基于特征融合与注意力机制的无人机图像小目标检测算法

李利霞[1]，王鑫[2,1,3]，王军[3]，张又元[4]

(1. 桂林电子科技大学计算机与信息安全学院，广西 桂林 541010；
2. 电子科技大学信息与软件工程学院，四川 成都 610000；
3. 桂林电子科技大学海洋工程学院，广西 北海 536000；
4. 兰州交通大学电子与信息工程学院，甘肃 兰州 730070)

摘　　要：由于无人机航拍图像目标物体尺寸太小、包含的特征信息少，导致现有的检测算法对小目标的检测效果不理想。针对该问题，在 YOLOv5 主干网络中融入多头注意力机制，可以有效整合全局特征信息。随着网络深度的不断加深，模型将更关注高层的语义信息，进而忽略对小目标检测至关重要的底层细节纹理特征，以致小目标的检测效果较差。因此，提出浅层特征增强模块来学习底层特征信息，达到增强小目标特征信息的目的。此外，为了加强特征融合的能力，设计了一种多级特征融合模块，将不同层级的特征信息进行聚合，使网络能够动态调节各输出检测层的权重。实验结果表明，该算法在公开数据集 VisDrone2021 平均均值精度达到 45.7%，相比原 YOLOv5 算法提升了 3.1%，对高分辨率图像的检测速度 FPS 达到 41 帧/秒，满足实时性，与其他主流算法相比该算法检测精度有明显提升。

关　键　词：特征融合；注意力机制；无人机航拍图像；小目标检测；YOLOv5

中图分类号：TP 391　　**DOI**：10.11996/JG.j.2095-302X.2023040658

文献标识码：A　　**文 章 编 号**：2095-302X(2023)04-0658-09

Small object detection algorithm in UAV image based on feature fusion and attention mechanism

LI Li-xia[1], WANG Xin[2,1,3], WANG Jun[3], ZHANG You-yuan[4]

(1. School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin Guangxi 541010, China;
2. School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan 610000, China;
3. School of Marine Engineering, Guilin University of Electronic Technology, Beihai Guangxi 536000, China;
4. School of Electronics and Information Engineering, Lanzhou Jiaotong University, Lanzhou Gansu 730070, China)

Abstract: The task of detecting small objects in UAV aerial images is a formidable challenge due to their diminutive size and insufficient amount of feature information. To surmount this predicament, a multi-head attention mechanism was incorporated into the YOLOv5 backbone network in order to seamlessly integrate global feature information. As

收稿日期：2022-11-18；**定稿日期**：2023-01-18

Received：18 November，2022；**Finalized**：18 January，2023

基金项目：广西科技重大专项(AA19254016)；广西硕士研究生创新项目(YCSW2021174)；北海市科技规划项目(202082033；202082023)

Foundation items：Guangxi Science and Technology Major Project (AA19254016); Guangxi Graduate Student Innovation Project (YCSW2021174); Beihai City Science and Technology Planning Project (202082033; 202082023)

第一作者：李利霞(1995–)，女，硕士研究生。主要研究方向为图像处理和物体识别。E-mail：20032202019@mails.guet.edu.cn

First author：LI Li-xia (1995–), master student. Her main research interests cover image processing and object recognition. E-mail：20032202019@mails.guet.edu.cn

通信作者：王鑫(1976–)，男，教授，博士。主要研究方向为图像处理、网络信息安全、物联网和数据挖掘等。E-mail：304379506@qq.com

Corresponding author：WANG Xin (1976–), professor, Ph.D. His main research interests cover image processing, network information security, internet of things, data mining and other research, etc. E-mail：304379506@qq.com

the network depth increased, the model tended to accentuate high-level semantic information at the expense of underlying detailed texture features vital for the detection of small objects. To address this issue, a shallow feature enhancement module was devised to acquire underlying feature information and augment small object feature information. Furthermore, a multi-level feature fusion module was developed to amalgamate feature information from different layers, thus enabling the network to dynamically adjust the weights of each output detection layer. Experimental results on the publicly available VisDrone2021 dataset demonstrated that the mean average precision of the proposed algorithm, attained a level of 45.7%, representing a 3.1% enhancement over the baseline YOLOv5 algorithm. Additionally, the proposed algorithm achieved a detection speed of 41 frames per second for high-resolution images, satisfying the requirement for real-time performance and exhibiting a noteworthy improvement in detection accuracy over other prevalent methods.

Keywords: feature fusion; attention mechanism; UAV aerial imagery; small object detection; YOLOv5

随着无人机应用场景的不断拓展，无人机航拍图像小目标检测算法引起研究人员的关注。由于无人机具有成本低、使用方便、能实现高分辨率影像采集等特点，被广泛应用于遥感图像、农业、抢险救灾、视频拍摄、灾情监视、工业目标检测等行业。无人机航拍图像中小目标的占比大，能够提供的分辨率有限[1]，导致无人机航拍图像的检测极具挑战性。

目标检测作为深度学习中的先行任务有着重要的研究意义[2]。目前主流的目标检测算法可以分为 2 类：①一阶段目标检测算法，主要有只用你看一次(you only look once，YOLO)[3]，和单激发多框探测器(single shot multibox detector，SSD)[4]算法，这类算法不使用候选框，可直接预测类别数据和位置数据，并在单次检测中得到最终结果，其优点是速度快；②二阶段目标检测算法，主要有基于区域的卷积神经网络(region-based convolutional neural networks，R-CNN)[5-6]和二阶段目标检测算法 D2Det[7]等。该算法在第一阶段创建候选框，在第二阶段对目标物体进行分类和回归，该类算法的检测精度较优，但检测速度较慢。针对小目标检测效果不理想，ZHAN 等[8]在 YOLOv5 的基础上增加检测层，该方法虽然有效提高了小目标的检测精度，但增加了模型的复杂度，导致检测速度变慢。LIM 等[9]提出了一种基于上下文与注意力的小目标检测算法，其更关注图像中的小物体，且在一定程度上提升了小目标的检测能力，但检测精度仍有待提高，在现实场景中无法应用于无人机航拍图像的检测。SONG 等[10]基于多尺度特征融合的小目标检测算法，其利用特征金字塔的思想，能更好地表达深层网络的语义特征和浅层网络的位置信息，从而提升对小目标的预测能力。但该方法存在一个问题，即骨干网的加深可导致网络参数量和计算量变大，从而使得模型的检测速度变慢。LIU 等[11]提出多分支并行金字塔网络结构，同时引入一种有监督的空间注意力模块来减弱背景噪声干扰，聚焦目标信息。尽管该方法对小目标检测效果有一定的改善，但仍存在检测精度较低的问题。胡俊等[12]将多模态技术应用到遥感图像的模型，并采用 RGB 和 IR (infra-red)图像互补增强特征信息的方法，能够提升模型对小目标检测的效果。由于网络需要对 2 种模态的图像特征信息进行融合，以致网络的计算速度受到影响。

基于上述文献的不足之处，本文提出一种基于特征融合与注意力机制的无人机图像小目标检测算法。该算法在 YOLOv5 的主干网络中融入多头注意力机制以聚焦有用的图像特征；通过融合高分辨率特征图来充分学习底层特征中的细节信息，再将融合后的特征图输出到检测层；为了加强网络对图像特征的提取，设计多级特征融合模块将颈部不同特征层的权重进行合理分配，以提高小目标的检测能力。

1　YOLOv5 网络结构

YOLOv5 是当前主流的目标检测算法之一，该算法结构简单、计算高效，具备良好的检测效果，并可同时兼顾检测精度和速度。因此，本文选择 YOLOv5 作为基线网络，在此基础上进行算法的改进，提高对小目标的检测能力。

YOLOv5 网络结构主要由 4 部分组成(图 1)：①首先通过 Mosaic 数据增强对图像数据进行预处理，并采用自适应锚框来计算不同训练集中的最佳锚框值；②主干网络主要包括 CBS 和 C3 结构，该结构能够从输入图像中提取丰富的特征信息；③颈

部将特征金字塔网络(feature pyramid networks，FPN)[13]与路径聚合网络(path aggregation network，PAN)[14]相结合，FPN 结构是自顶向下，将高层的强语义特征传递下来，以增强语义信息。PAN 结构是自底向上对 FPN 进行补充，将底层的特征信息进行传递。这 2 种方法的结合可加强网络特征整合的能力；④头部检测器，利用基于网格的锚框在不同尺度的特征图上进行结果的预测。

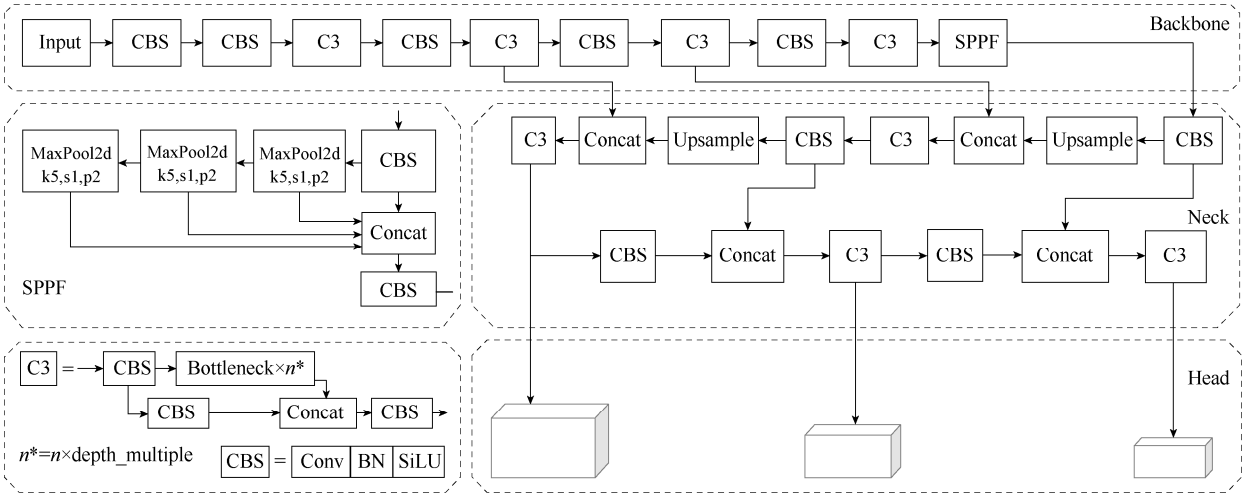


图 1　YOLOv5 网络结构

Fig. 1　YOLOv5 network structure

2　本文算法

为了提升无人机航拍图像小目标的检测性能，本文提出一种基于特征融合与注意力机制的无人机图像小目标检测算法，即 BA-YOLOv5s 算法，其网络结构如图 2 所示。为了保留有用的信息，抑制不重要的信息，本文在主干网络中融入多头注意力机制，即 BT-MHSA (bottleneck transformer muti-head self-attention)模块，该模块可以挖掘到更多有效的特征信息。为了更关注底层细节纹理信息，将浅层主干网络的高分辨特征图和颈部的 FPN 进行融合，再将融合后的特征信息输入到检测层，提升对小目标检测的预测能力。通过设计多级特征融合模块，将头部特征层的权重进行动态调整，从而提高无人机航拍图像小目标的预测能力。

2.1　主干网络的改进

在无人机航拍图像中，复杂环境容易造成目标之间存在的遮挡问题，不利于网络对图像特征的提取。主干网络主要进行图像特征信息的提取，但提取到的特征信息中存在无效信息，最终影响模型的检测效果。因此需要过滤掉图像中的噪声，将更多注意力集中在有用的特征信息上。

随着计算机视觉的发展，注意力机制和卷积操作在视觉任务上取得了巨大的成功。起初 VASWANI 等[15]提出自注意力机制(Self-attention)，以减少对外部信息的依赖，且更加擅长捕捉特征的内部相关性，并将其用在 Transformer 模块中，能够加强对特征的选择，极大地提升了模型的性能。为了兼顾卷积与自注意力的优点，PAN 等[16]提出一个混合模型 ACmix，将卷积与自注意力的优势进行结合。首先将输入特征经过 1×1 卷积映射，获得丰富的中间特征，接着将中间特征进行不同方式的聚合，从而有效发挥二者优势，且不用计算 2 次。为了进一步揭示自注意力与卷积的潜在关系，SRINIVAS 等[17]提出 BoTNet (bottleneck transformer network)结构，将卷积与多头注意力进行结合，用于提取丰富的图像细节信息与上下文信息，建立长距离依赖关系。卷积操作可以有效地捕获局部信息，多头注意力机制(multi-head self-attention，MHSA)是一种全局操作，可以将卷积操作获取的包含特征信息的特征图进行聚合，MHSA 具有捕获大范围图像特征信息的能力。基于此，本文将 BoTNet 结构进行改进，将 Bottleneck 模块改进为 MHSA 模块，即 BT-MHSA 模块。YOLOv5 的主干网络中融入多头注意力模块，如图 3 所示。即将 Bottleneck 中 3×3 的卷积改进成 MHSA，可以在目标检测中建立长距离依赖关系，而且对于航拍图像小目标检测表现出更好的检测效果，解决了航拍图像中的无效信息引起检测效果不佳的问题，且能更高效地提取图像特征信息。

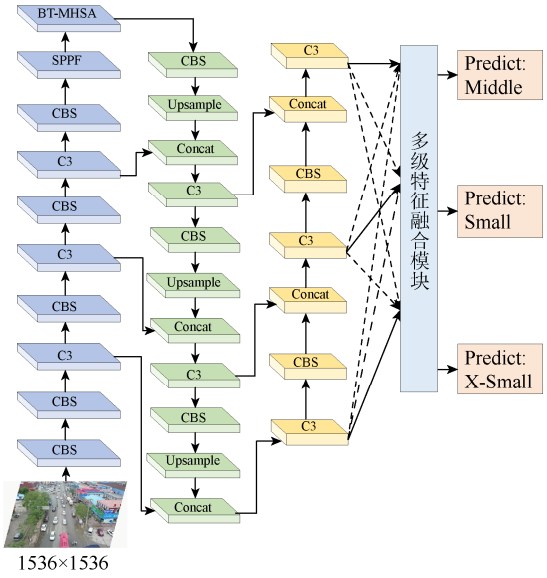


图 2　BA-YOLOv5s 网络结构图

Fig. 2　BA-YOLOv5s network structure

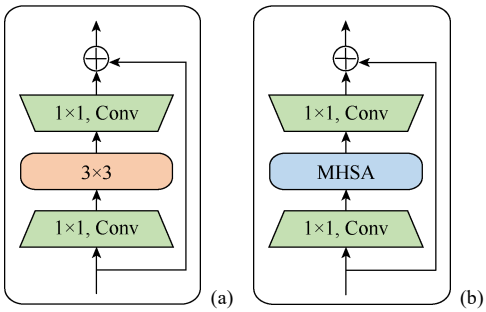


图 3　主干网络中融入多头注意力模块(a) Bottleneck 模块；(b) BT-MHSA 模块)

Fig. 3　Into the multi-attention module ((a) The Bottleneck module; (b)The BT-MHSA module)

MHSA 结构如图 4 所示，使用了 4 个注意力头，并且采用相对位置编码 R_h 和 R_w 以及 1×1 逐点卷积，使用逐点卷积的好处是可大幅降低模型的参数量和计算量。相对位置编码的注意力操作能够获得位置感知，在考虑了内容信息同时，还考虑了特征中不同位置像素之间的距离，从而能够有效地将目标之间的信息与位置感知相关联。

2.2　浅层特征增强模块

在无人机 VisDrone2021[18]数据集中，小目标占绝大多数。图 5 展示了数据集中目标物体的宽高分布，图中坐标原点附近颜色最深，说明数据集中小目标的数量最多，与本文的研究相符。在无人机的拍摄场景中，检测算法对小目标的检测效果不如大中目标。由于数据集中的小目标过多，使得目标检测算法无法充分发挥其能力。

由于无人机图像中小目标众多，主干网络经多次下采样后，底层细节信息容易被忽视，导致特征图保留更多的高层语义信息。FPN 自顶向下将顶层特征和底层细节进行融合，造成信息的冲突，使网络易忽略图像中的细节信息，最终使得小目标预测效果不佳。为了解决该问题，将 FPN 和主干网络的浅层特征图的细节纹理信息进行融合。浅层特征融合模块使得网络更关注底层特征信息，最终使得模型的检测效果提升。底层特征图分辨率更高、感受野更小，对小目标的表征能力更强，有利于检测小目标。在没有额外增加目标检测层的条件下，提高了对无人机航拍图像小目标的检测能力。

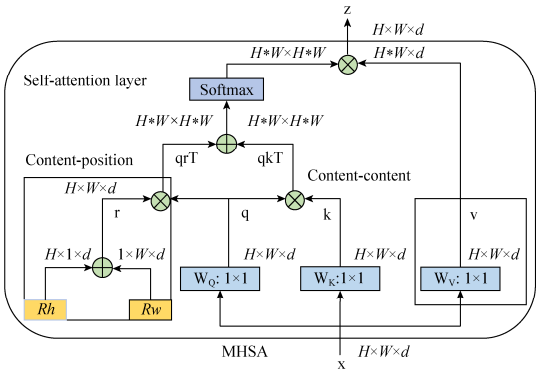


图 4　多头自注意力机制

Fig. 4　Multi-head self-attention mechanism

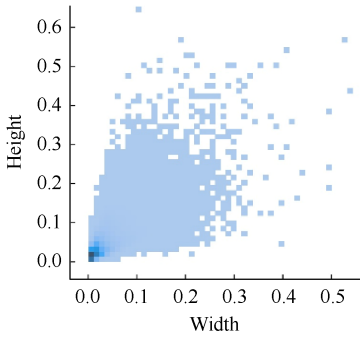


图 5　数据集中的目标的宽高分布

Fig. 5　Width and height distribution of objects in the dataset

2.3　多级特征融合模块

由于浅层网络包含更多的细节信息，对小目标检测具有重要的作用，但浅层特征包含的语义信息不足。为了解决此问题，FPN 将相邻特征层以自上而下的方式和横向连接顺序组合。PAN 和 FPN 结构，主要为了整合不同特征层间的特征信息。但在特征整合中忽略了不同特征层之间的关系，浅层更有益于小目标的检测，高层更适合大中目标的检测，因此每个检测头检测的目标尺度是不同的。

针对上述问题，多级特征融合模块将不同尺度特征层的权重进行合理分配，能够直接学习如何将其他层的特征进行空间过滤，优化不同特征层的权

重比例，以便只保留有用的信息。对于每一层，所有其他特征层的大小都被调整为相同分辨率的特征图，通过对网络模型的学习能够动态调整特征图的大小，最后找到最佳融合方式。在本文中，为了优化特征融合的能力，在 YOLOv5 模型的检测层中设计多级特征融合模块，也解决了 FPN 尺度不一致问题。多级特征融合模块主要分为调整多层特征图的大小和特征融合。

(1) 调整多层特征图的大小。颈部输出不同特征层 x^l，其中 l=1,2,3 表示不同特征层，本文将其他层 n (n 不等于 l)的特征 x^n 调整为与 x^l 相同的形状。不同的特征层具有不同的分辨率和通道数，通过上采样和下采样的策略相应地修改特征层的尺寸。对于上采样，首先使用 1×1 的卷积将特征的通道数压缩到第 l 级，然后采用线性插值来提升分辨率。对于 1/2 比例的下采样，只需使用步长为 2 的 3×3 卷积，同时修改通道数和分辨率。对于 1/4 比例的下采样，在步长为 2 的卷积前添加步长为 2 的最大池化层，通过这种方法统一特征图，为下一步特征融合做准备。

(2) 特征融合。令 $\boldsymbol{x}_{ij}^{n\to l}$ 表示从第 n 级调整到第 l 级的特征图上位置(i,j)处的特征向量。将相应级别 l 的特征进行融合，得到

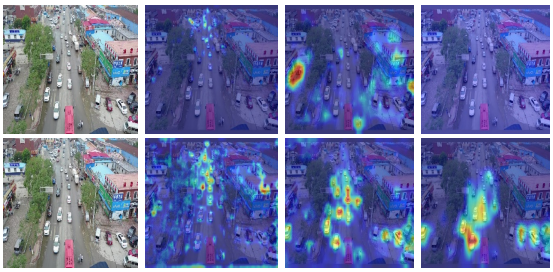
$$\boldsymbol{y}_{ij}^{l}=\alpha_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{1\to l}+\beta_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{2\leftarrow l}+\gamma_{ij}^{l}\cdot\boldsymbol{x}_{ij}^{3\to l} \tag{1}$$

其中，$\boldsymbol{y}_{ij}^{l}$ 为通道间输出特征映射 $\boldsymbol{y}^{l}$ 的第(i,j)的向量；α_{ij}^{l}，β_{ij}^{l} 和 γ_{ij}^{l} 分别为特征图上 3 个不同级别 l 空间权重的重要性，权重由网络自适应学习。注意 α_{ij}^{l}，β_{ij}^{l} 和 γ_{ij}^{l} 可以是简单的标量变量，且可在所有通道之间共享，其中 $\alpha_{ij}^{l}+\beta_{ij}^{l}+\gamma_{ij}^{l}=1$，$\alpha_{ij}^{l},\beta_{ij}^{l},\lambda_{ij}^{l}\in[0,1]$，可定义为

$$\alpha_{ij}^{l}=\frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}}+e^{\lambda_{\beta_{ij}}^{l}}+e^{\lambda_{\gamma_{ij}}^{l}}} \tag{2}$$

其中，α_{ij}^{l}，β_{ij}^{l} 和 γ_{ij}^{l} 可用 Softmax 函数来定义，且 $\lambda_{\alpha_{ij}}^{l}$，$\lambda_{\beta_{ij}}^{l}$ 和 $\lambda_{\gamma_{ij}}^{l}$ 分别为控制参数。使用 1×1 卷积分别从 $x^{1\to l}$，$x^{2\to l}$ 和 $x^{3\to l}$ 计算权重标量映射 λ_{α}^{l}，λ_{β}^{l}，λ_{γ}^{l}，因此可以通过标准反向传播来学习。使用这种方法可以将不同级别的特征在每个尺度上自适应的聚合。

在模型训练中，网络将学习到的权重参数合理分配给检测层，通过动态调节不同检测层权重的方式，增强对无人机航拍图像小目标的预测能力。为了更直观地展示多级特征融合模块的效果，图 6 展示了不同特征层之间的热力图，在图的第 1 行表示未加入，第 2 行表示加入了多级特征融合的热力图。图 6(b)~(d)表示来自颈部的 3 个特征图，分别表示浅层、中间层和高层语义信息。对比图 6 第 1 行与第 2 行可发现，图 6(b)第 1 行存在细节纹理信息被忽略，最终导致小目标的检测效果不理想，而第 2 行网络对浅层小目标的关注度较高。图 6(d)第 2 行通过动态调整权重，能够减少对冗余信息的关注。因此，本文采用多级特征融合模块可以将不同特征层进行动态调整，多级特征融合模块使得检测层兼顾丰富的语义信息和细节信息，从而提升目标检测的效果。



(a) (b) (c) (d)

图 6 多级特征融合的热力图

Fig. 6 Heat map of multi-level feature fusion ((a) Input; (b) Level-1; (c) Level-2; (d) Level-3)

3 实验

实验环境为 Intel Core i5-12500H 处理器，Ubuntu 18.04 操作系统，使用 NVIDIA Tesla T4、显存为 16 G 的 GPU 并行加速网络进行实验操作。本文 FPS 均在 RTX3070、显存为 8 G 的 GPU 并行加速网络上操作，所使用的编程语言为 Python3.7，使用 PyTorch1.9.1 版本的深度学习框架。

3.1 数据集

本文实验采用 VisDrone2021 数据集进行训练、验证、测试，该数据集在国内 14 个不同的城市进行拍摄并标记，共 8 629 张图像。该数据集的分配比例为训练集 6 471 张、验证集 548 张、测试集 1 610 张。图像种类包括：行人、自行车、人、汽车、卡车、敞篷三轮车、三轮车、面包车、摩托车、公交车 10 类。

3.2 参数设置与评价指标

首先本文在 YOLOv5s 模型进行改进，输入图像的分辨率为 1536×1536，数据增强采用 Mosaic，训练 100 个 epoch，初始学习率为 0.01，优化器为 SGD，采取 mAP50，精确度(precision，P)、召回率

(recall，R)、FPS 推理时间等作为模型性能的评价指标。本实验使用的平均精度均值(mean average precision，mAP)指在多类目标检测中，根据每个类的精确度和召回率绘制的 P-R 曲线。精确率和召回率的计算如下

$$P=\frac{TP}{TP+FP}\times 100\% \tag{3}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{4}$$

其中，*TP*(true positive)为预测为正样本的正样本数；*FP*(false positive)为预测为正样本的负样本数；*FN*(false negative)为预测为负样本的正样本数。P-R 曲线与坐标轴相交得到的面积就是平均精度，平均精度(average precision，AP)即

$$AP=\int_0^1 P(R)\,\mathrm{d}R \tag{5}$$

$$mAP=\sum_{i=1}^{N}\frac{AP(i)}{N} \tag{6}$$

3.3　实验分析

3.3.1　网络模型的选择

本文针对 YOLOv5 的 4 种模型规格进行实验，由小到大分别有 N，S，M 和 L 版本，其网络结构是一致的，但网络模型的深度和宽度不同，所得实验结果见表 1。Params 表示模型的参数量，Depth 用于控制模型深度，Width 用于控制模型宽度，GFLOPs 指模型每秒亿次的浮点运算操作，mAP 是指 IoU 阈值为 0.5 时，所有目标类别的平均检测精度。FPS 是网络模型每秒可以处理图像的数量，衡量模型的检测速度，FPS^{1536} 表示在 1536×1536 的高分辨率下进行模型的测试。通过实验数据分析可知，随着网络深度与宽度的增加，模型的 mAP 逐渐提升，但模型的 Params 及 GFLOPs 也逐渐增大，使得 FPS 逐渐降低。由此可见，网络规模的增加会使得网络的复杂度增高，从而导致模型的实时性变差。表中 YOLOv5n 的检测速度最快，深度与宽度最小，但 mAP 最差。而 YOLOv5l 的 mAP 最高，但检测速度最慢。相对于其他模型，YOLOv5s 的 mAP 较好且检测速度较快，能较好地做到精度与速度的平衡。因此，本文选择轻量化的 YOLOv5s 模型作为基线网络以提升检测效果。

3.3.2　消融实验

为了证明本文算法的性能，消融实验在 VisDrone 数据集中进行。模型的训练结果如图 7 所示，从图中可知，改进后的模型随着迭代次数的增加趋于平稳，说明模型的训练过程正常。其中，YOLOv5s 为基线网络，M1 为浅层特征增强模块，M7 为改进后的算法的训练结果图。

表 1　测试集上 YOLOv5 不同模型规格结果的比较

Table 1　Comparison of YOLOv5 model specification results on the test set

Model	Params (M)	Depth	Width	GFLOPs	mAP (%)	FPS^{1536} (帧/秒)
YOLOv5n	1.777	0.33	0.25	4.3	32.4	**81**
YOLOv5s	7.037	0.33	0.50	15.8	42.6	56
YOLOv5m	20.889	0.67	0.75	48.0	46.1	32
YOLOv5l	46.157	1.00	1.00	107.8	**48.2**	19

注：加粗数据为最优值

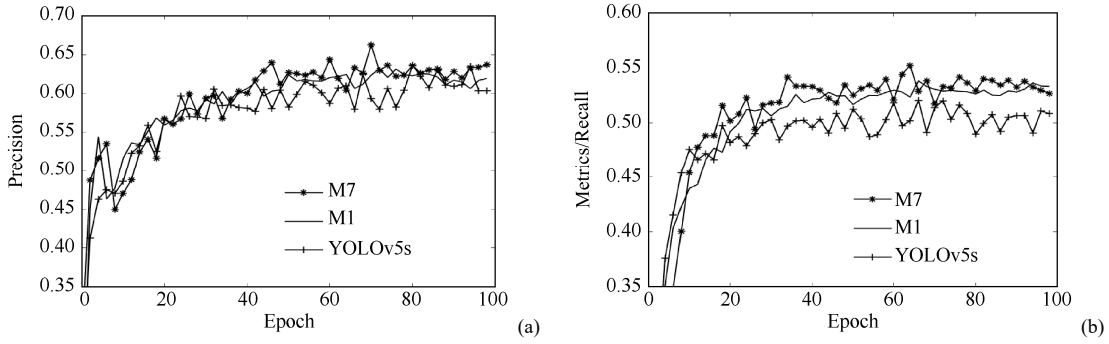


图 7　模型训练结过程中(a)准确率；(b)召回率)

Fig. 7　In model training process ((a) Accuracy; (b) Recall)

在同等条件下，为了测试本文算法的有效性，通过消融实验来验证各个模块对算法的影响，实验结果见表 2。

(1) 多头注意力机制的性能。为了证明多头注意力机制的效果，即 BT-MHSA 模块。在表 2 中，M1 模块在 YOLOv5s 的基础上 mAP 提升了 0.9%，

准确率提升了 0.8%，召回率提升了 0.7%。模型检测速度的提升是因为模型的参数量减少，进一步证明模型的检测速度与参数量和复杂度呈负相关。同时也认为是将 3×3 的卷积改进为多头注意力机制的原因，虽然卷积操作能够有效提取图像特征，但卷积操作本身需要较大的参数量，可致检测速度减慢。从实际应用场景考虑，本文只在主干网络末端引入 BT-MHSA 模块，以减少模型参数量的同时提升了检测精度。表 2 中，M6 较 YOLOv5s 的 mAP 提升了 1.8%，召回率和准确率分别提高了 2.5%和 0.1%，但模型的参数量增大，检测速度降低，主要是因引入了 MF 模块，MF 本身参数量较大，从而导致 M6 参数量的增多。

(2) 浅层特征增强模块的性能分析。对比分析表 2 中 YOLOv5s 和 M2 模型数据，可发现浅层特征增强模块(SP 模块)具有更好的检测效果。分析发现 M2 的 mAP 提高了 1.1%，而模型参数量的减少有利于模型检测速度的提升。模型的召回率和 mAP 提升的原因是由于浅层特征增强模块保留了更多的底层细节信息，从而提高了对小目标的检测能力。实验证明，在未额外增加检测层的情况下，对浅层特征增强模块的改进，可有效提高对小目标的检测能力。通过分析 M5 模型与 YOLOv5s 的实验结果发现 MF 模块与 BT-MHSA 模块的结合能更好地提高模型的 mAP，充分证明模型改进的有效性。

(3) 多级特征融合模块的性能分析。对比分析 M3 模块和 YOLOv5s 发现，M3 模块的 mAP 提升了 0.9%，召回率提升了 1.5%，准确率降低了 0.1%，模型参数量的增加导致推理速度变慢。分析认为多级特征融合模块(multi-level feature fusion，MF)在训练的过程中不断优化特征层的权重，同时需获得多个不同尺度的特征信息，并对其权重进行调节，虽然增强了特征整合能力，但导致参数量的增加，使得模型的推理速度变慢。在该模块的基础上引入浅层特征增强模块，即 M4 模型，其 mAP 比 YOLOv5s 增加了 1.3%，且召回率与精确率分别增加了 1.1%和 1.7%。M4 的 mAP 比 M5 提高了 0.3%，说明浅层特征增强模块与多级特征融合模块之间特征信息相互补充，模型的检测效果更佳。

表 2 VisDrone2021 测试集上的消融实验结果

Table 2 Ablation experiment results on the VisDrone2021 test set

Model	BT-MHSA	SP	MF	P (%)	R (%)	mAP (%)	Params (M)	FPS^{1536} (帧/秒)
YOLOv5s	-	-	-	53.2	43.5	42.6	7.037	56
M1	√	-	-	54.0	44.2	43.5	6.719	58
M2	-	√	-	52.7	45.3	43.7	5.388	60
M3	-	-	√	53.1	45.0	43.5	8.174	47
M4	-	√	√	54.9	44.6	43.9	9.159	44
M5	√	√	-	52.1	45.3	43.6	7.061	54
M6	√	-	√	53.3	46.0	44.4	9.747	43
M7	√	√	√	**55.6**	**46.5**	**45.7**	9.832	41

注：加粗数据为最优值

总之，本文算法在提高检测精度的同时可满足实时性。通过对比分析 YOLOv5s 的实验数据可知，本文算法的 mAP、精确度和召回率分别提升了 3.1%，2.4%和 3.0%，FPS 达到了 41 帧/秒。因此，本文算法可以准确、快速地检测出无人机图像中的目标。

为了更直观地展示本文算法在真实场景下的效果，从 VisDrone 数据集中选取部分图像进行检测，检测效果如图 8 所示。图 8(b)与(c)组第 1 列相比，发现图(c)组图像中各个类别的检测精度均有所提升；图 8(b)与(c)组第 2 列的检测结果中发现，在光照强度较低且目标密集的场景下，与 YOLOv5 算法相比，本文算法对图像中的小目标漏检和误检更少。在图 8(a)与(b)组的第 3 列图像中看出，本文算法对于被遮挡的小物体检测效果更好。从图 8(b)与(c)组的第 4 列图像中可发现，本文算法对图像中行人的检测精度有所提升。通过以上检测效果分析发现，本文算法具有明显的优势。

3.3.3 对比分析

为了验证改进算法的性能，本文选取多种先进的检测算法进行实验，见表 3。通过对比分析 VisDrone 测试集上 AP 与 mAP 的结果，本文算法在敞篷三轮车、自行车、公交车、汽车、行人和面包车的 AP 较 YOLOX 高出了 2.77%，2.87%，5.8%，2.64%，7.06%和 0.44%，但在摩托车和货车较 YOLOX 低了 0.83%和 0.21%，综合来看，YOLOX

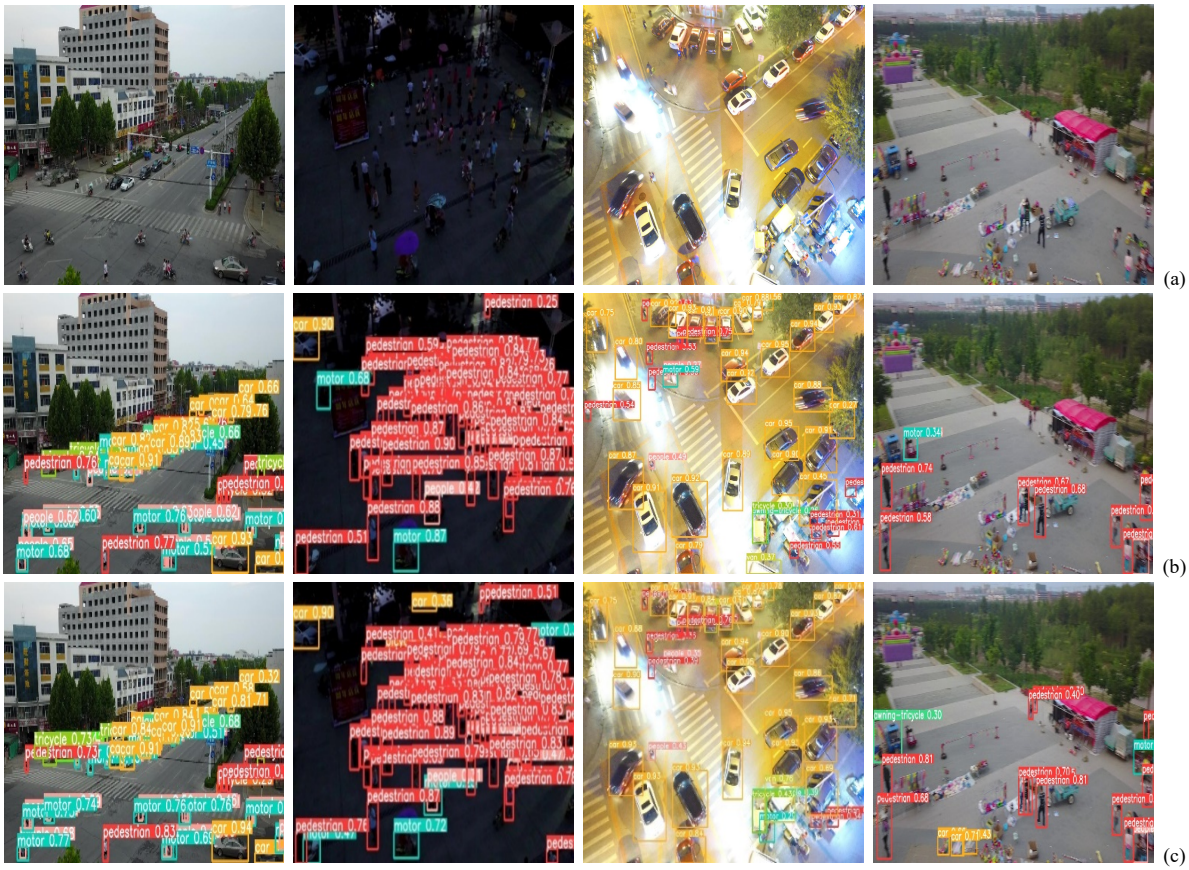


图 8　检测效果对比结果((a) VisDrone 图像；(b)原始 YOLOv5；(c)本文算法)

Fig. 8　Comparison results of detection effect ((a) Image of VisDrone; (b) Original YOLOv5; (c) Ours)

表 3　不同算法在 VisDrone 测试集上的对比分析

Table 3　Comparative analysis of different algorithms on the VisDrone test set

算法	输入尺寸	目标类别										mAP(%)
		Awn-tr	Bicycle	Bus	Car	Motor	Pedestrian	People	Tricycle	Truck	Van	
Faster R-CNN	640×640	8.73	5.86	43.79	44.16	16.83	12.55	8.10	8.53	30.42	20.45	19.94
YOLOv3	640×640	7.71	6.80	39.36	68.87	21.53	22.54	12.50	8.41	26.41	24.31	23.84
CenterNet	640×640	14.28	7.51	42.66	61.96	18.86	22.94	11.67	13.08	24.74	19.38	23.71
DMNet[19]	640×640	14.11	8.89	49.23	58.90	29.38	27.67	**18.93**	**20.32**	29.30	30.27	28.70
YOLOv4[20]	640×640	12.39	8.68	48.86	69.21	22.71	26.67	14.48	12.67	29.94	27.19	27.28
SSD	640×640	11.15	7.38	49.82	63.17	19.09	18.71	9.01	11.74	33.10	29.96	25.31
YOLOX	640×640	15.43	9.03	51.80	72.16	**29.33**	25.44	17.07	16.47	**39.21**	35.16	31.11
本文算法	640×640	**18.20**	**11.90**	**57.60**	**74.80**	28.50	**32.50**	18.80	17.60	39.00	**35.60**	**33.45**

注：加粗数据为最优值

算法的mAP较本文算法低了2.29%。相较于DMNet算法，人和三轮车分别较本文算法高了 0.13%和2.72%。表 3 中本文算法的 mAP 值表现最佳。因此，尽管在目标实例较小的情况下，本文算法也体现出更好的性能，相较于原算法能够获得更丰富的目标特征且能够挖掘上下文信息，在无人机航拍图像的检测中表现出较好的优势，且有效地提高了小目标的特征学习能力，使得模型在处理航拍图像任务中具有更大的优势。

4　结束语

本文针对无人机图像中目标尺寸小、特征提取不理想导致现有的目标检测器对小目标物体检测效果差这一问题，提出一种基于特征融合与注意力机制的无人机图像小目标检测算法。该算法首先使用重构的主干网络增强特征提取能力，充分捕获全局信息和丰富上下文信息，在提升检测效果的前提下，降低模型参数量。浅层特征增强模块使得模型

更关注小目标的特征信息，在未额外增加目标检测层的条件下，能够有效改善对小目标检测能力。此外，多级特征融合模块通过学习不同特征图之间的联系，动态调节不同层级的特征权重，进而提高模型的预测能力。实验结果表明，本文算法的 mAP、召回率、精确率均有提升，对高分辨率图像的检测速度达到了 41 帧/秒。本文算法虽然对航拍图像小目标检测有所改进，但仍有进步的空间，尤其是对小物体的误检、漏检。下一步计划研究如何更高效且轻量化提升对小目标的检测能力，更好地实现实时检测无人机航拍图像中的物体。

参考文献 (References)

[1] 江波, 屈若锟, 李彦冬等. 基于深度学习的无人机航拍目标检测研究综述[J]. 航空学报, 2021, 42(4): 524519. 1-524519. 15.
JIANG B, QU R K, LI Y D, et al. Object detection in UAV imagery based on deep learning: review[J]. Acta Aeronautica et Astronautica Sinica, 2021, 42(4): 524519. 1-524519. 15 (in Chinese).

[2] 周立旺, 潘天翔, 杨泽曦, 等. 多阶段优化的小目标聚焦检测[J]. 图学学报, 2020, 41(1): 93-99.
ZHOU L W, PAN T X, YANG Z X, et al. FocusNet: coarse-to-fine small object detection network[J]. Journal of Graphics, 2020, 41(1): 93-99 (in Chinese).

[3] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. [2022-05-26]. https://arxiv.org/abs/1804.02767.

[4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBoxsDetector[C]//The 14th European Conference on Computer Vision. Cham: Springer International Publishin, 2016: 21-37.

[5] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. New York: IEEE Press, 2015: 1440-1448.

[6] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99.

[7] CAO J, CHOLAKKAL H, ANWER R M, et al. D2Det: towards high quality object detection and instance segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 11485-11494.

[8] ZHAN W, SUN C F, WANG M C, et al. An improved Yolov5 real-time detection method for small objects captured by UAV[J]. Soft Computing, 2022, 26(1): 361-373.

[9] LIM J S, ASTRID M, YOON H J, et al. Small object detection using context and attention[C]//2021 International Conference on Artificial Intelligence in Information and Communication. New York: IEEE Press, 2021: 181-186.

[10] SONG Z Y, ZHANG Y, LIU Y, et al. MSFYOLO: feature fusion-based detection for small objects[J]. IEEE Latin America Transactions, 2022, 20(5): 823-830.

[11] LIU Y J, YANG F B, HU P. Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks[J]. IEEE Access, 2020, 8: 145740-145750.

[12] 胡俊, 顾晶晶, 王秋红. 基于遥感图像的多模态小目标检测[J]. 图学学报, 2022, 43(2): 197-204.
HU J, GU J J, WANG Q H. Multimodal small target detection based on remote sensing image[J]. Journal of Graphics, 2022, 43(2): 197-204 (in Chinese).

[13] 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. New York: IEEE Press, 2017: 2117-2125.

[14] LI H C, XIONG P F, AN J, et al. Pyramid attention network for semantic segmentation[EB/OL]. [2022-05-26]. https://arxiv.org/abs/1805.10180.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//The 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[16] PAN X R, GE C J, LU R, et al. On the integration of self-attention and convolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2022: 815-825.

[17] SRINIVAS A, LIN T Y, PARMAR N, et al. Bottleneck transformers for visual recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 16514-16524.

[18] 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. New York: IEEE Press, 2021: 2847-2854.

[19] 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. New York: IEEE Press, 2020: 737-746.

[20] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. [2022-05-26]. https://arxiv.org/abs/2004.10934.