DOI:10.3969/j.issn.1671-024x.2026.02.012

基于混合 CNN-Transformer 的堆垛纸箱检测方法

肖志涛[1,2],王 宇[2,3]
(1. 天津工业大学 生命科学学院, 天津 300387; 2. 天津工业大学 天津市光电检测技术与系统重点实验室, 天津 300387; 3. 天津工业大学 电子与信息工程学院, 天津 300387)

摘 要: 针对卷积神经网络(CNN)多层卷积和池化操作导致的空间信息处理不足和全局上下文信息缺乏的问题,提出了一种基于 CNN 与 Transformer 的多尺度密集堆垛纸箱检测方法。设计特征提取融合模块,结合窗口自注意力机制,增强模型对全局特征的建模能力;引入跨尺度连接,融合更多不同层级的语义信息,使模型具备更大的感受野和更好的特征融合能力;提出 BoxIoU 损失函数用于边界框回归,通过计算边界框的最小点距和宽高比评估边界框相似性,提高模型的检测精度。实验结果表明:在密集堆垛纸箱数据集(SCD)上,该方法的 mAP50 达到了 99.36%,mAP50-95 达到了 95.09%,具有良好的检测性能以及泛化能力。

关键词: 深度学习;堆垛纸箱;目标检测;自注意力机制
中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1671-024X(2026)02-0094-07

Stacked carton detection method based on hybrid CNN-Transformer

Xiao Zhitao[1,2], Wang Yu[2,3]
(1. School of Life Sciences, Tiangong University, Tianjin 300387, China; 2. Tianjin Key Laboratory of Optoelectronic Detection Technology and System, Tiangong University, Tianjin 300387, China; 3. School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China)

Abstract: In response to the problems of insufficient spatial information processing and lack of global contextual information caused by multi-layer convolution and pooling operations in convolutional neural networks(CNN), a multi-scale densely stacked carton detection method based on CNN and Transformer is proposed. A feature extraction and fusion module is designed, combined with a window self-attention mechanism, to enhance the model's ability to model global features. Cross-scale connections are introduced to fuse more semantic information at different levels, and the model has a larger receptive field and better feature fusion ability. A BoxIoU loss function is proposed for bounding box regression, which evaluates the similarity of bounding boxes by calculating the minimum point distance and aspect ratio of bounding boxes, improving the accuracy of the model. Experimental results show that on the SCD dataset of densely stacked cartons, the method achieves an mAP50 of 99.36% and an mAP50-95 of 95.09%, with good detection performance and generalization ability.

Key words: deep learning; stacked cartons; object detection; self-attention mechanism

随着电子商务的快速发展,快递物流行业面临着更大的挑战和更高的要求。仓储物流场景中,快递物流通常使用纸箱作为包装,并以堆垛的形式存放[1]。机器视觉技术已广泛应用于机器人自动化领域。因此,研究高精度的目标检测算法对于实现机器人自动化卸垛至关重要[2]。

在物流环境中进行纸箱检测存在一些主要挑战。例如,纸箱的堆叠、相互遮挡以及多样的包装形式都会降低检测的准确性和泛化能力[3]。基于卷积神经网络(CNN)的目标检测器已广泛应用于各种工业流水线及物流场景,如快速区域卷积神经网络(Faster R-CNN)[4]、掩码区域卷积神经网络(Mask R-CNN)[5]、单

收稿日期: 2024-05-11
基金项目: 京津冀基础研究合作专项项目(21JCZXJC00170)
通信作者: 肖志涛(1971—),男,博士,教授,主要研究方向为医学图像处理。E-mail:xiaozhitao@tiangong.edu.cn

阶段检测器（YOLO）[6-9]以及视网膜网络（RetinaNet）[10]等，并取得了一些显著成果[11]。张亚辉[12]将 Faster R-CNN 目标检测模型引入机器人抓取系统中完成目标纸箱识别任务。陶磊等[13]在 Mask R-CNN 中提出了添加上下文机制和改善损失函数等改进方案。但在密集目标检测任务中，两阶段检测器生成大量的候选框会增加计算量，并导致大量的冗余或重叠的候选框，降低检测性能。

相对于两阶段检测器，一阶段检测器具有不需要进行候选框生成和筛选的优势。王佳卓[14]在 RetinaNet 中添加分割头辅助检测头以提高检测精度，但在处理尺度变化时表现不如一些多尺度检测算法。最新的 YOLOv8 模型采用无锚框解决回归边界框尺寸多变的问题，能处理同一幅图像中目标尺寸不同的情况。但由于 YOLOv8 模型以 Darknet 作为特征提取网络，网络层数较浅，难以提取细致的目标特征，导致模型直接在底层特征上进行预测，无法充分利用上下文信息。因此，需要对 YOLOv8 模型进行改进以提升其全局特征建模能力。

Transformer 逐渐替代 CNN 成为计算机视觉领域中一个热门的研究方向[15-18]。研究人员发现Transformer 可以弥补 CNN 对捕捉全局特征信息和长距离依赖关系能力的不足。Carion 等[19]设计了一种使用 CNN 提取图像特征，并结合自注意力机制的 Transformer 模型进行目标检测的端到端模型 DETR。它通过将 CNN 提取的特征输入到 Transformer 模型中，实现了无需锚框和非极大值抑制（NMS）的目标检测。Dosovitskiy 等[20]提出了一种基于 Transformer 架构的图像分类模型Vision Transformer（ViT），直接使用纯 Transformer 进行分类任务，将 Transformer 模型成功应用于图像领域，提供了不依赖 CNN 的图像处理方法。Liu 等[21]根据 ViT 思想提出了一种新的通用骨干网 Swin Transformer，用于处理计算机视觉任务，在处理目标检测、语义分割和图像分类等任务中表现非常出色。由于更加注重全局特征的特性，其在处理局部纸箱细节如边缘、图标等方面不如一些专注于局部特征提取的模型，因此，在局部细节检测方面存在一定局限性。

本研究基于自注意力思想提出了一种一阶段检测模型，即 ST-YOLO。该模型在特征提取阶段利用 CNN 获取的局部特征和 Swin Transformer 中的全局特征，通过加权双向特征融合，并使用 BoxIOU 损失函数训练得到最优网络模型，以期增强模型对全局特征和局部特征的建模能力，提高密集纸箱检测和识别的准确性与稳定性。

1　网络结构

1.1　整体网络结构

本文提出的 ST-YOLO 整体结构以 YOLO 框架为主体，主要由特征提取网络（Backbone）、特征融合网络（Neck）和预测网络（Head）3 部分组成。整体网络结构如图 1 所示。

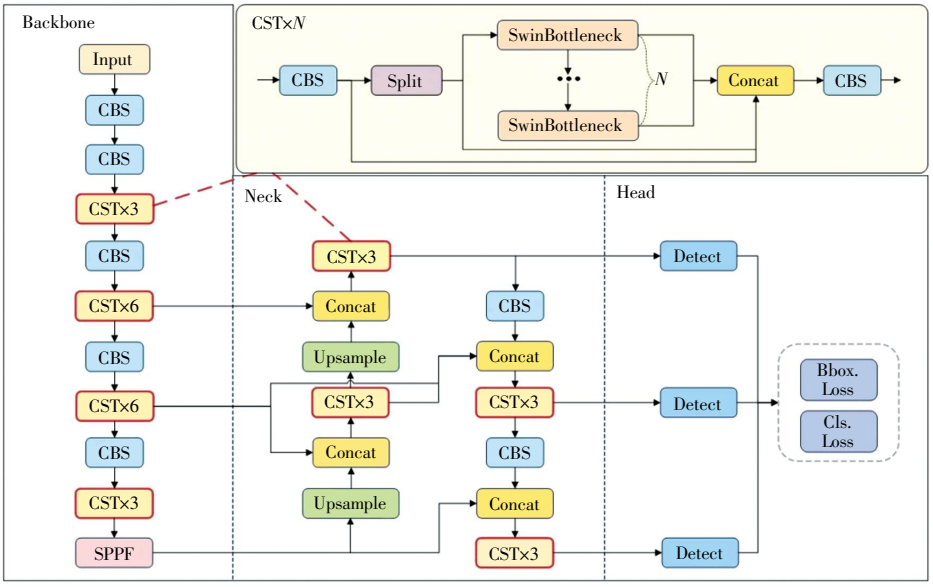


图 1　ST-YOLO 整体网络结构

Fig.1　Overall network architecture of ST-YOLO

图像输入到由普通卷积层、CST（CNN-Swin Transformer）模块以及快速空间金字塔池化（SPPF）模块构成的特征提取网络中。普通卷积层通过下采样操作压缩图像分辨率。CST 模块将特征图分别输入到由卷积

层和 Transformer Block 组成的分支中，另一路则作为旁路直接传递。两路分支得到的特征映射进行融合，将融合结果输出给下一个模块。通常在下采样卷积层之后会堆叠多个 CST 模块，经过多次下采样操作后，即可形成初步的骨干网络。

特征融合网络接收来自前级网络的多尺度特征图，通过加权双向特征金字塔网络（BiFPN）[22]实现特征融合。其中，自顶向下的路径主要是由上采样模块以及 CST 模块构成，自底向上的路径主要是由下采样模块和 CST 模块构成，并引入同尺度跨层连接，保留更多深层和浅层语义信息。

预测头接收到来自特征融合网络的特征输出，利用带有高效通道注意力的解耦头在不同尺度的特征图上实现最终的分类与回归，提出 BoxIOU 损失函数计算边界框回归损失，提升模型的鲁棒性和泛化能力。

1.2　CST 模块

针对堆垛纸箱的尺寸多变、堆叠和遮挡问题，需要进一步提升网络的全局特征提取能力。CNN 通过池化层对输入数据进行下采样操作，捕获输入图像的全局特征。以这种方式提取全局特征时，会导致输入数据的部分细微特征丢失，无法充分捕捉到局部区域和整体之间复杂的关联性。此外，CNN 卷积层中的卷积核是局部感知的，并采用参数共享的机制，即同一卷积核在整个输入图像上进行滑动。这种机制仅关注局部特征，忽略全局特征，导致模型无法正确区分纸箱之间的重叠部分或遮挡部分，造成漏检或误检的现象。

为了解决 CNN 特征局部性的问题，本文设计了基于混合 CNN-Transformer 思想的 CST 模块，利用多头自注意力机制和卷积融合模块提取特征。CST 模块基于 C2f 模块进行改进，将原本的 Bottleneck 模块替换为添加了 Transformer Block 的 SwinBottleneck 模块。CST 模块的整体结构如图 1 右上所示，主要由 N 个 SwinBottleneck 模块堆叠结合残差连接组成，优化了网络的梯度长度。SwinBottleneck 模块是基于 Swin Transformer 中的 Transformer Block 和卷积模块设计的卷积基础模块，通过注意力机制捕获特征的长距离依赖关系，与卷积层获取的边缘和纹理特征充分结合，以实现更为丰富的特征表示。

SwinBottleneck 模块的结构如图 2 所示。将输入特征图分为 2 路：第 1 路继续直接向下传递特征；而第 2 路分别输入到卷积模块 CBS 和第 1 个 Transformer Block 中。将第 1 个 Transformer Block 的输出与卷积模块 CBS 的输出融合后输入下 1 层的卷积模块中。最终将第 2 层中卷积模块 CBS 的输出、第 2 个 Transformer Block 的输出与第 1 路的原始特征融合后输出。

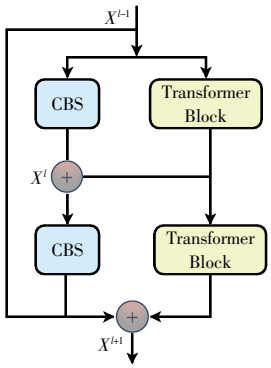


图 2　SwinBottleneck 模块结构

Fig.2　Structure of SwinBottleneck module

CBS 模块由卷积、批量归一化（BN）和 SiLU 激活函数组成。2 个 Transformer Block 都包含多头自注意力机制和前馈神经网络，用于对输入特征进行全局特征建模。为实现窗口间的交互，第 1 个 Transformer Block 使用基于窗口的多头自注意力模块（W-MSA），第 2 个 Transformer Block 使用基于移动窗口的多头自注意力模块（SW-MSA）串联组成。SwinBottleneck 模块的计算过程如式（1）和式（2）：

$$X^{l} = \mathrm{CBS}(X^{l-1}) + \mathrm{W\text{-}MSA}(X^{l-1}) \tag{1}$$

$$X^{l+1} = \mathrm{CBS}(X^{l}) + \mathrm{SW\text{-}MSA}(\mathrm{W\text{-}MSA}(X^{l-1})) + X^{l-1} \tag{2}$$

式中：l 为 图的层数；X^{l} 为第 l 层的特征；CBS 为卷积模块；W-MSA 为基于窗口的自注意力模块；SW-MSA 为基于滑动窗口的自注意力模块。

1.3　加权双向特征金字塔网络

YOLOv8 模型中的特征融合网络为路径聚合网络（PANet）[23]，通过直接求和操作，在特征金字塔中进行简单双向融合，没有考虑加权的相关性设计。由于不同特征具有不同的分辨率，对最终输出特征的贡献程度有所不同，直接求和操作可能无法有效地融合这些特征。本文采用 BiFPN 特征融合网络，其结构如图 3 所示。

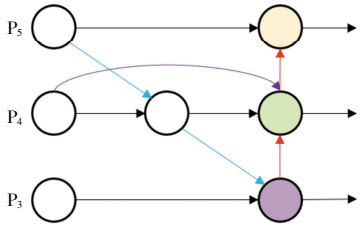


图 3　BiFPN 结构

Fig.3　Structure of BiFPN

BiFPN 为每个特征添加 1 个可学习的权重，在网络训练过程中逐渐学习每个特征的重要性以及不同特征之间的关系，并通过自顶向下和自底向上的方式进行特征双向融合。通过上采样和下采样统一特征分辨率尺度，以进行不同尺度间融合，同时在特征的原

始输入和输出节点之间添加横向连接,以缓解由于网络层数过多而导致的特征信息丢失问题。本文所采用的 BiFPN 网络使用快速归一化融合方法进行加权融合,计算过程如式(3)所示:

$$O = \sum_i \frac{w_i I_i}{\varepsilon + \sum_j w_j} \tag{3}$$

式中:I_i 为输入特征;O 为输出特征;w_i 和 w_j 为可学习的权重;ε 为一个很小的数,以确保分母不为零。

1.4　损失函数

除了改进网络结构,选择适合数据集特点的损失函数也会对模型的检测结果产生影响。在密集堆垛纸箱检测任务中,纸箱之间可能存在相互遮挡、重叠等情况。CIoU 损失函数[24]通过计算预测框与真实框的中心点距以及长宽比评价两者之间的差异性。当预测框与真实框的宽高比值相同时,CIoU 损失函数引入的相对比例的惩罚项无法起到作用。

MPDIoU 损失函数[25]通过计算预测框与真实框左上角点和右下角点之间的最小距离评估边界框相似性,包含纸箱重叠区域、中心点距离、宽度和高度偏差等。在堆垛纸箱检测任务中,宽高比较大的纸箱检测也是一个挑战。因此,本文提出 BoxIoU 损失函数,在 MPDIoU 损失函数的基础上,将边界框的宽高比添加到损失中,BoxIoU 损失函数的计算公式如式(4)—式(7):

$$\mathrm{IoU} = \frac{B_{\mathrm{gt}} \cap B_{\mathrm{prd}}}{B_{\mathrm{gt}} \cup B_{\mathrm{prd}}} \tag{4}$$

$$L_{\mathrm{BoxIoU}} = 1 - \mathrm{IoU} + \frac{d_1^2 + d_2^2}{h^2 + w^2} + \alpha\upsilon \tag{5}$$

$$\alpha = \frac{\upsilon}{(1 - \mathrm{IoU}) + \upsilon} \tag{6}$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w_{\mathrm{gt}}}{h_{\mathrm{gt}}} - \arctan\frac{w_{\mathrm{pre}}}{h_{\mathrm{pre}}}\right)^2 \tag{7}$$

式中:B_{gt} 为真实边界框;B_{prd} 为预测边界框;d_1 为真实边界框与预测边界框左上点之间的欧氏距离;d_2 为真实边界框与预测边界框右下点之间的欧氏距离;h 为图像的高;w 为图像的宽;α 为平衡参数,用于权衡边界框重叠度和形状差异的影响;υ 为修正因子,用于惩罚边界框的位置偏移以及边界框形状差异的程度;w_{gt}、h_{gt}、w_{pre} 和 h_{pre} 分别为真实框及预测框的宽和高。

2　实验结果与分析

2.1　实验数据

本文选择堆叠纸箱数据集(stacked carton dataset,SCD)[26]来证明所提出方法的有效性,该数据集共有 8 401 幅图片,其中训练集 7 401 幅图片,测试集 1 000 幅图片。将训练集进一步划分为训练集与验证集,比例为 9∶1。该数据集具有尺寸变化、不同颜色和图标、遮挡以及大宽高比等特点,有助于评估模型的鲁棒性和泛化能力,为堆垛纸箱检测方法研究提供了重要的基准。

2.2　评价指标

为评估 ST-YOLO 网络性能,本文用浮点运算次数(FLOPs)衡量模型的复杂度,用平均精度均值(mAP)评价模型的准确性以及鲁棒性,其中 mAP50 表示 IoU 阈值为 0.5 时的 mAP,mAP50-95 表示 IoU 阈值从 0.5 到 0.95 的平均 mAP。计算公式如式(8)所示:

$$\mathrm{mAP} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{AP}_i \tag{8}$$

式中:M 为检测类别;AP 为平均精度。

2.3　实验环境及训练参数

本文实验环境以 Pytorch1.10.1+Cuda10.2 为框架,硬件配置为 Intel Core i7-6800K CPU 和 4 块 NVIDIA GeForce GTX 1080Ti GPU,操作系统为 Ubuntu 18.04。

在训练过程中使用 Adam 优化器,初始学习率为 0.01,批次大小(batch size)为 32,训练轮次(Epoch)为 500,耐心值(Patience)设为 50,表示当损失值(Loss)在连续 50 个 Epoch 中均未下降时停止训练,避免过拟合并节省训练时间。

2.4　实验结果分析

2.4.1　与经典模型对比实验结果

为了验证 ST-YOLO 模型的优越性,本文使用SCD 对不同网络的性能进行了测试,并选择了Faster R-CNN、RetinaNet、YOLO 系列等经典 CNN 检测模型进行比较,此外还与最新的检测模型如 DEYOv2[27]、Gold-YOLO[28]等进行了对比,结果如表 1 所示。

表 1　基础模型的实验结果

Tab.1　Experimental results of base models

网络	FLOPs/G	mAP50/%	mAP50-95/%
Mask R-CNN	74.5	98.76	93.68
Faster R-CNN	82.1	98.42	93.19
RetinaNet	41.2	98.75	93.14
YOLOv8	**34.5**	98.62	93.20
DEYOv2	86.1	98.81	93.39
Gold-YOLO	54.6	98.87	93.57
ST-YOLO	42.7	**99.36**	**95.09**

由表 1 可以看出,两阶段检测器的检测性能优越,模型的检测精准率高于一阶段基础检测模型,但其模型复杂度相对较高,导致推理速度较慢。由于 Transformer 检测模型的参数较为复杂,当数据集较小时,模型容易过拟合,在测试集上的泛化能力较差。在一阶段检测器中,YOLOv8 模型的复杂度最低,并且模

型检测性能高于其他一阶段经典模型。但其网络结构简单，导致对重叠目标检测效果并不理想。本文 ST-YOLO 模型通过改进特征提取模块和优化损失函数，在堆垛纸箱检测数据上的准确性表现最好，mAP50 和 mAP50-95 上分别达到了 99.36%和 95.09%，并且参数量较小，计算复杂度相对较低。

2.4.2　与现有方法对比实验结果

为进一步分析本文提出算法的性能，取得广泛认可的比较结果，本文还将 ST-YOLO 模型与箱体检测领域现有方法在 SCD 上进行比较，包括张亚辉[12]、陶磊等[13]和王佳卓[14]等提出的改进模型。与改进模型的性能对比结果如表 2 所示。由表 2 可见，相对于基准模型，改进的 Faster R-CNN 和改进的 Mask R-CNN 在检测精度上有所提升，但增加了模型复杂度。改进的 RetinaNet 只在模型训练阶段添加分割头，推理时计算复杂度与原模型相同。本文提出的模型检测精准率最高，模型复杂度最小。

表 2　改进模型的实验结果

Tab.2　Experimental results of the improved models

网络	浮点运算总次数/10^9	mAP50/%	mAP50-95/%
Faster R-CNN[12]	99.8	98.53	93.31
Mask R-CNN[13]	82.3	98.83	93.75
RetinaNet[14]	56.5	98.82	93.18
ST-YOLO	42.7	99.36	95.09

图 4 显示了几种改进模型和 ST-YOLO 对堆垛纸箱的检测结果。

(a) 改进的 Faster R-CNN

(b) 改进的 Mask R-CNN

(c) 改进的 RetinaNet

(d) ST-YOLO

图 4　改进模型检测结果对比

Fig.4　Comparison of detection results of the improved model

图 4(a)的前 2 幅图中出现比较严重的误检现象，错误地将背景物体识别为目标纸箱，在后 2 幅图中出现比较严重的漏检问题。图 4(b)的 4 幅图中均出现少量漏检的情况。图 4(c)的后 2 幅图中出现个别漏检的情况。上述 3 种方法均是纯 CNN 模型，特征图在经过多层卷积和池化后会丢失部分细节信息，造成检测结果出现误检和漏检的情况。本文所提 ST-YOLO 模型融合了窗口自注意力机制，在提取特征时能保留更多的局部特征，在图 4(d)的检测结果中既没有出现误检的情况，也没有漏检的问题。此外，3 种现有方法预测框的置信度均低于本文所提出的 ST-YOLO 模型预测框的置信度。

2.4.3　消融实验结果

为研究单个改进点对基准模型的影响，本文进行了消融实验，分别将改进点单独添加到 YOLOv8 中，并进行各种改进点之间的组合实验，结果如表 3 所示。在物流场景中的密集纸箱目标检测任务中，将任意一个改进点单独添加到 YOLOv8 网络中都能提升模型的性能。本文还将 3 种优化方法进行了两两组合，用以改进模型。结果显示，组合后的改进点带来的性能提升更为显著。这表明本文所提出的优化方法对基准模型的性能提升是独立的，彼此之间并没有明显的抑制作用。

表 3　消融实验结果

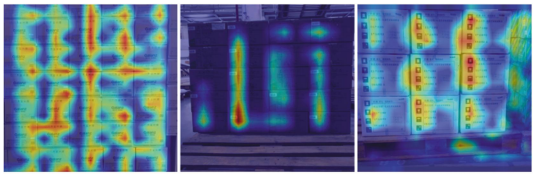
Tab.3　Results of the ablation experiments

CST	BiFPN	BoxIOU	mAP50/%	mAP50-95/%
—	—	—	98.62	92.20
√	—	—	99.04	94.23
—	√	—	98.81	93.71
—	—	√	98.74	93.54
√	√	—	99.25	94.75
√	—	√	99.16	94.62
—	√	√	98.94	94.07
√	√	√	99.36	95.09

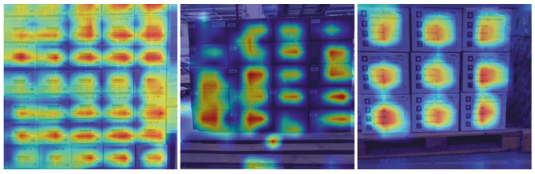
2.4.4　可视化对比

为了验证本文提出的混合 CNN-Transformer 架构的有效性，使用可视化热力图比较 ST-YOLO 与 YOLOv8 对密集纸箱的聚焦能力，如图 5 所示。

由图 5 左侧第 1 列热力图对比可知，相较于 ST-YOLO 模型，YOLOv8 模型的热力图颜色变化的分布不均匀，说明本文提出的模型对于堆垛纸箱特征信息的提取能力更为优异。对比右侧 2 列热力图，YOLOv8 模型由于其 CNN 卷积和池化的特性，其热力图的聚焦区域具有明显的局部性。而本文所提 ST-YOLO 模型通过 CNN-Transformer 融合模块提取堆垛纸箱特征，能够有效地补充卷积过程中遗失的微小细节信息，并保留更完整的全局特征，因此其热力图更能关注到每



（a）YOLOv8



（b）ST-YOLO

图 5　热力图对比

Fig.5　Comparison of heat maps

个纸箱的整体性。总体来说，本文所提模型在密集堆垛纸箱检测任务中的特征提取能力更具优越性。

为验证损失函数改进对于模型性能的影响，使用 BoxIOU 损失函数改进前后的网络模型分别对边界框预测的结果如图 6 所示。

（a）CIoU 损失函数

（b）BoxIOU 损失函数

图 6　不同损失函数检测结果对比

Fig.6　Comparison of detection results by different loss functions

图 6(a)为原始 YOLOv8 网络所用 CIoU 损失函数的检测效果，图 6(b)为 BoxIOU 损失函数对模型改进后的检测效果。对比左侧第 1 列 2 幅检测结果图，在纸箱被遮挡情况下，CIoU 损失函数模型的检测结果中出现漏检和误检情况，而 BoxIOU 损失函数模型的检测结果较为良好。对比第 2 列检测结果，在纸箱具有较大宽高比的情况下，CIoU 损失函数模型检测结果的召回率明显低于 BoxIOU 损失函数模型。综上所述，BoxIOU 损失函数能有效提高原网络模型的精确率和召回率，改善网络检测性能。

3　结　论

本文针对堆垛纸箱尺寸多变、密集堆叠和遮挡等特点，基于 CNN 与Transformer 相结合的思想，提出CST 特征提取模块，减少特征提取过程中部分纹理信息的丢失；利用加权双向特征金字塔网络替换原有的特征融合网络，添加权重学习特征的贡献度及不同特征之间的关系，融合更丰富的深层语义信息；最后设计 BoxIOU 损失函数计算边界框回归损失，考虑边界框宽高值及比例，提高预测边界框质量。实验结果表明：与常见的基础模型以及行业内现有的其他算法相比，本文提出的 ST-YOLO 改进模型在检测精度与鲁棒性方面均有优秀的表现，可以更好地适用于密集堆垛纸箱检测任务。在密集堆垛纸箱 SCD 数据集上，该方法的 mAP50 达到了 99.36%，mAP50-95 达到了 95.09%，具有良好的检测性能以及泛化能力。在未来工作中，将继续深入研究复杂物流仓储场景中的遮挡、密集堆叠目标检测算法的准确性和泛化能力。

参考文献：

[1] Alias C, Nikolaev I, Correa Magallanes E G, et al. An overview of warehousing applications based on cable robot technology in logistics[C]//2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). Singapore: IEEE, 2018: 232-239.

[2] 王成军，韦志文，严晨. 基于机器视觉技术的分拣机器人研究综述[J]. 科学技术与工程，2022，22(3)：893-902.
Wang Chengjun, Wei Zhiwen, Yan Chen. Review on sorting robot based on machine vision technology[J]. Science Technology and Engineering, 2022, 22(3): 893-902(in Chinese).

[3] 朱新龙，崔国华，陈赛旋，等. 视觉引导下机器人拆垛场景识别定位抓取方法[J]. 机床与液压，2023，51(3)：71-77.
Zhu Xinlong, Cui Guohua, Chen Saixuan, et al. Positioning and grasping method of robot destacking scene recognition based on vision guidance[J]. Machine Tool & Hydraulics, 2023, 51(3): 71-77(in Chinese).

[4] Ren S Q, He K M, 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] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2980-2988.

[6] Redmon J, Farhadi A. Yolov3: An incremental improvement [EB/OL]. (2018-04-08)[2024-05-11]. https://arxiv.org/abs/1804.02767.

[7] Bochkovskiy A, Wang C Y, Liao H M. YOLOv4: Optimal speed and accuracy of object detection[PP/OL]. V1.(2020-04-23)[2023-12-15]. https://doi.org/10.48550/arXiv.2004.10934.

[8] Jiang P Y, Ergu D, Liu F Y, et al. A review of yolo algorithm

developments[J]. Procedia Computer Science, 2022, 199: 1066-1073.

[9] Wang C Y, Bochkovskiy A, Liao H M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).Vancouver, BC, Canada: IEEE, 2023: 7464-7475.

[10] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2999-3007.

[11] Arkin E, Yadikar N, Xu X B, et al. A survey: Object detection methods from CNN to transformer[J]. Multimedia Tools and Applications, 2023, 82(14): 21353-21383.

[12] 张亚辉. 基于 Faster R-CNN 目标检测的机器人抓取系统研究[D]. 深圳: 中国科学院大学(中国科学院深圳先进技术研究院), 2019.
Zhang Yahui. Research on robot grasping system based on faster R-CNN object detection technology[D]. Shenzhen: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 2019(in Chinese).

[13] 陶磊, 李天剑, 胡欢. 基于改进 Mask R-CNN 的纸箱堆垛分割与定位方法[J]. 北京信息科技大学学报(自然科学版), 2020, 35(3): 85-88.
Tao Lei, Li Tianjian, Hu Huan. Carton detection and localization method based on the improved Mask R-CNN[J]. Journal of Beijing Information Science & Technology University (Science and Technology Edition), 2020, 35(3): 85-88(in Chinese).

[14] 王佳卓. 堆垛纸箱机器视觉定位算法研究[D]. 武汉: 华中科技大学, 2020.
Wang Jiazhuo. Positioning algorithm of machine vision for stacking cartons[D]. Wuhan: Huazhong University of Science and Technology, 2020(in Chinese).

[15] 李建, 杜建强, 朱彦陈, 等. 基于 Transformer 的目标检测算法综述[J]. 计算机工程与应用, 2023, 59(10): 48-64.
Li Jian, Du Jianqiang, Zhu Yanchen, et al. Survey of Transformer-based object detection algorithms[J]. Computer Engineering and Applications, 2023, 59(10): 48-64(in Chinese).

[16] 李翔, 张涛, 张哲, 等. Transformer 在计算机视觉领域的研究综述[J]. 计算机工程与应用, 2023, 59(1): 1-14.
Li Xiang, Zhang Tao, Zhang Zhe, et al. Survey of Transformer research in computer vision[J]. Computer Engineering and Applications, 2023, 59(1): 1-14(in Chinese).

[17] 石磊, 籍庆余, 陈清威, 等. 视觉 Transformer 在医学图像分析中的应用研究综述[J]. 计算机工程与应用, 2023, 59(8): 41-55.
Shi Lei, Ji Qingyu, Chen Qingwei, et al. Review of research on application of vision Transformer in medical image analysis[J]. Computer Engineering and Applications, 2023, 59(8): 41-55 (in Chinese).

[18] Han K, Wang Y H, Chen H T, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[19] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 213-229.

[20] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[PP/OL]. V2.(2020-10-22) [2021-06-03]. https://doi.org/10.48550/arXiv.2010.11929.

[21] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9992-10002.

[22] Tan M X, Pang R M, Le Q V. EfficientDet: Scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA:IEEE, 2020: 10778-10787.

[23] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 8759-8768.

[24] 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.

[25] Ma S L, Xu Y. MPDIoU: A loss for efficient and accurate bounding box regression[PP/OL]. V1.(2023-07-14)[2023-07-14]. https://doi.org/10.48550/arXiv.2307.07662.

[26] Yang J R, Wu S K, Gou L J, et al. SCD: A stacked carton dataset for detection and segmentation[J]. Sensors, 2022, 22 (10): 3617.

[27] Ouyang H D. DEYOv2:Rank feature with greedy matching for end-to-end object detection [PP/OL]. V2.[2023-06-15]. https://doi.org/10.48550/arXiv.2306.09165.

[28] Wang C C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism[PP/OL]. V5. [2023-09-20]. https://doi.org/10.48550/arXiv.2309.11331.

本文引文格式:

肖志涛, 王宇. 基于混合 CNN-Transformer 的堆垛纸箱检测方法[J]. 天津工业大学学报, 2026, 45(2): 94-100.
Xiao Zhitao, Wang Yu. Stacked carton detection method based on hybrid CNN-Transformer[J]. Journal of Tiangong University, 2026, 45(2): 94-100(in Chinese).

(责任编辑:程晓英)