

基于 SAM 视觉大模型的工业材料表面缺陷检测

黄亚如, 钟剑斌, 成颖怡

(广州软件学院, 广东 广州 510990)

摘要: 工业产品制造过程中时常存在各种缺陷, 不仅影响产品质量, 也给企业带来经济损失。传统人工检测低效且高成本, 迫切需要建立自动化高效准确的缺陷检测系统。基于深度学习的现有方法多数专注于特定缺陷类型, 缺乏通用性和灵活性, 且需大量标注数据训练。为解决这一挑战, 本文将视觉大模型 SAM (Segment Anything Model) 应用于工业材料表面缺陷检测任务, 通过对图像编码器和掩码解码器进行微调, 在电机磁瓦和钢轨两个缺陷数据集上进行实验评估, 结果显示微调后的 SAM 模型取得了优异表现, 最高平均交并比 (mIoU) 分别达到 0.512 和 0.655, 验证了其在面对复杂多变缺陷时的出色泛化能力, 尤其擅长检测微小缺陷, 为工业场景下提供了一种通用高效的新型缺陷检测方案。

关键词: 深度学习; 机器视觉; 缺陷检测; 视觉大模型

中图分类号: TP391.4 **文献标志码:** A

1 引言

在现代工业生产中, 工业产品在智能制造中通常会存在缺陷问题, 并且产品缺陷逐渐成为企业提高产品质量的瓶颈。在工业产品的生产制造过程中, 由于工艺、模具、设备故障和外界干扰等因素, 容易在产品表面出现斑点、凹坑、划痕、空隙和色差等缺陷。因此实现对工业产品缺陷的有效检测是非常必要的。在一般情况下, 工业产品上的缺陷由人工进行检测是很常见的, 缺陷检测性能在很大程度上依赖于手动定义的缺陷表示特征, 但是人工检测需要花费大量成本, 并且会出现效率低下, 漏检率不可控等情况。

如今大数据时代和工业 4.0 革命正在深刻影响现代制

造企业, 并随着计算机视觉和深度学习等相关技术的发展, 基于计算机视觉的工业产品的缺陷检测逐渐成为了当前的主流趋势。智能制造对传统产业的发展产生了重要影响, 将视觉检测技术应用于传统产业的升级转型, 贴合工业 4.0 的趋势。

在机器学习中, Daniel Weimer^[1] 等通过测试深度卷积神经网络 (CNN) 的设计配置及不同超参数设置对缺陷检测结果准确性的影响, 使用分层学习策略从大量训练数据中自动生成强大的特征, 进行缺陷检测。在人工智能技术方面, 特别是随机森林和多层感知器, A. Bustillo^[2] 等通过引入各向同性水平作为输入变量, 显著提高预测模型的准确性, 提出在摩擦工艺中预测表面粗

基金项目: 广州软件学院科研项目 (KY202405)

糙度和质量损失方面的方法。在深度学习中,马涛^[3]等提出了一种基于改进的 MFA-UNet 模型的铜制螺纹零件外表面缺陷检测算法,通过融合多尺度特征和注意力机制显著提升了检测效率和精度,满足了工业自动化质量检测的需求。在基于高分辨率网络(HRNet)的区域边界分割方法中,李响^[4]等通过图像预处理、直方图均衡化和自适应阈值法显著提升了边界分割的准确性和稳健性。在 ResNet-50 的网络结构基础上,丁龙飞^[5]等提出了一种结合 ResNet-50 网络结构和多重注意力机制的改进残差网络模型,用于提高电机磁瓦缺陷检测的准确率和效率,实验结果显示该模型在全类别平均准确率较原始 ResNet-50 网络提升了 2.17%,验证了其在工业应用中的有效性。

然而,工业生产中所遇到的产品缺陷类型繁多且复杂,这对现有的缺陷检测模型提出了巨大的挑战。当前的工业缺陷检测技术尽管在一定程度上取得了显著的进展,但其主要局限在解决特定类型的缺陷。面对这些挑战,工业界迫切需要一种面对新类型缺陷时无需重新训练,具有极高的灵活性和适应性的检测模型。为了解决这一难题,本文引入了近期兴起的视觉大模型 SAM (Segment Anything Model)^[6],相较于传统的缺陷检测模型,SAM 不仅能够有效识别已知的缺陷类型,更能应对未见过的新型缺陷,这对于工业生产中的多变性和复杂性提供了理想的通用缺陷检测方案。

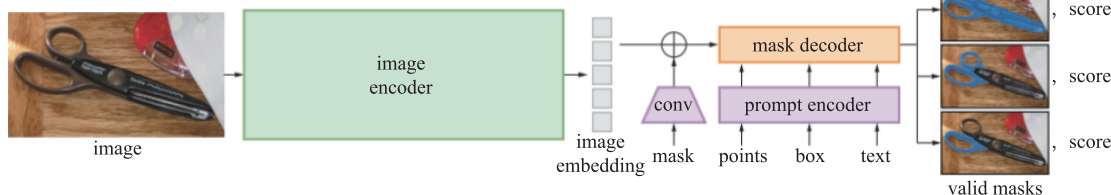


图2 SAM 结构图

对于一张尺寸为 $H \times W$ (H 、 W 分别为图像的高、宽),通道数为 3 的金属缺陷图像 I ,首先通过图像编码器进行处理,图像编码器将图片分割成若干图像块,并将每个图像块映射到高维特征空间,生成特征表示 E ,接着通过提供交互提示 T (如点击、边框或粗略掩码)来指导模型分割。这些提示通过提示编码器进行编码,生成提示特征表示 P ,图像特征表示 E 和提示特征表示 P 通过双向变换器进行融合。双向变换器通过双向注意力机制,实现图像特征和提示特征之间的信息交换和增强,最后融合特征表示 F 通过掩码解码器转换为最终的分割掩码 M 。

2 方法

2.1 SAM

SAM 模型由 Meta 公司在 2023 年发布,该模型基于 Meta AI 研究团队收集并标注的大规模图像数据集进行训练的,该数据集包含超过 1 100 万张图像和超过 11 亿个掩码标注。这样的数据规模和多样性保证了 SAM 模型在不同场景下的鲁棒性和泛化能力。该模型采用了多种分割方式,包括交互式 and 全自动分割,允许使用提示工程或自动生成策略来获得所需的分割掩码,如图 1 所示,能够灵活应对各种复杂的图像分割任务。

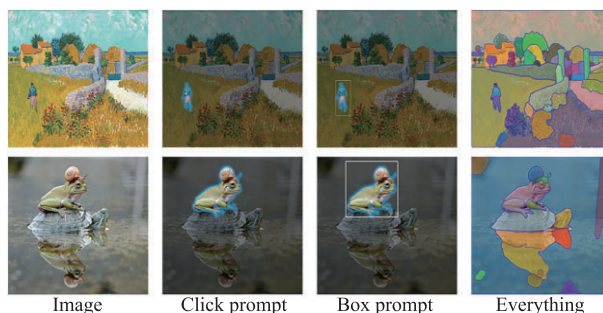


图1 SAM 公开提供的 demo^[7]

SAM 模型主要包含图像编码器 (Image Encoder)、提示编码器 (Prompt Encoder)、双向变换器 (Two-Way Transformer) 和掩码解码器 (Mask Decoder) 四个核心模块,如图 2 所示。

$$E = f_{\text{encoder}}(I) \quad (1)$$

$$P = f_{\text{prompt}}(T) \quad (2)$$

$$F = \text{Transformer}(E, P) \quad (3)$$

$$M = f_{\text{decoder}}(F) \quad (4)$$

通过以上步骤,在训练过程中经历数次迭代后,SAM 模型能够实现对金属缺陷图片的高精度分割,识别出图像中的各类缺陷。

2.2 模型训练

(1) 模型构建 本文以 SAM 作为主干,采用了 SAM-B 的基本配置。模型的整体结构如图 3 所示,包括

图像编码器、掩码解码器和冻结的提示编码器（Frozen prompt encoder）。

在模型的图像编码器部分，本文使用多个视觉 Transformer 网络，首先通过图像块嵌入（Patch Embedding）将输入图像划分为固定大小的图像块，并对每个图像块进行线性嵌入。然后，这些嵌入向量作为输入，传递给多个视觉 Transformer 模块。视觉 Transformer 模块通过多头自注意力机制和前馈神经网络来处理这些嵌入，提取图像特征。

在图像编码器生成图像特征后，这些特征被传递给掩码解码器。掩码解码器通过多层卷积神经网络（CNN）对图像特征进行处理和上采样，生成与输入图像相同大小的分割掩码，得到最终的分割结果。同时掩码解码器还接收来自提示编码器的嵌入，提示编码器将输入的提示（例如用户通过交互或者点击指定的感兴趣区域或特定特征）映射到高维特征空间，并生成嵌入。这些嵌入随后作为附加信息传递给掩码解码器，用于指导和优化掩码的生成。

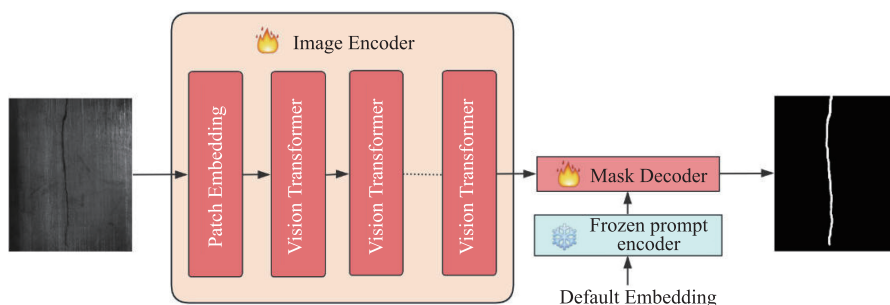


图3 模型整体结构

由于原始的 SAM 模型是为通用领域分割任务设计，并未针对工业缺陷检测进行优化，因此直接应用可能无法达到最佳效果。为了提升模型在工业缺陷检测中的性能，本文对构建的 SAM 模型进行了微调训练。模型微调是指在预训练模型的基础上，通过在特定任务或特定数据集上的进一步训练，使模型更好地适应特定任务的需求。微调的目的是利用预训练模型已经学习到的特征表示，在新的任务上进行少量参数调整，以达到更好的性能。

本文对图像编码器和掩码解码器进行了微调训练，并通过将提示编码器的所有参数设置为不参与梯度计算，冻结了提示编码器的参数。其目的是为了避免在训练过程中对其进行更新，减少训练过程中参数更新的计算量，从而更专注于图像编码器和掩码解码器的优化。通过上述步骤，本文构建并训练了适用于工业材料表面缺陷检测的 SAM-B 模型，增强其在不同工业缺陷场景下的泛化能力。

(2) 损失函数 针对本文研究的工业缺陷检测任务，在模型训练中使用 Dice 损失函数和交叉熵损失函数两种损失函数相结合。

1) Dice 损失函数是一种衡量两组数据相似度的统计方法，其定义于评估模型预测与实际标签之间的重叠程

度^[8]。计算公式为

$$Dice\ Loss = 1 - \frac{2 \sum_{i=1}^N p_i g_i + \epsilon}{\sum_{i=1}^N p_i + \sum_{i=1}^N g_i + \epsilon} \quad (5)$$

式中， p_i 是模型预测的第 i 像素值， g_i 是对应的真实标签值， N 是总的像素数。 ϵ 为平滑参数，是一个极小常数，用于避免分母为零的情况。

Dice 损失函数通过最大化预测结果和真实标签之间的重叠区域，有效地提升了模型在处理不平衡数据时的表现，特别是在缺陷检测中，缺陷区域通常较小且稀疏。在训练中使用该函数带有 sigmoid 激活和平方预测选项，以增强模型性能。

2) 交叉熵损失函数通过衡量预测概率分布与真实标签分布之间的差异来指导模型优化^[9]。对于二分类问题，交叉熵损失函数的定义为

$$Cross\ Entropy\ Loss = -\frac{1}{N} \sum_{i=1}^N [g_i \log(p_i) + (1 - g_i) \log(1 - p_i)] \quad (6)$$

式中， g_i 是第 i 个像素的真实标签（0 或 1）， p_i 是对应的预测概率值， N 是总的像素数。

交叉熵损失函数通过最大化真实类别的预测概率，

促进模型在分类任务中提高准确性。在训练中结合了 sigmoid 激活和二元交叉熵损失，为模型提供了稳定的训练目标。

在本实验中，结合 *Dice Loss* 和 *Cross Entropy Loss* 来训练模型。这种结合利用 *Dice Loss* 在处理不平衡数据方面的优势，以及 *Cross Entropy Loss* 在分类任务中稳定收敛的特点，以获得最佳的模型性能。具体的综合损失函数定义为两者的加权和

$$TotalLoss = \alpha \cdot DiceLoss + \beta \cdot CrossEntropyLoss \quad (7)$$

式中， α 和 β 是用于平衡两种损失的权重参数，*Cross Entropy* 的损失权重设置为 0.5，*Dice* 损失的损失权重设置为 0.5。

3 实验验证

3.1 数据描述

本实验采用两个数据集进行实验验证，第一个数据集来源于中国科学院自动化研究所的电机磁瓦表面缺陷数据集 (MT Defect Dataset)，共 1 344 张图像^[10]。根据缺陷类型的不同，所有图像分为了 6 个子数据集，分别是气孔、断裂、裂纹、磨损、凹凸 (研磨不均匀) 和正常 (无缺陷)。且每个子数据集都包含了原始采样图像与其对应的像素级标签。

第二个数据集来源北京交通大学李清勇教授团队制作的 RSDDs 数据集，RSDDs 数据集有以下两种类型的数据：第一种是从快车道采集的数据集，包含 67 张图像。第二种是从普通 / 重型运输轨道捕获的数据集，包含 128 张图像^[11]。

本文数据集均采用将训练集和验证集划分为 7 : 3 进行实验，具体数据集划分见表 1。

表 1 实验中使用的两个数据集的详细信息

数据集	产品类别	训练集	测试集
MT Defect	电机磁瓦	940	404
RSDDs	钢轨	136	59

3.2 实验设置与评估指标

本实验在 Ubuntu 22.04 操作系统环境下进行，开发基于 Python 3.8.0 和 Pytorch 2.0.1 框架。计算环境采用 NVIDIA Tesla V100-SXM-32GB GPU，并使用 CUDA 12.1 进行加速。在模型训练中，本文选择了 Adam 优化器，

将其学习率设定为 0.000 1。为了验证 SAM 在不同工业缺陷数据集上的表现，分别在 MT Defect 和 RSDDs 两个数据集上进行训练和测试，训练过程中的批量大小设定为 8。具体而言，每个数据集均训练 100 个 epoch，在训练过程中，所有图像均进行归一化处理，输入尺寸为 $1\,024 \times 1\,024$ 。为了保证实验的可重复性，设置了固定的随机种子。

为了全面评估 SAM 在工业材料表面缺陷检测中的性能，本文使用三种评估指标对 SAM 进行评估，包括平均交并比 (mIoU)、平均准确率 (mAcc)、平均精确率 (mPrecision)。

1) mIoU 是语义分割领域常用的评估指标之一，用于衡量预测结果与真实标注之间的重叠程度。对于每个类别 i ，其交并比 (IoU) 定义为预测区域和真实区域的交集与并集之比，mIoU 是所有类别 IoU 的平均值公式为

$$mIoU = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i + FN_i} \quad (8)$$

式中， TP_i 、 FP_i 和 FN_i 分别表示类别 i 的真阳性、假阳性和假阴性数量，

2) mAcc 用于评估分类结果的总体准确性，即所有像素中被正确分类的像素比例。对于每个类别 i ，准确率 (Acc) 定义为正确分类的像素数与该类别的真实像素数之比。mAcc 公式为

$$mAcc = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (9)$$

3) mPrecision 用于评估检测结果的精确度，即所有预测为某类别的像素中，真正属于该类别的像素比例。对于每个类别 i ，精确率 (Precision) 定义为正确分类的像素数与预测为该类别的像素总数之比。mPrecision 是所有类别精确率的平均值公式为

$$mPrecisionPr = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i} \quad (10)$$

3.3 实验结果分析

为了验证 SAM 视觉大模型在工业材料缺陷检测上的性能和泛化性，本文进行了对比实验，评估了零样本 (Zero-shot) 测试模型和微调后模型在 MT Defect 和 RSDDs 数据集上的表现。

Zero-shot 测试意味着在未经过任何训练或微调的情况下直接使用预训练模型进行推理，以评估模型在未见数据上的性能。Zero-shot 泛化在工业场景中尤为重要，

因为实际应用中常常会遇到大量未见过的缺陷类型，模型需要具备良好的泛化能力以应对这些新情况。为了提升模型在工业缺陷检测任务中的表现，本文对 SAM 模型进行了微调，微调后的模型在两个数据集上的验证 loss 曲线如图 4、5 所示，显示了随着训练的进行，损失逐渐减小，验证了微调过程的有效性。

表 2 显示了在 MT Defect 数据集上，Zero-shot 测试的 mIoU 为 0.431，mAcc 为 0.627，mPrecision 为 0.552；在 RSDDs 数据集上，Zero-shot 测试的 mIoU 为 0.583，mAcc 为 0.702，mPrecision 为 0.630。尽管 SAM 模型在未见过的数据上表现良好，但仍有提升空间。微调后，MT Defect 数据集上的 mIoU 提升至 0.512，mAcc 和 mPrecision 都有所提升；RSDDs 数据集上的 mIoU 提升至 0.655，其余指标同样显著提升。

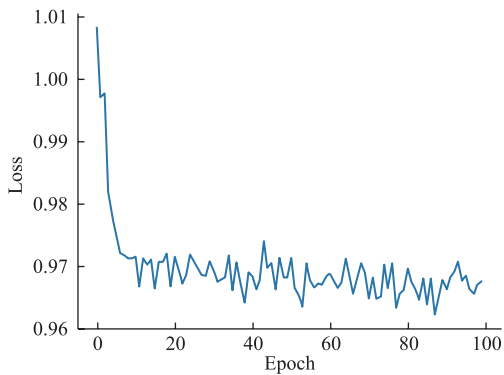


图 4 模型在 MT Defect 验证集上的 Loss 曲线图

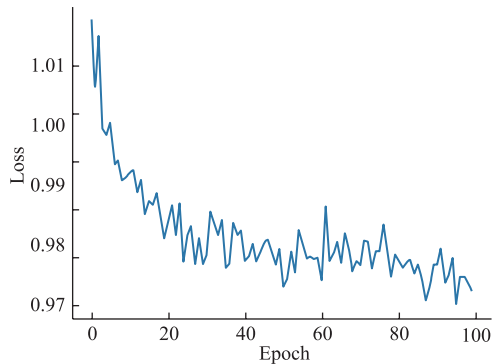


图 5 模型在 RSDDs 验证集上的 Loss 曲线图

表 2 在 MT Defect 和 RSDDs 两个数据集上不同方法的比较

Dataset	Model	mIoU	mAcc	mPrecision
MT Defect	SAM (Zero-shot)	0.431	0.627	0.552
	SAM (Finetune)	0.512	0.711	0.641
RSDDs	SAM (Zero-shot)	0.583	0.702	0.630
	SAM (Finetune)	0.655	0.798	0.726

这些结果表明，微调显著提升了模型在工业材料缺陷检测任务中的表现，提高了检测缺陷区域的准确性和精度。在实际应用中，通过微调使模型能够更好地适应具体的工业场景，从而提高缺陷检测的准确性和效率。

微调前后的 SAM 网络在两个数据集上的分割可视化结果如图 6 所示，从图中可以看出 SAM 即使在未经特定任务微调的情况下，依然展示出良好的零样本泛化能力。

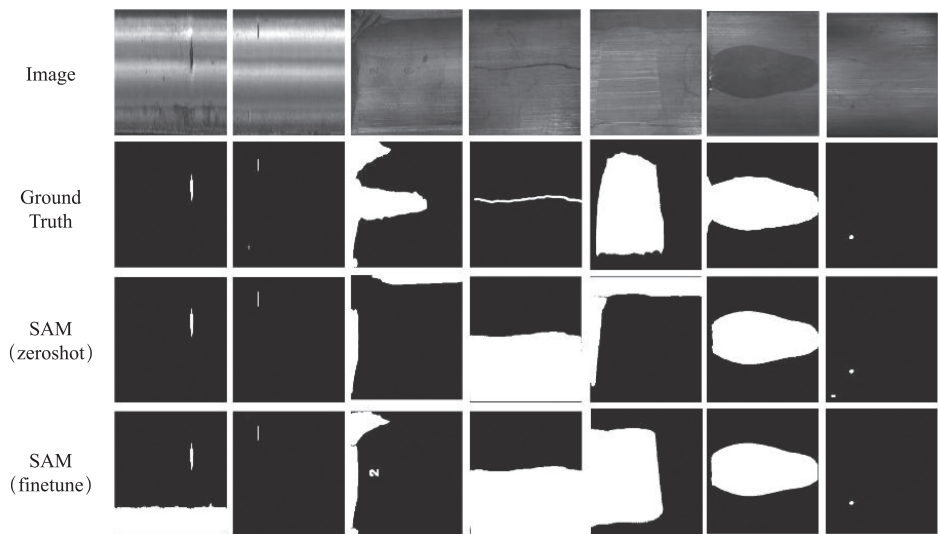


图 6 不同模型的分割结果对比

这使得其在面对包括小缺陷实例和其他干扰杂质的不同实例缺陷图像时，能够有效地分割出各个实例。通过特定缺陷数据集上进行微调后，SAM 能够更好地适应特定的缺陷检测任务，但同时在检测大面积缺陷时，由于光照、对比度等环境因素，模型并不能很好地识别出凹凸和裂纹等缺陷类别，在未来的工作中将优化这一点。

在真实的工业场景中，微小缺陷图像往往比其他出现的频率更高，同时，在面对单个缺陷图像时，SAM 可以完全分割出小缺陷类别。这表明 SAM 在细粒度实例分区方面的优势，使其能够准确检测和分割细微缺陷，说明在工业缺陷检测领域中具有巨大潜力。

4 总结与展望

本文利用电机磁瓦表面缺陷 (MT Defect) 数据集和钢轨 (RSDDs) 数据集，应用深度学习技术进行了缺陷检测实验，以评估 SAM 视觉大模型在工业产品缺陷检测中的性能。在这两个数据集上，微调后的 SAM 模型展现了优异的效果，最高的 mIoU 达到了 0.655。实验结果表明，SAM 视觉大模型在工业场景的通用缺陷检测中具有显著的潜力和应用前景。

然而，本文实验的工业场景较为单一，存在一定的局限性。未来的研究将致力于扩展和多样化工业缺陷数据集，涵盖更多种类和不同条件下的缺陷图像，以进一步提升 SAM 模型的泛化能力和检测精度。同时，开发自动化微调工具和流程，使 SAM 模型能够在不同工业应用场景中快速进行自适应优化，从而提高其在各类缺陷检测任务中的应用效率。IM

参考文献

- [1] WWEIMER D, SCHOLZ-REITER B, SHPITALNI M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection [J]. CIRP Annals - Manufacturing Technology, 2016, 65 (1) : 417-420.
- [2] BUSTILLO A, PIMENOV D Y, MATUSZEWSKI M, et al. Using artificial intelligence models for

the prediction of surface wear based on surface isotropy levels[J]. Robotics and Computer-Integrated Manufacturing, 2018 (53) : 215-227.

- [3] 马涛, 李敬兆. 基于 MFA-UNet 的铜制螺纹零件外表面缺陷检测 [J]. 现代制造工程, 2024 (5) : 113-120, 94.
- [4] 李喃, 刘华, 郝晋一, 等. 基于高分辨率网络的全钢子午线轮胎钢丝圈区域边界分割 [J/OL]. 济南大学学报 (自然科学版), 1-7[2024-10-28]. <https://doi.org/10.13349/j.cnki.jdxbn.20240522.001>.
- [5] 丁龙飞, 曾水玲. 基于多重注意力机制的电机磁瓦表面缺陷检测 [J]. 计算机技术与发展, 2022, 32 (12) : 194-199.
- [6] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023; 4015-4026.
- [7] SONG K, CUI W, YU H, et al. SAM Era: Can it segment any industrial surface defects?[J]. 计算机、材料和连续体 (英文), 2024, 78 (3) : 3953-3969.
- [8] LI X Y, SUN X F, MENG Y X, et al. Dice loss for data-imbalanced NLP tasks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.
- [9] MANNOR S, PELEG D, RUBINSTEIN R. The cross entropy method for classification[C]//Proceedings of the 22nd international conference on Machine learning. 2005; 561-568.
- [10] HUANG Y B, QIU C Y, GUO Y, et al. Surface defect saliency of magnetic tile[C]//2018 IEEE 14th International Conference on Automation Science and Engineering (CASE) . Munich: IEEE, 2018; 612-617.
- [11] GAN J, LI Q, WANG J, et al. A hierarchical extractor-based visual rail surface inspection system[J]. IEEE Sensors Journal, 2017, 17 (23) : 7935-7944.

收稿日期: 2024-06-17