

基于Swin Transformer和注意力机制的红外无人机检测算法



王思宇¹, 卢瑞涛^{1,2}, 黄攀³, 杨小冈¹, 夏文新¹, 李清格¹, 张震宇¹

1. 火箭军工程大学, 陕西 西安 710025

2. 光电控制技术重点实验室, 河南 洛阳 471000

3. 航空工业西安航空计算技术研究所, 陕西 西安 710068

摘要: 红外无人机目标检测在军民领域的应用前景广阔。由于无人机目标尺度较小, 空中环境复杂多变, 目前普遍存在检测率低和误报率高的现象。针对复杂场景下红外无人机目标检测不良等问题, 本文提出ST-YOLOA目标检测模型。首先, 使用Swin Transformer网络架构和协调注意力(CA)机制搭建STCNet骨干特征提取网络; 其次, 特征融合部分采用带残差结构的PANet路径聚合网络构建特征金字塔提升整体特征提取能力, 同时改进了上下采样方式以增强检测能力; 最后, 使用解耦检测头预测无人机目标的位置。试验结果表明, 本文提出的模型检测精度为92.8%, 检测速度达到了22帧/s, 这表明该模型与其他模型相比具有较好的检测效果, 且基本满足实时性检测要求, 对于多无人机目标场景下的检测具有现实意义。

关键词: 红外无人机; 目标检测; Swin Transformer; 协调注意力机制; STCNet

中图分类号: TP391.4

文献标识码: A

DOI: 10.19452/j.issn1007-5453.2024.02.005

随着无人机技术^[1-2]的快速发展, 无人机滥用的风险显著增加, 对公共安全^[3-4]构成了相当大的威胁。有效对抗非合作无人机已成为一个紧迫的社会保障问题。作为反无人机系统的关键组成部分, 实现非合作无人机的精确检测是进一步对抗的基础。红外无人机弱小目标检测技术是一种基于红外技术的非接触式无人机检测技术, 通过红外传感器对空中的无人机弱小目标进行实时监测和识别。

在早期研究中, 研究人员根据红外图像特征, 提出了背景估计的方法^[5], 结合红外图像目标与背景之间的差异特性, 部分基于局部对比度的方法被提出^[6-7]。董丽丽等^[8]对L. Itti等^[9]提出的人类视觉模型进行了修正, 使之更适用于红外弱小目标检测。此外, D. J. Gregoris等^[10]将小波变换引入红外弱小目标检测领域。吴一全等^[11]提出了一种基于轮廓波变换的方法, 在变换域中对背景频段进行抑制, 对目标频段进行增强, 以提高检测能力。

随着深度学习理论的蓬勃发展, 卷积神经网络在目标检测领域取得了巨大进展, 并表现出优异的检测性能。基

于深度学习的目标检测算法大致可以分为双阶段目标检测方法和单阶段目标检测方法^[12]。双阶段目标检测方法(如R-CNN^[13]、Fast R-CNN^[14]、Faster R-CNN^[15]等)先进行区域生成得到预选框, 再通过卷积神经网络进行样本分类和与边框定位回归, 检测精度高但检测效率较低; 单阶段目标检测方法(如YOLO^[16]、SSD^[17]、CenterNet^[18]等)则通过主干特征提取网络直接回归得到目标的位置信息及分类结果, 检测速度快, 但相比于双阶段检测方法容易出现误检、漏检的现象。目前, 基于Transformer的目标检测算法大多是在DETR^[19]算法的基础上进行改进, 也有一些是将通用的Transformer主干网络应用到目标检测任务中, 同样取得了很好的结果。Zhu Xizhou等^[20]提出了可形变的DETR模型, 融合了可形变卷积^[21]中稀疏空间采样的优势和Transformer强大的关系建模能力。此外, Dai Jifeng等^[22]提出了一种无监督预训练的UP-DETR目标检测算法。

尽管以上研究人员对红外目标多尺度、噪声杂波大、背景复杂等问题进行了研究, 但识别精度和计算速度仍然

收稿日期: 2023-07-10; 退修日期: 2023-11-29; 录用日期: 2024-01-05

基金项目: 国家自然科学基金(62276274); 航空科学基金(201851U8012); 陕西省自然科学基金(2023-JC-YB-528)

引用格式: Wang Siyu, Lu Ruitao, Huang Pan, et al. Infrared UAV detection algorithm based on Swin Transformer and attention mechanism [J]. Aeronautical Science & Technology, 2024, 35(02): 39-46. 王思宇, 卢瑞涛, 黄攀, 等. 基于Swin Transformer和注意力机制的红外无人机检测算法[J]. 航空科学技术, 2024, 35(02): 39-46.

限制了这些技术在无人机目标检测领域的应用。针对以上问题,本文选择了Swin Transformer和YOLO检测模型作为基础模型,提出了一种更符合现实复杂环境的、更适应红外图像的无人机目标检测模型。我们将其命名为ST-YOLOA。

1 ST-YOLOA目标检测模型

首先对ST-YOLOA目标检测模型进行总体概述,然后分别从特征提取(主干)、特征融合(颈部)、目标检测(损失)三个部分对ST-YOLOA模型的设计思路和网络架构进行详细论述。ST-YOLOA网络结构如图1所示。图1中,红色虚线框为注意力模块添加位置,Conv、BN和Silu分别表示卷积、批量归一化和SILU激活函数,Concat表示全连接操作,Cls、Reg和Obj分别表示分类、定位以及置信度得分, H 、 W 和 C 表示特征图的宽度、高度和通道数。

1.1 概述

(1) 骨干

在ST-YOLOA中,本文提出一个名为STCNet的骨干网络。它整合了Swin Transformer和CA注意力机制的优势。相比于传统的基于CNN的骨干特征提取网络在目标定位时仅利用区域提供的信息,STCNet具有动态注意和

全局建模能力,考虑远程依赖关系,具有良好的性能。STCNet网络采用分层架构,由Patch Embedding层、Swin Transformer Block层和CA-Patch Merging层三部分组成。

(2) 颈部

在ST-YOLOA的颈部,仍然使用PANet构建特征金字塔进行特征深度融合。此外,还在颈部引入SE和CBMA注意力机制,增强对目标信息的关注度,进一步提升模型性能。

(3) 损失

损失函数的目的主要是使模型定位更准确,识别精度更高。因此,ST-YOLOA中使用更加先进的EIOU Loss,加速收敛并提升模型性能。

1.2 特征提取部分

1.2.1 Patch Embedding 模块

Patch Embedding模块首先在特征提取网络前端对图像进行分块处理,将图像分割成 4×4 大小的互不重叠的图像块,使每个图像块的特征维数为 $4 \times 4 \times 3$,再通过线性变换的方式投影特征维度到任意维度,使得原始的二维图像转换成一系列一维的嵌入矢量,转换后的嵌入矢量输入三个stage特征提取层生成分层特征表示。其结构如图2所示, W 和 H 为输入特征图的长和宽, d 为通道维数, N 为批量大小。

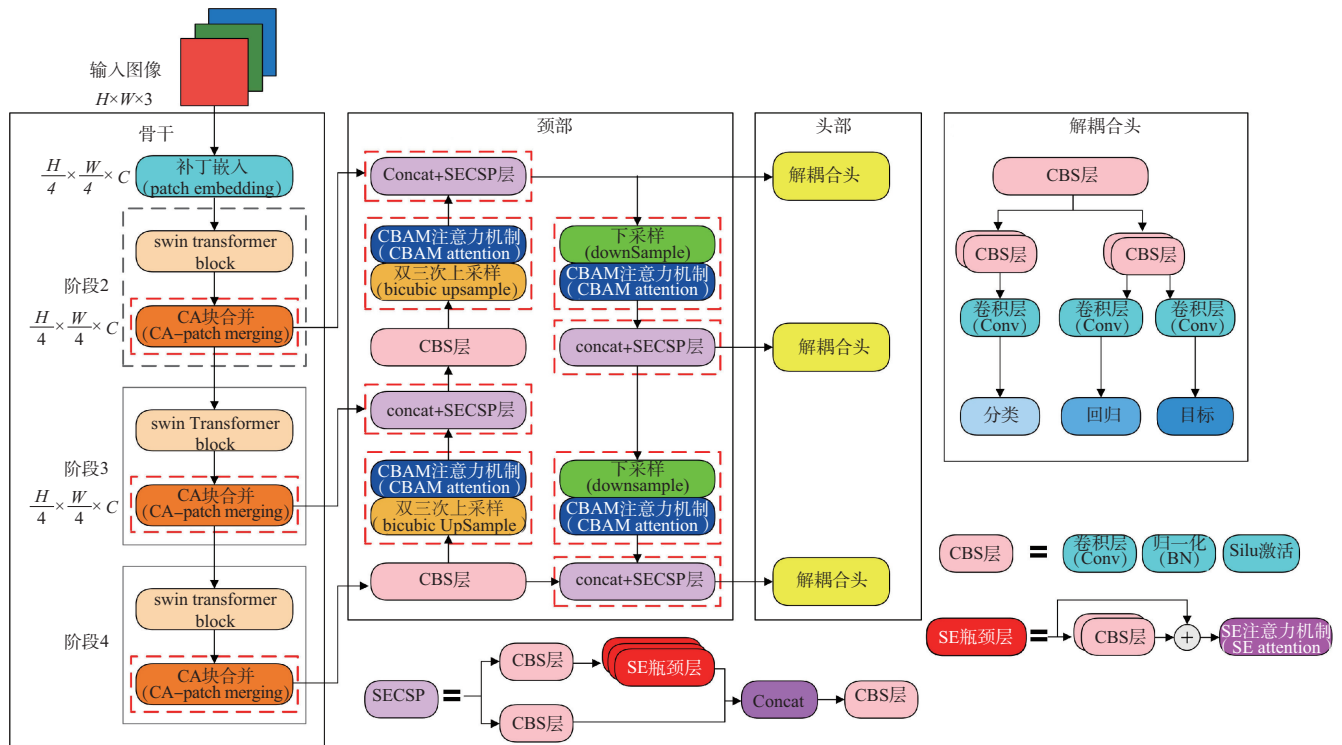


图1 ST-YOLOA网络结构
Fig.1 ST-YOLOA network structure

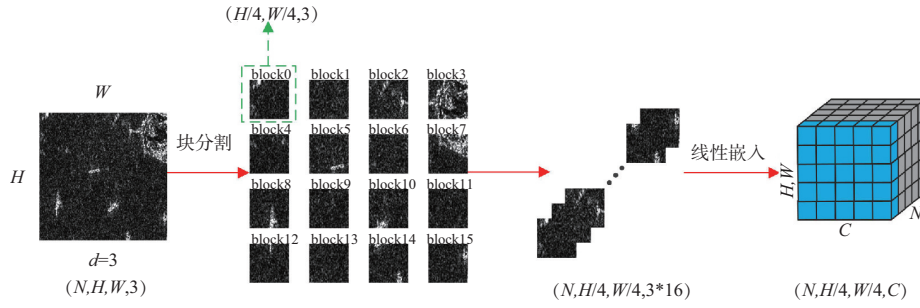


图2 Patch Embedding结构图

Fig.2 Patch Embedding structure diagram

1.2.2 Swin Transformer 模块

Swin Transformer 模块采用移动窗口的形式计算像素之间的注意力,有助于前层窗口的连接,在降低原有注意力计算复杂度的同时克服了缺少全局效应的缺点,显著增强了建模效果。

如图3所示,Swin Transformer模块中多头自注意力机制是基于移位窗口构建的。每个模块含有两个连续的滑动transformer模块。每个滑动Transformer模块均由归一化(layer norm, LN)层、多头自注意力模块、残差连接和多层感知器(MLP)模块组成,其中MLP含有两个使用GELU非线性激活函数的全连接层。两个连续Swin Transformer模块分别采用基于窗口的多头自注意力(W-MSA)模块和基于移位窗口的多头自注意力(SW-MSA)模块,在解决了不同窗口之间无法进行信息交流问题的同时减少了计算量。基于这种窗口分割机制,连续Swin Transformer模块计算公式为

$$\hat{z}^i = \text{W-MSA}(\text{LN}(z^{i-1})) + z^{i-1} \quad (1)$$

$$z^i = \text{MLP}(\text{LN}(\hat{z}^i)) + \hat{z}^i \quad (2)$$

$$\hat{z}^{i+1} = \text{SW-MSA}(\text{LN}(z^i)) + z^i \quad (3)$$

$$z^{i+1} = \text{MLP}(\text{LN}(\hat{z}^{i+1})) + \hat{z}^{i+1} \quad (4)$$

式中, $\hat{z}^i$ 和 z^i 分别为第 i 个窗口W-MSA模块和MLP模块的

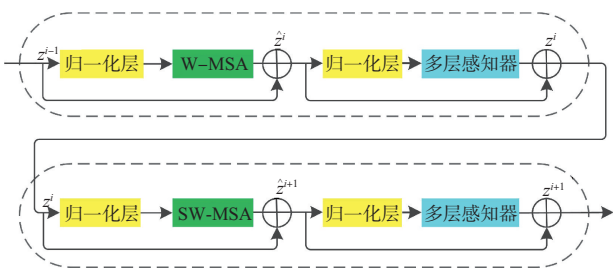


图3 Swin Transformer 模块

Fig.3 Swin Transformer blocks

输出。

1.2.3 CA-Patch Merging 层

Patch Merging层在骨干网络中起池化作用,能够降低特征图分辨率和调整通道数,从而形成分层设计,同时也能够节省一定的计算量,其工作过程如图4所示。

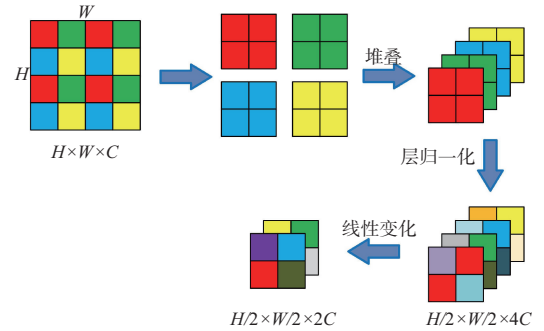


图4 Patch Merging层原理图

Fig.4 Schematic diagram for Patch Merging layer

考虑到Swin Transformer的上下文编码能力有限,本文在Patch Merging层之后添加了CA注意力机制。CA注意力将通道注意力工作过程分解为两个一维特征编码过程,而后沿空间中两个方向进行特征聚合。CA注意力机制结构如图5(a)所示。它首先将特征图在高度和宽度两个维度分别使用尺寸为 $(H,1)$ 和 $(1,W)$ 的卷积核进行全局平均池化

$$\begin{cases} z_c^h(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i) \\ z_c^w(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \end{cases} \quad (5)$$

上述变换分别获得空间中宽度和高度两个方向的特征图,接着CA注意力先对特征图进行拼接操作,而后进行 F_1 变换得到特征图 f

$$f = \delta(F_1([z^h, z^w])) \quad (6)$$

式中, F_1 为 1×1 卷积变换函数; $[,]$ 表示拼接操作; δ 为非线

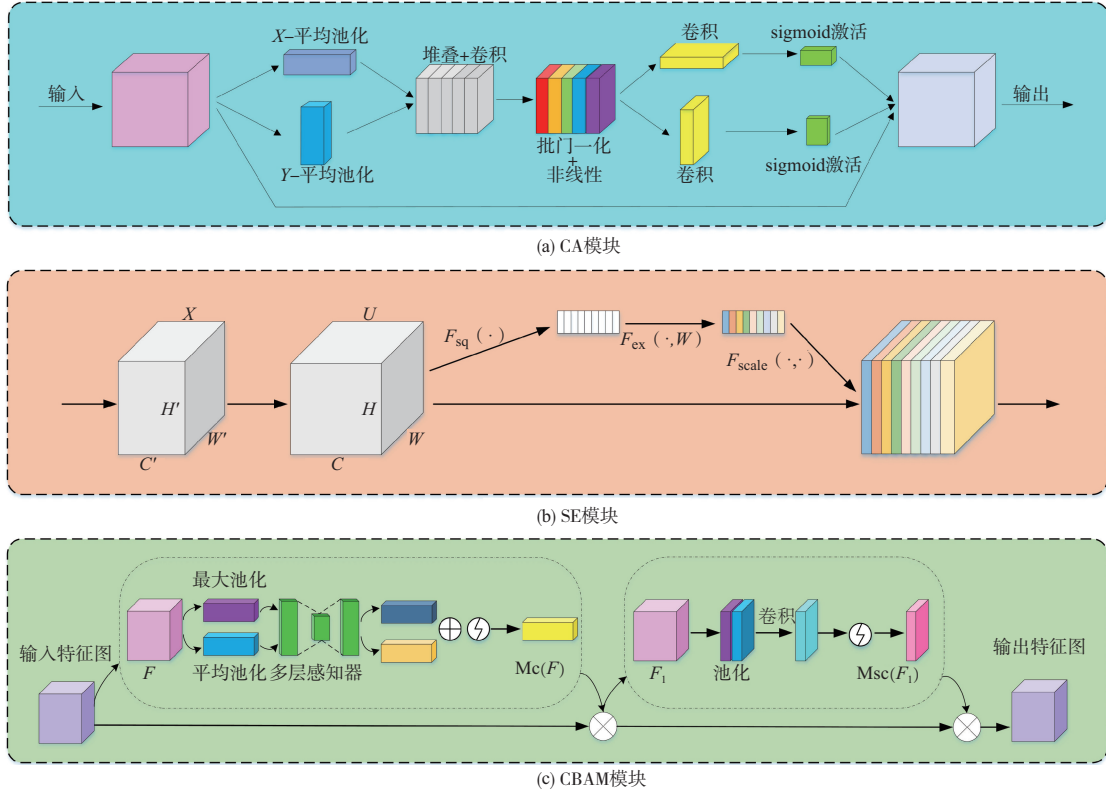


图5 注意力机制模块结构图

Fig.5 The module structure of attention mechanism

性激活函数。

随后特征图 f 按原来的高宽方向进行卷积,经sigmoid激活函数激活,得到特征图注意力权重 g^h 和 g^w ,其公式如下

$$\begin{cases} g^h = \sigma(F_h(f^h)) \\ g^w = \sigma(F_w(f^w)) \end{cases} \quad (7)$$

式中, σ 为sigmoid激活函数。

最后CA注意力机制通过乘法加权计算,得到带有注意力权重的特征图输出

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \quad (8)$$

它通过精确的位置信息对长距离依赖性进行编码,使本文模型能够有效地利用全局上下文信息。与此同时,CA兼具通道域和空间域的注意机制。它的引入能够更好地捕获方向感知和位置敏感信息,以便更加准确地定位识别感兴趣对象,进一步提高特征表达能力。

1.3 特征融合部分

1.3.1 改进的CSPLayer-SECSP

CSPLayer主要分为两部分,其结构参照图1。主干部分由卷积分支和残差分支组成,其余部分通过一个 1×1 卷积层直接连接到CSPLayer输出部分。残差模块的瓶颈层

(bottleneck)结构是CSPnet的重要组成部分,其堆叠部分由一个 1×1 卷积和一个 3×3 卷积组成,并在最后使用捷径连接将初始输入直接添加到卷积层的输出中。CSP网络的特征提取过程大都是在残差模块中进行的,其应用极大地缓解了模型的梯度消失问题。

CSPLayer的使用,使得模型过度考虑周围上下文信息,而造成局部信息干扰。为解决这个问题,本文在瓶颈层模块中引入SE注意力机制,对特征信息进行选择性强调,弱化干扰信息,进一步提升对目标特征的关注度。与此同时,瓶颈层中的 3×3 卷积层需要处理大量的参数运算,也造成了大量参数冗余。SE对特征图顺着空间维度进行特征压缩,将全局性的空间信息挤压到通道描述中,压缩后第 c 通道的输出特征图 z_c 为

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i,j) \quad (9)$$

式中, x_c 为输入; H 和 W 分别为空间中的高度和宽度两个方向。这一过程大大降低了网络中的冗余参数。SE的结构如图5(b)所示。

1.3.2 改进的上、下采样过程

不同尺度的特征图分辨率不同,在特征融合前需进行

下采样或上采样操作来缩小或放大特征图,以进行不同尺度特征图之间的特征融合。在特征图的压缩和延展过程中,会带来语义信息丢失和引入局部干扰的问题。CBAM 通过在空间和通道上同时进行全局平均和全局最大的混合池化能够专注于更深层次的特征语义信息,它的引入使模型更具鲁棒性。其具体工作过程如下

$$\begin{cases} M_c(F) = \sigma [\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))] = \\ \sigma [W_1(W_0(F_{\text{avg}}^c)) + W_1(W_0(F_{\text{max}}^c))] \\ M_s(F) = \sigma (f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) = \\ \sigma (f^{7 \times 7}([F_{\text{avg}}^s; F_{\text{max}}^s])) \end{cases} \quad (10)$$

式中, $M_c(F)$ 和 $M_s(F)$ 分别为一维通道注意和二维通道注意; σ 为 sigmoid 函数; $W_0 \in R^{C_r \times C}$, $W_1 \in R^{C \times C_r}$; $f^{7 \times 7}$ 为 7×7 大小的卷积核。

反卷积上采样可以近似看作卷积的逆操作,它通过引入训练参数进行学习,能够较好地还原特征图,但容易产生棋盘效应,使图像中出现像素块。插值法则不需要学习任何参数,根据已知像素点对未知的点进行预测估计,从而可以扩大图像的尺寸,达到上采样的效果。因此,本文使用双三次插值算法代替反卷积进行上采样,在保留图像原始信息的基础上减少了大量参数运算。

1.4 目标检测部分

考虑到红外图像中复杂环境下无人机目标需提取更多特征信息的现实问题,常用的耦合检测头会影响模型性能,无法满足复杂环境红外图像无人机检测任务。故 ST-YOLOA 网络模型将分类和回归任务分开,采用解耦的检测头进行目标检测实现目标位置和边界框的预测输出,大幅提升了收敛速度,提高了模型的精度。

损失函数方面,由于红外图像无人机目标检测是单类别检测任务,因此损失函数仅由定位损失(reg)和置信度损失(obj)两部分组成,数学表达式为

$$\text{Loss} = \frac{\lambda L_{\text{reg}} + L_{\text{obj}}}{N_{\text{pos}}} \quad (11)$$

式中, λ 为平衡系数; N_{pos} 为正样本的 Anchor Point 数; L_{obj} 为置信度损失,本文采用二分类交叉熵损失(BCEloss)作为损失函数以增加数值稳定性; L_{reg} 为定位损失,采用 Efficient-IOUloss(EIOUloss)损失函数

$$L = L_{\text{IOU}} + L_{\text{dis}} + L_{\text{asp}} = 1 - \text{IOU} + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \quad (12)$$

式中, L_{IOU} 为重叠损失; L_{dis} 为中心距离损失; L_{asp} 为宽高损失。EIOUloss 综合考虑了边界框回归的重叠面积、中心点

距离和纵横比,并将纵横比的损失项拆分成预测的宽高分别与最小外接框宽高的差值,有效地解决边界框回归任务中的样本不平衡问题,加速收敛、提高回归精度,使模型进一步得到优化。

综合上述分析,本文提出的 ST-YOLOA 检测模型适用于复杂环境下的红外图像无人机目标检测,具有特征提取能力强、高低层特征信息利用率及信息融合度高、鲁棒性好等优点,更适用于现实条件下的无人机目标检测。

2 试验结果与分析

2.1 试验环境

本文试验数据基于自行构建的红外无人机数据集,利用红外相机拍摄无人机群多种飞行状态下的序列视频,对视频进行截取,得到无人机群飞行图像,再进行有效筛选,获取 6900 张红外图像的无人机群数据集,自行构建的数据集具备无人机目标多、变化尺度大、分布密集等特点,同时还具有云雾遮挡、建筑物背景复杂等干扰因素。

试验数据集按照 8:2 的比例随机划分训练集与测试集,并在训练集中随机选取 20% 作为验证集。同时,为了增加数据的多样性、保证模型具有更好的训练效果,本文使用 Mosaic、Mixup 两种数据增强方式对数据集进行数据增强操作。

本文试验超参数参照参考文献[23]~[25],主要设置为:训练周期为 300 次迭代,模型最大学习率为 0.01、最小学习率为 0.0001,优化器选用“SGD”随机梯度下降,权值衰减为 0.0005,学习率下降方式采用“COS”余弦下降,并采用多线程数据读取加快数据读取速度。测试时的预测概率阈值为 0.5,置信度为 0.001,非极大值抑制阈值为 0.65。

2.2 对比试验

试验选择目标检测的平均精度(mAP)、检测速度(FPS)作为评价检测算法性能的指标。为客观评价 ST-YOLOA 模型的检测效果,本文与其他现有目标检测方法进行了对比试验。对比算法范围涵盖广泛,其中, Faster R-CNN 和 SSD 为经典目标检测模型, YOLOv5、YOLOX 是近年新公布的高性能目标检测模型, YOLOv7 为当前最先进的检测模型之一。同时, 本文将 Swin Transformer 在 YOLOv5 和 Faster R-CNN 框架下进行试验验证分别为 ST-YOLOv5 和 ST-Faster R-CNN, 对比试验结果见表 1。

虽然 SSD 检测速度更快,但试验证明其红外无人机目标检测能力较差,检测准确率较低。Faster R-CNN 检测准确率仅仅稍好于 SSD 算法,仍有较大提升空间,即使是基于

表1 不同算法性能对比

Table 1 Performance comparison between different algorithms

算法	mAP@0.50/%	mAP@0.50:0.95/%	FPS
YOLOv5	89.6	60.5	67
YOLOv7	86.3	28.9	40
YOLOX	85.9	28.8	33
SSD	56.7	23.2	87
Faster R-CNN	58.3	30.1	16
ST-YOLOv5	91.2	62.3	55
ST-Faster R-CNN	62.7	34.8	15
ST-YOLOA	92.8	65.7	22

Swin Transformer的Faster R-CNN算法精确率提升也极为有限。ST-YOLOA算法的mAP@0.50、mAP@0.50:0.95在对比试验中是最高的,检测速度相比于原始的YOLOX算法有所降低,但是mAP@0.50相比于原始的YOLOX提升了8.03%。YOLOv5、YOLOv7和ST-YOLOv5都具有不错的检测准确率和检测速度,但是和本文所提的ST-YOLOA算法相比,检测准确率还有较大的提升空间。

针对云雾背景、目标重叠、目标密集、尺度较小等复杂空中环境,对部分典型算法进行试验对比,结果如图6所示。可以发现,在云雾背景干扰的情况下(如测试1、测试2),ST-YOLOv5、YOLOv7和YOLOX算法均存在较多漏检的情况,当目标重叠且密集时(如测试3、测试4),此类算法难以实现红外无人机群目标的检测任务。综合来看,本文

提出的ST-YOLOA算法优于其他几种对比算法。

3 结束语

针对红外无人机目标检测不良等问题,本文提出了ST-YOLOA红外无人机目标检测模型。通过研究,得出以下结论:

(1) 本文提出STCNet骨干网络提取无人机目标特征信息。该网络不仅具有Swin Transformer出色的基于全局的信息建模能力,而且利用CA注意力捕获位置信息和通道关系,有效地抑制红外图像中特征提取不足的问题,在不同环境下获取更显著的特征信息,增强特征信息处理能力。

(2) 基于改进的PANet构建特征金字塔网络进行高低层特征深度融合,有效利用语义信息和定位信息,缓解了局部信息干扰和注意力分散的问题。使用双三线性插值上采样保留特征图原始信息,增强了检测的准确性。

(3) 将目标检测的分类与回归任务分开处理,有效减弱特征图中的噪声对检测精度的影响。使用EIOU作为定位损失函数,加速收敛,解决了样本不平衡问题,增强了模型泛化能力。在红外无人机数据集上试验并用测试集验证,本文提出的模型检测精度为92.8%,检测速度达到了22帧/s。未来将对模型进行轻量化处理,在保证算法检测精度的同时提升检测速度。

AST

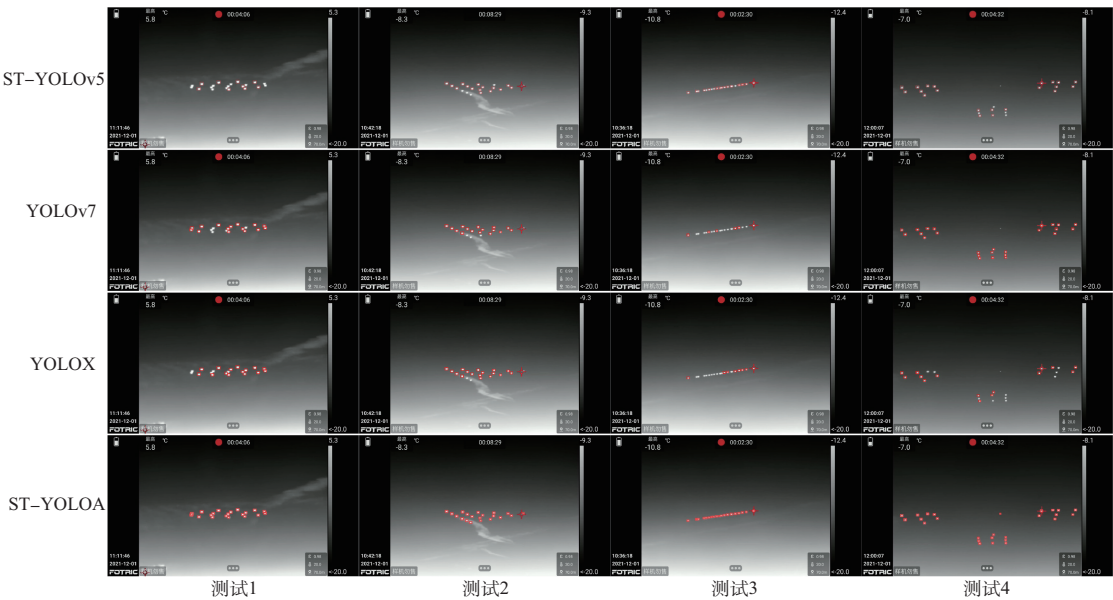


图6 检测结果对比图

Fig.6 Comparison of detection results

参考文献

- [1] 周龙, 赵令公, 王钢林, 等. 固定翼有人飞机无人化改造进展研究[J]. 航空科学技术, 2023, 34(8): 11-17.
Zhou Long, Zhao Linggong, Wang Ganglin, et al. Research on the progress of unmanned transformation of fixed wing manned aircraft[J]. Aeronautical Science & Technology, 2023, 34(8): 11-17. (in Chinese)
- [2] 陈宇恒, 陈进朝, 陈雪聪. 基于改进贪心算法的无人机集群协同任务分配[J]. 航空科学技术, 2022, 33(4): 13-18.
Chen Yuheng, Chen Jinchao, Chen Xuecong. Collaborative task allocation of unmanned aerial vehicle clusters based on improved greedy algorithm [J]. Aeronautical Science & Technology, 2022, 33 (4): 13-18. (in Chinese)
- [3] Svanström F, Alonso-Fernandez F, Englund C. Drone detection and tracking in real-time by fusion of different sensing modalities[J]. Drones, 2022, 6: 317.
- [4] Dewangan V, Saxena A, Thakur R, et al. Application of image processing techniques for UAV detection using deep learning and distance-wise analysis[J]. Drones, 2023, 7: 174.
- [5] 柯贤贵, 陈正鑫, 张越迁, 等. 基于多尺度宽激活残差注意力网络的图像去块效应[J]. 四川大学学报(自然科学版), 2022, 59(6): 89-98.
Ke Xiangui, Chen Zhengxin, Zhang Yueqian, et al. Image deblocking effect based on multi-scale wide activation residual attention networks[J]. Journal of Sichuan University (Natural Science Edition), 2022, 59(6): 89-98. (in Chinese)
- [6] Zhang Hong, Zhang Lei, Yuan Ding, et al. Infrared small target detection based on local intensity and gradient properties[J]. Infrared Physics & Technology, 2018, 89: 88-96.
- [7] 姚庆安, 张鑫, 刘力鸣, 等. 融合注意力机制和多尺度特征的图像语义分割[J]. 吉林大学学报(理学版), 2022, 60(6): 1383-1390.
Yao Qing'an, Zhang Xin, Liu Liming, et al. Image semantic segmentation based on fusion of attention mechanism and multi-scale features[J]. Journal of Jilin University (Science Edition), 2022, 60(6): 1383-1390. (in Chinese)
- [8] Dong Lili, Wang Bin, Zhao Ming, et al. Robust infrared maritime target detection based on visual attention and spatiotemporal filtering[J]. IEEE Transactions on Geoscience & Remote Sensing, 2017, 55(5): 3037-3050.
- [9] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259.
- [10] Gregoris D J, Simon K W, Tritchew S, et al. Wavelet transform-based filtering for the enhancement of dim targets in FLIR images[C]. Wavelet Applications. SPIE, 1994.
- [11] Wu Yiquan, Luo Zijuan, Wu Wenyi. A method of small target detection in infrared image based on nonsubsampling contourlet transform[J]. Journal of Image and Graphics, 2009, 14(3): 477-481.
- [12] Liu Li, Ouyang Wanli, Wang Xiaogang, et al. Deep Learning for generic object detection: A survey[J]. International Journal of Computer Vision, 2020, 128(2): 261-318.
- [13] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. IEEE Conference on Computer Vision & Pattern Recognition, 2014: 580-587.
- [14] Girshick R. Fast R-CNN[C]. IEEE International Conference on Computer Vision, 2015: 1440-1448.
- [15] Ren Shaoqing, He Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.
- [16] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. IEEE Conference on Computer Vision & Pattern Recognition, 2016: 779-788.
- [17] Liu Wei, Anguelov D, Erhan D, et al. SSD: Single shot multiBox detector[C]. European Conference on Computer Vision, 2016: 21-37.
- [18] Duan Kaiwen, Bai Song, Xie Lingxi, et al. CenterNet: Keypoint triplets for object detection[C]. IEEE/CVF International Conference on Computer Vision, 2019: 6569-6578.
- [19] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]. European Conference on Computer Vision, 2020: 213-229.
- [20] Zhu Xizhou, Su Weijie, Lu Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]. International Conference on Learning Representations, 2020.
- [21] Dai Jifeng, Qi Haozhi, Xiong Yuwen, et al. Deformable convo-

- lutional networks[C]. Proceedings of the IEEE International Conference On Computer Vision, 2017: 764-773.
- [22] Dai Zhigang, Cai Boolun, Lin Yugeng, et al. UP-DETR: Unsupervised pre-training for object detection with transformers[C]//Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition, 2021: 1601-1610.
- [23] An Quanzhi, Pan Zongxu, Liu Lei, et al. DRBox-v2: An improved detector with rotatable boxes for target detection in SAR images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(11): 8333-8349.
- [24] Yuan Yuan, Zhang Yuanlin. OLCN: An optimized low coupling network for small objects detection[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.
- [25] Wu Zhihao, Liu Chengliang, Wen Jie, et al. Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss[J]. IEEE Transactions on Image Processing, 2022, 32:682-693.

Infrared UAV Detection Algorithm Based on Swin Transformer and Attention Mechanism

Wang Siyu¹, Lu Ruitao^{1,2}, Huang Pan³, Yang Xiaogang¹, Xia Wenxin¹, Li Qingge¹, Zhang Zhenyu¹

1. Rocket Force University of Engineering, Xi'an 710025, China

2. Science and Technology on Electro-optic Control Laboratory, Luoyang 471000, China

3. AVIC Xi'an Aeronautics Computing Technique Research Institute, Xi'an 710068, China

Abstract: Infrared drone target detection has broad application prospects in both military and civilian fields, and it is a hot research topic in the field of computer vision. Due to the small scale of drone targets and the complex and ever-changing aerial environment, existing detection algorithms generally have low detection rates and high false alarm rates. Aiming at issues such as poor detection of infrared drone targets in complex scenarios, this article proposes a ST-YOLOA infrared unmanned aerial vehicle target detection model. Firstly, in order to improve model performance and effectively capture global information, an STCNet backbone feature extraction network is constructed using the Swin Transformer network architecture and coordinated attention (CA) mechanism; Secondly, in the feature fusion section, a PANet path aggregation network with residual structure is used to construct a feature pyramid to enhance the overall feature extraction ability, while improving the up and down sampling method to enhance detection ability; Finally, the decoupled detection head is used to predict the position of the drone target. The proposed model has a detection accuracy of 92.8% and a detection speed of 22frames/s, which is verified by experiments on an infrared drone dataset. This indicates that the model has better detection performance compared to other models, especially in complex environments, and basically meets the real-time detection requirements. It has practical significance for detection in multi drone target scenarios.

Key Words: infrared unmanned aerial vehicle; target detection; Swin Transformer; coordinated attention mechanism; STCNet

Received: 2023-07-10; **Revised:** 2023-11-29; **Accepted:** 2024-01-05

Foundation item: National Natural Science Foundation of China (62276274), Aeronautical Science Foundation of China (201851U8012), Shaanxi Natural Science Foundation(2023-JC-YB-528)