◎图形图像处理◎

改进YOLOv5的无人机影像小目标检测算法

谢椿辉[1,2],吴金明[1],徐怀宇[2]
1.中国科学院 上海高等研究院,上海 201210
2.上海科技大学 信息科学与技术学院,上海 201210

摘 要:无人机航拍影像具有目标尺度变化大、背景复杂等诸多特性,导致现有的检测器难以检测出航拍影像中的小目标。针对无人机影像中小目标误检漏检的问题,提出了改进YOLOv5的算法模型Drone-YOLO。增加了检测分支以提高模型在多尺度下的检测能力。设计了多层次信息聚合的特征金字塔网络结构,实现跨层次信息的融合。设计了基于多尺度通道注意力机制的特征融合模块,提高对小目标的关注度。将预测头的分类任务与回归任务解耦,使用Alpha-IoU优化损失函数定义,提升模型检测的效果。通过无人机影像数据集VisDrone的实验结果表明,Drone-YOLO模型较YOLOv5模型在AP50指标上提高了4.91个百分点,推理延时仅需16.78 ms。对比其他主流模型对于小目标拥有更好的检测效果,能够有效完成无人机航拍影像的小目标检测任务。

关键词:目标检测;无人机;小目标;注意力机制;特征融合;YOLO
文献标志码:A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2212-0336

Small Object Detection Algorithm Based on Improved YOLOv5 in UAV Image

XIE Chunhui[1,2], WU Jinming[1], XU Huaiyu[2]
1.Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China
2.School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

Abstract: UAV aerial images have many characteristics, such as large-scale changes and complex backgrounds, so it is difficult for the existing detectors to detect small objects in aerial images. Aiming at the problem of mistake detection and omission, a small object detection algorithm model Drone-YOLO is proposed. A new detection branch is added to improve the detection capability at multiple scales, meanwhile the model contains a novel feature pyramid network with multi-level information aggregation, which realizes the fusion of cross-layers information. Then a feature fusion module based on multi-scale channel attention mechanism is designed to improve the focus on small objects. The classification task of the prediction head is decoupled from the regression task, and the loss function is optimized using Alpha-IoU to improve the accuracy of detection. The experimental results of VisDrone dataset show that the Drone-YOLO has improved the AP50 by 4.91 percentage points compared with the YOLOv5, and the inference time is only 16.78 ms. Compared with other mainstream models, it has a better detection effect for small targets, and can effectively complete the task of small target detection in UAV aerial images.

Key words: object detection; unmanned aerial vehicle(UAV); small object; attention mechanism; feature fusion; YOLO

近些年,随着无人机(unmanned aerial vehicle,UAV)相关技术的发展,无人机凭借其轻便快速的特性,在农业、电网和城市巡检等领域得到了广泛的应用。但无人机航拍影像中,由于拍摄高度较高,图像中的各类感兴趣的目标如行人、自行车等尺度较小,且容易受环境干扰,导致难以被常规目标检测算法检测出来。因此提高算法对无人机航拍图像中小目标的检测能力成为了目标检测领域一个具有挑战性的研究方向。

近几年,卷积神经网络(convolutional neural network,CNN)在计算机视觉领域中取得了巨大的突破。虽然基

基金项目:中国科学院战略性先导科技专项(XDC02000000);SEANET规模试验验证评估与示范应用(XDC02070800)。
作者简介:谢椿辉(1998—),男,硕士研究生,研究方向为目标检测;吴金明,男,博士研究生,研究方向为计算机视觉;徐怀宇,通信作者,男,博士,教授,研究方向为人工智能。
收稿日期:2022-12-26 **修回日期:**2023-02-27 **文章编号:**1002-8331(2023)09-0198-09

于卷积神经网络的检测算法在常规目标检测任务中已经取得了巨大成功，但是在无人机场景下对于小目标的检测效果较差。首先，主要原因在于无人机航拍影像中物体的尺寸变化巨大，而卷积神经网络中单层特征图的表征能力有限。其次，航拍影像的背景复杂多变，小目标容易被复杂背景干扰，从而影响检测效果。

因此本文提出了面向无人机场景的目标检测算法，针对现有目标检测算法在无人机场景下难以检测出小目标的问题，重新进行网络结构的设计，融合多尺度、多层次的信息，提高网络的表征能力，同时设计多尺度通道注意力模块以提升对小目标的关注度。本文的贡献如下：

(1)提出了适应小目标的检测模型Drone-YOLO，网络结构方面增加检测头负责检测小目标，提高多尺度的检测能力。同时设计了一个多层次信息聚合的特征金字塔网络结构(multi-level feature pyramid network，ML-FPN)，将不同层级信息进行融合以提高网络的表征能力，充分利用浅层信息辅助网络检测小目标。

(2)设计了一个基于多尺度注意力机制的特征融合模块(multi-scale attentive feature fusion，MAFF)，采用多尺度信息生成通道注意力并融合多尺度信息。将其整合到网络结构中，提升对小目标区域的关注度与精确度，从而提升模型的检测效果。

(3)设计了一个平衡效率与精度的精简解耦头(simple decouple head，SD-Head)，将分类任务和回归任务分离，减少任务间差异带来的影响，从而提高目标检测的精度。

(4)使用Alpha IoU[1]优化了损失函数的计算方式，提升置信度的同时提高在小数据集以及存在数据噪声的情况下模型检测效果的鲁棒性。

1　相关工作

1.1　小目标检测

目标检测中对于小目标的定义有绝对尺寸和相对尺寸两种定义。绝对尺寸定义像素小于32×32的目标定义为小目标，而相对尺寸则定义长宽比例小于原图像尺寸的0.1倍的目标为小目标。小目标检测是目标检测中一个富有挑战的难题，提高检测小目标效果对如何利用图像特征提出了相当高的要求。近年来的许多工作提出了许多有用的方法来提高小目标检测的性能。

数据增强方面，Kisantal等人[2]通过复制粘贴小目标来提高在数据集中所占比例，从而提高小目标对网络的贡献，提升模型对小目标的检测效果。Chen等人提出了自适应采样[3]的方法，采用预训练好的语义分割网络，利用侵蚀算法和滤波器过滤噪声，提取到一个合适有效位置来复制粘贴物体，从而达到了数据增强的效果。尺度匹配[4]则是将网络预训练的数据与检测器所学习的数据集之间的特征进行对齐，以充分利用预训练的网络。Chen等人[5]提出了Stitcher方法，采用损失函数作为反馈，当小目标贡献过小时，则在下一次迭代中通过图片拼贴提高小目标占比以提高小目标训练效果。虽然数据增强在一定程度上提高了小目标的检测效果，但它只是简单地提高小目标的比例以及贡献，缺乏对深层语义信息的利用。

多尺度学习则是通过融合浅层细节信息与深层语义信息来提高网络表征能力，从而提高对于小目标的检测效果，但多尺度学习增加了许多参数，减慢了推理速度。最经典的多尺度学习的网络结构是特征金字塔结构FPN[6]，图像经过自下而上的特征提取后，再次自上而下进行特征融合，最后送入检测头进行回归预测。而拓展特征金字塔网络EFPN[7]则是在FPN的结构上增加了一个特征纹理迁移模块，用于超高分辨率的特征并同时提取可信的区域细节，增强小目标的特征信息，从而提升模型对小目标的检测效果。

小目标由于尺寸小、利用信息少，因此可以通过上下文来增强模型的检测能力。李青援等人[8]在SSD模型中引入一条自深向浅的递归反向路径，通过特征增强模块将深层包含上下文信息的语义特征增强到浅层。梁延禹等人[9]使用特征图的空间和通道间全局信息来增强浅层特征图中小目标的上下文信息。

最近针对小目标的目标检测算法中，TPH-YOLOv5[10]将Transformer引入到YOLOv5的预测头以提高网络的预测回归能力，同时使用注意力机制提高对小目标关注度。QueryDet[11]则是采用了一种查询机制来加速目标检测器的推理速度，利用低分辨率特征预测粗略定位以引导高分辨率特征进行更精确的预测回归。

1.2　注意力机制

注意力机制的核心在于让网络关注重点信息而忽略无关信息。其分类大致可以分为空间注意力机制、通道注意力机制和混合注意力机制。空间注意力机制其设计思想是在特征图中并不是每个区域都对任务有重要贡献，寻找特征图中最重要的位置进行处理才能提高任务的准确度。空间注意力机制的代表方法有STN[12]和DCN[13]等。通道注意力关注的是不同通道之间的关系，不同通道包含的信息是不同的，对相应的任务的影响也是不同的，因而学习每个通道的重要程度，针对不同任务增强或抑制不同的通道从而达到提高任务准确度的效果。通道注意力机制中代表性方法有SENet[14]、ECA[15]和CoordAttention[16]等。混合注意力机制则是混合了空间注意力机制和通道注意力机制的方法，其代表性方法有CBAM[17]、DANet[18]和CCNet[19]等。

2　Drone-YOLO 算法

本章将详细介绍本文所提出的Drone-YOLO模型，首先将简单介绍基准模型YOLOv5，然后详细叙述本文Drone-YOLO模型中改进的网络结构以及针对小目标检测而新设计的模块，最后将给出本文改进的损失函数的计算方式。

2.1　YOLOv5概述

YOLOv5是一个非常流行的目标检测框架，最早推出于2020年6月份。如图1所示，YOLOv5的网络结构较为简洁，其大致可以分为三部分：用于特征提取的骨干网络（Backbone）、用于特征融合的颈部（Neck）网络和用于目标类别和位置回归检测的头部（Head）网络。

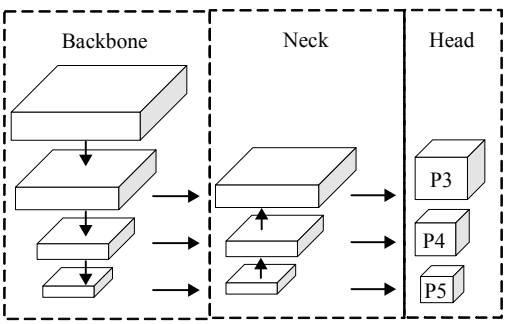


图1　YOLOv5模型结构简图

Fig.1　Architecture of YOLOv5

YOLOv5框架在数据管道的建立、模型训练、模型推理和模型部署等方面提供了非常易用的功能，在实际的目标检测开发中有着非常广泛的运用。其在训练部分采用了大量数据增强的技巧，不仅包含剪裁、翻转和缩放等几何变换，还采用了混合方法和马赛克等方法进行数据增强，这些方法的集成大大增强了模型的检测能力。

2.2　Drone-YOLO模型

本文Drone-YOLO模型的网络结构如图2所示，其中虚线箭头为ML-FPN结构新增的通路，虚线框为新增检测分支P2，灰色部分为本文新设计的模块，其中MAFF为基于多尺度注意力机制的特征融合模块，用于提高对小目标的关注度。Adaptive Fusion为自适应的融合模块，用于在ML-FPN结构中替换原有拼贴操作，保持参数不过多增大以及融合多层次信息。最后，模型的回归预测头更换为本文设计的精简解耦头SD-Head，以提高检测精度。

图2中Conv模块是一个基础模块，其包含了卷积、归一化和激活函数三个操作。本文结构图中的Conv均是Conv模块而非单纯的卷积操作。C3模块则是一个用于特征提取的模块，此模块不改变特征图的大小，采用了ResNet里的残差结构，其内部用于特征提取的BottleNeck数量可以进行调节。此外，Upsample为上采样操作，Concat为拼贴操作。

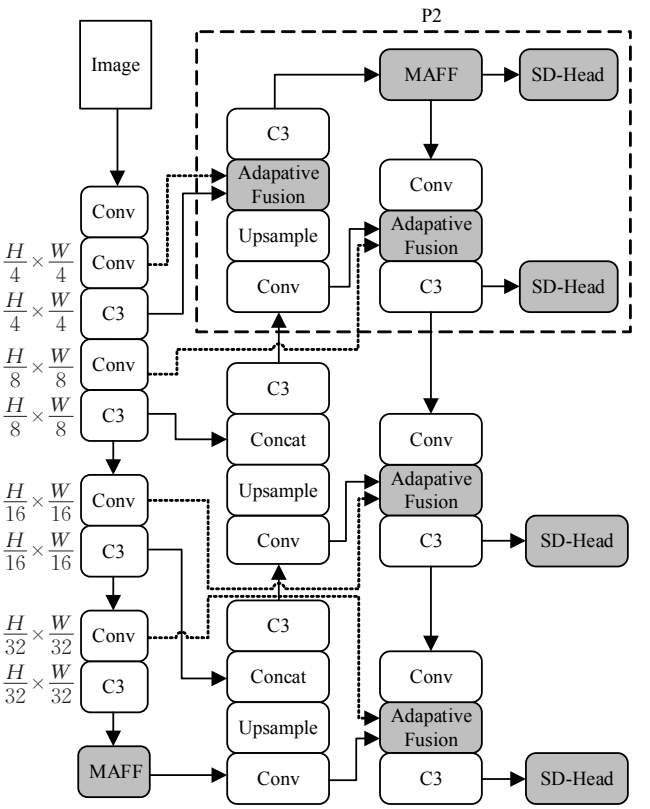


图2　Drone-YOLO模型结构图

Fig.2　Structure of proposed Drone-YOLO

2.2.1　新增小目标检测分支

如图2所示，虚线框内为新增的P2检测分支，此分支用于检测极小的目标。P2分支的输入大部分来自于浅层的卷积层，其包含许多形状、位置和大小等信息，而深层的特征图经过多次卷积池化后会损失许多信息，大目标的特征可能会掩盖过小目标的信息，造成误检漏检。所以引入了浅层信息的P2分支可以有效定位小目标的位置，从而能够更好地检测小目标。

同时，基于锚框的本文模型对于锚框设定是较为敏感的，而新增的P2检测分支进行预测回归时，锚框的大小设定为数据集进行*K*-means聚类分析得到的小目标尺寸，各个分支的锚框设定如表1所示。这样新增的P2分支能够减少由于物体过小而锚框过大导致的物体被忽略掉的情况，这能够有效缓解由于锚框设定所带来的误检和漏检情况。

表1　各个检测分支的锚框设定表

Table 1　Anchor settings for each detection branch

检测分支	锚框设定
P2	(1,4),(2,9),(5,6)
P3	(5,13),(10,10),(8,20)
P4	(19,17),(15,31),(34,42)
P5	(30,61),(62,45),(59,119)

2.2.2　多层次信息聚合网络

如图2所示，虚线为本文设计的多层次信息聚合网络（ML-FPN）结构新增的浅层信息的通道。添加浅层

信息通道的原因在于，随着网络的加深，小目标的信息会随之消失或降低，充分利用骨干网络的浅层信息成为了提高小目标检测精度的关键。

本文的思路是将浅层的特征图、中层与深层特征图进行融合，这样就同时保留了多层次的信息。同时为了不过多增大模型的参数量以及保持通道数不变，本文设计了一个自适应融合模块（adaptive fusion），其既能在不增加过多参数的情况下保持原有的通道数不变，也能进行多层次的特征融合，充分利用多层次信息。自适应融合模块的网络结构如图3所示。

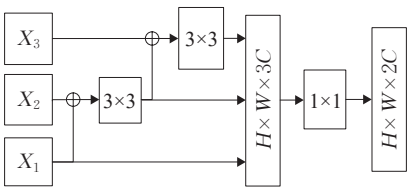


图3　自适应融合模块结构

Fig.3　Structure of proposed adaptive fusion module

模块的输入为三个形状大小以及通道数一致的特征图，大小假定为 $H\times W\times C$ ，其三个输入分别来自网络浅层（X_1）、中层（X_2）以及深层的来自上一个Conv模块的输出（X_3）。浅层特征图存在一条通路直接进行拼接，这条通路可以充分利用浅层信息，同时中层和深层的信息融合后经过3×3卷积后再与 X_1 进行拼贴，最后利用1×1的卷积进行通道降维。计算公式如式（1）~（4）：

$$W_1 = X_1 \tag{1}$$

$$W_2 = f^{3\times 3}(W_1 + X_2) \tag{2}$$

$$W_3 = f^{3\times 3}(W_2 + X_3) \tag{3}$$

$$W = f^{1\times 1}[W_1 : W_2 : W_3] \tag{4}$$

其中，$f^{1\times 1}$ 和 $f^{3\times 3}$ 分别表示1×1卷积和3×3卷积操作。此模块的直接通路可以充分利用充分浅层信息，而相加和卷积的操作混合了浅层、中层和深层的信息，多层次信息的融合可以提高对小目标的检测效果。

2.2.3　多尺度注意力特征融合模块

ML-FPN结构在骨干部分和P2分支部分添加了本文设计的基于多尺度注意力机制的特征融合模块MAFF来提高模型对于小目标的关注度，从而提高对小目标的检测效果，模块结构图如图4所示。

输入的特征图会经过三个不同核大小的最大池化层（MaxPool），以获得不同尺度的特征信息。其中，P2分支处的核大小设定为（1，3，5）以适应小目标的检测，而骨干网络处的核大小设定为（3，5，9）。

如图4所示，虚线框内最后输出了一个通道注意力权重，此权重即本文所提出的多尺度注意力机制。虚线框的输入为三个最大池化层输出的多尺度特征图，经过全局平均池化（GAP）后再次进行拼贴，最后通过全连接层（MLP）和Sigmoid函数形成了由多尺度信息的通道

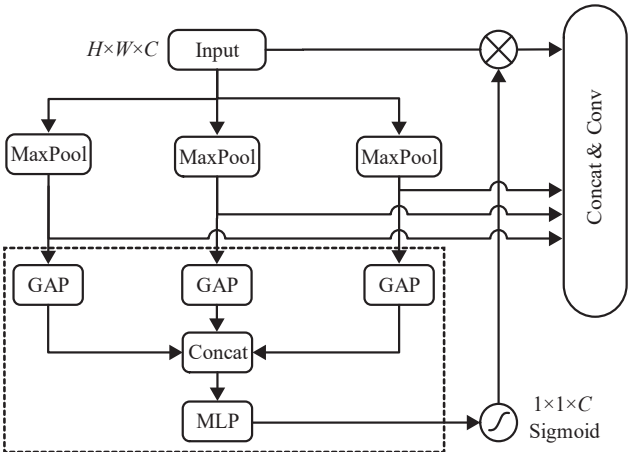


图4　MAFF模块结构

Fig.4　Structure of proposed MAFF module

注意力权重。此权重最后与原输入特征图相乘后与上述三个最大池化层的输出共同拼贴后卷积得到最终的输出。

此模块用不同核大小的最大池化层获取到了多尺度信息，同时多尺度信息指导生成了注意力权重，实现了多尺度特征间跨通道的信息交互，能有效提高模型对小目标的关注度。

2.2.4　精简的解耦头

在目标检测中，分类任务和回归任务之间是存在冲突的。如图5所示，YOLOv5最后的检测头中分类任务和回归任务是共享权重强耦合在一起的，而YOLOX[20]证明了将分类任务和回归任务解耦可以提高网络的检测效果。

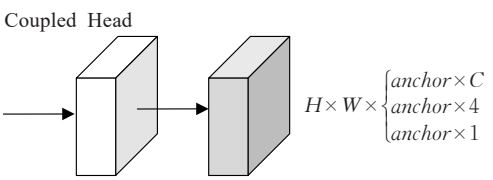


图5　YOLOv5的耦合头结构图

Fig.5　Structure of coupled head of YOLOv5

但YOLOX的解耦头新增了多个额外的卷积层，本文在综合考虑到精度与速度之间的平衡后，重新设计了一个简洁高效的解耦头SD-Head，在提升精度的同时，尽量减少模块带来的推理延时。本文设计的精简解耦头如图6所示。

解耦头的输入先用1×1卷积将通道数减少到256，再经过3×3卷积通道数减为128后，分别进入两条支路，一条用于分类任务，另外一条用于回归任务，回归任务又解耦成对于位置的回归任务和置信度的回归任务。检测头解耦后减少了任务之间的差异带来的预测偏差，从而提高了模型检测的精度。本文精简的解耦头采用了更少的卷积层，减少了通道数，平衡了解耦头的效率与精度。

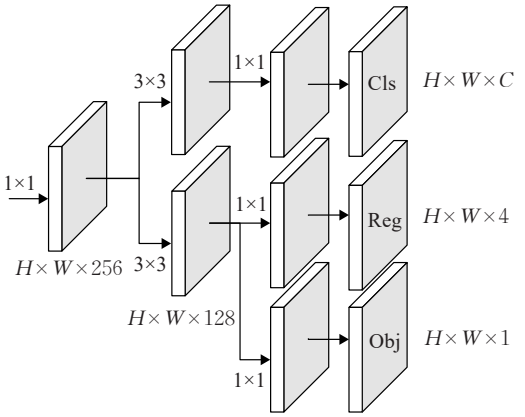


图6　精简解耦头结构图

Fig.6　Structure of proposed simple decouple head

2.3　改进的损失函数

模型的损失函数Loss总体上分为三部分,其计算公式如公式(5)所示:

$$\mathcal{L}=\mathcal{L}_{obj}+\mathcal{L}_{cls}+\mathcal{L}_{bbox} \tag{5}$$

其中,$\mathcal{L}_{obj}$ 是物体置信度的损失,采用的是二元交叉熵损失。$\mathcal{L}_{cls}$ 则是物体的分类损失,采用的是交叉熵损失。而 $\mathcal{L}_{bbox}$ 为预测框位置的损失,本文采用Alpha-IoU优化了预测框位置的损失 $\mathcal{L}_{bbox}$ 的计算方式,本文计算预测框的损失函数如公式(6)所示:

$$\mathcal{L}_{bbox}=1-IoU^{\alpha}+\frac{\rho^{2\alpha}(b,b^{gt})}{C^{2\alpha}}+(\beta v)^{\alpha} \tag{6}$$

其中,IoU 是预测框 (b) 和真实框 (b^{gt}) 的交并比,α 是可以调节的超参数,通过调节 α 可以调节预测框的精度。此外 ρ 表示计算预测框和真实框中心之间的欧几里德距离,C 表示最小包围预测框和真实框的框的对角线长度,v 是度量预测框和真实框的长宽比相似性参数,其计算公式如公式(7)所示:

$$v=\frac{4}{\pi}(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}) \tag{7}$$

其中,(w^{gt},h^{gt}) 和 (w,h) 分别是真实框和预测框的宽和高。β 则是一个权重,其计算公式如公式(8)所示:

$$\beta=\frac{v}{(1-IoU)+v} \tag{8}$$

本文方法在计算预测框位置损失中,将真实框与预测框之间的距离、重叠率、长宽比以及尺度等方面都考虑了进去,因此回归预测更为精准,在模型训练时能快速收敛。同时本文方法在小数据集和存在数据噪声情况下仍能保持较高的鲁棒性。

3　实验与结果分析

本文实验数据是采用由中国天津大学机器学习与数据挖掘实验室的AI SKYEYE团队收集的VisDrone数据集。其数据由各种不同型号的无人机,在不同场景以及各种天气和照明情况下进行收集,收集了共8 599张图片。VisDrone数据集的类别共有10类,其大约有540 000个标注信息。各个实例的类别分布如图7所示。

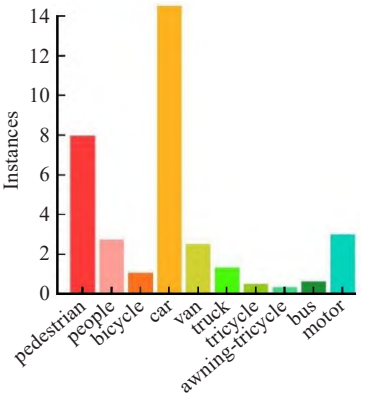


图7　VisDrone数据标签分布图

Fig.7　Label distribution of VisDrone dataset

VisDrone数据集中存在大量的小目标,各个物体的尺寸大小分布如图8所示。从图8中可以看出数据集中物体大部分的长宽比例小于原图像尺寸的0.1倍,满足小目标的相对尺寸定义。

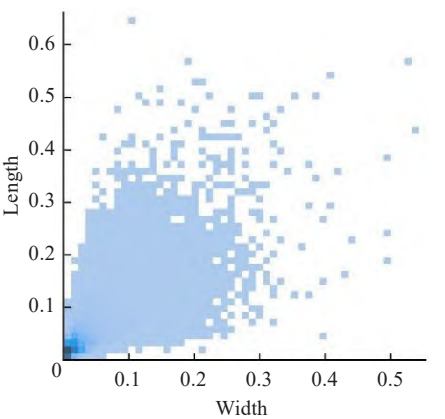


图8　VisDrone数据集长宽分布图

Fig.8　Length and width distribution of VisDrone dataset

VisDrone数据集共分为了三个部分,其中选取了6 471张图片作为训练数据集,548张图片为验证数据集,1 580张图片为测试数据集。VisDrone数据集的实例标注如图9所示,目标尺寸大多小于当前图像的0.1倍,满足小目标的定义,且标注信息计算所得锚框满足

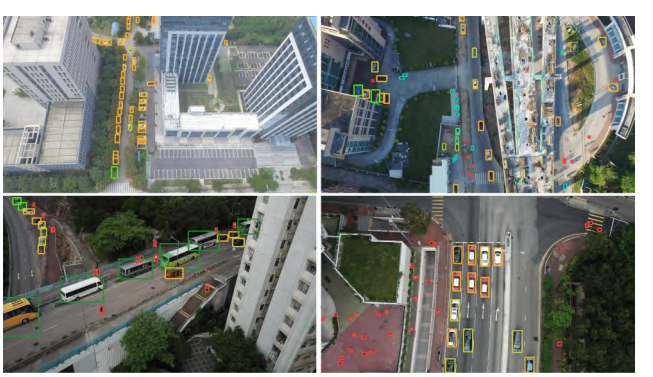


图9　VisDrone数据集示例图

Fig.9　Samples from VisDrone dataset

上文设定的锚框大小。

3.1　评价指标

目标检测中的平均精度均值(mean average precision，mAP)是评价检测性能的重要指标。本文采用了AP50和AP75作为评价指标，分别代表IoU阈值为0.5和0.75时各个类别AP的均值。AP计算的是精确度(precision)和召回率(recall)曲线下的面积，值越大表示模型检测精度越高，精确度、召回率和AP的计算公式如式(9)~(11)所示：

$$precision=\frac{TP}{TP+FP} \tag{9}$$

$$recall=\frac{TP}{TP+FN} \tag{10}$$

$$AP=\int P(r)\mathrm{d}r \tag{11}$$

其中，TP为正确分类的正样本数，FP为误报的负样本数，FN为漏报的正样本数。

3.2　实验环境

本文实验环境运行在Linux系统，使用2个RTX 2080Ti GPU进行训练推理，采用的深度学习框架是PyTorch 1.12，CUDA版本号为11.7。网络的预训练模型权重是在COCO数据集上进行训练得到的。设置的总的训练轮数为300 epochs，学习率为0.01，采用带动量的随机梯度下降法(SGD)作为优化器，动量设置为0.937，权重衰减系数为0.000 5。

3.3　实验结果与分析

3.3.1　检测分支对比实验

本文首先对检测分支数量和位置进行了实验。实验基于YOLOv5s模型进行实验，其中Neck结构为4层，从而引出4个检测分支P2~P5，实验中舍弃分支的方法为减少Head处的检测头不进行回归预测。实验数据为VisDrone数据集，测试图片大小固定为640×640。测试结果表如表2所示。

表2　不同检测分支组合的定量结果

Table 2　Quantitative results of different combinations of detection branches

包含的检测分支	AP50/%	参数量/10^6
P3，P4，P5	34.53	7.64
P2，P3，P4	34.62	7.55
P2，P3，P5	34.52	7.62
P2，P4，P5	34.53	7.63
P2，P3，P4，P5	35.06	7.70

从表2中第1~4项和第5项的对比可以看出包含四个检测头的第5项对比只包含三个检测分支的第1~4项精度均有提升，可以看出多一个检测分支可以提高模型对于尺度变化的鲁棒性，从而提高检测效果。从第1项与第2项的对比也可以看出，将深层检测分支P5替换成浅层检测分支P2能提高对于小目标的检测效果，原因在于P2检测分支拥有更多的浅层特征，包含更多形状、大小和位置等信息，对于小目标的定位更加精确，能有效提升小目标检测效果，并且锚框设定更适合小目标的检测。

综上，为了提高对小目标的检测效果，本文选择使用四个检测分支(P2~P5)，以尽量减少漏检情况的发生。

3.3.2　精简解耦头对比实验

为了验证本文设计精简解耦头(SDHead)的精度与效率，本文设计了与YOLOX的解耦头(DHead)的对比实验。实验数据为VisDrone数据集，测试图片大小固定为640×640，基准模型Baseline为预测头为耦合头的YOLOv5s模型，并依次替换为本文设计精简解耦头和YOLOX的解耦头，实验结果如表3所示。

表3　带不同解耦头的检测器性能比较

Table 3　Performance comparison of detector with different decouple heads

方法	AP50/%	参数量/10^6	FLOPs/10^9	延时/ms
Baseline	32.93	7.04	15.8	6.4
+DHead	34.15(+**1.22**)	7.63	18.2	7.9
+SDHead	34.09(+1.16)	7.15	17.9	7.2

从表3中可以看到本文设计的精简解耦头对比YOLOX模型的解耦头在精度上仅损失了0.06个百分点，但参数量减少了4.8×10^5，浮点运算减少了3.0×10^8，同时推理延时也减少了0.7 ms。实验结果说明了本文的精简解耦头在保持了精度的同时也能减少参数量与推理延时，能够满足实际的运用需求。

3.3.3　改进的损失函数超参数实验

本文改进的损失函数中，α是可以调节的超参数，通过调节α可以调整网络的检测精度，因此本文设计了实验以探索最优的α参数。实验数据为VisDrone数据集，测试图片大小固定为640×640，测试模型为YOLOv5s模型。实验结果如表4所示。

表4　超参数α对Drone-YOLO性能的影响

Table 4　Effect of hyperparameter α on performance of proposed Drone-YOLO

α	AP50/%	AP75/%
1	32.93	22.34
2	33.06	22.45
3	33.14	22.72
4	33.02	22.41
5	32.98	22.37
6	32.91	22.31

改进的损失函数的超参数α设置为1时，即为YOLOv5原本的CIoU损失函数。从表4中可以看出超参数α设置为3时可以达到最优的效果，AP50的提升约0.21个百分点，AP75的提升约0.38个百分点。表4中α在1到3范围内时精度值呈上升趋势，当α开始大于3时，精度值出现了下降的趋势，综上考虑本文的α超参数设定为3，以获取最优的检测效果。

3.3.4　综合对比实验

最后,为了综合测试本文模型的检测效果,将Drone-YOLO与基准模型YOLOv5s进行对比的同时,与主流模型进行了对比。模型的训练数据是VisDrone数据集,其最终测试结果是在VisDrone测试数据集上进行验证的,实验结果如表5所示。

表5　VisDrone 测试数据集实验结果

Table 5　Experiment results in VisDrone test dataset

方法	AP50/%	AP75/%	延时/ms
Light-RCNN[21]	39.56	23.24	52.14
Cascade-RCNN[22]	37.84	22.56	57.32
RetinaNet[23]	31.67	20.18	48.65
CornerNet[24]	41.18	25.02	36.45
YOLOX[20]	53.51	31.41	18.72
TPH-YOLOv5[10]	59.88	38.69	20.12
QueryDet[11]	48.14	28.75	25.55
YOLOv5	55.33	33.06	15.43
Ours	**60.24**	**39.91**	16.78

从实验结果表5中可以看出,与其他模型对比,本文模型Drone-YOLO的检测精度是最高的,且推理延时只为16.78 ms,可以满足实际的运用需求。本文模型与基准模型YOLOv5进行对比,在AP50上提升了4.91个百分点,AP75提升较为明显,提升了6.85个百分点。对比近年的针对小目标的检测器TPH-YOLOv5和QueryDet仍然保持了检测精度和速度上的优势。

为了验证分析模型对于小目标的检测效果以及实际的运用需求,本文测试了数据集部分类别的AP值,其测试数据集为VisDrone测试数据集,其实验结果如表6所示。

表6　VisDrone 测试数据集各类别的AP50结果

Table 6　AP50 experiment results of each category in VisDrone test dataset

单位:%

方法	pedestrian	person	bicycle	car
Light-RCNN[21]	17.02	4.83	5.73	32.29
Cascade-RCNN[22]	16.28	6.16	4.18	37.29
RetinaNet[23]	9.91	2.92	1.32	28.99
CornerNet[24]	20.43	6.55	4.56	40.94
YOLOX[20]	23.67	11.62	14.86	54.18
TPH-YOLOv5[10]	27.52	15.32	19.26	59.32
QueryDet[11]	20.56	8.72	6.64	45.51
YOLOv5	23.71	12.77	15.13	56.21
Ours	**28.63**	**16.75**	**19.64**	**59.53**

从表6中可以看出,本文模型在检测行人(pedestrian)、人(person)和自行车(bicycle)等小目标的效果上提升明显,对比基准模型YOLOv5,分别提升了4.92、3.98和4.51个百分点。并且在检测车(car)等体型较大的目标上的效果也有所提升,提升了3.32个百分点。

综上所述,本文Drone-YOLO模型对比基准模型YOLOv5提升明显,尤其是针对行人等小目标上的检测效果,从结果中可以看出本文改进的有效性。

3.3.5　改进模块的消融实验

为了更好地说明本文改进的模块与方法对于模型检测能力的提升,以及对于模型参数量、每秒浮点运算数(FLOPs)和推理时延的影响,本文进行了消融实验。本消融实验是在YOLOv5s模型上逐个添加本文改进模块与方法所得出的实验结果,首先添加新增的检测分支P2,然后添加多尺度注意力的特征融合模块MAFF,然后将网络的颈部结构替换为多层次信息聚合的特征金字塔网络结构ML-FPN,然后将耦合的预测头替换为精简的解耦头SD-Head,最后用Alpha-IoU优化损失函数。测试图片大小固定为640×640,在2个RTX 2080Ti GPU上进行推理。消融实验的测试结果基于VisDrone测试数据集,实验结果如表7所示。

表7　VisDrone 测试数据集消融实验

Table 7　Ablation study of proposed method on VisDrone test dataset

方法	AP50/%	参数量/10^6	FLOPs/10^9	延时/ms
Baseline	32.93	7.04	15.8	6.4
+P2分支	35.06(+**2.13**)	7.70	27.0	7.9
+MAFF	36.09(+1.03)	7.84	29.2	8.7
+ML-FPN	37.72(+**1.63**)	8.21	37.3	9.8
+SD-Head	38.87(+1.15)	8.33	39.4	10.6
+Alpha-IoU	39.08(+0.21)	8.33	39.4	10.6

表7中从上到下依次添加了本文改进的模块或方法,从消融实验结果表7上可以看出,本文设计的组件对模型检测小目标的准确度均有提升。首先,实验结果对比中新增检测分支P2的AP50指标的提升最为明显,提升了2.13个百分点。这说明了新增检测分支对小目标检测的有效性,也说明了新增P2检测分支锚框设定为小目标的尺寸,可以极大减少由于锚框设定过大而导致的漏检情况。其次,ML-FPN结构提升了AP50指标1.63个百分点,这说明多层次的信息融合,特别是充分利用浅层的形状大小信息可以提高对于小目标的定位效果,从而提高对小目标的检测效果。最后,本文模型添加的MAFF模块提升了AP50指标1.03个百分点,这说明了MAFF模块提升了对小目标的关注度,融合了多尺度信息,提升了网络的检测效果。精简解耦头对精度提升了1.15个百分点,这说明了分类任务和回归任务之间存在差异性,解耦两类任务可以提高目标检测的精度。虽然Alpha-IoU方法对AP50的提升不大,只有0.21个百分点,但是Alpha-IoU不增加网络参数和推理时延,且如图10所示,在训练过程中添加了Alpha-IoU方法的模型(实线)收敛较基准模型(虚线)更快。因此加入Alpha-IoU可以加速网络收敛,减少网络的训练时间。

3.4　可视化分析

从上述的实验结果表中可以看出本文改进模型Drone-YOLO在小目标的检测效果上优于YOLOv5和

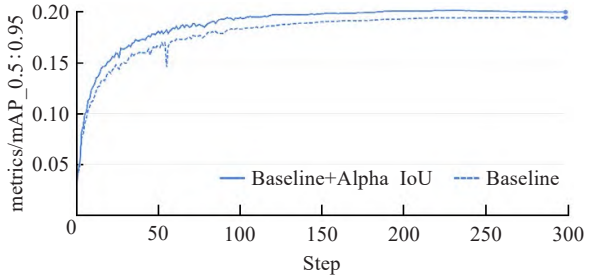


图10　mAP 0.5:0.95训练结果图

Fig.10　Training mAP curves with different method in mAP 0.5:0.95

其他主流的目标检测模型。图11展示了本文模型Drone-YOLO和基准模型YOLOv5的检测效果对比图。

(a)结果1

(b)结果2

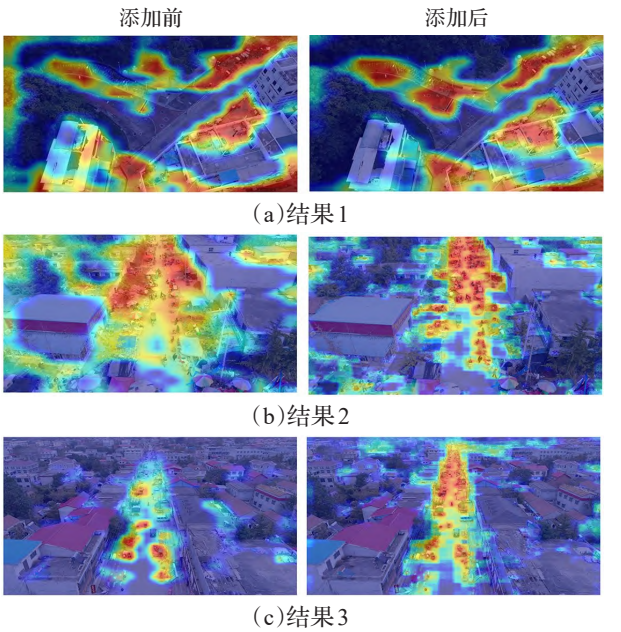
(c)结果3

图11　Drone-YOLO和YOLOv5检测结果对比图

Fig.11　Visual results comparisons between Drone-YOLO and YOLOv5

从图11的对比中可以看出本文模型Drone-YOLO比基准模型YOLOv5检测出了更多的物体，尤其是行人、自行车等小目标物体，并且对于同一个检测物的置信度也有所提高。图11(a)的对比中可以看出，物体非常小的时候，YOLOv5忽略掉了当前的行人目标，出现了漏检情况，而Drone-YOLO没有出现漏检行人的情况。同时，图11(b)的对比图中YOLOv5出现了误检情况，将黑色的棚子误判为了卡车，而Drone-YOLO没有误检。图11(b)和(c)的对比图中可以看处，对于远处的小目标，YOLOv5模型漏检了许多物体，而本文模型Drone-YOLO都可以检测出来，拥有良好的检测效果。

为了验证和分析本文设计的多尺度注意力机制的特征融合模块MAFF对于小目标检测效果提升的原因，本文采用Eigen-CAM[25]对本文模型添加模块前后进行热力图的分析和对比，对比分析如图12所示。从图12(a)可以看出，未添加MAFF模块的YOLOv5忽略掉了行人目标，关注到了无用的位置，添加MAFF模块后，模型关注到了行人目标，同时覆盖位置更加精准，降低了对不感兴趣目标的注意力。图12(b)和(c)对比图中可以看出来，添加MAFF模块后模型关注更加细致，同时也能够关注到远处的较小目标。



(a)结果1

(b)结果2

(c)结果3

图12　有/无MAFF模块的特征可视化

Fig.12　Feature visualization with/without MAFF module

为了测试本文模型对于极端场景的检测效果，对于一些常见的场景的进行了对比测试。如图13所示，图13第一行图中人较为模糊，YOLOv5漏检了行人和摩托目标，而Drone-YOLO则没有出现漏检情况。图13第二行图中存在遮挡和暗光情况，Drone-YOLO对比YOLOv5拥有更好的表现。

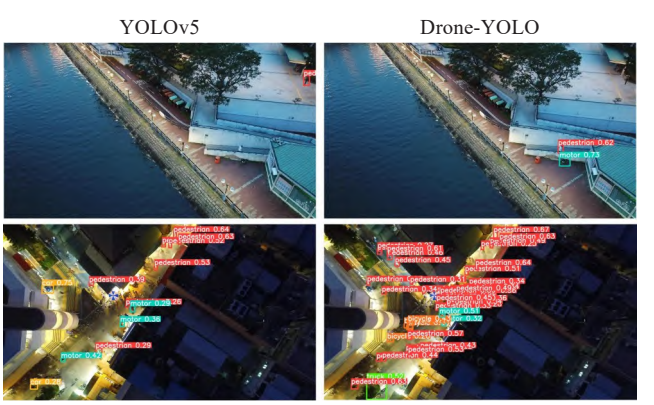


图13　困难场景下检测结果图

Fig.13　Detection results in difficult scenario

4　结束语

针对无人机场景下现有的目标检测器对于小目标检测效果差，存在误检漏检的问题，本文基于YOLOv5提出了改进的检测模型Drone-YOLO。首先，通过增加检测分支以及融合多层次、多尺度信息以适应无人机场

景下小目标的尺寸变化,其次采用了多尺度注意力模块以提升网络对物体的关注度,最后将分类任务与回归任务解耦以提升检测精度,优化损失函数以提升训练效率。在无人机数据集VisDrone的实验结果表明,本文Drone-YOLO模型对于小目标的检测效果优于其他主流模型,检测精度达到了60.24%,且检测时间只需16.78 ms,能够有效完成无人机场景下的小目标检测任务。

后续研究将继续进行网络结构的优化,采用模型剪枝或者知识蒸馏等方式来减少参数量,以提升模型在算力有限的情况下的部署与应用。

参考文献:

[1] HE J,ERFANI S,MA X,et al.Alpha-IoU:a family of power intersection over union losses for bounding box regression[C]//Advances in Neural Information Processing Systems,2021:20230-20242.

[2] KISANTAL M,WOJNA Z,MURAWSKI J,et al.Augmentation for small object detection[J].arXiv:1902.07296,2019.

[3] CHEN C,ZHANG Y,LV Q,et al.RRNet:a hybrid detector for object detection in drone-captured images[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.Los Alamitos:IEEE,2019:100-108.

[4] YU X,GONG Y,JIANG N,et al.Scale match for tiny person detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.Los Alamitos:IEEE,2020:1257-1265.

[5] CHEN Y,ZHANG P,LI Z,et al.Stitcher:feedback-driven data provider for object detection[J].arXiv:2004.12432,2020.

[6] LIN T Y,DOLLÁR P,GIRSHICK R,et al.Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:2117-2125.

[7] DENG C,WANG M,LIU L,et al.Extended feature pyramid network for small object detection[J].IEEE Transactions on Multimedia,2021,24:1968-1979.

[8] 李青援,邓赵红,罗晓清,等.注意力与跨尺度融合的SSD目标检测算法[J].计算机科学与探索,2022,16(11):2575-2586.
LI Q Y,DENG Z H,LUO X Q,et al.SSD object detection algorithm with attention and cross-scale fusion[J].Journal of Frontiers of Computer Science and Technology,2022,16(11):2575-2586.

[9] 梁延禹,李金宝.多尺度非局部注意力网络的小目标检测算法[J].计算机科学与探索,2020,14(10):1744-1753.
LIANG Y Y,LI J B.Small objects detection method based on multi-scale non-local attention network[J].Journal of Frontiers of Computer Science and Technology,2020,14(10):1744-1753.

[10] ZHU X,LYU S,WANG X,et al.TPH-YOLOv5:improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2021:2778-2788.

[11] 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.

[12] JADERBERG M,SIMONYAN K,ZISSERMAN A.Spatial transformer networks[C]//Advances in Neural Information Processing Systems,2015.

[13] DAI J,QI H,XIONG Y,et al.Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:764-773.

[14] HU J,SHEN L,SUN G.Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:7132-7141.

[15] WANG Q,WU B,ZHU P,et al.ECA-Net:efficient channel attention for deep convolutional neural networks[J].arXiv:1910.03151,2019.

[16] HOU Q,ZHOU D,FENG J.Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021:13713-13722.

[17] WOO S,PARK J,LEE J Y,et al.Cbam:convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision(ECCV),2018:3-19.

[18] FU J,LIU J,TIAN H,et al.Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019:3146-3154.

[19] HUANG Z,WANG X,HUANG L,et al.Ccnet:criss-cross attention for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision,2019:603-612.

[20] GE Z,LIU S,WANG F,et al.Yolox:exceeding yolo series in 2021[J].arXiv:2107.08430,2021.

[21] LI Z,PENG C,YU G,et al.Light-head R-CNN:in defense of two-stage object detector[J].arXiv:1711.07264,2017.

[22] 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.

[23] LIN T Y,GOYAL P,GIRSHICK R,et al.Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:2980-2988.

[24] LAW H,DENG J.Cornernet:detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision(ECCV),2018:734-750.

[25] MUHAMMAD M B,YEASIN M.Eigen-cam:class activation map using principal components[C]//2020 International Joint Conference on Neural Networks(IJCNN),2020:1-7.