基于位置可学习视觉中心机制的零售商品检测方法

吕晓华，魏铭辰，刘立波
（宁夏大学信息工程学院，宁夏 银川 750021）

摘 要：针对零售商品包装变形和重叠使得难以有效捕捉显著且多样化的特征信息，导致检测精度不高的问题，设计了位置可学习视觉中心（LLVC, location learnable visual center）机制，对 YOLOX-s 进行改进，取得了更高的检测精度。为有效应对商品包装变形和重叠现象，首先，通过轻量级多层感知机融合不同特征通道上的信息，以充分捕获全局上下文信息；接着，通过设计的 LLVC 增强局部特征表示能力，并利用空间信息为局部特征分配可学习的权重，提高辨别性局部特征的关注程度；最后，将交并比（IoU, intersection over union）损失函数替换为中心交并比（CIoU, centered intersection over union），并在此基础上引入功率参数 α，有效降低了漏检率。实验结果表明，所提方法在零售商品识别（RPC, retail product checkout）数据集上取得 91.3%的准确率，相比 YOLOX-s 提高了 2.2%，并优于目前主流的轻量级目标检测算法；同时每秒帧率（FPS, frame per second）为 97 frame/s，模型大小为 9.48 MB，能够在计算资源受限的场景下，准确且实时地进行零售商品检测。

关键词：零售商品检测；YOLOX-s；中心学习机制；损失函数；轻量级
中图分类号：TP18
文献标志码：A
doi: 10.11959/j.issn.2096−3750.2023.00366

Retail commodity detection method based on location learnable visual center mechanism

LYU Xiaohua, WEI Mingchen, LIU Libo
School of Information Engineering, Ningxia University, Yinchuan 750021, China

Abstract: To address the problem of low detection accuracy caused by the difficulty in effectively capturing significant and diversified feature information for packaging deformation and overlap products, a location learnable visual center (LLVC) mechanism was designed to improve YOLOX-s, achieving higher detection accuracy. To effectively deal with product packaging deformation and overlap phenomena, firstly, global context information was captured through a light-weight multi-layer perceptron to help the model better understand spatial information in product features. Secondly, the local feature representation ability was enhanced by the designed LLVC and the spatial information was used to allocate learnable weights for local features to increase the attention of discriminative local features. Finally, the intersection over union (IoU) loss function was replaced with centered intersection over union (CIoU) and power parameters were introduced on this basis to effectively reduce the missed detection rate. Experimental results show that the proposed method achieves an accuracy of 91.3% on the retail product checkout (RPC) dataset, which is 2.2% higher than YOLOX-s and better than current mainstream lightweight object detection algorithms. At the same time, frame per second (FPS) is 97 frame/s, and the model size is 9.48 MB. It can accurately and in real-time detect retail products in scenarios where computing resources are limited.

Key words: retail commodity detection, YOLOX-s, central learning mechanism, loss function, lightweight

收稿日期：2023−05−29 ；**修回日期：**2023−07−28
通信作者：刘立波，liulib@163.com
基金项目：国家自然科学基金资助项目（No.62262053）；宁夏科技创新领军人才计划项目（No.2022GKLRLX03）
Foundation Items: The National Natural Science Foundation of China (No.62262053), The Ningxia Science and Technology Innovation Leading Talent Plan (No.2022GKLRLX03)

0　引言

近年来，随着计算机视觉相关技术的快速发展，自助结算呈现愈加智能化的趋势。利用结算区顶部相机采集的图像，实现零售商品的精准和及时检测，是对自助结算系统下实时商品检测的有力支撑，也是对传统人工收银方式的有效补充，在提高结算效率和降低人工成本方面具有重要研究价值和现实意义。

目前，基于计算机视觉的零售商品检测方法可分为两类：基于人工特征的零售商品检测方法和基于深度学习的零售商品检测方法。基于人工特征的零售商品检测方法，通过人工设计的特征提取算法，将图像中的视觉属性（如颜色、纹理、形状和边缘等）转化为特征向量，送入支持向量机、决策树等分类器中进行零售商品识别[1-6]。该类方法尽管为零售商品检测提供了有效的解决方案，但过于依赖以往的经验，泛化性能有限且提取过程复杂，导致检测性能难以进一步提升。

基于深度学习的零售商品检测方法避免了复杂的人工特征提取工程，并且能够学习到更高级、更抽象的特征，逐渐成为该领域的主流。然而，基于深度学习的商品检测研究也面临着诸多极具挑战性的任务，其中商品包装变形和重叠尤其为难点[7]。Hurtik 等[8]基于 YOLOv3 通过四边形检测功能实现特征共享，同时利用物体轮廓信息来丰富特征表示，以此提高商品检测精度，然而对商品包装变形、商品之间重叠和复杂背景等情况不够敏感。Goldman 等[9]通过 Soft-IoU[10]层以及基于最大期望（EM, expectation-maximization）的高斯核聚类方法筛选重叠的探测框，提高重叠商品的检测精度，但其检测速度较慢，且未考虑商品包装变形带来的影响。Selvam 等[11]以 YOLOv5 为基础，在头部检测器生成 3 种不同尺度的特征图，以更好地检测小型、中型和大型商品。此外，Wang 等[12]基于 YOLOv4 设计了一种自助结算系统，利用区域增长算法对商品颜色进行统一划分，以增强特征表达能力。

这些方法虽然取得了较好的效果，但仍有诸多问题亟待研究。池化、步长卷积等操作使得感受野范围固定且缩小，无法同时有效获取包装变形和重叠商品的全局上下文信息；商品包装变形和重叠造成其特征图发生相应变化，导致网络学习到的细节特征无法呈现辨别性，进而影响模型分类性能，并且易将重叠物体的预测框剔除导致漏检。此外，出于成本考虑，零售商品检测模型对参数量大小容忍度较低，选用模型须保持轻量化和一定的实时性。YOLOX-s[13]作为一种轻量化和实时的目标检测算法，实现了检测精度和速度的良好平衡，引入的先进优化策略能够更好地处理目标之间的重叠情况，在实际应用中具有较大优势。因此，本文以 YOLOX-s 算法为基础进行改进。

虽然 YOLOX-s 具有优良的检测性能，但其网络结构中的 CSPDarknet[14]采用了多个下采样模块，如池化、步长卷积等，使得特征图的感受野范围缩小且固定，导致对包装变形和重叠商品的特征提取过程中难以充分获取全局上下文信息，限制了对包装变形和重叠商品的理解。因此，本文将轻量级多层感知机（Lightweight MLP, lightweight multilayer perceptron）[15]嵌入主干特征提取网络，首先通过深度卷积融合不同通道的特征图，随后利用通道多层感知机（Channel MLP, channel multilayer perceptron）拼接每个特征通道上的信息，以充分捕获全局上下文信息，提高模型对包装变形和重叠商品整体外观的理解能力。

为有效应对商品包装变形和重叠现象导致难以提取显著且多样性特征的问题，本文提出了位置可视化视觉中心机制，通过坐标注意力（CA, coordinate attention）机制[16]增强局部特征的表示能力和位置敏感性；同时借助可学习视觉中心（LVC, learnable visual center）机制[15]充分利用特征中的空间信息为局部特征分配可学习的权重，提高对局部辨别性特征的关注程度。

此外，针对检测过程中商品的预测框可能被剔除而漏检问题，本文将原 IoU[17]损失函数替换为 CIoU[18]，并在此基础上引入功率参数 α [19]，通过增加预测框与真实框比值的梯度权重拉进预测框与真实框之间的距离，以降低漏检率。

1　YOLOX-s 算法原理

YOLOX 算法是当前速度和精度较为均衡的目标检测算法之一，该算法考虑 YOLOv4[20]和 YOLOv5 可能存在一定的过度优化问题，因而以 YOLOv3[21]为基础模型进行了一系列的改进和优化。YOLOX 设计了多个深度与宽度不同的网络结构版本，其 YOLOX-s 版本的网络模型深度与宽度分别为 YOLOX 的 0.33 和 0.5，该网络模型结构如图 1 所示。

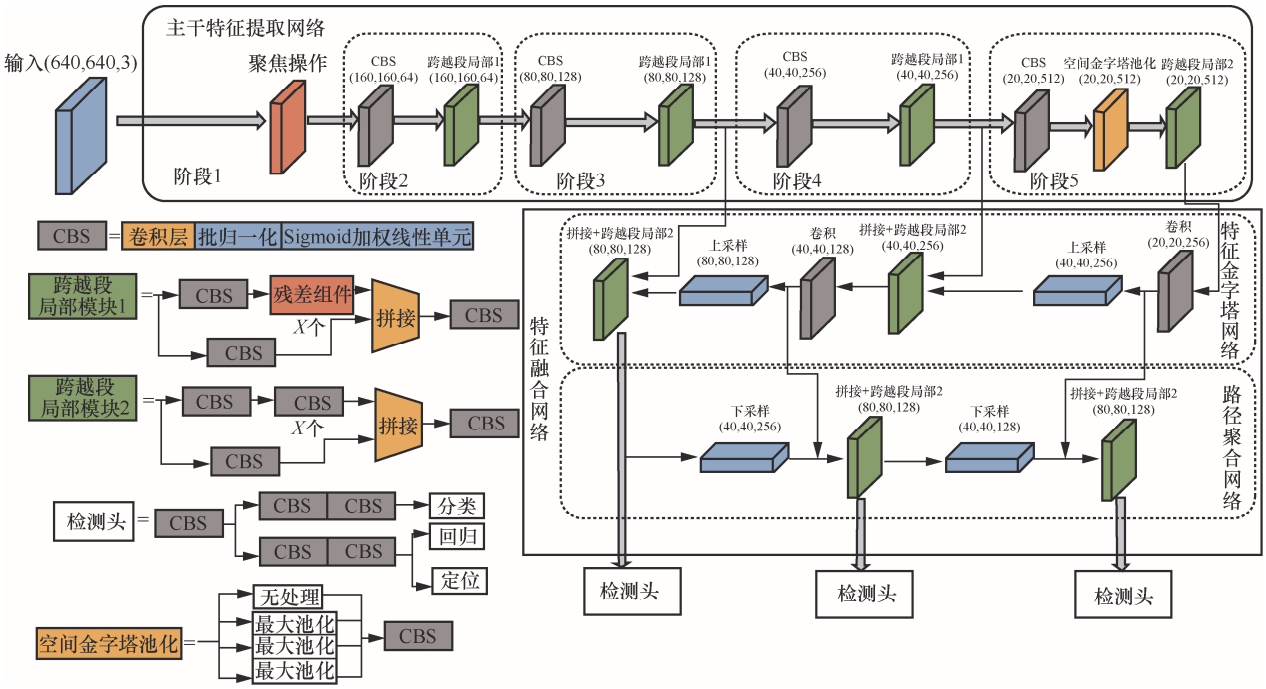


图 1　YOLOX-s 网络模型结构

对于输入图像，YOLOX-s 采用 MixUp 和 Mosaic 数据增强方法，将多张图片进行拼接训练，在一定程度上增强了模型的鲁棒性和泛化能力。

主干特征提取网络，采用了 CSPDarknet 结构，包括聚焦（Focus）、卷积批归一化和 Sigmoid 加权线性单元（CBS, convolution batch normalization sigmoid-weighted linear unit）、空间金字塔池化（SPP, spatial pyramid pooling）和跨阶段局部（CSP, cross stage partial）模块。其中，Focus 模块用于图像裁剪；CBS 模块用于特征图的加权处理，以提高网络的稳定性；CSP 模块将特征映射划分为两个分支进行编码并融合加权处理后的特征图，以提高计算速度和效率；SPP 模块采用 3 种大小各异的池化核，将池化后的特征图与原特征图拼接，以扩大感受野。提取最后 3 个 CSP 模块的输出作为特征融合网络的输入。

特征融合网络采用了特征金字塔网络（FPN, feature pyramid network）[22]和路径聚合网络（PAN, path aggregation network）[23]结构，其中 FPN 通过上采样操作将深层特征与浅层特征融合，自顶向下传递目标语义信息；PAN 通过下采样操作将深层特征传递给浅层特征，自底向上传递目标定位信息。

检测头采用了解耦头结构，缓解了分类任务和定位任务的冲突，加快了模型的收敛速度，并提高了检测精度。

2　YOLOX-s 算法的改进

为有效应对商品包装变形和重叠现象，首先，本文引入轻量级多层感知机，将其应用在主干特征提取网络中，以有效获取全局上下文信息，增强模型对商品整体外观的理解能力；接着，通过 LLVC 增强局部特征的表示能力，并利用特征中的空间信息为局部特征分配可学习的权重，以突出局部辨别性特征，提高检测包装变形和重叠商品的准确率；最后，利用改进的边界框损失函数，降低漏检率。改进后的网络结构如图 2 所示。

2.1　轻量级多层感知机的引入

考虑零售商品检测任务中商品蕴含着丰富的颜色、纹理、形状等信息，捕获全局上下文信息，有利于模型理解包装变形和重叠商品的整体外观，从而提高检测精度。由于 CSPDarknet 结构采用了多个下采样模块，如池化、步长卷积等，这些模块会将输入特征图的尺寸缩小，减小特征图中每个像素点周围的感受野范围，限制模型对包装变形和重叠商品全局上下文信息的理解和利用。为此，本文引入轻量级多层感知机，将其嵌入主干特征提取网络的 Dark3 和 Dark4 层，以有效捕获全局上下文信息，提高检测精度，同时为下一步输送更为丰富的特征信息。轻量级 MLP 结构示意图如图 3 所示。

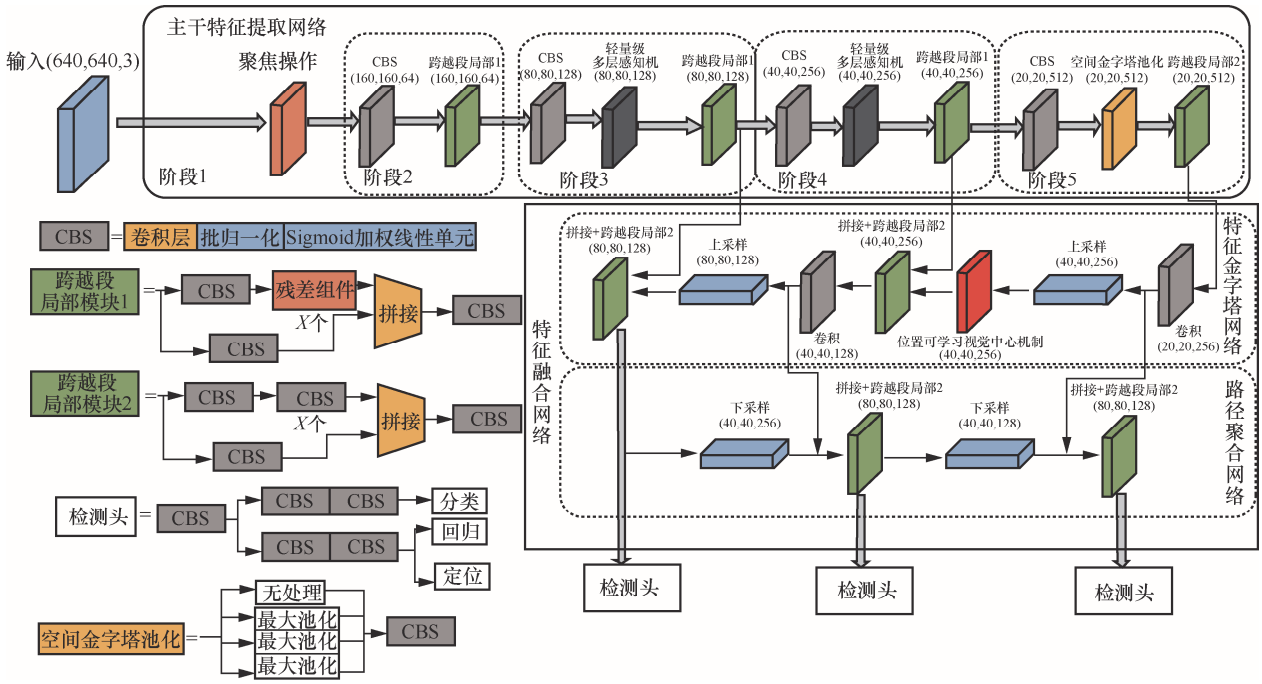


图 2　改进后的网络结构

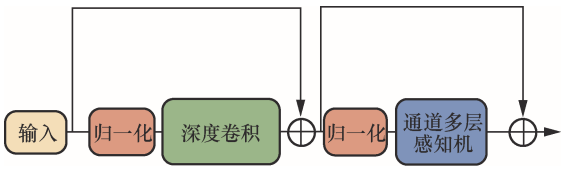


图 3　轻量级 MLP 结构示意图

首先，进行深度卷积[24]，对于一个 $H\times W\times C$ 的特征图，按通道划分为 C 个 $H\times W\times 1$ 的特征图，在每个特征图上采用不同的卷积核进行卷积，并将所有通道输出的特征向量按通道维度进行拼接，得到尺度为 $H\times W\times C$ 的特征图，将其与初始特征相加，得到的特征 $\bar{\boldsymbol{F}}\in\mathbb{R}^{H\times W\times C}$ 具有更大的感受野以及更广阔的上下文信息。计算式为

$$\bar{\boldsymbol{F}}=\mathrm{DConv}_{1\times1}(\mathrm{GN}(\boldsymbol{F}))+\boldsymbol{F} \tag{1}$$

其中，$\mathrm{GN}(\cdot)$ 表示组归一化操作，目的是规范化网络的输出特征图，使其具有相同的统计特性，以加快训练速度并提高模型的准确性；$\mathrm{DConv}_{1\times1}(\cdot)$ 表示卷积核为 1×1 的深度卷积。

接着，利用通道多层感知机[25]对特征 $\bar{\boldsymbol{F}}$ 进行进一步优化。具体地，采用全连接层将每个特征通道上的信息映射到一个向量中，并通过 GELU[26]激活函数进行非线性变换，将这些向量拼接，形成完整的图像特征向量，以提高不同通道特征图之间的关联性和相互作用性。最后，将优化的特征图与原特征图相加，以增强特征的泛化性和鲁棒性。计算式为

$$\mathrm{MLP}(\boldsymbol{F})=\mathrm{CMLP}(\mathrm{GN}(\bar{\boldsymbol{F}}))+\bar{\boldsymbol{F}} \tag{2}$$

其中，$\mathrm{CMLP}(\cdot)$ 表示通道感知机。

2.2　位置可学习视觉中心机制

零售商品检测任务由于其环境的复杂性，存在包装变形和重叠现象，造成部分关键区域信息难以捕捉。局部特征之间的空间关系蕴含丰富的特征空间信息，利用这些空间信息有利于网络聚焦于关键区域。然而，现有的特征融合网络利用上采样和下采样操作对不同尺度的特征进行融合，虽然输出的特征图携带高级语义信息和浅层空间信息，但是上采样和下采样操作缺乏根据空间信息重要性选取关键特征的能力，导致网络对零售商品局部判别性区域的关注程度不高。为此，本文设计了位置可学习视觉中心机制，将坐标注意力机制与可学习视觉中心机制融合，在增强局部特征表示能力的同时，利用上采样模块传递的特征空间信息为局部辨别性特征分配可学习的权重值，使网络更有效地聚焦于零售商品局部辨别性区域，以提高检测精度。位置可学习视觉中心机制结构示意图如图 4 所示。

LLVC 将上采样模块（UpSampling）输出的特征图 $\boldsymbol{F}\in\mathbb{R}^{H\times W\times C}$ 作为输入。为了帮助网络更好地学习特征表示，首先对输入特征进行 3 次卷积操作，

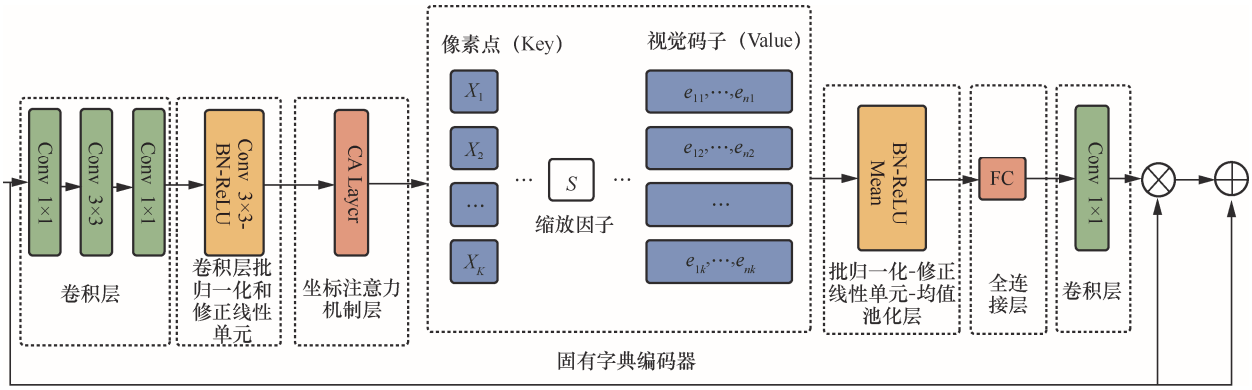


图 4　位置可学习视觉中心机制结构示意图

卷积核大小分别为 1×1、3×3 和 1×1。随后，通过具有批量归一化（BN, batch normalization）[27]和修正线性单元（ReLU, rectified linear unit）[28]激活函数的卷积层对 3 次卷积后的特征向量进行特征编码，将其标准化和非线性激活，在保证特征梯度稳定的同时增强其非线性表达能力。

由于 CA 机制能够对每个通道的特征图加权，以突出目标分类任务的关键特征，并可抑制无关特征。因此，本文使用 CA 机制增强商品辨别性局部特征的表示能力，将编码后的特征向量输入 CA 模块中，沿水平与垂直坐标方向对其进行全局平均池化，以获取 $C\times H\times 1$ 和 $C\times 1\times W$ 两种尺度且带有位置信息的特征映射。随后，对生成的特征映射分别进行特征编码，形成一对方向感知和位置敏感的注意力特征图，将它们与 CA 模块的原始输入特征图相乘，以增强特征的表示能力，使得输出的特征图 $\bar{\boldsymbol{F}}$ 携带权重数据。计算式为

$$\bar{\boldsymbol{F}}=\mathrm{CA}(\mathrm{CBR}(\mathrm{Conv}(\boldsymbol{F}))) \tag{3}$$

$$\mathrm{CA}=\boldsymbol{F}(i,j)\times x_c^h(i)\times x_c^w(j) \tag{4}$$

其中，Conv(·) 表示卷积层，CBR(·)表示具有批量归一化和 ReLU 激活函数的 3×3 卷积层，CA(·)表示坐标注意力机制，(i,j)表示特征图的位置坐标。

为了有效利用权重数据以突出局部辨别性特征，LLVC 将特征图 $\bar{\boldsymbol{F}}$ 输入固有字典编码器中，并将其分割成多个局部区域。对于每个局部区域构建一个 Key-Value 键值对集合。其中，局部区域内的所有特征向量作为视觉码字（Value），对应的空间位置作为像素点（Key）。在获得键值对集合后，利用一组比例因子 S，依次对 Key 和 Value 进行缩放，随后将其映射到原图相应的位置。在映射后的位置上，取所有 Key 的平均值作为新的 Key，并将该位置上所有的视觉码字通过加权平均的方式进行权重分配，将两者组合成新的 Key-Value 键值对集合，以提高局部特征表示能力。在获取所有视觉码字后，利用 BN-ReLU-均值池化（Mean, mean pooling）[29]融合整个图像对于 K 个视觉码字的完整信息，关联其局部特征与对应的全局空间信息，以便更好地理解和利用这些局部特征。上述过程可表示为

$$e=\sum_{k=1}^{K}\mathrm{BRM}\left(\sum_{i=1}^{N}\frac{e_k^{-s_k\bar{\boldsymbol{F}}_i-b_k{}^2}}{\sum_{j=1}^{K}e_j^{-s_k\bar{\boldsymbol{F}}_i-b_k{}^2}}(\bar{\boldsymbol{F}}_i-b_k)\right) \tag{5}$$

其中，BRM 由 ReLU、BN 层和 Mean 层组成，N 表示输入特征图的空间总数，$\bar{\boldsymbol{F}}_i$ 表示第 i 个像素点，b_k 表示第 k 个可学习视觉码字，s_k 表示第 k 个比例因子，K 是可学习视觉中心总数，$\bar{\boldsymbol{F}}_i-b_k$ 表示每个编码对应的像素位置信息。

经过上述过程得到特征图 e 后，进一步将其输入全连接层和 1×1 卷积之中，通过滤波器寻找具有高权重位置信息的局部判别性特征，提高网络模型对包装变形和重叠商品的辨别能力。计算式为

$$\boldsymbol{Z}=\boldsymbol{F}\otimes(\delta(\mathrm{Conv}_{1\times1}(e))) \tag{6}$$

其中，$\mathrm{Conv}_{1\times1}(\cdot)$ 表示 1×1 卷积，$\delta(\cdot)$ 表示 Sigmoid 函数，$\otimes$ 表示通道相乘。

最后，为了缓解网络模型梯度消失，将原特征图与具有局部区域位置信息的特征图进行通道相加。计算式为

$$\mathrm{LLVC}(\boldsymbol{F})=\boldsymbol{F}\oplus\boldsymbol{Z} \tag{7}$$

其中，$\oplus$ 表示通道相加。

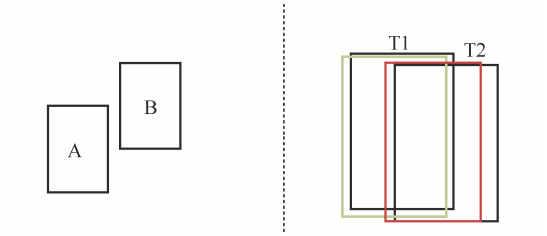
2.3　损失函数的改进

在原 YOLOX-s 网络中，采用 IoU 损失作为边界框损失函数评价预测框与真实框的距离，具体计算式为

$$\mathrm{IoULoss}=1-\mathrm{IoU} \tag{8}$$

$$\mathrm{IoU}=\frac{|A\bigcap B|}{|A\bigcup B|} \tag{9}$$

其中，A 和 B 分别代表预测框和真实框。预测框与真实框位置的不同情况如图 5 所示，当 A 与 B 不相交，如图 5(a)所示，此时 IoU 为 0，无法衡量两边界框的距离，也就不能进行梯度回传，从而导致学习训练失败。当出现物体重叠现象，如图 5(b)所示，其中 T1 和 T2 分别代表两个物体的真实框，绿色框代表 T1 的预测框，红色框代表 T2 的预测框，若此时 T1 与绿色预测框的 IoU 值大于其与红色预测框的 IoU 值，将剔除红色预测框，T2 无预测框可用，导致模型检测精度下降。



（a）预测框与真实框不相交　（b）真实框与多个预测框相交

图 5　预测框与真实框位置的不同情况

为此，本文将原 IoU 损失函数替换为 CIoU，CIoU 损失函数不仅考虑了预测框和真实框的交集、并集，还考虑了两个框之间的完整性和平滑性。当预测框与目标框不重叠时，CIoU 损失函数通过计算两边界框中心点距离和长宽比误差为边界框提供移动方向。重叠时，则通过计算真实框与预测框的中心点距离，根据距离长短选择所需的预测框，有效降低了漏检率，具体的计算式为

$$\mathrm{IoU}=1-\mathrm{IoU}(A,B)+\left(\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}+\beta\nu\right) \tag{10}$$

$$\beta=\frac{\nu}{(1-\mathrm{IoU})+\nu} \tag{11}$$

$$\nu=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{12}$$

其中，$\rho^2(b,b^{\mathrm{gt}})$ 表示预测框与真实框中心点的欧氏距离，c 表示两者矩形的对角线距离，β 为协调比例参数，ν 为边界框的长宽比参数，w 和 h 分别表示预测框的宽度和高度，w^{gt} 和 h^{gt} 分别表示真实框的宽度和高度。

此外，本文引入功率参数 α，通过增加预测框与真实框比值的梯度权重来进一步降低漏检率。最终的边界框损失函数 CIoU-α 计算式为

$$\text{CIoU-}\alpha=1-\mathrm{IoU}^{\alpha}(A,B)+\left(\frac{\rho^{2\alpha}(b,b^{\mathrm{gt}})}{c^{2\alpha}}+(\beta\nu)^{\alpha}\right) \tag{13}$$

3　实验与结果分析

为验证所提方法的性能，本文通过消融实验对所提方法分别进行验证。同时与 YOLOv3、YOLOv3-Tiny、YOLOv4-Tiny、YOLOv5-s、YOLOv7-Tiny、YOLOv8-s、YOLOX-Tiny 和 YOLOX-s 等主流的目标检测算法在 RPC 数据集中进行对比实验。此外，可视化分析了高精度、低参数量的模型。

3.1　实验环境

本文实验系统环境为 Ubuntu 16.04，采用主流的 PyTorch 框架进行训练及测试。软件环境为 cuda11.4 和 Python3.8.13。硬件配置为 NVIDIA Geforce RTX 3090 GPU、24 GB 显存；Intel Xeon 银牌 4210R CPU；用于嵌入深度学习模型的边缘设备为 NVIDIA® Jetson™ TX2。

实验设置统一训练参数，初始学习率为 0.01，选择随机梯度下降算法作为优化器，动量（momentum）设置为 0.9，功率参数设置为 3，权值衰减（weight decay）设置为 0.000 5，批大小（batch size）为 16。余弦退火能够帮助学习率逃离当前局部最优点，在经过多次训练后，模型收敛到一个更好的局部最优点，而预热能够帮助模型更快地收敛到合适的学习率范围。因此，学习率调度器采用余弦退火学习率[30]和预热学习率的组合，经过调整后的最终学习率为 0.002 5，最小学习率为所设置当前学习率的 5%，在最后 15 轮固定采用最小学习率。

3.2　数据集

RPC 数据集无论是图像数量或是商品类别都达到了该领域之最，并且相较于其他商品数据集的商品类别、个数、包装变形、重叠等因素均更接近真实场景。因此，选择 RPC 数据集评估所提模型，选取该数据集中 30 000 张结算环节下拍摄的图片作为实验数据，共 200 类商品。其中，按零售商品摆放杂乱级别，将商品类别为 3～5 种、数量 3～10 个且包装未变形、未重叠的图片划分为简单模式；商品类别为

5～8 种、数量 10～15 个且包装轻度变形和重叠的划分为中等模式，商品类别为 8～10 种、数量 15～20 个且存在包装大量变形和重叠的则划分为复杂模式，商品结算环节不同摆放难度示例如图 6 所示，共计 24 000 张，各难度数据占比为 1:1:1。在进行实验前，将该数据集图片标注信息的 XML 文件转换成 VOC 形式的 TXT 格式标注文件，同时将 30 000 张图片以 8:1:1 的比例分别作为训练集、测试集和验证集。

(a) 简单模式　(b) 中等模式　(c) 复杂模式

图 6　商品结算环节不同摆放难度示例

3.3　评价指标

本文采用均值平均精度（mAP, mean average precision）、模型参数总量（params）和每秒帧数（FPS）作为评价指标对算法检测精度进行定量评估。

mAP 由精确率（precision）和召回率（recall）求出。精确率又被称为查准率，用于衡量算法的准确度；召回率又称全查率，用于衡量算法的漏检率。计算式为

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{14}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{15}$$

其中，TP 表示正确分类样本数量，FP 表示误分样本数量，FN 表示样本误检数量。

平均精度（AP, average precision）用于计算单类别的检测精度，其表示精确率与召回率所围成曲线的面积，计算式为

$$\text{AP}=\sum_{i=1}^{n-1}(r_{i+1}-r_i)\text{Pinterp}(r_{i+1}) \tag{16}$$

其中，r_n 表示按升序排列的 Precision 插值段对应的 Recall 值，Pinterp 为插值操作。

因此，用于计算多个类别平均精度的均值平均精度计算式为

$$\text{mAP}=\frac{\sum_{j=1}^{M}\text{AP}_j}{M} \tag{17}$$

其中，M 代表数据集中所有类别个数。

3.4　消融实验

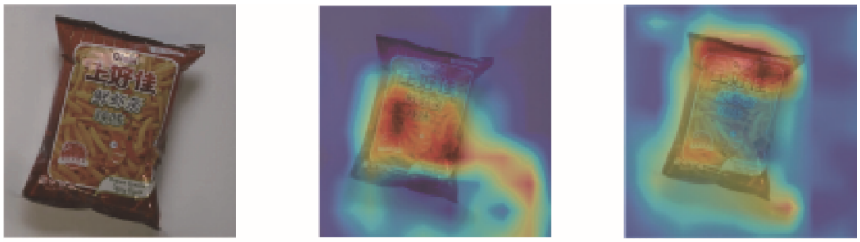
为验证各改进模块的有效性，在 YOLOX-s 中分别融入各改进模块，并在 RPC 数据集上展开消融实验分析。消融实验结果见表 1。

表 1　消融实验结果

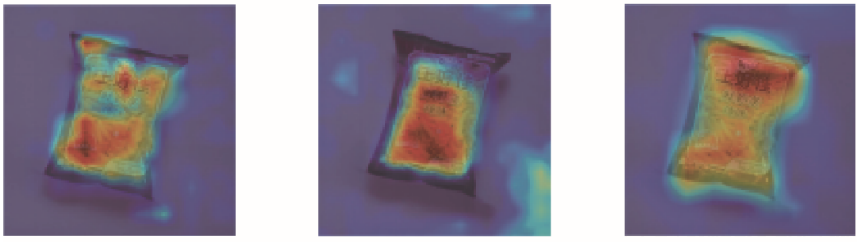
方法	mAP	params/MB	FPS/(frame·s^{-1})
YOLOX-s (baseline)	89.1%	9.0	102
YOLOX-s (MLP)	89.8%	9.21	100
YOLOX-s (LLVC)	90.1%	9.29	100
YOLOX-s (CIoU-α)	89.5%	9.0	102
YOLOX-s (MLP&LLVC&CIoU-α)	91.3%	9.48	97

从表 1 的结果可以看出，在基线模型基础上只嵌入轻量级多层感知机模块，模型的 mAP 为 89.8%，检测精度相较基线模型提升了 0.7%。嵌入位置可学习视觉中心机制后，模型的 mAP 达到 90.1%，相较于基线模型提升 1%。在 YOLOX-s 的基础上将 IoU 替换为 CIoU-α 损失函数，mAP 为 89.5%，相比基线模型提升了 0.4%。当轻量级多层感知机、位置可视化中心学习机制和 CIoU-α 损 3 种改进策略共同加入时，检测精度为 91.3%，相较于基线模型提升了 2.2%。为进一步验证本文模型的有效性，以可视化热图形式验证嵌入不同改进模块后，模型预测出的判别性区域位置。嵌入各改进模块的热图对比如图 7 所示，热图中红色高亮部分代表与预测类别和定位相关的区域，红色高亮区域覆盖于商品判别性区域的面积越多越准确，模型对商品的检测精度越高。

YOLOX-s 融合 MLP 模块后，感受到更多的区域信息，说明增大模型感受野后可有效获取到商品全局上下文信息。YOLOX-s 融合 LLVC 模块后，热图中红色高亮区域离散地分布于各个判别性区域、高亮区域被控制在商品范围内，验证了 LLVC 模块能够充分捕获细节特征信息，并且能够有效抑制冗余干扰。YOLOX-s 在采用 CIoU-α 损失函数后，相较于 YOLOX-s 红色高亮区域左移，被限制在商品范围内，表明采用 CIoU-α 损失函数后物体的定位能力得到了提升。在融合 MLP、LLVC 和 CIoU-α，热图中红色高亮区域覆盖更广且集中于商品的判别性区域，在 3 种改进策略相辅相成的作用下，该模型在检测精度方面得到了有效提升。虽然本文模型与基线模型相比，在检测速度上降低了 5 frame/s，模型参数量提高了 0.48 MB，其重要原



(a) 原图　(b) YOLOX-s　(c) YOLOX-s +MLP



(d) YOLOX-s +LLVC　(e) YOLOX-s +CIoU-α　(f) YOLOX-s + MLP +LLVC +CIoU-α

图 7　嵌入各改进模块的热图对比

因是轻量级多层感知机模块和位置可视化中心学习机制的引入，在一定程度上增加了计算量和参数量，但改进后的模型每秒检测图像数量为 97 张、模型参数量为 9.48 MB，仍满足现实场景下商品检测的实时性要求（FPS≥40 frame/s，params<8 GB）。因此，本文将 YOLOX-s（MLP&LLVC&CIoU-α）作为最终模型，以下简称 YOLOX-MLC。

3.5　对比实验

3.5.1　零售商品不同杂乱级别摆放

为检验改进方法在自助结算环节下零售商品检测的有效性和先进性，将 YOLOX-MLC 与基线模型在简单模式、中等模式和复杂模式 3 种杂乱级别摆放的 RPC 数据集下分别进行对比实验，零售商品不同杂乱级别摆放下检测精度对比见表 2。

表 2　零售商品不同杂乱级别摆放下检测精度对比

方法	mAP		
	简单模式	中等模式	复杂模式
YOLOX-s	87.1%	85.9%	82.6%
YOLOX-MLC	89.5%	88.9%	86.8%

由表 2 结果可以看出，YOLOX-MLC 在简单模式、中等模式和复杂模式 3 种杂乱级别摆放的 RPC 数据集下 mAP 分别为 89.5%、88.9%和 86.8%，相较于基线网络分别提升了 2.4%、3.0%和 4.2%。由此可知，摆放方式越接近现实场景，本文改进的模型检测效果越显著。

3.5.2　其他主流检测模型

为进一步验证 YOLOX-MLC 在 mAP、params 和 FPS 3 个评价指标上的综合优越性，分别选取 YOLOv3、YOLOv3-Tiny、YOLOv4-Tiny、YOLOv5-s、YOLOv7-Tiny、YOLOX-Tiny、YOLOX-s、YOLOv8-s 等主流方法，在 RPC 数据集上进行对比实验，不同模型检测精度对比见表 3。

表 3　不同模型检测精度对比

方法	mAP	params/MB	FPS/(frame·s^{-1})
YOLOv3	87.3%	63.0	74
YOLOv3-Tiny	72.2%	8.0	277
YOLOv4-Tiny	74.4%	6.1	270
YOLOv5-s	86.8%	7.3	156
YOLOv7-Tiny	85.2%	6.1	286
YOLOX-Tiny	81.9%	5.06	296
YOLOX-s	89.1%	9.0	102
YOLOv8-s	90.5%	12.5	151
YOLOX-MLC	91.3%	9.48	97

由表 3 可知，YOLOX-MLC 的 mAP 为 91.3%，相比于 YOLOv3、YOLOv5-s、YOLOv7-Tiny、YOLOX-s 和 YOLOv8-s 分别提升了 4.0%、4.5%、6.1%、2.2%和 0.8%，相比于 YOLOv3-Tiny、YOLOv4-Tiny 和 YOLOX-Tiny 分别提升了 19.1%、16.9%和 9.4%，在 9 种检测模型中拥有最高的 mAP，可以对零售商品进行准确定位和分类。在检测速度方面，YOLOX-MLC 相较于 YOLOv3-Tiny、YOLOv4-Tiny、YOLOv5-s、YOLOv7-Tiny、YOLOX-

Tiny、YOLOX-s 和 YOLOv8-s 分别下降了 65%、65%、38%、66%、67%、5%和 36%，相较于 YOLOv3 提升了 31%。在模型参数量方面，YOLOX-MLC 相较于 YOLOv3-Tiny、YOLOv4-Tiny、YOLOv5-s、YOLOv7-Tiny、YOLOX-Tiny、YOLOX-s 分别增加 1.48 MB、3.38 MB、2.18 MB、3.38 MB、4.42 MB、0.48 MB，相较于 YOLOv3 和 YOLOv8-s 降低了 53.52 MB 和 3.02 MB。虽然 YOLOX-MLC 在检测速度上不具备优势，但仍有 97 frame/s，且模型参数量为 9.48 MB，能够在资源受限的物联网设备上完成商品实时检测的工作。综上分析可知，YOLOX-MLC 实现了检测精度、检测速度和模型性能的充分平衡，更适用于应对零售商品检测任务中出现的商品包装变形和重叠现象。

3.6　模型可视化分析

为直观展示 YOLOX-MLC 在应对商品包装变形和重叠现象的优越性，选择在包含 3 类商品且不重叠、不变形的简单模式，包含 6 类商品且存在轻度变形和重叠现象的中等模式以及包含 9 类商品且存在大量变形和重叠现象的复杂模式下进行可视化结果分析测试，将 YOLOX-MLC 与检测精度高且模型参数量低的主流目标检测模型 YOLOX-Tiny、YOLOv7-Tiny、YOLOv5-s、YOLOX-s 和 YOLOv8-s 进行对比，RPC 数据集测试效果如图 8 所示。

图 8　RPC 数据集测试效果

通过对比发现，当处于不存在变形和重叠现象的简单模式时，所列举的模型均能对商品进行准确定位与识别；当处于商品轻度变形和重叠现象的中等模式时，YOLOX-Tiny 出现商品漏检以及识别错误，YOLOv5-s、YOLOv7-Tiny 和 YOLOX-s 均出现商品漏检现象；当处于存在大量变形以及重叠的复杂模式时，YOLOX-Tiny 出现大量的漏检商品，YOLOv5-s 和 YOLOv7-Tiny 漏检两件商品，YOLOX-s 出现商品漏检和识别错误。相较于以上目标检测模型，YOLOX-MLC 在不同杂乱级别摆放下均能够准确判断商品类别，并且有效解决了商品间重叠、同一商品形变造成准确率下降的问题。在图 8 中，YOLOv8-s 也呈现出良好的检测效果。为进一步验证本文方法在检测变形和重叠商品方面的优势，选择变形和重叠现象更加严重的两种图片，其中一种商品形状变化更剧烈，另一种商品重叠数量更多且重叠面积更大。变形和重叠现象更加严重的测试效果如图 9 所示。

图 9　变形和重叠现象更加严重的测试效果

4　结束语

本文提出一种基于位置可学习视觉中心机制的零售商品检测模型 YOLOX-MLC。在 YOLOX-s 网络的主干特征提取网络结构中嵌入轻量级多层感知机扩大网络的感受野，充分捕获商品的全局上下文信息，以增强对变形和重叠商品整体外观

的理解能力；其次，利用位置可学习视觉中心机制改进特征融合网络结构，增强对商品判别性特征提取能力，以提高变形、重叠商品的检测精度；最后，利用 CIoU-α 损失函数进一步提升商品检测精度。实验结果表明，在公开的 RPC 数据集上，YOLOX-MLC 的检测精度为 91.3%，相比于 YOLOX-s 提高 2.2%。在 3 种检测难度的数据集下检测精度分别为 89.5%、88.9%和 86.8%，相较于 YOLOX-s 分别提升了 2.4%、3.0%和 4.2%。YOLOX-MLC 的 FPS 为 97 frame/s，params 仅 9.48 MB，表明所提模型在计算资源有限的情况下，可以准确且实时地完成零售商品检测任务。

参考文献：

[1] SARAN A, HASSAN E, MAURYA A K. Robust visual analysis for planogram compliance problem[C]//Proceedings of 2015 14th IAPR International Conference on Machine Vision Applications (MVA). Piscataway: IEEE Press, 2015: 576-579.

[2] RAY A, KUMAR N, SHAW A, et al. U-PC: Unsupervised planogram compliance[C]//European Conference on Computer Vision. Cham: Springer, 2018: 598-613.

[3] GEORGE M, FLOERKEMEIER C. Recognizing products: a per-exemplar multi-label image classification approach[C]//European Conference on Computer Vision. Cham: Springer, 2014: 440-455.

[4] HIGA K, IWAMOTO K, NOMURA T. Multiple object identification using grid voting of object center estimated from keypoint matches[C]//Proceedings of 2013 IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2014: 2973-2977.

[5] BAO R, HIGA K, IWAMOTO K. Local feature based multiple object instance identification using scale and rotation invariant implicit shape model[C]//Computer Vision-ACCV 2014 Workshops. Cham: Springer International Publishing, 2015: 600-614.

[6] YÖRÜK E, ÖNER K T, AKGÜL C B. An efficient Hough transform for multi-instance object recognition and pose estimation[C]//Proceedings of 2016 23rd International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2017: 1352-1357.

[7] WEI Y C, TRAN S, XU S X, et al. Deep learning for retail product recognition: challenges and techniques[J]. Computational Intelligence and Neuroscience, 2020: 8875910.

[8] HURTIK P, MOLEK V, VLASANEK P. YOLO-ASC: you only look once and see contours[C]//Proceedings of 2020 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2020: 1-7.

[9] GOLDMAN E, HERZIG R, EISENSCHTAT A, et al. Precise detection in densely packed scenes[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 5222-5231.

[10] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 2818-2826.

[11] SELVAM P, KOILRAJ J A S. A deep learning framework for grocery product detection and recognition[J]. Food Analytical Methods, 2022, 15(12): 3498-3522.

[12] WANG H I, MIYAZAKI L K, FALHEIRO M S, et al. Designing a self-payment cashier for bakeries using YOLO V4[C]//Proceedings of 2021 14th IEEE International Conference on Industry Applications (INDUSCON). Piscataway: IEEE Press, 2021: 260-265.

[13] GE Z, LIU S, WANG F, et al. YOLOX: exceeding YOLO series in 2021[J]. arXiv preprint, 2021, arXiv: 2107.08430.

[14] WANG H I, MIYAZAKI L K, FALHEIRO M S, et al. Designing a self-payment cashier for bakeries using YOLO V4[C]//Proceedings of 2021 14th IEEE International Conference on Industry Applications (INDUSCON). Piscataway: IEEE Press, 2021: 260-265.

[15] QUAN Y, ZHANG D, ZHANG L, et al. Centralized feature pyramid for object detection[J]. arXiv preprint, 2022, arXiv: 2210.02093.

[16] HOU Q B, ZHOU D Q, FENG J S. Coordinate attention for efficient mobile network design[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 13708-13717.

[17] YU J H, JIANG Y N, WANG Z Y, et al. UnitBox: an advanced object detection network[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 516-520.

[18] ZHENG Z H, WANG P, LIU W, 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.

[19] HE J, ERFANI S, MA X, et al. Alpha-IoU: a family of power intersection over union losses for bounding box regression[J]. Advances in Neural Information Processing Systems, 2021, 34: 20230-20242.

[20] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv preprint, 2020: arXiv: 2004.10934.

[21] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv preprint, 2018, arXiv: 1804.02767.

[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 936-944.

[23] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.

[24] HOWARD A G, ZHU M, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, 2017, arXiv: 1704.04861.

[25] TOLSTIKHIN I, HOULSBY N, KOLESNIKOV A, et al. MLP-mixer: an all-MLP architecture for vision[J]. Advances in Neural Information Processing Systems, 2021(34): 24261-24272.

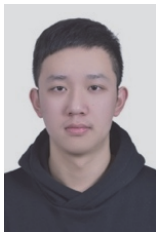
[26] HENDRYCKS D, GIMPEL K. Gaussian error linear units (GELUs)[J]. arXiv preprint, 2016, arXiv: 1606.08415.

[27] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning. New York: ACM Press, 2015: 448-456.

[28] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
[29] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. arXiv preprint, 2016, arXiv:1607.06450.
[30] LOSHCHILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts[J]. arXiv preprint, 2016, arXiv:1608.03983.

[作者简介]

吕晓华（2000− ），男，宁夏大学信息工程学院硕士生，主要研究方向为基于深度学习的细粒度商品检测、增量学习。



魏铭辰（1993− ），男，宁夏大学信息工程学院硕士生，主要研究方向为基于深度学习的细粒度商品检测。

刘立波（1974− ），女，博士，宁夏大学教授、博士生导师，主要研究方向为智能信息处理、计算机视觉。