doi:10.6041/j.issn.1000-1298.2023.10.021

基于改进 YOLO v4 和 ICNet 的番茄串检测模型

刘建航[1]　何鉴恒[1]　陈海华[2]　王晓政[1]　翟海滨[3]

(1. 中国石油大学(华东)海洋与空间信息学院, 青岛 266555; 2. 中国科学院计算技术研究所, 北京 100094;
3. 国家计算机网络应急技术处理协调中心, 北京 100029)

摘要: 针对深层神经网络模型部署到番茄串采摘机器人,存在运行速度慢,对目标识别率低,定位不准确等问题,本文提出并验证了一种高效的番茄串检测模型。模型由目标检测与语义分割两部分组成。目标检测负责提取番茄串所在的矩形区域,利用语义分割算法在感兴趣区域内获取番茄茎位置。在番茄检测模块,设计了一种基于深度卷积结构的主干网络,在实现模型参数稀疏性的同时提高目标的识别精度,采用 K-means++聚类算法获得先验框,并改进了 DIoU 距离计算公式,进而获得更为紧凑的轻量级检测模型(DC-YOLO v4)。在番茄茎语义分割模块(ICNet)中以 MobileNetv2 为主干网络,减少参数计算量,提高模型运算速度。将采摘模型部署在番茄串采摘机器人上进行验证。采用自制番茄数据集进行测试,结果表明,DC-YOLO v4 对番茄及番茄串的平均检测精度为 99.31%,比 YOLO v4 提高 2.04 个百分点。语义分割模块的 mIoU 为 81.63%,mPA 为 91.87%,比传统 ICNet 的 mIoU 提高 2.19 个百分点,mPA 提高 1.47 个百分点。对番茄串的准确采摘率为 84.8%,完成一次采摘作业耗时约 6 s。

关键词: 番茄串; 采摘机器人; 深度学习; YOLO v4; ICNet; 采摘模型

中图分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1000-1298(2023)10-0216-09　　**OSID**:

Development of Detection Model for Tomato Clusters Based on Improved YOLO v4 and ICNet

LIU Jianhang[1]　HE Jianheng[1]　CHEN Haihua[2]　WANG Xiaozheng[1]　ZHAI Haibin[3]

(1. *College of Oceanography and Space Information, China University of Petroleum (East China), Qingdao* 266555, *China*
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100094, *China*
3. *National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing* 100029, *China*)

Abstract: For the deep neural network model deployed to embedded devices (such as tomato clusters picking robots), there are some problems, such as slow running speed, low recognition rate of picking targets, inaccurate positioning and so on, an efficient model for tomato clusters detection was proposed and verified. The model was composed of two modules: detection and semantic segmentation. Target detection was responsible for extracting the rectangular region where the tomato cluster was located, and then using the semantic segmentation algorithm to obtain the tomato stem position in the rectangular region. In the tomato detection module, a backbone network based on deep convolution structure was designed to improve the accuracy of crop recognition while realizing the sparsity of model parameters. K-means++ clustering algorithm was used to obtain a priori frame, and DIoU distance calculation formula was improved to obtain a more compact lightweight detection model (DC-YOLO v4). In the semantic segmentation module (ICNet), MobileNetv2 was used as the backbone network to reduce the amount of parameter calculation and improve the operation speed of the model. The model was deployed on the tomato clusters picking robot for verification. The self-made tomato data set was used for testing. The results showed that the average detection accuracy was 99.31% on tomato test set, outperforming YOLO v4 by 2.04 percentage points. The mIoU and mPA achieved 81.63% and 91.87% on tomato stem set, exceeding ICNet by 2.19 percentage points and 1.47 percentage points, respectively. The accurate picking rate of tomato clusters was 84.8%, it took 6s to complete a picking operation.

Key words: tomato clusters; picking robot; deep learning; YOLO v4; ICNet; picking model

收稿日期: 2022-09-01　修回日期: 2022-10-16

基金项目: 山东省自然科学基金项目(ZR2020MF005)

作者简介: 刘建航(1978—),男,副教授,博士,主要从事人工智能研究,E-mail: liujianhang@upc.edu.cn

通信作者: 翟海滨(1983—),男,高级工程师,博士,主要从事智能信息处理研究,E-mail: zhaihaibin@163.com

0　引言

我国农作物采摘主要以手工为主，采摘工作季节性强、劳动强度大、成本高[1]。随着人工智能和农业机械相互结合，使农作物采摘智能化成为可能，采摘机器人在提高采摘率、推动现代化农业发展等方面具有重要意义[2]。

番茄作为中国种植的主要经济作物，其采收方式分为粒收与串收，串收有着较高的采摘效率，并且串收番茄更容易保存和运输。快速且精确地采摘番茄串是目前串型番茄采摘机器人的重点研究内容。番茄串检测以及番茄茎定位主要通过计算机视觉实现。因此，视觉感知系统的性能直接影响番茄串的采摘率[3]。文献[4]采用颜色分量运算和彩色空间转换实现图像阈值分割和目标特征提取，同时对末端执行器进行了设计，实现了串型番茄的采摘，但采摘时间较长，成熟番茄串果实识别成功率为 90%。文献[5]借鉴 AdaBoost 学习算法在人脸识别中的成功应用[6-7]，提出了基于 Haar-like 特征及其编码和 AdaBoost 学习算法的番茄识别方法。实验结果表明，单幅图像的处理时间为 15 s，正确识别率为 93%。文献[8]提出使用 Mask R－CNN 模型对果园中重叠绿色苹果进行识别和分割，将残差网络与密集连接卷积网络相结合作为骨干网络提取特征，对 120 幅苹果图像进行检测，结果表明，平均检测准确率为 97.31%，但由于数据集太少，仍需增加样本集和丰富样本多样性以更具说服力。文献[9]使用双目视觉技术对番茄进行识别，根据番茄颜色特征用拟合曲线对番茄分割，并通过双目视觉测量原理计算出番茄的三维坐标，测量误差低于 4%，但仍有待进一步优化提升检测精度。文献[10]提出一种番茄果实串采摘点识别方法，该方法对垂直向下的番茄果实串采摘点识别效果较好，但不能对其他姿态的番茄果实进行识别。文献[11]提出了一种基于改进型 YOLO 的复杂环境下番茄果实快速识别方法，能够提取多特征信息，模型对番茄检测精度为 97.13%。

综上，国内外研究人员针对番茄串的识别和定位问题提出的研究方法尚未达到理想的精度和工业级实时性的要求，对多样的特征变化鲁棒性不足。因此，难以满足实际需求。为进一步提高农作物的识别率和采摘率，本文以番茄为研究对象，提出一种视觉感知模型，模型包括检测和语义分割 2 个模块，即番茄串检测和番茄茎分割。采用一种基于深度卷积结构的主干网络，取代残差块结构中的普通卷积运算，降低主干网络的计算量，从而获得更为紧凑的主干特征提取网络，通过 K-means + + 聚类算法获得先验框，并改进 DIoU 距离计算公式，获得更为紧凑的轻量级检测模型（DC－YOLO v4），在实现模型参数稀疏性的同时提高识别精度。将 MobileNetv2 作为 ICNet 分割模型的主干网络，以有效减少计算量，达到实时分割效果。

1　材料与方法

1.1　图像采集

番茄数据集、番茄茎数据集采集于黄河三角洲农业高新技术示范园区的设施农业测试验证平台（山东省广饶县）。通过 Intel RealSense D435 型深度相机采集番茄样本，图像分辨率为 4 032 像素 × 3 024 像素，如图 1a 所示，相机安装在末端执行器上方 5 cm 处，通过图 1b 所示的方式，在移动端遥控机器人进行番茄样本采样，模拟番茄采摘机器人的实际工作场景。

(a) 深度相机安装位置　　(b) 采摘机器人

图 1　采摘机器人采集番茄样本

Fig. 1　Picking robot collects tomato samples

采摘机器人的主要结构如图 2 所示，包括机械臂、可移动装置、机器人控制系统、深度相机和末端执行器 5 部分。默认状态下，末端执行器的安装位置距地面 10 cm。

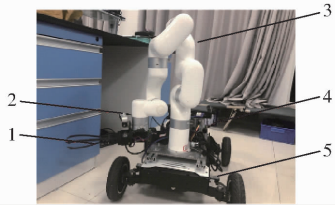


图 2　采摘机器人主要结构

Fig. 2　Main structure of picking robot

1. 末端执行器　2. 深度相机　3. 机械臂　4. 机器人控制系统　5. 可移动装置

番茄植株种植在桁架上，行距约 0.4 m，高约 2 m，为保证数据集样本的多样性，分别采集不同光照强度、不同果实数量、不同拍摄角度的番茄串样本共 2 000 幅，番茄茎样本图像 1 000 幅。采集的部分番茄样本如图 3 所示。

1.2　番茄检测网络

1.2.1　YOLO 目标检测网络

番茄采摘机器人的视觉感知模型包括目标检测和语义分割两部分[12]。针对番茄检测模型，本文借

图 3 温室环境下采集的番茄样本

Fig. 3 Tomato samples collected in greenhouse environment

鉴 YOLO 系列的模型结构[13],其突出特点是快速和精确。与 Two-Stage(如 Faster R-CNN)使用 Region proposal 区域建议特征提取方式不同,YOLO 的工作原理[14]如下:①对输入图像的全局区域进行训练。②利用主干特征提取网络完成番茄样本的特征初次提取。③融合加强特征提取网络,增大感受野的同时反复提取特征信息。④采用 Bounding box 预测方式,预测目标类别、置信度和预测框。

YOLO 系列网络模型中,YOLO v1 存在网络模型检测精度差、目标定位不准确等问题[15];YOLO v2 中加入了锚框和批量归一化,并通过更改网络模型结构等操作提升了训练模型性能,但不适用于检测目标重叠的情况[16];YOLO v3 中引入了多尺度融合训练、残差结构、改变网络模型结构等操作,使得训练模型性能得到了极大提升,但其主干网络深度达 53 层且采取了多尺度融合,导致检测速度慢[17];YOLO v4 本质上继承了 YOLO v3 的结构,主干网络更改为 CSPDarkNet53 优化特征提取性能,采用 Mish 激活函数使梯度下降过程更为平滑,相较于 ReLU、Sigmod 等激活函数,Mish 在处理负值时不会完全截断,保证了特征信息流动[18],同时加入了更多目前流行的技巧(如 Mosaic 数据增强、标签平滑、CIOU 等)。但实际上,在检测精度和速度方面并没有明显提升,未达到工业级番茄检测的要求。

1.2.2 改进的 YOLO v4 网络模型

在剖析 YOLO v4 网络结构的基础上,设计了一个基于深度卷积结构的主干网络,用于对番茄串图像的初步特征提取。深度卷积结构如图 4 所示。

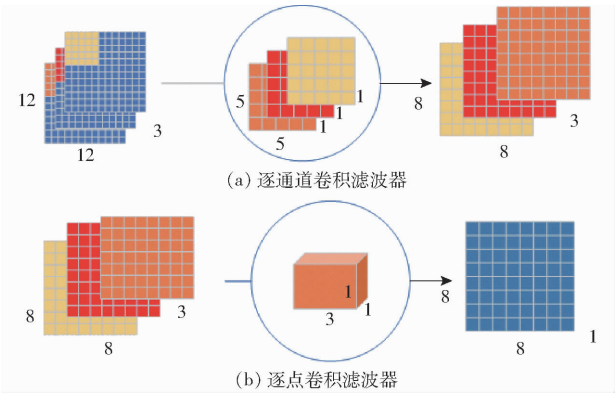


图 4 深度卷积结构

Fig. 4 Depth convolution structure

番茄检测模块由 DarkNetBN_Mish 模块、主干网络、空间金字塔池化(Spatial pyramid pooling,SPP)、像素聚合网络(Pixel aggregation network,PANet)和 YOLO Head 构成。如图 5 所示,将深度卷积结构替换主干网络中 Resblock_body 的普通卷积,降低主干网络的计算量。

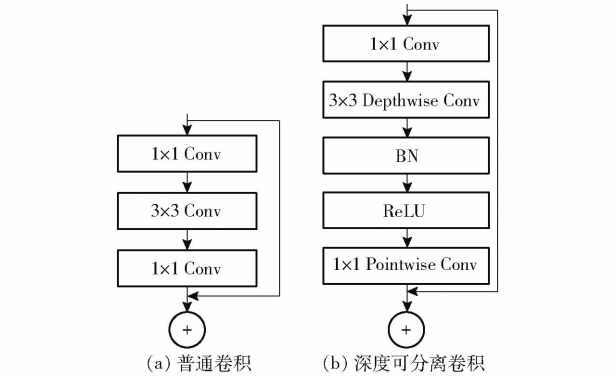


图 5 改进后的 Resblock_body

Fig. 5 Improved Resblock_body

基于深度卷积结构的主干网络提取输入图像的特征信息,并将特征信息通过卷积传递到 DarkNetConv2D_BN_Mish 模块中,对输入图像进行归一化和非线性操作,SPP 和 PANet 负责对特征信息加强提取。深度卷积结构处理 3 个通道的特征信息,最后,通过卷积核尺寸为 1×1×3 的卷积核将 3 个通道的属性进行融合,传递给加强特征提取网络。相较于普通卷积,深度卷积结构产生的网络参数少,有效解决了深度学习网络重复学习特征信息造成计算量大的问题,提高了运算速度。网络模型的参数如表 1 所示。可以看出 DC-YOLO v4 在参数量、处理速度、模型内存占用量等方面均优于一些主流模型的主干网络。

YOLO v4 使用 K-means 设计先验框尺寸,但是它存在预先人为确定 k 个初始聚类中心的缺点,导致生成的先验框不稳定,难以反映真实框尺寸情况。之后提出的K-means++针对这一问题,进行了一

表 1　不同网络模型的主干网络参数

Tab. 1　Backbone network parameters of different network models

模型	参数量	模型内存占用量/MB	处理速度/(f·s^{-1})
YOLO v4	60 040 001	244.29	24.4
MobileNetv3 − YOLO v4	39 989 933	149.01	27.8
YOLO v5 − m	21 375 645	81.54	28.6
CenterNet	32 665 432	124.61	22.4
YOLO v6	11 008 515	41.99	30.0
DC − YOLO v4	10 801 149	41.20	32.0

系列改进，不再预先人为确定初始聚类中心，具体实现流程如图 6 所示。

本文采用改进的交并比(GIoU)计算公式，通过引入检测框宽高的比例因子 v_s，避免 GIoU 在某些情况下退化成 IoU 的问题，改进的 GIoU 表达式为

$$GIoU = IoU - v_s \frac{C - (A \cup B)}{C} \tag{1}$$

其中

$$v_s = \frac{\max\left(\frac{w_{gt}}{h_{gt}}, \frac{w_{bb}}{h_{bb}}\right)}{\min\left(\frac{w_{gt}}{h_{gt}}, \frac{w_{bb}}{h_{bb}}\right)} \tag{2}$$

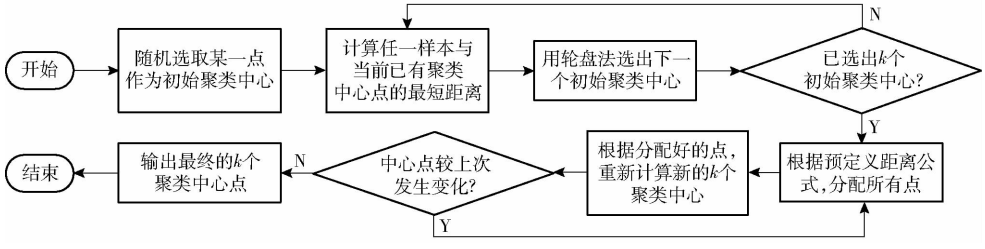


图 6　K-means + + 算法流程图

Fig. 6　K-means + + algorithm flow chart

式中　w_{gt}、h_{gt}——真实框的宽、高

w_{bb}、h_{bb}——预测框的宽、高

C——两框最小外接矩形的面积

$A \cup B$——两框并集的面积

并将式(1)作为 K-means + + 的距离计算公式，提高了网络预测精度。

在网络训练前对数据集进行了聚类处理，共得到 9 种尺寸的 Anchor box，如图 7 所示，其尺寸分别为(18,20)，(28,34)，(40,45)，(59,50)，(45,69)，(75,79)，(126,55)，(55,138)，(266,295)。相较于 K-means 聚类结果，采用 K-means + + 得到的锚框拟合程度更好，便于模型的训练。

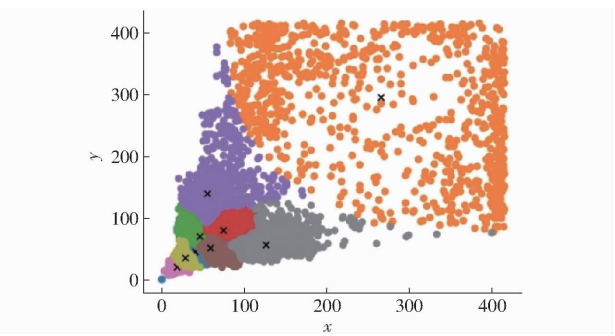


图 7　9 种尺寸的聚类中心分布图

Fig. 7　Distribution of cluster centers of nine sizes

1.3　番茄茎分割网络

1.3.1　ICNet 语义分割网络

番茄串检测问题大部分采用传统图像处理与机器学习相结合的方式，会受到图像本身噪声等多种因素的制约，为了解决番茄串检测中的局限性，本文将基于深度学习的语义分割算法应用于番茄串分割领域。ICNet 网络模型[19]是基于高分辨率图像的实时语义分割网络。它利用处理低分辨率图像的效率以及高分辨率图像的高质量。思路是使低分辨率图像先通过全语义感知网络来取得大概的语义预测图，然后提出级联特征融合单元和级联标签指导策略整合中等和高分辨率特征，这逐渐提炼了粗糙的语义预测图。ICNet 的网络架构如图 8 所示。它使用 PSPNet 的金字塔池化模块融合多尺度上下文信息，并将网络结构划分为 3 个分支，分别为低分辨率、中分辨率和高分辨率。配合 ResNet50 使用 3 个分支进行特征融合形式的训练，前 2 个分支增加辅助训练，增加模型收敛。对于每个输出特征，在训练时会以真实标签的 1/16、1/8、1/4 来指导各分支训练，使得梯度优化更加平滑，随着每个分支学习能力的增强，预测没有被某一分支主导。

分支 1 将原图下采样到 1/4 尺寸，然后经过连续 3 次下采样降维到原图的 1/32，使用空洞卷积层扩展感受野的同时不缩小尺寸，最终输出 1/32 原图的特征图。分支 1 的卷积层数多但特征图尺寸小，速度快，且第 2 个分支与第 1 个分支共享前 3 层卷积的权值。

分支 2 将 1/2 尺寸的原图作为输入，经过卷积后降维到 1/8 原图，得到 1/16 尺寸的特征图，再将第 1 个分支中由低分辨率图像提取出的特征图通过级联特征融合单元得到最终输出。

分支 3 以原图像为输入，经 3 次卷积后得到原图 1/8 尺寸的特征图，再将处理后的输出和分支 2

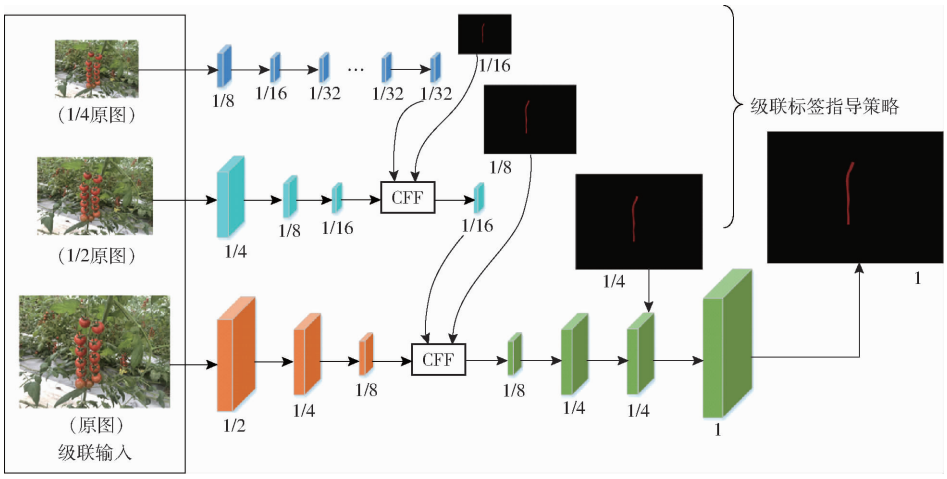


图 8 ICNet 网络结构

Fig. 8 ICNet network structure

的输出通过 CFF 融合。分支 3 的图像分辨率大,但卷积层数少,耗时较少。

ICNet 的损失函数是通过构建多分支 loss 实现,损失函数表达式为

$$L = -\sum_{t=1}^{\tau} \lambda_t \frac{1}{y_t x_t} \sum_{y=1}^{y_t} \sum_{x=1}^{x_t} \mathrm{lb} \frac{\mathrm{e}^{F^t_{n',y,x}}}{\sum_{n=1}^{N} \mathrm{e}^{F^t_{n,y,x}}} \tag{3}$$

式中 τ——分支数量,取 3

x_t、y_t——分支的特征图尺寸

$F^t_{n,y,x}$——位置(n,y,x)的值

n'——相关的真实标签

λ_t——每个分支的损失权重

$F^t_{n',y,x}$——真实标签(n,y,x)的值

通常,高分辨率分支权重 λ_3 设置为 1,中分辨率和低分辨率分支的权重 λ_2 和 λ_1 分别设置为 0.4 和 0.16。

1.3.2 改进的 ICNet 语义分割网络

在一些经典的深度学习语义分割算法中,主要采用 VGG 系列或者 ResNet 系列作为主干特征提取网络,虽然二者都能够提取图像的深层信息,但是对于部署到嵌入式设备上而言,其网络模型的参数量过大,分割速度慢。因此,采用 MobileNetv2 替换 ResNet,取消传统的卷积计算,采用深度卷积以及 1×1 的逐点卷积来提取图像特征,可以成倍减少卷积层的时间复杂度和空间复杂度。同时还引入了倒残差结构,先升维后降维,增强梯度的传播,显著减少推理期所需的内存占用量。倒残差结构如图 9 所示。

在残差结构中,首先通过 1×1 卷积实现降维,再通过 3×3 卷积提取通道特征,最后使用 1×1 卷积实现升维。但在倒残差结构中,先通过 1×1 卷积实现升维,再通过 3×3 的逐通道卷积提取

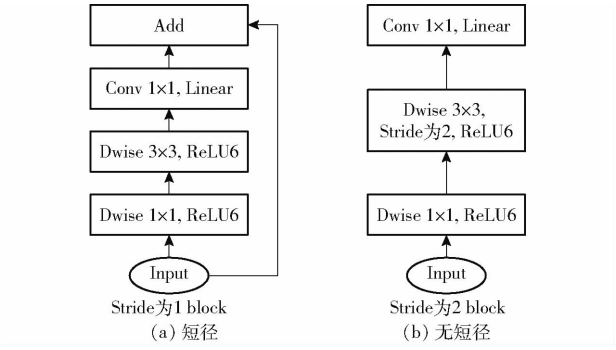


图 9 倒残差结构

Fig. 9 Inverted residuals

特征,最后使用 1×1 卷积实现降维。调换了降维和升维的顺序,并将 3×3 的标准卷积换为逐通道卷积,呈两头小、中间大的菱形结构。其次,改变了之前所采用的激活函数。残差结构中通常采用 ReLU 激活函数,但是,在倒残差结构中,采用 ReLU6 作为激活函数,最后 1 个卷积使用的是线性激活函数。用 ReLU6 替换 ReLU,目的是为了保证在嵌入式设备低精度也能保有很好的数值分辨率。如果对 ReLU 的输出值不加限制,那么输出范围就是零到正无穷,无法精确描述其数值,这将带来精度损失。ReLU6 激活函数如图 10 所示。最后 1 个卷积使用线性激活,则是线性瓶颈结构的内容。瓶颈结构是指将高维空间映射到低维空间,缩减通道数;膨胀层则相反,其将低维空间映射到高维空间,增加通道数。沙漏型结构和梭型结构,都可看做是 1 个膨胀层和 1 个瓶颈结构的组合。瓶颈结构和膨胀层本质上体现的都是 1×1 卷积。线性瓶颈结构是末层卷积使用线性激活的瓶颈结构。ReLU 容易导致逐通道卷积部分的卷积核失活,即卷积核内数值大部分为零,这是因为在变换过程中,需要将低维信息映射到高维空间,再经 ReLU 重新映射回低维空间。若输出的维度

相对较高，则变换过程中信息损失较小；若输出的维度相对较低，则变换过程中信息损失很大。因此，末层采用线性激活来避免这一问题。

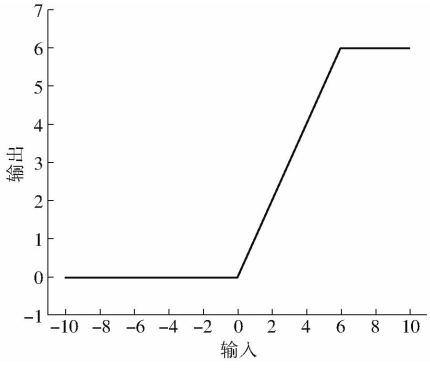


图 10　ReLU6 激活函数

Fig. 10　ReLU6 activation function

2　网络模型训练与评价指标

2.1　验证平台

主机操作系统为 Ubuntu 16.04，中央处理器为 Intel Core i9 - 10920X GPU@ 3.50 GHz，运行内存 32 GB，显卡为 Nvidia Quadro P2200（5 GB/戴尔）。神经网络在 Anaconda 3 虚拟环境下训练，采用 Pytorch 1.2.0 深度学习框架，配置安装 Python 3.8 编程环境、GPU 并行计算架构 Cuda 10.0 和神经网络 GPU 加速库 Cudnn 10.0。

2.2　番茄检测网络模型训练

采用 PASCAL VOC 2007 数据集的预训练权重训练，训练图像分辨率为 416 像素 ×416 像素，每个批次处理 8 幅图像，总迭代次数为 1 000，前 450 次采用冻结训练加快训练速度，训练学习率为 0.001，每迭代 100 次，学习率降低 0.1，后 550 次的解冻训练学习率为 0.000 1。可以看出前 200 次迭代中网络快速拟合，200 次迭代之后损失函数基本稳定，番茄检测网络开始收敛。图 11 反映了损失函数的变化趋势。

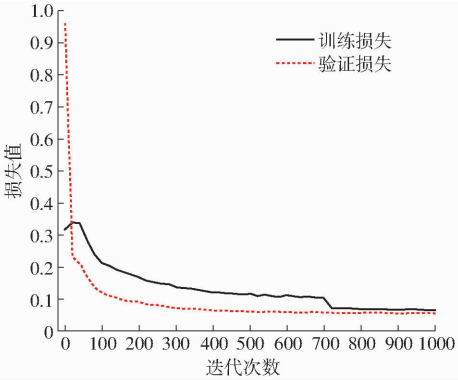


图 11　目标检测模型训练曲线

Fig. 11　Target detection model training curves

2.3　番茄串语义分割网络模型训练

采用 PASCAL VOC 2007 数据集的预训练权重训练，输入图像分辨率为 512 像素 ×512 像素，格式为 JPG，对应的标签图像格式为 PNG，类别数为 2，下采样倍数为 16，每个批次处理 8 幅图像，总迭代次数为 500，前 100 次为冻结训练，学习率为 0.000 5，后 400 次的解冻训练学习率为 0.000 005。由图 12 可以分析出，在前 100 次迭代中网络快速拟合，100 次迭代后损失函数基本稳定，番茄串语义分割检测网络开始收敛。

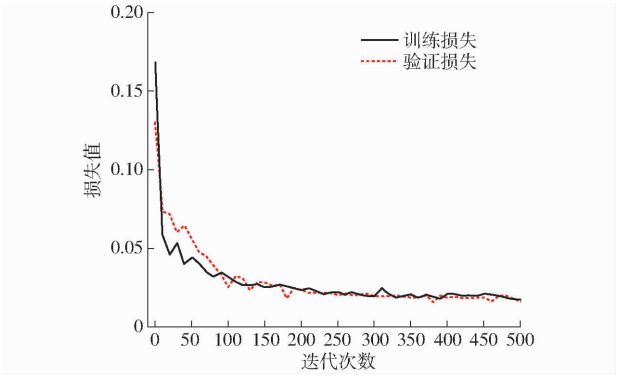


图 12　语义分割模型训练曲线

Fig. 12　Semantic segmentation model training curves

2.4　评价标准

为了客观分析 DC - YOLO v4 对番茄数据集以及 ICNet 模型对番茄串数据集的语义分割性能，本文引入平均交并比（mIoU）、准确率（Precision）、召回率（Recall）、平均精度均值（mAP）、综合评价指标（F1 值）、类别平均像素准确率（mPA）和检测时间（Time）等评价指标。本文的目的是快速准确识别番茄并分割番茄茎，因此把平均交并比、平均精度均值和检测时间作为主要评价指标。利用 IoU 阈值为 0.5 的平均精度来测定番茄识别模型的准确性。此度量标准用于测量目标检测器的精度，因为它平衡了精度和召回率。

3　结果分析

3.1　番茄检测效果

本文设计的检测模块借鉴了 YOLO 系列的架构，融合了深度卷积结构，因此有必要与传统的 YOLO 系列算法的番茄识别性能进行对比分析。同时，使用批量为 8、尺寸为 416 像素 ×416 像素的图像，对经过训练的 MobileNet - YOLO v4、YOLOX、YOLO v5 - m、YOLO v6 进行测试和比较，在测试模型中获得的结果存在差异，测试结果如图 13 所示。DC - YOLO v4 模型对番茄和番茄串的识别正确率高于 YOLO v4 模型，YOLO v4 模型深度图中存在大

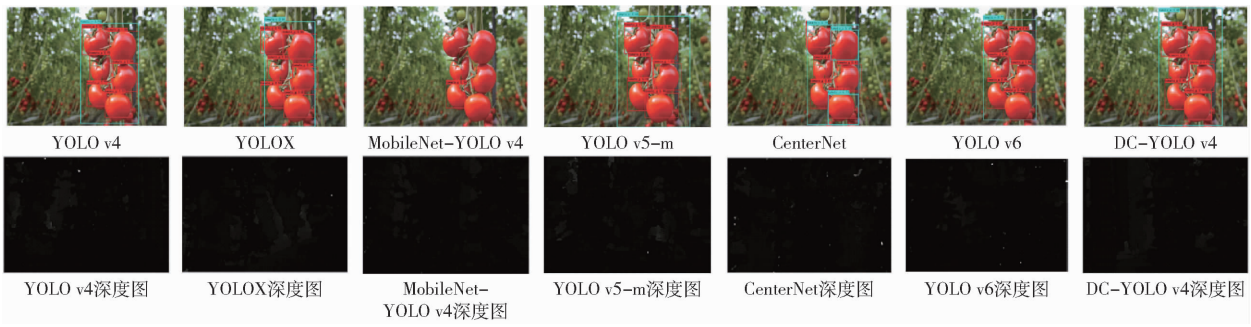


图 13 不同检测模型测试效果

Fig. 13 Test results of different test models

量噪点，导致其识别精度不足，误检率高。MobileNet－YOLO v4 检测模型在实际应用中，对番茄串的识别不敏感，且在深度图中，DC－YOLO v4 模型的番茄串轮廓更为光滑。YOLOX 模型[20]是由旷视科技在 2021 年提出的全新检测模型，DC－YOLO v4 模型与 YOLOX 模型在实际测试中，并无明显区别。YOLO v5－m 模型的检测速度快，但丧失了一定的识别准确度，虽然能够获取图像的高级特征，但这些特征具有平移不变性[21]，不利于对目标信息的区域采样。

为论证本研究提出的 DC－YOLO v4 模型的有效性，又与 YOLO v5 模型系列中最为主流的 YOLO v5－m、YOLO v6 以及 CenterNet 检测模型比较。YOLO v6 模型是美团视觉智能部研发的一款目标检测框架，致力于工业应用。CenterNet 模型[22]是无锚框目标检测器，由于没有复杂的 Anchor 操作，检测速度优于 Two-Stage 及预锚框系列，算法性能良好，对小目标检测具有优势[23]。CenterNet 模型只通过 FCN（全卷积）的方法实现了对于目标的检测与分类，即使没有 Anchor 与 NMS 等操作，它在高效的同时精度也较好。可以将其结构进行简单修改就可以应用到农业场景下的番茄目标检测之中。

表 2 展示了不同网络模型对番茄串和番茄的检测性能，DC－YOLO v4 模型的 mAP 最大，对于番茄和番茄串的识别准确率及召回率最高，单幅图像预测时 DC－YOLO v4 模型比 YOLO v4 模型的 mAP 高 2.04 个百分点。比 MobileNet v3－YOLO v4 模型的 mAP 高 1.08 个百分点。原因是卷积层较多，计算量大，检测速度偏慢，神经网络层数过深，因此检测精度较低。与 DC－YOLO v4 模型相比，CenterNet 模型难以对纹理特征进行有效提取，mAP 低于 DC－YOLO v4 模型 2.34 个百分点，并且检测时间差，不满足工业条件下的实时性要求。

表 2 不同识别模型性能比较

Tab. 2 Performance comparison of different recognition models

模型	准确率/%（tomato_c/tomato_g）	召回率/%（tomato_c/tomato_g）	F1 值/%（tomato_c/tomato_g）	mAP/%	时间/ms
YOLO v4	92.00/95.90	88.46/91.67	90/94	97.27	7.62
MobileNetv3－YOLO v4	94.30/94.30	93.94/93.94	94/96	98.23	6.74
YOLOX	88.46/91.51	88.46/95.10	88/93	95.69	5.58
YOLO v5－m	92.59/97.18	96.15/97.64	94/97	98.03	7.06
YOLO v6	98.30/97.41	98.99/96.23	98/97	98.97	6.93
CenterNet	96.15/93.48	96.15/91.98	96/95	96.97	19.00
DC－YOLO v4	98.86/97.56	98.48/97.35	99/97	99.31	6.32

另外，DC－YOLO v4 模型的召回率稍低于 YOLO v5－m 模型与 YOLO v6 模型，原因是 YOLO v5－m 模型的 Backbone 是基于 CSPNet 搭建的，而 YOLO v6 模型的 Backbone 则是引入了 RepVGG 结构[24]，二者的主干检测网络较为复杂，对于单番茄果实的特征提取能力强。相对于 YOLOX 模型，虽然 DC－YOLO v4 模型的检测时间增加 0.74 ms，实时性略低于 YOLOX 模型，但是检测精度提高 3.62 个百分点。可以看出，DC－YOLO v4 模型能同时兼顾实时性和准确性，满足工业条件下采摘机器人的需求。

3.2 番茄茎分割效果

为了更好地展现改进的 ICNet 模型性能提升的直观效果，本研究还选取目前有代表性的主流语义分割网络 DeepLab_v3＋[25]、U－Net[26]和 PSPNet[27]进行实际测试实验。对比实验结果如图 14 所示，相较于 ICNet，改进后的 ICNet 能够完整分割出番茄茎，较好地保存逐像素点含有的位

置信息和语义信息，U－Net 只能捕捉大致外形，且包含大量噪点，PSPNet 缺少分割细节，不能很好地表征目标特征，DeepLab_v3 + 在实际测试中，效果与改进后的 ICNet 无明显差异。根据本研究提出的量化指标，结合表 3 可以得出，改进的 ICNet 网络与其他网络相比分割性能有了一定的提高，本文提出的改进 ICNet 网络 mIoU 和 mPA 分别为 81.63% 和 91.87%，相较于 ICNet 模型，mIoU 和 mPA 分别提升 2.19 个百分点和 1.47 个百分点；相较于 DeepLab_v3 + 模型，mIoU 和 mPA 分别提升 7.04 个百分点和 3.51 个百分点；相较于 U－Net 模型，mIoU 和 mPA 分别提升 7.74 个百分点和 4.88 个百分点；相较于 PSPNet 模型，mIoU 和 mPA 分别提升 9.71 个百分点和 4.66 个百分点。结果表明，改进 ICNet 网络相较于其他网络在番茄茎分割上更有优势。

图 14　不同分割模型测试效果

Fig. 14　Test results of different segmentation models

表 3　不同分割模型性能比较

Tab. 3　Performance comparison of different segmentation models

模型	mIoU/%	mPA/%	时间/ms
ICNet	79.44	90.40	8.16
DeepLab_v3 +	74.59	88.36	9.74
U－Net	73.89	86.99	12.58
PSPNet	71.92	87.21	10.71
改进的 ICNet	81.63	91.87	8.21

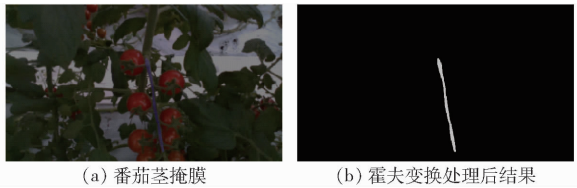
3.3　温室中视觉感知模型验证

为了验证本文提出的农作物采摘视觉感知模型在实际应用场景下的性能，将模型部署到山东中科智能农业机械装备创新技术中心自主研发的番茄采摘机器人系统中进行采摘实验。如图 15 所示，采摘机器人核心组成部件包括众为创造 xARM 型六轴机械臂、Intel RealSense D435 型深度相机、可移动吊轨以及工控机。

图 15　吊轨采摘机器人

Fig. 15　Rail picking robot

在实际应用中，当完成检测任务后，控制系统会给机械臂发送一个前移指令，末端执行器带动 RealSense D435 型深度相机向番茄茎大概方位移动，拉近感受视野，使 ICNet 能够实时分割出视频流数据中的番茄茎。如图 16 所示，彩色图中的蓝色掩膜为 ICNet 模型在视频流中的分割效果，为了满足工程级的实时性要求，采用霍夫变换将其转换为二值图进行处理，加快系统的处理速度。



(a) 番茄茎掩膜　　(b) 霍夫变换处理后结果

图 16　实际场景中的分割效果

Fig. 16　Segmentation effect in real scene

本文共进行了 80 次采摘实验，由于吊轨采摘机器人每次仅能采摘一串红色番茄，故只统计了红色番茄串的采摘成功率，最终的平均采摘成功率为 84.8%。RealSense D435 型深度相机检测到番茄串后，会计算并返回感兴趣区域的的中心点，控制系统驱动末端执行器移动到中心点前的 10 cm 处，经过 ICNet 模型处理后，得到分割番茄茎，末端执行器会根据计算得到的采摘点进行采摘，最后将采摘的番茄串放入收纳篮中，完成上述采摘流程平均用时 6 s。影响工作时间的主要原因是番茄茎与背景颜色相近，对于番茄茎的形状特征提取能力差，同时枝叶的遮挡也增加了番茄茎提取的难度。

4　结论

(1) 番茄和番茄串测试集上的实验结果表明，

检测模块对番茄的识别准确率为 98.86%,召回率为 98.48%,F1 值为 99%,对番茄串的识别准确率为 97.56%,召回率为 97.35%,F1 值为 97%,模型平均精度为 99.31%,模型平均识别单幅图像需要 6.32 ms。相比于本研究中选用的一些目标检测对比模型,在性能上有明显的提升,DC - YOLO v4 模型的 mAP 相比于 YOLO v4、MobileNet v3 - YOLO v4、YOLOX、YOLO v5 - m、CenterNet、YOLO v6 模型提高 2.04、1.08、3.62、1.28、2.34、0.34 个百分点。

(2)番茄茎测试集上的实验结果表明,改进的 ICNet 分割模型对番茄茎的平均召回率为 91.87%,mIoU 为 81.63%,mPA 为 91.87%,模型平均分割单幅图像需要 8.21 ms,改进的 ICNet 模型的 mPA 相比于 ICNet、DeepLab_v3 +、U - Net 和 PSPNet 分别提升 1.47、3.51、4.88、4.66 个百分点。

(3)将检测模型部署到番茄采摘机器人上,在温室环境下对番茄串进行采摘论证,与人工检验进行对比,结果表明,机器人的准确采摘率为 84.8%,平均完成一次采摘动作用时 6 s。本文的研究结果可以为复杂温室环境下的其他农作物采摘提供技术支撑。

参 考 文 献

[1] 张天柱,吴卫华. 应着手规划我国的番茄产业[J]. 农产品加工·学刊,2009(6):108 - 110,114.
ZHANG Tianzhu, WU Weihua. Start with planning Chinese tomato industry[J]. Academic Periodical of Farm Products Processing, 2009(6):108 - 110,114. (in Chinese)

[2] 李寒,张漫,高宇,等. 温室绿熟番茄机器视觉检测方法[J]. 农业工程学报,2007,23(增刊):328 - 334,388.
LI Han, ZHANG Man, GAO Yu, et al. Green ripe tomato detection method based on machine vision in greenhouse[J]. Transactions of the CSAE, 2007,23(Supp.):328 - 334,388. (in Chinese)

[3] 卢军,王贤锋,后德家. 水果采摘机器人视觉系统研究进展[J]. 湖北农业科学,2012,51(21): 4705 - 4708.
LU Jun, WANG Xianfeng, HOU Dejia. Development of machine vision system for fruit harvesting robots[J]. Hubei Agricultural Sciences, 2012,51(21): 4705 - 4708. (in Chinese)

[4] JI C, ZHANG J, YUAN T, et al. Research on key technology of truss tomato harvesting robot in greenhouse[J]. Applied Mechanics & Materials, 2014, 442:480 - 486.

[5] 赵源深,贡亮,周斌,等. 番茄采摘机器人非颜色编码化目标识别算法研究[J]. 农业机械学报,2016,47(7):1 - 7.
ZHAO Yuanshen, GONG Liang, ZHOU Bin, et al. Object recognition algorithm of tomato harvesting robot using noncolor coding approach[J]. Transactions of the Chinese Society for Agricultural Machinery, 2016,47(7):1 - 7. (in Chinese)

[6] VIOLA P A, JONES M J. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2001.

[7] PAPAGEORGIOU C P, OREN M, POGGIO T. A general framework for object detection[C]// IEEE Computer Society, 1998.

[8] JIA W K, TIAN Y Y, LUO R, et al. Detection and segmentation of overlapped fruits based on optimized Mask R - CNN application in apple harvesting robot[J]. Computers and Electronics in Agriculture, 2020, 172: 1 - 7.

[9] 张瑞合,姬长英,沈明霞,等. 计算机视觉技术在番茄收获中的应用[J]. 农业机械学报,2001,32(5):50 - 52,58.
ZHANG Ruihe, JI Changying, SHEN Mingxia, et al. Application of computer vision to tomato harvesting[J]. Transactions of the Chinese Society for Agricultural Machinery, 2001,32(5):50 - 52,58. (in Chinese)

[10] 梁喜凤,章艳. 串番茄采摘点的识别方法[J]. 中国农机化学报,2016,37(11):131 - 134,149.
LIANG Xifeng, ZHANG Yan. Recognition method of picking point for tomato cluster[J]. Journal of Chinese Agricultural Mechanization, 2016,37(11):131 - 134, 149. (in Chinese)

[11] 刘芳,刘玉坤,林森,等. 基于改进型 YOLO 的复杂环境下番茄果实快速识别方法[J]. 农业机械学报,2020,51(6):229 - 237.
LIU Fang, LIU Yukun, LIN Sen, er al. Fast recognition method for tomatoes under complex environments based on improved YOLO[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(6):229 - 237. (in Chinese)

[12] HAN J, ZHANG D, CHENG G,et al. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning[J]. IEEE Trans. Geosci. Remote Sens.,2015,53(6): 3325 - 3337.

[13] 朱逢乐,郑增威. 基于图像和卷积神经网络的蝴蝶兰种苗生长势评估[J]. 农业工程学报,2020,36(9):185 - 194.
ZHU Fengle, ZHENG Zengwei. Image-based assessment of growth vigor for *Phalaenopsis aphrodite* seedlings using convolutional neural network[J]. Transactions of the CSAE, 2020, 36(9):185 - 194. (in Chinese)

[14] BOCHKOVSKIY A, WANG C Y, LIAOH Y M. YOLO v4: optimal speed and accuracy of object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE,2020.

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

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

[17] REDMON J, FARHADI A. YOLO v3: an incremental improvement[J]. arXiv Preprint, arXiv: 1804. 02767v1, 2018.

[18] MISRA D. Mish: a self regularized non-monotonic neural activation function[C]// arXiv Preprint, arXiv:1908.08681v3, 2020.

(下转第 254 页)