DOI: 10.13382/j.jemi.B2306431

基于特征聚合与多元协同特征交互的航拍图像小目标检测*

陈朋磊[1] 王江涛[1,2] 张志伟[1] 何 程[1]
(1. 淮北师范大学物理与电子信息学院 淮北 235000;2. 智能计算及应用安徽省重点实验室 淮北 235000)

摘 要:针对无人机航拍图像目标尺寸太小、包含的特征信息较少,导致现有的检测算法对小目标检测效果不理想的问题,提出一种基于特征聚合与多元协同特征交互的无人机航拍图像小目标检测算法。首先,针对主干网对特征提取不足的问题,采用 Swin Transformer 作为 RetinaNet 主干网络,以增强算法对全局信息的提取能力。其次,为提高网络对远处目标即小目标的检测能力,设计出一种高效的小目标特征聚合网络(SFANet),实现对浅层特征图小目标细节信息的充分整合。最后,为进一步提高网络对多尺度目标的检测性能,使低层特征信息流向高层,提出全新的多元协同特征交互模板(MCFIM)。在公开无人机航拍数据集 VisDrone2019-DET 上的实验结果表明,所提算法相较于原 RetinaNet 基线网络检测精度提高 7.6%,对于小目标具有更好的检测效果。

关键词: 小目标检测;航拍图像;小目标特征聚合网络;多元协同特征交互模块
中图分类号: TP391.4 **文献标识码**:A **国家标准学科分类代码**: 520.6040

Small object detection in aerial images based on feature aggregation and multiple cooperative features interaction

Chen Penglei[1] Wang Jiangtao[1,2] Zhang Zhiwei[1] He Cheng[1]
(1. School of Physics and Electronic Information, Huaibei Normal University, Huaibei 235000, China;
2. Ahui Province Key Laboratory of Intelligent Computing and Applications, Huaibei 235000, China)

Abstract: Aiming at the problem that the target size of the UAV aerial image is too small and contains less feature information, which leads to the unsatisfactory detection effect of the existing detection algorithm on small objects, a UAV aerial photography based on feature aggregation and multi-collaborative feature interaction is proposed. First of all, in view of the insufficient feature extraction of the backbone network, Swin Transformer is selected as the RetinaNet backbone network to enhance the global information extraction ability of the algorithm. Secondly, in order to improve the detection ability of remote targets, a small target feature aggregation network is proposed, which can fully integrate the details of small targets in shallow feature maps. Finally, in order to further improve the detection performance of multi-scale targets, a new multiple collaborative feature interaction module is proposed to make the low-level feature information flow to the high-level. Experimental results on VisDrone2019-DET, a public UAV aerial photo data set, show that compared with the original RetinaNet baseline network detection precision increased by 7.6%, the proposed algorithm has better detection effect for small targets.

Keywords: small object detection; aerial images; small target feature aggregation network; multiple collaborative feature interaction module

0 引 言

近年来,无人机因其兼顾机动性和稳定性而成为热门装备,在安防监控[1]、航拍[2]、快递[3]、农业生产[4]等诸多应用中,基于无人机航拍图像的目标检测日益成为当前的研究热点[5]。值得注意的是尽管在过去几年中卷积神经网络(convolutional neural networks, CNN)[6]对自

收稿日期: 2023-04-13 Received Date: 2023-04-13
*基金项目:国家自然科学基金(61976101)、安徽省高校自然科学研究重点项目(2023AH050319)、安徽省高校优秀科研创新团队项目(2023AH010044)资助

然图像的目标检测得到了极大的发展,但在无人机航拍图像上的目标检测,准确性和效率方面仍然受到限制。正因为小目标检测是无人机图像处理的核心问题之一,它涉及到对于细节信息的敏感性、对于小尺寸目标的检测准确率、对于大尺寸目标的鲁棒性等多方面的挑战[7]。

传统的无人机图像小目标检测方法大多采用基于手工设计特征[8]的目标检测算法,然而这类方法需要人工去进行设计与微调,其鲁棒性和泛化能力受限。随着深度学习[9]技术的快速发展,无人机图像处理中使用深度学习的技术进行目标检测成为主流[10]。以是否需要区域建议为依据,当前的无人机图像目标检测方法可分为两类:一类是基于区域建议的无人机航拍图像目标检测,如 Faster R-CNN[11]、Mask R-CNN[12]、Cascade R-CNN[13]等,以上算法利用候选区域的相关信息,对预先设定的目标区域进行分类和回归,以获取目标的位置和类别信息。另一类是基于单阶段检测的无人机航拍图像目标检测算法,这类方法直接通过网络输出目标的位置和类别,具有速度快、精度高等优点,典型的算法包括 YOLO[14]、SSD[15]、RetinaNet[16]等。然而,这些算法在处理较小目标时表现欠佳,往往会出现目标漏检、定位不准确等问题。

传统目标检测任务有很多手工设计痕迹,所以不是端到端的网络。Deformable DETR[17]运用到了 Transformer 强大的功能以及全局关系建模能力来取代目标检测中人工设计痕迹来达到端到端的目的。但是其存在收敛慢以及 Transformer 中的 Self-Attention 的计算复杂度是平方级别的,造成的特征图分辨率不能太高的问题,这就导致了小目标检测性能很差。QueryDet[18](一种查询机制,而非框架),使用了新的查询机制来加速特征金字塔的对象检测器的推理速度和减少计算量、提高准确性。可以提高小物体的检测准确性,但对于非常小的物体,仍然存在一定的限制。为了平衡标注样本的稀缺和保持多尺度分层表示的困难,PyraMIM[19]提出了一种新颖的金字塔式掩膜图像建模框架,用于航空场景的自监督预训练。在没有手动标注的情况下,PyraMIM[19]能够在预训练期间建立金字塔表示,这些表示可以无缝地适应航空物体检测,从而提高性能。但是上述算法均在特征提取部分忽视了小目标的特征表示,网络的浅层中丰富的细节信息并未得到充分利用,特别是底层中模糊遮挡的小目标信息丢失较为严重。

由于无人机航拍图像中的小目标特征信息较少,大小和形状也较为不规则,通用的目标检测方法往往难以达到预期的检测效果,特别是对于重叠或者模糊场景下的目标检测,出现糟糕的检测效果的可能性大大增加[20]。本文关注到如果头部网络没有改变的前提下,小目标检测性能不理想的原因很有可能是由于特征表示[21-22]不足。探索元素之间信息流的信息瓶颈机制[23]可以帮助解释这种现象,即不完善的颈部网络可能会导致一些与任务相关的信息丢失[24],特别是对于非判别性[25]的情况,信息丢失问题更为重要。

为解决航拍图像目标检测所存在的以上问题,本文提出了一种基于特征聚合与多元协同特征交互的航拍图像小目标检测算法。针对现有的颈部网络一般都是基于多个卷积的粗累加,其特征融合能力有限,对底层小目标细节信息的整合极为不友好,为此,本文算法在 RetinaNet 网络的基础上,通过设计小目标特征聚合网络(small target feature aggregation network, SFANet)获得丰富的底层骨干小目标特征,从而使模型更适合于密集的小目标预测任务。针对底层中的小目标在高层中容易被视为背景负样本,网络中缺乏向上融合的路径,Liu 等[26]提出的 PANet 使用多个卷积层自下而上的融合多尺度特征,融合时使用了元素相加的简单方式,未充分考虑不同特征层对小目标检测的不同贡献,这可能导致在某些情况下小目标检测性能的下降。基于此,本文提出了多元协同特征交互模块(multiple collaborative feature interaction module, MCFIM),使得低层的信息能够自适应的传递给高层,实现浅层细节特征和高层语义信息更为高效的特征协同交互,提高了航拍图像中小目标的检测精度和鲁棒性。

1　RetinaNet 目标检测框架概述

为了深入学习目标的信息并获得良好的检测效果,模型往往需要训练大量的样本。为了增强模型的鲁棒性,数据集中同一类型的目标往往具有多种特征。然而,不同的样本在训练过程中对大多数检测模型的影响是相同的,这意味着容易样本和困难样本的损失权重是相同的。虽然在目标检测中容易样本的损失较低,但大量的容易样本仍然会在总损失中占主导地位,这将会降低模型检测困难样本的能力。为了减少样本不平衡对模型的影响,Lin 等[16]提出了 Focal loss 并将其应用于他们的 RetinaNet 模型。

RetinaNet 是一个将主干网、特征金字塔网络(feature pyramid network, FPN)和特定任务子网结合在一起的检测网络,Focal loss 应用于子网以提高模型的检测能力。RetinaNet 的结构可以分为 3 个部分,如图 1 所示。在其第 1 部分中,输入图像由 CNN 处理以获得相应的特征图;在第 2 部分中,将特征图输入到 FPN 中,提取并重组不同尺度的特征;最后将 FPN 输出的 3 种不同尺度的 feature maps 送入“box+class”子网得到目标的类别和位置。

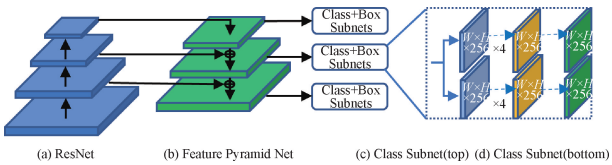

图 1　RetinaNet 框架图

Fig. 1　The architecture of RetinaNet

2　本文算法

受 Retinanet 启发,本文所提出的小目标检测架构如图 2 所示。为提高算法主干网络的全局信息提取能力,使用了 Swin Transformer 替代原先的 Backbone。此外,为充分利用浅层特征提取模块中包含的大量小目标特征信息,减少特征图中小目标细节信息的流失,设计了 SFANet,并舍弃 P7 预测特征层以减少模型的参数量。最后,提出了 MCFIM 模块,该模块在特征金字塔的基础上进一步加强了不同层级特征的交互,从而增强网络对多尺度目标的检测能力,有效地整合不同层次的特征信息,提高目标检测的准确率和召回率。

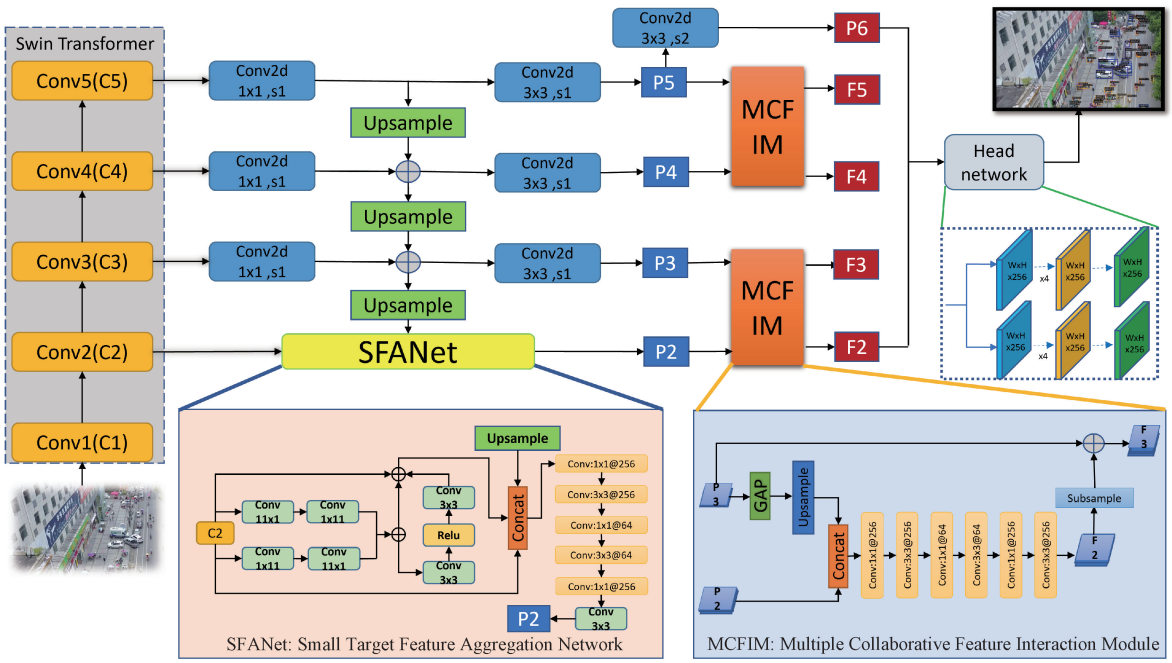

图 2　本文算法架构细节

Fig. 2　The architectural details of the algorithm in this paper

2.1　Backbone 的选取

由于 Transformer 模型在序列建模和转导任务方面取得了成功,研究人员开始将其应用于计算机视觉领域。然而,在目标检测中不同的图像和目标在尺度上存在较大差异,Transformer 模型适应这些差异较大的图像目标效果并不理想。针对上述问题,Liu 等[27] 提出了 Swin Transformer,基本框架如图 3 所示。

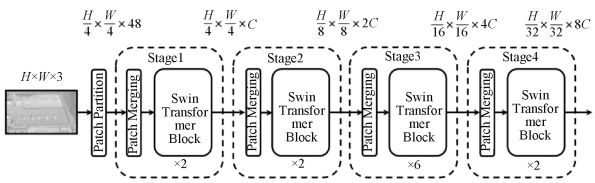

图 3　Swin Transformer 基本框架

Fig. 3　The basic framework of Swin Transformer

由于视角、尺度、光照条件和物体高密度的变化,无人机航拍图像中的物体检测比一般物体检测更具挑战性。FPN 虽然十分完美,但使用 CNN 如 ResNet 所构建的特征提取网络仍然存在一些问题,比如卷积具有局部性,导致难以获取图像的全局语义信息,而对于小尺寸、信息较少的目标而言,上下文信息显得格外重要。为解决上述问题,关键在于网络中全局的语义信息的引入,让特征提取网络更为注重相关目标的上下文信息。因此,本文选用 Swin Transformer 作为特征提取网络,有效地结合了 CNN 和 Transformer 的各自优点,使得模型可以更好地适应不同尺寸的目标和图像,以此来引入 CNN 所需有关图像的全局语义信息。

2.2　小目标特征聚合网络

无人机航拍图像目标检测存在一定难度,原因在于图像中要检测的目标尺寸过小且排列密集,而背景图像较大。另外,由于原始的 RetinaNet 模型下采样倍数较大,导致在较深的特征图上难以学习到航拍图像中小目

标的特征信息。为了解决这一问题,通过设计小目标特征聚合网络(SFANet)获取丰富的底层骨干网络中小目标特征信息,从而使模型更适合于密集的小目标预测任务,小目标特征聚合网络具体细节架构如图4所示。

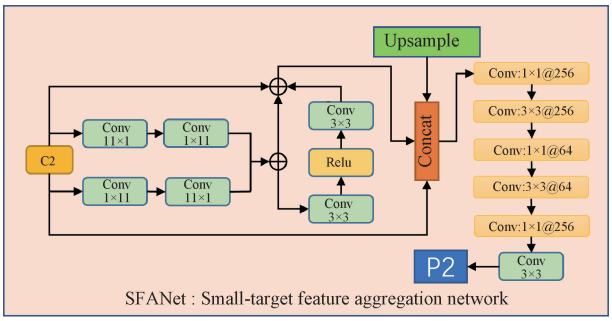

图4　小目标特征聚合网络架构图

Fig. 4　The architecture diagram of small target feature aggregation network

SFANet明确减少了小目标的特征损失,提高了模型的定位能力。首先,使用尽可能大的卷积核显著性扩展有效感受野,但本文并不是采用大核直接进行卷积,这是与大部分分类网络不同的地方,因为直接采用大核卷积会带来巨大的计算量。另外,由于底层特征图尺寸相对较大,如果采用和颈部网络上层特征图一样的方式通过1×1的卷积进行通道调整,也将会带来极大的计算成本。因此本文采用的是11×1+1×11和1×11+11×1对称式的大核卷积,仅需较少的计算量就可以充分学习到底层小目标的细节信息。

SFANet结构公式表示如下:

$$C_{2\_1} = f^{3\times3}(R(f^{3\times3}(f^{1\times11}(f^{11\times1}(C_2)))+f^{11\times1}(f^{1\times11}(C_2))))) \tag{1}$$

$$P_2 = f^{3\times3}(f^{1\times1}(f^{3\times3}(f^{1\times1}(f^{3\times3}(f^{1\times1}(Cat(C_2,C_{2\_1},C_{3\_1})))))))) \tag{2}$$

式中:$f^{1\times1}$和$f^{3\times3}$表示标准的1×1和3×3卷积,$f^{11\times1}$和$f^{1\times11}$分别为11×1和1×11的大核卷积,$R(\cdot)$代表ReLU激活函数,C_2表示特征提取网络中第2层特征图,$Cat(\cdot)$代表通道级联操作,C_{3_1}特征图由FPN模块经过上采样操作得到80×80尺度的特征图后,再进行一次上采样操作得到。

具体来说,RetinaNet模型结构中的FPN并未充分利用特征提取模块(Backbone模块)中的浅层特征图,而这些浅层特征图中包含更多小目标的特征信息。因此,在FPN模块经过上采样操作得到80×80尺度的特征图后,再进行一次上采样操作,将生成的160×160特征图与Backbone模块中浅层的160×160尺度的特征图经过SFANet后的特征图进行通道级联操作,通过叠加卷积层来减少信息损失,并将得到的特征图输入到预测模块中,以得到一个针对小目标的预测层。这里为了减少模型的复杂度和参数量,本文中舍弃了深层的5×5尺度的预测特征层,以保持预测特征层个数与原RetinaNet模型一致。

2.3　多元协同特征交互

在特征金字塔中,高层语义所代表的大目标信息通过自上而下的路径向低层传递,因此低层中的大目标就不会被误判为背景负样本。然而,网络中由于缺乏向上融合的路径,低层中的中小目标在高层中仍然容易被视为背景负样本,这将不利于对无人机航拍图像的多尺度目标的检测。基于此,本文提出的MCFIM模块,通过相邻两层特征之间的自适应协同交互,将低层的小目标信息传递给高层,从而避免高层特征中小目标被误判为背景而导致漏检。值得一提的是,此模块不仅提高了检测性能,同时还保持了网络的高效性和可训练性。

FPN输出的预测特征层有P_2、P_3、P_4、P_5和P_6,发现当对给定的{P_2、P_3、P_4、P_5、P_6}5层特征层进行相邻两两协同交互之后会出现较明显的特征冗余现象,所以在交互的时候舍弃P_6特征层,只用剩下的4层特征进行两两协同交互。因此MCFIM包含了两部分,分别是高层协同交互和低层协同交互,即给定特征图{P_2、P_3、P_4、P_5}$\in R^{C\times H\times W}$,特征协同交互后得到的特征图为{$F_2$, F_3, F_4, F_5}。

本文发现在协同交互过程中,直接相加并不是一种合理的跨尺度融合方法,因为不同尺度的特征图存在语义信息缺口。与加法相比,通道级联操作保留了更多的特征信息,但也增加了模型参数的数量和计算量。为此,每隔一段时间采用1×1卷积降维,缓解卷积瓶颈。在此基础上,实现了叠加卷积层,消除了插值引起的混叠效应,减少了通道中的信息损失,增强了特征表征能力,具体的交互过程如图5所示。

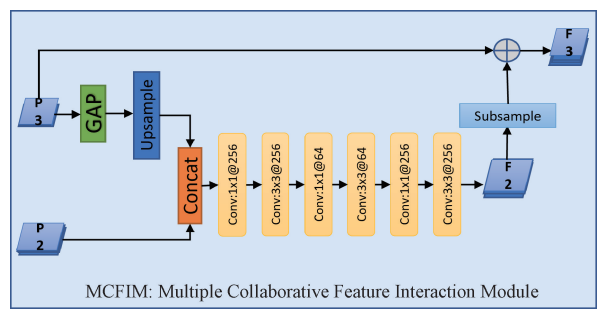

图5　多元协同特征交互模块架构图

Fig. 5　The structure of multiple collaborative feature interaction modules

以底层的协同交互为例,公式表示如下:

$$Z = f^{3\times3}(f^{1\times1}(f^{3\times3}(f^{1\times1}(f^{3\times3}(f^{1\times1}(X)))))) \tag{3}$$

$$F_2 = Z(\mathrm{Concat}(Upsample(Gap(P_3)), P_2)) \quad (4)$$

$$F_3 = Subsample(F_2) + P_3 \quad (5)$$

式中：Z 表示叠加的卷积层，$f^{1\times1}$ 和 $f^{3\times3}$ 表示标准的 1×1 和 3×3 卷积，X 为通道级联后的特征图。$Gap(\cdot)$ 表示对特征图进行全局平均池化操作，$Upsample(\cdot)$ 表示采用最近邻插值上采样操作，$Concat(\cdot)$ 表示通道级联操作，P_2 和 P_3 分别为 FPN 输出的底层两个特征图，$Subsample(\cdot)$ 表示对特征图进行 2 倍下采样操作。

全局平均池化操作可生成具有丰富语义信息的特征图，而低层特征与高层特征的协同交互能够增强高层特征的细节信息，避免将高层特征中的中小目标误判为背景而产生漏检。最终生成的这 5 层特征包含了丰富的全局语义信息和小目标细节信息，有助于更好地检测无人机航拍图像中不同尺度的目标。

2.4　损失函数

与 RetinaNet 一致，本文算法的损失函数依然由两部分组成，一部分是分类损失，另一部分是回归损失，计算公式如下：

$$Loss = \frac{1}{N_{pos}}\sum_i L_{cls}^i + \frac{1}{N_{pos}}\sum_j L_{reg}^j \quad (6)$$

其中，N_{pos} 代表正样本的个数，i 为所有的正负样本，j 为所有的正样本，L_{cls} 表示用 Focal loss 作为分类损失，L_{reg} 代表用 Smooth-L1 Loss 函数作为回归损失。Focal loss 是一种用于解决类别不平衡的损失函数，特别是对于存在大量容易被误分类的负样本的情况。其中它们的定义形式如下：

$$L_{cls} = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (7)$$

$$Smooth_{L1} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{其他} \end{cases} \quad (8)$$

式中：p_t 表示模型对于样本的预测概率；α_t 是类别权重；γ 是一个可调的超参数，控制了易分类样本权重下降速度的快慢。

3　实验结果与分析

3.1　数据集

本文实验采用 VisDrone2019-DET 数据集，该数据集涵盖了拍摄于国内 14 个不同城市的 8 629 张图像，并标注了 10 种类型的目标，包括行人、自行车、人、汽车、卡车、敞篷三轮车、三轮车、面包车、摩托车和公交车。训练集、验证集和测试集的图片数量分别为 6 471、548 和 1 610 张，本文采用测试集来评估所提出的算法。此外，为了验证算法的扩展性，本文还在 NWPU VHR-10 数据集上进行了实验，按 6∶2∶2 的比例将包含目标的 650 幅图像划分为训练集、验证集和测试集，涵盖了飞机、舰船等 10 个类别。

3.2　参数设置及评价指标

本文在 NVIDIA GeForce RTX 3090 上使用了 Pytorch 中的 mmdetection 代码库进行实验，使用 CUDA 版本为 11.3，编译语言为 Python3.7。训练最小批次为 2，输入图像尺寸大小设置为 1 080×720，使用 SGD 优化器，动量因子为 0.9，初始学习率为 0.001 25，权重衰减因子为 0.000 1，总共训练 20 个 epoch。

选取平均精准率均值（mean average precision，mAP），mAP0.5，mAP0.75，mAPs，参数量 Params 以及检测速率（frame per second，FPS）作为评价指标。AP 是一种针对类别精度的评价指标，mAP 为不同类别的 AP 值相加再除以数据集的类别数而得到。而 mAP0.5、mAP0.75 分别以 IoU=0.5 和 0.7 时进行计算，mAPs 为小目标（像素小于 32×32）检测的平均准确度。平均精度 AP 和平均精度均值 mAP 计算过程如式（9）~（12）所示：

$$precision = \frac{TP}{TP + FP} \quad (9)$$

$$recall = \frac{TP}{TP + FN} \quad (10)$$

$$AP = \int_0^1 precision(recall)\,\mathrm{d}(recall) \quad (11)$$

$$mAP = \frac{\sum_i^N AP_i}{N} \quad (12)$$

式中：N 表示数据集类别个数；TP 表示正确识别为正样本的正样本数；FP 表示错误识别为正样本的负样本数；FN 表示错误识别为负样本的正样本数。

3.3　VisDrone 数据集的实验结果

1）消融实验

在训练过程中记录了本文算法在 VisDrone 数据集上训练损失值，其损失曲线如图 6 所示。其中横坐标表示训练的轮次（iter）数，纵坐标表示训练损失值。从损失函数曲线图的收敛情况看，当模型训练 1 200 轮左右，也即 20 个 epoch 时，网络的损失函数趋于收敛，训练结果较为理想。

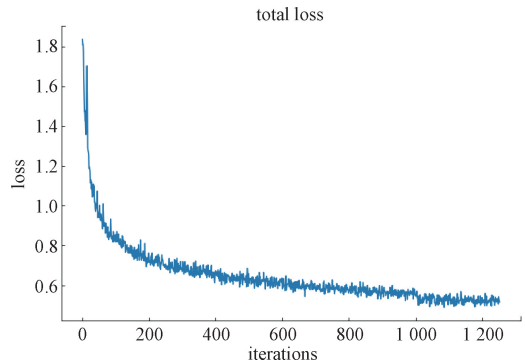

图 6　损失函数曲线

Fig. 6　Loss function curve

为了验证本文所提基于特征聚合与多元协同特征交互的无人机航拍图像目标检测算法的有效性,针对不同模型进行了消融实验,结果如表1所示,所有实验均在VisDrone2019数据集上进行测试。在RetinaNet的主干网中采用Swin Transformer来取代ResNet50得到模型a;在模型a的基础上新增小目标特征聚合网络得到模型b,去除最深预测特征层以减少模型参数;在保证模型b的前提下加入多元协同特征交互模块得到最终模型c。

表1　VisDrone数据集消融实验结果

Table 1　Ablation results of VisDrone dataset

模型	Baseline	swin	SFANet	MCFIM	mAP/%	mAP0.5/%	mAPs/%	参数量	FPS
RetinaNet	√				12.6	23.7	4.3	32.16	24.8
模型a	√	√			13.5	25.8	5.2	35.83	27.2
模型b	√	√	√		14.4	28.5	7.5	39.06	19.3
模型c	√	√	√	√	15.8	31.3	9.0	39.08	18.6

由表1可以看出,在用swin Transformer替换ResNet50作为主干后,模型a提高了模型主干的特征提取,mAP相比于基准模型提高了0.9%,mAP0.5相比于基准模型提高了2.1%。模型b由于小目标特征聚合网络的增加,mAP相比于基准模型提高了1.8%,mAP0.5相比于基准模型提高了4.8%,mAPs相比于基准模型提高了3.2%,由此可见,小目标特征聚合网络对航拍图像小目标有很强的检测能力。由于多元协同特征交互模块的引入,模型c相比模型b的mAP提高了1.4%,mAP0.5相比模型b提高了2.8%,mAPs相比相比模型b提高了1.5%,验证了多元协同特征交互模块的可靠性。综上,在本文最终模型上达到的检测效果最佳,mAP相较于基准模型RetinaNet提升3.2%,mAP0.5提升7.6%,mAPs提升4.7%,进一步验证了本文算法的有效性。

2)对比试验及其可视化结果

在本节中,将提出的算法与当前经典的RetinaNet[16]、Faster R-CNN[11]、CenterNet[28]、Refinedet[29]、RFB-Net[30]、De-DETR[17]及QueryDet[18]模型进行了比较。平均精度均值(mAP),mAP0.5,mAP0.75的对比结果如表2所示。所有实验同样在VisDrone2019数据集上进行,根据表2实验结果来看,本文算法的mAP、mAP0.5、mAP0.75均高于其他经典的目标检测算法。综合来看,本文算法与其他的先进算法的比较中,具有一定的优势,特别是相较于经典的检测算法CenterNet[28]和Faster R-CNN[11],在检测精度上有大幅提高,可见更适用于无人机航拍图像小目标的检测任务。

针对不同模型算法计算量的分析,表2中展示了其他先进算法与本文算法计算量的对比。在相同的实验条件下,分别计算其他先进算法的FLOPs与本文算法进行对比,FLOPs是指模型执行的浮点运算的总数量,它通常用于衡量模型的计算复杂性,较高的FLOPs表示模型需要更多的计算资源。得益于本文算法模块设计的轻量化,相比于基线算法RtinaNet[16]的FLOPs只提高了13.07,相较于基于相同基线改进的算法QueryDet[18],本文算法FLOPs降低17.09。在与其他先进算法的FLOPs比较,本文算法计算量最低,特别是FLOPs相较于经典的双阶段目标检测算法Faster R-CNN[11]低280.07。在上述计算量的对比试验和分析中,可以看出本文算法在计算量方面的的优势,可以更好地实现航拍图像小目标的实时性检测。

为了直观地感受本文算法对航拍图像中小目标的检测能力,图7展示了本文算法和基线在3组不同场景下无人机航拍图像检测结果的细节对比,第1行采用的是RetinaNet基线算法,第3行采用本文算法对其检测结果可视化,在对比中验证了本文算法的优异性能。从第1组远处小目标、第2组夜间场景下行人遮挡的典型困难检测任务和第3组无人机俯视图像中摩托车等小目标与

表2　与其他算法对比试验结果

Table 2　Test results compared with other algorithms

算法	Backbone	mAP/%	mAP0.5/%	mAP0.75/%	FLOPs/G
RetinaNet[16]	ResNet50	12.8	24.00	12.50	**152.16**
Faster R-CNN[11]	VGG-16	11.3	18.20	10.60	445.50
CenterNet[28]	ResNet50	12.4	22.70	12.40	173.87
RefineDet[29]	ResNet50	14.8	27.86	13.17	459.27
RFB-Net[30]	ResNet50	14.3	22.17	12.06	376.92
De-DETR[17]	ResNet50	14.9	30.10	12.60	327.61
QueryDet[18]	ResNet50	15.4	30.80	13.70	182.32
本文算法	Swin Transformer	**15.8**	**31.30**	**14.50**	165.23

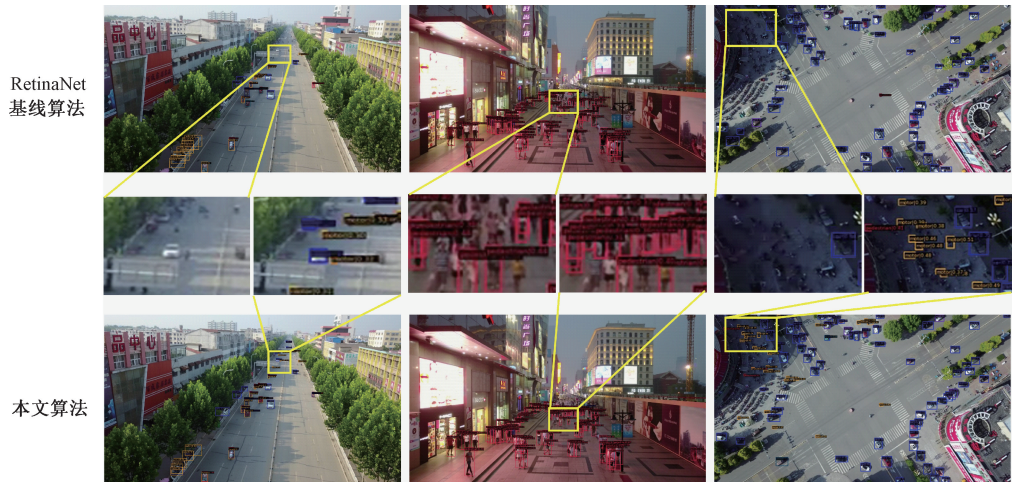

图 7　本文算法与基线在不同场景下检测结果对比

Fig. 7　Comparison of detection results between the algorithm in this paper and the baseline in different scenarios

背景高度混杂都可以看出,针对无人机图像中小尺度、背景复杂和排列密集的目标,RetinaNet 基线算法存在部分小目标漏检的情况,因此无论从主观视觉检测效果还是客观试验数据结果来看,本文算法对无人机航拍图像中小目标具有强大的检测能力。

3.4　NWPU VHR-10 数据集上实验结果

与 VisDrone2019-DET 数据集相比,NWPU VHR-10 数据集中目标的尺寸和特征都存在较大的差异,同时尺度变化也更加明显。为了验证本文所提算法的扩展性和有效性,在 NWPU VHR-10 数据集上进行了消融和对比实验。模型 a 在 Backbone 选取上,采用 Swin Transformer 来取代 ResNet50;模型 b 在模型 a 的基础上新增小目标特征聚合网络,去除最深预测特征层以减少模型参数;最终模型 c 在模型 b 的基础上添加多元协同特征交互模块。如表 3 所示,所有实验均在 NWPU VHR-10 的数据集上进行测试,由数据结果可知,本文算法相较于基线 mAP0.5 提高了 5.3%,增强了对尺度变化较大的遥感图像目标的检测性能。

表 3　NWPU VHR-10 数据集上消融实验结果

Table 3　Ablation results on the NWPU VHR-10 dataset

模型	Baseline	swin	SFANet	MCFIM	mAP0.5/%
RetinaNet	√				87.9
模型 a	√	√			89.8
模型 b	√	√	√		91.6
模型 c	√	√	√	√	93.2

在 NWPU VHR-10 数据集上,本文算法和其他主流算法的实验结果对比如表 4 所示,涉及 10 个目标类别。从表中可以看出,本文算法在田径场、舰船、储油罐和车辆的检测精度方面均优于其他模型,特别是对于像素较大的物体(如飞机、棒球场、田径场、储油罐),本文模型的检测精度高达 95%以上;对于像素较小的物体(如舰船和车辆),本文模型的检测精度也超过了 93%,达到所有模型最高检测水平。这表明本文模型具备较强的小目标检测和多尺度检测能力,适用于航拍图像和遥感图像目标的检测任务。

表 4　与其他算法在 NWPU VHR-10 数据集上比较

Table 4　Comparison with other algorithms on the NWPU VHR-10 dataset

目标类别	Faster R-CNN[11]	CenterNet[28]	ODDP[31]	FMSSD[32]	De-DETR[17]	YOLOX-L[33]	PyraMIM[19]	本文算法
飞机	97.7	99.8	90.8	99.7	99.6	**99.9**	99.4	99.0
棒球场	94.1	92.0	90.6	98.2	98.2	**98.7**	98.3	96.3
篮球场	78.4	78.5	90.9	**96.8**	93.2	96.7	91.3	92.7
桥梁	72.6	**81.5**	78.8	80.1	75.1	61.9	77.8	76.4
田径场	97.0	71.8	89.6	97.7	97.1	94.7	98.1	**98.5**
港口	84.0	75.1	89.8	77.6	92.2	**96.1**	87.1	89.6
舰船	72.8	87.7	88.3	89.9	92.1	86.7	91.1	**93.8**
储油罐	81.8	92.0	86.2	90.3	92.5	96.6	96.2	**98.4**
网球场	83.4	85.3	81.2	86.0	90.2	**92.8**	92.6	92.6
车辆	56.2	77.5	78.8	88.2	93.2	88.8	92.2	**94.3**
mAP0.5/%	81.8	84.1	86.7	90.4	90.5	91.3	91.4	**93.2**

对比表格中参考的先进算法,特别是 YOLOX_L[33],由于网络层数的加大,网络对感受野的不断提高,对存在数量较少的大型目标(如飞机、棒球场和篮球场等)检测性能的提高,但这种粗鲁的提高神经网络深度的方式在略微提高大型目标检测的同时,不可避免的带来参数量的加大和计算成本的增加。这对于航拍图像中存在的较多的小目标检测是得不偿失的,在最终综合性能 mAP0.5 和航拍图像中存在数量极多的小目标(如舰船、储油罐和车辆等)的检测性能可以看出,本文算法相对于其他先进算法的检测能力具有较大的提升。

本文算法在 NWPU VHR-10 数据集上的检测结果可视化如图 8 所示。可以看出,本文算法对小目标,尤其是形变和尺度变化大的目标具有良好的检测效果。

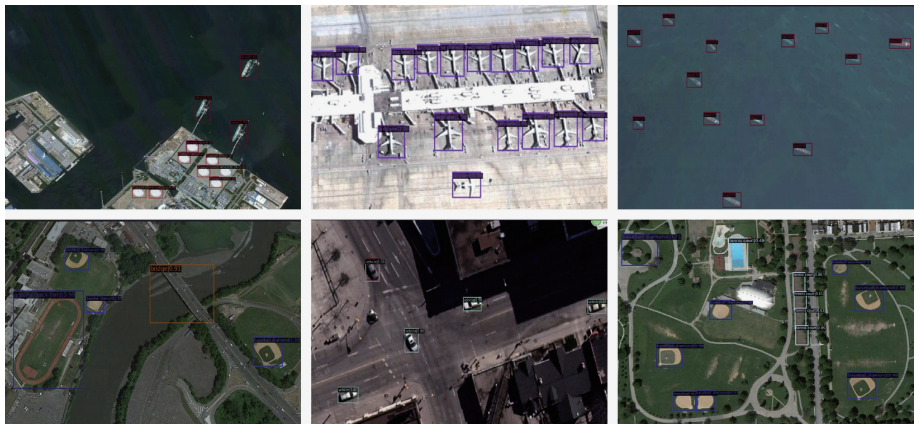

图 8　本文算法在 NWPU VHR-10 数据集上检测结果

Fig. 8　Detection results of the proposed algorithm on the NWPU VHR-10 dataset

4　结　论

本文提出了一种基于特征聚合和多元协同特征交互的航拍图像小目标检测算法。针对无人机航拍图像小目标的特征较少,特别在重叠或者模糊的场景下,通用的目标检测器在小目标检测上精度低的问题,设计了小目标特征聚合网络(SFANet)缓解小目标在颈部网络中的信息流失,极大丰富了小目标在网络中的特征表示,提高了网络对小目标的检测能力。提出的多元协同特征交互模块(MCFIM),能有效地将相邻的两层特征进行信息交互,从而避免高层特征的小目标被当成背景而造成漏检,可以较大程度上提高模型对小目标的检测能力。

实验结果证实,本文算法对航拍图像中小尺寸目标具有更强的检测能力,平均精度均值达到最优水平,同时有效减少了漏检和误检率,能很好的应对航拍图像的检测任务。在未来的工作中,将继续研究无人机航拍图像的小目标检测算法,在保证小目标实时性检测的前提下进一步提升精准度。

参考文献

[1] KALEEM Z, REHMANI M H. Amateur drone monitoring: State-of-the-art architectures, key enabling technologies, and future research directions[J]. IEEE Wireless Communications, 2018, 25(2): 150-159.

[2] 刘军黎,刘晓锋,邱洁,等. YOLOX-IM:一种无人机航拍视频的轻量化交通参数提取模型[J]. 国外电子测量技术,2023,42(1):159-169.
LIU J L, LIU X F, QIU J, et al. YOLOX-IM: A lightweight traffic parameter extraction model for UAV aerial images [J]. Foreign Electronic Measurement Technology, 2023,42(1):159-169.

[3] FRACHTENBERG E. Practical drone delivery [J]. Computer, 2019, 52(12): 53-57.

[4] ASLAN M F, DURDU A, SABANCI K, et al. A comprehensive survey of the recent studies with UAV for precision agriculture in open fields and greenhouses[J]. Applied Sciences, 2022, 12(3): 1047.

[5] 史雨馨,朱继杰,凌志刚. 基于特征增强 YOLOv4 的无人机检测算法研究[J]. 电子测量与仪器学报,2022,36(7):16-23.
SHI Y X, ZHU J J, LING ZH G. Research on UAV detection method based on feature enhanced YOLOv4 algorithm[J]. Journal of Electronic Measurement and Instrumentation,2022,36(7):16-23.

[6] BHARATI P, PRAMANIK A. Deep learning techniques—R-CNN to mask R-CNN: A survey [J]. Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019, 2020: 657-668.

[7] HUANG Y, CHEN J, HUANG D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery[C]. Proceedings of the AAAI Conference on Artificial Intelligence,2022, 36(1): 1026-1033.

[8] 徐晓光,李海. 多尺度特征在 YOLO 算法中的应用研

究[J]. 电子测量与仪器学报,2021,35(6):96-101.
XU X G, LI H. Application research of multi-scale features in YOLO algorithm[J]. Journal of Electronic Measurement and Instrumentation, 2021,35(6):96-101.
[9] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553):436.
[10] LIU W, QUIJANO K, CRAWFORD M M. YOLOv5-Tassel: Detecting tassels in RGB UAV imagery with improved YOLOv5 based on transfer learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 8085-8094.
[11] REN S Q, HE K M, 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.
[12] HE K M, GEORGIA G, PIOTR D, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.
[13] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018: 6154-6162.
[14] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[J]. Computer Science, 2015, DOI: 10. 1109 /CVPR. 2016. 91.
[15] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
[16] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017 (99): 2999-3007.
[17] ZHU X, SU W, LU L, et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]. International Conference on Learning Representations,2021.
[18] YANG C, HUANG Z, WANG N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 13668-13677.
[19] ZHANG C, LIU T, JU Y, et al. Pyramid masked image modeling for transformer-based aerial object detection[C]. 2023 IEEE International Conference on Image Processing (ICIP). IEEE, 2023: 1675-1679.
[20] ZHAO H, ZHANG H, ZHAO Y. YOLOv7-sea: Object detection of maritime UAV images based on improved YOLOv7[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 233-238.
[21] DONG Z, HANWANG Z, JINHUI T, et al. Self-regulation for semantic segmentation[J]. Computer Science, 2021, arXiv preprint arXiv:2108. 09702.
[22] DING J, XUE N, LONG Y, et al. Learning RoI transformer for detecting oriented objects in aerial images[J]. Computer Science, 2018, arXiv preprint arXiv:1812. 00155.
[23] SAXE A M, BANSAL Y, DAPELLO J, et al. On the information bottleneck theory of deep learning[J]. Journal of Statistical Mechanics: Theory and Experiment, 2019, 2019(12): 124020.
[24] LIU Q, TAN Z, CHEN D, et al. Reduce information loss in transformers for pluralistic image inpainting[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11347-11357.
[25] ZHANG D, SUN Y, YE Q, et al. Recursive discriminative subspace learning with ℓ_1-norm distance constraint[J]. IEEE Transactions on Cybernetics, 2018, 50(5): 2138-2151.
[26] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.
[27] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
[28] ZHOU X, WANG D, KRHENBUHL P. Objects as points[J]. Computer Science, 2019, DOI: 10. 48550/arXiv. 1904. 07850.
[29] ZHANG S, WEN L, LEI Z, et al. RefineDet++: Single-shot refinement neural network for object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(2): 674-687.
[30] LIU S, HUANG D. Receptive field block net for accurate and fast object detection[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 385-400.
[31] 王生生,王萌,王光耀. 基于深度神经网络剪枝的两阶段遥感图像目标检测[J]. 东北大学学报(自然科学版),2019,40(2):174-179.
WANG SH SH, WANG M, WANG G Y. Deep neural network pruning based two-stage remote sensing image

object detection[J]. Journal of Northeastern University Natural Science, 2019, 40(2): 174-179.

[32] WANG P, SUN X, DIAO W, et al. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 58(5): 3377-3390.

[33] GE Z, LIU S, WANG F, et al. YOLOX: Exceeding YOLO series in 2021[J]. Computer Science, 2021, arXiv preprint arXiv:2107.08430.

作者简介

陈朋磊,2022年于淮北师范大学获得学士学位,现为淮北师范大学研究生,主要研究方向为目标检测、深度学习。

E-mail: 12211060763@chnu.edu.cn

Chen Penglei received his B. Sc. degree from Huaibei Normal University in 2022. Now he is a M. Sc. candidate in Huaibei Normal University. His main research interests include object detection and deep learning.

王江涛(通信作者),2002年于青岛大学获得学士学位,2005年于青岛大学获硕士学位,2008年于南京理工大学获得博士学位,现为淮北师范大学教授,主要研究方向为图像处理、模式识别、计算机视觉。

E-mail: jiangtaowang@chnu.edu.cn

Wang Jiangtao (Corresponding author) received B. Sc. degree from Qingdao University in 2002, M. Sc. degree from Qingdao University in 2005 and Ph. D. degree from Nanjing University of Technology in 2008, respectively. Now he is a professor in Huaibei Normal University. His main research interests include image processing, pattern recognition and computer vision.