智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

改进RT-DETR的金属表面缺陷检测算法

李冰, 王月, 张易牧, 魏乐涛, 颉卓凡, 叶猛, 翟永杰

引用本文:
李冰, 王月, 张易牧, 等. 改进RT-DETR的金属表面缺陷检测算法[J]. 智能系统学报, 2025, 20(6): 1404-1419.
LI Bing, WANG Yue, ZHANG Yimu, et al. Metal surface defect detection algorithm based on improved RT-DETR algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1404-1419.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202502021

您可能感兴趣的其他文章

双向特征融合与注意力机制结合的目标检测
Target detection based on bidirectional feature fusion and an attention mechanism
智能系统学报. 2021, 16(6): 1098–1105 https://dx.doi.org/10.11992/tis.202012029

融合迁移学习的AlexNet神经网络不锈钢焊缝缺陷分类
Welding defect classification of stainless steel based on AlexNet neural network combined with transfer learning
智能系统学报. 2021, 16(3): 537–543 https://dx.doi.org/10.11992/tis.202005013

一种基于深度学习目标检测的长时目标跟踪算法
A long-term object tracking algorithm based on deep learning and object detection
智能系统学报. 2021, 16(3): 433–441 https://dx.doi.org/10.11992/tis.201910029

基于注意力机制的显著性目标检测方法
Salient object detection method based on the attention mechanism
智能系统学报. 2020, 15(5): 956–963 https://dx.doi.org/10.11992/tis.201903001

基于小样本学习的LCD产品缺陷自动检测方法
An automatic small sample learning-based detection method for LCD product defects
智能系统学报. 2020, 15(3): 560–567 https://dx.doi.org/10.11992/tis.201904020

基于改进的Faster R-CNN高压线缆目标检测方法
Object detection of high-voltage cable based on improved Faster R-CNN
智能系统学报. 2019, 14(4): 627–634 https://dx.doi.org/10.11992/tis.201905026

DOI: 10.11992/tis.202502021
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250926.1031.002

改进 RT-DETR 的金属表面缺陷检测算法

李冰[1,2], 王月[1], 张易牧[1], 魏乐涛[1], 颉卓凡[1], 叶猛[1], 翟永杰[1,2]
(1. 华北电力大学 自动化系, 河北 保定 071003; 2. 保定市电力系统智能机器人感知与控制重点实验室, 河北 保定 071003)

摘 要: 针对金属表面缺陷检测任务中检测目标小、尺度变化大、背景复杂等问题, 提出了一种基于 RT-DETR(real-time detection Transformer) 的改进模型——HAS-DETR(high accurancy for small object-DETR)。HAS-DETR 通过在骨干网络中引入复合差分卷积, 增强对小目标的特征提取能力; 构建双重多尺度特征融合模块, 有效捕获全局语义信息与细节特征, 解决目标尺度变化大的问题; 设计全局多尺度注意力机制, 替代 AIFI(attention-based intra-scale feature interaction) 模块中的多头注意力机制, 提高模型在复杂背景和多尺度目标场景中的鲁棒性和精确度。在金属表面缺陷数据集上, HAS-DETR 在 mAP50 和 mAP50-95 上分别较 RT-DETR 提升了 6.5% 和 4.5%; 在公开 ADPPP 数据集上, mAP50 提升了 2%, mAP50-95 提升了 1.3%。实验结果表明: HAS-DETR 在保持较高检测效率的同时, 有效提升了在复杂背景中对小目标的检测精度, 具有良好的实际应用前景。

关键词: 深度学习; 金属表面缺陷; 小目标; RT-DETR; 特征融合; 注意力机制; 差分卷积; 目标检测
中图分类号: TP183 **文献标志码**: A **文章编号**: 1673−4785(2025)06−1404−16

中文引用格式: 李冰, 王月, 张易牧, 等. 改进 RT-DETR 的金属表面缺陷检测算法 [J]. 智能系统学报, 2025, 20(6): 1404–1419.
英文引用格式: LI Bing, WANG Yue, ZHANG Yimu, et al. Metal surface defect detection algorithm based on improved RT-DETR algorithm[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1404–1419.

Metal surface defect detection algorithm based on improved RT-DETR algorithm

LI Bing[1,2], WANG Yue[1], ZHANG Yimu[1], WEI Letao[1], XIE Zhuofan[1], YE Meng[1], ZHAI Yongjie[1,2]
(1. Department of Automation, North ChinaElectricPower University, Baoding 071003, China; 2. Baoding Key Laboratory of Intelligent Robot Perception and Control in Electric Power System, Baoding 071003, China)

Abstract: To address the challenges posed by small detection targets, significant scale variations, and complex backgrounds in metal surface defect detection tasks, an improved model based on RT-DETR (real-time detection transformer) has been proposed. This model is referred to as HAS-DETR (high accuracy for small object-DETR). HAS-DETR enhances the feature extraction capability for small targets by introducing a multiple differential convolution module (MDConv) into the backbone network. A double multiscale feature fusion module is constructed to effectively capture global semantic information and detailed features, addressing the problem of scale variations. Additionally, a global multiscale attention mechanism has been developed to replace the multihead attention mechanism in the AIFI (attention-based intra-scale feature interaction) module. This modification has been shown to enhance the model's robustness and accuracy in complex backgrounds and multiscale target scenarios. On the metal surface defect dataset, HAS-DETR has been demonstrated to achieve improvements of 6.5% in mAP50 and 4.5% in mAP50-95 compared to RT-DETR. On the public ADPPP dataset, the model demonstrates a 2.0% enhancement in mAP50 and a 1.3% improvement in mAP50-95. Experimental results demonstrate that HAS-DETR significantly enhances the detection accuracy for small objects in complex backgrounds while maintaining high detection efficiency. These findings indicate that HAS-DETR has strong potential for practical industrial applications.

Keywords: deep learning; metal surface defects; small target; RT-DETR; feature fusion; attention mechanism; difference convolution; object detection

金属材料广泛应用于汽车、航空航天、电子设备、家居用品制造等产业, 其表面质量对产品的安全和性能具有重要影响。在现代工业环境中, 由于工作人员操作不当、工业生产环境恶劣

收稿日期: 2025−02−27. **网络出版日期**: 2025−09−26.
基金项目: 国家自然科学基金项目(62373151); 国家自然科学基金联合基金重点支持项目(U21A20486); 中央高校基本科研业务费专项资金项目(2023JC006); 河北省自然科学基金面上项目(F2023502010).
通信作者: 翟永杰. E-mail: zhaiyongjie@ncepu.edu.cn.

©《智能系统学报》编辑部版权所有

等影响因素, 金属表面容易产生各种缺陷, 例如划痕、凹坑、擦伤、白点等。这些缺陷不仅会影响产品美观, 更可能导致产品失效、性能下降以及企业生产效益受损[1]。因此, 金属表面缺陷检测已成为工业领域的研究热点, 而如何进一步提高检测精度则是当前亟待解决的关键问题[2]。

传统的金属表面缺陷检测往往依靠人工目检, 在大规模、高速度的生产中, 存在效率低、劳动强度大、误检率高等问题, 难以满足现代制造业对快速、高精度检测的迫切需求。人工检测需要大量的人力资源, 不仅增加了生产成本, 也易受工作人员经验不足、身体疲劳、情绪不稳定等因素影响造成检测错误, 进而影响最终产品的质量[3]。

随着计算机视觉和深度学习技术的不断发展, 以及工业相机在各个领域的推广, 自动检测技术得到广泛应用, 但金属表面缺陷所具有的检测目标小、尺度变化大等特点依旧给自动化检测技术带来了很大挑战。

目前, 基于深度学习的目标检测算法主要分为以 R-CNN(region-based convolutional neural networks)[4] 系列为代表的两阶段算法和以 YOLO(you only look once)[5] 系列、SSD(single shot multiBox detector)[6] 系列为代表的单阶段算法。在两阶段算法中, 向宽等[7] 将 Faster R-CNN 算法应用到铝材表面缺陷检测上, 在 Faster R-CNN 的基础上应用感兴趣区域校准(region of interest align, ROI align)算法和 K-means 优化锚框, 提升了铝材缺陷检测精度。Wang 等[8] 提出了一种集成多级特征的 Faster R-CNN 算法, 有效解决了金属表面多样化和随机缺陷的检测问题。Fang 等[9] 提出带有 Mix-NMS(mix non-maximun suppression)的注意力级联 R-CNN(attention cascade R-CNN with Mix-NMS, ACRM)算法, 可以对金属表面缺陷进行稳健的分类和定位。

在单阶段算法中, Wang 等[10] 在 DETR(detection Transformer)模型中引入 STF(span-sensitive texture fusion) 模块, 恢复丢失的细节信息并提高检测速度, 有效提升对金属表面缺陷的检测效率。刘浩瀚等[11] 采用多支路并行卷积和空间可分离卷积改进 YOLOv3, 改善了对复杂缺陷的特征提取能力。凌强等[12] 针对金属双极板表面缺陷对比不明显、种类繁多等问题, 加入 NAM 注意力机制和深度可分离卷积, 降低模型复杂度的同时保持检测率。孙卫波等[2] 在 YOLOv7 中采用 PConv(partial convolution)替换 Backbone 部分的卷积, 并引入 SimAM(simple attention mechanism)注意力机制、动态蛇形卷积和 BiFPN(bi-directional feature pyramid network)特征融合模块, 减少计算冗余的同时, 增强卷积神经网络的特征表达能力。Zhang 等[13] 为解决现有算法仅能检测单一金属类别的表面缺陷问题, 基于 YOLOv8 提出 DEFECT-YOLO 模型, 该模型能够对多种金属类型缺陷进行高精度检测。

现有的目标检测算法, 虽然在一定程度上提升了检测精度, 降低了计算冗余, 但仍然存在一些问题: 1)过度依赖锚框设计, 无法灵活处理复杂目标形状或比例, 容易导致漏检或边界框预测不准确; 2)需要进行阈值筛选和 NMS(non-maximum suppression)非极大值抑制, 检测实时性不高; 3)针对背景纹理复杂、检测目标小、目标尺度变化大的金属表面缺陷数据集, 检测精度无法满足实际工业场景需求。

为了解决上述的问题, 本文选择 RT-DETR (real-time detection Transformer)[14] 算法作为基线模型, 提出了一种改进后的 HAS-DETR 模型, 能够对尺寸微小、尺度变化大的金属表面缺陷实现有效检测。本文的主要贡献如下:

1) 设计复合差分卷积模块, 增强对小目标的特征提取能力, 进而提高模型对小目标的检测精度。

2) 构建双重多尺度特征融合模块, 有效捕获全局语义信息和细节特征, 解决因目标尺度变化大而导致的检测困难问题。

3) 用全局多尺度注意力机制替代 AIFI(attention-based intrascale feature interation)模块中的多头注意力机制, 增强模型在复杂背景和多尺度目标检测场景中的鲁棒性和精确度, 能够更好地捕获复杂上下文关系。

1　RT-DETR

RT-DETR 是基于 DETR(detection Transformer)[15] 的实时端到端目标检测框架。DETR 采用 Transformer 架构进行目标检测, 其最大的特点是采用自注意力机制来捕捉图像中的长程依赖关系。DETR 去除了传统目标检测方法中的锚框生成和非极大值抑制步骤, 可以直接从图像中预测对象。虽然 DETR 在理论层面展现出极大的优越性, 但训练和推理速度较慢, 尤其在大规模图像和实时检测任务中, 计算消耗较大[16], 不适合现

实的工业场景。

如图 1 所示，RT-DETR 的网络架构主要分为 3 个部分：骨干网络 (Backbone)、高效混合编码器 (Efficient hybrid encoder) 和解码器 (RTDETRDecoder)。

骨干网络主要用于从输入图像中初步提取多尺度特征，不同尺度的特征图包含了不同层次的图像信息，供后续的编码器和解码器处理。RT-DETR 提供了多种主干网络的选择，其中 RT-DETR-R18 版本通过采用较浅的 ResNet-18 网络，在保持较高检测精度的同时，提高了推理速度。因此，本文选择 RT-DETR-R18 版本作为基线模型。

高效混合编码器由 AIFI 和 CCFM(CNN-based cross-scale feature fusion module) 这两个模块组成，对主干网络输出的多尺度特征进行交互和融合。AIFI 模块通过自注意力机制在同一尺度的特征上进行交互，利用 Transformer 编码器增强图像的高级语义表示，能够捕捉图像中的长程依赖关系。CCFM 通过卷积神经网络将来自不同尺度的特征进行融合，产生丰富的多尺度特征图，以更好地检测不同大小的目标。

RT-DETR 的解码器将编码器输出的特征与查询匹配，生成边界框和类别标签。通过 IoU-aware（intersection over union-aware）查询选择和多次迭代优化，解码器调整预测框的位置和置信度，直到得到最终检测结果。

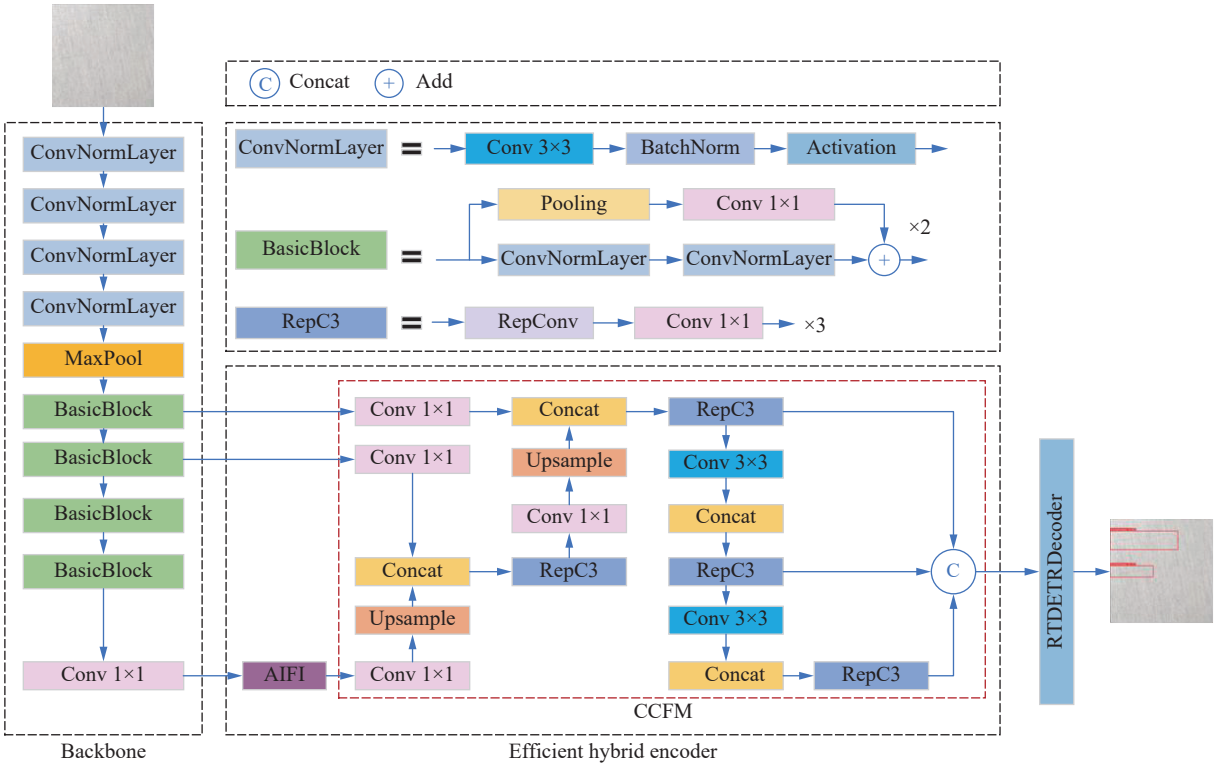


图 1　RT-DETR 网络架构

Fig. 1　Framework of RT-DETR

2　基于 RT-DETR 的改进算法

针对现有模型对金属表面缺陷检测精度低的问题，本文在 RT-DETR-R18 的基础上，提出一种高效的 HAS-DETR（high accurancy for small object-DETR）模型，其总体框架如图 2 所示。本文设计了一种复合差分卷积 (multiple differential convolution, MDC) 替换骨干网络中的标准卷积，增强对小目标的特征提取能力；设计了一种双重多尺度特征融合模块 (double multi-scale feature fusion, DMFF) 替换 RT-DETR 中的 CCFM，利用多尺度特征提取器 (multi-scale feature extractor, MSFE) 解决金属表面缺陷尺度变化大的问题，同时通过小波变换进行上、下采样；设计了一种全局多尺度注意力机制 (global multi-scale attention, GMSA) 替换 AIFI 模块中的多头注意力机制，提取更丰富的全局信息。

2.1　复合差分卷积

在 RT-DETR-R18 模型中，骨干网络采用标准卷积 (vanilla convolution) 进行特征提取和学习，但在背景复杂的小目标检测任务中，标准卷积的优势并不明显[17]。标准卷积层在没有任何约束的情

况下搜索广阔的解空间（甚至是从随机初始化开始），限制了模型的表达能力或建模能力[18]。此外，梯度信息对区分小目标区域至关重要，利用图像边缘信息来辅助目标检测十分有效[19]。为此，本文提出一种复合差分卷积 (multiple differential convolution, MDConv)，用于替代骨干网络中的标准卷积层，提升小目标的细节提取能力。复合差分卷积结构如图 3 所示。

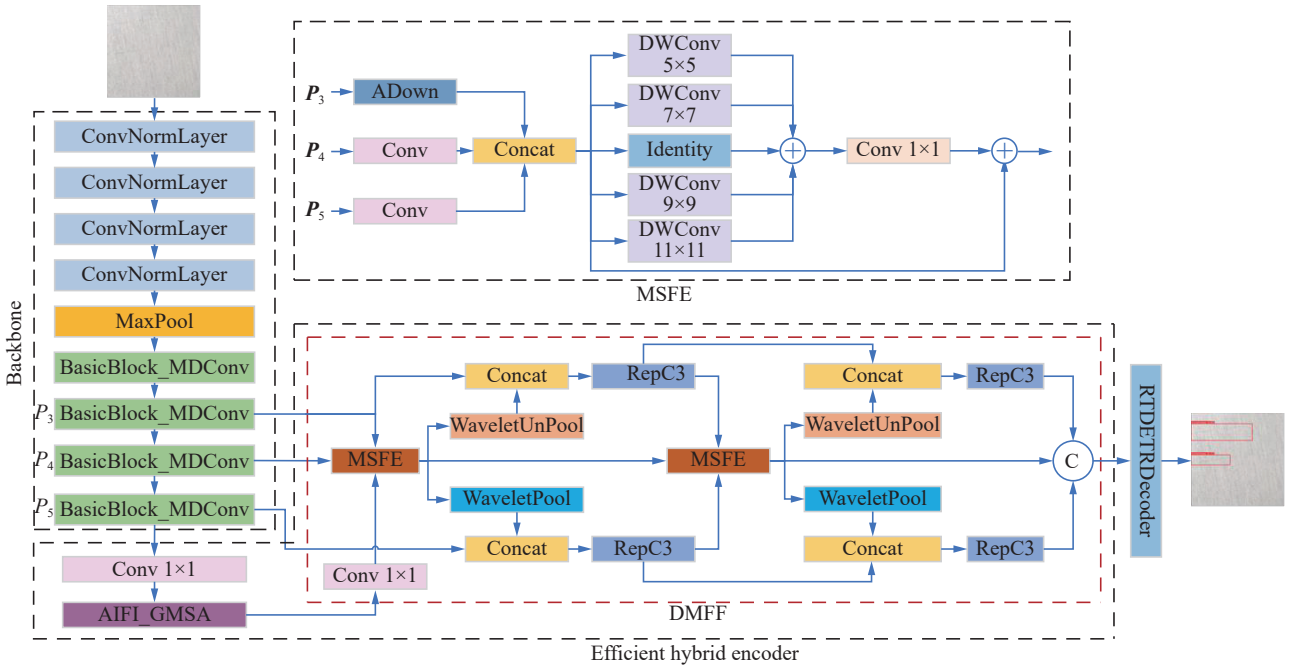


图 2　HAS-DETR 网络架构

Fig. 2　Framework of HAS-DETR

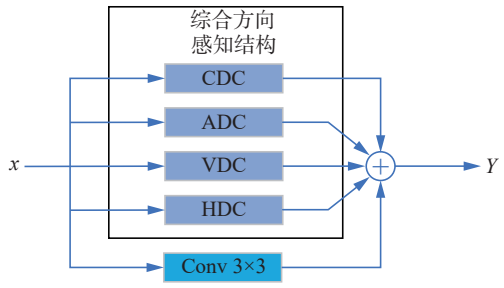


图 3　复合差分卷积

Fig. 3　Structure of MDConv

与传统的卷积操作不同，差分卷积 (differential convolution)[20] 侧重于计算相邻像素之间的差异（即梯度信息），而不仅是对输入图像进行加权求和。这种卷积方法有助于提取图像中的细微变化，尤其是针对目标的边缘、纹理等高频特征。

常见的差分卷积有垂直差分卷积 (vertical difference convolution, VDC)、水平差分卷积 (horizontal difference convolution, HDC)、角差分卷积 (angular difference convolution, ADC)、中心差分卷积 (central difference convolution, CDC) 等。垂直差分卷积、水平差分卷积分别侧重于提取垂直方向上、水平方向上的细节信息；角差分卷积关注像素间的角度变化，在捕捉图像中的斜边、曲线或者其他复杂的结构特征方面表现出色；而中心差分卷积通过计算中心像素和周围像素的差异，能够有效提取图像中小目标的边缘特征。

设计一个综合方向感知结构，包含这 4 个不同方向上的差分卷积，能够同时捕获目标边缘、纹理、对角线特征等多个特征。这对于小目标尤为重要，因为小目标具有细微的局部变化，且仅在某些方向上变化显著。

在复合差分卷积中，通过部署综合方向感知结构和一个标准卷积来进行特征提取，兼顾梯度信息和整体结构信息。这样既能提取细节特征，又能捕捉目标的整体形状和颜色。

由于卷积的可加性[18]，当多个大小相同的二维卷积核对同一输入进行卷积，并在对应位置求和时，可以将这些卷积核直接相加，得到一个等效的单一卷积核，生成与多个卷积核求和后相同的输出。将这个性质应用到复合差分卷积中，在给定输入 x 的情况下：

$$Y = \mathrm{MDConv}(x) = \sum_{i=1}^{5}(x * W_i) = x * \sum_{i=1}^{5} W_i = x * W_{\mathrm{sum}}$$

式中：Y 表示输出，MDConv(·) 表示复合差分卷积运算，W_i 表示垂直差分卷积、水平差分卷积、角差分卷积、中心差分卷积、标准卷积的卷积核，* 表示卷积运算，W_{sum} 表示 5 个并行卷积转换后的卷

积核。

基于该卷积性质，复合差分卷积在提升小目标特征提取能力的同时，又避免了多次卷积带来的参数量和计算量的增加。

2.2　双重多尺度特征融合模块

金属表面缺陷检测目标较小，导致特征难以有效提取，此外，各类缺陷间的尺度变化大也给检测任务带来了很大的困难[21]。为了解决这个问题，本文设计了一种双重多尺度特征融合模块 DMFF，如图 2 所示。DMFF 通过两轮特征融合操作，综合利用多分辨率特征的优势，有效捕捉全局语义信息和细节特征。

在 DMFF 中设计了一种多尺度特征提取器 MSFE，包含两个核心阶段。在第 1 阶段特征融合中，将 Backbone 网络提取的多层特征（浅层 $\boldsymbol{P}_3$、中层 $\boldsymbol{P}_4$、深层 $\boldsymbol{P}_5$）输入至多尺度特征提取器 MSFE 中，生成第一轮融合特征 $\boldsymbol{F}^{(1)}$。接着，将 $\boldsymbol{F}^{(1)}$ 进行小波上采样（WaveletUnPool）[22]，与原始浅层特征 $\boldsymbol{P}_3$ 拼接，交由 RepC3 模块重参数化，得到新的浅层特征 $\boldsymbol{p}'_3$；同时将 $\boldsymbol{F}^{(1)}$ 进行小波下采样（WaveletPool）[22]，与深层特征 $\boldsymbol{P}_5$ 拼接并重参数化，得到新的深层特征 $\boldsymbol{p}'_5$。在第 2 阶段特征融合中，将第 1 轮融合特征 $\boldsymbol{F}^{(1)}$ 视为新的中层特征 $\boldsymbol{p}'_4$，与新生成的浅层特征 $\boldsymbol{p}'_3$、深层特征 $\boldsymbol{p}'_5$ 一起输入至第 2 个 MSFE，重复融合过程，最终输出 3 组包含多尺度信息的特征图 $\boldsymbol{p}''_3$、$\boldsymbol{p}''_4$、$\boldsymbol{p}''_5$，供检测头使用。DMFF 通过两轮自顶向下和自底向上的双向特征融合，使得较浅层的细节信息与较深层的语义信息能够相互补充。

MSFE 结构如图 4 所示，其主要目标是在不同尺度间融合上下文信息与细节信息。由于浅层特征具有较高的分辨率，包含丰富的细节信息，而 ADown（adaptive downsampling）[23] 模块相较普通的卷积操作，能够保留更多的特征信息，因此采用 ADown 操作对浅层特征进行下采样。ADown 结构如图 5 所示。

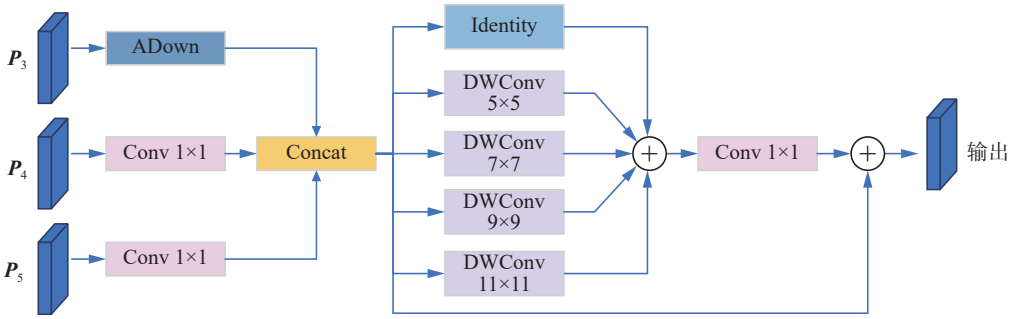


图 4　MSFE 结构

Fig. 4　Structure of MSFE

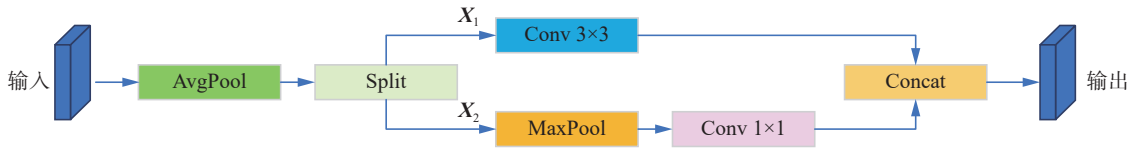


图 5　ADown 结构

Fig. 5　Structure of ADown

ADown 首先将输入进行平均池化，降低特征图的尺寸，再通过 Split 操作，将特征图在通道维度上分成两部分，每个部分包含原始输入的一半特征。

$$\boldsymbol{X}_1, \boldsymbol{X}_2 = \text{split}((\text{AvgPool}(\boldsymbol{X}_{P_3}), \text{ratio} = 0.5)$$

式中：$\boldsymbol{X}_{P_3}$ 表示输入特征，$\boldsymbol{X}_1$、$\boldsymbol{X}_2$ 分别表示包含一半特征的子特征图。

一部分子特征图 $\boldsymbol{X}_1$ 经过 3×3 的卷积进行下采样；另一部分子特征图 $\boldsymbol{X}_2$ 经过最大池化，保留显著特征，再通过 1×1 的卷积调整通道数。两部分特征图在通道维度上进行拼接，形成输出。$\boldsymbol{X}_{\text{AD}}$ 经过 ADown 操作的输出特征。

$$\boldsymbol{X}_{\text{AD}} = \text{Concat}(\text{Conv}_{3\times3}(\boldsymbol{X}_1), \text{Conv}_{1\times1}(\text{MaxPool}(\boldsymbol{X}_2)))$$

在 MSFE 中，浅层特征通过 ADown 操作减少特征图空间维度的同时，保留更多的特征细节信息。中层特征和深层特征采用 1×1 的卷积调整通道数。接着，通过 Concat 操作将这 3 种包含不同信息的特征拼接到一起。

$$\boldsymbol{X}_{\text{C}} = \text{Concat}(\boldsymbol{X}_{\text{A}}, \boldsymbol{X}_{P_4}, \boldsymbol{X}_{P_5})$$

式中：$\boldsymbol{X}_{\text{C}}$ 表示拼接后的特征；$\boldsymbol{X}_{P_4}$、$\boldsymbol{X}_{P_5}$ 分别表示中层特征和深层特征。

拼接后的特征经过一组并行的深度卷积来捕捉跨多个尺度的上下文信息，同时经过一个 Identity 函数进行恒等映射，将经过并行深度卷积和恒等映射后的输出进行相加，再通过 1×1 的卷积

调整通道数。

$$\boldsymbol{X}' = \boldsymbol{X}_{\mathrm{C}} + \sum_{i=1}^{4} \mathrm{DWConv}_{k_i \times k_i}(\boldsymbol{X}_{\mathrm{C}})$$

$$\boldsymbol{X}'' = \mathrm{Conv}_{1\times1}(\boldsymbol{X}')$$

$$k_i = (i+1)\times2+1, i = 1,2,3,4$$

式中：$\boldsymbol{X}'$表示并行深度卷积与恒等映射的求和；$\boldsymbol{X}''$表示调整通道后的输出；k_i表示深度卷积的卷积核大小。

最后，MSFE 通过跳跃连接，将调整后的输出与原始拼接特征再次相加，这样不仅保留了原始特征信息，缓解梯度消失问题，还可以增强多尺度信息的表达能力。$\boldsymbol{X}_{\mathrm{Out}}$表示输出特征图：

$$\boldsymbol{X}_{\mathrm{Out}} = \boldsymbol{X}'' + \boldsymbol{X}_{\mathrm{C}}$$

为进一步减少信息损失并提升特征表达能力，DMFF 引入小波池化进行上采样和小波反池化进行下采样。

小波下采样通过小波变换将输入特征分解为低频分量和高频分量，不仅保留了全局信息，还提取了丰富的局部细节特征。相比传统的最大池化和平均池化方法，小波下采样能够更精准地捕获多尺度特征，尤其对小目标和边缘信息的保留效果更好[24]。

小波上采样利用小波逆变换将分解后的多频信息进行上采样复原，通过重构的高频分量重新注入细节信息，增强还原特征的完整性和表达力。与传统上采样方法相比，小波上采样能够还原全局结构、精准复现特征中的局部细节，同时增强模型对尺度变化的适应性和鲁棒性[22]。

DMFF 的设计通过两轮特征融合策略，充分利用多尺度特征的表达能力，实现高效的上下文信息交互；利用 MSFE 结合多层特征，增强了对不同尺度目标的感知能力；通过引入小波池化和小波反池化，进一步减少信息损失，提升特征保真度和细节传递效果。

2.3　全局多尺度注意力机制

尺度变化大的对象会在不同的图像区域以不同的尺寸出现，全局信息提供了关于目标尺度和位置的更广泛的线索，使得模型能够通过上下文进行尺度推断，因此全局信息对尺度变化大的检测对象非常重要[25]。为此，设计一个全局多尺度注意力机制 (global multi-scale attention, GMSA)，来提取更丰富的全局信息，如图 6 所示。

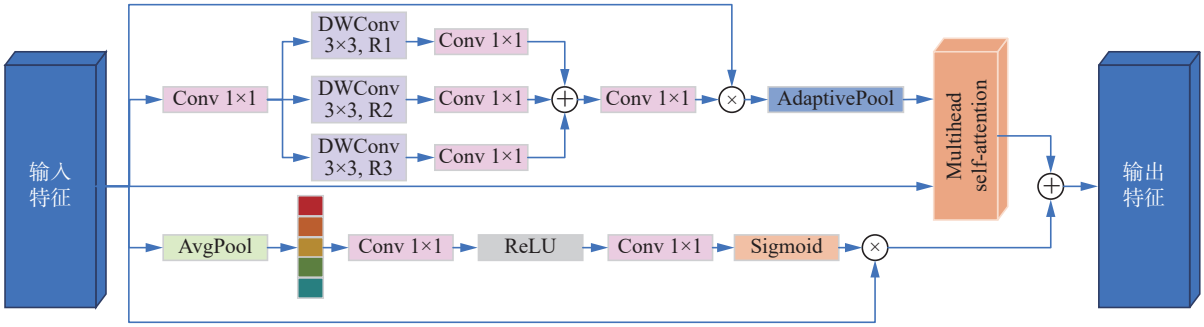


图 6　GMSA 结构

Fig. 6　Structure of GMSA

GMSA 结构如图 6 所示，整体由多尺度上下文分支与通道注意力分支构成，分别从空间和通道两个维度建模全局信息，最终融合得到增强后的特征图。

多尺度分支中，输入特征图 $\boldsymbol{X}\in\mathbf{R}^{C\times H\times W}$ 首先通过 1×1 的卷积调整通道数，得到中间特征；将中间特征分别输入到 3 个并行、扩张率不同的深度可分离空洞卷积中提取多尺度上下文信息。3 个尺度分支分别经过 1×1 的卷积恢复原始通道维度，并进行求和操作，生成一个融合了多尺度信息的特征图。融合后的特征图通过一个 1×1 卷积进行进一步处理，并与原始输入特征 $\boldsymbol{X}$ 进行逐元素相乘，实现特征增强，得到多尺度融合增强特征图 $\boldsymbol{X}_{\mathrm{O}}$。最后，$\boldsymbol{X}_{\mathrm{O}}$ 通过自适应平均池化层，将融合后的多尺度特征图压缩为固定大小的输出 $\boldsymbol{F}$，供后续注意力机制使用。

$$\boldsymbol{X}_i = \mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}^{R_i}(\mathrm{Conv}_{1\times1}(\boldsymbol{X})))$$

$$\boldsymbol{X}_{\mathrm{O}} = \left(\mathrm{Conv}\left(\sum_{i=1}^{3}\boldsymbol{X}_i\right)\right)\otimes\boldsymbol{X}$$

$$\boldsymbol{F} = \mathrm{AdaptivePool}(\boldsymbol{X}_{\mathrm{O}})$$

式中：$\boldsymbol{X}_i$表示深度可分离空洞卷积经过调整通道数后的输出；R_i表示深度可分离空洞卷积的扩张率，i=1, 2, 3, R_1=(1, 3, 5)、R_2=(3, 5, 7)、R_3=(5, 7, 9)；⊗表示逐元素相乘。

基于多尺度增强的特征 $\boldsymbol{F}$ 包含丰富的多尺度上下文信息，在计算自注意力时可替换输入 $\boldsymbol{X}$。

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = (\boldsymbol{X}\boldsymbol{W}^q,\boldsymbol{F}\boldsymbol{W}^k,\boldsymbol{F}\boldsymbol{W}^v)$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示多头注意力中的查询张量、键张量、值张量；$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$分别表示用于生成查

询张量、键张量、值张量的线性变换的权重矩阵。

计算查询 $\boldsymbol{Q}$ 与键 $\boldsymbol{K}$ 的点积以生成注意力权重，对点积结果进行 Softmax 归一化，得到注意力分布：

$$\boldsymbol{A}=\mathrm{Soft\,max}\left(\frac{\boldsymbol{Q}\times\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)$$

式中：d_k 为键张量 $\boldsymbol{K}$ 的通道维度。

使用注意力分布加权值张量 $\boldsymbol{V}$，得到注意力输出：

$$\boldsymbol{X}_{\mathrm{M}}=\boldsymbol{A}\cdot\boldsymbol{V}$$

该过程通过全局多头注意力建模长距离依赖，提升了模型对复杂缺陷上下文的理解能力。

在通道分支中，首先对输入特征图 $\boldsymbol{X}$ 进行自适应平均池化，压缩空间维度，得到每个通道的全局响应向量。该向量通过一个 1×1 的卷积降低通道数，以便进行计算，再经过 ReLU 激活函数来增强非线性表达。随后，通过另一个 1×1 的卷积恢复通道数，并经 Sigmoid 激活函数进行归一化至 [0, 1] 范围内，得到每个通道的权重值。将权重值与原始输入进行逐元素相乘，得到包含通道信息的特征图。

$$\boldsymbol{X}_{\mathrm{R}}=\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathrm{AvgPool}(\boldsymbol{X})))$$

$$\boldsymbol{X}_{\mathrm{c}}=(\mathrm{Sigmod}(\mathrm{Conv}_{1\times1}(\boldsymbol{X}_{\mathrm{R}})))\otimes\boldsymbol{X}$$

式中：$\boldsymbol{X}_{\mathrm{R}}$ 表示通道特征的非线性表示张量，$\boldsymbol{X}_{\mathrm{c}}$ 表示通道分支的输出特征图。

最后，通过求和运算将多尺度注意力输出与通道注意力输出相融合，得到最终的输出特征：

$$\boldsymbol{X}_{\mathrm{Out}}=\boldsymbol{X}_{\mathrm{M}}+\boldsymbol{X}_{\mathrm{c}}$$

GMSA 有效地融合多尺度特征，利用多头自注意力机制捕获复杂的上下文关系，并通过通道分支提升特征的表达能力。

3　实验结果与分析

3.1　评价指标及实验环境

在检测金属表面缺陷时，本文使用多个指标来综合评价模型效率，包括精确率 (precision, P)、召回率 (recall, R)、均值平均精度 (mean average precision, mAP)、参数量 (parameters, Para)、检测帧率 (frames per second, FPS) 等[26]。其中，mAP 是检测金属表面缺陷的核心评价指标，反映了模型在进行目标检测时的整体性能。

本实验采用 Ubuntu 20.04.6 LTS 操作系统，GPU 型号为 NVIDIA GeForce RTX 3090 Ti，运行版本为 CUDA12.2，深度学习框架为 torch-2.4.1+cu121，Python 版本为 3.8。为保证实验公平性，所有实验均不加载预训练权重。实验时的主要参数设置见表 1。

表 1　实验设置参数

Table 1　Setting parameters of the experiment

实验参数	数值
训练轮次（epoch）	300
批次（batch）	4
线程（workers）	1
初始学习率（lr0）	0.000 1
优化器（optimizer）	AdamW

设置训练总轮次为 300，旨在保证模型能够充分拟合。由于 RT-DETR 架构在目标检测中的学习复杂性，较长的训练周期有助于其充分捕捉跨尺度和全局语义特征。在多次实验中发现，轮次超过 250 后性能趋于收敛，300 轮为一个平衡点，既保证了训练充分，又避免了过拟合风险。受限于显存资源，批次大小设置为 4。尽管小批次会产生训练波动，但配合 AdamW 优化器和学习率调度策略，仍能实现稳定训练。数据加载线程设置为 1，以控制实验的可重复性和系统资源的最优利用。在服务器资源受限的场景下，该设置可确保每轮加载稳定、训练结果一致。学习率设置为 0.000 1，结合训练初期较小的批次和较深的网络结构，是一个保守而稳定的选择。在初期实验中发现，过大的学习率容易导致损失函数的值不收敛，而过小则收敛速度明显下降。因此选择 0.000 1 作为折中，能够保证前期稳定训练，后期逐步收敛。

3.2　实验数据集

本实验使用海康威视工业相机现场采集铝型金属表面图像，并通过 Lableme 标注表面缺陷制作铝型金属表面缺陷数据集。采集图像时使用同轴光源，从而使金属工件表面不平整的部位在图像中呈现暗色，凸显缺陷；同时加入圆偏振镜片，利用偏振光特性减少金属工件表面带来的反光影响。采集到的图像包含划痕 (scratches)、擦伤 (bruises)、凹坑 (pits)、白点 (whitespots) 这 4 种金属表面缺陷，图像的分辨率为 2 592×1 944。采集图像示例如图 7 所示。

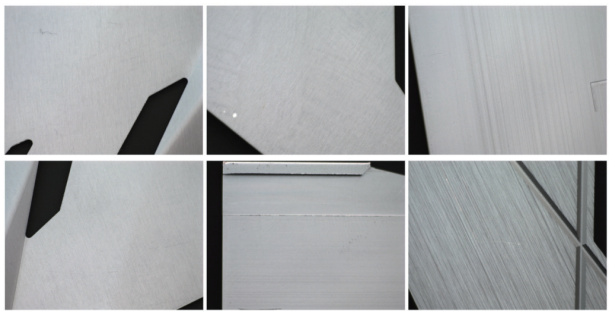
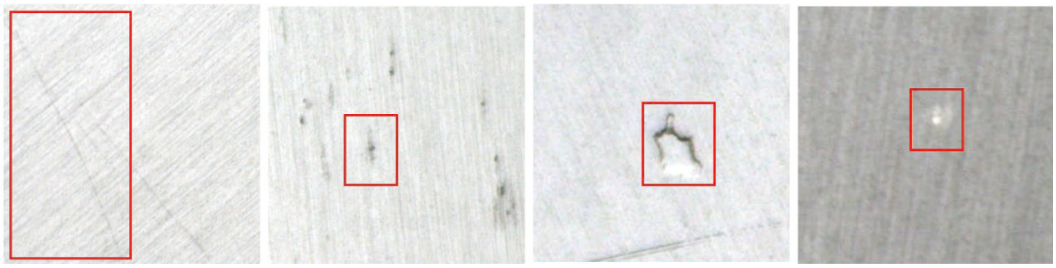


图 7　采集图像示例

Fig. 7　Examples of acquired images

为防止发生过拟合，增强模型的鲁棒性和泛化能力，对采集到的图像进行旋转、切割等数据增强操作，得到分辨率为 640×640 的图像共 5 538 张。所有图像使用 LabelImg 工具进行详细的标注，并将标注好的数据集按 7∶2∶1 的比例划分为训练集、测试集、验证集。缺陷示例如图 8 所示。



(a) 划痕　(b) 擦伤　(c) 凹坑　(d) 白点

图 8　缺陷示例

Fig. 8　Examples of defects

3.3　实验结果分析

在采集到的金属表面缺陷数据集上，评估 HAS-DETR 模型的性能。表 2 对比了基线模型 RT-DETR 模型和本文所提模型 HAS-DETR 的各项评价指标。如表 2 所示，HAS-DETR 模型在 4 类缺陷的检测精度上总体优于 RT-DETR 模型，其平均精度 mAP50 从 55.5% 提高至 62%。具体而言，划痕类别的检测精度提升了 6.9%，凹坑类别提升了 1.5%，白点类别显著提升了 18.8%，而擦伤类别的检测精度差异不明显，这可能是由于擦伤类别形状多变，容易误检导致。实验结果也表明，HAS-DETR 模型在白点这类尺寸虽小但形状相对规则、颜色突出明显的缺陷方面具有显著优势。

表 2　各类缺陷的评价指标

Table 2　Evaluation indicators for various defects　%

模型	类别	评价指标			
		P	*R*	mAP50	mAP50-95
RT-DETR	整体	61.5	58.1	55.5	23.0
	划痕	52.7	40.5	41.1	18.6
	凹坑	65.3	65.3	61.6	26.1
	擦伤	59.8	53.8	52.4	20.1
	白点	68.2	72.8	67.1	27.1
HAS-DETR	整体	65.5	63.3	62.0	27.5
	划痕	57.1	49.8	48.0	24.7
	凹坑	65.9	62.0	62.6	27.6
	擦伤	61.3	56.5	51.5	22.9
	白点	77.7	84.8	85.9	34.9

从整体结果来看，HAS-DETR 相较于 RT-DETR 在所有评估指标上均取得了明显提升，其中 mAP50 从 55.5% 提升至 62%，mAP50-95 提升了 4.5%，表明其在精度与泛化能力方面均优于基线模型。HAS-DETR 模型在精确率 *P*、召回率 *R* 这两个指标上也有明显提升，分别提高了 4% 和 5.2%。这表明，HAS-DETR 模型在减少漏检的同时有效提升了检测结果的准确性，进一步验证了其在金属表面缺陷检测任务中的优越性能。

为了验证本文提出的 HAS-DETR 算法在真实场景中的效果，在金属表面缺陷数据集上随机挑选具有代表性的图像进行检测，涵盖划痕、凹坑、擦伤等典型小目标缺陷类别，并将检测结果与基线 RT-DETR 进行对比。检测效果如图 9 所示。观察图 9 可知，RT-DETR 模型在图像 a 中漏检了一个细小的划痕缺陷，在图像 b 中漏检了一个凹坑缺陷和一个划痕缺陷，在图像 c 中漏检了一个擦伤缺陷，在图像 d 中漏检了一个划痕缺陷和一个擦伤缺陷，而 HAS-DETR 模型则没有出现这些漏检问题。在图像 e、f 中，尽管 RT-DETR 模型和 HAS-DETR 模型均成功检测出所有缺陷，但可以明显发现，HAS-DETR 生成的检测框具有更高的置信度。通过对比可以看出，HAS-DETR 模型在金属表面缺陷检测任务中取得了显著的性能提升。

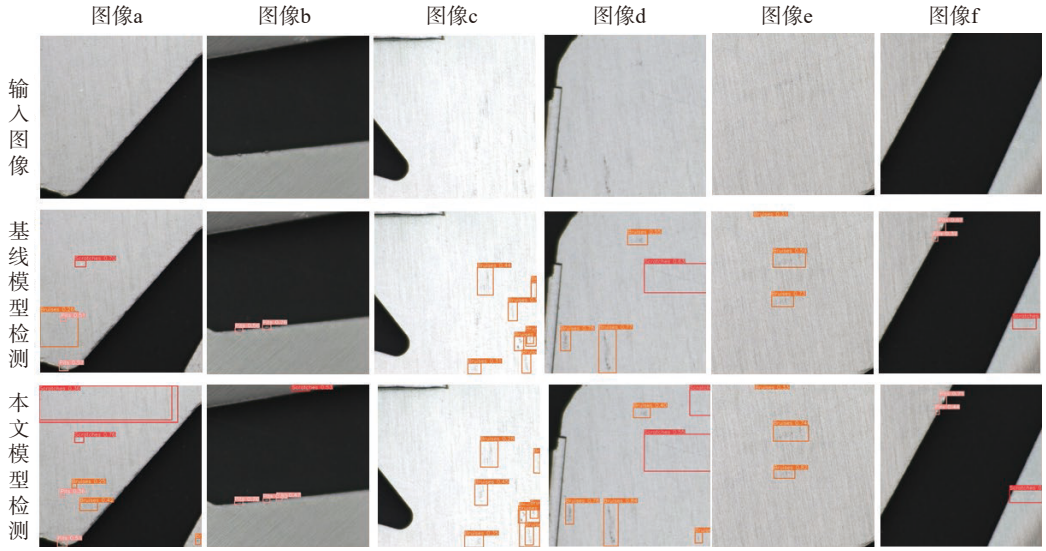


图 9　RT-DETR 和 HAS-DETR 检测结果对比

Fig. 9　Comparison of RT-DETR and HAS-DETR detection results

3.4　消融实验

为了验证本文提出的复合差分卷积、双重多尺度特征融合模块、全局多尺度注意力机制对小目标检测的有效性，选用 RT-DETR 模型作为消融实验的基准模型，依次加入各个模块进行消融实验，所有实验均在自制的金属表面缺陷数据集上进行，并保持相同的实验配置。消融实验结果如表 3 所示。

表 3　消融实验结果

Table 3　Ablation test results

模型	MDConv	GMSA	DMFF	评价指标				
				mAP50/%	mAP50-95/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(f/s)
RT-DETR	×	×	×	55.5	23.0	**19.88**	**57.0**	55.87
	√	×	×	59.0	24.7	**19.88**	**57.0**	67.11
	×	√	×	60.1	23.9	19.91	57.1	**109.89**
	×	×	√	60.6	24.6	20.95	61.9	50.00
	√	√	×	61.2	25.8	19.92	57.2	69.93
	×	√	√	61.2	26.4	20.98	62.1	106.38
	√	×	√	61.1	26.7	20.95	62.0	88.49
HAS-DETR	√	√	√	**62.0**	**27.5**	20.99	62.2	84.75

注：×为不添加模块，√为添加模块，加粗数据为最优值。

采用 MDConv 代替 BasicBlock 中的标准卷积后，模型的 mAP50 由 55.5% 提升至 59%，mAP50-95 由 23% 提升至 24.7%，参数量和计算量保持不变，推理速度上升。这表明 MDConv 能够有效提升对小目标的特征提取能力，从而显著改善模型的检测精度，而对模型的计算开销没有额外影响。用 GMSA 注意力机制替换 AIFI 模块中的多头注意力机制之后，mAP50 提升至 60.1%，mAP50-95 达到 23.9%，FPS 上升至 109.89 帧/s。这表明，GMSA 能够通过捕获复杂的上下文关系提升模型在复杂背景中的鲁棒性和检测精度，又能够提高推理速度，具有很高的实用性。在单独引入 DMFF 后，mAP50 和 mAP50-95 分别提升至 60.6% 和 24.6%，参数量有 1.07×10^6 的略微增加，推理速度下降至 50 帧/s。这表明，DMFF 通过多尺度特征融合显著提升了模型对目标尺度变化的适应性，但会导致推理速度下降，更加适用于精度优先的应用场景。

在组合模块的实验中，同时引入 GMSA 与 DMFF 可进一步将 mAP50 提升至 61.2%，mAP50-95 达到 26.4%，但推理速度略低于单独引入 GMSA 的情况。在同时引入 MDConv、GMSA 和 DMFF 的 HAS-DETR 模型中，mAP50 和 mAP50-95 都达到了最好的效果，分别为 62% 和 27.5%，参数量和计算量相较 RT-DETR 有略微提升，但推理速度上升至 84.75 帧/s。HAS-DETR 的实验结果表明，

3 种模块协同作用能够显著提升检测精度, 同时在推理速度上也保持较好的性能, 适用于对精度和实时性均有较高要求的应用场景。

3.4.1　MDConv 模块的消融实验

为了更加清晰地验证 MDConv 模块的性能, 以及其解决尺度变化大、背景复杂问题的能力。本文设计 MDConv 模块的变体 Ⅰ、Ⅱ、Ⅲ、Ⅳ、Ⅴ、Ⅵ进行消融实验。变体 Ⅰ 使用模型 RT-DETR 中原有的 3×3 卷积, 作为基线方法。在变体 Ⅱ 中, 将垂直差分卷积、水平差分卷积与 3×3 的卷积进行串联融合, 用于初步验证差分操作在局部方向感知增强中的作用。在变体 Ⅲ 中, 在变体 Ⅱ 基础上引入中心差分卷积, 以探究中心响应对目标边界和纹理的敏感性效果。在变体 Ⅳ 中, 在变体 Ⅲ 的基础上引入角差分卷积, 使用完整的 MDConv 进行特征提取。在变体 Ⅴ、Ⅵ 中, 分别引入 DySnakeConv[27] 和 ShiftwiseConv[28] 进行特征提取, 与 MDConv 的检测效果进行对比。

从表 4 结果可以看出, 变体 Ⅱ、Ⅲ 在引入差分卷积结构后, mAP50 分别为 57.8% 和 58.1%, 已较基线结构取得一定性能提升, 说明差分信息对于增强局部边缘、方向梯度等特征是有效的, 能够提升模型处理复杂纹理背景的能力。变体 Ⅳ 作为完整的 MDConv 模块结构, 综合融合各方向差分特征, 并在训练后期通过加权卷积融合统一表示, 取得了最高的 mAP50(59.0%)和 mAP50-95(24.7%), 显著优于其他变体, 验证了 MDConv 在提升检测准确性方面的显著效果。

表 4　MDConv 模块的消融实验结果

Table 4　Ablation results of the MDConv module

变体	评价指标				
	mAP50/%	mAP50-95/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(帧/s)
Ⅰ	55.5	23.0	**19.88**	**57.0**	**55.87**
Ⅱ	57.8	24.2	**19.88**	**57.0**	112.36
Ⅲ	58.1	23.1	**19.88**	**57.0**	68.03
Ⅳ	**59.0**	**24.7**	**19.88**	**57.0**	67.11
Ⅴ	58.7	23.2	27.86	60.8	97.09
Ⅵ	58.1	23.5	15.93	59.8	86.21

注: 加粗数据为最优值。

值得注意的是, MDConv 的参数量和计算量与变体 Ⅰ、Ⅱ、Ⅲ 保持一致, 这是因为在 MDConv 中, 虽然训练阶段引入多个分支, 这些分支均基于相同的输入输出维度, 通过不同的卷积结构增强特征表达能力。在推理阶段, 所有分支的卷积权重进行加权融合, 合并为一个等效的单一卷积。因此, 在最终的推理阶段仅保留这一合并后的卷积层, 使得模型的参数量和计算复杂度不会因为多分支而增加, 从而实现训练阶段的表达增强与推理阶段的高效兼顾。

在与现有代表性动态卷积方法的对比中, DySnakeConv 虽然在 mAP50 上达到 58.7%, 但在 mAP50-95 上仍不及 MDConv, 同时由于结构复杂, 参数量显著上升至 27.86×10^6, 存在参数量利用率不高的问题。ShiftwiseConv 则表现出一定的轻量化优势, 但在精度方面略逊一筹, mAP50-95 仅为 23.5%。

3.4.2　GMSA 模块的消融实验

为系统地评估所提出 GMSA 模块中各关键组成部分对模型性能的影响, 以及其解决检测目标尺度变化大、背景纹理复杂等问题的能力, 本文设计了 6 种模型变体, 分别记为 A、B、C、D、E 和 F, 开展消融实验与对比分析。在变体 A 中, 去除 GMSA 模块中的多尺度部分, 仅保留 GMSA 架构中的基于全局自注意力的计算过程, 旨在验证多尺度特征对于检测性能的影响。在变体 B 中, 保留多尺度结构, 将原有的 3 个并行的深度可分离空洞卷积替换为普通 3×3 卷积进行特征提取, 以评估空洞卷积在捕捉不同感受野特征方面的作用。在变体 C 中, 去除 GMSA 中的通道分支, 保留多尺度分支, 探究通道注意力对模型整体检测性能的影响。在变体 D 中, 使用完整的 GMSA 模块, 以便与其他变体进行对比。同时, 为进一步对比 GMSA 的性能优势, 引入两种主流的注意力机制 DAttention[29]、CascadedGroupAttention[30] 作为变体 E、F, 进行横向对比。

通过表 5, 可以观察到完整的 GMSA 模块在 mAP50、mAP50-95 指标上取得最优性能, 且推理速度达到 109.89 f/s, 这说明 GMSA 模块在提升检测准确率的同时也兼顾了检测速度。其中 GMSA 模块 mAP50 指标达到 60.1%, 相比去除多尺度结构的变体 A 提升了 1.7 个百分点, 验证了

多尺度信息融合对目标检测性能的显著提升作用。将深度可分离空洞卷积替换为普通卷积导致 mAP50-95 降低 0.6%, 说明空洞卷积对于多尺度上下文建模具有重要作用。去除通道分支对性能有较大影响, mAP50 降至 56.5%, 验证了通道注意力在特征选择与抑制冗余方面具有关键作用。与 DAttention 和 CascadedGroupAttention 相比, GMSA 模块在保持相似参数与计算量的同时, 在精度上保持领先, 展示出更优的结构设计与实际效能平衡。

表 5　GMSA 模块的消融实验结果

Table 5　Ablation results of the GMSA module

变体	评价指标				
	mAP50/%	mAP50-95/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(帧/s)
A	58.4	22.8	**19.78**	57.1	**169.49**
B	56.7	23.3	20.02	57.2	136.99
C	56.5	22.2	19.91	57.1	119.05
D	**60.1**	**23.9**	19.91	57.1	109.89
E	59.2	23.7	19.71	**57.0**	120.48
F	59.0	23.4	19.88	57.2	123.46

注: 加粗数据为最优值。

3.4.3　DMFF 模块的消融实验

为验证提出的双重多尺度特征融合模块 DMFF 在金属表面缺陷检测任务中的有效性, 设计了 a、b、c、d、e 共 5 种变体结构, 逐步引入 DMFF 的关键子模块, 并开展消融实验。在变体 a 中, 直接采用 RT-DETR 中原有的 CCFM, 作为基线方法。在变体 b 中, 将原有的 CCFM 模块替换为 MSFE 模块, 验证 MSFE 模块对整体检测性能的直接影响。在变体 c、d 中, 在变体 b 的基础上, 分别加入小波上采样、小波下采样, 以评估小波上、下采样对整体模型的贡献。在变体 e 中, 集成 MSFE、小波上采样与下采样, 构成完整的 DMFF 模块, 以评估完整结构下的检测性能与效率表现。

如表 6 所示, 相较基线模型, 引入 MSFE 模块后, 模型在 mAP50 和 mAP50-95 上均取得明显提升, 分别提高 2.2% 和 2.5%, 说明了 MSFE 在特征层级融合与尺度建模方面的有效性。然而, 计算开销亦随之增加, 浮点运算速度从 57.0 s^{-1} 上升至 66.1 s^{-1}, FPS 下降至 52.63 帧/s。在此基础上, 分别引入小波上采样与小波下采样。变体 c 提升 mAP50 至 58.5%, FPS 明显提升至 65.79 帧/s, 体现上采样对局部细节恢复和推理加速有较好促进作用, 但 mAP50-95 反而下降至 23.7%, 说明其对检测边界与小目标支持有限。变体 d 取得 59.9% 的 mAP50 和 25.1% 的 mAP50-95, 以及 103.09 帧/s 的 FPS, 相较基线模型有大幅提升, 验证了小波下采样在增强深层特征语义上的优势。最后, 变体 e 集成完整 DMFF 模块, 在 mAP50 和 mAP50-95 上均达到最优值, 验证了多尺度特征融合在检测精度方面的优势, 但推理速度下降至 50 帧/s, 显示了完整结构在精度优先场景中的应用潜力。DMFF 模块的参数量和浮点运算速度均低于变体 b、c、d, 表明完整 DMFF 在保持结构紧凑的同时, 充分发挥了多尺度增强与小波融合机制的协同作用。

表 6　DMFF 模块的消融实验结果

Table 6　Ablation results of the DMFF module

变体	评价指标				
	mAP50/%	mAP50-95/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(帧/s)
a	55.5	23.0	**19.88**	**57.0**	**55.87**
b	57.7	**25.5**	22.24	66.1	52.63
c	58.5	23.7	21.21	65.3	65.79
d	59.9	25.1	21.98	62.8	103.09
e	**60.6**	24.6	20.95	61.9	50.00

注: 加粗数据为最优值。

综上所述, 为有效应对金属表面缺陷检测中常见的尺度变化大与背景复杂问题, 本文从模块结构设计上进行了多层次的针对性优化, 并通过系统的消融实验验证了其有效性。

一方面, 多尺度建模能力在多个模块中被重点强化。MDConv 模块通过引入垂直、水平、中心、角差分卷积, 增强了对不同方向和粒度的边缘特征提取能力, 在不增加参数与计算量的前提下显著提升了检测性能。GMSA 模块融合多尺度空洞卷积与通道注意力机制, 显著提升了模型对多尺度上下文的感知能力。DMFF 模块则在多尺度融合基础上引入小波上、下采样策略, 进一步提升模型对局部细节与全局语义信息的建模能力, 在保持较低参数量的同时取得了最优检测精度。

另一方面, 背景纹理复杂问题亦通过多种机制得以改善。MDConv 的方向差分操作增强了模型对边界与梯度变化的敏感性, 有效削弱了背景干扰。GMSA 模块中的通道注意力分支进一步提升了模型的特征选择能力, 抑制了冗余信息干扰。同时, 基于 Transformer 架构的全局建模能力使模型具备较强的长距离依赖理解能力, 对背景复杂、结构变化多样的缺陷区域具有更强的适应性。

3.5　对比实验

为进一步验证 HAS-DETR 检测金属表面缺陷的性能, 本实验采用自建的金属表面缺陷数据集, 选取当前主流的目标检测模型作为对比, 涵盖了 YOLO 系列中的 YOLOv5[31]、YOLOv6[32]、YOLOv7[33]、YOLOv8[34]、YOLOv9[23]、YOLOv10n[35]、YOLOv11[36] 和 YOLOv12[37] 等模型, 以及 R-CNN 系列中的 Cascade R-CNN[38] 和 Grid R-CNN[39] 模型。其中, 设置 Cascade R-CNN 和 Grid R-CNN 的训练轮次为 50, 其余模型均为 300, 其他参数保持一致。对比实验结果如表 7 所示。

表 7　对比实验结果

Table 7　Results of comparative experiments

模型	评价指标						
	mAP50/%	mAP50-95/%	*P*/%	*R*/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(帧/s)
YOLOv5	43.7	17.2	49.4	46.7	**2.50**	7.10	94.34
YOLOv6	42.0	15.9	45.4	47.0	4.23	11.80	111.11
YOLOv7	48.6	15.4	57.8	46.7	36.49	103.20	81.30
YOLOv8n	50.4	20.9	55.4	48.3	3.00	8.10	117.65
YOLOv9s	40.3	15.9	44.4	46.0	60.50	263.90	49.02
YOLOv10n	46.0	17.2	46.4	51.0	2.69	8.20	172.41
YOLOv11	42.6	16.3	50.6	42.7	2.58	6.30	103.09
YOLOv12	37.8	14.1	44.2	40.7	2.51	**5.80**	**227.27**
Casecade R-CNN	58.3	25.9	59.5	37.6	69.16	162.10	120.40
Grid R-CNN	57.1	27.3	60.3	43.5	247.72	64.47	189.10
RT-DETR	55.5	23.0	61.5	58.1	19.88	57.00	55.87
HAS-DETR	**62.0**	**27.5**	**65.5**	**63.3**	20.99	65.20	84.75

注: 加粗数据为最优值。

分析表 7, 不难发现, 在 mAP50 上, 本文提出的模型 HAS-DETR 达到了 62%, 远超过 YOLO 系列模型, 以及精度较高的 Cascade R-CNN(58.3%)和 Grid R-CNN(57.1%)模型。在 mAP50-95 上, HAS-DETR 模型同样表现优异, 达到了 27.5%, 明显超过 YOLO 系列模型和 R-CNN 系列模型。这表明, HAS-DETR 在复杂场景中的精度具有显著优势, 尤其是在细粒度目标检测任务中, 能够提供更准确的检测结果。HAS-DETR 在精度 *P* 和召回率 *R* 这两个指标上同样表现突出, 分别为 65.5% 和 63.3%, 优于所有对比模型。特别是在召回率 *R* 上, HAS-DETR 相较于 Cascade R-CNN(37.6%)和 YOLOv12(40.7%)有明显提升, 表明其在捕捉目标方面有更强的能力。这使得 HAS-DETR 在多目标场景中的表现更稳健, 能够有效减少漏检和误检。

在计算复杂度方面, HAS-DETR 的浮点运算速度为 65.2 s^{-1}, 相较基线模型 RT-DETR 的 57.0 s^{-1}, 有轻微的上涨, 但相较 Cascade R-CNN 162.1 s^{-1} 和 YOLOv9 263.9 s^{-1}, 并综合考虑 HAS-DETR 的高精度表现, 这个计算开销是合理的。与之相比, YOLOv12 这样的低计算量模型, 更加适合资源受限的环境。在推理速度方面, HAS-DETR 的 FPS 为 84.75 帧/s, 虽然低于部分对比模型, 但相较基线模型 RT-DETR, 仍有很大提升。

为了更加直观地对比 HAS-DETR 和 YOLO

模型的检测精度，根据 epoch 迭代，绘制了各模型在训练过程中的 mAP50 曲线，每 30 个 epoch 取一次值，如图 10 所示。

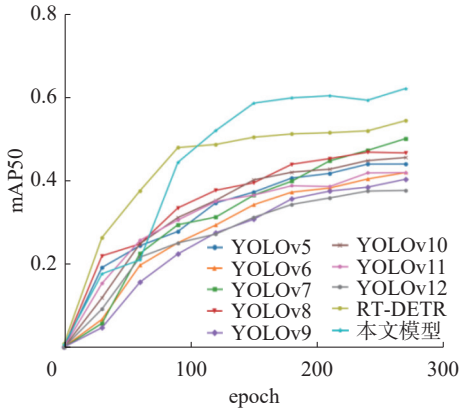


图 10　mAP50 曲线

Fig. 10　mAP50 curves

根据图 10 可以发现，在训练的初始阶段，即 0~50 个 epoch 内，HAS-DETR 的精度增长率明显快于大部分 YOLO 系列模型，尤其相较 YOLOv6、YOLOv9 等模型。在 50~150 个 epoch 的中期训练阶段，HAS-DETR 的 mAP50 曲线始终保持高速增长，并逐渐拉开与其他模型的差距。在 150 个 epoch 附近，HAS-DETR 模型的 mAP50 指标趋于平稳并达到 62% 的峰值，这一结果显著优于基线模型 RT-DETR(55.5%) 和 YOLO 系列中精度最高的 YOLOv7(48.6%)。这一表现说明，HAS-DETR 模型在提升精度的同时，具备更好的训练稳定性和更强的泛化能力。

HAS-DETR 的 mAP50 曲线始终保持高速增长，并逐渐拉开与其他模型的差距。在 150 个 epoch 附近，HAS-DETR 模型的 mAP50 指标趋于平稳并达到 62% 的峰值，这一结果显著优于基线模型 RT-DETR(55.5%) 和 YOLO 系列中精度最高的 YOLOv7(48.6%)。这一表现说明，HAS-DETR 模型在提升精度的同时，具备更好的训练稳定性和更强的泛化能力。

与 YOLO 系列和 R-CNN 系列模型相比，本文所提出的 HAS-DETR 模型的整体性能具有明显优势。YOLOv5、YOLOv6 等轻量化模型虽然推理速度较快，但其精度与 HAS-DETR 模型相差较大，难以满足复杂场景下小目标的检测需求。而 Cascade R-CNN、YOLOV8 等相对精度较高模型的精度也明显低于 HAS-DETR 模型。

3.6　泛化实验

为验证 HAS-DETR 模型的泛化能力，本文选取广东工业智造大数据创新大赛——智能算法赛《铝型材表面瑕疵识别》的初赛开源数据集 ADPPP(aluminum profile surface detection database) 对模型的检测性能进行验证。ADPPP 数据集专为铝材表面缺陷检测设计，共包含 1 885 张图像，涵盖 10 类不同的表面瑕疵。数据集按照 8∶1∶1 的比例划分为训练集、验证集和测试集。实验中的运行环境和模型参数与对比实验保持一致，实验结果如表 8 所示。

表 8　泛化实验结果

Table 8　Results of generalization experiments

模型	评价指标						
	mAP50/%	mAP50-95/%	*P*/%	*R*/%	参数量/10^6	浮点运算速度/(10^9 s^{-1})	检测帧率/(帧/s)
YOLOv5	59.1	35.2	68.3	58.2	**2.50**	7.1	169.49
YOLOv6	58.3	33.5	63.3	54.0	4.23	11.8	384.62
YOLOv7	55.6	34.4	54.2	57.3	36.49	103.2	50.76
YOLOv8n	58.3	36.0	64.0	56.4	3.00	8.1	**476.19**
YOLOv9s	61.2	37.8	66.1	**57.7**	60.50	263.9	227.27
YOLOv10n	53.4	32.6	53.7	49.1	2.69	8.2	384.62
YOLOv11	60.4	36.7	70.8	54.4	2.58	6.3	256.41
YOLOv12	56.5	32.8	62.0	54.6	2.51	**5.8**	204.08
RT-DETR	62.4	39.4	70.0	57.4	19.88	57.0	178.57
HAS-DETR	**64.4**	**40.7**	**70.1**	57.4	20.99	65.2	178.57

注：加粗数据为最优值。

由表 8 可知，HAS-DETR 模型在泛化实验中各项指标均表现出较强的鲁棒性和跨场景适应能力。针对 ADPPP 数据集，HAS-DETR 模型相较基线模型 RT-DETR，在 mAP50 和 mAP50-95 上分别

提升 2.0% 和 1.3%; 相较 YOLO 系列模型则实现全面领先, 在 mAP50 和 mAP50-95 上领先幅度分别高达 11.0% 和 5.5%。在精确率与召回率方面, HAS-DETR 分别达到 70.1% 和 57.4%, 整体优于多数 YOLO 系列模型, 尤其在精确率指标上处于当前最优水平, 显示其在误检控制方面具备更强的泛化能力。而在召回率方面, HAS-DETR 与 RT-DETR 相当, 明显优于 YOLOv6(54.0%)和 YOLOv10n(49.1%), 进一步证明其具备更好的目标完整性识别能力。在推理速度方面, HAS-DETR 与 RT-DETR 持平, 低于部分 YOLO 系列模型, 但在满足实时性要求的同时保持了良好的检测精度, 适合在对精度要求较高的实际场景中部署。

综上所述, HAS-DETR 在 ADPPP 数据集上的综合表现表明, 其在检测精度、误检控制与检测速度之间取得良好平衡, 展现出极强的泛化能力和跨场景适用性, 是金属表面缺陷检测任务中兼顾性能与实用性的有力方案。

4 结束语

针对金属表面缺陷数据集检测目标小、尺度变化大、背景复杂的问题, 本文在 RT-DETR 模型的基础上提出 HAS-DETR 模型, 有效提升了检测精度。通过在 Backbone 中设计复合差分卷积, 增强对小目标的特征提取能力。通过构建双重多尺度特征融合模块, 有效捕捉全局语义信息和细节特征, 解决了目标尺度变化大的问题。同时, 设计全局多尺度注意力机制代替 AIFI 模块中的多头注意力机制, 能够捕获复杂的上下文关系, 在处理复杂背景和多尺度目标时, 表现出更好的鲁棒性和精确度。实验结果表明, HAS-DETR 在检测精度方面取得显著提升。

本文所提出的模型的检测精度虽然在基线模型的基础上有很大提升, 但在实际工业场景中还需要进一步优化。下一步工作将重点研究如何在保证检测实时性的同时, 进一步提升检测精度, 以期为工业应用场景提供更加高效且可靠的解决方案。

参考文献:

[1] 李宗祐, 高春艳, 吕晓玲, 等. 基于深度学习的金属材料表面缺陷检测综述[J]. 制造技术与机床, 2023(6): 61−67.
LI Zongyou, GAO Chunyan, LV Xiaoling, et al. A review of surface defect detection for metal materials based on deep learning[J]. Manufacturing technology & machine tool, 2023(6): 61−67.

[2] 孙卫波, 丁卫. 改进 YOLOv7 的带钢表面缺陷检测算法[J]. 工业控制计算机, 2024, 37(8): 94−96,101.
SUN Weibo, DING Wei. Improved YOLOv7 strip surface defect detection algorithm[J]. Industrial control computer, 2024, 37(8): 94−96,101.

[3] 马鸽, 邓开宏, 李国章, 等. 基于改进 YOLOv5s 模型的金属表面缺陷检测方法[J]. 广州大学学报 (自然科学版), 2024, 23(4): 9−19.
MA Ge, DENG Kaihong, LI Guozhang, et al. Metal surface defect dectection method based on an improved YOLOv5s model[J]. Journal of Guangzhou University (natural science edition), 2024, 23(4): 9−19.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580−587.

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

[6] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 21−37.

[7] 向宽, 李松松, 栾明慧, 等. 基于改进 Faster RCNN 的铝材表面缺陷检测方法[J]. 仪器仪表学报, 2021, 42(1): 191−198.
XIANG Kuan, LI Songsong, LUAN Minghui, et al. Aluminum product surface defect detection method based on improved Faster RCNN[J]. Chinese journal of scientific instrument, 2021, 42(1): 191−198.

[8] WANG H, WANG J, LUO F. Research on surface defect detection of metal sheet and strip based on multi-level feature Faster R-CNN[J]. Mechanical science and technology for aerospace engineering, 2020, 20(4): 94−107.

[9] FANG Junting, TAN Xiaoyang, WANG Yuhui. ACRM: attention cascade R-CNN with mix-NMS for metallic surface defect detection[C]//2020 25th International Conference on Pattern Recognition. Milan: IEEE, 2021: 423−430.

[10] WANG Chenglong, XIE Heng. MeDERT: a metal surface defect detection model[J]. IEEE access, 2023, 11: 35469−35478.

[11] 刘浩翰, 孙铖, 贺怀清, 等. 基于改进 YOLOv3 的金属表面缺陷检测[J]. 计算机工程与科学, 2023, 45(7): 1226−1235.

LIU Haohan, SUN Cheng, HE Huaiqing, et al. Metal surface defect detection based on improved YOLOv3[J]. Computer engineering & science, 2023, 45(7): 1226−1235.

[12] 凌强, 刘宇, 王春举, 等. DN-YOLOv5 的金属双极板表面缺陷检测算法[J]. 哈尔滨工业大学学报, 2023, 55(12): 104−112.
LING Qiang, LIU Yu, WANG Chunju, et al. DN-YOLOv5 algorithm for detecting surface defects of metal bipolar plates[J]. Journal of Harbin Institute of Technology, 2023, 55(12): 104−112.

[13] ZHANG Heng, FU Wei, WANG Xiaoming, et al. An efficient model for metal surface defect detection based on attention mechanism and multi-scale feature[J]. The journal of supercomputing, 2024, 81(1): 40.

[14] ZHAO Yian, LYU Wenyu, XU Shangliang, et al. DETRs beat YOLOs on real-time object detection[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 16965−16974.

[15] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[M]//Computer Vision– ECCV 2020. Cham: Springer International Publishing, 2020: 213−229.

[16] 董适, 赵国瑞, 苟豪, 等. 基于改进 RT-Detr 的黄瓜果实选择性采摘识别方法[J]. 农业工程学报, 2025, 41(1): 212−220.
DONG Shi, ZHAO Guorui, GOU Hao, et al. Identifying cucumber fruits during selective picking using improved RT-Detr[J]. Transactions of the Chinese society of agricultural engineering, 2025, 41(1): 212−220.

[17] 陶健. 基于空洞卷积与空间注意力的遥感影像小目标检测方法[J]. 测绘与空间地理信息, 2024, 47(10): 104−107,111.
TAO Jian. Small target detection method in remote sensing images based on atrous convolution and spatial attention[J]. Geomatics & spatial information technology, 2024, 47(10): 104−107,111.

[18] CHEN Zixuan, HE Zewei, LU Zheming. DEA-net: single image dehazing based on detail-enhanced convolution and content-guided attention[J]. IEEE transactions on image processing, 2024, 33: 1002−1015.

[19] ZHANG Luping, LUO Junhai, HUANG Yian, et al. MDIGCNet: multidirectional information-guided contextual network for infrared small target detection[C]//IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. [S.l.]: IEEE, 2024: 2063−2076.

[20] YU Zitong, ZHAO Chenxu, WANG Zezheng, et al. Searching central difference convolutional networks for face anti-spoofing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 5294–5304.

[21] 李琼, 考月英, 张莹, 等. 面向无人机航拍图像的目标检测研究综述[J]. 图学学报, 2024, 45(6): 1145−1164.
LI Qiong, KAO Yueying, ZHANG Ying, et al. Review on object detection in UAV aerial images[J]. Journal of graphics, 2024, 45(6): 1145−1164.

[22] WILLIAMS T, LI R. Wavelet pooling for convolutional neural networks[C]//6th International conference on learning representations. Vancouver: [s. n.], 2018.

[23] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: learning what you want to learn using programmable gradient information[EB/OL]. (2024−02−29) [2025−01−02]. https://arxiv.org/abs/2402.13616.

[24] YAO Ting, PAN Yingwei, LI Yehao, et al. Wave-ViT: unifying wavelet and transformers for visual representation learning[M]//Computer Vision–ECCV 2022. Cham: Springer Nature Switzerland, 2022: 328−345.

[25] 吴铁钰, 杨光, 邹丽. RSG-YOLO: 用于检测道路坑洼的高效神经网络[J]. 计算机技术与发展, 2025, 35(2): 199−206.
WU Tieyu, YANG Guang, ZOU Li. RSG-YOLO: an efficient neural network for road pothole detection[J]. Computer technology and development, 2025, 35(2): 199−206.

[26] 孙己龙, 刘勇, 周黎伟, 等. 基于 DCNv2 和 Transformer Decoder 的隧道衬砌裂缝高效检测模型研究[J]. 图学学报, 2024, 45(5): 1050−1061.
SUN Jilong, LIU Yong, ZHOU Liwei, et al. Research on efficient detection model of tunnel lining crack based on DCNv2 and Transformer Decoder[J]. Journal of graphics, 2024, 45(5): 1050−1061.

[27] QI Yaolei, HE Yuting, QI Xiaoming, et al. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 6047−6056.

[28] LI Dachong, LI Li, CHEN Zhuangzhuang, et al. ShiftwiseConv: small convolutional kernel with large kernel effect[EB/OL]. (2024−01−23) [2025−03−13]. https://arxiv.org/abs/2401.12736.

[29] XIA Zhuofan, PAN Xuran, SONG Shiji, et al. Vision transformer with deformable attention[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 4784−4793.

[30] LIU Xinyu, PENG Houwen, ZHENG Ningxin, et al. EfficientViT: memory efficient vision transformer with cas-

caded group attention[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouve: IEEE, 2023: 14420−14430.

[31] JOCHER G, STOKEN A, BOROVEC J, et al. Ultralytic/yolov5: v5.0-YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations[EB/OL]. (2021−10−12) [2025−01−02]. https://github.com/ultralytics/yolov5.

[32] LI Chuyi, LI Lulu, JIANG Hongliang, et al. YOLOv6: a single-stage object detection framework for industrial applications[EB/OL]. (2022−09−07) [2025−01−02]. https://arxiv.org/abs/2209.02976.

[33] 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. Vancouver: IEEE, 2023: 7464−7475.

[34] VARGHESE R, M S. YOLOv8: a novel object detection algorithm with enhanced performance and robustness [C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems. Chennai: IEEE, 2024: 1−6.

[35] WANG Aowang, CHEN Hui, LIU Lihao, et al. YOLOv10: real-time end-to-end object detection[EB/OL]. (2024−10−30)[2025−01−02]. https://arxiv.org/abs/2405.14458.

[36] KHANAM R, MUHAMMAD H. YOLOv11: An Overview of the Key Architectural Enhancements[EB/OL]. (2024−10−23) [2025−01−02]. https://arxiv.org/abs/2410.17725.

[37] TIAN Yunjie, YE Qixiang, DOERMAN D. YOLOv12: Attention-Centric Real-Time Object Detectors[EB/OL]. (2025−02−18) [2025−03−13]. https://arxiv.org/abs/2502.12524.

[38] CAI Zhaowei, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6154−6162.

[39] LU Xin, LI Buyu, YUE Yuxin, et al. Grid R-CNN[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 7355−7364.

作者简介：

李冰，副教授，博士，主要研究方向为模式识别与计算机视觉。主持中央高校基金面上项目 2 项、主持横向科研项目 5 项。发表学术论文 30 余篇，获发明专利授权 4 项。E-mail：li_bing@ncepu.edu.cn。

王月，硕士研究生，主要研究方向为电力视觉及目标检测。E-mail：2011616203@qq.com。

翟永杰，教授，博士，主要研究方向为电力视觉。主持国家自然科学基金面上项目 2 项、河北省自然科学基金项目 2 项。编著教材 1 部，著作 4 部。发表学术论文 30 余篇。E-mail：zhaiyongjie@ncepu.edu.cn。