文章编号:1671-6833(2025)05-0069-08

基于 Transformer 多元注意力的钢材表面缺陷视觉检测

韩慧健, 邢怀宇, 张云峰, 张 锐

(山东财经大学 计算机科学与技术学院,山东 济南 250014)

摘 要: 针对钢材表面缺陷尺度不一和现有检测算法多尺度特征处理能力较差、精度不高的问题,提出一种混合采样与多元注意力协同的钢材表面缺陷检测方法。首先,构建高效通道特征提取主干网络模块,在复杂的钢材表面背景下着重提取缺陷特征;其次,提出一种双重注意力协同的特征金字塔,扩大网络感受野,更好地捕获多尺度缺陷特征,提高对小目标的检测性能;最后,设计出一种 Transformer 混合采样策略,动态感知缺陷区域,提高模型的整体检测性能。在 NEU-DET 数据集上进行实验,结果表明:相较于基准算法 DETR,所提改进算法的平均精度均值提高 6.1 百分点,达到 81.4%,提升了模型对钢材表面缺陷检测的精度;此外,检测帧率为 44.2 帧/s,所提算法在检测速度和检测性能之间取得了较好的平衡。

关键词: 缺陷检测; 注意力机制; Transformer; 混合采样; DETR

中图分类号: TP391;TP18 **文献标志码**: A **doi**: 10.13705/j.issn.1671-6833.2025.05.009

工业生产的钢材表面往往会出现裂纹、划痕等缺陷,产生这些缺陷的原因大致有两种:一是由于炼钢连铸坯的原因,坯料上有未消除的裂纹、皮下气泡及金属夹杂物等杂质,从而在钢材表面形成缺陷;二是轧钢原因引起的,主要是加热和冷却制度的影响。这些缺陷不仅影响美观,还会降低钢材的抗腐蚀性,降低产品质量。采用人工目视的检测方法来筛选优良产品是工厂的常用手段,但人工检测效率低,工人的主观判断也使得检测标准无法统一,进而导致检测结果不稳定。

近年来,随着深度学习的发展,基于深度学习的目标检测算法被广泛应用于钢材表面缺陷检测中。具体来说,以 YOLO(you only look once)[1-2] 为代表的一阶段算法能够以较快的速度直接预测出目标位置和类别;而以 Faster R-CNN[3] 为代表的二阶段算法在生成候选框的基础上再回归出目标区域,具有更高的精度。但是这些算法在检测钢材表面缺陷时表现较差,检测率较低。因此,钢材表面缺陷检测仍然具有很大的改进空间。

钢材表面缺陷检测效果差主要是由模型网络本身的局限性以及缺陷特点所导致。为了获得较强的语义信息和较大的感受野,检测网络不断堆叠下采样层,使得缺陷信息在向前传播过程中逐渐丢失,限制了小目标的检测性能。同时,在复杂的钢材表面背景下,缺陷特征易受干扰,不同尺度的钢材表面缺陷检测具有挑战性。此外,大多数模型算法在检测速度和精度中无法得到较好的平衡。

针对以上问题,本文提出了一种基于 Transformer[4] 混合采样与多元注意力协同的钢材表面缺陷检测模型。首先,构建高效通道特征提取网络 ECA-ResNet,在复杂的钢材表面背景下着重提取缺陷特征;其次,为了扩大网络感受野以更好地捕获多尺度缺陷特征,同时捕捉全局和局部缺陷特征信息,提出双重注意力协同特征金字塔 DAFPN;最后,为了在 Transformer 中实现以较小的计算量提取更精细的特征信息,设计出一种 Transformer 混合采样策略,分别通过可变形注意力机制和深度可分离卷积提高空间缺陷特征提取能力,以较小的计算量提取更精细的缺陷特征。

1 相关工作

1.1 钢材表面缺陷检测

随着目标检测的发展,钢材表面缺陷检测领域也得到了实质性的发展。许多研究者开始将目标检

收稿日期:2025-02-25;**修订日期**:2025-04-02

基金项目:国家自然科学基金资助项目(61972227);山东省自然科学基金青年基金资助项目(ZR2023QF161)

作者简介:韩慧健(1971—),男,山东济南人,山东财经大学教授,博士,博士生导师,主要从事认知智能、数据挖掘可视化和区块链金融等研究,E-mail:Hanhuijian@sdufe.edu.cn。

引用本文:韩慧健,邢怀宇,张云峰,等. 基于 Transformer 多元注意力的钢材表面缺陷视觉检测[J]. 郑州大学学报(工学版),2025,46(2):69-76. (HAN H J,XING H Y,ZHANG Y F,et al. Visual detection of steel surface defects based on Transformer multi-attention [J]. Journal of Zhengzhou University(Engineering Science), 2025,46(2):69-76.)

测算法应用于缺陷检测中。Ferguson 等[5]提出了一种使用卷积神经网络和 X 射线相结合进行缺陷检测的解决方案,将基于掩模区域的网络用于缺陷检测,该方法可以在输入图像中同时执行多个缺陷检测和相同缺陷的分割。Fu 等[6]介绍了一种用于钢材表面缺陷检测和分类的多尺度混合卷积神经网络方法,该方法利用了预先训练的模型和迁移学习技术,允许使用有限的样本进行训练,并提高了模型的泛化能力。以上方法在缺陷检测领域中取得了重大突破,但仍然存在检测定位不够准确以及检测速度过慢等不足。近几年,钢材缺陷检测得到了进一步发展,He 等[7]使用预先训练的 ResNet 提取多尺度特征,并使用多级特征融合网络(multi-view feature fusion network,MFN)对不同尺度的特征进行融合,得到的特征包含更多的位置信息,该方法的主要缺点是定位非常粗糙。Liu 等[8]设计了 MSC-DNet 模型,该模型擅长精确定位缺陷,并能够准确检测中等尺度和大尺度缺陷,但对小尺度缺陷的检测能力不足。

1.2 DETR 目标检测

DETR[9]是一种基于 Transformer 的端到端目标检测算法,为目标检测引入了全新思路,它将 CNN 和 Transformer 模型相融合,借助 Transformer 的全局特征关注机制使得 DETR 在全局特征学习方面表现出色。

DETR 在没有使用传统锚框和非极大值抑制的情况下直接输出目标检测结果,极大地简化了目标检测模型的设计和训练过程。但同时 DETR 存在训练速度慢、查询模糊、小目标检测效果差等缺点。为了解决上述问题,研究者们提出了 DETR 的几种变体。Deformable DETR[10]利用多尺度特征提高了注意力机制的效率并加速了训练收敛;DAB-DETR[11]引入了 4D 参考点,并逐层迭代优化预测框;DN-DETR[12]通过添加查询去噪损失来解决 DETR 的缓慢收敛问题;DINO[13]通过应用对比学习和混合查询选择进一步改进 DN-DETR。本文以 DETR 为基础网络,并在此基础上加以改进。

2 本文方法

2.1 方法概述

本文基于改进 DETR 算法的钢材表面缺陷检测网络总体结构如图 1 所示。算法整体结构由高效通道特征提取网络 ECA-ResNet、双重注意力协同特征金字塔模块 DAFPN、Transformer 混合采样模块和预测头构成。

在预处理部分,针对钢材表面缺陷原始图像,首先,通过高效通道特征提取网络提取缺陷图像的特征;其次,通过双重注意力协同特征金字塔模块捕获多尺度缺陷特征,并进行多尺度特征融合,提高小目标检测效果;再次,对特征图进行位置编码,这里主要通过 Transformer 编码器进行混合采样操作,动态感知缺陷区域,丰富空间特征信息,并将采样结果传入 Transformer 解码器,解码器的输出结果将传递

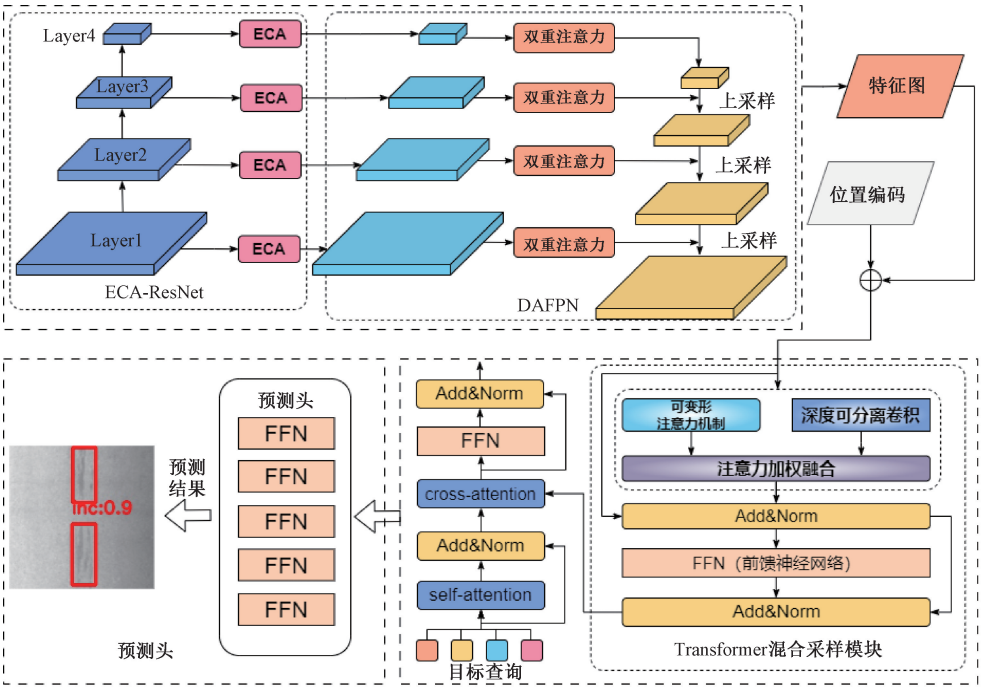

图 1 本文算法的总体结构

Figure 1 General structure of the proposed algorithm

给前馈网络生成检测结果;最后,将检测结果与真实框二分匹配输出预测集,输出缺陷的类别、位置、置信度等信息。以下是该模型算法各个结构的详细介绍。

2.2　高效通道特征提取主干网络模块

在 ResNet 特征提取主干网络中,不同通道之间可能存在一定的相关性和一些冗余信息。受 ECA-Net[14] 注意力模块的启发,在特征提取主干 ResNet 中添加高效通道注意力机制,着重提取图像中的缺陷特征信息(如颜色、纹理和形状)。高效通道特征提取主干网络模块如图 2 所示。

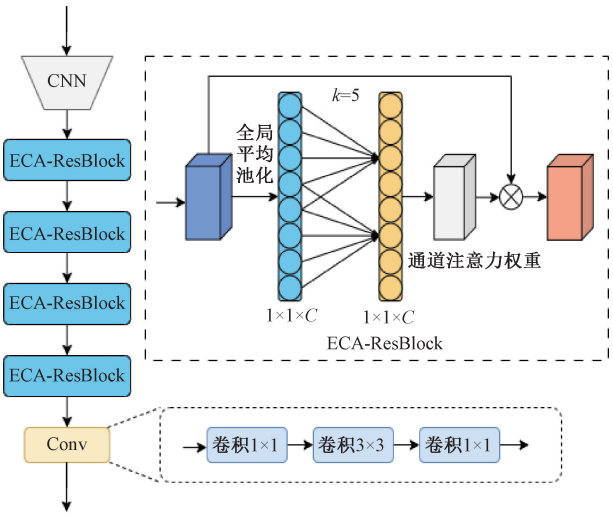

图 2　高效通道特征提取主干网络模块

Figure 2　Efficient channel feature extraction backbone network module

图 2 中,首先,分别对主干 Layer1、Layer2、Layer3、Layer4 这 4 层输出的特征图进行全局平均池化,得到一个全局的特征描述;其次,经过全局平均池化之后,ECA 对每个通道及其邻近通道的特征信息进行分析,进而捕获局部跨通道的交互信息,为了提高效率,ECA 使用快速一维卷积生成通道权重,使得所有通道共享相同的学习参数;最后,不同的输入得到不同的通道权重,实现更高效的通道信息之间的交互。与此同时,在特征提取网络最后加入 Conv 模块,该模块旨在提取更深层次的语义信息,从而提高模型的检测效果。该卷积块最初使用 1×1 卷积减少通道数,随后通过 3×3 卷积缩小特征图大小,最后使用另一个 1×1 卷积增加通道数。通道信息交互公式[14] 如下所示:

$$\boldsymbol{\omega}_i = \delta\left(\sum_{j=1}^{k} \boldsymbol{\omega}_i^j \boldsymbol{y}_i^j\right), y_i^j \in \Omega_i^k。\tag{1}$$

式中:$\boldsymbol{\omega}_i$ 表示通道 $\boldsymbol{y}_i$ 的权重;Ω_i^k 表示 $\boldsymbol{y}_i$ 的 k 个相邻通道的集合;k 为 1D 卷积的卷积核数,因此,k 的个数是由输入的通道数来动态决定的。

2.3　双重注意力协同特征金字塔网络模块

钢材表面缺陷存在类间差异不明显、类内差异较大等突出特点,因此增加了模型检测的难度。针对多尺度缺陷,尤其是小尺度(检测目标在图像中像素小于 32×32)[15] 缺陷检测任务,本文在主干网络与 Transformer 之间引入改进的 FPN 网络结构以实现特征融合。在卷积神经网络(CNN),浅层特征图分辨率高,但语义信息有限且噪声较多;深层特征图分辨率低,语义信息丰富,却对细节特征感知欠佳。为减少小尺寸目标特征丢失,特别选取 $P1$、$P2$、$P3$ 和 $P4$ 这 4 个有效的特征层,并将其输入到改进的 FPN 网络中,双重注意力结构如图 3 所示。由于传统的 FPN 经过 1×1 卷积和上采样操作会丢失部分信息,严重影响尺度缺陷的检测效果,因此在 FPN[16] 横向传输路径上加入双重注意力机制。双重注意力机制由通道注意力和空间注意力模块组成。在通道注意力子模块上,先是进行不同数据维度转换,捕获各通道之间的依赖关系;在空间注意力子模块上,类似于 SE[17] 模块,先进行 7×7 卷积对通道进行降维操作,再通过 7×7 卷积对通道进行升维操作,实现空间维度的信息融合。在整个过程中,输入特征为 $F_1 \in \mathbf{R}^{C\times H\times W}$,经过通道注意力后按元素乘法操作得到 F_2:

$$F_2 = M_C(F_1) \otimes F_1。\tag{2}$$

经过空间注意力后按元素乘法操作得到 F_3:

$$F_3 = M_S(F_2) \otimes F_2。\tag{3}$$

式中:M_C 和 M_S 分别代表通道和空间注意力操作;⊗为按元素乘法操作。

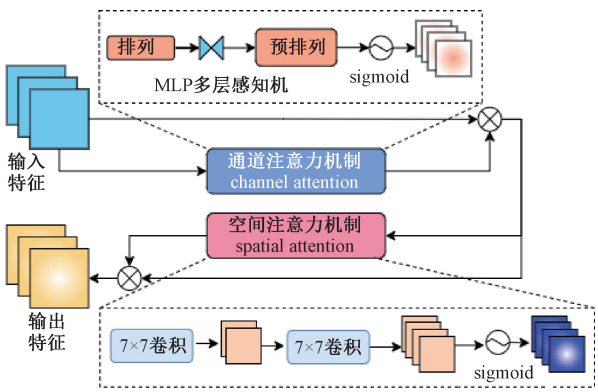

图 3　双重注意力结构

Figure 3　Dual attention structure

2.4　混合采样模块

为更好地提取全局上下文信息和动态感知局部重要特征信息,本文在编码器中设计一种混合采样策略,引入可变形注意力机制与深度可分离卷积模块。可变形注意力机制进行稀疏采样,动态感知缺陷区域,降低模型计算量;深度可分离卷积则提取更

丰富的全局空间特征信息。本文将注意力机制与卷积融合,实现以较小的计算量,提高模型对全局空间信息和局部重要特征的提取能力。Transformer 混合采样模块如图 4 所示。特征图按位置添加可学习的位置编码后送入 Transformer 编码器中的混合采样模块,分别对特征图进行序列化操作和卷积操作,操作后得到的特征图传入到可变形注意力机制与深度可分离卷积这两条分支,分别进行稀疏采样和全局采样。整个可变形注意力模块的输入由 3 部分组成:图像特征 $\boldsymbol{x}$ 、需计算注意力的向量 $\boldsymbol{x}_k$ 以及基准点 $\boldsymbol{p}_k$,其中 $\boldsymbol{x}_k$ 表示在图像特征 $\boldsymbol{x}$ 中对应基准点 $\boldsymbol{p}_k$ 位置的向量。

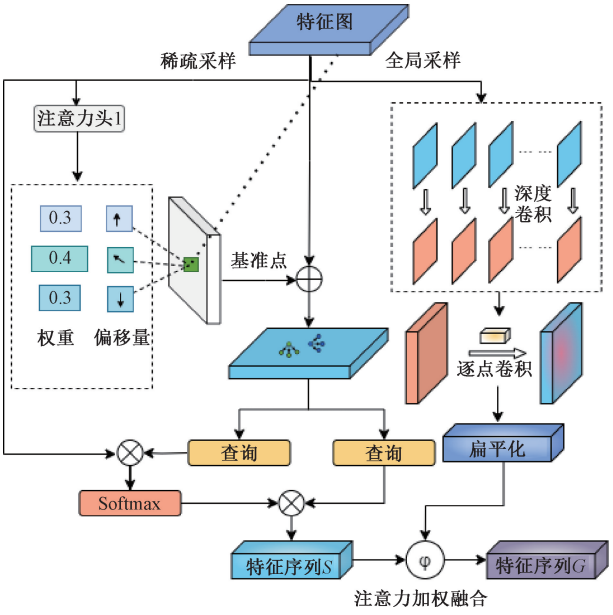

图 4 Transformer 混合采样模块

Figure 4 Transformer hybrid sampling module

对于每一个注意力头,其首先通过对 $\boldsymbol{x}_k$ 进行全连接操作获得 l 个相对于基准点 $\boldsymbol{p}_k$ 的偏移量 $\Delta\boldsymbol{p}_{mkl}$ 以及每个偏移点的权重 A_{mkl} ,然后从图像特征 $\boldsymbol{x}$ 中根据基准点以及偏移量提取对应位置的特征值;其次,利用权重 A_{mkl} 对这 l 个特征值加权求和,获得单个注意力头的输出;最后,将 m 个注意力头的输出通过全连接层合并从而获得最终的输出。图 4 中所示的 L 为 3, m 为 2。以上过程可用如下公式[10]表示:

$$S = \mathrm{DeformAttn}(\boldsymbol{x}_k, \boldsymbol{p}_k, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{l=1}^{L} A_{mkl} \cdot \boldsymbol{W}_v \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{p}_k + \Delta\boldsymbol{p}_{mkl}) \right] 。 \quad (4)$$

式中: $\boldsymbol{W}_m \in \mathbf{R}^{C \times C_v}$ 、$\boldsymbol{W}_v \in \mathbf{R}^{C_v \times C}$ 表示注意力头的可学习权重投影矩阵且 $C_1 = C/m$; l 为预设的偏移点个数; $\Delta\boldsymbol{p}_{mkl}$ 表示每个偏移点相对基准点 $\boldsymbol{p}_k$ 的偏移量; A_{mkl} 表示每个偏移点的权重因子,计算公式如下所示:

$$A_{mkl} = \exp\left(\frac{\boldsymbol{x}_k \boldsymbol{W}_{kl}}{\sqrt{C_1}} \right) 。 \quad (5)$$

在深度可分离卷积模块中,通过将逐点卷积和深度卷积相结合,对通道进行压缩和扩张,在减少模型参数量的同时提取空间缺陷特征 T ,学习全局特征信息。将深度可分离卷积提取的特征 T 进行序列化操作变为 $T1$:

$$T1 = \mathrm{Flatten}(T) 。 \quad (6)$$

采用注意力加权方式生成融合权重[5],自适应地对不同特征进行加权,将 $T1$ 与 S 通过权重进行加权融合,生成新的特征序列 G :

$$G = \mathrm{Softmax}(f_1(T1, S)) \cdot T1 + \mathrm{Softmax}(f_2(T1, S)) \cdot S 。 \quad (7)$$

式中: f_1 、f_2 分别表示点积注意力机制与双线性注意力机制;Softmax 用于注意力权重归一化。

3 实验结果与分析

3.1 数据集

本文算法实验使用的钢材表面缺陷数据集为 NEU-DET 数据集。该数据集共计 1 800 张图像,包括氧化、斑块、裂纹、麻点、夹杂以及划痕。为了提高数据集的质量和数量,增强模型的泛化能力,采用了翻转、随机裁剪、色温变换等数据增强操作,最终生成 15 000 张钢材表面缺陷图像。将数据集随机划分为训练集、验证集、测试集,比例为 7∶1∶2。

3.2 实验环境

为了减小无关因素的影响,本研究进行的所有实验都在一台 NVIDIA GeForce RTX 3090 设备上完成。模型的训练轮次为 300,批处理大小为 4。优化器与 DETR 相同,使用 AdamW 进行优化。学习率为 0.000 1,动量为 0.9,学习率在第 100 轮后衰减为原来的 1/2,在 150 轮后衰减为原来的 1/10。本文使用 DETR 模型的数据增强方法来确保实验之间的一致性。

3.3 公共评价指标

本文的实验结果采用了目标检测领域常用的评价指标,包括精确率 P、召回率 R、平均精度均值 mAP 、参数量以及帧率。在计算平均精度均值时,分别使用了 IoU 阈值为 0.5 和 0.5 至 0.95 的情况,以评估模型在不同 IoU 下的表现。计算公式如下所示:

$$P = \frac{TP}{TP + FP}; \quad (8)$$

$$R = \frac{TP}{TP + FN}; \tag{9}$$

$$AP = \int_0^1 P(R)\,\mathrm{d}R; \tag{10}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}。 \tag{11}$$

式中:*TP* 为真正例,即模型正确检测到的目标;*FP* 为假正例,指模型误检测为目标的负样本;*FN* 为假反例,即实际存在目标但未被模型检测到的样本;*R* 衡量的是在所有实际为正样本的目标中,被模型正确识别出来的比例[18];*AP* 基于 *P-R* 曲线计算得到,*P-R* 曲线通过不同的置信度阈值,描绘了模型的精度和召回率的变化关系,随着阈值的提高,模型的召回率通常会下降,而精度会提高;*mAP* 为各个类别 *AP* 的加权平均值,作为一个综合性指标,能够全面反映模型在不同类别上的表现,从而有效评估模型的总体性能。除此之外,利用参数量来反映模型大小,利用帧率来评价模型的检测速度,帧率表示每秒可以检测到的图片数量,它反映了模型在实际应用中的实时检测能力[19-20]。

3.4　对比实验

为了评估模型的性能,本研究在 NEU-DET 数据集上进行实验,比较了主流目标检测器、同类改进模型以及本文提出的算法的表现,并对结果进行了分析,表 1 为对比实验结果。根据表 1 的结果,Faster R-CNN[3] 在 *mAP* @ 0.5 上的表现为 0.722,然而其帧率仅为 8.9 帧/s。虽然该模型能够保持一定的检测精度,但两阶段检测器存在较多计算冗余,导致其在速度上的表现不理想。相比之下,YOLOv7[21] 和 YOLOv8[22] 的 *mAP* @ 0.5 分别提升至 0.746 和 0.753,在精度上显著优于 Faster R-CNN,同时它们的检测速度分别达到了 42.7 帧/s 和 58.3 帧/s,能够满足实时检测的需求。YOLOv10[23] 的检测精度与 YOLOv8 相当,但由于整体模型结构更加轻便,其检测速度优势尤为突出。帧率高达 61.4 帧/s。

在 DETR 系列算法中,基准模型 DETR 的 *mAP* @ 0.5 达到 0.743。对比通用目标检测算法,针对钢材表面缺陷改进的模型性能表现良好;MSC-DNet[8] 利用上下文增强模块丰富了多尺度信息,平均精度均值 *mAP* @ 0.5 达到了 0.779,但是速度方面表现不佳;DEA_RetinaNet[25] 中添加的通道注意力机制与自适应特征融合模块也增强了特征融合,*mAP* @ 0.5 达到了 0.780,但同时牺牲了速度。

表 1　对比实验结果

Table 1　Comparison of experimental results

算法	*R*	*mAP*@0.5	*mAP*@0.5:0.95	参数量/10^6	帧率/(帧·s^{-1})
Faster R-CNN[3]	0.669	0.722	0.407	41.4	8.9
YOLOv7[21]	0.672	0.746	0.426	36.2	42.7
YOLOv8[22]	0.690	0.753	0.438	43.7	58.3
YOLOv10[23]	0.686	0.757	0.438	**24.4**	**61.4**
DETR[9]	0.682	0.743	0.444	40.2	18.9
Sparse-DETR[24]	0.716	0.789	0.473	40.9	29.2
Deformable-DETR[10]	0.705	0.785	0.464	40.0	31.0
DINO[13]	0.713	0.776	0.455	47.2	32.8
MSC-DNet[8]	0.700	0.779	0.469	45.6	13.6
DEA_RetinaNet[25]	0.704	0.780	0.473	41.3	12.9
本文算法	**0.732**	**0.814**	**0.482**	36.3	44.2

本文所提改进后模型在保证一定速度的前提下提升了检测精度,*mAP* @ 0.5 高达 0.814,帧率为 44.2 帧/s。同时,召回率的提升和参数量的下降可以看出改进后的模型在性能上更具优势,满足工业对钢材表面缺陷检测精度和实时性的要求。图 5 为不同算法的 *P-R* 曲线。由图 5 可知,相较于 DETR 算法,本文算法的 *mAP* 得到了显著提升。

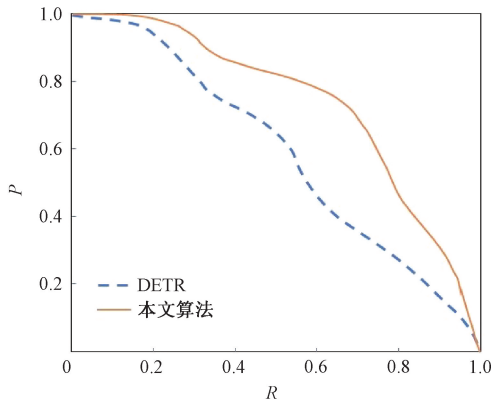

图 5　不同算法的 *P-R* 曲线

Figure 5　*P-R* curves of different algorithms

3.5　消融实验

为了进一步反映各模块在钢材表面缺陷检测的有效性,设计消融实验进行验证。通过在原 DETR 网络模型中依次嵌入对应的改进模型进行训练及测试,使用 *R*、*mAP* @ 0.5、*mAP* @ 0.5∶0.95、参数量和帧率为定量评价指标,实验结果如表 2 所示。

单独使用 ECA 通道注意力时,实验结果和 DETR 相比,*R* 提升了 1.9 百分点,*mAP* @ 0.5 和 *mAP* @ 0.5∶0.95 分别提升了 1.6 百分点和 0.7 百分点,帧率上升为 27.6 帧/s。完整模型与没有采用 ECA 通道注意力相比,*mAP* @ 0.5 提升了 1.2 百分点。

表 2 消融实验

Table 2 Ablation experiment

ECA	DAFPN	HSM	R	mAP@0.5	mAP@0.5:0.95	参数量/10^6	帧率/(帧·s^{-1})
			0.682	0.753	0.444	40.2	18.9
√			0.701	0.769	0.451	40.2	27.6
	√		0.714	0.780	0.468	40.9	21.0
		√	0.691	0.771	0.459	37.8	36.7
√	√		0.714	0.785	0.472	39.6	28.9
√		√	0.706	0.769	0.457	37.7	40.1
	√	√	0.725	0.782	0.475	37.1	36.5
√	√	√	**0.732**	**0.814**	**0.482**	**36.3**	**44.2**

说明 ECA 通道注意力对检测精度并没有带来很大的提升,但有效提升了检测速度。

单独使用 DAFPN 时,实验结果与 DETR 相比,R 提升了 3.2 百分点,mAP@0.5 和 mAP@0.5:0.95 分别提升了 2.7 百分点和 2.4 百分点。完整模型相比没有采用 DAFPN 模块的模型,其 mAP@0.5 和 mAP@0.5:0.95 都有所上升。说明在 FPN 模块中融入空间和通道双重注意力机制可以避免上采样小像素丢失问题,优化系统对钢材表面缺陷的检测性能。

采用 HSM 模块代替 Transformer,相比 DETR,其 mAP@0.5 和 mAP@0.5:0.95 分别提升 1.8 百分点和 1.5 百分点,参数量下降了 2.4×10^6,帧率上升到 36.7 帧/s。对于完整模型来说,使用 HSM 模块,mAP@0.5 和 mAP@0.5:0.95 分别提升了 0.9 百分点和 1.0 百分点,参数量下降到 36.3×10^6,帧率上升到 44.2 帧/s。因此,在 Transformer 的编码器部分使用可变形注意力机制替换自注意力机制,使用稀疏采样代替全局计算,可以有效提升检测性能,同时可以显著减少参数量,提升检测速度。

当使用 ECA-ResNet 为骨干网络,引入 DAFPN 和 HSM 时,召回率、mAP@0.5 和 mAP@0.5:0.95 分别比基准方法高出 5.0 百分点、6.1 百分点和 3.8 百分点,对钢材表面缺陷检测的性能较好,有效提升了钢材表面缺陷检测中小尺度检测精度较低的问题。在检测精度提升的同时,参数量下降了 3.9×10^6,帧率上升了 25.3 帧/s。

3.6 实验结果

为进一步突出模型的对比效果,选取 NEU-DET 数据集中特征不明显的部分图像进行了可视化实验,实验效果对比如图 6 所示。由图 6 可以看出,在斑块检测中,本文算法共检测出 3 个目标,基准算法 DETR 只检测出了 1 个目标;在麻点检测中,本文算法检测出 2 个目标,DETR 仅检测出 1 个目标。可视化的结果验证了召回率的提升。本文提出的算法

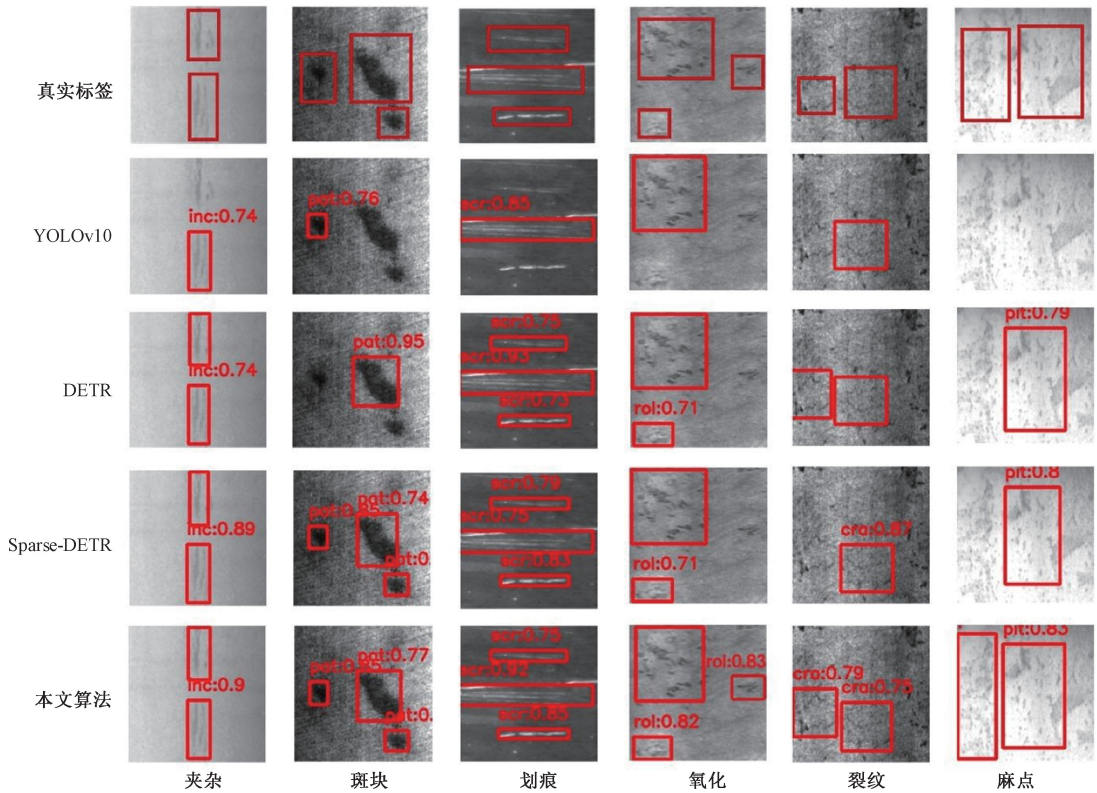

图 6 实验效果对比

Figure 6 Comparison of experimental effect

几乎不存在漏检(检测器未能检测到图像中实际存在的目标)的现象。这是由于改进后特征提取主干网络与多尺度层级特征的融合,有效减少了部分缺陷漏检现象发生,提升了模型的整体检测效果。同时,与其他主流目标检测探测器相比,本文算法在缺陷检测中展现良好的检测性能。

4　结论

本文针对钢材表面缺陷检测问题,提出了一种基于 Transformer 混合采样与多元注意力协同的钢材表面缺陷检测方法。该方法主要包括以下 3 个方面:构建出高效通道特征提取网络 ECA-ResNet,提高复杂背景下感知和提取缺陷特征的能力;提出双重注意力协同特征金字塔网络 DAFPN,解决小尺度缺陷漏检误检问题;设计出混合采样模块代替传统 Transformer,动态感知缺陷区域和全局信息,以较小计算量提取更全面、更精细的特征信息。在 NEU-DET 数据集上的实验结果表明,与其他检测算法相比,所提改进后模型在精度与速度方面均表现良好。相较于原始的 DETR 算法,改进方法在召回率、*mAP* @ 0.5 和 *mAP* @ 0.5:0.95 分别高出 5.0 百分点、6.1 百分点和 3.8 百分点,有效地提高了检测精度;参数量下降了 3.9×10^6,有效地降低了模型复杂度。同时,该算法的帧率为 44.2 帧/s,满足工业检测的实时性要求。此外,改进算法还改善了小目标类别中漏检和误检现象,表现出比其他主流算法更好的性能。所提改进算法可以更好地应用于钢材表面缺陷检测任务,未来可进一步优化算法以提高检测精度和速度,同时保持模型轻量化。

参考文献:

[1] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2025-02-08]. https://doi.org/10.48550/arXiv.1804.02767.

[2] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2025-02-08]. https://doi.org/10.48550/arXiv.2004.10934.

[3] 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.

[4] LIU Z, HU H, LIN Y T, et al. Swin Transformer V2: scaling up capacity and resolution[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 11999-12009.

[5] FERGUSON M K, RONAY A, LEE Y T, et al. Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning[J]. Smart and Sustainable Manufacturing Systems, 2018, 2(1): 137-164.

[6] FU G Z, ZHANG Z G, LE W W, et al. A multi-scale pooling convolutional neural network for accurate steel surface defects classification[J]. Frontiers in Neurorobotics, 2023, 17: 1096083.

[7] HE Y, SONG K C, MENG Q G, et al. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(4): 1493-1504.

[8] LIU R Q, HUANG M, GAO Z M, et al. MSC-DNet: an efficient detector with multi-scale context for defect detection on strip steel surface[J]. Measurement, 2023, 209: 112467.

[9] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformers[C]//Computer Vision-ECCV 2020. Cham: Springer, 2020: 213-229.

[10] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[EB/OL]. (2020-10-08)[2025-02-08]. https://doi.org/10.48550/arXiv.2010.04159.

[11] LIU S L, LI F, ZHANG H, et al. DAB-DETR: dynamic anchor boxes are better queries for DETR[EB/OL]. (2022-06-28)[2025-02-08]. https://doi.org/10.48550/arXiv.2201.12329.

[12] LI F, ZHANG H, LIU S, et al. DN-DETR: accelerate DETR training by introducing query denoising[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2239-2251.

[13] ZHANG H, LI F, LIU S L, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection[EB/OL]. (2022-03-07)[2025-02-08]. https://doi.org/10.48550/arXiv.2203.03605.

[14] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference onComputer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11531-11539.

[15] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[EB/OL]. (2020-05-26)[2025-02-08]. https://doi.org/10.48550/arXiv.2005.12872.

[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.

Piscataway: IEEE, 2017: 936-944.

[17] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[18] 肖进胜, 赵陶, 周剑, 等. 基于上下文增强和特征提纯的小目标检测网络[J]. 计算机研究与发展, 2023, 60(2): 465-474.
XIAO J S, ZHAO T, ZHOU J, et al. Small target detection network based on context augmentation and feature refinement[J]. Journal of Computer Research and Development, 2023, 60(2): 465-474.

[19] 魏明军, 王镆涵, 刘亚志, 等. 基于特征融合和混合注意力的小目标检测[J]. 郑州大学学报(工学版), 2024, 45(3): 72-79.
WEI M J, WANG M H, LIU Y Z, et al. Small object detection based on feature fusion and mixed attention[J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(3): 72-79.

[20] 薛均晓, 武雪程, 王世豪, 等. 基于改进YOLOv4的自然人群口罩佩戴检测方法[J]. 郑州大学学报(工学版), 2022, 43(4): 16-22.
XUE J X, WU X C, WANG S H, et al. A method on mask wearing detection of natural population based on improved YOLOv4[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(4): 16-22.

[21] 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. Piscataway: IEEE, 2023: 7464-7475.

[22] REIS D, KUPEC J, HONG J, et al. Real-time flying object detection with YOLOv8[EB/OL]. (2023-05-17)[2025-02-08]. https://doi.org/10.48550/arXiv.2305.09972.

[23] WANG A, CHEN H, LIU L H, et al. YOLOv10: real-time end-to-end object detection[EB/OL]. (2024-05-23)[2025-02-08]. https://doi.org/10.48550/arXiv.2405.14458.

[24] ROH B, SHIN J, SHIN W, et al. Sparse DETR: efficient end-to-end object detection with learnable sparsity[EB/OL]. (2021-11-29)[2025-02-08]. https://doi.org/10.48550/arXiv.2111.14330.

[25] CHENG X, YU J B. RetinaNet with difference channel attention and adaptively spatial feature fusion for steel surface defect detection[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 70: 2503911.

Visual Detection of Steel Surface Defects Based on Transformer and Multi-attention

HAN Huijian, XING Huaiyu, ZHANG Yunfeng, ZHANG Rui

(School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China)

Abstract: Addressing the challenges posed by the varying scales of steel surface defects and the limited multi-scale feature processing capabilitied and accuracy of existing detection algorithms, in this study a steel surface defect detection method that integrates hybrid sampling and multi-attention collaboration was proposed. Firstly, an efficient channel feature extraction backbone was constructed to emphasize defect feature extraction against the complex background of steel surfaces. Secondly, a dual-attention collaborative feature pyramid was introduced to expand the network's receptive field, thereby enhancing the capture of multi-scale defect features and improving the detection performance for small targets. Finally, a Transformer-based hybrid sampling strategy was designed to dynamically perceive defect regions, thereby boosting the overall detection performance of the model. Experimental comparisons on the NEU-DET dataset revealed that, compared to the baseline DETR algorithm, the improved algorithm achieved a 6.1 percentage point increase in mean average precision, reaching 81.4%, thereby enhancing the model's accuracy in detecting steel surface defects. Additionally, with a detection speed of 44.2 frame/s, the proposed algorithm strikes a commendable balance between detection speed and performance.

Keywords: defect detection; attention mechanism; Transformer; hybrid sampling; DETR