基于跨模态数据增强的红外时敏目标检测技术

王思宇，杨小冈*，卢瑞涛，李清格，范继伟，朱正杰

(火箭军工程大学 导弹工程学院，陕西 西安 710025)

摘　要：目前红外时敏目标检测技术在无人巡航、精确打击、战场侦察等领域应用广泛，但有些高价值目标图像的获取难度高且成本昂贵。针对红外时敏目标图像数据匮乏、缺少用于训练的多场景多目标数据、检测效果不佳等问题，文中提出一种基于跨模态数据增强的红外时敏目标检测技术，跨模态数据增强方法为两阶段模型。首先在第一阶段通过基于 CUT 网络的模态转换模型将包含时敏目标的可见光图像转换为红外图像，其次在第二阶段模型中引入 coordinate attention 注意力机制，随机生成大量红外目标图像，实现了数据增强效果。最后提出一种基于 SE 模块和 CBAM 模块改进的 Yolov5 目标检测架构，实验结果表明，文中提出的 Yolov5(CSP-A) 目标检测技术与原网络相比，准确率提升了 7.36%，召回率提升了 5.43%，平均精度提升了 2.74%。有效提高了红外时敏目标的检测精度，实现了红外时敏目标精确检测。

关键词：红外时敏目标；　数据增强；　模态转换；　目标检测

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.3788/IRLA20220876

0　引　言

红外时敏目标是指打击机会受限于时间窗口，且具有极高军事价值的舰船飞机等红外目标。红外时敏目标检测技术在无人巡航、精确打击、战场侦察等领域应用广泛。为满足红外时敏目标检测精度需求，基于深度学习的方法得益于强大的算力、深层网络结构以及大量的标注数据在目标检测领域[1]取得了巨大进展。由于具备大量的可见光遥感数据集，当前的时敏目标检测研究主要集中在可见光领域[2]，受限于数据获取难度较大、标注成本较高，针对红外时敏目标检测的研究较少，而通过对数据进行处理生成“新数据”，则成为扩大数据集同时提高模型泛化能力的一项重要手段[3]。

研究人员通过设计合理的神经网络模型结构，利用大量已标注的数据集计算损失函数，从而实现对目标任务的数据挖掘，通过对模型参数进行迭代优化，最终得出基于任务的深度学习模型。数据作为深度学习的驱动力，在目标检测模型训练中起到至关重要的作用，数据增强作为一种常规的增加训练数据的手段，可以有效防止模型在训练过程中的过拟合问题，并且在一定程度上提高了模型的检测精度以及泛化能力。

目前较多领域存在数据集规模较小、分布不均匀等情况。有些高价值目标图像的获取难度高且成本昂贵，为解决此类问题，部分学者对原始样本进行数据增强，从而扩充数据集目标多样性及丰富度[4]。在图像数据增强技术中，如何在扩充宏观数据集数量的同时丰富其目标微观特征数量，则成为了研究的主要关注点。

传统的数据增强方法主要有几何变换、颜色变换等有监督的数据增强方法。通过平移、旋转、缩放、裁剪、噪声、模糊、填充等方式实现数据集中的样本增强。Taylor 等人[5]将图像裁剪运用到包含 101 类目标的数据集中，将精度提升了 13.82%。Zhong 等人[6]提出了一种基于随机擦除的数据增强方法，使得深度学习模型学习更深层次的特征。Ma 等人[7]将椒盐、

收稿日期：2022−12−06；　**修订日期：**2023−02−15

基金项目：国家自然科学基金项目 (62276274)；航空科学基金项目 (201851U8012)

作者简介：王思宇，男，博士生，主要从事视觉导航、目标检测、图像处理等方面的研究。

导师(通讯作者)简介：杨小冈，男，教授，博士生导师，主要从事视觉导航、目标检测、图像处理等方面的研究。

高斯等噪声加入到训练集中进行图像分类训练，结果表明，该方法对遥感分类任务并没有取得明显的精度提升。因此可以发现基于数据变形的数据增强方法虽然操作简单，但是在复杂任务场景下对深度学习模型的效能提升有限。

基于深度学习的智能数据增强方法主要体现在生成对抗网络，通过模型学习生成新的训练数据，从而产生更好的模型。Gulrajani 等人提出了 WGAN[8] 从而解决了模型训练过程中的梯度消失问题，使得生成的图像更加真实。为了去除数据集所带的不平衡性，Zheng 等人[9] 提出 DCGAN 作为一种数据增强工具，模型从大多数类中学习有效特征并为少数类生成图像。通过神经风格迁移进行数据增强，可以通过选择一组 k 个风格，并将它们应用于训练集中的所有图像。Zhong 等人[10] 利用 CycleGAN 将标记的训练图像进行风格转换，并与原始训练样本一起形成增强训练集。

传统的目标检测算法主要利用人工特征提取的方法，因此获取的图像信息较为片面，针对背景复杂的场景检测效果不佳，自从 Alexnet 在计算机视觉领域取得成功，卷积神经网络在图像目标检测领域取得了巨大的进展。现阶段基于深度学习的目标检测算法主要包括两类：单阶段检测和两阶段检测，其中单阶段检测算法主要为 SSD[11] 和 Yolo 系列[12] 算法，两阶段算法主要包括 R-CNN[13]、Fast R-CNN[14] 等。

目前基于深度学习的红外目标检测算法相对较少，时敏目标为金属壳体，表面具有附加涂层，通过反射太阳光产生辐射，目标表面温度会高于背景温度，因此红外目标具有较强的红外特性。由于成像质量受天气影响，成像视角较高，红外图像中的时敏目标具有外观模糊、细节信息丢失严重、边界不清晰等特性。因此，常规的深度学习目标检测算法效果较差。Ju 等人[15] 提出了一种端到端网络 ISTDet，该方法将图像滤波与目标检测相结合，在抑制背景的同时增强目标的响应。Yao 等人[16] 以图像序列的形式加入时域特征，使网络能够学习图像序列中的时空相关特征，从而实现了红外目标实时检测。Lu 等人[17] 将局部对比度机制与信杂比的计算相结合，在增强图像中疑似红外弱小目标区域的同时也提高图像的信杂比。Jiang 等人[18] 提出一种 scSE-IYOLOv4 的目标检测算法，通过在 YOLOv4 主干网络中嵌入 scSE 模块提高了目标检测精度。

针对红外时敏目标图像数据匮乏、缺乏颜色和纹理特征导致检测效果较差等问题。文中提出了一种基于跨模态数据增强的红外时敏目标检测技术，首先跨模态数据增强两阶段模型的第一阶段将包含多种时敏目标的可见光图像迁移为红外图像数据，第二阶段则在此基础上对单张红外图像进行生成式模型训练，实现样本随机生成。然后在 Yolov5 模型中引入 SE 模块和 CBAM 模块，增强红外时敏目标的特征提取。与同类模型相比可以发现，文中算法有效提升了红外时敏目标的检测准确率。

文中的主要贡献有：

1) 提出了一种跨模态红外时敏目标数据增强方法，通过将风格迁移模型与目标生成式模型相结合，利用可见光图像数据集实现红外时敏目标数据增强。

2) 提出一种基于 coordinate attention 注意力机制的生成器结构，增强图像目标的特征提取，同时丰富了目标的细节纹理，从而实现随机红外时敏目标样本生成。

3) 提出了一种改进的 Yolov5 目标检测模型，在 CSP 网络中增加 SE 和 CBAM 注意力机制，增强了网络的特征表达，更好的实现红外时敏目标检测。

1 跨模态数据增强两阶段模型

目前，数据增强算法被广泛应用在基于地面视角的可见光图像领域，然而由于空地红外图像领域的特殊性，针对包含时敏目标的空地红外图像数据集较为匮乏，相关的增强算法研究也极其有限。尽管在部分基于深度学习的目标检测和图像分类任务中使用几何变换、颜色变换等有监督的数据增强方法进行辅助训练，然而，利用可见光遥感图像数据集实现空地红外时敏目标数据增强的相关研究未见报道。

文中提出了一种跨模态红外时敏目标数据增强两阶段模型。图 1 描述了两阶段模型的整体过程，主要包括可见光图像到红外图像的模态转换网络，以及针对单张图像的目标数据增强网络。在第一阶段中，文中利用模态转换模型将包含时敏目标的可见光遥感影像转化为红外图像，从而生成初步的红外空地图

像数据集；在第二阶段中，文中将转换后的红外图像输入到多尺度金字塔结构中，实现了多尺度下的对抗性随机样本生成。在忽略背景影响的同时，随机生成多目标样本，从而实现红外时敏目标数据增强。

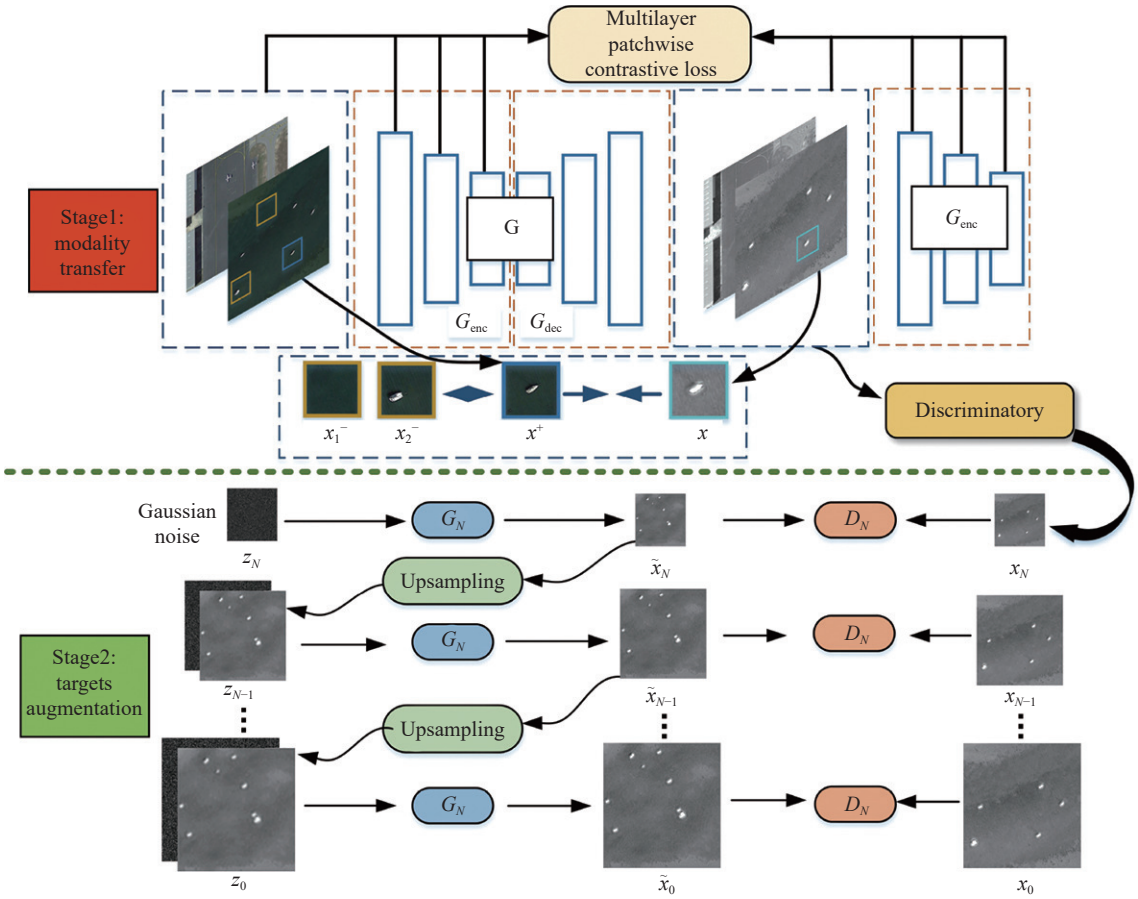


图 1 红外时敏目标数据增强两阶段模型概述

Fig.1 Overview of two-stage model for IR time-sensitive target data augmentation

1.1 可见光红外图像模态转换模型

不同感光元件对不同波长电磁波的敏感程度不同，红外传感器能够捕获物体的热辐射能量，因此红外图像可以根据辐射差异将目标与其背景区分开来。而可见光传感器可以通过接收物体的反射光来对场景细节、纹理特征等进行描述，因此可见光图像具有高分辨率和清晰度的纹理细节。文中利用基于深度学习的图像风格迁移算法实现从可见光图像到红外图像的模态转换。模态转换的效果取决于用于训练的可见光-红外数据集，利用不同波段的红外数据集训练，模态转换的效果则随之改变。

CUT(Contrastive Learning for Unpaired Image-to-Image Translation) 模型通过将对比学习的思想引入 CycleGAN 当中，使该框架能够在未配对的图像到图像的风格迁移中实现简单快速的单边转换。得益于他们的工作，文中推导出了一个用于可见光红外图像模态转换的网络。

1.1.1 模态转换网络架构设计

完整的可见光红外模态转换过程如图 1 中的阶段一所示。该模型仅学习单方向的模态转换映射关系，结合对比学习框架将输入域和目标域之间的互信息最大化，在一定程度上降低训练时间的同时提高训练效率。

文中生成器模型使用的是一种简单的编码解码架构，被分为两个部分编码器G_{enc}和解码器G_{dec}。将输入域为$\mathbb{A} \subset \mathbb{R}^{H\times W\times C}$的可见光图像转换为目标域为$\mathbb{B} \subset \mathbb{R}^{H\times W\times 3}$的伪红外图像数据，从而实现初步的红外数据增强，未配对的数据集为可见光数据$A = \{a \in \mathbb{A}\}$，红外数据$B = \{b \in \mathbb{B}\}$，目标域的输出图像可以表示为：

$$\hat{b} = G(z) = G_{dec}(G_{enc}(a)) \tag{1}$$

式中：G_{enc}为生成器中的编码器；G_{dec}为解码器；a表示输入域中的可见光图像；$\hat{b}$表示生成的伪红外目标与图像。

1.1.2 损失函数设计与分析

可见光红外模态转换的损失函数设计具体步骤如下：

1) 对抗损失。利用对抗损失函数使得生成的伪红外图像在视觉上更接近目标域中的真实红外图像。对抗损失函数如公式 (2) 所示。

$$L_G(G,D,A,B)=\mathbb{E}_{b\text{-}B}\log D(b)+\mathbb{E}_{a\text{-}A}\log(1-D(G(a))) \quad (2)$$

式中：G为生成器；D为判别器；$\mathbb{E}$为期望。

2) 对比损失。利用一个噪声对比估计框架 (Noise Contrastive Estimation, NCE)[19] 来最大化输入和输出之间的互信息，该框架包含 3 个信号，查询样本、正样本以及负样本。在输入图像a和生成图像$\hat{b}$的相同位置设置图像块分别为正样本$x^+\in\mathbb{R}^K$，查询样本$x\in\mathbb{R}^K$，同时在输入图像中随机选取N个负样本图像块$x^-\in\mathbb{R}^{N\times K}$，$x_n^-\in\mathbb{R}^{N\times K}$表示第$n$个负样本。

通过上述方法构造对比学习中的正负样本从而实现互信息的求取，通过使查询样本与正样本相互关联、与负样本形成对比实现互信息最大化。利用生成样本、正样本以及N个负样本计算交叉熵损失函数如公式 (3) 所示，其中τ取常数 0.07。

$$\ell(x,x^+,x^-)=-\log\left[\frac{\exp(x\cdot x^+/\tau)}{\exp(x\cdot x^+/\tau)+\sum_{n=1}^{N}\exp(x\cdot x^-/\tau)}\right] \quad (3)$$

该模型的目标是在一个特定的位置匹配相应的输入-输出图像块，同时可以利用输入中的其他图像块作为负样本。文中选择了L层感兴趣层，并通过一个两层的 MLP(Multilayer Perceptron) 网络H_l传递特征图。对比损失函数公式如公式 (4) 所示：

$$L_C(G,H,A)=\mathbb{E}_{a\text{-}A}\sum_{l=1}^{L}\sum_{s=1}^{S_l}\ell\left(\hat{z}_l^s,z_l^s,z_l^{S\setminus s}\right) \quad (4)$$

式中：$z_l=H_l(G_{enc}^l(a))$，$l\in\{1,2,\cdots,L\}$为每层的索引值；$s\in\{1,\cdots,S_l\}$，其中S_l为每一层中的空间位置数。相对应的特征为$z_l^s\in\mathbb{R}^{C_l}$，其他特征为$z_l^{S\setminus s}\in\mathbb{R}^{(S_l-1)\times C_l}$，$C_l$为每层的通道数，同时将输出图像$\hat{b}$编码为$\hat{z}_l=H_l(G_{enc}^l(G(a)))$。

该模型的总损失函数表达式为：

$$L=L_G+L_C(G,H,A)+L_C(G,H,B) \quad (5)$$

1.2 对抗性随机样本生成模型

在所有的深度生成模型方法中，由 Goodfellow 等人[20] 提出的生成对抗网络 (Generative Adversarial Networks, GAN) 是非常具有代表性的方法之一。GAN 通过一种无监督式的方法进行学习训练。其主要由两个神经网络模块构成。生成器 G 和判别器 D，两组模型通过博弈学习从而生成全新的数据。

1.2.1 多尺度生成对抗网络架构

SinGAN 无条件生成模型，不同于传统的生成对抗网络模型，该模型仅利用单张图像进行训练，从而获取图像的内部结构纹理信息。利用单张图像的不同尺度，实现多层生成器以及判别器架构的模型训练，其整体结构如图 1 中的阶段二所示。

该模型的训练和推理都以一种由粗尺度到细尺度的方式进行。

首先构建多尺度生成器$\{G_0,\cdots,G_N\}$，然后对图像$x:\{x_0,\cdots,x_N\}$由粗尺度到细尺度依次进行模型训练，其中生成的低尺度图像通过一个因子r上采样至高尺度图像。图像样本的生成从粗尺度开始，然后依次通过所有生成器，达到细尺度图像生成，其中每个尺度的输入中都包含噪声。结构中的生成器和判别器具有相同的感受野，因此随着金字塔结构的进程，所能提取到的特征也由整体结构信息变为细节信息，该模型在获取图像的全局属性同时学习了图像的细节纹理信息。生成样本图像$\tilde{x}_n$表示如下：

$$\begin{cases}\tilde{x}_n=G_n(z_n) & n=N\\ \tilde{x}_n=G_n(z_n,(\tilde{x}_{n+1})\uparrow^r) & n<N\end{cases} \quad (6)$$

式中：G_n为第初始层的生成器；z_n为高斯噪声；$\uparrow^r$为图像上采样。

在金字塔的首层即$n=N$，仅使用噪声作为输入生成样本图像，由于该层中的有效感受野通常为输入图像一半，因此G_n主要用于生成图像的总体布局和目标的全局结构。在其他层中即$n<N$，不同尺度的生成器G_n都添加了以前尺度没有生成的细节。除了噪声空间z_n，每个生成器G_n增加一个来自上层尺度图像的上采样版本，模型训练过程如算法 1 所示。

算法 1：样本图像生成训练

Input：输入图像x_n，输入噪声z_n，多尺度架构层数$N+1$，第n层生成器G_n第n层判别器D_n

Output: 生成图像$\tilde{x}_n$

1: **for** $n \leftarrow N$ to 0 **do**
2: **if** $n = N$ **then**
3: 初始层生成的图像为: $\tilde{x}_n = G_N(z_N)$
4: 损失函数计算公式如下:
5: $L = L_{\text{adv}}(G_N, D_N) + \alpha \|G_N(z_N^{rec}) - x_N\|^2$
6: **else**
7: 生成第n层的图像为:
8: $\tilde{x}_n = G_n(z_n, (\tilde{x}_{n+1})\uparrow^r)$
9: 损失函数计算公式如下:
10: $L = L_{\text{adv}}(G_n, D_n) + \alpha \|G_n(z_n^{rec}, (\tilde{x}_{n+1}^{rec})\uparrow^r) - x_n\|^2$
11: **end if**
12: **end for**

1.2.2 生成器模型改进

针对红外时敏目标数据增强而言, 用于训练的原始数据均来自于公开遥感数据集, 图像中各类舰船飞机等时敏目标的特征不明显。为了提取到更多的图像细节信息, 主要对用于特征提取的生成器结构进行改进。图 2 给出了改进后的生成器结构。由左向右进行样本图像生成, 将噪声图像z_n和上一尺度生成并上采样后的图像$\tilde{x}_{n+1}\uparrow^r$输入一个由3×3卷积、批归一化和 LeakyReLU 激活函数构成的卷积层, 然后通过一个 coordinate attention 模块[21]后, 经过 3 个相同的卷积层, 再经过一个 coordinate attention 模块, 最后经过一个3×3卷积和一个 Tanh 激活函数后输出生成的样本图像。在最粗的尺度文中选用了 32 个通道, 每上升 4 层, 则通道数变为原来的 2 倍。每个尺度都存在一个和生成器相对应的判别器, 它的结构与生成器相似, 包括 5 个卷积核大小为3×3全卷积层, 通道数与生成器相同。

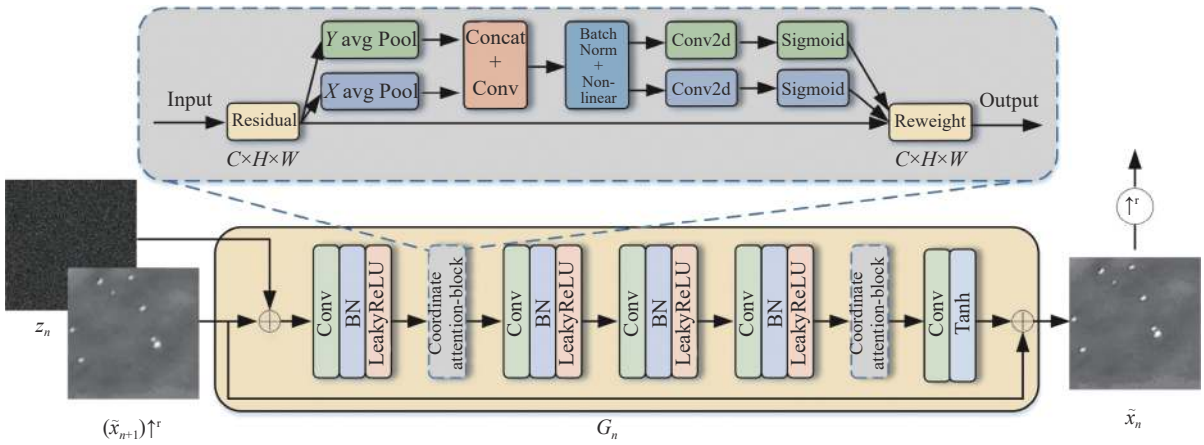


图 2 单尺度生成模型结构

Fig.2 Single-scale generative model structure

1.2.3 模型训练

方法从最粗糙的尺度到最精细的尺度依次训练我们的多尺度金字塔架构, 每一层被训练好后模型参数将会固定不再改变, 第n层的网络损失函数由对抗损失和重建损失构成, 目标函数如公式 (7) 所示:

$$\min_{G_n} \max_{D_n} L_{\text{adv}}(G_n, D_n) + \alpha L_{\text{rec}}(G_n) \tag{7}$$

式中: L_{adv}为对抗损失; L_{rec}为重建损失; G_n和D_n分别为第n层的生成器和判别器; α为重建损失的权重。

该方法选择$\{z_N^{\text{rec}}, z_{N-1}^{\text{rec}}, \cdots, z_0^{\text{rec}}\} = \{z^*, 0, \cdots, 0\}$作为输入噪声图生成原始图像$x$, 其中$z^*$为固定的噪声图, 重建损失函数计算公式如下:

$$L_{\text{rec}} = f(x) = \begin{cases} \left\|G_n(0, (\tilde{x}_{n+1}^{\text{rec}})\uparrow^r) - x_n\right\|^2 & n < N \\ \|G_n(z^*) - x_n\|^2 & n = N \end{cases} \tag{8}$$

2 红外时敏目标检测技术

2.1 目标检测网络架构

为了展现所提出的跨模态数据增强两阶段模型的有效性。文中采用基于通道注意力机制改进的 Yolov5(CSP-A) 目标检测模型对红外时敏目标进行检测。目标检测网络如图 3 所示。

改进的 Yolov5s(CSP-A) 网络模型主要包含以下结构: Input、Backbone、Neck 和 Prediction。Input 模块主要进行数据增强, 自适应图像尺度变换以及锚框计

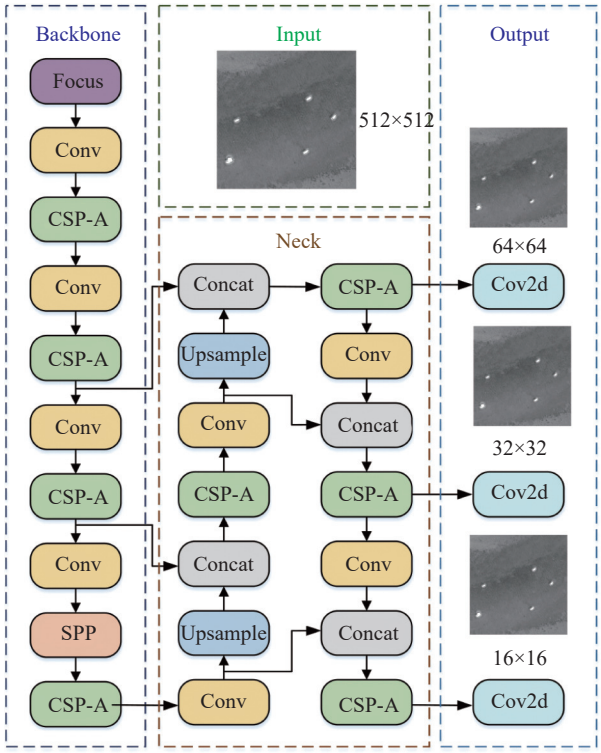


图 3 Yolov5s 目标检测总体架构

Fig.3 Overall structure of Yolov5s target detection

算。Backbone 部分主要包括 Focus、改进的跨阶段局部网络 (Cross Stage Partial Advance Network, CSP-A)、Conv 卷积块、空间金字塔池化层 (Spatial Pyramid Pooling, SPP)。主干网络部分主要用于特征提取，其中 Focus 模块对图像进行降采样操作，将特征层扩充为原有的 4 倍，从而确保了特征的充分提取。CSP-A 网络通过将梯度的变化从头到尾地集成到特征图中，在减少了计算量的同时可以保证准确率。Conv 卷积块由卷积层、批归一化层和一个 ReLU 激活函数构成，实现各层之间的连接。SPP 模块使用最大池化层和四个不同大小的卷积核来实现多个尺度上的特征融合。Neck 层则采用了特征金字塔网络 (Feature Pyramid Network, FPN) 和路径聚合网络 (Path Aggregation Network, PAN)。FPN 网络从上到下对图像进行上采样，并将提取出的特征与主干网络融合，而 PAN 则从下到上对图像进行降采样，将提取出的特征与 FPN 相融合。Prediction 部分则包括锚框的损失函数以及非极大值抑制 (Non Max Suppression, NMS)。

输出中包含了 3 个尺度的特征图，用于检测大、中、小目标。利用 NMS 消除冗余的检测框，同时保留置信度最高的预测框的信息，从而完成目标检测。

2.2 改进特征提取层网络结构

虽然 YOLOv5 在精度和准确率上具有良好的性能，但在检测多尺度红外时敏目标方面仍存在一定问题。CSP-A 是基于 CSPNet 的思想设计的结构，文中所提的具体改进主要为在 CSP 网络中增加 SE (Squeeze-and-Excitation)[22] 和 CBAM (Convolutional Block Attention Module)[23] 注意力机制，更好的实现红外时敏目标检测。注意机制主要作用于特征图，增强了网络的特征表达能力。添加注意力机制后的 CSP-A 模块如图 4 所示。

SE 模块首先通过全局平均池化操作将输入的每

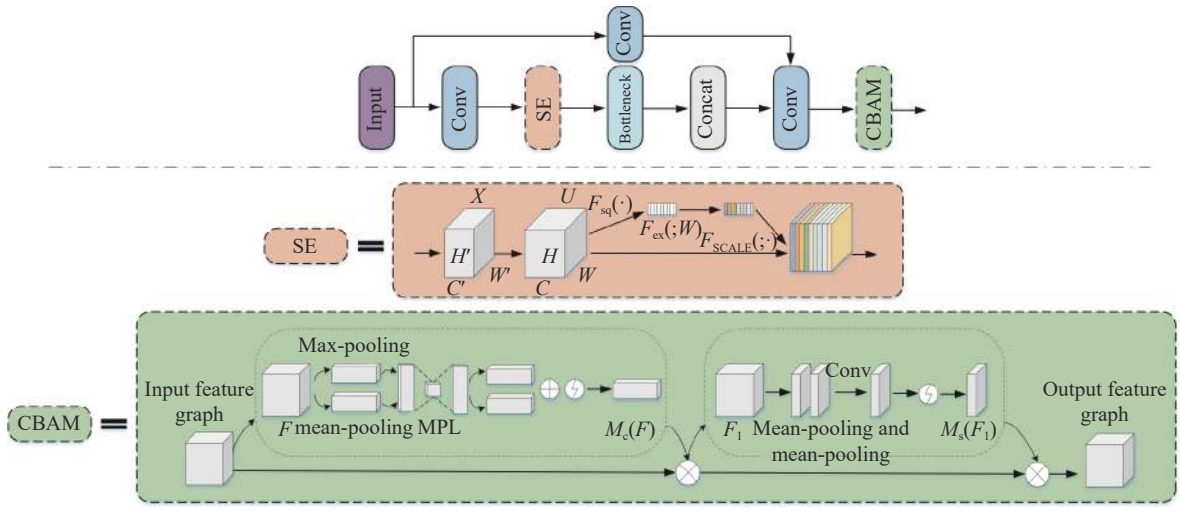


图 4 添加注意力机制的 CSP-A 架构

Fig.4 The CSP-A structure of adding attention mechanism

个通道压缩为一个实数，随后经过全连接层和激活函数实现每个通道的权重更新，最后将更新后的权重与元时输入相乘，实现原有特征图的重构，从而获取更有用的特征。

CBAM 模块结合了通道注意力机制和空间注意力机制，使更多有用的特性信息得以保留。在 CBAM 中输入为$H\times W\times C$的特征图 F 分别通过平均池化和最大池化生成两个 1×1×C 的特征图，然后经过多层感知机，特征图经过相加和 Sigmoid 激活函数生成一维通道注意力特征图 $M_c(F)$，将 F 与 $M_c(F)$ 相乘从而得到通道注意调整特征图 F_1。F_1 再进行最大池化和平均池化，得到两个 $H\times W\times 1$ 的特征图。并对池化后生成的两个二维向量进行拼接和卷积，最终生成二维空间注意图 $M_s(F_1)$，并将其与特征图 F_1 相乘。

3 实验验证与结果分析

3.1 实验环境及数据集

文中的实验环境为一台搭载 Intel Xeon E5-2667 CPU 与 4 块 NVIDIA GeForce RTX 2080 Ti GPU 的深度学习工作站。操作系统为 Ubuntu 18.04，开发软件为 PyCharm2020，编程语言为 Python，依赖深度学习框架 PyTorch 1.6，并且配置了 CUDA 10.2、cuDNN 7.6.5、和其他常用的深度学习和图像处理库。

文中整理了 DIOR[24]、DOTA[25]、LEVIR[26] 遥感影像目标检测数据集中的部分图像作为文中的原始数据集共计 100 张，尺寸统一大小为512×512。实验数据集主要包含舰船、飞机两类时敏目标数据。

3.2 实验评估指标

文中的跨模态数据增强方法为两阶段模型。为了评估第一阶段可见光红外模态转换模型性能，文中选用均值 (Mean)、标准差 (Standard Deviation)、方差 (Variance)、信息熵 (Information Entropy) 对比度 (Contrast ratio)、和平均梯度 (Mean Gradient)6 个指标评价生成的红外图像质量，从而评估可见光图像转换为红外图像转换效果。

为验证文中数据增强中第二阶段的随机样本生成模型的有效性，以及 CSP-A 改进 YOLOV5 网络的效果，文中利用 SSD、Fast R-CNN、原始 YOLOV5、以及基于 CSP-A 改进的 YOLOV5 目标检测算法进行模型及其性能测试，采用平均准确度 mAP、召回率 Recall rate、准确率 Precision rate 和 F_1 指数作为评估指标，在增强前后的红外时敏目标数据集上进行验证测试，召回率、准确率和 F_1 指数的定义如下：

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

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

$$F_1=\frac{2\times Recall\times Precision}{Recall+Precision} \tag{9}$$

式中：TP 为实际存在的红外时敏目标个数；FN 为漏检存在目标个数；FP 为误检出的虚假目标个数。

3.3 实验结果及分析

3.3.1 模态转换实验

文中首先利用 800 张可见光-红外图像对模态转换模型进行训练，图像被归一化为256×256大小后被输入到训练网络中，训练样本图像对如图 5 所示。文中所使用的红外数据的波长范围为 0.75~1.1 μm，即近红外短波图像。通过计算 100 张测试集中原始红外图像与转换后的红外图像的均值、标准差、方差、信息熵、对比度和平均梯度，对六类指标进行定量分析，平均结果如表 1 所示。

(a) 可见光图像数据
(a) Visible image data

(b) 红外图像数据
(b) Infrared image data

图 5 模态转换训练数据集

Fig.5 Modality transfer partial training dataset

生成红外图像的均值、标准差和方差与原始红外图像相比数值略高于原始红外图像，结果表明仿真图像与原始图像在亮度、边缘清晰度等方面具有较高的相似性。信息熵代表图像平均信息量，平均梯度反映图像中的微小细节反差和纹理变化特征，结果表明生成红外图像在这两个指标上与原始红外图像也较为

表 1 模态转换效果对比

Tab.1 Comparison of modal transformation effects

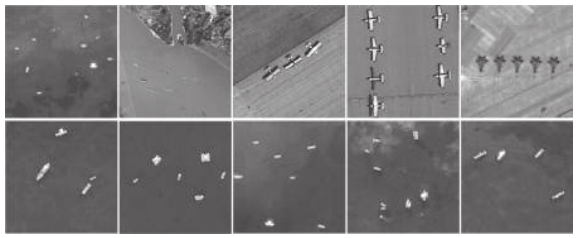
	Mean	Standard deviation	Variance	Information entropy	Contrast ratio	Mean gradient
Original IR images	126.049 31	43.281 21	0.029 37	7.231 47	147.147 56	5.135 38
Transfer IR images	137.884 15	45.599 76	0.032 64	7.188 54	43.930 82	3.158 52

接近。图像对比度一定程度上反映了图像的清晰程度，在这方面生成红外图像与原始图像相比还有较大的提升空间。

通过表中的数据可知与原始红外图像相比，转换后的红外图像具有较高的相似性，同时该模型的输入输出的分辨率相同，因此不会产生图像失真的效果，从而验证了文中模态转换模型的有效性。

将原始数据集中的 100 张数据进行模态转换测试，部分实验结果如图 6 所示。

(a) 原始可见光图像
(a) Original visible image



(b) 模态转换后红外图像
(b) Infrared image after modality transfer

图 6 可见光红外图像转换结果

Fig.6 Visible infrared image conversion results

利用文中的模态转换模型可以在不损失尺寸、结构、视场的前提下将遥感可见光图像转换为红外图像，且不存在失真、噪声、畸变等问题。通过实验结果可以看出，生成红外时敏目标具有较好的纹理细节和红外特性，与背景有着较为明显的区分。

3.3.2 随机样本生成实验

采用对抗性随机样本生成模型生成红外时敏目标图像，将图 6 中的部分转换后的红外时敏目标数据输入到模型的训练网络训练，本模型共进行了 9 个尺度的 GAN 训练 (包括生成器和判别器)，每次训练为 2 000 次，共训练 18 000 次，生成器和判别器的学习率均为 0.000 5。

文中将 100 张原始数据扩充为 500 张数据，实现了数据集的 5 倍扩充，生成的部分结果如图 7 所示，可以发现，文中算法不仅可以生成和原始数据同一方向的数据，而且可以生成多种类、多位置的红外时敏目标数据。

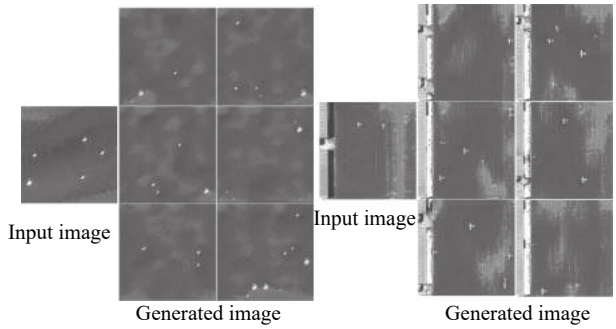


图 7 部分生成图像数据

Fig.7 Partial generated image data

为了验证跨模态数据增强模型的数据增广效果，将原始数据集和扩充后数据集分别在所改进的 Yolov5 (CSP-A) 网络上进行训练测试。数据增强前后测试评估结果如表 2 所示。

表 2 数据增强前后性能对比

Tab.2 Performance comparison before and after data augmentation

Dataset	Precision	Recall	mAP@0.5	F_1
Origin	0.786 3	0.910 5	0.892 4	0.843 9
Augmentation	0.932 0	0.970 4	0.980 6	0.950 8

分析表 2 可知，相比于使用原始数据训练深度学习检测网络，文中所提数据增强算法，对正样本的检测能力提升明显，检测准确率提升了 14.57%，召回率提升了 5.99%，平均精度也提升了 8.82%。

3.3.3 不同模型性能对比测试

为了验证文中提出的改进 Yolov5(CSP-A) 算法针对红外时敏目标检测效果, 利用多种基于深度学习目标检测算法在增强前后的数据集上进行对比实验。对比算法主要包括 SSD、Fast R-CNN、Yolov5 以及文中改进 Yolov5(CSP-A) 算法。图 8 为不同场景下的时敏目标检测结果, 通过可视化结果分析可知, 文中所提的 Yolov5(CSP-A) 算法具有较强的鲁棒性, 漏检和误检率较低, 检测准确率较高。实验量化结果如表 3 所示。

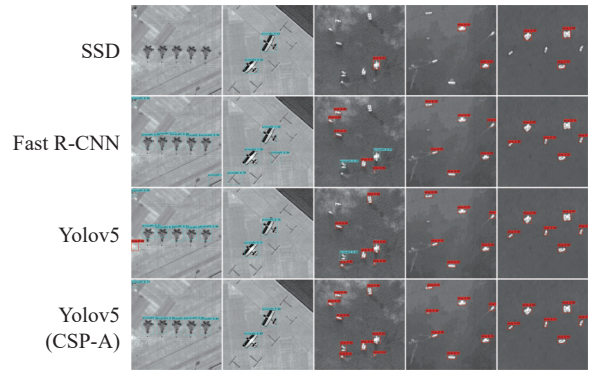


图 8 典型红外时敏目标检测结果对比

Fig.8 Comparison of typical IR time-sensitive target detection results

表 3 不同检测方法的对比实验

Tab.3 Comparison experiments of different detection methods

Method	Precision rate	Recall rate	mAP@0.5	mAP@0.5(ship)	mAP@0.5(aircraft)	F_1
SSD	0.3564	0.8423	0.8271	0.7693	0.8848	0.5009
Fast R-CNN	0.4328	0.8534	0.8327	0.8564	0.8180	0.5743
Yolov5	0.8584	0.9161	0.9532	0.9687	0.9376	0.8863
Yolov5 (CSP-A)	0.9320	0.9704	0.9806	0.9807	0.9805	0.9508

由表 3 可知, 相比于 SSD、Fast R-CNN 和 Yolov5, 文中算法在准确率、平均精度以及 F_1 指数上都具有较大的提升。与原始 Yolov5 网络相比, 准确率提升了 7.36%, 召回率提升了 5.43%, 平均精度提升了 2.74%, F_1 指数提升了 6.45%, 部分检测结果如图 9 所示。

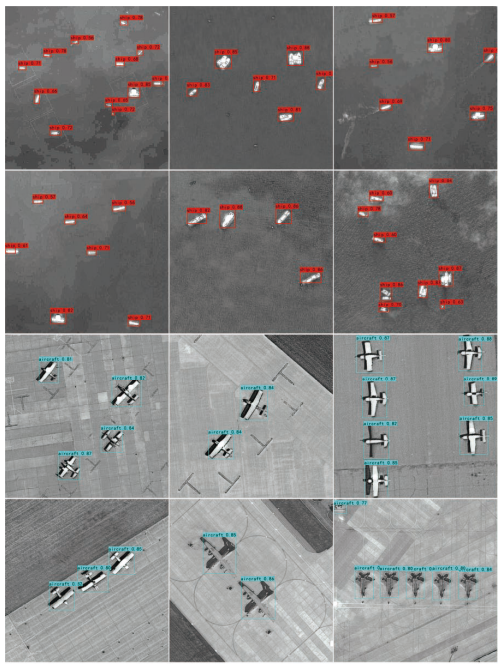


图 9 部分图像检测结果

Fig.9 Test results on partial image

3.3.4 消融实验

为了验证改进 Yolov5(CSP-A) 模型中各关键模块的有效性, 笔者进行了消融实验, 文中以 Yolov5 网络为基准模型。通过不同结构的网络检验更改添加 SE 和 CBAM 注意力机制模块等策略对模型检测效果的影响。每次实验均采用相同实验设备和训练参数, 分别在实验划分出来的测试集进行测试。

由表 4 可以得出, Yolov5 基准模型的测试效果最差, 使用单一的自注意力机制模块可以在一定程度上提升模型的检测效果, 仅增加 SE 模块使得 AP 提升了 2.48%, 仅增加 CBAM 模块使得 AP 提升了 1.92%。序号 4 添加了 SE、CBAM 注意力机制模块至特征提取和特征融合部分, 使得测试的平均精度提升了 4.29%, 表明该方法可以有效提取红外时敏目标特征。

表 4 消融实验结果

Tab.4 Ablation experiment results

Number	SE	CBAM	mAP@0.5
1	-	-	0.937 6
2	-	√	0.956 8
3	√	-	0.962 4
4	√	√	0.980 5

可以看出，文中方法可以以较高置信度检测出图中的时敏目标。通过以上数据分析可知，文中提出的 Yolov5(CSP-A) 算法检测红外时敏目标可以做到低漏警率，且实现了较好的检测效果。

4 结 论

针对红外时敏目标数据匮乏和检测效果不佳的问题，文中提出了一种跨模态数据增强的红外时敏目标检测技术。在两阶段模型数据增强方面，首先利用模态转换网络将包含时敏目标的可见光遥感图像转换为具备红外特性的目标图像，其次在样本随机生成模型中引入 coordinate attention 注意力机制，最后提出基于改进 CSP 模块的 Yolov5 检测技术。多组实验结果表明，文中算法在红外时敏目标数据集中检测准确率高达 98.06%，解决红外时敏目标数据匮乏的问题的同时具有较好的目标检测能力，下一步拟对不同光谱条件下的红外图像数据进行分析实验，提升算法的准确性以及适应能力。

参考文献：

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[2] Yu X, Hong S, Yu J, et al. Research on a ship target data augmentation method of visible remote sensing image [J]. ***Chinese Journal of Scientific Instrument***, 2020, 41(11): 261-269. (in Chinese)

[3] Ma Y, Tang P, Zhao L, et al. Review of data augmentation for image in deep learning [J]. ***Image Graphics***, 2021, 26(3): 487-502. (in Chinese)

[4] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. ***Communications of the ACM***, 2017, 60(6): 84-90.

[5] Taylor L, Nitschke G. Improving deep learning with generic data augmentation[C]//2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, 1542-1547.

[6] Zhong Z, Zheng L, Kang G, et al. Random erasing data augmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13001-13008.

[7] Ma D, Tang P, Zhao L. SiftingGAN: Generating and sifting labeled samples to improve the remote sensing image scene classification baseline in vitro [J]. ***IEEE Geoscience and Remote Sensing Letters***, 2019, 16(7): 1046-1050.

[8] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[EB/OL]. (2017-12-25) [2022-12-06]. https://arxiv.org/abs/1704.00028.

[9] Zheng Z, Zheng L, Yang Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 3754-3762.

[10] Zhong Z, Zheng L, Zheng Z, et al. Camera style adaptation for person re-identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5157-5166.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]//Proceedings of the IEEE European Conference on Computer Vision, 2016: 21-37.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// 2014 IEEE Comference on Computer Vision and Pattern Recognition, 2014: 580-587.

[14] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[15] Ju M, Luo J, Liu G, et al. ISTDet: An efficient end-to-end neural network for infrared small target detection [J]. ***Infrared Physics & Technology***, 2021, 114: 103659.

[16] Yao S, Zhu Q, Zhang T, et al. Infrared image small-target detection based on improved FCOS and spatio-temporal features [J]. ***Electronics***, 2022, 11(6): 933.

[17] Lu X F, Bai X F, Li S X, et al. Infrared small target detection method based on the improved weighted enhanced local contrast measurement [J]. ***Infrared and Laser Engineering***, 2022, 51(8): 20210914. (in Chinese)

[18] Jiang R Q, Peng Y P, Xie W X, et al. Improved YOLOv4 small target detection algorithm with embedded scSE module [J]. ***Journal of Graphics***, 2021, 42(4): 546-555. (in Chinese)

[19] Owens A, Wu J, McDermott J H, et al. Ambient sound provides supervision for visual learning[C]//European conference on computer vision. Springer, Cham, 2016: 801-816.

[20] Goodfellow I, Pouget-abadie J, Mirza M, et al. Generative adversarial networks [J]. ***Communications of the ACM***, 2020, 63(11): 139-144.

[21] Hou Q, Zhou D, Feng J. Coordinate attention for efficient

mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13713-13722.

[22] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[23] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.

[24] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: A survey and a new benchmark [J]. ***ISPRS Journal of Photogrammetry and Remote Sensing***, 2020, 159: 296-307.

[25] Xia G S, Bai X, Ding J, et al. DOTA: A large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3974-3983.

[26] Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers [J]. ***IEEE Transactions on Geoscience and Remote Sensing***, 2021, 60: 1-14.

Infrared time-sensitive target detection technology based on cross-modal data augmentation

Wang Siyu，Yang Xiaogang*，Lu Ruitao，Li Qingge，Fan Jiwei，Zhu Zhengjie

(Missile Engineering Institute, PLA Rocket Force University of Engineering, Xi'an 710025, China)

Abstract:

Objective Infrared time-sensitive targets refer to infrared targets such as ships and aircraft, which have high military value and the opportunity of attack is limited by the time window. Infrared time-sensitive target detection technology is widely used in military and civilian fields such as unmanned cruise, precision strike, battlefield reconnaissance, etc. The target detection algorithm based on deep learning has made great progress in the field of target detection due to its powerful computing power, deep network structure and a large number of labeled data. However, the acquisition of some high-value target images is difficult and costly. Therefore, the infrared time-sensitive target image data is scarce, and the multi-scene and multi-target data for training is lacking, which makes it difficult to ensure the detection effect. Based on this, this paper proposes an infrared time-sensitive target detection technology based on cross-modal data enhancement, which generates "new data" by processing the data, expands the infrared time-sensitive target data set, and improves the model detection accuracy and generalization ability.

Methods We propose an infrared time-sensitive target detection technology based on cross-modal data enhancement. The cross-modal data enhancement method is a two-stage model (Fig.1). First, in the first stage, the visible light image containing time-sensitive targets is converted into infrared images through the mode conversion model based on the CUT network, and then the coordinate attention mechanism is introduced into the second stage model to randomly generate a large number of infrared target images, realizing the data enhancement effect. Finally, an improved Yolov5 target detection architecture based on SE module and CBAM module is proposed (Fig.3).

Results and Discussions The proposed cross-modal infrared time-sensitive target data enhancement method combines the style migration model with the target generation model, and uses the visible light image data set to achieve infrared time-sensitive target data enhancement. We can convert remote sensing visible image into infrared image without losing size, structure and field of view, without distortion, noise, distortion and other problems. It can be seen from Fig.6 that the generated infrared time-sensitive target has good texture details and

infrared characteristics, and is clearly distinguished from the background. An improved Yolov5 target detection model is proposed. SE and CBAM attention mechanisms are added to the CSP network to enhance the feature expression of the network and better achieve infrared time-sensitive target detection. It can be seen from the analysis of Tab.2 that compared with using the original data to train the deep learning detection network, the data enhancement algorithm proposed in this paper has significantly improved the detection ability of positive samples, the detection accuracy rate, the recall rate, and the average accuracy have increased by 14.57%, 5.99%, and 8.82% respectively. It can be seen from Tab.3 that compared with SSD, Fast R-CNN and Yolov5, the algorithm in this paper has a great improvement in accuracy, average accuracy and F1 index. Compared with the original Yolov5 network, the accuracy rate, the recall rate, the average accuracy, and the F1 index have increased by 7.36%, 5.43%, 2.74%, and 6.45% respectively. Some test results are shown (Fig.9).

Conclusion Due to the lack of infrared time-sensitive target data and poor detection effect, we proposes a cross-modal data enhancement infrared time-sensitive target detection technology. In the aspect of two-stage model data enhancement, firstly, the visible light remote sensing image containing time-sensitive targets is converted into the target image with infrared characteristics using the mode conversion network. Secondly, the coordinate attention mechanism is introduced into the sample random generation model. Finally, the Yolov5 detection technology based on the improved CSP module is proposed. Multiple sets of experimental results show that the detection accuracy of the algorithm in this paper is up to 98.06% in the infrared time-sensitive target data set, which solves the problem of the lack of infrared time-sensitive target data and has good target detection ability.

Key words: infrared time-sensitive targets; data augmentation; modal transformation; target detection

Funding projects: National Natural Science Foundation of China (62276274); Aviation Science Foundation (201851U8012)