

引用格式：冒国韬，邓天民，于楠晶. 基于多尺度分割注意力的无人机航拍图像目标检测算法[J]. 航空学报, 2023, 44(5): 326738. MAO G T, DENG T M, YU N J. Object detection in UAV images based on multi-scale split attention[J]. Acta Aeronautica et Astronautica Sinica, 2023, 44(5): 326738 (in Chinese). doi: 10.7527/S1000-6893.2021.26738

基于多尺度分割注意力的无人机航拍图像目标检测算法

冒国韬¹, 邓天民^{1,2,*}, 于楠晶³

1. 重庆交通大学 交通运输学院, 重庆 400074
2. 重庆大学 自动化学院, 重庆 400044
3. 重庆交通大学 航运与船舶工程学院, 重庆 400074

摘要：随着无人机(UAV)遥感技术的发展,无人机航拍图像目标检测逐渐成为无人机应用领域的一项核心技术,在交通规划、军事侦查及环境监测等领域具有重要应用价值。针对无人机图像中小目标实例多、背景复杂及特征提取困难的问题,提出一种基于多尺度分割注意力的无人机航拍图像目标检测算法 MSA-YOLO。首先,利用嵌入在骨干网络瓶颈层的多尺度分割注意力单元建立多尺度特征间的远程依赖关系,从而强化关键特征的表达能力并抑制背景噪声干扰;其次,设计了一种自适应加权特征融合方法,该方法动态的优化各输出特征层权重,实现浅层特征与深层特征的深度融合;最后,在VisDrone公开数据集上的实验结果表明:该方法取得了34.7%的平均均值精度(mAP),相比于基线算法YOLOv5提高了2.8%,在复杂背景下仍能显著提升无人机图像目标检测性能。

关键词：无人机图像; 计算机视觉; 目标检测; 注意力机制; 自适应加权特征融合

中图分类号: V279; TN911.73; TP391.41 文献标识码: A 文章编号: 1000-6893(2023)05-326738-11

随着航空遥感技术的发展,无人机在军事侦查、环境监测及交通规划等领域引起了广泛关注,无人机图像目标检测技术作为无人机图像应用的关键技术之一,能够拓宽无人机的场景理解能力,在军事和民用领域具有重要的应用价值。然而,传统目标检测方法由于手工特征设计繁琐、鲁棒性差及计算冗余等原因,难以满足无人机图像目标检测的需求。近年来,以卷积神经网络(Convolutional Neural Network, CNN)为代表的深度学习方法在计算机视觉领域迎来了迅速发展^[1],基于深度学习的目标检

测方法凭借其强大的自适应学习能力和特征提取能力,在检测性能上远超传统的目标检测方法,因此越来越多学者开始利用深度学习的方法进行无人机图像目标检测。当前基于深度学习的无人机图像目标检测方法可依据是否需要区域建议分为2类:

1) 基于区域建议的无人机图像目标检测算法,此类方法通过提取若干候选区域的特征信息来对预设的候选目标区域进行分类与回归,进而获取目标的类别与位置,其中较为典型的有Faster R-CNN^[2]、Mask R-CNN^[3]、Cascade R-

收稿日期: 2021-12-03; 退修日期: 2021-12-20; 录用日期: 2021-12-31; 网络出版时间: 2022-01-12 11:42

网络出版地址: <https://hkxb.buaa.edu.cn/CN/Y2023/V44/I5/326738>

基金项目:国家重点研发计划(SQ2020YFF0418521);重庆市技术创新与应用发展专项(cstc2020jscx-dxwtBX0019);川渝联合实施重点研发项目(cstc2020jscx-cylhX0007)

*通信作者. E-mail: dtianmin@cqjtu.edu.cn

CNN^[4]等。近年来,许多学者基于这类算法提出了针对无人机图像目标的检测方法。例如,Liu等^[5]针对无人机图像小目标可获取特征信息少的问题,基于Faster R-CNN网络设计了一种多分支并行特征金字塔网络(Multi-branch Parallel Feature Pyramid Networks, MPFPN)以捕获更丰富的小目标特征信息,此外,通过引入监督空间注意力模块(Supervised Spatial Attention Module, SSAM)减弱背景噪声的干扰,有效提升了对无人机图像小目标的检测性能,但对于训练图像中从未标注的物体存在误检的情况。Lin等^[6]在Cascade R-CNN网络的基础上提出了多尺度特征提取骨干网络Trident-FPN,同时引入注意力机制设计了一种注意力双头检测器,有效改善了由于无人机图像目标尺度差异大对目标检测器带来的不利影响,但区域建议网络较大的计算开销还有待改善。

2) 基于回归的无人机图像目标检测算法,该类方法在不进行区域建议的情况下完成端到端的目标检测,直接通过初始锚点框对目标定位并预测类别,典型的有YOLO(You Only Look Once)系列算法^[7]、单击多盒检测器(Single Shot MultiBox Detector, SSD)^[8]及RetinaNet^[9]等。为达到无人机图像目标实时检测的目的,已有研究人员将基于回归的目标检测算法应用于无人机图像领域。例如,Zhang等^[10]提出一种基于YOLOv3的深度可分离注意力引导网络,通过引入注意力模块并将部分标准卷积替换为深度可分离卷积,有效提升了对无人机图像中小目标车辆的检测效果。Wang等^[11]提出了一种高效的无人机图像目标检测器SPB-YOLO,首先利用设计的条形瓶颈(Strip Bottleneck, SPB)模块来提高对不同尺度目标的检测效果,其次,通过基于路径聚合网络(Path Aggregation Network, PA-Net)^[12]提出的特征图上采样策略,提高了检测器在无人机图像密集检测任务中的表现。裴伟等^[13]提出了一种基于特征融合的无人机图像目标检测方法,通过引入不同分类层的特征融合机制以高效的结合网络浅层和深层的特征信息,有效改善了SSD目标检测算法存在的漏检和重复检测问题,但由于更多的网络层次和深度增加了

较大的计算开销,严重影响了目标检测实时性。

由于大视场下的无人机航拍图像目标往往呈现稀疏不均的分布,搜索目标将会花费更高的成本。此外,无人机航拍图像的待检目标具有小尺度、背景复杂、尺度差异大及排列密集等特征,通用场景的目标检测方法很难取得理想的检测效果。基于此,本文提出一种多尺度分割注意力单元(Multi-Scale Split Attention Unit, MSAU),分别从通道和空间2个维度自适应的挖掘不同尺度特征空间的重要特征信息,抑制干扰特征信息,通过将其嵌入基础骨干网络,使网络更具指向性的提取任务目标区域的关键信息;进一步的,本文结合加权特征融合思想提出一种自适应加权特征融合方法(Adaptive Weighted feature Fusion, AWF),通过动态调节各个特征层的重要性分布权重,实现浅层细节信息与深层语义信息的高效融合。最后,结合以上提出的MSAU和AWF两种策略,本文设计了一种基于多尺度分割注意力的无人机航拍图像目标检测算法(Multi-scale Split Attention-You Only Look Once, MSA-YOLO)。

1 本文方法

MSA-YOLO算法的核心思想是尽可能保证目标检测器实时检测性能的前提下,着重关注如何挖掘有益于无人机图像目标检测的关键特征信息,通过提出的多尺度分割注意力单元MSAU和自适应加权特征融合AWF来提升基准模型YOLOv5在无人机图像目标检测任务中的表现。MSA-YOLO算法的框架结构如图1所示,嵌入在骨干网络瓶颈层(Bottleneck Layer)中的多尺度分割注意力单元MSAU主要包括多尺度特征提取模块、通道注意力模块及空间注意力模块3个部分,首先通过多尺度特征提取模块提取出丰富的多尺度特征信息,随后利用并行组合的混合域注意力为多尺度特征层的不同特征通道和区域赋予不同的注意力权重,从大量多尺度特征信息中筛选出对无人机图像任务目标更重要的信息;自适应加权特征融合AWF利用可学习的权重系数对3个特征尺度的特征层进行加权处理并实现自适应的特征融合,进

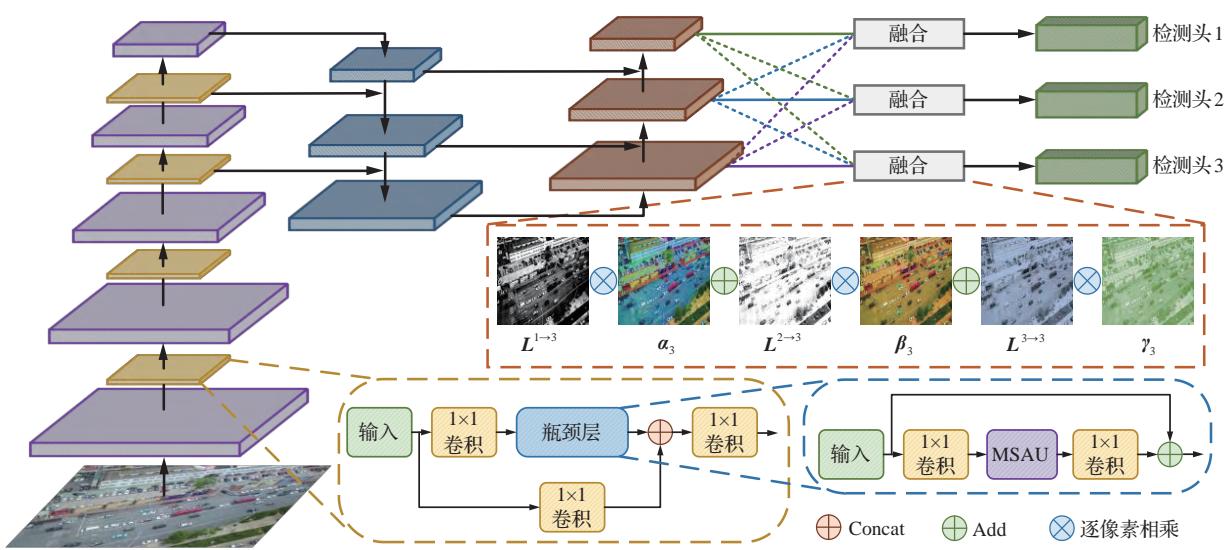


图1 MSA-YOLO算法框架结构图

Fig. 1 Architecture of MSA-YOLO algorithm

而结合丰富的上下文信息强化目标检测器的表征能力。

1.1 多尺度分割注意力单元

在特征提取过程中,采用固定尺寸的卷积核只能提取到目标局部的特征信息,无法通过不同感受野挖掘丰富的上下文信息,为有效利用不同尺度的特征空间信息,本文设计了一种多尺度特征提取模块(Multi-scale Feature Extraction Module, MFEM),通过多尺度卷积的方式来获取不同尺度的特征信息。MFEM的多尺度特征提取过程如图2所示,假定多尺度特征提取模块MFEM的输入特征空间为 $X = [X_1, X_2, \dots, X_c] \in \mathbb{R}^{C \times H \times W}$,通过split切片操作将输入特征空间 X 的通道平均切分为 n 个部分,若 C 表示该输入特征的通道数,则切片后各个部分 X_i 的通道数为 $C' = C/n$,为了降低模块的参数量,本文采用不同分组数量 G_i 且不同卷积核尺寸 $k_i \times k_i$ 的分组卷积提取多尺度的特征信息 $F_i \in \mathbb{R}^{C' \times H \times W}$

$$F_i = F_{\text{conv}}^{k_i \times k_i}(X_i, G_i) \quad i = 0, 1, \dots, n-1 \quad (1)$$

式中: $F_{\text{conv}}^{k_i \times k_i}(X_i, G_i)$ 表示对特征图 X_i 进行分组数量为 G_i 且卷积核尺寸为 $k_i \times k_i$ 的卷积操作,为保证模型较小的计算开销,本文将输入特征空间切分为4个部分,则设置 $n=4$,分组卷积核尺寸 k_i 分别为3、5、7、9,分组数量 G_i 分别为1、2、3、4。

各个部分的特征图 X_i 在分别经过不同尺寸

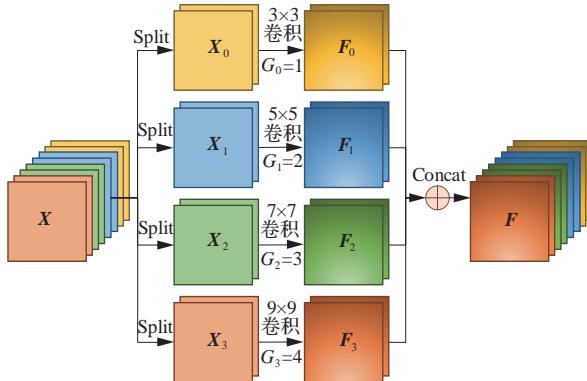


图2 多尺度特征提取模块流程图

Fig. 2 Flow chart of multi-scale feature extraction module

的卷积核后获得了不同尺度的感受野,并提取出不同尺度的特征信息 F_i ,对 F_i 进行Concat操作可以得到最终融合后的特征空间 $F \in \mathbb{R}^{C \times H \times W}$

$$F = \text{Cat}([F_0, F_1, F_2, \dots, F_{n-1}]) \quad (2)$$

式中:Cat(\cdot)表示对所有的特征图进行Concat操作。

本文的多尺度特征提取模块在一定程度上弥补了卷积核尺寸单一对网络特征提取能力的不利影响,对输入特征空间进行均匀分割,再分别利用4种不同感受野的卷积核捕捉不同尺度的特征空间信息,最后将获得的4种不同尺度的特征信息进行融合,使得融合后的特征空间 F 具备丰富的多尺度上下文信息,有利于交错复杂的无人机图像检测任务。

注意力机制中,所有特征信息会根据学到的注意力权重进行加权处理,相关性较低的特征信息被赋予较低的权重,反之则被赋予较高的权重,以此弱化不重要信息的干扰,并分离出重要信息。按照注意力域的不同,一般可将注意力机制分为通道域注意力机制、空间域注意力机制及混合域注意力机制。通道注意力机制关注特征图通道之间的远程依赖关系,空间域注意力机制聚焦于特征图中对分类起决定作用的像素区域,混合域注意力机制则同时利用到空间域和通道域的信息,每个通道特征图中的每个元素都对应一个注意力权重。这些即插即用的注意力模型可以无缝集成到各种深度学习网络中用以指导目标检测任务。

为更好地提取无人机图像目标的特征信息,弱化无关背景信息的干扰,本文结合通道域注意力和空间域注意力,提出了一种并行组合的混合域注意力,一方面沿着通道维度获取通道间的远程相互依赖关系,另一方面通过强调空间维度感兴趣的区域进一步挖掘特征图的上下文信息。本文设计的混合域注意力由挤压激励模块(Squeeze-and-Excitation Module, SEM)^[14]和空间注意力模块(Spatial Attention Module, SAM)^[15]并行连接组成。通道注意力旨在通过生成一种可以维持通道间相关性的注意力权重图来挖掘输入与输出特征通道之间的远距离依赖关系,SEM和SAM的网络结构如图3所示。

假设通道注意力模块SEM的输入特征空间为 $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c] \in \mathbf{R}^{C \times H \times W}$, C 表示该输入人

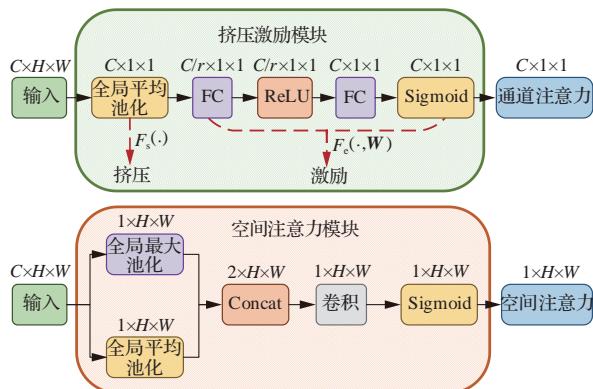


图3 SEM和SAM的网络结构图

Fig. 3 Network structure of SEM and SAM

特征的通道数, $H \times W$ 表示输入特征的尺度大小,输入特征空间的第 c 个通道用 $\mathbf{X}_c \in \mathbf{R}^{H \times W}$ 表示。 $F_s(\cdot)$ 表示挤压(Squeeze)映射, $F_e(\cdot, \mathbf{W})$ 表示激励(Excitation)映射, $F_s(\cdot)$ 通常采用全局平均池化(Global Average Pooling, GAP)实现,对输入空间特征 \mathbf{X} 进行 $F_s(\cdot)$ 映射后获得全局特征空间 $\mathbf{Z} \in \mathbf{R}^{C \times H \times W}$ 的第 c 个特征 \mathbf{Z}_c :

$$\mathbf{Z}_c = F_s(\mathbf{X}_c) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \mathbf{X}_c(i, j) \quad (3)$$

进一步的,利用 $F_e(\cdot, \mathbf{W})$ 激励操作来降低计算开销,获得高效的自适应学习注意力图。首先使用参数为 \mathbf{W}_0 ,降维系数为 r 的全连接(Fully Connected, FC)层进行降维操作获得维度为 $C/r \times 1 \times 1$ 的特征,通过ReLU函数对特征进行激励操作 δ ,经过参数为 \mathbf{W}_1 的FC层后恢复原始的维度 $C \times 1 \times 1$,最后利用sigmoid激活函数进行归一化后得到各个通道的注意力权重 \mathbf{S} ,即

$$\mathbf{S} = F_e(\mathbf{Z}, \mathbf{W}) = \sigma(\mathbf{W}_1 \delta(\mathbf{W}_0 \mathbf{Z})) \quad (4)$$

空间注意力模块旨在利用输入特征的空间信息生成空间注意力权重图,并对输入特征进行空间域注意力加权,进而增强重要区域的特征表达。空间注意力模块的输入特征空间与通道注意力模块的输入特征空间 $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c] \in \mathbf{R}^{C \times H \times W}$ 相同,分别沿着通道维度采用全局最大池化(Global Max Pooling, GMP)和全局平均池化(Global Average Pooling, GAP)压缩后得到 $\mathbf{X}_{\text{avg}} \in \mathbf{R}^{1 \times H \times W}$ 和 $\mathbf{X}_{\text{max}} \in \mathbf{R}^{1 \times H \times W}$ 这2个特征图,对2个特征图进行concat操作后采用感受野较大 7×7 卷积核进行卷积操作 $F_{\text{conv}}^{7 \times 7}$,最后通过sigmoid激活函数 σ 进行归一化后得到空间域注意力权重图 $\mathbf{M} \in \mathbf{R}^{1 \times H \times W}$:

$$\mathbf{M} = \sigma(F_{\text{conv}}^{7 \times 7}(\text{Cat}([\mathbf{X}_{\text{avg}}', \mathbf{X}_{\text{max}}'])))) \quad (5)$$

空间注意力模块SAM将输入特征的每个通道进行相同的空间加权处理,忽视了通道域的信息交互;而通道注意力模块SEM则忽视了空间域内部的信息交互,将一个通道内的信息进行全局加权处理。因此,本文将通道注意力模块与空间注意力模块通过并行的方式连接,旨在从全局特征信息出发,沿着通道与空间2个维度深入挖掘输入特征内部的关键信息,进而筛选出任务相关的重要信息,弱化不相关信息

的干扰。相比于级联连接的组合方式,这种并行组合的方法无需考虑通道注意力模块与空间注意力模块的先后顺序,2种注意力模块都直接对初始输入特征空间进行学习,不存在特征学习过程中互相干扰的情况,从而使混合域注意力的效果更稳定^[16]。

混合域注意力同时考虑了空间注意力和通道注意力,在一定程度上丰富了特征信息,但无法有效地挖掘和利用不同尺度的特征空间信息。鉴于此,本文设计了一种能够有效地建立多尺度

注意力间长期依赖关系的多尺度分割注意力单元MSAU,如图4所示。MSAU主要由多尺度特征提取模块MFEM、通道注意力模块SEM及空间注意力模块SAM组成,输入特征空间 X 通过多尺度特征提取模块捕捉不同尺度的特征信息,得到多尺度特征空间 F ,随后,不同尺度的特征图分别通过通道注意力模块和空间注意力模块得到多尺度注意力权重,最后利用并行组合的通道与空间2个维度的多尺度注意力进行注意力加权后得到最终输出的特征空间 Y 。

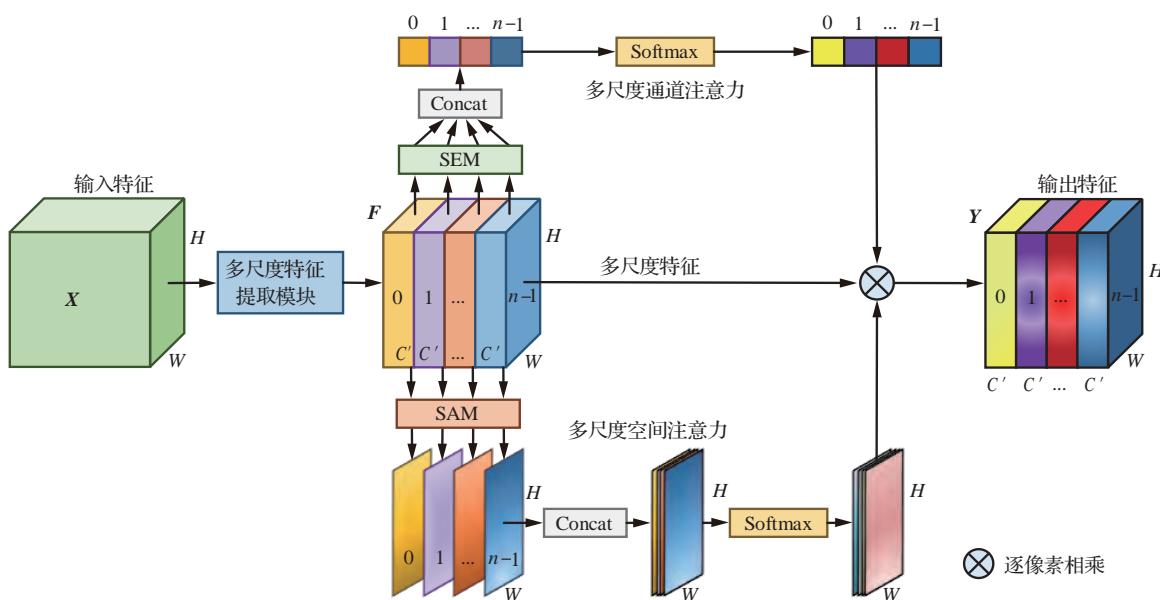


Fig. 4 Architecture of multi-scale split attention unit

假设多尺度分割注意力单元MSAU的输入特征空间为 $X=[X_1, X_2, \dots, X_n] \in \mathbf{R}^{C \times H \times W}$,经过多尺度特征提取模块提取特征后得到多尺度特征空间 $F \in \mathbf{R}^{C \times H \times W}$,随后不同尺度特征图 F_i 利用通道注意力模块来获得多尺度通道注意力权重 S_i :

$$S_i = \text{SEM}(F_i) \quad i = 0, 1, \dots, n-1 \quad (6)$$

式中:SEM(•)代表利用通道注意力模块SEM挖掘特征图的通道注意力; S_i 为 F_i 的通道注意力权重值,因此整个多尺度通道注意力 S 可以表示为
 $S = S_0 \oplus S_1 \oplus \dots \oplus S_{n-1}$ (7)

式中: \oplus 表示Concat操作; S 为多尺度通道注意力权重。

为建立通道间的远程依赖关系,实现多尺度

通道注意力之间的信息交互,进一步利用Softmax函数对通道注意力 S_i 进行重新标定得到最终的通道注意力权重 H_i :

$$H_i = \text{Softmax}(S_i) = \frac{\exp(S_i)}{\sum_{i=0}^{n-1} \exp(S_i)} \quad (8)$$

式中:Softmax(•)表示Softmax操作,用于获取多尺度通道的重标定权重 H_i 。

类似的,可以利用空间注意力模块捕捉不同尺度特征图 F_i 的多尺度空间注意力权重 M_i :

$$M_i = \text{SAM}(F_i) \quad i = 0, 1, \dots, n-1 \quad (9)$$

式中:SAM(•)代表利用空间注意力模块SAM捕捉特征图的空间注意力; M_i 为 F_i 的空间注意力权重值,因此整个多尺度空间注意力 M 可表示为
 $M = M_0 + M_1 + \dots + M_{n-1}$ (10)

式中: + 表示 add 操作; \mathbf{M} 为多尺度空间注意力权重。

随后, 利用 Softmax 函数 Softmax(\cdot) 对空间注意力 \mathbf{M}_i 进行重新标定得到最终的空间注意力权重 \mathbf{P}_i :

$$\mathbf{P}_i = \text{softmax}(\mathbf{M}_i) = \frac{\exp(\mathbf{M}_i)}{\sum_{i=0}^{n-1} \exp(\mathbf{M}_i)} \quad (11)$$

最后, 将 SEM 和 SAM 学习到的多尺度通道注意力权重向量 \mathbf{H}_i 和多尺度空间注意力权重图 \mathbf{P}_i 与多尺度特征空间 $\mathbf{F} \in \mathbf{R}^{C \times H \times W}$ 进行注意力加权 F_{scale} 得到输出的多尺度特征空间 \mathbf{Y}_i :

$$\mathbf{Y}_i = \mathbf{F}_i \otimes \mathbf{H}_i \otimes \mathbf{P}_i \quad i = 0, 1, \dots, n-1 \quad (12)$$

式中: \otimes 表示特征加权乘法运算符号。Concat 操作能在不破坏原始特征图信息的前提下, 完整地维持特征表示, 因此, 最终得到的多尺度分割注意力单元 MSAU 的输出 \mathbf{Y} 可表示为

$$\mathbf{Y} = \text{Cat}([\mathbf{Y}_0, \mathbf{Y}_1, \dots, \mathbf{Y}_{n-1}]) \quad (13)$$

MSAU 首先利用多尺度特征提取模块有效提取了不同尺度特征空间的多维特征, 随后将其分别输入并行组合的混合域注意力, 为不同尺度特征空间赋予了不同的重要性权重。这种方法不仅能考虑到多尺度特征信息, 同时使网络能够有选择地处理关键信息, 对目标区域投入更多注意力资源, 以获取更多待检目标的细节信息。同时, 不同尺度特征空间的多尺度注意力权重会在模型训练过程中根据每轮输入特征空间的重要性差异进行自适应的、精确的调整更新, 通过将其嵌入骨干网络, 进而利用丰富的特征空间以指导无人机图像目标检测任务。

1.2 自适应加权特征融合

浅层网络提取目标纹理边缘特征, 具有更多的细节内容描述; 深层网络则提取目标丰富的语义特征, 但同时削弱了对小目标位置信息和细节信息的感知, 以致丢失小目标在特征图中的特征信息^[17]。PANet 将不同深度特征信息以平等关系跨层融合, 忽略了不同特征层之间的关系, 直接使用 3 个特征尺度的输出特征进行目标预测, 但不同深度特征层对任务目标的贡献其实是不同的, 浅层网络特征在小目标检测过程中占据着更重要的位置。针对以上问题, 本节设计了一种

自适应加权特征融合方法 AWF, 通过为各尺度特征层赋予不同比例权重, 有效利用了 3 个不同尺度特征层的浅层和深层特征, 自适应的强化特征金字塔中对任务目标检测更重要的特征信息, 进而融合丰富的特征信息以指导无人机图像小目标检测任务。

AWF 在进行最终的特征融合时采用了加权再相加的方式, 因此, 需要确保参与融合的特征层分辨率相同, 且通道数也应相同。对于特征金字塔的输出特征 $\mathbf{L}^n \in \mathbf{R}^{C^n \times H^n \times W^n}$, 其中 $n \in 1, 2, 3$, 通过上采样或下采样将特征金字塔输出特征 $\mathbf{L}^m \in \mathbf{R}^{C^m \times H^m \times W^m}$ 的特征图分辨率和通道数都调整为与 \mathbf{L}^n 相同, $\mathbf{L}^{m \rightarrow n} \in \mathbf{R}^{C^n \times H^n \times W^n}$ 表示调整后的特征。对于上采样, 首先使用 1×1 卷积层来调整特征的通道数, 然后通过双线性插值来提高分辨率; 对于下采样, 则使用步长为 2 的最大池化层和 3×3 卷积层同时改变特征的分辨率和通道数。将调整后的特征通过 Concat 操作进行整合后可表示为整个特征金字塔的输出特征 $\mathbf{L} \in \mathbf{R}^{3C^n \times H^n \times W^n}$:

$$\mathbf{L} = \text{Cat}([\mathbf{L}^{1 \rightarrow n}, \mathbf{L}^{2 \rightarrow n}, \mathbf{L}^{3 \rightarrow n}]) \quad (14)$$

随后, 使用 Softmax 函数 Softmax(\cdot) 和 1×1 卷积层 $F_{\text{conv}}^{1 \times 1}$ 得到权重矩阵 $\mathbf{W} \in \mathbf{R}^{4 \times H^n \times W^n}$:

$$\mathbf{W} = \text{softmax}(F_{\text{conv}}^{1 \times 1}(F_{\text{conv}}^{1 \times 1}(\mathbf{L}))) \quad (15)$$

最后, 沿着通道维度将权重矩阵 \mathbf{W} 切割为 $\alpha_n^*, \beta_n^*, \gamma_n^* \in \mathbf{R}^{1 \times H^n \times W^n}$, 再沿着通道维度进行扩展后得到特征金字塔调整后特征 $\mathbf{L}^{m \rightarrow n}$ 对应的重要权重参数 $\alpha_n, \beta_n, \gamma_n \in \mathbf{R}^{C^n \times H^n \times W^n}$, 这些重要性权重参数来自前面特征层经过卷积后的输出, 并通过网络的梯度反向传播变为了可自适应学习的参数。将其与对应特征 $\mathbf{L}^{m \rightarrow n}$ 加权融合后得到新的融合特征 \mathbf{F}^n :

$$\mathbf{F}^n = \mathbf{L}^{1 \rightarrow n} \alpha_n + \mathbf{L}^{2 \rightarrow n} \beta_n + \mathbf{L}^{3 \rightarrow n} \gamma_n \quad (16)$$

由于加权特征融合的权重参数均源自前面 3 个尺度特征层的输出, 因此可学习的权重参数和特征是息息相关的, 数据集实例样本的特点则是影响贡献衡量标准的主要因素, 针对小目标实例居多的无人机航拍图像, 则认为浅层网络中丰富的纹理和边缘特征对无人机航拍目标检测任务具有更大的贡献, 更有利于提取小目标的类别及位置信息, 因此浅层网络特征层则会被赋予更高的权重值, 而这样一个有效的权重系数可以经过

不断优化的训练过程产生。在模型训练过程中,AWF根据各尺度特征层对当前任务目标的贡献大小来动态的调节其权重值,充分挖掘了不同深度特征层的多维特征,可以更好地监督网络的特征融合过程,使融合后的特征兼顾强大的语义信息和丰富的几何细节信息。

值得一提的是,这种自适应加权的特征融合方法并不是能够完全适用于任何目标检测任务,在数据集整体实例的像素大小或各类目标实例的特征未呈现出一种较为显著的趋势时,可能很难达到较为理想的效果。

2 实验结果与分析

2.1 实验数据与参数设置

1) 实验平台:本文实验采用的硬件配置为Nvidia RTX3060 GPU 和 Intel i5-10400 2.90 GHz CPU,软件环境为Windows10系统下的Pytorch深度学习框架。

2) 数据集:本文实验所采用的数据来源于VisDrone无人机图像目标检测公开数据集^[18]。该数据集包括行人(指具有行走或站立姿势的人)、人(指具有其他姿势的人)、汽车、货车、公共汽车、卡车、摩托车、自行车、遮阳蓬三轮车及三轮车共10个类别。VisDrone数据集由288个视频剪辑而成,分为 1360×765 和 960×540 像素2种不同的图像尺寸,总计提供了由不同高度的无人机捕获的10 209幅静态图像,其中包括6 471幅训练集图像、548幅验证集图像及3 190幅测试集图像,共计260万个目标实例样本。

3) 评价指标:为评估本文所提算法的有效性,选取模型规模、参数数量及每秒浮点运算次数(Floating Point Operations, FLOPs)来评价模型的复杂程度,选取平均均值精度(mean Average Precision, mAP)作为模型对多个目标类别综合检测性能的评价指标,采用平均精度(Average Precision, AP)来评价模型对单个目标类别的检测性能。

2.2 消融实验

为了验证所提的多尺度分割注意力单元

MSAU和自适应加权特征融合 AWF在无人机图像目标检测任务中的有效性,本文在VisDrone测试集上进行了一系列的消融实验,以YOLOv5为基线算法,mAP、模型规模、参数量及浮点运算次数为评价指标,最终结果如表1所示。

表1 VisDrone测试集上的消融实验结果

Table 1 Results of ablation experiment on VisDrone test set

方法	mAP	模型规模/MB	参数量/B	FLOP/G
基线算法	31.9	89	46.68	114.4
基线算法+MSAU	34.1	104	54.21	140.9
基线算法+AWF	32.8	94	48.89	119.8
MSA-YOLO	34.7	108	56.28	146.1

消融实验的结果表明,将提出的多尺度分割注意力单元MSAU嵌入基线算法的骨干网络后,算法的模型规模和参数量分别增加了15 MB和7.53M($1M=10^6$),同时浮点运算次数增加到140.9G($1G=10^9$),取得了34.1%的mAP,检测精度的提升也从侧面反映出了MSAU捕获不同尺度特征信息的能力,正是由于其精准高效的挖掘了特征空间在多尺度上的特征信息,因此能在确保模型较小计算复杂度的同时有效提升对无人机航拍图像目标的检测效果;进一步的,在基线算法基础上采用所提的自适应加权特征融合AWF方法,相比基线算法仅增加了2.21M的参数量和5.4G的浮点运算次数,并取得了32.8%的mAP,AWF在自适应地融合了网络深层与浅层的丰富语义信息和几何信息后,能够较为充分的捕获无人机图像目标的特征信息。同时,由于AWF添加了3个特征融合层,且每个特征融合层都利用到前面各个特征尺度的输出特征,给网络带来了一定的计算开销,但相比于基线算法采用Concat的特征融合操作,加权再相加的特征融合方式可使融合后的特征空间维持在更低的通道数,因此保持了良好的实时性能;与基线算法相比较,本文所提MSA-YOLO算法的参数量和浮点运算次数分别增加了9.6 M和31.7 G,模型规模由于参数量的增高而上升到108 MB,mAP则比基线算法提高了2.8%,达到34.7%。综上所述,MSA-YOLO算法在维持较小计算开销的前提下获得了更好的检测性能,可以有效地指导无

人机图像目标检测任务。

2.3 各算法在 VisDrone 测试集上的检测性能

为证明 MSA-YOLO 算法对无人机图像各目标检测的有效性,本文在 VisDrone 测试集中与各种先进的无人机图像目标检测算法进行对比分析,表 2^[19-24]为各算法对 VisDrone 测试集 10 类目标的 AP 值与 mAP 值。从表 2 中可以看出,MSA-YOLO 算法与其他先进算法相比取得了最优的综合性能,比次优的 CDNet 高出 0.5% 的 mAP。对于汽车、卡车及公共汽车等目标类别取得了最优的检测性能,分别达到了 76.8%、41.4% 及 60.9% 的 AP 值,对于行人、货车、及摩托车等纵横比较大且实例个数较少的目标类别则分别达到了 33.4%、41.5% 及 31.0% 的较优 AP 值,在目标实例个数较少的情况下能够较为充分的挖掘其特征信息,由此可见本文提出的 MSA-YOLO 算法在处理无人机图像目标检测任务时具有较大优势,其检测效果是十分可观的。

为了验证 MSA-YOLO 算法在实际场景中的检测效果,选取 VisDrone 测试集中实际检测较

为困难的图像进行测试,部分检测结果如图 5 所示,可以看出,本文方法对不同拍摄角度下背景复杂且分布密集的无人机图像展现出了较为优异的检测性能,能够有效地抑制图像背景噪声信息的干扰,更具选择性的挖掘有利于无人机图像目标检测任务的重要特征信息。为进一步评价基线算法和 MSA-YOLO 算法在处理无人机图像目标检测任务时的性能差异,本文在 VisDrone 测试集中随机选取了小目标样例图片进行测试,并可视化对比分析,如图 6 所示。

本文分别抽取了晴天和夜间的小目标样例并对比了 2 种算法的检测结果,可以看出,MSA-YOLO 算法有效提升了基线算法对小尺度目标的检测效果。通过图 6(a)与图 6(b)的对比发现,基线算法错将站立的行人检测为人,且存在大量行人目标漏警的情况,而 MSA-YOLO 算法则能够精准的进行识别。对比图 6(c)和图 6(d)可以看出,在夜间低照度的情况下,基线算法受到背景噪声信息的干扰出现了部分漏警,MSA-YOLO 算法则通过弱化噪声干扰、强化网络感兴趣的多尺度特征,从大量多尺度特征信息中分离出了有利于无人机图像目标检测的信息,在面对复杂的

表 2 不同算法在 VisDrone 测试集上的 AP 与 mAP 对比

Table 2 Comparison of AP and mAP of different algorithms on VisDrone test set

方法	骨干网络	AP/%										mAP /%
		行人	人	自行车	汽车	火车	卡车	三轮车	遮阳蓬	公共三轮车	汽车	
Faster R-CNN ^[19]	ResNet-50	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7	21.7
Faster R-CNN ^[19]	ResNet-101	20.9	14.8	7.3	51.0	29.7	19.5	14.0	8.8	30.5	21.2	21.8
Cascade R-CNN ^[19]	ResNet-50	22.2	14.8	7.6	54.6	31.5	21.6	14.8	8.6	34.9	21.4	23.2
RetinaNet ^[19]	ResNet-50	13.0	7.9	1.4	45.5	19.9	11.5	6.3	4.2	17.8	11.8	13.9
CenterNet ^[20]	Hourglass-104	22.6	20.6	14.6	59.7	24.0	21.3	20.1	17.4	37.9	23.7	26.2
YOLOv4 ^[21]	CSPDarknet	24.8	12.6	8.6	64.3	22.4	22.7	11.4	7.6	44.3	21.7	30.7
DMNet ^[22]	ResNet-50	28.5	20.4	15.9	56.8	37.9	30.1	22.6	14.0	47.1	29.2	30.3
DBAI-Det ^[23]	ResNeXt-101	36.7	12.8	14.7	47.4	38.0	41.4	23.4	16.9	31.9	16.6	28.0
CDNet ^[22]	ResNeXt-101	35.6	19.2	13.8	55.8	42.1	38.2	33.0	25.4	49.5	29.3	34.2
HR-Cascade++ ^[22]	HRNet-W40	32.6	17.3	11.1	54.7	42.4	35.3	32.7	24.1	46.5	28.2	32.5
MSC-CenterNet ^[22]	Hourglass-104	33.7	15.2	12.1	55.2	40.5	34.1	29.2	21.6	42.2	27.5	31.1
YOLOv3-LITE ^[24]	Darknet-53	34.5	23.4	7.9	70.8	31.3	21.9	15.3	6.2	40.9	32.7	28.5
MSA-YOLO	CSPDarknet	33.4	17.3	11.2	76.8	41.5	41.4	14.8	18.4	60.9	31.0	34.7

注:粗体字表示最优结果。



图5 MSA-YOLO在VisDrone测试集上的部分检测结果

Fig. 5 Partial detection results of MSA-YOLO on VisDrone test set

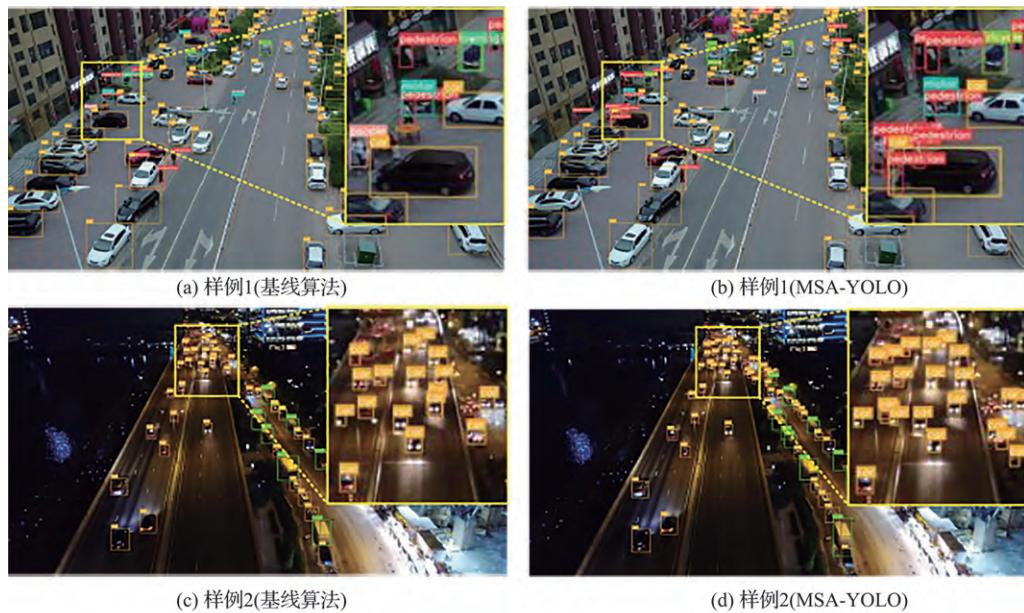


图6 VisDrone测试集上的小目标检测效果对比

Fig. 6 Comparison of small object detection effect on VisDrone test set

背景信息时表现出了较强的抗干扰能力,有效改善了夜间的漏警情况。总体而言,在处理无人机图像目标检测任务时,MSA-YOLO算法相比于基线算法有更明显的优势,对于小尺度、背景复杂及排列密集的无人机图像目标具备更强的辨识能力,有效避免了出现虚警、漏警等现象。

3 结 论

本文提出一种基于多尺度分割注意力的无人机图像目标检测算法 MSA-YOLO。针对无人机图像背景复杂混乱的特点,提出了多尺度分割注意力单元 MSAU,在多个尺度上沿着空间和

通道维度提取无人机图像目标的关键特征信息,同时弱化不相关的背景噪声信息,有益于提高无人机图像目标检测性能。针对无人机图像小尺度目标实例多,缺乏有效特征信息的问题,提出了自适应加权特征融合AWF方法,通过自适应学习的方式动态调节各输入特征层的权重,充分强调浅层细粒度特征信息在特征融合过程中的重要性,有效改善目标检测器对小目标细节位置信息的感知能力。在VisDrone数据集上的实验结果表明,相比于现有的先进无人机图像目标检测方法,MSA-YOLO算法在行人、货车及摩托车类别上分别取得了第五、第三及第二的检测效果,而在汽车、卡车及公共汽车这3种目标类别则上取得了最优的检测效果,能很好的应对无人机图像目标检测任务。

参考文献

- [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.
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 (in Chinese).
- [2] 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.
- [3] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2980-2988.
- [4] CAI Z W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018; 6154-6162.
- [5] 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.
- [6] LIN Q Z, DING Y, XU H, et al. ECascade-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images[C]//2021 7th International Conference on Automation, Robotics and Applications (ICARA). Piscataway: IEEE Press, 2021: 268-272.
- [7] 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. Piscataway: IEEE Press, 2016: 779-788.
- [8] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector[C]//Proceedings of the 14th European Conference on Computer Vision (ECCV). Berlin: Springer, 2016: 21-37.
- [9] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2) : 318-327.
- [10] ZHANG Z Y, LIU Y P, LIU T C, et al. DAGN: A real-time UAV remote sensing image vehicle detection framework [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(11): 1884-1888.
- [11] WANG X R, LI W H, GUO W, et al. SPB-YOLO: An efficient real-time detector for unmanned aerial vehicle images[C]//2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC). Piscataway: IEEE Press, 2021: 99-104.
- [12] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.
- [13] 裴伟,许晏铭,朱永英,等.改进的SSD航拍目标检测方法[J].软件学报,2019,30(3):738-758.
PEI W, XU Y M, ZHU Y Y, et al. The target detection method of aerial photography images with improved SSD [J]. Journal of Software, 2019, 30 (3) : 738-758 (in Chinese).
- [14] 赵辉,李志伟,张天琪.基于注意力机制的单发多框检测器算法[J].电子与信息学报,2021,43(7):2096-2104.
ZHAO H, LI Z W, ZHANG T Q. Attention based single shot multibox detector [J]. Journal of Electronics & Information Technology, 2021, 43 (7) : 2096-2104 (in Chinese).
- [15] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 1352-1368.
- [16] 王美华,吴振鑫,周祖光.基于注意力改进CBAM的农作物病虫害细粒度识别研究[J].农业机械学报,2021,52(4):239-247.
WANG M H, WU Z X, ZHOU Z G. Fine-grained identification research of crop pests and diseases based on improved CBAM via attention [J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(4) : 239-247 (in Chinese).
- [17] LIU S T, HUANG D, WANG Y H. Learning spatial fusion for single-shot object detection[DB/OL]. arXiv preprint: 1911.09516, 2019.

- [18] ZHU P F, WEN L Y, DU D W, et al. Vision meets drones: Past, present and future [DB/OL]. arXiv preprint: 2001.06303, 2020.
- [19] 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. Piscataway: IEEE Press, 2021: 3257-3266.
- [20] ALBABA B M, OZER S. S_yNet: An ensemble network for object detection in UAV images[C]//2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2020: 10227-10234.
- [21] ALI S, SIDDIQUE A, ATEŞ H F, et al. Improved YOLOv4 for aerial object detection[C]//2021 29th Signal Processing and Communications Applications Conference (SPCOM).
- [22] 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). Piscataway: IEEE Press, 2021: 2847-2854.
- [23] 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). Piscataway: IEEE Press, 2019: 213-226.
- [24] ZHAO H P, ZHOU Y, ZHANG L, et al. Mixed YOLOv3-LITE: A lightweight real-time object detection method [J]. Sensors (Basel, Switzerland), 2020, 20(7): 1861.

(责任编辑: 李丹)

Object detection in UAV images based on multi-scale split attention

MAO Guotao¹, DENG Tianmin^{1,2,*}, YU Nanjing³

1. School of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China

2. School of Automation, Chongqing University, Chongqing 400044, China

3. School of Shipping and Naval Architecture, Chongqing Jiaotong University, Chongqing 400074, China

Abstract: With the development of Unmanned Aerial Vehicle (UAV) remote sensing technology, UAV aerial image object detection has become a core technology in the field of UAV applications such as traffic planning, military reconnaissance and environmental monitoring. To overcome the problem of difficulty in feature extraction due to many instances of small objects and complex background in UAV images, this paper proposes an object detection algorithm for UAV aerial images based on multi-scale split attention, i.e., MAS-YOLO. Firstly, the multi-scale split attention unit embedded in the bottleneck layer of the backbone net-work is used to establish the long-range dependency relationship between different scales of attention, so as to enhance the expression ability of key features and suppress the interference of background noise. Secondly, an adaptive weighted feature fusion method is designed, which dynamically optimizes the weight of each output feature layer and realize the deep fusion of shallow and deep features. Finally, experimental results on the VisDrone public data set show that the proposed method achieves 34.7% mean Average Precision (mAP), which is 2.8% higher than that of the baseline algorithm YOLOv5, and can also significantly improve the performance of UAV image object detection in complex background.

Keywords: unmanned aerial vehicle image; computer vision; object detection; attention mechanism; adaptive weighted feature fusion

Received: 2021-12-03; **Revised:** 2021-12-20; **Accepted:** 2021-12-31; **Published online:** 2022-01-12 11:42

URL: <https://hxkb.buaa.edu.cn/CN/Y2023/V44/I5/326738>

Foundation items: National Key Research & Development Program of China (SQ2020YFF0418521); Chongqing Science and Technology Development Foundation (cstc2020jscx-dxwtBX0019); Joint Key Research & Development Program of Sichuan and Chongqing (cstc2020jscx-cylhX0007)

* Corresponding author. E-mail: dtianmin@cqjtu.edu.cn