本文引用格式：刘印,龚长友,徐国栋.基于改进 YOLOv10 的轻量化目标检测算法[J].自动化与信息工程,2025,46(1):29-35.
LIU Yin, GONG Changyou, XU Guodong. Lightweight object detection algorithm based on improved YOLOv10[J]. Automation & Information Engineering, 2025,46(1):29-35.

# 基于改进 YOLOv10 的轻量化目标检测算法

刘印[1]　龚长友[2]　徐国栋[1]
（1.西南林业大学，云南 昆明 650224
2.新疆生产建设兵团兴新职业技术学院，新疆 铁门关 841007）

**摘要：**针对目标检测算法部署在边缘设备的轻量化需求，提出一种基于改进 YOLOv10 的轻量化目标检测算法（CMD-YOLO 算法）。该算法利用跨尺度特征融合模块对 YOLOv10 算法的网络结构进行改进，减少了算法模型的参数量与计算量；采用基于 Mamba 的线性注意力机制改进的部分自注意力模块替换传统的部分自注意力模块，进一步降低了算法模型的参数量；利用空间深度转换卷积模块替换部分传统卷积模块，增强了算法模型对下采样细节信息的提取能力；利用动态上采样器 DySample 替换传统的上采样模块，在保持上采样精度的同时，降低了算法模型的计算延迟。实验结果表明，CMD-YOLO 算法与 YOLOv10-n 算法相比，在检测精度略微提升的同时，算法模型参数量降低了 30.5%，计算量下降了 19%，权重文件缩小了 29.3%，计算延迟减少了 8.8%，能够满足目标检测算法部署在边缘设备中的轻量化需求。

**关键词：**目标检测算法；YOLOv10 算法；跨尺度特征融合模块；Mamba 线性注意力机制；空间深度转换卷积模块；动态上采样器

**中图分类号：**TP391.41　　**文献标志码：**A　　**文章编号：**1674-2605(2025)01-0004-07

**DOI：**10.3969/j.issn.1674-2605.2025.01.004　　开放获取

# Lightweight Object Detection Algorithm Based on Improved YOLOv10

LIU Yin[1]　GONG Changyou[2]　XU Guodong[1]
(1.Southwest Forestry University, Kunming 650224, China
2.Bingtuan Xingxin Vocational and Technical College, Tiemenguan 841007, China)

**Abstract:** Aiming at the lightweight requirements of deploying object detection algorithms on edge devices, a lightweight object detection algorithm based on improved YOLOv10 (CMD-YOLO algorithm) is proposed. This algorithm utilizes a cross-scale feature fusion module to improve the network structure of YOLOv10 algorithm, reducing the parameter and computational complexity of the algorithm model; Adopting a Mamba based linear attention mechanism to improve the partial self attention module and replace the traditional partial self attention module, further reducing the parameter count of the algorithm model; Replacing some traditional convolution modules with spatial depth conversion convolution modules enhances the algorithm model's ability to extract downsampling detail information; By using the dynamic UpSampler DySample to replace the traditional upsampling module, the computational delay of the algorithm model is reduced while maintaining upsampling accuracy. The experimental results show that compared with the YOLOv10-n algorithm, the CMD-YOLO algorithm has slightly improved detection accuracy, reduced model parameters by 30.5%, decreased computational complexity by 19%, reduced weight files by 29.3%, and reduced computational latency by 8.8%, which can meet the lightweight requirements of object detection algorithm deployment in edge devices.

**Keywords:** object detection algorithm; YOLOv10 algorithm; cross-scale feature fusion module; Mamba-like linear attention mechanism; space to depth Conv module; dynamic UpSampler

## 0　引言

随着人工智能技术的快速发展，基于深度学习的目标检测算法被广泛应用于自动驾驶、机器人和无人机等场景。目前，基于深度学习的目标检测算法主要

分为一阶段和两阶段两类。两阶段目标检测算法（如R-CNN 算法）虽然检测精度较高[1]，但模型体积较大、检测时间较长，难以部署在性能受限的嵌入式设备中。而一阶段目标检测算法（如 YOLO 系列算法）不仅更轻量，且实时性更高、检测速度更快，检测精度也在迭代提升，适合部署在嵌入式设备中。

2016 年，YOLO 算法首次被提出[2]，其通过直接回归目标框为实时目标检测奠定了基础。2020 年，Ultralytics 公司推出了 YOLOv5 算法[3]，其采用 C3 特征提取结构，实现了轻量化性能。YOLOv5 算法分为 S、M、L 和 X 四个版本，可以根据项目需求选择合适的版本，以实现检测精度与速度的平衡。然而，YOLOv5 算法在基于锚框生成预测框时，会产生大量冗余框，增加了计算负担。2023 年，Ultralytics 公司又推出了 YOLOv8 算法[4]，其利用 C2f 获得更多的梯度流信息，并采用无锚方法加快了检测速度，减少对计算能力的需求。然而，YOLOv8 算法需要后处理操作与非极大值抑制依赖，限制了检测速度的提升。

2024 年 5 月，WANG 等[5]提出了 YOLOv10 算法，通过一致的双重分配策略，解决了对后处理操作的需求，消除了非极大值抑制的依赖，实现了真正的端到端检测，减少了算法检测时间，达到了迄今为止最快的检测速度。YOLOv10 算法发布时间较短，目前针对该算法的改进研究较少，尤其是轻量化改进的相关研究更少。

本文提出一种基于改进 YOLOv10 的轻量化目标检测算法，在提升检测精度的同时，减少算法的参数量与计算量，以满足部署在嵌入式设备的轻量化需求。

## 1　YOLOv10 算法

通过调整网络的宽度和深度，YOLOv10 算法可得到网络结构相同但规模不同的 6 个版本[6]，规模由大到小分别为 YOLOv10-x、YOLOv10-l、YOLOv10--b、YOLOv10-m、YOLOv10-s 和 YOLOv10-n。YOLOv10 算法主要由主干网络（Backbone）、颈部网络（Neck）和头部网络（Head）3 部分组成[7]，网络结构如图 1 所示。

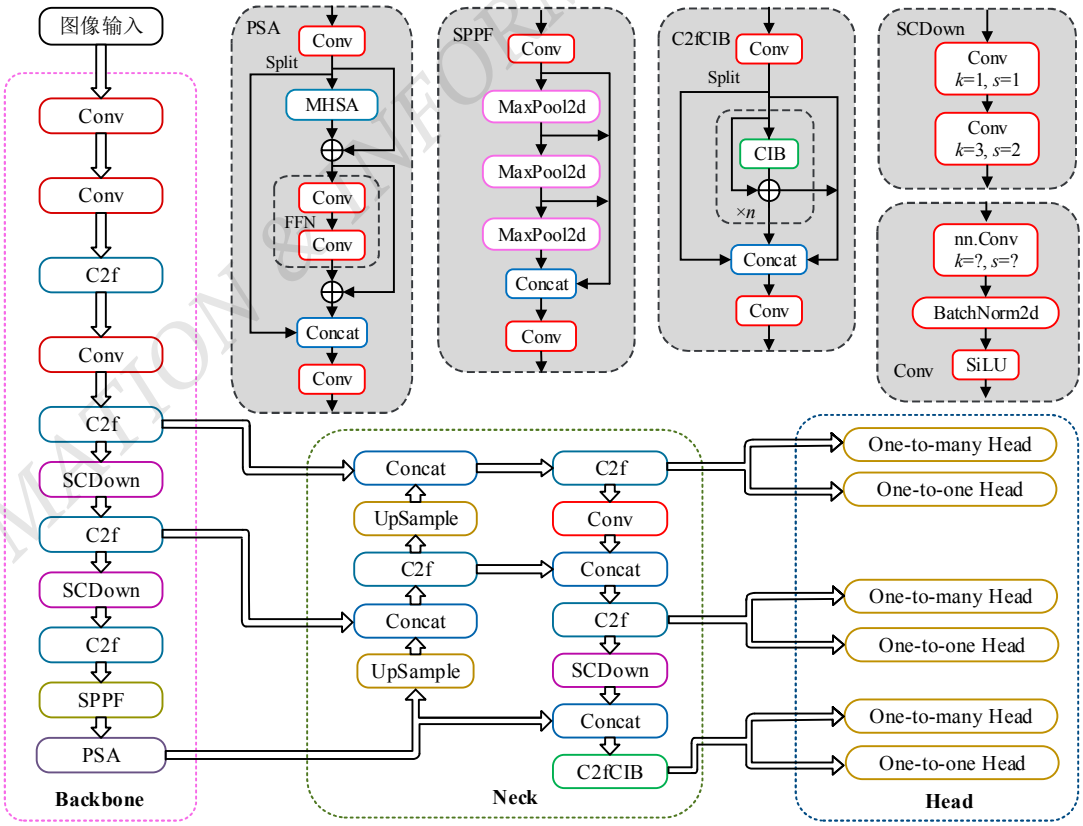


图 1　YOLOv10 算法网络结构

在 YOLOv10 算法网络结构中：Conv 为卷积模块，*k* 为卷积核大小，*s* 为步长；C2f 为包含 2 个 Conv 和 *n* 个带有残差连接的卷积模块；SCDown 为空间-通道分离下采样模块；SPPF 为快速空间金字塔池化模块；Split 为特征分层；PSA 为部分自注意力模块；Concat 为特征连接模块；UpSample 为上采样模块；MaxPool2d 为最大池化层；MHSA 为多头自注意力模块；FFN 为前馈神经网络；CIB 为紧凑倒置块；BatchNorm2d 为批归一化层；SiLU 为激活函数。

YOLOv10 算法团队提出了整体效率-精度驱动的模型设计策略，从效率和精度的角度优化模型架构。为了提高效率，YOLOv10 算法引入了轻量化分类头、SCDown 模块、C2fCIB 模块，以减少计算冗余，实现更高效的算法架构。为了提高精度，YOLOv10 算法采用大核卷积并引入 PSA 模块，在低成本的情况下，提升了模型性能。但在实际工业生产场景中，特别是在高实时性要求的嵌入式设备上，模型的计算量与参数量直接影响其硬件部署。虽然 YOLOv10-n 已经是 YOLOv10 算法的最小版本，但其计算量与参数量相较于其他算法仍需进一步改进。

## 2 CMD-YOLO 算法

本文对 YOLOv10 算法网络结构的改进主要包括：利用跨尺度特征融合模块（cross-scale feature fusion module, CCFM）对 YOLOv10 算法的网络结构进行重构；在 YOLOv10 算法的主干网络与颈部网络中，利用空间深度转换卷积（space to depth conv, SPDConv）模块替换步长为 2 的传统卷积（Conv）模块；在 YOLOv10 算法的主干网络中，利用基于 Mamba 的线性注意力机制（Mamba-like linear attention, MLLA）改进的部分自注意力（PSAMLLA）模块，替换传统的部分自注意力模块（PSA）；在 YOLOv10 算法的颈部网络中，利用动态上采样器 DySample 替换传统的上采样（UpSample）模块。改进后的 YOLOv10（CMD-YOLO）算法网络结构如图 2 所示。

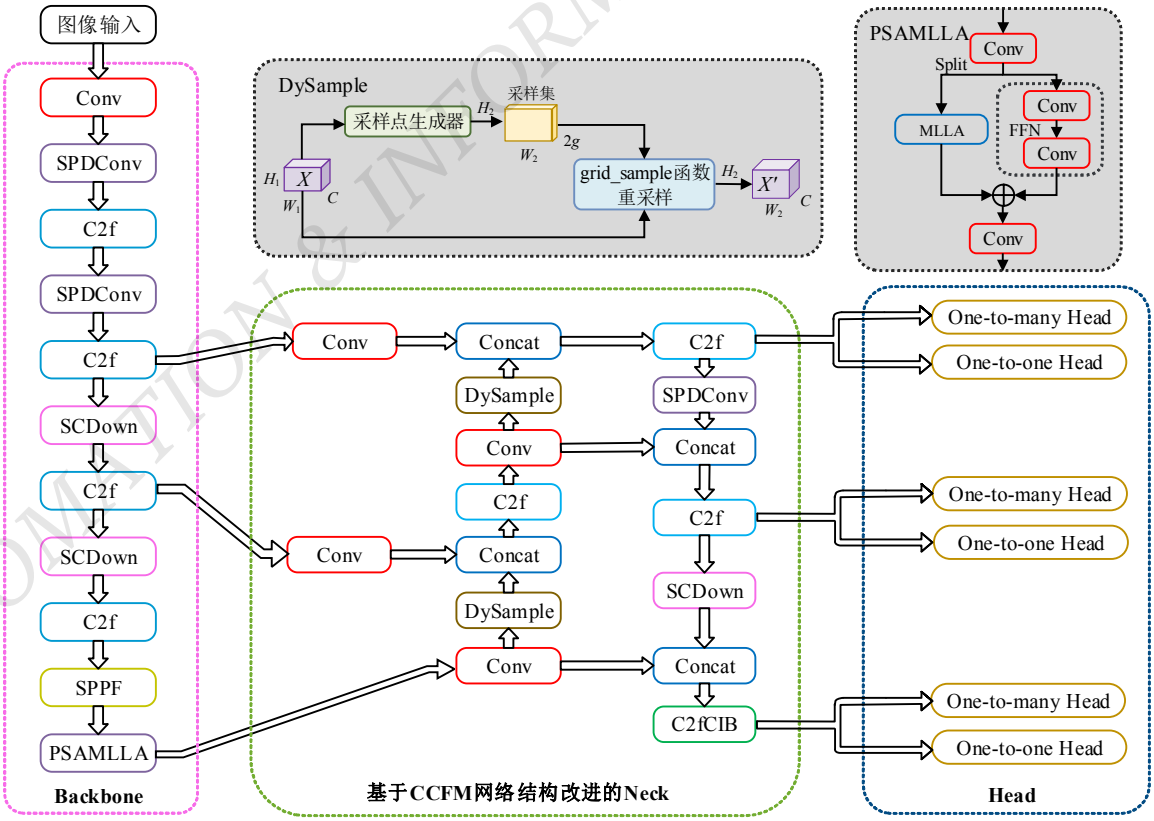


图 2 CMD-YOLO 算法网络结构

## 2.1 CCFM

多尺度特征是影响目标检测算法模型性能的重要因素之一。本文利用 CCFM 对 YOLOv10 算法的颈部网络结构进行改进，通过融合操作整合不同尺度的特征，以增强网络对尺度变化的适应性和对小尺度目标的检测能力。CCFM 网络结构如图 3 所示。

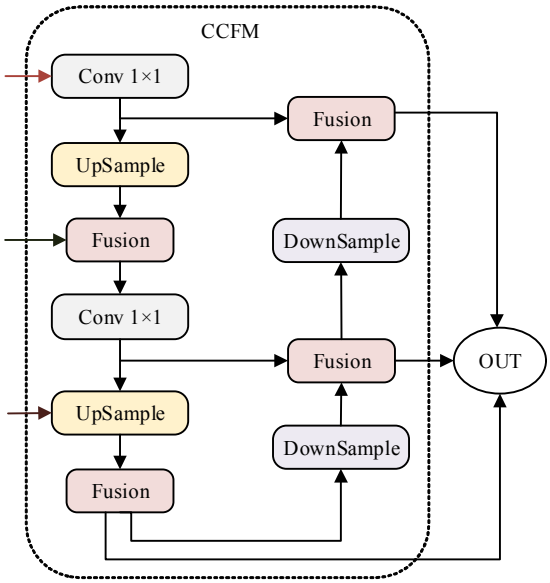


图 3　CCFM 网络结构

首先，分别对 YOLOv10 算法主干网络输出的 3 个不同尺度的初始特征图进行一次 1×1 卷积操作，以提取、融合初始特征图信息，这些卷积层均采用轻量化的结构设计，以降低算法模型的计算复杂度；然后，在特征融合中段增加一次 1×1 卷积操作，实现不同尺度特征图的横向融合，有效增加了算法模型的非线性特征表达能力；接着，横向融合后的特征图通过一系列自下而上的卷积层传递，进一步增强特征表达能力；最后，经过处理后的特征图被输入到检测头进行预测。CCFM 可有效整合细节特征和上下文信息，从而提高算法模型性能。

## 2.2 PSAMLLA 模块

由于 YOLOv10 算法中 PSA 模块的计算量较大，CMD-YOLO 算法的主干网络用 PSAMLLA 模块替换传统的部分自注意力模块（PSA），在保持检测精度的同时，进一步降低算法模型的计算量。

PSAMLLA 模块的网络结构如图 2 所示。首先，利用卷积层提取图像特征；然后，将图像特征分为两部分：一部分图像特征输入 MLLA 模块，在保持并行计算和快速推理的同时，提供必要的位置信息；另一部分图像特征输入 FFN 模块，进行特征变换和维度扩展；最后，将这两部分输出的特征合并，通过卷积层进行特征提取、融合，生成该层网络的输出。PSAMLLA 模块在处理图像数据时，能有效提取、融合特征信息。此外，CMD-YOLO 算法将原始 Mamba 线性注意力算法网络中需要预输入的图像大小改为实时计算，在保持相似浮点运算数的同时，提高了算法模型性能。

MLLA 模块网络结构如图 4 所示。

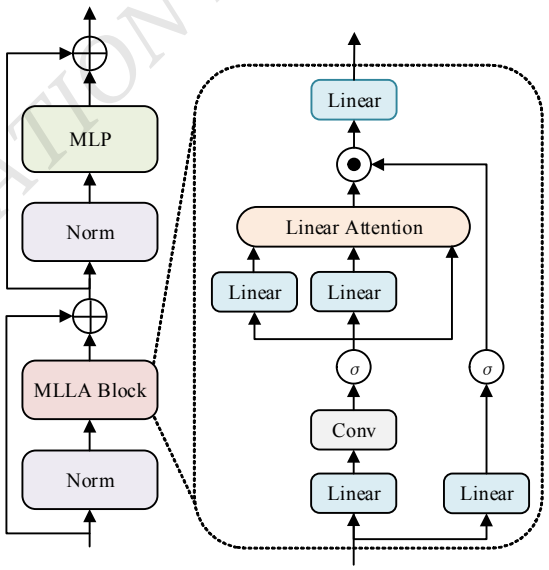


图 4　MLLA 模块网络结构

## 2.3 SPDConv 模块

YOLO 系列算法的传统步长卷积模块在提取小目标特征时，连续下采样会导致细节信息丢失，降低算法模型性能。本文在 YOLOv10 算法的主干网络与颈部网络中，利用 SPDConv 模块替换步长为 2 的传统 Conv 模块。SPDConv 模块由一个空间到深度（SPD）层和一个非步长卷积（Conv）层组成[8]，原理图如图 5 所示。

SPDConv 模块通过对输入图像进行下采样和拼接，利用卷积操作提取特征，在不大幅降低算法精度

的前提下，减少计算复杂度，进一步提高了算法模型性能。

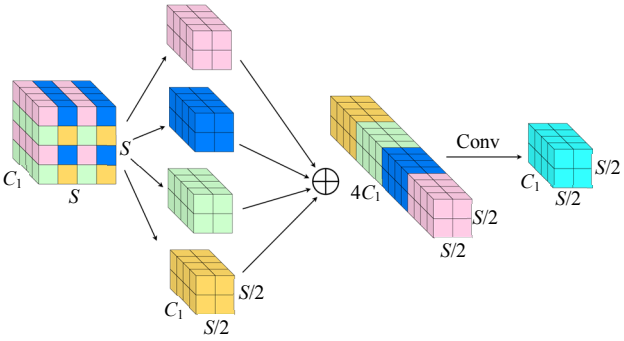


图 5　SPDConv 模块原理图

### 2.4　动态上采样器 DySample

针对传统上采样方法无法提供充足细节信息的问题，CMD-YOLO 算法利用轻量且高效的动态上采样器 DySample，替换原颈部网络中的传统上采样器。动态上采样器 DySample 通过动态调整采样位置，使算法模型能够更灵活地处理图像数据，并增强图像的细节和纹理信息。

动态上采样器 DySample 的上采样流程如图 2 所示。给定一个大小为 $H_1\times W_1\times C$ 的特征映射 $X$ 和大小为 $H_2\times W_2\times 2g$ 的点采样集，由 grid_sample 函数利用点采样集中的位置信息，将特征映射 $X$ 重新采样，生成大小为 $H_2\times W_2\times C$ 的特征映射 $X'$。DySample 采样点生成器相比基于内核的动态上采样器，不需要高分辨率引导特征输入，在提升算法模型精度的同时，降低了参数量，且不影响帧率，具有更小的参数量、浮点运算数、GPU 内存占用和延迟。

## 3　实验与结果分析

### 3.1　实验数据集

本实验采用 VOC2012 数据集进行 CMD-YOLO 算法的性能验证。该数据集包含了 17 125 幅标注图像，均为 JPG 格式，涵盖了 20 种常见的物体类别。VOC2012 数据集按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集[9]，其中训练数据集包含 13 700 幅图像，验证集包含 1 712 幅图像，测试集包含 1 713 幅图像。

### 3.2　实验环境与评估指标

实验运行环境配置：操作系统为 Windows 11；GPU 为 NVIDIA RTX 4060Ti，显存 为 16 GB；CPU 为 Intel Core i5-13600KF；开发环境使用 CUDA 12.6，PyTorch-GPU 2.4.1，Python 3.9.19 版本，采用 PyCharm 进行训练。数据集训练参数配置如表 1 所示。

表 1　数据集训练参数配置

| 优化器 | 批大小/幅 | 循环次数/次 | 线程/个 | 学习率 | 输入/像素 | AMP |
|---|---|---|---|---|---|---|
| SGD | 32 | 300 | 8 | 0.01 | 640 | False |

采用 YOLO 系列算法的评估指标，包括平均精度（mean average precision, mAP）、精度（precision, P）、计算量（GFLOPs）、召回率（recall, R）、参数量（Params）对 CMD-YOLO 算法进行性能评估[10]。为确保实验的公平性，所有检测算法网络均不使用官方的预训练权重，都从零开始训练。本实验以数据集所有类的 IoU 阈值为 0.5 时的平均精度 mAP@0.5、IoU 阈值为 0.5:0.05:0.95 时的平均精度 mAP@0.5:0.95 来评估模型的检测精度。精度 $P$ 和召回率 $R$ 的计算公式分别为

$$P=\frac{TP}{TP+FP} \tag{1}$$

$$R=\frac{TP}{TP+FN} \tag{2}$$

式中：$TP$ 为真阳性，$FP$ 为假阳性，$FN$ 为假阴性。

mAP 是所有样本 AP 的平均值，其值越高表示模型性能越好。AP 与 mAP 的计算公式分别为

$$P_{\mathrm{AP}}=\int_0^1 PRdR \tag{3}$$

$$P_{\mathrm{mAP}}=\frac{1}{n}\sum_{i=1}^{n}P_{\mathrm{AP}(i)} \tag{4}$$

通过消融实验与对比实验，评估本文提出的 CMD-YOLO 算法模型性能[11]。

### 3.3　消融实验

本文分别采用 CCFM、PSAMLLA、SPDConv、DySample 作为独立的变量模块来验证 YOLOv10 算

法的改进效果，消融实验结果如表 2 所示。其中，权重文件用来衡量算法模型的大小，计算量用来衡量算法模型的计算负担，参数量用来评估算法模型的参数量。

表 2　CMD-YOLO 不同模块的消融实验

| 算法 | mAP@0.5/% | mAP@0.5:0.95/% | 参数量/B | 计算量/GFLOPs | 计算延迟/ms | 权重文件/MB |
|---|---|---|---|---|---|---|
| YOLOv10-n | 59.8 | 44.9 | 2 714 840 | 8.4 | 3.4 | 5.8 |
| YOLOv10-n+CCFM | 60.5 | 45.1 | 1 932 952 | 7.1 | **3.1** | 4.2 |
| YOLOv10-n+CCFM+PSAMLLA | **60.7** | **45.2** | 1 915 928 | 7.1 | 3.3 | 4.2 |
| YOLOv10-n+CCFM+PSAMLLA+SPDConv | 60.6 | 45 | 1 882 648 | **6.8** | 3.3 | **4.1** |
| YOLOv10-n+CCFM+PSAMLLA+SPDConv+DySample | **60.7** | **45.2** | **1 886 808** | **6.8** | **3.1** | **4.1** |

由表 2 可知：YOLOv10 算法在引入 CCFM 模块后，参数量降低了 28.8%，计算量降低了 15.5%，权重文件缩小了 27.6%，计算延迟减少了 0.3 ms，且检测精度略有提升（mAP@0.5 提高了 0.7%，mAP@0.5:0.95 提高了 0.2%）；利用 PSAMLLA 模块替换 PSA 模块后，参数量降低了 0.9%，检测精度略微提高（mAP@0.5 提高了 0.2%，mAP@0.5:0.95 提高了 0.1%），计算延迟增加了 0.2 ms；利用 SPDConv 模块替换主干网络和颈部网络中步长为 2 的 Conv 模块后，虽然检测精度略有降低（mAP@0.5 降低了 0.1%，mAP@0.5:0.95 降低了 0.2%），但参数量与计算量大幅下降（参数量降低了 1.7%，计算量降低了 4.2%），且权重文件缩小了 0.1 MB；利用动态上采样器 DySample 替换颈部网络中的上采样器后，尽管算法参数量提高了 0.2%，但 mAP@0.5 提高了 0.1%，mAP@0.5:0.95 提高了 0.2%，计算延迟减少了 0.2 ms。综上所述，本文提出的 CMD-YOLO 算法相较于 YOLOv10-n 算法，在不影响原有检测精度的同时，算法模型参数量下降了 30.5%，计算量下降了 19%，权重文件缩小了 29.3%，计算延迟减少了 8.8%。

### 3.4　对比实验

为验证 CMD-YOLO 算法的同级性能，将 CMD-YOLO 算法与 YOLOv5-n、YOLOv6-n、YOLOv8-n、YOLOv10-n 算法进行对比实验，结果如表 3 所示。

表 3　CMD-YOLO 与主流算法轻量化版本对比实验

| 算法 | 参数量/B | 计算量/GFLOPs | 计算延迟/ms | 权重文件/MB |
|---|---|---|---|---|
| YOLOv5-n | 2 512 364 | 7.2 | 3.2 | 5.3 |
| YOLOv6-n | 4 240 124 | 11.9 | 3.1 | 8.7 |
| YOLOv8-n | 3 014 748 | 8.2 | 3.8 | 6.3 |
| YOLOv10-n | 2 714 840 | 8.4 | 3.4 | 5.8 |
| CMD-YOLO | **1 886 808** | **6.8** | **3.1** | **4.1** |

由表 3 可知：CMD-YOLO 算法在参数量、计算量及权重文件方面均优于其他算法；在计算延迟方面，其与 YOLOv6-n 算法并列榜首，优于其他算法。综上所述，CMD-YOLO 算法不仅在参数量与检测精度之间取得了平衡，还在检测速度上表现出优势，能够满足边缘嵌入式设备在目标检测场景下的轻量化部署需求。

## 4　结论

本文针对目标检测算法在性能受限的边缘嵌入式设备中部署困难的问题，提出了一种基于改进 YOLOv10 的轻量化目标检测算法——CMD-YOLO 算法。首先，利用 CCFM 对 YOLOv10 算法的网络结构进行改进，使算法模型的参数量与计算量均有所下降，在加快检测速度的同时略微提升了检测精度；利用 PSAMLLA 模块替换 PSA 模块，在保持相似的浮点运算数的同时，提升了算法模型的检测精度；利用 SPDConv 模块替换步长为 2 的传统卷积模块，进一步降低了算法模型的参数量与计算量；利用动态上采样

器 DySample 替换颈部网络中的传统上采样器，在略微增加算法模型参数量的同时，提升了检测精度，降低了计算延迟。经实验验证，CMD-YOLO 算法可以满足边缘嵌入式设备部署目标检测算法的轻量化需求。但现实环境复杂多样，本文所提算法无法适用所有场景，且该算法网络结构中仍包含部分传统卷积模块，有待进一步优化。



## 参考文献

[1] 王珂,赵慧,张成,等.基于改进的 YOLOv5 人脸口罩识别算法[J].信息化研究,2022,48(6):38-45.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

[3] NELSON J, SOLAWETZ J. YOLOv5 is here: State-of-the-art object detection at 140 FPS[EB/OL]. Roboflow Blog, (2020-06-26)[2025-01-02]. https://blog.roboflow.com/yolov5-is-here.

[4] 胡峻峰,李柏聪,朱昊,等.改进 YOLOv8 的轻量化无人机目标检测算法[J].计算机工程与应用,2024,60(8):182-191.

[5] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection[J]. arXiv preprint arXiv:2405.14458, 2024.

[6] 牛鑫宇,毛鹏军,段云涛,等.基于 YOLOv5s 室内目标检测轻量化改进算法研究[J].计算机工程与应用,2024,60(3):109-118.

[7] 鲁鑫,郭业才.基于改进 YOLOv4 的烟条拉线头缺陷检测[J].科学技术与工程,2022,22(21):9199-9206.

[8] 娄瑶迪,岳俊峰,周迪斌,等.基于 Efficient-YOLO 的轻量化轴承缺陷检测[J].计算机系统应用,2024,33(2):265-275.

[9] 马斌,张亚.针对电动车头盔佩戴的 YOLOv5s 改进算法研究[J].无线互联科技,2023,20(15):90-93.

[10] 黄家興,南新元,张文龙,等.基于改进 YOLOv5 的轻量化口罩检测算法研究[J].计算机仿真,2023,40(5):541-547.

[11] 苏山杰.基于深度学习的复杂道路场景下遮挡中小目标检测算法研究[D].重庆:重庆交通大学,2023.

作者简介：

刘印，男，1998 年生，在读硕士研究生，主要研究方向：深度学习目标检测。E-mail: 553841609@qq.com

龚长友，男，1996 年生，硕士研究生，主要研究方向：混合动力汽车能量管理。E-mail: changyou_key@163.com

徐国栋，男，1976 年生，硕士研究生，副教授，主要研究方向：神经网络。E-mail: ynxugd@126.com

（上接第 28 页）

[18] LU S, CHEN M, LIU Y, et al. Adaptive NN tracking control for uncertain MIMO nonlinear system with time-varying state constraints and disturbances[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022,34(10):7309-7323.

[19] ZHANG Y, SUN J, LIANG H, et al. Event-triggered adaptive tracking control for multiagent systems with unknown disturbances[J]. IEEE Transactions on Cybernetics, 2018,50(3):890-901.

[20] WANG X, WANG H, HUANG T, et al. Neural-network-based adaptive tracking control for nonlinear multiagent systems: The observer case[J]. IEEE Transactions on Cybernetics, 2021, 53(1):138-150.

[21] SUN Y, CHEN B, LIN C, et al. Adaptive neural control for a class of stochastic nonlinear systems by backstepping approach[J]. Information Sciences, 2016,369:748-764.

[22] HADDAD W M, CHELLABOINA V S, NERSESOV S G. Impulsive and hybrid dynamical systems: stability, dissipativity, and control[M]. Princeton University Press, 2006.

[23] KHALIL HK. Nonlinear systems third edition[M]. Upper Saddle River, NJ: Prentice Hall, 2002.

[24] CHEN B, LIU X, LIU K, et al. Direct adaptive fuzzy control of nonlinear strict-feedback systems[J]. Automatica, 2009,45(6): 1530-1535.

[25] CHEN M, GE S S. Adaptive neural output feedback control of uncertain nonlinear systems with unknown hysteresis using disturbance observer[J]. IEEE Transactions on Industrial Electronics, 2015,62(12):7706-7716.

作者简介：

罗振发，男，1999 年生，硕士研究生，主要研究方向：自适应控制、神经网络。E-mail: m18379124404@163.com