智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

注意力优化的轻量目标检测网络及应用

吴珺, 董佳明, 刘欣, 王春枝

引用本文:
吴,董佳明,刘欣,王春枝. 注意力优化的轻量目标检测网络及应用[J]. 智能系统学报, 2023, 18(3): 506–516.
WU Jun,DONG Jiaming,LIU Xin,WANG Chunzhi. Lightweight object detection network and its application based on the attention optimization[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(3): 506–516.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202206014

DOI: 10.11992/tis.202206014
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20230328.1626.002.html

注意力优化的轻量目标检测网络及应用

吴珺[1,2], 董佳明[1], 刘欣[1], 王春枝[1]
(1. 湖北工业大学 计算机学院, 湖北 武汉 430068; 2. 武汉理工大学 材料科学与工程学院, 湖北 武汉 430070)

摘 要: 本文以轻量化改进 YOLO 网络为主要目标, 选取具有代表性的 (squeeze and excitation, SE) 通道注意力模块和比较新颖的 (coordinate attention, CA) 空间注意力模块与 YOLOv5s 目标检测网络进行融合, 提出新的轻量网络模型 YOLOv5s-CCA (YOLOv5s-C3-coordinate attention) 和 YOLOv5s-CSE(YOLOv5s-C3-squeeze-and-excitation)。通过进一步探索, 论证出 SE 和 CA 注意力模块在 YOLOv5s 目标检测网络中最优插入位置的策略, 实验论证了在轻量化网络模型中 CA 优于 SE 注意力模块。本文所提出的 YOLOv5s-CCA 网络模型在 PASCAL VOC 2012 数据集和 Global Wheat 2020 数据集中实现了网络轻量化并且精度较原始网络有所提升; 并证实了 YOLOv5s-CCA 具有一定的通用性和泛化性, 为其在实际生产与生活中进行轻量化部署提供了可靠的数据支撑和一定参考价值。

关键词: 目标检测; 深度学习; 计算机视觉; 轻量化网络; 空间注意力; 通道注意力; 一阶目标检测网络; 损失函数
中图分类号: TP18 **文献标志码**: A **文章编号**: 1673−4785(2023)03−0506−11

中文引用格式: 吴珺, 董佳明, 刘欣, 等. 注意力优化的轻量目标检测网络及应用 [J]. 智能系统学报, 2023, 18(3): 506–516.
英文引用格式: WU Jun, DONG Jiaming, LIU Xin, et al. Lightweight object detection network and its application based on the attention optimization[J]. CAAI transactions on intelligent systems, 2023, 18(3): 506–516.

Lightweight object detection network and its application based on the attention optimization

WU Jun[1,2], DONG Jiaming[1], LIU Xin[1], WANG Chunzhi[1]
(1. School of Computer Science, Hubei University of Technology , Wuhan 430068, China; 2. School of Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, China)

Abstract: Taking the lightweight improved YOLO network as the main target, the new lightweight network models YOLOv5s-CCA (YOLOv5s-C3-coordinate attention) and YOLOv5s-CSE (YOLOv5s-C3-squeeze-and-excitation) are put forward in this paper by selecting the representative SE (squeeze-and-excitation) channel attention module and relatively novel CA (coordinate attention) spatial attention module to fuse with YOLOv5s object detection network. By further exploration, the strategy for the optimal insertion position of the SE and CA attention modules in YOLOv5s object detection network is demonstrated. The experiment proves that CA is superior to SE attention module in the lightweight network model. The YOLOv5s-CCA network model proposed in this paper realizes the goal of network lightweight in both PASCAL VOC 2012 and Global Wheat 2020 data sets, and its accuracy is improved compared with the original network. It is confirmed that YOLOv5s-CCA has certain universality and generalization, which provides reliable data support and certain reference value for its lightweight deployment in actual production and life.

Keywords: object detection; deep learning; computer vision; lightweight network; coordinate attention; squeeze-and-excitation; one-stage object detection network; loss function

随着万物互联理念的提出, 物联网设备得到了高速发展, 使得物联网正快步进入人工智能+物联网时代。计算机视觉已被广泛应用于农业、工业、医学等多个领域; 在 5G 时代能够实现智能视觉全面融入物联网。但是对于目前大部分物联网内的移动设备而言, 其设备的计算能力与存储

收稿日期: 2022−06−08. **网络出版日期**: 2023−03−29.
基金项目: 国家自然科学基金项目 (61602161, 61772180); 湖北省重点研发项目 (2020BAB01); 湖北工业大学研究生基金项目 (2021046).
通信作者: 吴珺. E-mail: wujun@whut.edu.cn.

空间受成本和相关芯片供应紧缺等因素的影响, 使得复杂的视觉检测网络模型无法有效地部署到资源受限的小型处理器上, 并进行高效的实时检测。

目前广泛使用的基于卷积神经网络的目标检测方法大致可以分为两大类, 一类是基于二阶段的卷积神经网络目标检测方法, 如 RCNN[1]、SPPNet[2]、Fast RCNN[3]、Faster RCNN[4]、FPN[5]、Mask RCNN[6] 等。另一类就是基于一阶段的卷积神经网络目标检测方法, 如 YOLO 系列[7-10]、SSD[11] 等。

YOLO 由 Redmond 等[7] 于 2015 年提出, 它是深度学习时代的第一个基于一阶段的目标检测网络; 它完全抛弃了之前基于二阶段目标检测网络的检测范式: 候选框检测+验证; 原作者进行了一系列改进, 提出 YOLOv2[9] 和 YOLOv3[10], 在保持高检测速度的前提下, 进一步提高了检测精度。相关研究人员也在 YOLOv3 的基础之上改进出了 YOLOv4[8,11] 与 YOLOv5。陈科圻等[12] 针对基于深度学习目标检测两个主要算法流派的奠基过程进行了回顾, 包括以 R-CNN 系列为代表的两阶段算法和以 YOLO、SSD 为代表的一阶段算法; 进一步以多尺度目标检测的实现为核心, 重点诠释了图像金字塔、构建网络内的特征金字塔等典型策略。毛莺池等[13] 提出一种基于 Faster R-CNN 的多任务增强裂缝图像检测 ME-Faster RCNN 方法, 即将图片输入 ResNet-50 网络提取特征; 然后将所得特征图输入多任务增强 RPN 模型, 同时改善 RPN 模型的锚盒尺寸和大小以提高检测识别精度, 生成候选区域; 最后将特征图和候选区域发送到检测处理网络, 达到提升平均 IoU 和 mAP 值的良好效果。邵江南等[14] 针对长时目标跟踪所面临的目标被遮挡、出视野等常常会导致跟踪漂移或丢失的问题, 提出一种深度长时目标跟踪算法 LT-MDNet 在跟踪精度和成功率上都展现了极强的竞争力, 并且在目标被遮挡、出视野等情况下保持了优越的跟踪性能和可靠性。赵文清等[15] 对 SSD 模型深层特征层与浅层特征层进行特征融合, 然后将得到的特征与深层特征层进行融合; 其次在双向融合中加入了通道注意力机制, 增强了语义信息; 最后提出了一种改进的正负样本判定策略, 降低目标的漏检率; 该方法在对目标进行检测时, 目标平均准确率有较大提高。田永林等[16] 以分类任务为切入, 介绍了典型视觉 Transformer 的基本原理和结构, 并分析了 Transformer 与卷积神经网络在连接范围、权重动态性和位置表示能力三方面的区别与联系; 提出了视觉 Transformer 的一般性框架, 并分析了视觉 Transformer 在特征学习、结果产生和真值分配等方面给上层视觉模型设计带来的启发和改变。郭璠等[17] 在 YOLOv3 算法的基础上, 提出了目标检测的通道注意力方法和基于语义分割引导的空间注意力方法, 形成 YOLOv3-A 算法; 该算法对小目标检测性能的改善尤为明显, 精度和召回率都有所提升。

本文将 YOLO 目标检测网络模型作为研究重点, 在确保一定检测精度的前提下以网络轻量化为目标, 引入空间注意力机制(coordinate attention, CA)和通道注意力机制(squeeze and excitation, SE)等多种注意力机制得到优化目标检测网络。该网络模型不但具有一定目标检测精度, 而且通过减少目标检测网络的参数量和计算量达到减少网络检测耗时的效果, 实现了轻量化的目标检测。本文的主要研究工作如下: 首先, 回顾注意力模块在计算机视觉领域的发展, 选取其中具有代表性的 SE 注意力模块和比较新颖的 CA 注意力模块与 YOLOv5 目标检测网络进行融合, 并进行了相应的消融实验, 探索出了通道注意力模块和空间注意力模块在目标检测网络中最优插入位置的策略。其次, 将本文提出的 YOLOv5s-CCA 目标检测网络在 PASCAL VOC 2012 数据集进行实验获得了良好结果。进一步在 Global Wheat 2020 数据集实验也表现出色, 这也证实了该模型具有一定的通用性和良好的泛化性, 为 YOLO 系列目标检测网络在实际生产与生活中进行轻量化部署提供了可靠的数据支撑。YOLOv5s-CCA 目标检测网络在检测精度和检测速度上能达到较好的平衡, 更适合部署在小型处理器上, 使其在农业、工业、医学图像领域得到更广泛的应用。

1　注意力机制

1.1　注意力机制及其应用分析

目前可以将注意力机制大致分为软注意力和强注意力两类。首先, 软注意力是一种可微的确定性注意力, 这样就可以通过目标检测网络计算梯度来调节注意力的权重。因此软注意力更注重通道[18] 和区域[19]。然而强注意力[20-22] 则是一种基于动态随机预测生成的不可微注意力, 更关注目标检测网络里的每个点; 因此需要通过强化学习来完成网络训练。

通常在目标检测任务所使用的数据集当中, 图像中的一些目标可能存在各种各样的角度和姿

态，为了提高模型的泛化能力和鲁棒性，Jaderberg 等[18] 在 2015 年提出了 spatial transformer 网络，通过使用 spatial transformer 模块处理输入图像，使图像中的目标在空间上被摆正，进而使模型对各种姿态、扭曲变形的目标，都能有较好的识别检测能力。Wang 等[23] 在 2017 年将残差结构与注意力结合，提出了残差注意力网络。这是一种在非常深的结构中采用混合注意力机制的卷积网络。残差注意力网络由多个注意力模块组成，这些模块会产生注意力感知功能。随着模块的深入，来自不同模块的注意力感知功能会自适应地变化，从而提高分类网络的表现。Jie 等[19] 在 2018 年提出了一个简单高效且富有创造力的 SENet，并靠着这个模型获得了 ImageNet 的冠军。其主要创新点就是提出了由挤压和激励两部分组成的注意力模块。针对 SE 模块没有考虑空间信息，卷积注意力机制 (convolutional block attention module, CBAM) 虽然同时考虑了通道信息和空间信息，但是却只考虑了局部区域的空间信息。Hou 等[24] 于 2021 年提出了 CA 注意力机制，CA 注意力是通过对水平方向和垂直方向上分别进行平均池化，再使用转换器对空间信息进行编码，最后把空间信息通过加权的方式融合进通道中，这样就实现了 CA 注意力机制对空间信息和通道信息的全面考虑。

1.2　SE 注意力

从图 1 的结构中可以看出，SE 模块是由挤压和激励两部分组成，分别用于全局信息嵌入和通道关系的自适应重新校准。

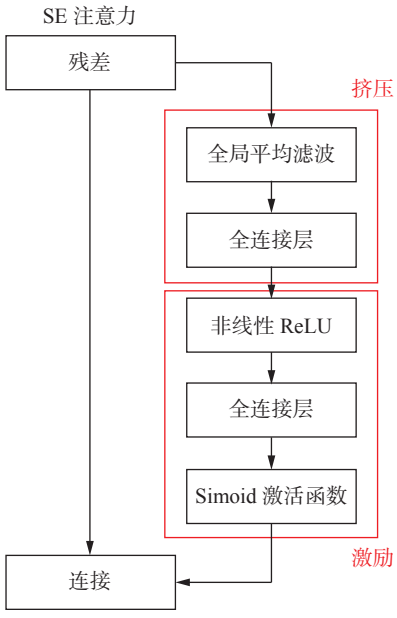


图 1　SE 注意力模块

Fig. 1　Squeeze and excitation module

第 1 步，对于给定的输入 $\boldsymbol{X}=[x_1\ x_2\ \cdots\ x_c]$，对于第 c 个通道进行挤压的操作可以用下面的公式进行表示：

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j)$$

式中：z_c 表示第 c 个通道的输出；H 和 W 表示输入 X 的高和宽。为了获得通道的平均池化值，z_c 通过计算每个通道内所有的特征值的全局平均值。

第 2 步，激励部分，旨在完全捕获通道方面的依赖关系，可以表述为

$$\hat{X}=X\cdot\sigma(T_2(\mathrm{ReLU}(T_1(z))))$$

式中：“·”指的是 channel-wise 乘法，σ 是 Sigmoid 函数，T_1 和 T_2 是两个线性变换，可以通过学习来捕捉每个通道的重要性。

SE 模块在很多轻量化的 mobilenet 网络中表现优异，但是由于其只考虑了通道信息，没有考虑位置信息，这使得其有很大的改进空间。

1.3　CA 注意力

CA 注意力模块是基于 SE 注意力和 CBAM 注意力改进而来。SE 注意力模块只考虑了通道信息，没有考虑空间信息。CBAM 注意力模块对每个位置的通道上进行池化，由于经过几层卷积和下采样后特征图的每个位置只包含原图的一个局部区域，因此这种做法只考虑了局部区域信息。

CA 注意力模块结构图如图 2 所示，其中高度平均池化表示沿 H 方向的全局平均池化层，宽度平均池化表示沿 W 方向的全局平均池化层，Concat 表示拼接操作，Conv2d 表示普通二维卷积操作，BatchNorm 表示批量归一化操作等。

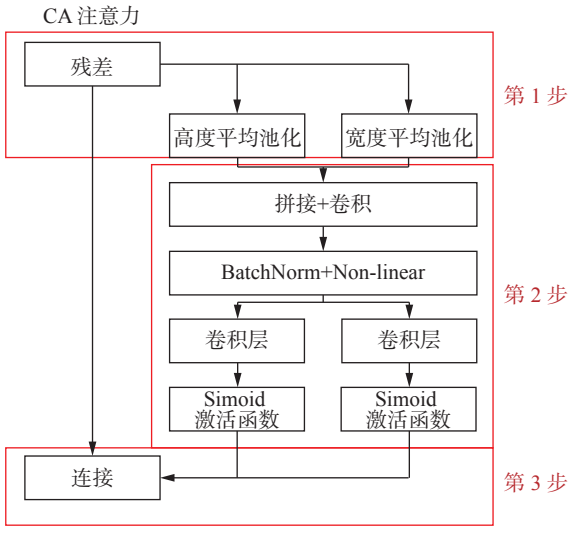


图 2　CA 注意力模块

Fig. 2　Coordinate attention module

具体实现过程如下:

1)利用两个一维全局池化核(*H*,1)和(1,*W*),将沿垂直和水平方向的特征图分别聚合为两个单独的方向注意力特征图,沿 *H* 方向的第 *c* 个通道的输出为

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i<W}x_c(h,i)$$

同样沿 *W* 方向的第 *c* 个通道的输出为

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j<H}x_c(j,w)$$

式中:*H*表示该层特征图的高度;*W*表示该层特征图的宽度;$z_c^h(h)$表示沿 *H* 方向的第 *c* 个通道的输出结果;$z_c^w(w)$表示沿 *W* 方向的第 *c* 个通道的输出结果;$x_c(h,i)$表示输入特征图 $\boldsymbol{X}$ 沿 *H* 方向的输入;$x_c(j,w)$表示输入特征图 $\boldsymbol{X}$ 沿 *W* 方向的输入。

2)将具有嵌入特定方向信息的这两个特征图分别编码为 2 个注意力图,该过程为坐标注意力生成。对应产生的位置信息会被保存起来,存放在注意力图内。位置信息是指特征图沿 *H* 方向提取的信息和沿 *W* 方向提取的信息:

$$\boldsymbol{f}=\delta\left(F_1\left(\left[z^h,z^w\right]\right)\right)$$

式中:[.,.] 表示沿空间维度的拼接操作;F_1表示卷积操作;δ表示 Sigmoid 激活函数,其中$\boldsymbol{f}\in\mathbf{R}^{C/r\times(H+W)}$表示在水平方向和垂直方向编码空间信息的中间特征图。*r* 是一个控制注意力模块大小的超参。然后将$\boldsymbol{f}$沿 *H* 和 *W* 两个方向拆分为$\boldsymbol{f}^h\in\mathbf{R}^{C/r\times H}$和$\boldsymbol{f}^w\in\mathbf{R}^{C/r\times W}$两个特征图,然后使用$F_h$和$F_w$两个卷积操作将$\boldsymbol{f}^h$和$\boldsymbol{f}^w$两个特征图的通道数转化为与输入特征 $\boldsymbol{X}$ 具有相同通道数的注意力权重$\boldsymbol{g}^h$和$\boldsymbol{g}^w$。

$$\boldsymbol{g}^h=\sigma\left(F_h\left(\boldsymbol{f}^h\right)\right)$$
$$\boldsymbol{g}^w=\sigma\left(F_w\left(\boldsymbol{f}^w\right)\right)$$

3)通过乘法将两个注意力权重$\boldsymbol{g}^h$和$\boldsymbol{g}^w$都应用于输入特征图 $\boldsymbol{X}$ 上,得到注意力模块的输出 $\boldsymbol{Y}=[y_1\ y_2\ \cdots\ y_c]$,以强调注意区域。

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j)$$

式中:$y_c(i,j)$表示第 *c* 个通道的输出;$x_c(i,j)$表示第 *c* 个通道的输入;$g_c^h(i)$表示第 *c* 个通道上沿 H 方向的注意力权重;$g_c^w(j)$表示第 *c* 个通道上沿 *W* 方向的注意力权重。简单说来,CA 注意力模块是通过沿 *H* 方向和沿 *W* 方向进行平均池化,再通过转换器对空间信息进行相关编码,最后融合通道的加权信息和对应的空间信息。

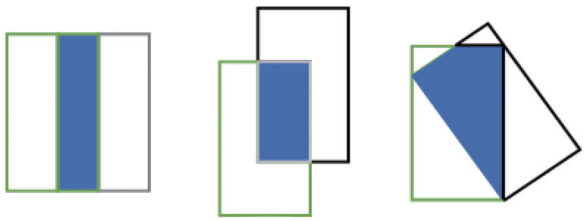
2　损失函数

本文提出的 YOLOv5-CCA 目标检测网络使用的是 CIoU(complete-intersection over union) 损失函数。本节将会介绍常用的损失函数,并梳理从以 IoU(intersection over union)作为损失函数到以 CIoU 作为损失函数的发展历程,系统地介绍各自的优缺点。

IoU 损失函数计算公式为

$$L_{\mathrm{IoU}}=1-\mathrm{IoU}$$

以 IoU 作为边界框的回归损失函数会出现以下两个问题:1) 如果 2 个框没有相交,根据定义,IoU=0 就不能反映 2 个框之间的距离大小和重合程度。同时因为损失函数为 0,没有梯度回传,无法进行学习训练。2)IoU 无法精确地反映两者的重合度大小。如图 3 所示,假设 3 种情况 IoU 都相等,但看得出来它们的重合度是不一样的,图 3(a) 的回归效果最好,图 3(b) 的回归效果相对差一点,图 3 (c) 的回归效果最差。



(a) 回归效果最好　(b) 回归效果相对差一点　(c) 回归效果很差

图 3　预测框与目标框可能出现的 3 种状态

Fig. 3　Three states appeared about the prediction box and the target box

2019 年 Hamid 等[25] 提出了优化边界框的新思想。由于 IoU 是比值的概念,对目标物体的比例是不敏感的。然而检测任务中边界框的回归损失函数 (MSE 损失函数、L1-smooth 损失函数等) 优化和 IoU 优化不是完全等价的,而且 Ln 范数对物体的比例也比较敏感,IoU 无法直接优化没有重叠的部分。于是提出将 GIoU 应用于边界框的回归损失函数中,其中 GIoU 的计算公式为

$$\mathrm{GIoU}=\mathrm{IoU}-\frac{|A_{\mathrm{c}}-U|}{|A_{\mathrm{c}}|}$$

式中:A_{c} 表示预测框与真实框的最小闭包区域面积,即同时包含了预测框和真实框的最小框的面积,如图 4 中蓝色边框包围的区域。U 表示预测框与真实框并集部分的面积,如图 4 中绿色框与红色框并集的面积。$|A_{\mathrm{c}}-U|$表示预测框与真实框的最小闭包区域面积减去预测框与真实框并集部分的面积,如图 4 中黄色部分的面积。

GIoU 损失函数的计算公式为

$$L_{\mathrm{GIoU}}=1-\mathrm{GIoU}$$

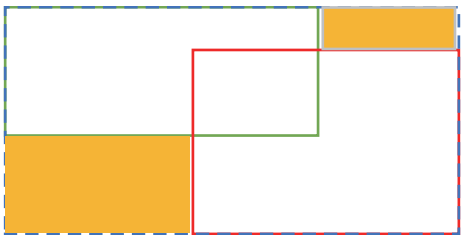


图 4　GIoU 损失作为边界框损失函数的原理图

Fig. 4　Schematic of the loss function GIoU loss

同时 GIoU 也存在些许不足：1) 当预测框与真实框的高宽相同，且处于同一水平面时，GIoU 就退化为 IoU；2) 在训练过程中预测框在水平或垂直方向优化困难，导致收敛速度慢、回归不够准确。Zheng 等[26] 在 GIoU 的基础之上提出了 DIoU(distance-IoU)。DIoU 的惩罚项是基于中心点的距离和对角线距离的比值，避免了 GIoU 在两框距离较远时，产生较大的最小闭包区域面积，使得损失函数值较大而难以优化的问题。

DIoU 的计算公式为

$$\mathrm{DIoU}=\mathrm{IoU}-\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}$$

式中：b表示预测框的中心点；b^{gt}表示真实框的中心点；ρ表示b和b^{gt}之间的欧氏距离。如图 5 中蓝色框对角线的距离。

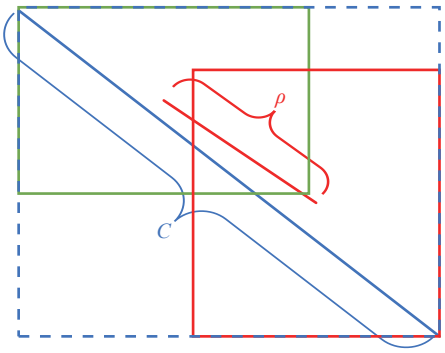


图 5　DIoU 损失作为边界框损失函数的原理图

Fig. 5　Schematic of the loss function DIoU loss

DIoU 损失函数的计算公式为

$$L_{\mathrm{DIoU}}=1-\mathrm{DIoU}$$

DIoU 作为边界框的回归损失函数具有以下优点：

1) 与 GIoU 损失类似，DIoU 损失在预测框与目标框不重叠时仍然可以为边界框提供移动方向；

2) DIoU 损失通过最小化预测框与目标框中心点的欧氏距离因此比 GIoU 损失收敛速度要快；

3) 对于包含预测框与目标框在水平方向和垂直方向上这种情况，DIoU 损失可以很快的回归，而 GIoU 损失几乎会退化为 IoU 损失；

4) DIoU 还可以替换普通的 IoU 评价策略，应用于 NMS 中，使得 NMS 得到的结果更合理和有效。

由于 CIoU 没有将预测框与目标框的宽高比考虑其中，于是 CIoU[26] 在 DIoU 的基础上将预测框与目标框的宽高比考虑了进去。

CIoU 的计算公式为

$$\mathrm{CIoU}=\mathrm{IoU}-\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}-\alpha v$$

其中α是权重函数，定义为

$$\alpha=\frac{v}{(1-\mathrm{IoU})+v}$$

v是用来度量宽高比的相似性，定义为

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2$$

CIoU 损失函数的计算公式为

$$L_{\mathrm{CIoU}}=1-\mathrm{IoU}+\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\alpha v$$

CIoU Loss 的梯度与 DIoU Loss 类似，但还要考虑 v 的梯度。高宽比在 [0,1] 时，w^2+h^2的值通常很小，这将会导致梯度爆炸的情况出现，因此在具体计算时会将$\frac{1}{w^2+h^2}$的结果替换为 1。综合比较各个损失函数，为了更好地比较检测网络的边界框尺度数据获得更精准的检测信息，因此本节中所有涉及到目标检测网络的实验，其所使用的损失函数均为 CIoU Loss。

3　YOLOv5s-CCA 目标检测网络

由于 YOLOv5 项目一直保持更新，所以本文所有关于 YOLOv5s 的介绍均基于官方第五版。

本章节所有实验均基于 YOLOv5s 目标检测网络，其结构如图 6 所示，其中左侧数字表示 Backbone 部分每个模块的编号，方便描述注意力模块插入的位置。首先为了验证 CA 模块性能优于 SE 模块，本实验在 Backbone 中编号为 2 的 C3 模块后面分别添加一个 SE 模块和 CA 模块进行训练，如图 7 所示。其中图 7(a) 是原始的 Backbone，图 7(b) 是在编号为 2 的 C3 后面插入 SE 注意力模块，图 7(c) 是在编号为 2 的 C3 后面插入 CA 注意力模块。然后将该 CA 模块移动到编号为 4 的 C3 模块后面，后续依次移动到编号为 6 和编号为 9 的 C3 模块后面，如图 8 所示。最终结果均符合预期，在同一位置插入 CA 模块的 mAP 均高于插入 SE 模块的 mAP。同时，还发现随着注意力模块的后移，mAP 的值越高。因此，为了验

证注意力模块在 Backbone 中插入的位置对 mAP 的影响, 以 CA 模块为基础设计了相应的实验。分别将 CA 模块移动到编号为 2、4、6、9 的 C3 模块之前, 如图 9 所示, 观察其对结果的影响。

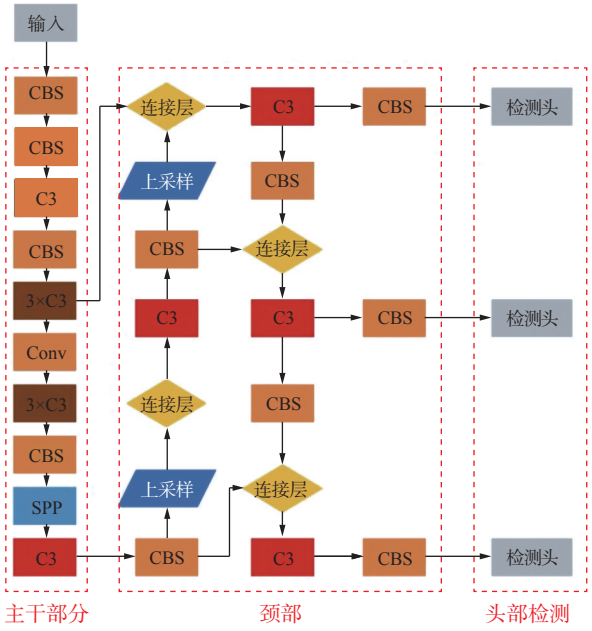


图 6　YOLOv5s 网络结构

Fig. 6　YOLOv5s network structure diagram

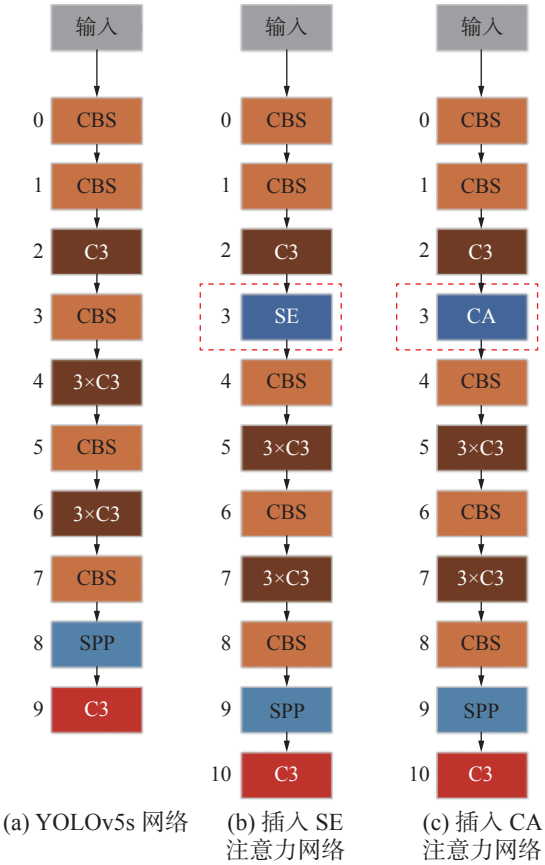


图 7　YOLOv5s 网络与插入注意力优化的网络结构对比

Fig. 7　Comparison figures among YOLOv5s and with attention network

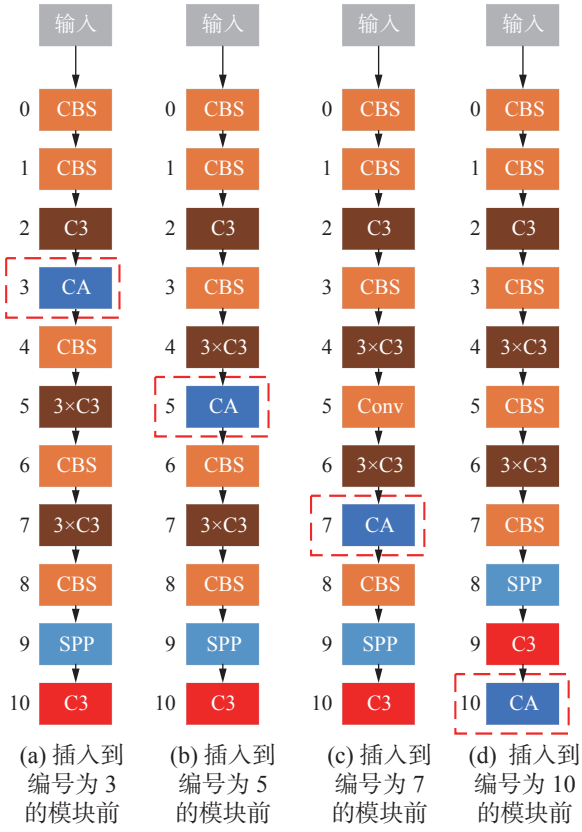


图 8　CA 注意力模块插入到 C3 模块后面不同位置

Fig. 8　CA module inserted in different positions after the C3

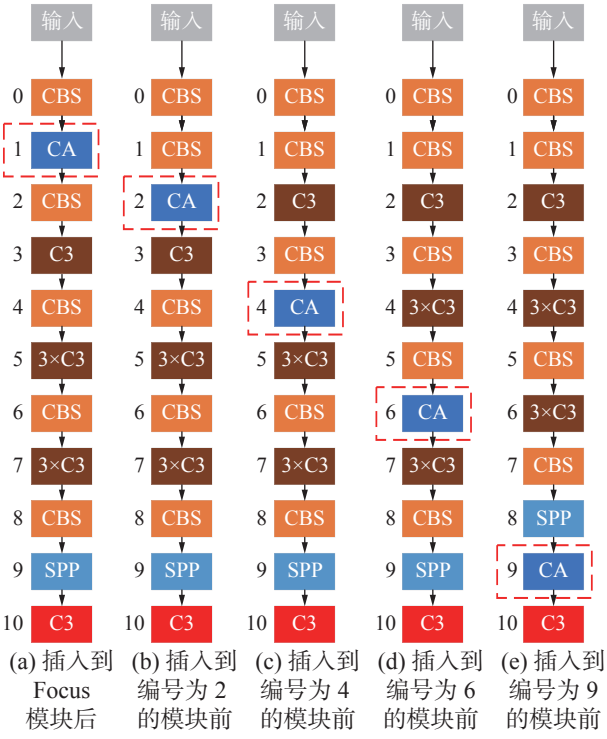


图 9　CA 注意力模块插入到 C3 模块之前不同位置

Fig. 9　CA module inserted in different positions before C3

4　实验及结果分析

4.1　实验环境及数据集

本文实验的硬件平台: CPU 采用 Intel (R)

Xeon (R) CPU E5-2678 v3 @ 2.50 GHz, GPU 采用 NVIDAI GTX 1070Ti , 操作系统为 Ubuntu 16.04, 开发语言是 Python 3.8, 深度学习训练框架采用 Pytorch 1.8.0。训练和测试网络时, 图片输入大小均为 640 像素×640 像素, 测试时使用的 batch size 均为 32, 所有实验均迭代训练到网络模型收敛为止。训练过程中所用到的超参设置如表 1 所示。

表 1　实验中超参数值的设置

Table 1　Setting of hyperparameter values in experiment

超参	数值
学习率	0.003 2
余弦退火超参数	0.12
动量	0.843
权重衰减参数	0.000 36
warmup的epochs数	2
warmup时的动量	0.5
边界框损失函数参数	0.029 6
分类损失函数参数	0.243
目标损失函数参数	0.301
IoU阈值	0.6

实验用到的数据集为 PASCAL VOC2012, 该数据集包含 20 类物体, 每张图片都有标注, 标注的物体包括人、动物、交通工具、家具等在内的 20 个类别, 平均每张图片有 2.4 个目标。

4.2　结果分析

将第 3 节讨论的基于注意力机制优化目标网络模型分别进行实验, 获得的实验结果如表 2 所示。当 Backbone 是 MobileNetV2, 检测头是 SSDLite-320 时, 添加 CA 注意力模块可以提高 mAP。同时在网络结构简单的 YOLOv3-tiny 中加入 CA 注意力模块, 也能提高 mAP。然而, 在 YOLOv5s 中加入 SE 注意力模块或 CA 注意力模块都会导致 mAP 的下降。这个实验的结果表明, 在不同的目标检测模型中, 注意力模块并不是随意插入就能提高 mAP, 而是需要具体情况具体分析。

表 2　不同目标检测模型添加注意力模块的结果

Table 2　Results of different attention optimal models

主干网络	检测头	参数量	mAP/%
MobileNetV2	SSDLite320	4.3	71.7
MobileNetV2 + SE	SSDLite320	4.7	71.7
MobileNetV2 + CBAM	SSDLite320	4.7	71.7
MobileNetV2 + CA	SSDLite320	4.8	73.1
CSPDakrNet53	YOLO	7.1	84.7
CSPDakrNet53 + SE	YOLO	7.2	81.7
CSPDakrNet53 + CA	YOLO	7.1	81.8
YOLOv3-tiny	YOLO	8.7	63.4
YOLOv3-tiny + CA	YOLO	8.8	64.4

如表 3 所示, 第 1 列表示注意力模块在 Backbone 当中的位置编号, 第 2 列表示 SE 模块在不同编号位置时 mAP 的结果, 第 3 列表示 CA 模块在不同编号位置时 mAP 的结果, 第 4 列表示在同一个位置插入 CA 模块与插入 SE 模块 mAP 的差值。

表 3　SE 模块与 CA 模块 mAP 结果对比

Table 3　Compared mAP results between SE and CA　%

编号	SE	CA	mAP
1	65.3	71.3	+6.0
3	73	73.5	+0.5
5	76.8	77.1	+0.3
7	79.8	80.3	+0.5
10	81.7	81.8	+0.1

从表 3 中可以先清楚地看到, 添加 CA 模块的网络 mAP 均优于 SE 模块的。尤其是将注意力模块插入到编号为 1 的实验, CA 注意力模块的 mAP 是 71.3%, 而 SE 注意力模块却只有 65.3%。CA 注意力模块表现尤为突出, 这说明空间注意力机制起到了作用, 并且空间注意力模块只在网络输入最开始的地方作用最大。而且随着层数的增加, 添加 SE 注意力模块和 CA 注意力模块的 mAP 都有提升, 这说明随着通道数的增多, 通道注意力机制占据主导作用。

结合第 3 节 YOLOv5-CCA 目标检测网络设计可以在 Backbone 不同编号位置插入 CA 注意力模块, 对应编号如图 8 中标注的 1、3、5、7、10 处; 和图 9 中标注的 2、4、6、9 处。从而获得不同编号位置插入 CA 注意力模块的 mAP 的结果如图 10 所示, 横坐标表示插入 Backbone 的位置编号, 纵坐标表示 mAP 的数值, mAP 的值越大表示网络模型的精度越高, 可以很清楚地看到: 在编号为 1、3、5、7、10 处插入 CA 注意力模块得到的性能表现均优于 2、4、6、9 处插入 CA 注意力模块。

这表明 CA 注意力模块放在残差结构之后能获得更好的性能表现。出现上述现象的原因可能是 CA 注意力模块的空间注意力机制使得后续残

差模块在融合不同尺寸大小特征图时造成了一定的阻碍, 使得残差结构表现变差。

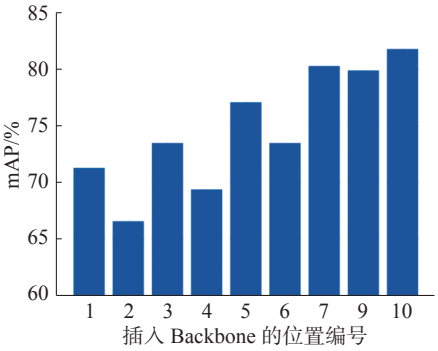


图 10　在 Backbone 不同编号位置插入 CA 模块得到 mAP 的结果

Fig. 10　mAP results by inserting CA models at different position

5　应用分析

欧美等发达国家早在三十多年前就已将计算机视觉技术应用在包括农业生产在内的很多领域, 时至今日, 计算机视觉在农业各个精细化分支下都取得了重大进展。在农作物种子质量检测的应用中, Zapotoczny 等[27] 于 2011 年通过利用神经网络的方法对春、冬季不同质量等级的 11 个小麦品种进行分类实验, 并取得了分类准确度高达 100% 的惊人效果。万鹏等[28] 于 2008 年提出了一套基于计算机视觉技术识别大米粒形的装置, 利用该装置代替人眼识别完整米粒及碎大米粒形, 并取得了完整米粒识别准确率为 98.67%、碎米识别准确率为 92.09% 的好结果。在农作物病虫草害的监测与防治的应用中, Wang 等[29] 于 2020 年通过利用害虫图像自动采集装置, 在田间诱捕害虫, 并采集害虫图像, 建立了 Pest24 数据集, 该数据集包含了中国农业部规定的 24 种主要农作物害虫, 共计 25 378 张图片, 平均每张图片中包含 2.3 个类别和 7.6 个实例对象。刘浏[30] 构建了两种大规模害虫图像标准数据集 Multi-class Pest Dataset 2018 (MPD2018) 和 AgriPest, MPD-2018 和 AgriPest 分别包含 88 670 与 49 707 张害虫图像, 以及对应的 582 170 和 264 728 个害虫目标数据标注, 并提出了一种基于混合全局与局部特征的农作物害虫图像检测方法。在农产品自动化收获的应用中, Tian 等[31] 针对果实大小、颜色、集群密度和其他会随着生长周期而变化的特征的问题, 将 YOLOv3 网络和 DenseNet 网络相结合, 提出了 YOLOv3-DenseNet 模型极大地改善了苹果在重叠和遮挡的情况下的检测性能。武星等[32] 为了提高采摘机器人在复杂环境下检测目标果实的检测速度, 提出了轻量化目标检测网络 Light-YOLOv3 用于苹果的实时检测。

全球小麦穗数据集(Global Wheat 2020)[33] 是第一个用于从田间光学图像中检测小麦头部的大规模数据集。它包括来自各大洲的大量栽培品种。小麦是一种在世界各地都会种植的主要粮食作物, 因此对小麦表型的研究获得了全世界相关领域研究者的兴趣。针对室外拍摄的野生小麦进行小目标检测是十分具有挑战性的。比如图像经常会出现小麦堆叠分布、伴随光照风速等自然因素、小麦生长的个体形态特征等诸多不利原因都会导致单个小麦头的识别变得困难。

小麦在全球种植广泛, 但是不同地域培育的品种不同, 其种植条件和种植密度不同, 因此设计小麦的目标检测模型, 使得它更具备泛化性和应用性。

Global Wheat 2020 数据集一共包含 4 700 张高分辨率的 RGB 图像和 190 000 个标记的小麦头, 这些图像来自 2016~2019 年在 10 个不同地点的 9 个机构收集的数据集, 涵盖来自欧洲、北美洲、亚洲和澳大利亚的基因型。其中训练集有 2 676 张图片, 验证集有 748 张图片, 测试集有 1 276 张图片。如图 11 所示 Global Wheat 2020 数据集中小麦穗的分布是十分杂乱无章的, 而且重叠遮挡部分也是很多的, 这对目标检测网络提出了巨大的挑战。

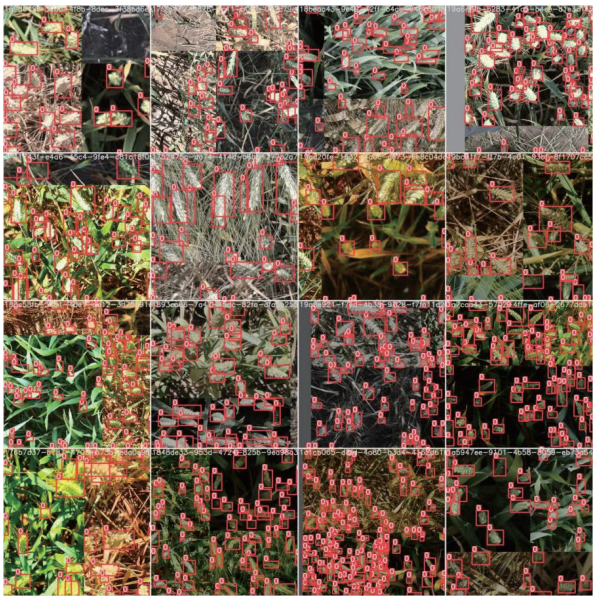


图 11　Global Wheat 2020 部分标签数据

Fig. 11　Global Wheat 2020 data tags

从表 4 的实验结果可以看出, 使用本文提到的 Slice-Concat 结构和将首个卷积层的步长变为 2 的改进策略, 能在不过分损失精度的前提下大

幅提升检测速度以及缩减模型的计算量，为其在实际应用场景中进行轻量化部署提供了一定的理论依据。从 YOLO5s 与 YOLO5s-Ghost 的实验结果对比也可以看出，使用 Ghost 模块也能有效地对 YOLOv5s 网络进行轻量化，使用 Ghost 模块以后使得 YOLOv5s 网络模型的参数量和计算量分别降低了 47.8% 和 50%。从图 12 对 Global Wheat 2020 数据集中部分图片的预测结果可以看出，本文提到的一些改进方法对麦穗的识别效果还是不错的，从表 4 的结果也可以看出，本文所提到的轻量化方法以及注意力插入策略在 Global Wheat 2020 数据集中的表现与在 PASCAL VOC 2012 数据集中的表现一致，这也证实了该方法及策略具有一定的通用性，为 YOLO 系列目标检测网络在实际生产与生活中进行轻量化部署提供了可靠的数据支撑。

表 4　本文实验在 Global Wheat 2020 数据集上鲁棒性验证结果

Table 4　Result of robustness verification Global Wheat 2020

模型	mAP/%	推理/ms	时间/ms	速率/(f·s^{-1})	参数规模/M
YOLOv3	95.3	38.7	44.3	22.6	61.49
YOLOv3-SC	93.6	9.1	15.8	63.3	61.49
YOLOv3-S2	92.7	8.4	11.3	88.5	61.52
YOLO5s	95.0	6.9	28.7	34.8	7.05
YOLOv5s-CCA	92.4	11.9	22	45.5	3.68
yolov5s-Ghost	91.3	12.1	24	46.2	3.76
YOLOv3-tiny	89.7	6.8	14.4	69.4	8.67
YOLOv3-tiny-CA	90.4	7.9	16	62.5	8.76

从图 12 对 Global Wheat 2020 数据集中部分图片的预测结果可以看出，本文提到的一些改进方法对麦穗的识别效果还是不错的，从表 4 的结果也可以看出，本文所提到的轻量化方法以及注意力插入策略在 Global Wheat 2020 数据集中的表现与在 PASCAL VOC 2012 数据集中的表现一致，这也证实了该方法及策略具有一定的通用性，为 YOLO 系列目标检测网络在实际生产与生活中进行轻量化部署提供了可靠的数据支撑。

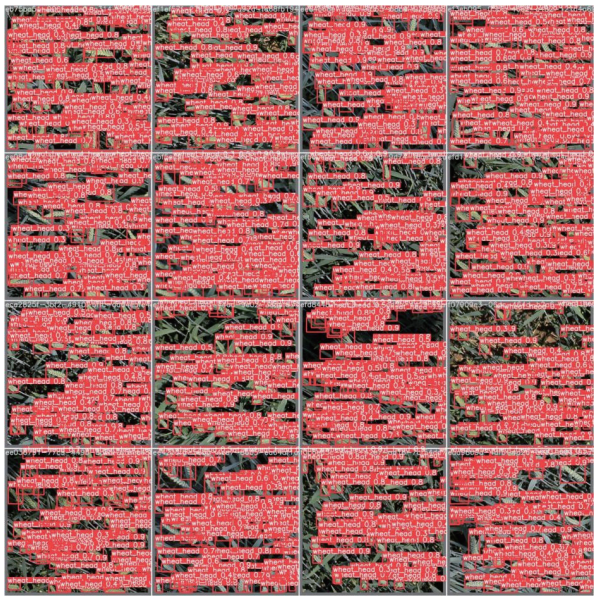


图 12　Global Wheat 2020 部分预测结果

Fig. 12　Prediction result of Global Wheat2020

6　结束语

本文提出的 YOLOv5s -CCA 目标检测网络在 PASCAL VOC 2012 数据集进行实验获得了良好结果。进一步在 Global Wheat 2020 数据集实验也表现出色，这也证实了该模型具有一定的通用性和良好的泛化性，为 YOLO 系列目标检测网络在实际生产与生活中进行轻量化部署提供了可靠的数据支撑。本文通过在不同目标检测网络模型中插入注意力模块，发现在不同的目标检测模型中，注意力模块并不是随意插入就能提高 mAP，而是需要具体情况具体分析。然后选择在 YOLOv5s 的 Backbone 的不同位置插入 SE 注意力模块和 CA 注意力模块训练模型，从而探索出了通道注意力模块和空间注意力模块在目标检测网络中最优的插入位置，即空间注意力模块应该在输入图像初试阶段就使用，可以获得最佳性能表现。通道注意力模块则是随着通道数量的增加，性能表现会越来越好，所以通道注意力模块只需要在通道数最多的那一层添加就能获得最佳性能表现。本文提到的所有优化方法及策略具有一定的通用性，为 YOLO 系列目标检测网络在特定领域进行轻量化部署与应用提供了一定参考价值。

参考文献：

[1]　GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich

feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580−587.

[2] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904–1916.

[3] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2016: 1440−1448.

[4] REN Shaoqing, HE Kaiming, 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.

[5] 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. Honolulu: IEEE, 2017: 936−944.

[6] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.

[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. Las Vegas: IEEE, 2016: 779−788.

[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020−04−23)[2022−06−08].https://arxiv.org/abs/2004.10934.

[9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6517−6525.

[10] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018−04−08) [2022−06−08]. https://arxiv.org/abs/1804.02767.

[11] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. Scaled-YOLOv4: scaling cross stage partial network[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13024−13033.

[12] 陈科圻, 朱志亮, 邓小明, 等. 多尺度目标检测的深度学习研究综述 [J]. 软件学报, 2021, 32(4): 1201–1227.
CHEN Keqi, ZHU Zhiliang, DENG Xiaoming, et al. Deep learning for multi-scale object detection: a survey[J]. Journal of software, 2021, 32(4): 1201–1227.

[13] 毛莺池, 唐江红, 王静, 等. 基于 Faster R-CNN 的多任务增强裂缝图像检测方法 [J]. 智能系统学报, 2021, 16(2): 286–293.
MAO Yingchi, TANG Jianghong, WANG Jing, et al. Multi-task enhanced dam crack image detection based on Faster R-CNN[J]. CAAI transactions on intelligent systems, 2021, 16(2): 286–293.

[14] 邵江南, 葛洪伟. 一种基于深度学习目标检测的长时目标跟踪算法 [J]. 智能系统学报, 2021, 16(3): 433–441.
SHAO Jiangnan, GE Hongwei. A long-term object tracking algorithm based on deep learning and object detection[J]. CAAI transactions on intelligent systems, 2021, 16(3): 433–441.

[15] 赵文清, 杨盼盼. 双向特征融合与注意力机制结合的目标检测 [J]. 智能系统学报, 2021, 16(6): 1098–1105.
ZHAO Wenqing, YANG Panpan. Target detection based on bidirectional feature fusion and an attention mechanism[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1098–1105.

[16] 田永林, 王雨桐, 王建功, 等. 视觉 Transformer 研究的关键问题: 现状及展望 [J]. 自动化学报, 2022, 48(4): 957–979.
TIAN Yonglin, WANG Yutong, WANG Jiangong, et al. Key problems and progress of vision transformers: the state of the art and prospects[J]. Acta automatica sinica, 2022, 48(4): 957–979.

[17] 郭璠, 张泳祥, 唐琎, 等. YOLOv3-A: 基于注意力机制的交通标志检测网络 [J]. 通信学报, 2021, 42(1): 87–99.
GUO Fan, ZHANG Yongxiang, TANG Jin, et al. YOLOv3-A: a traffic sign detection network based on attention mechanism[J]. Journal on communications, 2021, 42(1): 87–99.

[18] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[EB/OL]. (2015−02−05) [2022−06−08].https://arxiv.org/abs/1506.02025.

[19] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8): 2011–2023.

[20] ZHAO Bo, WU Xiao, FENG Jiashi, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE transactions on multimedia, 2017, 19(6): 1245–1256.

[21] VOLODYMYR M, NICOLAS H, ALEX G, et al. Recurrent models of visual attention[EB/OL]. (2014−06−24) [2022−06−08].https://arxiv.org/abs/1406.6247v1.

[22] WU Jun, ZHU Jiahui, TONG Xin, et al. Dynamic activation and enhanced image contour features for object detection[J]. Connection Science, 2022, 12: 1–21.

[23] WANG Fei, JIANG Mengqing, QIAN Chen, et al. Residual attention network for image classification[C]//2017 IEEE Conference on Computer Vision and Pattern Recog-

nition. Honolulu: IEEE, 2017: 6450−6458.

[24] HOU Qibin, ZHOU Daquan, FENG Jiashi. Coordinate attention for efficient mobile network design[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 13708−13717.

[25] HAMI D R, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 658−666.

[26] ZHENG Zhaohui, WANG Ping, LIU Wei, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(7): 12993–13000.

[27] ZAPOTOCZNY P. Discrimination of wheat grain varieties using image analysis and neural networks. Part I. Single kernel texture[J]. Journal of cereal science, 2011, 54(1): 60–68.

[28] 万鹏, 孙瑜, 孙永海. 基于计算机视觉的大米粒形识别方法 [J]. 吉林大学学报 (工学版), 2008, 38(2): 489–492.
WAN Peng, SUN Yu, SUN Yonghai. Recognition method of rice kernel shape based on computer vision[J]. Journal of Jilin university (engineering and technology edition), 2008, 38(2): 489–492.

[29] WANG Qijin, ZHANG Shengyu, DONG Shifeng, et al. Pest24: a large-scale very small object data set of agricultural pests for multi-target detection[J]. Computers and electronics in agriculture, 2020, 175: 105585.

[30] 刘浏. 基于深度学习的农作物害虫检测方法研究与应用 [D]. 合肥: 中国科学技术大学, 2020.
LIU Liu. Research and applications on agricultural crop pest detection techniques based on deep learning[D]. Hefei: University of Science and Technology of China, 2020.

[31] TIAN Yunong, YANG Guodong, WANG Zhe, et al. Apple detection during different growth stages in orchards using the improved YOLO-V3 model[J]. Computers and electronics in agriculture, 2019, 157: 417–426.

[32] 武星, 齐泽宇, 王龙军, 等. 基于轻量化 YOLOv3 卷积神经网络的苹果检测方法 [J]. 农业机械学报, 2020, 51(8): 17–25.
WU Xing, QI Zeyu, WANG Longjun, et al. Apple detection method based on light-YOLOv3 convolutional neural network[J]. Transactions of the Chinese society for agricultural machinery, 2020, 51(8): 17–25.

[33] DAVID E, MADEC S, SADEGHI-TEHRAN P, et al. Global wheat head detection (GWHD) dataset: a large and diverse dataset of high-resolution RGB-labelled images to develop and benchmark wheat head detection methods[J]. Plant phenomics, 2020: 3521852.

作者简介:

吴珺, 副教授, 博士, 主要研究方向为深度学习及多模态数据分析、大数据分析及应用、智能方法优化。主持国家自然科学基金及湖北省自然科学基金; 参与研发各类省部级项目 5 项, 并发表学术论文 16 篇。

董佳明, 硕士研究生, 主要研究方向为目标检测、大数据技术。

刘欣, 硕士研究生, 主要研究方向为目标检测、智能方法。