计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):333−340 [doi: 10.15888/j.cnki.csa.008576]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于注意力 YOLOv5 模型的自动水果识别①

曹秋阳[1], 邵叶秦[2], 尹　和[1]

[1](南通大学 信息科学技术学院, 南通 226019)
[2](南通大学 交通与土木工程学院, 南通 226019)
通信作者: 邵叶秦, E-mail: hnsyk@ntu.edu.cn

摘　要: 近年来, 人工智能在各个领域有着广泛的应用. 针对超市及菜市场人工称重操作耗时、计价流程繁杂的问题, 本文提出一种基于注意力 YOLOv5 模型的水果自动识别算法. 首先, 为了提升仅有局部特征不同, 全局特征相似水果的识别准确率, 本文在 YOLOv5 的 SPP (spatial pyramid pooling) 层后增加 SENet (squeeze-and-excitation networks), 采用注意力机制自动学习每个特征通道的重要程度, 进而按照重要程度强化对水果识别任务有用的特征并抑制没有用的特征; 其次, 针对水果识别预测框与目标框重叠时, *GIOU* 不能准确表达边框重合关系问题, 本文将原有的边框回归损失函数 *GIOU* 替换为 *CIOU*, 同时考虑目标框与预测框的高宽比和中心点之间的关系, 从而使水果预测框更加接近真实框, 提升预测精度. 实验结果表明, 改进后的模型在常见场景下水果识别能力有明显提升, 平均精度 mAP 达 99.10%, 识别速度 FPS 达到 82, 能够满足实际应用需要.

关键词: YOLOv5; 水果识别; CIOU; 注意力机制; 目标检测; 深度学习

引用格式: 曹秋阳,邵叶秦,尹和.基于注意力 YOLOv5 模型的自动水果识别.计算机系统应用,2022,31(7):333–340. http://www.c-s-a.org.cn/1003-3254/8576.html

Automatic Fruit Recognition Based on Attention YOLOv5 Model

CAO Qiu-Yang[1], SHAO Ye-Qin[2], YIN He[1]

[1](School of Information Science and Technology, Nantong University, Nantong 226019, China)
[2](School of Transportation and Civil Engineering, Nantong University, Nantong 226019, China)

Abstract: In recent years, artificial intelligence has been widely used in various fields. To address time-consuming manual weighing and complicated pricing procedures in supermarkets and vegetable markets, this study proposes an automatic fruit recognition model based on attention YOLOv5. First, to improve the recognition accuracy of fruits with different local features but similar global features, the study adds squeeze-and-excitation networks (SENet) after the spatial pyramid pooling (SPP) layer of YOLOv5 and uses the attention mechanism to automatically learn the importance of each feature channel. Further, the useful features for fruit recognition tasks according to the importance are strengthened and those useless are suppressed. Second, when the fruit recognition prediction frame overlaps the target frame, *GIOU* cannot accurately express the overlapping relationship of the frames. In response, this study replaces the original frame regression loss function *GIOU* with *CIOU* and considers the relationships of aspect ratio and center point between the target frame and the prediction frame. In this way, the fruit prediction frame is closer to the real frame, and thereby the prediction accuracy is improved. Experimental results show that the improved model has significantly improved fruit recognition ability in common scenarios with a mean average precision (mAP) of 99.10% and a recognition speed of 82 FPS, which can meet the needs of practical applications.

Key words: YOLOv5; fruit recognition; CIOU; attention mechanism; object detection; deep learning

① 基金项目: 国家自然科学基金面上项目 (61671255); 南通市科技项目 (MS12020078); 国家级大学生创新训练项目 (202110304050Z, 202110304047Z)
收稿时间: 2021-10-08; 修改时间: 2021-11-08; 采用时间: 2021-11-19; csa 在线出版时间: 2022-03-18

近年来, 随着科学技术的快速发展, 人工智能给人们生活带来了便捷和智能化的服务. 水果自动识别在超市、菜市场、果园等很多场景有着重要的应用. 超市以及菜市场可以结合水果称重, 自动计算水果的价格, 提高顾客购买的效率. 果园可以通过水果的检测与识别, 估计水果的收成, 并利于机械化自动采摘.

目前, 越来越多的国内外研究人员聚焦果蔬识别. 彭红星等[1] 提出一种改进的 single shot multibox detector (SSD) 水果检测模型, 将 SSD 模型主干网络 VGG16 替换为 ResNet-101 网络, 并通过随机梯度下降算法以及迁移学习思想优化 SSD 模型, 在 4 种水果上的检测精度达到 88.4%. 王辉等[2] 在 Darknet-53 网络的基础上使用组归一化代替原先的批量归一化, 继而引入 YOLOv3[3] 算法构建水果检测模型, 实现水果的准确识别. Bargoti 等[4] 设计了基于 Faster-RCNN 的目标检测模型实现自然环境下 3 种水果的检测. Liu 等[5] 提出了 single shot detector 方法, 用于对象的检测和识别, 在保证准确率的同时提高了效率. 这些方法普遍存在如下问题: (1) 数据集中水果种类过少; (2) 模型倾向于对象的全局信息, 容易忽略某些关键及重要的水果局部信息; (3) 目标框与预测框重合时未考虑它们之间的相互关系, 容易出现预测结果不精确问题.

因此, 本文采用包括不同光照、不同角度等的 15 种水果组成的数据集, 并使用基于注意力的 YOLOv5 模型实现水果的准确分类和识别. 具体来说, 该模型在主干网络后增加注意力机制 squeeze-and-excitation networks (SENet), 通过神经网络计算通道注意力权重, 以增强水果的重要特征, 减弱不重要的特征, 使提取的特征更具代表性且保留局部的重要信息, 提升水果识别的准确率. 同时, 将原先的 *GIOU* 损失函数替换为包括边框长宽比信息和中心点位置关系的 *CIOU* 损失函数, 使预测框更加接近真实框. 实验证明, 本文基于注意力的 YOLOv5 模型在准确率及速度上都优于目前最新的水果识别算法.

1 YOLOv5 模型

YOLOv5 是由 Ultralytics LLC 公司提出的深度神经网络模型. 相比于早期的 YOLO 模型[3,6], YOLOv5 模型体积小、速度快、精度高, 受到工业界的青睐. 具体来说, 对比于 YOLOv4, YOLOv5 进行了如下改进. 首先, 对输入图片经过 Focus 切片操作, 保留了更完整的图片下采样的信息; 其次, 采用 CSPDarknet-53 主干网络进行特征提取, 分别在主干网络以及 Neck 部分设计了两种 CSP 结构用来调整残差组件的数量以及卷积层数量; 最后, 在 Neck 部分输出小、中、大 3 层特征. 虽然 YOLOv5 主干网络后的 spatial pyramid pooling (SPP) 层解决了输入图像特征尺寸不统一的问题, 但是没有对特征图进行通道间的加权融合. 为此, 本文通过软自注意力的方式融合图像特征, 强调有效特征, 提高水果识别的准确率.

2 基于注意力 YOLOv5 模型的自动水果识别

本文实现基于注意力 YOLOv5 模型的自动识别水果, 流程如图 1 所示. 首先, 将数据集进行预处理, 接着输入主干网络提取特征, 并使用 SENet 注意力模块得到一个与通道对应的一维向量作为评价分数; 其次, 将评价分数通过乘法操作作用到 feature map 的对应通道上, 得到用于水果识别的有效特征; 然后, 经过 feature pyramid networks (FPN)[7] 和 path aggregation network (PAN)[8] 结构将特征融合并获得语义信息更强, 定位信息更准的特征图; 最后, 经过类别分类与预测框回归计算得到精准检测结果.

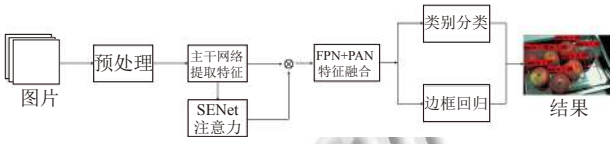


图 1 本文方法的处理流程图

2.1 预处理

2.1.1 Mosaic 数据增强

Mosaic 数据增强的方式参考了 CutMix[9] 数据增强思想. CutMix 数据增强将两张图片进行拼接, 而 Mosaic 采用 4 张图片的拼接, 增加数据量的同时可以丰富检测物体的背景, 如图 2 所示.

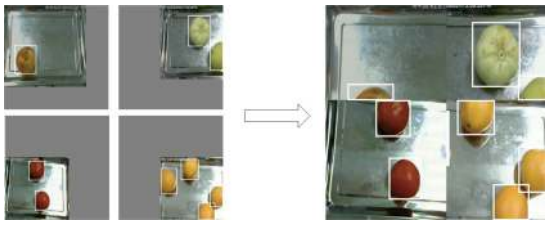


图 2 Mosaic 数据增强

2.1.2 自适应锚框

在 YOLO 系列算法中, 通常对不同的数据集都会设定初始长宽的锚框. 在 YOLOv3、YOLOv4 中, 初始

锚框都是通过单独算法得到的, 常用的是 K-means 算法. 本文将这种功能嵌入至代码中, 实现了每次训练可以自适应的计算不同训练集中的最佳初始锚框. 本文的初始锚框为 [10, 13, 16, 30, 33, 23]、[30, 61, 62, 45, 59, 119]、[116, 90, 156, 198, 373, 326], 经过计算本文最佳初始锚框为 [111, 114, 141, 121, 127, 141]、[150, 149, 159, 169, 195, 212]、[256, 173, 173, 292, 326, 298].

2.1.3 自适应缩放图片

数据集的大小往往都是大小不一, 需要对其尺寸归一化. 然而, 实际项目中的很多图片长宽比不一致, 缩放并填充后, 两端填充部分较多, 存在很多冗余信息, 影响模型速度及效果. 本文方法对原始图像进行自适应填充最少的灰度值, 使得图像高度或宽度两端的灰度值最少, 计算量也会随之减少, 速度也得到提升. 具体步骤如下.

(1) 图像缩放比例. 假设原始图像为 1000×800, 缩放至 416×416. 将 416×416 除以原始图像相应宽高, 得到系数分别为 0.416 和 0.52, 取其较小值 0.416.

(2) 缩放后的尺寸. 将原始图片宽高乘以较小的系数 0.416, 则宽为 416, 高为 332.

(3) 灰边的填充值. 先将 416−332=84, 并采用取余的方式得需要填充的像素值 84%32=20 (32 是由于网络经过了 5 次下采样, 2 的 5 次方为 32), 两端各 10 个像素. 在测试过程中采用灰色填充, 训练过程依旧使用原始的 resize 操作以提高物体的检测、计算速度.

2.2 主干网络

2.2.1 特征提取网络

为了在水果图像上提取丰富的特征, 受到 YOLOv5 的启发, 本文使用 CSPDarknet-53 作为主干网络. CSPDarknet-53 可以增强卷积网络的学习能力, 降低内存消耗.

CSPDarknet-53 主干网络包括 Focus、Mosaic、多次卷积、残差结构等, 其中 CSP1_X 用来调整残差组件的数量, 如图 3 所示. Neck 中的 CSP2_X 则是用来对卷积层数量的调整, 如图 4 所示. CSPDarknet-53 提取的特征后续用于得到通道注意力.

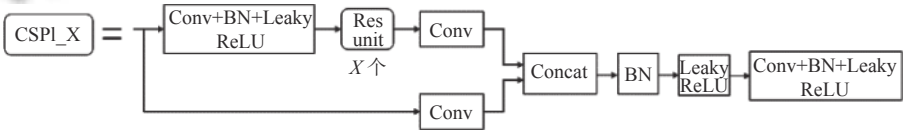


图 3 CSP1_X 结构

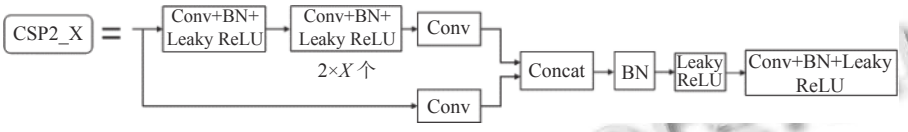


图 4 CSP2_X 结构

2.2.2 SELayer

为了得到不同特征通道的权重, 强化重要通道, 减弱次要通道, 本文使用 SENet[10] 注意力机制学习通道权重. SENet 可以学习通道之间的相关性, 生成通道注意力. 虽然计算量有所增加, 但是提取的特征更加有效. 图 5 是 SENet 模型示意图. 首先, 使用全局平均池化作为 Squeeze 操作; 其次, 使用两个全连接层得到通道间的相关性, 同时减少参数与计算量; 然后, 通过 Sigmoid 归一化权重; 最后, 通过 Scale 操作将归一化后的权重作用在原始通道的特征上. 本文是将 SELayer 嵌入至 SPP[11] 模块, 如图 6 所示. SPP 作为一种 Inception 结构, 嵌入了水果多尺度信息, 聚合了不同感受野上的特征, 因此使用 SELayer 能够对卷积特征通道重新加权, 增强重要特征之间的相互依赖, 可以学习到不同通道特征的重要程度, 从而产生更好的效果并提升识别性能.

针对全局特征差别不大 (大小、形状、颜色等), 某些局部特征有差异的水果, 注意力机制 SENet 能够增强水果的重要特征, 减弱不重要的特征, 使得提取的水果特征更加具有代表性且保留局部重要信息. 如图 7 特征图所示, 本文选取前 16 张特征图, 青苹果与番石榴的大小、形状、颜色等全局特征相似, 而部分区域颜色、表面纹理以及根蒂等有所不同. 如图 7(b)、图 7(e) 所示, 在没有进行 SENet 操作前, 两者特征信息类似, 特征像素未体现出特征的重要程度, 经过 SENet 操作后, 如图 7(c)、图 7(f) 所示, 根据特征重要程度将特征像素进行重新加权计算, 一方面减弱了周边不重要的信息, 另一方面突出了两种水果局部纹理、形状等重要特征, 有利于准确识别出青苹果与番石榴.

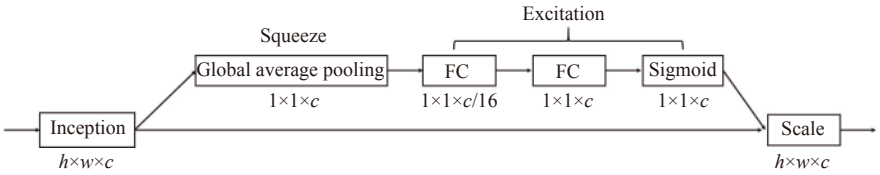


图 5 SENet 结构

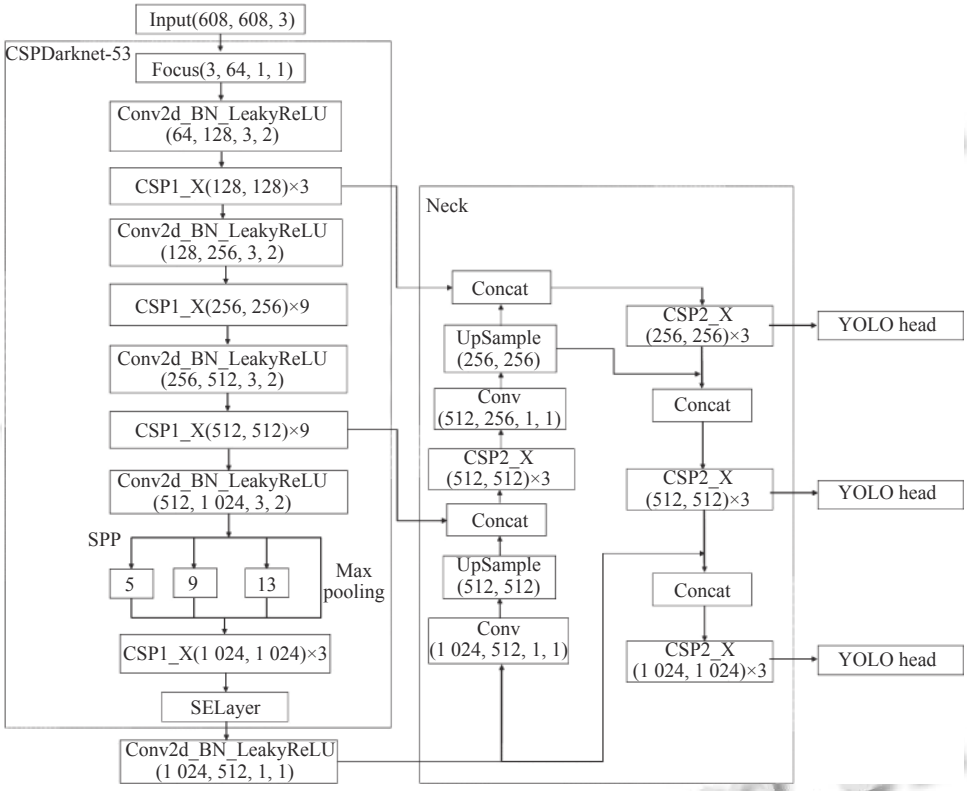
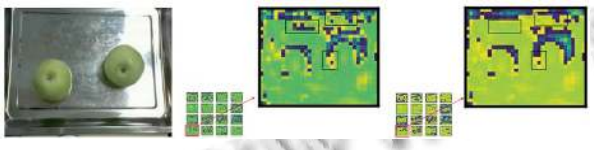
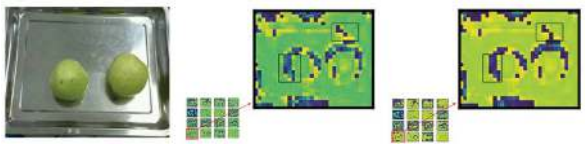


图 6 改进 YOLOv5 模型结构



(a) 青苹果 (b) 青苹果 SPP 层特征图 (c) 青苹果 SENet 层特征



(d) 番石榴 (e) 番石榴 SPP 层特征图 (f) 番石榴 SENet 层特征图

图 7 特征图对比图

2.3 FPN+PAN 特征融合

为了获得更强的语义信息以及更为精准的位置信息实现水果准确识别, 本文采用特征金字塔 FPN+PAN 提取多层次的特征, 顶层特征包含丰富的语义信息, 而底层特征具有精准的位置信息, 如图 8 所示, 其中, (a) 区域为 FPN 部分, (b) 区域为 PAN 部分.

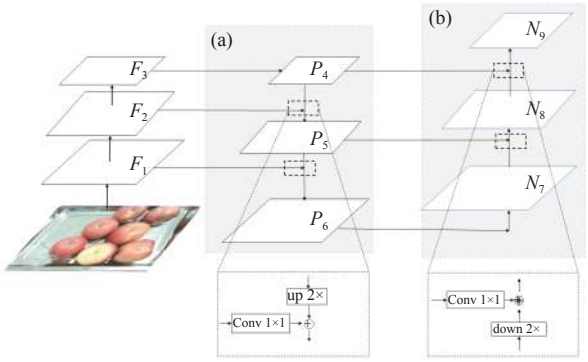


图 8 FPN+PAN 结构图

FPN 设计了自顶向下和横向连接的结构, 这样的好处是既利用了顶层语义特征 (利于分类), 又利用了

底层的高分辨率信息 (利于定位), 如图 9 所示.

本文在 FPN 后增加自底向上的特征金字塔 PAN, 将底层的特征信息通过下采样的方式进行融合, 将底层定位信息传送至顶层, 这样的操作是对 FPN 的补充, 将底层的强定位特征传递上去.

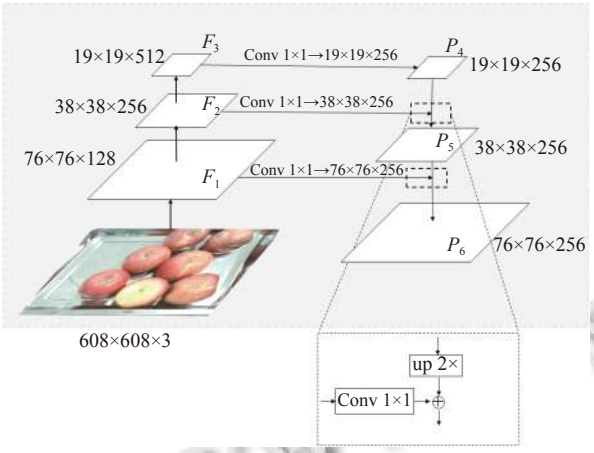


图 9 FPN 结构图

通过组合 FPN+PAN 两个模块, 对不同的检测层进行参数的聚合, 增强语义信息的同时, 提高目标的定位精度从而全面的提升模型的鲁棒性和准确率.

2.4 损失函数

2.4.1 *GIOU*

YOLOv5 采用 GIOU_Loss[12] 作为 bounding box 的损失函数. 具体来说, 对于两个 bounding box A、B (如图 10), 首先, 算出 A、B 的最小外接矩形 C; 其次, 计算 C 中没有覆盖 A 和 B 的面积 (即差集) 占 C 总面积的比值; 最后, 用 A 与 B 的 *IOU* 减去这个比值:

$$GIOU = IOU - \frac{C-(A\cup B)}{C}$$

$$GIOU_{\text{Loss}} = 1 - \left(IOU - \frac{C-(A\cup B)}{C}\right)$$

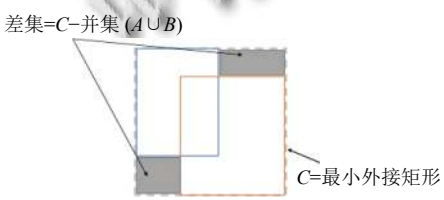


图 10 *GIOU* 示意图

相比于 *IOU*, *GIOU* 一方面解决了当预测框与目标框不相交 (*IOU*=0) 时损失函数不可导的问题; 另一方面, 当两个预测框大小相同、*IOU* 相同时, *IOU* 损失函数无法区分两个预测框相交的不同之处, *GIOU* 则缓解这种情况的发生.

但是, 如图 11 所示, 当预测框与目标框重叠时, 则 *GIOU* 的值与 *IOU* 值相同, 它们的效果一致, 因此难以区分两者相对的位置关系.

(a) (b) (c)

图 11 目标框与预测框重叠, *GIOU*=*IOU*=0.85

2.4.2 *CIOU*

针对 GIOU_Loss 损失函数所产生的问题, 本文采用 CIOU_Loss[13] 替换了 GIOU_Loss. GIOU_Loss 解决了边框不重合的问题, 而 CIOU_Loss 在其基础上不仅考虑了边框重合问题, 而且将边框高宽比和中心的位置关系等信息也考虑进去, 使得预测框的回归速度与精度更高.

CIOU 是将真实框与预测框之间的距离、重叠率、边框尺度以及惩罚因子均考虑进去, 使得目标边框回归更加稳定, 有效的解决 *IOU* 在训练过程中发散的问题, 如图 12 所示.

式 (1) 为 *CIOU* 公式:

$$CIOU = IOU - \frac{\rho^2(b,b^{gt})}{c^2} - \alpha v \tag{1}$$

其中, $\rho^2(b,b^{gt})$即图 10 中预测框与真实框中心点之间的欧式距离 d, c 表示同时包含真实框与预测框最小闭包矩形框的对角线距离.

式 (2) 为惩罚项αv中α的公式:

$$\alpha = \frac{v}{1-IOU+v} \tag{2}$$

式 (3) 为惩罚项αv中v的公式:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{3}$$

其中, w^{gt}和h^{gt}分别表示真实框的宽和高, w和h分别表示预测框的宽和高.

式 (4) 为 *CIOU* 在回归时 Loss 的计算公式:

$$CIOU_{\text{Loss}} = 1 - IOU + \frac{\rho^2(b,b^{gt})}{c^2} + \alpha v \tag{4}$$

如图 13 所示, 目标框与预测框重合时, *CIOU* 值也不相同. c 值相同时, 通过目标框与预测框中心点的欧式距离与对角线的比值 d, 有效度量两者位置关系, 损失函数能够有效收敛.

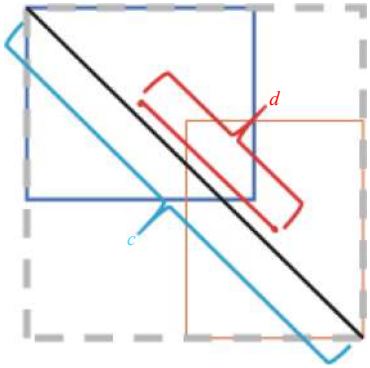


图 12 *CIOU* 示意图

(a) *CIOU*=0.80 (b) *CIOU*=0.83 (c) *CIOU*=0.79

图 13 目标框与预测框重叠, *CIOU* 值不同

3 实验

3.1 数据采集与预处理

本文的水果数据集部分来自于网上公开数据集, 部分来自于手机拍摄的数据, 所用数据均为模拟称重时俯拍的水果图片. 水果类别共有 15 种, 共计 13 676 张, 训练集、验证集、测试集的比例为 8:1:1 (训练集 10 940 张, 验证集和测试集均为 1 368 张), 具体类别及数量如表 1 所示.

表 1 数据集表

水果分类	数量	训练集	验证集	测试
Apple (苹果)	1 050	844	103	103
Banana (香蕉)	999	809	95	95
Carambola (杨桃)	1 050	814	118	118
Kiwi (猕猴桃)	960	746	107	107
Mango (芒果)	907	715	96	96
Pitaya (火龙果)	985	815	85	85
Guava (番石榴)	913	717	98	98
Muskmelon (香瓜)	899	713	93	93
Orange (橙子)	943	769	87	87
Peach (桃子)	883	695	94	94
Pear (梨)	833	649	92	92
Persimmon (柿子)	794	650	72	72
Plum (李子)	848	704	72	72
Pomegranate (石榴)	794	652	71	71
Tomato (番茄)	818	650	84	84

3.2 实验配置

本文实验是在深度学习开发框架 PyTorch 下进行, 工作站的配置为 Ubuntu 16.04.6、内存 64 GB、显存 12 GB、GPU 为 NVIDIA TITAN Xp、CUDA 10.2 版本以及 CUDNN 7.6.4.

3.3 模型训练

模型训练过程中, epoch 共 100 次, 学习率为 0.01, batch_size 为 16, 权重衰减数为 0.000 5. 训练过程中, 模型训练集损失函数损失值 (box、objectness、classification)、验证集损失值 (val box、val objectness、val classification)、查准率 (precision)、召回率 (recall) 以及平均精度 (mAP@0.5、mAP@0.5:0.95) 如图 14. 图 15 给出 15 类水果在验证集上的 P-R 曲线图.

3.4 *CIOU* 效果验证

为了证明 *CIOU* 的有效性, 我们进行了对比实验. 在 YOLOv5 模型的基础上, 将 *GIOU* 损失函数改为对应的 *CIOU* 损失函数. 实验结果如表 2 所示.

从表 2 中可以看出, 利用 *CIOU* 作为边框回归损失函数, 模型 mAP 值为 97.72%, 提升 1.57%, 证明了 *CIOU* 损失函数的有效性.

3.5 SELayer 效果验证

为了证明 SELayer 的有效性, 我们同样进行了对比实验. 在 YOLOv5+CIOU 模型的基础上增加注意力模块 SELayer. 实验结果如表 3 所示.

从表 3 中, 可以看出, 在 YOLOv5+CIOU 的基础上增加 SENet 注意力机制模块, 即本文基于注意力 YOLOv5 模型, mAP 值为 99.10%, 提升了 1.38%, 精度提升的同时, 模型的速度并没有下降, 证明了 SELayer 的有效性.

如图 16 所示, 在形状、颜色、纹理、大小类似的两种水果中, 图 16(a) 为苹果, 图 16(b) 为番石榴, 模型能够准确识别.

3.6 模型鲁棒性检验

为了验证本文方法的鲁棒性, 本文检测了 15 种水果, 并分别考虑了光照、遮挡等因素. 如图 17–图 20.

(1) 不同光照. 如图 17、图 18.

(2) 有遮挡. 如图 19.

(3) 同类别不同品种. 如图 20.

通过对比发现, 本文模型在遮挡、不同光照、多目标等情况下水果的识别效果更好、鲁棒性更好, 输出的预测框相比更符合目标水果.

3.7 与最新方法的对比

为了验证方法的有效性, 除了对比 YOLOv5 模型, 本文对比了最新主流的 Faster-RCNN、YOLOv4 模型, 如表 4 所示.

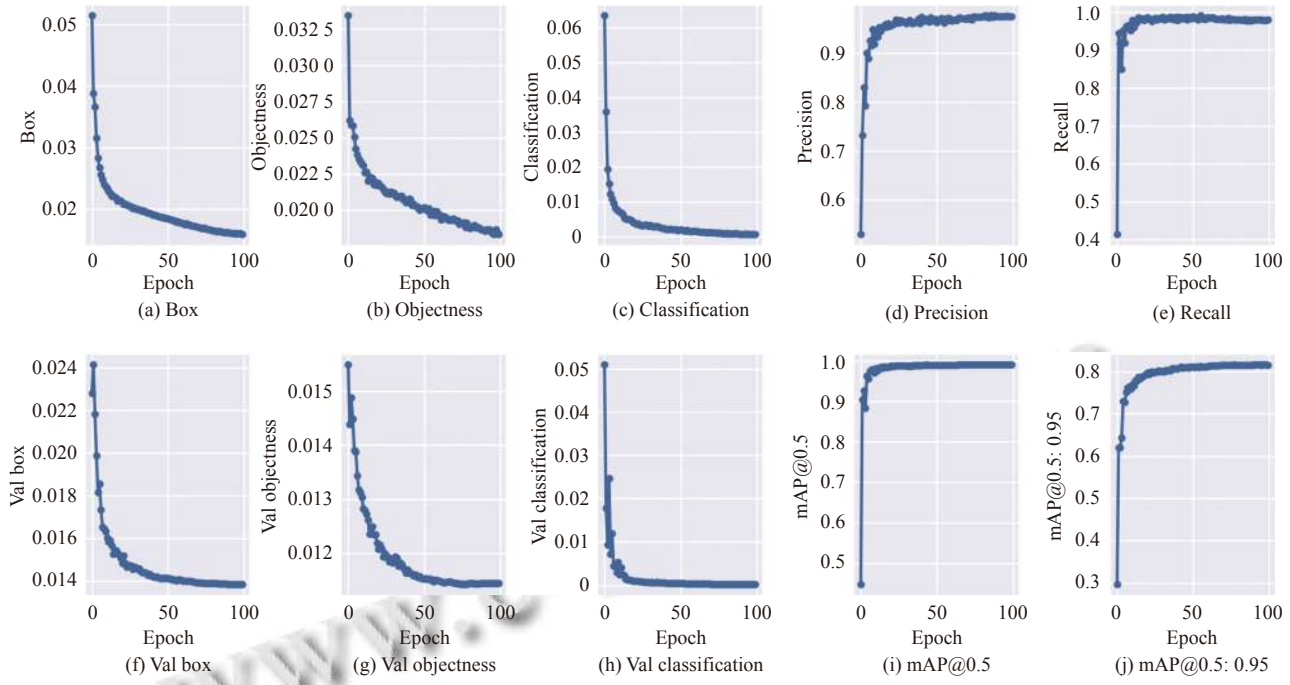


图 14 各项性能指标

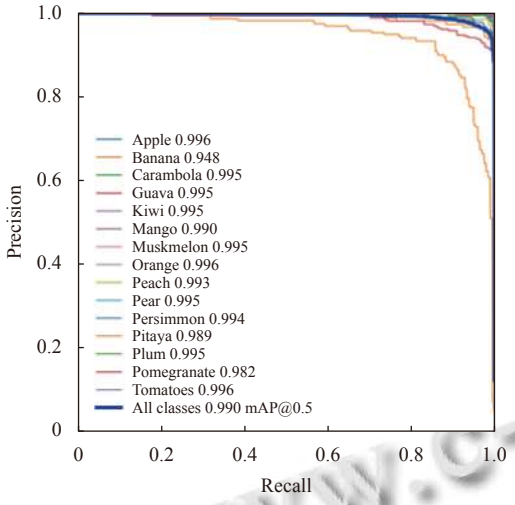


图 15 P-R 曲线图

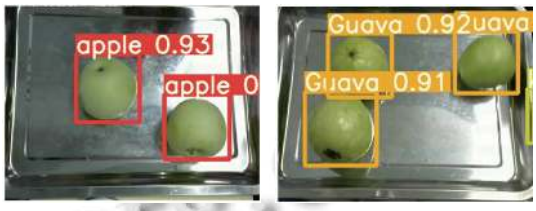
表 2 *CIOU* 效果验证性能对比

模型	mAP (%)	FPS (帧/s)
YOLOv5 (GIOU)	96.15	82
YOLOv5 (CIOU)	97.72	82

表 3 SELayer 效果验证性能对比

模型	mAP (%)	FPS (帧/s)
YOLOv5 (CIOU)	97.72	82
YOLOv5 (CIOU)+SELayer (ours)	99.10	82

由表 4 可见, Faster-RCNN 的 mAP 为 95.49%, YOLOv4 的 mAP 为 95.39%, YOLOv5 的 mAP 为 96.15%, 而本文方法的 mAP 为 99.10%, 识别速度到 82 帧/s, 在准确率及速度上都优于其他主流的对比方法.



(a) 苹果 (b) 番石榴

图 16 SELayer 效果对比

(a) YOLOv5 测试结果

(b) 本文模型测试结果

图 17 光照较强下测试效果对比

4 结语

本文采用基于注意力 YOLOv5 算法模型实现 15 类水果的自动识别. 实验表明, 本文方法是鲁棒的, 并且在水果识别准确率和识别速度上都优于主流 Faster-

RCNN、YOLOv4 和传统 YOLOv5 算法. 在后续的研究中, 将考虑更多种类的水果, 在保证水果种类的多样性时也能够保证模型的泛化能力以及识别准确率与速度.

(a) YOLOv5 测试结果

(b) 本文模型测试结果

图 18　光照较弱下测试效果对比

图 19　有遮挡测试效果对比

(a) YOLOv5 测试结果

(b) 本文模型测试结果

图 20　同类别不同品种测试效果对比

表 4　模型性能对比

模型	mAP (%)	FPS (帧/s)
Faster-RCNN	95.49	5
YOLOv4	95.39	43
YOLOv5 (GIOU)	96.15	82
YOLOv5 (CIOU)	97.72	82
YOLOv5 (CIOU)+SELayer (ours)	99.10	82

参考文献

1 彭红星, 黄博, 邵园园, 等. 自然环境下多类水果采摘目标识别的通用改进 SSD 模型. 农业工程学报, 2018, 34(16): 155–162. [doi: 10.11975/j.issn.1002-6819.2018.16.020]

2 王辉, 张帆, 刘晓凤, 等. 基于 DarkNet-53 和 YOLOv3 的水果图像识别. 东北师大学报 (自然科学版), 2020, 52(4): 60–65.

3 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

4 Bargoti S, Underwood J. Deep fruit detection in orchards. 2017 IEEE International Conference on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 3626–3633.

5 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

6 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv: 2004.10934, 2020.

7 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 2117–2125.

8 Liu S, Qi L, Qin HF, *et al*. Path aggregation network for instance segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768.

9 Yun SD, Han DY, Chun S, *et al*. Cutmix: Regularization strategy to train strong classifiers with localizable features. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 6023–6032.

10 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

11 He KM, Zhang XY, Ren SQ, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]

12 Rezatofighi H, Tsoi N, Gwak JY, *et al*. Generalized intersection over union: A metric and a loss for bounding box regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019. 658–666.

13 Zheng ZH, Wang P, Liu W, *et al*. Distance-IoU loss: Faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993–13000. [doi: 10.1609/aaai.v34i07.6999]

(校对责编: 牛欣悦)