

铁路货车故障图像智能识别数据集构建技术

刘丹丹, 金佳鑫, 庞博

哈尔滨市科佳通用机电股份有限公司, 黑龙江 哈尔滨 150000

DOI: 10.61369/TACS.2025140023

摘 要 : 随着大数据和机器视觉技术的迅速发展与应用, 深入探究铁路货车故障图像的智能识别与检测, 推动人工检车向智能检车方向发展, 成为提高铁路货车故障诊断效率和质量的重要问题。但是, 由于铁路货车故障场景具有复杂性和多样性, 且受铁路运输安全性、连续性的特殊影响, 数据集构建是目前亟待解决的瓶颈问题。本文结合铁路货车故障检测实际需求, 围绕数据预处理、数据标注、数据划分、数据增强四个环节, 探讨数据集构建的关键技术, 提出数据集质量评估与应用方案, 旨在加快铁路货车故障诊断智能化发展进程, 为提高复杂场景下故障识别精度与效率提供参考。

关 键 词 : 铁路货车故障; 图像智能识别; 数据集构建技术; 技术应用

Technology for Constructing an Intelligent Recognition Dataset of Railway Wagon Fault Images

Liu Dandan, Jin Jiaxin, Pang Bo

Harbin Kejia General Mechanical and Electrical Co., Ltd., Harbin, Heilongjiang 150000

Abstract : With the rapid development and application of big data and machine vision technology, in-depth research on the intelligent recognition and detection of railway wagon fault images, and the promotion from manual inspection to intelligent inspection have become important issues to improve the efficiency and quality of railway wagon fault diagnosis. However, due to the complexity and diversity of railway wagon fault scenarios, as well as the special requirements of railway transportation for safety and continuity, dataset construction has become a bottleneck problem urgently to be solved. Combined with the actual needs of railway wagon fault detection, this paper discusses the key technologies of dataset construction around four stages: data preprocessing, data annotation, data partitioning and data augmentation, and puts forward a scheme for dataset quality evaluation and application. It aims to accelerate the intelligent development of railway wagon fault diagnosis and provide a reference for improving the accuracy and efficiency of fault recognition in complex scenarios.

Keywords : railway freight car faults; image intelligent recognition; dataset construction technology; technology application

引言

在高质量发展背景下, 铁路货运需求不断增长, 对故障诊断效率与质量提出了更高要求。而传统人工车检对人工依赖性强, 难以兼顾检测效率与质量。在深度学习、机器学习等图像识别技术的支持下, TFDS 货车图像检测系统逐步推广和应用, 但存在样本不均衡、覆盖场景不全面、数据冗余等问题。由此, 构建规范化的数据集, 对提高故障推理性能与识别精度, 实现全天候、高精度的故障检测具有重要意义。

一、铁路货车故障图像智能识别数据集构建的基础前提

数据集构建需以明确的需求分析、规范的标准制定与充足的原始数据采集为基础, 三者相互支撑, 保障构建工作有序开展。

(一) 数据集构建需求分析

需求分析是主要前提。从故障检测场景与模型训练需求出发, 明确适配需求。故障场景包含制动系统、转向架、车轮等关

键部件的故障, 覆盖常见与罕见故障类型, 避免模型过拟合^{[2][7]}。场景需覆盖晴雨天、昼夜等, 提升模型抗干扰能力^{[1][9]}。数据需满足准确、完整、一致、清晰的质量要求, 且质量控制贯穿全过程^{[6][10]}。需适配 YOLO、DETR、DINO 等模型, 兼顾关键点检测的速度与精度, 统一图像尺寸、标注格式并划分数据集^[8]。

(二) 数据集构建标准制定

制定规范化的构建标准。采集标准明确高清相机参数、拍摄要求及含关键信息的命名规则^{[2][7]}。分类采用“部件-故障-场

景”三级规则并分配唯一编码^{[5][8]}。标注标准明确故障区域框选、标注格式及故障分级规范^{[6][10]}。存储标准采用分层管理、高可靠设备及分布式存储，建立数据管理系统^{[1][2][5]}。

（三）原始图像数据采集

整合原始图像采集渠道。对接 TFDS、TVDS 系统提取故障图像作为核心数据源，补充正常部件负样本^{[2][7]}。现场人工采集现有系统未覆盖的罕见故障、复杂场景及低清晰度图像，记录相关信息^[8]。借助公开数据集与相关文献图像补充样本，预处理标注后纳入^{[3][4][6]}。采集后需剔除模糊、失真图像，保留合格样本用于后续预处理^[10]。

二、铁路货车故障图像智能识别数据集构建的关键技术环节

数据集构建的核心是数据预处理、标注、划分、增强四大环节，四者依次递进、相互关联，共同决定数据集质量与实用性。

（一）数据预处理技术

原始图像存在噪声、分辨率不一致、失真等问题，难以直接用于模型训练。预处理的核心是消除无效信息、提升图像质量、统一格式、突出故障特征，为后续工作奠定基础。

1. 图像去噪处理

采集过程中，图像易受干扰产生高斯噪声、椒盐噪声，导致模糊^[1]。常用高斯滤波、中值滤波、双边滤波算法，其中双边滤波在去噪同时能保留边缘细节，适用于铁路货车故障图像去噪，避免破坏故障边界特征^[9]。

2. 图像尺寸标准化处理

原始图像分辨率与尺寸差异大，会降低模型训练效率^[8]。需遵循“不破坏故障特征、不失真”原则，采用双线性插值算法调整尺寸，统一为640×480等指定规格，确保图像平滑无锯齿，便于模型批量处理^{[8][9]}。

3. 图像光照均衡化处理

不同光照下的图像亮度差异大，故障特征不明显。常用直方图均衡化、CLAHE 算法，CLAHE 算法将图像分区域均衡化，可解决全局均衡化的过亮过暗问题，适合夜间、阴天等光照不足场景，能有效突出故障特征，贴合数字图像识别技术应用需求^[1]。

4. 图像失真校正处理

拍摄角度偏差、镜头畸变会导致图像几何失真，影响故障定位^[7]。常用几何变换校正、镜头畸变校正算法，通过平移、旋转等变换恢复图像真实形态，如校正倾斜的转向架图像，便于故障框选标注。

5. 图像分割处理

图像背景会干扰模型学习，分割目的是提取含故障的感兴趣区域，减少无效干扰。常用阈值分割、边缘检测、语义分割算法，其中语义分割可精准分割故障区域、非故障部件与背景，适合复杂故障场景，优化处理效果^[4]。

6. 数据格式转换处理

格式统一是数据集构建的重要环节^[10]。原始图像格式往往存

在差异。需统一格式要求，转换为 JPG、PNG 等支持格式。同时将灰度值归一化至 [0,1] 范围，减少数值干扰。批量命名图像并关联标注文件，便于后续操作。

（二）数据标注技术

铁路货车故障图像类型多、精度要求高，数据标注要对经过预处理图像添加故障类型、区域等标注信息，使模型能有效学习故障特征。

1. 标注工具选择

常用标注工具有 LabelImg、Labelme、VIA 等。LabelImg 操作简单，适合目标检测标注。LabelMe 支持多种标注类型，适配复杂故障区域。VIA 支持批量与多人协作标注，提升大规模数据集标注效率^[9]。在实际应用中，采用“LabelImg+VIA”的组合方式，结合模型训练平台开发专用标注模块^[8]。

2. 标注流程实施

按规范流程标注，确保准确性与一致性。一是标注前培训，统一标注人员理解，考核合格后方可参与^[9]。二是初步标注，框选故障区域、标注相关信息，模糊图像暂存审核^[7]。三是双人审核，分别负责检查标注准确性与规范性，不合格标注退回修正^[10]。四是抽样修正，确保标注完整准确，关联存储图像与标注文件，可引入自动审核机制提升效率^[3]。

3. 标注难点与解决方案

针对标注三大难点，制定针对性解决方案。故障区域小、边界模糊，可采用高分辨率图像，通过增强技术突出边界，放大局部区域框选，安排经验丰富人员标注^[9]。故障类型相似易混淆，需制定详细区分标准，加强样本对比培训，审核时重点核查^{[6][10]}。大规模标注效率低、一致性差，可采用多人协作标注，引入半自动化标注技术，建立一致性评估机制，提高效率^[6]。

（三）数据划分技术

按“随机划分、比例合理、分布均匀”原则，将数据集划分为训练、验证、测试集，分别用于模型训练、验证与性能评估，确保划分合理，避免数据偏差。

1. 划分比例确定

常用比例为训练集70%~80%、验证集10%~15%、测试集10%~15%。大规模数据集采用80%、10%、10%比例，小规模数据集采用70%、15%、15%比例，且需确保每类故障与场景样本在三类数据集中分布均匀，确保训练效果^{[8][10]}。

2. 划分方法选择

优先采用分层随机划分法，按故障、场景类型分层，在各子集中随机划分，确保分布均匀。随机划分法操作简单但可能分布不均。时序划分法仅适用于特定训练需求^[9]。实际应用中，采用“分层随机划分+K折交叉验证”方式，提升模型训练稳定性^[8]。

3. 划分后验证

验证内容包括：样本分布与原始数据集一致、无模糊误标等无效数据、数据集适配模型训练。验证合格后分别存储，建立说明文档，记录划分相关信息，便于后续复用^[10]。

（四）数据增强技术

数据集常存在样本不足、罕见故障稀缺等问题，易导致模型

过拟合^{[6][8]}。数据增强通过变换现有标注数据生成新样本，扩大规模、丰富多样性，提升模型泛化能力，需遵循“不改变故障特征、不引入虚假故障”原则，分为传统与深度学习两类方法^[9]。

1. 传统数据增强方法

通过几何、像素变换生成新样本，操作简单高效^[10]。几何变换包括旋转、平移、缩放等，模拟不同拍摄角度与尺寸的故障图像；像素变换包括亮度、对比度调整与噪声添加，模拟不同光照与环境干扰，该方法适用于样本量扩充^{[11][9]}。

2. 深度学习数据增强方法

传统方法样本多样性有限，深度学习方法可生成更真实多样的样本^{[6][9]}。DiD 通过扩散模型和多尺度图像块技术恢复低光照图像，生成高质量罕见故障样本。VAE 通过编码解码生成多样化场景样本。迁移学习利用预训练模型提取特征，适配不同场景样本，提升模型识别效果^[8]。

3. 数据增强策略实施

采用“传统 + 深度学习”组合策略^[9]。样本量少、场景不全时，重点采用传统方法。罕见故障稀缺时，采用 DiD、DDIM 和其它扩散模型方法。控制增强强度，审核提高样本质量，剔除不合格样本，合并原始与增强样本，形成最终数据集^{[8][10]}。

三、铁路货车故障图像智能识别数据集的质量评估与典型应用

高质量数据集需通过科学评估保障实用性，同时结合工程场景落地应用，实现技术价值转化。

（一）质量评估

从完整性、准确性、一致性、实用性出发，制定可量化评估指标与实操方案。完整性评估故障与场景覆盖率，准确性核查标

注精度与图像清晰度，一致性规范标注与分类标准，实用性通过模型适配性与识别效果验证^{[6][8][10]}。评估后针对短板优化，确保数据集支撑模型训练需求。

（二）典型应用

该数据集在铁路货车智能检测领域实现典型应用。

一是 TFDS 系统升级，基于数据集训练的识别模型，嵌入原有检测平台，实现转向架裂纹、制动系统磨损等故障自动识别，检测效率提升 30% 以上，漏检率降至 1% 以下^{[2][8]}。

二是货车检修智能化，在编组站、检修站部署适配数据集的终端检测设备，辅助检修人员快速定位故障，降低人工劳动强度。

三是模型迭代优化，通过应用反馈的故障样本，反哺数据集更新，形成“数据集 - 模型 - 应用 - 优化”的闭环^{[1][5]}。

未来有望覆盖更多铁路货车班组，推动货车检测向智能化转型，彰显数据集的工程应用价值。

四、结论与展望

综上所述，构建铁路货车故障图像智能识别数据集，关系到智能故障诊断时长、检测精度与效率。在理论建构中，开展需求分析、标准制定、原始数据采集工作，充分运用数据预处理、标注、划分、增强等核心技术，有效保障数据集质量。未来实践应用中，要不断扩大覆盖范围，重视复杂与罕见场景样本的搜集，引入半监督、数据交互平台等其它更优方法，提高标注水平和质量，并依托数据质量评价体系，持续优化数据集，推动数据集构建技术向标准化、智能化、规模化方向发展，从而提高货车故障智能检测水平，为铁路货运安全运行提供坚实保障。

参考文献

- [1] 邓一辉. 大数据时代下数字图像识别技术浅析 [J]. 中国自动识别技术, 2024, (06): 55-57.
- [2] 王德明, 高善兵, 王仲辉, 等. TFDS 货车图像数据应用系统 [J]. 铁路计算机应用, 2024, 33(09): 59-67.
- [3] 陈世民. 基于关联规则算法的铁路供电系统故障数据自动识别方法 [J]. 电工技术, 2024, (16): 137-139.
- [4] 张森, 杨苹, 刘泽健, 等. 大型海上风电机组叶片故障图像识别方法 [J]. 可再生能源, 2024, 42(06): 767-773.
- [5] 章亚彬. 基于图像识别的铁路信息化建设探讨 [J]. 通讯世界, 2024, 31(05): 142-144.
- [6] 于飞, 隋正伟, 邱凤婷, 等. SAR 图像智能解译样本数据集构建进展综述 [J]. 网络安全与数据治理, 2023, 42(S1): 97-105.
- [7] 贾建坤. 铁路客车故障轨旁图像检测系统 (TVDS) 的运用研究分析 [J]. 轨道交通装备与技术, 2023, (04): 27-31.
- [8] 杨凯, 张森, 祁苗苗. 铁路车辆监测图像识别模型训练及验证平台研究 [J]. 铁路计算机应用, 2023, 32(06): 26-30.
- [9] 李云鹏, 葛宝臻, 田庆国, 等. 基于立体像对数据集的非对称离焦数据集构建 [J]. 光学学报, 2022, 42(14): 138-146.
- [10] 李红蕾. 构建图形图像数据集的方法概述 [J]. 计算机产品与流通, 2020, (08): 97.