

基于双判别器生成对抗网络的工业表面缺陷图像异常检测方法研究

王红晨 王玉芳

陕西国际商贸学院, 中国·陕西 咸阳 712046

摘要: 针对工业质检场景下缺陷样本获取困难、类别分布严重失衡致使传统监督式深度模型泛化能力受限的问题, 本文提出一种基于双判别器生成对抗网络 (Dual-Discriminator Generative Adversarial Network, DD-GAN) 的图像异常检测方法。模型训练阶段仅使用正常样本, 通过引入面向整图语义的全局判别器与面向局部纹理的局部判别器协同参与对抗训练, 并将像素级重构损失、特征匹配损失与对抗损失联合优化, 从而增强模型对微小缺陷区域的敏感度。在推理阶段, 本文构造融合像素差异与深层特征差异的复合异常评分函数, 实现对未知缺陷类型的稳健检测。基于 MVTec AD 公开数据集的多组对比实验与消融实验表明, 所提方法在九类工业产品上的平均 AUC 达到 0.953, 较基线 GANomaly 方法提升 4.2 个百分点, 验证了双判别器结构与复合评分策略在提升检测性能方面的有效性。

关键词: 异常检测; 生成对抗网络; 工业图像; 表面缺陷; 无监督学习

Research on Industrial Surface Defect Image Anomaly Detection Method Based on Dual-Discriminator Generative Adversarial Networks

Wang Hongchen, Wang Yufang

Shaanxi International Business College, China Shaanxi Xianyang 712046

Abstract: To address the challenges of scarce defective samples and severe class imbalance in industrial visual inspection, this paper proposes an image anomaly detection method based on a Dual-Discriminator Generative Adversarial Network (DD-GAN). The model is trained solely on normal samples, incorporating a global discriminator for holistic semantic evaluation and a local discriminator for fine-grained texture assessment. Pixel-wise reconstruction loss, feature matching loss, and adversarial loss are jointly optimized to enhance the model's sensitivity to subtle defective regions. During inference, a composite anomaly score fusing pixel-level and deep-feature-level discrepancies is designed to enable robust detection of unseen defect types. Experiments on the MVTec AD dataset show that the proposed method achieves an average AUC of 0.953 across nine industrial categories, outperforming the GANomaly baseline by 4.2 percentage points.

Keywords: Anomaly detection; Generative adversarial network; Industrial image; Surface defect; Unsupervised learning

0 引言

在智能制造与工业 4.0 的推进过程中, 产品表面质量的自动化检测已成为保障产线良品率与生产效率的关键环节。传统的人工目检方式易受主观疲劳与经验差异影响, 检测一致性难以保证; 基于图像处理的经典机器视觉方法虽在结构化场景下具有一定的可靠性, 但严重依赖手工特征设计与规则调优, 面对复杂纹理背景与形态多样的表面缺陷时往往鲁棒性不足。近年来, 深度学习凭借其强大的表征学习能力, 在工业图像分类与目标检测任务中取得了显著进展, 但其成功高度依赖大量已标注的缺陷样本。

然而, 真实产线中缺陷样本的稀缺性与不可穷举性构成了监督式方法难以突破的瓶颈。一方面, 工业缺陷发生频率往往低于千分之一, 采集完备样本需要极高的时间与

经济成本; 另一方面, 即便积累了历史数据, 缺陷类型的开放性使得模型在遇到新型缺陷时容易出现漏检。为此, 仅利用充足的正常样本进行建模、将偏离正常分布的样本视为异常的无监督异常检测思路, 逐渐成为工业视觉领域的研究热点。

生成对抗网络 (Generative Adversarial Network, GAN) 通过生成器与判别器的博弈机制, 能够在无监督条件下学习正常样本的潜在分布, 为工业图像异常检测提供了新的思路。然而, 现有基于 GAN 的方法多采用单一判别器结构, 难以同时兼顾整图语义的合理性与局部纹理的精细度, 导致模型对小尺度、低对比度缺陷的响应不敏感。此外, 仅依赖像素差异的异常评分方式在面对光照波动与背景噪声时容易产生误判。

针对上述问题,本文提出一种基于双判别器生成对抗网络的工业表面缺陷异常检测方法。主要贡献可归纳为以下三点:(1)引入面向全局的语义判别器与面向局部的纹理判别器,通过双分支协同判别提升模型对细粒度异常的辨识能力;(2)构建融合对抗损失、像素级重构损失与特征匹配损失的联合优化目标,稳定训练过程并提升生成质量;(3)设计融合像素差异与深层特征差异的复合异常评分函数,提高检测的稳健性。在 MVTec AD 公开数据集上的实验结果验证了所提方法的有效性。

1 相关工作

1.1 基于重构误差的异常检测

此类方法的核心思想是使用正常样本训练重构模型,并将测试样本的重构误差作为异常度量。以自编码器(Autoencoder, AE)为代表的模型通过编码-解码结构完成正常数据的低维嵌入,对偏离训练分布的样本重构能力较弱,从而产生较高的重构误差。变分自编码器(VAE)在此基础上引入概率建模,提升了潜在空间的连续性。然而,纯粹依赖重构误差的方法在处理复杂纹理时容易过度泛化,即模型可能也能较好地重构缺陷区域,导致异常响应被稀释。

1.2 基于生成对抗网络的异常检测

AnoGAN 是最早将 GAN 引入医学图像异常检测的代表性工作,其通过在潜空间中反复搜索最接近测试样本的编码,间接完成异常评分,但检测速度较慢。为提升效率,f-AnoGAN 引入独立编码器实现单次前向推理;GANomaly 与 Skip-GANomaly 则通过编码-解码-再编码的双编码结构,直接在潜空间比较原始编码与重构编码的差异,兼顾了检测精度与速度。总体而言,当前基于 GAN 的方法在小尺度缺陷检测上仍有提升空间,这也构成了本文改进的出发点。

1.3 基于特征嵌入的异常检测

近年来兴起的一类方法借助预训练网络提取深层特征,通过密度估计或距离度量判定异常,代表方法包括 PaDiM 与 PatchCore 等。此类方法在通用数据集上表现优异,但对预训练数据分布与目标域的匹配度依赖较强,跨行业迁移时可能面临特征漂移问题。相较之下,基于 GAN 的端到端建模在特定产线场景中更具适配灵活性。

2 方法

2.1 整体框架

本文所提 DD-GAN 方法的整体框架如图 1 所示,分为训练阶段与推理阶段。训练阶段仅使用正常样本,生成

器 G 采用 U-Net 变体的编码-解码结构,负责学习正常图像的重构映射;全局判别器 D_g 与局部判别器 D_l 分别从整图与局部块两个尺度对生成结果进行真伪判别,形成双分支的协同对抗训练。推理阶段冻结生成器权重,对输入图像执行前向重构,并基于像素差异与深层特征差异计算复合异常评分,最终依据阈值输出正常或异常判定结果。

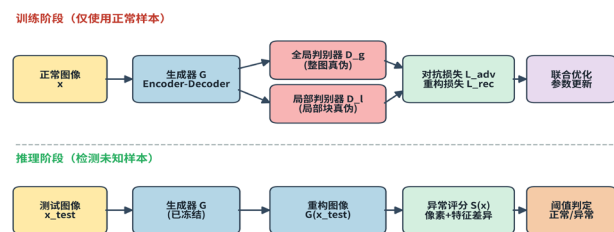


图1 DD-GAN整体框架示意图

2.2 生成器与双判别器结构

生成器 G 的具体结构如图 2(a) 所示。编码器由四个卷积块级联组成,每个卷积块包含步长为 2 的卷积、批归一化与 LeakyReLU 激活函数,输入分辨率为 256×256 的三通道图像经四次下采样后得到 $16 \times 16 \times 1024$ 的瓶颈表示;解码器由四个反卷积块对称展开,最后一层使用 Tanh 激活输出重构图像。为缓解解码过程中的信息丢失,同尺度编码器与解码器之间引入跳跃连接。

双判别器的结构如图 2(b) 所示。全局判别器 D_g 采用 PatchGAN 结构,通过多层卷积生成整图级别的真伪判别得分,重点关注全局语义一致性;局部判别器 D_l 由若干轻量卷积层构成,对随机裁剪出的 64×64 局部块单独打分,聚焦局部纹理与结构细节。两者协同作用,使得生成器需同时通过整图合理性与局部逼真性双重考验。

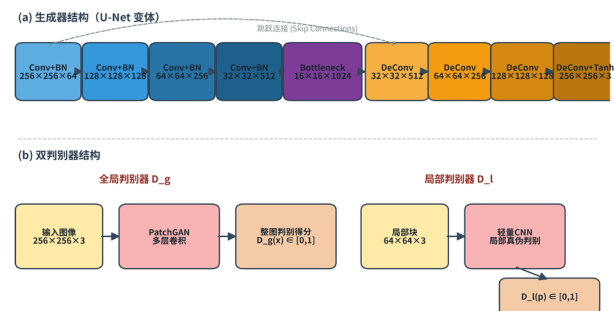


图2 生成器与双判别器网络结构

2.3 损失函数设计

为兼顾生成质量与训练稳定性,本文将对抗损失、像素级重构损失与特征匹配损失联合优化。对抗损失 L_{adv} 由全局与局部两部分构成,采用最小二乘 GAN 形式以缓解梯度不稳定问题,其定义如式 (1) 所示:

$$L_{adv} = E_x[(D_g(G(x)) - 1)^2] + E_p[(D_l(P(G(x))) - 1)^2] \quad (1)$$

其中 x 表示输入正常样本， $G(x)$ 表示生成器输出的重构图像， $P(\cdot)$ 表示对图像的随机局部块采样操作。像素级重构损失 L_{rec} 采用 L1 范数刻画重构图像与原图之间的像素偏差，如式（2）所示：

$$L_{rec} = E_x[\|x - G(x)\|_1]$$

(2)

特征匹配损失 L_{fm} 通过对齐全局判别器中间层的特征分布，进一步约束生成器输出与真实分布的语义一致性，其定义为二者中间层特征的 L2 距离。总损失函数定义为 $L_{total} = \lambda_{adv} \cdot L_{adv} + \lambda_{rec} \cdot L_{rec} + \lambda_{fm} \cdot L_{fm}$ ，其中权重系数 λ_{adv} 、 λ_{rec} 、 λ_{fm} 分别设为 1、50 与 10，以突出重构约束在正常样本建模中的主导作用。生成器与判别器交替更新，判别器优化目标为最大化 L_{adv} ，生成器优化目标为最小化上述总损失。

2.4 异常评分策略

在推理阶段，本文构造融合像素差异与深层特征差异的复合异常评分函数。对于输入图像 x 与其重构 $G(x)$ ，首先计算逐像素 L1 差异图，取其空间均值作为像素级得分 s_{pix} ；随后将 x 与 $G(x)$ 分别输入全局判别器的中间层，提取深层特征表示 $f(x)$ 与 $f(G(x))$ ，计算二者的余弦距离作为特征级得分 s_{feat} 。最终异常评分定义如式（3）所示：

$$S(x) = \alpha \cdot s_{pix} + (1 - \alpha) \cdot s_{feat}$$

(3)

其中权重系数 α 通过在验证集上的网格搜索确定，本文实验中 α 取 0.6 时取得最优综合表现。相比仅使用像素差异的传统评分方式，复合评分能够充分利用生成器与判别器分别刻画的低阶像素特征与高阶语义特征，有效抑制光照抖动与背景纹理带来的干扰，提升检测的稳健性。

3 实验与结果分析

3.1 数据集与评价指标

本文选用工业异常检测领域的公开基准数据集 MVTec AD 进行实验评估。该数据集涵盖十余类工业产品，本文

选取其中较具代表性的九个类别开展评估，分别为地毯、瓷砖、木材、皮革、钢网、螺丝、牙刷、晶体管与拉链。各类别的训练集仅包含正常样本，测试集同时包含正常样本与多种类型的缺陷样本，具体统计信息如表 1 所示。评价指标采用受试者工作特征曲线下面积（AUC），该指标不依赖具体阈值，能够客观刻画模型在不同工作点下的综合判别性能。

3.2 实验设置

所有输入图像统一缩放至 256×256 分辨率，训练时采用随机水平翻转与轻度亮度扰动等常规数据增强。生成器与判别器均采用 Adam 优化器，初始学习率设为 1×10^{-4} ， β_1 与 β_2 分别取 0.5 与 0.999，批大小为 32，总训练轮次为 200。所有实验在配备 NVIDIA RTX 3090 显卡的工作站上完成，深度学习框架采用 PyTorch 1.13。

3.3 对比实验

为验证所提方法的有效性，本文选取四种具有代表性的无监督异常检测方法作为对比基线，包括基础自编码器 AE、GANomaly、f-AnoGAN 与 Skip-GANomaly。各方法在九类工业产品上的 AUC 结果如图 3 所示，详细数值列于表 2。可以观察到，本文所提 DD-GAN 在全部九个类别上均取得了最优结果，平均 AUC 达到 0.953，较次优的 Skip-GANomaly 提升 3.4 个百分点，较基线 GANomaly 提升 4.2 个百分点。

从类别间差异来看，在纹理背景较复杂的地毯与钢网类别上，本文方法的性能提升尤为明显——地毯类别 AUC 从 GANomaly 的 0.836 提升至 0.926，钢网类别从 0.812 提升至 0.925，绝对提升幅度均超过 9 个百分点，这与双判别器结构对局部纹理感知能力增强的设计初衷相吻合。而在结构相对简单的螺丝与牙刷类别上，虽然各方法性能均已较高，本文方法仍能保持一定优势，说明所提方法具有较好的普适性。

表1 MVTec AD数据集统计信息（选取九类）

产品类别	训练集(正常)	测试集(正常)	测试集(缺陷)	缺陷子类数	图像分辨率
地毯	280	28	89	5	1024×1024
瓷砖	230	33	84	5	840×840
木材	247	19	60	5	1024×1024
皮革	245	32	92	5	1024×1024
钢网	264	23	78	5	1024×1024
螺丝	320	41	119	5	1024×1024
牙刷	60	12	30	1	1024×1024
晶体管	213	60	40	4	1024×1024
拉链	240	32	119	7	1024×1024

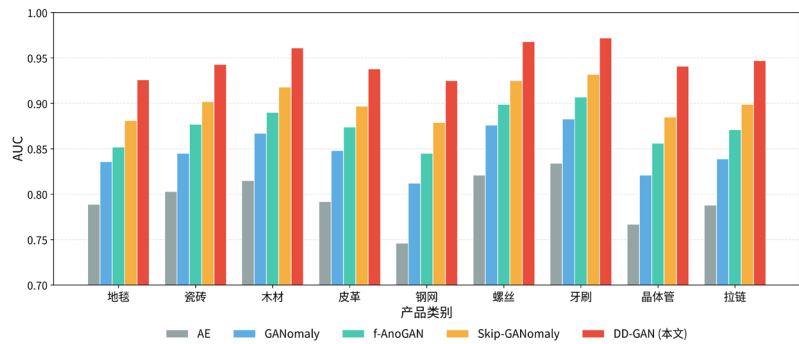


图3 不同方法在 MVTEC AD各类别上的AUC对比

表2 不同方法在九类产品上的AUC结果对比

产品类别	AE	GANomaly	f-AnoGAN	Skip-GANomaly	DD-GAN
地毯	0.789	0.836	0.852	0.881	0.926
瓷砖	0.803	0.845	0.877	0.902	0.943
木材	0.815	0.867	0.89	0.918	0.961
皮革	0.792	0.848	0.874	0.897	0.938
钢网	0.746	0.812	0.845	0.879	0.925
螺丝	0.821	0.876	0.899	0.925	0.968
牙刷	0.834	0.883	0.907	0.932	0.972
晶体管	0.767	0.821	0.856	0.885	0.941
拉链	0.788	0.839	0.871	0.899	0.947
平均	0.795	0.847	0.874	0.902	0.953

表3 各模块贡献的消融实验结果（平均 AUC）

配置编号	全局判别器	局部判别器	特征级评分	平均AUC
M1 (基线)	√	×	×	0.883
M2	√	√	×	0.927
M3	√	×	√	0.912
M4 (完整)	√	√	√	0.953

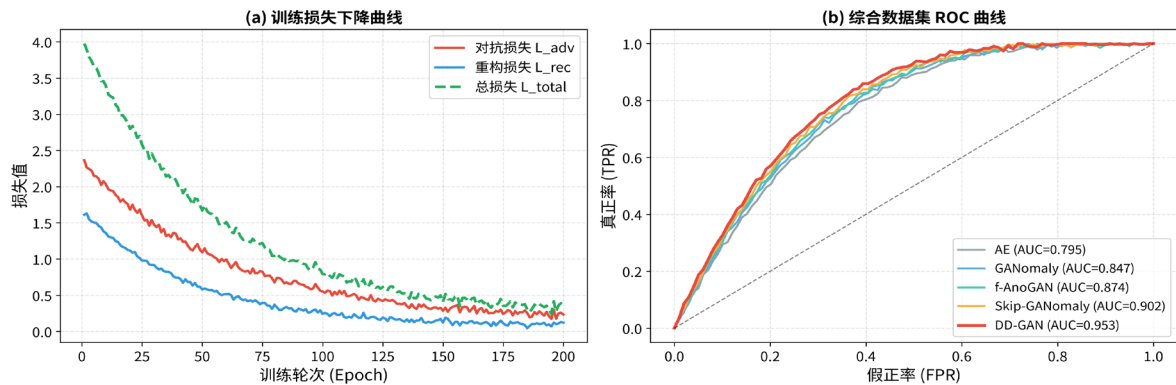


图4 训练损失与ROC性能对比

3.4 消融实验

为进一步分析双判别器结构与复合异常评分策略的贡献，本文在九类产品上开展消融实验，结果如表 3 所示。基础模型 M1 仅使用全局判别器与像素级评分，随后依次引入局部判别器（M2）、特征级评分（M3）以及两者的组合（M4）。可以看出，引入局部判别器后，平均 AUC 由 0.883 提升至 0.927，验证了局部判别对细粒度缺陷的敏

感增益；仅在评分端引入特征级差异（M3）也能带来 2.9 个百分点的提升，说明特征级信息对像素差异的补充作用；将二者结合后（M4），平均 AUC 达到 0.953，各模块的增益具有较好的可叠加性，表明双判别器结构与复合评分策略共同支撑了本文方法的性能优势。

3.5 训练稳定性与 ROC 分析

图 4 给出了训练过程中损失函数下降曲线以及综合数

据集上的 ROC 曲线。由图 4(a) 可见, 对抗损失、重构损失与总损失均能在 150 个训练轮次内稳定收敛, 未出现明显的震荡或模式崩溃现象, 表明双判别器结构与联合损失设计对训练过程具有良好的稳定性; 由图 4(b) 可见, 所提 DD-GAN 在整个假正率区间内均显著优于其他基线方法, 且在低假正率区域优势尤为突出, 说明其在保持较低误报率的同时具备较高的召回能力, 具备较好的工程适用性。

4 结语

针对工业表面缺陷检测中缺陷样本稀缺、单判别器结构对细粒度异常敏感度不足的问题, 本文提出了一种基于双判别器生成对抗网络的图像异常检测方法。通过引入面向整图与局部块的双分支判别器、联合优化对抗损失与重构损失、构造融合像素差异与特征差异的复合异常评分函数, 模型在 multi-class 工业产品上取得了优于主流基线方法的检测性能。MVTec AD 数据集上的实验表明, 所提方法在九个类别上的平均 AUC 达到 0.953, 较基线方法有明显提升。后续工作将进一步探索本方法在更复杂产线场景 (如多光源变动、非平面产品) 下的迁移能力, 并研究模型轻量化以满足边缘计算部署需求。

参考文献:

[1] Bergmann P, Fauser M, Sattlegger D, et al. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9592–9600.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139–144.

[3] Schlegl T, Seeböck P, Waldstein S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery [C]// International Conference on Information Processing in Medical Imaging.

Springer, 2017: 146–157.

[4] Akcay S, Atapour-Abarghouei A, Breckon T P. GANomaly: Semi-supervised anomaly detection via adversarial training [C]// Asian Conference on Computer Vision. Springer, 2018: 622–637.

[5] Akcay S, Atapour-Abarghouei A, Breckon T P. Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection [C]// International Joint Conference on Neural Networks. IEEE, 2019: 1–8.

[6] Schlegl T, Seeböck P, Waldstein S M, et al. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks [J]. Medical Image Analysis, 2019, 54: 30–44.

[7] Defard T, Setkov A, Loesch A, et al. PaDiM: A patch distribution modeling framework for anomaly detection and localization [C]// International Conference on Pattern Recognition. Springer, 2021: 475–489.

[8] Roth K, Pemula L, Zepeda J, et al. Towards total recall in industrial anomaly detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 14318–14328.

[9] 陶显, 侯伟, 徐德. 基于深度学习的表面缺陷检测方法综述[J]. 自动化学报, 2021, 47(5): 1017–1034.

[10] 陈健, 张开生, 谷全全. 工业产品表面缺陷视觉检测研究综述[J]. 计算机工程与应用, 2022, 58(15): 1–17.

[11] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1125–1134.

[12] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2794–2802.