工程科学学报
Chinese Journal of Engineering

基于卷积与Transformer融合框架的列车轮对轴承损伤识别方法

邓飞跃 蔡毓龙 王锐 郑守禧

Train wheelset bearing damage identification method based on convolution and transformer fusion framework

DENG Feiyue, CAI Yulong, WANG Rui, ZHENG Shouxi

引用本文:
邓飞跃, 蔡毓龙, 王锐, 郑守禧. 基于卷积与Transformer融合框架的列车轮对轴承损伤识别方法[J]. 北科大：工程科学学报, 2024, 46(10): 1834–1844. doi: 10.13374/j.issn2095–9389.2024.01.02.003
DENG Feiyue, CAI Yulong, WANG Rui, ZHENG Shouxi. Train wheelset bearing damage identification method based on convolution and transformer fusion framework[J]. *Chinese Journal of Engineering*, 2024, 46(10): 1834–1844. doi: 10.13374/j.issn2095–9389.2024.01.02.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2024.01.02.003

工程科学学报, 第 46 卷, 第 10 期: 1834−1844, 2024 年 10 月
Chinese Journal of Engineering, Vol. 46, No. 10: 1834−1844, October 2024
https://doi.org/10.13374/j.issn2095-9389.2024.01.02.003; http://cje.ustb.edu.cn

基于卷积与 Transformer 融合框架的列车轮对轴承损伤识别方法

邓飞跃✉, 蔡毓龙, 王　锐, 郑守禧

石家庄铁道大学机械工程学院, 石家庄 050043
✉通信作者, E-mail: dengfy@stdu.edu.cn

摘　要　针对传统机器视觉方法在列车轮对轴承损伤检测中存在的图像特征提取不敏感、专家经验要求高以及识别准确率偏低等问题, 本文提出了一种基于卷积与 Transformer 融合框架的列车轮对轴承损伤识别方法. 首先, 发展了一种图像增强类别重组的预处理方法, 消除不同类别数据样本不均衡的影响, 提高图像数据集质量; 其次, 基于卷积与自注意力融合思想, 设计了 VGG 与 Transformer 双分支并行融合网络(VGG and Transformer parallel fusion network, VTPF-Net),综合获取图像全局轮廓特征与局部细节特征信息; 再次, 构建了多尺度膨胀空间金字塔卷积(Multiscale dilation spatial pyramid convolution, MDSPC)模块, 利用多尺度膨胀卷积递进融合充分挖掘特征图中多尺度语义特征; 最后, 基于 NEU-DET 图像缺陷数据集与自建列车轮对轴承图像数据集进行了实验分析. 结果表明, 所提模型对 NEU-DET 数据中 6 类缺陷图像与轮对轴承 4 类故障图像的识别准确率分别为 99.44% 与 98%, 能够较为准确识别不同损伤类型图像样本, 在不明显增加模型复杂度基础上各项评价指标要显著优于当前 CNN 模型、自注意力机制 ViT 模型以及 CNN-Transformer 融合模型.

关键词　轮对轴承; 损伤识别; 卷积网络; Transformer 网络; 多尺度特征
分类号　TH133.3;TH17

Train wheelset bearing damage identification method based on convolution and transformer fusion framework

DENG Feiyue✉, *CAI Yulong*, *WANG Rui*, *ZHENG Shouxi*

School of Mechanical Engineering, Shijiazhuang Tiedao University, Shijiazhuang 050043, China
✉Corresponding author, E-mail: dengfy@stdu.edu.cn

ABSTRACT　To address the issues of image feature insensitivity, high requirement of expert experience, and low recognition accuracy of traditional machine vision methods in train wheelset bearing damage detection, this paper proposes an identification method based on the framework of convolutional and transformer fusion networks for identifying damage to train wheelset bearings. First, due to the complexity of train-bearing images, their category imbalance is more severe; an image preprocessing method called image enhancement category reorganization is used to improve the quality of the acquired image dataset and eliminate the effects of the imbalance dataset. Second, a convolutional neural network (CNN) has high model construction and training efficiency due to adopting a local sensing field and weight-sharing strategy, which can only sense local neighborhoods but has limited ability to capture global feature information. Transformer is a network model based on a self-attention mechanism. With strong parallel computing ability, it can learn the remote dependencies between image pixels in the global scope and has a more powerful global information extraction ability. However, the ability to mine the local features of the image is not sufficient. Therefore, this paper presents a VGG and transformer parallel fusion

收稿日期: 2024−01−02
基金项目: 国家自然科学基金资助项目(12272243); 河北省研究生案例库项目资助(KCJPZ2023037)

network that integrates the global contour features and local details of the image based on the fusion of convolution and self-attention. Furthermore, a multiscale dilation spatial pyramid convolution (MDSPC) module is constructed to fully mine the multiscale semantic features in the feature map using multiscale dilation convolution progressive fusion. The proposed method effectively solves the problem of feature information loss due to the mesh effect caused by the expansion convolution. Additionally, embedding coordinate attention (CA) after the MDSPC module can obtain remote dependencies and more precise positional relationships of feature images from two spatial directions, which can more accurately focus on specific regions in the feature map. Finally, experimental analyses were conducted using the NEU-DET image defect and self-constructed train wheelset bearing image datasets. The experimental results demonstrate that the proposed model has an accuracy of 99.44% and 98% for recognizing six types of defects and four types of images of wheelset bearings in NEU-DET data, respectively. The feature extraction capability of the proposed model was verified using model visualization methods. Compared with existing CNN models, ViT model with self-attention mechanism, and CNN-transformer fusion model, the proposed method shows significantly better evaluation metrics and accurately identifies different types of image samples without significantly increasing the model complexity.

KEY WORDS wheelset bearing; damage identification; convolutional network; transformer network; multi-scale feature

铁路交通运输具有运量大、成本低、占地少、绿色环保等优点，对促进国民经济发展具有举足轻重的作用. 轮对轴承作为高速列车走行部核心部件之一，长时间在恶劣的工况环境下运行，极易发生划伤、剥离、凹痕等多种损伤[1]. 因此，开展有效的高速列车轮对轴承故障运维研究，对保障列车安全、平稳运营具有重要的学术价值与经济意义.

当前，油液监测、温度监测、轨边声学监测、振动监测等方法被广泛用于轴承故障检测[2−4]，但用于列车轮对轴承故障检测仍存在较多限制：油液监测设备复杂，用于列车轮对轴承监测耗时长、成本高；轴箱内置的温度传感器误报警现象频发，难以监测轴承早期故障；轨边声学监测虽然布置、监测较为方便，但检测成本较高，车速升高后受多普勒效应影响准确率会明显降低；振动监测受安装空间狭小、缺少电源等条件限制，车载振动监测系统安装较为困难. 因此，现实场景中列车轮对轴承仍需要返厂拆解进行维修，根据人工目测及人为经验检测轴承表面缺陷损伤.

采用机器视觉的轴承缺陷损伤检测相比人工目测或凭借经验方法，具有速度快、成本低、自动化等优点，更符合企业现代化、智能化生产要求. 陈金贵等[5]采用阈值分割方法分析轴承图像后，采用改进的 Niblack 算法识别轴承滚子表面缺陷类型；陈昊等[6]先对轴承图像进行奇异值分解，消除灰度值异常影响，再通过金字塔分层细化方法和光流误差估计，确定轴承缺陷区域故障类型；王恒迪等[7]利用中值滤波、阈值分割、边缘检测处理列车轴承缺陷图像，根据正常与缺陷图像的灰度差值来确认轴承表面缺陷. 石炜等[8]基于二值化处理、形态学滤波和图像标记等算法处理列车轴承图像，采用分类决策树识别轴承不同类型缺陷；杨加东等[9]通过图像阈值分割和几何特征提取处理轴承损失图像，利用 BP 神经网络对列车轴承损伤类型进行分类. 上述机器视觉检测方法要求复杂的图像预处理操作，图像特征提取困难，决策树、BP 网络等传统机器学习模型自学习能力较差，往往难以满足列车轴承损伤检测的实际需求.

伴随人工智能技术快速发展，基于深度学习的图像检测方法得到了广泛关注，该方法不依赖较多的图像预处理操作及人为经验，能够自动从图像中学习相关特征信息，实现图像关键区域的快速识别. 蒋兴群等[10]提出了一种基于改进 YOLO-V3 模型，能对锚框尺度进行调整优化，准确识别了风机叶片的表面损伤；Dong 等[11]设计了一种金字塔特征融合与全局语义注意力网络模型，实现了缺陷图像较为精细地检测识别；Zheng 等[12]提出了链式不对称空间金字塔池化网络模型构建方法，通过链式不对称结构扩大网络感受野范围，在钢表面缺陷检测中取得了较好效果. 基于深度学习的图像检测方法应用前景巨大，但相比其他工业损伤图像，高速列车轮对轴承表面缺陷检测难度更大：(1)由于灰尘、油脂、锈蚀因素影响，轴承图像中噪声强烈，缺陷对比度低；(2)轴承缺陷异常复杂，部分类间缺陷在纹理、边缘方面相差不大，而类内缺陷在形状、大小方面差异很大；(3)轴承损伤随机性大，不同类型缺陷图像规模不一，类别不均衡较为严重. 因此，基于深度学习方法准确识别轴承缺陷仍存在诸多挑战，相关研究在轴承故障诊断领域仍鲜有报道.

为解决上述问题，本文提出采用图像增强类别重组方法提高列车轴承图像数据质量，并构建

了一种基于卷积与 Transformer 融合的网络模型框架, 用于列车轮对轴承损伤识别, 通过实验分析证实了所提方法的有效性. 所提方法的主要优势如下:

(1)通过图像增强类别重组方法处理轮对轴承缺陷图像, 一方面能够扩充图像样本数量; 另一方面对轴承不同类型缺陷图像进行数量均衡化处理, 解决样本不均衡问题.

(2)设计了 VGG 与 Transformer 双分支并行融合网络(VGG and transformer parallel fusion network, VTPF-Net), 同时嵌入了卷积运算与自注意力运算, 综合获取图像全局轮廓特征与局部细节特征信息.

(3)结合不同尺度膨胀卷积与坐标注意力(Coordinate attention, CA)机制, 构建了多尺度膨胀空间金字塔卷积(Multiscale dilation spatial pyramid convolution, MDSPC)模块, 更有效挖掘特征图中多尺度语义特征信息.

1　图像增强类别重组方法

深度神经网络模型虽然具有较强的图像特征学习能力, 但往往需要大量的图像数据来驱动网络模型训练. 为此常采用图像增强方式扩充图像样本数量, 本文采用的图像增强具体操作如下:

(1)针对每一类型样本图像, 随机选取 50% 样本进行水平翻转, 剩余 50% 样本进行垂直翻转;

(2)将水平、垂直翻转后图像与原始图像构成新的数据集, 样本数量增加一倍;

(3)在新数据集中随机选取 50% 样本, 然后平均分为 3 份, 分别对每一份进行对比度、色度与饱和度随机调整, 将调整后的新图像加入到数据集中, 相比原始图像, 样本数量再增加一倍.

相比传统的单一图像增强操作, 本文选取不同比例的样本图像, 分别进行不同方式的图像增强操作, 可以在扩充样本数量的同时, 进一步丰富图像内容, 改善图像质量. 真实工况中列车轮对轴承缺陷随机性较大, 不同损伤类型的图像样本数量差别较大. 样本不均衡会造成网络模型往往更倾向训练数量多的样本类型, 忽视数量较少的样本类型, 导致模型性能及泛化能力下降. 因此, 本文在图像增强基础上通过类别重组方法均衡各类型样本数量, 具体步骤如下:

步骤一: 类型排序与计数. 每一类型中的图像样本进行序号排序, 并设定对应标签类型;

步骤二: 随机排序重组. 根据各标签类型中样本数量最大值, 对每一种类型进行随机排序, 重组为一个新序列;

步骤三: 取余排序重组. 新序列与对应类型标签的样本数相除取余, 并对余数进行排序重组;

步骤四: 生成新样本. 样本数量最多的类型保持不变, 其余类型根据余数索引, 选取对应样本进行扩充.

为了能更明确说明类别重组方法的实现过程, 举例说明如下: 现有 3 种类型的图像集, 标签类型分别为 a、b、c, 对应样本数量为 4、3、2, 随机排序后取余数, 然后根据余数复制新样本. 标签 a 中样本数量最多, 保持不变; 标签 b 中余数新增是 0, 扩充样本序号 0; 标签 c 中余数新增是 0、1, 扩充样本序号 0、1. 3 种类型的原始样本数量彼此不同, 经过类别重组后, 各类型样本数量得到均衡, 整个过程如图 1 所示. 该方法通过随机取余数, 能够较为随机的均匀扩充样本数量, 操作简单便捷.

2　所提模型框架

2.1　VTPF-Net

卷积神经网络(Convolutional neural network, CNN)由于采用了局部感受野与权值共享策略, 模型构建和训练效率很高, 在图像识别领域应用非常广泛. VGG 网络作为典型的 CNN, 采用了相同大小的卷积核和池化核, 结构简洁, 使用尺寸为 3×3 卷积核能够有效提取图像中深层次的纹理、形状等细节特征信息. VGG16 作为最具代表性的 VGG 模型, 由 13 个卷积层、5 个最大池化层和 3 个全连接层构成[13], 网络结构如图 2(a) 所示. 虽然 VGG 网络在图像识别任务中表现出色, 但仍存在一些缺陷: 每一层卷积中感受野范围固定, 仅能感知局部的邻近区域, 对于全局特征信息的捕捉能力有限, 难以并行计算、长距离依赖性学习较差等等[14].

Transformer 是一种基于自注意力机制的网络模型, 并行计算能力强, 能够在全局范围内学习图像像素之间的远程依赖关系, 因此具有较为强大的全局信息提取能力[15−16]. Visual transformer (ViT) 基于 Transformer encoder 结构, 在大规模图像识别研究中展现出了较为优越的性能[17]. Transformer encoder 基本模块如图 2(b) 所示, 主要由多头注意力层和多层感知器(Multilayer perceptron, MLP)层两部分组成. 多头注意力层包括层归一化(Norm)和多头注意力机制. 注意力机制利用缩放点积注意力来实现查询矩阵 $\boldsymbol{Q}$ 到键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$ 的映射, 从而得到注意力值:

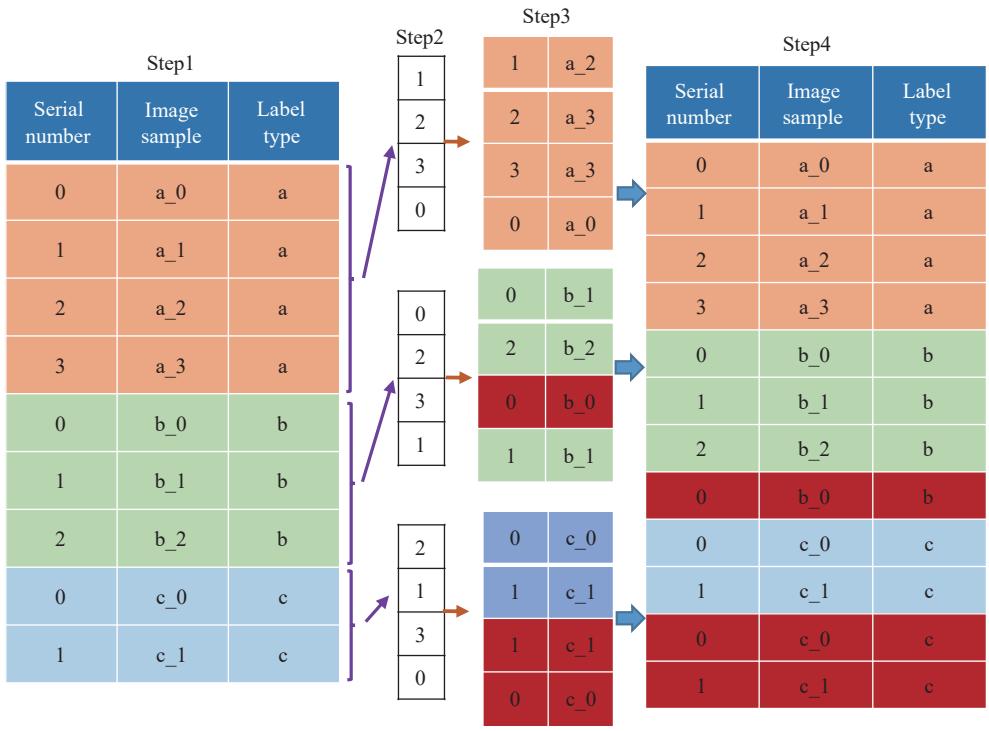

图 1 类别重组实例

Fig.1 Living example of class recombination

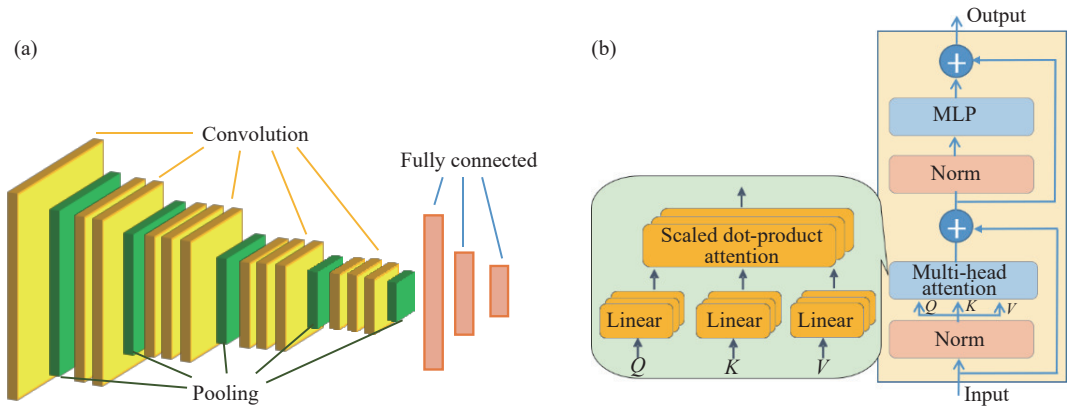

图 2 网络模型结构. (a) VGG16 结构; (b) transformer encoder 结构

Fig.2 Network model structure: (a) VGG16 structure; (b) transformer encoder structure

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Soft max}\left(\frac{\boldsymbol{QK}}{\sqrt{d}}\right)\boldsymbol{V} \tag{1}$$

$$\begin{cases} \boldsymbol{Q} = X_{\text{f}}\boldsymbol{W}^{Q} \\ \boldsymbol{K} = X_{\text{f}}\boldsymbol{W}^{K} \\ \boldsymbol{V} = X_{\text{f}}\boldsymbol{W}^{V} \end{cases} \tag{2}$$

式中: $\boldsymbol{W}^{Q}$、$\boldsymbol{W}^{K}$、$\boldsymbol{W}^{V}$分别是 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$对应的权重矩阵. $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 点积, 使用 Softmax 归一化计算权重系数, 然后再与 V 进行点乘得到结果. 多头注意力机制通过引入多组注意力权重, 使得模型可以同时关注输入中的不同位置和特征, 相比单头注意力机制具有更大的灵活性和表现能力[18]. MLP 层由层归一化和 MLP 组成. Transformer 模型中图像特征是通过自注意力机制计算得到, 每个位置的特征图表示都是基于整个输入序列的. 这种全局运算方式会忽视局部细节信息, 因此针对细粒度特征信息的捕捉能力不足.

基于上述分析, 本文将 CNN 与 Transformer 两者的优势进行结合, 提出了双分支并行融合 VTPF-Net, 结构如图 3 所示. Transformer encoder 分支具有较好的全局信息建模能力, 能够充分提取特征图中全局特征信息; VGG 分支具有较好的特征图局部特征提取能力. 需要指出的是: 在 VTPF-Net 中为了保证双分支输出特征图维度匹配, 保留了 VGG16 中较为重要的卷积层和池化层, 去掉后面

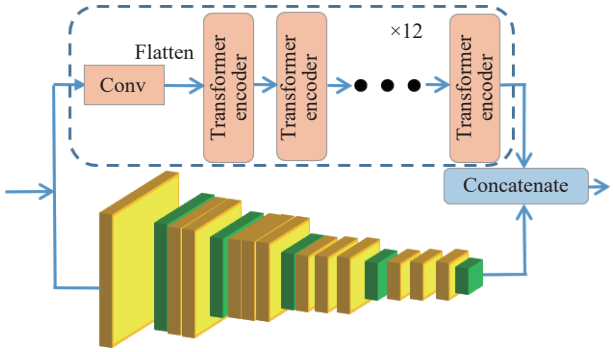

图 3　VTPF-Net 结构

Fig.3　Structure of the VTPF-Net

3 个全连接层.

2.2　多尺度膨胀空间金字塔卷积（MDSPC）

相比大尺寸卷积核卷积操作, 膨胀卷积 (Dilated convolution, Di-Conv) 操作同样能够实现较大的感受野范围, 并且不会带来额外运算负担. 基于上述思想, Chen 等[19] 提出了膨胀空间卷积池化金字塔 (Atrous spatial pyramid pooling, ASPP) 网络, 通过并行地应用不同尺度的膨胀卷积操作, 来捕获图像中不同级别的上下文信息, 实现特征图多尺度信息的有效感知. 然而, 膨胀卷积也会引起较为严重地网格效应[20], 例如: 膨胀率为 2 的 3×3 膨胀卷积感受野范围虽然与常规卷积 5×5 的感受野范围相同, 但是膨胀卷积的特征图中有部分像素点处于卷积的视觉盲区, 无法参与卷积过程, 导致特征信息的大量丢失. 基于此, 本文提出了 MDSPC 构建方法, 通过将多尺度膨胀卷积分支递进融合在一起, 改进了 ASPP 中膨胀卷积分支相互独立的结构, 有效解决了膨胀卷积网格效应造成的特征信息丢失问题. MDSPC 结构如图 4(a) 所示, 图中每个分支的膨胀卷积的卷积核大小固定为 3×3, 膨胀率逐渐增大, 各个分支之间通过 Concatenate 操作递进融合.

为了提高模型对关键图像目标的检测精度, 在 MDSPC 结构中还嵌入了 CA 机制, 具体结构如图 4(b) 所示. 不同尺度空洞卷积融合后的特征图分别在水平和垂直方向上执行全局池化操作, 从而获取特征图沿一个空间方向上的相关关系, 保存另一个空间方向的位置信息. 接着通过张量拼接和 ReLU 非线性激活操作得到水平和垂直方向上注意力张量, 然后切分为 2 个单独的张量, 利用单位卷积和 Sigmoid 激活得到注意力权重更新矩阵. 最终将获取的注意力权重与输入特征图相乘, 完成 CA 机制的施加. MDSPC 通过嵌入 CA 可以从两个空间方向上获取特征图像的远程依赖关系和较为精确的位置关系, 从而能够更准确地关注特征图中的特定区域.

样本图像经过图像增强类别重组方法处理后, 输入至所提的网络模型框架, 具体结构及详细参数如图 5 所示. 输入 VTPF-Net 模型的图像样本是三维矩阵格式, 在 Transformer 分支中, 先经过尺寸为 16×16 的卷积操作将图像进行 Patch 划分, 后进行展平操作转换为二维矩阵格式输入至串行连接的 Transformer encoder, 最后通过 view 操作进行维度重构, 输出特征图维度为 (768, 14×14); 在卷积分支中, 图像样本直接输入至 VGG 模型分支中, 输出特征图维度为 (512, 14×14), 两分支输出特征图经过 Concatenate 操作进行融合, 最终 VTPF-Net 输出特征图维度为 (1270, 14×14), 然后输入至 MDSPC, 最后经过 MLP 分类层进行多标签分类, 从而识别分类结果.

3　实验分析

3.1　数据集介绍

为了验证所提方法有效性, 本文采用 2 类图像数据分别进行实验测试, 第一类是 NEU-DET 图像数据集, 第二类是自建的列车轮对轴承图像数据集. NEU-DET 数据集是东北大学公开发布的钢材

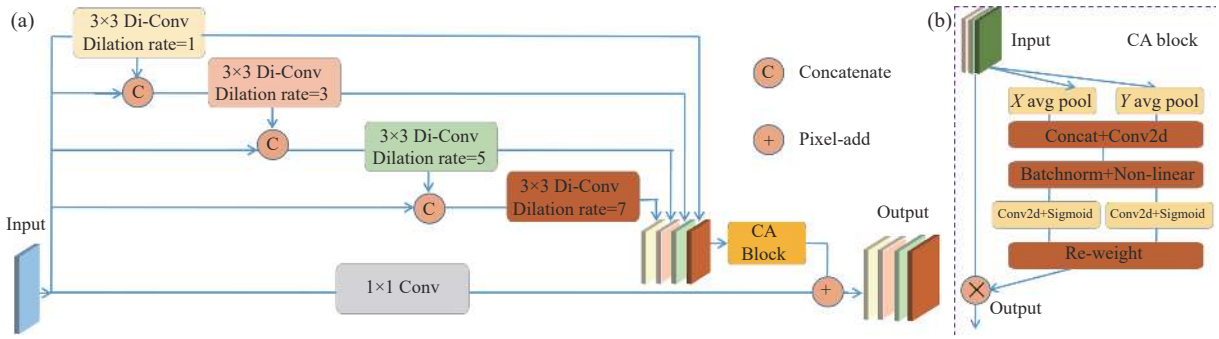

图 4　MDSPC 的网络结构. (a) MDSPC 结构 (b) CA 结构

Fig.4　Network structure of MDSPC: (a) MDSPC structure; (b) CA structure

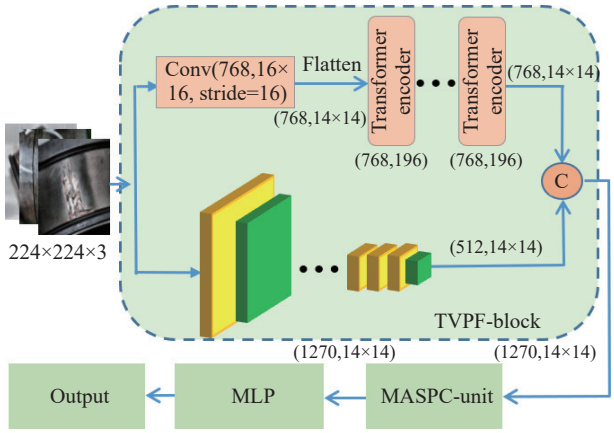

图 5 所提模型框架

Fig.5 Framework of the proposed model

表面缺陷图像数据集[21]. 该数据集包含 6 类表面缺陷图像, 分别是斑块(Pa)、轧制氧化皮(Rs)、划痕(Sc)、开裂(Cr)、内含物(In)和点蚀(Ps), 每一种缺陷包含 300 张像素大小为 200×200 的灰度图像. 列车轮对轴承图像数据集是自建的数据集, 所拍摄的各类轴承全部来自中车石家庄车辆有限公司列车轮对轴承检修流水线. 数据集共包含 4 种类型轴承样本图像, 分别是正常、划伤、剥落与电蚀, 这些轴承损伤均是在列车长时间运行过程中所造成的, 部分图像如图 6 所示. 其中, 正常样本数量最多为 150 张、划痕 130 张、剥落 120 张、电蚀 100 张.

3.2 数据预处理

鉴于 NEU-DET 数据中样本图像质量较为理想, 并且每一类缺陷图像样本数量相同, 因此无需采用本文所提的图像增强类别重组方法进行预处理, 可以直接输入网络模型进行分类识别. 列车轮对轴承出现故障损伤具有随机性强、概率性小等特点, 相比正常轴承图像样本, 各个缺陷图像样本数量较少, 并且不同缺陷类型图像数量相差很大. 在每类图像中随机选取 20% 图像样本作为测试集, 剩余作为训练集. 模型训练和测试过程中, 为保证测试分析的准确性, 测试集图像保持不变, 仅对训练集图像进行图像增强类别重组预处理, 其中轴承正常图像经过增强操作后, 样本数量扩充为 360 张, 划痕、剥落与电蚀样本分别扩充至 312 张、288 张、240 张, 然后通过类别重组方法平

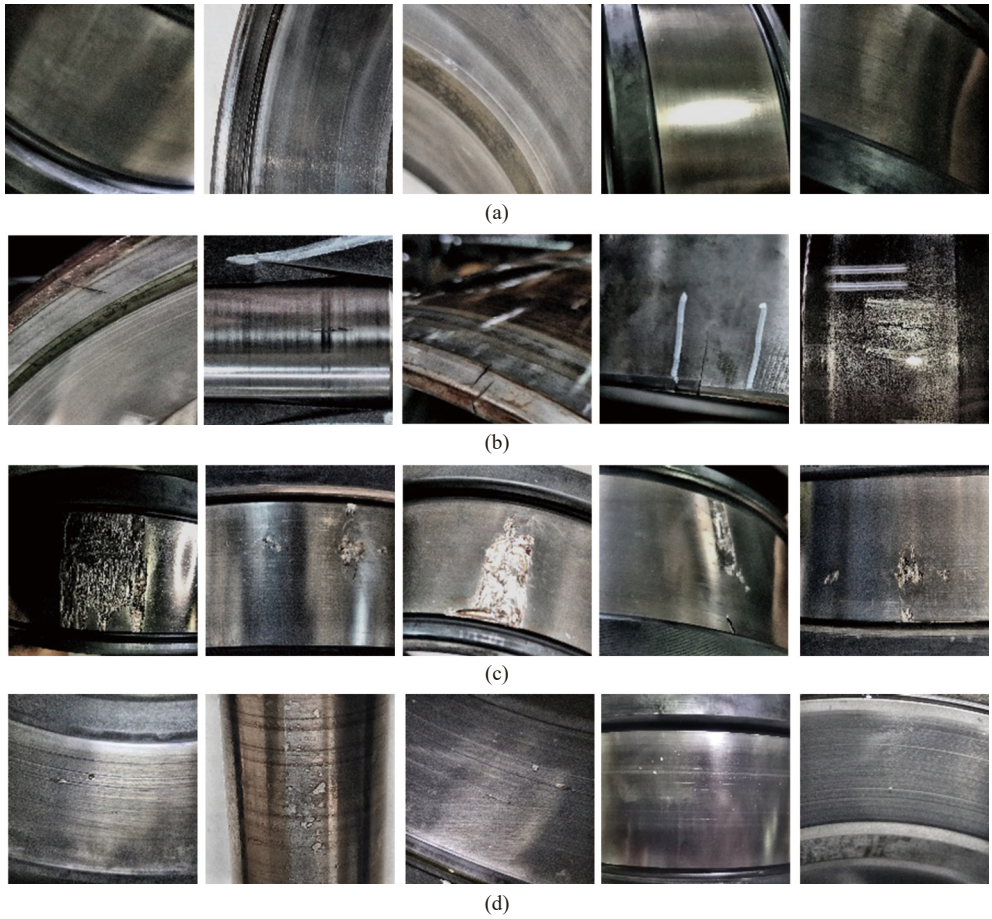

图 6 列车轴承图像. (a) 正常; (b) 划痕; (c) 剥落; (d) 电蚀

Fig.6 Train bearing images: (a) normal; (b) scratch; (c) spalling; (d) pitting

衡每一类图像样本数量, 每一类别均包含 360 张图像, 训练集共有 1440 张图像. 针对上述两个数据集, 分别在训练集中随机选取 12.5% 图像样本作为验证集.

3.3　实验环境和评价指标

采用 Pytorch1.7.1 深度学习框架和 Python3.8.0 搭建实验分析环境, 计算机 CPU 为 AMD Ryzen 7, GPU 为 NVIDIA GeForce RTX 3070 8G. 为了更好地评价模型测试结果, 本文选取文献 [21] 中准确率(Acc)、精准率(Pre)、召回率(Re)和 F1 分数等 4 类指标作为分析所提方法测试结果评价指标.

4　实验结果分析

4.1　缺陷分类实验

使用 NEU-DET 数据集进行实验分析, 所提网络模型训练过程中准确率 δ 和损失 Δ 随迭代次数增加的变化过程如图 7(a)所示, 从图中可知, 模型训练 30 次迭代之后, 模型的准确率和损失值变化趋于收敛, 模型训练过程没有出现过拟合现象, 性能可靠稳定. 模型测试结果的混淆矩阵如图 7(b)所示, 从中可知, 缺陷类别 Pa、Sc、Cr、Rs 中图像均能够正确识别, 点蚀(Ps)和内含物(In)两类缺陷部分样本特征较为相似, 因此各有 1 张图像被彼此错误识别, 测试准确率为 99.44%. 测试结果显示, 本文所提模型能够较为准确地识别 6 类不同缺陷图像.

采用自建的列车轮对轴承图像数据集进行分析, 模型训练过程中准确率 δ 和损失 Δ 的变化曲线如图 8(a)所示, 从图中可知, 模型训练在迭代 30 次后基本达到收敛状态. 模型训练及验证完成后, 进行测试分析, 测试结果的混淆矩阵如图 8(b)所示. 从中可知, 列车轴承剥落故障和正常两个类别的图像能够完全准确识别. 电蚀和划痕 2 类轴承图像中部分缺陷特征不明显, 各有 1 张图像被错误识别, 测试集整体识别准确率为 98%. 上述分析表明, 所提模型能够较为准确地识别列车轴承不同缺陷类型图像.

4.2　模型可视化与热力图分析

针对列车轮对轴承图像数据集, 使用 t 分布–随机近邻嵌入(t-Distributed stochastic neighbor em-

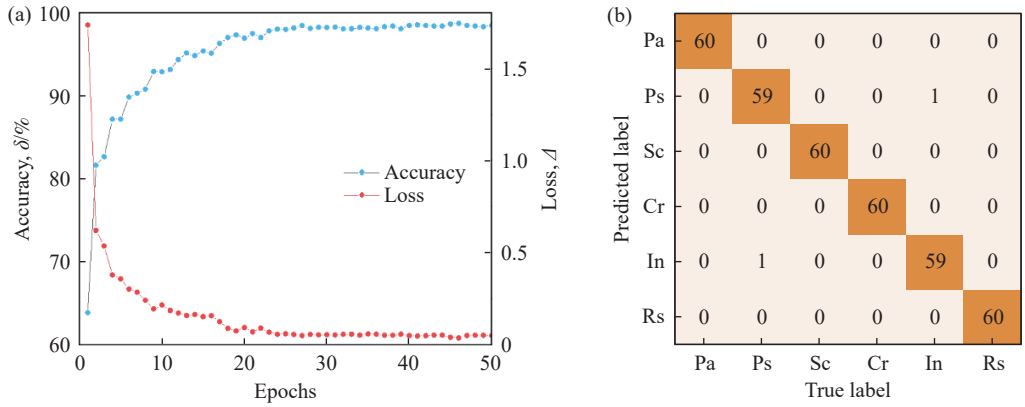

图 7　NEU-DET 数据集实验结果. (a) 模型性能曲线; (b) 混淆矩阵结果

Fig.7　Experimental results of the NEU-DET dataset: (a) model performance curve; (b) confusion matrix results

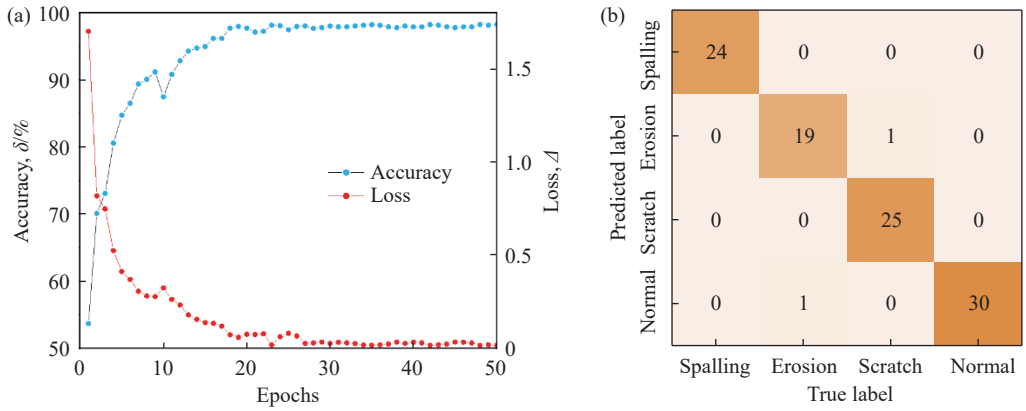

图 8　列车轮对轴承数据集实验结果. (a) 模型性能曲线; (b) 混淆矩阵结果

Fig.8　Experimental results of the train wheelset bearing dataset: (a) model performance curve; (b) confusion matrix results

bedding, *t*-SNE)方法, 将高维图像特征映射到二维空间, 通过可视化分析来说明模型不同阶段图像特征学习、聚类过程. 选取所提网络模型中 4 个不同阶段开展分析, 分别为模型输入端、VGG 与 Transformer 分支后、MDSPC 后, 可视化结果如图 9 所示. 不同类型的图像特征在模型输入端完全混合在一起, 难以区分. 伴随网络深度增加, 在 VGG 分支与 Transformer 分支后, 不同类型的图像特征点逐渐分离, 相同类型图像特征点聚类在一起, 在 MDSPC 后, 不同类型图像特征实现了较为准确地聚类分离.

下面进一步研究所提模型捕捉列车轮对轴承缺陷图像特征的能力, 采用梯度加权类激活映射(Gradient-weighted class activation map, Grad-CAM)[22] 方法开展图像激活热力图分析, 即通过对图像不同区域受关注程度的变化来对模型识别过程做出可解释性分析, 颜色越红的图像区域代表在识别过程中影响越大. 针对 4 种不同类型图像, 分别在 VTPF-Net 后和 MDSPC 后开展热力图分析, 结果如图 10 所示. 从图中可知, 在模型 VTPF-Net 阶段热力图颜色覆盖范围较大, 不仅包括图像缺陷区域, 还有其他区域. 而在 MDSPC 后热力图颜色已经主要聚焦到图像的缺陷区域, 模型关注的图像特征转向缺陷区域. 这说明 MDSPC 模块可以有效提升模型的整体性能, 一方面利用其多尺度膨胀卷积递进融合充分挖掘特征图中的多尺度语义特征; 另一方面通过 CA 机制提高了对关键图像目标的检测精度, 能够更准确地关注特征图中的特定区域.

4.3 消融实验及复杂度分析

为了进一步验证本文所提网络模型构成的合理性, 以及 MDSPC、VTPF-Net、VGG 与 Transformer 等模块对网络模型性能的影响, 设计了包含不同模块以及各个模块组合的消融实验, 使用 Acc、Pre、Re、F1 分数值作为评价指标, 针对 NEU-DET 数据和列车轮对轴承图像数据集进行测试分析, 消融实验结果如表 1 所示. 仅采用 VGG、Transformer 模块时, 模型测试结果较差, 分别结合 MDSPC 模块后, 测试结果准确性显著提升. VTPF-Net 融合了 VGG 与 Transformer 的优点, 其测试结果要明显优于单一的 VGG、Transformer 模型, 并且结合 MDSPC 模块后取得了最优的测试结果, 这种组合也正是本文所提网络模型结构. 本文所提 VTPF-Net 包含

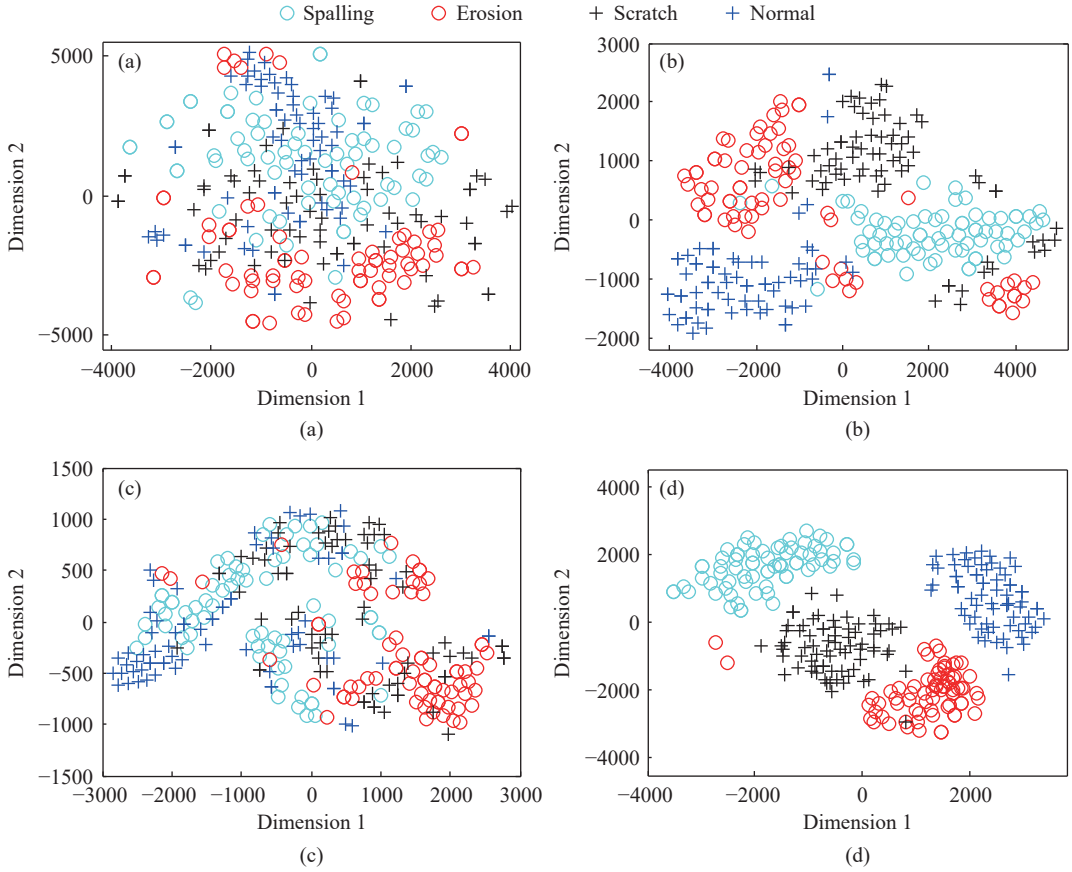

图 9 可视化结果. (a) 输入; (b) VGG 分支输出; (c) transformer 分支输出; (d) MDSPC 分支输出

Fig.9 Visualization results: (a) input; (b) output of VGG branch; (c) output of transformer branch; (d) output of MDSPC branch

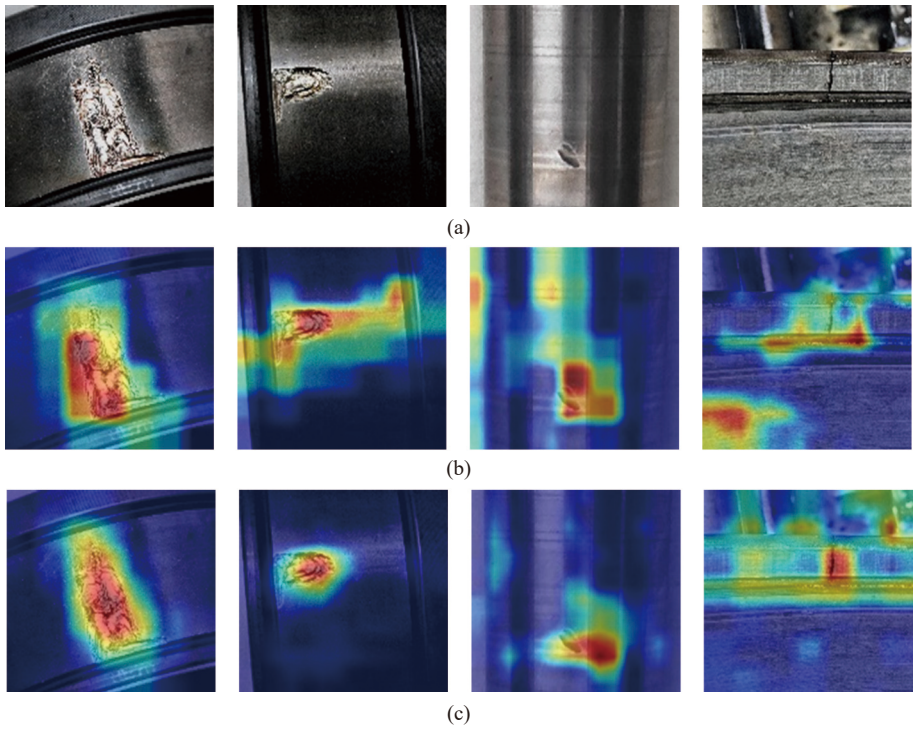

图 10　Grad-CAM 结果. (a) 原图; (b) VTPF-Net; (c) MDSPC

Fig.10　Grad-CAM results: (a) original images; (b) VTPF-Net; (c) MDSPC

表 1　消融实验结果

Table 1　Ablation experiment results

VGG	Transformer	VTPF-net	MDSPC	NEU-DET dataset/%				Train wheelset bearing image dataset/%			
				Acc	Pre	Re	F1-score	Acc	Pre	Re	F1-score
√				92.22	92.22	92.46	92.17	91.00	90.86	90.85	90.96
	√			91.28	91.28	91.51	91.27	90.00	89.54	90.02	89.95
√			√	95.54	95.54	95.66	95.53	95.00	94.54	94.91	95.03
	√		√	94.99	94.99	95.09	94.96	94.00	93.91	93.78	94.00
		√		97.77	97.77	97.83	97.76	96.00	95.50	96.03	96.01
		√	√	**99.44**	**99.44**	**99.44**	**99.44**	**98.00**	**97.79**	**97.94**	**98.00**

了 VGG 与 Transformer 网络, 利用卷积运算与自注意力运算能够综合获取图像局部细节特征与全局轮廓特征信息, 同时结合 MDSPC 模块增强了对特征图多尺度语义信息的挖掘能力, 提高了对关键图像目标的检测精度, 进一步增强了模型的性能.

下面对所提网络模型中各个模块的复杂度及效率进行分析, 分别选取模型训练参数量(Params)、浮点运算次数(Floating-point operations, FLOPs)作为复杂度评价指标, 模型训练时间作为效率评价指标, 针对列车轮对轴承图像数据集开展分析, 结果如表 2 所示. 从表中可知, 所提 VTPF-Net 相比单一的 VGG 及 Transformer 网络, 模型训练时 Params 及 FLOPs 至多增加了 0.9%、14.3%, 训练时间增加了 17.76%. 结合 MDSPC 模块后, 所提模型虽然 Params 及 FLOPs 进一步增加, 相比单一的 VGG 及 Transformer 网络增加了 38.2%、20.13%, 模型训练时间为 1591.23 s. 分析可知, 相比传统的 VGG 及 Transformer 模型, 所提网络模型复杂度增加, 模型训练效率下降, 但模型复杂度增加并不显著, 模型训练时间也在可接受范围内.

4.4　对比实验分析

为验证本文所提模型的综合性能, 选取基于 CNN 框架的 VGG16[13]、GoogleNet[23]、ResNet50[24], 基于自注意力机制的 ViT 模型[17] 以及文献 [25] 中提出的 CNN-Transformer 融合模型进行对比试验. 分别针对 NEU-DET 数据集和列车轮对轴承图像数据集进行对比分析, 各个模型对比测试结果如表 3 所示, 基于传统 CNN 的 VGG16、GoogleNet 和

表 2 模型复杂度分析

Table 2 Model complexity analysis

VGG	Transformer	VTPF-Net	MDSPC	Train wheelset bearing image dataset		
				Params/10^6	FLOPs/10^9	Time/s
√				83.96	15.61	1151.34
	√			84.33	14.83	1324.61
		√		84.52	16.95	1355.74
		√	√	**84.72**	**25.2**	**1591.23**

表 3 对比方法实验结果

Table 3 Experimental results of comparative methods

Dataset	Model	Acc/%	Pre/%	Re/%	F1-score
NEU-DET dataset	VGG16	92.22	92.22	92.46	92.17
	GooleNet	92.89	92.89	92.76	92.82
	ResNet50	95.88	95.88	96.17	95.44
	ViT	91.28	91.28	91.51	91.27
	Literature [25]	97.8	97.8	98.03	97.79
	Proposed method	**99.44**	**99.44**	**99.44**	**99.44**
Train wheelset bearing image dataset	VGG16	91.00	90.86	90.85	90.96
	GooleNet	93.00	93.08	93.03	92.98
	ResNet50	95.00	94.74	95.19	95.00
	ViT	90.00	89.54	90.02	89.95
	Literature [25]	97.00	96.75	97.19	97.00
	Proposed method	**98.00**	**97.79**	**97.94**	**98.00**

ResNet50 模型中, VGG16 和 GoogleNet 网络结果较为相近, ResNet50 的测试结果相对更好一些. VGG16 与 GoogleNet 采用的是常规卷积操作, ResNet50 在此基础上采用了残差结构, 解决了网络深度增加引起的梯度消失和模型性能退化问题, 所以模型测试结果有所提升. ViT 模型结果较差, 这是因为 ViT 模型性能的提升建立在大量数据样本基础上, 该实验中图像样本数量有限, 导致 ViT 模型结果与其他对比方法相比稍微差一些. CNN-Transformer 模型融合了 CNN 与 Transformer 的优势, 模型构建基本思想与本文所提方法相同, 并且引入了通道注意力 (Squeeze and excitation, SE) 模块, 在对比方法中测试效果最好, 但由于该模型采用的是较为简单的浅层 CNN 结构, 未考虑多尺度特征信息的挖掘, 测试结果低于本文所提模型. 上述对比实验分析进一步证实了本文所提模型的优越性.

5 结论

基于深度学习与机器视觉的列车轴承缺陷图像识别技术在轨道交通智能运维领域具有重要的研究意义和应用价值. 本文提出了一种图像增强类别重组方法与卷积 Transformer 融合框架用于列车轮对轴承表面缺陷检测识别, 相关结论如下.

(1) 图像增强类别重组方法对轴承图像进行预处理后, 提高了数据集多样性与模型泛化能力, 消除了不同类别图像样本数量不均衡的问题.

(2) 构建的 VTPF-Net 能够同时进行卷积运算与自注意力运算, 融合了卷积与 Transformer 网络的优势, 能够有效获取图像不同语义下的全局与局部特征信息, 可以更准确地用于图像特征提取与识别.

(3) MDSPC 通过将多尺度空洞卷积分支递进融合在一起, 更有效挖掘特征图多尺度特征, 克服了网格效应造成的信息丢失, 同时引入 CA 机制实现了对特征图中特定区域的重点关注.

(4) 通过对 NEU-DET 数据集与自建列车轮对轴承图像数据集进行测试, 表明所提模型能够有效用于不同类型数据集的准确识别, 各项评价指标要优于当前的 CNN 模型、基于自注意力机制的 ViT 模型以及 CNN-Transformer 融合模型.

参 考 文 献

[1] Deng F Y, Wang H L, Gao R Y, et al. Vibration characteristics analysis of the inner race fault of axlebox bearing under wheel-rail excitation. *J Hebei Univ (Nat Sci Ed)*, 2023, 43(6): 561
(邓飞跃, 王红力, 高瑞洋, 等. 轮轨激励条件下轴箱轴承内圈故障振动特性分析. 河北大学学报(自然科学版), 2023, 43(6): 561)

[2] Zhang S B, He Q B, Zhang H B, et al. Doppler correction using short-time MUSIC and angle interpolation resampling for wayside acoustic defective bearing diagnosis. *IEEE Trans Instrum Meas*, 2017, 66(4): 671

[3] Sun R B, Yang Z B, Zhai Z, et al. Sparse representation based on parametric impulsive dictionary design for bearing fault diagnosis. *Mech Syst Signal Process*, 2019, 122: 737

[4] Krishna B M V, Vishwakarma D M. A Review on vibration-based fault diagnosis in rolling element bearings. *Int J Appl Eng Res*, 2018, 13(8): 6188

[5] Chen J G, Chen H, Zhang B. Research on surface defect detection of bearing roller based on improved niblack algorithm. *Modul Mach Tool Autom Manuf Tech*, 2018(12): 82
(陈金贵, 陈昊, 张奔. 基于改进 Niblack 算法的轴承滚子表面缺陷检测. 组合机床与自动化加工技术, 2018(12): 82)

[6] Chen H, Zhang B, Li M, et al. Surface defect detection of bearing roller based on image optical flow. *Chin J Sci Instrum*, 2018, 39(6): 198
(陈昊, 张奔, 黎明, 等. 基于图像光流的轴承滚子表面缺陷检测. 仪器仪表学报, 2018, 39(6): 198)

[7] Wang H D, Li S, Deng S E, et al. Research on visual inspection algorithm of bearing outer ring side defects. *Mach Des Manuf*, 2017(12): 169
(王恒迪, 李莎, 邓四二, 等. 轴承外圈侧面缺陷的视觉检测算法研究. 机械设计与制造, 2017(12): 169)

[8] Shi W, Zhang Y X, Li J N. Research on machine vision detection method for surface defects of train roller bearings. *Mach Des Manuf*, 2022(4): 183
(石炜, 张袁祥, 李嘉楠. 列车滚子轴承表面缺陷机器视觉检测方法研究. 机械设计与制造, 2022(4): 183)

[9] Yang J D, Xie M, Wang L H, et al. Surface defect detection and classification based on BP neural network. *Mach Tool Hydraul*, 2017, 45(16): 160
(杨加东, 谢明, 王丽华, 等. 基于 BP 神经网络的表面缺陷检测分类. 机床与液压, 2017, 45(16): 160)

[10] Jiang X Q, Liu B, Song L, et al. Surface damage detection and recognition of wind turbine blade based on improved YOLO-v3. *Acta Energiae Solaris Sin*, 2023, 44(3): 212
(蒋兴群, 刘波, 宋力, 等. 基于改进 YOLO-v3 的风力机叶片表面损伤检测识别. 太阳能学报, 2023, 44(3): 212)

[11] Dong H W, Song K C, He Y, et al. PGA-net: Pyramid feature fusion and global context attention network for automated surface defect detection. *IEEE Trans Ind Inform*, 2020, 16(12): 7448

[12] Zheng Z Z, Hu Y H, Zhang Y, et al. CASPPNet: A chained atrous spatial pyramid pooling network for steel defect detection. *Meas Sci Technol*, 2022, 33(8): 085403

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J/OL]. *arXiv preprint* (2015-03-10) [2024-01-02]. https://arxiv.org/abs/1409.1556

[14] Liu Y T, Zhao J J, Luo Q Y, et al. Automated classification of cervical lymph-node-level from ultrasound using Depthwise Separable Convolutional Swin Transformer. *Comput Biol Med*, 2022, 148: 105821

[15] Tian T L, Song C, Ting J, et al. A french-to-english machine translation model using transformer network. *Procedia Comput Sci*, 2022, 199: 1438

[16] Bai P R, Wang R, Liu Q Y, et al. DS-YOLOv5: A real-time detection and recognition model for helmet wearing. *Chin J Eng*, 2023, 45(12): 2108
(白培瑞, 王瑞, 刘庆一, 等. DS-YOLOv5: 一种实时的安全帽佩戴检测与识别模型. 工程科学学报, 2023, 45(12): 2108)

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J/OL]. *arXiv preprint* (2015-06-03) [2024-01-02]. http://arxiv.org/abs/2010.11929

[18] Alexakos C T, Karnavas Y L, Drakaki M, et al. A combined short time Fourier transform and image classification transformer model for rolling element bearings fault diagnosis in electric motors. *Mach Learn Knowl Extr*, 2021, 3(1): 228

[19] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans Pattern Anal Mach Intell*, 2018, 40(4): 834

[20] Yang L, Gu Y G, Bian G B, et al. TMF-net: A transformer-based multiscale fusion network for surgical instrument segmentation from endoscopic images. *IEEE Trans Instrum Meas*, 2023, 72: 5001715

[21] Lu Y, Jiang M F, Wei L Y, et al. Automated arrhythmia classification using depthwise separable convolutional neural network with focal loss. *Biomed Signal Process Contr*, 2021, 69: 102843

[22] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization // 2017 *IEEE International Conference on Computer Vision* (*ICCV*). Venice, 2017: 618

[23] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions // 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Boston, 2015: 1

[24] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 770

[25] Tang D L, Yang Z, Cheng H, et al. Metal defect image recognition method based on shallow CNN fusion transformer. *China Mech Eng*, 2022, 33(19): 2298
(唐东林, 杨洲, 程衡, 等. 浅层卷积神经网络融合 Transformer 的金属缺陷图像识别方法. 中国机械工程, 2022, 33(19): 2298)