**引用格式：**陈路, 李阳, 周昊昱, 等. ACDet：强化自我注意力机制的药品包装轮廓检测方法 [J]. 电子科技大学学报, 2025, 54(6): 935-944.
CHEN L, LI Y, ZHOU H Y, et al. ACDet: Enhanced self-attention mechanism for pharmaceutical packaging contour detection[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(6): 935-944.

# ACDet：强化自我注意力机制的药品包装轮廓检测方法

陈 路[1]，李 阳[1]，周昊昱[2]，王钧慷[3]，钱伟中[1]，张新昱[1]，陈丽竹[1]，高 勇[4*]

（1. 电子科技大学 航空航天学院，成都 611731；2. Meta Platforms Inc., Menlo Park 94025；3. 四川省医学科学院·四川省人民医院，成都 610072；
4. 电子科技大学 电子科学与工程学院，成都 611731）

**摘要：**提出了一种基于卷积神经网络的物品矢量检测识别方法：ACDet（self-attention and concatenation based detector），旨在解决照度变化下密集无序药品包装轮廓的高效检测问题。该方法采用组合图像增强技术提升模型学习物品外观特征的能力，对计算模块 C2F-A（C2F with attention）采用多条梯度流输出来进行多维度的强化自我注意力增强，包括特征维度和空间维度。设计的 WConcat（weighted concatenation）模块可以对不同层次的特征图进行加权拼接并捕捉更关键的特征图，从而使网络具备更好的认知能力。在医药案例数据集（cancer pathological and pharmaceutical dataset, CPPD）实验中实现了 81.0% 的 mAP（mean average precision），79.5% 的 SmoothmAP，平均领先其他 YOLO（you only look once）架构的模型 5.5%～16.6%，在公开数据集平均领先 0.7%～6.9%。同时，零样本测试中复核成功率达到 99.9%。研究结果显示，ACDet 能克服复杂检测场景难题，实现网络鲁棒性提升及轻量化，为工业智能化生产提供了技术支持。

**关键词：**YOLO；药品包装轮廓；动态照度；视觉检测
**中图分类号：**TP309.7 **文献标志码：**A DOI: 10.12178/1001-0548.2024132

# ACDet: Enhanced self-attention mechanism for pharmaceutical packaging contour detection

CHEN Lu[1], LI Yang[1], ZHOU Haoyu[2], WANG Junkang[3], QIAN Weizhong[1],
ZHANG Xinyu[1], CHEN Lizhu[1], and GAO Yong[4*]

(1. School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China;
2. Meta Platforms Inc., Menlo Park 94025, USA; 3. Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu 610072, China;
4. School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

**Abstract:** This article proposes ACDet (self-attention and concatenation based detector), an object vector detection and recognition method based on convolutional neural networks. This method aims to efficiently detect dense and unordered pharmaceutical packaging contours under varying lighting conditions. By employing combined image enhancement techniques, the method enhances the model's ability to learn the appearance features of objects. The C2F-A (coarse-to-fine with attention) computational module utilizes multiple gradient flows for multidimensional self-attention enhancement, encompassing both feature and spatial dimensions. The WConcat (weighted concatenation) module facilitates weighted concatenation of various levels of feature maps, capturing more critical features, thereby enhancing the network's cognitive ability. In experiments on the CPPD (cancer pathological and pharmaceutical dataset) for pharmaceutical cases, ACDet achieved 81.0% mAP (mean average precision) and 79.5% SmoothmAP, outperforming other YOLO (you only look once) architecture models by an average of 5.5% to 16.6%, and leading by 0.7% to 6.9% on public datasets. Additionally, zero-shot testing achieved a review success rate of 99.9%. The research results suggest that the proposed ACDet can overcome complex detection scenarios, enhance network robustness, and support intelligent industrial production.

**Key words:** YOLO; pharmaceutical packaging contour; dynamic lighting; visual detection

---

收稿日期：2024−06−03
基金项目：四川省科技计划（24SYSX0210）；跨域飞行交叉技术实验室项目（2024-KF03004）
作者简介：陈路，博士生，主要从事自动控制、计算机视觉检测等方面的研究。
*通信作者 E-mail: gaoyong@uestc.edu.cn

视觉检测能确保产品的质量和一致性，在当今的工业制造和质量控制中至关重要。随着自动化技术的进步，视觉检测系统也变得更加智能和高效，可提高生产线的效率和可靠性。在深度学习和 AI 的发展影响下，视觉检测技术变得更加精准且检测效率进一步提高，大幅提升了复杂场景中的检测能力和适应性。在本文的医药案例中，不同种类及数量的药品是根据需求全自动配送到不同楼层药房的，类似于机场的行李托运。通过药品包装轮廓检测来复核每次配送的药物数量及种类，关乎着配送效率与安全[1]。医药行业对复核成功率要求极高，因为它是药品自动配送中最后的检验步骤，成功率需达到 99.9%。药品包装轮廓外观和检测环境都具有多样性和复杂性[2]，这些因素都加剧了检测的困难。现有的物品包装轮廓检测方法难以完全适应本文案例，因此迫切需要优化物品通用视觉检测方法，以匹配智能制造的高标准需求。

当前，图像处理与计算机视觉技术的迅速进步为各类物品检测带来了新契机。深度学习技术的崛起，特别是卷积神经网络（convolutional neural network, CNN）的成熟应用，显著提升了图像处理的准确性和效率。在物品外观检测中，这些技术能够自动学习图像中的关键特征，如形状、纹理和结构等，从而增强检测的精度和适应性，有效弥补传统方法在复杂包装盒图像处理上的局限性。此外，计算机视觉领域中的注意力机制对提高检测任务的精度和效率也起到了关键作用。引入注意力机制能使模型更聚焦于物品的关键特征，进而降低误检率。同时，迁移学习和增强学习等技术的融入，也为物品检测领域注入了新的活力，提升了模型的泛化能力和检测效果。在本文的医药案例中结合这些先进的技术来研究解决方案，可以使药物包装检测更加准确、灵活地适应各种场景，为自动化配送药品提供有力保障，不仅降低了人工检测成本和人为错误的可能性，也提升了医疗自动化行业的智能化与可靠性。

本文提出的系统结构包括数据增强与 ACDet 两个模块。基于 CPPD 数据集的特点对输入模型的数据集进行增强，包括亮度增强和额外组合增强，以此来提升模型的多角度学习能力。ACDet 的 C2F-A 模块通过多维度的强化自我注意力机制对拼接后的特征图像进行增强，WConcat 模块对多层次的图像特征进行加权融合。

# 1 相关研究

## 1.1 基于深度学习的目标检测算法发展

随着深度学习的崛起，尤其是卷积神经网络（CNN）的普及，目标检测领域经历了重大变革。深度学习能通过学习标注数据集中的目标特征来识别图像，无须人工设计特征。CNN 的设计包含 Backbone（主干网络）、Neck（特征融合层）和 Head（头部网络）3 部分。对于不同检测平台和算力需求，可选择适合的 Backbone，如低算力平台可选用 SqueezeNet[3]、MobileNet[4-6] 或 ShuffleNet[7-8]，而高算力平台则可采用 VGG[9]（visual geometry group）、ResNet[10] 或 DenseNet[11]。Neck 部分由多个上下路径构成，典型代表包括 FPN[12]（feature pyramid network）、PANet[13]（path aggregation network）和 BiFPN[14]（bi-directional feature pyramid network）。Head 部分则通常分为单阶段和双阶段目标检测模型。

典型的双阶段模型如 R-CNN（region-based convolutional neural network）系列，其主要工作原理是采用选择性搜索（selective search, SS）算法生成候选框，将其作为输入，通过卷积神经网络提取其特征，最后执行边界框回归及类别预测。但该算法存在明显的耗时长、计算量较大、占用内存较大等问题[15]。为了解决上述问题，文献 [16] 于 2015 年提出了 Fast R-CNN，其核心思路为将整张图片作为特征提取网络的输入，再利用选择性搜索算法生成的候选框映射到特征图上，同时引入感兴趣区域池化层（ROI pooling），提升了检测速度与精度。Faster R-CNN 改变了生成候选框的逻辑，将选择性搜索算法替换为区域建议网络，减少了大量的训练时间[17]。R-FCN 在全卷积网络架构中引入位置敏感的得分图，实现了高效且准确的目标定位和分类，提升了检测效率与精度[18]。Libra R-CNN 通过引入平衡注意模块、平衡采样模块和 IoU 平衡损失，显著提升了检测精度和鲁棒性。它通过对多尺度特征进行综合利用，解决了样本不平衡和特征不一致的问题，在复杂场景中表现出色[19]。

单阶段模型如 YOLO 系列，该模型直接将输入图像划分为 7×7 大小的网格，同时利用网格中的预测信息执行预测和边界框回归。也利用了卷积神经网络进行图像特征提取，大幅度提升了检测速度[20-22]。SSD（single shot multibox detector）是一种

实时目标检测模型，通过在不同尺度的特征图上同时进行分类和定位，实现了高效的多尺度目标检测。其一阶段检测架构避免了候选区域生成步骤，显著提高了检测速度，同时保持了较高的准确性[23]。文献 [24] 提出了 RetinaNet，通过引入焦点损失（focal loss）有效解决了目标检测中正负样本不平衡的问题，从而显著提高了检测精度。其单阶段检测架构兼具速度和性能优势，在处理复杂场景和小目标检测方面表现优异。近年来，因为对头部的检测更简单、灵活、快速，无锚单阶段目标探测器也成了新的研究方向[25]。2019 年文献 [26] 提出 CenterNet，通过直接预测目标中心点的位置和大小，实现了高效且精确的目标检测。简化了检测流程，避免了烦琐的候选框生成步骤。CornerNet 通过预测目标边界的角点来进行目标检测，其创新的角点检测机制和角点组合策略显著提升了检测精度，特别在处理目标重叠和复杂背景方面表现出色[27]。FCOS（fully convolutional one-tage object detection）则采用了中心度评分和多级预测特征直接预测每个像素点的边界框和类别[28]。

### 1.2　医药案例检测场景存在的问题

尽管深度学习算法在提高目标检测性能方面取得了显著进展，但在药品包装轮廓检测案例中的实际场景下仍然存在如下问题。

1）药物的类型数不胜数（几千种），并且某一种药盒有多种摆放方式，即一种药品有多种形态的存在。目标的多样性使得模型难以通用化检测所有种类的药盒。

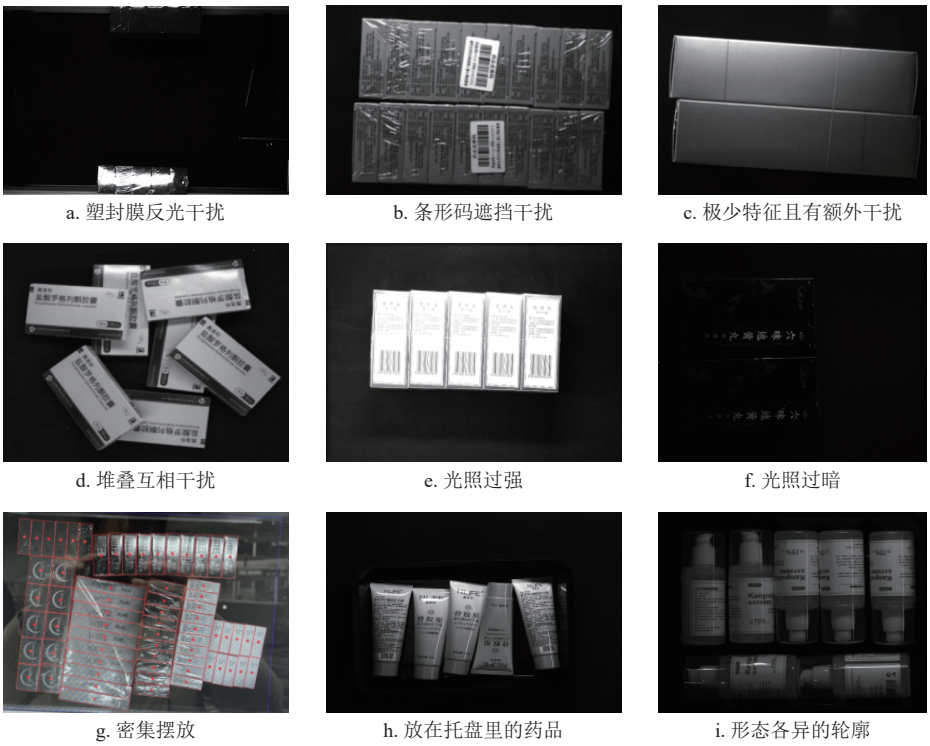
2）观察包装的角度受限：部分图像受限于观察角度，仅呈现了药品包装的某一侧面或角度。这增加了在有限视野内进行药品包装全方位识别的难度，因此要求模型具备处理单一视角下的物品的检测和识别能力。

3）为了满足防水和运输的需求，药品包装表面会有塑封膜等额外的干扰类包装，会导致反光等各类干扰药品包装轮廓检测的复杂情况。

4）工业场景是开放的环境（不同时段工厂接受的日照不同，不同区域灯光部署也不同），光照条件会一直变化。且因图像采集设备差异、工厂采集区域自动配送设备振动的干扰，成像质量也不尽相同。

5）由于全自动配送，药品进入检测区域的形态是随机多样的。如由于全自动配送途中形成的药品互相堆叠，会造成检测时药品之间相互干扰，且有排列密集的药盒甚至是装在托盘里的药盒，这些干扰项都会增加检测难度。

以上具体情况如图 1 所示。另一方面，深度学习模型的训练需要大规模标注的数据集来支撑，而在药物包装领域，获取高质量标注的数据集可能是一项昂贵且耗时的任务。



a. 塑封膜反光干扰　b. 条形码遮挡干扰　c. 极少特征且有额外干扰

d. 堆叠互相干扰　e. 光照过强　f. 光照过暗

g. 密集摆放　h. 放在托盘里的药品　i. 形态各异的轮廓

图 1　药品包装轮廓检测被干扰的 9 种情况

因此，当前药品包装轮廓检测领域需要进一步解决的问题包括：提高模型的鲁棒性；加强对不同场景和条件下的适应性；寻找更有效的数据增强方法。未来的研究方向应当在解决这些问题的基础上来推动目标检测技术在药品包装和其他领域的更广泛应用。

## 2 密集药盒快速识别网络 ACDet

在工业应用场景下，YOLO 模型能很好地兼顾精度和速度两方面，因此受到广泛的应用。YOLOv8 主要从两个方面进行了改进，分别是多层级特征融合模块 C2F-A、加权拼接融合模块 WConcat。该模块都聚焦于使不同层级的特征来实现完美融合，改进的模型基本结构如图 2 所示。

### 2.1 YOLO 模型分析

#### 2.1.1 YOLOv8 模型分析

在 YOLOv8 中，Head 部分在边界框回归中换用了 Anchor-Free 的回归模式。Anchor-Free 回归模式消除了传统目标检测中需要预定义的先验框（Anchor）的需求。由于不再受到先验框的限制，Anchor-Free 模型对目标的形状和大小变化更具有适应性。这使得模型更能够捕捉各种目标的特征，尤其是在目标尺度变化较大的场景中。还能够提供更精准的目标定位，特别是在目标边界模糊或者存在遮挡的情况下。这对于一些要求高精度检测的任务非常重要。此外，Head 部分还采取了目前主流的解耦 Head 设计，有利于边界框和类别回归分支使用不同损失函数来进行独立优化，提高模型的精度和训练的稳定性。

对于目标检测领域而言，Backbone 的设计对模型性能起到至关重要的作用。YOLOv8 在其 Backbone 中采用了 C2F（coarse-to-fine）结构，与之前的 C3 结构相比进行了一系列的改进以提高特征提取的效果。相较于 C3 结构，C2F 结构引入更大感受野的 DarknetBottleneck，并在梯度流的设计上借鉴了 ELAN（efficient layer aggregation networks）结构，以更好地构造丰富的梯度流[29]。具体来说，在 YOLOv8 中，C2F 结构替代了 C3 结构的设计，DarknetBottleneck 被引入以获得更大的感受野，能够有效地增加网络对输入信息的感知范围，有助于提取更全局和语义丰富的特征。此外，C2F 结构在梯度流的设计上也进行了改进。通过减少残差连接路径中的一个 ConvBlock，模型可以更加有效地传播梯度信息，从而促使网络更好地学习特定任务的特征。

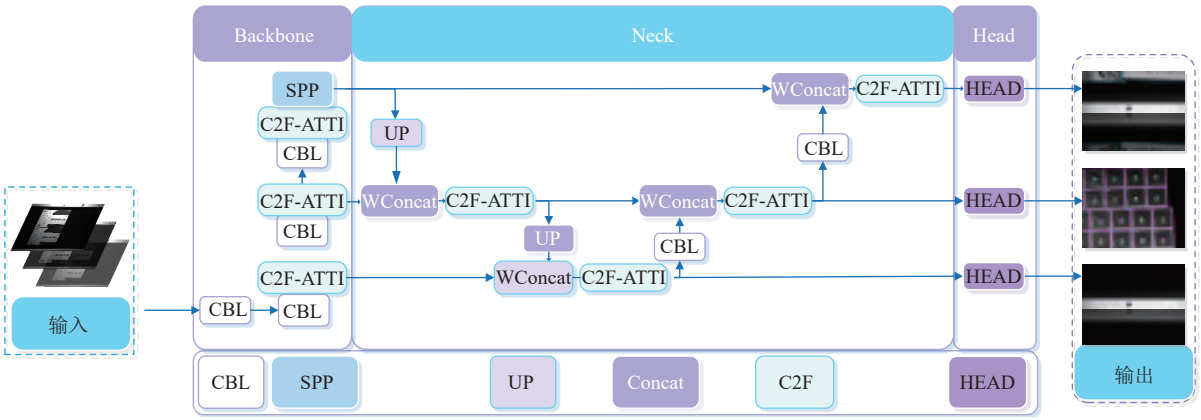



图 2 ACDet 基本结构

与此同时，C2F 结构还通过调整不同尺度模块的通道数，使得模型能够更好地适应不同尺度下的目标检测任务。这种通道数的灵活调整有助于在不同层次上捕捉目标的细节和整体信息，提高模型对目标的识别和定位能力。此外，C2F 的第一个模块省去了一个 Conv 操作，并引入了 Split 操作，进一步增强了网络的表达能力。综合而言，YOLOv8 中采用的 C2F 结构兼顾了感受野的增大、梯度流的改进以及通道数的调整，使得 Backbone 网络在特征提取方面更加强大。这一系列的设计改进在实际应用中为更精准、高效的目标检测提供了有力支持。但是由于 YOLOv8 特征提取和融合机制在多尺度信息处理上存在一定的局限性，其在处理小目标检测和密集场景时，精度和鲁棒性有所下降。

#### 2.1.2 YOLOv10 模型分析

2024 年提出的 Yolov10 是首个无非最大值抑制（non-maximum suppression, NMS）的目标检测模型，通过消除 NMS 并优化模型架构，显著提高

了推理速度和效率。通过引入一致的双重分配策略，这一策略结合了一对一和一对多分配方法，优化了模型训练和推理过程。在推理阶段采用一对一匹配，避免了非极大值抑制的后处理步骤。这使得推理效率显著提高，因为不需要对大量冗余预测进行过滤。在训练阶段采用一对多匹配，为模型提供丰富的监督信号，有助于优化模型性能和加速收敛。这种方法能够弥补一对一匹配中监督信号不足的问题[30]。

YOLOv10 对 Backbone 采用更高效的卷积神经网络架构，以提高特征提取能力和参数效率。Neck 部分通过增强特征金字塔网络和路径聚合网络的设计，能够更好地融合多尺度特征，提高了对不同尺寸目标的检测能力。它采用多种数据增强方法来提高模型泛化能力，同时引入自适应的损失函数调整策略，使模型在不同训练阶段能够动态调整优化目标，提升收敛速度和最终性能。但 YOLOv10 的高性能在很大程度上依赖于大规模标注数据集，在小规模数据集上的表现可能需要更多的微调和优化，以达到理想效果。

### 2.2　改进的计算模块：C2F-A

在药品包装轮廓检测场景下需要重新设计计算模块以克服上文提到的问题。本文提出了一种针对药盒轮廓检测而改进的多层级特征融合模块，称为 C2F-A，利用跳层连接将多个堆叠的 Bottleneck 的输出拼接在一起，构造了更为丰富的梯度流。然而，观察到直接拼接的方式可能导致不同通道之间信息的重叠和干扰，影响了各输出的有效利用。

为了解决特征直接拼接的缺陷，引入 CBAM（convolutional block attention module）模块[31]，通过对拼接后的特征图进行强化注意力操作。本文命名这一改进为 C2F-A，意指通过引入加权机制使得每个通道的信息得到更合理、更有效地利用。C2F-A 模块通过对卷积层提取的各通道特征进行权重重分配，从而动态调整通道的重要性，强化关键信息并抑制冗余特征。C2F-A 模块在对特征图进行全局平均池化和最大池化之前首先强化关键特征，指数放大辨识度高的关键特征，缩小辨识度低的非关键特征。然后通过自学习调整第 1、2 全连接层的权重。最终提高了网络对于关键特征的感知能力，同时抑制了无用特征信息的影响，这样便可在一定程度上减少前文提到的无序堆叠等情况干扰而产生的非关键特征影响，但是不能完全解决。同时需要持续收集各类极端特殊的复杂样本，让模型得到充分训练与学习，目前 CPPD 的数据量依然不足，需要持续扩充完善。整体计算流程如下。

令输入特征图为$F \in \mathbb{R}^{C\times H\times W}$，其中 $C$ 是特征图的通道数，$H$是高度，$W$是宽度。如式 (1) 所示，将$F$归一化到 [−1,1] 得到$F^*$。再通过 Softmax 函数将 $F^*$映射成为指数的概率分布 $F_{\text{en}}$，如式 (2) 所示。该步骤使关键特征值的概率分布指数增长，非关键特征值的概率分布指数减少。最后如式 (3) 所示，通过增强指数 e 将$F^*$与$F_{\text{en}}$线性加权相加后得到$F_{\text{in}}$再进行全局池化处理，其中增强参数设置为 55。该方法可有效过滤无用特征，并增强关键特征，有效提高后续注意力机制的性能。

$$F^* = \tanh F = \frac{e^{F} - e^{-F}}{e^{F} + e^{-F}} \tag{1}$$

$$F_{\text{en}} = \frac{e^{F_i}}{\sum_{k=1}^{C} e^{F_k}} \tag{2}$$

$$F_{\text{in}} = F^* + eF_{\text{en}} \tag{3}$$

全局平均池化和最大池化后生成两个不同的特征描述符，如式 (4) 和式 (5) 所示：

$$F_{\text{in-avg}} = \text{AvgPool}(F_{\text{in}}) \tag{4}$$

$$F_{\text{in-max}} = \text{MaxPool}(F_{\text{in}}) \tag{5}$$

第 1 个全链层采用 ReLu 激活函数来增加特征非线性，有助于模型学习更复杂的通道相关性。第 2 个全链层则恢复强化后的原始特征值。$W_1$、$W_2$分别为两个全链层权重，加权后通过 Sigmoid 激活函数确保安全生成注意力机制的权重向量$\boldsymbol{M}(F)$。通道和空间注意力权重向量分别如式 (6) 和式 (7) 所示。其中 $f_{7\times7}$为 7×7 的卷积操作，$\sigma$是 Sigmoid 激活函数。

$$\begin{aligned}\boldsymbol{M}_c(F) = \sigma(&W_2(F_{\text{in-avg}} + F_{\text{in-max}}) + \\ &W_1(\text{ReLu}(F_{\text{in-avg}}) + \text{ReLu}(F_{\text{in-max}})))\end{aligned} \tag{6}$$

$$\begin{aligned}\boldsymbol{M}_s(F) = \sigma(f_{7\times7}([&W_1(\text{ReLu}(F_{\text{in-avg}}) + W_2(F_{\text{in-avg}}); \\ &W_1(\text{ReLu}(F_{\text{in-max}}) + W_2(F_{\text{in-max}})])))\end{aligned} \tag{7}$$

式中，两个全链层的权重$W_1$、$W_2$采用小批量梯度下降法自学习更新，初始权值由人工设置。损失函数 $L$ 选择 Cross-Entropy Loss。权重自学习更新如式 (8) 所示：

$$W_{b+1} = W_b - \eta\frac{\partial L}{\partial W_b} \tag{8}$$

其中学习率每隔 10 个epoch缩小一半，初始学习率为 0.6。动态学习率如式 (9) 所示，其中$n$为迭代次数：

$$\eta_{e+10} = \eta_e \times 0.5^{\frac{n}{10}} \tag{9}$$

改进后的 C2F-A 模块结构如图 3 所示。

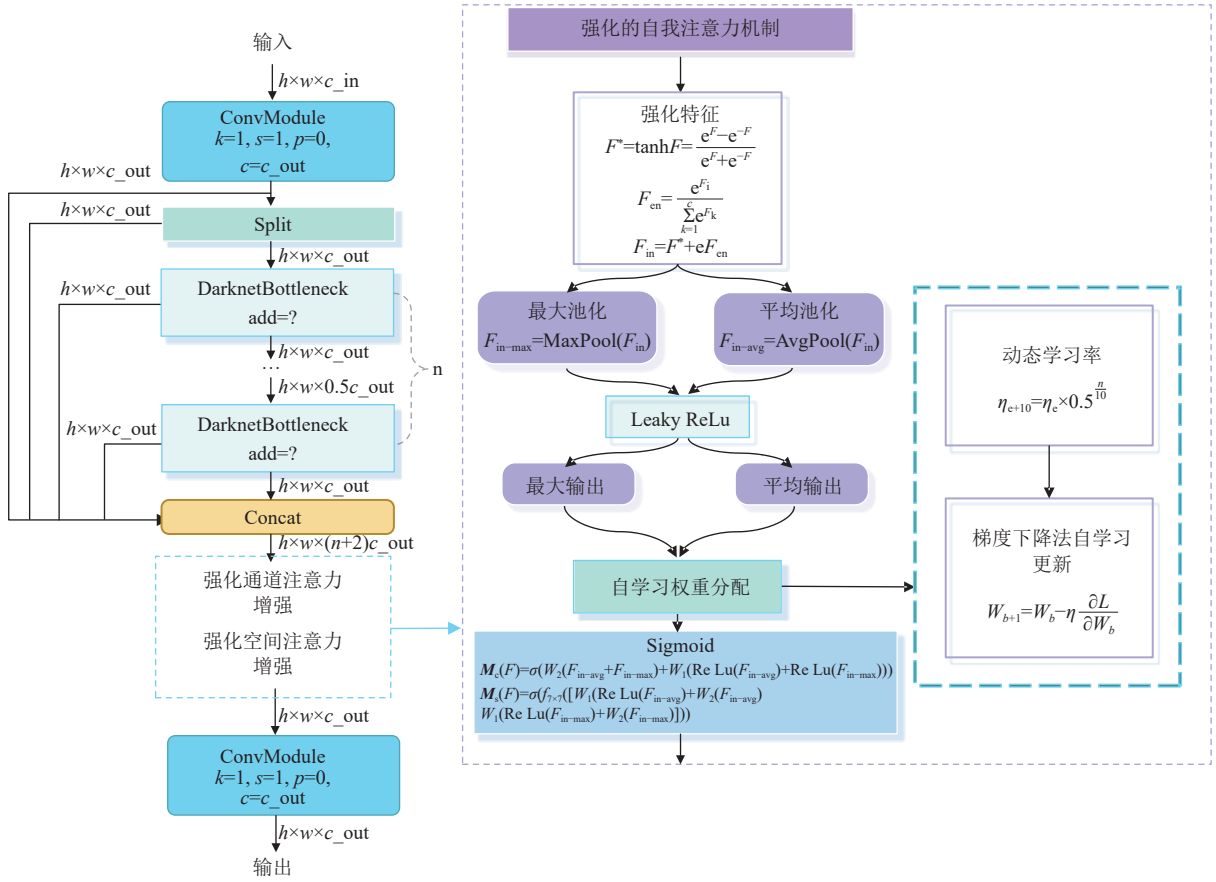



图 3　C2F-A 模块结构示意图

## 2.3　改进的 WConcat 模块

路径聚合特征金字塔网络 PAFPN（path aggregation feature pyramid network）作为一种广泛应用于目标检测任务的特征融合模块，通过直接拼接上下游路径的特征图来实现多尺度信息的整合[13]。然而，直接拼接的方式在处理不同路径特征时可能导致一些通道间的信息冗余和互相干扰，特别是对于某些重要通道的贡献可能被掩盖。为了解决这一问题，本文提出了一种改进的加权拼接融合模块 WConcat。

在 WConcat 中，引入加权拼接的思想，通过嵌入通道自注意力模块，对拼接后的特征图进行通道级别的加权。这使得模型能够自适应地关注来自众多通道中的特定通道的信息，提高网络对关键特征的感知能力。在这个过程中，对直接拼接方式的不足进行了有针对性地改进，使模型更加准确地捕捉不同尺度特征之间的关联性，提高目标的检测性能。

在具体的特征融合阶段，当来自上层与本层的特征图进行 Concat 操作之后，将通道自注意力模块嵌入其后，可以有效地进行特征增强。特别值得注意的是，在 Concat 后引入通道自注意力的设计选择，是为了克服在通道拼接后各通道学习的特征不一致的问题。这样可以赋予包含重要信息的通道更大的权重，而忽略掉那些无足轻重的通道。这样的注意力机制有助于提高网络对拼接后的特征图的关注程度，使得网络更聚焦于对目标的敏感特征，从而提升检测性能。其计算原理如式 (10)：

$$f_{wc} = W\{C(f_1, f_2)\} \tag{10}$$

式中，$C$为 Concat 模块的等价函数，为自学习产生的加权函数，$f_1, f_2$是来自上层进入 Concat 模块的特征图，$f_{wc}$是经过增强后的特征图。改进后的 ACDet 的 Neck 结构如图 4 所示。

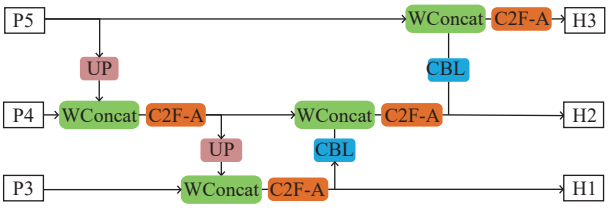



图 4　ACDet 的 Neck 结构

### 2.4　损失函数及评价指标Smooth mAP

#### 2.4.1　损失函数

由于 YOLO 同时解决了一个预测边界框坐标的回归任务和一个预测各分类分数的分类任务，因此使用了两种不同的损失函数。对于回归任务 CIoULoss（complete intersection over union loss），它是一种专门用于优化目标检测中边界框回归问题的损失函数。CIoU Loss 考虑了边界框的重叠面积（IoU）、中心点距离和宽高比惩罚项，这使得它在定位精度和收敛速度上具有显著优势。相比于传统的 IoU Loss，CIoU Loss 更全面地衡量了预测框和真实框之间的差异，从而提升了目标检测模型在训练和推理中的精度和稳定性。具体如式 (11) 所示：

$$\text{CIoU Loss} = 1 - \text{CIoU} = 1 - \text{IoU} + \frac{d^2}{c^2} + \alpha v \tag{11}$$

式中，IoU是预测框和真实框的交并比（intersection over Union）；$d$ 是预测框和真实框中心点的欧氏距离；$c$ 是预测框和真实框外接矩形框的对角线长度；$\alpha$是平衡距离比和宽高比的系数；$v$ 是衡量宽高比一致性的参数。

此外还使用了 DFL（distribution focal loss），它通过引入焦点机制平衡难易样本的贡献，并采用分布学习方法，显著提高了目标检测中的定位精度和分类性能，同时增强了模型在处理小目标和密集目标时的鲁棒性和稳定性[32]。DFL 在噪声和异常值下表现出较高的鲁棒性，适用于各种复杂场景，如本文提到的照度变化下的密集药盒检测场景。有助于提升目标检测模型的整体表现，如式 (12) 所示：

$$\text{DFL} = -((1-y)\log(1-\delta) + y\log\delta) \tag{12}$$

式中，$\delta$是预测的边界框；$y$是真实的边界框。

对于分类任务则使用带 Logits 损失函数的二元交叉熵的 BCE 损失，该损失函数将模型输出层的 Llogits 通过 Sigmoid 激活函数进行转换，并计算二元交叉熵损失。这一损失的计算方式使得在结果数值上更加稳定，无须手动添加 Sigmoid 激活函数到模型的输出中。

#### 2.4.2　评价指标Smooth mAP

在目标检测任务中，常见的性能评价指标包括精度、召回率、平均精度等。为了更好地评价模型的性能，本文设计了一种基于平滑mAP的评价指标 Smooth mAP，它是当前的mAP值与上一epoch的mAP值的加权组合，分别为 0.6 及 0.4。平均精密度（AP）被定义为精密度−召回率曲线下的面积，IoU设置为 0.5。Smooth mAP具体定义如式 (13) 所示：

$$\begin{aligned}\text{Smooth mAP} = {} & \alpha \times \text{mAP}_{\text{current}} + (1-\alpha)/2 \times \\ & \text{mAP}_{\text{previous1}} + (1-\alpha)/2 \times \text{mAP}_{\text{previous2}}\end{aligned} \tag{13}$$

式中，$\alpha$的值为 0.6，表示当前epoch的$\text{mAP}_{\text{current}}$的权重；$(1-\alpha)/2$的值为 0.2，表示上一个epoch的$\text{mAP}_{\text{previous1}}$和上两个epoch的$\text{mAP}_{\text{previous2}}$的权重。

## 3　实验与分析

### 3.1　测试数据集

本文使用工业黑白相机收集了 9 048 张图像，每张图像的原始大小为 4 320×4 320 像素，包含了不同形状和比例的物体。为了适应模型训练和实际应用的需求，将图像的分辨率降低到了 640×480 像素。降低分辨率可以减少模型的计算负担，加快训练速度，这对于大型数据集和复杂模型的训练尤为重要。此外，此分辨率是多种工业相机和视觉系统常见的输出尺寸，因此在此分辨率下进行研究，有助于提升在工业场景下的通用性。同时，在较低的分辨率下进行开发，可以减少对硬件算力的依赖，使得研究成果更容易在实际的工业环境中部署和应用。

这些药物包装图像对 7 个不同形状的包装类型进行了注释，形成了完全注释的 CPPD 数据集，共计 16 848 个实例。各类别的实例统计数量如图 5 所示，展现了数据集的多样性和平衡性。通过这种细致的标注，CPPD 为本工作提供了一个具有挑战性和实用价值的旋转目标检测数据集。

此外，为了验证 ACDet 在更大规模的通用数据集上的泛化能力，本文在 COCO 数据集上进行了实验[33]。该数据集包含 20 万张各类通用目标的标注样本，图像分辨率为 640×640，这与 CPPD 填充后的尺寸相同。

实验硬件为 RTX3060 GPU，零样本测试为 RTX3050 GPU。在 Python 3 环境下进行，采用 Ultralytics 提出的 PyTorch 深度学习框架，并使用 TensorBoard 来监控训练过程[34]。Batch size 取 16，训练 100 个epoch。在训练过程的最后 10 个epoch

关闭 Mosaic 增强。

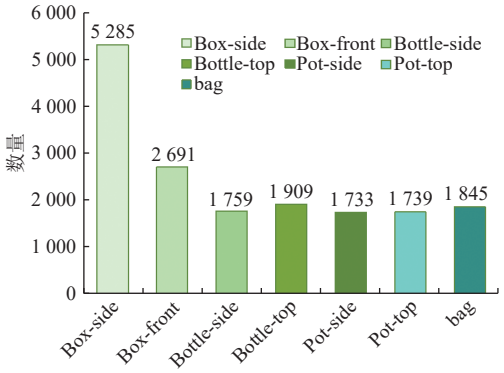



图 5 CPPD 数据集统计情况

### 3.2 ACDet 与其他模型表现对比

首先在某医药配送场地的 10 台设备上进行实际零样本测试，来验证 ACDet 的真实表现。统计时间超过 10 天，累计检测次数超过 10 万次。成功率判断标准为每次检测系统接收到的药品期望数量与实际检测数量进行对比，一致则认为成功。平均复核成功率 99.91%，统计如图 6 所示。

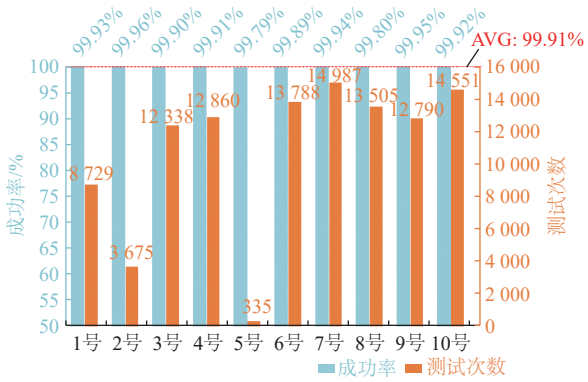



图 6 零样本测试结果

随后，为了验证 ACDet 的性能，表 1 和图 7 展示了本文方法与其他 YOLO 模型，分别是 YOLOv5、YOLOv6、YOLOv8 和 YOLOv10，在 CPPD 数据集上的性能对比。本文的方法取得了较优的性能，其中 Smooth mAP提升了 4.66%～13.85%，mAP提升了 6.31%～19.41%。为了保持最快的推理速度，所有模型均采用 Nano 级结构设计，参数规模极为精简，可充分体现模型轻量化特征。在本文的实验环境中，所有模型推理出一张输入图像的检测结果的平均时间在 1.8 ms，满足药盒检测场景下的速度要求。

通过表 2 看出，本文方法在 COCO 数据上测试，性能与大规模数据表现优异的 YOLOv10 非常接近，Smooth mAP相较 YOLOv5、YOLOv6 和 YOLOv8 提升了 0.6%～7.5%，mAP提升了 0.8%～6.3%。

**表 1 不同模型在 CPPD 数据集上的性能表现和参数量**

| 模型 | Smooth mAP/% | mAP/% | FLOPS/M | 平均推理时间/ms |
|---|---|---|---|---|
| YOLOv5 | 66.49 | 66.95 | 2.58 | 1.78 |
| YOLO6 | 62.67 | 62.15 | 4.26 | 1.83 |
| YOLO8 | 74.86 | 75.25 | 3.08 | 1.79 |
| YOLOv10 | 64.69 | 65.10 | 2.71 | 1.77 |
| Ours | **79.52** | **81.56** | 3.69 | 1.79 |

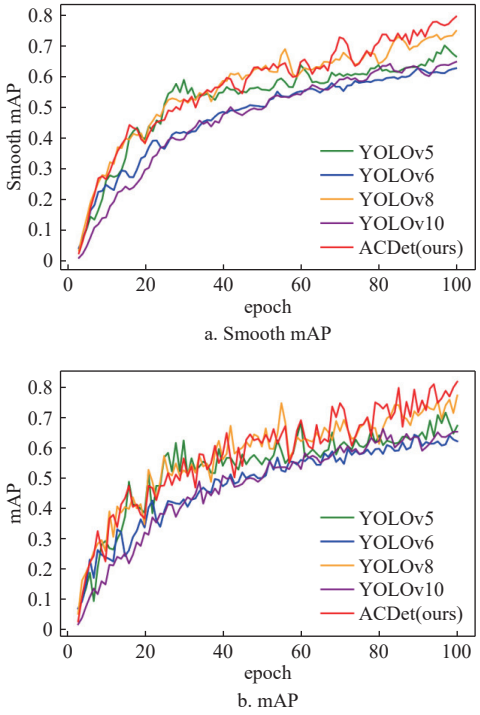



图 7 不同模型在 CPPD 数据集上的性能表现和参数量

**表 2 在 COCO 数据集上的测试结果**

| 模型 | Smooth mAP/% | mAP/% | FLOPS/M | 平均推理时间/ms |
|---|---|---|---|---|
| YOLO5 | 47.2 | 48.9 | 2.58 | 1.70 |
| YOLO6 | 49.5 | 50.4 | 4.26 | 1.75 |
| YOLO8 | 50.7 | 51.8 | 3.08 | 1.71 |
| YOLOv10 | **51.3** | **52.4** | 2.71 | 1.70 |
| Ours | 51.0 | 52.2 | 3.69 | 1.70 |

### 3.3 特征加权融合模块的消融实验

依赖于 CPPD 数据集，本文接下来研究了上文提出的模型结构在药盒检测场景下的检测性能。如上所述，C2F-A 模块针对特征通道拼接后的特征图进行增强，即对不同的特征图进行加权，突出强调有用的特征。为了探究它的效果，本文会研究该模块对提升本文方法的检测准确度的正面影响。

表 3 和图 8 展示了 C2F-A 模块在 CPPD 数据集上的正面作用。在 YOLOv8 模型中引入 C2F-A 模块后，模型的Smooth mAP值比基线模型提高了 3.46%，mAP值比基线模型提高了 3.05%，表明 C2F-A 模块的引入对模型的检测性能有明显的提升。

在 YOLOv8 模型中引入 WConcat 模块。模型的Smooth mAP 值比基线模型提高了 3.97%，mAP值比基线模型提高了 3.28%，表明 WConcat 模块在特征融合网络阶段的应用对性能提升有较大的作用。当两个模块都被激活后，模型的Smooth mAP值比基线模型提高了 5.98%，mAP值比基线模型提高了 6.31%，这表明本文提出的两个特征融合模块能够给药盒检测场景的模型带来巨大的性能提升。

**表 3　YOLOv8 融合不同模块在 cppd 数据集上的性能表现**

| 方法 | Smooth mAP/% | mAP/% | FLOPS/M |
|---|---|---|---|
| Yolov8 | 74.86 | 75.25 | 3.08 |
| +C2F-A | 77.80 | 76.94 | 3.32 |
| +wconcat | 77.71 | 78.53 | 3.45 |
| +C2F-A+wconcat | **79.52** | **81.56** | 3.69 |

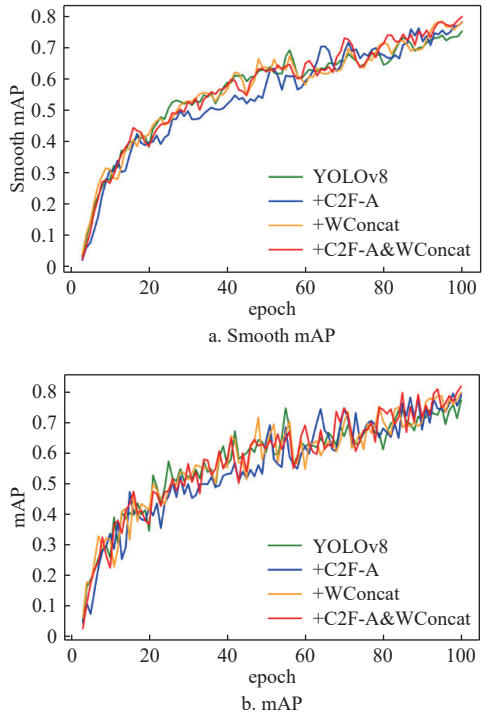



图 8　YOLOv8 融合不同模块在 CPPD 数据集上的性能表现

## 4　结束语

本文分析了实际工业环境下的药品轮廓检测案例中的难点，并提出了几点针对性地改进。通过在不同数据集上的多次实验，验证了本文方法精度较高，能够满足实际场景下的应用要求。具体来说，本文构建了一种基于自我注意力机制强化的卷积神经网络快速检测模型 ACDet，该模型仅依靠灰度图像在变化照度下对密集排列、多角度、多种类的物品进行高效检测。与其他使用彩色信息简化图像处理过程的模型相比，减少了计算资源的消耗并实现了快速检测；并设计了一种额外图像增强方法，从亮度、旋转、随机图像合成等方面对数据源进行有效增强，提高模型多角度学习物品外观特征的能力及鲁棒性，从而能适应各类检测场景及目标；改进了核心特征提取计算模块 C2F-A，对特征图通道和空间进行多维度强化自我注意力增强，提升了模型的自我认知能力；设计 WConcat 模块，对输出的多层次特征图进行自学习加权拼接，并且确认了模块的最佳使用时机。在零样本、CPPD 和公开数据集上进行了充分的实验，验证了 ACDet 可以兼顾检测准确率和效率。

另一方面，C2F-A 模块特征聚合能力较强，用在主干网络部分会导致一定程度的性能损失。后续研究将进一步优化网络，从多个方面提升精度，并研究基于 Transformer 架构的更准确的物品轮廓检测算法。最后，为了验证模型的泛化性和鲁棒性，后续会在其他公开数据集或案例中进行更多的对比测试并优化网络。

## 参 考 文 献

[1] MARQUES C M, MONIZ S, DE SOUSA J P, et al. Decision-support challenges in the chemical-pharmaceutical industry: Findings and future research directions[J]. Computers & Chemical Engineering, 2020, 134: 106672.

[2] KUMAR G. Pharmaceutical drug packaging and traceability: A comprehensive review[J]. Universal Journal of Pharmacy and Pharmacology, 2023, 2(1): 19-25.

[3] IANDOLA F N, HAN S, MOSKEWICZ M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[EB/OL]. [2024-05-15]. https://doi.org/10.48550/arXiv.1602.07360.

[4] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications[EB/OL]. [2024-05-15]. https://doi.org/10.48550/arXiv.1704.04861.

[5] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S. l. ]: IEEE, 2019: 1314-1324.

[6] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]//Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition. New York: IEEE, 2018: 4510-4520.
[7] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design [C]//Proceedings of the European Conference on Computer Vision. Munich: ECCV, 2018: 116-131.
[8] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 6848-6856.
[9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2024-05-15]. https://doi.org/10.485501/arXiv:1409.1556, 2014.
[10] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2016: 770-778.
[11] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 2261-2269.
[12] 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. New York: IEEE, 2017: 936-944.
[13] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 8759-8768.
[14] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 10781-10790.
[15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2014: 580-587.
[16] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2015: 1440-1448.
[17] REN S, HE K, 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, 2016, 39(6): 1137-1149.
[18] DAI J F, LI Y, HE K M, et al. R-FCN: Object detection via region-based fully convolutional networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Red Hook, NK: Curran Associates Inc., 2016: 379-387.
[19] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: Towards balanced learning for object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 821-830.
[20] 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. New York: IEEE, 2016: 779-788.
[21] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2017: 6517-6525.
[22] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. [2024-05-15]. https://members.loria.fr/GSimon/files/cours/article10.pdf.
[23] LIU W, ANGUELOV D, ERHAN D, et al. Ssd: Single shot multibox detector[C]//European Conference on Computer Vision. Amsterdam: Springer International Publishing, 2016: 21-37.
[24] 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. New York: IEEE, 2017: 2999-3007.
[25] DUAN K W, BAI S, XIE L X, et al. CenterNet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 6568-6577.
[26] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European conference on computer vision (ECCV). Munich: Springer International Publishing, 2018: 734-750.
[27] LAW H, TENG Y, RUSSAKOVSKY O, et al. CornerNet-Lite: Efficient keypoint based object detection[EB/OL]. [2024-05-15]. https://arxiv.org/abs/1904.08900.
[28] TIAN Z, SHEN C, CHEN H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 9627-9636.
[29] WANG C Y, BOCHKOVSKIY A, LIAO H M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2023: 7464-7475.
[30] WANG A, CHEN H, LIU L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.
[31] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). Munich: Springer International Publishing, 2018: 3-19.
[32] LI X, LV C Q, WANG W H, et al. Generalized focal loss: Towards efficient representation learning for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3139-3153.
[33] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//Proceedings of the European Conference on Computer Vision (ECCV). Zurich: Springer International Publishing, 2014: 740-755.
[34] GUO G, ZHANG Z. Road damage detection algorithm for improved YOLOv5[J]. Scientific Reports, 2022, 12(1): 15523.

编 辑 张 莉