\left(\mathcal{N}_p,\mathcal{N}_{gt}\right) = \exp\left(-\frac{\sqrt{W_2^2\left(\mathcal{N}_p,\mathcal{N}_{gt}\right)}}{13}\right), \tag{6}$$

Therefore, the regression loss function can be redesigned as:

$$\mathcal{L}_{NWD} = 1 - NWD\left(\mathcal{N}_p,\mathcal{N}_{gt}\right), \tag{7}$$

The main advantage of Normalized Wasserstein Distance (NWD) is that it can provide a gradient for optimizing the network in two special cases. One is when there is no overlap between the predicted bounding box and the ground truth. The other is the predicted box contains the ground truth box completely or vice versa. Therefore, NWD-loss is suitable for our infrared dim-small target detector.

2 Experiments

In this section, we evaluate the effectiveness of our method in the scenarios under complex backgrounds. First of all, we describe the experimental setting, which includes the datasets, evaluation metrics, network implementation details and comparison methods. Next, we conduct ablation studies to examine the effectiveness and practicability of each module in our proposed framework. Then, the visual and numerical comparison between ACE-STDN and state-of-the-art methods further demonstrates that ACE-STDN can accurately detect infrared small targets. Finally, we show detection results on visible images to verify the generalization ability of our ACE-STDN.

2.1 Experimental Setting

Dataset. We use the public dataset [29] for small infrared moving target detection under clutter background to build our training set, validation set, and test set. This dataset contains a group of data for one or multiple fixed-wing UAV targets via outfield recording and data post-processing. The data acquisition scenario covers sky background and complex field background. This dataset includes 22 image sequences, 30 trajectories, 16177 frames and 16944 targets. The size of images is 416 × 416. In addition, we conduct experiments on the dataset published in Ref. [23], in which the training set has 8525 images and the test set has 545 images. For the sake of convenience, all the label formats are converted into YOLOv5-form, and we denote above two datasets as *ATDT* and *SIRST* respectively.

Evaluation Metrics. The evaluation metrics used in this paper are Precision, Recall, F_1-score, and Average Precision (AP). Precision represents the credibility of detection results, while Recall reflects whether the detection algorithm locates all the infrared small targets. F1-score is used to measure the relationship between them.

$$F_1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} , \tag{8}$$

$$Precision = \frac{TP}{TP + FP} , \tag{9}$$

$$Recall = \frac{TP}{TP + FN} , \tag{10}$$

TP means that a positive target is predicted as real. FP means that a negative target is predicted as real. TF means that a positive target is predicted as fake.

Implement Details. We adopt the training protocol of YOLOv5 in our proposed ACE-STDN algorithm. During training, the data augmentation methods like image flipping, mosaic, and random perspective are applied to expand the training dataset. Our ACE-STDN approach is trained by the Adam optimizer with 200 epochs. The start learning rate is 10^{-3} and the batch size is 8. ACE-STDN predicts the bounding boxes at two scales (tiny and small), and three anchors at each scale. We implement ACE-STDN on Pytorch 1.11.0 and run it on the NVIDIA TITAN V for training and testing.

Comparison Methods. This paper chooses seven detection methods based on deep learning to compare with ACE-STDN in two aspects. The first group aims to prove the availability and superiority of the combination of self-attention mechanism and convolution operation. We compare our method with a pure CNN-based network - YOLOv5, a pure transformer framework - ViT [30], and a method combined both convolution and self-attention mechanisms - TPH-YOLOV5 [31]. In the second group, we select a background suppression-based method TopHat [1], an HSV-based method LCM [3], two optimization-based method IPI [5] and NRAM [6].

2.2 Ablation Study

To examine the effectiveness of each module in our proposed framework, we conduct ablation experiments on different settings, including hybrid backbone, adaptive contrast enhancement module, and 2D Gaussian Distribution Regression Loss. As shown in Table 1, we conduct quantitative mAP evaluation on module selection. By combining the three modules, Model D (ACE-STDN) obtains the best result, demonstrating the effectiveness of the proposed improvements. Compared with Model A which uses CSPDarknet53 as backbone, Model B which uses CIoU in regression loss, and Model C which is a straight detection framework, the detection results of Model D (ACE-STDN) are the best. Model A and D demonstrate the superiority of the proposed hybrid backbone with TSConv blocks. This feature extraction structure can capture global dependency and have strong feature representation potential using the inverted bottleneck. Although the AP only increases 0.74% using NWD-based loss, its main contribution is to decrease the missing alarm rate and increase the positional precision of infrared small targets. From Model C to Model D, ACESN promotes the detection rate of dim targets by adjusting the contrast, which make ACE-STDN suitable and robust for more circumstances.

Table 1 Ablation study on ATDT
表1　消融实验结果

Model	HB	ACESN	NWD	AP
A		√	√	88.25%
B	√	√		93.02%
C	√		√	92.48%
D	√	√	√	93.76%

2.3 Comparison

In this section, we compare ACE-STDN with state-of-the-art methods through visual and numerical evaluation to further verify the effectiveness of our method.

The effectiveness of hybrid usage of self-attention and convolution. As shown in Fig. 8, we selected seven typical infrared small target scenes and compared the detection results with three other methods. In case (c) and case(f), ACE-STDN can detect small-dim infrared targets precisely compared to YOLOv5. A low false alarm is of great importance in the real application. ViT has a poor performance in this detection task, especially in case(a), in which the target has a relatively big size and notable intensity among the dataset. In the hard case (e) and (f), where the target is both small and dim, ACE-STDN shows outstanding detecting ability and stability.

According to Table 2, it is clear that the mean precision of our method achieves 93.76%, which surpasses all other detection models. For example, it outperforms the original YOLOv5s by 6.49%. Although TPH-YOLOv5 shows the feasibility of unifying convolution and self-attention compared with ViT, it sacrifices its computational efficiency greatly. As small targets dominate the infrared dataset, ACE-STDN with its customized modules has significant advantages. The major one is the appropriate placement of convolution and self-attention blocks. In shallow layers, TSConv can effectively solve the problem that information is lost during the downsampling. The transformer encoder block is integrated into the last layer of the HB to extract long-range information and overcome redundancy and dependency.

Table 2 Comparison with generic detection method on ATDT

表2 不同机制的检测方法在ATDT数据集上的表现

Model	Average Precision	Inference Time
YOLOv5	87.27%	5.3 ms
ViT	59.63%	5.5 ms
TPH-YOLOv5	74.95%	8.7 ms
ACE-STDN	93.76%	4.8 ms

Comparison with specialized methods for infrared small target detection. As shown in Fig. 9, four methods are selected to compare the detection results in different scenes. From the figures, we can observe that filter-based TopHat is the most sensitive one to cluster. And if the target is an area target, both TopHat and LCM fail to work. The detection result of NRAM is not satisfactory and the predicted results are relatively small compared to the ground truth, which is because of the various constraints guided by prior knowledge. Although our proposed method is not a segmentation algorithm, it can accurately detect and locate targets of different sizes.

In order to illustrate the effectiveness of ACE-STDN straightforwardly, we use numerical methods for quantitative evaluation. From Table 3, we can see that our method makes a balance between accurate detection and reduction of false alarm. On the contrary, classic methods exhibit high Precision and low Recall, because these methods suppress the background and targets at the same time. For those dim targets, the ACE module in our method prevents them from being flooded with background clusters.

2.4 Supplementary Result on Multiclass Object detection

In addition to evaluating ACE-STDN on infrared datasets containing targets of just one type, the model is also trained on the multiclass dataset to verify its outstanding detection ability. The results are shown in Fig. 10-11. In addition, ACE-STDN can detect and classify objects of different sizes, modalities, and densities, which shows its robustness and generalization ability. In Fig. 10, ACE-STDN successfully distinguishes three flying objects under different complex background. Fig. 11 shows the generality in visible dataset. In the challenging situation with size-diverse planes or helicopters, ACE-STDN outputs correct detection and classification predictions. This is useful in practical application when

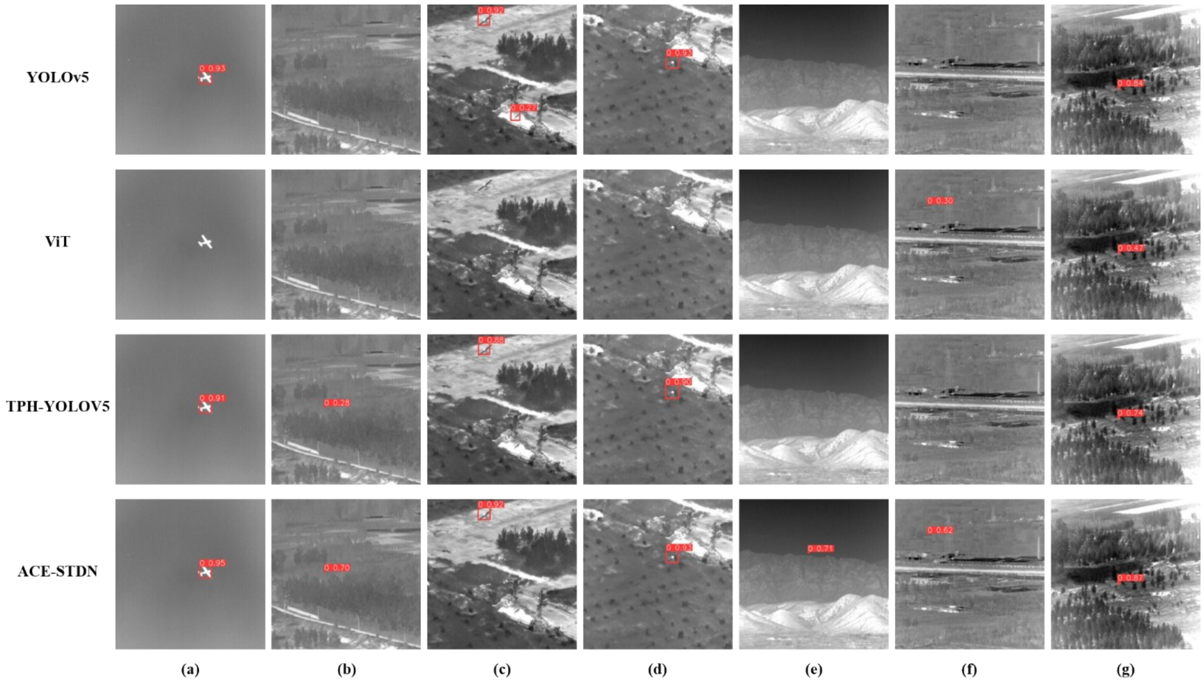

Fig. 8 Illustration of detection results on ATDT

图8 在ATDT数据集上的检测结果

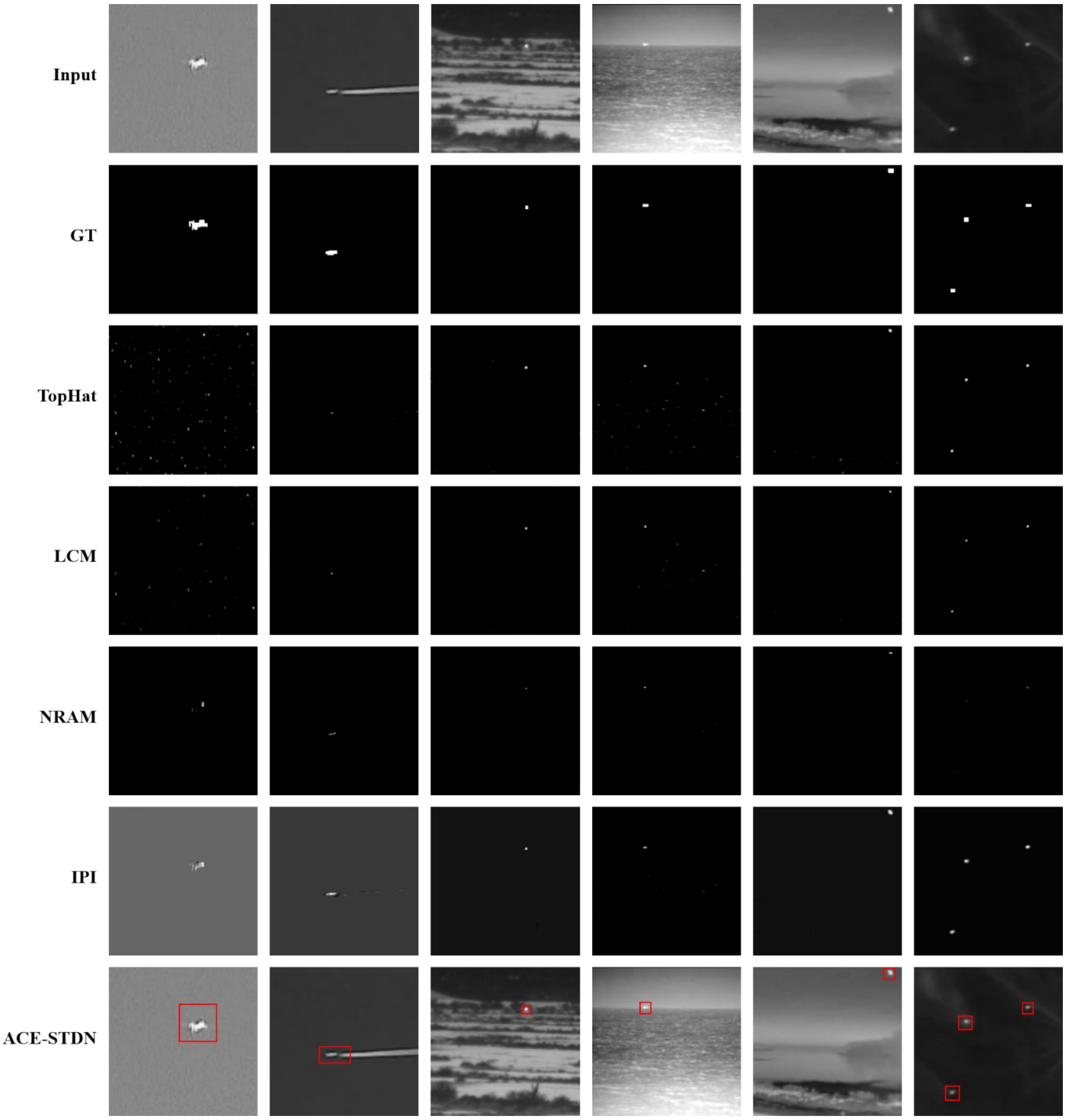

Fig. 9　Illustration of detection results on SIRST
图 9　在SIRST数据集上的检测结果

Table 3　Comparison with generic detection method on SIRST
表 3　不同的红外小目标检测方法在SIRST数据集上的表现

Method	Precision	Recall	F_1-score
TopHat	0. 6873	0. 0818	0. 1461
LCM	0. 6201	0. 1443	0. 2341
NRAM	0. 7549	0. 1544	0. 2563
IPI	0. 7640	0. 1813	0. 2931
ACE-STDN	0. 8537	0. 8362	0. 8448

close targets appear in the camera's sight.

3　Conclusion

We proposed a new infrared target detection model ACE-STDN with contrast enhancement adaptively. With the benefit of preprocessing subnetwork, it can work well in detecting dim small targets in infrared images. In addition, the hybrid backbone is designed to improve feature representation, which proves that proper mix-usage of self-attention and convolution is superior to the pure mechanisms. And the 2D Gaussian distribution-based regression loss is suitable for infrared small target detection concerning the relative position between two bounding boxes. Evaluations on both single-class and multiclass datasets, both infrared and visible datasets demonstrate the outstanding performance of ACE-STDN as it achieves a better balance between precision and recall. In summary, ACE-STDN provides a new choice for small-dim target detection in IP systems. In the future, we plan to speed up the network for real-time detection tasks.

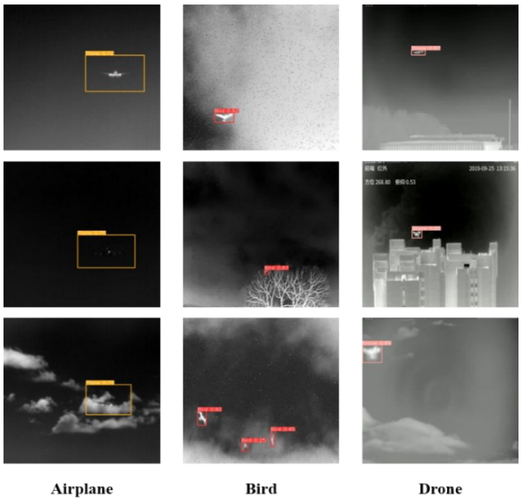

Fig. 10 Illustration of detection results on a multiclass infrared dataset

图 10 在多类别红外数据集上的检测结果

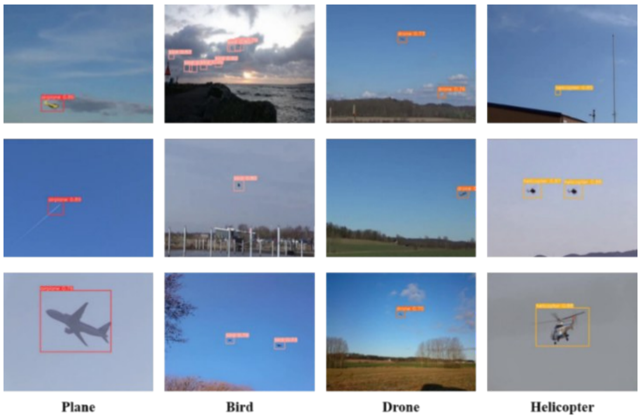

Fig. 11 Illustration of detection results on a multiclass RGB dataset

图 11 在多类别可见光数据集上的检测结果

References

[1] Tom V T, Peli T, Leung M, *et al.* Morphology-based algorithm for point target detection in infrared backgrounds[C]. Signal & Data Processing of Small Targets. International Society for Optics and Photonics, 1993.

[2] Wang X, Peng Z, Zhang P, *et al.* Infrared Small Target Detection via Nonnegativity-Constrained Variational Mode Decomposition [J]. *IEEE Geoscience and Remote Sensing Letters*, 2017, **14**(10): 1700 - 1704.

[3] Chen, Philip C L, Li H, *et al.* A Local Contrast Method for Small Infrared Target Detection [J]. *IEEE Transactions on Geoscience & Remote Sensing*, 2014, **52**(1): 574 - 581.

[4] Bai X, Bi Y. Derivative Entropy-Based Contrast Measure for Infrared Small-Target Detection[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, PP (99): 1-15.

[5] Gao C, Meng D, Yang Y, *et al.* Infrared Patch-Image Model for Small Target Detection in a Single Image[J]. *IEEE Transactions on Image Processing*, 2013, **22**(12): 4996 - 5009.

[6] Zhang L, Peng L, Zhang T, *et al.* Infrared Small Target Detection via Non-Convex Rank Approximation Minimization Joint l2, 1 Norm [J]. *Remote Sensing*, 2018, **10**(11): 1821.

[7] Lin T Y, Dollar P, Girshick R, *et al.* Feature Pyramid Networks for Object Detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 936-944.

[8] Li Y, Chen Y, Wang N, *et al.* Scale-Aware Trident Networks for Object Detection[C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6053-6062.

[9] Bai Y, Zhang Y, Ding M, *et al.* SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network[C]. European Conference on Computer Vision (ECCV), 2018.

[10] Li J, Liang X, Wei Y, *et al.* Perceptual Generative Adversarial Networks for Small Object Detection[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 1951-1959.

[11] Noh J, Bae W, Lee W, *et al.* Better to Follow, Follow to Be Better: Towards Precise Supervision of Feature Super-Resolution for Small Object Detection[C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9724-9733.

[12] Bell S, Zitnick C L, Bala K, *et al.* Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2874-2883.

[13] Hu H, Gu J, Zhang Z, *et al.* Relation Networks for Object Detection [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 3588-3597.

[14] Xu T, Du D K, He Z, *et al.* PyramidBox: A Context-assisted Single Shot Face Detector[C]. European Conference on Computer Vision (ECCV), 2018.

[15] Zhang S, Zhu X, Lei Z, *et al.* S^{3}FD: Single Shot Scale-invariant Face Detector[C]. IEEE International Conference on Computer Vision (ICCV), 2017: 192-201.

[16] Zhang S, Zhu X, Lei Z, *et al.* FaceBoxes: A CPU real-time face detector with high accuracy[C]. IEEE International Joint Conference on Biometrics (IJCB), 2017: 1-9.

[17] Yu J, Jiang Y, Wang Z, *et al.* UnitBox: An Advanced Object Detection Network[C]. ACM Proceedings of the 24th ACM international conference on Multimedia, 2016: 516-520.

[18] Zheng Z, Wang P, Liu W, *et al.* Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

[19] Xu C, Wang J, Yang W, *et al.* Dot Distance for Tiny Object Detection in Aerial Images[C]. IEEE/CVF Computer Vision and Pattern Recognition Workshops (CVPRW), 2021: 1192-1201.

[20] Xu C, Wang J, Yang W, *et al.* Detecting Tiny Objects in Aerial Images: A Normalized Wasserstein Distance and a New Benchmark[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021, **190**: 79-93.

[21] Wang H, Zhou L, Wang L. Miss Detection vs. False Alarm: Adversarial Learning for Small Object Segmentation in Infrared Images[C]. IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[22] Li B, Xiao C, Wang L, *et al.* Dense Nested Attention Network for Infrared Small Target Detection[J]. *IEEE Transactions on Image Processing*, 2022.

[23] Dai Y, Wu Y, Zhou F, *et al.* Asymmetric Contextual Modulation for Infrared Small Target Detection[C]. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021: 950-959.

[24] Tong X, Sun B, Wei J. EAAU-Net: Enhanced Asymmetric Attention U-Net for Infrared Small Target Detection [J]. *Remote Sensing*, 2021, 13.

[25] Liu F, Gao C, Chen F, *et al.* Infrared Small-Dim Target Detection with Transformer under Complex Backgrounds[J]. *arXiv e-prints*, 2021.

[26] Liu Z, Mao H, Wu C Y, *et al.* A ConvNet for the 2020s[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 11966-11976.

[27] Tan M, Pang R, Le Q V. EfficientDet: Scalable and Efficient Object Detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 10778-10787.

[28] Sungho Kim, Joo-Hyoung Lee. Robust Scale Invariant Target Detection Using the Scale-Space Theory and Optimization for IRST[J]. *Pattern Analysis & Applications*, 2011, **14**: 57-66.

[29] Hui B, Song Z, Fan H, *et al.* A Dataset for Infrared Image Dim-Small Aircraft Target Detection and Tracking under Ground / Air Background[DS/OL]. *Science Data Bank*, 2019.

[30] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]. International Conference on Learning Representations. 2021.

[31] Zhu X, Lyu S, Wang X, *et al.* TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios[C]. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021: 2778-2788.