DOI:10.19651/j.cnki.emt.2416923

基于轻量化 YOLOv7 的带式输送机输送带撕裂检测算法*

安龙辉 王满利 张长森
(河南理工大学物理与电子信息学院 焦作 454000)

摘 要:针对矿井下输送机特殊运行环境下输送带撕裂的检测难题,提出一种线激光辅助下的改进 YOLOv7 轻量化检测算法。首先,针对输送带撕裂以小目标为主,不需要最大的检测层,从而简化网络模型,达到减小模型体积和减少参数量的目的。此外,采用动态非单调 FM 的 Wise-IoU 损失函数,使模型更加关注普通质量的样本,提高模型检测性能。然后,使用 LAMP 剪枝方法,提高模型的计算速度并降低计算复杂度,实现检测网络的轻量化,采用通道知识蒸馏无损提高模型精度,最后使用 TensorRT 加速模型,达到更快的检测速度。实验结果表明,与基准模型相比,改进后模型的参数量和计算量分别减少了 86.8%、49.2%,mAP@0.5:0.95 达到了 62.4%,并且检测速度提升 151.0 fps,模型大小从 71.3 MB 减少到 12.8 MB。经过改进后的模型,提高了对输送带撕裂故障检测的准确性和实时性。

关键词:YOLOv7;目标检测;Wise-IoU;LAMP 剪枝;通道知识蒸馏;TensorRT
中图分类号:TN911.73;TP391.41 **文献标识码**:A **国家标准学科分类代码**:460.4030

Conveyor belt tear detection algorithm of belt conveyor based on lightweight YOLOv7

An Longhui Wang Manli Zhang Changsen
(School of Physics and Electronic Information, Henan Polytechnic University, Jiaozuo 454000, China)

Abstract: To solve the problem of conveyor belt tear detection in the special operating environment of underground mines, a lightweight detection algorithm based on line laser assistance and improved YOLOv7 is proposed. Firstly, considering that the conveyor belt tear is mainly small targets, the largest detection layer is not needed, thus simplifying the network model to reduce the model size and the number of parameters. In addition, the dynamic non-monotonic FM-based Wise-IoU loss function is adopted to make the model pay more attention to common quality samples and improve the model detection performance. Then, the LAMP pruning method is used to improve the model's computing speed and reduce the computing complexity, achieving the lightweight of the detection network. The channel knowledge distillation is used to improve the model accuracy without loss, and finally, the model is accelerated by TensorRT to achieve faster detection speed. The experimental results show that compared with the benchmark model, the improved model has a parameter number and computing volume reduced by 86.8% and 49.2%, respectively, mAP@0.5:0.95 reached 62.4%, and the detection speed was improved by 151.0 fps, the model size was reduced from 71.3 MB to 12.8 MB. After the improvement, the model has improved the accuracy and real-time detection of conveyor belt tear faults.

Keywords: YOLOv7;object detection;Wise-IoU;LAMP pruning;channel knowledge distillation;TensorRT

0 引 言

随着工业自动化的发展,带式输送机被广泛应用于物流、矿山和工厂生产线等领域,用于物料的输送和运输。然而,长时间的运行和高强度的工作环境容易导致输送带撕裂,这可能给生产过程带来严重的安全隐患和经济损失[1]。文献[2-3]指出带式输送机智能化的发展趋势,及时准确地检测和识别输送带撕裂成为了保障生产安全和提高生产效

收稿日期:2024-09-18
* 基金项目:国家自然科学基金(52074305)、河南省科技攻关项目(242102221006)、河南省研究生教育改革与质量提升工程(YJS2024AL026)项目资助

率的重要任务。

由于输送带撕裂的形式多样,且容易受到环境光线、湿度、物料等因素的影响,其检测精度和可靠性是一个难点问题。在以往的研究中,研究者们一直致力于提取更有效的撕裂特征,以提升输送带故障检测的效能。例如,Guo 等[4]提出了一种新的多分类条件循环 GAN 方法来生成和区分输送带损伤表面图像。Yu 等[5]设计了一种双波段红外检测方法,利用损伤区域和正常区域之间的灰度差异来识别撕裂。Wang 等[6]改进了 SSD(single shot multiBox detector,SSD)模型,提出的模型能够实现危险源的在线检测,从而防止输送带纵向撕裂。Yang 等[7]提出了一种基于红外光谱分析的输送带纵向撕裂预警方法,通过红外辐射场的频域特征系数 T 判断输送带是否存在纵向撕裂。Che 等[8]提出了视听融合检测输送带损伤方法。Ren 等[9]提出了一种基于直线激光的传送带纵向撕裂检测方法。

上述方法对于输送带撕裂检测取得一定的效果,但是部分算法设计复杂,需要设置特殊的阈值,不能满足实时性和通用性的需求。随着深度学习技术的不断进步,基于深度学习的目标检测算法展现出了卓越的识别性能。然而,深度学习模型普遍具有复杂的层次结构和庞大的参数集,这导致了模型体积庞大以及参数的冗余性,进而增加了推理所需的时间。这种推理时间的延长使得模型难以在终端设备上实现实时部署,这对于工业领域中及时发现输送带撕裂等紧急情况至关重要。为了在计算和存储资源受限的设备上实现高效的推理,已经开发出了一系列轻量级模型。例如,文献[10]将部分卷积(partial convolution,PConv)集成到 YOLOv7 网络中,在降低模型参数的同时保证检测准确性,同时在 YOLOv7 中添加了简化的 SimSPPF(simplified spatial pyramid pooling-fast,SimSPPF)模块来替换 SPPCSPC(spatial pyramid pooling cross stage partial channel,SPPCSPC),与原始 YOLOv7 相比,改进后的 Lightweight-YOLOv7 参数数量减少了 10.06 M。文献[11]基于 Mobilenetv3 对 YOLOv7 主干网络进行轻量化设计,减少 YOLOv7 模型的计算量、参数量,尽管引入 ECA 机制、使用 AlphaIo U 损失函数提升性能,但是 Faster-YOLOv7 最终准确率仍低于 YOLOv7 模型 3.2%,表明更换 Mobilenetv3 为主干在使模型轻量化的同时,对模型的准确度影响较大。文献[12]在 YOLOv7 颈部,通过设计 Slim-Neck,以缓解多次聚合导致网络参数量大的问题。相较于 YOLOv7,最终 SlimNeck-YOLOv7 模型参数量减少 10.2%,平均检测速度提升 10 FPS。文献[13]采用 GhostNet,实现了轻量级的模型设计和有保证的检测精度。轻量级优化后,模型的参数数、GFLOPs 和权重分布分别下降了 36.6%、40% 和 34.7%。上述轻量化方法尚未使模型达到理想状态,模型的参数减少幅度有限,检测速度的进一步提升仍有待实现。此外,简单地用参数更少的主干网络替换现有网络可能会对模型的性能产生不利影响。鉴于此,有必要使用一种新的轻量化技术,该技术应在尽可能不影响模型精度的情况下,大幅增强模型的轻量化效果,以便在嵌入式和移动平台上高效地执行目标检测任务。YOLOv7[14-15]是一种高度通用的深度学习算法,广泛用于目标检测。该算法的速度和准确性使其成为实时输送带故障检测的理想候选者。

本文提出一种基于改进 YOLOv7 的带式输送机输送带撕裂识别方法。本研究的贡献如下:

1)由于要检测的物体大小很固定,小目标为主,不用很大的参数也可以检测到,深层特征图感受野大适合检测大目标,锚框大小用不到最大的检测层,因此去掉大目标检测层。

2)本文采用 Wise-IoU 损失函数作为网络的边界框损失函数,通过平衡低质量样例和高质量样例的学习权重,以提高检测器的整体性能。

3)为了减少神经网络中不必要的参数和连接,来优化模型的效率和性能,使用 LAMP(layer-adaptive magnitude-based pruning,LAMP)剪枝方法。

4)使用 TensorRT 加速,大幅提高推理速度。

1 YOLOv7

本文采用 YOLOv7 模型,对其进行优化以提高性能。该模型主要由 3 个关键部分构成:输入、主干和头部。

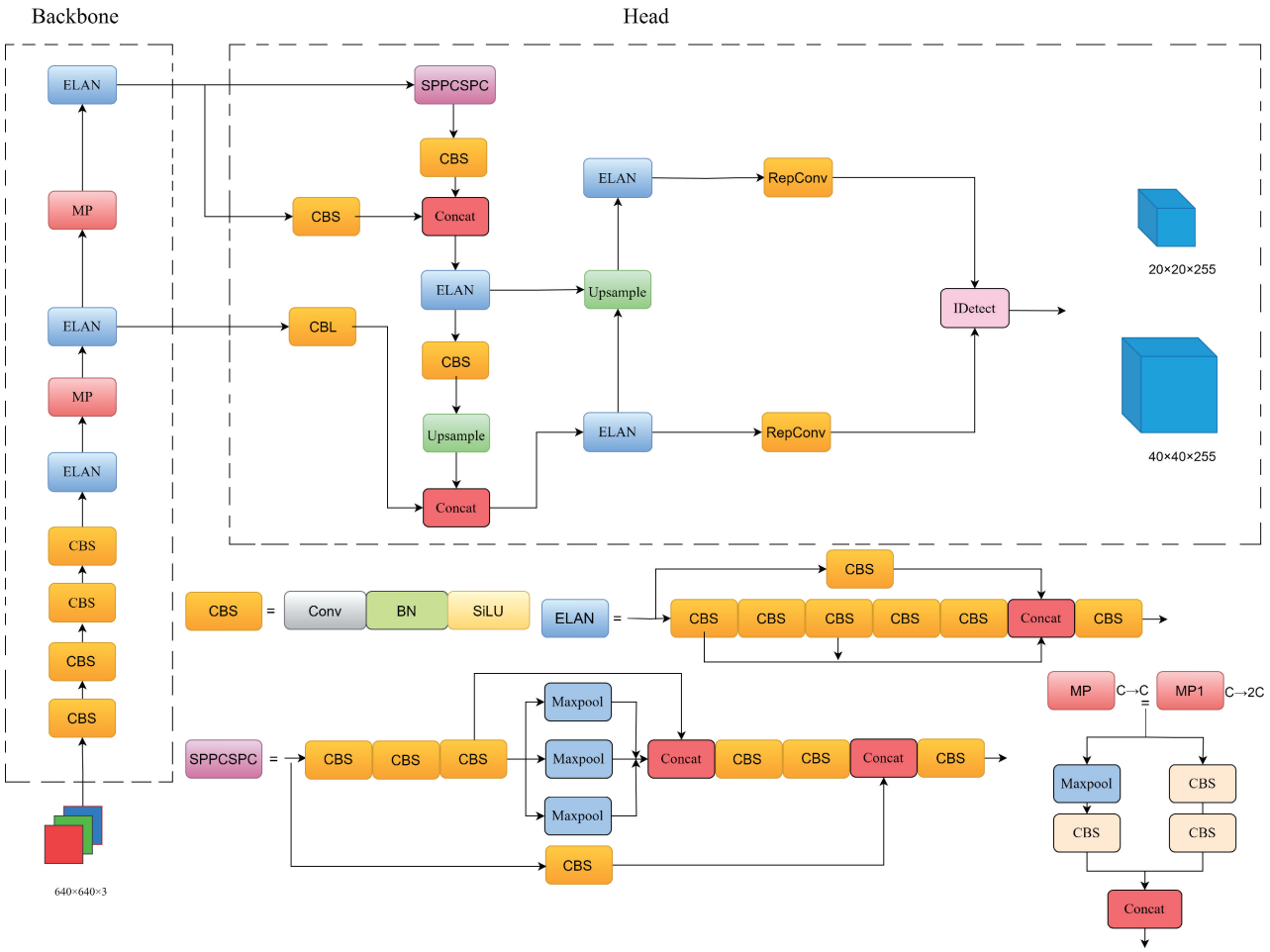
YOLOv7 的主干部分负责特征提取,由 3 个组件构成:CBS(convn-bn-silu)、ELAN(extended latent attention network)和 MP(max-pooling)。CBS 模块通过 3 种不同的卷积核大小和步长设计,使其能够在不同尺度上捕捉特征。ELAN 模块控制网络中最短和最长梯度路径。MP 模块通过两个分支组成,执行最大池化操作以降低特征图的空间维度。

头部组件由 4 个模块组成:SPPCSPC、特征金字塔网络(feature pyramid network,FPN)[16]、重参数化结构(RepConv)以及检测头。SPPCSPC 模块通过最大池化获取不同接收域以适应不同分辨率的图像,FPN 增强了网络对不同特征层的集成能力,RepConv 在推理时将分支的参数重参数化到主分支上,从而减少计算量和内存消耗。将 YOLOv7 去掉大目标检测层,简化后的网络结构如图 1 所示。

2 改进 YOLOv7 的方法

2.1 损失函数的改进

YOLOv7 模型沿用了与 YOLOv5 相同的损失函数设计,该损失函数主要由 3 个核心部分构成:定位损失、目标置信度损失以及类别损失。在坐标损失方面使用了 CIoU 损失函数进行计算;而目标置信度损失和分类损失则采用交叉熵损失进行计算。在训练数据集中,不可避免地存在一些低质量样本。这些样本可能影响模型的泛化能力,导



注：CBS 为卷积＋标准化＋激活函数；Upsample 为上采样模块；ELAN 为高效聚合层网络；MP 为 Maxpool(最大池化)＋CBS；SPPCSPC 为改进后的空间金字塔池化结构；RepConv 为重参数化卷积；Concat 为特征拼接；C→2C 表示通道数变为原来的 2 倍；C→C 表示通道数不变。

图 1　YOLOv7 简化后网络结构

Fig. 1　Simplified network structure of YOLOv7

致预测结果存在偏差。

本文使用了一种基于 IoU 的损失函数，该函数具有动态非单调 FM 特性，称为 Wise-IoUv3[17]。

Wise-IoUv1：根据距离度量构建了距离注意力，得到了具有两层注意力机制的 WIoUv1；其中，第 1 层注意力用于计算样本之间的距离，第 2 层注意力则用于惩罚低质量样本，如式(1)所示。

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{1}$$

其中，L_{WIoUv1} 为 WIoUv1 函数，L_{IoU} 为 IoU 损失函数，R_{WIoU} 为 WIoU 的惩罚项。R_{WIoU} 公式如式(2)所示。

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}\right) \tag{2}$$

其中，x 和 y 是锚框左上角的坐标，x_{gt} 和 y_{gt} 是目标框左上角的坐标，* 表示计算图分离操作，W_g 和 H_g 表示最小边界框的宽和高。

Wise-IoUv3：利用 β 构造了一个非单调聚焦系数并将其应用于 WIoUv1 就得到了具有动态非单调 FM 的 WIoUv3，它提供了一种明智的梯度增益分配策略。这种策略降低了高质量锚盒的竞争力，同时也减少了低质量样例产生的有害梯度。这使得 WIoU 可以专注于普通质量的锚盒，并提高探测器的整体性能，如式(3)～(5)所示。

$$\beta = \frac{L_{IoU}^*}{\overline{L_{IoU}}} \in [0, +\infty) \tag{3}$$

$$L_{WIoUv3} = rL_{WIoUv1} \tag{4}$$

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}} \tag{5}$$

2.2　LAMP 剪枝

本文所使用的剪枝方法是 LAMP[18]，可简述为：首先为卷积网络的每一层的权值进行平方并排序，计算每一层的 LAMP 分数，该分数是权重大小的重新缩放版本，它包含了由修剪引起的模型级别 Frobenius 失真(L2 失真)的分层最小化的解，并且不需要任何超参数调整或繁重的计算，然后全局进行排序和剪枝，将所有连接的权值集合并

排序,将 0 分配给权重最小的连接,直到达到所需的稀疏度阈值即剪枝率,如图 2 所示。

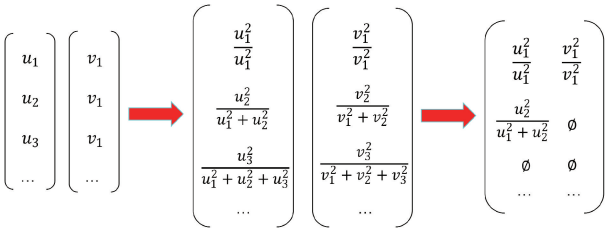


图 2 LAMP 流程

Fig. 2 LAMP process

为了给完全连接层和卷积层的 LAMP 评分一个统一的定义,每个权重张量被展开(或平坦化)为一个一维向量。对于这些展开向量中的每一个,本文假设(不损失一般性)权重按照给定的索引映射按升序排序,即 $|\boldsymbol{W}[u]|\leqslant|\boldsymbol{W}[v]|$ 在 $u<v$ 时成立,其中 $\boldsymbol{W}[u]$ 表示由索引 u 映射的权重张量 $\boldsymbol{W}$。然后将权重张量 $\boldsymbol{W}$ 的第 u 个指标的 LAMP 评分定义为,如式(6)所示。

$$\text{score}(u;\boldsymbol{W}):=\frac{(\boldsymbol{W}[u])^2}{\sum\limits_{v\geqslant u}(\boldsymbol{W}[v])^2} \tag{6}$$

LAMP 分数衡量属于同一层的所有幸存连接中目标连接的相对重要性,其中在同一层中具有较小权值的连接已经被修剪。因此,两个具有相同权重的连接具有不同的 LAMP 分数,这取决于所使用的索引图。一旦计算出 LAMP 分数,全局修剪具有最小 LAMP 分数的连接,直到满足所需的全局稀疏性约束;这个过程相当于用自动选择的分层稀疏度执行 MP,如式(7)所示。

$$(\boldsymbol{W}[u])^2>(\boldsymbol{W}[v])^2\Rightarrow\text{score}(u;\boldsymbol{W})>\text{score}(v;\boldsymbol{W}) \tag{7}$$

剪枝后的模型会损失一部分精度,而重新训练剪枝后的网络,可有效降低模型损失的精度。因为剪掉对网络贡献很小的权重连接,微调后的网络泛化能力得到加强,网络精度会回升甚至可能超过剪枝前的精度。多次剪枝流程如图 3 所示。

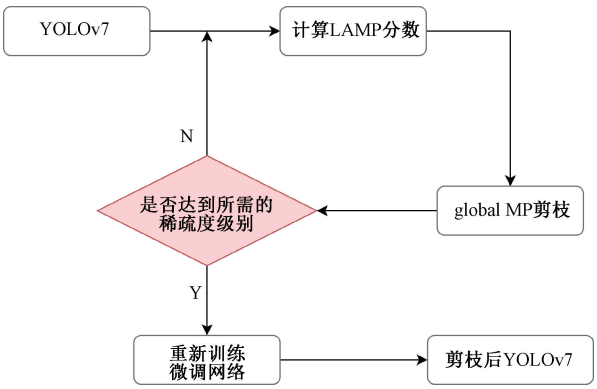


图 3 剪枝流程

Fig. 3 Pruning process

2.3 通道式蒸馏

知识蒸馏(knowledge distillation,KD)已经证明了是一种十分有效的将大模型的知识迁移到小模型的手段。本文使用通道知识蒸馏(channel-wise knowledge distillation,CWD)[19],在激活层的每个通道上提取出软标签,然后再将学生网络和教师网络进行损失计算,可以充分利用不同通道关注不同的特征这一特性,通道分布蒸馏(右上)通过最小化 KL 散度,将学生特征映射的每个渠道与教师网络的特征映射对齐,通道蒸馏方法如图 4 所示。

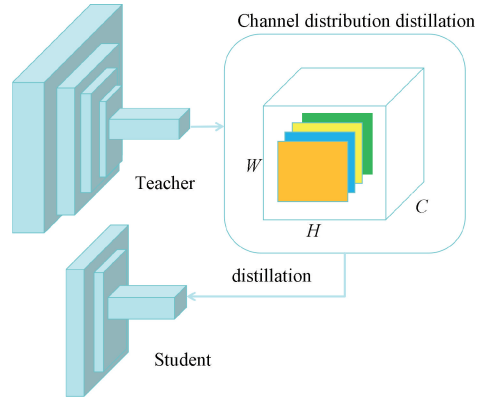


图 4 通道蒸馏方法

Fig. 4 Channel distillation method

使用了 KL 散度来比较 student 网络的 channel 和其对应的 teacher 网络中的 channel 的差异,如式(8)~(10)所示。

$$\varphi(\phi(y^T),\phi(y^S))=\varphi(\phi(y_c^T),\phi(y_c^S)) \tag{8}$$

$$\phi(y_c)=\frac{\exp\left(\frac{y_{c,i}}{T}\right)}{\sum\limits_{i=1}^{W\cdot H}\exp\left(\frac{y_{c,i}}{T}\right)} \tag{9}$$

$$\varphi(y^T,y^S)=\frac{T^2}{C}\sum_{c=1}^{C}\sum_{i=1}^{W\cdot H}\phi(y_{c,i}^T)\cdot\log\left[\frac{\phi(y_{c,i}^T)}{\phi(y_{c,i}^S)}\right] \tag{10}$$

式中:ϕ 是 softmax,c 代表通道 channel,i 代表 channel 中的 location,T 代表知识蒸馏中常用的超参数温度,作用是调节训练过程中的对负样本的的重视程度。如果 student 网络和 teacher 网络对 channel 数不一致,可使用 1×1 的卷积对通道数进行调整。KL 散度是一种非对称的指标。式(10)可以看出,如果 $\phi(y_{c,i}^T)$ 值较大,$\phi(y_{c,i}^S)$ 也应该和前者差不多大时才能最小化 KL 散度的值。另一种情况,当 $\phi(y_{c,i}^T)$ 较小时,整个 KL 散度的值也会较小,训练过程中也就不会对 $\phi(y_{c,i}^S)$ 过于重视,因为此时当前的 location 并不是正样本。使用 KL 散度有利于 student 网络学习到和 teacher 网络一致的前景概率分布,而背景区域则在学习过程中产生的影响较小。

3 实验及结果分析

3.1 实验设备

实验运行环境为 CentOs Linux release 7.6.1810

(Core)操作系统,CPU 使用 Intel(R) Xeon(R) Gold 6226R CPU@2.90 GHz,搭载了 NVIDIA GeForce RTX 3090 显卡,显存容量为 24 GB。使用的编程语言 Python3.8.5,深度学习框架为 Torch 1.13.1、CUDA 11.6,TensorRT 版本为 8.6.1.6。

3.2　数据集及超参数设置

线激光作为强光源可以忽略环境光的影响,在图像上呈现稳定、高质量的激光线。同时,它可以将纵向撕裂的输送带特征转换为稳定的线状特征,提取线性特征可以帮助检测输送带纵向撕裂。激光线的形状大致为一条长直线,图像上的激光条纹是平滑和连续的表明输送带正常。当存在输送带磨损时,激光条纹表现为不均匀、不平滑,表明该处出现输送带磨损。当存在纵向撕裂时,输送带会发生一定程度的变形,激光线会因部分偏移或弯曲而断裂,线条特征相对明显。

实验所用数据集来自某煤矿井下输送带工作时的监控视频,逐帧提取视频每一帧的图像,筛选掉大量相似图像和不合格图像,为了有效地提高算法的鲁棒性,对采集到的数据集进行添加噪声、对比度调整等数据增强方法。数据集经过数据增强后一共包含 2 593 张输送带撕裂图像,最后 LabelImg 标注软件进行手动标注。数据集中使用 tear 标注输送带撕裂故障,wear 标注输送带磨损,如图 5 所示。

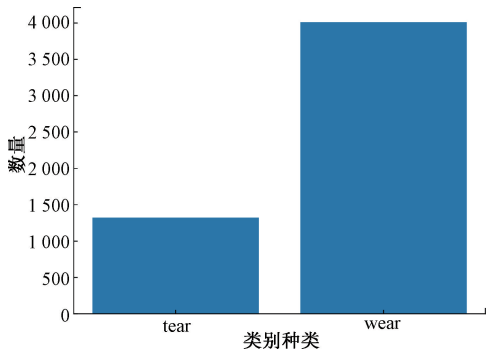


图 5　输送带撕裂和磨损类别数量

Fig. 5　Number of belt tear and wear categories

并将数据集图片按照 8∶1∶1的比例划分为训练集、测试集和验证集。随机选取 2 078 张图像作为训练集,259 张图像作为验证集,剩下 256 张图像作为测试集,部分数据集如图 6 所示。

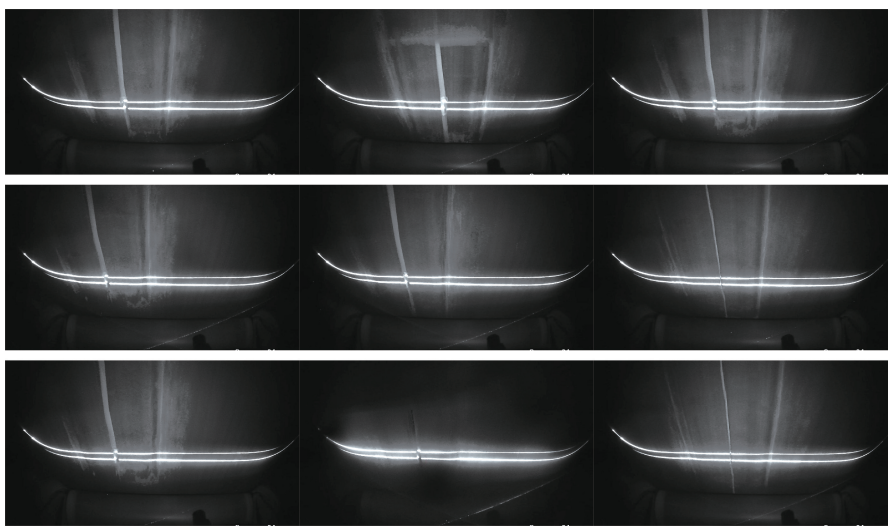


图 6　输送带数据集

Fig. 6　Conveyor belt dataset

为了评估改进模型的性能,本文使用以下评估指标,包括平均精度(mean average precision, mAP)、参数(parameters)、推理速度(fps)、每秒千兆浮点运算(giga floating-point operations per second,GFLOPs)、模型体积(model volume)。

mAP@0.5 和 mAP@0.75 分别是当预测框与标注框的 IOU 大于 0.5 时或大于 0.75 时,就认为这个对象预测正确。mAP@0.5∶0.95 是多个 IOU 阈值下的 mAP,会在区间[0.5,0.95]内,以 0.05 为步长,取 10 个 IOU 阈值,分别计算这 10 个 IOU 阈值下的 mAP,再取平均值。mAP@0.5∶0.95 越大,表示预测框越精准,因为取到了更多 IOU 阈值大的情况。

本文网络训练阶段的超参数设置如表 1 所示,其他训练超参数采用默认值。

3.3　实验结果分析

1)简化 YOLOv7 对模型的影响

将简化后的 YOLOv7 模型与原模型得对比,如图 7 所示,在模型训练初期,学习率较高,Loss 曲线在前 50 轮迅速收敛,简化后的 YOLOv7 模型收敛速度快于原模型,随着迭代的进行,Loss 曲线逐渐变缓并于 250 轮次左右达到

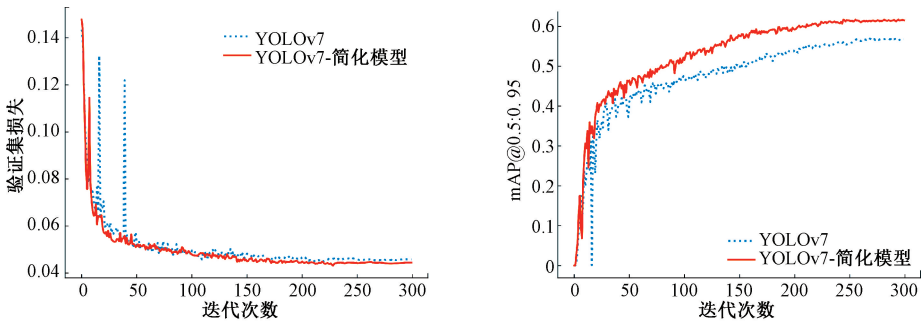
表 1　网络训练超参数

Table 1　Network training hyperparameters

参数名称	参数设置
训练次数(epochs)	300
批次大小(batch-size)	32
初始学习率(learn-rate)	0.01
学习率动量(momentum)	0.937

收敛,Loss 值在 0.045 附近波动,模型达到稳定状态。同时可以看出简化后的模型损失比基准模型更低。图 7 所示简化后得到的 YOLOv7 模型在验证集上 mAP@0.5:0.95 指标高于基准模型,精度得到极大的提升。

使用 XGrad-CAM[20] 可视化方法,对简化后的模型进行可视化分析,网络模型感兴趣的特征信息可以从视觉特征图中看到。为了验证简化后的网络对小目标特征的关注,本文将小目标检测层的输出特征图可视化,比较 YOLOv7



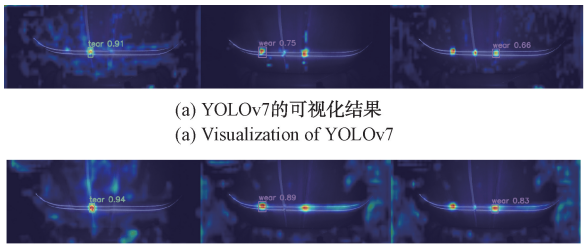
(a) 模型验证集损失对比
(a) Model validation set loss comparison

(b) 模型mAP@0.5:0.95对比
(b) Model mAP@0.5:0.95 comparison

图 7　模型验证集损失和 mAP@0.5:0.95 对比

Fig. 7　Model validation loss compared to mAP@0.5:0.95

简化前后对模型的影响,如图 8 所示。



(a) YOLOv7的可视化结果
(a) Visualization of YOLOv7

(b) 简化后YOLOv7的可视化结果
(b) Visualization of the simplified YOLOv7

图 8　对简化后的模型可视化分析

Fig. 8　Visual analysis of the simplified model

从图 8 中可以看出,相比 YOLOv7 对撕裂特征对比可以看出,简化后的 YOLOV7 模型对撕裂特征的关注更加完整;对磨损特征对比可以看出简化后的模型减少了对不重要的特征的关注,这表明模型更加关注目标的有效特征,抑制了无用或无效的特征。随着网络层数的增加和感受野变大,微观信息丢失,使关于小物体的信息被聚合到一个点上,如果它增加,聚合后的特征就会减少。实验数据以小目标为主,去除大目标检测层,简化网络结构使模型更加关注小目标,提高了网络模型的性能。

2)模型使用 WIoUv3 对模型的影响

不同版本 WIoU 的 mAP@0.5:0.95 对比,如表 2 所示。采用 WIoUv3 替换 CIoU 损失函数,在保证推理速度没有下降的情况下,使模型的 mAP@0.5:0.95 提升了 0.7%,在识别撕裂和磨损目标上分别提高了 0.2% 和 1.1%,虽然使用 WIoUv1 比使用 WIoUv3 在模型识别撕裂目标上高 0.3%,但是在识别磨损目标上却低 1.5%。与简化后 YOLOv7 模型对比,使用 WIoUv3 训练的模型在输送带撕裂和磨损类别的平均精度均取得了提高。

表 2　使用不同 WIoU 版本前后平均精度对比

Table 2　Comparison of average accuracy before and after using different WIoU versions

损失函数	mAP@0.5:0.95/%	AP@0.5:0.95/%(tear)	AP@0.5:0.95/%(wear)
CIoU	62.1	65.1	59.1
WIoUv1	62.2	65.6	58.7
WIoUv2	62.0	65.2	58.7
WIoUv3	62.8	65.3	60.2

3)模型使用 LAMP 剪枝对模型的影响

本文对损失函数改进后的简化 YOLOv7 模型,采用 LAMP 剪枝方法,设置不同剪枝率,(剪枝率即减少的 GFLOPs/剪枝前 GFLOPs),剪枝后,使用与原来相同的训练参数进行微调,通过模型效果选择最佳剪枝率。不同剪枝率进行剪枝并进行微调后的模型参数变化如表 3 所示。

由表 3 可知,通过不同比例的剪枝,YOLOv7 模型的参数量和计算量均有不同程度的下降。模型性能方面,当剪枝率是 20%或 30%时,模型的 mAP@0.5:0.95 提升明显,当剪枝率大于 30%时,平均精度下降较快;随着剪枝率的增加,模型的参数减少,模型体积越小,检测速度越快。

表 3　不同剪枝率下模型对比实验

Table 3　Model comparison experiments under different pruning rates

不同剪枝率/%	Parameters/M	FLOPs/G	mAP@0.5:0.95/%	AP@0.5:0.95/%(tear)	AP@0.5:0.95/%(wear)	Model Size/MB	FPS/fps (bs=8)
0	18.6	106.0	62.8	65.3	60.2	36.2	54.1
10	15.4	95.1	63.3	66.0	60.5	30.0	56.8
20	12.4	84.4	64.1	67.4	60.8	24.2	59.0
30	9.7	74.2	64.3	67.9	60.7	18.9	60.6
40	7.0	63.2	62.7	65.4	59.9	13.7	67.4
50	4.8	52.4	61.5	65.0	58.0	9.5	70.1
60	3.3	42.4	59.9	62.3	57.5	6.5	74.3
70	2.1	31.5	57.1	59.8	54.3	4.2	83.5
80	1.3	20.9	54.4	57.5	51.2	2.8	96.6
90	0.6	10.1	52.0	55.3	48.7	1.5	116.6

当剪枝率是 50%，各卷积层通道变化如图 9 所示，可以看出，通道数较多的卷积层，其通道数量大幅减少，比如第 8 层的最后一个卷积通道数由原来的 1 024 变为 131，表明该剪枝算法有效。相比剪枝前模型，以 mAP@0.5:0.95 降低 1.3%为代价，参数和计算量分别下降了 74.2%和 50.6%，模型大小减少了 26.7 MB，帧率提高 16.0 fps。

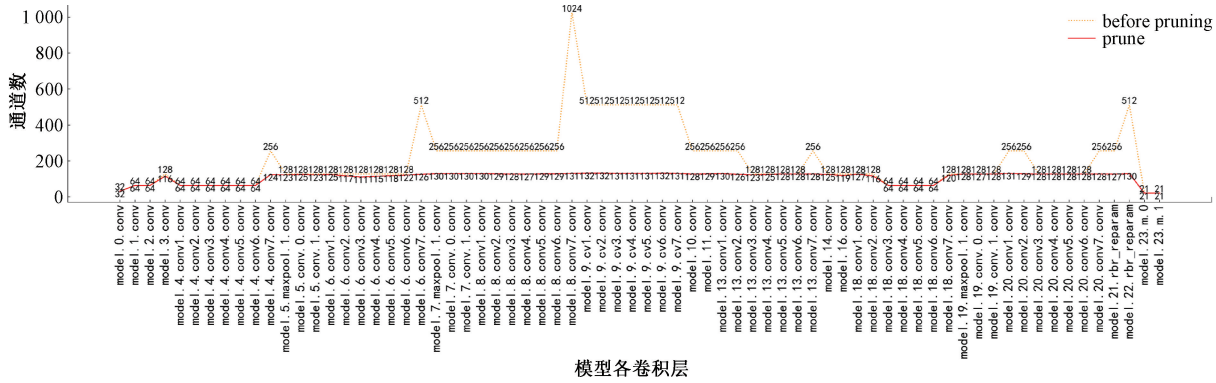


图 9　各卷积层通道变化对比

Fig. 9　Comparison of channel changes in each convolutional layer

4)通道式蒸馏对模型的影响

为了模型轻量化，确保模型满足实时监测的需求，尽可能的压缩模型，但剪去更多的通道会造成精度损失，因此选择剪枝率大于 30%的模型进行知识蒸馏，使用教师模型的知识恢复损失的精度。

选择有更高精度改进后的 YOLOv7 模型作为教师模型分别使用 CWD 和 MGD(masked generative distillation，MGD)[21]知识蒸馏方法，对剪枝后的 YOLOv7 模型作为学生模型进行知识蒸馏，实验结果如表 4 所示。

由表 4 所示，使用 CWD 和 MGD 蒸馏方法使大部分剪枝后的模型平均精度得到不同程度的回升，但 CWD 蒸馏方法对模型的平均精度提高最为明显，证明该方法的有效性。在保证精度损失不大的同时，尽可能的使模型轻量化，因此选择 50%剪枝率的模型，使用 CWD 蒸馏方法使模型平均精度比正常训练微调后模型提高 0.9%，并且输送带撕裂和磨损的检测精度分别提高 0.7%和 1.1%。

5)消融实验

为了验证本文提出的改进后的 YOLOv7 模型在性能方面的可行性，反映算法中不同改进策略对模型检测性能的影响，基于 YOLOv7 进行了消融实验，实验选取 YOLOv7 网络作为对比基准，通过加入不同的模块来验证模型性能。消融实验结果如表 5 所示。

最终得到了改进后的 YOLOv7 模型，相比基准模型，mAP@0.5:0.95 增加 5.7%，参数量和 GFLOPs 分别下降了 86.8%和 49.2%，最后使用 TensorRT 加速，推理速度达到 205.0 fps，推理速度大幅提升，相比加速前提高将近 3 倍，模型体积仅变大 3.3 MB，相比基准模型推理速度提高 151.0 fps，模型体积减少了 58.5 MB。实现了大幅减少模型参数量和计算量，有效的提高模型的平均精度和检测速度。

3.4　与不同检测模型的比较

改进后 YOLOv7 与已有 YOLOv7 轻量化后模型在改

表 4 知识蒸馏对模型的影响

Table 4 Effects of knowledge distillation on the model

模型	Parameters/M	FLOPs/G	mAP@0.5:0.95/%	AP@0.5:0.95/%(tear)	AP@0.5:0.95/%(wear)	Model Size/MB	FPS/fps (bs=8)
teacher	18.6	106.0	62.8	65.3	60.2	36.2	54.1
40%	7.0(−62.4%)	63.2	62.7	65.4	59.9	13.7	67.4
40%-CWD	7.0(−62.4%)	63.2	62.8(+0.1)	65.4(+0)	60.2(+0.3)	13.7	67.4
40%-MGD	7.0(−62.4%)	63.2	62.2(−0.5)	63.9(−1.5)	60.5(+0.6)	13.7	67.4
50%	4.8(−74.2%)	52.4	61.5	65.0	58.0	9.5	70.1
50%-CWD	4.8(−74.2%)	52.4	62.4(+0.9)	65.7(+0.7)	59.1(+1.1)	9.5	70.1
50%-MGD	4.8(−74.2%)	52.4	61.8(+0.3)	64.3(−0.7)	59.2(+1.2)	9.5	70.1
60%	3.3(−82.3%)	42.4	59.9	62.3	57.5	6.5	74.3
60%-CWD	3.3(−82.3%)	42.4	60.8(+0.9)	64.3(+2.0)	57.4(−0.1)	6.5	74.3
60%-MGD	3.3(−82.3%)	42.4	61.4(+1.5)	65.6(+3.3)	57.2(−0.3)	6.5	74.3
70%	2.1(−88.7%)	31.5	57.1	59.8	54.3	4.2	83.5
70%-CWD	2.1(−88.7%)	31.5	59.9(+2.8)	63.3(+3.5)	56.4(+2.1)	4.2	83.5
70%-MGD	2.1(−88.7%)	31.5	58.5(+1.4)	61.6(+1.8)	55.5(+1.2)	4.2	83.5
80%	1.3(−93.0%)	20.9	54.4	57.5	51.2	2.8	97.1
80%-CWD	1.3(−93.0%)	20.9	58.1(+3.7)	62.6(+5.1)	53.5(+2.3)	2.8	97.1
80%-MGD	1.3(−93.0%)	20.9	58.0(+3.6)	62.4(+4.9)	53.5(+2.3)	2.8	97.1
90%	0.6(−96.8%)	10.1	52.0	55.3	48.7	1.5	116.6
90%-CWD	0.6(−96.8%)	10.1	53.1(+1.1)	56.1(+0.8)	50.2(+1.5)	1.5	116.6
90%-MGD	0.6(−96.8%)	10.1	51.6(−0.4)	53.9(−1.4)	49.4(+0.7)	1.5	116.6

表 5 改进后 YOLOv7 消融实验结果

Table 5 Results of improved YOLOv7 ablation experiments

简化网络结构	WIoUv3	Lamp	CWD	TensorRT	Parameters/M	FLOPs/G	mAP@0.5:0.95/%	Model Size/MB	FPS/fps (bs=8)
×	×	×	×	×	36.5	103.2	56.7	71.3	54.0
√	×	×	×	×	18.6	106.0	62.1	36.2	54.8
√	√	×	×	×	18.6	106.0	62.8	36.2	54.1
√	√	√	×	×	4.8	52.4	61.5	9.5	70.1
√	√	√	√	×	4.8	52.4	62.4	9.5	70.1
√	√	√	√	√	—	—	—	12.8	205.0

进之处的区别如表 6 所示。为了评估改进后 YOLOv7 的性能,对比了 YOLOv7-tiny、使用 fasternet[22] 和 efficientViTM0[23] 更换 YOLOv7 主干网络,以及 Lightweight-YOLOv7、Faster-YOLOv7,SlimNeck-YOLOv7 轻量化目标检测方法,比较结果如表 7 所示。

YOLOv7-tiny 是 YOLOv7 的小模型,具有固定简化网络结构的 YOLOv7-tiny 网络的模型体积小,图像处理速度比改进后的 YOLOv7 快 57.7 fps,但是 YOLOv7-tiny 的模型体积和速度,是以牺牲检测精度为代价。YOLOv7 更换 fasternet 和 efficientViTM0 轻量级主干网络后模型参数量和计算量分别减少 33.4%、60.7%和 34.0%、61.9%,检测速度分别提高 38.5 fps 和 31.6 fps,但模型的平均精度下降较多,分别下降 5.8%和 4.8%,对模型影响较大。Lightweight-YOLOv7、Faster-YOLOv7 和 SlimNeck-YOLOv7 相较于原版 YOLOv7 参数量分别减少 29.0%、37.8%、14.0%,计算量分别下降 24.3%、65.5%、12.4%,推理速度分别提升 5.7、39.8、2.2 fps,其中 Faster-YOLOv7 由于更换 Mobilenetv3 为网络主干参数量和计算量下降明显,推理速度达到 93.8 fps。YOLOv7-tiny、fasternet、efficientViTM0、Lightweight-YOLOv7,Faster-YOLOv7 和 SlimNeck-YOLOv7 与改进后的 YOLOv7 网络相比,mAP@0.5:0.95 分别低 11.7%、11.5%、10.5%、10.7%、12.1%和 10.1%。使用 TensorRT 加速后模型推理速度比 YOLOv7-tiny 快 86.7 fps。表明本文所使用的

表 6　与轻量级网络核心实现的对比

Table 6　Comparison with the lightweight network core implementation

模型	轻量化核心实现
YOLOv7-tiny	YOLOv7 的缩放版本，使用 SiLU 作为激活函数
fasternet	FasterNet 模块的设计
efficientViTM0	高效的注意力机制和改进的前馈神经网络(FFN)
Lightweight-YOLOv7	将部分卷积(PConv)集成到主干 ELAN 中，SimSPPF 模块来替换 SPPCSPC
Faster-YOLOv7	基于 Mobilenetv3 对 YOLOv7 主干网络进行轻量化设计
SlimNeck-YOLOv7	通过鬼影混洗卷积(GSConv)与 VoVGSCSP 构建 Slim-Neck
改进后 YOLOv7	简化网络结构，并模型进行剪枝和蒸馏

表 7　与轻量级网络对比实验

Table 7　Comparison experiments with lightweight networks

模型	Parameters/M	FLOPs/G	mAP@0.5：0.95/%	AP@0.5：0.95/%(tear)	AP@0.5：0.95/%(wear)	Model Size/MB	FPS/fps (bs=8)
YOLOv7-tiny	6.0	13.0	50.7	54.9	46.6	11.7	118.3
fasternet	24.3	40.6	50.9	53.7	48.2	48.1	92.5
efficientViTM0	24.1	39.3	51.9	54.6	49.2	48.1	85.6
Lightweight-YOLOv7	25.9	78.1	51.7	53.7	49.7	51.2	59.7
Faster-YOLOv7	22.7	35.6	50.3	53.9	46.8	45.0	93.8
SlimNeck-YOLOv7	31.4	90.4	52.3	55.2	49.4	61.6	56.2
YOLOv7	36.5	103.2	56.7	59.4	54.0	71.3	54.0
改进后 YOLOv7	4.8	52.4	62.4	65.7	59.1	9.5	70.1
改进后 YOLOv7＋TensorRT	—	—	—	—	—	12.8	205.0

方法有效的实现了对模型参数量和计算量的减少，加快推理速度，同时也保证了模型精度的提高。

为了进一步验证本文所提出的改进后的 YOLOv7 模型检测性能，本文与目前主流目标检测模型 YOLOv3[24]、YOLOv4[25]、YOLOv5l、YOLOv8l[26]、YOLOv9[27]、RT-DETR[28]，以及基于 RTDETR 改进后的 RTDETR-IHD[29] 和 YOLOv8m-HTL[30] 在测试集上进行对比实验。如表 8 所示。

表 8　输送带数据集上各模型对比实验

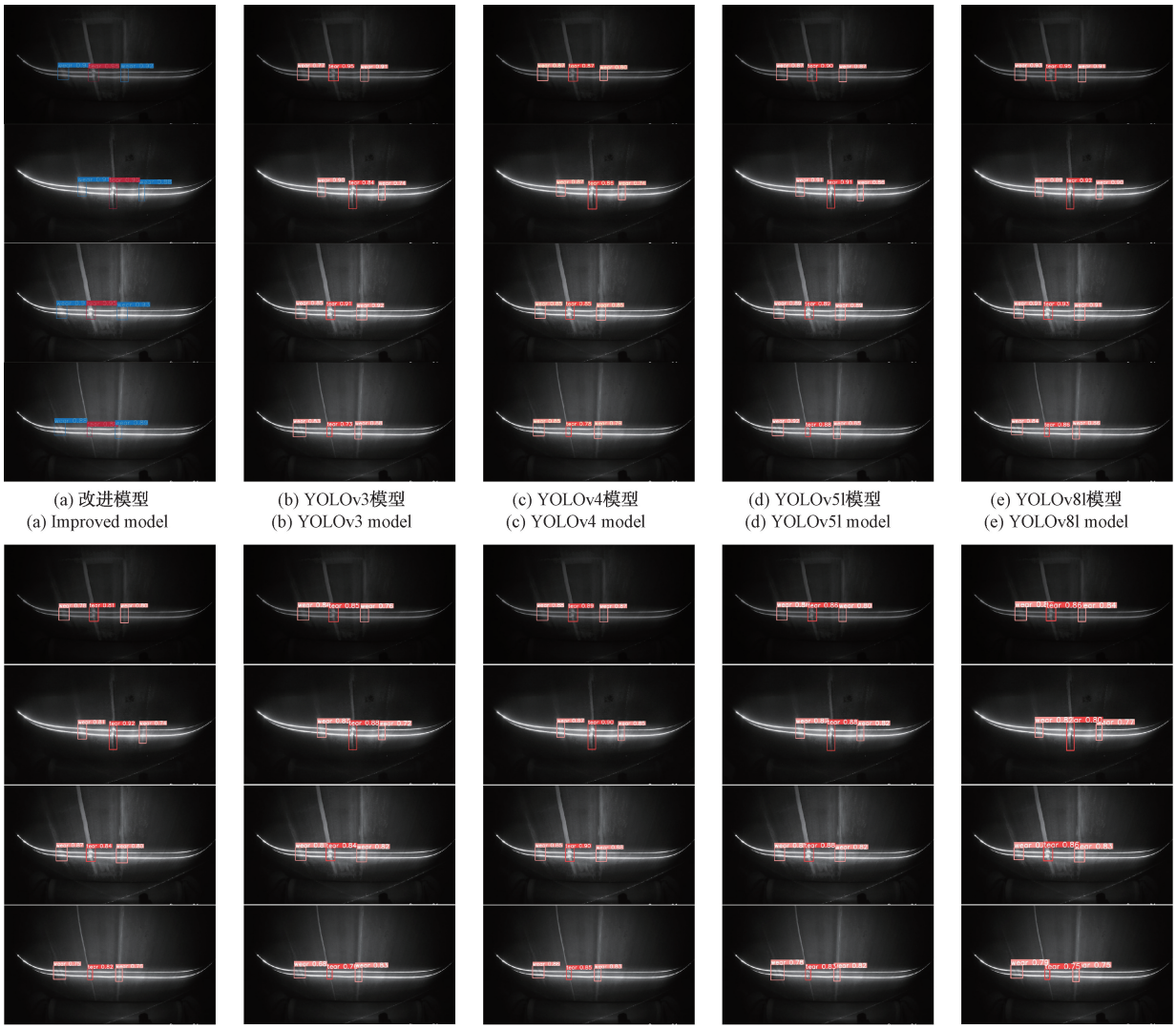
Table 8　Comparison experiments of each model on the conveyor belt dataset

模型	Parameters/M	FLOPs/G	mAP@0.5：0.95/%	AP@0.5：0.95/%(tear)	AP@0.5：0.95/%(wear)	Model Size/MB	FPS/fps (bs=8)
YOLOv3	61.5	154.6	55.8	59.6	52.0	117.7	46.5
YOLOv4	63.9	141.9	52.4	55.5	49.3	244.4	33.1
YOLOv5l	46.1	107.7	58.2	61.9	54.5	88.5	56.2
YOLOv7	36.5	103.2	56.7	59.4	54.0	71.3	54.0
YOLOv8l	43.6	164.8	58.0	60.9	55.0	83.6	48.2
YOLOv9	50.7	236.6	57.1	59.5	54.8	98.0	29.2
RT-DETR	41.9	129.5	60.0	61.7	58.4	82.1	36.5
RTDETR-IHD	16.4	49.3	60.3	62.9	57.6	32.0	50.1
YOLOv8m-HTL	34.0	85.2	56.2	60.0	52.3	65.3	62.4
改进后 YOLOv7	4.8	52.4	62.4	65.7	59.1	9.5	70.1
改进后 YOLOv7＋TensorRT	—	—	—	—	—	12.8	205.0

通过分析表 8 可以看出,本文提出改进后的 YOLOv7 模型在参数量、mAP@0.5:0.95、模型体积,推理速度均有显著的优势,计算量比 RTDETR-IHD 高 3.1 G。在平均精度方面,YOLOv3、YOLOv4、YOLOv5l、YOLOv8l、YOLOv9、RT-DETR、RTDETR-IHD 和 YOLOv8m-HTL 对比改进后的 YOLOv7 分别低 6.6%、10%、4.2%、5.7%、4.4%、5.3%、2.4%、2.1% 和 6.2%,改进后 YOLOv7 对输送带撕裂和磨损的检测精度均高于其他检测模型,YOLOv4 的检测精度最低。改进后的 YOLOv7 模型推理速度最快,YOLOv9 的推理速度最慢。表明改进后的 YOLOv7 模型在保持较小的模型尺寸的同时,实现了较高的检测精度。

3.5　不同模型下输送带检测结果

为进一步展示改进后的 YOLOv7 模型在实际场景下的性能,对输送带撕裂故障图像进检测,检测结果如图 10 所示。



(a) 改进模型　(b) YOLOv3模型　(c) YOLOv4模型　(d) YOLOv5l模型　(e) YOLOv8l模型
(a) Improved model　(b) YOLOv3 model　(c) YOLOv4 model　(d) YOLOv5l model　(e) YOLOv8l model

(f) YOLOv7-tiny模型　(g) YOLOv9模型　(h) RT-DETR模型　(i) RTDETR-IHD模型　(j) YOLOv8m-HTL模型
(f) YOLOv7-tiny model　(g) YOLOv9 model　(h) RT-DETR model　(i) RTDETR-IHD model　(j) YOLOv8m-HTL model

图 10　对比不同算法的检测结果

Fig. 10　Compare the detection results of different algorithms

从图 10 中可以看出,由于在线激光辅助下的图像特征明显,算法都能正确检测,而改进后的 YOLOv7 对撕裂故障的识别的置信度比其他算法都高,在图 10(d)中 YOLOv5l 对左侧磨损目标置信度略高于本文模型 0.4,但对于其他磨损目标高于其他算法。综上所述,提出的改进 YOLOv7 模型可实现对输送带撕裂和磨损精准检测。

4　结　　论

本文基于 YOLOv7 算法进行改进,简化网络模型,去除大目标检测层,使用 WIoUv3 损失函数提高模型锚框的质量,改善检测效果,使用 LAMP 剪枝方法,在不减少模型检测精度情况下,有效减少模型参数和计算量,使用 CWD

蒸馏方法，改善由于模型剪枝造成的精度上的损失，使用 TensorRT，大幅提高推理速度，在自制的输送带撕裂数据集上，该模型的 mAP@0.5:0.95 分别达到了 62.4%，检测速度可达 205.0 fps，模型大小减少 82.0%。改进后的 YOLOv7 模型使计算更加高效的同时，还保持较高的检测精度，满足了检测所需的轻量化网络结构和实时检测要求。然而，尽管该模型在实时性取得一定的进展，但是仍存在性能上的瓶颈，本文还可以改进剪枝方法，在几乎不影响模型检测精度情况下，进一步压缩模型体积，从而寻找到更适合在嵌入式等工业级设备上部署的深度检测网络。

参考文献

[1] 曹虎奇. 煤矿带式输送机撕带断带研究分析[J]. 煤炭科学技术，2015，43(S2)：130-134.
CAO H Q. Study and analysis on on belt breakage and tear in coal mine belt conveyors[J]. Coal Science and Technology，2015，43(S2)：130-134.

[2] 程德强，钱建生，郭星歌，等. 煤矿安全生产视频 AI 识别关键技术研究综述[J]. 煤炭科学技术，2023，51(2)：349-365.
CHENG D Q，QIAN J SH，GUO X G，et al. Review on key technologies of AI recognition for videos in coal mine[J]. Coal Science and Technology，2023，51(2)：349-365.

[3] 王海军，王洪磊. 带式输送机智能化关键技术现状与展望[J]. 煤炭科学技术，2022，50(12)：225-239.
WANG H J，WANG H L. Status and prospect of intelligent key technologies of belt conveyor[J]. Coal Science and Technology，2022，50(12)：225-239.

[4] GUO X Q，LIU X H，KROLCZYK G，et al. Damage detection for conveyor belt surface based on conditional cycle generative adversarial network[J]. Sensors，2022，22(9)：3485.

[5] YU B CH，QIAO T ZH，ZHANG H T，et al. Dual band infrared detection method based on mid-infrared and long infrared vision for conveyor belts longitudinal tear[J]. Measurement，2018，120：140-149.

[6] WANG Y M，MIAO CH Y，MIAO D，et al. Hazard source detection of longitudinal tearing of conveyor belt based on deep learning[J]. PLoS One，2023，18(4)：e0283878.

[7] YANG R Y，QIAO T ZH，PANG Y S，et al. Infrared spectrum analysis method for detection and early warning of longitudinal tear of mine conveyor belt[J]. Measurement，2020，165：107856.

[8] CHE J，QIAO T ZH，YANG Y，et al. Longitudinal tear detection method of conveyor belt based on audio-visual fusion[J]. Measurement，2021，176：109152.

[9] REN X D，WEN J，WANG X H，et al. The study of detection for longitudinal rip of coal conveyor belt based on MATLAB[C]. 2020 Chinese Control And Decision Conference(CCDC)，2020：3139-3144.

[10] ZHOU Y T，PIAO J CH. A lightweight YOLOv7 algorithm for steel surface defect detection[C]. 2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence(PRAI)，2023：285-289.

[11] 唐俊，李敬兆，石晴，等. 基于 Faster-YOLOv7 的带式输送机异物实时检测[J]. 工矿自动化，2023，49(11)：46-52，66.
TANG J，LI J ZH，SHI Q，et al. Real time detection of foreign objects in belt conveyors based on Faster-YOLOv7[J]. Industrial Automation，2023，49(11)：46-52，66.

[12] 冯恒健，韩李涛，张鹏飞，等. 基于改进 YOLOv7 的高效行人检测方法[J]. 计算机应用，2024，44(S1)：290-296.
FENG H J，HAN L T，ZHANG P F，et al. Efficient pedestrian detection method based on improved YOLOv7[J]. Computer Applications，2024，44(S1)：290-296.

[13] WANG Q Y，DONG H B，HUANG H Y. Swin-Transformer-YOLOv5 for lightweight hot-rolled steel strips surface defect detection algorithm[J]. PLoS One，2024，19(1)：e0292082.

[14] 卢开喜，段先华，陶宇诚，等. KThin-YOLOV7：轻量级的焊接件表面缺陷检测[J]. 电子测量技术，2024，47(7)：9-18.
LU K X，DUAN X H，TAO Y CH，et al. KThin-YOLOV7：Lightweight inspection of surface defects on welded parts[J]. Electronic Measurement Technology，2024，47(7)：9-18.

[15] WANG C Y，BOCHKOVSKIY A，LIAO H Y 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)，2023：7464-7475.

[16] 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(CVPR)，2017：936-944.

[17] LYU ZH W，ZHANG X G，HU J D，et al. Visual detection method based on line lasers for the detection of longitudinal tears in conveyor belts[J]. Measurement，2021，183：109800.

[18] LEE J，PARK S，MO S，et al. Layer-adaptive sparsity for the magnitude-based pruning[J]. ArXiv

preprint arXiv:2010.07611,2020.

[19] SHU CH Y, LIU Y F, GAO J F, et al. Channel-wise knowledge distillation for dense prediction[C]. 2021 IEEE/CVF International Conference on Computer Vision(ICCV), 2021: 5291-5300.

[20] FU R G, HU Q Y, DONG X H, et al. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns [J]. ArXiv preprint arXiv: 2008.02312, 2020.

[21] YANG ZH D, LI ZH, SHAO M Q, et al. Masked generative distillation [C]. Computer Vision-ECCV 2022, Cham: Springer Nature Switzerland, 2022: 53-69.

[22] CHEN J, KAO SH H, HE H, et al. Run, don't walk: chasing higher flops for faster neural networks[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2023: 12021-12031.

[23] LIU X Y, PENG H W, ZHENG N X, et al. EfficientViT: Memory efficient vision transformer with cascaded group attention[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2023: 14420-14430.

[24] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv preprint arXiv: 1804.02767,2018.

[25] BOCHKOVSKIY A, WANG CH Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. ArXiv preprints arXiv:2004.10934,2020.

[26] REIS D, KUPEC J, HONG J, et al. Real-time flying object detection with YOLOv8[J]. ArXiv preprint arXiv:2305.09972,2023.

[27] WANG C Y, YEH I H, LIAO H P. YOLOv9: Learning what you want to learn using programmable gradient information [J]. ArXiv preprints arXiv: 2402.13616,2024.

[28] ZHAO Y, LYU W Y, XU SH L, et al. DETRs beat YOLOs on real-time object detection[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965-16974.

[29] ZHU W M, ZHU CH M. Improved safety helmet detection model based on RTDETR: RTDETR-IHD[C]. 2024 7th International Conference on Computer Information Science and Application Technology (CISAT), 2024: 895-899.

[30] WANG ZH Y, YUAN G W, ZHOU H, et al. Foreign-object detection in high-voltage transmission line based on improved YOLOv8m [J]. Applied Sciences, 2023, 13(23): 12775.

作者简介

安龙辉,硕士研究生,主要研究方向为目标检测、图像处理。

E-mail:alh1117@163.com

王满利(通信作者),博士,副教授,主要研究方向为人工智能、图像处理等。

E-mail:wml920@163.com

张长森,博士,教授,主要研究方向为人工智能、工业物联网等。

E-mail:zhangchangsen@hpu.edu.cn