聚焦小目标的航拍图像目标检测算法

张 智[1],易华挥[1],郑 锦[2]
(1. 中国民航大学计算机科学与技术学院,天津 300300;2. 北京航空航天大学计算机学院,北京 100191)

摘 要: 与通用目标检测不同,无人机(Unmanned Aerial Vehicle,UAV)航拍图像目标检测主要面临两个难题:(1)远距离观察下存在大量小尺寸目标,难以与背景区分;(2)大量区域中目标密集且存在严重遮挡. 因此,将通用目标检测器直接应用于航拍图像会导致检测精度下降. 本文提出一种聚焦小目标的航拍图像目标检测算法(Focusing on Small objects Detector in aerial images,FocSDet). 针对小目标,通过密集高级组合(Dense Higher-Level Composition,DHLC)模式连接双Swin-Transfomer骨干网络,并和特征金字塔(Feature Pyramid Networks,FPN)结合,构建小目标特征聚合网络作为FocSDet的骨干网络,可丰富单层特征表达并提升对图像全局信息的利用,在不损失大目标语义信息的同时得到对小目标更好的特征描述,有效提升了小目标检测能力;针对区域密集遮挡,提出任务平衡样本分配策略,区别于现有样本分配策略只依赖定位位置,本文所提出的策略中样本匹配质量评价分数由定位位置信息和预测分类分数共同构成. 基于该新评价分数不断迭代更新样本分配和监督网络优化,取得了更高质量的预测结果. 最后,在检测头的分类和回归分支中引入层注意力构成增强检测头,进一步提升了小目标的检测性能. 在Visdrone无人机数据集、CARPK航拍数据集上的实验表明,本文提出的FocSDet相较于现有方法ATSS和VFNET,在Visdrone上平均精度(Average Precision,AP)分别提升2%和0.6%,小目标AP_s分别提升2.6%和1.2%;在CARPK上AP分别提升2.2%和1.7%,小目标AP_s分别提升5.2%和5.0%.

关键词: 航拍图像;目标检测;小目标特征聚合网络;任务平衡样本分配;增强检测头
基金项目: 国家重点研发计划(No.2020YFB1600101);国家自然科学基金(No.61876014)
中图分类号: TP391.4; **文献标识码:** A **文章编号:** 0372-2112(2023)04-0944-12
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220313

Focusing on Small Objects Detector in Aerial Images

ZHANG Zhi[1], YI Hua-hui[1], ZHENG Jin[2]
(1. *School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*;
2. *School of Computer, Beihang University, Beijing* 100191, *China*)

Abstract: Different from general object detection in natural images, object detection in unmanned aerial vehicle (UAV) aerial images mainly faces these challenges such as large number of small objects in remote observation, which is difficult to distinguish from the background, and dense objects with serious occlusion in lots of areas. Therefore, the direct application of general object detector to aerial images will lead to the decline of detection performance. In this paper, an aerial image object detection algorithm focusing on small objects (FocsDet) is proposed. For small objects, a small object feature aggregation network is designed, which connects the dual Swin-Transfomer backbone network through dense higher-level composition (DHLC) mode and combines with feature pyramid networks (FPN), so as to improve the utilization of global image information, enrich single-layer feature expression, and obtain better feature description of small objects without losing semantic information of large objects. It effectively improves the detection performance of small object. For regional dense occlusion, a task-balance label assignment is proposed, in which the label matching quality evaluation score is composed of location cost and classification cost, which is different from the existing evaluation score which only depends on location cost. Based on the evaluation score, label assignment and supervision network optimization are updated iteratively, so as to achieve better prediction results. Finally, layer attention is introduced into the classification and regression branches of the detection head to form enhanced detection head, which further improves the detection performance of small objects. Experiments on Visdrone dataset and CARPK dataset show that compared with the existing methods such as ATSS and VF-

收稿日期:2022-03-26;修回日期:2022-05-10;责任编辑:朱梅玉

NET, the average precision (AP) of FocsDet is improved by 2% and 0.6%, AP_s is improved by 2.6% and 1.2% on Vsidrone dataset respectively. On CARPK dataset, AP increases by 2.2% and 1.7%, and AP_s increases by 5.2% and 5.0% respectively.

Key words: aerial images; object detection; small object feature aggregation network; task-balance label assignment; enhanced detection head

Foundation Item(s): National Key Research and Development Program of China (No.2020YFB1600101); National Natural Science Foundation of China (No.61876014)

1 引言

近年来,深度神经网络快速发展,通用目标检测器(例如:Faster R-CNN[1],YOLOv4[2])取得了巨大成功,尤其是其中的单阶段 anchor-free 检测器(例如:FSAF[3],CenterNet[4],FCOS[5]),相比 anchor-based 的检测器,它们的参数更少,也更简洁,更符合端到端设计理念,因此在现实中应用更广. 然而,与其他类型的通用目标检测器所面临的问题类似,尽管 anchor-free 检测器在自然图像如 MS COCO[6]上表现十分优异,但在面对航拍图像如 VisDrone[7]时检测性能明显下降. 主要原因是远距离拍摄的航拍图像中存在大量小尺寸目标,同时存在大量区域目标密集且目标被遮挡. 这些因素对目标检测造成了极大的干扰,降低了检测精度.

小目标的存在和区域密集遮挡是航拍图像面临的典型问题. 为了验证航拍图像中目标尺寸的分布,本文以无人机数据集 Visdrone 为研究对象,对 Visdrone 和自然图像数据集 COCO 就目标尺寸分布情况进行了分析比较. 将目标尺寸分为 Extre-Small(小于 16×16 像素)、Small(大于 16×16 且小于 32×32 像素)和 Normal(大于 32×32 像素)三类. 从表 1 可以明显看出,相较于 COCO 数据集中的目标尺寸分布,Visdrone 数据集(训练集)中小于 32×32 像素的目标在数据集中的占比高达 44.70%(Extre-Small 为 12.05%, Small 为 32.65%),远远高于 COCO 中的 21.73%(Extre-Small 为 7.35%, Small 为 14.38%). 表 1 中 Visdrone 数据集中加粗部分为表中 10 类目标的合计统计情况,COCO 数据集中加粗部分是全部 80 类目标的合计统计情况. 在 people 单类中,Visdrone 的小目标占比更高达 77.44%,小于 16×16 像素的目标占比为 34.25%. 从图 1(*a*)可以直观看出 Visdrone 验证集和测试集中小目标的占比均高过 50%. 这些数据直观说明了航拍图像目标尺寸小的特点.

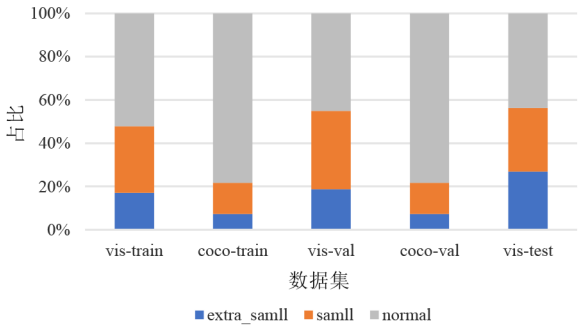
表 1　COCO 和 Visdrone 数据集中目标尺寸分布情况

数据集	数据类型	Extre-Small (小于 16×16 像素)		Small (大于 16×16 且小于 32×32 像素)		Normal (大于 32×32 像素)	
		目标个数	目标占比	目标个数	目标占比	目标个数	目标占比
Visdrone	pedestrain	18 599	23.44%	32 648	41.15%	28 090	35.41%
	people	9 269	34.25%	11 688	43.19%	6 102	22.55%
	bicycle	1 350	12.88%	4 146	39.56%	4 984	47.56%
	car	19 912	13.75%	34 352	23.71%	90 602	62.54%
	van	2 188	8.77%	5 989	24.00%	16 779	67.23%
	trunk	747	5.80%	2 189	17.00%	9 939	77.20%
	tricycle	465	9.66%	1 192	24.77%	3 155	65.57%
	awning-tricycle	286	8.81%	786	24.21%	2 174	66.97%
	bus	310	5.23%	869	14.66%	4 747	80.10%
	motor	5 758	19.42%	11 667	39.35%	12 222	41.23%
	上述 10 类合计	**38 972**	**12.05%**	**105 526**	**32.65%**	**178 794**	**55.30%**
COCO	person	19 788	7.54%	35 459	13.51%	207 218	78.95%
	bicycle	237	3.33%	1 110	15.61%	5 766	81.06%
	car	5 923	13.5%	11 268	25.69%	26 676	60.81%
	bus	37	0.61%	247	4.07%	5 785	95.32%
	truck	224	2.25%	1 086	10.89%	8 663	86.86%
	80 类合计	**63 248**	**7.35%**	**123 705**	**14.38%**	**673 408**	**78.26%**

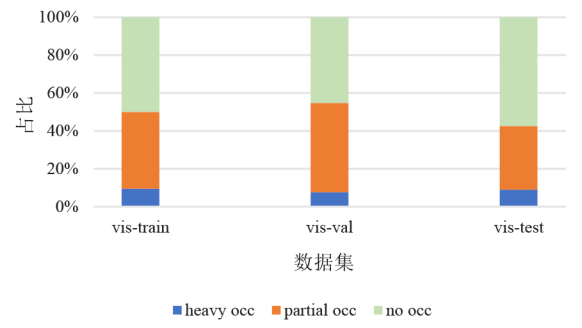
注:特别需要说明的是,COCO 将面积小于 32×32 像素的目标定义为小目标,而本文统计的小目标要求更加严格,将长、宽尺寸均小于 32 像素且面积小于 32×32 像素的目标定义为小目标;在类别方面选取了 Visdrone 全部 10 个类别,COCO 数据集中统计了全部 80 个类别的分布,但表 1 中只列出 Visdrone 数据集所包含的 5 个类别.

为了验证Visdrone数据集中目标被遮挡的情况，本文利用Visdrone数据集自带的遮挡标签对其进行了分析. 遮挡可分为部分遮挡和严重遮挡，其中部分遮挡定义为遮挡面积小于50%，严重遮挡为遮挡面积大于50%. 经过统计，Visdrone数据集（训练集）中共有图像6 471张，全部目标353 550个，被部分遮挡目标142 873个，严重遮挡目标33 804个，被遮挡比例近50%，平均每张图像有27个目标被遮挡.

在Visdrone数据集（验证集）中，平均每张图像有高达40个目标被遮挡. 从图1（*b*）可以直观看到Visdrone数据集中的遮挡比例. 这些数据不仅说明Visdrone数据集中目标分布密集，也证明了其存在大量遮挡的情况.



（*a*）Visdrone和COCO数据集目标尺寸分布情况



（*b*）Visdrone数据集遮挡分布情况

图1 Visdrone数据集目标分布特点分析

针对上述特点，航拍图像需要解决小目标、密集遮挡目标的准确检测问题. 针对航拍小目标检测，研究者们探索如何在网络中获得更鲁棒的特征，在有效检测大目标的同时提升对小目标的特征表达，进而提升小目标的检测精度. 一种典型的做法是对各种尺度特征进行融合和增强，但这类方法并未充分考虑图像全局信息和小目标的上下文信息，对小目标检测的准确性提升有限. 针对密集区域中目标相互遮挡造成的检测框置信度低、漏检误检严重的问题，研究者们主要通过定位图像密集区域进行增强检测，即获得密集目标更高置信度的检测框以提升检测性能. 但是这类方法依赖密集区域的定位效果，同时不能实现端到端的训练和推理过程. 因此，现有方法在面对航拍图像中小目标、密集遮挡目标的准确检测时依然表现不佳. 提出一个可以更好地解决上述问题，检测精度更高、通用性更强的航拍图像目标检测算法是十分迫切的.

针对上述需求，本文提出聚焦小目标的航拍图像目标检测算法（FocSDet）. FocSDet中设计了可提取更鲁棒的小目标特征表示的骨干网络-小目标特征聚合网络、更高检测置信度的任务平衡样本分配策略，以及匹配前两者的增强检测头. 具体网络结构如图2所示. 在骨干网络方面，如图2左侧红色虚线框所示，通过使用DHLC（Dense Higher-Level Composition）[8]模式将两个Swin-Transfomer网络（辅助骨干网络和引导骨干网络）进行连接，利用Swin-Transfomer-Block来获取图像全局信息；同时引入特征金字塔[9]（Feature Pyramid Networks，FPN）获得高层语义和低层细节特征，共同形成FocSDet的骨干网络-小目标特征聚合网络. 该聚合网络在不损失大目标语义信息的同时，得到对小目标更好的特征表示. 在样本分配策略方面，提出任务平衡样本分配策略，如图2右侧绿色虚线框所示. 针对ATSS[10]等方法在样本分配时只考虑定位位置信息，在密集遮挡的情况下容易产生冗余检测框，进而影响遮挡目标检测性能的弊端，引入预测分类分数，与定位分数共同构成样本匹配质量评价分数. 基于该评价分数动态更新样本分配，可生成更高质量的样本，使得分类和定位任务在网络的训练阶段达到更好的平衡，进而提升密集遮挡下的检测精度. 在检测头方面，如图2紫色虚线框所示，通过为分类和回归分支同时引入层注意力机制，为分类和回归分支分别提供更重要的分类语义特征信息和位置回归特征信息，匹配小目标特征聚合网络和任务平衡样本分配策略，进一步提升航拍图像小目标检测性能. 本文的主要贡献如下.

（1）提出小目标特征聚合网络，利用DHLC模式构建双Swin-Transfomer[11]骨干网络，进而与FPN相结合，将多个高层特征和低层特征进行融合，生成更丰富的特征表示和全局信息. 在不损失大目标语义信息的同时，得到对小目标更鲁棒的特征表示. 该特征聚合网络可直接用于各类通用检测器，对小目标均有明显的检测效果提升.

（2）提出任务平衡样本分配策略，在现有方法只利用定位位置信息的基础上引入预测分类分数，平衡定位任务和分类任务，构成可动态更新的样本匹配质量评价分数，根据该评价分数生成更高质量的检测候选框，减少冗余检测框对非极大值抑制（Non-Maximum Suppression，NMS）的干扰，提升在密集遮挡情况下的目标检测性能.

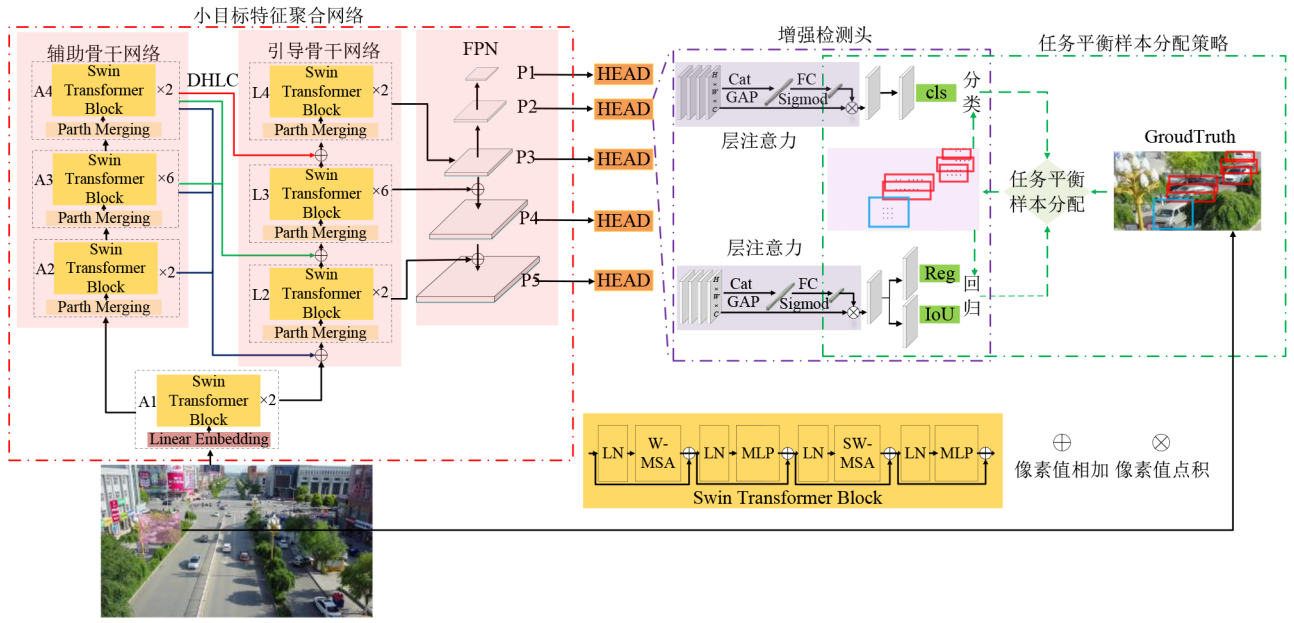


图2　FocSDet的网络结构

(3)提出增强检测头,为分类和回归分支同时引入层注意力机制,该增强检测头更适配小目标特征聚合网络和任务平衡样本分配策略,进一步增强了对小目标的检测能力.

(4)在Visdrone无人机数据集、CARPK航拍数据集上进行了实验验证. 结果表明,本文提出的聚焦小目标的航拍图像目标检测算法(Focusing on Small objects Detector in aerial images, FocSDet),相较于ATSS[10]和VFNET[12]方法,在VisDrone数据集上,AP分别提升2%和0.6%,小目标AP_s分别提升2.6%和1.2%;在CARPK[13]数据集上,AP分别提升2.2%和1.7%,小目标AP_s分别提升5.2%和5.0%.

2　相关工作

相比自然图像,航拍图像的目标检测面临更多挑战. 一方面,相比其他高分辨率的图像,航拍数据集中小目标占比近50%,极小目标占比20%左右,小尺寸目标更多. 另一方面,由于相机视点多变及场景的复杂性,往往存在某些区域目标分布稀疏,但某些区域目标分布密集,而密集区域往往同时存在大面积遮挡和小目标等情况. 因此,本文从小目标检测网络、密集遮挡解决策略两方面论述相关研究工作.

2.1　小目标检测网络

从AlexNet开始,计算机视觉骨干网络的主流一直是通过卷积的方式提取特征. 卷积的方式具有以下特点:(1)空间不变性;(2)局部性. 前者帮助网络识别不同位置的目标,而后者相比前馈神经网络,大幅降低了计算量,这二者给卷积带来精度和时间的高效性. 但是,随着卷积网络的加深,小目标特征信息逐步减弱. 例如,目标检测领域最受欢迎的ResNet[14]网络,在COCO数据集的榜单中作为重要的骨干网络,采用ResNet101性能明显优于ResNet50. 但是通过表2可以看出,在Visdrone数据集中ResNet101整体性能与ResNet50整体性能基本一致,在大目标检测精度明显提升的情况下,多种方法采用ResNet101进行小目标检测的精度反而不如ResNet50. 这表明单纯加深网络,对小目标检测性能提升基本无效. 原因在于当卷积网络变深后,输出的特征图变小,在经过多层池化后得到的更小的特征图上感知不到原图上的目标偏移,最终检测效果变差.

通过上述实验可以看出,现有骨干网络主要采用卷积提取特征,存在着提取的特征全局信息不足、当网络加深时小目标重要的细节特征丢失加剧等问题. 单纯加深网络深度并不能解决小目标检测的问题,如何提取更鲁棒的小目标特征表示是提升小目标检测精度的关键. 针对图像目标尺寸小的特点,研究者通过获得对小目标更有效的特征信息来提高检测性能. 例如,文献[17]通过拓展数据集内小目标数量,以及多次采用小目标图像进行训练,即对小目标这类困难样本再挖掘以补充特征来提高对小目标的检测性能;文献[18]在上采样后添加融合因子,控制深层信息向浅层传递的比例,避免形成的特征包含太多的语义信息、进而干扰小目标的检测效果;文献[19]通过生成更细粒度的特征表示来提升小目标的召回率;文献[20,21]针对航拍图像中的小目标,设计对特征进行不同尺度或比例的融合和增强,挖掘对小目标有效的信息并提取对小

表2 ResNet不同深度网络在Visdrone上的表现

方法	Backbone	AP	AP_s	AP_m	AP_l
ClusDet (2019 ICCV)[15]	ResNet50	26.7	17.6	38.9	51.4
ClusDet (2019 ICCV)[15]	ResNet101	26.7	17.2	39.3	54.9
DMNet (2020 CVPR)[16]	ResNet50	28.2	19.9	39.6	55.8
DMNet (2020 CVPR)[16]	ResNet101	28.5	20.0	39.7	57.1
FCOS (2020 TIPAM)[5]	ResNet50	22.5	14.4	32.1	35.2
FCOS (2020 TIPAM)[5]	ResNet101	22.5	13.8	32.7	37.1
ATSS (2020 CVPR)[10]	ResNet50	24.1	15.5	34.6	36.1
ATSS (2020 CVPR)[10]	ResNet101	24.3	15.4	35.2	38.9

注：ClusDet和DMNet为航拍无人机检测算法.其中AP，AP_s，AP_m和AP_l分别代表所有目标、小目标、中等目标和大目标的平均检测精度.

目标更鲁棒的特征，从而降低小目标的漏检和误检. 但是，这些方法一方面没有更好地考虑对小目标有效的全局特征信息，另一方面无法直接应用于各类检测器，在通用性方面有待加强.

2.2 密集遮挡解决策略

航拍图像中目标的分布通常呈现出非均匀的特点，即有些区域目标稀疏，而有些区域目标密集，目标密集区域中通常还伴随着严重的目标被遮挡情况. 这给通用检测器带来了十分巨大的挑战. 为了提出更有效且更通用的解决方案，本文对现有方法如何在遮挡密集区域产生更高质量的检测框进行了分析.

现有的单阶段目标检测器RetinaNet[22]，FCOS，ATSS等，实现目标检测时需要同时优化目标分类与目标定位两个子任务，而在样本分配过程中却往往只使用定位位置这一种信息. 例如，YOLO[23]和RetinaNet利用锚框(anchor box)和真值框(Ground Truth，GT)的交并比(Intersection over Union，IoU)，将其中IoU大于0.5的候选框作为正样本，小于0.4的作为负样本；FCOS和CenterNet根据中心点之间的距离，将中心点位于GT的候选框视为正样本；ATSS则通过定位位置信息的统计情况进行正负样本划分. 这种只考虑定位位置信息的样本分配策略和网络的优化目标存在着较大的不对齐，进而容易产生大量冗余的检测框，为后续NMS处理带来了更多的困难. 在航拍图像中目标密集且存在遮挡的情况下，不对齐这一问题被进一步放大. 如果对于每一个目标，能生成尽可能少甚至唯一的高质量预测检测框，那么不对齐问题以及后续由密集遮挡形成的冗余检测框导致的误检漏检问题将得到较大的改善.

针对这一情况，如何进行更好的正负样本分配成为问题的关键. 在目标分布不均匀及密集区域大面积遮挡的场景下，研究者们探索了如何使网络更关注密集遮挡区域，并在这一区域获得目标的高质量检测框. 例如，文献[24]和ClusDet[15]分别通过训练一个困难预测区域网络和设计群集网络获得图像中的困难区域；而DMNet[16]更进一步提出不用单独训练网络，而是采用生成密度掩膜图的方法来定位航拍图像中密集遮挡区域，然后对这部分密集遮挡区域和整幅图像分别利用通用检测器进行检测，最后将两次的检测结果融合，生成密集遮挡区域内目标的高质量检测框，从而提升目标的检测精度. 但这些方法都需要对遮挡区域定位后进行二次检测，并不能实现端到端的训练和检测推理.

为了进一步提取对小目标有效的特征，同时更易于适配各类目标检测器，本文提出一种小目标特征聚合网络，获得更鲁棒的小目标特征和更有效的全局信息；提出无需提前定位密集遮挡区域就可直接生成更高检测置信度的任务平衡样本分配策略；同时，为了更好地匹配小目标特征聚合网络和任务平衡样本分配策略，提出增强检测头. 这三部分共同构成了端到端的聚焦小目标的航拍图像目标检测算法(FocSDet)，该算法可在一定程度上有效解决航拍图像目标小、密集遮挡等问题，提升航拍图像目标检测精度.

3 聚焦小目标的航拍图像目标检测算法(FocSDet)

3.1 算法概述

本文在通用目标检测器ATSS的基础上，分别从骨干网络、样本分配策略以及检测头三方面改进，提出了聚焦小目标的航拍图像目标检测算法(FocSDet). 在骨干网络方面，提出小目标特征聚合网络，将双Swin-Transfomer骨干网络利用DHLC模式来连接，并与特征金字塔(FPN)相结合. 通过这一方法，在引导骨干网络的低层特征基础上增加辅助骨干网络的高层特征，在保留大目标语义信息的同时，丰富了对小目标的细节特征描述. 在样本分配策略方面，提出任务平衡样本分配策略，针对现有方法只利用定位位置信息、生成低质量冗余检测框的问题，引入预测分类分数构建可动态更新的样本匹配质量评价分数，用于监督网络优化，生成更高质量的检测候选框，减弱冗余检测框对NMS的干扰，进而提升对密集遮挡的检测性能. 在检测头方面，为适配小目标特征聚合网络和任务平衡样本分配策略，提出增强检测头，为分类和回归分支同时引入层注意力机制，进一步提升FocSDet对小目标的检测能力.

3.2　小目标特征聚合网络

ResNet等骨干网络提取的特征有如下特点:低层特征语义信息比较少,位置信息精确;相比之下,高层特征语义信息丰富,但位置信息不够精确. 即随着网络层数加深,位置信息逐步模糊,进而造成小目标检测性能变弱. 为此,本文引入经典特征金字塔(FPN)进行特征融合,将高层语义信息逐渐传向低层,使得不同特征层的尺度更为平滑,在一定程度上解决目标多尺度变化问题,提升小目标的检测效果. 然而,尽管FPN看起来十分完美,但直接使用CNN构成的如ResNet这样的特征提取网络仍然存在问题:(1)卷积的局部性特点会造成对图像全局语义信息获取不足,而上下文相关信息对目标小、信息少的小尺寸目标更为重要;(2)单一主干网络相较于多主干网络,在特征提取的丰富程度方面有所不及.

针对上述第一个问题,需要增加全局语义信息,使特征提取网络更关注相关目标及上下文,因此,本文采用Swin-Transfomer作为特征提取网络,将CNN与Transformer各自的优势有效地结合,即通过Transformer引入CNN所需要的图像全局语义信息. 其中,Swin-Transformer由多个Swin-Transformer-Block构成,Swin-Transformer-Block内部结构如图2橙色虚线框所示. 首先对特征图进行层归一化(LayerNorm,LN)操作,然后经过基于窗口的多头自注意力模块(Window Multi-head Self-Attention,W-MSA),即将特征图切成小窗口并计算注意力,随后进行窗口合并. 再通过一层LN和全连接层(MultiLayer Perceptron,MLP),随后再次进行LN操作,经过基于移动窗口的多头自注意力模块(Shifted Windows Multi-head Self Attention,SW-MSA). 相较于W-MSA,SW-MSA可引入前一层相邻非重叠窗口之间的连接和特征信息. 最后再通过一层LN和一层MLP. Swin-Transformer-Block通过多头自注意力模块(Multi-head Self-Attention,MSA)对每一层提取的整体特征建立全局的远距离依赖,引入CNN不具备的远距离全局语义信息,同时更关注目标特征,减弱航拍图像复杂的背景噪声干扰.

为了解决上述第二个问题,即单一主干网络提取的特征不够丰富的问题,本文构建了更为优异的小目标特征聚合网络. 采用双Swin-Transfomer骨干网络,并采用DHLC模式来进行连接,即辅助骨干网络的每一层都与引导骨干网络中的更低层相连,连接效果如图2红色虚线框内的小目标特征聚合网络所示,引导主干网络的L2层特征由辅助主干网络的A1~A4特征通过采样和特征相加构成;引导主干网络的L3层特征由辅助主干网络的A3和A4特征和自身L2特征通过采样和特征相加构成;引导主干网络的L4层特征由辅助主干网络的A4特征和自身L3特征通过采样和特征相加构成. 具体实现如式(1)所示:

$$g^l(x)=\sum_{i=l+1}^{L}U\left(w_i(x^i)\right),l\geqslant 1 \tag{1}$$

其中,$g^l(x)$代表复合连接,$x=\{x^i|i=1,2,\cdots,L\}$作为辅助骨干网络的特征输入,L代表辅助骨干网络的总输入层数,w代表1×1卷积和1个BN层,$U(\cdot)$代表上采样操作.

在构成双Swin-Transfomer骨干网络后,结合FPN,将引导骨干网络的L4作为P3,一方面下采样依次生成P2、P1,另一方面上采样与L3特征进行特征融合构成P4;类似地,将P4特征上采样与引导骨干网络的L2特征进行特征融合构成P5. 最终形成了小目标特征聚合网络.

通过采用DHLC模式来构建双Swin-Transfomer骨干网络,即在引导骨干网络的低层特征基础上增加辅助骨干网络的高层特征,将多个高层和低层特征融合在一起,以生成更丰富的特征表示,增强了引导骨干网络的特征提取表达能力. 与此同时,Swin-Transfomer的多头注意力机制,使得其能够获得全局语义信息及更多与目标相关的上下文信息. 最后,再与FPN相结合,共同构成小目标特征提取网络. 该网络可以在不损失大目标语义信息的同时,得到对小目标更好的特征表示.

3.3　任务平衡样本分配策略

本文提出了任务平衡样本分配策略,该策略在ATSS只针对定位信息的样本分配策略基础上,引入预测分类信息来迭代更新样本分配和监督网络优化,通过更优的样本分配来进一步平衡网络的分类定位子任务,进而实现更高质量的预测结果,该策略的示意图如图3所示. 以图3为例,左上角图像中五个红色框为car类型的GroudTruth,从上往下依次为GT1,GT2,GT3,GT4,GT5,蓝色框GT6为van类型的GroudTruth,粉色小圆点代表一轮迭代中得到的所有预测框bbox的中心点. 为了更好地获得正负样本,具体样本分配策略如下:以GT6内一个bbox中心点为例,并将颜色由粉色更改为绿色,其中绿色框为该绿色中心点的预测bbox6,IOU6为bbox6与GT6的IOU,P(cls6)为对应van类型的预测分数,通过式(2)联合两者得到样本匹配质量评价分数Cost6.

$$\mathrm{Cost}=\left(P(\mathrm{cls}_g)\right)^{1-\alpha}\cdot\left(\mathrm{IoU}(g,\mathrm{bbox})\right)^{\alpha} \tag{2}$$

其中,α为计算样本匹配质量评价分数的加权超参数,g为目标的GroudTruth.

同理可得绿色中心点对应GT1~GT5的质量评价分数Cost1~Cost5. 进一步地,图像中所有预测中心点对应G1~G6的.

Cost值共同构成样本匹配质量评价分数矩阵(Cost Matrix). 对Cost Matrix进行筛选,计算所有预测中心点

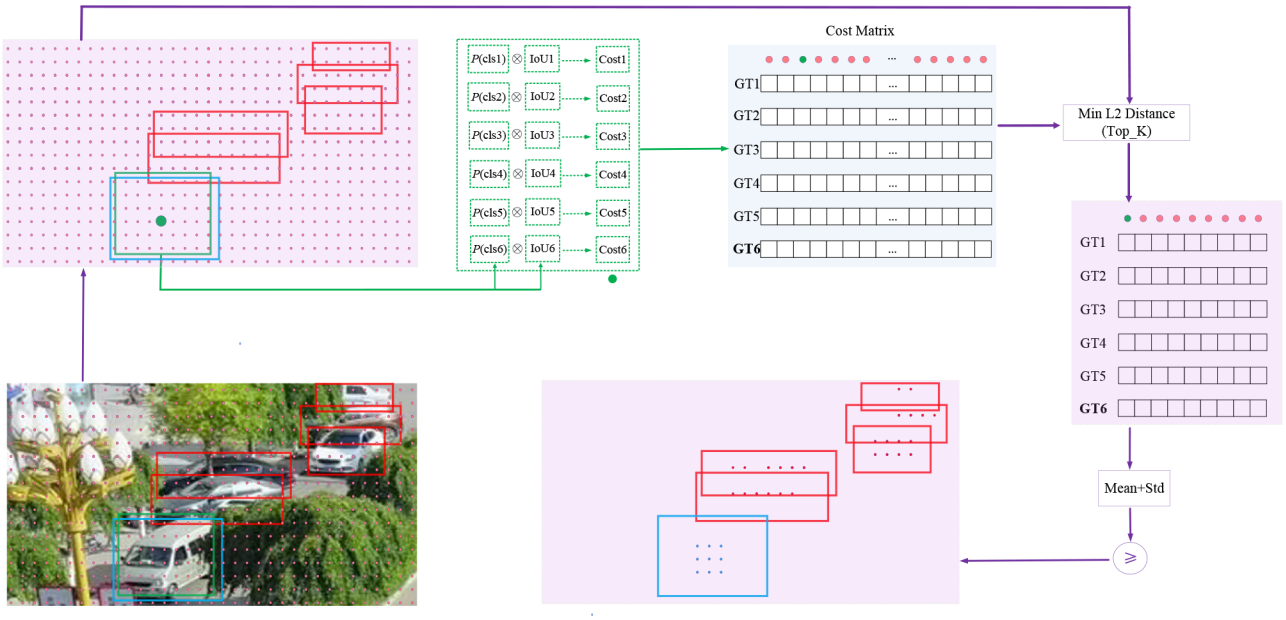


图3 任务平衡样本分配策略示意图

与GT中心点的L2距离. 对每个GT,从Cost Matrix中选取L2距离最小的k个预测中心点对应的Cost值. 计算每个GT对应的这k个预测中心点对应Cost值的均值和方差,并对均值和方差求和得到对应GT的Cost阈值,如式(3)所示:

$$t_g = \mathrm{Mean}(\mathrm{Cos\,t}_g) + \mathrm{Std}(\mathrm{Cos\,t}_g) \tag{3}$$

其中,Mean(·)和Std(·)分别代表求均值和方差,$\mathrm{Cos\,t}_g$代表某个GT对应的L2距离最小的k个预测中心点对应Cost值的集合,t_g代表该GT的Cost阈值.

从筛选后的Cost Matrix中选取所有预测中心点的Cost值大于对应GT的Cost阈值的中心点. 在图3右下角图像中,预测中心点由原先的大量粉色点更新为满足阈值的少量的红色点和蓝色点. 即将最后符合条件的car类型的预测中心点更改为红色,van类型的预测中心点更改为蓝色. 并将满足阈值中心点对应的预测框bbox作为正样本,剩余的作为负样本. 可见,结合定位位置信息和分类分数信息两者,共同得到样本匹配质量评价分数,从而预测得到更准确的正负样本,实现更高质量的预测结果.

具体任务平衡样本分配策略见算法1.

3.4 增强检测头和损失函数

针对航拍图像目标小且易被遮挡的特点,本文分别在骨干网络和样本分配策略方面,提出小目标特征聚合网络和任务平衡样本分配策略,有效提升了航拍图像小目标的检测性能. 为了进一步适配主干网络和样本分配策略,本文提出增强检测头,使得对小目标的检测性能得到进一步提升.

如图2右侧所示,增强检测头在分类和回归两个检测头分支中各增加一个层注意力. 层注意力具体为将FPN特征经过4个卷积,得到4个不同尺度的感受野特征,然后经过特征通道相加操作(Concat,Cat)、全局平均池化(Global Average Pooling,GAP)、Sigmoid激活函数来计算这4个特征的权重,最后与原特征进行点乘的特征拼接. 一方面,通过增强检测头,为分类和回归分支分别提供更重要的分类语义特征信息和位置回归特征信息,克服直接进行分类和回归造成的两者任务干扰. 另一方面,利用增强检测头,对小目标特征聚合网络提取的特征进行再一次区分,为任务平衡样本分配策略得到的正负样本进行特征信息再挖掘,进而得到更高的分类分数和更精准的定位位置.

针对损失函数设计,在分类部分采用可平衡前景和背景不均衡问题的Focal Loss[22],在回归部分,与IoU只关注重叠区域不同,采用不仅关注重叠区域,还关注其他的非重合区域的全局交并比指标(Generalized Intersection over Union,GIoU)[25,26]. 本文设计的损失函数与ATSS和FOCS不同之处在于,将其回归分支中用于抑制低质量检测框的centerness分支替换为Generalized Focal Loss(GFL)方法中使用的IoU分支[27]. 最终损失函数见式(4):

$$\mathrm{Loss} = \underbrace{\lambda_1 L_{\mathrm{cls}}}_{\text{classification}} + \underbrace{\lambda_2 L_{\mathrm{GIoU}} + \lambda_3 L_{\mathrm{IoU}}}_{\text{regression}} \tag{4}$$

其中,λ_1值设置为1.0,λ_2值设置为2.0,λ_3值设置为1.0,L_{cls}代表分类损失,L_{GIoU}代表定位损失,L_{IoU}代表定位质量评价损失.

算法1　任务平衡样本分配策略

输入:G为图像中GT的集合;
　　L为图像特征金字塔层数;
　　k为每一层金字塔选取bbox数量的超参数,默认为9;
　　α为计算样本匹配质量评价分数中分类和回归部分的加权超参数.

输出:Ps为正样本集合;
　　Ns为负样本集合.

FOR 每一次迭代 t:
　更新输入:B_t为所有bbox的集合
　　P_t为B_t中所有bbox的预测分类分数集合
　FOR 每一个$g \in G$:
　　cls_g为g的类别标签
　　分别构建对于g的候选正样本bbox,预测分类分数和样本评价分数阈值空集合

$$B_g \leftarrow \varnothing, P_g \leftarrow \varnothing, \mathrm{Cos\,t}_g \leftarrow \varnothing$$

　　计算样本匹配质量评价分数

$$\mathrm{Cos\,t}_t = \underbrace{\left(P_t(\mathrm{cls}_g)\right)^{1-\alpha}}_{\text{classification}} \cdot \underbrace{\left(\mathrm{IoU}(g, B_t)\right)^{\alpha}}_{\text{regression}}$$

　　FOR 特征金字塔的l层:$l \in [1, L]$
　　Tmep$B_l \leftarrow$ 从$B_{l,t}$中选择bbox的中心点与g的中心点的L2距离最小的k个
　　Tmep$P_l \leftarrow$ 从$P_{l,t}$中选择与TmepB_l中所有bbox相对应的预测分类分数
　　Tmep Cos $\mathrm{t}_l \leftarrow$从Cos t_t中选择TmepB_l中所有bbox相对应的样本匹配质量评价分数
　　$B_g = B_g \cup \mathrm{Tmep}B_l, P_g = P_g \cup \mathrm{Tmep}P_l$,
　　Cos t_g = Cos $\mathrm{t}_g \cup$ Tmep Cos t_l
　　END FOR
　　计算Cos t_g的均值:$m_g = \mathrm{Mean}(\mathrm{Cos\,t}_g)$
　　计算Cos t_g的方差:$v_g = \mathrm{Std}(\mathrm{Cos\,t}_g)$
　　计算g的Cos t_g阈值:$t_g = m_g + v_g$
　　FOR 每一个bbox $b \in B_t$:
　　cos $\mathrm{t}_b \leftarrow$从Cos t_g中选择与b相对应的样本匹配质量评价分数
　　IF cos $\mathrm{t}_b \geqslant t_g$ and b的中心点在g内:
　　Ps = Ps $\cup b$
　　END IF
　　END FOR
　END FOR
　Ns = B_t − Ps
　Return Ps,Ns
END FOR

4　实验

4.1　实验设置

本文实验基于MMDetection工具箱完成. 使用在ImageNet[28]数据集上的预训练模型进行训练,实验环境为Titan RTX. 输入图像在单尺度训练时保持短边800像素、长边1 333像素;在多尺度训练时长边依然保持1 333像素,短边范围为480~800像素,批量大小均为4. 除与Swin-Transformer的对比实验外,本文方法使用随机梯度下降算法(Stochastic Gradient Descent, SGD)作为优化器,动量为0.9,权重衰减为0.000 1,学习率为0.05,训练24个循环. 在与Swin Transformer相关的实验中,本文的配置与Swin-Transformer的目标检测实验配置保持一致,即优化器采用AdamW,betas为(0.9, 0.999),权重衰减为0.05,学习率为0.000 1,训练36个循环.

4.2　VisDrone数据集上的实验结果

为了展示所提出方法的有效性,在VisDrone-DET[7]数据集上进行大量实验,并与其他目标检测方法进行对比实验,主要分为以下四个部分:小目标特征聚合网络对比实验、任务平衡样本分配策略对比实验、增强检测头对比实验、与其他优秀目标检测方法的对比实验.

4.2.1　小目标特征聚合网络对比实验

为了验证本文所提出的小目标特征聚合网络的性能,本文选取了Resnet101,Swin_T和小目标特征聚合网络三种骨干网络,在FCOS,ATSS,GFLV2[29],VFNet和本文所提出的FocSDet五种目标检测方法上进行对比实验,表3展示了各种骨干网络在Visdrone上的表现对比结果,其中加粗值代表该值为当列精度最高值.

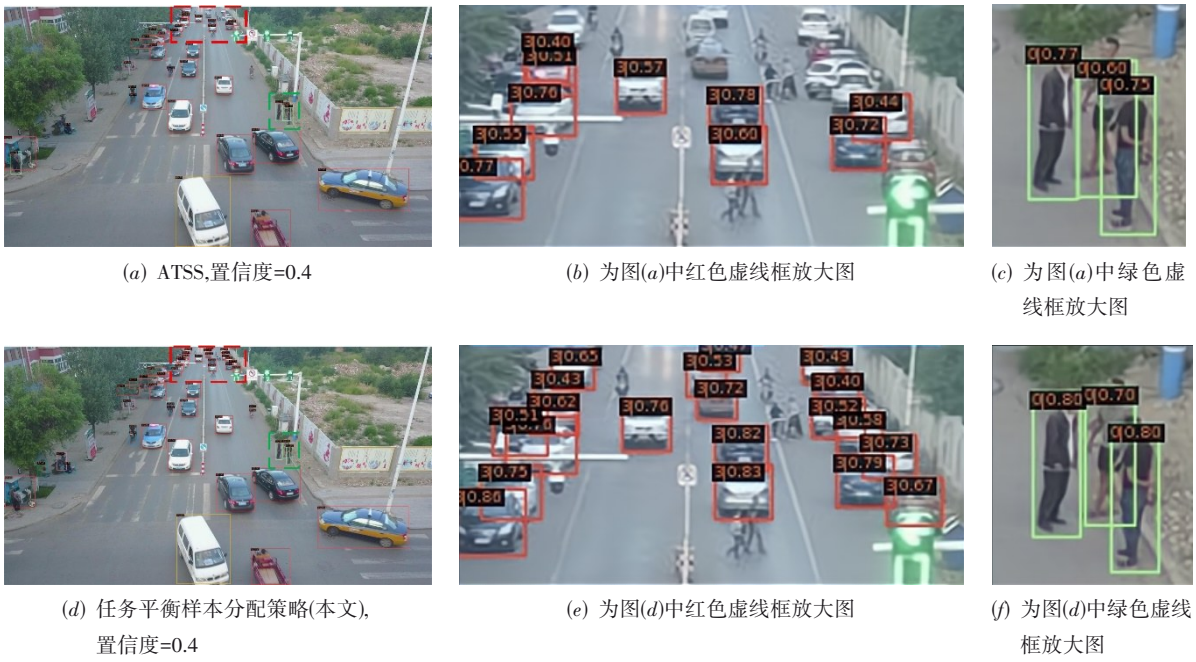
从表3中可以看出,相较于ResNet101+FPN,五种目标检测方法在采用小目标特征聚合网络后,在小目标AP_s上均获得4%左右的大幅提升,整体AP提升5%左右. 与2021年提出的Swin_T[11]骨干网络进行比较,FCOS,ATSS,GFLV2,VFNet和FocSDet在采用小目标特征聚合网络后,小目标AP_s分别提升1.6%,1.3%,1.5%,0.8%和1.3%,整体AP分别提升1.6%,1.4%,1.2%,1.2%和1.4%. 通过实验结果可以明显看出,本文所提出的FocSDet,其小目标特征聚合网络可以在不损失大目标语义信息的同时,得到对小目标更好的特征表示. 同时也通过将该网络应用于不同的检测器,表明了该网络的通用性和鲁棒性,对数据集中各种尺寸的目标均有明显的检测效果提升.

4.2.2　任务平衡样本分配策略对比实验

为了展示任务平衡样本分配策略可生成更高质量检测框的效果,将部分对比结果在图4中展示. 其中图4(a)和图4(d)分别为采用ATSS算法和本文提出的任务平衡样本分配策略在置信度为0.4时的结果. 从图4的(b)和(e)可以看出,图(b)中只正确检测出10个车辆目标,而图(e)中则正确检测出20个车辆目标,此外,本文方法对同一辆车的检测置信度有明显提升. 在绿色虚线框对比图中也可以看出,图(c)中从左到右的

表3 各种骨干网络在Visdrone上的表现对比 单位:%

网络类型	方法									
	FCOS (2020 TIPAM)[5]		ATSS (2020 CVPR)[10]		GFLV2 (2021 CVPR)[29]		VFNet (2021 CVPR)[12]		FocSDet (本文方法)	
	AP_s	AP	AP_s	AP	AP_s	AP	AP_s	AP	AP_s	AP
ResNet101+FPN	13.8	22.6	15.4	24.3	16.9	25.2	16.5	25.6	16.7	25.3
Swin_T+FPN	16.2	25.6	18.2	27.4	19.3	28.1	20.1	29.0	20.8	29.4
小目标特征聚合网络	**17.8**	**27.2**	**19.5**	**28.8**	**20.8**	**29.3**	**20.9**	**30.2**	**22.1**	**30.8**



(*a*) ATSS,置信度=0.4 (*b*) 为图(*a*)中红色虚线框放大图 (*c*) 为图(*a*)中绿色虚线框放大图

(*d*) 任务平衡样本分配策略(本文),置信度=0.4 (*e*) 为图(*d*)中红色虚线框放大图 (*f*) 为图(*d*)中绿色虚线框放大图

图4 任务平衡样本分配策略效果对比图

人的检测置信度为0.77,0.60,0.75,而采用任务平衡样本分配策略的图(*f*)中对应的检测置信度为0.80,0.70和0.80,均有明显提升. 图4直观展示了任务平衡样本分配策略对密集遮挡问题的解决情况. 在复杂场景中,任务平衡样本分配策略提升了目标检测的准确性,同一检测目标的检测置信度也得到了提升.

图5展示了3组不同场景下图像分别采用现有ATSS算法、任务平衡样本分配策略进行检测的结果,验证了任务平衡样本分配策略的优异性能. 从第一组夜间遮挡、第二组车辆连续遮挡和第三组小目标行人遮挡的典型困难场景中都可以看出,ATSS存在部分目标漏检的情况,而任务平衡样本分配策略得到的检测框质量更高,漏检也更少,证明了任务平衡样本分配策略可生成更高质量的预测结果.

为了进一步验证任务平衡样本分配策略的有效性,本文对Visdrone数据集中的遮挡目标进行了平均召回率(Average Recall,AR)的计算,具体结果如表4所示,其中加粗值代表该值为当列精度最高值. 从表4中可以看出,在均使用小目标特征聚合网络的基础上,采用任务平衡样本策略,相比ATSS方法,针对遮挡目标中小目标的AR_s提升1.5%,整体AR提升1.3%. 实验数据证明了本文所提出的任务平衡样本分配策略可生成更高质量检测框,提升了密集遮挡目标的检测能力.

4.2.3 增强检测头对比实验

表5展示了增强检测头的作用,其中加粗值代表该值为当列精度最高值. 可以看出,在采用了小目标特征聚合网络和任务平衡样本分配策略后,算法AP为30.5%、AP_s为21.9%的优秀表现基础上,本文提出的增强检测头,将AP和AP_s又再次分别提升了0.3%和0.2%,证明了增强检测头对小目标特征聚合网络和任务平衡样本分配策略的适配性,以及有效性.

4.2.4 与其他方法的对比实验

为了更客观地展示所提出的FocSDet的有效性,在VisDrone-DET[7]数据集上与近两年各类优异的目标检测方法进行比较,具体表现对比如表6所示. 其中加粗值代表该值为当列精度最高值. 本文所提出的FocSDet,即使不使用多尺度训练、模型集成等技巧,仍表现

图5　不同场景下任务平衡样本分配策略效果展示图,第一行为ATSS算法,第二行为局部放大图,第三行为本文任务平衡样本分配策略;置信度的取值均为0.4

表4　任务平衡样本分配策略　　单位:%

方法	骨干网络	AR	AR_s	AR_m	AR_l
FCOS	ResNet50	30.6	24.9	39.5	40.0
	ResNet101	30.4	24.2	40.0	44.4
	小目标特征聚合网络	37.5	30.6	49.2	60.1
ATSS	ResNet50	33.6	27.5	43.5	49.2
	ResNet101	34.2	27.5	44.7	48.8
	小目标特征聚合网络	39.8	33.2	51.0	61.5
FocSDet(本文方法)	小目标特征聚合网络+任务平衡样本分配策略	**41.1**	**34.7**	**51.8**	**62.0**

表5　增强检测头在Visdrone上的验证　　单位:%

模型	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
任务平衡样分配策略	30.5	51.9	30.8	21.9	41.5	49.7
任务平衡样本分配略+增强检测头(FocSDet)	**30.8**	**52.5**	**31.0**	**22.1**	**42.0**	**50.3**

最好. 相较于2021 CVPR oral的VFNet方法,即使在VFNet使用本文提出的小目标特征聚合网络的情况下,FocSDet的AP_s仍高出1.2%,整体AP值高出0.6%. 证明了本文所提出的方法对航拍小目标检测的优越性能.

表6　与各种目标检测算法在Visdrone对比　　单位:%

模型	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS (2020 TIPAM)[5]	27.2	46.4	27.1	17.8	38.5	43.6
ATSS (2020 CVPR)[10]	28.8	48.7	29.3	19.5	40.2	49.2
GFLV2 (2021CVPR)[29]	29.3	50.0	29.3	20.8	40.7	45.9
VFNet (2021CVPR)[12]	30.2	50.6	30.5	20.9	**42.0**	47.3
本文方法(FocSDet)	**30.8**	**52.5**	**31.0**	**22.1**	**42.0**	**50.3**

4.3　CARPK数据集上的实验结果

为了进一步验证本文所提出的FocSDet的有效性,本文在CARPK数据集上进行了比较. CARPK包含来自4个不同停车场的近90 000辆汽车的图像,这些图像是在大约40 m高的无人机视图中收集的. 图像集由每辆车的边界框注释,所有标记的边界框都记录了左上角点和右下角点.

为了验证所提出方法的鲁棒性和泛化性能,所有算法只在Visdrone数据上训练,不再在CARPK数据集上训练,直接对1 488张标注图像进行推理检测. 实验结果如表7所示. 其中加粗值代表该值为当列精度最高值.

从表7的实验结果看出,本文所提出的FocSDet,在整体AP上精度高于GFLV2方法1.5%,小目标AP_s更是高VFNet方法5%. 这两个数据集的验证,充分证明了本文所提出的检测方法对航拍图像小目标具有较强的泛化性和鲁棒性.

表7 与各类目标检测算法在CARPK上的表现对比 单位:%

模型	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FCOS (2020 TIPAM)[5]	69.8	95.5	89.7	2.50	71.3	79.2
ATSS (2020 CVPR)[10]	70.7	95.9	90.3	6.70	72.1	78.3
GFLV2 (2021 CVPR)[29]	71.4	94.9	90.8	3.50	72.7	80.9
VFNet (2021CVPR)[12]	71.2	94.9	90.8	6.90	72.4	**81.1**
FocSDet(本文方法)	**72.9**	**97.8**	**92.7**	**11.9**	**73.9**	80.9

5 总结与展望

针对航拍图像中存在的目标尺寸小、难以与背景区分,以及大量区域中目标密集且存在严重遮挡的情况,本文提出小目标特征聚合网络、任务平衡样本分配策略以及配合两者的增强检测头,共同构成了一个聚焦小目标的航拍图像目标检测算法(FocSDet). 相较于ATSS和VFNET等现有优秀方法,FocSDet在Vsidrone上AP分别提升2%和0.6%,小目标AP_s分别提升2.6%和1.2%;在CARPK上AP分别提升2.2%和1.7%,小目标AP_s分别提升5.2%和5.0%. 大量实验结果表明,FocSDet在保持大目标检测精度的同时能有效提升航拍图像中小目标的检测性能,并利用任务平衡样本分配策略进一步提升了密集遮挡目标的检测精度,在面对密集遮挡且存在大量小目标的情况下,FocSDet优于现有检测算法. 同时,本文将提出的小目标特征聚合网络、任务平衡样本分配策略用于多个检测方法,以及在两个大型航拍数据集上进行了验证,表明了FocSDet通用性更强,泛化性能更优.

此外,虽然本文提出的FocSDet有效地提升了航拍图像中小目标的检测精度,但是在不经过训练而直接在CARPK数据集上进行测试时,发现小目标的精度仍然相对较低. 如何进一步提升航拍图像中小目标的检测精度,以及提高模型对小目标的泛化性能,是后续研究的方向.

参考文献

[1] 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 and Machine Intelligence, 2017, 39(6): 1137-1149.

[2] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2022-03]. https://arxiv.org/abs/2004.10934.

[3] ZHU C C, HE Y H, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 840-849.

[4] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points[EB/OL]. (2019-04-16)[2022-03]. https://arxiv.org/abs/1904.07850.

[5] TIAN Z, SHEN C H, CHEN H, et al. FCOS: A simple and strong anchor-free object detector[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933.

[6] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision - ECCV 2014. Zurich: Springer, 2014: 740-755.

[7] CAO Y R, HE Z J, WANG L J, et al. VisDrone-DET2021: The vision meets drone object detection challenge results [C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal: IEEE, 2021: 2847-2854.

[8] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 2261-2269.

[9] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 936-944.

[10] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 9756-9765.

[11] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2022: 9992-10002.

[12] ZHANG H Y, WANG Y, DAYOUB F, et al. VarifocalNet: An IoU-aware dense object detector[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 8510-8519.

[13] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 4165-4173.

[14] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770-778.

[15] YANG F, FAN H, CHU P, et al. Clustered object detection in aerial images[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2020: 8310-8319.

[16] LI C L, YANG T, ZHU S J, et al. Density map guided object detection in aerial images[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle: IEEE, 2020: 737-746.

[17] ÜNEL F Ö, ÖZKALAYCI B O, ÇIĞLA C. The power of tiling for small object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE, 2020: 582-591.

[18] GONG Y Q, YU X H, DING Y, et al. Effective fusion factor in FPN for tiny object detection[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2021: 1159-1167.

[19] BAI Y C, ZHANG Y Q, DING M L, et al. SOD-MTGAN: Small object detection via multi-task generative adversarial network[C]//Computer Vision - ECCV 2018. Munich: Springer, 2018: 210-226.

[20] LIANG X, ZHANG J, ZHUO L, et al. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1758-1770.

[21] LONG H, CHUNG Y, LIU Z B, et al. Object detection in aerial images using feature fusion deep networks[J]. IEEE Access, 2019, 7: 30980-30990.

[22] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 2999-3007.

[23] 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 (CVPR). Las Vegas: IEEE, 2016: 779-788.

[24] ZHANG J Y, HUANG J Y, CHEN X K, et al. How to fully exploit the abilities of aerial image detectors[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul: IEEE, 2020: 1-8.

[25] 侯志强, 刘晓义, 余旺盛, 等. 使用GIoU改进非极大值抑制的目标检测算法[J]. 电子学报, 2021, 49(4): 696-705.
HOU Z Q, LIU X Y, YU W S, et al. Object detection algorithm for improving non-maximum suppression using GIoU[J]. Acta Electronica Sinica, 2021, 49(4): 696-705. (in Chinese)

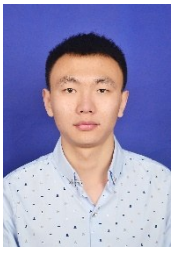
[26] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2020: 658-666.

[27] LI X, WANG W H, WU L J, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 21002-21012.

[28] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

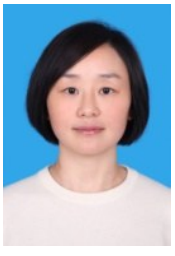
[29] LI X, WANG W H, HU X L, et al. Generalized focal loss V2: Learning reliable localization quality estimation for dense object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 11627-11636.

作者简介



张　智　男,1993年12月出生于山西省大同市. 现为中国民航大学计算机科学与技术学院实验师. 主要研究方向为视频图像目标检测.
E-mail: zhangz@cauc.edu.cn

易华挥　男,1997年11月生于四川省南充市. 现为中国民航大学本科学生. 主要研究方向为目标检测.
E-mail: huahui_yi@163.com



郑　锦(通讯作者)　女,1978年10月出生于四川省乐山市. 现为北京航空航天大学计算机学院副教授、博士生导师. 主要研究方向为计算机视觉、视频图像处理等.
E-mail: JinZheng@buaa.edu.cn