

文章编号 1004-924X(2023)18-2723-13

联合自注意力和分支采样的无人机图像目标检测

张云佐^{1,2*}, 武存宇¹, 刘亚猛¹, 张 天¹, 郑宇鑫¹

(1. 石家庄铁道大学 信息科学与技术学院, 河北 石家庄 050043;

2. 河北省电磁环境效应与信息处理重点实验室, 河北 石家庄 050043)

摘要: 无人机图像目标检测在诸多领域被广泛应用, 但受制于图像背景复杂、目标密集, 目标尺度变化剧烈, 现有的无人机图像目标检测算法检测效果不够精准。为解决此类问题, 提出了一种联合自注意力和分支采样的无人机图像目标检测方法。首先, 设计了自注意力和卷积相融合的嵌套残差结构以实现全局信息和局部信息的有效结合, 让模型聚焦于待测目标, 从而淡化复杂背景的影响。其次, 设计了一种基于分支采样的特征融合模块以弥补目标信息丢失。最后, 引入浅层细粒度特征图, 新增了针对微小目标的改进检测头以缓解尺度剧烈变化, 并基于此提出一种特征增强模块, 用于捕获更多具有鉴别性的小目标特征。经实验验证, 本文所提算法在多种场景中性能良好。其中 s 模型在 VisDrone2019 数据集上的 mAP₅₀ 和 mAP 分别达到 59.3% 和 37.1%, 相较于基线模型增长了 5.6% 和 5.4%, 在 UAVDT 数据集上的 mAP₅₀ 和 mAP 分别达到 44.1% 和 24.9%, 相较于基线模型提高了 5.8% 和 3.2%。

关键词: 无人机图像; 自注意力; 分支采样; 多尺度; 特征融合

中图分类号: V249; TP391 **文献标识码:** A **doi:** 10.37188/OPE.20233118.2723

Joint self-attention and branch sampling for object detection on drone imagery

ZHANG Yunzuo^{1,2*}, WU Cunyu¹, LIU Yameng¹, ZHANG Tian¹, ZHENG Yuxin¹

(1. School of Information Science and Technology, Shijiazhuang Tiedao University,
Shijiazhuang 050043, China;

2. Hebei Key Laboratory of Electromagnetic Environmental Effects and Information Processing,
Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

* Corresponding author, E-mail: zhangyunzuo888@sina.com

Abstract: Object detection on drone imagery is widely used in many fields. However, due to the complexity of the image background, the dense small objects and the dramatic scale changes, the existing object detection on drone imagery methods are not accurate enough. In order to solve this problem, we propose an accurate object detection method for drone imagery joint self attention and branch sampling. Firstly, a nested residual structure integrating self attention and convolution is designed to achieve the effective combination of global and local information, which makes the model to focus on the object area and ignore in-

收稿日期: 2022-12-29; 修订日期: 2023-02-09.

基金项目: 国家自然科学基金资助项目 (No. 61702347, No. 62027801); 河北省自然科学基金资助项目 (No. F2022210007, No. F2017210161); 河北省高等学校科学技术研究项目资助 (No. ZD2022100, No. QN2017132); 中央引导地方科技发展资金项目资助 (No. 226Z0501G)

valid features. Secondly, we design a feature fusion module based on branch sampling to mitigate the loss of object information. Finally, an improved detector for small objects is added to alleviate the problem of sharp scale changes. Furthermore, we propose a feature enhancement module to obtain more discriminative small object features. The experimental results show that the proposed algorithm performs well in various scenarios. Specifically, the mAP_{50} and mAP of the s model on the VisDrone2019 reached 59.3% and 37.1% respectively, an increase of 5.6% and 5.4% compared with the baseline. The mAP_{50} and mAP on the UAVDT reached 44.1% and 24.9% respectively, an increase of 5.8% and 3.2% compared with the baseline.

Key words: UAV image; self attention; branch sampling; multi-scale; feature fusion

1 引 言

作为信息化时代的新型技术产物,无人机在遥感测绘、城市管理、灾害预警等领域都展现出了巨大的价值和应用前景^[1]。但无人机高空巡航、飞行高度不定等作业方式,使其所捕获的图像通常存在背景复杂、包含大量密集微小目标、目标尺度变化剧烈等特点^[2-3],因此无人机图像目标检测仍然是一项具有挑战性的任务。

传统目标检测方法通常可分为三步:首先通过区域选择器以遍历的方式选出候选区域;然后利用 HOG (Histograms of Oriented Gradients)^[4], Haar^[5] 等特征提取器进行特征提取;最后使用 AdaBoost^[6]、支持向量机^[7] 等分类器对提取到的特征进行分类。但该类方法通过穷举候选框来得到感兴趣区域,不仅时间复杂度高,而且会产生大量窗口冗余^[8]。此外手工设计的特征提取器泛化能力不足以应对航拍图像中的复杂场景和多类检测任务^[9]。

得益于硬件和算力的发展,基于深度学习的无人机图像目标检测算法逐渐代替传统方法成为了主流。与传统方法相比,基于深度学习的方法因其出色的特征表达和学习能力促进了无人机航拍图像目标检测的发展。Yang 等^[10]提出了一种集群检测网络 ClusDet,将聚类 and 检测过程统一到了端到端框架中,同时通过隐式地建模先验上下文信息提高尺度估计的准确性。Yu 等^[11]对于无人机数据集中类别分布不均衡的问题进行了研究,并采用双路径方式分别处理头部类和尾部类,这种处理方式有效提高了尾部类的检测效果。Liu 等^[12]设计了一种针对高分辨率图像的检测模型 HRDNet。该方法利用深层主干网络和浅层主

干网络分别对低分辨率特征图和高分辨率特征图进行处理,解决了检测高分辨率特征图时计算开销过大的问题。Wu 等^[13]从提高无人机目标检测鲁棒性的角度展开研究,通过对抗学习方式区分有效的目标特征和干扰因素,提高了单类目标检测的鲁棒性。Youssef 等^[14]将多层级联 RCNN 与特征金字塔进行融合,在个别类别上提升了精度,但整体效果下降。Li 等^[15]提出了一种感知生成对抗网络模型,用于实现小目标的超分辨率表示,使小目标具有与大目标相似的表达,从而缩减尺度差异。Tang 等^[16]设计了一种无锚框的检测器,并将原始高分辨率图像分割为多个子图像进行检测,这使得算法在精度上得到提高,但这也带来了更多的计算负荷。Mekhalfi 等^[17]通过胶囊网络对目标之间的关系进行建模,提高了网络对于拥挤、遮挡情况下目标的解析能力。Chen 等^[18]提出了场景上下文特征金字塔,强化了目标与场景之间的关系,抑制了尺度变化带来的影响,此外在 ResNeXt 结构的基础上,引入了膨胀卷积增大感受野。这些方法从不同角度入手对密集微小目标检测任务进行优化,但此类方法没有考虑到复杂背景对航拍图像目标检测任务的影响,以及微小目标信息随网络层数增加而丢失的问题。

为了实现对无人机视角中微小目标的精准检测,本文提出了一种联合自注意力和分支采样的无人机图像目标检测方法。首先,通过嵌套残差结构将自注意力机制集成到主干网络中,实现对于全局信息和局部信息的有效结合,提升网络对全局信息和上下文信息的捕获能力。其次,针对现存特征融合模块的缺陷,设计了一种基于分支采样的特征融合模块,通过分支并行策略分别实现上、下采样操作,而后对不同分支的采样结

果进行融合,有效强化模型的空间位置信息和特征表达能力,并缓解了小目标信息丢失问题。此外,将包含细粒度特征的浅层特征图引入到特征融合序列中,改善小目标检测效果。最后,构建了一种基于反残差结构的特征增强模块以提取具有鉴别性的小目标特征。

2 算法设计

目前主流目标检测算法可以分为两大类:以 R-CNN 系列^[19-20]为代表的双阶段算法和以 Yolo 系列^[21-22]为代表的单阶段算法,其中双阶段算法的精度更高,但检测速度较慢,而单阶段算法凭借简洁快速的优点得到了更广泛的应用。

本文选择 YOLOv5 算法作为基线模型,改进后的网络结构图如图 1 所示,其中,主干部分由 CSPDarkNet53 和本文提出卷积和自注意力相结合的嵌套残差模块(Nested Residual combining

Convolution and Self-attention, NRCS) 构成, CSPDarkNet53 用于提取局部特征, NRCS 模块获取特征之间的全相关性,实现局部特征和全局特征的结合。Neck 部分通过本文所提出的双分支采样的特征融合模块(Double-Branch Sampling Feature Fusion model, DBS-FU)实现特征融合。该结构通过多分支并行策略生成更精细的特征以提高特征融合效果。Head 中引入浅层特征,构建新的检测头(Head 1)。每个检测头通过编码目标信息生成具有 $S^2 \times B \times (4+1+C)$ 维度的张量。 S^2 为特征图中包含的网格数; B 为每个网格上预设的预测框数量,在本文中被设置为 3; $S^2 \times B \times (4+1+C)$ 中的数字 4 表示预测框的坐标信息(x, y, h, w), 数字 1 表示置信度; C 表示待检测目标的类别数量。同时使用本文提出的反残差结构的特征增强模块(Feature Enhancement module of In-Residual, FE-IR)对新增的检测头进行特征增强。

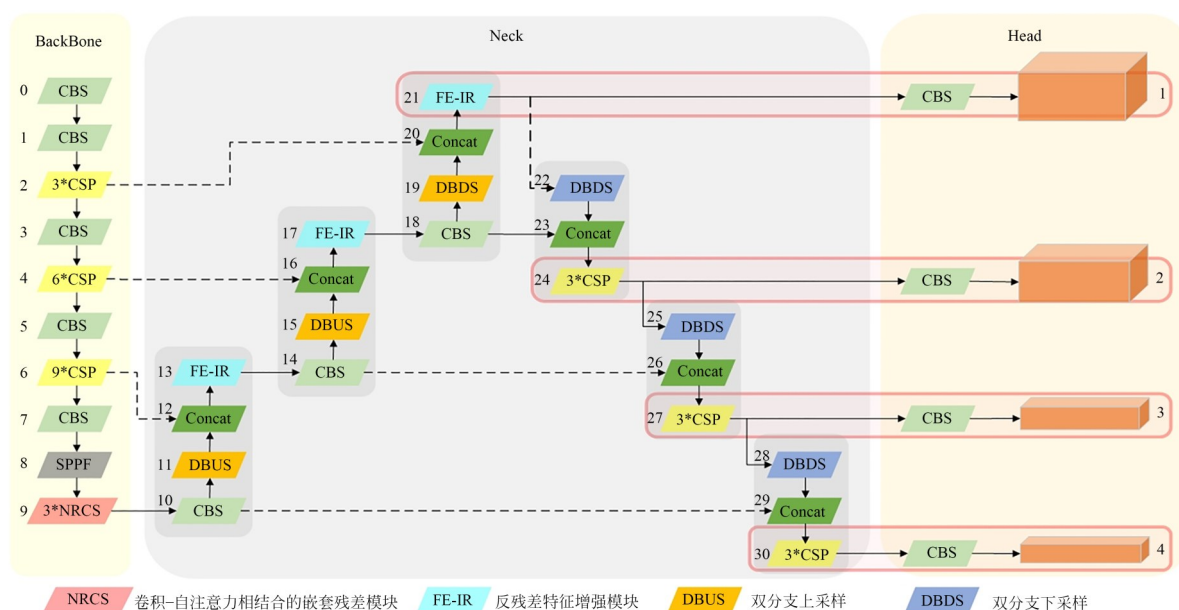


图 1 本文所提算法整体结构图

Fig. 1 Overall structure of the algorithm proposed in this paper

2.1 集成自注意力的特征提取网络

无人机所捕获的高分辨率航拍图像背景复杂,在通常情况下背景信息远超目标信息。如果不予以处理,主干网络在特征提取过程中就会提取到大量的无效背景信息。因此,需要让模型聚焦在有效的目标信息上。

自注意力机制最初被广泛应用在自然语言处理领域中。Dosovitskiy 等^[23]首次将自注意力机制引入到了计算机视觉领域,并取得了一定成效。随着对自注意力机制研究的不断深入,自注意力机制的作用已经在计算机视觉领域得到了广泛认可。

在基于卷积神经网络的主干网络中,通过共享的卷积核进行特征提取,减少冗余参数的同时也带来了两种归纳偏好,即平移不变性和局部相关性,这使得卷积神经网络在提取底层特征和视觉结构方面具有很大优势,但这也导致卷积神经网络受限于局部感受野,无法捕获全局依赖关系。而自注意力机制则通过全局相似度计算像素之间的联系,为不同像素点赋予不同权重,从而获得全局相关性。自注意力相似度计算公式如下:

$$Q = XW^q, K = XW^K, V = XW^V, \quad (1)$$

其中: X 表示输入序列, W^q, W^K, W^V 分别表示三个随机初始化矩阵,并且它们可以通过注意力进行动态调整, X 与它们相乘得到 Q, K, V 。

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

其中: Q 点乘 K^T ,计算每个输入序列和其他序列之间的相似程度。除以 $\sqrt{d_k}$ 用于防止内积过大导致梯度消失。其中 d_k 为隐藏层维度。通过 $Softmax$ 进行归一化计算得到权重系数后再与 V 相乘即可得到通过自注意力机制调整后的输入序列。

卷积神经网络擅长于对局部信息进行提取,而自注意力机制更擅长于对全局信息进行建模。综上,基于卷积神经网络和自注意力机制互补的思想,设计了卷积和自注意力相结合的嵌套残差模块 NRCS,其结构如图 2 所示。该结构是一种嵌套残差结构,外层残差边对局部信息进行恒等映射,内层残差结构中使用多头自注意力模块(Multi Head Self Attention module, MHSA)计算全局相关性。与卷积结构相比,所提出的 NRCS 模块通过结合局部信息和全局信息提高模型对上下文信息的捕获能力。

此外,无人机图像分辨率远超普通图像,Sun 等^[24]研究发现,对分辨率较高的特征图进行全局相关性计算会带来过多的计算量和显存开销,并导致训练过程难以收敛。实验中发现目标检测任务中浅层的高分辨率特征图更依赖卷积的局部特征提取能力。而对经过多次下采样的深层低分辨率特征图进行全局相关性建模不仅可以保证主干网络对局部信息和全局信息进行有效提取,而且所需的内存开销和计算量也更符合应

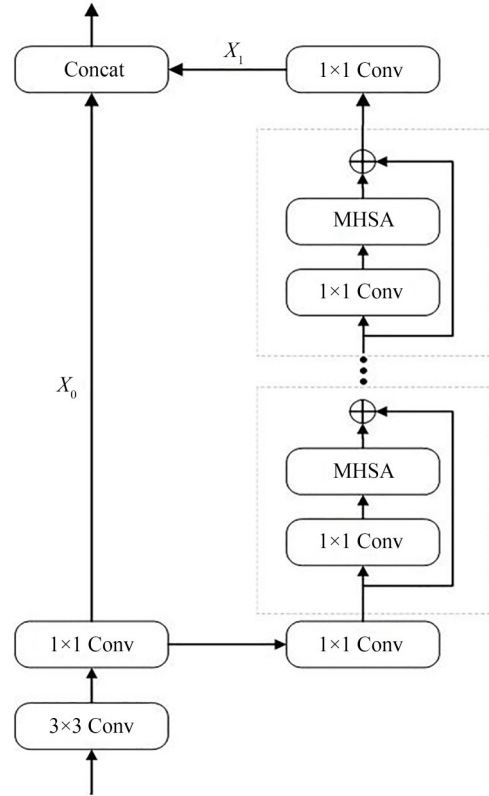


图 2 NRCS 结构图

Fig. 2 Structure diagram of NRCS

用需求。因此,本文方法仅在主干网络末端使用 NRCS 结构,其余部分仍然使用卷积神经网络。这样做既保留了卷积神经网络强大的局部特征提取能力,也通过自注意力获取了全局信息。同时在深层低分辨率特征图使用 NRCS 模块可以有效减少模型的参数量。

2.2 分支采样特征融合

在特征融合阶段,原始网络结构中通过构建双向特征金字塔实现信息交互,其中第一个特征金字塔通过线性插值进行上采样,将浅层特征图中的空间位置信息传递到深层特征图。第二个特征金字塔则通过卷积进行下采样,将融合后的信息逐步传递回浅层特征图,保证每层特征层都有足够的信息进行预测,但这种设计有两个明显的缺陷。首先,在深层特征图向浅层特征图传递信息的时候使用的采样方法单一,导致深层语义信息并未有效传递。其次,在浅层特征图向深层特征图适配的过程中会出现严重的信息衰减的问题。而小目标携带的信息量本就远小于常规目标,因此信息衰减问题对小目标检测来说更加

致命。针对这些缺陷,设计了一种基于双分支采样的特征融合模块 DBS-FU。具体来讲,DBS-FU 包含 DBUS 和 DBDS 两个主要模块。其结构如图 3 所示。

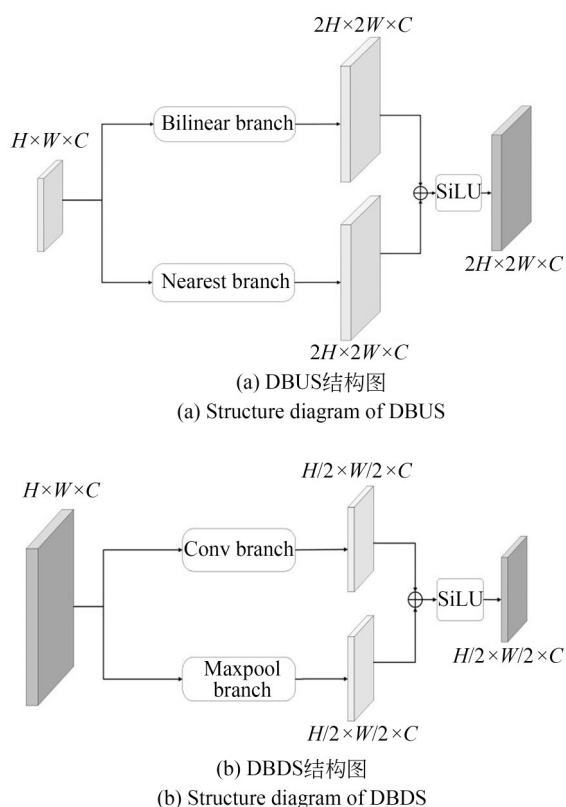


图 3 DBS-FU 结构图

Fig. 3 Structure diagram of DBS-FU

如图 3(a) 所示, DBUS 模块通过构建 Bilinear 和 Nearest 两条并行分支分别实现上采样, 其中 Bilinear 分支和 Nearest 中的采样倍数 (Scale_factor) 均设为 2。得到上采样结果后使用归一化 (Batch Normalization, BN) 分别对不同的上采样结果进行处理, 防止梯度消失。最后将各分支采样结果进行逐元素加和, 并使用 SiLU 激活函数引入更多非线性因素, 得到最终的采样结果。相较于单一的采样方式, 该结构通过生成高质量的上采样结果强化浅层特征图中的语义信息, 从而提高模型对混淆目标的分类能力。该过程定义如下:

$$\text{Branch_Bi} = \text{BN}(\text{Bilinear}(x)), \quad (3)$$

$$\text{Branch_Ne} = \text{BN}(\text{Nearest}(x)), \quad (4)$$

$$\text{DBUS} = \text{SiLU}(\text{Branch_Bi} \oplus \text{Branch_Ne}), \quad (5)$$

其中: Branch_Bi 和 Branch_Ne 对应不同分支的上采样结果, BN 表示批处理归一化层, $\oplus$ 表示逐元素加和。

如图 3(b) 所示, DBDS 通过构建 Conv 和 Maxpool 两条并行分支分别实现下采样, 其中, Conv 分支中卷积核大小为 3, 填充量为 1, 步长为 2; Maxpool 分支中池化核大小为 2, 填充量为 0, 步长为 2。Conv 分支提取局部感受野内的整体特征, Maxpool 分支则专注于池化核内最突出的信息。这两种特性对小目标来说同样重要, 任意一种单一采样方式都会导致某种信息的丢失。而分支并行策略可以有效缓解这一问题。每条分支都代表了原始特征图中一部分特征。将下采样结果融合后, 可以对深层特征图中的空间位置信息进行进一步强化, 提高对小目标的定位能力, 并保留更多的上下文信息。其公式如下:

$$\text{Branch_Conv} = \text{BN}(\text{Conv}(x)), \quad (6)$$

$$\text{Branch_Max} = \text{BN}(\text{maxpooling}(x)), \quad (7)$$

$$\text{DBDS} = \text{SiLU}(\text{Branch_Conv} \oplus \text{Branch_Max}), \quad (8)$$

其中: Branch_Conv 和 Branch_Max 对应不同分支的下采样结果, BN 表示批处理归一化层, $\oplus$ 表示逐元素加和。

2.3 针对微小目标的改进检测头

在基于卷积的深度学习网络对图像进行处理的过程中, 特征图的分辨率会随着层数加深而逐渐变小。深层特征图包含丰富的语义信息, 但空间位置信息逐渐丧失。相比而言, 浅层特征图保留着较多的细粒度信息, 更适合预测无人机图像中大量的密集小目标。因此, 本文选择将较浅层的特征图集成到特征图融合序列中。所提算法将最终用于预测的特征图分为 $S \times S$ 个单元格, 每个单元格仅预测一个中心点落入该单元格内的目标。当多个目标的中心点落入同一个单元格时会导致大量目标漏检。而浅层高分辨率特征图将包含更密集的 S , 这降低了多个目标落入同一网格区域的概率, 有效提高密集小目标的检测精度。同时, 我们通过将高分辨率特征图的通道数由 128 提升至 256, 以提高特征融合过程中的浅层特征图权重, 该方法可有效提高模型对微小目标的敏感性。

此外, 在输出最终的检测结果前, 需要先对

特征进行降维,但对无人机图像中大量存在的小目标来说,低维特征不足以保留足够的有用信息。并且特征通道数不足可能会阻碍梯度回传,不利于模型训练。因此本文设计了一个基于反残差的特征增强模块 FE-IR。该结构先对特征进行升维,并利用深度卷积(Depthwise Convolution, DWConv)对高维特征进行特征提取以保证代表性。同时,我们将跳连路径建立在升维后的特征上,将增强后的特征映射至下一层。激活函数 ReLU 会将分布小于 0 的特征截断,导致信息损失。因此本文选择在深层模型上效果更好的 Swish 作为激活函数,以提高 FE-IR 模块性能。本文将该模块嵌入到检测层之前,用于对小目标特征进行增强。其结构图如图 4 所示。

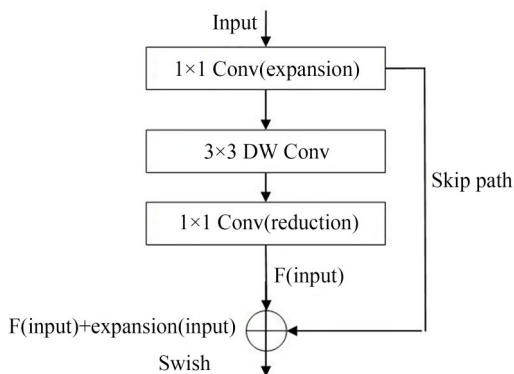


图 4 FE-IR 模块结构图

Fig. 4 Structure diagram of FE-IR

2.4 损失函数

损失函数用于度量网络的预测结果与真实标签之间的距离,通过使损失函数最小化,不断修正网络参数,实现精准预测。所提算法使用的损失函数由三部分组成:置信度损失表示检测框内是否存在目标;回归损失用于衡量预测框与真实框的误差;分类损失表示预测类别和真实标签的差异。总体损失函数定义如下:

$$Loss = A Loss_{Obj} + B Loss_{Rect} + C Loss_{Cls}, \quad (9)$$

其中: $Loss_{Obj}$, $Loss_{Rect}$, $Loss_{Cls}$ 分别表示置信度损失、回归损失和分类损失。 A, B, C 表示不同损失所占权重。

本文在计算回归损失时考虑到预测值与真实值中心点坐标、重叠面积和宽高比之间的相关性,通过 CIoU 处理回归损失。定义如下:

$$Loss_{CIoU} = 1 - IoU + \frac{\rho^2}{c^2} + \lambda v, \quad (10)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{x_{l2} - x_{l1}}{y_{l2} - y_{l1}} - \arctan \frac{x_{p2} - x_{p1}}{y_{p2} - y_{p1}} \right). \quad (11)$$

其中: ρ 为预测框和真实框的中心点距离, c 为两者的最小包围矩形的对角线长度, v 为两者的宽高比相似度, λ 为 v 的影响因子。

本文置信度损失使用二值交叉熵(Binary cross entropy, BCE)。置信度损失定义如下:

$$Loss_{Obj} = -\sum o_i \ln(\text{Sigmoid}(c_i)) + (1 - o_i) \ln(1 - \text{Sigmoid}(c_i)), \quad (12)$$

其中: o_i 表示预测框 i 内是否存在目标, c_i 表示预测值。

BCE 损失不仅适用于二分类任务,也可以通过多个二元分类叠加实现多标签分类,故本文分类损失同样采用 BCE 损失,分类损失定义如下:

$$Loss_{Cls} = -\sum_{i \in pos} \sum_{j \in cls} O_{ij} \ln(\text{Sigmoid}(C_{ij})) + (1 - O_{ij}) \ln(1 - \text{Sigmoid}(C_{ij})), \quad (13)$$

其中: O_{ij} 表示预测框 i 内的目标是否为当前类别, C_{ij} 表示预测值。

3 实验与分析

3.1 实验环境

本文实验的运行环境为:操作系统 CentOS7, 显卡型号 NVIDIA Tesla V100S-PCIE-32GB, 处理器 Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz。采用 Pytorch 深度学习框架。实验过程中,所有模型训练均采用随机梯度下降算法,初始学习率为 0.001,动量因子为 0.9, Batchsize 为 2。

3.2 实验数据集

VisDrone2019 数据集^[25]由 AISKEYEY 团队制作,其中包含 288 个视频片段,共包括 261 908 帧和 10 209 个静态图像,官方提供的数据集中选取了 6 471 张图像作为训练集,548 张图像作为验证集。数据集中包含了 11 个类别,分别是行人、汽车、人、公交车、自行车、卡车、面包车、带棚三轮车、三轮车、摩托车以及 others,其中 others 是非有效目标区域,本次实验中予以忽略。训练过程中,输入图像分辨率为 $1\,536 \times 1\,536$ 。

UAVDT 数据集^[26]由 Du 等在 2018 年欧洲计

算机视觉会议提出。数据集中包含 50 个视频片段。本文按照通用标准,将前 30 个视频片段中的 22 919 张图像用于训练,后 20 个视频片段中的 17 457 张图像用于测试。该数据集共包含汽车、货车、公交车 3 类检测目标,涉及十字路口、高速公路、街区等常见场景。训练过程中,输入图像分辨率为 800×800 。

3.3 评估指标

为了精准量化所提模型在无人机航拍图像上的检测性能,本文采用多类目标检测任务中公认度最高的评估指标:平均精度均值(mean Average Precision, mAP)。其中 mAP 是通过按 0.05 为步长在 0.5~0.95 中取 10 个 IoU 阈值再取平均精度的过程,mAP₅₀ 和 mAP₇₅ 则是分别以 0.5 和 0.75 的作为 IoU 阈值计算得到的平均精度,mAP 计算过程如下:

$$Precision = \frac{TP}{TP + FP}, \quad (14)$$

$$Recall = \frac{TP}{TP + FN}, \quad (15)$$

$$mAP = \frac{1}{n} \sum_{k=1}^n J(Precision, Recall)_k, \quad (16)$$

其中:TP(True Positives)表示真正例,FP(False Positive)表示假正例,FN(Flase Negatives)表示假反例, n 表示类别数量, $J(Precision, Recall)_k$ 表示平均精度函数。

3.4 对比实验

为了适应不同的需求,本文通过调整网络的深度和宽度获得了两种不同规模的模型:m 和 s。其中 m 模型精度更高,但检测速度相对较低。s 模型检测速度优于 m 模型,但精度相较于 m 模型有所下降。我们将两种模型在 VisDrone2019 数据集上的表现和其他目标检测算法进行比较,并在 UAVDT 数据集上验证所提方法的泛化能力。

3.4.1 VisDrone 数据集实验结果

为了直观地展示本文所提算法相较于基线模型的性能提升,我们在表 1 中显示了两种基线模型和本文所提算法对 VisDrone2019 数据集中各类目标的检测性能。实验结果表明,本文所提算法的两种模型在全部类别上均实现提升,其中,行人、人、摩托车等小目标提升效果明显。

表 1 基线模型和本文算法在 VisDrone2019 上的性能对比

Tab. 1 Comparison of the results of the baseline and the algorithm in this paper on VisDrone2019 (%)

Method	指标	Pedestrian	People	Bicycle	Car	Van	Truck	Tricycle	Awning-tricycle	Bus	Motor	ALL
Baseline-s	mAP ₅₀	66.8	53.7	35.8	88.4	56.3	45.2	42.4	21.5	64.6	64.3	53.9
	mAP	33.9	22.5	16.6	57.7	39.9	30.6	24.3	13.6	46.8	31.6	31.7
Ours-s	mAP ₅₀	72.1	58.0	42.7	90.5	61.0	53.8	48.6	25.2	72.4	68.8	59.3
	mAP	38.4	26.1	21.6	67.6	45.0	36.0	29.2	16.0	54.6	36.4	37.1
Baseline-m	mAP ₅₀	69.4	56.0	39.7	89.4	59.1	52.6	46.8	26.3	71.4	66.0	57.7
	mAP	34.2	24.5	19.7	67.5	44.2	34.3	28.4	16.2	51.9	33.8	35.4
Ours-m	mAP ₅₀	73.6	60.3	45.4	91.1	62.0	55.2	50.6	29.7	79.4	70.3	61.8
	mAP	39.4	27.1	23.2	69.1	46.5	39.1	30.4	18.3	59.2	36.9	38.9

为了客观地展示本文所提方法的性能优势,我们将两种型号的模型与其他最先进的目标检测方法在 VisDrone2019 数据集上的性能表现进行了比较和分析。结果如表 2 所示,其中所提方法的 m 模型的 mAP₅₀,mAP 和 mAP₇₅ 分别达到 62.1%,38.9% 和 39.5%,在所有对比算法中达到最优。此外,所提算法的 s 模型的 mAP₅₀,mAP

和 mAP₇₅ 分别达到 59.3%,37.1% 和 38.3%,其中 mAP 和 mAP₇₅ 达到次优。mAP₅₀ 仅次于所提算法的 m 模型和 HRDNet。值得注意的是,HRDNet 算法的输入图像分辨率远超本文所提方法,尽管这会提高精度,但这会严重影响检测速度。本文所提方法的两种模型在输入图像分辨率远低于 HRDNet 的情况下,整体性能仍然取

表 2 不同算法模型在 VisDrone2019 上的结果对比

Tab. 2 Comparison of results of different algorithm models on VisDrone2019 (%)

Method	mAP ₅₀	mAP	mAP ₇₅	AP-small	AP-mid	AP-large
ClusDet ^[10]	56.2	32.4	31.6	—	—	—
DSHNet ^[11]	51.8	30.3	30.9	—	—	—
HRDNet ^[12]	<u>62.0</u>	35.5	35.1	—	—	—
Yolov5-s	53.7	31.7	31.4	22.0	43.7	49.5
Yolov5-m	58.6	35.4	36.8	27.6	47.6	52.4
Yolov7 ^[22]	57.3	35.6	37.1	26.5	46.0	50.3
YoloX	53.5	31.4	31.7	22.5	41.5	48.5
MobileNetv3	55.4	32.9	32.9	24.5	44.3	49.5
MobileViT	55.5	33.3	33.7	24.9	44.2	41.8
mSODANet ^[27]	55.9	36.9	37.4	—	—	—
Ours-s	59.3	<u>37.1</u>	<u>38.3</u>	<u>29.2</u>	<u>48.5</u>	<u>51.6</u>
Ours-m	62.1	38.9	39.5	31.2	51.3	52.5

注:加粗字体表示最优值,下划线表示次优值。

得了最优和次优的结果。因此,本文所提方法更适合处理无人机图像目标检测。

3.4.2 UAVDT 数据集实验结果

为了探究本文所提算法的泛化能力,我们测试了两种型号的基线模型和本文所提算法在 UAVDT 数据集中的检测性能。结果如表 3 所示,相较于基线模型,本文所提算法在全部类别

表 3 基线模型和本文方法在 UAVDT 上的性能对比

Tab. 3 Comparison of the results of the baseline and the algorithm in this paper on UAVDT (%)

Method	指标	Car	Truck	Bus	All
Baseline-s	mAP ₅₀	73.0	15.2	26.6	38.3
	mAP	41.2	8.1	15.9	21.7
Ours-s	mAP ₅₀	73.3	19.4	39.7	44.1
	mAP	40.9	10.9	23.0	24.9
Baseline-m	mAP ₅₀	72.1	8.55	40.5	40.4
	mAP	39.6	4.86	22.9	22.5
Ours-m	mAP ₅₀	73.7	18.0	40.9	44.2
	mAP	42.3	11.2	24.1	25.8

上的检测精度均实现提升。

表 4 列举了本文所提方法和其他对比算法在 UAVDT 数据集上的性能表现。实验结果表明,本文所提方法的 m 模型在 UAVDT 数据集上的 mAP₅₀, mAP 和 mAP₇₅ 分别达到 44.7%, 25.3% 和 28.1%, 相较于其他对比方法, m 模型的三项指标均达到最优, 而 s 模型的 mAP₅₀ 和 mAP 也达到次优。综上, 本文所提方法具有较高的检测精度和泛化能力。

表 4 不同算法在 UAVDT 数据集上的结果对比

Tab. 4 Comparison of results of different algorithm models on UAVDT (%)

Method	mAP ₅₀	mAP	mAP ₇₅
ClusDet ^[10]	26.5	13.9	12.5
DMNet ^[28]	24.6	14.7	16.3
GLSAN ^[29]	30.5	19.0	21.7
CDMNet ^[30]	29.1	16.8	18.5
DSHNet ^[11]	30.4	17.8	19.7
Yolov5-s	38.3	21.7	22.6
Yolov5-m	41.4	22.6	24.1
PRDet ^[31]	34.1	19.8	21.3
UFPMP-Det ^[32]	38.7	24.6	<u>28.0</u>
Ours-s	<u>44.1</u>	<u>24.9</u>	27.3
Ours-m	44.7	25.3	28.1

注:加粗字体表示最优值,下划线表示次优值。

3.5 消融实验

本文提出的 DBS-FU 模块通过并行分支分别进行采样,以获取高质量的特征融合结果。为了验证单一采样和并行采样方式的性能差异,我们在 VisDrone2019 数据集上进行了相关测试,结果如表 5 所示。当使用基线模型默认的 Nearest 上采样和 Conv 下采样时, mAP₅₀ 和 mAP 分别为 53.7% 和 31.7%。相较于 Nearest, Bilinear 利用相邻像素分别在两个方向上进行插值。在替换 Bilinear 后,模型性能有所提升。而本文所提出的 DBUS 模块通过不同分支分别进行上采样,并在批处理归一化后对不同的上采样结果进行逐元素加,再由激活函数 SiLU 引入更多非线性因素。相较于单一方式的采样结果, DBUS 所得到

表 5 不同采样方式性能对比
Tab.5 Performance comparison of different sampling methods (%)

Method	Up-sampling			Down-sampling		
	Nearest	Bilinear	DBUS	Conv	Pooling	DBDS
mAP ₅₀	53.7	54.1	<u>54.5</u>	53.7	53.5	54.8
mAP	31.7	32.1	<u>32.8</u>	31.7	32.0	33.2

注:加粗字体表示最优值,下划线表示次优值。

的特征图中携带更多的信息量,在特征融合过程中可以向浅层特征图传递更丰富的语义信息,从而提高模型对目标的分类性能。因此,当使用 DBUS 时,模型精度优于任意一种单一采样方法。在下采样时,使用卷积或池化任意一种下采样方式都会导致另一部分信息缺失,而本文所提出的 DBDS 模块同时使用两条采样分支捕获不同信息,有效减少信息损耗。

表 6 消融实验结果
Tab.6 Ablation experiment result

	Baseline	Fourhead	Our-Four-head	NRCS	DBUS	DBDS	FE-IR(3)	Parm	GFLOPs	mAP ₅₀ (%)	mAP(%)	FPS
1	✓							7.028M	16.0	53.7	31.7	35.6
2	✓	✓						7.262M	19.3	54.8	32.4	24.7
3	✓		✓					6.911M	19.1	56.1	34.3	25.6
4	✓		✓	✓				6.583M	19.3	56.9	35.1	27.1
5	✓		✓		✓			6.942M	21.4	56.3	34.6	27.1
6	✓		✓			✓		6.982M	21.6	56.8	35.2	26.9
7	✓		✓		✓	✓		7.013M	23.4	57.1	35.2	26.5
8	✓		✓	✓	✓	✓		6.686M	23.6	58.3	35.9	26.9
9	✓		✓	✓	✓	✓	✓	7.875M	26.8	59.7	37.1	22.1

为了验证本文所提出的各个模块对模型整体性能的影响,本文选取 6.0 版本的 YOLOv5-s 作为基线模型,通过消融实验对所提出的各个模块的有效性进行验证。结果如表 5 所示。对于消融实验结果的具体分析如下:

在未添加任何改进模块时,基线模型的 mAP₅₀ 和 mAP 分别为 53.7% 和 31.7%。在此基础上,引入浅层高分辨率特征图构建新的检测头, mAP₅₀ 和 mAP 增长至 54.8% 和 32.4%,但参数量有所上升。当提高特征融合序列中浅层特征图权重占比,并通过加大深层特征图中卷积步距来平衡参数时, mAP₅₀ 和 mAP 进一步增长至 56.1% 和 34.3%,并且参数量下降。NRCS 模块通过结合局部信息和全局信息,降低复杂背景对模型的影响。数据表明在加入 NRCS 模块后, mAP₅₀ 和 mAP 增长至 56.9% 和 35.1%,参数量

明显下降。考虑到现有特征融合方法无法高效融合多尺度信息,信息衰减严重,构建 DBS-FU 特征融合模块,该模块通过并行分支生成高质量特征图,提高信息传递效率。数据表明分别加入 DBUS 模块和 DBDS 模块后, mAP₅₀ 和 mAP 均出现上升,证明采用分支并行策略生成的采样结果在特征融合过程中效果更优。为了对小目标特征进行增强,设计了 FE-IR 模块,数据表明加入 3 组 FE-IR 模块后, mAP₅₀ 和 mAP 增长至 59.7% 和 37.1%,但检测速度有所下降。鉴于 FE-IR 模块即插即用,使用者可以根据需求灵活增加或减少 FE-IR 模块的数量。综上所述,消融实验表明本文所提的四个改进点都可以提高无人机图像目标检测的精度。

3.6 算法有效性展示

为客观表现本文所提算法相较于其他算法

的优势,本文在 VisDrone2019 和 UAVDT 数据集中选取了不同场景下的图片作为测试对象,不同检测结果中差异较大的区域被黄色线框标注(彩图见期刊电子版)。

如图 5 所示,我们在 VisDrone2019 数据集中选取了四种常见的无人机图像,分别为高空视角、低空视角、密集小目标场景和夜间场景。图 5(a)展示无人机在城市高空捕获的鸟瞰图像,该类图像分辨率较高,但背景信息复杂,容易对待测目标造成干扰。如图 5(a)结果所示,其他检测算法不仅将图像右上角中的形似物体(民房、小屋等)误判为目标,而且部分待测目标并未被检测,而本文所提算法可以通过引入全局信息对复杂背景进行有效抑制。图 5(b)中展示了无人机在低空作业环境下捕获的图像,该类图像中目标尺寸差异明显。观察结果可发现,此类图像中漏

检情况较为频繁,而所提方法在检测不同尺寸的目标时更具有优势。图 5(c)展示了密集小目标场景中模型的检测性能,其他算法中将特征图分为若干网格单元,每个网格单元预测一个目标。当目标尺寸较小且排列密集时,多个目标落入同一网格单元的概率将大大增加,导致大量目标未被检测。而所提方法通过改进后的检测头引入浅层特征层增加了网格单元的数量,减少了多个小尺寸目标落入同一网格的概率。图 5(d)中展示了夜间场景的检测结果。目标检测模型极易受到光线不足带来的影响。该类图像中,可用特征较少,目标难以与背景分离。其他算法的检测结果中大量受到遮挡或处于暗处的目标未能被检测。而本文所提算法可以提取到精细和具有鉴别性的特征,从而提高在此类场景中的检测能力。

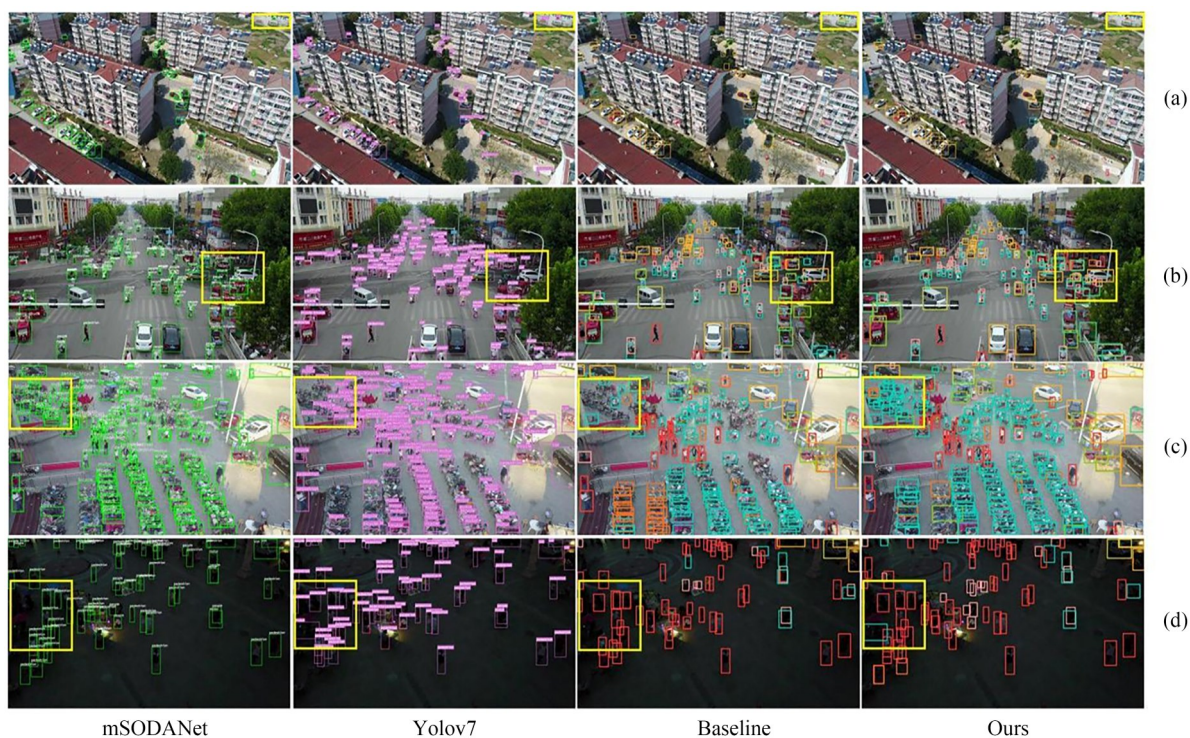


图 5 不同算法在 VisDrone2019 数据集上的检测结果对比

Fig. 5 Comparison of detection results of different algorithms on VisDrone2019

图 6 中展示了不同算法在 UAVDT 数据集上的检测结果。结果表明,本文所提方法街道、路口、夜晚多种复杂场景中取得了显著的检测结果。

不同数据集上的验证结果表明,本文所提方法具有抗干扰性强、检测精度高、泛化能力强等优势,更适用于无人机图像目标检测任务。



图 6 不同算法在 UAVDT 数据集上的检测结果对比

Fig. 6 Comparison of detection results of different algorithms on UAVDT

4 结 论

无人机图像中存在背景复杂、尺度变化剧烈、包含大量密集微小目标等问题,这对无人机图像目标检测任务带来了一定挑战性。针对上述问题,本文提出了一种联合自注意力和分支采样的无人机图像目标检测方法。首先,该方法在现有的目标检测算法基础上,构建卷积和自注意力相结合的嵌套残差模块 NRCS,通过结合局部信息和全局信息提高模型对上下文信息的捕获

能力,缓解复杂背景带来的干扰问题。其次,针对现有特征融合模块的缺陷,通过分支并行策略构建 DBS-FU 特征融合模块,提高特征融合质量。最后,将经过权重调整后的浅层细粒度特征图引入到特征融合序列中,设置对应的检测头以适应尺度变化的影响,并利用提出的特征增强模块 FE-IR 捕获更具有鉴别能力的小目标特征。实验结果表明,本文所提方法在不同场景中均表现出了良好的检测性能。

参考文献:

- [1] 曲优,李文辉. 基于锚框变换的单阶段旋转目标检测方法[J]. 吉林大学学报(工学版), 2022, 52(01): 162-173.
QU Y, LI W H. Single-stage rotated object detection network based on anchor transformation [J]. *Journal of Jilin University (Engineering and Technology Edition)*, 2022, 52(1): 162-173. (in Chinese)
- [2] 李超,黄新宇,王凯. 基于特征融合和自学习锚框

的高分辨率图像小目标检测算法[J]. 电子学报, 2022, 50(7):1684-1695.

LI C, HUANG X Y, WANG K. Small object detection of high-resolution images based on feature fusion and learnable anchor[J]. *Acta Electronica Sinica*, 2022, 50(7): 1684-1695. (in Chinese)

- [3] 刘芳,韩笑. 基于多尺度深度学习的自适应航拍目标检测[J]. 航空学报, 2022, 43(5):463-474.
LIU F, HAN X. Adaptive aerial object detection based on multi-scale deep learning [J]. *Acta Aeronautica et Astronautica Sinica*, 2022, 43(5): 463-

474. (in Chinese)
- [4] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection [C]. 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. June 20-25, 2005, San Diego, CA, USA. IEEE, 2005: 886-893.
 - [5] LIENHART R, MAYDT J. An Extended Set of Haar-Like Features for Rapid Object Detection[C]. *Proceedings of International Conference on Image Processing*. September 22-25, 2002, Rochester, NY, USA. IEEE, 2002: 1.
 - [6] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR*. December 8-14, 2001, Kauai, HI, USA. IEEE, 2003: 1.
 - [7] 林浩, 李雷孝, 王慧. 支持向量机在智能交通系统中的研究应用综述[J]. *计算机科学与探索*, 2020, 14(6):901-917.
XU D G, WANG L, LI F. Review of typical object detection algorithms for deep learning[J]. *Computer Engineering and Applications*, 2021, 57(8): 10-25. (in Chinese)
 - [9] 张财广, 熊博莅, 匡纲要. 光学卫星遥感图像舰船目标检测综述[J]. *电波科学学报*, 2020, 35(5): 637-647.
ZHANG C G, XIONG B L, KUANG G Y. A survey of ship detection in optical satellite remote sensing images [J]. *Chinese Journal of Radio Science*, 2020, 35(5):637-647. (in Chinese)
 - [10] YANG F, FAN H, CHU P, *et al.* Clustered object detection in aerial images [C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 27-November 2, 2019, Seoul, Korea (South). IEEE, 2020: 8310-8319.
 - [11] YU W P, YANG T, CHEN C. Towards resolving the challenge of long-tail distribution in uav images for object detection [C]. 2021 *IEEE Winter Conference on Applications of Computer Vision (WACV)*. January 3-8, 2021, Waikoloa, HI, USA. IEEE, 2021: 3257-3266.
 - [12] LIU Z M, GAO G Y, SUN L, *et al.* HRDNet: high-resolution detection network for small objects [C]. 2021 *IEEE International Conference on Multimedia and Expo (ICME)*. July 5-9, 2021, Shenzhen, China. IEEE, 2021: 1-6.
 - [13] WU Z Y, SURESH K, NARAYANAN P, *et al.* Delving into robust object detection from unmanned aerial vehicles: a deep nuisance disentanglement approach [C]. 2019 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 27-November 2, 2019, Seoul, Korea (South). IEEE, 2020: 1201-1210.
 - [14] YOUSSEF Y, ELSHENAWY M. Automatic vehicle counting and tracking in aerial video feeds using cascade region-based convolutional neural networks and feature pyramid networks[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2021, 2675(8): 304-317.
 - [15] LI J N, LIANG X D, WEI Y C, *et al.* Perceptual Generative Adversarial Networks for Small Object Detection [C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 21-26, 2017, Honolulu, HI, USA. IEEE, 2017: 1951-1959.
 - [16] TANG Z Y, LIU X, YANG B J. PENet: object detection using points estimation in high definition aerial images [C]. 2020 *19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. December 14-17, 2020, Miami, FL, USA. IEEE, 2021: 392-398.
 - [17] LAMINE MEKHALFI M, BEJIGA M B, SORESINA D, *et al.* Capsule networks for object detection in UAV imagery [J]. *Remote Sensing*, 2019, 11(14): 1694.
 - [18] CHEN C Y, GONG W G, CHEN Y L, *et al.* Object detection in remote sensing images based on a scene-contextual feature pyramid network [J]. *Remote Sensing*, 2019, 11(3): 339.
 - [19] GIRSHICK R. Fast R-CNN [C]. 2015 *IEEE International Conference on Computer Vision (ICCV)*. December 7-13, 2015, Santiago, Chile. IEEE, 2016: 1440-1448.
 - [20] 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.
 - [21] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal Speed and Accuracy of Object Detection [EB/OL]. 2020: arXiv:

2004. 10934. <https://arxiv.org/abs/2004.10934>
- [22] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: Trainable Bag-of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors [EB/OL]. 2022; arXiv: 2207.02696. <https://arxiv.org/abs/2207.02696>
- [23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. 2020; arXiv: 2010.11929. <https://arxiv.org/abs/2010.11929>
- [24] SUN Z Q, CAO S C, YANG Y M, *et al.* Rethinking transformer-based set prediction for object detection [C]. 2021 *IEEE/CVF International Conference on Computer Vision (ICCV)*. October 10-17, 2021, Montreal, QC, Canada. IEEE, 2022: 3591-3600.
- [25] DU D W, ZHU P F, WEN L Y, *et al.* VisDrone-DET2019: the vision meets drone object detection in image challenge results[C]. 2019 *IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. October 27-28, 2019, Seoul, Korea (South). IEEE, 2020: 213-226.
- [26] DU D W, QI Y K, YU H Y, *et al.* *The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking*[M]. Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 375-391.
- [27] CHALAVADI V, JERIPOTHULA P, DATLA R, *et al.* mSODANet: a network for multi-scale object detection in aerial images using hierarchical dilated convolutions [J]. *Pattern Recognition*, 2022, 126: 108548.
- [28] 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)*. June 14-19, 2020, Seattle, WA, USA. IEEE, 2020: 737-746.
- [29] DENG S T, LI S, XIE K, *et al.* A global-local self-adaptive network for drone-view object detection[J]. *IEEE Transactions on Image Processing*, 2021, 30: 1556-1569.
- [30] DUAN C Z, WEI Z W, ZHANG C, *et al.* Coarse-grained density map guided object detection in aerial images[C]. 2021 *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. October 11-17, 2021, Montreal, BC, Canada. IEEE, 2021: 2789-2798.
- [31] LENG J X, MO M, ZHOU Y H, *et al.* Pareto re-focusing for drone-view object detection [C]. *IEEE Transactions on Circuits and Systems for Video Technology*. September 28, 2022, IEEE, 2022: 1320-1334.

通讯作者:



张云佐(1984—),男,河北石家庄人,博士,副教授,博士生导师,2016年于北京理工大学获博士学位,主要从事计算机视觉、人工智能、大数据方面的研究。E-mail: zhangyunzuo888@sina.com

作者简介:



武存宇(1998—),男,山西太原人,硕士研究生,2021年于陕西理工大学获得学士学位,主要从事图像处理,目标检测方面的研究。E-mail: wucunyu1410@sina.com