DOI:10.19651/j.cnki.emt.2415864

面向无人机监控的动态多尺度目标检测模型的研究与实现*

张 宇[1,4] 王延吉[1,4] 马 辉[1,3] 闫 锴[2] 李大舟[1,4]

(1.沈阳化工大学计算机科学与技术学院 沈阳 110142; 2.沈阳科技学院信息与控制工程系 沈阳 110167;
3.沈阳化工大学网络与信息化中心 沈阳 110142; 4.辽宁省化工过程工业智能化技术重点实验室 沈阳 110142)

摘 要: 在无人机侦察、安防监控以及自动驾驶等领域中,目标检测技术面临巨大的挑战,图像中的目标往往具有多尺度属性,尤其是小尺寸目标检测难,以及目标很容易受到不同程度的遮挡。针对这些亟待解决的问题,本文提出了一种创新的动态多尺度目标检测模型:YOLO-DDE。首先,本文了提出了 CEMA 和 CED 卷积模块,增强了骨干网络对多尺度信息的处理能力精细特征提取能力,从而实现在复杂场景下更加精确的识别效果。此外,本文通过对 FPAN 网络结构进行创新性重构,提出了 DFPN 结构,此结构采用纵向跨尺度融合技术,显著提升了模型的尺度特征融合效果。最后,引入了动态检测头,提出了 DD-Head 结构,强化了模型对下游任务处理的能力。综上所述,本文提出的 YOLO-DDE 模型以其动态多尺度结构,为目标检测技术的性能提升提供了新的可能性。本文在 PASCAL VOC 数据集上进行了消融实验和对比试验,与当前主流先进模型 YOLOv8 相比,本文模型 YOLO-DDE 在评价指标 map50 和 map50-95 上分别提升了 1.8%和 3.2%,并且本文还在 VisDrone、HIT-UAV、FAIR1M2.0 数据集上进行了泛化性实验,验证了模型具有很强的泛化能力。

关键词: 注意力机制;多尺度;解耦头;可变形卷积;DFPN
中图分类号: TN911.73 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Research and implementation of dynamic multi-scale target detection model for UAV surveillance

Zhang Yu[1,4] Wang Yanji[1,4] Ma Hui[1,3] Yan Kai[2] Li Dazhou[1,4]

(1. School of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China;
2. Department of Information and Control Engineering, Shenyang Institute of Science and Technology, Shenyang 110167, China;
3. Network and Informatisation Centre, Shenyang University of Chemical Technology, Shenyang 110142, China;
4. Key Laboratory of Industrial Intelligent Technology of Chemical Process of Liaoning Province, Shenyang 110142, China)

Abstract: In the fields of UAV reconnaissance, security monitoring, and autonomous driving, target detection technology faces significant challenges. Targets in images often exhibit multi-scale attributes, making detection of small-sized targets particularly difficult, and targets are prone to various degrees of occlusion. To address these pressing issues, this paper proposes an innovative dynamic multi-scale target detection model: YOLO-DDE. Firstly, novel CEMA and CED convolutional modules are introduced to enhance the backbone network's ability to handle multi-scale information and extract fine features, thus achieving more precise recognition in complex scenes. Additionally, the FPAN network structure is innovatively restructured into the DFPN structure, which employs longitudinal cross-scale fusion technology to significantly improve the model's scale feature fusion effect. Finally, a dynamic detection head is introduced, proposing the DD-Head structure, which strengthens the model's ability to handle downstream tasks. In summary, the proposed YOLO-DDE model, with its dynamic multi-scale structure, provides new possibilities for improving target detection technology performance. Experiments on the PASCAL VOC dataset were conducted to validate the proposed model. Compared to the current state-of-the-art model YOLOv8, the YOLO-DDE model achieves a 1.8% and 3.2% improvement in evaluation metrics map50 and map50-95, respectively. Furthermore, generalization experiments on the VisDrone, HIT-UAV, and FAIR1M2.0 datasets validate the model's strong generalization ability.

Keywords: attention mechanism; multi-scale; decoupled head; deformable convolution; DFPN

0 引 言

近年来,随着深度学习理论的不断深入和计算能力的显著提升,基于深度学习的目标检测性能已经远远超越传统方法。根据候选框的生成方式,目标检测方法可以分为单阶段(如 SSD[1]、YOLO 系列[2]等)和双阶段(如 R-CNN

收稿日期:2024-04-17
* 基金项目:辽宁省教育厅科学研究项目(LJKZ0449)资助

系列[3-4]系列等)两大类。

单阶段 YOLO 系列的目标检测算法已经在各个领域(如无人机侦察、安防监控和自动驾驶等)得到了广泛应用[5]。在无人机监控领域为了应对多样化目标检测、复杂背景和动态场景等挑战,研究人员已经开始开发一系列创新技术,以提高目标检测算法在各种挑战下的性能表现。

赵耘彻等[6]通过在 YOLOv4 中增加小目标检测层来提升模型对多尺度目标检测能力,但是基准模型 YOLOv4 的能力有限。YOLOv8[2]是 YOLO 系列较新版本,通过改进网络结构和优化训练策略,在速度和准确性方面取得了更好的成果。

Zhang 等[7]提出了一种名为 Drone-YOLO 的改进检测模型。该模型在 YOLOv8 中引入了 RepVGG[8]模块作为下采样层,还添加了大视野检测头,加强多尺度特征学习能力。但是 Drone-YOLO 模型仅关注尺度特征问题,忽略了细节特征提取问题,导致模型面对复杂背景检测任务时性能无法达到预期的水平。本文不仅关注了尺度特征问题,还通过引入注意力机制提高模型的细节提取能力。

Huangfu 等[9]提出了 LW-YOLO,该模型在 YOLOv8 中直接插入 SE 注意力机制[10],增强小目标特征提取能力,使用轻量化的 GSConv[11]模块代替了普通卷积和 C2F 模块的中的 Bottleneck 模块。但这种设计在面对无人机监测领域中小目标众多且密集的场景时,表现可能不如预期。本文不对特征融合层进行轻量化设计,而是进一步加强了 Neck 层的特征融合能力。

Li 等[12]在 YOLOv8 的特征融合层中通过三次跨层连接来提高不同尺度间的特征融合,并用 GhostblockV2[13]结构代替部分卷积模块 C2F,WIoU 替换原有的边界框损失函数。模型因此在参数量下降的基础上,还得到了性能的提升。但是由于他们实验只在单一的数据集进行,无法验证模型的泛化性,不能得知他们在所有的小目标类别中表现是否可以达到预期。本文则在多个数据集上进行了泛化实验,验证了模型的有效性和泛化性。

Lou 等[14]提出了 DCS-YOLOv8 检测模型。该模型使用 MDC 模块来执行下采样操作,MDC 模块结合了深度可分离卷积[15]、最大池化和步幅为 2 的 3×3 卷积。此外还提出 DC 模块,通过堆叠深度可分离卷积和普通卷积而组成。但在特别是面对复杂背景和多样化目标时,这种混合结构的性能无法完全满足需求。本文的设计的核心卷积模块则不存在此问题,相反在面对复杂背景和多样化目标时表现更为优越。

本文模型与上述最新的无人机监控模型有一定的联系,首先都是在 YOLOv8 模型的基础上进行了修改,其次都在 Backbone 层的核心模块上进行了创新设计,最后都在 Neck 层的特征融合部分进行了努力。

综上所述,现有改进大多专注于小目标检测,忽略了多样化目标场景中较大目标的处理需求。此外,在复杂和动态场景下,由于适应性和动态表达能力的不足,模型的检测效果可能不理想。更重要的是,各改进模块之间缺乏足够的关联性和协调性。为了解决这些问题,本文以动态多尺度为核心理念对 YOLOv8 进行改进,提出了一种新的动态多尺度模型(yolo dynamic multiscale target detection model,YOLO-DDE)。具体改进包括:

在 Backbone 层中,分别入 Intern-Image[16]核心算子可变形卷积 v3(deformable convolution v3,DCNv3)和高效多尺度注意力机制[17](efficient multi-scale attention,EMA)。本文结合 EMA 的高效多尺度特性和 DCNv3 的动态适应性提出了高效多尺度卷积(convolution efficient multi-scale attention,CEMA)结构和高效多尺度可变卷积(deformable convolution efficient multi-scale,CED)结构。本文在 EMA 与常规卷积块进行组合时进行了多次对比实验,在如何发挥他们 CEMA 与 CED 应有的性能时也做了相关实验。

在 Neck 层中,基于原有的 FPN-PAN[18]设计了一种新特征融合结构双向特征金字塔(double feature pyramid networks,DFPN),采用上、下采样跨空间融合的结构来弥补不相邻特征层之间语义信息的丢失,增强特征复用性,加强模型对多个尺度目标特征的敏感性。

在 Head 层中,引入了 Dyhead[19]结构,并对 Dyhead 进行了一定改进,提出动态解耦头(dynamic decoupling head,DD-Head)进一步加强其动态特性,然后与原有的解耦头进行融合,形成了一种新的动态解耦头。

为充分验证模型有效性和优越性,在数据集 PASCAL VOC 上进行了对比实验和消融实验,在 VisDrone、HIT-UAV、FAIR1M2.0 进行了泛化性实验,结果表明由于本文提出的动态多尺度思想,YOLO-DDE 的表现更好。

1　YOLO-DDE 算法设计

图像经过预处理后输入本文模型,先通过改进的 Backbone 提取特征,提取完的特征被送入本文提出的 DFPN 架构进一步处理和整合。处理和整合后的特征传递到本文的 DD-Head 模块,最后输出结果图片。本文提出的 YOLO-DDE 模型结构如图 1 所示。

1.1　CEMA 模块

注意力机制通过动态地调整特征权重和关注重要信息,可以有效提高模型性能、降低计算成本、处理长距离依赖关系,并增强模型的解释性,因此在各种深度学习任务中得到广泛应用。

现在研究者普遍认为,目前提出的注意机制主要有 3 种类型:通道注意、空间注意和两者兼而有之。SE 是通道注意力的代表,核心思想是在每个通道上学习到一个注意力权重,以便模型可以自动地选择对于特定任务更重要的特征。

CBAM[20]用于自动学习图像中不同区域的注意力权重,以改善模型在各种视觉任务中的表现。通过将通道注

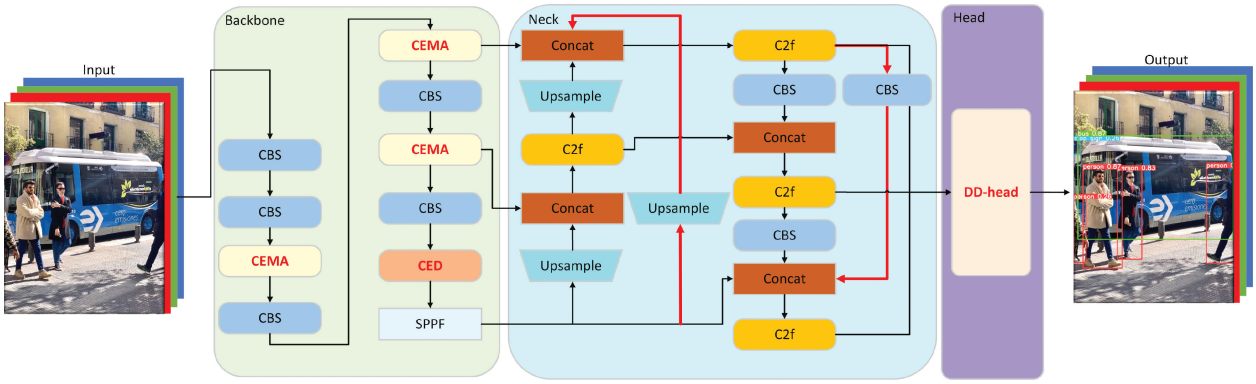


图 1　本文提出的 YOLO-DDE 模型结构

意力和空间注意力结合起来,CBAM 可以有效地增强卷积(convolutional neural network,CNN)模型对于输入图像的表征能力。

SGE[21]将通道维度分组为多个子特征,改善了不同语义子特征表示的空间分布。虽然对通道的降维和分组能获得更优的性能,但不可避免地降低了检测器的处理效率从而增大延迟。

EMA 是一种高效多尺度注意力机制,综合了以上注意力的优点,该注意力机制专注于保留每个通道的信息并降低计算开销,它将部分通道重塑为批处理维度,并将通道维度分组成多个子特征,使得空间语义特征在每个特征组内得到良好分布。EMA 总体结构如图 2 所示。

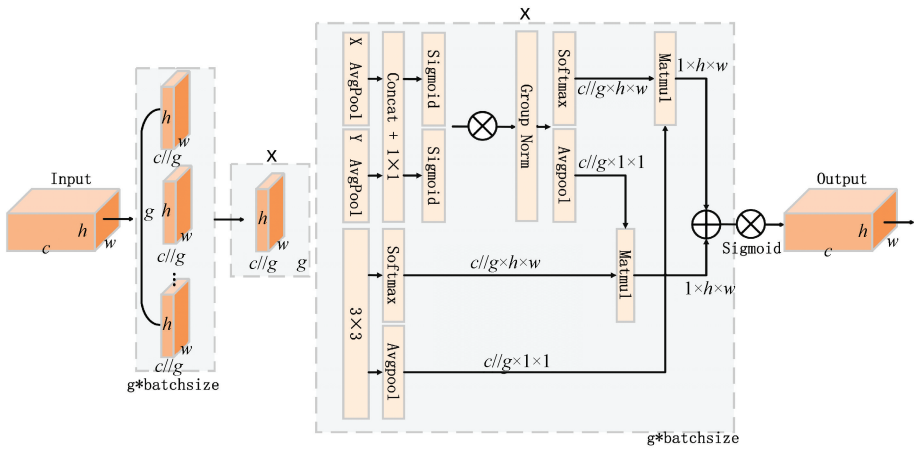


图 2　EMA 注意力结构

对于输入特征 $\boldsymbol{F}\in\boldsymbol{R}^{C\times H\times W}$,EMA 按通道数将其划分为 G 个子特征,输入特征:

$$\boldsymbol{F}=[\boldsymbol{F}_0,\boldsymbol{F}_i\cdots,\boldsymbol{F}_{G-1}],\boldsymbol{F}_i\in\boldsymbol{R}^{C//G\times H\times W} \tag{1}$$

其中,分的组数 G 要远小于通道数 C,才能利用学习到的注意力权重来增强所有子特征的表征能力。

在 1×1 卷积分支中,对 F_i 进行水平和垂直方向上的平均池化操作:

$$\boldsymbol{F}_{ih}=agp\_h(\boldsymbol{F}_i) \tag{2}$$

$$\boldsymbol{F}_{iw}=agp\_w(\boldsymbol{F}_i) \tag{3}$$

将水平和垂直方向上的结果拼接,并经过一个 1×1 的卷积操作进行信息交互,然后再将其分割为两部分:

$$\boldsymbol{F}'_{ih},\boldsymbol{F}'_{iw}=Split(\mathrm{Conv}_{1\times1}(\mathrm{concat}(\boldsymbol{F}'_{ih},\boldsymbol{F}'_{iw})),\boldsymbol{H},\boldsymbol{W}) \tag{4}$$

将分组的输入张量 $\boldsymbol{F}_i$,与经过 sigmoid 函数处理的水平信息和垂直信息分别相乘,然后进组归一化:

$$\boldsymbol{X}_1=\boldsymbol{GN}(\boldsymbol{F}_i\times\mathrm{sigmoid}(\boldsymbol{F}'_{ih})\times\mathrm{sigmoid}(\boldsymbol{F}'^{\mathrm{T}}_{iw})) \tag{5}$$

在 3×3 卷积分支中对输入的分组张量 F_i 进行卷积操作:

$$\boldsymbol{X}_2=\mathrm{Conv}_{3\times3}(\boldsymbol{F}_i) \tag{6}$$

跨纬度交互,得到最终权重 $\boldsymbol{W}$:

$$\boldsymbol{W}=\mathrm{softmax}(\boldsymbol{X}_2)\times agp(\boldsymbol{X}_1)+\mathrm{softmax}(\boldsymbol{X}_1)\times agp(\boldsymbol{X}_2) \tag{7}$$

使用权重 W 对输入张量 F_i 进行加权,得到最终的输出张量 F_{io}:

$$\boldsymbol{F}_{io}=\boldsymbol{F}_i\times\mathrm{sigmoid}(\boldsymbol{W}) \tag{8}$$

YOLOv8 的 C2F 模块受 YOLOv7 的 ELAN[2]模块启发,拥有更丰富的梯度流,但也损失了一定的空间尺度捕获能力。本文在 C2F 的瓶颈模块中以串联的方法添加 EMA 注意力机制,设计出一种新的 CSP 结构 CEMA,如图 3 所示,从而加强模型的特征分离与聚合的能力,捕捉更

多空间尺度信息。本文对 CEMA 的瓶颈结构 EBottleneck 进行了 4 种设计，如图 4 所示。不同种 EBottleneck 的性能表达如表 1 所示，由表 1 可知，采取图 4(b)设计最好。其中实验配置采用 2.2 小节中表 3 的参数设置。

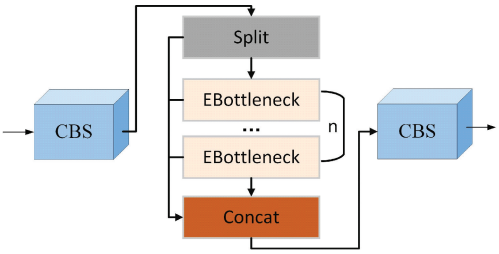


图 3　CEMA 模块结构

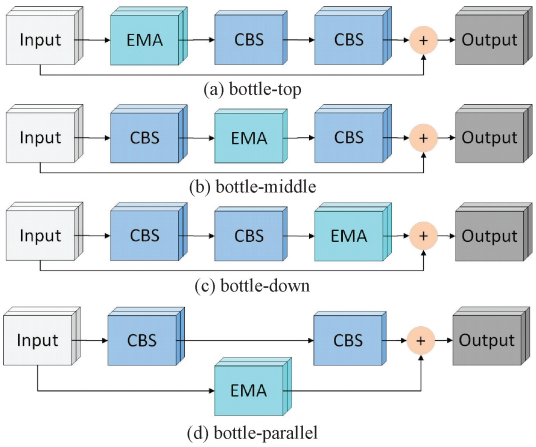


图 4　EBottleneck 模块设计图

表 1　不同种 EBottleneck 的性能表达

模型	Parames/10^6	Flops/G	mAP50/%	mAP50-95/%
Base	11.14	28.7	82.5	62.1
a	11.15	28.7	82.6	62.9
b	11.15	28.7	82.9	62.9
c	11.15	28.7	82.5	62.5
d	11.15	28.7	82.4	62.3

1.2　CED 模块

随着 Transformer 在大规模语言模型中的显著成功，Vision Transformer[22]也席卷了计算机视觉领域，击败了 CNN[23]。常规 CNN 已经比不上大规模参数的 ViTs，于是上海 AI 实验室设计了一个新的基于 CNN 的基础模型 InternImage，该模型可以有效地扩展到大规模的参数和数据。

InternImage 的核心算子对比不同的核心算子，如图 5 所示。图 5(a)为多头自注意力[24]，在需要高分辨率输入的下游任务中计算和内存成本较高；图 5(b)使用局部窗口自注意力[25]，减少成本；图 5(c)使用大核卷积；图 5(d)是一种可变形卷积。

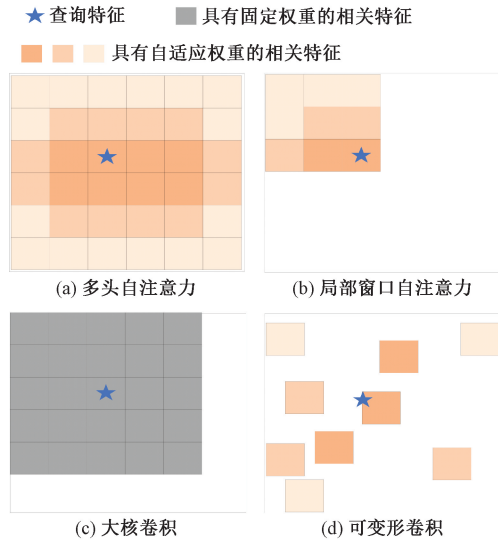


图 5　DCNv3 与其他算子的比较

与最近改进的具有非常大核的 CNN(如 31×31[26])不同，InternImage 的核心算子 DCNv3 是一个动态稀疏卷积，其卷积核大小为 3×3，具有以下特点：

1)其采样偏移量灵活，可以从给定数据动态学习适当的接受域(可以是长范围或短范围)；

2)根据输入数据自适应调整采样偏移量和调制标量，实现像 ViTs 那样的自适应空间聚合，降低了正则卷积的过归纳偏置；

3)卷积窗口是一种常见的 3×3，避免了大卷积核带来的优化问题和昂贵的成本。

DCNv3 算子的计算公式如下：

$$y(p_0)=\sum_{g=1}^{G}\sum_{p_n\in R}w_g\cdot x_g(p_0+p_n+\Delta p_{gn})\cdot\Delta m_{gn} \tag{9}$$

其中，G 为聚合组的总数。对于第 g 组，$w_g\in R^{C\times C'}$ 表示组的位置无关投影权值，其中 $C'=C/G$ 表示组维度。$m_{gk}\in R$ 表示第 g 组中第 k 个采样点的调制标量，沿 K 维度用 softmax 函数归一化。$x_g\in R^{C'\times H\times W}$ 表示切片输入特征图。Δp_{gk} 是第 g 组中网格采样位置 pk 对应的偏移量。

本文研究在 YOLOV8 的骨干网络中引入了 DCNv3 设计了 CED 模块，如图 6 所示。特征图流入 EDBottleneck 先经过第一个 DC3 模块，然后 EMA 注意力进行特征捕获，DC3 继续处理，最后进行残差处理。本文研究经过实验发现 CED 处于骨干网络最后一个 C2F 的位置时，效果收益最好，如表 2 所示，上标表示 CED 在骨干网络中的位置。实验配置采用 2.2 小节中表 3 的参数设置。

1.3　DFPN 模块

YOLOv8 使用的 FPAN 结构是对传统 FPN 的补充。传统 FPN 使用自下而上的形式来传递深层语义特征，通

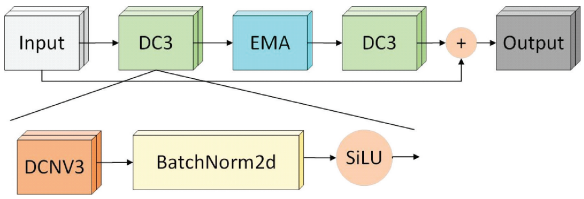


图 6　EDBottleneck 结构

表 2　CED 在骨干网络中不同位置的性能表达

模型	Parames/ 10^6	Flops/ G	mAP50/ %	mAP50-95/ %
Base	11.14	28.7	82.5	62.1
CED^{all}	9.69	24.7	81.9	61.8
$CED^{1,2}$	11.03	26.9	81.2	61.6
$CED^{3,4}$	9.81	26.7	82.7	62.6
CED^{4}	10.69	28.4	83.4	63.3

过对 B3 到 P3、P4 到 P3 和 B4 到 P4、P5 到 P4 的融合,加强语义学习,但造成了一定程度的定位信息丢失。FPAN 是对 FPN 背后自下而上结构的补充,利用 A3 到 A4、P4 到 A4 和 A4 到 A5、P5 到 A5 的融合加强对定位特征的学习,达到互补的效果。但是,将这种结构应用于多尺度尤其小目标和遮挡目标的复杂背景检测时,存在改进的空间:一方面,由于对大规模特征映射的关注不足,检测模型可能会忽略一些有用的特征,降低检测质量;另一方面即使考虑 B、P、A 特征的融合和补充,特征的重用率也很低,原始特征经过较长的上采样和下采样路径后会丢失一些信息。

因此,本文引入了 Bi-PAN-FPN[27] 的思想提高多尺度特征融合的概率和次数,以获得更高的检测精度。受此思想的启发,本文研究重新关注 FPN 和 PAN,设计了新的特征融合结构 DFPN。增加额外的路径,在 FPN 和 PAN 分别增加跨层连接,P5 层通过上采样与 P3 融合,A3 层通过下采样与 A5 层连接,对双向过程中损失的特征信息进行补充,并提高特征复用率特征融合网络结构如图 7 所示。

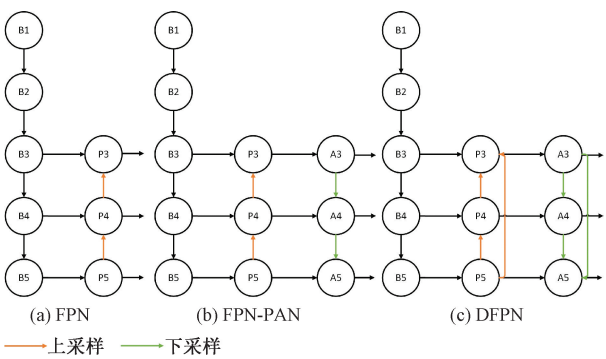


图 7　特征融合网络结构

1.4　DD-Head 模块

目标检测中定位与分类相结合的复杂性导致了各种检测头方法的蓬勃发展,一个高效的检测头可以有效提高目标检测的性能,特别是在复杂场景中,可以更精准的定位目标并进行分类。

DAI 等提出了一种创新的动态头部框架 DyHead,如图 8 所示。该框架在感知特征层级、空间的感知定位以及任务相关的输出通道之间实施了多样化的注意力机制。这种综合性的方法显著增强了目标检测头部的表征能力。

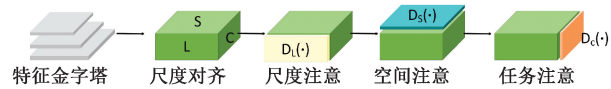


图 8　动态头部框架

为了增强 YOLOv8 在尺度感知、空间感知、任务感知方面的功能,本文结合了 DyHead 与 YOLOv8 的解耦头设计,提出了一种创新的动态解耦头部结构 DD-Head。如图 9 所示。

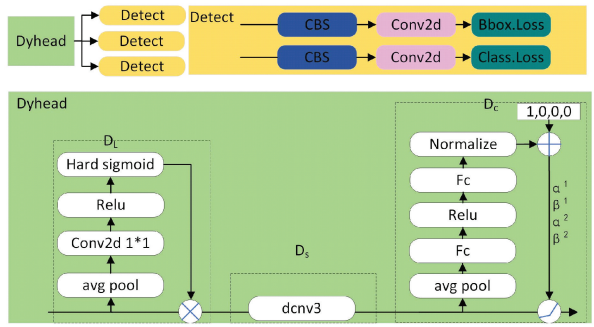


图 9　DD-Head 动态解耦头模块

Neck 层网络输出的多层特征图尺度大小统一后,用 concat 拼接操作将其输入 Dyhead 中,再进行解耦输出。对于给定的特征张量 $F \in R^{L\times S\times C}$,Dyhead Block 可以描述为

$$W(F) = D_C(D_S(D_L(F)\cdot F)\cdot F)\cdot F \tag{10}$$

式中:$D_C(\cdot)$、$D_S(\cdot)$ 和 $D_L(\cdot)$ 分别表示任务感知注意力、空间感知注意力、尺度感知注意力。

本研究还对 DyHead 进行了进一步的改进。由于 DyHead 中的 D 模块依赖于 DCNv2[28] 来增强空间特征感知能力,因此本文研究认为其性能仍有提升空间。因此,本文研究采用了新一代的 DCNv3 来替换 DCNv2,以期提高模型的空间特征感知能力,进一步优化下游任务的处理。

2　实验结果与分析

2.1　数据集

本研究采用 PASCAL VOC2007 和 2012[29] 数据集进行对比和消融实验,Visrdrone[30]、HIT-UAV[31]、FAIR1M2.0[32] 数据集进行模型泛化实验。

PASCAL VOC 2007 与 2012 数据集共汇集了来自日常场景的成千上万张图像,覆盖了 20 个不同的对象类别,包括人类、动物、交通工具和日用品等。VisDrone,是一个

专注于无人机视角下的目标检测和跟踪的复杂数据集。HIT-UAV 是哈尔滨工业大学无人机数据集，用于目标检测和跟踪等计算机视觉任务的研究和评估。FAIR1M2.0 数据集是由脸书研究院创建的大规模航拍图像数据集。

2.2　实验环境

本研究实验在训练过程中使用主流先进模型 YOLOv8 作为基线模型，所用操作系统为 ubuntu20.04，使用 Python3.8、CUDA11.3、Pytorch1.11.0 深度学习框架开发环境，在 NVIDIA GEForce RXT 2080Ti 显卡上进行训练。训练过程中的重要参数设置如表 3 所示。

表 3　训练参数设置表

Parameters	Setup
Epoch	300
Batch size	32
worker	12
OPtimizer	SGD
Momentum	0.937
Initial Leraning Rate	0.01
Final Leraning Rate	0.000 1
Weight-decay	0.000 5
NMS IoU	0.7

2.3　模型评价指标

本文实验结果使用目标检测常用的评估指标：参数量（Parames/10^6）、计算量（Flops/G）、和平均精度均值（mean average precision，mAP），mAP 计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$AP = \int_0^1 P(r)\mathrm{d}r \tag{13}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{14}$$

式中：TP 指图片中的目标被识别为正确的目标；FP 指能识别出目标所在位置，但是目标的类别识别错误；FN 表示正确目标未被识别出来，被认为是其他目标，即漏检；N 代表类别数；APi 为每个类 Precision-Recall 曲线下面的面积，面积越大，证明检测模型越优秀；mAP 为多个类别 APi 的平均值。

本研究使用 map50 与 map50-95 作为精度指标。其中 map50 是指在交并比（intersection over union，IoU）为 0.5 时的 mAP，通常用来评估模型在较低要求下的性能，即模型预测的框与真实框的重叠程度较低时的性能。

map50-95 是指在 IOU 从 0.5～0.95 范围内的 mAP，可以用来评估模型在不同要求下的性能，从较宽松的 0.5 IOU 到较严格的 0.95 IOU 的性能表现。

2.4　对比实验结果分析

在本研究中，通过广泛的实验评估了本文提出的 YOLO-DDE 模型的性能。这些实验旨在通过与其他先进的目标检测模型比较，验证本文模型的有效性和优越性。实验结果如表 4 所示，详细展示了模型性能的比较分析。

表 4　YOLO-DDE 在 PASCAL VOC 数据集上与其他先进模型的对比实验

模型	Parames/ 10^6	Flops/ G	mAP50/ %	mAP50-95/ %
Faster-RCNN-VGG16	136.8	369.8	73.6	—
SSD	24.1	61.2	76.8	—
YOLOv5s	7.07	16.1	78.6	53.7
YOLOv6s	13.14	30.6	84.1	63.5
YOLOv7-tiny	6.07	13.3	79.0	53.3
YOLOv8n	3.16	8.9	78.2	57.7
YOLO-DDEn	**3.00**	**8.5**	79.8	59.2
YOLOv8s	11.14	28.7	82.5	62.1
文献[7]模型	10.9	30.7	83.0	63.4
文献[8]模型	10.7	31.6	83.1	64.1
文献[9]模型	10.3	33.1	82.9	63.0
文献[10]模型	11.1	33.7	82.8	63.8
YOLO-DDEs	10.47	31.8	**84.3**	**65.3**

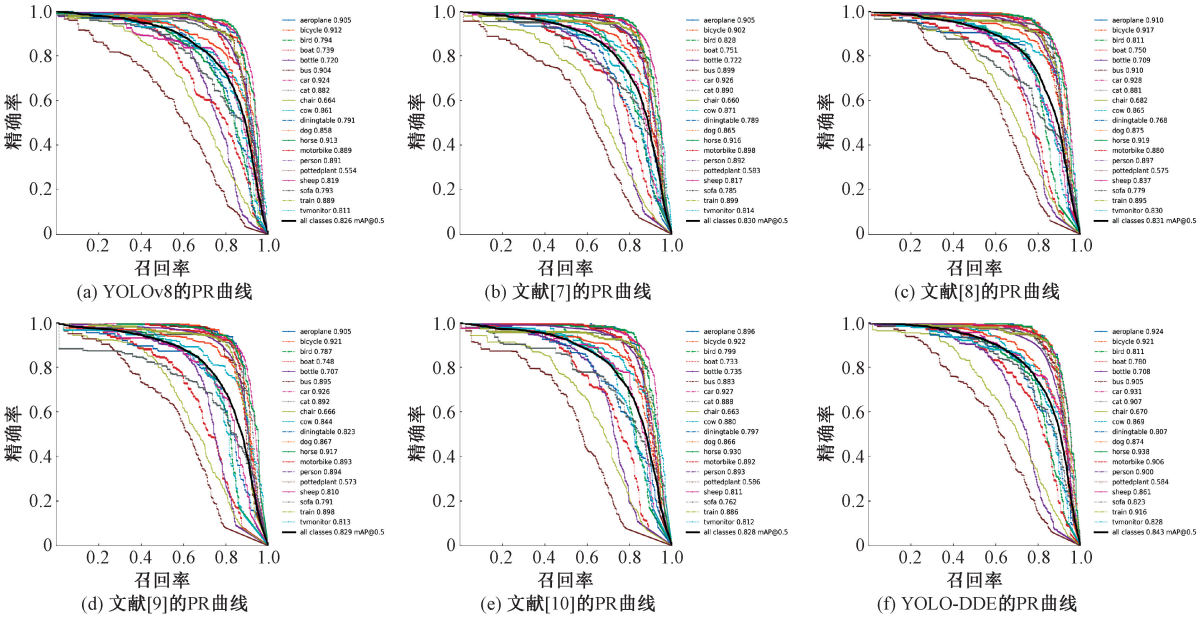
首先，将本文提出的 YOLO-DDE 模型与传统的目标检测算法，进行了比较。结果表明，本文提出的 YOLO-DDE 模型在参数量和计算效率方面具有明显优势，同时在各项性能指标上也显著超越了对比模型。

随后将本文提出的 YOLO-DDE 模型进一步与 YOLO 系列中的多个经典版本和最新已有改进 YOLOv8 模型进行了详细的比较。首先，本文模型 YOLO-DDE 与 YOLOv8n、YOLOv7-tiny 以及 YOLOv5s 等模型进行了性能对比，YOLO-DDEn 以更少的参数和计算资源消耗实现了更高的检测精度。然后，将本文提出模型 YOLO-DDEs 与 YOLOv8s 及 YOLOv6s、文献[7]、文献[8]、文献[9]和文献[10]模型进行对比，结果同样显示，YOLO-DDE 在保持较低计算成本的同时，拥有了更出色的检测性能。

图 10 为主流先进模型和本文提出模型 YOLO-DDE 的 PR 曲线对比图，图 10 展示每个类以及总的 mAP 值和 PR 曲线。从图 10 中曲线与坐标轴所围成的面积可知，本文模型对每个类的识别能力都有不同程度的提升，体现出了模型良好的泛化能力。

通过可视化对比进一步体现 YOLO-DDE 的优越性，如图 11 所示，本文模型同其他模型相比可以很好的检测的各个类别，并且拥有更好的性能表现。

通过这些比较实验，不仅证实了本文针对多尺度和有遮挡问题提出的改进策略和技术的有效性和创新性，还展



(a) YOLOv8的PR曲线　(b) 文献[7]的PR曲线　(c) 文献[8]的PR曲线

(d) 文献[9]的PR曲线　(e) 文献[10]的PR曲线　(f) YOLO-DDE的PR曲线

图 10　主流先进模型和本文提出模型 YOLO-DDE 在 Pascal VOC 07+12 数据集上的 PR 曲线

(a) YOLOv5s检测结果　(b) YOLOv6s检测结果　(c) YOLOv8s检测结果　(d) 文献[7]检测结果

(e) 文献[8]检测结果　(f) 文献[9]检测结果　(g) 文献[10]检测结果　(h) YOLO-DDEs检测结果

图 11　YOLO-DDE 与不同模型可视化检测结果对比

示了其在模型轻量化和计算效率方面的显著优势。这些实验结果为未来在高性能、高效率目标检测模型的研究和应用提供了坚实的基础。

2.5　消融实验

为证明本文提出的各项改进措施对无人监控目标检测算法性能的提升作用,以 YOLOv8s 为基准算法,依此对其添加相应的改进措施,进行了一系列消融实验,充分验证本文提出的各个模块的有效性,实验结果如表 5 所示。

通过对表中的实验结果进行分析,可以得出以下结论:

1)Backbone 中引入 CEMA 模块,在几乎不增加计算量和参数量的情况下,模型的 mAP 得到提升。因此 CEMA 模块能够提高骨干网络细节特征提取能力。

2)在 YOLOv8s 的 Backbone 中同时应用 CEM 和 CED,模型的 mAP 显著提高。本文采用 GradCAM[33] 热力图方式对 Backbone 的特征提取进行可视化,如图 12 所示。

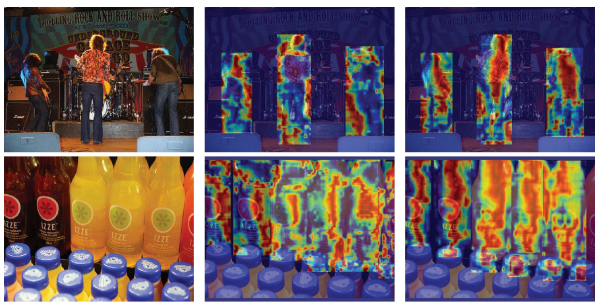
3)将 Neck 部分的特征融合结构替换为本文提出的 DFPN,提高特征融合能力。这证明了 DFPN 在多尺度特征融合方面具有更出色的能力。

4)当检测头采用本文提出的 DD-Head 时,加强下游任务处理能力,mAP 得到了显著的提升。这验证了 DD-Head 的在下游任务中的优越性。

5)在基线模型的基础上添加了所有本文改进措施。相较于基准算法,该方案在 mAP50 上提升了 1.8%,在 mAP50-95 上提升了 3.2%。参数量减少了 0.67 M,总浮点运算量增加了 3.1 G。该改进算法在计算量略微增加的

表 5　基于本文提出模型 YOLO-DDEs 在 PASCAL VOC 数据集上的消融实验

	CEMA	CED	DFPN	DD-Head	Parames/10^6	Flops/G	mAP50/%	mAP50-95/%
基线					11.14	28.7	82.5	62.1
实验 1	√				11.15	28.7	82.9(+0.4)	62.9(+0.8)
实验 2	√	√			10.69	28.4	83.4(+0.9)	63.3(+1.2)
实验 3			√		11.55	30.1	82.9(+0.4)	62.9(+0.8)
实验 4				√	10.51	30.6	83.8(+1.3)	64.2(+2.1)
本文	√	√	√	√	10.47	31.8	84.3(+1.8)	65.3(+3.2)



(a) 原图　(b) 基线热力图　(c) 改进后热力图

图 12　原骨干网络与改进骨干网络热力图可视化对比

同时，其他各项指标均优于 YOLOv8s 基准模型。

2.6　模型泛化性验证

为了深入评估改进模型的泛化能力，本研究进一步在 VisDrone，FAIR1M2.0，HIT-UAV 数据集上进行了模型泛化性能的验证。实验在 2.2 表 1 的条件下进行，实验结果如表 6 所示，改进后的模型在上述数据集上得到了显著的性能提升，这一结果充分证明了改进模型在处理复杂场景下小尺寸目标和遮挡问题时的优越性。

表 6　在多个数据集上的泛化性实验　%

数据集	模型	P	R	mAP50	mAP50-95
VisDrone	YOLOV8s	50.8	38.6	39.3	23.3
	YOLO-DDEs	51.4	38.9	40.3	24.3
FAIR1M 2.0	YOLOV8s	36.9	37	31.2	22.1
	YOLO-DDEs	39.9	38.3	33.8	24.1
HIT-UAV	YOLOV8s	90.3	69.9	79	49.2
	YOLO-DDEs	88.4	76.7	81.2	52.2

为了更直观的展示本文算法泛化能力，本文展示了在不同数据集上的可视化效果，如图 13 所示。

图 13　YOLO-DDE 与基线模型在不同数据集的效果对比

图 13 中 4 个数据集有很多尺度不一的目标,尤其小目标难以检测,且目标之间还有不同程度的遮挡现象,检测难度比较大。在前 3 个数据集中,YOLOv8 虽然将所有的目标都检测了出来,但是在精度上依然存在不足,而改进后的 YOLO-DDE 在精度上比 YOLOv8 有全方位的提升。在 FAIR1M2.0 数据集中 YOLOv8 只检测出了 3 个目标,并且还有误检的现象,而本文模型检测出了 7 个目标。

3 结　论

本文通过以 YOLOv8 为基础,结合了多个先进技术的优点,提出了一系列创新性的改进措施,成功开发了动态多尺度模型 YOLO-DDE。通过本文 CEMA 和 CED 模块在骨干网络的应用,以及本文 DFPN 架构和 DD-Head 模块在颈部和头部的应用,不仅显著提升了模型在复杂场景下的目标检测精度,而且保证了模型的参数量与计算量并没有大幅增加。

参考文献

[1] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]. Computer Vision-ECCV,2016.

[2] VIJAYAKUMAR A, VAIRAVASUNDARAM S. YOLO-based object detection models: A review and its applications [J]. Multimedia Tools and Applications,2024:1-40.

[3] GIRSHICK R. Fast R-CNN[C]. Santiago: IEEE International Conference On Computer Vision,2015.

[4] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. Venice: IEEE International Conference on Computer Vision,2017.

[5] 张阳婷,黄德启,王东伟,等.基于深度学习的目标检测算法研究与应用综述[J]. 计算机工程与应用,2023,59(18):1-13.

[6] 赵耘彻,张文胜,刘世伟.基于改进 YOLOv4 的无人机航拍目标检测算法[J]. 电子测量技术, 2023,46(8):169-175.

[7] ZHANG Z. Drone-YOLO: An efficient neural network method for target detection in drone images[J]. Drones, 2023,7(8):526.

[8] DING X, ZHANG X, MA N, et al. Repvgg: Making vgg-style convnets great again[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021.

[9] HUANGFU Z, LI S. Lightweight you only look once v8: An upgraded you only look once v8 algorithm for small object identification in unmanned aerial vehicle images[J]. Applied Sciences,2023,13(22):12369.

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

[11] LI H, LI J, WEI H, et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles [J]. ArXiv preprint arXiv: 2206.02424,2022.

[12] LI Y, FAN Q, HUANG H, et al. A modified YOLOv8 detection network for UAV aerial-image recognition[J]. Drones,2023,7(5):304.

[13] HAN K, WANG Y, TIAN Q, et al. Ghostnet: More features from cheap operations[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020.

[14] LOU H, DUAN X, GUO J, et al. DC-YOLOv8: small-size object detection algorithm based on camera sensor[J]. Electronics,2023, 12(10):2323.

[15] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile visionapplications [J]. ArXiv preprint arXiv:1704.04861,2017.

[16] WANG W, DAI J, CHEN Z, et al. Internimage: Exploringlarge-scale vision foundation models with deformable convolutions [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2023.

[17] OUYANG D, HE S, ZHANG G, et al. Efficient multi-scale attention module with cross-spatial learning[C]. Rhodes Island: CASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),2023.

[18] LIU S, QI L, QIN H, et al. Path aggregation network forinstance segmentation[C]. Salt Lake City: IEEE Conference on Computer Vision and Pattern Recognition,2018.

[19] DAI X, CHEN Y, XIAO B, et al. Dynamic head: Unifyingobject detection heads with attentions[C]. Nashville: IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021.

[20] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Munich: European Conference on Computer Vision(ECCV), 2018.

[21] LI X, HU X, YANG J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks [J]. Arxiv preprint arxiv: 1905.09646, 2019.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers

for image recognition at scale[J]. Arxiv preprint arxiv:2010.11929,2020.

[23] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems,2017,30.

[25] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Montreal: IEEE/CVF International Conference on Computer Vision,2021.

[26] DING X, ZHANG X, HAN J, et al. Scaling up your kernels to 31×31: Revisiting large kernel design in cnns[C]. New Orleans: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2022.

[27] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]. Seattle: IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020.

[28] ZHU X, HU H, LIN S, et al. Deformable convnets v2: More deformable, better results[C]. Long Beach: IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019.

[29] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes(voc) challenge[J]. International Journal of Computer Vision, 2010, 88: 303-338.

[30] DU D, ZHU P, WEN L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]. Seoul: IEEE/CVF International Conference on Computer Vision Workshops,2019.

[31] SUO J, WANG T, ZHANG X, et al. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection[J]. Scientific Data,2023,10(1):227.

[32] SUN X, WANG P, YAN Z, et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022,184:116-130.

[33] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-basedlocalization [C]. Venice: IEEE International Conference on Computer Vision,2017.

作者简介

张宇,博士,讲师,硕士研究生导师,主要研究方向为深度学习与计算机视觉。

E-mail:zytriumph@163.com

王延吉,硕士,主要研究方向为图形图像处理、目标检测。

E-mail:13963745853@163.com

马辉,硕士,工程师,主要研究方向为图像处理与网络优化。

E-mail:mahui@syuct.edu.cn

闫锴(通信作者),硕士,副教授,主要研究方向为深度学习与计算机视觉。

E-mail:18041382316@163.com

李大舟,博士,副教授,硕士研究生导师,主要研究方向为深度学习与计算机视觉、数据挖掘技术。

E-mail:lidazhou@syuct.edu.cn